Algorithmes d'optimisation et d'analyse des problèmes multidimensionnels, non linéaires, en Biologie et Biophysique

Benjamin Parent

n° d'ordre : 59

École Centrale de Lille

THÈSE

présentée en vue d'obtenir le grade de

DOCTEUR

en Automatique, Informatique Industrielle

par

Benjamin Parent

Ingénieur de l'Institut Supérieur d'Électronique et du Numérique de Lille

Algorithmes d'optimisation et d'analyse des problèmes multidimensionnels non linéaires en Biologie et Biophysique

Doctorat délivré par l'École Centrale de Lille

Soutenue le 29 Octobre 2007 devant le jury constitué de :

Mme A. Imberty	Directeur de Recherches *Centre de Recherches sur les Macromolécules Végétales, Grenoble*	*Rapporteur*
M. A. Richard	Professeur *Centre de Recherche en Automatique de Nancy*	*Rapporteur*
M. T. Bastogne	Maître de conférences *Centre de Recherche en Automatique de Nancy*	*Rapporteur*
M. A. Varnek	Professeur *Laboratoire d'Infochimie de Strasbourg*	*Membre*
M. D. Horvath	Chargé de Recherches *Unité de Glycobiologie Structurale et Fonctionnelle, Lille*	*Membre*
M. M. Davy	Chargé de Recherches *Lab. d'Automatique, Génie Info. et Signal, Lille*	*Membre*
M. B. Vandenbunder	Directeur de Recherches *Institut de Recherches Interdisciplinaires, Lille*	*Directeur*
M. J.-P. Richard	Professeur *Lab. d'Automatique, Génie Info. et Signal, Lille*	*Directeur*
M. G. Lippens	Directeur de Recherches *Unité de Glycobiologie Structurale et Fonctionnelle, Lille*	*Invité*

« *En todo amar y servir* »
Saint Ignace de Loyola

Remerciements

À la fin de cette thèse et de sa rédaction, force m'est d'avouer que j'ai expérimenté le sentiment de gratitude ; ce paragraphe se veut être un résumé des profus et profonds remerciements que je souhaite exprimer.

Je suis, tout d'abord, extrêmement honoré de compter, parmi les membres du jury, les personnes suivantes : Madame A. Imberty et Messieurs A. Richard et T. Bastogne, qui ont accepté d'être rapporteurs pour cette thèse, ainsi que Messieurs A. Varnek et M. Davy.

Cette thèse interdisciplinaire n'a pu voir le jour et se concrétiser que grâce à l'implication de personnes à l'esprit particulièrement ouvert, je tiens donc à remercier toutes les personnes qui ont participé au pilotage de ce travail : B. Vandenbunder, J.-P. Richard, mes directeurs, D. Horvath, G. Lippens et A. Kökösy, qui m'ont encadré au jour le jour. Votre sens humain et votre honnêteté remarquable ont modelé ma personne autant que mon travail. Pour avoir été habiles à diriger ces recherches et parce que votre passion est contagieuse, je vous exprime chaleureusement toute ma gratitude.

De même, je tiens à remercier ma famille d'accueil scientifique qu'est l'équipe de G. Lippens : Fanny, Isabelle, Laziza, Nathalie, Alain, Arnaud, Dries, Gérard, Jean-Michel et Xavier. Cette formidable aventure, avant tout scientifique, s'est écrite dans un *jour-après-jour* avec des mots humains.

À l'équipe de modélisation des rythmes circadiens, je voudrais également exprimer toute ma reconnaissance pour toutes les discussions et les groupes de travail passionnants. Je remercie également l'équipe de F.-Y. Bouget qui nous a accueillis chaleureusement et nous a initiés patiemment au B.A.-ba des rythmes circadiens.

Mes remerciements vont également à l'ensemble des personnes du LAGIS qui m'ont apporté les réponses ou les pistes à poursuivre quand j'en avais besoin.

Je remercie aussi l'équipe OPAC du LIFL, et en particulier Emilia et Alexandru pour leur disponibilité et leur simplicité.

Je voudrais enfin exprimer ma gratitude au corps enseignant de l'école ISEN Lille qui a forgé en moi ce goût de la recherche et m'a ensuite offert l'opportunité de charges d'enseignements pendant ces trois années de thèse.

Je terminerai par les premières personnes qu'il faut remercier : la famille. Vous avez mis en moi cette graine de curiosité arrosée de passions… Elle est maintenant devenue grande et insatiable. C'est elle, avec votre soutien, qui me fait avancer, même après les nuits blanches passées devant les problèmes épineux. Vous avez vécu

ma thèse seize heures par jour sans vous lasser de croire que j'aboutirai et les mots, décidément, ne seront jamais suffisants pour vous dire toute ma reconnaissance. Je pense en premier lieu à mes parents et à ma femme : Anne ; quant à toi, Joseph, même du ventre de ta maman et depuis que tu en es sorti, tu m'as bien aidé aussi à ta façon. Tu représentes tout ce en quoi je crois, et puisque j'ai la naïveté de penser que travailler, c'est croire en demain, je te dédie ce travail.

Table des matières

Introduction

Ceci est une thèse sur la complexité du vivant !

Cette complexité apparaît déjà à l'échelle moléculaire, pour laquelle la détermination de la forme tridimensionnelle des molécules est encore un challenge majeur pour la biochimie. C'est pourquoi une grande partie de la thèse est dédiée à l'étude de méthodes computationnelles permettant d'accélérer et/ou de compléter les approches expérimentales destinées à mieux comprendre la fonction des molécules et leurs interactions.

Car la fonction d'une molécule repose sur ses interactions. Par ailleurs, tout le fonctionnement de la *cellule*, brique de base des organismes vivants, repose sur ces interactions. La complexité, déjà présente au niveau de la molécule unique, explose alors lorsqu'il s'agit d'intégrer plusieurs molécules (quels sites de fixation ? quels modes d'interactions ? quelles affinités ? combien d'acteurs ? lesquels ?). En ce sens, le sujet principal de cette thèse aurait pu être : comment aborder *in silico* l'étude de l'*interactome*, c'est-à-dire de l'ensemble des interactions qui se jouent sur la scène cellulaire.

Entre les interactions moléculaires et l'organisation générale de la cellule, il existe (au moins) un niveau fonctionnel de hiérarchisation intermédiaire : celui des modules fonctionnels. Depuis l'ère de la génomique où de nombreux génomes ont été entièrement séquencés, on sait en effet que le graphe des interactions moléculaires n'est pas purement aléatoire avec des interactions tous azimuts, mais qu'au contraire, les molécules travaillent par familles à l'accomplissement de tâches spécifiques qui leurs sont dédiées ; c'est ce que nous avons appelé les *modules fonctionnels*. Inutile de souligner encore une fois le niveau de complexité qui caractérise ces modules et leurs interfaçages...

Pour modéliser de tels réseaux, extrêmement complexes, on s'expose immédiatement à la difficulté de la *mesure* des quantités au sein d'organismes vivants, étant donnés, d'une part, leur aspect microscopique et, d'autre part, leur fragilité. L'arme de choix est alors la biologie, qui attaque les modules fonctionnels par une approche descendante de type *boîte blanche*. Cependant, l'avènement de la génomique, de pair avec la bioinformatique, a permis de réaliser des avancées remarquables. De plus, les développements, depuis un siècle, de la physique et de la biochimie ont donné le jour à de nouveaux outils permettant d'accéder à une masse de plus en plus considérable de données, qui inonde dorénavant la communauté scientifique. Le traitement et l'interprétation de ces données, issues d'expériences bruitées et pas toujours

reproductibles, posent maintenant de nouveaux dilemmes. C'est pourquoi on voit apparaître sur la scène, des scientifiques issus des mathématiques, de l'informatique (calcul scientifique et calcul formel), de la physique théorique, de l'automatique des systèmes, etc.

Le but de cette thèse fut de tirer sur ce voile des modules fonctionnels.

La difficulté de l'interdisciplinarité de ce travail s'est aussi ressentie dans l'exercice de rédaction, pour lequel les coutumes et exigences diffèrent parfois. Ce manuscrit peut paraître long... il a été rédigé dans le but de pouvoir être repris par une autre personne dans le même contexte interdisciplinaire. Par ailleurs, certains chapitres — voire certaines sections — sont plus adressés à tel ou tel corps de spécialistes. C'est pourquoi la structure est maintenant détaillée.

Afin de pouvoir se rattacher à des éléments connus, nous avons entamé la question par l'approche ascendante, c'est-à-dire par la modélisation des molécules qui obéissent toujours aux lois de la mécanique (quantique et/ou newtonienne). Une grande partie de notre travail a alors été d'intégrer les bases de chimie nécessaires à la problématique ; pour que ce travail puisse être accessible par d'autres « non-spécialistes », nous avons donc souhaité consacrer le premier chapitre à un mini-exposé de ces quelques rudiments. Certainement, ce chapitre paraîtra superficiel aux chimistes, cependant, il fonde le modèle mathématique permettant de reformuler le problème physique en une question d'optimisation. Y sont rapidement décrits : ce qu'est une molécule, les principaux termes de vocabulaire utilisés ultérieurement et quelques principes du repliement des molécules. En particulier, nous avons insisté sur le fait que, contrairement aux problèmes classiques de recherche opérationnelle, nous ne cherchons pas *une* solution satisfaisante à un problème, mais bien *toutes* les solutions minimisant le critère énergétique.

Après cette première présentation, purement chimique, de la molécule, le deuxième chapitre présente les différentes étapes d'intégration de la molécule dans l'ordinateur : il faut encoder l'information sur les atomes, la géométrie de la molécule, il faut pouvoir décrire sa flexibilité ; et puisque cette flexibilité dépend de la forme du « paysage énergétique », il faut pouvoir estimer cette énergie interne. Ce chapitre permet de donner un aperçu des approches utilisées dans la littérature et de justifier nos choix. Il s'achève sur la définition du cadre précis des recherches menées. Ces dernières font l'objet des chapitres 3 et 4.

Pour traiter les interactions entre plusieurs molécules, il a fallu s'attacher au

cas particulier où « plusieurs = 1 », c'est-à-dire prédire *in silico* la géométrie des molécules d'intérêt, véritable base de leurs fonctions. Cette étape est appelée « d'*échantillonnage conformationnel* ». La phase de *docking* (ancrage d'une petite molécule dans le site actif d'une plus grande) apparaît alors comme une généralisation naturelle où, à la flexibilité des deux molécules, on ajoute les degrés de liberté correspondant au positionnement de chacun des partenaires.

Le troisième chapitre est exclusivement consacré à notre travail concernant l'optimisation de l'échantillonnage conformationnel d'une seule molécule. Il détaille l'ensemble des algorithmes développés, les résultats et les avancées, la parallélisation de la stratégie et deux applications. Le quatrième chapitre présente nos premiers développements en vue de la prédiction des affinités entre plusieurs molécules. Il concerne essentiellement la gestion des degrés de liberté du positionnement relatif des molécules et s'achève sur les développements envisagés pour le futur.

Ainsi, bien que les chapitres 1 et 2 fassent partie intégrante de notre travail, dans le sens où ils traduisent une formation à un nouveau domaine, ils ne sont pas indispensables à la compréhension des stratégies développées. Ces dernières ont été volontairement rassemblées dans les chapitres 3 et 4.

Dans une deuxième partie (chapitre 5), nous nous sommes attachés à une modélisation plus abstraite des modules fonctionnels (approche descendante) et avons montré en particulier comment les dynamiques d'interactions moléculaires peuvent entraîner différents comportements à l'échelle du module fonctionnel. Il existe plusieurs exemples typiques de comportements : bistabilité (mémoire, commutateur), multi-stabilité (différentiation cellulaire), oscillations (horloges internes), arythmie, hystérèse, voire parfois phénomènes chaotiques. Pour notre part, nous nous sommes concentrés sur la modélisation d'un module d'horloge qui permet aux organismes de synchroniser leur métabolisme sur le rythme du jour et, ainsi, d'anticiper les périodes de lumière et celles de pénombre.

NB : pour faciliter la lecture (éventuellement non linéaire) de ce manuscrit, nous avons ajouté en annexe une liste des abréviations employées, page 237.

Liste des symboles utilisés

symbole	signification
$\triangleq$	égal, par définition
$\sharp A$	cardinal de l'ensemble A
$B(a,r)$	boule de centre a et de rayon r
$A \perp B$	A est orthogonal à B pour le produit scalaire considéré
$\binom{n}{p}$	coefficient binomial de Newton égal à $\frac{n!}{p!(n-p)}$
$\mathrm{Re}(z)$, $\mathrm{Im}(z)$	respectivement, parties réelle et imaginaire du complexe ou du quaternion z
$\langle u \vert v \rangle$	produit scalaire dans l'espace vectoriel considéré
$\delta(x = x_0)$	mesure ou distribution (selon le contexte) de Dirac en x_0
$\mathcal{L}^2(\mathbb{R}^n)$	espace des fonctions de carré intégrable sur $\mathbb{R}^n$
$d(A,B)$	distance entre les points A et B dans l'espace considéré et selon la distance considérée. Parfois, la notation $d_{1,2}$ est utilisée pour dénoter la distance euclidienne entre les centres des atomes numéros 1 et 2.
${}^\top V, {}^\top X$	transposé du vecteur V ou de la matrice X
$\mathrm{tr}(X)$	trace de la matrice X
$\det(X)$	déterminant de la matrice X

Première partie

La modélisation moléculaire

Chapitre 1

Introduction à la chimie et biochimie

1.1 Introduction

La compréhension des mécanismes du repliement tridimensionnel et des interactions des molécules est, d'une part, particulièrement prometteuse, car ceux-ci interviennent dans de nombreux processus biologiques et leurs dysfonctionnements sont incriminés directement dans le développement de certaines maladies (Alzheimer, vache folle, etc.). D'autre part, elle continue de défier les scientifiques depuis plus de cinq décénnies.

En cherchant à modéliser l'arrimage entre molécules, nous avons développé une suite de programmes qui se distinguent par deux points très importants. Tout d'abord, contrairement à de nombreuses autres approches[1], nous considérons toutes les molécules, sans restriction, de manière générique. Nous pensons que, s'il existe un champ de force pour décrire les interactions à l'échelle atomique, il doit pouvoir s'appliquer aussi bien aux protéines qu'aux petites molécules organiques, ou qu'aux bases d'ADN. Aussi nous rappelons, dans ce premier chapitre, les quelques notions élémentaires de chimie dont nous avons besoin.

Notre travail se différencie aussi par l'approche multimodale et l'envie de caractériser, même de façon sommaire, *tous* les états probables. Ainsi, nous ne cherchons pas *la* structure la plus stable, mais tentons de décrire la molécule en solution avec sa flexibilité, tout en restant dans une description statique. Nous présentons donc succinctement les bases théoriques qui dictent la conformation des molécules.

Cette présentation à l'usage du lecteur étranger à la chimie peut être omise par

[1] certains ne considèrent qu'un seul type de molécules, d'autres, qu'une seule molécule (Jin *et al.*, 1999).

le chimiste qui trouvera, au besoin, des références aux paragraphes correspondants dans la suite des chapitres.

Cette partie s'articule autour de la section principale 1.3 qui, après un rapide aperçu de ce qu'est une molécule dans la section 1.2, présente la ou les structures des molécules (1.3.1 et 1.3.2), le pourquoi (1.3.3) et le comment (1.3.4) physiques de cette structuration.

1.2 Qu'est-ce qu'une molécule ?

1.2.1 Cas général

La molécule se présente comme un système d'atomes reliés entre eux par des liaisons dites de *covalence* (figure 1.1).

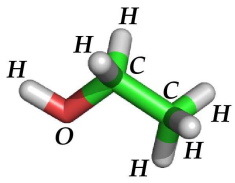


FIG. 1.1: premier exemple de molécule : l'éthanol qui succèdera peut-être aux carburants actuels.

Ces liaisons sont le fait de la mise en commun d'*orbitales électroniques* : les noyaux atomiques sont en effet entourés de un ou plusieurs nuages électroniques qui occupent des orbitales dites *liantes* ou *non-liantes*, selon qu'elles sont respectivement partiellement remplies (un seul électron cherchant à se lier) ou entièrement remplies (par un doublet d'électrons complémentaires). Ces liaisons covalentes peuvent se rompre et se former, c'est le cas des *réactions chimiques* (figure 1.2).

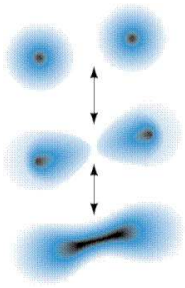


FIG. 1.2: formation et dissociation du dihydrogène.

Une molécule peut donc être interprétée, d'un point de vue topologique, comme

un *graphe* où les atomes sont les sommets et les liaisons covalentes, les arêtes. Ce graphe peut comporter des cycles (figures 1.3).

FIG. 1.3: exemples de molécules cycliques : la caféine et la molécule de fullerene.

Dans cette représentation, chaque type atomique se distingue par certaines caractéristiques (voir tableau 1.1) comme son nombre de voisins — appelés substituants — son rayon de covalence, son électronégativité, etc. Conventionnellement, on attribue une couleur aux principaux types atomiques.

Atome	**symbole**	**nombre de liaison(s)**	**rayon de covalence (en Å)**	**couleur**
Carbone	C	4	0,77	vert ou noir
Azote	N	3	0,75	bleu
Phosphate	P	3	1,06	marron
Oxygène	O	2	0,73	rouge
Soufre	S	2	1,02	jaune
Hydrogène	H	1	0,37	blanc
Fluor	F	1	0,71	bleu ciel
Chlore	Cl	1	1,00	vert
Brome	Br	1	1,14	bordeau
Iode	I	1	1,33	violet

TAB. 1.1: caractéristiques des principaux atomes rencontrés.

Cependant la molécule reste un objet tri-dimensionnel et tous les voisins d'un atome donné ne sont pas forcément équivalents. Ainsi, par exemple, si les quatre substituants d'un carbone tétrahédrique sont différents, la molécule et son image dans un miroir ne seront pas superposables et auront des propriétés physico-chimiques différentes. L'atome responsable est alors dit *asymétrique*, la molécule *chirale* et les deux molécules images l'une de l'autre sont des *stéréoisomères*. C'est le cas, par exemple, de la carvone, dont une molécule est à l'origine de l'odeur de fenouil de l'aneth, tandis que son stéréoisomère donne une odeur de menthe. La chiralité peut aussi apparaître lorsqu'il n'y a que trois substituants, mais qu'il existe un nuage électronique forçant une géométrie tétrahédrique ; un exemple de telles molécules est donné figure 1.4.

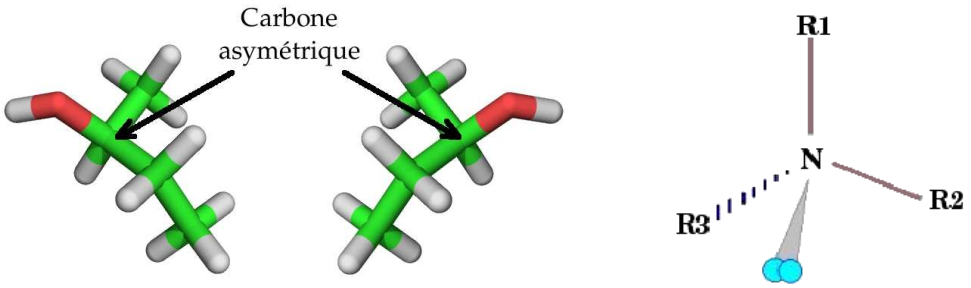


FIG. 1.4: (gauche) le carbone indiqué par une flêche, est asymétrique, les deux stéréoisomères sont chimiquement différents. (droite) De même, l'azote, qui a une structure tétraédrique due à ses trois substituants et son doublet électronique non-liant, est asymétrique.

1.2.2 Exemples biologiques

En chimie, on partitionne généralement l'étude des molécules en deux grandes sections que sont la *chimie organique* et la *chimie inorganique* ou *minérale*. La première concerne l'étude des composés dits organiques ou carbonés, car ils sont principalement constitués de carbone et d'hydrogène. La deuxième étudie tous les composés non-organiques (minéraux, métaux, complexes métalliques, etc.).

Enfin, la *biochimie* — en intersection non nulle avec ces deux domaines — s'intéresse aux réactions qui ont lieu dans et au voisinage des cellules (et éventuellement au niveau de leurs parois). Si la biochimie est en grande partie organique, on compte toutefois de nombreux éléments métalliques intervenant dans des processus biologiques.

Il faut également noter un élément qui distingue la chimie classique de la biochimie : dans la première, les « réactions » sous-entendent des modifications *covalentes* de la molécule, tandis que la deuxième répertorie également des interactions beaucoup plus faibles et réversibles (repliement, arrimage de molécules, etc.).

Parmi les molécules du vivant, on peut répertorier les suivantes (liste non-exhaustive) :

L'ADN ou acide désoxyribonucléique constitue le support du génome ; il est, non pas une, mais deux molécules enroulées en forme de double hélice (première structure proposée par Watson et Crick en 1953). Chacune des deux molécules est une succession de motifs appelés nucléotides. Il en existe quatre (figure 1.5) :

- G ou Guanine
- C ou Cytosine
- A ou Adénine

– T ou Thymine

Ces nucléotides s'apparient avec les nucléotides du deuxième brin selon le schéma A···T et G···C, formant ainsi une sorte de négatif.

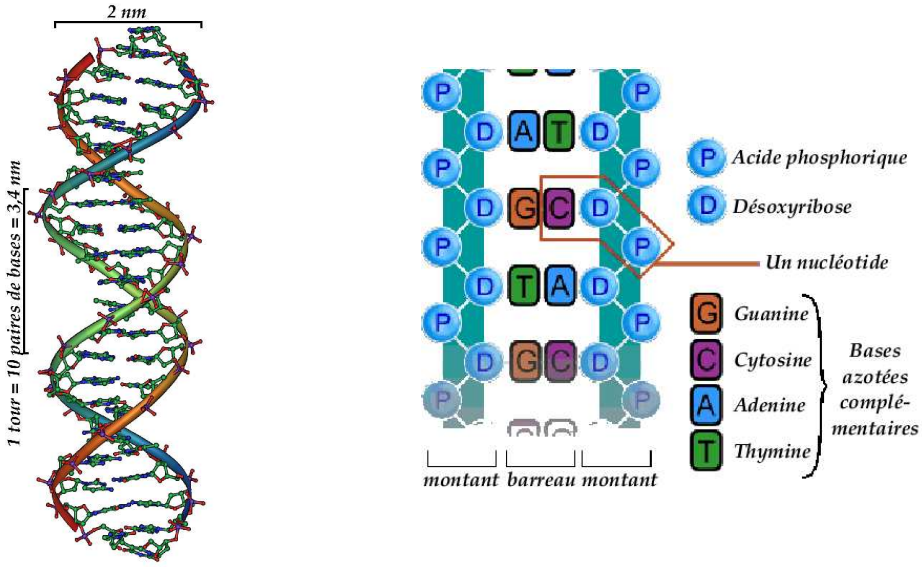


FIG. 1.5: structure de l'ADN : répétition des paires appariées de nucléotides en double hélice.

Les études théoriques concernant la modélisation de la structure tridimensionnelle de l'ADN (Ruscio et Onufriev, 2006; Sun *et al.*, 2005; Lauria *et al.*, 2004; Cui et Simmerling, 2002; Packer et Hunter, 2001; Hobza *et al.*, 1998) sont assez complètes, prenant en compte de nombreux paramètres, mais rencontrent la difficulté des grandes longueurs de brins d'ADN (de 50 à 250 millions de bases pour les chromosomes humains) ainsi que celle de l'enroulement de la double hélice sur d'autres structures (comme les protéines histones). En revanche, les divers niveaux de compactage permettent d'aborder la question à différentes échelles : pour un aperçu des références, voir Lavelle et Benecke (2006).

Par ailleurs, la grande quantité de données disponibles dans ce domaine, grâce essentiellement aux travaux de la génomique, a donné lieu à des études statistiques sur des chromosomes entiers qui ont mis en évidence des autocorrélations à longues distances entre les séquences ainsi que l'existence de structures particulières et qui ont permis d'expliquer leurs implications (Audit *et al.*, 2002; Vaillant *et al.*, 2005).

L'ARN ou acide ribonucléique est semblable à l'ADN (succession de nucléotides, excepté la thymine qui est remplacée par de l'uracile de symbole U), mais diffère par sa stabilité beaucoup plus faible, sa structure généralement simple-brin et sa taille

moindre (de 50 à 5000 nucléotides).

L'ARN diffère aussi de l'ADN par ses fonctions étendues ; on le retrouve ainsi dans le cytoplasme. Par sa stabilité limitée, l'ARN a plutôt un rôle temporaire de transport d'information tandis que l'ADN « *stocke* » le matériel génétique. Mais il peut également remplir certaines fonctions *effectives* des biomolécules au même titre que les protéines et les enzymes.

En guise d'exemple, citons

- l'*ARN messager*, qui est une copie (on parle de *transcription*) d'un gène de l'ADN : son rôle est d'acheminer l'information génétique du noyau vers les ribosomes du cytoplasme ;
- au niveau des *ribosomes* (eux-même constitués d'ARN et de protéines) qui permettent de « *traduire* » l'ARN messager en protéine, chaque triplet de nucléotides (appelé *codon*) est lu ; un autre ARN — l'ARNt ou *ARN de transfert* — est alors recruté, effectue la conversion nucléotide vers acide aminé et déclenche la polymérisation du nouvel acide aminé sur la protéine en cours de fabrication.
- l'*ARN de transfert* est lui-même un ARN très court (70 à 100 nucléotides) comportant un acide aminé.

L'exercice de prédiction des géométries de l'ARN (Mathews et Turner, 2006) bénéficie de la tendance des nucléotides à s'apparier : A avec U et G avec C.

Les protéines sont des assemblages séquentiels[2] d'acides aminés, formant une chaîne et reliés entre eux par des liaisons dites peptidiques[3]. Il existe vingt acides aminés très courants (et d'autres plus exotiques) tous bâtis sur le même modèle schématisé sur la figure 1.6 (exception faite de la proline, figure 1.7) et représentés par une lettre majuscule de l'alphabet latin.

La partie qui varie d'un acide aminé à l'autre est appelée *chaîne latérale* de l'acide aminé (*side chain*), tandis que l'enchaînement des motifs répétés « NH-CH-CO » forme le *squelette* (*backbone*). De plus, le carbone au point d'embranchement de la chaîne latérale est généralement dénommé « carbone alpha » ou C_α ; les autres carbones de la chaîne sont ensuite comptabilisés C_β, C_γ (ou $C_\gamma^{1,2}$ s'il y en a plusieurs), etc. Notons aussi que la chaîne principale des acides aminés n'est pas symétrique,

[2] excepté le cas des liaisons cystéine-cystéine

[3] ce sont les liaisons entre azote et carbone d'un groupement O=C—N-H ; lors de la mise en commun des orbitales électroniques, des électrons se délocalisent, stabilisant par résonance la liaison qui acquiert un caractère de double liaison et qui ne peut plus subir de libre rotation autour de son axe.

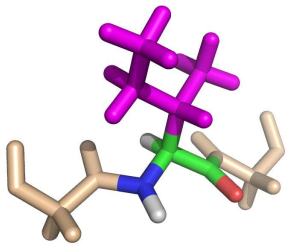


FIG. 1.6: structure d'un acide aminé ; la chaîne principale est représentée avec les couleurs habituelles des atomes, la chaîne latérale est en magenta (ici, une isoleucine), en beige : les acides aminés suivant et précédent.

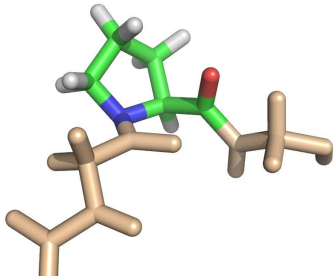


FIG. 1.7: la proline diffère des dix-neuf autres acides aminés.

de sorte que la séquence L-I-V-E, par exemple, n'a pas le même sens (biologique) que E-V-I-L : il y a un sens de lecture. L'extrémité initiale[4], dans la biosynthèse de la séquence, est dite *N-terminal* (par opposition au brin *C-terminal*[5]) et dénote le début de la chaîne d'acides aminés (respectivement la fin). C'est aussi le sens conventionnel pour l'écriture de la séquence.

Enfin, la proline est construite sur le même principe que les dix-neuf autres acides aminés, mais son azote est covalemment lié au dernier carbone de sa chaîne latérale, ce qui en fait un acide aminé cyclique et donc beaucoup plus rigide (figure 1.7). De plus, les deux états stables de la liaison peptidique sont moins déséquilibrés énergétiquement que dans le cas des autres acides aminés, de sorte que la proline existe sous deux formes dites *cis* et *trans* (figure 1.8).

Par définition, un *peptide* est une chaîne d'acides aminés reliés par des liaisons peptidiques ; toute protéine est donc un peptide. Cependant, les chimistes réservent habituellement le terme « peptide » pour les courtes séquences de moins de 50 à 100 résidus n'ayant — en général — pas de fonction biologique (figure 1.9), par opposition aux plus grandes protéines (figure 1.10, extraite de (Dobson *et al.*, 1998)).

Encore une fois, la simulation de la conformation des protéines est très large-

[4]avec le groupement NH_2 libre
[5]avec le groupement CO-OH libre

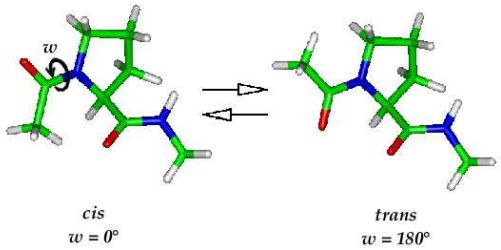


FIG. 1.8: la proline possède deux états stables dits *cis* et *trans*.

FIG. 1.9: deux représentations d'un peptide d'une longueur de 20 acides aminés ; dans la deuxième représentation, le squelette, formant deux hélices et un brin recouvrant l'ensemble, est mis en évidence.

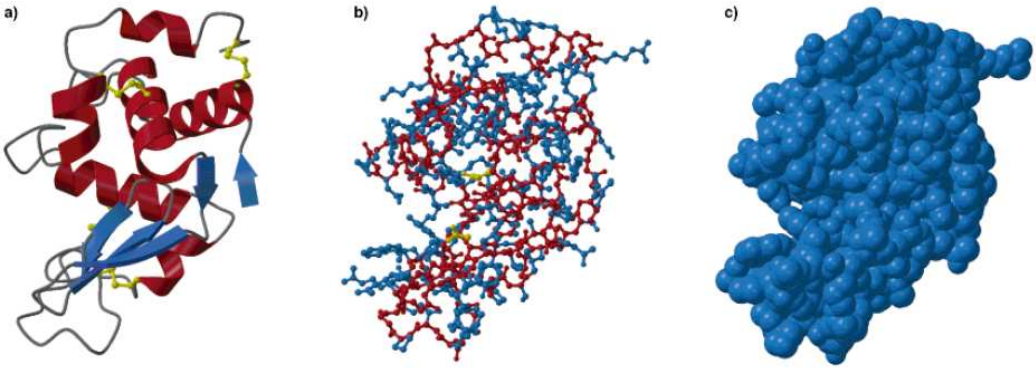


FIG. 1.10: différentes représentations du lysozyme : a) mise en évidence des éléments de structure sous forme de rubans (hélices en rouge, feuillets en bleu), les liaisons entre résidus cystéines sont représentées en jaune. b) Schématisation en boules et bâtonnets, les résidus participant au site actif sont en jaune. c) Représentation par des sphères pour souligner l'occupation spatiale de la molécule.

ment étudiée et tire profit de cette apparente séparation entre les degrés de liberté appartenant à la chaîne principale et ceux des chaînes latérales (voir chapitre 2, § 2.2.5).

Les enzymes sont des molécules (protéines ou ARN) qui catalysent, c'est-à-dire qui accélèrent, (jusqu'à des millions de fois), certaines réactions chimiques.

Chaque enzyme est extrêmement spécifique à sa cible (appelée *substrat*) grâce à son site actif. Celui-ci peut être présent de manière statique à la surface de l'enzyme, ou bien apparaître dynamiquement lors de l'assemblage des acteurs (complexes moléculaires ou activation par un ligand).

L'activité des enzymes et leur dépendance aux conditions environnementales en font des outils clefs dans les boucles de régulation génique, comme nous le verrons au chapitre 5.

Les *kinases* sont un exemple d'enzymes qui catalysent la *phosphorylation* (ajout d'un groupement phosphate) de certains acides aminés, elles appartiennent à la famille des *transférases* qui servent à lier des groupements fonctionnels sur certaines molécules de transport. Il existe également des *polymérases*, qui catalysent la synthèse des séquences d'ADN ou d'ARN, des *protéases* qui facilitent la dégradation des protéines, des *isomérases* qui accélèrent la transition des molécules entre leurs différents stéréoisomères, etc.

1.3 La structure des molécules

1.3.1 La flexibilité des molécules... un fait

Contrairement à ce que suggèrent les différentes figures en amont, il est faux de concevoir une molécule comme un solide indéformable, avec une structure figée. En réalité, une certaine flexibilité apparaît à différents niveaux :

La molécule oscille autour de sa conformation d'équilibre sous l'effet des chocs subis par son environnement (principalement des molécules d'eau, mais également des autres molécules). La force et la fréquence de ces chocs stochastiques entrent dans la notion de température, c'est pourquoi notre corps fonctionne différemment à 35°C , à 37°C et à 39°C . La température détermine les vitesses de réactions (nous en reparlerons lors de l'étude des rythmes circadiens, chapitre 5), mais, lorsqu'elle est trop importante, elle est aussi responsable de la déstructuration des protéines

(on parle de *dénaturation*). Les oscillations autour du point d'équilibre sont trop importantes pour que la protéine garde sa fonction physiologique.

La molécule interagit avec son environnement (sinon elle ne sert à rien...) et ces interactions reposent sur sa flexibilité (Karplus et Kuriyan, 2005), il y a alors déformation des structures pour obtenir le complexe final. C'est le cas lorsque les deux acteurs s'adaptent géométriquement l'un à l'autre, ou quand les mouvements de la molécule mettent à jour un site actif (Hornak et Simmerling, 2007), mais cela peut aussi survenir quand un ligand force l'ouverture du site dans lequel il vient se lier. Enfin, il faut également citer le cas de l'*allostérie* où l'interaction de deux partenaires moléculaires au niveau d'un site de fixation change la structure en d'autres sites, modifiant ainsi l'activité du complexe (la figure 1.11 fournit un exemple avec l'Aspartate TransCarbamylase ou ATCase).

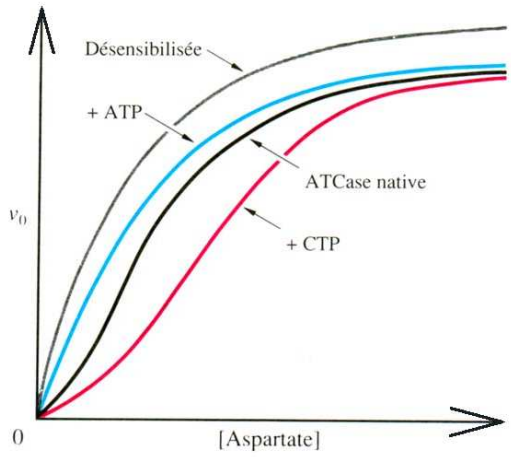


FIG. 1.11: exemple de modification allostérique de l'ATCase : la vitesse de réaction, qui est fonction de la concentration en aspartate, est modifiée par la présence des différents ligands, tous en compétition pour se fixer dans le site actif (figure extraite de http ://www.unine.ch/bota/bioch/cours/enzyme2.html, consulté en août 2007).

La molécule se dénature et agrège. Ci-dessus, a été introduite la notion de conformation d'équilibre ; cependant, le processus de repliement des molécules est complexe (démarrage immédiat pendant la synthèse, modifications ultérieures possibles, existence de *chaperones* qui encapsulent la molécule le temps de son repliement). La balance entre les différentes conformations stables d'une molécule est très dépendante de la température et de l'environnement chimique. Ainsi, l'albumine du blanc d'œuf change complètement d'aspect après cuisson (coagulation) parce que sont rassemblées des conditions environnementales très différentes des conditions

natives (la température dénature les structures et la concentration induit l'agrégation). La figure 1.12 ci-dessous présente un autre exemple qui est l'agrégation de la protéine humaine Tau en longs filaments, découverts dans le cerveau de patients décédés des suites de la maladie d'Alzheimer. Les récentes études laissent présumer une structuration pathologique en agrégats alors que la protéine native n'a pas de structure ; cependant, ni les mécanismes, ni les causes et conséquences de tels comportements moléculaires ne sont encore bien compris.

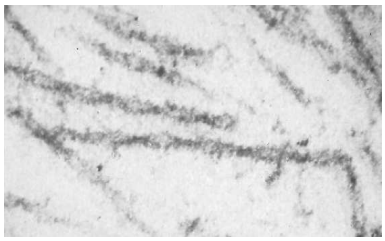


FIG. 1.12: filaments de protéines tau agrégée.

1.3.2 Les niveaux de structuration

Le degré de détail adopté pour décrire la structure d'une molécule permet différents niveaux de caractérisation.

On entend par *structure primaire*, la donnée de la formule brute de la molécule, c'est-à-dire, uniquement ce qui concerne les types atomiques entrant dans la composition et leur graphe de liaison. Ainsi, pour une protéine (ou un brin d'ARN ou d'ADN), toute la structure primaire est contenue dans la séquence de ses acides aminés (respectivement de ses nucléotides), sans aucune autre forme d'information. Attention, la structure primaire précise également les éventuelles asymétries que comporte la molécule.

La forme géométrique globale d'une molécule, qu'on appelle également *conformation* définit sa *structure tertiaire*, alors que, pour les protéines, on définit également la *structure secondaire* qui désigne seulement des sous-unités de structures qui la composent (figure 1.13). C'est le cas par exemple des hélices ou des feuillets que l'on trouve dans les protéines et dont la géométrie est stabilisée par des interactions à moyennes ou longues distances. La structure tertiaire correspond donc à l'arrangement des sous-structures secondaires entre-elles.

Lorsqu'il s'agit de complexes ou de très grandes molécules partitionnées en domaines, la *structure quaternaire* fait référence à l'organisation de tous ces domaines et partenaires dans l'espace (figures 1.14).

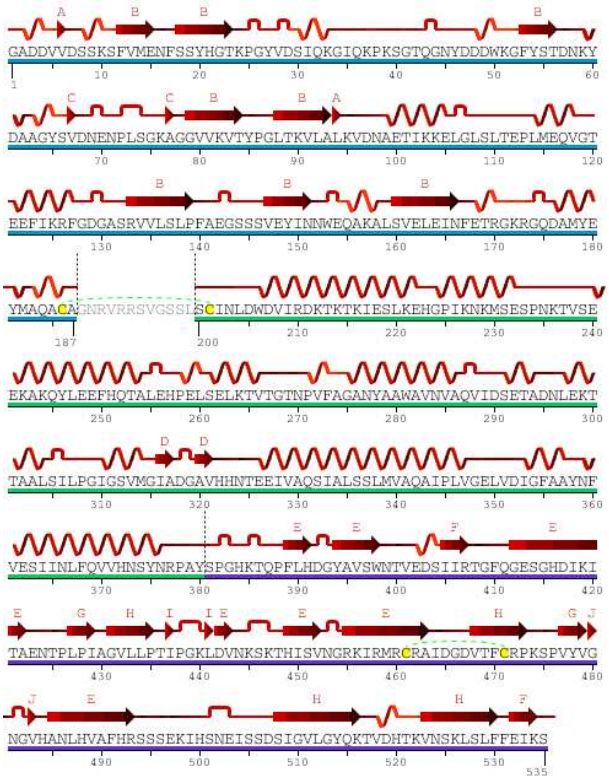


FIG. 1.13: structures primaire (séquence des acides aminés) et secondaire (éléments de structuration indiqués en rouge) de la protéine humaine « PIN1 ». Les flèches indiquent les feuillets β, les « ressorts » représentent les hélices et les créneaux schématisent les tournants.

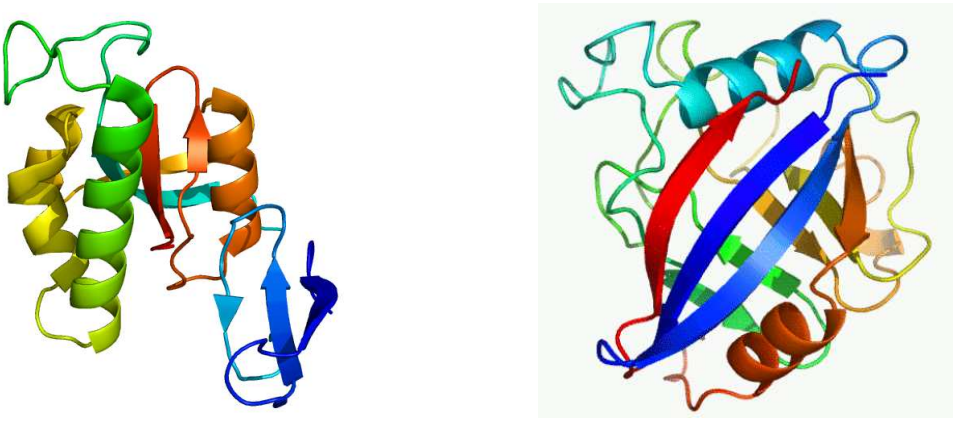


FIG. 1.14: structures tertiaire et/ou quaternaire; (gauche) PIN1, la partie bleue correspond au domaine « WW » dit de liaison, le reste étant le domaine catalytique (effectif). (droite) Cyclophiline B, intervenant dans le système immunitaire.

1.3.3 L'interprétation énergétique

> Physics is mathematical not because we know so much about the physical world, but because we know so little ; it is only its mathematical properties that we can discover.
>
> Bertrand Russell

Formellement, la structure tridimensionnelle de la molécule, lorsqu'elle est au repos, devrait pouvoir se déduire des états propres de l'hamiltonien quantique dans l'équation de Schrödinger. Cependant, même sous l'hypothèse simplificatrice de Born et Oppenheimer qui tirent parti du très grand rapport de masse entre noyaux et électrons ($> 10^3$) pour supposer ces derniers infiniment plus rapides, l'équation de Schrödinger ne reste numériquement envisageable que pour quelques centaines d'atomes.

Pourtant, un ensemble de règles établies plus ou moins empiriquement fait un peu de lumière sur les mécanismes sous-jacents et un certain nombre d'approximations et de modèles vont nous permettre de formaliser toutes les interactions ; c'est ce qui fait de la modélisation moléculaire un pont entre les disciplines de la physique statistique, de la mécanique quantique et newtonienne.

Plutôt que de raisonner en termes de *forces*, qui est un modèle typiquement newtonien, on parle plus généralement d'*interactions* et on considère dorénavant les *potentiels* desquels dérivent les forces[6] (équation (1.1)).

$$\begin{aligned} dV &= -F.d\ell, \\ F &= -\mathrm{grad}(V), \end{aligned} \tag{1.1}$$

où F est la force, $d\ell$ un déplacement élémentaire et V le potentiel.

La somme de toutes les contributions (forces électromagnétiques, effets quantiques et modèles empiriques des phénomènes supplémentaires) constitue l'énergie potentielle du système. À cette énergie s'ajoute la partie cinétique :

$$\begin{aligned} \text{(newtonien)} \quad E_c &= \sum_{i\in\mathcal{P}} \frac{1}{2} m_i \mathcal{V}_i^2 = \sum_{i\in\mathcal{P}} \frac{p_i^2}{2m_i}, \\ \text{(quantique)} \quad E_c &= \sum_{i\in\mathcal{P}} -\frac{\hbar^2}{2m_i} \triangle, \end{aligned}$$

[6] Pour que de tels potentiels existent, il faut des forces *conservatives*, c'est-à-dire de rotationnel nul.

où $\mathcal{P}$ est l'ensemble des particules de la molécule, m_i est la masse de la particule i, $\mathcal{V}_i$, sa vitesse et p_i son impulsion[7].

Cependant, il est illusoire de vouloir décrire individuellement *toutes* les particules d'une solution (une mole d'un composé chimique — c'est-à-dire le nombre d'atomes dans 12 grammes de ^{12}C — contient $N_A = 6 \times 10^{23}$ molécules, où N_A est le nombre d'Avogadro...), de plus, les innombrables chocs stochastiques que subissent les molécules rendent les études dynamiques difficiles : seuls des résultats statistiques sur de multiples et longues trajectoires peuvent être extraits de telles simulations. C'est pourquoi nous allons voir que nous pouvons nous restreindre à la seule partie potentielle de l'énergie interne (Bryngelson *et al.*, 2004).

1.3.3.1 Une description statique

> Mathematics are well and good but nature keeps dragging us around by the nose.
>
> Albert Einstein

Ce nombre astronomique de 600 mille milliards de milliards de molécules par mole a permis le développement d'outils spécifiques, apanage de la physique statistique. En particulier, L. Boltzmann a proposé une interprétation probabiliste de l'énergie interne résumée dans l'équation (1.2).

$$\Pr(\text{état d'énergie } E) = \frac{1}{Z} \exp\left(-\frac{E}{k_B T}\right). \tag{1.2}$$

Le préfacteur $\frac{1}{Z}$ étant un facteur de normalisation, calculé de sorte à avoir une densité de probabilité qui s'intègre à 1 sur l'ensemble des états accessibles Ω ; T est la température absolue en Kelvins, E est l'énergie exprimée en Joules et k_B est la constante de Boltzmann ($\approx 1,38 \times 10^{-23}$). Cette équation est fondamentale pour la suite de cet exposé et constitue la base de la compréhension actuelle de la stéréochimie.

Remarque : on note parfois β la *température inverse* égale à $1/RT$ où $R = k_B N_A \approx 8,3$ J.mol^{-1}.K^{-1} est la constante des gaz parfaits. Si on utilise une énergie exprimée en kcal.mol^{-1}, on obtient $\beta \approx \frac{503,5}{T}$.

Ainsi, certains états sont plus souvent visités que d'autres — ils sont dits *préférentiels* — et la prépondérance de ces états est quantifiée par l'équation (1.2). Éventuellement, seule une fraction des molécules peut se trouver dans l'état actif,

[7] $\hbar$ est la constante de Planck réduite ($\approx 1,05.10^{-34}$J.s) et Δ l'opérateur laplacien.

ce qui réduirait son activité. Un système qui ne possèderait que deux états A et B (configurations cis et trans d'une double liaison par exemple, ou bien conformations repliée et dépliée d'une molécule) d'énergies respectives E_A et E_B serait représenté par des sous-populations de chacun des deux états, proportionnelles aux ratios suivants :

$$\begin{aligned} \Pr(A) &= \frac{e^{-\beta E_A}}{e^{-\beta E_A} + e^{-\beta E_B}} \\ &= \frac{1}{1 + e^{-\beta(E_B - E_A)}}, \end{aligned} \tag{1.3}$$

$$\Pr(B) = 1 - \Pr(A) = \frac{e^{-\beta(E_B - E_A)}}{1 + e^{-\beta(E_B - E_A)}}. \tag{1.4}$$

Remarque : on voit sur cet exemple, que les niveaux de population de chaque état ne dépendent que de la différence énergétique, ce qui était prévisible, puisque tout potentiel est défini à une constante additionnelle[8] près : équation (1.1).

Dans un espace de phase continu, Ω, décrit par des degrés de liberté continus Θ, on interprète l'équation de Boltzmann en terme de *densité* de probabilité (équations (1.5) et (1.6)) :

$$\begin{aligned} \Pr\left(\Theta \in [\theta; \theta + d\theta[\right) &= p(\theta) d\theta \\ &= \frac{1}{Z} \exp\left[-\beta E(\theta)\right] d\theta, \end{aligned} \tag{1.5}$$

$$\Pr\left(\Theta \in \mathcal{D}\right) = \frac{1}{Z} \int_{\mathcal{D}} e^{-\beta\theta} d\theta. \tag{1.6}$$

1.3.3.2 L'énergie libre

En principe, l'énergie qui apparaît dans l'équation de Boltzmann (1.2) et (1.6) n'est pas l'énergie potentielle, mais plutôt l'*énergie libre*. Reprenons l'exemple précédent, d'une molécule qui peut être soit dans son état replié natif N, soit dans un état dénaturé D (paysage d'énergie en une dimension, figure 1.15) ; l'état natif replié sera généralement d'énergie inférieure à n'importe quel état déplié, mais il n'y a pas un seul état déplié, de sorte que ce qu'on a appelé « état déplié D » est en fait un ensemble (souvent énorme) d'états $\mathcal{D}_D \subset \Omega$ (figure 1.16).

[8] Le choix de cette constante est bien souvent dicté par la précision de l'ordinateur afin d'éviter tout problème dans le calcul numérique de l'exponentielle.

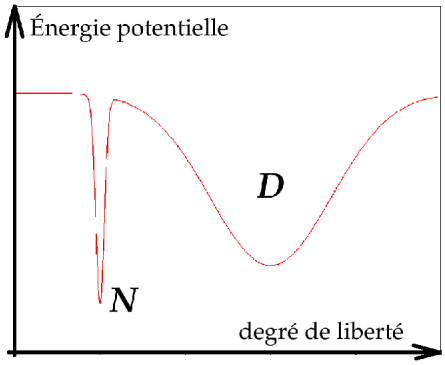


FIG. 1.15: un système à deux états, l'état natif N (puits de gauche) est énergétiquement favorable par rapport à l'état dénaturé D (puits de droite).

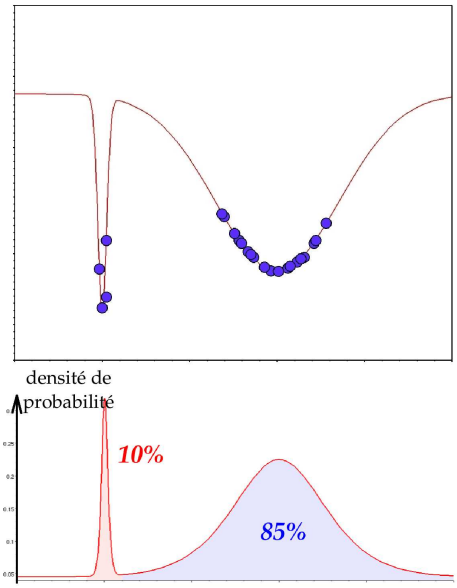


FIG. 1.16: bien que l'état natif soit meilleur en énergie, la largeur du puits de potentiel peut favoriser d'autres états sous-optimaux (dépendamment de la température).

La véritable probabilité de l'« état déplié » est donc

$$\Pr(D) = \int_{\mathcal{D}_D} \frac{1}{Z} e^{-\beta E(\theta)} d\theta, \tag{1.7}$$

$$\text{où} \quad Z = \int_{\Omega} e^{-\beta E(\theta)} d\theta. \tag{1.8}$$

Bien que l'état natif soit énergétiquement favorable, la largeur du puits de potentiel peut favoriser l'état dénaturé et ce d'autant plus que la température sera élevée (paramètre β des équations). C'est ce que représente l'*entropie* d'un état (mesure du désordre, généralement noté S).

Si on souhaite rassembler tous les états dénaturés en un *super*état, on ne peut plus utiliser l'énergie interne, mais on définit l'énergie libre d'un domaine $\mathcal{D}$ par :

$$G(\mathcal{D}) \triangleq -\frac{1}{\beta} \ln \left(\int_{\mathcal{D}} e^{-\beta E(\Theta)} d\Theta \right), \tag{1.9}$$

$$\text{de sorte que} \quad \Pr(\mathcal{D}) = \frac{1}{Z} e^{-\beta G(\mathcal{D})}. \tag{1.10}$$

Dans notre exemple, si on note V_N et V_D les *volumes* respectifs des domaines $\mathcal{D}_N$ et $\mathcal{D}_D$, alors, les probabilités des états N et D sont données par l'équation (1.11) et (1.12) :

$$\Pr(N) = \int_{\mathcal{D}_N} \frac{1}{Z} e^{-\beta E(\theta)} d\theta$$

$$\Pr(N) \propto \frac{1}{Z} V_N e^{-\beta E_N}, \tag{1.11}$$

$$\text{de même} \quad \Pr(D) \propto \frac{1}{Z} V_D e^{-\beta E_D}, \tag{1.12}$$

$$\text{et} \quad Z \propto V_N e^{-\beta E_N} + V_D e^{-\beta E_D}, \tag{1.13}$$

où le coefficient de proportionnalité (que l'on notera α) est le même dans les trois équations (1.11), (1.12) et (1.13).

Il vient alors les énergies libres suivantes :

$$\begin{aligned} G(\mathcal{D}_N) &= -\tfrac{1}{\beta} \ln \left[Z.\Pr(N) \right] = E_N - T k_B \ln(V_N) - \tfrac{1}{\beta} \ln(\alpha), \\ G(\mathcal{D}_D) &= \underbrace{E_D}_{\text{énergie interne}} - T \times \underbrace{k_B \ln(V_D)}_{\text{entropie}} \underbrace{-\frac{1}{\beta} \ln(\alpha)}_{\text{constante}}. \end{aligned} \tag{1.14}$$

Remarque : de même que pour l'énergie interne, l'énergie libre est définie à

une constante près, de sorte que le facteur en $\ln(\alpha)$ peut être retranché des deux équations (1.14).

On retrouve alors la formule, plus courante, de l'énergie libre, où S désigne l'entropie du domaine :

$$G = E - T.S. \tag{1.15}$$

À titre d'exemple, dans son cours[9], Levitt propose l'étude du taux de conformations hélicoïdales d'une protéine lors de simulations de dynamiques moléculaires à différentes températures. Plus celle-ci est élevée, plus les géométries dénaturées prennent le pas sur les conformations natives (figure 1.17).

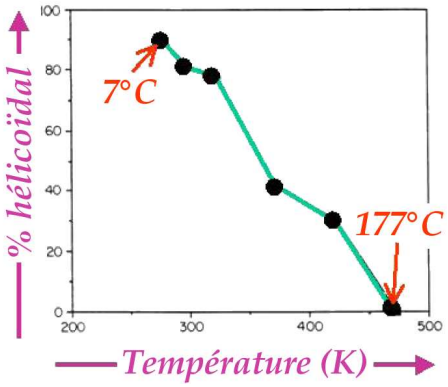


FIG. 1.17: taux de conformations hélicoïdales en fonction de la température lors de dynamiques moléculaires (extrait du cours en ligne de Levitt : http ://csb.stanford.edu/levitt/, consulté en juillet 2007).

Notons également un autre facteur important, qui est la dimension N_{ddl} de l'espace de phase, puisque le volume V évolue en $L^{N_{\text{ddl}}}$, où L est la taille du puits dans chacune des dimensions.

Voici l'exemple d'un système à deux états ayant les caractéristiques suivantes :

$$\begin{aligned} \Delta E &= 10 \text{ kcal.mol}^{-1}, \\ L_2/L_1 &= 10, \\ T &= 300 \text{ K}, \\ \rho_{N_{\text{ddl}}} &= \Pr(\text{Etat } E_1)/\Pr(\text{Etat } E_2). \end{aligned}$$

[9] http ://csb.stanford.edu/levitt/, consulté en juillet 2007.

N_{ddl}	$\rho_{N_{\text{ddl}}}$
1	50×10^{-6}
5	5×10^{-3}
10	500
20	5×10^{12}

Enfin, pour illustrer l'effet de l'entropie, la figure 1.18 présente, pour différentes températures, la position moyenne de la molécule dans son espace de phase (espérance mathématique). À mesure que la température augmente, tous les états deviennent équiprobables dans la formule de Boltzmann (1.2).

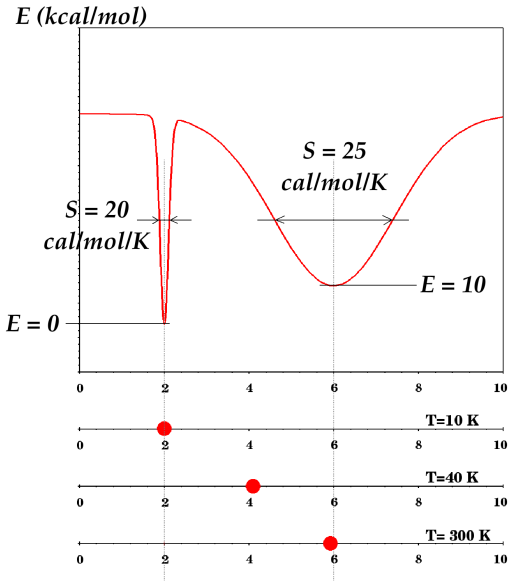


FIG. 1.18: paysage énergétique et position moyenne pour différentes températures. Plus la température est élevée, plus les solutions sous-optimales sont favorisées.

1.3.3.3 L'hypothèse thermodynamique

> Les lois de la Nature ne sont que les pensées mathématiques de Dieu.
>
> Euclide

Les expériences de dénaturation et de repliement de molécules ont conduit la communauté scientifique à accepter l'*hypothèse thermodynamique* mise en avant par Anfinsen (1973) (voir également Govindarajan, 1998) et initialement énoncée de la manière suivante :

> les molécules adoptent, dans leur milieu physiologique normal, la structure tridimensionnelle qui minimise leur énergie libre. [...] Autrement dit, la géométrie d'une molécule est entièrement déterminée par les interactions qu'elle abrite.

Cette reformulation succinte de ce qui a été présenté au paragraphe précédent (sous réserve que les notions d'énergie libre et d'entropie soient correctement assimilées, voir équations (1.9) et (1.14)) soulève cependant un nouveau problème.

L'interprétation de Anfinsen oblige en effet à redéfinir la notion d'« *état* ». Alors qu'un état représentait précédemment une géométrie possible ou un point (sans dimension) dans l'espace de phase, il s'agit maintenant d'un *sous-domaine caractéristique* de l'espace de phase (voir figure 1.19). Cependant, la façon dont sont réunies les conformations — ou, équivalemment, la partition de l'espace de phase en domaines caractéristiques — est laissée au libre arbitre du chimiste.

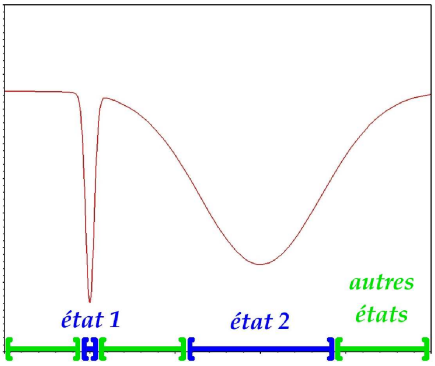


FIG. 1.19: un « *état* »représente maintenant un ensemble de conformations possibles.

D'un point de vue chimique, deux conformations correspondant à des caractéristiques chimiques similaires doivent clairement être rassemblées en un seul état. Mais en chimie structurale, deux géométries qui divergent nettement, même si les caractéristiques chimiques sont conservées, seront différentiées par deux états distincts.

Ce problème n'est pas minime et pèse sur la modélisation ; de plus, on ne peut pas le contourner à moindre frais par une définition mathématique du type partitionnement en *classes d'équivalence* où la relation d'équivalence serait par exemple donnée par une des équations (1.16) et (1.17).

$$\theta_1 \sim \theta_2 \iff \arg\min_{\theta \in B(\theta_1,R)} E(\theta) = \arg\min_{\theta \in B(\theta_2,R)} E(\theta), \tag{1.16}$$

$$\theta_1 \sim \theta_2 \iff \theta_1 \text{ et } \theta_2 \text{ sont dans le même bassin d'attraction.} \tag{1.17}$$

$B(\theta_i, R)$ étant la boule de centre θ_i et de rayon R.

Dans le premier cas (1.16), la définition de R reste le point sensible : un minimum très étroit peut « être ou ne pas être » physiquement pertinent. Néanmoins, l'idée a été réutilisée dans certains algorithmes qui n'utilisent plus l'énergie potentielle de chaque conformation, mais calculent celle de l'optimum local le plus proche (dans un domaine permis, voir Schug *et al.*, 2005a). Dans le deuxième cas, l'hypersurface d'énergie potentielle extrêmement accidentée multiplie le nombre de minima locaux ; même au fond des puits les plus profonds, de nombreux minima restent présents (voir figures 1.20 et 1.21 tirée du chapitre de Karplus et Shakhnovitch 1992). Gfeller *et al.* définissent un état comme un « *bassin* » tel qu'il peut être mis en évidence par une dynamique moléculaire.

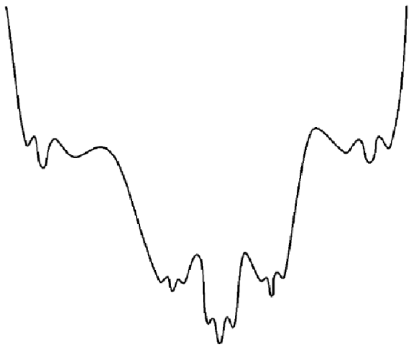


Diagram of a hierarchically-structured energy surface. At low resolution, there is only one energy well. Closer examination shows that this is subdivided into three shallower energy wells, each of which is further subdivided.

FIG. 1.20: figure tirée de Given et Gilson (1998) présentant un profil hiérarchisé d'énergie potentielle (voir légende).

1.3.4 Le processus de repliement

Après avoir rappelé les quelques résultats importants de l'approche statique, où le système est supposé avoir atteint un certain équilibre statistique (i.e. thermodynamique), nous présentons brièvement les contraintes mécaniques dues à l'aspect dynamique des molécules.

1.3.4.1 Le paradoxe de Lévinthal

Dieu a écrit l'Univers dans un langage mathématique.

Galilée

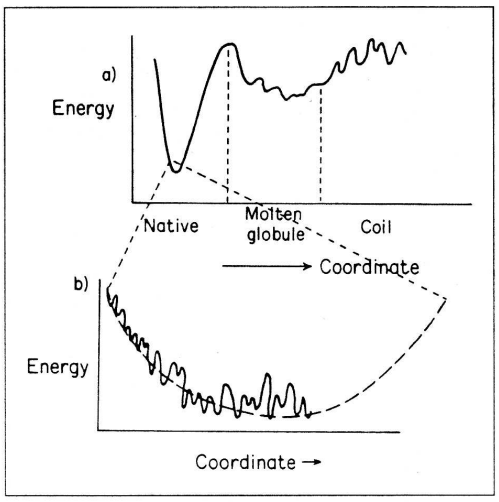


FIG. 1.21: à chaque échelle apparaissent de nouveaux minima.

En mars 1969 a eu lieu une conférence à l'université de l'Illinois ayant comme sujet « Mössbauer Spectroscopy in Biological Systems ». À cette époque, il était communément accepté que les protéines se repliaient progressivement, formant peu à peu les motifs structuraux de leurs géométries finales à mesure qu'elles échantillonnaient leurs espaces de phase comme une bille qui roulerait sur une nappe (la figure 1.22 est d'origine).

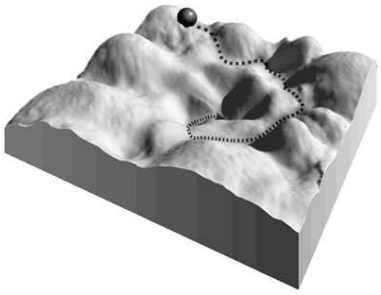


FIG. 1.22: le système (bille ou protéine) évolue sur l'hypersurface, explorant différentes vallées et tombant éventuellement dans un minimum qui peut être le minimum global.

La métaphore est esthétique, mais une recherche *aléatoire* du minimum absolu n'est pas concevable. C'est l'objet du séminaire de Levinthal (Levinthal, 1969) qui présente l'analyse grossière suivante : si une petite protéine comporte une centaine d'acides aminés où chacun possède trois états stables, alors, la protéine complète doit avoir $3^{100} \approx 10^{48}$ minima ! Même si la protéine évolue très rapidement d'un état à un autre (au moins supérieur à la femtoseconde), il faut plus de 10^{25} années pour

tout explorer (par comparaison, l'univers a seulement 15 milliards d'années...).

Cet événement reste toutefois plus probable que de voir un des singes de Borel taper une pièce de Shakespeare sur une machine à écrire (Borel, 1913)... même en 15 milliards d'années.

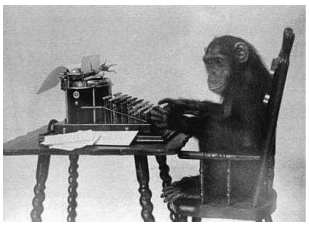


FIG. 1.23: un million de singes dactylographes tapant sur un million de machines à écrire peuvent-ils réinventer Hamlet, par hasard ?

En réalité, il n'y a pas de *paradoxe* dans le sens où l'expérience s'accorde avec ces probabilités très faibles, mais à très très hautes températures. À température ambiante, la moindre ΔE bouleverse les niveaux de population. Le paysage d'énergie est donc nécessairement « conçu » de manière à attirer rapidement la molécule vers sa géométrie native (Zwanzig *et al.*, 1992).

1.3.4.2 Représentations du paysage

L'article de Dill (1997) propose un certain nombre de figures (voir figures 1.24 à 1.27) faisant coller l'interprétation en termes de paysages d'énergie aux phénomènes expérimentalement connus.

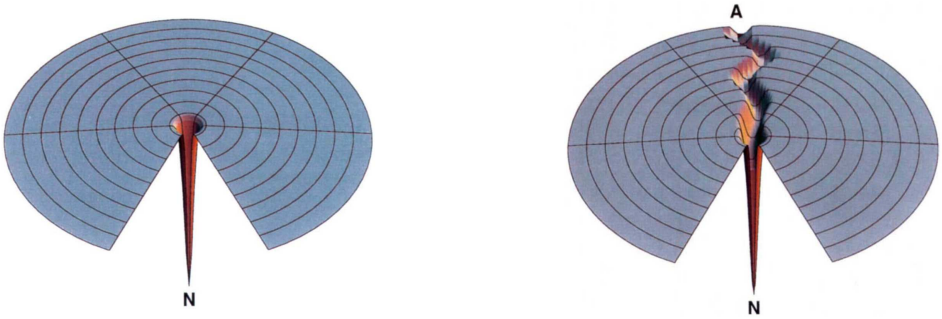


FIG. 1.24: (gauche) le paysage d'énergie vu par Levinthal : N représente la conformation native que la protéine recherche aléatoirement. (droite) L'existence de *chemin de repliement* permet de guider les molécules d'une conformation dénaturée (A) vers leur état natif.

Bien entendu, ces figures ne sont que des schématisations du véritable paysage d'énergie : il faut imaginer ces mêmes hypersurfaces dans des espaces de dimensions

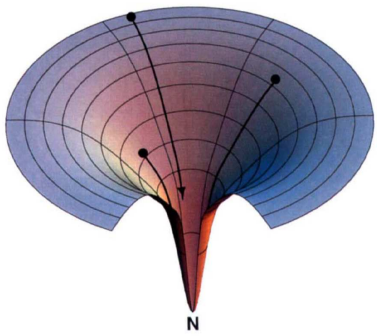


FIG. 1.25: un paysage en forme d'entonnoir permet d'accélérer le repliement des molécule ; tous les degrés de liberté évoluent de manière concertée vers l'état natif.

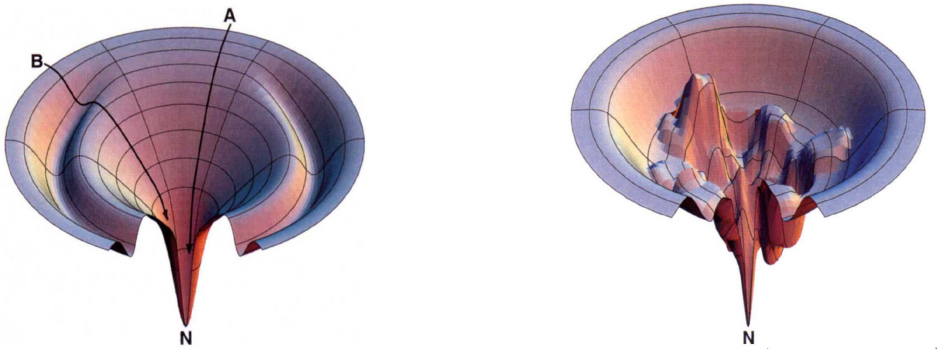


FIG. 1.26: des modèles d'entonnoirs non-parfaits permettent d'expliquer les différentes dynamiques observées (relaxations multi-exponentielles, dynamiques lentes ou rapides), ainsi que l'existence de structures métastables intermédiaires (globules fondus).

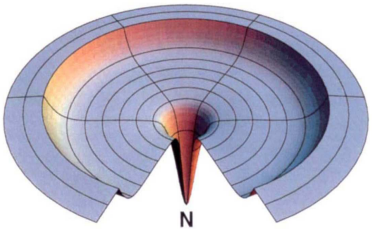


FIG. 1.27: ce paysage d'énergie présente un état natif énergétiquement favorable en compétition avec un *ensemble* d'états entropiquement favorisés.

bien supérieures (10, 100, 1000, 10000). Pour aborder le problème de la représentation et tenter de reproduire fidèlement un paysage réel, plusieurs auteurs ont proposé des solutions.

Sauf à choisir un nombre restreint (une ou deux) de variables représentatives (Schug *et al.*, 2005b), la solution retenue consiste à rassembler les états afin d'obtenir un ensemble discret (§ 1.3.3 page 37) permettant une représentation où les états sont reliés selon les barrières énergétiques qui les séparent (figure 1.28 reprises de Frauenfelder et Leeson, 1998). Les figures 1.29 et 1.30 (extraites de Krivov et Karplus, 2004) exposent le principe avec trois bassins principaux (A, B et C) eux-même composés de clusters de conformations préférentielles (points noirs), puis un cas réel avec une protéine formant deux feuillets β en épingle.

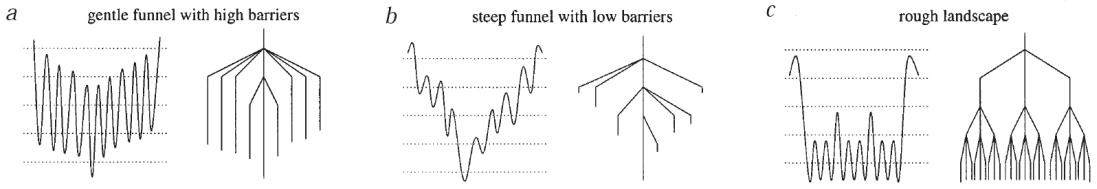


FIG. 1.28: principe de représentation des états les plus échantillonnés.

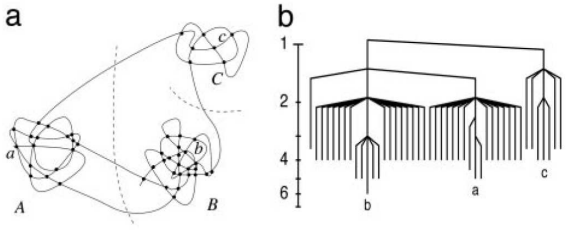


FIG. 1.29: exemple avec trois états, présentant chacun plusieurs minima.

Ce type de graphes a été amélioré afin de prendre en compte l'importance de chacun des puits (énergie et entropie), ce qui permet une compréhension accrue des chemins de repliement et met en évidence les points d'embranchement des différents minima (Rylance *et al.*, 2006).

Enfin, Gfeller *et al.* (2007) proposent d'illustrer le paysage par un graphe pondéré des états (les poids correspondant aux probabilités de Boltzmann), clusterisés selon un critère de similarité et dont les arcs sont établis en fonction des transitions qui s'opèrent au cours de simulations de dynamique moléculaire (figure 1.32).

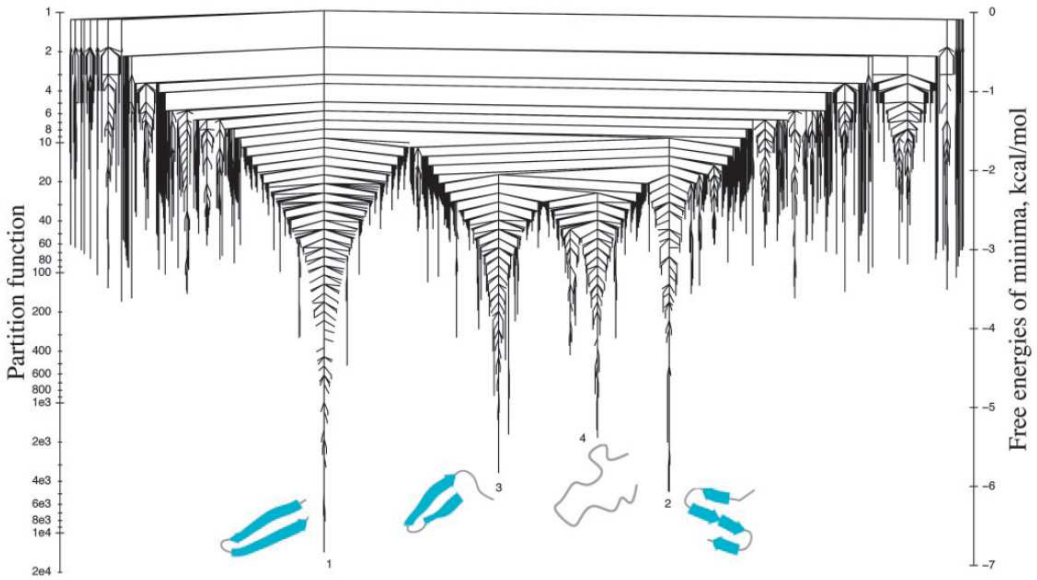


FIG. 1.30: cas réel avec l'épingle β de la protéine G (l'abscisse n'a pas sens physique), figure extraite de Krivov et Karplus (2004).

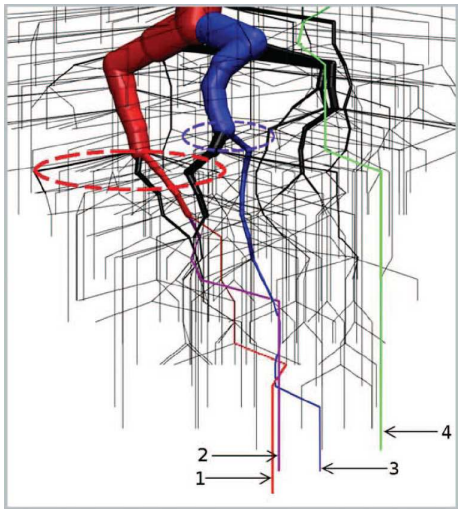


FIG. 1.31: chemins de repliement et largeur des bassins d'attraction d'une protéine ; en rouge, violet, bleu et vert sont respectivement représentés les quatre premiers minima (tirée de Rylance *et al.*, 2006).

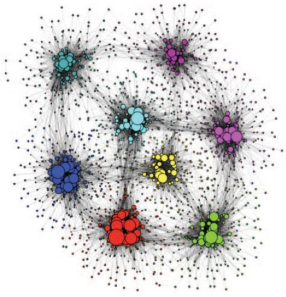


FIG. 1.32: représentation des états visités et des transitions par un graphe pondéré (extrait de Gfeller *et al.*, 2007).

1.3.4.3 Dans quelles conditions la molécule se replie-t-elle ?

> La santé est un état d'équilibre instable, qui comporte bien des oscillations.
>
> Maurice Halbwachs, Les causes du suicide

Tout d'abord, comme nous l'avons évoqué à la section 1.3.1, l'environnement chimique est déterminant pour le repliement des structures. Ainsi, *in vivo*, les molécules sont généralement dans l'eau, mais se replient au cours de leur synhèse et sont parfois aidées par des chaperones. *In vitro*, la molécule peut être étudiée dans différents solvants et à différentes températures, afin d'observer sa dénaturation. Enfin, lorsqu'elles sont en très forte concentration, nous avons vu que la dénaturation n'était pas toujours réversible.

Autrement dit, les molécules biologiquement actives sont dans un état d'équilibre qui n'est en général que localement stable ; ce qui nous fait adhérer à l'affirmation de Halbwachs.

Cependant, même si certaines protéines ne se replient pas dans le même état selon l'environnement, on continue à croire que les petits sous-éléments de structures (secondaires) restent, eux, relativement bien conservés malgré les conditions parce que relativement stables. De même, jusqu'à une certaine taille (plusieurs milliers d'atomes), la dénaturation des petites molécules reste réversible.

De plus, même si, d'après Boltzmann (équation (1.2)), tous les puits de potentiel seront peuplés selon leurs énergies et entropies, les temps de commutation d'un état à un autre peuvent être d'un ordre de grandeur supérieur aux temps biologiques.

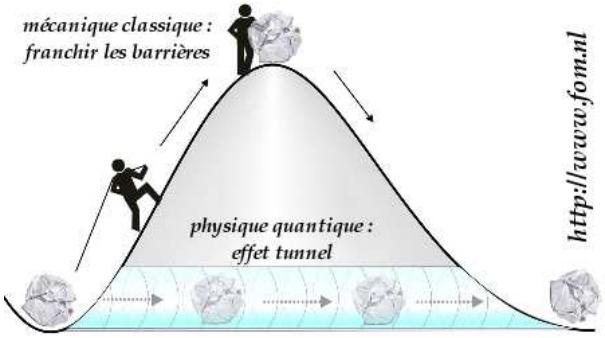


FIG. 1.33: certains phénomènes quantiques, comme les franchissements de barrières de potentiel, se différencient complètement des comportements prédits par la mécanique classique.

1.3.4.4 Interconversions et temps d'attente

Même si le temps nécessaire à la transition entre deux états est en réalité extrêmement court (quelques femtosecondes), c'est le temps moyen d'attente dans chacun des états qui affecte la dynamique (voir figure 1.34).

Alors que la formule de Boltzmann (équation (1.2), page 34) détermine les niveaux de population asymptotiques de chacun des puits selon leurs énergies libres, il existe des estimateurs pour les temps moyens de transition d'un état A à un état B ($\tau_{A\to B}$), basés sur des modèles probabilistes (équation (1.18)).

$$\tau_{A\to B} = \beta h \times \exp\left(\beta \Delta G\right), \tag{1.18}$$

où ΔG est la hauteur de la barrière : $\Delta G = G_{\max} - G_A$ et h est la constante de Planck ($\beta h \approx 16$ fs à $T = 300$ K).

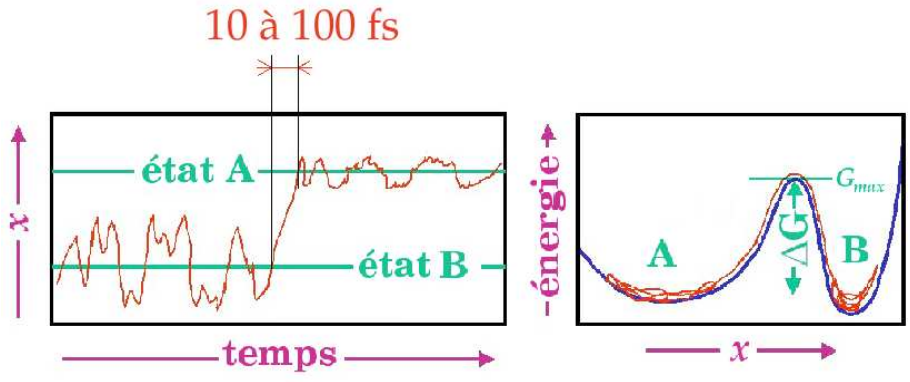


FIG. 1.34: exemple de trajectoire d'un système échantillonnant deux états avec son paysage d'énergie correspondant.

Enfin, notons que, dans les cas plus complexes, la rugosité du paysage d'énergie influence les temps de repliement des molécules (Chavez *et al.*, 2004).

1.3.4.5 Un repliement hiérarchisé

Dans ce paragraphe, nous tentons de donner des éléments de réponse à la question : « Qu'est-ce qui casse l'apparente complexité (§ 1.3.4.1) du repliement des molécules ? ». Cette question est primordiale, autant pour la compréhension du processus de repliement lui-même que pour pouvoir anticiper ou répondre aux difficultés que posera l'étape de modélisation.

1) Les chemins préférentiels. La notion de « *chemins préférentiels* » (figure 1.24 droite) subodore l'existence, non seulement d'un état prépondérant, mais également d'itinéraires énergétiquement favorables qui permettent de drainer *efficacement* — tel un entonnoir — la molécule vers sa conformation native. Toutefois, étant donné le nombre de degrés de liberté, il faut réinterpréter la figure 1.25 : certaines *grappes* de variables évoluent rapidement et de manière concertée vers des sous-éléments structuraux.

Une idée qui a été introduite dernièrement et qui étaye cette hypothèse, est celle des *contacts non-natifs* — c'est-à-dire non-définitifs et absents de la structure finale — qui apparaissent au cours du repliement et qui pourraient accélérer la convergence vers la géométrie native. Dans le principe, la molécule (ou bien un motif) ne *diffuse* plus dans son espace de phase en dimension N_{ddl}, mais évolue dans un sous-espace (ou une sous-variété) de dimension inférieure.

Cependant, la faiblesse des interactions entrant en jeu dans les processus de repliement fait qu'il existe de multiples chemins menant d'une géométrie quelconque à la géométrie native (Zhou et Karplus, 1999). Notons également que la présence de ces « faux-contacts » accroît la difficulté de la prédiction *in silico* des conformations moléculaires (Paci *et al.*, 2002).

2) La hiérarchie des structures (primaire jusque quaternaire) reflète peut-être une hiérarchie du repliement qui permet cette fois une réduction drastique de la complexité. Les éléments de structure secondaire (hélices, tournants) se forment relativement rapidement et simultanément, tandis que, sur une échelle de temps plus longue, se forme la structure tertiaire où les éléments locaux se positionnent les uns par rapport aux autres. Enfin, l'ensemble achève son repliement pour former les complexes finaux. En d'autres termes, on imagine que les degrés de liberté sont plus

ou moins indépendants lorsqu'ils sont topologiquement éloignés (au moins dans les premières phases du repliement).

Supputons cela par un calcul innocent : s'il faut un temps $\tau(N)$ pour replier une petite molécule de taille N, il est possible qu'il ne faille pas un temps $\tau(N)$ pour une grande molécule de taille N, mais plutôt un temps qui évoluerait en $\alpha\tau(M)+\beta\tau\left(\frac{N}{M}\right)$ où M est la taille moyenne d'un motif et $\frac{N}{M}$, le nombre attendu de ces motifs ; le facteur α traduit l'écart autour de la valeur moyenne ($\alpha < M$) et β représente un facteur d'échelle traduisant le rapport des temps caractéristiques entre les échelles au niveau des atomes et au niveau des structures secondaires (évoluant comme M ou M^γ avec $1 \leq \gamma < 2$). Le repliement des motifs se faisant en parallèle, on peut donc s'attendre à des repliements encore plus courts.

La figure 1.35 donne l'évolution du temps de repliement en fonction du nombre N de degrés de liberté (τ, supposée exponentielle, est en rouge). Le cas vert correspond à des motifs d'une dizaine de degrés de liberté, le cas bleu correspond à un niveau hiérarchique supplémentaire où les motifs s'arrangent en *super*-motifs de taille 10×10 avant de se compacter dans la géométrie finale.

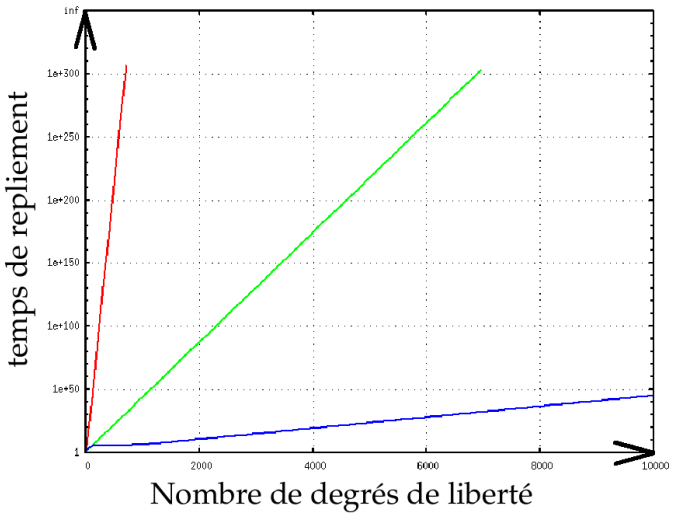


FIG. 1.35: en hiérarchisant le processus de repliement, on peut réduire drastiquement la complexité.

Les approches dites « *divide and conquer* » exploitent clairement cette idée afin de concevoir de nouvelles stratégies de recherche opérationnelle. Citons Takahashi (1999) dans le domaine de l'échantillonnage conformationnel, où un premier algorithme est en charge de détecter les régions prometteuses de l'espace de phase, tandis qu'un second algorithme optimise localement les géométries ainsi proposées.

Remarque : ces deux premières approches sont antagonistes (Zhou et Karplus,

1999) puisque l'une préconise l'existence d'interactions non-natives qui disparaissent par la suite tandis que l'autre suppose la formation immédiate des motifs structuraux présents dans la conformation finale. La question n'est pas vraiment résolue à l'heure actuelle (Baldwin et Rose, 1999) et des expériences sur des molécules différentes montrent des résultats différents. Ainsi, il a souvent été fait référence dans la littérature sur le repliement des protéines, à un « *collapsus hydrophobe* » au cours duquel les résidus hydrophobes s'effondreraient sur eux-même, formant un noyau compact (Mok *et al.*, 2007). Dans cette conception, la structure tertiaire devance la formation des structures secondaires, qui n'apparaissent qu'ultérieurement au collapsus. Au contraire, de nombreuses expériences et simulations ont pu montrer l'existence d'intermédiaires de repliement et d'états de transition (*globules fondus*) qui étayent plutôt la thèse du repliement hiérarchique (Honeycutt et Thirumalai, 1990; Mu *et al.*, 2006).

3) Des séquences pas tout à fait aléatoires... Enfin, notons que l'élaboration d'une séquence protéique (ou d'ARN ou d'ADN), qui est le fruit d'une évolution darwinienne comptant des milliers d'essais et d'échecs, a subi une double pression de sélection, puisque d'une part, la *fonction* chimique est importante pour assurer la pérennité, mais d'autre part, que le *temps de repliement* peut aussi devenir un facteur pénalisant (Dobson, 2003). Ainsi, les séquences de la Nature ne sont pas complètement aléatoires, mais vérifient le critère de posséder une fonction *et* de pouvoir l'acquérir dans un temps biologiquement acceptable (milliseconde - seconde).

Parfois, la fonction prime sur le temps de repliement. Ainsi, il a été fait référence à une protéine dont les temps de repliement et la stabilité auraient été accrus en mutant certains acides aminés, mais qui, dans ce cas, perdent leur fonction biologique (Jäger *et al.*, 2006).

1.4 Quelles méthodes existe-t-il pour l'observer ?

> The human observer, whom we have been at pains to keep out of the picture, seems irresistibly to intrude into it
>
> Rosenfeld, 1965

La variété des méthodes expérimentales ne couvre pas encore l'ensemble des questions que l'on se pose sur la structure et la dynamique des molécules. Pour

aborder physiquement une molécule, les obstacles sont nombreux et, pour entrer dans son intimité, les scientifiques ont dû

- gérer des échelles de taille de l'ordre de l'angström (10^{-10}m),
- gérer un nombre de 1 (de plus en plus d'expériences sur molécules uniques) à $N_A = 6 \times 10^{23}$ molécules,
- gérer des échelles de temps allant de la femtoseconde (10^{-15}s) à la seconde, en particulier, le rapport entre les temps de mesure et les temps caractéristiques des phénomènes est particulièrement important pour l'interprétation,
- gérer l'aspect dynamique : vibrations, mouvements, chemins de repliement souvent multiples, diffusion...

Nous terminons ce chapitre en citant deux articles : le premier, d'où la figure 1.36 est extraite, provient de Dobson *et al.* (1998) ; il illustre schématiquement les différents éléments pouvant être observés dans la molécule. Le deuxième, plus récent, est dû à Sali *et al.* (2003) ; il présente les différentes approches expérimentales et computationnelles envisageables pour extraire des informations des molécules.

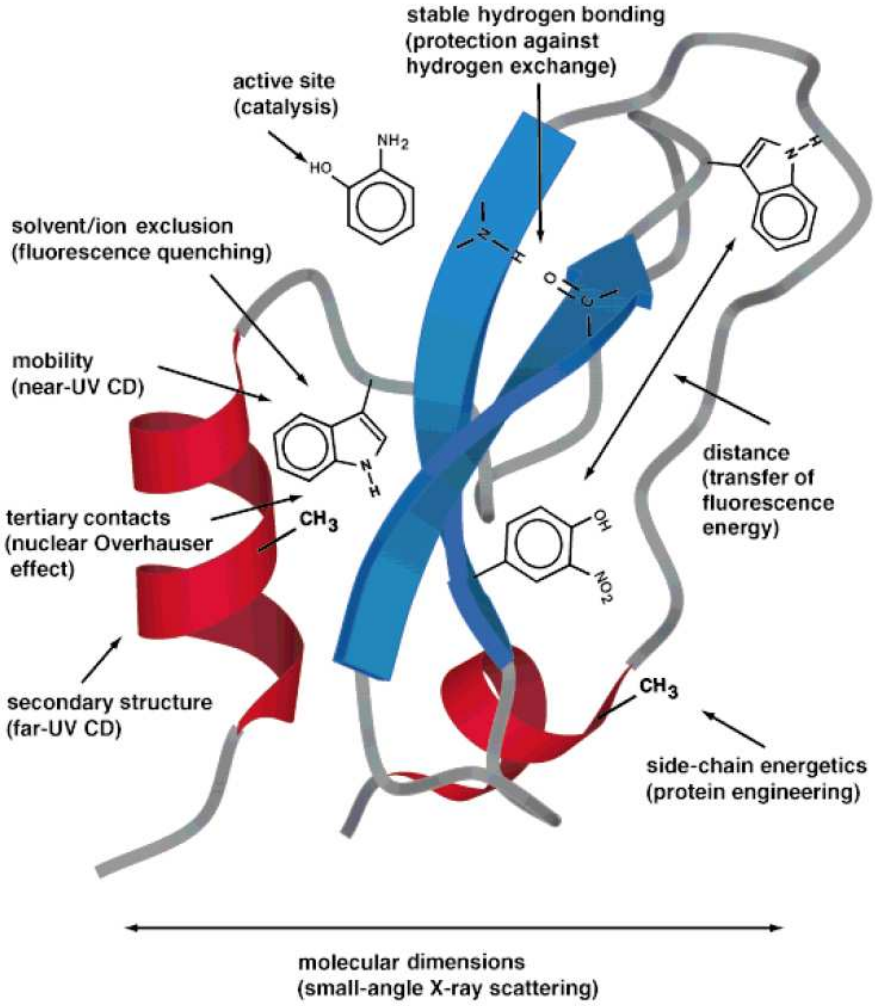


FIG. 1.36: différents éléments observables dans la molécule et méthodes expérimentales utilisées.

Chapitre 2

La modélisation moléculaire

2.1 Introduction

> In a peculiar way that scientists are still trying to understand, nature can easily solve a problem — how to fold proteins into the proper configuration — that eludes the most powerful computers and the most powerful minds.
>
> George Johnson, Designing Life : Proteins 1, Computer 0, The New York Times

Après avoir rappelé les aspects purement chimiques qui vont nous intéresser à propos des molécules, nous nous plaçons maintenant dans une approche computationnelle de ces molécules et considérons les descriptions possibles, tant en ce qui concerne les données chimiques, que l'encodage de la flexibilité et que l'estimation de l'énergie.

Ainsi, nous verrons comment importer une molécule générique dans l'ordinateur, quelles solutions ont été proposées afin de saisir l'essentiel de la flexibilité tout en limitant la complexité et verrons enfin les principes généraux des champs de forces qui permettent d'estimer l'énergie potentielle interne à la façon des approches newtoniennes. Enfin, nous définirons et cernerons la problématique étudiée et donnerons les principales hypothèses de travail ; ces précisions nous permettent de donner un cadre formel à la modélisation moléculaire.

2.2 Comment intégrer la molécule *in silico* ?

Afin d'encoder la structure primaire de la molécule dans un format informatique, il faut tout d'abord sauvegarder la liste des atomes et leurs types, ainsi que le graphe des liaisons et les ordres correspondants. Cela correspond à la structre primaire de la molécule. Certains auteurs se sont arrêtés à cette description, notamment pour la conception d'algorithmes très rapides, traitant des bases de données très fournies : c'est l'objet du § 2.2.1.

Si l'on désire élever le niveau de description aux aspects géométriques, il faut compléter ces données topologiques par les coordonnées cartésiennes. Plusieurs méthodes sont possibles pour stocker cette information : codage absolu (§ 2.2.2), relatif (§ 2.2.3) ou codage des distances interatomiques (§ 2.2.4), selon l'approche de Crippen et Havel (1988).

Certains auteurs simplifient le problème dans le cas particulier des protéines et réduisent la description des acides aminés à une seule entité unifiée (§ 2.2.5). Enfin, la dernière simplification possible après cela est de négliger les types particuliers de ces acides aminés-boules, pour ne garder que l'information sur leurs natures hydrophobes ou polaires (§ 2.2.6).

2.2.1 Les approches topologiques

Devant la forte complexité que représente la reconstruction de la géométrie d'une molécule, certains chercheurs ont tenté de court-circuiter cette étape en élaborant des prédicteurs de l'activité chimique des molécules sur la base de leurs structures topologiques. Bien entendu, ces predicteurs n'ont pas la fiabilité des approches tridimensionnelles, cependant, la nécessité de cribler très rapidement d'immenses bases de données de composés pharmaceutiques exclut immédiatement l'approche géométrique. En réalité, comme nous le verrons, ces deux approches sont complémentaires.

Ce point de vue *topologique* a donné naissance à une nouvelle matière de la chémoinformatique, appelée *QSAR* (pour Quantitative Structure-Activity Relationship). Elle repose sur une hypothèse simple, mais pas toujours vérifiée, qui veut que des structures proches aient des activités similaires. Si cette forme de « *continuité* » est généralement vérifiée pour les structures 3D, l'affirmation est plus délicate pour les structures topologiques dont la similarité n'entraîne pas toujours la similarité géométrique.

Néanmoins, si les approches QSAR ne permettent qu'une sensibilité médiocre

(voir table 2.1), leur force est dans le nombre restreint de faux positifs (bonne spécificité : $C \approx 0$ et $\frac{A}{A+C} \approx 1$). Ceci permet d'écarter rapidement de nombreux composés, qui auraient dû être synthétisés et testés sur paillasse sans ce premier filtrage. Le temps gagné en laboratoire s'est traduit par un appui important des entreprises pharmaceutiques, cependant, la dissémination des méthodes et la mutualisation des moyens (informatique et bases de données) sont plus que restreintes.

	Prédite active	Prédite inactive	Validité
Active	vrai positif (A)	faux négatif (B)	sensibilité $= \frac{A}{A+B}$
Inactive	faux positif (C)	vrai négatif (D)	-
Validité	spécificité $= \frac{A}{A+C}$	-	

TAB. 2.1: types d'erreurs de prédiction.

2.2.2 Les coordonnées cartésiennes

Ayant vu rapidement les approches topologiques QSAR, nous présentons maintenant les descriptions géométriques.

La plus simple façon de coder la conformation d'une molécule consiste à mémoriser toutes les coordonnées cartésiennes de ses atomes. C'est l'approche la plus communément adoptée.

2.2.3 La description vectorielle

Toutefois, il peut s'avérer utile de faire un codage *relatif* des atomes, ce qui rend la description indépendante du référentiel qu'on se donne. C'est par exemple le cas de la description vectorielle, utilisée pour démontrer certains résultats théoriques sur le cyclohexane (Gathen et Gerhard, 2003) : on montre en effet que le cyclohexane possède deux conformations rigides dites « chaises » et une sous-variété de dimension 1 de conformations « bateau », voir figure 2.1.

2.2.4 L'analyse en distance ou *Distance Geometry*

L'approche par *distance geometry* développée par Blumenthal et Menger (1970) et formalisée par Crippen et Havel (1988) permet de coder différemment le pro-

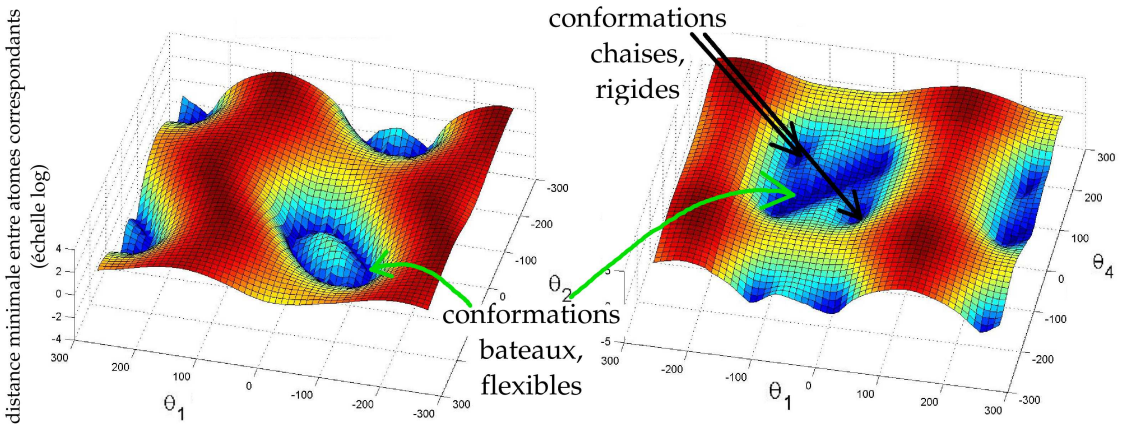


FIG. 2.1: étude du cyclohexane : (en z) distance entre atomes correspondants (s'annulant quand le cycle est fermé), tracée en fonction de différents degrés de liberté ; figure réalisée avec Matlab.

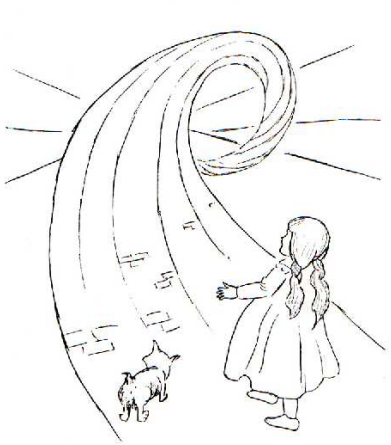


FIG. 2.2: "My goodness, Toto, I don't think we're in $\mathbb{R}^n$ anymore !", tirée de (Crippen et Havel, 1988).

blème afin de l'aborder sous un angle différent. Elle propose la reconstruction de la géométrie 3D à partir des distances interatomiques.

Ce type d'algorithmes a été très utilisé afin de remonter des données expérimentales indirectes aux véritables structures moléculaires. Il permet de ne retenir, des coordonnées atomiques, que les $N_{\text{atomes}}(N_{\text{atomes}} - 1)/2$ distances interatomiques, ce qui a donné lieu à certaines heuristiques de recherche originales.

Cette approche géométrique, ainsi que celle de Gathen (2003) permettent, au moins théoriquement, certaines résolutions exactes, comme celle du cyclohexane (figure 2.3).

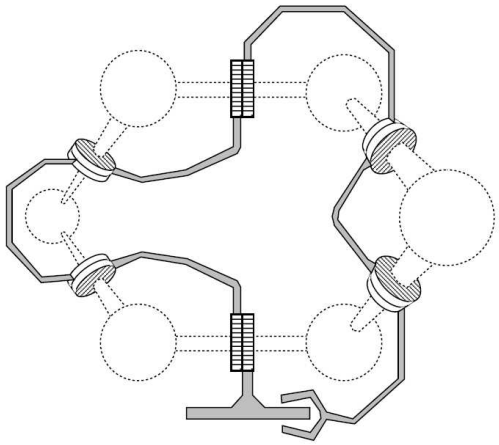


FIG. 2.3: le cyclohexane, vu comme un robot à six degrés de liberté (extrait de Nikitopoulos 2001).

2.2.5 La description « résidus unifiés »

D'autres méthodes existent, afin d'« alléger » le problème, qui consistent à fusionner chaque acide aminé en une seule entité indivisible. C'est le cas des approches dites par *résidus unifiés* dans le cas des protéines (Huang *et al.*, 1995; Hoffmann et Knapp, 1996; Liwo *et al.*, 1999; Pillardy *et al.*, 2001). Elles substituent les acides aminés par des billes ou des ellipsoïdes (voir figure 2.4) et utilisent alors un champ de forces adapté[1] et/ou moyenné sur les degrés de liberté omis.

Pourtant, cette approche n'est pas adaptée au niveau de précision que nous recherchons, mais concerne plutôt les études de repliement global de grandes protéines dont elle implémente en quelque sorte, le concept de repliement hiérarchique

[1] par exemple comme les champs de forces ff (anciennement parm) implémentés dans CHARMM (Brooks *et al.*, 1983).

(voir 1.3.4.5) ; de plus, elle est nécessairement restreinte au cas des protéines.

Remarquons également que, si les atomes unifiés sont « latéraux », ils ne sont pas subsidiaires et les réarrangements des chaînes latérales sont primordiaux dans le repliement global de la protéine et dans ses interactions avec d'autres acteurs (Najmanovich *et al.*, 2000). Ainsi, une étude récente du laboratoire a démontré que des différences minimes dans la chaîne latérale d'une valine, entre les ligands cyclosporine A et son homologue pharmaceutique Debio-025, déterminaient l'interaction ou non, de la cyclophilline B avec la calcineurine[2].

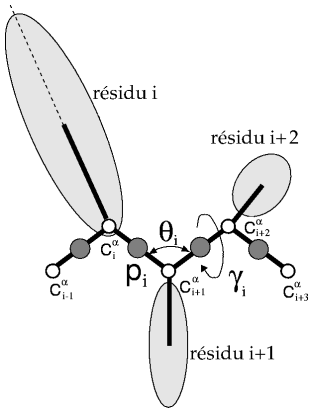


FIG. 2.4: unification de chaque résidu en un seul solide, représenté par un ellipsoïde.

2.2.6 Le modèle « hydrophobe-polaire » sur grilles 2D et 3D

Cherchant toujours à simplifier, certains auteurs ont même proposé de concevoir la protéine comme un collier de perles (chaque perle représentant un acide aminé) s'inscrivant dans une grille (en deux ou trois dimensions selon les études). Chaque perle est localisée sur une intersection de la grille et est séparée d'un pas de la perle qui la précède et de celle qui la suit. Dernière hypothèse : l'effet hydrophobe[3] domine tous les autres effets et dicte seul le repliement des protéines. Le but est alors d'incruster le collier sur la grille de sorte à maximiser le nombre de contacts hydrophobes entre les résidus, tout en évitant de positionner deux résidus sur le même noeud.

Il existe des résultats sur le bien fondé d'une telle restriction aux caractéristiques d'hydrophobie et d'hydrophilie (Huang *et al.*, 1995), cependant, ce type de simplifications engendre une forte perte d'information et nécessite des approfondissements.

[2]article soumis

[3]les résidus polaires sont plus stables au contact du solvant, contrairement aux résidus hydrophobes, voir § 2.4.2.4.

Il peut toutefois constituer une première étape d'exploration des conformations du squelette protéique tandis qu'une deuxième étape devrait raffiner la structure en prenant en compte les chaînes latérales. Notons surtout qu'il transforme le problème initial en un cas scolaire de combinatoire, ce qui permet d'avancer quelques conclusions théoriques sur la complexité (Hart et Istrail, 1995; Crescenzi *et al.*, 1998) et sur la caractérisation du paysage d'énergie (Baldwin et Rose, 1999; Bryngelson *et al.*, 2004) qui offre également un exercice générique palpitant pour les méthodes d'optimisation classiques : Monte Carlo et algorithmes génétiques (Unger et Moult, 1993b; Khimasia et Coveney, 1997), Monte Carlo séquentiel i.e. couplé aux chaînes de Markov (Grassberger, 2004), paradigme des fourmis (Shmygelska et Hoos, 2003, et 2005), etc.

2.3 Comment décrire la flexibilité des molécules ?

Ayant explicité le codage des informations topologiques et géométriques, nous passons maintenant à une étape de compréhension de la molécule en décrivant les diverses façons de saisir sa flexibilité, autrement dit, quels sont les degrés de liberté qui permettent de modeler sa géométrie.

La plus évidente est d'autoriser chacun des N_{atomes} atomes à bouger indépendamment des autres, dans toutes les directions (§ 2.3.1). Toutefois, un certain nombre de propriétés géométriques sont relativement bien conservées au cours du temps — telles que les longueurs et les angles de valence — de sorte qu'il est possible de restreindre les degrés de liberté au seules angles de torsions des liaisons interatomiques (§ 2.3.2). Nous montrerons que cette description allège considérablement la complexité, tout en captant l'essentiel de la flexibilité moléculaire.

Enfin, notons que la description « hydrophobe-polaire » sur grille 2D et 3D, donne lieu à un codage particulièrement simple des conformations moléculaires : la géométrie de la protéine est encodée par une chaîne de caractères qui indique, à chaque acide aminé, si le squelette « tourne » à gauche, à droite, en haut, en bas, ou continue dans la même direction[4].

2.3.1 Codage absolu et relatif des coordonnées cartésiennes

On peut coder en absolu les $3N_{\text{atomes}}$ coordonnées cartésiennes atomiques, c'est ce que font la majorité des auteurs : citons par exemple Goto et Osawa (1989, 92,

[4]pour une grille 2D, trois cas subsistent : gauche, droite, tout droit

93) et Braden (2002), ainsi que la majorité des logiciels de modélisation moléculaire : CHARMM (MacKerell *et al.*, 1998), Accelerys (Accelerys, 2005), etc.

Cette description offre une approche au plus près de la réalité où chaque atome subit de ses voisins des tensions et des répulsions. C'est aussi la plus simple, qui permette de coder indifféremment une ou plusieurs molécules et, en particulier, de simuler explicitement le solvant.

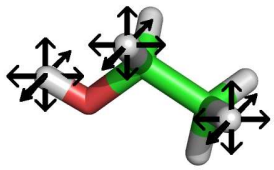


FIG. 2.5: tous les atomes peuvent se mouvoir indépendamment dans les trois dimensions.

Remarque : un codage relatif des positions atomiques permet d'accélérer l'intégration de certaines équations de la dynamique ou l'optimisation de conformations moléculaires de la façon suivante : lorsqu'une force est appliquée sur un atome, classiquement elle se propage dans la chaîne des atomes comme le long d'un ressort au fil des pas d'intégration. Un positionnement relatif des atomes offre un cadre idéal pour propager les contraintes de longueurs et d'angles de valence et ainsi d'accélérer le calcul.

2.3.2 Les degrés de liberté torsionnels

Par ailleurs, nous savons que les longueurs de liaisons, de même que les angles de valence, adoptent des valeurs plus ou moins standard en fonction des atomes et de l'environnement chimique (voir tableau 2.2, en aval, page 69). En utilisant ces tables de valeurs, on peut donc commencer à reconstruire la géométrie de la molécule, cependant, il manque encore une information : celle des valeurs d'angles de torsion.

Bien que l'on puisse encore trouver des statistiques[5], voir figure 2.6, il n'existe plus de tables de valeurs standards pour la raison que ces torsions sont relativement flexibles en comparaison des autres degrés de liberté. On voit ici que la flexibilité d'une molécule peut être, en grande partie, saisie par la description de ses angles de torsion. Ainsi, certains auteurs ont limité le nombre de degrés de liberté en adoptant

[5]dans le cas des protéines, on dispose des densités de probabilité empiriques des couples d'angles (ϕ, ψ) (torsions du squelette) par résidu : ce sont les statistiques de Ramachandran (1968)

une description torsionnelle ; parmi ceux-ci, nous pouvons par exemple citer Schulze-Kremer et Tiedemann (1994), Vieth *et al.* (1998a), Jin (1999), Day *et al.* (2002), Vengadesan et Gautham (2003), Schug *et al.* (2004 a et b).

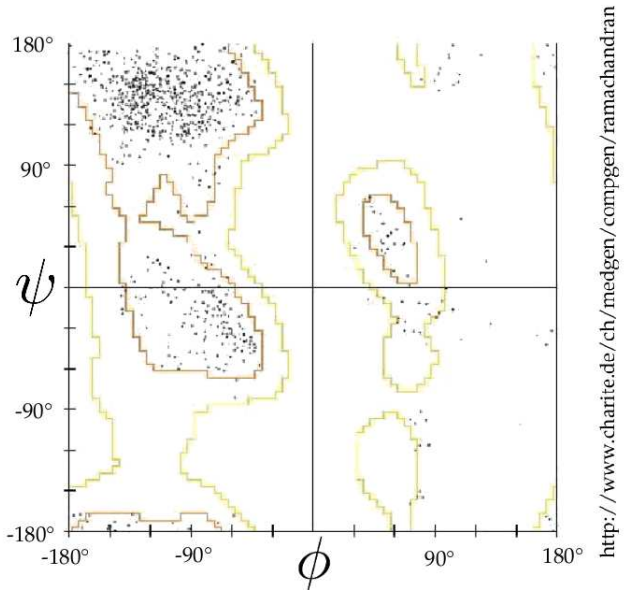


FIG. 2.6: répartitions *a postériori* des angles (ϕ, ψ) pour la conformation des squelettes protéiques.

Pour adopter une telle démarche, il faut stocker la liste des torsions actives et, pour chacune d'elles, connaître la liste des atomes mis en mouvement. De plus, certains degrés de liberté ont une période plus petite que 2π, il est donc utile de détecter les éventuelles symétries de la molécule (voir § 3.4.2.1).

Enfin, notons que dans le cas des cycles, la description torsionnelle pose un problème, puisque chaque degré de liberté est sensé mettre en rotation un ensemble d'atomes — soit à droite, soit à gauche de la liaison —, alors que dans un cycle, il n'est plus possible de faire cette distinction... La solution proposée est alors la suivante : soit le cycle est considéré comme un bloc rigide (aucun degré de liberté), soit l'utilisateur précise une liaison particulière qui sera *formellement* coupée (figure 2.7). Comme cette liaison existe toujours physiquement, il y a une pénalisation énergétique forte qui favorise les configurations telles que les distances et les angles de valence soient proches des valeurs standards. Cet artéfact permet d'aborder l'optimisation de la géométrie des cycles de la même manière que le reste de la molécule.

Nous pouvons, par un rapide calcul, estimer le gain en complexité, apporté par une telle démarche : dénombrons les liaisons utiles... Sur l'ensemble des molécules

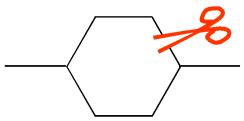


FIG. 2.7: la coupure formelle d'une liaison permet d'échantillonner les conformations du cycle.

étudiées, la valence moyenne des atomes est environ[6] $< V > \approx 2,33$; une molécule ayant N_{atomes} atomes compte donc *a priori* $N_{\text{atomes}} \times < V > /2$ liaisons. Parmi celles-ci, $6,7\%$ participent à une liaison multiple, et $49,5\%$ impliquent un atome d'hydrogène (insensible aux rotations autour de son unique liaison de valence). Finalement, le nombre de degrés de liberté N_{ddl} escompté sera au maximum égal à :

$$\begin{aligned} N_{\text{ddl}} &\leq (1 - 6,7\% - 49,5\%) \times \frac{< V >}{2} N_{\text{atomes}}, \\ &\leq 0,51 \times N_{\text{atomes}}, \end{aligned} \tag{2.1}$$

à comparer aux $3N_{\text{atomes}}$ degrés de liberté en coordonnées cartésiennes, soit un gain d'un facteur 6 environ.

Remarque : l'inéquation (2.1) constitue une majoration ; cette borne est atteinte dans le cas de la « cyclodextrine » qui est une des molécules que nous avons traitées ($\frac{N_{\text{ddl}}}{N_{\text{atomes}}} = 0,49$), cependant, dans le cas général, on a plutôt l'encadrement $0,23 \leq \frac{N_{\text{ddl}}}{N_{\text{atomes}}} \leq 0,30$ (voir graphique 2.8).

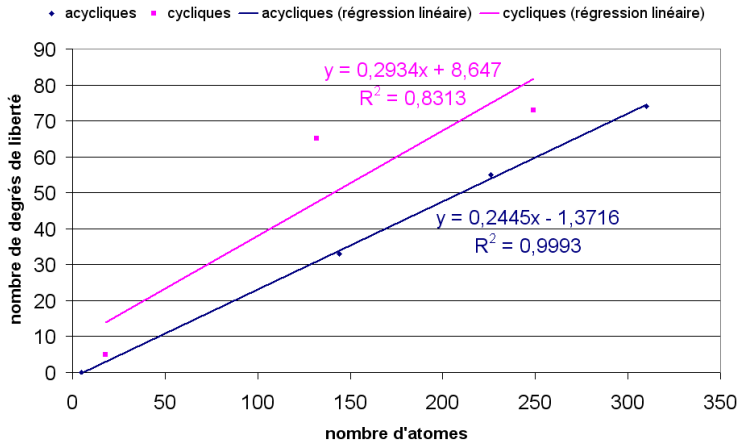


FIG. 2.8: évolution du nombre de degrés de liberté en fonction du nombre d'atomes.

Les méthodes par résidus unifiés, qui rejettent les degrés de liberté des chaînes latérales jugés peu influents, simplifie nettement le problème et offre un gain d'un

[6] comptée avec les ordres de multiplicité

facteur[7] 2,3 environ. Elles permettent donc d'aborder des protéines de plus grandes tailles.

Les hybridations. Nous avons présenté succintement les outils possibles pour décrire la flexibilité des molécules mais les solutions envisageables ne sont ni figées ni cloisonnées. Ainsi, certains auteurs ont utilisé les avantages de plusieurs approches en les hybridant entre elles.

Nous avons maintenant achevé les deux premières étapes : celle de codage des informations topologiques et géométriques des molécules, puis celle de compréhension chimique de la molécule avec la mise en évidence de ses degrés de liberté. Nous allons voir maintenant comment cette géométrie, modelée par ces degrés de liberté, définit différents niveaux d'énergie.

2.4 Le hamiltonien moléculaire

L'approximation de Born et Oppenheimer permet de découpler l'équation de Schrödinger électronique de l'équation atomique. Le *calcul quantique ab initio* permet de reconstruire la fonction d'onde électronique sur la base des coordonnées atomiques et ainsi d'estimer précisément l'énergie interne de la conformation (Miller, 2005). À l'inverse, certains chercheurs ont développé des méthodes de champ de forces utilisant des fonctions simplifiées et paramétrées *a posteriori* afin de reproduire certaines données empiriques[8]. Ces méthodes donnent accès à des estimateurs — moins précis mais plus legers à manier — de l'énergie en fonction des coordonnées atomiques (Jorgensen et Tirado-Rives, 2005).

Enfin, certaines méthodes[9] sont purement empiriques et proposent d'utiliser un score de *fitness* défini sur la base de connaissances expérimentales sur un ensemble de petits peptides (typiquement, des bases de rotamères, voir Shetty *et al.*, 2003). Les géométries conformes aux densités de probabilités *a posteriori* des angles de torsion sont alors favorisées (Dill *et al.*, 1996; Canutescu *et al.*, 2003). Cette hypothèse de travail repose sur l'observation que les nombreuses analogies de séquences entre les différentes protéines connues sont (beaucoup) plus fréquentes que ne l'auraient été les similarités dans un ensemble purement aléatoire de séquences[10].

[7] en pondérant par les fréquences de chacun des acides aminés.

[8] les méthodes semi-empiriques offre un intermédiaire où certaines intégrales du calcul *ab initio* sont estimées par des fonctions paramétrées expérimentalement.

[9] concernant principalement les protéines

[10] En effet, il n'existe pas moins de 20^{15} séquences possibles de peptides d'une quinzaine d'acides

Nous avons adopté une approche d'estimation par un champ de forces qui permet des temps de calculs nettement réduits.

Parmi les différents éléments constituant la molécule, nécessaires au calcul de son énergie interne, on distingue les liaisons et angles de valence, les torsions et les paires d'atomes non liés (figure 2.9).

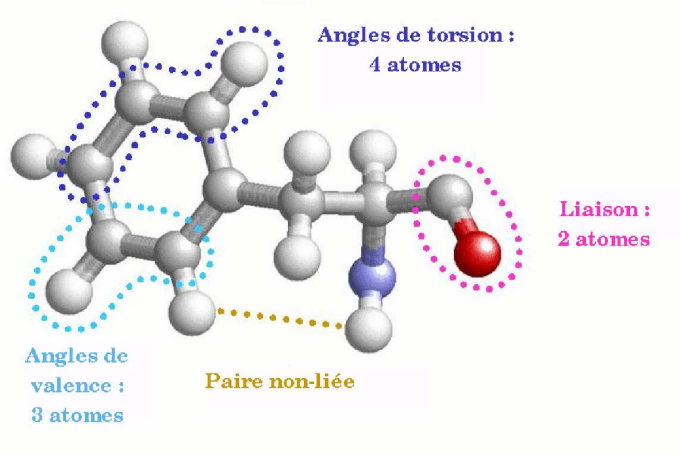


FIG. 2.9: éléments nécessaires au calcul de l'énergie interne de la molécule.

2.4.1 Contributions dominantes

L'approche de modélisation des interactions par un champ de forces semi-empirique tel que le CVFF (Hagler *et al.*, 1974; Hagler et Lifson, 1974), permet d'interpréter toutes les contraintes physico-chimiques en termes de contributions énergétiques dont les niveaux définissent la stabilité des conformations.

Les différents champs de forces fournis dans la littérature (ou vendus) reprennent plus ou moins la même philosophie (Jorgensen et Tirado-Rives, 2005) ; nous détaillons ci-après, à titre d'aperçu, les modèles des contributions qui constituent le champ de forces que nous avons utilisé : le *Consistent Valence Force Field* (CVFF).

Chaque contribution du champ de forces intervient avec des paramètres internes et des coefficients de pondération dépendant des atomes impliqués et de leur environnement. Ces paramètres sont estimés sur la base de données expérimentales concernant un jeu de molécules limité, aussi, chaque champ de forces se distingue par des contributions décrites par des fonctions particulières et un ensemble de paramètres qui lui est propre.

---

aminés, pris parmi les 20 acides aminés naturels

2.4.1.1 Les énergies de valence

Les liaisons de covalence résultent de, ou plus exactement *formalisent*, la mise en commun d'orbitales électroniques de deux atomes. Elles peuvent être de différents ordres (simples, doubles ou triples) selon le nombre d'orbitales mises en commun. Le cas le plus courant, celui des liaisons simples, offre en particulier une assez bonne flexibilité de rotation autour de l'axe portant les deux atomes (figure 2.10), tandis que les liaisons multiples ne présentent que deux états de torsion stables dits *cis* et *trans* (lorsque les quatre atomes voisins sont dans un même plan), séparés par des barrières de potentiel très prononcées (figures 2.12 et 2.13).

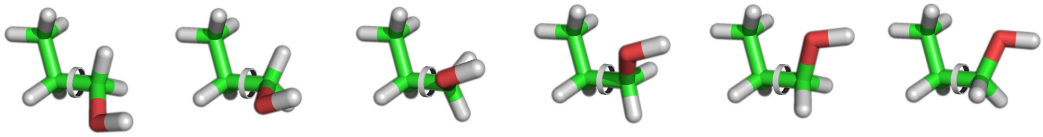


FIG. 2.10: les liaisons simples offrent un degré de liberté torsionnel permettant de modeler localement la géométrie de la molécule.

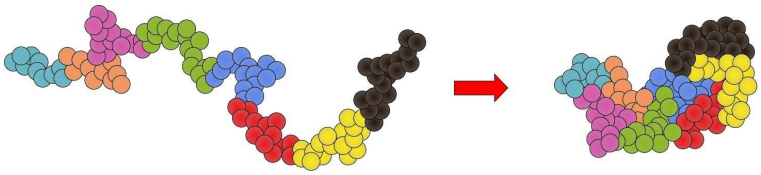


FIG. 2.11: en jouant sur les degrés de liberté torsionnel, on peut modeler la géométrie moléculaire.

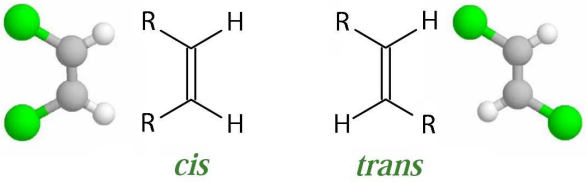


FIG. 2.12: conformations *cis* et *trans* d'une double liaison, les boules vertes représentant des groupements quelconques.

Comme nous l'avons fait remarquer (§ 2.3), les longueurs de ces liaisons, $d_{1,2}$ (entre l'atome 1 et l'atome 2), sont relativement bien conservées et ne dépendent que du contexte chimique des atomes impliqués. Les tables des valeurs standards font partie du bagage des connaissances empiriques des chimistes, qui assimilent généralement ces liaisons à l'image intuitive d'un ressort mécanique entre les deux

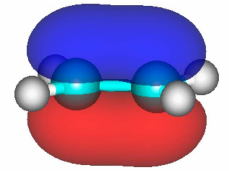


FIG. 2.13: la rigidité des liaisons multiples est issue de la présence d'orbitales supplémentaires.

atomes. Ce modèle est repris dans la méthodologie des champs de forces, en introduisant un potentiel harmonique décrivant la déformation des liaisons selon un modèle de type *masse/ressort* (figure 2.14 et équation (2.2)).

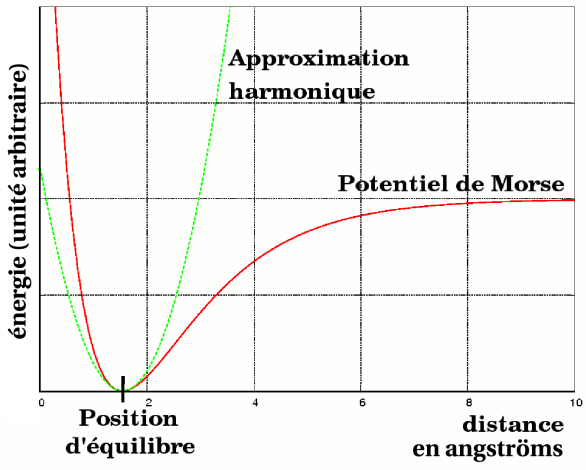


FIG. 2.14: modélisation des interactions de liaisons : potentiel de Morse (rouge) et potentiel harmonique (vert).

$$E_{\text{liaison}} = K_\ell \left(d_{1,2} - \ell^0\right)^2 . \tag{2.2}$$

En pratique, les très faibles déviations autour de ℓ^0 justifient le modèle harmonique et l'estimation des deux constantes est faite de manière à reproduire au mieux les données expérimentales, cependant, il existe d'autres modèles prenant en compte le profil complet des énergies de déformation (potentiel de Morse, figure 2.14, potentiel de Hook, etc.). K_ℓ et ℓ^0 sont fonctions des atomes 1 et 2 et de l'ordre de la liaison (voir tableau 2.2 pour des exemples).

Les angles de valence. Les orbitales libres ou liantes tendent à occuper l'espace autour des atomes, de sorte à être les plus éloignées possibles les unes des autres (règle de Gillespie, voir figure 2.15). Les angles entre les liaisons covalentes $\theta_{1,2,3}$

Type de liaison	longueur de la liaison en Å	énergie de dissociation en kcal.mol^{-1}
C−N	1,47	73,6
C−O	1,43	86,0
C−H	1,09	98,7
C−C	1,54	83,2
C=C	1,33	146,7
C≡C	1,20	200,5

TAB. 2.2: longueurs de liaisons de covalence pour différents types atomiques

— définis par trois atomes — oscillent également autour de valeurs nominales déterminées expérimentalement et sont, encore une fois, modélisées par un potentiel harmonique (équation (2.3)).

FIG. 2.15: La règle de Gillespie impose la valeur des angles de valence.

$$E_{\text{angle}} = K_{\text{a}}(\theta_{1,2,3} - \theta^0)^2. \tag{2.3}$$

Remarquons que ces constantes dépendent maintenant des trois types atomiques intervenant dans l'angle de valence. De plus, les constantes de raideurs sont plus faibles que dans le cas des liaisons, ce qui signifie que des déformations locales sont envisageables si elles permettent un réarrangement globalement favorable.

Les constantes de raideurs de ces ressorts (liaisons et angles de valence) sont telles que les forces et les fréquences de vibration dominent toutes celles des autres contributions. Ainsi, les algorithmes de simulation de dynamiques moléculaires doivent nécessairement adopter un pas d'intégration de l'ordre de la femtoseconde pour être plus rapides que la période de vibration des liaisons. À moins que le simulateur ne permette de propager des contraintes et ainsi conserver fixes les longueurs de liaison et les angles de valence : voir par exemple les algorithmes de dynamiques moléculaires SHAKE (Van-Gunsteren et Berendsen, 1977) et RATTLE (Andersen, 1983).

2.4.1.2 Les énergies non covalentes

Les contributions que nous avons vues jusqu'ici mettent en scène des atomes voisins dans la topologie de la molécule, elles définissent donc une première catégorie d'interactions, par opposition aux interactions entre atomes non voisins, que nous abordons maintenant.

Interactions coulombiennes. D'une part, certains atomes de la molécule peuvent être chargés électriquement, d'autre part, les différences de charges et de masses entre les noyaux atomiques entraînent différents niveaux d'*électronégativité* (capacité à attirer à soi les électrons d'une liaison) ; ceci fait apparaître une polarisation de la liaison et donc des charges partielles dans la molécule (figure 2.16).

FIG. 2.16: polarisation des liaisons : (gauche) électronégativités égales, liaison apolaire, (doite) différents niveaux d'électronégativité impliquent un déplacement du doublet liant.

La présence de ces charges se traduit par des interactions coulombiennes, attractives ou répulsives selon le signe des charges et dont l'énergie potentielle s'exprime sous la forme :

$$E_{\text{Coulomb}} = \frac{\delta_1 \delta_2}{4\pi\varepsilon_i d_{1,2}}, \tag{2.4}$$

où $d_{1,2}$ est la distance entre les deux sites interagissant, ε_i est la constante diélectrique du solvant[11] et $\delta_{1,2}$ les valeurs de charges.

En réalité, une évolution du terme E_{Coulomb} en $1/d_{1,2}^2$ a été utilisée afin de prendre en compte l'hypothèse d'une dépendance linéaire de ε_i en fonction de la distance entre les atomes impliqués. Cette approximation permet essentiellement de s'affranchir de la racine carrée dans le calcul de $d_{1,2}$.

Par exemple, l'eau, H_2O, est une molécule polaire, la charge partielle négative étant placée sur l'atome d'oxygène et la charge partielle positive répartie sur les deux atomes d'hydrogène, ce qui leur permet de se lier entre elles par des liaisons dites *ponts hydrogène*. Alors qu'une liaison covalente nécessite une centaine de kcal.mol^{-1} pour être rompue, de telles *ponts* ne requièrent que quelques kcal.mol^{-1} (voir tableau 2.3 pour des ordres de grandeur). Les ponts hydrogène expliquent que l'eau soit liquide à la température ambiante, alors que le méthane (molécule apolaire la

[11] Dans le vide, $\varepsilon_i = \varepsilon_0 \approx 8,85$ C^{2}J^{-1}m^{-1}

plus simple) est gazeux. La formation des ponts hydrogène, natifs ou non, est donc un élément important dans l'étude des structures moléculaires.

Type de pont hydrogène	énergie en kcal.mol^{-1}
O-H$\leftrightsquigarrow$N	6,9
O-H$\leftrightsquigarrow$O	5,0
N-H$\leftrightsquigarrow$N	3,1
N-H$\leftrightsquigarrow$O	1,9
HO-H$\leftrightsquigarrow$OH$_3^+$	4,3

TAB. 2.3: énergies impliquées dans les ponts hydrogène

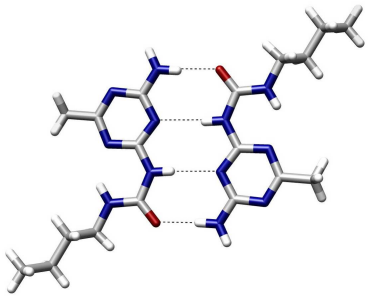


FIG. 2.17: les ponts hydrogène jouent un rôle important dans la dynamique et la stabilité des molécules (ici en pointillés).

Les termes de Van der Waals. Cette contribution comporte deux effets (figure 2.18) : l'un, attractif, peut s'interpréter grâce aux inductions électromagnétiques entre les dipôles, qui apparaissent suite aux faibles fluctuations au sein des nuages électroniques. L'autre est répulsif et modélise la très grande énergie qu'il est nécessaire de fournir pour tenter d'interpénétrer deux nuages électroniques (équation (2.5)).

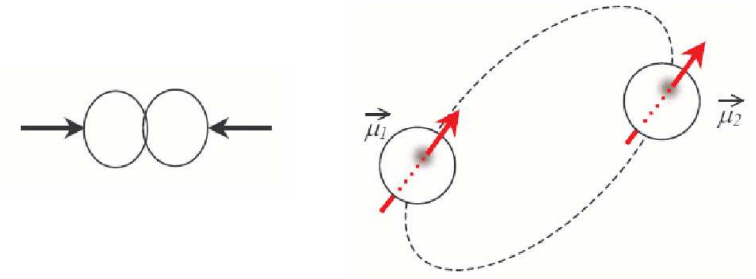


FIG. 2.18: (gauche) recouvrement d'orbitales électroniques impossible : fusion nucléaire à froid. (droite) induction des moments magnétiques.

$$E_{\mathrm{VdW}} = \frac{A}{d_{1,2}^{12}} - \frac{B}{d_{1,2}^{6}}. \tag{2.5}$$

Ce terme en $1/d^{12}$ joue essentiellement un rôle de « garde-fou » pour prévenir et interdire les recouvrements d'orbitales qui n'ont lieu que dans des conditions extrêmes[12] et ne découle pas de principes clairement formalisés, de sorte que l'exposant 12 est parfois remplacé par une autre valeur :

- 14 dans le champ de forces MMFF94 (Halgren, 1996),
- 9 dans le CFF (Maple *et al.*, 1994).

Les constantes A et B dépendent spécifiquement des types atomiques mis en jeu et ont été paramétrées sur la base de données expérimentales disponibles pour un ensemble représentatif de *petites* molécules (Hagler *et al.*, 1974; Hagler et Lifson, 1974). Le profil des contributions de Van der Waals est représenté figure 2.19 où l'on voit que les deux effets antagonistes définissent une distance optimale.

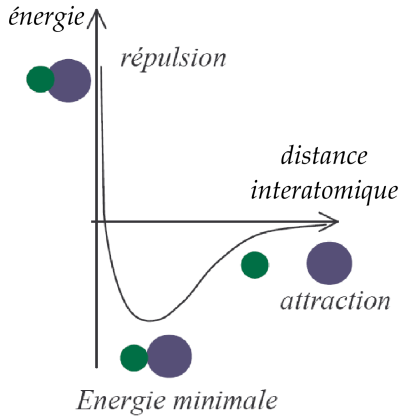


FIG. 2.19: évolution de la contribution Van der Waals en fonction de la distance interatomique

2.4.2 Modélisation, approximations et corrections : les autres contributions

Les différentes contributions que nous venons de voir définissent un premier estimateur de l'énergie interne de la molécule en fonction de ses coordonnées atomiques. Pour décrire entièrement le champ de forces CVFF, il faut néanmoins ajouter un dernier terme : le terme correctif de torsion, concernant les quadruplets d'atomes

[12] au cœur des étoiles et dans certains accélérateurs de particules.

consécutifs (§ 2.4.2.1). Ce terme est à part car il n'est pas issu de lois électromagnétiques ni de principes fondamentaux, il est simplement justifié par les meilleurs résultats empiriques qu'il permet d'obtenir...

Par ailleurs, la frontière entre énergies de valence et non covalente n'est pas aussi nette que ne le laisse sous-entendre la séparation en paragraphes précédente. En effet, les termes énergétiques non covalents ont des paramètres spécifiques lorsque les atomes impliqués sont en position dite « 1-4 », c'est-à-dire, lorsque les atomes sont séparés par exactement 3 liaisons.

Enfin, l'interaction de la molécule avec le solvant est déterminante mais très coûteuse à simuler explicitement. Aussi, verrons-nous (au § 2.4.2.2) qu'il existe des modèles continus permettant d'estimer une sorte d'effet moyen.

2.4.2.1 Les termes de torsion

Cette contribution traduit la modification de l'énergie lors de la rotation d'un fragment d'une molécule autour d'une liaison. Il s'agit d'un terme correctif, qui n'a pas d'interprétation théorique directe, mais qui se justifie par les résultats empiriques plus cohérents qu'il permet d'obtenir.

Chaque quadruplet d'atomes topologiquement consécutifs et non coplanaires définissent deux plans entre lesquels apparaît un angle dit de torsion (figure 2.20). Une liaison comporte donc plusieurs torsions. Le potentiel énergétique est alors donné par une formule empirique (équation (2.6)) :

$$E_{\text{tors}} = K_t \left(1 + \cos(n\phi - \phi^0)\right), \tag{2.6}$$

où ϕ est l'angle de torsion, ϕ^0, n et K_t sont des constantes. ϕ^0 prend les valeurs 0 ou π, n vaut 2, 3 ou 4 et K_t prend des valeurs relativement modestes : $|K_t| < 20\text{kcal.mol}^{-1}$.

2.4.2.2 Modéliser le solvant

L'environnement de la molécule est déterminant pour son repliement. La présence de solvant modifie les interactions au moins de deux manières (hydrophobie et ponts hydrogène) et, bien qu'il puisse être modélisé explicitement en simulant toutes les molécules, certains modèles existent qui permettent une prise en compte implicite et beaucoup moins coûteuse en temps de calcul[13] : ce sont les modèles de solvant continu. L'obtention d'un tel modèle se fait en moyennant sur toutes les positions

[13] remarquons que certains auteurs omettent simplement le solvant afin d'économiser le temps de calcul (Takahashi *et al.*, 1999)

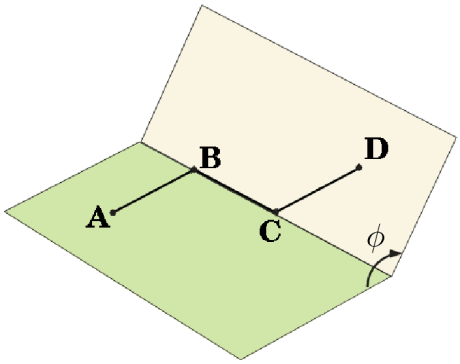


FIG. 2.20: Tout quadruplet d'atomes consécutifs : (A, B, C, D) forme un *angle de torsion* entre les plans (A, B, C) et (B, C, D)

possibles des molécules d'eau, c'est pourquoi on parle parfois de « potentiel de champ moyen » ou PMF (*Potential of Mean Force*).

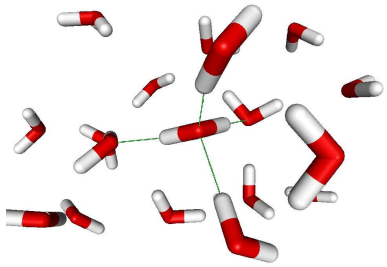


FIG. 2.21: simulation des molécules d'eau et de la formation des ponts d'hydrogène.

Parmi les modèles de solvants continus, le plus répandu dans les simulations de repliements des protéines est certainement le modèle de Born généralisé (ou GB, voir Still *et al.*, 1990). Il est moins coûteux en temps de calcul, mais moins précis que des solveurs de type Poisson-Boltzmann (Honig et Nicholls, 1995). Cependant nécessite malgré tout le calcul d'intégrales telles que les surfaces accessibles au solvant. Enfin, nous avons opté pour un modèle encore plus simple, précédemment développé et implémenté par Horvath (1997) que nous détaillons ci-après.

Il existe des études comparatives des approches par solvant explicite et implicite, notamment celles de Zhou *et al.* de 2002 et 2003. Celle de 2003 met en évidence les défauts pouvant apparaître lors du couplage d'un champ de forces avec un modèle de solvant implicite et en particulier, l'apparition de minima non pertinents dans le paysage d'énergie potentielle. On continue à croire toutefois que les modèles implicites, en limitant les frictions, ont plutôt tendance à lisser le paysage. Ainsi, citons les travaux de Tsui *et al.* (2000), où le solvant implicite a permis d'accélérer grandement la convergence (facteur 20) et ceux de Millar (1997) et Williams (1999) dans

lesquels les simulations explicites ont échoué à prédire le repliement vers un état correct alors que le modèle implicite y est parvenu.

L'impact de l'approximation continue n'est donc pas clair. Ni tout à fait néfaste, ni tout à fait bénéfique sur les résultats, elle permet néanmoins de réduire considérablement les temps de calculs et offre des simulations plus reproductibles. Nous adoptons ce modèle, mais garderons à l'esprit les éléments discutés ici.

2.4.2.3 La désolvatation

L'eau est un solvant particulièrement polaire ; de fait, les molécules d'eau tendent à s'organiser autour des groupements polarisés. L'avantage énergétique dû au repliement de la molécule doit donc contrebalancer l'énergie nécessaire à l'exclusion des molécules d'eau enfouies au cœur de la molécule. Le terme de *désolvatation*, basé sur un modèle de solvant continu développé par Horvath (1997), pénalise l'arrivée d'un atome de volume V_2 au voisinage (distance $d_{1,2}$) d'un atome de charge Q_1, par un terme évoluant en

$$E_{\text{Desolv}} = K_{\text{D}} \frac{Q_1^2 V_2}{d_{1,2}^4}. \tag{2.7}$$

Ainsi le solvant tend à limiter la portée des effets électromagnétiques, c'est ce qu'illustre la figure 2.22. Alors que dans le vide, il existe toujours une force attractive entre les charges de signes opposés, dans l'eau, le scénario est différent : en écartant les deux charges, on franchit une barrière énergétique lorsqu'il devient possible d'introduire des molécules de solvant.

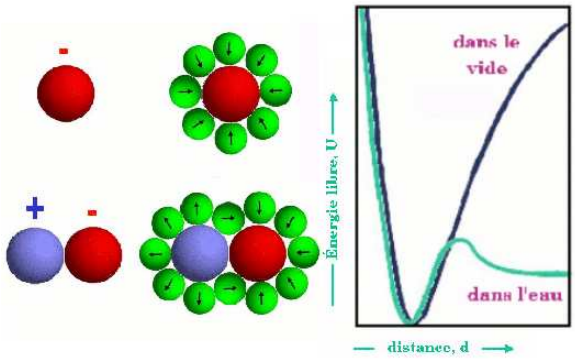


FIG. 2.22: contribution énergétique de la solvatation des groupements polaires

Enfin, l'eau, en s'intercalant entre les groupements polaires de la molécule, stabilise les états intermédiaires et ainsi catalyse la rupture des ponts hydrogène (figure 2.23).

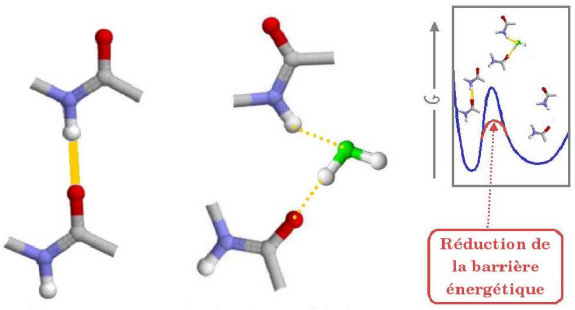


FIG. 2.23: Les molécules d'eau catalysent la rupture des ponts hydrogène

2.4.2.4 Les contacts hydrophobes

L'effet hydrophobe est d'une toute autre nature. Lorsqu'on simule explicitement toutes les molécules d'eau autour de la molécule d'intérêt, cet effet est occulté, mais dans la mesure où l'on cherche à approximer ces molécules discrètes et polarisées par un milieu continu, il faut tenir compte d'artéfacts notamment dus aux aspects dynamiques. Ainsi, les molécules d'eau, polarisées, sont beaucoup plus contraintes lorsqu'elles sont au voisinage de groupements apolaires qu'au voisinage de sites polarisés — elles sont en quelque sorte gelées. Or ce manque de liberté (on parle de *frustration*) traduit un rétrécissement dans l'espace de phase qui engendre une pénalisation entropique. Cet effet tend à rassembler les sites apolaires de la (ou des) molécule(s) pour former des clusters hydrophobes ; c'est aussi le phénomène qui explique pourquoi l'huile (hydrophobe) qui minimise la surface de contact avec l'eau, tend à ne former qu'une seule tâche circulaire à la surface de l'eau (figure 2.24).

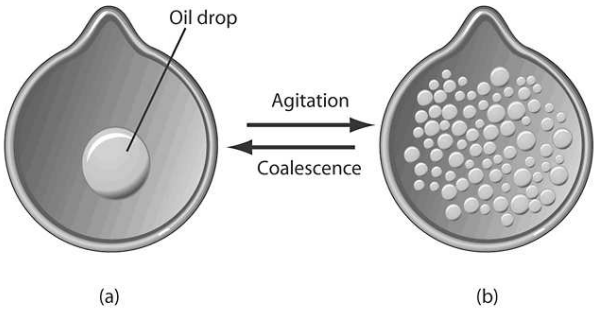


FIG. 2.24: l'effet hydrophobe tend à rassembler les éléments apolaires.

Cet effet apparaît comme une force uniquement parce qu'on cherche à moyenner sur toutes les positions du solvant.

Certains champs de forces prévoient un terme énergétique binaire selon que le contact est établi ou non, mais, pour éviter les effets indésirables des fonctions

discontinues, nous avons modélisé l'effet hydrophobe par une fonction continue de la distance interatomique d de la forme :

$$E_{\text{Hphob}} = K_H \min(0, d - 5), \tag{2.8}$$

qui est également un terme que nous avons ajouté aux termes classiques du CVFF.

2.4.2.5 Le lissage des singularités

La présence de singularités dans le paysage énergétique peut représenter un inconvénient majeur lors de l'implémentation et surtout lors de l'optimisation (instabilités). Or ces singularités apparaissent lorsque deux atomes se retrouvent exactement au même endroit ($d_{1,2} = 0$), ce qui n'est, finalement, pas plus aberrant que l'interpénétration des orbitales ($d_{1,2} \ll 1$). De plus, les atomes sont ici modélisés par des « billes », ce qui n'a plus de sens lorsque $d_{1,2}$ est petit car les particules sont délocalisées. Ceci a motivé la définition d'une nouvelle fonction « *distance*[14] » : $\mathfrak{d}_{1,2}$ ne s'annulant plus en 0 (équation (2.9) et figure 2.25).

$$\mathfrak{d}_{1,2} = \max\left[d_{1,2}; K_{\text{smooth}} + \left(1 - \frac{K_{\text{smooth}}}{3}\right) d_{12}\right]. \tag{2.9}$$

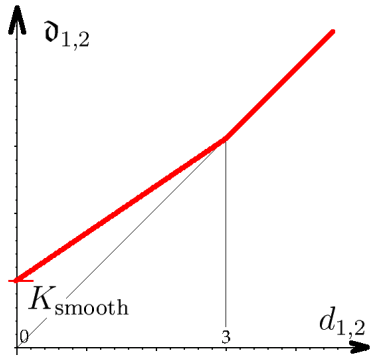


FIG. 2.25: écrêtage des singularités par redéfinition de la fonction distance.

Par ce biais, nous assurons la continuité de la fonction énergie sur le paysage global ; ce dernier étant compact[15] (au sens mathématique), nous sommes sûrs de l'existence d'au moins un minimum global dans le domaine.

[14] il ne s'agit plus d'une distance au sens mathématique
[15] ensemble borné et topologiquement fermé

2.4.2.6 La troncature des interactions à longues distances

Contrairement au nombre de liaisons de covalence, le nombre de paires d'atomes non liés évolue en $o(N^2_{\text{atomes}})$, ce qui fait des énergies non covalentes les plus gourmandes en temps de calcul. Aussi est-il courant de négliger les contributions impliquant des atomes plus éloignés qu'une certaine distance (ici 10Å).

Loncharich et Brooks (1989) ont comparé plusieurs méthodes de *cutoff* et montré que cette méthode donnait des résultats acceptables sur des simulations de dynamique moléculaire (voir également (Vieth *et al.*, 1998b)).

2.4.3 Résumé des contributions et exemple

Les différentes contributions à l'énergie totale passées en revue ci-dessus interviennent toutes avec des coefficients de pondération et des paramètres internes. Comme signalé plus haut, l'estimation de ces constantes est réalisé de sorte à reproduire au mieux les données expérimentales observées sur un jeu de petites molécules à l'équilibre :

- la cristallographie par rayons-X donne la position moyenne des atomes de la molécule quand elle est sous forme solide,
- la RMN permet aussi d'avoir indirectement de telles informations en solution, mais l'effet de moyenne sur l'ensemble de Boltzman peut introduire des erreurs de part l'anharmonicité du potentiel,
- le calcul quantique *ab initio* donne accès à de (très) bonnes approximations de l'énergie, de son gradient et de son Hessienne, en tout point de l'espace de phase,
- les spectres de vibrations fournissent les valeurs propres de l'Hessienne aux voisinages des points d'équilibre,
- l'analyse thermodynamique de données macroscopiques permet d'extraire des informations sur le paysage d'énergie telles que les températures de mixture, la stabilité des minima, les niveaux d'entropie, etc.

De fait, ce champ de force n'est pas exact — il n'est qu'une somme de modèles des véritables phénomènes quantiques — et, comme nous allons voir au chapitre 3, il devra être remis en question lorsque les molécules traitées seront plus grandes et/ou hors de leur point d'équilibre. En particulier, il repose essentiellement sur des paires atomiques et ne prend qu'implicitement en compte les dipôles ou multipôles d'ordres supérieurs.

Pour l'instant, cette fonction énergie peut être considérée comme une boîte noire, renvoyant pour toute conformation d'entrée une valeur de sortie que l'on cherche à minimiser.

La figure 2.26 récapitule les différentes contributions intervenant dans le calcul de l'énergie. La figure 2.27 présente le profil énergétique du butane en fonction de son angle de torsion central. Les conformations « *décalées*[16] » sont moins énergétiques que les conformations dites « *eclipsées* », mais la conformation opposant les deux groupements méthyles apparaît encore plus stable. À droite, représentation en bâtonnets et en sphères de la conformation la plus stable.

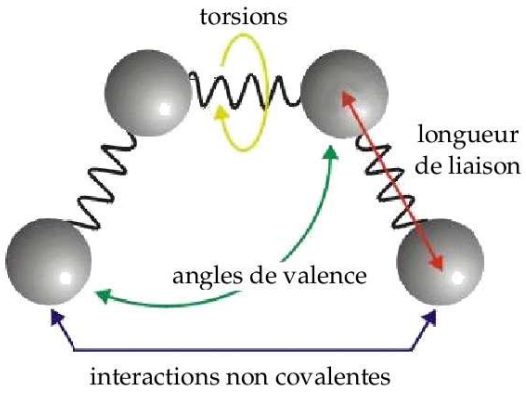


FIG. 2.26: résumé des différentes contributions intervenant dans l'estimation de l'énergie interne.

Considérons l'exemple du propane (figure 2.28) de dimension deux, qui nous permet de représenter son paysage énergétique comme une surface de $\mathbb{R}^3$: le paysage d'énergie (figure de droite) présente un minimum local et un minimum global.

2.4.4 Les champs de forces

La forme des fonctions utilisées dans le champ de forces varie en fonction du niveau de détail adopté. Ainsi, les champs de forces pour les petites molécules organiques diffèrent de ceux qui sont uniquement dédiés aux protéines, qui, eux-même, n'ont pas la même expression lorsqu'ils sont en *all-atom* ou en résidus unifiés. Comme nous l'avons dit, la détermination des *paramètres* de champs de forces est faite afin de reproduire différentes données expérimentales (Kosinsky *et al.*, 2004) ; là encore, différents jeux de paramètres sont obtenus selon l'ensemble de molécules utilisé. Parmi ces champs de forces, nous pouvons citer

[16] lorsque les substituants ne sont pas en vis-à-vis.

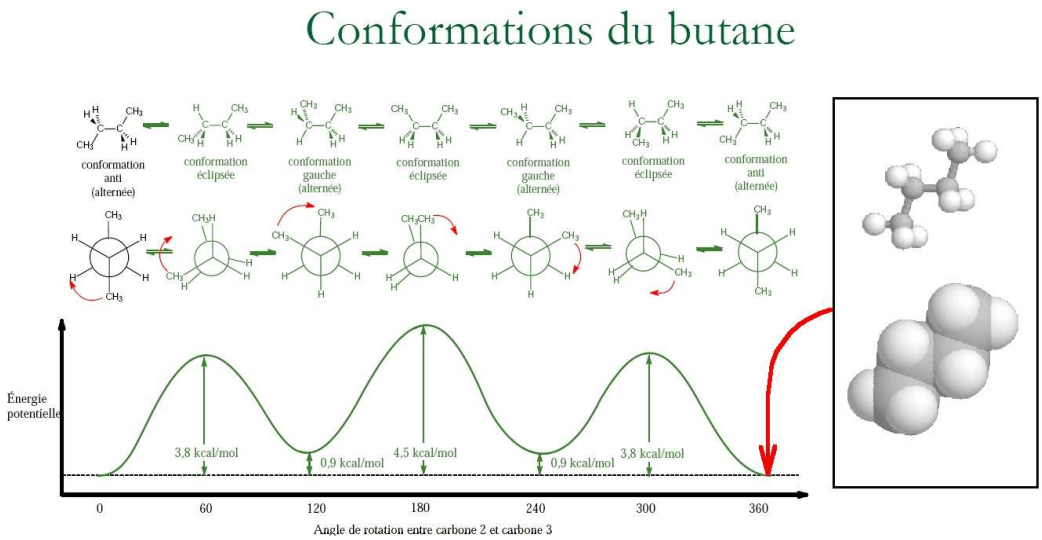


FIG. 2.27: (gauche) profil énergétique du butane. (droite) conformation la plus stable.

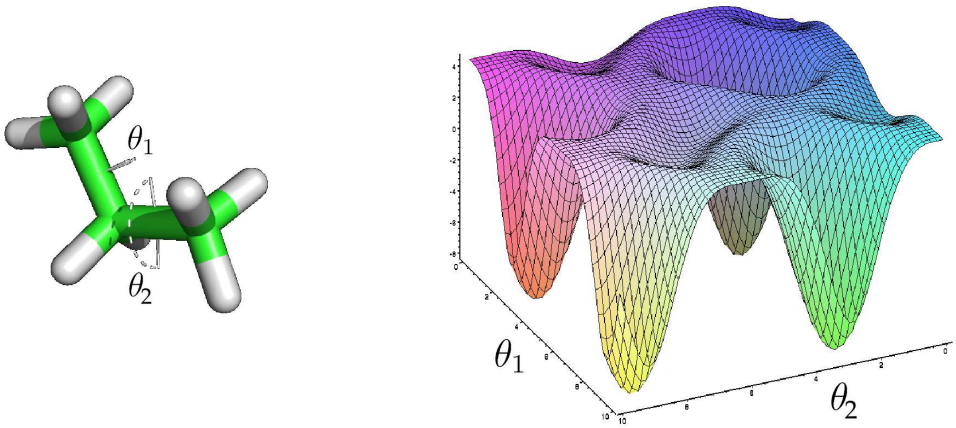


FIG. 2.28: la molécule de propane et son paysage d'énergie correspondant en fonction des deux degrés de liberté θ_1 et θ_2 (figures réalisées avec PyMol et Matlab).

- CVFF (Hagler *et al.*, 1974; Hagler et Lifson, 1974), que nous avons utilisé, et CFF (Maple *et al.*, 1994),
- PFF01 (Herges et Wenzel, 2004),
- MMFF (Halgren, 1996),
- MM2/3/4,
- GROMOS (Kutzner *et al.*, 2007),
- OPLS (Jorgensen et Tirado-Rives, 2005),
- CHARMM (Brooks *et al.*, 1983), paramétrisation, voir MacKerell *et al.* (1998),
- AMBER (Cornell *et al.*, 1995), délivrent ff94 (parm94), ff99 (parm99),
- ECEPP : atomique mais réservé aux protéines (Momany *et al.*, 1975),
- EEF1 (Lazaridis et Karplus, 1999; Krivov et Karplus, 2004),
- UNRES (Pillardy *et al.*, 2001), pour résidus unifiés.

La littérature fournit en outre un certain nombre de revues (Jorgensen et Tirado-Rives, 2005; Mackerell, 2004) dont certaines réalisent des comparatifs : (Hobza *et al.*, 1998; Varma, 2001). D'autres sont consacrées à la définition de fonctions de score pour le *docking* (Vieth *et al.*, 1998b).

2.5 La problématique et les hypothèses

Ayant codé les données du problème et décrit ses degrés de liberté, nous venons de présenter l'outil pour traduire le problème chimique en une question mathématique : une estimation de l'énergie. Nous précisons maintenant le cadre des études menées et décrites aux chapitres 3 et 4.

2.5.1 Quel algorithme cherche-t-on ?

Nous sommes intéressés par la modélisation des molécules dans un cas général (pas de restriction aux protéines ou autres cas particuliers) en vue de la prédiction des interactions et des affinités et, plus généralement, de l'estimation des propriétés macroscopiques. Pour cela, nous avons adopté une description statique (voir discussion § 2.5.3), à l'échelle atomique — soit une précision de l'ordre du picomètre — et n'avons donc pas considéré de simplifications de type « résidus unifiés » inadaptées aux échelles de taille des interactions et au traitement du cas général. Afin de pouvoir aborder des exemples réels de *docking* moléculaires, il a semblé judicieux *a priori* de considérer les seules hypothèses simplificatrices suivantes, qui optimisent le rapport précision/coût de calcul :

- remplacer le hamiltonien quantique des approches *ab initio* par un hamiltonien moléculaire de type champ de force : ici, le CVFF ;
- éluder la simulation explicite du solvant en approximant ses principaux effets par un modèle continu implicite ;
- décrire la flexibilité de la molécule par ses seuls degrés de liberté torsionnels.

Par ailleurs, l'ordre de taille des problèmes considérés correspond à celui de l'interaction d'un ligand organique (tout au plus quelques centaines d'atomes) avec un site actif de protéine (quelques milliers d'atomes), soient environ de 1 à 200 degrés de liberté.

Le logiciel est destiné aux biochimistes, souhaitant prédire, expérimenter ou valider des hypothèses et complémenter et interpréter leurs données expérimentales, tout autant qu'aux chimistes de l'industrie pharmaceutique désireux d'estimer les affinités et les activités potentielles de leurs ligands.

2.5.2 Une ou plusieurs molécules ?

La littérature différencie les études selon que sont traitées une ou plusieurs molécules. Pourtant, la première étape pour une simulation de *docking* est l'échantillonnage conformationnel ; et inversement, le *docking* peut être vu comme une généralisation des études de repliement où les degrés de liberté regroupent ceux des partenaires en scène et ceux de leurs positionnements relatifs. Toutefois, la complexité de chaque question spécifique a motivé le partitionnement en plusieurs domaines. Par ordre de complexité croissante, voici donc les prédictions possibles :

- l'échantillonnage conformationnel,
- les interactions de type site-ligand où une petite molécule (le *ligand* de quelques centaines d'atomes) se fixe telle une clef dans une serrure, dans un site actif d'une protéine ou d'un complexe,
- les mouvements plus amples de parties de protéines mettant à jour des sites actifs (voir figure 2.29),
- l'allostérie, où l'arrimage dans un site modifie la géométrie et l'activité globale de la molécule,
- les dimérisations où les deux acteurs et leurs surfaces d'interaction peuvent être de tailles plus importantes (Jin et Harrison, 2002, complexe calcineurine - cyclophiline),
- les multimérisations,
- les assemblages extrêmement complexes, tels les moteurs moléculaires (Elston

et al., 1998; Aksimentiev *et al.*, 2004), voir figure 2.30.

Bien entendu, la limite entre ces domaines est artificielle et dénote les différences d'approches computationnelles. En réalité, tout le continuum de complexité existe entre les extrêmes.

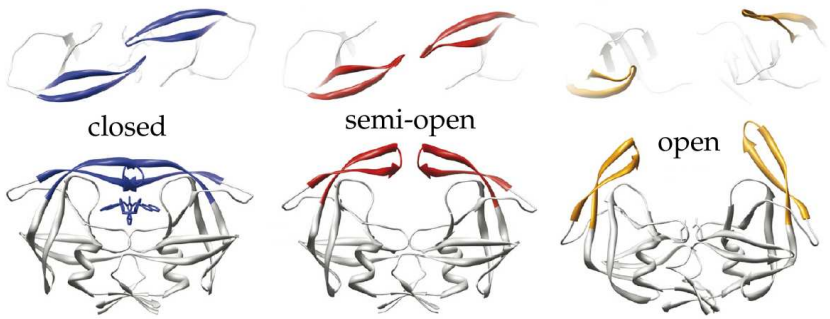


FIG. 2.29: la flexibilité de la protéase HIV-1 rend les simulations *in silico* difficiles (tirée de Hornak et Simmerling, 2007).

Ainsi, dans notre cas, le *docking* est vu comme une généralisation de l'échantillonnage conformationnel, ce qui justifie l'étude, dans une première phase (chapitre 3), d'une molécule unique.

2.5.3 Approches dynamiques $\mathcal{VS}$ statiques

Comme nous l'avons vu au chapitre précédent, la description statique, avec l'équation de Boltzmann (1.2) permet une caractérisation complète des niveaux de peuplement *asymptotiques* de chacun des états. Toutefois, l'approche dynamique est plus riche d'informations car elle donne accès aux « hauteurs des barrières énergétiques » et donc aux temps d'attente espérés dans chaque état, de même que les chemins de repliement (Snow *et al.*, 2005). Le lecteur peut consulter Karplus et Kuriyan (2005) pour une présentation des principaux concepts et des idées actuelles dans ce domaine et Iftimie *et al.* (2005) pour les dynamiques quantiques.

L'inconvénient de l'approche dynamique réside dans sa complexité accrue : en effet, les simulateurs, même s'ils reposent sur l'ergodicité des trajectoires (Tupper, 2005), ne peuvent espérer simuler plus qu'une centaine de microsecondes (Pande *et al.*, 2003) et souffrent de l'hétérogénéité des chemins de repliement.

Hornak et Simmerling (2003) ont adopté une démarche intermédiaire, appelée « *low barrier molecular dynamics* », combinant dynamique moléculaire et approche statique afin de situer et caractériser les états de transition. Ces états de transi-

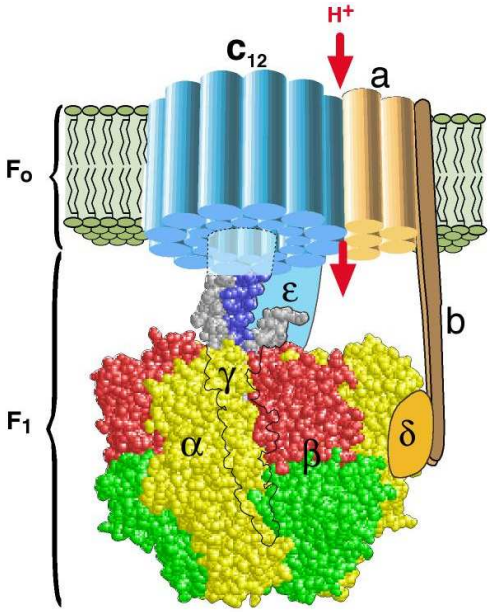


FIG. 2.30: moteur moléculaire : le potentiel hydrogène (pH) est converti en force motrice de rotation, entraînant le rotor central (douze hélices bleues) dans son stator (complexe F_1 et partie orange). Cette force génère des changements conformationnels cycliques et entraîne la synthèse d'ATP. Ce moteur fonctionne également dans le sens inverse, pompant les protons et consommant de l'ATP (figure extraite de Elston *et al.*, 1998, site : http ://www.soe.ucsc.edu/∼hongwang/ATP_synthase.html, consulté en août 2007).

tion sont particulièrement importants (Onuchic *et al.*, 1996; Shoemaker *et al.*, 1999; Baldwin et Rose, 1999) et peuvent permettre de localiser, même dans une approche statique, de nouveaux minima (Kolossvary et Guida, 1996).

Pour notre part, nous nous sommes restreints à une approche statique, mais avons cherché à caractériser la globalité de l'espace de phase en décrivant les principaux minima du paysage énergétique.

2.5.4 Que serait l'algorithme idéal ?

Afin de mieux comprendre la visée du présent travail, illustrons ce que serait un algorithme idéal... Car contrairement à l'idée souvent véhiculée, l'échantillonnage conformationnel et le *docking* ne doivent pas se limiter à déterminer *in silico* une structure tridimensionnelle des édifices moléculaires : comme nous l'avons vu au chapitre précédent, *tous* les minima peuplés du paysage d'énergie sont utiles à la compréhension de la fonction moléculaire. Le but de la modélisation moléculaire est donc de caractériser tous ces minima, en termes d'énergie, de volume et de forme des puits (on peut même chercher à caractériser les états de transition, Baldwin et Rose, 1999, Shoemaker *et al.*, 1999b). En un mot, le but ultime est de reconstruire la densité de probabilité sur tout l'espace de phase. Une caractéristique chimique macroscopique, $\mathcal{C}$, s'obtient alors comme la moyenne pondérée (l'espérance mathématique) des caractéristiques de toutes les géométries possibles $C(\theta)$ (équation (2.10)).

$$\text{caractéristique} \quad \mathcal{C} = \mathbb{E}_p[C(\Theta)] = \int_{\theta\in\Omega} C(\theta)p(\theta)d\theta. \tag{2.10}$$

Il existe plusieurs barrières à cela dont la première est celle de la taille extraordinaire de l'espace de phase Ω. Les très nombreuses évaluations de l'énergie sur cet espace nous forcent à utiliser un modèle de champ de forces dont les approximations sont également un facteur limitant pour une bonne estimation. Comme cela ne peut donc être fait directement, on repense la densité de probabilité en s'inspirant de la méthode d'approximation des intégrales de Monte Carlo (équation (2.11)).

$$\int_\Omega f(x)p(x)dx \approx \frac{1}{N_{\text{éch}}} \sum_{x_i\in\mathcal{E}_p} f(x_i), \tag{2.11}$$

où $\mathcal{E}_p$ représente un échantillonnage de Ω selon la loi de probabilité de densité p et

$N_{\text{éch}}$ son cardinal. Autrement dit, on approxime la densité par :

$$p(x) \approx \frac{1}{N_{\text{éch}}} \sum_{x_i \in \mathcal{E}_p} \delta(x = x_i), \tag{2.12}$$

où $\delta(x = a)$ est la mesure de dirac en a.

Cette approximation est d'autant plus précise que l'échantillonnage $\mathcal{E}_p$ est important ($N_{\text{éch}} \to +\infty$). Et c'est sur cette base que reposent implicitement tous les algorithmes de modélisation moléculaire (d'où l'origine contrôlée de l'appellation : échantillonnage conformationnel).

2.5.5 Formalisation de l'échantillonnage conformationnel

Pour aller plus avant dans cette formalisation, remarquons qu'il n'est pas forcément possible d'obtenir un échantillonnage $\mathcal{E}_p$, représentatif de l'espace, selon une densité p connue *a posteriori*, c'est pourquoi on utilise une astuce de calcul afin d'utiliser d'autres lois de distribution π, dont la plus courante est la distribution uniforme sur tout Ω.

En posant g la fonction telle que $g(x) = \frac{p(x)}{\pi(x)} f(x)$ et en lui appliquant le théorème de l'équation (2.11), il vient :

$$\begin{aligned} \frac{1}{N_{\text{éch}}} \sum_{x_i \in \mathcal{E}_\pi} g(x_i) &\xrightarrow{N_{\text{éch}} \to +\infty} \int_\Omega g(x)\pi(x)dx \\ &\xrightarrow{N_{\text{éch}} \to +\infty} \int_\Omega f(x)p(x)dx. \end{aligned}$$

Autrement dit, cela revient à échantillonner l'espace selon une densité π que l'on maîtrise mieux et à pondérer les échantillons $f(x_i)$ par des *poids* ω_i définis de la manière suivante :

$$\omega_i = \frac{p(x_i)}{\pi(x_i)}, \tag{2.13}$$

$$\frac{1}{N_{\text{éch}}} \sum_{x_i \in \mathcal{E}_\pi} \omega_i f(x_i) \xrightarrow{N_{\text{éch}} \to +\infty} \int_\Omega f(x)p(x)dx. \tag{2.14}$$

Pour pouvoir appliquer une telle astuce, il faut s'assurer, dans les poids ω_i, que π ne s'annule pas là où p est non nulle... Autrement dit, le support de p doit être

inclus dans celui de π, ce qui nécessite un échantillonnage plus vaste[17].

L'approximation (2.12) devient alors :

$$p(x) \approx \frac{1}{N_{\text{éch}}} \sum_{x_i \in \mathcal{E}_\pi} \omega_i \delta(x = x_i). \tag{2.15}$$

Ainsi, la prédominance des minima pertinents du paysage énergétique est maintenant explicitement mise en évidence par la pondération par ces facteurs de Boltzmann, alors qu'elle est implicitement prise en compte lors de dynamiques moléculaires qui revisitent de nombreuses fois les états peuplés.

Enfin, remarquons que dans le cas d'une distribution uniforme,

$$\pi(x) = \frac{1}{\mathcal{V}_\Omega}, \tag{2.16}$$

où $\mathcal{V}_\Omega$ est le volume total de l'espace. Dans l'équation (2.14), le facteur $\omega_i / N_{\text{éch}}$ peut alors s'exprimer

$$\frac{\omega_i}{N_{\text{éch}}} = p(x_i) \frac{\mathcal{V}_\Omega}{N_{\text{éch}}}, \tag{2.17}$$

et $\mathcal{V}_\Omega / N_{\text{éch}}$ représente le volume élémentaire de l'échantillon, qu'il faut rapprocher du volume élémentaire dx dans les intégrales (équation (2.11)). L'échantillon x_i représentant un puits de potentiel est donc caractérisé par son facteur de Boltzmann $p(x_i) = e^{-\beta E_i}/Z$, mais pondéré par le volume de ce puits, ce qui permet de faire le lien avec l'entropie introduite dans le premier chapitre.

2.6 Conclusion

Dans ces deux premiers chapitres, nous avons exposé les principaux éléments nécessaires à la compréhension de la problématique et à la justification de nos choix. Nous avons décrit la molécule et la façon de l'intégrer dans l'ordinateur. Nous avons également présenté ce qui fait de cette thématique un problème d'optimisation original : la nécessité de localiser *tous* les minima et des dimensions d'espaces de recherche particulièrement importantes. Enfin, nous avons posé les fondements mathématiques de l'échantillonnage conformationnel et du *docking*.

[17] remarquons que dans le cas d'une probabilité de Boltzmann, $e^{-\beta E}$ ne s'annulant jamais, on devra faire l'approximation $E = +\infty$ en dehors de l'ensemble $\mathcal{E}_\pi$

Remarque sur la complémentarité des approches. Loin de nous l'idée que les méthodes computationnelles puissent concurrencer les approches expérimentales ! Car la compréhension des fonctions moléculaires et de leur insertion dans des graphes d'interaction globaux est d'une complexité telle que ces méthodes apparaissent le plus souvent comme complémentaires.

Ainsi, l'alignement d'une séquence protéique sur des bases de données de structures connues reste le meilleur moyen et le plus rapide pour extraire des informations structurales sur la molécule (Vinga et Almeida, 2003, revue sur l'alignement des séquences).

L'échantillonnage conformationnel permet souvent d'affiner les données expérimentales parfois lacunaires ou imprécises. On peut également, connaissant la structure (Yang *et al.*, 2006) ou/et la fonction (Sommer *et al.*, 2004), essayer de prédire les chemins de repliement. De même, les modèles de type Go utilisent la connaissance des contacts natifs (Taketomi *et al.*, 1975).

D'autres auteurs insèrent les informations expérimentales dans les heuristiques de recherche, c'est le cas notamment de Clore *et al.* (1986) qui propagent des contraintes de distances interatomiques dans les simulations. De même, Dandekar et Argos (1997) dont l'algorithme génétique en charge de déterminer la structure tertiaire de protéines accepte les informations extraites d'expériences telles que l'existence de ponts disulfure ou d'interactions site-ligand, la préservation du collapsus hydrophobe ou de « *cages* » à ion metallique, etc. Les algorithmes SHAKE (Van-Gunsteren et Berendsen, 1977, annoncent un gain de temps d'un facteur 3) et RATTLE (Andersen, 1983) propagent également des contraintes de distances afin d'accélérer les simulations de dynamiques moléculaires.

Chapitre 3

Échantillonnage conformationnel d'une seule molécule

3.1 Introduction

La première étape pour comprendre les modes d'interactions de deux molécules, est de mettre en évidence, pour chacun des acteurs, sa ou ses structures préférentielles, ses états de transition, sa dynamique. Avant de traiter deux molécules simultanément, nous étudions donc le cas d'une seule.

Toute molécule possède, comme nous l'avons vu, un certain nombre de degrés de liberté lui permettant de modeler sa géométrie en fonction des différentes interactions intra ou intermoléculaire. L'existence d'une conformation optimale, bien qu'en compétition avec d'autres géométries lorsque la température augmente, découle de la formule de Boltzmann (équation (1.2), page 34) et correspond au minimum absolu de l'hypersurface d'énergie potentielle.

Nous avons choisi, pour capturer la flexibilité des molécules, de décrire ses degrés de liberté torsionnels, ce qui constitue un bon compromis entre taille de l'espace de phase (nombre N_{ddl} de degrés de liberté) et principales sources de flexibilité (voir § 2.3.2, p. 62).

Nous disposons donc d'un premier modèle physique permettant de comprendre le problème biochimique comme une question mathématique : trouver le N_{ddl}-uplet d'angles de torsions qui minimise la fonction énergie.

Ce problème de recherche opérationnelle — trouver un minimum d'une fonction coût — est bien connu des informaticiens et des automaticiens, cependant, ce qui caractérise le problème présent, c'est :

- la taille de l'espace de recherche (avec 1 à 200 degrés de liberté pour les molécules que nous avons traitées),
- les très fortes irrégularités de la fonction cible rendant les études locales et globales très fastidieuses et le nombre prodigieux de minima locaux qui voue toute approche déterministe à l'échec,
- enfin, nous ne cherchons pas *une* solution correcte, mais *le* minimum absolu *et tous* les minima pertinents.

Avant de nous concentrer sur notre implémentation de l'algorithme (§ 3.4 et suivants), nous présenterons quelques stratégies utilisées dans la littérature (§ 3.2) et apportons quelques précisions (§ 3.3) concernant la complexité théorique du problème, la précision que l'on peut attendre du calcul ainsi que le temps de calcul carctéristique d'une évaluation. Ces éléments ont été déterminants dans nos choix.

NB : la fonction cible que l'on cherche à optimiser sera appelée *fitness*, par référence à la recherche opérationnelle.

3.2 Les stratégies existantes

Les différentes stratégies de recherche qui ont été développées peuvent être hiérarchisées selon plusieurs critères dont nous avons retenu un petit nombre listés ci-dessous. Bien souvent, elles dépendent des problèmes auxquels elles ont été appliquées, mais les idées ont été fréquemment reprises, donnant lieu à des adaptations. Enfin, notons que de nombreuses hybridations entre les approches ont rendu la classification plus difficile.

Nous présentons maintenant les principales stratégies de recherche opérationnelle ainsi que les approches existantes de la modélisation moléculaire. Pour ne pas alourdir la rédaction, nous avons résumé la classification des différentes idées dans le tableau 3.1 (page 100) et ne détaillons que les particularités utiles à nos développements futurs.

Critères retenus pour la classification des méthodes :

- espace de recherche discret ou continu,
- optimisation déterministe ou stochastique,
- stratégie d'intensification ou de diversification,
- heuristique nécessitant une solution initiale (voire plusieurs) ou aucune,
- intégration d'un mécanisme de sélection ou non,
- gestion d'une unique solution ou d'une population d' « *individus* »,
- stratégie parallélisable ou séquentielle.

On trouve, dans ce domaine, un certain nombre de revues réalisant l'état de l'art, auxquelles nous renvoyons le lecteur pour plus de précisions : (Neumaier, 1997; Neumaier, 2004).

3.2.1 Algorithmes déterministes

Le premier et le plus simple algorithme d'échantillonnage consiste à explorer exhaustivement tout l'espace de phase, c'est à dire toutes les conformations possibles d'une molécule. Cette stratégie, rapidement écartée étant donnée la croissance exponentielle de la taille de l'espace de recherche en fonction du nombre de degrés de liberté, reste pourtant la seule méthode de recherche qui trouve en temps fini le minimum absolu d'un espace discret (même s'il faut 10^{25} années, voir paradoxe de Levinthal, § 1.3.4.1 p. 41).

Le deuxième type d'algorithmes déterministes, est celui des méthodes par gradient (« steepest descent » ou « hill climbing ») qui réalisent une optimisation locale d'une solution préexistante en explorant le paysage d'énergie dans la direction du gradient, c'est-à-dire en suivant la plus grande pente (Morris *et al.*, 1998; Thomsen, 2003). Comme nous l'avons fait remarquer, il s'agit d'une stratégie de recherche locale offrant une aptitude limitée à l'exploration (même s'il en existe une version *multistart*). Elle souffre de la nécessité d'être initialisée avec une solution de départ et reste bloquée dans le minimum local avoisinant. Elle ne peut donc suffire seule dans un paysage comportant énormément de minima, mais offre cependant un outil très performant lorsqu'elle est hybridée avec d'autres heuristiques.

Les méthodes par voisinages variables (Teghem, 2003), se basant sur une définition multiple, voire adaptative, de la notion de voisinage, permettent de modifier la stratégie en fonction de la configuration du paysage et, ainsi, d'éviter le piège des minima locaux. La recherche reste toutefois limitée à certaines régions de l'espace de solutions, elle ne permet pas une exploration diversifiée et reste une méthode très lente.

La classe des stratégies branch and bound (Androulakis *et al.*, 1995; Klepeis et Floudas, 2001) permettent de localiser le minimum en procédant par découpage et en restreignant progressivement l'espace de recherche. Des exemples de telles approches sont données par les méthodes par intervalles, les intersections par des hyperplans (*cutting planes*), la programmation linéaire pour les fonctions cibles convexes, le cas des fonctions cibles se présentant sous la forme de différences de fonctions convexes, etc.

3.2.2 Algorithmes stochastiques sans mécanisme de sélection

Ils comptent principalement les heuristiques de marche aléatoire et les méthodes de bruitage.

Les marches aléatoires constituent le compromis le plus simple entre les méthodes par gradient (bloquées dans le moindre minimum local) et la recherche exhaustive (qui aboutit en 10^{25} années) : nous avons là une heuristique qui fait fi des barrières énergétiques, tout en préservant une convergence *asymptotique*. C'est aussi la première méthode qu'on peut qualifier de *anytime*, c'est-à-dire qu'elle propose une solution temporaire à tout instant. Ce sont là ses seuls avantages car en plus d'être lente, elle peut revisiter plusieurs fois les mêmes régions de l'espace.

Les méthodes de bruitage, qui consistent à ajouter ou oublier des termes dans la fonction cible (figure 3.1), sont conçues afin d'éviter d'immobiliser la recherche dans les minima locaux. Il s'agit en réalité plus d'une astuce d'optimisation (à rapprocher du *smoothing*, voir § 2.4.2.5 p. 77) à utiliser en conjugaison d'une autre heuristique que d'une méthode de recherche à part entière. Cette stratégie prend toutefois un sens particulier dans le cadre de l'échantillonnage conformationnel si on la rapproche des mécanismes des molécules *chaperones* (§ 1.3.1) qui isolent la protéine à replier et modifient temporairement l'environnement chimique et donc le paysage d'énergie potentielle.

Parmi les méthodes de bruitage, on peut ranger les stratégies qui approximent grossièrement la fonction cible de sorte qu'elle devient beaucoup moins coûteuse à calculer, c'est le cas par exemple des sous-estimateurs convexes (Dill *et al.*, 1996), des approximations par réseaux de neurones (Antes *et al.*, 2005) ou par des estimateurs Pareto (Ultsch, 2003). Parfois elle est plus coûteuse mais présente certains avantages, concernant par exemple l'abaissement des barrières énergétiques : c'est le cas de la stratégie STUN : Stochastic TUNneling (Schug *et al.*, 2005a) qui atténue les fortes énergies (fonction logarithme), ou de l'heuristique Basin Hopping Technique qui utilise comme fonction cible la meilleure énergie dans un voisinage de l'échantillon (Nayeem *et al.*, 1991; Schug *et al.*, 2005a). Enfin, Coleman et Wu (1996) ont proposé d'utiliser un critère similaire à l'énergie libre sur un voisinage des solutions (méthodes par continuation), ce qui lisse d'autant plus le paysage que ce voisinage est grand ; au fil de l'algorithme, le voisinage est rétréci et la fonction modifiée converge vers la fonction cible initiale.

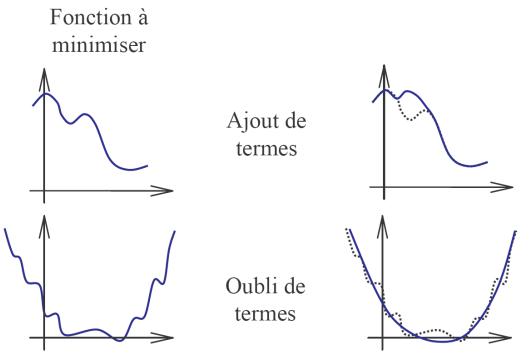


FIG. 3.1: principe des méthodes de bruitage : modification temporaire de la fonction cible.

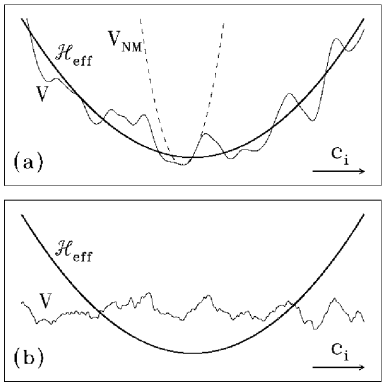


FIG. 3.2: approximation du paysage énergétique par des fonctions simplifiées

3.2.3 Algorithmes stochastiques avec mécanismes de sélection sur solution unique

Un certain nombre d'algorithmes implémentent des stratégies de sélection : le traitement d'une solution dépend maintenant de son *fitness* et la visite des régions de l'espace de phase n'est plus uniquement le fruit du hasard, mais dépend également des solutions antérieurement explorées.

Citons en premier lieu la stratégie « tabous » (Glover, 1989; Glover, 1990; Glover *et al.*, 1995) qui se décline de différentes façons : la première est basée sur la marche aléatoire et consiste à interdire certains mouvements afin d'éviter de revisiter certaines régions connues. Ainsi, si la marche aléatoire propose un déplacement $(d\theta_1, \ldots, d\theta_{N_{\text{ddl}}})$, la stratégie tabous peut n'autoriser que les déplacements vérifiant $d\theta_1 > 0$. Une autre implémentation propose de stocker temporairement les dernières solutions échantillonnées dans une « liste taboue » et de réutiliser cette liste afin de rejeter certains mouvements ou certaines solutions dans des régions considérées comme connues. Bussi *et al.* (2006) ont utilisé une méta-dynamique qui consiste à construire au fur et à mesure, sur le paysage d'énergie, une nouvelle contribution pénalisant les régions déjà échantillonnées (bruitage). Cela permet d'aplanir le paysage, ce qui rend la recherche beaucoup plus exploratoire (la fréquence d'échantillonnage des états devient théoriquement linéaire avec l'énergie au lieu d'une dépendance classiquement exponentielle comme dans la formule de Boltzmann), de plus, le terme de pénalisation offre une image en négatif de l'énergie libre. De la mème façon, Schug *et al.* (2005b), ont proposé une stratégie, appelée energy landscape paving, intermédiaire entre le Monte Carlo (Cf. ci-dessous), les méthodes de bruitage et l'utilisation de tabous, appliquée à l'échantillonnage conformationnel. Elle consiste à explorer les régions de basses énergies du paysage mais utilise une fonction cible modifiée, prenant en compte le temps passé dans chaque minimum afin de forcer constamment la recherche vers de nouvelles régions.

La stratégie de Monte Carlo est également dérivée de la marche aléatoire mais diffère par l'existence d'un critère d'acceptation de chacun des pas, dépendant de la température et des énergies des solutions initiale et finale. Ce critère dit de Metropolis-Hastings peut s'écrire, dans le cas d'une minimisation d'une fonction f, comme (équation (3.1)) :

$$\Pr(\text{accepter un pas de } X \text{ à } Y) = \min\left[1; \exp\left(-\frac{f(Y) - f(X)}{k_B T}\right)\right], \qquad (3.1)$$

où T est un paramètre de température autorisant l'exploration ou au contraire forçant l'intensification. Il est équivalent à $\min\left(1; \frac{\pi(Y)}{\pi(X)}\right)$ lorsque la densité cible π n'est pas exprimée sous la forme $\frac{1}{Z}\exp\left(-\frac{f}{k_B T}\right)$.

Le recuit simulé (Kirkpatrick *et al.*, 1983) s'inspire de concepts de la physique statistique et du procédé de fabrication du même nom, selon lequel les atomes s'arrangent de façon plus stable lorsque la température est augmentée puis diminuée très progressivement. Ainsi, l'algorithme repose sur un ou plusieurs cycles de Monte Carlo avec des montées en température suivies de refroidissements lents (Teghem, 2003; Schug et Wenzel, 2004). Cette stratégie permet, en principe, de sortir des minima locaux et de franchir certaines barrières (lorsque la température est suffisamment haute) et elle assure une convergence asymptotique vers le minimum global. En pratique, l'existence de très fortes barrières énergétiques — comme c'est le cas dans le repliement moléculaire — borne malgré tout la recherche dans des régions restreintes de l'espace de phase ; de plus, il est nécessaire de disposer d'une solution initiale, qui peut influence grandement le résultat final.

Nayeem *et al.* (1991) ont utilisé la méthode intermédiaire du « Basin Hopping Technique » et ont comparé les résultats au recuit simulé : l'approche BHT semble supérieure au recuit simulé, en particulier en ce qui concerne la découverte de minima diverses.

Schug et Wenzel (2004a) ont reporté une version parallèle de recuit simulé où plusieurs solutions sont optimisées indépendamment sur différents processeurs tandis qu'une machine maîtresse gère la convergence des solutions et la répartition des tâches.

3.2.4 Algorithmes stochastiques avec mécanismes de sélection sur un ensemble de solutions

D'autres algorithmes ayant puisé leur inspiration dans les systèmes biologiques naturels, font appel à un ensemble de solutions qu'ils gèrent et font évoluer simultanément ; le devenir d'une solution (appelée individu) ne dépend alors plus simplement de son passé ou de son *fitness*, mais également de l'ensemble de la population. Un premier exemple est celui du paradigme des fourmis (Teghem, 2003), basé sur le recrutement d'individus dans les régions intéressantes de l'espace de phase. Typiquement, cela se fait en mémorisant temporairement (notion de phéromones volatiles) les dernières solutions intéressantes visitées, afin de tirer parti de leur expérience et de proposer des pistes pour les recherches futures. Inversement, le taux d'erreur

qui se traduit par un certain nombre de fourmis déambulant aléatoirement, permet d'explorer globalement le paysage. Ce type de stratégie a été appliqué au problème de l'échantillonnage conformationnel dans le cas du modèle hydrophobe-polaire sur grilles 2D et 3D (Shmygelska et Hoos, 2003, et 2005).

De même, l'heuristique des essaims d'abeilles (Kennedy et Spears, 1998) reproduit certains comportements individuels en espérant voir émerger les comportements collectifs des insectes sociaux qui trouvent inmanquablement la nourriture dans leur paysage propre.

Vengadesan et Gautham (2003) ont proposé d'utiliser un ensemble de « carrés latins » mutuellement orthogonaux afin d'échantillonner l'espace de phase ; à chaque itération, la connaissance du paysage énergétique en N^2 points (où N est la taille de l'espace de phase) leur permet de choisir N^2 nouvelles solutions potentiellement meilleures.

La stratégie, beaucoup plus populaire, des algorithmes génétiques (Holland, 1975) copie les modes de reproduction observés (croisement de chromosomes, mutations accidentelles) et de sélection naturelle (la loi du plus fort) afin de faire émerger les meilleurs individus (Darwin, 1859). De nombreux livres exposent cette stratégie (Goldberg, 1989; Davis, 1991; Michalewicz, 1994; Renders, 1995; Bäck, 1996, etc.) qui offre un véritable cadre de développement pour incorporer toutes les heuristiques complémentaires et astuces vues précédemment. Sa présentation sous forme de squelette algorithmique laissant beaucoup de liberté, ainsi que sa facilité à les adaptater aux différents types de problèmes, ont fait la renommée de cette heuristique. De plus, la possibilité de les paralléliser à plusieurs niveaux (parallélisation des évaluations individuelles, de l'évaluation de la population, modèle des îles, etc.) leur a permis un nouvel essor avec l'avènement du calcul distribué.

Les mécanismes utilisés dans les AGs permettent d'éviter le piège des minima locaux ; de plus, la notion d'héritage des *schémas*[1] (fragments de solutions) prend un sens particulier avec les notions d'éléments structuraux (ou structures secondaires pour les protéines, voir § 1.3.2 p. 31) et de repliement hiérarchique (voir § 1.3.4.5 p. 49). En effet, des sous-parties de solutions correctement repliées peuvent être préservées à travers les mécanismes de croisement et mutation et, ainsi, être disséminées à travers la population en offrant un avantage concurrentiel aux individus, qui maximisent alors la probabilité de recombiner les éléments structuraux pour obtenir une solution globale.

Bien qu'il ait été montré que les algorithmes génétiques (AGs) ne sont pas bien

[1]terme introduit par Holland dans les études théoriques des comportements des AGs.

adaptés à l'optimisation de fonction (De Jong, 1993) (d'où les nombreuses hybridations), ils ont malgré tout été largement utilisés dans les problèmes de modélisation en chimie en général (voir les références de la revue de Leardi de 2001) et de l'échantillonnage conformationnel en particulier (Schulze-Kremer, 1995; Takahashi *et al.*, 1999; Jin *et al.*, 1999; Damsbo *et al.*, 2004; Djurdjevic et Biggs, 2006).

Vieth *et al.* (1998b) ont comparé les AGs à deux autres heuristiques pour le *docking* semi-flexible (Monte Carlo et dynamiques moléculaires) ; la conclusion de cette étude est que les AGs échantillonnent plus souvent dans les régions aberrantes. C'est malheureusement là un des défauts des AGs qui, de plus, n'évitent pas le rééchantillonnage de solutions déjà rencontrées... En conséquence, les AGs sont très gourmands en temps de calcul.

Remarquons enfin que la stratégie de croisement des AGs peut être comprise et implémentée de plusieurs façons. Contrairement à la méthode classique, Glover (1997) a proposé d'utiliser, comme recombinaison, des barycentres de deux individus sélectionnés parmi les meilleurs (heuristique du « scatter search »). Dans ce cas, si la population est autour d'un même minimum, on intensifie la recherche en n'utilisant que des poids positifs (enveloppe convexe) tandis qu'on tend à diversifier si on autorise des poids négatifs. Si la population est répartie dans différents minima, on relie ainsi les puits par des chemins linéaires (en les dépassant éventuellement) pour en découvrir de nouveaux (stratégie dite par recombinaison de chemins ou « path relinking »).

Une autre stratégie évolutionnaire utilise le principe de mutation en réalisant des perturbations des individus selon une loi normale gaussienne dont la matrice de covariance est adaptativement ajustée : c'est la CMA : Covariance Matrix Adaptation (Hansen et Ostermeier, 1996; Hansen et Ostermeier, 2001). Cette heuristique est très en vogue actuellement (Auger *et al.*, 2004) mais se comporte d'autant mieux que le paysage est peu accidenté.

La dernière stratégie est celle des méthodes particulaires appartenant à la classe des heuristiques de Monte Carlo avec chaînes de Markov (Del Moral et Doucet, 2002; Davy *et al.*, 2003; Grassberger, 2004). Elles ont été initialement inventées afin d'estimer des intégrales complexes sur des domaines de grandes dimensions (voir la méthode de Monte Carlo § 2.5.4 p. 85) en générant un n-échantillon selon une densité de probabilité dite *cible*, liée à la fonction objectif. Ces méthodes combinent élégamment les heuristiques de recuit simulé et d'algorithmes génétiques, puisqu'elles utilisent des mécanismes similaires aux mutations/sélections pour diversifier et intensifier la recherche et peuvent introduire un paramètre de température qui diminue

au cours de l'algorithme.

3.2.5 Les dynamiques moléculaires

La popularité des dynamiques moléculaires est telle que chaque idée nouvelle dans le domaine de l'optimisation a reçu son pendant en dynamique moléculaire ; aussi est-il difficile de la classifier dans une catégorie particulière.

Avec des termes stochastiques dits de Langevin, elles font plutôt partie des stratégies *aléatoires* puisqu'on ajoute aux forces usuelles des termes modélisant les chocs stochastiques ayant lieu en solution. Un échantillonnage suffisamment long permet, en théorie, de caractériser le paysage global d'énergie.

Le principal inconvénient des dynamiques moléculaires, c'est qu'elles peuvent difficilement être menées sur des temps plus longs qu'une microseconde[2] (même après plusieurs mois de calculs), durée qui commence seulement à être statistiquement pertinente. De plus les espaces de phase de très grandes dimensions ou comportant des hautes barrières énergétiques restent difficiles à échantillonner (Cui et Simmerling, 2002) et les simulations restent dépendantes des conditions initiales. Ceci n'est pas seulement un problème de moyenne sur une trajectoire trop courte (la puissance des ordinateurs permettant des simulations de plus en plus longues), mais plutôt sur l'unicité de la trajectoire, alors que les observables dans les tubes à essais sont moyennées sur un grand nombre de trajectoires.

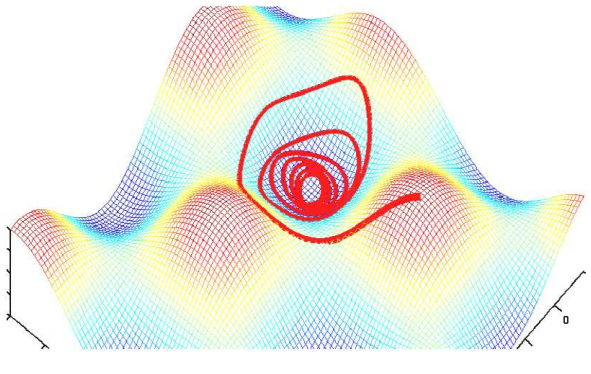


FIG. 3.3: schématisation d'une trajectoire dynamique de la molécule dans son espace de phase, selon les équations de Newton et un champ de forces donné, à partir d'une géométrie et de vitesses atomiques initiales données.

Si la parallélisation de la simulation d'une unique trajectoire paraît difficile, il

[2]Un déploiement masssivement parallèle mené par Pande *et al.*, 2003, reporte un total de quelques centaines de microsecondes.

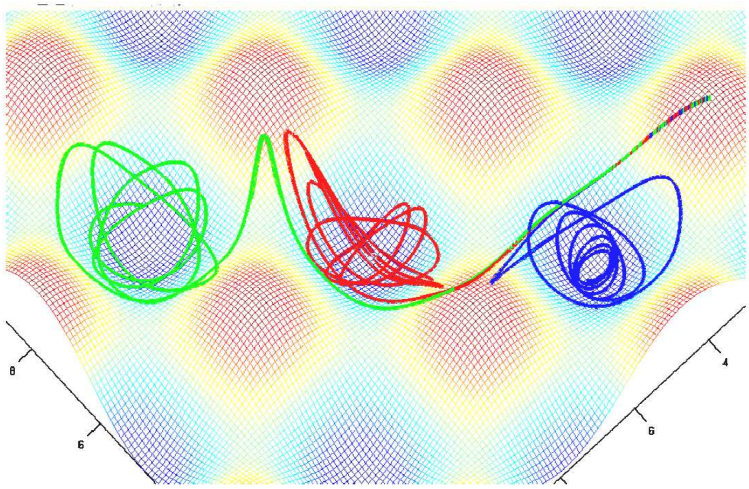


FIG. 3.4: mise en évidence de la dépendance aux conditions initiales : ici les trois simulations aboutissent dans trois minima différents.

est possible de simuler de nombreuses petites trajectoires, ce qui rend possible la parallélisation d'une telle approche (voir Pande, 2003 et Kim, 2004).

Les inconvénients de la dynamique moléculaire ont poussé les scientifiques à améliorer cette stratégie : ainsi, la stratégie de « replica exchange » tire parti de l'avènement des clusters de processeurs et des grilles d'ordinateurs et simule de multiples trajectoires à différentes températures en parallèle avec des échanges de solutions possibles selon un critère semblable au critère de Métropolis-Hastings (Garcia et Onuchic, 2003; Mu *et al.*, 2006; Roitberg *et al.*, 2007). Cette stratégie est aussi dénomée « parallel tempering method » (Schug *et al.*, 2004).

Le « multicanonical molecular dynamics » (Kim *et al.*, 2004) adapte la stratégie du « multicanonical sampling » pour assurer l'échantillonnage des régions de plus fortes énergies lors de simulations de dynamiques moléculaires. Kamiya et Higo (2001) ont également reporté une combinaison de recuit simulé avec une dynamique multicanonique.

3.2.6 Résumé des heuristiques

Le tableau 3.1 résume la classification des principales stratégies de recherche commentées en amont. Les intitulés des colonnes renvoient aux critères retenus dans l'introduction (page 90).

NB : « O/N » signifie « Oui ou Non » ; le symbole « / » indique que la méthode n'est pas concernée par le critère ; enfin, les étoiles : S*, indiquent que des versions « multistart » existent et sont parallélisables.

Stratégie	**C**ont ou **D**isc	**D**ét ou **S**toch	**I**ntens ou **G**lob	Sol init **O**/**N**	Sélection **O**/**N**	**I**ndiv ou **P**op	Parallélisable **S**éq
R. exhaust.	D	D	/	N	N	I	//
Gradient	C	D	I	O	N	I	S*
Branch & B	C	D	G	N	N	I	//
Monte Carlo	D	S	G	O	O	I	S*
Tabous	D	S	G	O	O	I	S*
Recuit Sim.	D	S	G	O	O	I	S*
Basin H. T.	D	S	G+I	O	O	I	S*
Fourmis	C	S	équilibre	O	O	P	//
AGs	D	S	équilibre	N	O	P	//
CMA	C+D	S	I	O	O	P	//
particulaire	D	S	équilibre	N	O	P	//
Dynamique M.	C	S	/	O	N	I	S
Replica exch	C	S	/	O	O	P	//

TAB. 3.1: classification des stratégies de recherche selon principaux critères.

3.3 Premières caractéristiques

3.3.1 Résultats sur la complexité

Je ne me découragerai jamais

Sainte Thérèse de l'Enfant-Jésus, 11 ans

Hart et Belew (1991) ont démontré que l'optimisation d'une fonction quelconque est un problème NP-difficile. Formellement, en considérant la classe des fonctions $f : \{0,1\}^{N_{\text{ddl}}} \to \mathbb{Z}$, qui se calculent en temps polynomial, ils prouvent que le problème de savoir s'il existe un point P de l'espace tel que $f(P) < \lambda$ (λ donné) est NP-complet. La conclusion de cette étude est que l'analyse théorique ou expérimentale des AG ne peut se faire qu'en regard de la classe de fonctions utilisée pour l'optimisation (à ce sujet, il existe des générateurs de problèmes multimodaux aléatoires pour AGs : voir (De Jong *et al.*, 1997)).

Prédire la structure d'une protéine est également NP-complet, comme l'ont prouvé plusieurs auteurs pour différents modèles combinatoires (Ngo et Marks, 1992; Unger et Moult, 1993a; Fraenkel, 1993; Crescenzi *et al.*, 1998). Et même en considérant les géométries connues d'autres séquences, il a été montré que l'étape d'alignement de séquence est déjà NP-difficile (Lathrop, 1994; Calland, 2003)...

La complexité pour l'évaluation de la fonction énergie est, elle, beaucoup plus faible puisqu'elle n'évolue qu'en $o(N_{\text{ddl}}^2)$. En effet, la première étape de reconstruction de la géométrie se fait en $o(N_{\text{ddl}}^2)$, car le nombre de rotations à pourvoir est N_{ddl} et la taille moyenne des fragments à tourner est d'ordre N_{ddl}. On tourne des blocs de plus en plus petits autour des liaisons de valence en commençant par le centre de la molécule. Ensuite, la génération de la matrice des distances interatomiques requiert également un effort qui évolue en $o(N_{\text{ddl}}^2)$, mais l'évaluation des termes d'énergie (liée et non-liée) est linéaire grâce à la troncature à longue distance (§ 2.4.2.6 p. 78).

3.3.2 Précision du calcul pour l'estimation de l'énergie

> Le calcul sur nombres flottants n'est ni associatif, ni commutatif...
>
> anonyme

Estimons le cumul des erreurs dans la rotation des fragments pour la reconstruction des géométries : l'atome en bout de chaîne (une molécule linéaire représentant le pire cas) a subi au maximum $N_{\text{ddl}}/2$ transformations, donnant lieu à chaque fois à une numérisation sur 32 bits (c'est-à-dire une erreur d'environ 10^{-8} maximum dans chacune des trois directions soit $e \triangleq \sqrt{3} \times 10^{-8}$Å au total). Dans le pire des cas, et pour $N_{\text{ddl}} \lesssim 200$, on a donc une erreur de position du dernier atome de l'ordre de $100 \times e$.

Pour avoir la précision de l'énergie par rapport à celle des angles de torsion, il faut multiplier les précisions :

$$\left|\frac{\partial E}{\partial \Theta}\right| = \left|\frac{\partial E}{\partial \mathfrak{d}}\right| \times \left|\frac{\partial \mathfrak{d}}{\partial d}\right| \times \left|\frac{\partial d}{\partial \Theta}\right|, \tag{3.2}$$

où $\mathfrak{d}$ est la pseudo distance utilisée pour lisser le paysage (section 2.4.2.5).

La pire situation advient dans les termes de Van der Waals en A/d^{12} (équation (2.5), p. 72), lorsque des atomes s'interpénètrent : d est proche de 0, $\mathfrak{d}$ vaut alors K_{smooth} (typiquement 1Å) et $A/\mathfrak{d}^{12}$ est de l'ordre de 10^6 à 10^7. Pourtant, les très hautes énergies de telles conformations ne sont là que pour signifier les aberrations dûes aux artéfacts de la modélisation. Ces conformations sont immédiatement détectées et écartées[3] dans l'algorithme et leurs énergies, jamais comparées.

Pour être *viable*, on impose donc que d soit supérieure à une certaine valeur

[3]fœtus *non-viable* ou *mort-né*, dans le vocabulaire des algorithmes génétiques

($d_{\min}$= 0,6Å dans notre cas). À cette distance, on peut estimer toutes les différentielles de l'équation (3.2) :

$$\begin{array}{rcccccc} \left|\frac{\partial E}{\partial \Theta}\right| & = & \left|\frac{\partial E}{\partial \mathfrak{d}}\right| & \times & \left|\frac{\partial \mathfrak{d}}{\partial d}\right| & \times & \left|\frac{\partial d}{\partial \Theta}\right| \\ & = & \frac{12A}{\mathfrak{d}_{\min}^{13}} & \times & \left(1-\frac{K_{\text{smooth}}}{3}\right) & \times & 100e \\ & \approx & 10^7 e, & & & & \end{array} \tag{3.3}$$

soit une précision de l'ordre de 0,1 à 1 kcal.mol^{-1}, ce qui est acceptable, mais non négligeable.

3.3.3 Temps caractéristique

Les temps donnés ci-dessous sont issus de tests sur station de travail HP xw6200 Xeon 3,4 GHz.

Le temps de chargement *offline* de la molécule (lecture des atomes, reconstruction du graphe de connectivité, etc.) se corrèle avec le nombre d'atomes et prend environ 2ms pour une molécule de 300 atomes. Ce temps est identique pour toutes les implémentations de nos algorithmes. C'est aussi, à peu près, le temps qu'il faut compter pour créer et écrire un fichier moléculaire de sortie.

De même, le temps nécessaire pour reconstruire la géométrie d'une molécule de 300 atomes à partir de son vecteur d'angles de torsions est environ de 600μs. L'évaluation de ses termes d'énergie est négligeable et prend au maximum une dizaine de microsecondes.

3.4 Implémentation d'un algorithme génétique

3.4.1 Principe général

Introduits pour la première fois par John Holland[4], les AGs cherchent à reproduire à la fois les mécanismes de croisements d'individus, mais également la pression de sélection qui existe pour la survie et la pérennité des espèces. Le but est de faire émerger, selon la « loi du plus fort », des solutions de plus en plus adaptées.

Les points de l'espace de phase sont donc interprétés comme des *chromosomes* représentant les solutions potentielles ; des opérateurs de croisement et de mutation

[4]première introduction des bases des AGs par John Holland en 1962 : Outline for adaptative systems with programs roving cellular computer, qui a débouché sur le livre fondateur de 1975.

simulent les recombinaisons génétiques et les mutations accidentelles observées dans les organismes naturels (voir figure 3.5).

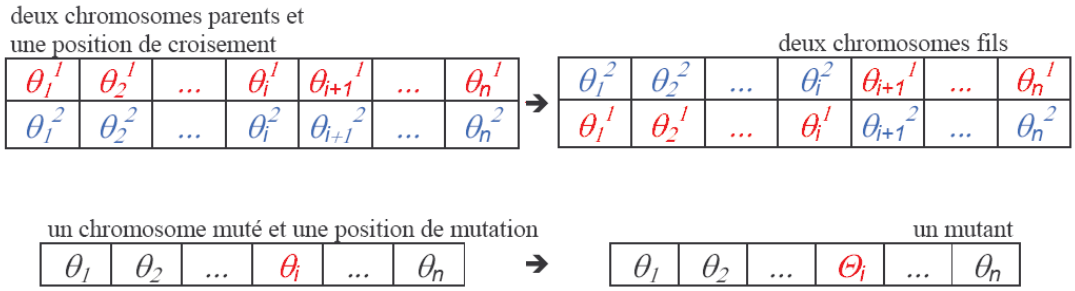


FIG. 3.5: opérateurs de croisement et mutation.

L'algorithme alterne alors, à chaque itération, étape de recherche de nouvelles solutions (diversification) et étape de sélection des individus de meilleures énergies (intensification). Dans la première, les recombinaisons permettent d'agrandir la population, tandis que dans la seconde, on rejette les moins bonnes solutions afin de garder constante la taille de la population (figure 3.6).

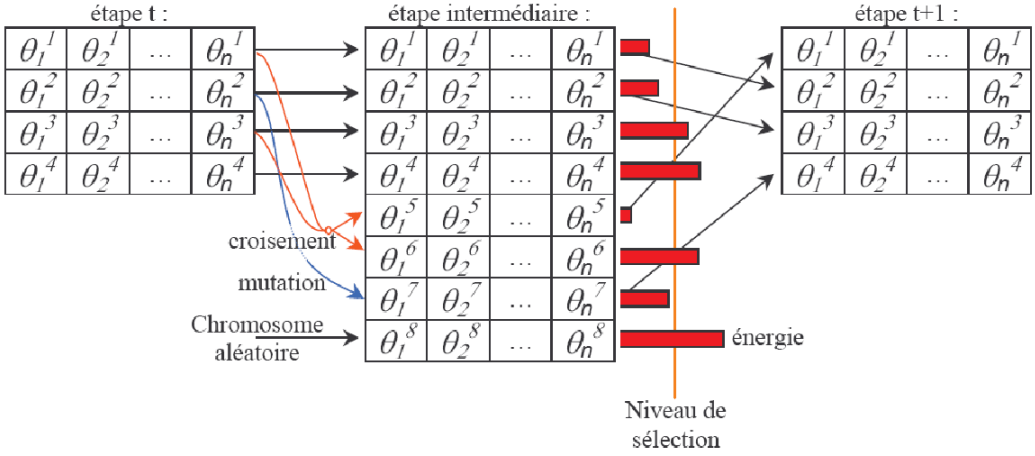


FIG. 3.6: évolution de la population au cours d'une génération ; les θ_i représentent les valeurs d'angles de torsions.

L'opérateur de croisement n'engendre pas de nouvelles valeurs d'angles de torsion (croisement des torsions parentales) et génère donc une exploration limitée ; aussi l'opérateur de mutation permet-il de compenser ce manque. Et même si, comme dans la Nature, ces mutations aveugles aboutissent fréquemment à des individus non-viables (comportant des mauvais contacts, c'est-à-dire des atomes interpénétrés),

elles permettent, d'un point de vue théorique, d'assurer l'ergodicité de la chaîne de Markov (Vose, 1999).

3.4.2 Implémentation

L'un des points sensibles lors de l'implémentation d'un AG, est le choix des différents paramètres opérationnels (taille de population, taux de croisement et mutation, etc.). Ils doivent être calibrés, selon le problème traité, afin d'obtenir une convergence rapide et efficiente vers les minima utiles. Dans ce contexte, nous avons souhaité laisser le choix de ces VALEURS paramétrables (mises en évidence par des PETITES MAJUSCULES dans le texte), en remettant à plus tard la recherche du meilleur réglage et l'étude de l'influence de chacun d'entre eux.

3.4.2.1 Le codage des données

Les angles de torsions sont codés en nombre de PAS D'ÉCHANTILLONNAGE. La précision de ce pas, initialement fixée à 1°, a été abaissée ultérieurement à $0,4°$. Un chromosome est donc un N_{ddl}-uplet d'entiers entre 0 et 899 (0 et 359 initialement) représentant la liste des angles de torsions utiles de la molécule.

En général, les degrés de liberté sont 2π périodiques, mais certains fragments possèdent des symétries qui réduisent la période (π, $2\pi/3$, etc.). Une détection automatique de ces symétries est donc implémentée pour éviter d'avoir plusieurs chromosomes différents codant une même géométrie.

Pour les petits fragments aux extrémités des chaînes de la molécule, il a semblé intéressant d'augmenter ce pas d'échantillonnage en fonction de la TAILLE DE CES FRAGMENTS, ce qui réduit l'espace de recherche[5]. Pour cela, on munit chaque degré de liberté d'une pondération, dépendant linéairement de la taille du fragment qu'il entraîne, et borné entre deux paramètres[6] : MIN et MAX. Ceci nous donne une idée de l'importance relative des degrés de liberté, car un des inconvénients majeurs de la description par angles de torsion est en effet que certaines torsions sont très mobiles et peu influentes tandis que d'autres occupent des positions critiques et sont très rigides. Par cet artéfact, on rétablit en partie l'homogénéité des degrés de liberté. La précision des pas d'échantillonnage est alors définie à partir de ces poids.

[5] éventuellement, certains degrés de liberté comme la rotation des méthyles, peuvent être complètement désactivés, comme dans Damsbo *et al.*, 2004.

[6] les torsions participant à un cycle ont le poids maximum

3.4.2.2 *Fitness*

Comme nous l'avons vu, une molécule est d'autant plus stable que son énergie est basse. Le *fitness* d'un chromosome est donc pris comme étant l'opposé du critère énergie.

L'évaluation de ce *fitness* passe alors par celle de toutes les contributions que nous avons exposées au chapitre précédent. En particulier, les distances inter-atomiques sont calculées à partir des coordonnées cartésiennes, ce qui oblige à reconstruire les géométries 3D des conformères échantillonnés. En outre, l'accès aux différentes constantes du champ de forces se fait lors de l'initialisation de l'algorithme par des fichiers de données empiriques.

3.4.2.3 Gestion de la population

La TAILLE DE LA POPULATION (notée N_{pop}) est fixée au début de l'algorithme. Une population trop grande sur un espace de dimension réduite génère trop de recombinaisons tous azimuts et ralentit l'évolution, tandis que Nix et Vose (1992) ont montré que le nombre de minima locaux dans l'espace de phase déterminait une taille de population critique en dessous de laquelle la probabilité d'une convergence prématurée augmentait dramatiquement. Par ailleurs, la recherche est parallélisée sur PLUSIEURS « continents » (ou « îles » selon les auteurs[7]) en admettant, de temps à autres, des « *migrations* » (inspiré de Spears, 1994). La FRÉQUENCE DE MIGRATION ne doit pas être trop faible, sinon les continents seraient totalement indépendants ; au contraire, si les individus migrent trop souvent, on perd l'intérêt de diversification de la recherche multiple et le coût de communication augmente.

Afin de limiter le nombre de solutions aberrantes (comportant des mauvais contacts) dans la population initiale, nous avons sélectionné N_{pop} individus parmi UN NOMBRE beaucoup plus important d'échantillons. Ceci consiste donc à initialiser la population avec, typiquement, plusieurs milliers d'étapes de Monte Carlo.

3.4.2.4 Gestion de l'évolution

La littérature reporte de nombreux mécanismes de croisement : croisements multi-points (recommandés par Khimasia et Coveney, 1997), croisements uniformes, croisements avec trois parents ou plus (Jin *et al.*, 1999), croisements systématiques (qui consistent à réaliser tous les croisements possibles des deux parents et à ne

[7] voir (Günter, 1992; Mühlenbein, 1992; Lin *et al.*, 1996; Vertanen, 1998; Whitley *et al.*, 1999)

garder que le meilleur enfant, voir König and Dandekar, 1999). Dans certains cas, le mécanisme de croisement est adaptatif (Spears, 1992). Pour une taxonomie complète, le lecteur est invité à consulter l'article de Herrera *et al.* (2003a).

Nous avons opté pour des croisements à un et deux points (le choix de l'un ou de l'autre se faisant selon une PROBABILITÉ donnée), qui ne sont applicables que lorsqu'on est sûr que les enfants seront différents des parents (le choix des partenaires est aléatoire sur l'ensemble des accouplements ainsi autorisés).

Le TAUX DE CROISEMENTS et le TAUX DE MUTATIONS ALÉATOIRES font aussi partie des paramètres à définir.

3.4.2.5 Le mécanisme de sélection naturelle

Le modèle standard d'AG (De Jong *et al.*, 1994; Prebys, 1999) préconise qu'à chaque génération, la population soit remplacée par ses enfants (heuristique dite (λ,μ)) ; cependant, il existe une autre stratégie consistant à mélanger parents et enfants pour ne conserver que la meilleure partie (heuristique $(\lambda+\mu)$). Ce dernier type d'algorithme est qualifié de « steady state » parce que sa convergence est moins hésitante : tout minimum trouvé est conservé. Néanmoins, son évolution est plus intensive car la population « campe » plus sur ses positions et explore moins l'espace.

Notre mécanisme de sélection et de type $(\lambda+\mu)$, il se fait de façon déterministe selon le rang des individus, en triant la population par énergies croissantes. Là encore, de nombreuses solutions étaient possibles (sélections stochastiques, par tournoi, roulette, ou utilisant des probabilités dépendant des énergies, etc.), nous avons retenu plusieurs stratégies pour pouvoir ajuster la balance entre intensification et diversification :

- l'ensemble de la population est filtrée par similarité (Damsbo *et al.*, 2004) : lorsque deux solutions sont jugées trop proches (selon un CRITÈRE DE SIMILARITÉ à définir, voir paragraphe ci-dessous), la moins bonne est remplacée par un chromosome aléatoire. Cette stratégie s'interprète comme un partage des ressources (« food sharing », voir Spears, 1994), où les individus ne peuvent pas tous *butiner* (à l'image de la stratégie des abeilles) au même endroit. Cela force la diversité intra-population en introduisant du « sang neuf » lorsqu'un minimum est trop représenté.
- Périodiquement (à INTERVALLE donné en nombre de générations), la sélection n'est plus faite sur la population complète, mais au sein de chaque famille : parents-enfants ou muté-mutant (mode de sélection dit *intra-familial*). Cela

permet de réduire la consanguinité globale de la population.

– Un remède proposé par Kubota et Fukuda (1997) afin d'éviter l'alternative (λ,μ)-(λ+μ), est d'introduire dans un AG steady state, un mécanisme de vieillissement afin d'autoriser les bons individus à vivre plusieurs générations (stabilisation de la convergence) sans pour autant occuper définitivement les places. Par ailleurs, si une solution n'a pas été disqualifiée par son mauvais *fitness* pendant un certain nombre de générations, on estime qu'elle a eu suffisamment de temps pour répandre son matériel génétique dans la population. Une LIMITE D'ÂGE bien choisie permet alors un bon compromis entre exploitation et exploration de l'espace de phase.

Afin de comparer les individus entre-eux pour pouvoir éliminer les redondances, il a fallu définir une *topologie* sur l'espace de phase. Comme les stratégies de superposition (voir chapitre suivant, § 4.2.1) n'avaient pas encore été étudiées et que la détection des symétries n'était pas implémentée, cette topologie devait tenir compte des symétries internes de la molécule. Pour cela nous avons utilisé des descripteurs géométriques à deux points (voir travaux de Horvath, 2003). Par la suite, ayant pris en compte les symétries dans le codage des données, les comparaisons ont été reportées directement sur les angles de torsion. Cela permet une comparaison beacoup plus rapide (pas de reconstruction de la géométrie) et représente un gain d'espace mémoire étant donné que les descripteurs utilisés étaient relativement volumineux.

Le NIVEAU DE SIMILARITÉ MAXIMAL toléré peut être initialement défini par l'utilisateur, mais par la suite, l'algorithme utilise un compromis entre ce niveau initial et la similarité moyenne au sein de la molécule ; ainsi, pour une molécule très peu flexible où les régions intéressantes sont très restreintes, la pression sera relâchée laissant plus de liberté aux individus. Au contraire, pour une molécule flexible où la population peut se diversifier, la contrainte sera adaptativement renforcée.

3.4.2.6 Contrôle de la convergence

Les solutions jugées intéressantes (c'est-à-dire dans une fenêtre énergétique donnée au-dessus du meilleur minimum rencontré jusqu'alors), sont stockées au cours du déroulement de l'algorithme. Lorsque l'évolution est jugée stagnante (concrètement, le nombre de générations depuis la dernière amélioration significative dépasse un SEUIL FIXÉ), on génère une « *apocalypse* » sur l'ensemble du continent et on réinitialise la population. Néanmoins, les meilleures solutions sont préservées (stratégie d'*élitisme*) à la fois des apocalypses et du *vieillissement*, mais ne se reproduisent

qu'en mode de sélection intra-familiale. Le NOMBRE de ces *immortels* est paramétrable (positif ou nul) car, s'il est important de garder quelques solutions correctes pour redémarrer une population, il ne faut pas entraîner celle-ci dans les mêmes minima locaux que précédemment (Kubota et Fukuda, 1997) ; il y a donc un compromis à trouver.

Enfin, la condition générale d'arrêt de l'algorithme sur chaque continent est définie par un double critère : soit le nombre total de générations dépasse un CERTAIN SEUIL, soit l'évolution est bloquée pendant TROP LONGTEMPS, malgré les apocalypses.

3.4.3 Les hybridations avec d'autres heuristiques

L'algorithme tel que présenté ci-dessus implémente la majorité des stratégies classiques des AGs, mais il n'utilise aucune compréhension physique du problème ; or, c'est généralement lorsqu'on arrive à introduire un minimum de connaissance *a priori* qu'on parvient à diriger la recherche et, ainsi, accélérer et fiabiliser l'algorithme. De plus, les AGs, qui sont principalement un outil d'exploration, sont connus pour bénéficier grandement de stratégies d'hybridation intensifiant les recherches.

3.4.3.1 Gradient conjugué

Pour intensifier cette recherche, nous avons soumis les solutions intéressantes à une optimisation par gradient conjugué afin d'une part, d'accélérer la recherche, mais également pour trouver les géométries stables avoisinants les points échantillonnés. En effet, les différents termes en $1/d^n$ sont tels que des conformations « presque » correctes, proches de minima intéressants sont parfois rejetées à cause d'une énergie dominée par un seul terme provenant d'un mauvais contact et pouvant être facilement corrigé.

La littérature utilise le terme d'optimisation « *lamarckienne*[8] » (Morris *et al.*, 1998) — par opposition aux idées de Darwin sur l'évolution des espèces — car les individus, en apprenant de leur environnement, perpétuent dans les générations suivantes leurs acquis. Ce type d'hybridation a souvent été implémenté avec succès (voir par exemple Khimasia et Coveney, 1997, qui la recommandent), cependant, pour éviter une convergence prématurée de l'AG et une perte de temps dispropor-

[8]Jean Baptiste Lamarck : biologiste français des XVIII et XIX^{e} siècles, ayant prôné la théorie selon laquelle les acquis d'un être biologique pouvaient se recopier dans son génome au fur et à mesure de son apprentissage et ensuite être transmis au générations futures

tionnée, la stratégie n'est appliquée qu'aux bonnes solutions et avec une certaine PROBABILITÉ paramétrable (contrairement à Damsbo *et al.*, 2004, qui l'appliquent systématiquement).

3.4.3.2 Explorateurs indépendants

La plupart du temps, l'opérateur de mutation, qui ne modifie qu'un seul codon du chromosome à la fois, génère des conformères totalement erronés — et ce d'autant plus que la géométrie commence à se structurer avec des imbrications de chaînes entre-elles... De là, l'idée que si les mutations se faisaient, non-plus sur un seul codon, mais en modifiant de façon concertée les différents angles de valence, on obtiendrait un opérateur plus efficace.

Car l'opérateur de mutation possède à la fois un intérêt marginal et primordial : marginal parce qu'il se base sur les erreurs de la nature et conduit souvent à des échecs ; mais primordial parce qu'il assure la convergeance asymptotique vers le minimum global. L'idée est la même que dans l'heuristique des colonies de fourmis (Teghem, 2003; Shmygelska et Hoos, 2005), où ce sont les erreurs qui assurent l'adaptation à l'environnement.

Nous nous sommes alors inspirés d'une heuristique de modélisation moléculaire : le « torsional angle driving » (Accelerys, 2005) pour définir une nouvelle stratégie qui consiste à choisir, comme pour une mutation classique, un codon particulier et une valeur cible de l'angle correspondant. On opère ensuite la mutation en forçant la valeur cible grâce à l'addition, dans la fonction énergie, d'un terme de contrainte très important ne s'annulant qu'au voisinage du point recherché[9] (de type harmonique : $\alpha(\theta - \theta_{\text{cible}})^2$). On laisse alors la molécule se relaxer, par gradient conjugué, dans ce nouveau paysage d'énergie (figures 3.7 et 3.8). Après cette exploration, la solution est à nouveau minimisée afin de se relaxer vers l'optimum voisin. L'intérêt est de conserver des structures viables au cours de la mutation (mutation que nous qualifierons de *dirigée*).

Comme cette nouvelle heuristique est coûteuse en temps de calcul, elle provoquerait une rupture dans l'évolution du continent ; aussi a-t-elle été implémentée sous forme d'un explorateur indépendant — processus fils autonome qui se sépare du continent et revient une fois le calcul terminé (de la même façon qu'un immigrant). Le nombre de ces explorateurs est pourtant limité puisque un seul ne peut exister à la fois pour l'ensemble des continents.

[9] ce qu'on peut rapprocher des méthodes de bruitage § 3.2.2, p. 92.

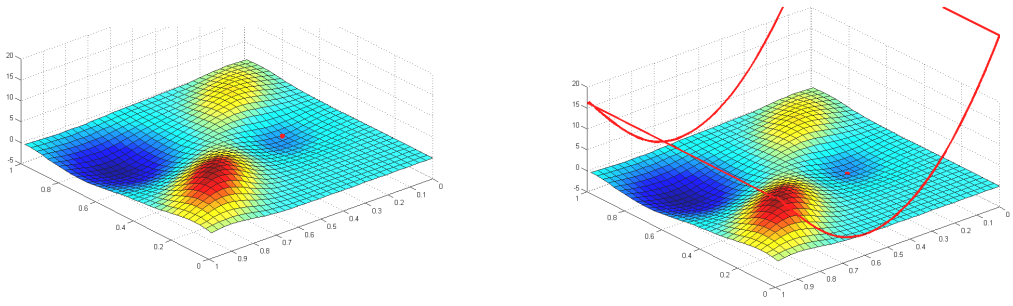


FIG. 3.7: (gauche) quand aucune mutation ou optimisation locale ne peut améliorer une solution, l'adjonction (droite) d'un terme harmonique supplémentaire permet de forcer l'exploration d'autres régions.

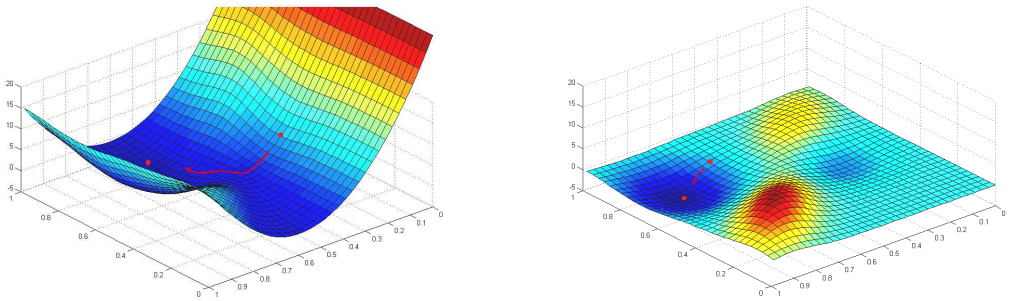


FIG. 3.8: cette exploration reste dans des régions de basses énergies, tout en préservant le principe des mutations. Elle s'achève par une optimisation locale dans le paysage initial.

3.4.3.3 Introduction de tabous

Puisque les AGs ne peuvent pas éviter de re-visiter les régions déjà échantillonnées, nous avons mis en place une politique tabous : au fur et à mesure que les solutions sont stockées dans les fichiers de résultats, elles sont reprises par l'algorithme en guise de représentants des régions déjà explorées. Les individus de la population courante se rapprochant trop de ces « ancêtres » (selon un NOMBRE DE DIFFÉRENCES MINIMAL défini directement sur les angles de torsions) sont rejetés et remplacés. Cette heuristique tabous (Glover *et al.*, 1995) est coûteuse puisqu'elle se base sur des comparaisons d'individus et évolue donc en $o(N_{\text{pop}} \times N_{\text{ancêtres}})$ ($N_{\text{ancêtres}}$ étant le nombre de solutions stockées) toutefois, elle force la diversité et l'exploration de *terra incognita*.

3.4.3.4 Distributions de probabilités biaisées

Pour générer les valeurs d'angles de torsions dans l'initialisation des chromosomes et lors de mutations, on utilise classiquement une densité uniforme obtenue grâce au générateur de nombres aléatoires. Or, en jouant sur ces densités, on peut introduire toute forme de connaissance *a priori* pour entraîner la recherche vers telle ou telle région plus prometteuse. Nous avons alors retenu deux mécanismes pour le choix de ces régions (détaillés ci-après).

Lois marginales. Pour une molécule en solution, la véritable densité de probabilité (notée p_{Θ_i}) d'*une unique* torsion Θ_i, indépendamment des autres degrés de liberté, est donnée par la moyenne des probabilités sur les autres torsions (ce sont les lois marginales, voir équation (3.4)).

$$\begin{aligned} p_{\Theta_i}(\theta) &= \frac{\Pr(\Theta_i \in [\theta; \theta + d\theta[)}{d\theta} \\ &= \int p(\theta_1, \ldots, \theta_{i-1}, \theta, \theta_{i+1}, \ldots, \theta_{N_{\text{ddl}}}) d\theta_1 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_{N_{\text{ddl}}}. \quad (3.4) \end{aligned}$$

On aimerait disposer de ces densités pour générer les chromosomes, malheureusement, elles ne sont accessibles qu'*a posteriori* et, de plus, la densité globale n'est en général pas égale au produit des densités marginales (en effet, les degrés de liberté ne sont pas indépendants). Cependant, toute information, même fragmentaire, permet de tirer des valeurs d'angles « *en moyenne* » plus intéressantes (c'est l'idée des densités de Ramachandran pour les squelettes protéiques). Enfin, pour ne pas occulter

certaines régions de l'espace de recherche, ces densités sont toujours mélangées avec une densité uniforme selon un PARAMÈTRE réglable.

Biais *a priori*. Tout d'abord, nous savons *a priori* que les conformations dites *décalées* sont plus souvent adoptées que les conformations *eclipsées* (figure 3.9). Cela provient de l'existence de tensions locales (i.e. entre atomes topologiquement proches) qui dominent les autres termes énergétiques et sont donc déterminantes pour les densités de probabilités boltzmaniennes. Tout se passe comme si — en première approximation — les densités marginales p_{Θ_i} ne dépendaient que des premiers atomes mis en mouvement par Θ_i. On évalue alors, pour chaque valeur d'angle de la torsion, un « Hamiltonien local » simplifié[10] qui est transformé en probabilités par l'équation de Boltzmann (1.2). Ceci constitue une forme de connaissance *a priori*, innée pour nos chromosomes (voir Strizhev *et al*, 2006).

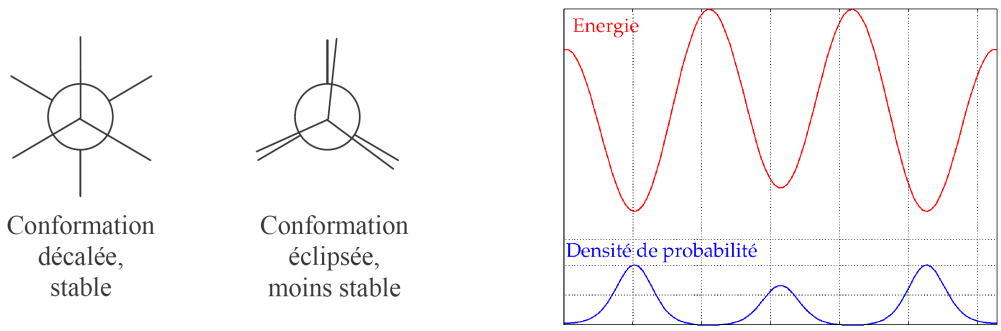


FIG. 3.9: densités de probabilités non-uniformes pour minimiser les tensions locales.

Biais *a posteriori*. La deuxième source d'information intégrée dans les densités de probabilités provient de l'expérience qu'a acquis la population du paysage énergétique[11]. Pour cela, on réalise des statistiques par torsion afin de mettre en évidence les régions intéressantes, voir figure 3.10). Ces acquis de la population forment une connaissance *a posteriori*, que l'on interprète comme un traditionalisme (à rapprocher de Liwo *et al.*, 1999).

Comme cette dernière stratégie est *auto-cohérente* (plus l'algorithme converge, plus les régions connues sont probablement visitées), la technique n'est appliquée que sur une seule île et seulement si le nombre de solutions accumulées est suffisant (typiquement 100). En effet, l'analyse des résultats montrera que cette forme d'in-

[10] reprenant ainsi l'idée des méthodes de bruitage qui omettent certains termes.
[11] à l'image de la stratégie des fourmis qui indroduit les phéromones comme effet mémoire

tensification de la recherche présente aussi le risque d'une convergence prématurée dans des minima locaux.

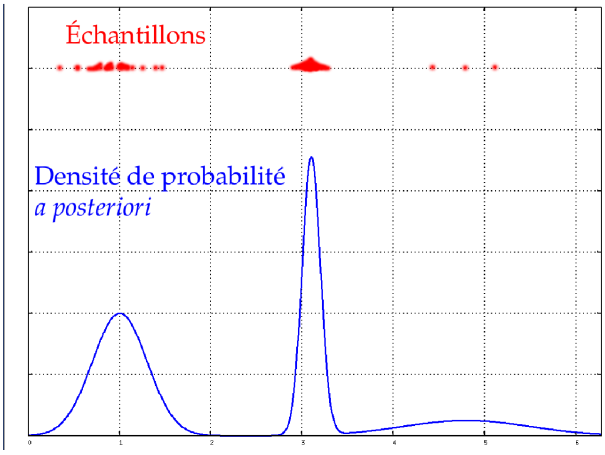


FIG. 3.10: densités de probabilités non-uniformes par apprentissage.

3.4.4 Méta-optimisation

Nous disposons donc maintenant d'un algorithme génétique que l'on peut qualifier de générique dans le sens où ses paramètres opérationnels (taux de mutations/croisements, taille de population, condition d'arrêt, etc.) sont paramétrables et les diverses stratégies (parallélisation, élitisme, gradient conjugué, biais dans les densités de probabilités, filtrage par dissimilitude) peuvent être relativisées voire totalement désactivées.

Le problème qui nous intéresse maintenant est de savoir comment régler tous ces paramètres afin d'obtenir une convergence « *satisfaisante* » de l'algorithme. Étant données les bonnes performances potentielles des AG et la forte dépendance de leurs résultats vis-à-vis de ces réglages, il n'est pas étonnant que cette question soit au coeur des recherches dans ce domaine.

Même en limitant le nombre de valeurs par paramètre (de deux à cinq valeurs pour un total de dix-sept paramètres, voir tableau 3.2), on trouve 10^9 réglages possibles ! Il nous faut donc définir ce que « convergence satisfaisante » signifie, c'est-à-dire trouver un moyen de comparaison entre les algorithmes (une fonction des paramètres que nous appelerons « méta-*fitness* »).

Les deux principales approches que l'on trouve pour la recherche des paramétrages optimaux, sont soit des tentatives de descriptions purement analytiques des

Valeurs possibles	Paramètre
2, 3 ou 4	nombre d'îles
5, 10, 25 ou 50	période de migration (en nombre de générations)
500, 800 ou 1000	nombre maximum de générations sans succès avant arrêt global
50, 75 ou 100	nombre maximum de générations sans succès avant apocalypse
50, 100, 150 ou 200	taille de population
0 ou 1	nombre d'élites immortels
20, 50, 100 ou 200	âge maximum toléré
1, 2, 5 ou 10	fréquence de sélection intrafamiliale (en nombre de générations)
1% ou 10%	fréquence de mutations
40, 70 ou 100%	taux de croisements
0, 33, 67, 100%	taux de croisements à deux points
10, 30 ou 50%	probabilité d'application d'une relaxation par gradient conjugué
75, 80, 85 ou 90%	niveau de similarité maximal dans la population
20, 30, 40, 50 ou 60	taille du voisinage tabou autour des individus déjà rencontrés
10, 30 ou 50%	niveau de mélange de la densité uniforme par rapport aux densités biaisées
6, 8, 10, 12	taille minimale des fragments définissant une torsion active
3, 5, 10, 20	taille des fragments au-dessus de laquelle les torsions ont toute la même pondération (en nombre d'atomes)

TAB. 3.2: paramètres de contrôle de l'algorithme et ensemble des valeurs possibles.

AGs (à l'aide des chaînes de Markov), soit des mises en évidence de certains comportements en s'appuyant sur des résultats expérimentaux.

3.4.4.1 Les chaînes de Markov

Les AG codés binaires (et plus généralement, ceux qui s'appliquent à des espaces de recherche discrets), peuvent être modélisés par une chaîne de Markov finie, discrète (Nix et Vose, 1992; De Jong *et al.*, 1994; Spears et De Jong, 1996). Pour une population de N_{pop} individus de longueur N_{ddl}, chaque composante pouvant prendre N_{steps} valeurs, nous avons de l'ordre de ${N_{\text{steps}}}^{N_{\text{ddl}}}$ états possibles et donc $M = \binom{{N_{\text{steps}}}^{N_{\text{ddl}}}}{N_{\text{pop}}}$ combinaisons[12] possibles de populations différentes qui formeront les *états* de la chaîne de Markov. Nous voyons immédiatement que la matrice de transition, de dimension M^2, devient rapidement impossible à gérer informatiquement, lorsque N_{pop}, N_{ddl} et/ou N_{steps} prennent des valeurs physiquement utiles !

Une visualisation possible (De Jong *et al.*, 1994; Spears et De Jong, 1996) est de dessiner la matrice de transition sous la forme d'une image carrée, en remplaçant les probabilités par des niveaux de gris. Il faut ensuite ordonner les états (numérotation

[12]en accord avec la nouvelle règle, les coefficients binomiaux $\frac{n!}{p!(n-p)!}$ sont notés $\binom{n}{p}$

a priori) de façon subtile si l'on veut voir apparaître les régions attractives de l'espace de recherche (figure 3.11).

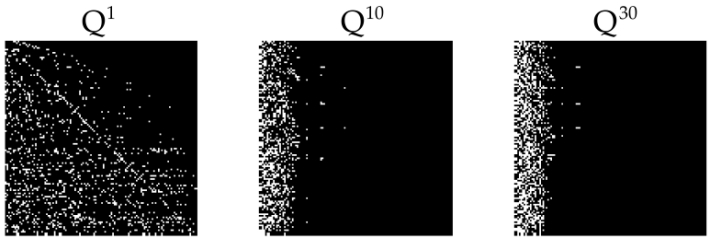


FIG. 3.11: représentation des puissances de la matrice de transition, où l'on visualise l'apparition de régions attractrices (extrait de Spears, 1996).

Les quelques tentatives de mises en pratique de telles études mettent en évidence certains comportements des AGs (Spears, 1992; Spears, 1994; Ochoa *et al.*, 1999), mais la quantification des phénomènes est à mettre en relation avec la taille de l'espace de recherche ou la classe de fonctions « *fitness* » utilisée. Dans De Jong *et al.* (1994), l'espace d'états est tellement réduit qu'à titre de comparaison, une recherche aléatoire montre de meilleurs résultats !

La recherche du paramétrage optimal par une analyse théorique semble donc être une approche difficile étant donné la taille de nos espaces de recherche, le nombre de stratégies implémentées et la dépendance du paramétrage optimal au problème traité.

Il existe une autre approche consistant à considérer le « méta-problème » comme une optimisation classique pouvant être faite en-ligne (auto-adaptation : voir Sawai et Adachi, 2002, logique floue : voir Herrera et Lozano, 2001 et 2003b) ou hors-ligne (Grefenstette, 1986; Djurdjevic et Biggs, 2006) — c'est cette dernière approche que nous avons retenue.

3.4.4.2 Le *fitness* d'un algorithme

L'évaluation d'un AG pose deux principaux problèmes : le premier est que nous ne recherchons pas simplement le minimum absolu du paysage d'énergie potentielle, mais le maximum de minima significatifs. Ce que la plupart des auteurs proposent (Spears, 1992; Ochoa *et al.*, 1999) — juger un AG en utilisant le *fitness* du meilleur individu jamais produit (« *best-so-far* ») — n'est donc pas applicable dans notre cas.

Le second problème concerne la non-reproductibilité des résultats. L'algorithme étant fortement stochastique, le bon fonctionnement d'un AG particulier peut aussi bien être dû à la qualité des réglages qu'être le fruit de la chance : or, ce qui nous

intéresse, c'est de trouver le paramétrage qui nous assure le plus de chances d'obtenir une convergence efficace.

De façon plus formelle, on peut considérer la réalisation d'un AG comme un événement aléatoire dont le méta-*fitness* dépendrait. Ce dernier doit donc être vu comme une *variable aléatoire* dont la moyenne est le véritable critère qu'il faut optimiser.

Le méta-*fitness* (μF) doit donc prendre en compte les critères suivants :

- l'énergie du meilleur chromosome (au plus bas est cette énergie, au meilleur sera l'AG),
- les minima pertinents (leur nombre et leurs énergies),
- le temps de calcul nécessaire à produire ces solutions.

Un certain nombre de critères et d'indices sont proposés par Wehrens *et al.* (1998) pour évaluer la qualité des AGs, prenant en compte les effets stochastiques et les aspects multimodaux.

Afin de répondre aux deux premiers critères, nous avons emprunté à la physique statistique, la *fonction de partition* des solutions retournées (qui donne l'énergie libre d'un ensemble de molécules). En rajoutant une pénalité pour le temps de calcul on obtient le méta-*fitness* utilisé :

$$\mu F = +k_B T . \log \left[\sum_{\text{échantillons}} \exp\left(-\frac{E_i}{k_B T} \right) \right] - \alpha . t_{\text{CPU}}. \tag{3.5}$$

Le choix du paramètre α est fait de sorte qu'une heure de calculs ait le même poids que 10 kcal.mol^{-1}, un algorithme qui ne serait pas descendu de 10 kcal.mol^{-1} après une heure de calculs sera donc défavorisé par rapport à l'AG qui se serait arrêté tout de suite.

Afin d'évaluer l'espérance de μF (notée $\mathbb{E}[\mu F]$), nous avons utilisé la moyenne des valeurs sur plusieurs réalisations :

$$\mathbb{E}[\mu F] = \frac{1}{N_{\text{runs}}} \sum_{i \leq N_{\text{runs}}} \mu F_i. \tag{3.6}$$

Comme la réalisation d'un AG prend entre 30 minutes et quelques jours, nous nous sommes limités à $N_{\text{runs}} = 3$ réalisations.

Nous disposons donc maintenant d'un critère précis qu'il reste à optimiser en jouant sur les paramètres. Même si, en utilisant $\mathbb{E}[\mu F]$ plutôt que μF, on a une meilleure idée de l'impact d'un jeu de paramètres, gardons à l'esprit que la reproductibilité d'une expérience reste quand même un point très sensible.

3.4.4.3 Méta-algorithme d'optimisation

Nous cherchons à minimiser un critère de coût, aussi peut-on donc appliquer toutes les heuristiques de recherche que nous avons vues dans la section 3.2, avec la donnée supplémentaire que l'évaluation du méta-*fitness* peut prendre jusqu'à 48 heures.

Cette méthodologie a déjà été utilisée (Grefenstette, 1986; Schulze-Kremer et Tiedemann, 1994; Jin *et al.*, 1999; Nùnez-Letamendia, 2003; Djurdjevic et Biggs, 2006). Elle présente la particularité de faire un réglage hors-ligne des paramètres, contrairement à l'approche en-ligne qu'ont implémentée d'autres chercheurs.

Une remarque importante, faisant référence à l'article de Hart et Belew (1991), est qu'il faut se rappeler que l'optimisation des paramètres va dépendre de la molécule traitée. Aussi, la méta-optimisation devra-t-elle être appliquée à chaque molécule.

Enfin, pour méta-optimiser les paramètres opérationnels, nous avons considéré un algorithme génétique extrêmement simplifié : un chromosome est un n-uple de paramètres de réglage de l'AG de base ; à chaque itération, une population de dix *méta*-individus permet de générer, par croisements (un point) et mutations, dix enfants qui sont évalués ; parmi les dix parents et dix enfants, ne sont alors conservés que les dix meilleurs pour former la nouvelle population. Etant donné le coût de l'évaluation du méta-*fitness*, on évite de générer des jeux de paramètres déjà testés. Enfin, la condition d'arrêt a lieu lorsqu'aucune nouvelle géométrie n'a été trouvée par l'AG d'échantillonnage conformationnel depuis quatre méta-individus.

Cette « machinerie » est gérée par des scripts shell et awk qui lancent les algorithmes génétiques et récupèrent les solutions renvoyées. Afin de distinguer les deux couches algorithmiques, nous appelerons C_SG_A (pour « Conformational Sampling Genetic Algorithm »), l'algorithme d'échantillonnage conformationnel (hybridé et paramétré) et méta-algorithme génétique (ou μG_A) l'algorithme en charge de trouver le meilleur paramétrage et la meilleure stratégie d'hybridation des différentes heuristiques. La figure 3.12 représente le schéma global de l'un et l'autre.

3.4.5 Résultats

3.4.5.1 Les molécules de tests

L'ensemble des stratégies présentées ainsi que l'algorithme génétique paramétrable ont été implémentés en Fortran 77 et validés sur un certain nombre de petites

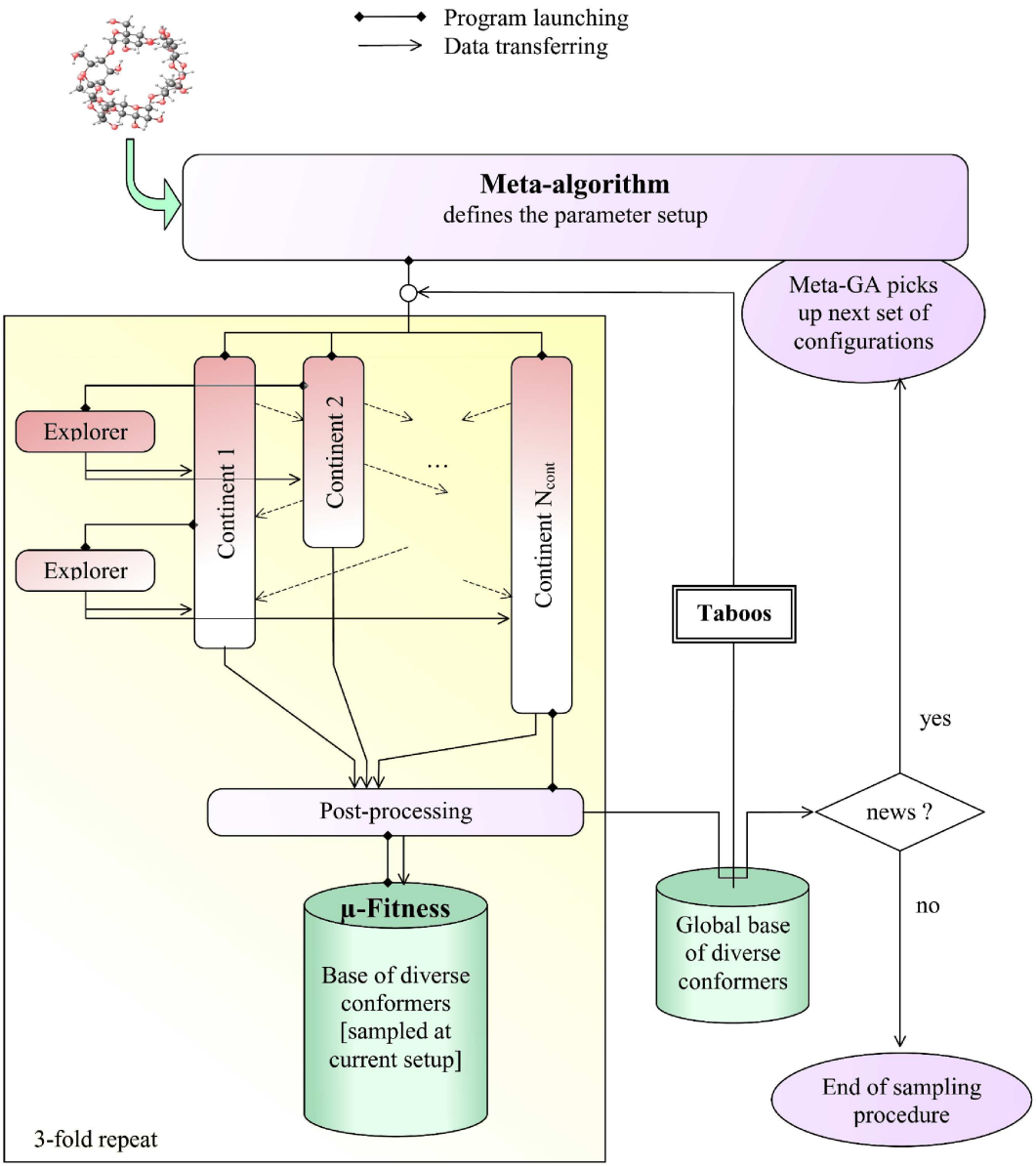


FIG. 3.12: schéma global de l'échantillonnage conformationnel, faisant apparaître les trois répétitions des AGs impliquées dans la méta-boucle d'optimisation (figure parue dans le journal Soft Computing, voir annexe C).

molécules tests (petit peptide à huit degrés de liberté, molécule organique « p3sem », molécule dendritique). Des molécules plus originales ont également été traitées : un poly-cycle et un dodécaèdre de carbone qui comportent plusieurs cycles adjacents, ont permis de tester différentes stratégies de coupure formelle des liaisons (figure 3.13). Dans tous ces cas, les structures prédites coïncident avec les structures expérimentales. La qualité des prédictions du C_SG_A pour la fourchette de molécules $0 \leq N_{\text{ddl}} \leq 30$ en fait déjà un outil potentiel pour générer les structures 3D et calculer des descripteurs géométriques sur les bases de données pharmaceutiques.

Les temps de simulation par AG seul (sans biais, ni explorateurs) se situaient alors typiquement entre une demi-heure et deux jours, sur un quadriprocesseur Silicon Graphics, R12000, 360 MHz. Avec les hybridations, ces temps sont maintenant considérablement réduits.

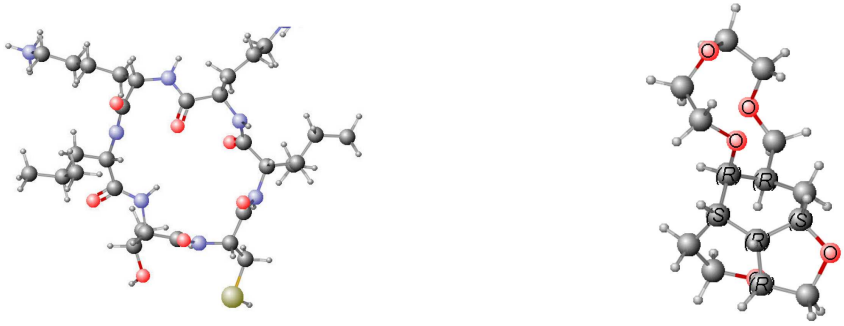


FIG. 3.13: quelques exemples simples de molécules tests.

Les différentes hybridations et stratégies complémentaires nous ont permis d'aborder des molécules plus grandes (jusqu'à 65 degrés de liberté) comportant certaines particuliarités :

- la « fillipine » (Volpon et Lancelin, 2000) est une molécule cyclique présentant d'une part une succession de doubles liaisons en résonnance et, d'autre part, un réseau de ponts hydrogène rigidifiant la structure (figure 3.14).
- La « cyclodextrine » est un macrocycle comportant six cycles de glucose s'orientant comme autour d'un cylindre (figure 3.15). Étant donnée sa structure, elle est utilisée comme vecteur pharmaceutique pour véhiculer certains médicaments instables. Chacun de ces cycles ainsi que le cycle global ont été ouverts afin de permettre un échantillonnage global des conformations, ce qui représente 65 degrés de liberté au total.

Le comportement du C_SG_A sur de telles molécules fût satisfaisant dans le sens où il a découvert le minimum expérimental d'une part et, d'autre part, il a permis de localiser d'autres minima peuplés à haute température. Avant que ne soit

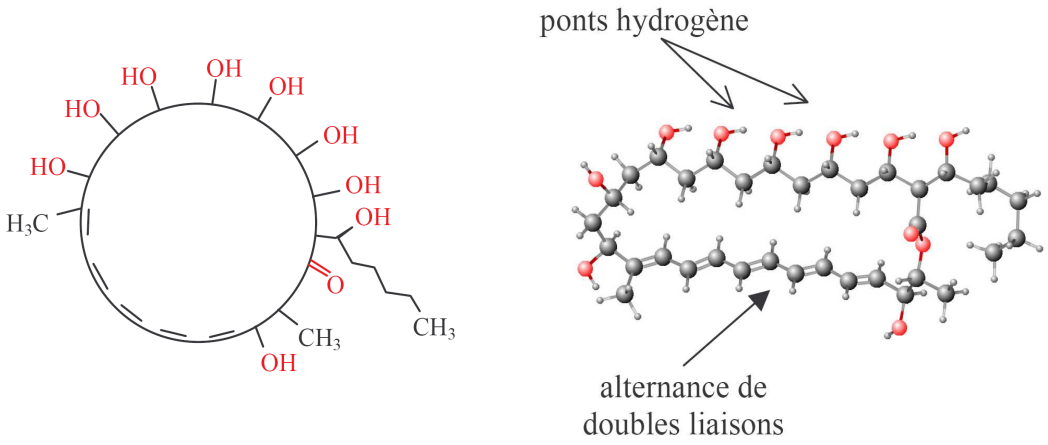


FIG. 3.14: la fillipine, formule topologique (gauche) et structure tridimensionnelle (droite).

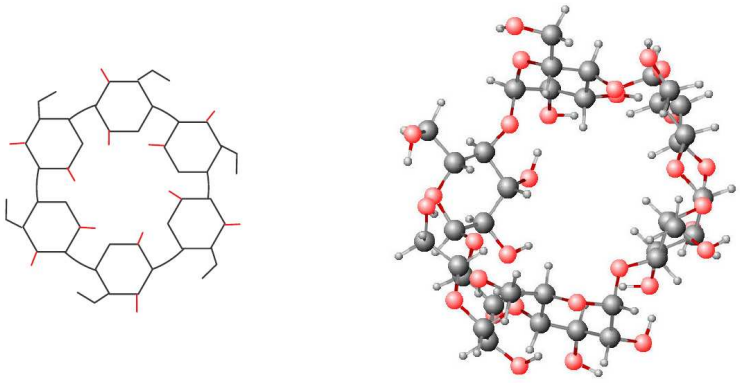


FIG. 3.15: la cyclodextrine, formule topologique et structure tridimensionnelle.

implémentée la stratégie des explorateurs indépendants, une telle recherche nécessitait environ une semaine de calculs. Maintenant, la cyclodextrine est devenue un problème « facile », soluble en moins d'un jour.

Les nombreuses contraintes covalentes rendent son paysage d'énergie particulièrement étroit et difficile à échantillonner pour des stratégies de recuit simulé ou de dynamique moléculaires. Les AGs offrent donc une alternative particulièrement attrayante, cependant, les outils d'optimisation locale par gradient (lamarckisme, explorateurs indépendants) ont été particulièrement précieux dans ce genre de paysage et il est probable[13] que des molécules similaires, mais non-cycliques, nécessitent beaucoup plus de temps de calculs.

3.4.5.2 Vers un traitement automatique des molécules ?

La convergence du μG_A permet de connaître *a posteriori* le meilleur (ou du moins « un bon ») réglage du $C_S G_A$ pour la molécule traitée. La première remarque que nous pouvons faire, c'est que ce réglage permet un échantillonnage nettement amélioré de la conformation d'une molécule (comme en témoigne la figure 3.16) : en termes d'énergies mais également en reproductibilité.

Malheureusement, ces paramètres sont inconnus avant l'échantillonnage... Dans ce contexte, nous avons implémenté une stratégie d'apprentissage des réglages en fonction de certaines caractéristiques topologiques de la molécule. Ceci nous permet de classer la molécule entrante avec les molécules déjà traitées ; ensuite, le μG_A est initialisé avec une population comportant des jeux de paramètres correspondant à des molécules connues (voir figure 3.17).

3.4.5.3 Analyse des résultats

La cyclodextrine nous a servi de modèle pour tester l'impact des différentes heuristiques d'hybridation et pour analyser la convergence du méta-algorithme. Nous avons mis en évidence un certain nombre de « comportements[14] » que nous avons relatés dans l'article[15] qui est paru dans le jounal *Soft Computing - A Fusion of Foundations, Methodologies and Applications* en janvier 2007 (mis en annexe C).

Deux remarques préliminaires peuvent être faites : la première est que les essais ne sont que faiblement reproductibles. La stochasticité des résultats (malgré la

[13] et même certain *a posteriori*.
[14] des *schémas* de paramètres au sens de Holland.
[15] voir Parent *et al.*, 2007

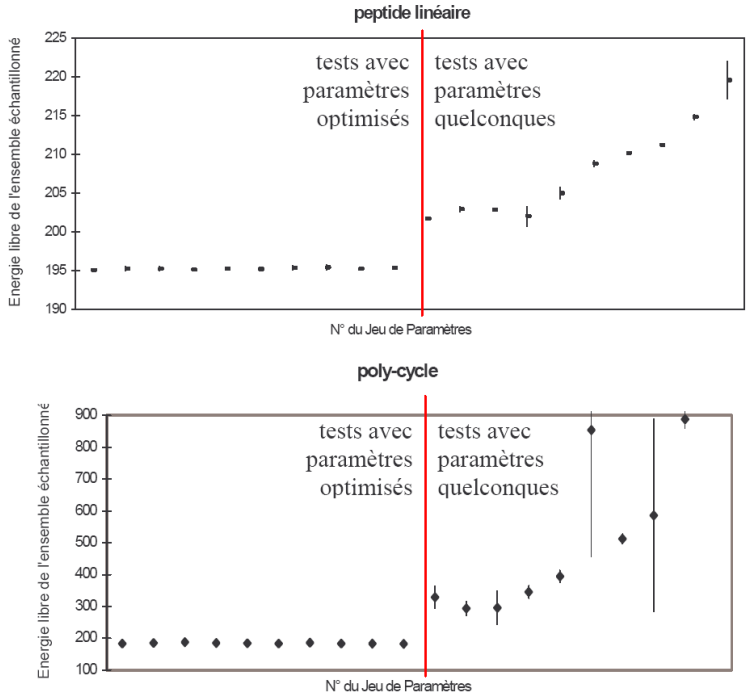


FIG. 3.16: énergie libre (avec barres d'erreur sur trois réplicats) de l'ensemble des solutions retournées en fonction du jeu de paramètres. Partie gauche : dix exécutions avec la dernière population de paramètres, partie droite : dix exécutions avec des paramètres aléatoires (première population de paramètres).

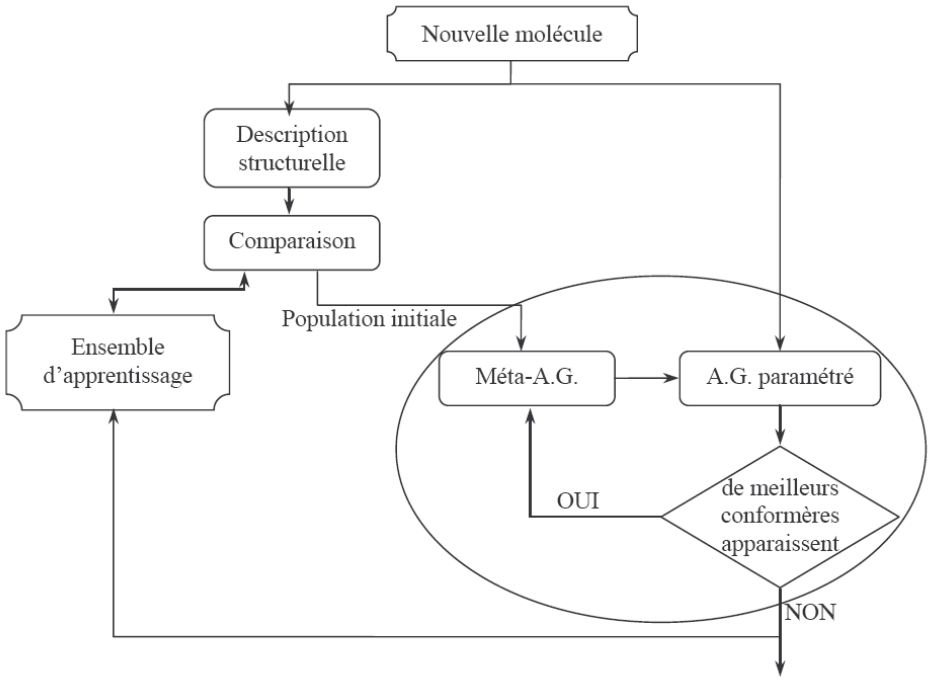


FIG. 3.17: schéma de fonctionnement pour l'assignation de paramètres initiaux dans le μG_A.

moyenne sur trois essais) peut être imputée à la condition d'arrêt du μG_A. Néanmoins, un certain nombre de tendances peuvent être mises en évidence. Deuxièmement, les conformations retournées par l'algorithme (toutes stratégies confondues) correspondent bien à la géométrie attendue, connue expérimentalement. Les différences d'énergies et de structures sont principalement attribuables à des réarrangements spatiaux des groupements latéraux. Puisque les aspects géométriques sont correctement prédits, nous pouvons nous concentrer, dans un premier temps, sur les résultats purement algorithmiques.

L'analyse est faite à deux niveaux : d'une part concernant le choix d'application des différentes stratégies et, d'autre part, pour le réglage des paramètres internes à l'AG :

3.4.5.4 Comportement en fonction des stratégies d'hybridations

Les jeux de tests ont été générés de la manière suivante : différentes combinaisons des stratégies ont été appliquées à partir d'une même méta-population aléatoire initiale. Les combinaisons sont basées sur un mode de fonctionnement « par défaut » pour lequel toutes les heuristiques sont autorisées ; puis tour à tour, les stratégies

sont désactivées :

- « Default » : toutes les stratégies sont activées,
- « No Taboo » : on autorise les recherches aux voisinages des points déjà échantillonnés,
- « No Explorer » : le mécanisme de mutations dirigées (torsional angle driving) est inactif,
- « No Tradition » : le principe d'apprentissage et de biais des probabilités vers les régions *a posteriori* intéressantes est désactivé, mais on conserve la stratégie de biais a priori par minimisation de l'énergie locale,
- « Flat distribution » : aucune stratégie de modification des probabilités n'est permise.

Pour chacune des politiques proposées, trois tests (μG_A) ont été réalisés (les choix des paramètres internes, sauf pour la première méta-population est donc fait automatiquement par le μG_A).

La stratégie d'exploration se révèle d'une grande utilité pour générer de bonnes structures ; de plus, comme elle est régulièrement appliquée entre deux points a priori quelconques de l'espace de recherche, elle n'entraîne pas de convergence prématurée de la population comme pourrait le faire l'introduction d'immortels (effet de dérive, voir Kubota et Fukuda, 1997). Sur le graphe figure 3.18, on voit clairement que sans la procédure de mutation dirigée, les énergies et le nombre de conformères retournés sont beaucoup moins bons (au moins dans deux cas sur les trois). Il est probable qu'en lui laissant plus de temps, l'AG finirait par trouver ces mêmes minima ; cependant, en comparant les temps de calculs (figure 3.19), on s'aperçoit que la stratégie améliore également la vitesse de convergence.

L'introduction de tabous ralentit l'évolution mais améliore la diversité au sein de la population (la figure 3.18 montre en effet que désactiver les tabous génére un nombre restreint de minima). L'affirmation est d'autant plus vraie que le paysage d'énergie potentielle pour la cyclodextrine ne doit comporter que quelques rares et étroits minima (dus à la présence des multiples cycles). L'utilisation de tabous serait peut-être davantage recommandée pour des molécules plus flexibles et dont le paysage d'énergie, moins accidenté, nécessiterait une heuristique de recherche plus globale. Autoriser la revisite des minima connus dans la stratégie « No Taboo », augmente les chances d'optimiser localement la structure ; cependant, comme nous le verrons ultérieurement, la décision de transformer un individu — potentiellement

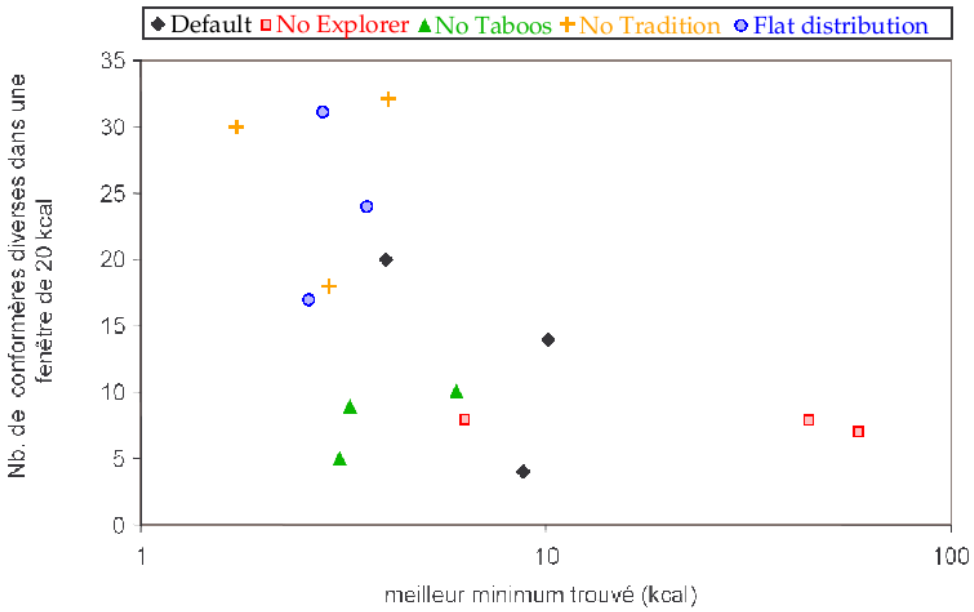


FIG. 3.18: nombre de solutions pertinentes trouvées et meilleure énergie trouvée pour les tests des différentes stratégies.

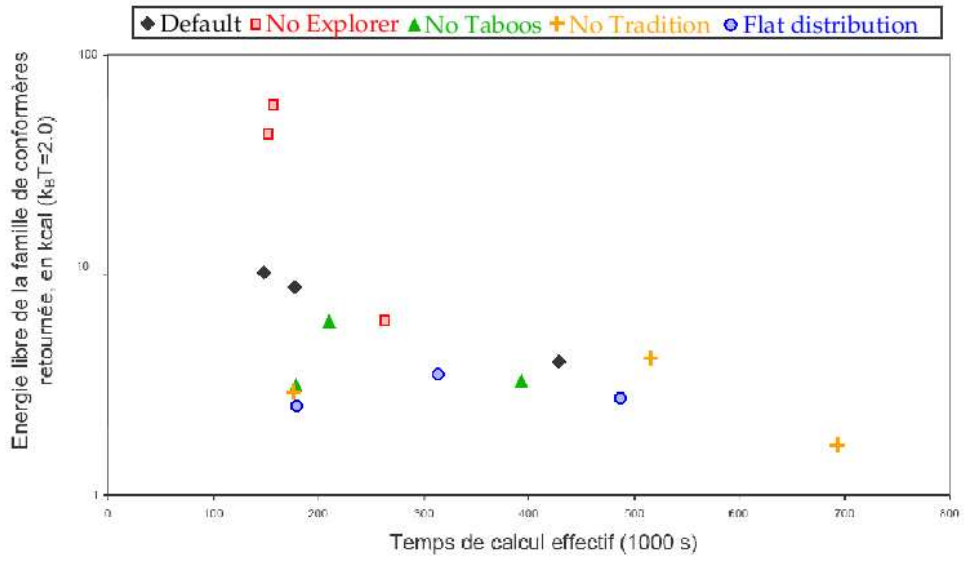


FIG. 3.19: comparaison des temps de calculs des différents tests et énergie libre de l'ensemble des conformères retournés.

attracteur (puisqu'il répand son matériel génétique par croisements et mutations) — en tabou répulsif est un point critique de la stratégie.

Les méthodes d'apprentissage (propagation des solutions connues pour modifier les densités de probabilités) semblent responsables de convergences prématurées de l'algorithme. En effet, en observant les deux graphiques précédents, on s'aperçoit que les stratégies « No tradition » et « Flat distribution » nécessitent beaucoup plus de temps de calcul mais génèrent des solutions meilleures en nombre et en énergies. Bien que la méthode ne soit appliquée que sur une île, il semble que l'introduction de bonnes solutions trop tôt dans l'évolution d'une population soit une mauvaise stratégie.

Le problème ne vient pas tant de l'*information* disséminée (toutes les stratégies ont généré la bonne conformation de la cyclodextrine), mais plutôt du déclenchement abusif du critère d'arrêt. En effet, en biaisant la recherche, on accélère la découverte de minima (figure 3.21) mais on s'expose au risque de longues périodes de stagnations au cours desquelles l'algorithme risque de se terminer. Autrement dit, les stratégies convergent *différemment* ; la figure 3.20 schématise les deux types de profils : l'algorithme bleu converge lentement mais sûrement, tandis que le rouge construit plus rapidement des solutions intermédiaires et intensifie la recherche dans ces régions, mais stagne ensuite, au risque de déclencher le critère d'arrêt.

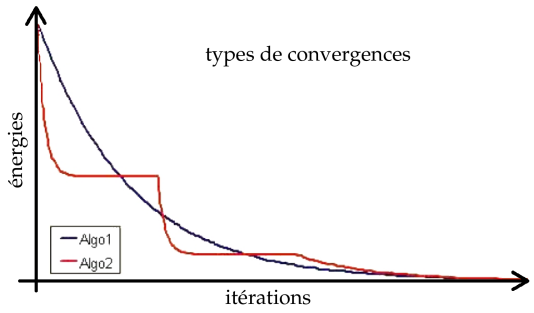


FIG. 3.20: schéma de deux profils d'énergie libre de la population en fonction du nombre de générations.

L'impact des mécanismes de modification des distributions de probabilités a également été mis en évidence à travers des étapes de recherche Monte-Carlo (figure 3.21). On observe que cette heuristique accélère systématiquement la convergence par rapport à une stratégie sans biais ; cependant, à mesure que l'algorithme génère de nouvelles solutions (au fil des générations du μG_A), le pool de solutions disponibles augmente et multiplie le nombre de recombinaisons possibles des $N_{\text{ancêtres}}$.

C'est pourquoi, après 14 générations (courbes jaunes) la convergence est plus lente qu'après 3 (courbe violette).

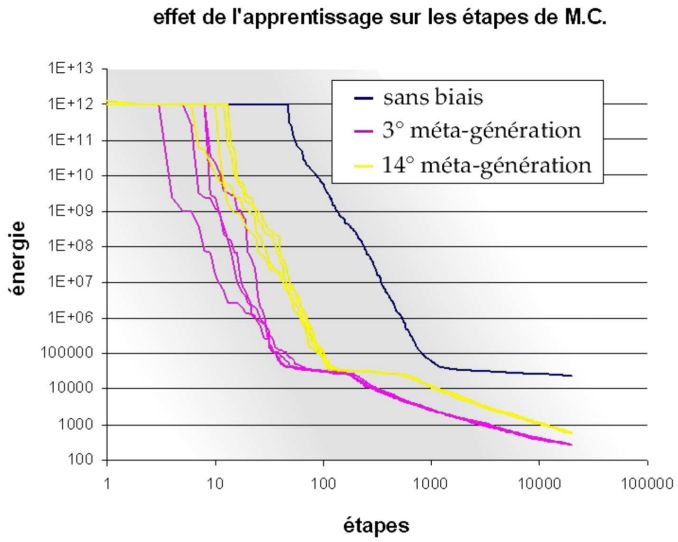


FIG. 3.21: énergie en fonction du nombre d'itérations (échelle log-log) dans des étapes de Monte-Carlo et mise en évidence de l'effet de l'introduction de biais dans les probabilités.

La probabilité de trouver les bonnes solutions se corrèle avec le temps de calcul (en échelle logarithmique), (Cf. figure 3.19), mis à part les deux échecs de la stratégie « No Explorer ». Les stratégies appliquées ne modifient donc pas la vitesse d'exploration de l'espace de phase ; par contre, elles peuvent prévenir des terminaisons trop hâtives de la recherche et éviter à l'évolution de s'enfermer dans certaines régions de l'espace. Cet effet est surtout visible pour les stratégies de modification des densités de probabilités comme discuté ci-dessus.

3.4.5.5 Convergence du μG_A et étude des paramètres internes

Pour analyser les résultats, nous avons utilisé le logiciel Pipeline Pilot software (Scitegic, 2005), qui propose en particulier le greffon de statistique intitulé « Learn Good from Bad » permettant d'estimer l'impact d'un paramètre par rapport aux résultats moyens qu'il induit. Le principe est le suivant : l'ensemble des tests réalisés (les méta-individus) est trié selon le critère de méta-*fitness*, les 10 premiers pourcents de chaque stratégie sont comptabilisés comme réussis (« good ») et les 90 derniers pourcents sont considérés comme des échecs (« bad ») ; ensuite, l'outil évalue l'*avantage* (paramètre entre -1 et 1) d'appliquer chacune des valeurs particulières

aux paramètres.

Notons P_i le i-ième paramètre du méta-individu P et $\mathcal{A}_i$ l'ensemble des valeurs qu'il peut adopter. Par ailleurs, notons $\mathcal{G}$ et $\mathcal{B}$ les sous-ensembles constitués des méta-individus marqués respectivement comme Good et Bad (qui constituent donc une partition de l'ensemble total).

On a d'une part, la probabilité d'avoir un bon méta-individu qui est donnée par :

$$\Pr(P \in \mathcal{G}) = \frac{\sharp\mathcal{G}}{\sharp\mathcal{G} + \sharp\mathcal{B}} = 10\%, \tag{3.7}$$

où « $\sharp$ » représente le cardinal des ensembles, et d'autre part la probabilité de l'événement $(P_i = a)$, pour une valeur $a \in \mathcal{A}_i$:

$$\Pr(P_i = a) = \frac{\sharp(P_i = a)}{\sharp\mathcal{G} + \sharp\mathcal{B}}. \tag{3.8}$$

Si la définition des ensembles $\mathcal{G}$ et $\mathcal{B}$ était indépendante de l'événement $P_i = a$, on aurait alors

$$\Pr(P_i = a | P \in \mathcal{G}) = \Pr(P_i = a). \tag{3.9}$$

Or *a posteriori*, on a

$$\begin{aligned} \Pr(P_i = a | P \in \mathcal{G}) &= \frac{\Pr\left[(P_i = a) \text{ et } (P \in \mathcal{G})\right]}{\Pr(P \in \mathcal{G})} \\ &= \frac{\sharp\left[\mathcal{G} \cap (P_i = a)\right]}{\sharp\mathcal{G}}. \end{aligned} \tag{3.10}$$

L'avantage de l'événement $(P_i = a)$ est alors calculé à partir de la différence de ces deux probabilités. Si une valeur de paramètre n'apporte rien à la qualité de la convergence (événements indépendants), les deux calculs doivent redonner le même résultat et l'avantage sera nul.

Ainsi, le taux de mutation, qui ne peut prendre que les valeurs 1% et 10%, prend plus fréquemment la valeur 10% dans le sous-ensemble des « good » que dans l'ensemble total des simulations (toutes stratégies d'hybridation confondues). Une fréquence de mutation élevée semble donc avantageuse.

Il est à noter toutefois que des phénomènes parasites peuvent se manifester, du fait de l'utilisation séquentielle des solutions précédemment échantillonées pour biaiser les probabilités ou pour définir les zones taboues ; les tendances générales, qui peuvent toutefois être analysées, sont maintenant présentées.

Les grandes populations sont garantes d'un meilleur succès, comme le montre la figure 3.22 ci-dessous. Ceci est à peu près évident, cependant, la complexité en temps de l'algorithme augmente avec la taille de la population ; ainsi, les trop grosses populations sont défavorisées par la méta-évolution lorsque le problème posé est suffisamment simple grâce à la pénalisation proportionnelle au temps de calcul introduit dans l'équation (3.5) du méta-*fitness*.

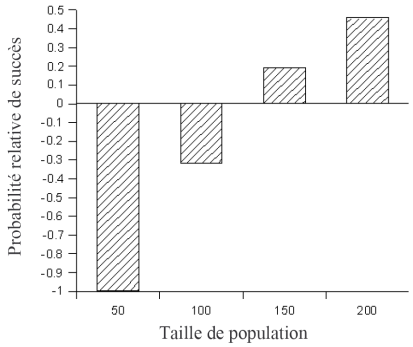


FIG. 3.22: probabilité relative de succès en fonction de la taille de population (toutes stratégies confondues).

Le paramètre de vieillissement semble jouer un rôle plus important pour les stratégies « No Explorers » et « No taboos » (figure 3.23). Dans le premier cas, le torsional angle driving étant désactivé, il est intéressant de voir que le bon compromis de limite d'âge se situe vers 100 générations (relativement grande valeur) tandis que les autres valeurs (sauf 10000) sont clairement défavorables. Pour la stratégie sans tabou, la recherche est intensifiée par rapport aux « No Explorers » ; dans ce contexte, on voit émerger des valeurs plus petites d'âge maximum.

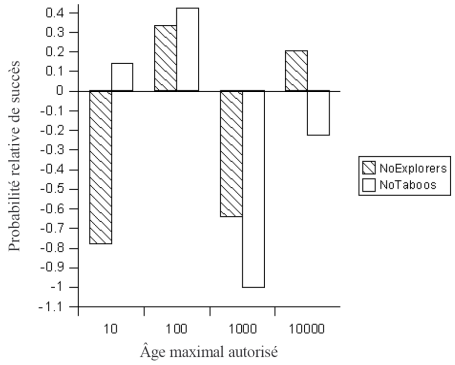


FIG. 3.23: avantages du paramètre de vieillissement pour les stratégies « No Explorers » et « No taboos ».

L'application fréquente d'une heuristique de gradient conjugué (0,3 à 0,5) paraît être utile en général, bien que l'interprétation des graphes soit surtout cohérente dans les stratégies sans explorateurs et sans tabou (figure 3.24). Les explorateurs utilisant pleinement l'idée de gradient conjugué, la stratégie « No Explorers » tend à compenser ce manque en préférant les grandes valeurs. La stratégie « No Taboos » quant à elle défavorise expressément les trop grandes probabilités d'application de gradient, ce qui permet aux individus de ne pas retomber dans les minima déjà occupés et compense ainsi le mécanisme même des tabous.

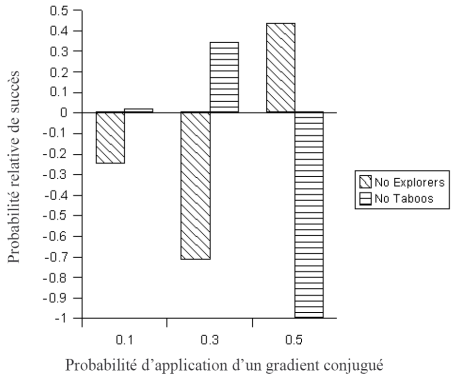


FIG. 3.24: avantage du taux de gradient conjugué sur le succès des stratégies « No Explorers » et « No Taboos ».

La période des apocalypses ne doit pas excéder 1 par 75 générations et cela est d'autant plus vrai pour la stratégie sans tabou pour laquelle il y a moins d'individus aléatoires introduits au cours de l'évolution. La tendance est donc à compenser le manque de « sang neuf » par des redémarrages plus fréquents.

Un filtrage par dissimilitude plutôt sévère semble être favorable dans presque toutes les stratégies (sauf la stratégie sans tabou), sachant que le critère est adaptativement relaxé lorsque la population converge.

Un résumé des valeurs préconisées (ou déconseillées) est fourni, tableau 3.3, afin de faciliter la lecture pour une réutilisation éventuelle (seuls les paramètres pour lesquels il a été possible de conclure y apparaissent).

Valeurs possibles	Paramètre
2, 3 ou 4	nombre d'îles
5, 10, 25 ou 50	période de migration (en nombre de générations)
500, 800 ou 1000	nombre maximum de générations sans succès avant arrêt global
50, 75 ou 100	nombre maximum de générations sans succès avant apocalypse
50, 100, 150 ou 200	taille de population
0 ou 1	nombre d'élites immortels
20, 50, 100 ou 200	âge maximum toléré
1% ou 10%	fréquence de mutations
10, 30 ou 50%	probabilité d'application d'une relaxation par gradient conjugué
75, 80, 85 ou 90%	niveau de similarité maximal dans la population

TAB. 3.3: résumé des valeurs préconisées (vertes) et déconseillées (rouges) pour certains paramètres de contrôle de l'algorithme.

3.5 Vers une validation à plus grande échelle

3.5.1 Les molécules utilisées

En vue d'appliquer la procédure sur des problèmes de plus grandes tailles et puisque nous bénéficions d'un outil générique, nous avons établi un nouveau jeu de molécules parmi lesquelles figurent :

- un mini peptide (code PDB 1UAO), avec $N_{\text{ddl}} = 32$ degrés de liberté,
- le peptide « Tryptophan zipper » (code PDB[16] 1LE1), également utilisé par Okur *et al.* pour tester l'extensibilité du champ de force[17] pour des molécules plus grandes, $N_{\text{ddl}} = 54$,
- une « proto »-hélice, covalemment modifiée, appelée CRH, $N_{\text{ddl}} = 72$,
- le peptide « Tryptophan cage » (code PDB 1L2Y), $N_{\text{ddl}} = 73$,
- le domaine WW de la protéine humaine « PIN1 », $N_{\text{ddl}} = 140$.

Toutes ces molécules ont la particularité de se structurer en solution et d'avoir été étudiées expérimentalement[18], de sorte que des données sont disponibles et peuvent servir pour comparer nos résultats. Nous avons également gardé la cyclodextrine ($N_{\text{ddl}} = 65$) afin de s'assurer que les développements futurs n'allaient pas se faire au détriment des performances précédemment validées.

[16] Protein Data Bank, http ://www.rcsb.org/pdb/index.html

[17] les champs de forces ff94 et ff99 intégrés à Amber.

[18] ces deux conditions restreignent beaucoup les choix possibles et il existe assez peu d'exemples utilisables

3.5.1.1 Détail des molécules

Le mini-peptide 1UAO (figure 3.25) comporte 10 acides aminés et est un des plus petits assemblages peptidiques connus pour se structurer en solution.

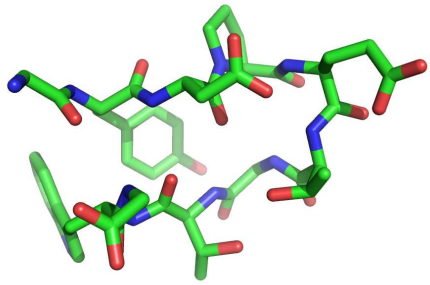


FIG. 3.25: strucure spatiale du peptide 1UAO.

Le « tryptophan zipper » est la plus petite épingle connue ayant un tel niveau de rigidité. Il appartient à une famille de peptides conçus artificiellement (Cochran *et al.*, 2001), voir figure 3.26. Il est composé de 12 acides aminés dont quatre tryptophanes qui s'intercalent (à la manière d'une fermeture éclair) et stabilisent nettement la conformation grâce à des interactions de type aromatique-aromatique. Étant donné l'importance qu'a pris ce petit peptide dans la littérature, à la fois dans les études expérimentales et par simulations, une revue des principaux articles le concernant (modélisation et approches expérimentales) est proposée en annexe B.

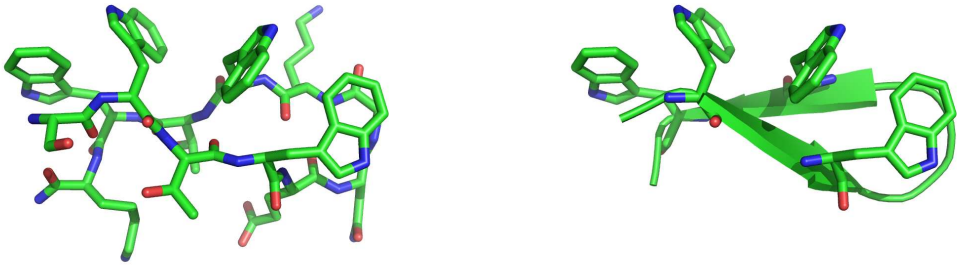


FIG. 3.26: (gauche) structure en bâtonnets de 1LE1, (droite) mise en évidence de la stabilisation du tournant grâce aux interactions entre les tryptophanes.

Le CRH (Conformationally Restrained Helix) est une chaîne polypeptidique de 17 acides aminés ayant subi une modification covalente qui crée un cycle à une extrémité de la molécule (figure 3.27). Ce cycle contraint la conformation du premier pas de l'hélice et induit la conformation hélicoïdale sur toute la chaîne. Pour son échantillonnage, nous avons ouvert le cycle.

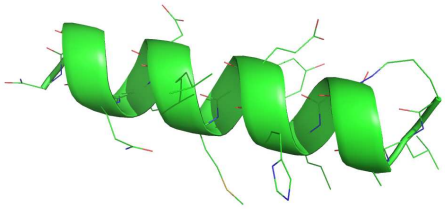


FIG. 3.27: structure de l'hélice CRH, la modification covalente apparaît à l'extrémité droite.

Le « tryptophan cage » est un motif polypeptidique obtenu par mutations et simplifications de structures existantes (Neidigh *et al.*, 2002). Ses 20 acides aminés se structurent de telle sorte que l'unique résidu tryptophane soit enfoui au cœur de l'édifice et ait un accès réduit au solvant. Ce collapsus hydrophobe (mis en évidence par RMN par Mok *et al.*, 2007) est à l'origine de son repliement extrêmement rapide : 4 ms (Kubelka *et al.*, 2004). Par ailleurs, il comporte trois motifs structuraux : deux hélices dont une ne forme qu'un seul tour et un brin étendu couvrant l'ensemble (figure 3.28). Ce peptide a souvent servi de modèle pour des simulations : Schug *et al.* en 2004(b) ont réalisé une simulation all-atom par « tempering method », la même équipe en 2005(b) a appliqué et comparé plusieurs méthodes.

FIG. 3.28: structure géométrique du « tryptophan cage ».

Enfin, le domaine de liaison de PIN1, comportant 34 acides aminés, se présente sous la forme de trois feuillets β maintenus par des ponts hydrogène (figure 3.29) (Nguyen *et al.*, 2005; Jäger *et al.*, 2006). Comme il contient deux tryptophanes bien conservés, on le désigne par le nom « WW ».

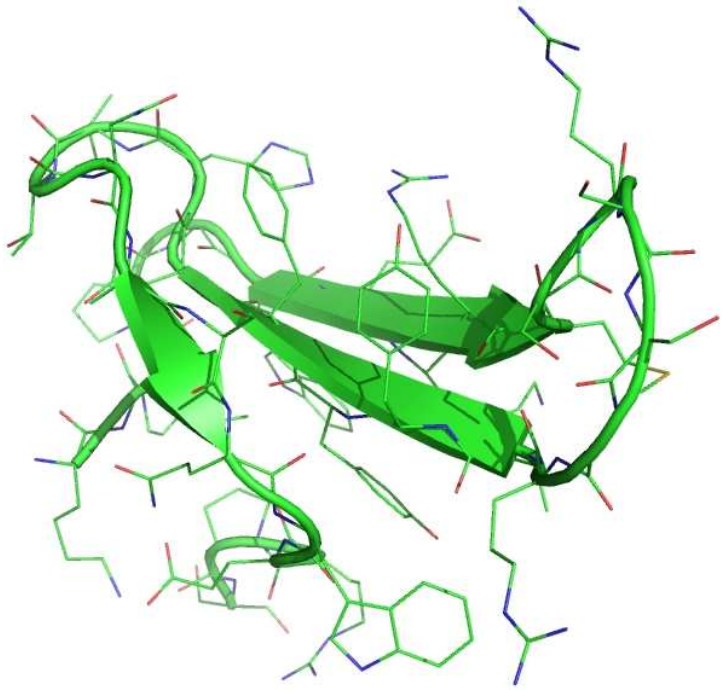


FIG. 3.29: structure du domaine WW de la PIN1.

3.5.1.2 Un échantillonnage partiel

Enfin, puisqu'il y a relativement peu de systèmes « abordables » (structure clairement définie, données expérimentales disponibles et taille restreinte), l'idée[19] de faire un échantillonnage partiel de molécules plus grandes, où seuls certains degrés de liberté seraient activés, a été implémentée (introduisant, au besoin, des coupures formelles de certaines liaisons). La figure 3.30 présente quatre exercices d'échantillonnage sur des parties de la PIN qui ont été soumis à l'algorithme.

3.5.2 Premiers constats

3.5.2.1 Un besoin d'intensification

Avec ces molécules, les volumes des espaces de phase deviennent difficiles à gérer et les temps de calculs de plus en plus long (plusieurs semaines). Il faut donc reconnaître que la stratégie pour de tels problèmes commence à saturer. Toutefois l'algorithme est toujours capable d'échantillonner *en largeur* les paysages énergétiques et, en particulier, il visite systématiquement la région native (sauf pour le

[19] nous remercions le Docteur L. Serrano pour l'idée originale.

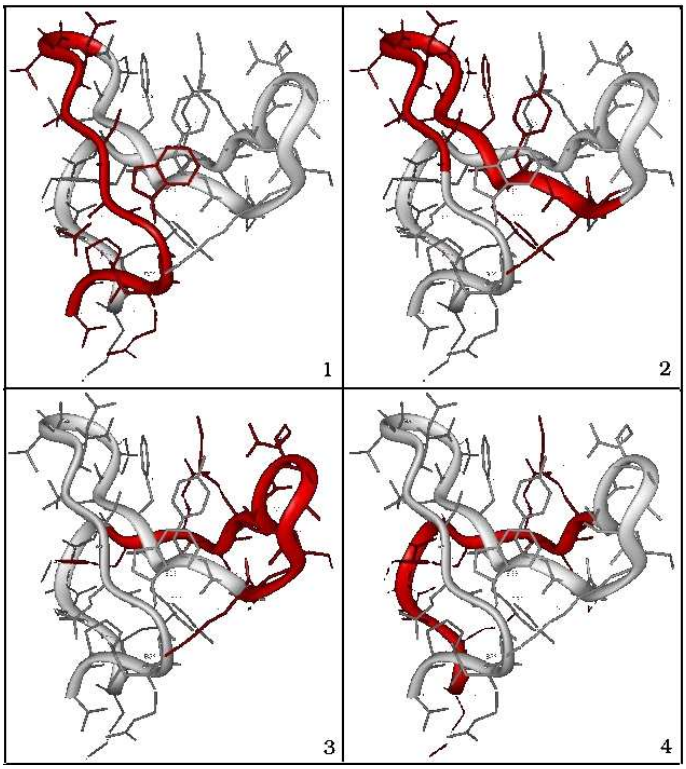


FIG. 3.30: PIN1 échantillonnée par morceaux : les parties blanches sont fixées tandis que les degrés de liberté des parties rouges sont optimisés.

cas de PIN1). Malheureusement, nous avons observé que l'algorithme trouve, en un temps convenable, des solutions plus ou moins proches de la structure native, mais avec quelques réarrangements qui expliquent des différences énergétiques parfois importantes. Ainsi, ces solutions sont rejetées car plus énergétiques que d'autres géométries, non-natives, mais sans mauvais contact. Nous allons donc chercher à intensifier la recherche dans les vallées visitées.

La première modification a été d'autoriser, lorsqu'on génère des individus aléatoires, à reprendre des morceaux d'anciennes solutions en réalisant un croisement d'ancêtres. Cela permet de réintroduire des gênes (plusieurs codons contigus) potentiellement favorables et d'intensifier la recherche autour des solutions précédemment échantillonnées. Cette stratégie a été implémentée suite à la frustration de voir apparaître parmi les solutions, deux moitiés de molécules bien repliées ; toutefois, un nouveau PARAMÈTRE DE CONTRÔLE a été introduit pour modérer cet effet.

Introduire une solution trop bonne, tôt dans l'évolution d'une population, est vivement déconseillé car cela entraîne généralement une convergence prématurée suite à la dissémination du chromosome à travers la population. Au contraire, cette stratégie s'est montrée efficace en introduisant « discrètement » des morceaux de solutions,

tout en préservant la progression de la population. En outre, elle permet d'intensifier la recherche autour des solutions ré-utilisées.

3.5.2.2 Interprétation des résultats expérimentaux

Deuxième constat : pour des molécules de cette taille, nous voyons apparaître des géométries plus stables que la géométrie native (de meilleures énergies). La conformation expérimentale ne correspond donc pas au minimum absolu de l'espace de phase, ce qui contredit l'hypothèse thermodynamique exposée au chapitre 1.

Il existe à cela plusieurs explications, la première étant que la « géométrie expérimentale » est en fait issue d'un processus de détermination complexe. Seules les *données* sont expérimentales, elles sont interprétées et des algorithmes[20] sont en charge de trouver des géométries qui satisfont à ces contraintes. Il n'y a généralement pas une seule solution, mais une famille de solutions qui dénote la flexibilité de la molécule. Enfin, les géométries trouvées sont généralement minimisées selon le critère énergétique estimé à partir d'un champ de forces quelconque.

De ce processus de détermination de la structure, il découle plusieurs sources d'erreurs potentielles (en plus des difficultés inhérentes aux technologies, aux techniques de synthèse et de purification) :

- l'effet de moyenne sur l'ensemble de Boltzmann au cours de l'expérience peut rendre l'interprétation difficile. Ainsi, deux sous-populations de conformations distinctes peuvent générer des contraintes expérimentales impossibles à concilier.
- La minimisation selon un champ de forces différent du nôtre peut entraîner quelques différences se traduisant par une pénalisation énergétique.

Lorsque la molécule est petite, le minimum local peut être retrouvé par une simple optimisation par gradient, cependant, pour des problèmes de plus grandes tailles, cela ne suffit plus (voir figure 3.31). Pour résoudre ce problème, nous avons donc soumis les solutions natives à un processus de recuit simulé qui permet de visiter le voisinage de la « géométrie native ». De cette façon, nous reconstruisons un *ensemble* de solutions natives et nous caractérisons correctement leurs énergies.

Malgré cela, l'algorithme retournait encore des solutions de plus basses énergies que l'énergie du natif optimisé. Nous reviendrons sur ce point ultérieurement ; nous détaillons maintenant rapidement la stratégie d'échantillonnage local utilisée, qui

[20] typiquement basés sur le *distance geometry*, voir section 2.2.4 ou le *torsional angle driving*, section 3.4.3.2.

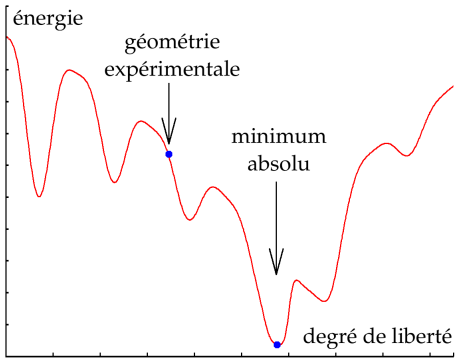


FIG. 3.31: la solution expérimentale diffère légèrement du minimum absolu et une optimisation par gradient ne suffit pas toujours à le retrouver...

a dû être adaptée pour prendre en compte l'aspect extrêmement rugueux du paysage énergétique. Puis nous présenterons l'heuristique d'intensification, basée sur les potentiels de forces moyennes.

3.5.3 Détails de l'échantillonneur local

Une simple optimisation par gradient conjugué avait initialement été envisagée en réalisant une minimisation 3D grâce au logiciel insight (Accelerys, 2005), mais la rugosité de la fonction énergie rend ce type d'approche inefficace sur des systèmes de cette taille.

Notre stratégie s'inspire donc essentiellement du recuit simulé avec un ou plusieurs cycles de chauffe et refroidissement (voir paragraphe page 3.2.3). Les pas sont générés en modifiant toutes les variables avec une densité uniforme sur l'hypercube entourant la solution courante. La taille de cet hypercube est adaptée au cours de la recherche afin de tenir compte du comportement local de la fonction énergie. Enfin, puisque le paysage énergétique est extrêmement accidenté, il a été nécessaire de coupler ce recuit simulé avec une relaxation par gradient, appliquée systématiquement après chaque pas. Cette optimisation, bien que limitée en nombre d'itérations, reste la partie la plus gourmande en ressources.

Des tests ont été faits sur l'application d'une stratégie originale appelée leapfrog (Ishwaran, 1999). Elle consiste à diviser les sauts d'une conformation à l'autre en plusieurs petits pas, en modifiant progressivement la trajectoire en fonction du gradient en chacun de ces pas. Cette stratégie est supposée donner moins de solutions aberrantes.

Malgré cela, la rugosité de l'hypersurface d'énergie est telle que les « sauts de grenouille » aboutissent le plus souvent à des énergies beaucoup trop grandes pour

être acceptées. L'application d'un gradient conjugué après chaque saut fournit les mêmes résultats que dans la stratégie initiale, de sorte que l'heuristique a été abandonnée. De plus, les évaluations du gradient au cours des sauts ralentissent d'autant la progression globale.

Cet échantillonneur « peuple » la région autour de la conformation déterminée expérimentalement et fournit en particulier l'énergie du meilleur minimum local avoisinant.

3.5.4 La fragmentation

Étant données les performances de l'algorithme sur les molécules plus grandes, nous avons cherché à améliorer la balance entre diversification et intensification en faveur de cette dernière.

En s'inspirant de la stratégie « divide and conquer » et de ce qui a été fait sur les modifications des densités de probabilité pour chaque variable du vecteur de torsions, nous avons développé une nouvelle heuristique basée sur la fragmentation des molécules. L'idée est de fractionner le grand problème en petites tâches, plus simples, puis de réunir les éléments afin de construire une solution globale. L'hypothèse sous-jacente est que les degrés de liberté ne sont que peu influencés par les atomes topologiquement éloignés. Cette hypothèse est certainement vraie en première approximation (par exemple dans les hélices et les tournants des protéines), mais est sujette à caution puisque le repliement global de la molécule peut permettre à deux extrémités topologiquement éloignées de se rapprocher et d'interagir. Cette stratégie n'est toutefois pas nouvelle, puisqu'elle reprend les idées de l'utilisation statistique de bases de données de molécules connues. C'est le cas des couples d'angles (ϕ, ψ) de torsions des squelettes protéiques (Ramachandran et Sasisekhan, 1968) ou des bases de rotamères (Shetty *et al.*, 2003). L'avantage de notre approche, c'est qu'elle ne fait pas intervenir de connaissances sur d'autres molécules, mais apporte de l'information sur le comportement local de la molécule étudiée.

Nous présentons maintenant chacune de ces deux étapes de fragmentation et de reconstruction de la géométrie globale.

3.5.4.1 Méthode de fragmentation

De la même manière que nous avions estimé les densités marginales de chaque torsion, nous avons généralisé l'approche à l'estimation de densités marginales à *plusieurs* variables (figure 3.32).

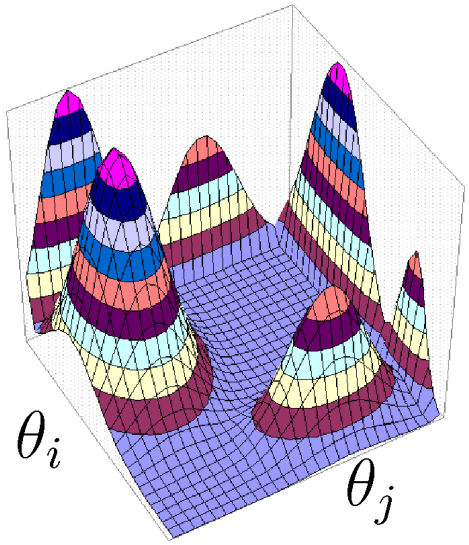


FIG. 3.32: schéma de densités marginales à une et deux variables.

Ces marginales concernent des sous-ensembles de k torsions ($k = 4$, 5 ou 6 dans les tests), topologiquement contiguës en choisissant préférentiellement celles qui ont les poids les plus importants (voir pondération des degrés de liberté, page 104). Elles définissent un fragment de molécule qui est échantillonnable par le $C_S G_A$, cependant, si l'on procède de la sorte, les effets de bords seront importants et l'échantillonnage sera biaisé. Pour éviter cela, chaque fragment est regarni des atomes qui l'entourent (topologiquement proches). Ainsi, aux k torsions, que nous qualifierons de *clefs*, nous avons ajouté tous les atomes *environnementaux*, dans un ellipsoïde basé sur la distance topologique : on choisit, dans le fragment F, deux atomes (a_1, a_2) impliqués dans des torsions clefs, qui maximisent la distance topologique $d_t(a_1, a_2)$ (ce choix n'est pas toujours unique). Un atome a de la molécule est alors inclu dans l'environnement de F (notée $\bar{F}$) si la somme des distances aux foyers est inférieure à $d_t(a_1, a_2)$ plus un paramètre à définir (voir figure 3.33 et équation (3.11)).

$$a \in \bar{F} \Leftrightarrow d_t(a, a_1) + d_t(a, a_2) \leq d_t(a_1, a_2) + 2d_0, \tag{3.11}$$

où nous avons testé les valeurs $d_0 = 4$ et $d_0 = 6$.

La procédure de fragmentation, résumée sur la figure 3.34, est entièrement automatisée.

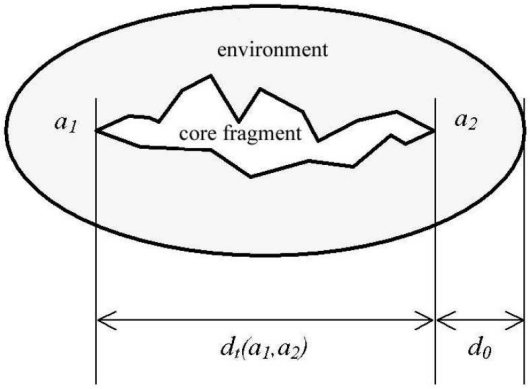


FIG. 3.33: définition de l'environnement d'un fragment.

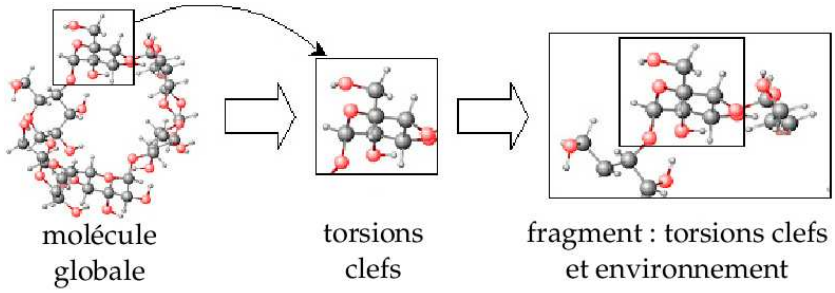


FIG. 3.34: exemple de la construction d'un fragment de la cyclodextrine.

3.5.4.2 Réunion des fragments

En théorie, la densité marginale pour le k-uple de torsions clefs de F (p_F) nécessite l'échantillonnage sur toutes les variables qui ne sont pas dans F (Koehl et Delarue, 1996), mais d'après notre hypothèse, seuls les degrés de liberté hors de F mais dans $\bar{F}$ interviennent :

$$\begin{aligned} p_F(\theta_{i_1}, \ldots, \theta_{i_k}) &= \int_{\bar{F}-F} p_{\bar{F}}(\theta_{i_1}, \ldots, \theta_{i_k}, \theta_{j_1}, \ldots, \theta_{j_m}) d\theta_{j_1} \ldots d\theta_{j_m} \\ &\approx \int_{\bar{F}-F} p(\theta_1, \ldots, \theta_{N_{\text{ddl}}}) d\theta_{j_1} \ldots d\theta_{j_m}, \end{aligned} \tag{3.12}$$

où $(\theta_{j_1}, \ldots, \theta_{j_m})$ sont les variables environnementales.

Enfin, pour reconstruire $p_F(\theta_{i_1}, \ldots, \theta_{i_k})$, nous n'échantillonnons pas les torsions environnementales à torsions clefs fixées, mais échantillonnons toutes les torsions clefs et environnementales et utilisons l'approximation de Monte Carlo (équation (2.12) rappelée ici) :

$$p_F(x) \approx \frac{1}{N_{\text{éch}}} \sum_{x_i \in \mathcal{E}_{p_F}} \delta(x = x_i), \tag{3.13}$$

où $\mathcal{E}_{p_F}$ est un échantillonnage de l'espace selon la densité p_F, de cardinal $N_{\text{éch}}$.

En échantillonnant le fragment regarni : $\bar{F}$, nous pouvons donc estimer la densité marginale de F. Cette stratégie peut même éventuellement ne servir qu'à écarter les régions aberrantes de l'espace de phase ; c'est particulièrement le cas lors de l'échantillonnage des petits cycles (lorsqu'ils sont englobés dans $\bar{F}$), nous le verrons clairement à travers l'exemple de la cyclodextrine.

Pour reconstruire des solutions globales, nous avons repris la méthodologie des densités biaisées *par variable* : l'algorithme est exécuté sur la molécule entière, mais choisit les fragments en respectant les densités marginales estimées précédemment. Une probabilité uniforme est toujours mélangée afin d'éviter l'interdiction de recherches dans certaines régions de l'espace (selon un PARAMÈTRE opérationnel).

3.5.4.3 Résultats

Afin d'analyser le bien fondé de la méthode, nous avons voulu vérifier que, pour chacun des N_{frgs} fragments, la conformation native avait bien été retrouvée parmi les solutions envisagées par l'algorithme d'échantillonnage local. En effet, dans ce cas il ne reste plus, à l'échantillonneur global, qu'à trouver, pour chaque fragment,

la bonne configuration parmi les N_{sols} proposées. Ceci donne des tailles d'espaces de recherche pour chaque fragment F_i de l'ordre de $\left(\frac{360}{\text{pas}}\right)^{N_{\text{ddl}}(\bar{F}_i)}$ et une taille d'espace global de l'ordre de ${N_{\text{sols}}}^{N_{\text{frgs}}}$, qu'il faut comparer à $\left(\frac{360}{\text{pas}}\right)^{N_{\text{ddl}}}$ sans fragmentation.

Prenons, par exemple, le cas de la « tryptophan cage » (N_{ddl}= 73) avec $(k, d_0) = (6, 6)$: il y a 15 fragments comportant entre 6 et 24 degrés de liberté chacun. Le travail d'échantillonnage des fragments (c'est-à-dire la taille des espaces de recherche) est donc d'un ordre 10^{43} (avec un pas de 6°), tandis que la recombinaison des solutions ($21 \leq N_{\text{sols}} \leq 9117$) demande un travail en 10^{41}. Sans fragmentation, le nombre total de conformations envisageables est de l'ordre de 10^{130}...

En réalité, le calcul ci-dessus n'offre qu'un ordre de grandeur car on utilise également les niveaux d'énergies des solutions partielles pour pondérer leurs probabilités (équation (2.15), p. 87). Ceci nous permet d'évaluer un « *facteur d'enrichissement* » apporté par la procédure de fragmentation, qui est défini de la manière suivante : c'est le rapport de la nouvelle probabilité de la conformation native par rapport à une distribution complètement aléatoire.

Cette analyse est faite pour différentes valeurs du couple de paramètres (k, d_0) et pour les différentes molécules dont nous disposons. Le facteur d'enrichissement est classé selon cinq catégories comme indiqué dans la légende (figures 3.35 et 3.36).

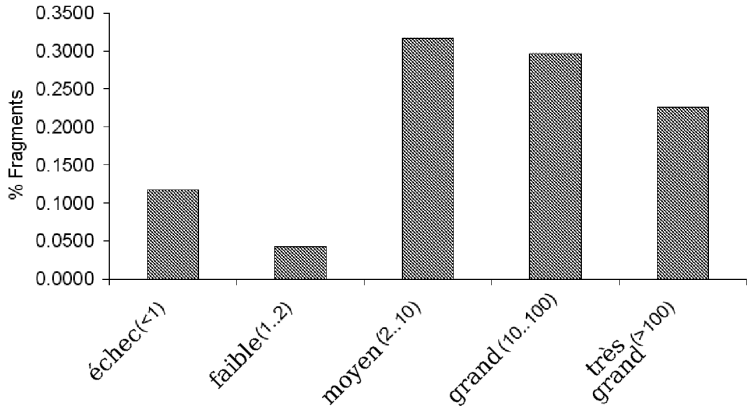


FIG. 3.35: répartition du facteur d'enrichissement pour la « tryptophan cage », avec les paramètres $(k, d_0) = (5, 4)$.

Selon toute attente, l'opération de fragmentation est d'autant plus réussie que la taille du fragment est grande. Étonnamment, augmenter d_0 indépendamment de k, ne semble pas particulièrement intéressant pour les fourchettes de valeurs que nous avons considérées.

Pour la cyclodextrine en particulier, on voit que l'échantillonnage local de ses cycles de glucose permet d'acquérir une connaissance précise qui rend la stratégie

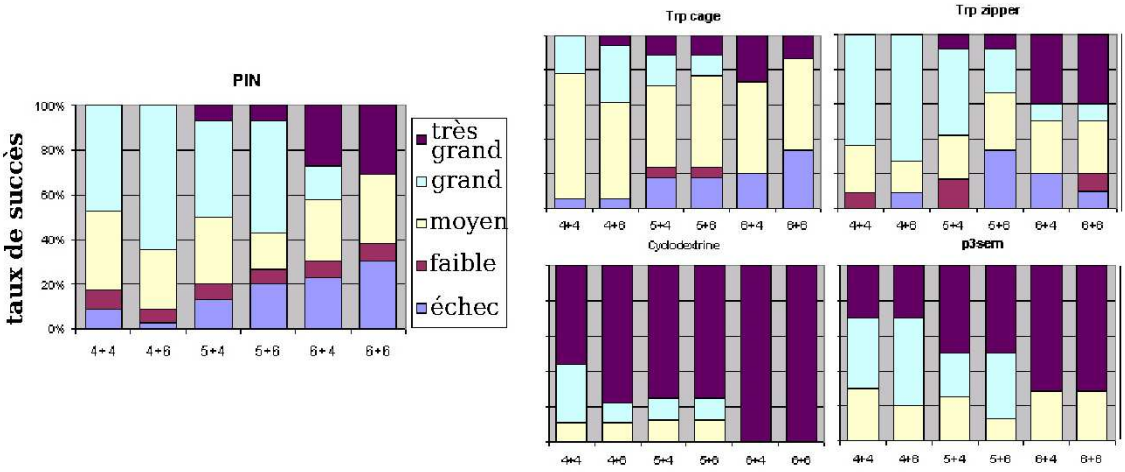


FIG. 3.36: taux de réussite en fonction des paramètres (notés $(k + d_0)$) pour cinq molécules (voir légende).

prometteuse. La même conclusion peut être faite sur « p3sem » qui est une petite molécule organique. Pour les plus grandes molécules, la stratégie reste tout à fait encourageante ; on note toutefois que la fragmentation de structures de type feuillets β présente plus de risques, surtout lorsque les fragments sont grands. Cela s'explique par le fait que les fragments ne peuvent pas former les ponts hydrogène qui stabilisent la structure générale de la molécule. Nous notons également que dans ce cas, les fragments de plus petites tailles échouent moins souvent : l'algorithme échantillonne des espaces plus petits (la fonction énergie a moins d'amplitude) et met plus facilement en évidence les régions probables et les régions aberrantes.

3.6 Parallélisation de l'algorithme

Afin de réduire les temps de calculs, on peut également envisager l'utilisation de matériel informatique plus performant. Une des thématiques importantes de ces dernières années, est de devancer l'optimisation des composants informatiques, en regroupant les ressources existantes et en les faisant calculer de concert. Un énorme travail d'orchestration a été réalisé dans ce domaine permettant une utilisation quasi-transparente de « *grilles* » d'ordinateurs à travers différentes couches qui correspondent à différents niveaux d'abstraction (Cahon *et al.*, 2004).

Afin de pouvoir avancer vers ce type de déploiements, nous avons démarré un projet commun, surnommé Dock@GRID pour « **conformational sampling and molecular docking on grids** », avec l'équipe OPAC du Laboratoire d'In-

formatique Fondamentale de Lille (LIFL) et le Commissariat à l'Énergie Atomique (CEA) ayant donné lieu à un financement ANR[21] fin 2005 (voir son site : http ://dockinggrid.gforge.inria.fr/index.html, consulté en août 2007) et impliquant les personnes suivantes :

- Sylvaine Roy, Ingénieur Chercheur CEA iRTSV/LBIMCEA[22],
- El Gazali Talbi, Professeur, LIFL, responsable de l'équipe OPAC,
- Nouredine Melab, Professeur, LIFL,
- Alexandru-Adrian Tantar, doctorant, LIFL,
- Jean-Charles Boisson, doctorant, LIFL,
- Gaël Evan, ingénieur de recherches, LIFL,
- Dragos Horvath, Chargé de Recherches, UGSF,
- Benjamin Parent, doctorant, UGSF.

3.6.1 L'environnement de GRID5000

GRID5000 est un exemple de grilles de calcul, elle est répandue à travers toute la France sur neuf sites : Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Sophia-Antipolis et Toulouse, et est soutenue par le CNRS et l'INRIA. La connection des unités de calcul est assurée par le réseau académique français RENATER[23]. Cette grille est munie des environnements suivants :

- Condor[24],
- MW (Master-Worker),
- ParadisEO[25] (Parallel distributed Evolving Objects).

Le système condor permet d'administrer des parcs hétérogènes d'ordinateurs en mode multi-utilisateurs. Il gère automatiquement le recrutement de ressources supplémentaires, les disponibilités des machines (en scrutant l'activité des périphériques : claviers, souris) et libère les machines lorsqu'un utilisateur s'y connecte physiquement. Enfin, il autorise de nombreux points de contrôle permettant de vérifier et de sauvegarder les calculs en cours, afin de pouvoir les reprendre en cas d'interruption ou d'échecs.

Le logiciel MW fait partie de ce qu'on appelle les « *middlewares* », car il offre un niveau d'abstraction intermédiaire. Il permet un développement facilité d'appli-

[21] http ://www.gip-anr.fr, consulté en août 2007
[22] http ://www-dsv.cea.fr/lbim/iacg, consulté en août 2007.
[23] http ://www.renater.fr, consulté en août 2007.
[24] http ://www.cs.wisc.edu/condor/condorg, consulté en août 2007.
[25] http ://paradiseo.gforge.inria.fr/index.php, consulté en août 2007.

cations de type « maîtres/esclaves » grâce à un ensemble de classes C++. De plus, il assure la gestion des échecs (calculs, transmission, ou libération de la ressource) et relance au besoin les processus sur d'autres machines.

Enfin, ParadisEO est une librairie C++ opensource (sous GPL[26]) offrant un cadre de développement transparent pour les applications sur la grille. Elle est le dernier étage d'abstraction de l'architecture (voir figure 3.37). Elle fournit un grand nombre d'heuristiques de recherches parallèles, tant pour l'intensification locale de solutions que pour l'exploration globale. Presque toutes les stratégies classiques de la littérature sont déjà implémentées, mais il est possible d'ajouter ses propres heuristiques (et en particulier, sa propre fonction de *fitness*...) ainsi que tout opérateur adapté à la physique du problème.

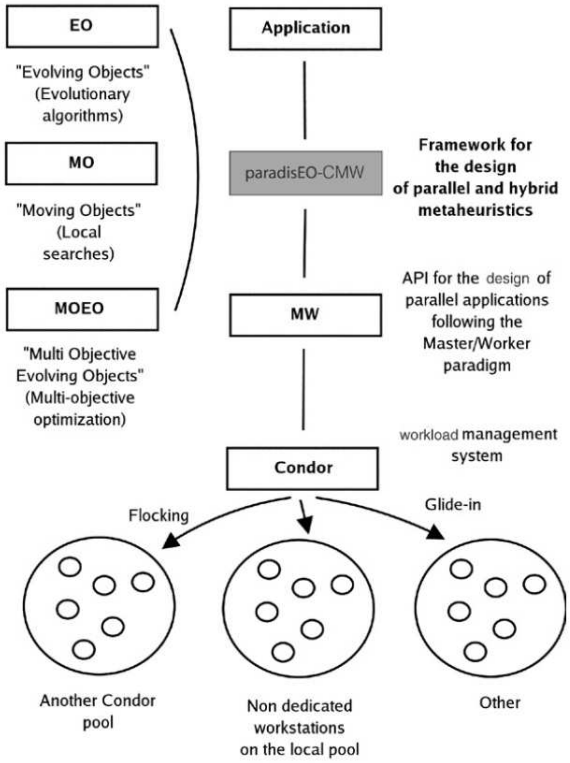


FIG. 3.37: différents niveaux d'abstraction dans l'architecture utilisée.

La première étape a été de traduire notre programme principal en C++, ce qui a été l'objet du stage de DEA de Samuel Hoareau. Cependant, un énorme travail d'adaptations au langage hiérarchisé en classes fût — et est encore — nécessaire pour pouvoir fonctionner optimalement. L'article de Tantar *et al.*, paru en 2007 dans « Future Generation Computer Systems », présente les premières validations

[26]Gnu General Public License

du code (annexe E).

L'algorithme a été testé sur une fourchette de 1 à 80 ordinateurs (et récemment sur 200). Dans un cas idéal, le gain de temps est donné par le nombre de processeurs utilisés ; le temps nécessaire pour un tel algorithme parallélisé sur N_{CPU} machines est alors réduit d'un facteur N_{CPU}. Mais concrètement, une trop forte parallélisation multiplie les coûts de commuication et réduit les performances. Pour évaluer cela, on définit le critère de « SpeedUp » comme étant le rapport de la somme des temps de calcul sur chacun des ordinateur utilisé, par le temps nécessaire pour exécuter l'algorithme sur une seule machine. La figure 3.38 présente l'évolution de ce critère pour la cyclodextrine et la « tryptophan cage » en fonction du nombre de CPU utilisées.

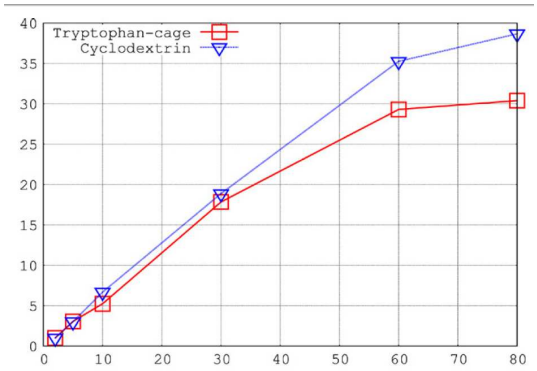


Fig. 3.38: SpeedUp en fonction du nombre d'ordinateurs utilisés, pour la cyclodextrine et 1L2Y.

En parallèle de ces travaux, nous avons commencé à concevoir une stratégie de plus bas niveau (utilisant MW), permettant un déploiement des heuristiques précédemment exposées ainsi que du méta-AG. Puisque chaque processeur gérait plusieurs îles (ou continents), nous avons repensé nos algorithmes sous la forme d'un modèle planétaire, où chaque planète-processeur recrée localement un modèle insulaire...

3.6.2 Une stratégie dédiée à la grille : le modèle planétaire

Afin de limiter les temps de communications et d'adapter la parallélisation à la structure matérielle sous-jacente, nous avons généralisé notre implémentation n'utilisant que quelques îles (et un explorateur indépendant), par un modèle « planétaire » abritant chacun plusieurs îles (et un explorateur) ; une planète correpondant à un processeur de calcul.

3.6.2.1 Une optimisation asynchrone des paramètres opérationnels

Ce modèle nous permet de paralléliser l'évaluation du méta-*fitness* mais nous oblige à gérer les méta-individus de manière asynchrone pour éviter les temps d'inactivité. Les croisements et mutations de vecteurs de paramètres opérationnels se font donc « à la demande », en gardant en lice les réglages les plus productifs. Dès qu'une planète a accompli son travail (d'après son critère d'arrêt sur toutes ses îles ou bien à l'approche de la fin du temps alloué par la grille), les solutions échantillonnées rejoignent le pool universel de solutions (en écartant les redondances). Le méta-*fitness* est alors calculé et le méta-individu classé parmi ses semblables ; le processeur signale alors son inactivité dans l'attente d'un nouveau méta-chromosome à évaluer, proposé par le dispatcheur central.

3.6.2.2 La panspermie

Les îles communiquent entre-elles, de façon limitée, grâce au processus d'émigration, tandis que les planètes travaillent en autarcie totale. Néanmoins, les C_5G_A sont initialisés avec un fichier réunissant quelques solutions précédemment échantillonnées qui leur sert soit de tabous, soit d'attracteurs. Cette stratégie, baptisée « *panspermie* » en accord avec la théorie selon laquelle la vie sur la Terre aurait été inséminée par des micro-organismes extra-terrestres, est appliquée,

- en utilisant l'heuristique tabous, pour forcer la diversification (voir critère de distance, équation (3.14)),
- ou par le biais des croisements d'ancêtres (§ 3.5.2), pour attirer la recherche dans une zone à caractériser finement.

3.6.2.3 Stratégie d'intensification

Or, comme nous l'avons vu, l'algorithme est capable de localiser rapidement les régions prometteuses de l'espace de recherche — dont la région native — mais échoue à caractériser correctement leurs énergies. En effet, le paysage d'énergie est tellement accidenté, que certains détails de la géométrie engendrent parfois de grandes différences énergétiques[27]. En d'autres termes, la découverte du minimum absolu d'une région donnée est loin d'être triviale et nécessite d'importants efforts d'intensification. Nous avons donc dédié certaines planètes à une recherche spécifique autour de solutions connues, ce qui est réalisé en initialisant directement les populations avec

[27] l'idéal serait de caractériser systématiquement un petit domaine autour des conformations proposées ; cette idée est en cours de développement.

toutes les solutions du pool universel appartenant à une même sous-région restreinte de l'espace.

Ce clustering est effectué selon le critère de distance ci-dessous, équation (3.14) (utilisant la pondération des degrés de liberté) et une limite de distance $\mathfrak{D}_{\max}$:

$$\mathfrak{D}(\Theta^0, \Theta^1) = \sum_{i \leq N_{\text{ddl}}} \omega_i \Delta(\theta_i^0, \theta_i^1), \tag{3.14}$$

où la fonction Δ renvoie l'angle entre ses arguments, en prenant en compte la 2π-périodicité de l'espace de départ. Les conformations les moins énergétiques sont choisies comme centres pour les clusters, qui peuvent éventuellement évoluer en fonction de l'apparition de minima plus profonds dans le voisinage considéré.

On autorise alors une « région prometteuse » à être intensifiée un certain nombre de fois (paramètre N_{intens} que l'utilisateur doit définir) tandis que les autres planètes doivent éviter toute recherche dans cette zone.

La définition des paramètres $\mathfrak{D}_{\max}$ et N_{intens} est un point particulièrement sensible de notre stratégie car

- trop grands, les clusters seraient difficiles à échantillonner alors que trop petits, ils deviennent rapidement très nombreux et difficiles à gérer ;
- avec des petites valeurs de N_{intens}, la recherche risque de manquer le minimum absolu et l'échantillonnage pourrait être incomplet (la région devenant ensuite taboue), tandis que les grandes valeurs de N_{intens} réquisitionnent beaucoup de ressources informatiques ne pouvant plus être utilisées à d'autres tâches.

Par défaut, nous avons fixé $N_{\text{intens}} = 5$, toutefois, si le cluster évolue par suite de l'apparition de nouveaux minima plus profonds, l'intensification reprend ; la région n'est déclarée taboue qu'après N_{intens} recherches infructueuses.

3.6.2.4 Résultats

Pour tester le modèle planétaire, nous l'avons appliqué à trois problèmes : le « triptophan zipper » (1LE1), le « triptophan cage » (1L2Y) et un des tournants du domaine WW de la PIN (échantillonnage partiel, voir § 3.5.1.2). Dans les deux derniers cas, nous avons réussi à localiser reproductiblement le minimum natif en l'espace de quelques jours sur un nombre restreint de machines (20 à 30 nœuds, le nombre de nœuds réservé étant paramétrable par l'utilisateur), voir figure 3.39.

Le cas du 1LE1, bien que ne comportant que 54 degrés de liberté, est plus pernicieux que les autres exemples. En effet, appartenant à la famille des structures

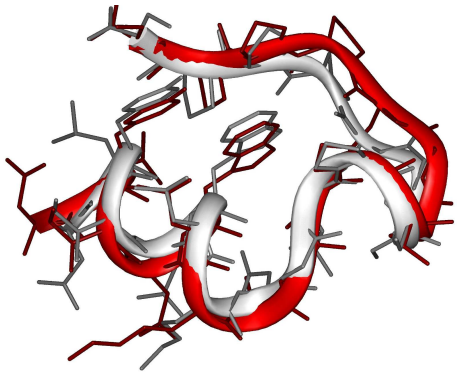


FIG. 3.39: conformation native (blanche) et meilleure conformation renvoyée par l'algorithme (rouge).

β, son paysage énergétique s'apparente moins à un entonnoir que celui des structures dites α (Muñoz *et al.*, 1997).

Dans de rares cas (deux sur plusieurs dizaines de simulations), l'algorithme est capable de reproduire parfaitement la structure expérimentale, tant concernant son squelette que ses chaînes latérales (figure 3.40). Malheureusement, la majorité des simulations se sont arrêtées avant de découvrir ce minimum. Parmi les conformations renvoyées, il y a des géométries dont le squelette est correctement prédit, mais où les chaînes latérales ne correspondent pas à la géométrie proposée par la PDB (figure 3.41). Bien que les arrangements géométriques et les interactions des groupements aromatiques sont encore à l'étude d'un point de vue théorique et mal pris en compte par les champs de forces (Guvench et Brooks, 2005), ces géométries restent plus énergétiques que la conformation native. Autrement dit, l'algorithme échoue à localiser le minimum absolu.

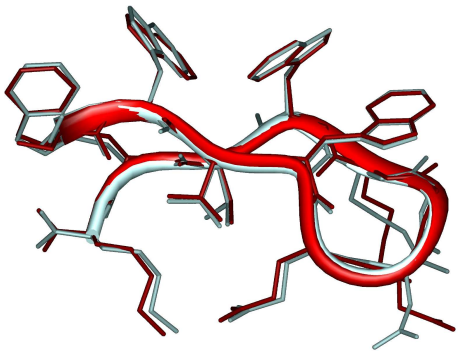


FIG. 3.40: la géométrie native trouvée par l'algorithme (structure expérimentale en blanc).

Nous pensons toutefois que ces géométries ne sont pas aberrantes et sont *probablement* présentes en solutions, mais correspondent à des états beaucoup moins

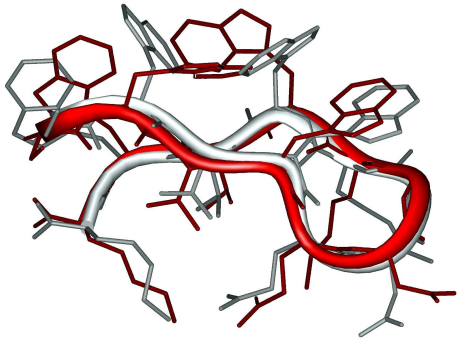


FIG. 3.41: structure presque correcte découverte par l'algorithme mais classée en 79^{e} position derrière d'autres géométries dénaturées (les interactions des tryptophanes diffèrent des prédictions d'autres auteurs).

FIG. 3.42: la meilleure solution renvoyée par l'algorithme correspond à une conformation dénaturée.

peuplés qui échappent éventuellement aux méthodes expérimentales. Les conformations 3.41 et 3.42 recréent en effet des interactions entre cycles aromatiques. De plus, le positionnement des tryptophanes n'est pas clairement connu : comme les structures expérimentales sont issues de minimisations selon des champs de forces semi-empiriques, le positionnement prédit dépend du modèle choisi. Ainsi Yang *et al.* (2004) ont proposé une structure légèrement différente de la structure initiale (Cochran *et al.*, 2001) où les tryptophanes s'arrangent plutôt dans une conformation où les tranches des uns font face aux cycles de leurs voisins (1HRX est alors remplacé par 1LE1 dans la PDB, voir figure 3.43).

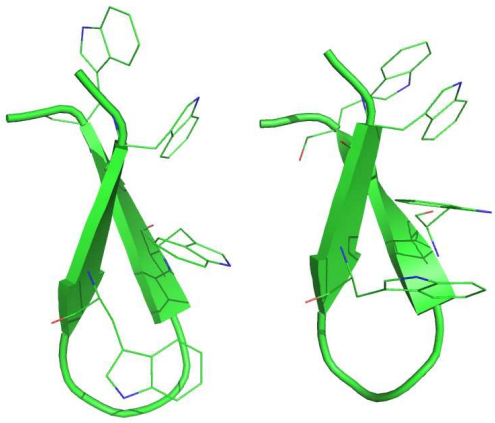


FIG. 3.43: structure tridimensionnelle de 1LE1 (gauche) qui remplace celle de 1HRX (droite). Des études plus récentes (Yang *et al.*, 2004) indiquent que les résidus tryophanes se positionnent plutôt en forme de « T » (tranche contre face).

La découverte de la géométrie 3.40 n'est qu'une simple question de temps de recherche, cependant, en autorisant plus de temps ou en fixant des plus grandes valeurs de N_{intens}, on augmente les temps de calculs proportionnellement au nombre de clusters à traiter (typiquement 10^5 à 10^6 pour 1L2Y et 1LE1). Une fois le cluster déclaré tabou — par exemple, un cluster centré sur la conformation 3.41 — plus aucune géométrie ne pourra être trouvée dans le domaine correspondant. De plus, la comparaison des cas 1L2Y et 1LE1 montre que la balance optimale entre intensification et exploration dépend de la molécule (et pas nécessairement du nombre de degrés de liberté).

Par ailleurs, les géométries presque correctes ne sont pas en tête du classement par énergies... Ainsi, la géométrie présentée (en rouge) dans la figure 3.41 est en position 79 dans la liste (comportant plusieurs centaines de milliers de solutions). Les meilleures énergies sont obtenues pour des conformations encore plus dénaturées (figure 3.42).

3.6.3 Interprétation chimique

Nous discutons ici rapidement la différence de complexité que peuvent présenter les molécules et essentiellement les structures α (Yang et Honig, 1995a) en comparaison des structures β (Yang et Honig, 1995b) et de leur épingles, voir figure 3.44. Les structures α hélicoïdales ont été étudiées expérimentalement longtemps avant les motifs de type β ; cela s'explique par des différences de stabilité et de temps de repliement[28] (et aussi par leur tendance à agréger). Ainsi, les difficultés rencontrées par nos algorithmes (bien que 1LE1 ait moins de degrés de liberté) est déjà présent dans la structure même de la molécule.

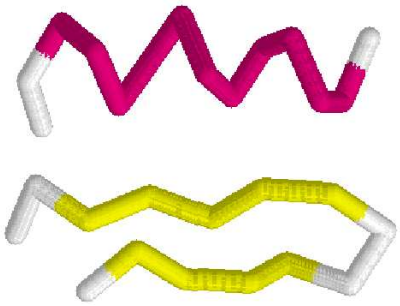


FIG. 3.44: squelette des structures secondaires « hélice » et « épingle »

Dans le cas d'une hélice, des ponts hydrogène relient des acides aminés topologiquement proches (typiquement entre l'acide aminé i et i+3). La perte entropique dûe au « gel » des quatre acides aminés n'est pas compensée par la stabilisation qu'apporte le pont hydrogène, mais une fois le premier pas d'hélice initié, chaque nouvel acide aminé qui se positionne apporte un nouveau pont hydrogène qui compense la perte entropique de sa rigidification. D'un point de vue algorithmique, nous interprétons cela comme des corrélations entre variables à courtes distances topologiques (figure 3.45). Par ailleurs, le processus peut être initié n'importe où dans l'hélice et éventuellement indépendamment en plusieurs endroits (Muñoz *et al.*, 1997).

Inversement, dans la formation d'un tournant entre deux feuillets β, le processus est nécessairement initié au niveau du tournant. Le reste de la structuration se fait alors séquentiellement en gelant, à chaque étape, deux acides aminés qui établissent alors un pont hydrogène. Les barrières d'énergie libre sont donc plus grandes dans ce cas. Les variables interagissent maintenant avec d'autres qui leur sont topologiquement éloignées (figure 3.46).

[28] Muñoz *et al.* (1997) annoncent des temps de repliement 30 fois plus longs, cependant, Nguyen *et al.* (2005) en modifiant le domaine WW de la PIN ont obtenu des temps de repliements inférieurs à la microseconde.

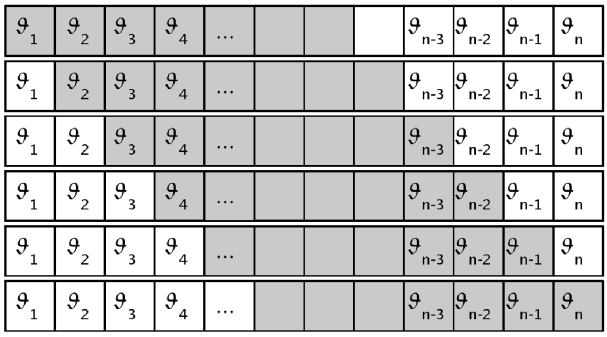


FIG. 3.45: Schémas intéressants pour former une hélice, ces schémas peuvent être découverts et se former en parallèle, s'héritent indépendamment et se concatènent facilement

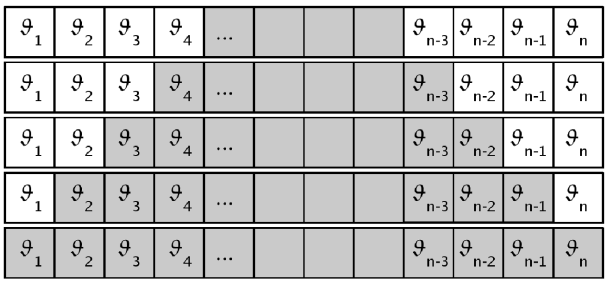


FIG. 3.46: Schémas intéressants pour former une épingle, ces schémas sont imbriqués et ne peuvent être découverts que séquentiellement

Pour aider à la formation de ce type de structures, Wenzel (2006) a récemment évoqué l'effet de la désolvatation et parle d'une compétition avec la formation des ponts hydrogène. Cet effet peut en effet diriger le collapsus hydrophobe (les tryptophanes se protégeant mutuellement) vers la géométrie que l'on connaît plutôt que vers une conformation hélicoïdale.

Enfin, remarquons que Wenzel (2006), comme Guvench et Brooks (2005) prédisent un squelette relativement bien conservé, mais les positions des chaînes latérales et en particulier celles des tryptophanes sont beaucoup plus floues et mal prédites, ce qui est conforme à nos prédictions. Ces derniers auteurs proposent même l'introduction d'un nouveau terme de champ de forces pour faciliter la convergence vers des structures plus proches de celles supposées par les expériences

L'ensemble de ces résultats a été soumis et accepté à la conférence « Congress on Evolutionary Computation » se tenant à Singapour fin septembre 2007 (annexe G).

3.7 Des défauts dans le champ de forces ?

En réalité, il est un point que nous avons laissé de côté (§ 3.5.2) et que nous détaillons maintenant : celui des conformations non-natives ayant des énergies plus basses que la géométrie expérimentale. Cette partie apparaît ici afin de préserver l'unité de la présentation des stratégies d'échantillonnage, toutefois, ce problème a dû être abordé dès les premiers tests sur les molécules plus grandes (§ 3.5).

Plusieurs phénomènes, que nous listons ici, peuvent intervenir, pour lesquels nous présentons à chaque fois les arguments qui pondèrent les hypothèses :

1. première hypothèse : la précision de l'estimation de l'énergie est insuffisante et ne permet pas de distinguer des différences énergétiques significatives du niveau de bruit de calculs. Ce pourrait être le cas si ces conformations non-natives étaient faiblement favorisées, alors que nous observons des différences énergétiques jusqu'à 30kcal.mol^{-1}.
2. La conformation native, qui est issue de l'interprétation des données expérimentales et d'une optimisation selon un champ de force différent de celui que nous avons utilisé, peut générer certaines tensions locales que des réarrangements minimes pourraient effacer. L'échantillonneur local (section 3.5.3) qui optimise relativement bien la géométrie native, devrait dans ce cas mettre en évidence des conformations plus stables. Cette situation est observée, mais n'est pas systématique : des conformations dénaturées continuent de concurrencer les énergies natives. Or, d'après l'hypothèse thermodynamique (section 1.3.3.3), la géométrie observée expérimentalement doit posséder un franc avantage énergétique.
3. La géométrie native correspond éventuellement à un minimum sous-optimal, mais entropiquement favorisé. Ce cas est tout-à-fait possible, bien que nous ayons choisi nos molécules tests pour leurs structures clairement définies. De plus, les familles de solutions proposées dans la PDB indiquent que les conformations natives sont rigides (faible variabilité). Néanmoins, cette hypothèse est une des raisons pour lesquelles nous avons développé un AG fonctionnant sur l'énergie libre d'hypercubes dans l'espace de phase (voir section 3.7.3).
4. La dernière hypothèse, est que les modèles utilisés pour l'estimation de l'énergie interne de la molécule sont approximatifs et saisissent mal certains effets[29] C'est ce dernier point que nous étudions ci-après.

[29] en particulier, le modèle de solvant continu est sujet à caution et une simulation dans un solvant explicite serait un gage d'une meilleure fiabilité de l'estimation.

Le point numéro 3 soulève à la fois un problème difficile et un faux problème. C'est un faux problème car l'échantillonnage par algorithmes génétiques est pertinent : si un état est entropiquement favorisé, il correspondra à une large zone de l'espace de recherche et sera sur-représenté dans la population de solutions proposées ; l'estimation d'une caractéristique macroscopique sur la base de cette population prendra donc implicitement en compte cet avantage entropique. Ce qui est maladroit, c'est de comparer *la* meilleure structure prédite avec *la* meilleure conformation expérimentale[30]. Une façon simple de s'affranchir de ce problème, est de comparer les molécules sur la base de propriétés macroscopiques globales. Ainsi, une stratégie qui est malheureusement restée au stade de projet, aurait été de reconstruire, sur la base de l'*ensemble* des conformations échantillonnées par l'AG, les spectres attendus de Résonnance Magnétique Nucléaire (RMN). Une comparaison de ce spectre prédit avec le spectre réel aurait alors pu trancher en faveur ou en défaveur du champ de forces (point numéro 4). De plus, cette comparaison sur des données expérimentales brutes court-circuite l'inconvénient mentionné au point 2.

3.7.1 La culpabiblité du champ de forces

La définition d'un champ de force est sans doute l'étape la plus difficile dans le domaine de la modélisation moléculaire. C'est une somme d'approximations plus ou moins précises aux domaines de validité limités et la détermination des paramètres est particulièrement difficile.

Pour remettre en question le champ de force, nous évoquons également les travaux de Kremer et Tiedemann (1994) qui ont également implémenté des AGs capables de localiser les minima absolus de l'espace de phase, mais pour lesquels ces minima ne correspondent pas au minimum natif... Plus récemment, Zhou (2003) a montré que plusieurs modèles de solvants implicites couplés à des champs de forces de type OPLS ou AMBER pouvaient prédire des minima erronés pour la structure d'une protéine.

Les paramètres sont dérivés pour reproduire le comportement local des molécules autour de leurs conformations natives et sont souvent validés par des dynamiques moléculaires qui restent des échantillonneurs locaux. De plus, les molécules utilisées dans l'ensemble d'apprentissage sont bien souvent de petites tailles. Notre étude, elle, porte sur des molécules de plus grandes tailles et l'échantillonnage de l'espace est conçu pour visiter des régions aussi diverses que possibles...

[30] on perd dans ce cas la notion de nombre de solutions.

Citons enfin Okur *et al.* (2003), qui ont tenté d'évaluer la transférabilité des champs de forces (AMBER) des petites molécules vers des systèmes plus grands (1LE1 en l'occurence) en utilisant des clusters d'ordinateurs afin d'assurer un échantillonnage exhaustif du paysage. La bonne caractérisation du paysage qu'ils ont obtenu leur a permis de mettre en évidence les tendances et les défauts de leur champ de force (OPLS).

3.7.2 Un optimiseur de champs de forces...

Rappelons que le champ de forces que nous utilisons, le CVFF, est complété par un modèle de solvant continu, qu'il utilise une distance modifiée pour atténuer les singularités et qu'il est appliqué en « *all-atom* » à des molécules quelconques (sucres, peptides, etc.).

Le nombre de paramètres qui définissent ce champ de forces (près de 4000) est tel qu'il est inconcevable de vouloir les modifier tous. En particulier, la plupart de ces paramètres dépendent des types atomiques mis en jeu ; si un type atomique n'apparaît pas ou n'est pas suffisamment représenté dans les molécules étudiées, son optimisation ne sera pas possible.

Afin de sélectionner un jeu de paramètres *les plus sujets à caution*, nous nous sommes inspirés de Vieth *et al.* (1998a), qui proposent la constante diélectrique, le modèle de solvant, l'échelle pour les charges de surface, certains rayons Van der Waals atomiques et le cutoff pour l'estimation des énergies concernant les paires d'atomes non-liés.

Parmi ceux-ci, nous avons retenu

- la constante diélectrique ε,
- le facteur de pondération des répulsions de Van der Waals,
- certains rayons Van der Waals (carbones, oxygènes et hydrogènes dans les situations les plus fréquentes),

auxquels nous avons également ajouté

- le coefficient d'influence hydrophobe,
- le coefficient de Gilson-Honig pour l'influence de la désolvatation,
- le paramètre de smoothing pour lisser les singularités dans les calculs.

soit un total de quinze paramètres.

En modifiant ces paramètres de champ de forces...

- nous remodellons le paysage énergétique, le but étant de restaurer l'avantage en énergie libre de la région native face au reste des conformations ;

– nous perturbons les lois générales qui régissent le repliement (*in silico*) des molécules. Notre approche se doit donc d'être aussi générale que possible, c'est pourquoi nous avons considéré l'effet des modifications de ces paramètres sur un maximum de molécules.

3.7.2.1 Définition du score d'un champ de force

Nous disposons déjà d'un outil pour échantillonner localement la région native avec l'échantillonneur local basé sur le recuit simulé (section 3.5.3). Nous avons aussi un outil performant pour caractériser la globalité du paysage : la machinerie des AGs métissés et méta-optimisés. Pour se définir un critère d'évaluation du champ de force, Okur *et al.* (2003) ont proposé d'utiliser non-seulement la différence énergétique entre solutions natives et non-natives (discriminées selon un critère de RMSD[31] au natif), mais également la pente de la régression linéaire entre énergies et RMSD. Ce dernier terme permet de favoriser les paysages énergétiques se comportant comme des entonnoirs. Cependant nous avons préféré nous restreindre à la physique du problème en ne gardant que la différence en énergie libre des deux simulations, ce qui revient à maximiser la probabilité du domaine natif : $\mathcal{D}_{\text{natif}}$

$$
\begin{array}{rrrrcl}
 & & & \Pr(\mathcal{D}_{\text{natif}}) & = & \frac{1}{Z}\int_{\mathcal{D}_{\text{natif}}} e^{-\beta E(\theta)} d\theta^n. \\
\text{Posons} & G_{\text{natif}} & \text{tel que} & e^{-\beta G(\mathcal{D}_{\text{natif}})} & \triangleq & \Pr(\mathcal{D}_{\text{natif}}), \\
\text{et} & G_{\text{total}} & \text{tel que} & e^{-\beta G_{\text{total}}} & \triangleq & \int_{\Omega} e^{-\beta E(\theta)} d\theta^n = Z. \\
 & & \text{Alors} & \Pr(\mathcal{D}_{\text{natif}}) & = & \exp\left[\beta(G_{\text{total}} - G_{\text{natif}})\right]. \\
 & & & \text{Critère de } \textit{fitness} & \triangleq & \Delta G = G_{\text{total}} - G_{\text{natif}}
\end{array}
$$

3.7.2.2 Une stratégie d'optimisation

Disposant dorénavant d'un critère pour évaluer la pertinance d'un paysage énergétique pour chacune des molécules traitées, nous pouvons maintenant optimiser les paramètres proposés ci-dessus. Pour le choix de la stratégie, nous maîtrisons celle des AGs, mais pouvons toutefois citer les auteurs suivants

– Koretke *et al.* (1998) utilisent le recuit simulé pour l'optimisation de fonctions énergies dédiées à l'échantillonnage conformationnel.
– Okur *et al.* (2003) qui ont également opté pour un AG à la recherche de paramétrages plus pertinents des champs de forces de AMBER,

[31] Root Mean Squared Deviation : déviation standard des coordonnées atomiques après une superposition optimale des deux molécules.

– Antes *et al.* (2005) optimisent une fonction régulière (peu rugueuse) pour le *docking* par un va-et-vient constant entre apprentissage par réseau de neurones sur un ensemble de points connus (évalués par FlexX) et recherche de nouveaux points à tester qui minimisent la fonction approximée.

En s'inspirant de ces recherches, nous avons mis en place une stratégie que nous détaillons maintenant.

En échantillonnant à la fois l'espace entier par $C_S G_A$ piloté par le μG_A, et la région avoisinnant la conformation native par l'échantillonneur local, nous obtenons un ensemble de conformations caractéristique du paysage courant pour chacune des molécules. On modifie alors les paramètres du champ de forces afin de minimiser les énergies libres des solutions natives par rapport à celles des solutions globales.

Étude de la faisabilité d'une coévolution. En théorie, il est possible de suivre les minima locaux de l'espace de phase au fur et à mesure que les paramètres évoluent (voir calculs ci-dessous et équation (3.18), qui donnent l'évolution de la position du minimum en fonction de la variation des paramètes). Cependant, les irrégularités du paysage rendent l'évaluation de l'Hessienne difficile et peu rigoureuse. De plus, les disparitions et surtout les apparitions de nouveaux minima (bifurcation lorsque l'Hessienne n'est plus inversible) font échouer l'approche. Nous avons même renoncé à un suivi progressif des solutions de type recuit simulé ou dynamique moléculaire au cours des modifications du paysage, car l'échantillonnage global du champ de forces modifié est de toute façon nécessaire pour localiser les éventuels nouveaux minima.

L'énergie dépend des variables θ et des paramètres p :

$$E \ : (p, \theta) \longrightarrow E(p, \theta). \tag{3.15}$$

Si θ_0 est un minimum local pour p_0, alors

$$\frac{\partial E}{\partial \theta}(p_0, \theta_0) = 0. \tag{3.16}$$

Alors le couple $(p_0 + dp, \theta_0 + d\theta)$ est encore un minimum si

$$\frac{\partial E}{\partial \theta}(p_0 + dp, \theta_0 + d\theta) = 0, \tag{3.17}$$

or

$$^t\left(\frac{\partial E}{\partial \theta}(p_0+dp,\theta_0+d\theta)\right) = {}^t\left(\frac{\partial E}{\partial \theta}(p_0,\theta_0)\right) + \frac{\partial^2 E}{\partial p \partial \theta}(p_0,\theta_0)\times dp + \frac{\partial^2 E}{\partial \theta^2}(p_0,\theta_0)\times d\theta,$$

ainsi,

$$d\theta = -\left(\frac{\partial^2 E}{\partial \theta^2}(p_0,\theta_0)\right)^{-1}\left(\frac{\partial^2 E}{\partial p \partial \theta}(p_0,\theta_0)\right) dp. \tag{3.18}$$

Une optimisation séquentielle. Ne pouvant pas faire coévoluer les solutions dans leurs paysages en même temps que les paysages eux-mêmes, nous avons cherché à optimiser, pour les solutions échantillonnées, les paramètres du champ de forces, jusqu'à obtenir des ΔG positifs, puis avons relancé l'échantillonnage dans les nouveaux paysages. Cette recherche est assurée par un AG simpliste (semblable au μG_A, voir page 117) qui doit ré-évaluer systématiquement, pour chaque jeu de paramètres, les énergies de toutes les conformations de toutes les molécules. Les paramètres sont choisis parmi un ensemble de valeurs discret que nous fournissons à l'AG.

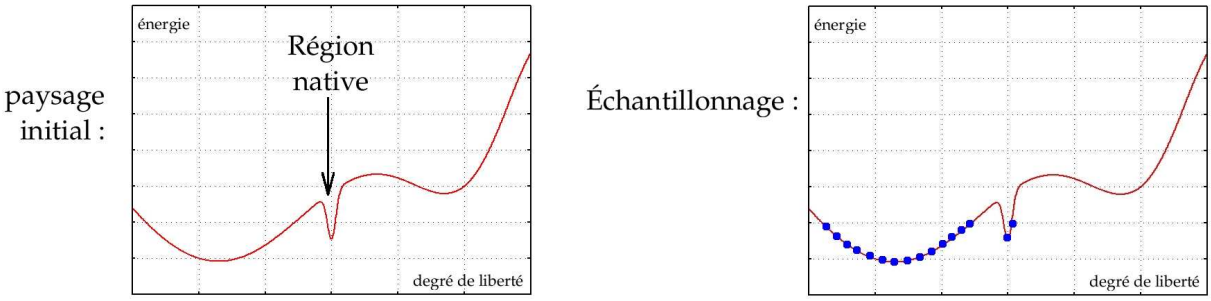


FIG. 3.47: le paysage initial est échantillonné par le C_5G_A.

3.7.2.3 Résultats

Après une dizaine d'allers et retours entre échantillonnage des molécules et optimisation des paramètres de champ de force, voici les conclusions de cette étude :

- certaines molécules, comme la cyclodextrine, sont systématiquement et correctement prédites (la géométrie native est trouvée et est classée en pôle position dans le classement par énergies), ce qui indique que les modifications du champ de forces n'ont pas été faites au détriment des plus petites molécules pour lesquelles il était initialement conçu ;

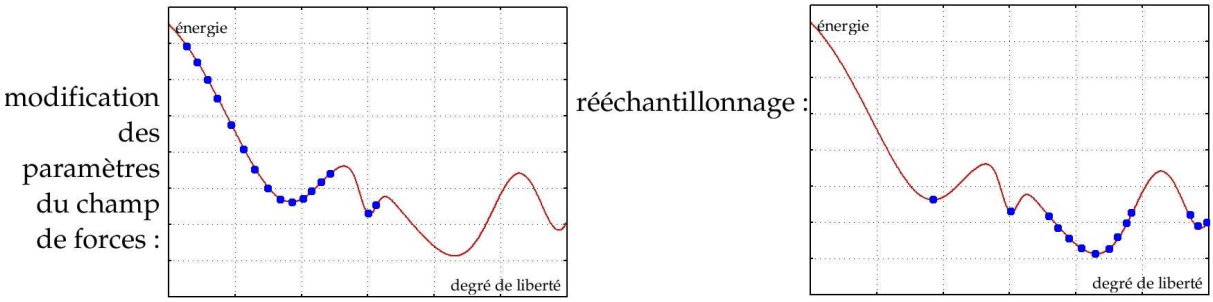


FIG. 3.48: en retouchant les paramètres du champ de force, il est possible de favoriser les solutions natives. Après modification, il est nécessaire de rééchantilllonner pour découvrir les éventuels nouveaux minima.

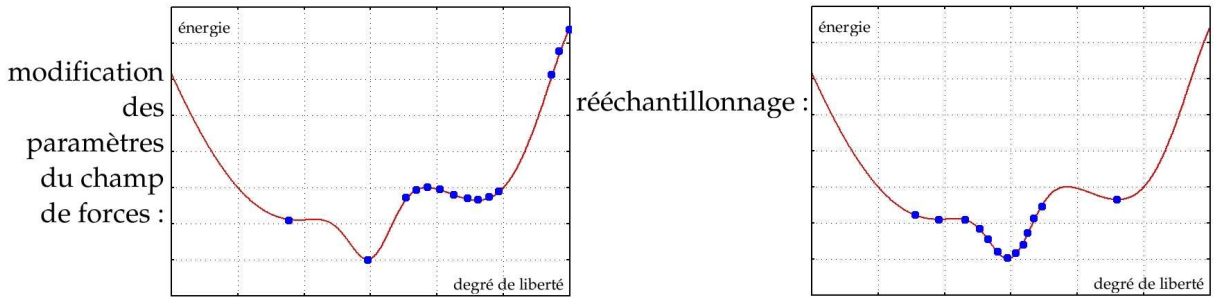


FIG. 3.49: le cycle reprend jusqu'à obtenir un paysage qui favorise la région native.

- pour d'autres molécules, la solution native est échantillonnée et figure dans le fichier de résultats, mais n'est pas classée parmi les meilleures conformations : c'est le cas par exemple de la « tryptophan cage » (sauf dernier paramétrage) et de l'hélice covalemment modifiée pour lesquelles des géométries dénaturées possèdent des énergies plus basses ;
- pour d'autres molécules enfin, les géométries natives ne sont jamais échantillonnées... c'est en particulier le cas de la PIN pour qui l'énergie du minimum expérimental n'a jamais été égalée ; le « tryptophan zipper » fait aussi partie de ces molécules, mais depuis, l'intensification des efforts de calculs grâce à la grille d'ordinateurs a permis de meilleurs résultats.

La figure suivante (3.50), qui apparaît sur le poster présenté par D. Horvath lors de la « Computational Chemistry Gordon Research Conference » (Parent *et al.*, 2006), résume les solutions trouvées par l'algorithme qui furent les plus proches du natif (RMSD incluant tous les atomes). Chaque colonne présente une molécule (dans l'ordre : cyclodextrine, 1L2Y, CRH et 1LE1), chaque nouvelle ligne correspond à un nouveau paramétrage du champ de force. Les structures vertes sont les géométries natives, tandis que les jaunes correspondent aux solutions prédites. Enfin, sont indiqués les rangs de ces conformations dans leur classement selon les énergies croissantes ainsi que leur RMSD au natif.

Ainsi, la cyclodextrine est systématiquement correctement prédite avec un RMSD ne dépassant pas 1,5Å ; le mauvais classement de certaines géométries repose alors sur des différences minimes. Les trois dernières versions de champ de forces ont permis de trouver la conformation native de la « tryptophan cage », de plus, dans le dernier cas, elle est classée en première position. L'helice CRH est correctement repliée dans les deux derniers cas, mais a reculé dans le classement. Enfin, comme nous l'avons évoqué au paragraphe 3.6.3 et comme le suggère la dernière colonne, le « tryptophan zipper » constitue un problème difficile. Toutefois, dans la dernière version de champ de forces, le squelette semble enfin se rapprocher du natif.

Il faut rester prudent avec cette analyse qui présente — à tort — *une seule* géométrie par molécule. Le principal résultat est d'avoir réussi à optimiser le ΔG qui fait intervenir une notion d'ensemble et cela, simultanément pour toutes les molécules. Le champ de forces ainsi obtenu promet donc un échantillonnage plus représentatif des paysages, prenant en compte les profondeurs des puits et leurs largeurs.

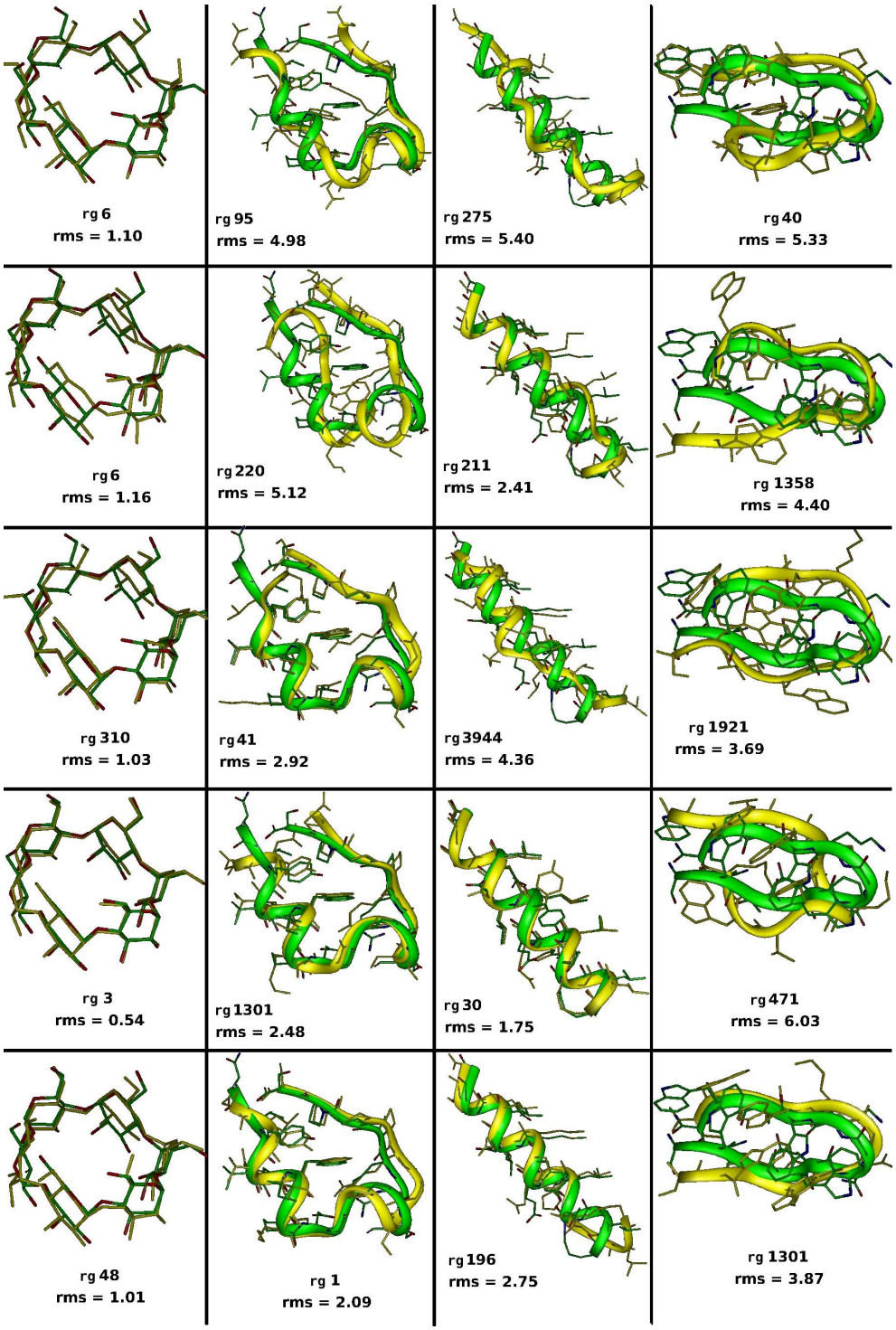


FIG. 3.50: conformations trouvées par l'algorithme (jaunes) les plus proches des géométries expérimentales (vertes) pour chaque molécule et chaque paramétrage de champ de forces ; sont indiqués les rangs dans le classement par énergies et les RMSD entre ces conformations et le natif.

3.7.3 Derniers développements : comment gérer l'entropie

3.7.3.1 Introduction

En abordant des molécules de cette taille, nous avons dû repenser nos stratégies d'échantillonnage et nous avons dû réétudier le modèle de champ de forces. Dans ces deux directions, les résultats sont très encourageants et nous pouvons maintenant aborder les cas les plus difficiles des structures β.

Les molécules cycliques ou partiellement échantillonnées ainsi que les structures α, sont des cas plus simples et nous sommes en mesure de résoudre des problèmes de plus de 70 degrés de liberté. À titre de comparaison, le problème scolaire d'échantillonnage conformationnel : le neuropeptide [Met]-enkephaline, comporte 24 degrés de liberté (Jin *et al.*, 1999; Day *et al.*, 2002; Vengadesan et Gautham, 2003).

Malgré cela, les géométries natives restent en concurrence avec des géométries dénaturées d'énergies comparables. La dernière hypothèse évoquée au point 3, section 3.7, est que l'entropie peut modifier la balance énergétique. Ainsi, ce n'est pas l'énergie potentielle qui dicte la conformation des molécules, mais bien l'énergie libre[32].

Par ailleurs, ce qui limite la capacité exploratrice des algorithmes, c'est le nombre $\left(\frac{360}{\text{pas}}\right)^{N_{\text{ddl}}}$ de conformations envisageables. Nous avons réduit le nombre N_{ddl} de degrés de liberté en adoptant une description torsionnelle de la molécule, mais nous ne pouvons pas augmenter à souhait la taille du pas pour la discrétisation de l'espace de phase. Cela est une conséquence de l'échantillonnage qui nous fait perdre l'information présente entre les points de l'espace. Dans l'approximation des intégrales par la méthode de Monte Carlo (section 2.5.4), les échantillons représentent un volume élémentaire $d\theta^{N_{\text{ddl}}}$; mais pour un pas plus grand, il est intrinsèquement faux de représenter un volume $(\text{pas})^{N_{\text{ddl}}}$ par un représentant ponctuel... Il serait plus judicieux de pouvoir évaluer l'énergie libre sur cette boîte.

Partant de ces réflexions et puisque nous disposons d'une grille de calcul, nous nous sommes inspirés des travaux de Takahashi *et al.* (1999) pour imaginer une stratégie d'échantillonnage à deux niveaux : un premier AG gère des régions de solutions selon un découpage grossier, tandis que l'évaluation du *fitness* dans ces boîtes repose sur une estimation de l'énergie libre réalisée par un $C_S G_A$ confiné à cette région. Cette approche est particulièrement adaptée à un calcul distribué où

[32] souvent les auteurs utilisent le terme d'énergie libre à la place d'énergie potentielle, ceci est une conséquence de l'intégration de l'effet du solvant qui repose sur un calcul de moyenne (PMF, voir section 2.4.2.2).

va pouvoir être parallélisée l'évaluation de chaque région de conformations (que l'on appellera une $\mathcal{R}$-conformation, par opposition aux $\mathcal{P}$-conformations ponctuelles).

La méthode d'exploration de chacune des $\mathcal{R}$-conformations est identique à celle du $C_S G_A$, à la différence qu'elle ne se fait plus sur le tore et ne peut donc plus profiter de la périodicité de l'énergie selon ses variables.

3.7.3.2 Détail de la stratégie

L'espace est donc découpé en hyper-parallélépipèdes — « $\mathcal{R}$-conformations » — dont la longueur $\Delta\theta_i$, dans chaque dimension i, dépend de la pondération du degré de liberté correspondant et d'un paramètre noté $\delta\vartheta$. Ces longueurs sont choisies de façon à avoir un nombre entier de divisions (noté D_i) dans chaque dimension.

Une $\mathcal{R}$-conformation est donc un N_{ddl}-uplet d'entiers : $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_{N_{\text{ddl}}})$, où chaque $\mathcal{R}_i \in [1; D_i]$ indique quelle sous-division est considérée. La 2π-périodicité de l'espace de phase se traduit alors par le fait que les $\mathcal{R}_i$ sont calculés *modulo* D_i.

Chacune des $\mathcal{R}$-conformations $(\mathcal{R}_1, \ldots, \mathcal{R}_{N_{\text{ddl}}})$ est alors représentée par un échantillonnage de la boîte $\{(\theta_1, \ldots, \theta_{N_{\text{ddl}}}) | \forall i, (\mathcal{R}_i - 1)\Delta\theta_i \leq \theta_i \leq \mathcal{R}_i \Delta\theta_i\}$, avec un pas donné par un paramètre N_{s}. Cet échantillonnage est réalisé par les $C_S G_A$.

Avant de pouvoir définir une stratégie pour explorer l'ensemble des $\mathcal{R}$-conformations, nous avons cherché à étudier le comportement des $C_S G_A$ sur des sous-domaines de l'espace complet. En effet, la fonction de *fitness* d'une région $\mathcal{R}$ est donnée par une *estimation* de son énergie libre approximée par un algorithme stochastique. Il s'agit donc d'une variable aléatoire qu'il faut rendre la plus reproductible possible ; cela peut être fait en augmentant les ressources de calculs dédiées à l'exploration de chaque boîte, ou en réglant les paramètres $\delta\vartheta$ et N_{s}.

Afin d'évaluer la reproductibilité de l'échantillonnage, nous avons considéré plusieurs molécules dans différentes conformations (des conformations de basses énergies, en particulier, la conformation native) et avons évalué l'écart-type de l'énergie libre sur 5 exécutions indépendantes de l'algorithme. Ce travail est encore en cours de validation : plusieurs jeux de paramètres sont étudiés, prenant également en compte différentes stratégies de filtrage par dissimilarité (qui est aussi un facteur important dans l'estimation de l'énergie libre). Des résultats préliminaires semblent indiquer qu'il est possible de trouver un paramétrage tel que les écarts types soient tous inférieurs à 4kcal.mol^{-1}.

3.8 Applications

Nous n'avons présenté, jusqu'à présent, que le développement et la validation de stratégies d'échantillonnage conformationnel sur des exemples connus. L'utilité de cette suite d'algorithmes est de pouvoir aider à la compréhension des mécanismes de repliement, mais également de fournir aux expérimentateurs un outil pour complémenter leurs données, souvent partielles et parfois imprécises, qui concernent des molécules dont la structure n'est pas toujours connue. C'est pourquoi l'Hamiltonien moléculaire comporte des termes supplémentaires pouvant être utilisés lorsque certaines données sont disponibles. Ces données peuvent être de plusieurs types :

- la distance entre deux atomes est estimée, ou du moins bornée dans une fourchette (grâce notamment à la RMN, voir Van de Ven, 1995),
- l'angle d'une torsion est connu ou estimé.

On pourrait, de la même façon, pénaliser la violation de toute forme de contraintes expérimentales[33].

Les contraintes expérimentales de distance en particulier sont intégrées par le biais de termes harmoniques qui ont donc un effet semblable à la coupure de liaisons. Tout se passe dans la simulation, comme si il existait une liaison entre les deux atomes impliqués, jusqu'à ce que la fourchette de distances précisées soit respectée. Ceci tend à accélérer considérablement la convergence de l'algorithme.

Par ailleurs, les molécules étudiées expérimentalement sont généralement d'un ordre de taille supérieure à ce que nous pouvons traiter, c'est une des raisons pour lesquelles nous avons développé la possibilité de faire un échantillonnage partiel de la molécule.

Nous avons donc cherché à aborder des cas réels de molécules inconnues ou partiellement connues. Nous présentons ici un début d'étude de deux cas : le premier se rapporte à l'exploration des conformations d'un tournant entre deux feuillets β. Le deuxième concerne la prédiction du positionnement des deux brins terminaux d'une protéine dont le reste de la structure est connue.

3.8.1 Tournant de PIN1

Le facteur limitant dans le processus de repliement du domaine WW de la protéine humaine PIN1, est la formation de la boucle du premier tournant. Comme toute structure biologique, cette molécule a subi la pression de sélection de milliers

[33] comme par exemple la colinéarité de certains liaisons N-H dans les mesures de couplages dipolaires.

de générations, on est donc en droit de se demander pourquoi l'évolution darwinienne n'a pas sélectionné de meilleures séquences, plus rapides à se former et plus stables. Jäger *et al.* (2006) ont considéré la question en mutant la protéine afin de remplacer cette boucle par des séquences connues pour se replier de façon plus robuste (voir aussi Nguyen *et al.*, 2005).

Les mutants obtenus sont effectivement plus stables et plus rapides, mais ils perdent partiellement leur fonction biologique puisqu'ils n'interagissent plus avec leurs partenaires habituels. Ainsi, la pression de sélection a favorisé la fonction au prix d'un temps de repliement plus long.

Nous avons voulu mettre cela en évidence en étudiant la boucle du premier tournant de la PIN1 sauvage et de ses mutants.

Pour cela, nous avons réalisé un échantillonnage partiel (voir section 3.5.1.2) du domaine WW sauvage, noté S, et des deuxième et septième mutants (les plus stables) proposés par Jäger *et al.*, notés M_2 et M_7. Dans les trois cas, les degrés de liberté appartenant aux acides aminés du tournant ont été échantillonnés (voir tableau 3.4 et figure 3.51).

Molécule	atomes impliqués (et nombre)	acides aminés impliqués (et nombre)	nombre de degrés de liberté
S	164 à 305 (142)	19 à 27 (9)	39
M_2	232 à 349 (118)	15 à 23 (8)	35
M_7	142 à 251 (110)	15 à 23 (7)	32

TAB. 3.4: caractéristiques de l'échantillonnage de PIN1 sauvage (S) et des mutants 2 (M_2) et 7 (M_7).

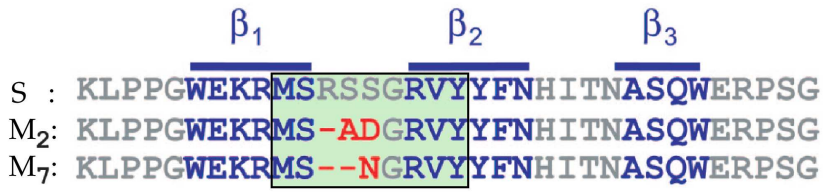


FIG. 3.51: séquences des domaines WW sauvage et mutants, en vert figurent les acides aminés optimisés.

Pour analyser les résultats, chaque conformation échantillonnée a été reportée sur un graphe donnant son RMSD[34] à la géométrie cristalline et son énergie (figures 3.52 pour M_7 et 3.53 pour S). Il est alors possible de tracer une énergie libre en fonction du RMSD (en rouge sur les figures).

[34]ce RMSD prend en compte tous les atomes.

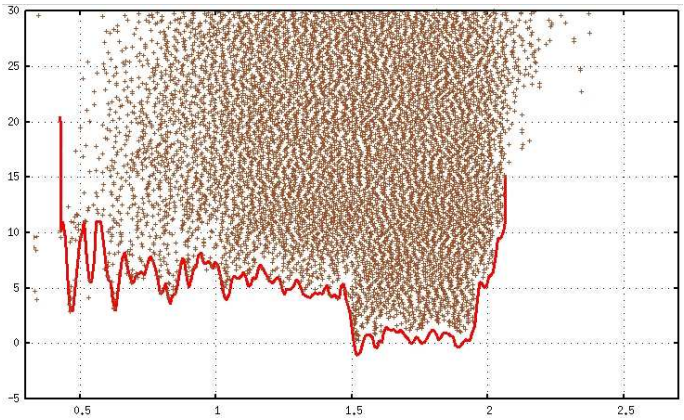


FIG. 3.52: marron : ensemble des conformations échantillonnées de la PIN mutante en fonction du RMSD au natif et des énergies internes. En rouge : énergie libre en fonction du RMSD.

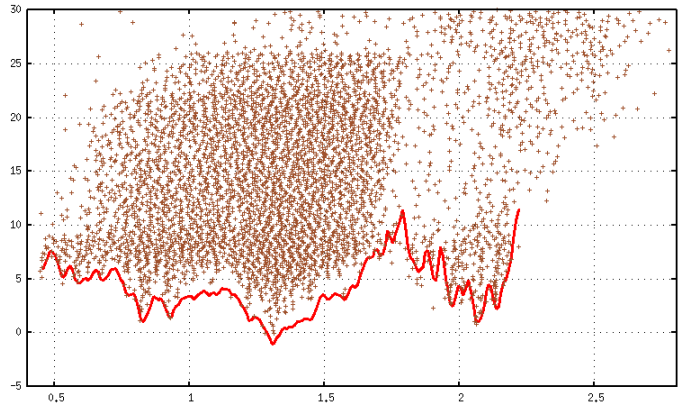


FIG. 3.53: idem avec la PIN sauvage.

Remarque : les résultats ne sont pas présentés pour M_2 qui n'a pas convergé vers la structure expérimentale.

On observe alors des profils énergétiques différents, où le mutant possède globalement un unique puits ne dépassant pas 1,9Å, tandis que la PIN sauvage possède deux puits bien distincts, le deuxième étant autour de 2,1Å.

L'utilisation du critère RMSD n'est peut-être pas pertinente dans ce cas, car dans un rayon de 2Å il est possible de trouver une assez grande variété de structures. Nous avons donc extrait les structures de plus basses énergies afin de les visualiser (figures 3.54 pour M_7 et 3.55 pour S).

En vert, figurent les structures cristallines de S et M_7. En orange, nous avons indiqué les meilleures solutions retournées par l'algorithme ; elles correspondent à des géométries à 1,53Å pour M_7 et 1,30Å pour S. Enfin, les structures violettes sont les géométries les plus différentes du natif dans une fenêtre de 1 kcal.mol^{-1} au dessus de la meilleure énergie (1,92Å pour M_7 et 2,06Å pour S).

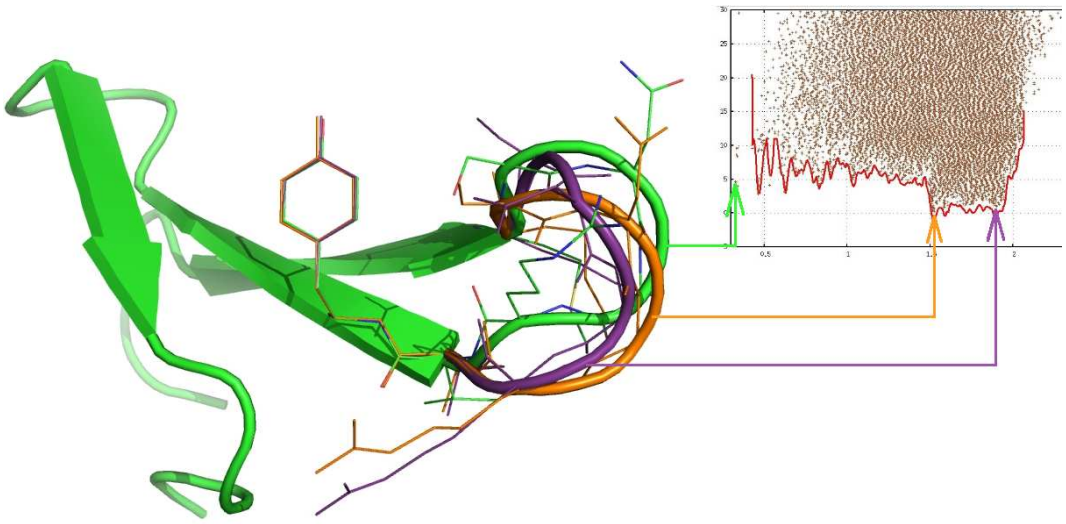


FIG. 3.54: mutant de la PIN. Vert : structure cristallographique ; orange : meilleure structure découverte par l'algorithme (de meilleure énergie que le natif) ; violet : autre minimum à 1,9Å. Les différences s'expliquent surtout par des réarrangements des chaînes latérales.

Pour l'instant, les tests sur cette partie de protéine ne permettent pas de conclure plus précisément sur son mécanisme d'interaction.

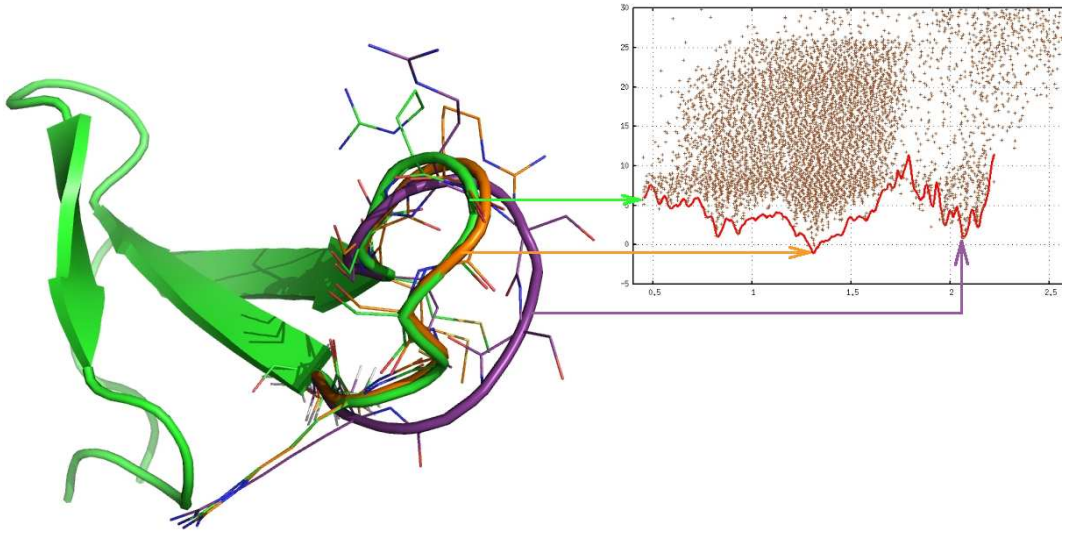


FIG. 3.55: PIN native. Vert : conformation native ; orange : meilleure structure découverte par l'algorithme (de meilleure énergie que le natif) ; violet : géométrie très différente et d'énergie comparable aux autres minima à 2,1Å du natif.

3.8.2 La cyclophilline

La « *cyclophilline B* » se lie à l'héparine. Pour étudier cette interaction et, en particulier, mettre en évidence le site de fixation, des études par RMN ont été menées. La structure de la cyclophilline B a été déterminée par diffraction de rayons X (Jin et Harrison, 2002), cependant, lors de la purification, les deux brins terminaux ont été coupés par protéolyse. Or, les résultats de RMN prédisent justement que le site de fixation implique ces brins terminaux. Il est donc nécessaire de déterminer le positionnement de ces brins.

Nous avons alors proposé de modéliser cette partie de la cyclophilline, en gardant le reste de la protéine (dont la structure est connue) fixe. L'hypothèse est que les imprécisions du champ de forces sur une molécule si grande seront compensées par les quelques contraintes expérimentales disponibles. Les études par RMN de la cyclophilline B avec héparine ont en effet permis de restreindre des distances interatomiques impliquant certains atomes de ces brins (total de 19 contraintes de distances exploitables).

Les brins N-ter et C-ter ont été reconstruits manuellement en utilisant l'interface de conception de PyMol[35], dans une conformation quelconque. Nous avons alors autorisé 116 degrés de liberté à être optimisés, impliquant plus de 400 atomes parmi près de 3000 (voir figure 3.56). Ces degrés de liberté concernent principalement les

[35] http ://www.pymol.org/

brins terminaux, mais également quelques chaînes latérales du reste de la protéine susceptibles d'interagir (parties sur fond rouge dans la figure).

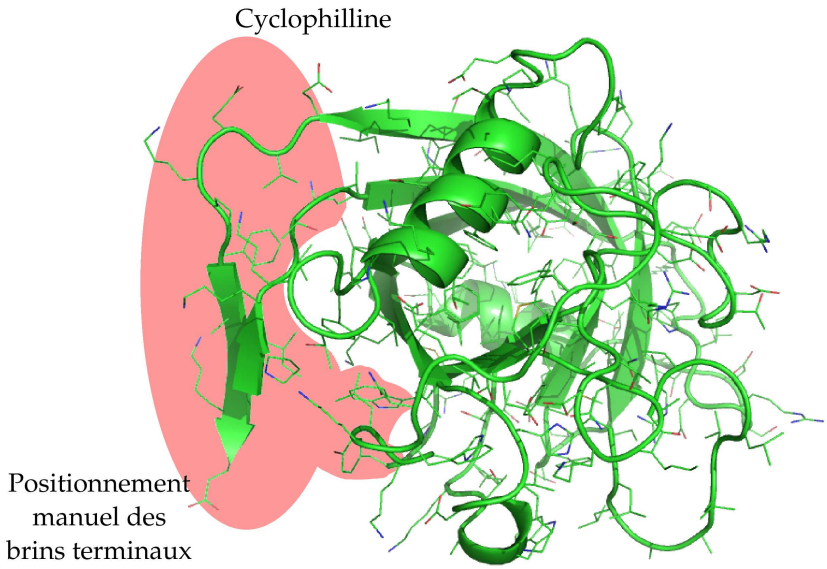


FIG. 3.56: structure de la cyclophilline B, les brins terminaux ont été positionnés manuellement. Les atomes sur fond rouge correspondent aux parties que nous avons optimisées.

Malgré le nombre important de degrés de liberté, l'algorithme arrive à localiser des solutions de basses énergies dans le sous-domaine respectant les contraintes expérimentales. L'intérêt de l'utilisation de notre algorithme, est qu'il est conçu pour renvoyer un ensemble de solutions, permettant de caractériser la flexibilité des brins dans la limite des contraintes expérimentales. La figure 3.57 montre la meilleure solution trouvée.

Cette étude a ainsi permis de valider le principe d'intégration de connaissances expérimentales par le biais de contraintes énergétiques. Elle s'insère dans le cadre d'une étude plus complète sur l'interaction de l'héparine avec la cyclophilline et fait l'objet d'un article récemment accepté dans le « Journal of Biological Chemistry » (Hanoulle *et al.*, à paraître).

3.9 Conclusion

Après avoir présenté les stratégies envisagées dans la littérature et en avoir choisie une parmi les plus adaptées, nous avons détaillé l'implémentation d'un algorithme génétique original, comportant de nombreux paramètres de contrôle et hybridé avec

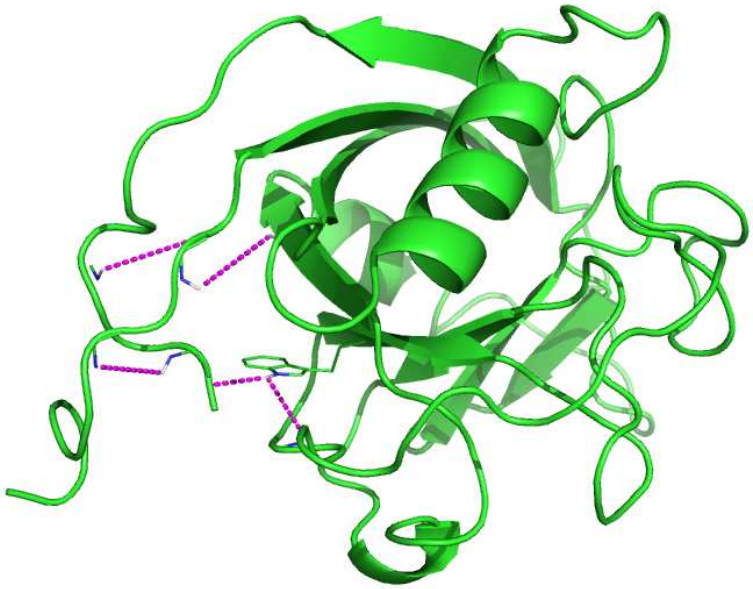


FIG. 3.57: meilleure solution retournée : en pointillés magenta, figurent les contraintes expérimentales de distances.

plusieurs stratégies complémentaires. Cette machinerie complexe est entièrement gérée par une deuxième couche algorithmique qui assure une stratégie de recherche efficace et reproductible[36] (comme l'ont indiqué les résultats à la sous-section 3.4.5).

Cette suite d'algorithmes offre un outil de recherche efficace et spécifiquement adapté à la problématique d'une recherche multimodale en grandes dimensions ce qui, généralement n'est abordé que par des simulations de dynamiques moléculaires.

Par ailleurs, nos simulations ont permis d'illustrer la difficulté que représente la définition d'une balance correcte entre diversification et exploration. S'il est clair que nos algorithmes ont de bonnes propriétés exploratrices, il était important qu'ils restent capables d'intensifier les recherches dans les régions de basses énergies, afin d'éviter que celles-ci ne soient mal caractérisées et donc délaissées au profit d'autres régions, d'énergies comparables, mais mieux connues.

Lorsque nous avons appliqué notre stratégie à des molécules plus grandes, nous avons dû développer quelques heuristiques complémentaires comme le principe de fragmentation. La principale adaptation a alors été de définir une politique de parallélisation des îles de recherche.

La capacité exploratrice des algorithmes sur des problèmes de cette taille ont enfin permis d'accréditer une idée plutôt discrète dans la littérature : celle du do-

[36] la notion d'*exploration optimale* restant difficile à définir...

maine d'applicabilité des champs de forces. Ces champs de forces ont été paramétrés sur un ensemble, nécessairement incomplet, de petites molécules. Même s'ils restent applicables à de plus grandes molécules, ils ne permettent pas nécessairement de caractériser le paysage énergétique loin de l'état natif (c'est pourquoi les simulations de dynamiques moléculaires ne les remettent pas en cause), en particulier ils peuvent prédire l'existence de faux minima en dehors de la région native. Après avoir mis en évidence ce fait, nous avons cherché à perturber quelques uns des paramètres du champ de forces, afin de rétablir l'équilibre (au sens thermodynamique du terme) entre la région native et le reste de l'espace de phase.

Enfin, le traitement de deux exemples concrets (donc complexes), nous a permis d'illustrer et de valider la stratégie de recherche sur une portion de molécule (le reste étant fixe) et l'utilisation de contraintes expérimentales pour la recherche.

Chapitre 4

Vers des stratégies de prédiction des affinités entre ligands et cibles macromoléculaires

4.1 Introduction

Nous avons abordé jusqu'à présent le problème de la prédiction de la géométrie d'une seule molécule. Cette première phase a dû être approfondie par l'étude et l'optimisation de certains paramètres du champ de force, afin d'obtenir une estimation de la fonction énergie qui soit plus fiable. Désormais, nous souhaitons généraliser notre approche du repliement au cas de deux molécules en abordant le *docking*. Ceci nécessite l'incorporation des degrés de liberté du positionnement relatif des partenaires.

Pour entreprendre des simulations de *docking*, il faut connaître des acteurs susceptibles d'interagir. Or, quelque soit la propriété chimique d'une molécule que l'on cherche à déterminer — électro ou hydro-philie/phobie, présence de sites actifs, activité biologique et en particulier, affinité pour d'autres acteurs — il faut, en principe, passer par une étape de prédiction de la structure tridimensionnelle, seule garante de la fonction. Dans une optique pharmaceutique, les molécules sont issues d'énormes bases de données de cibles thérapeutiques potentielles, impossibles à traiter de manière systématique, ce qui a motivé le développement d'algorithmes moins précis mais très rapides, exploitant uniquement les données topologiques des molécules et immédiatement accessibles sans calcul préalable. Ces méthodes QSAR (Quantitative Structure-Activity Relationship) tentent de mettre en évidence des corrélations,

parmi les molécules, entre certains *indices* topologiques et certaines mesures de l'activité. Elles ont pour but d'écarter, dès les premiers stades du filtrage de ces bases de données pharmaceutiques, les cibles « visiblement » inactives, afin d'économiser le temps de synthèse en laboratoire. Les étapes, plus précises mais plus coûteuses de repliement et de *docking* sont alors laissées pour les stades ultimes du processus de filtrage (voir figure 4.1). Avant de présenter l'approche adoptée pour aborder l'échantillonnage d'un complexe de deux molécules, nous proposons donc un aperçu des méthodes QSAR à travers la contribution que nous y avons apportée.

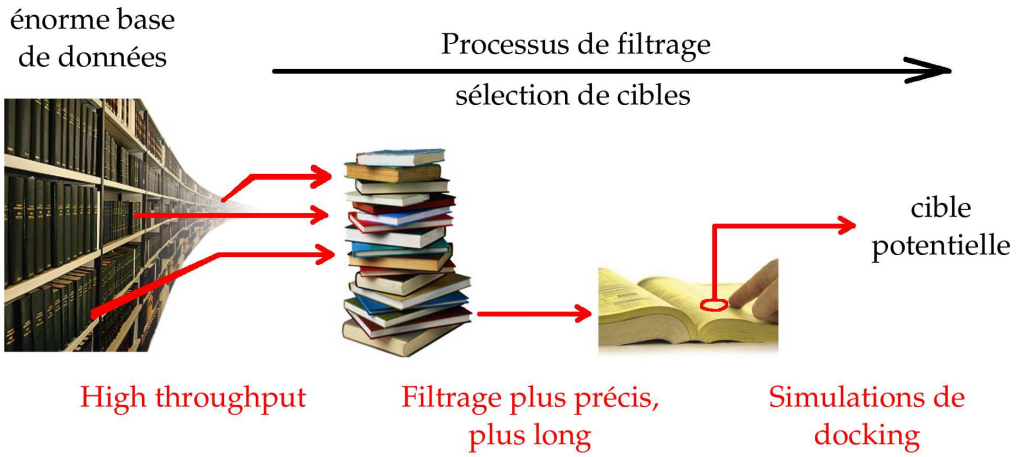


FIG. 4.1: différentes étapes de filtrage des bases de données moléculaires, des plus rapides aux plus précises.

La recherche de complémentarités entre deux molécules n'est pas très différente de la recherche de similarités, c'est pourquoi une partie du chapitre est dédié à la superposition de deux molécules (dans le cas de deux molécules identiques différant par leur conformation, puis dans le cas général). Dans tous les cas (superposition ou *docking*), il s'agit de positionner une molécule par rapport à l'autre, ce qui est le fil conducteur de ce chapitre.

4.2 La comparaison des structures

Le cas le plus simple, est de chercher à *comparer* deux conformations différentes d'une même molécule. Ceci est fait grâce au RMSD qui a déjà servi à plusieurs niveaux et en premier lieu à l'analyse des résultats des algorithmes mis en place. Nous verrons ensuite une stratégie pour relater des structures issues de composés

chimiques différents, ce qui offre un outil précieux lors de la recherche de substituts pharmaceutiques d'une molécule connue.

Énoncé : étant données deux molécules (ensembles d'atomes avec leurs graphes de liaisons et coordonnées cartésiennes), comment superposer *intelligemment* leurs structures tridimensionnelles ? Autrement dit, comment trouver les meilleures translation et rotation afin de mettre en correspondance les groupements fonctionnels similaires des deux molécules ?

4.2.1 La déviation standard moyenne

Dans le cas le plus simple, une telle superposition implique deux molécules identiques qui ne diffèrent que par leurs conformations ; on parle alors de déviation standard moyenne (RMSD : Root Mean Squared Deviation), parfois normalisée (Carugo et Pongor, 2001), mais fondée sur le même calcul (équation (4.1)). Cette déviation standard définit une *distance* dans l'espace des conformations[1].

Notre approche, présentée ici, diffère de l'ancienne démarche (Kabsch, 1976; Kabsch, 1978) en ce sens qu'elle utilise les quaternions plutôt que le calcul matriciel pour dérouler le calcul et obtenir une formule presque directe (contrairement à Mc Lachlan, 1982). Cette idée a déjà été appliquée en stéréovision (voir article de Horn, 1987). L'annexe A présente rapidement les quaternions et rappelle le principal résultat qui nous sera utile. On peut alors faire apparaître une forme bilinéaire dans l'espace $\mathbb{H}$ des quaternions et montrer que le calcul du RMSD revient à un calcul de plus petite valeur propre. Récemment, Coutsias *et al.* (2004) ainsi que Kneller (2005) ont montré l'équivalence des approches matricielles et par quaternions.

Nous détaillons ici les étapes du calcul.

4.2.1.1 Définition du critère

On considère donc, parmi les deux molécules, l'une fixe : $\mathfrak{M}^O$, constituée des atomes $\{A_i^0, i \leq N_{\text{atomes}}\}$ et l'autre mobile : $\mathfrak{M}$ (atomes $\{A_i, i \leq N_{\text{atomes}}\}$), tandis qu'on cherche à minimiser le RMSD entre les deux en jouant sur la translation et la rotation de $\mathfrak{M}$.

[1] ce qui se démontre sur l'espace quotienté par le groupe des isométries affines positives de $\mathbb{R}^3$ (Steipe, 2002)

L'RMSD se définit alors par :

$$\text{RMSD} \triangleq \sqrt{\frac{1}{N_{\text{atomes}}} \sum_{i \leq N_{\text{atomes}}} \|A_i^0 - A_i\|^2}. \tag{4.1}$$

La superposition de certains atomes peut être plus ou moins importante (hydrogènes ou atomes lourds, etc.) et parfois peut ne pas nous intéresser du tout, c'est le cas par exemple lorsqu'on superpose des protéines sur la base de leurs squelettes uniquement ; dans ce contexte, il est intéressant de pouvoir fixer des poids (entre 0 et 1) pour chaque atome, dans le critère à minimiser qui s'écrit alors (après élévation au carré par souci de simplicité) :

$$\varepsilon \triangleq \text{RMSD}^2 = \sum_{i \in I} \omega_i \|\mu(A_i) - A_i^0\|^2, \tag{4.2}$$

$$\text{où} \quad \sum_i \omega_i = 1, \tag{4.3}$$

où nous avons noté μ la transformation (translation rotation) appliquée à $\mathfrak{M}$.

4.2.1.2 Translation

Pour déterminer la translation optimale ($\mu(A_i) = A_i + t$), dérivons l'expression de ε par rapport à t :

$$\varepsilon(t) = \sum_i \omega_i \left\langle A_i - A_i^0 + t | A_i - A_i^0 + t \right\rangle, \tag{4.4}$$

$$\frac{1}{2}\frac{d\varepsilon}{dt}(t) = \sum_i \omega_i \, {}^\top(A_i - A_i^0 + t), \tag{4.5}$$

où le transposé du vecteur V est le vecteur ligne noté ${}^\top V$.

Cette dérivée s'annule lorsque

$$t = \sum_i \omega_i (A_i^0 - A_i). \tag{4.6}$$

La translation optimale est donc celle qui superpose les barycentres (pondérés par les poids ω_i) des deux molécules (conforme à Kabsch 1976).

4.2.1.3 Rotation

Afin de décrire l'ensemble des rotations applicables à $\mathfrak{M}$, on pourrait utiliser les angles d'Euler, ou bien la détermination d'un axe et d'un angle de rotation, mais nous allons utiliser (équivalemment) les quaternions[2].

On rappelle que tout quaternion Q, normé ($Q_0^2+Q_1^2+Q_2^2+Q_3^2=1$), définit une isométrie de $\mathbb{R}^3$ dans lui-même par la relation :

$$\mu_Q(A) = QA\bar{Q}, \tag{4.7}$$

où l'on identifie, lorsqu'il n'y a pas d'ambiguïté, le vecteur de $\mathbb{R}^3$: ${}^\top(x,y,z)$ au quaternion pur $(0,x,y,z)$ et le réel r au quaternion réel $(r,0,0,0)$.

De plus, si Q est écrit sous la forme[3] $\cos(\alpha/2)+\sin(\alpha/2)\vec{u}$ ($\vec{u}\in\mathbb{R}^3$ normé), alors μ_Q est la rotation d'axe (orienté) porté par $\vec{u}$ et d'angle α.

Calcul de $\varepsilon(Q)$. On pourrait, comme pour la translation, dériver ε par rapport à Q, il faut cependant prendre en compte le fait que le quaternion de la rotation doit respecter la condition de normalité ; on devrait alors introduire un terme supplémentaire de type multiplicateur de Lagrange. Ici, nous allons commencer par simplifier l'expression du critère.

Notons les parties réelle et imaginaire de Q :

$$\begin{aligned} Q &= \gamma + q, \\ \gamma &\in \mathbb{R}, \quad q\in\mathbb{R}^3, \\ \gamma^2+\|q\|^2 &= 1. \end{aligned} \tag{4.8}$$

$\mu_Q(A)$ s'écrit alors :

$$QA\bar{Q} = A + (\gamma^2 - 1 - \|q\|^2)A + 2\gamma(q\wedge A) + 2\langle q|A\rangle\, q. \tag{4.9}$$

À ce point, nous passons en coordonnées relatives et définissons M et D :

$$M = \frac{A+A^0}{2}, \tag{4.10}$$

$$D = A - A^0. \tag{4.11}$$

[2] voir annexe A

[3] cette décomposition est unique si on impose $\alpha\in[0;2\pi]$ et $|Q_0|\neq 1$

Notons également $|Q|$ la norme ou le module du quaternion Q.

Il en découle (après moult calculs...) :

$$\begin{aligned}\|QA\bar{Q} - A^0\|^2 &= \left((|Q|^2+1)^2 - 4\gamma^2\right)\|M\|^2 + \frac{1}{4}\left((|Q|^2-1)^2 + 4\gamma^2\right)\|D\|^2 \\ &\quad + \left(|Q|^4 - 1\right)\langle D|M\rangle + 4\gamma\langle q|M\wedge D\rangle \\ &\quad -4\langle q|M\rangle^2 + \langle q|D\rangle^2 . \end{aligned} \tag{4.12}$$

Ce qui, pour un quaternion de norme 1 nous donne :

$$\begin{aligned}\|QA\bar{Q} - A^0\|^2 &= ({}^\top qq)\|2M\|^2 + \gamma^2\|D\|^2 + 2\gamma\,{}^\top q(2M\wedge D) \\ &\quad -{}^\top q(2M)\,{}^\top(2M)q + {}^\top qD^\top Dq. \end{aligned} \tag{4.13}$$

Pour $\varepsilon(Q)$, nous obtenons alors :

$$\begin{aligned}\varepsilon(Q) &= ({}^\top qq)\sum_i \omega_i\|2M_i\|^2 + \gamma^2\sum_i \omega_i\|D_i\|^2 + 2\gamma\,{}^\top q(\sum_i \omega_i 2M_i\wedge D_i) \\ &\quad -{}^\top q\sum_i \omega_i(2M_i)^\top(2M_i)q + {}^\top q\sum_i \omega_i {D_i}^\top D_i q. \end{aligned} \tag{4.14}$$

Posons naturellement les matrices et vecteur suivants :

$$\left\{\begin{array}{lcll}\Lambda &=& \sum_i \omega_i 2M_i\wedge D_i & \in\mathbb{R}^3, \\ N &=& \sum_i \omega_i(2M_i)^\top(2M_i) & \in\mathcal{M}_3(\mathbb{R}), \\ \Delta &=& \sum_i \omega_i {D_i}^\top D_i & \in\mathcal{M}_3(\mathbb{R}).\end{array}\right. \tag{4.15}$$

Remarquons, au passage que

$$\sum_i \omega_i\|2M_i\|^2 = \mathrm{tr}(N), \tag{4.16}$$

$$\sum_i \omega_i\|D_i\|^2 = \mathrm{tr}(\Delta). \tag{4.17}$$

où « tr » représente la trace de la matrice. Ainsi, nous avons :

$$\begin{aligned}\varepsilon(Q) &= {}^\top q\left[\Delta - N - \mathrm{tr}(N).Id_3\right]q + \mathrm{tr}(\Delta)\gamma^2 + 2{}^\top\Lambda.\gamma.q \\ &= {}^\top Q\begin{pmatrix}\mathrm{tr}(\Delta) & {}^\top\Lambda \\ \Lambda & \Delta - N - \mathrm{tr}(N).Id_3\end{pmatrix}Q, \end{aligned} \tag{4.18}$$

$$\varepsilon(Q) = {}^\top QXQ. \tag{4.19}$$

Interprétation du résultat. L'expression de $\varepsilon(Q)$ obtenue en (4.19) nous épargne tout travail de dérivation car le critère : $\mathrm{RMSD}^2 = \min_{|Q|=1} \varepsilon(Q)$ apparaît comme la *norme opérateur* de la matrice symétrique positive X.

Voyons tout d'abord quelques propriétés de la matrice X :

1. X est symétrique, elle est donc diagonalisable dans une base orthonormée de vecteurs propres ;

2. elle est positive (en effet, $\forall Q, \varepsilon(Q) \geq 0$, ce qui se vérifie aussi avec l'équation (4.18)), ses quatre valeurs propres sont donc positives ou nulles ;

3. elle n'est pas forcément définie (il existe une valeur propre nulle), en effet, pour $\mathfrak{M} = \mathfrak{M}^0$, X a une colonne de zéros.

Notons d_i, $i = 1 \ldots 4$, les valeurs propres de X de telle sorte que

$$0 \leq d_1 \leq d_2 \leq d_3 \leq d_4. \tag{4.20}$$

Le minimum sur la sphère unité de $\left|{}^{\top}Q.X.Q\right|$ est, par définition, la « norme opérateur » de X et est donné par la plus petite des valeurs propres (en module). Ainsi,

$$\varepsilon = \min_{|Q|=1} \varepsilon(Q) = d_1. \tag{4.21}$$

Résolution finale. Il nous reste donc à déterminer la plus petite valeur propre de X, ce qui peut être fait en utilisant la méthode de la puissance sur X^{-1} (si X n'est pas inversible, son déterminant est nul et sa plus petite valeur propre est $0 \ldots$) ; cependant, une autre solution a été envisagée : elle consiste à calculer le polynôme caractéristique de X, puis à déterminer la première racine en utilisant l'algorithme de Newton initialisé à 0 (pratiquement, 5 à 10 itérations suffisent).

La rotation optimale est alors obtenue par le quaternion propre correspondant à d_1.

Dernière remarque : pour réduire le temps de calcul, on peut réexprimer la matrice X directement en terme des coordonnées atomiques des deux molécules

(x_i, y_i, z_i) et (x_i^0, y_i^0, z_i^0) :

$$X = \sum_i \omega_i \begin{pmatrix} (x_i - x_i^0)^2 & & & \\ +(y_i - y_i^0)^2 & 2(z_i y_i^0 - y_i z_i^0) & 2(x_i z_i^0 - z_i x_i^0) & 2(y_i x_i^0 - x_i y_i^0) \\ +(z_i - z_i^0)^2 & & & \\ & (x_i - x_i^0)^2 & & \\ 2(z_i y_i^0 - y_i z_i^0) & +(y_i + y_i^0)^2 & -2(x_i y_i^0 + y_i x_i^0) & -2(x_i z_i^0 + z_i x_i^0) \\ & +(z_i + z_i^0)^2 & & \\ & & (x_i + x_i^0)^2 & \\ 2(x_i z_i^0 - z_i x_i^0) & -2(x_i y_i^0 + y_i x_i^0) & +(y_i - y_i^0)^2 & -2(y_i z_i^0 + z_i y_i^0) \\ & & +(z_i + z_i^0)^2 & \\ & & & (x_i + x_i^0)^2 \\ 2(y_i x_i^0 - x_i y_i^0) & -2(x_i z_i^0 + z_i x_i^0) & -2(y_i z_i^0 + z_i y_i^0) & +(y_i + y_i^0)^2 \\ & & & +(z_i - z_i^0)^2 \end{pmatrix}. \quad (4.22)$$

Le polynôme caractéristique de X est alors de la forme :

$$P_X(\lambda) = \lambda^4 - \mathrm{tr}(X)\lambda^3 + A\lambda^2 + B\lambda + \det(X), \quad (4.23)$$

où A et B sont des expressions volumineuses des coordonnées, mais simples à implémenter.

Pour trouver le quaternion propre correspondant, c'est-à-dire la rotation qu'il faut appliquer à $\mathfrak{M}$, il suffit de réaliser un pivot de Gauss sur la matrice $X - d_1 I$ pour trouver un vecteur propre.

Ceci achève notre calcul. Nous voyons maintenant les résultats en termes de temps de calculs.

4.2.1.4 Résultats et performances.

Le calcul du RMSD à proprement parler est négligeable devant le temps nécessaire à la reconstruction de la géométrie ou même à la lecture des fichiers de coordonnées. Il faut compter environ 600μs pour reconstruire la géométrie d'une molécule de 300 atomes (voir paragraphe 3.3.3) et seulement 40μs pour estimer son RMSD avec une autre géométrie (en nombres flottants 64 bits). Enfin, pour déterminer la translation-rotation qui superpose les deux structures[4], il faut ajouter un temps de calcul d'environ 10μs.

Ces temps sont donnés à titre indicatif pour une molécule d'environ 300 atomes, qui est l'ordre de grandeur des molécules étudiées. En réalité, les temps de calculs se corrèlent avec le nombre d'atomes de la molécule (la figure 4.2 donne les temps obtenus sur une station de travail HP xw6200 Xeon 3,4 GHz).

[4]Le calcul de la valeur propre ne nécessitant pas celui du vecteur propre correspondant

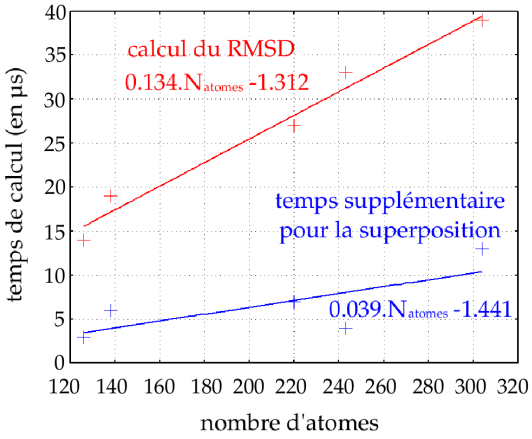


FIG. 4.2: temps de calcul (en μs) nécessaire à l'estimation du RMSD (courbe rouge) et temps supplémentaire pour déterminer la superposition optimale (courbe bleue).

À titre de comparaison, McLachlan (en 1982) a rapporté un temps de superposition (sans reconstruction de la géométrie) de 3ms pour des molécules « de tailles utiles » sur IBM 370/165.

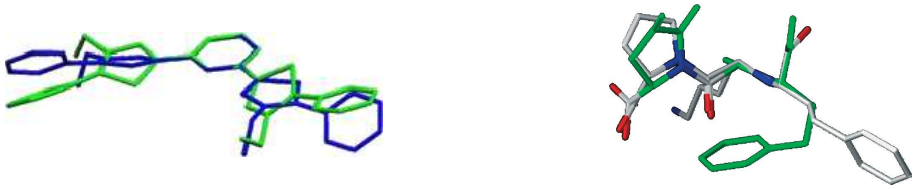


FIG. 4.3: exemples de superpositions de différentes conformations moléculaires.

4.2.2 Un score de superposition pharmacophorique flou

Dans le cas général, on cherche à relater deux molécules n'ayant pas nécessairement la même liste d'atomes... L'approche précédente n'est donc plus valable et la notion d'RMSD n'a plus de sens. Afin de réutiliser les résultats, on peut essayer de mettre en évidence des couples d'atomes à apparier (pris dans chacune des molécules) ; ces atomes (voire groupes d'atomes) remplissant des fonctions particulières sont appelés *pharmacophores* (figures 4.4).

La difficulté dans l'utilisation de ces pharmacophores est de gérer les équivalences : une charge négative peut remplacer une autre charge négative, un cycle aromatique peut remplacer un autre cycle aromatique, voire (dans une certaine mesure) un groupement hydrophobe. Pour formaliser tout cela, on abandonne les types précis des atomes pour un nombre restreint de catégories pharmacophoriques (notées $\mathfrak{Tp}$) telles que celles présentées dans le tableau 4.1.

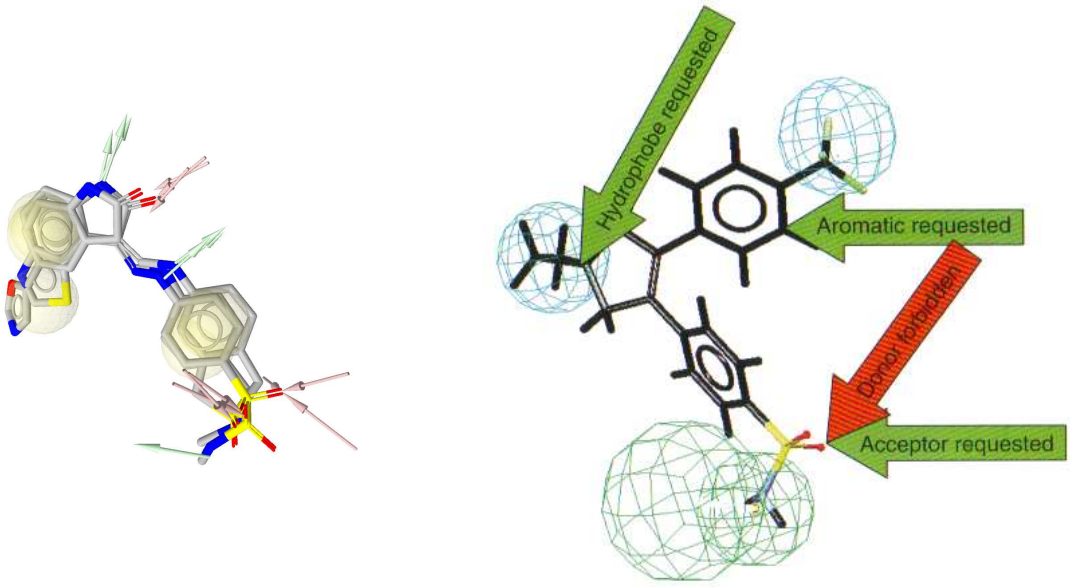


FIG. 4.4: les différents groupes fonctionnels de la molécule forment des motifs pharmacophoriques ; la figure de droite est extraite de Oprea (2005).

type pharmacophorique	abréviation
aromatique	Ar
donneur d'hydrogène	HD
accepteur d'hydrogène	HA
hydrophobe	Hp
charge négative	NC
charge positive	PC

TAB. 4.1: principaux types pharmacophoriques avec leurs abréviations.

4.2.2.1 Définition du score

On modélise alors les pharmacophores par des sources générant en tout point de l'espace un « champ pharmacophorique » gaussien en fonction de leur type (équation (4.24) et figure 4.5)

$$\forall \text{ pharmacophore } \Phi, \text{ de type } T \text{ au point } A \text{ et } P \text{ un point de l'espace,} \quad F_A(P) = k'_T.e^{-\alpha'_T.d^2(A,P)}, \tag{4.24}$$

où k'_T et α'_T sont des constantes caractérisant le pharmacophore.

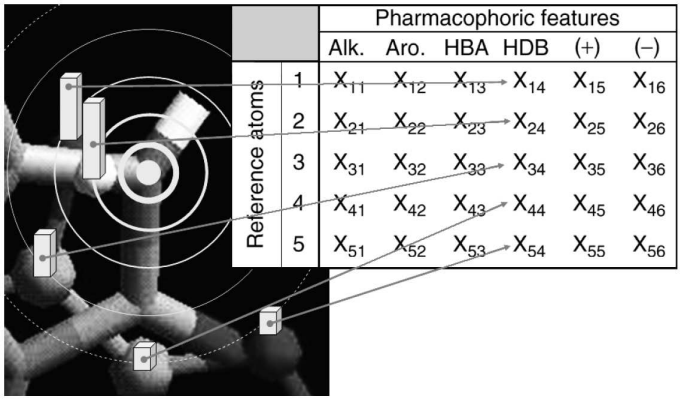


FIG. 4.5: chaque atome génère des champs pharmacophoriques. Figure extraite de Oprea (2005).

Une molécule $\mathfrak{M}^0$ est constituée d'un ensemble de pharmacophores Φ_i^0, indicés par $i \in I^0$, de types T_i^0 et de coordonnées A_i^0. Le champ total de type T, généré par $\mathfrak{M}^0$ est donc :

$$F_T^0(P) = k'_T \sum_{i \in I^0} \delta(T_i^0 = T) e^{-\alpha'_T d^2(A_i^0, P)}. \tag{4.25}$$

Puisqu'il existe des pharmacophores éventuellement équivalents, remplaçons dès à présent la fonction de Dirac $\delta(T_i^0 = T) \in \{0, 1\}$ par une pondération $\omega(T_i^0, T) \in [0, 1]$.

Pour mesurer le « degré de similitude pharmacophorique » de deux molécules ($\mathfrak{M}^0$ et $\mathfrak{M}$), on étudie le produit scalaire de leurs champs pharmacophoriques totaux :

$$\langle F_T^0 | F_T \rangle = k'^2_T \sum_{i \in I^0, j \in I} \omega(T_i^0, T)\omega(T_j, T) \int_{\mathbb{R}^3} \exp\left\{-\alpha'_T \left[d^2(A_i^0, P) + d^2(A_j, P)\right]\right\} d^3P. \tag{4.26}$$

Or on montre aisément que

$$\int_{\mathbb{R}^n} \exp\left\{-\alpha'_T\left[d^2(A_i^0,P)+d^2(A_j,P)\right]\right\} d^nP = \left(\frac{\pi}{2\alpha'_T}\right)^{\frac{n}{2}} e^{-\frac{\alpha'_T}{2}d^2(A_i^0,A_j)}.$$

En posant, pour tout $T \in \mathfrak{Tp}$

$$k_T^2 = k_T'^2\sqrt{\frac{\pi}{2\alpha'_T}}^3, \qquad \alpha_T = \frac{\alpha'_T}{2}, \tag{4.27}$$

on a

$$\langle F_T^0|F_T\rangle = k_T^2 \sum_{i,j} \omega(T_i^0,T)\omega(T_j,T)e^{-\alpha_T d^2(A_i^0,A_j)}. \tag{4.28}$$

En particulier,

$$\begin{aligned}\langle F_T^0|F_T^0\rangle &= k_T^2 \sum_{(i,j)\in(I^0)^2} \omega(T_i^0,T)\omega(T_j^0,T)e^{-\alpha_T d^2(A_i^0,A_j^0)} \\ \|F_T^0\|^2 &= k_T^2 \sum_i \omega(T_i^0,T)\left(1+2\sum_{j>i}\omega(T_j^0,T)e^{-\alpha_T d^2(A_i^0,A_j^0)}\right).\end{aligned} \tag{4.29}$$

On définit alors un critère « normalisé » $\mathfrak{C}$, basé sur le produit scalaire (qui prend donc en compte la colinéarité des champs) mais qui fasse également intervenir une comparaison sur les normes (équation (4.30)) :

$$\mathfrak{C}_T = \frac{2\langle F_T^0|F_T\rangle}{\|F_T^0\|^2+\|F_T\|^2}. \tag{4.30}$$

Ce critère vérifie les propriétés suivantes :

- $\mathfrak{C}_T \geq 0$, car le produit scalaire ne fait intervenir que des fonctions positives ;
- $\mathfrak{C}_T = 0$ si et seulement si $F_T^0 \perp F_T$ au sens du produit scalaire dans $\mathcal{L}^2(\mathbb{R}^3)$;
- d'après l'inégalité de Cauchy-Schwarz,

$$\mathfrak{C}_T \leq \frac{2\|F_T^0\|.\|F_T\|}{\|F_T^0\|^2+\|F_T\|^2} = \tanh\left(2\arg\tanh\frac{\|F_T^0\|}{\|F_T\|}\right) \leq 1,$$

- et $\mathfrak{C}_T = 1$ si et seulement si $\|F_T^0 - F_T\|^2 = 0$ et donc $F_T^0 = F_T$.

Nous avons donc autant de critères que de types pharmacophoriques T et il est possible de construire un score global en sommant (éventuellement avec des pondérations) tous ces critères, mais on peut tout autant considérer une approche multi-critère.

Nous exposons ici une expression de la dérivée par rapport aux degrés de liberté de translation et rotation du critère $\mathfrak{C}_T$ (4.30), qui pourrait servir à l'implémentation d'un algorithme de gradient conjugué.

Les normes des champs F_T et F_T^0 sont invariantes par isométrie (puisqu'elles reposent sur des distances internes), il suffit donc de dériver le produit scalaire. Celui-ci s'écrivant comme somme de termes simples, notons f la fonction

$$f(t,q) = \exp\left[-\alpha d^2(A, qB\overline{q} + t)\right], \tag{4.31}$$

où A et B sont deux points donnés de l'espace, t un vecteur quelconque de $\mathbb{R}^3$ et q un quaternion de norme 1.

Rappelons également que

$$\frac{\partial}{\partial v}\langle u|u\rangle = 2^\top u \left(\frac{\partial u}{\partial v}\right).$$

Ainsi, par exemple,

$$\frac{\partial f}{\partial t} = -2\alpha f(t,q)^\top \left(qB\overline{q} + t - A\right). \tag{4.32}$$

Calculons $\frac{\partial(qB\overline{q})}{\partial q}$:

$$\begin{aligned}
(q+dq)B(\overline{q+dq}) - qB\overline{q} &\approx dqB\overline{q} + qB\overline{dq} \\
&\approx 2\ \mathrm{Im}(dqB\overline{q}) \\
&\approx 2\left[(B\wedge\vec{q} - q_0B)\wedge\vec{dq} + \langle q|B\rangle\vec{dq} - (B\wedge\vec{q} - q_0B)dq_0\right] \\
&\approx \underbrace{2\Big(q_0B - B\wedge\vec{q}\,\Big|\,\langle q|B\rangle I_{\mathbb{R}^3} - \Lambda_{q_0B - B\wedge\vec{q}}\Big)}_{\triangleq W_{(B,q)}} \times \begin{pmatrix} dq_0 \\ \vec{dq} \end{pmatrix} \\
&\approx W_{(B,q)}dq.
\end{aligned} \tag{4.33}$$

Où $W_{(B,q)}$ est donc une matrice 3×4.

Ainsi,

$$\frac{\partial\left(\|qB\overline{q} + t - A\|^2\right)}{\partial q} = 2^\top \left(qB\overline{q} + t - A\right) W_{(B,q)}. \tag{4.34}$$

De sorte que,

$$\frac{\partial f}{\partial q} = -4\alpha.f(t,q)\underbrace{{}^\top\left(qB\overline{q} + t - A\right)}_{1\times 3}\underbrace{W_{(B,q)}}_{3\times 4}. \tag{4.35}$$

Et finalement

$$\frac{\partial \mathfrak{C}_T}{\partial (t,q)} = \frac{-4\alpha_T}{\|F_T^0\|^2 + \|F_T\|^2} k_T^2 \times \sum_{i,j} \omega(T_i,T)\omega(T_j,T) e^{-\alpha_T d^2(A_i^0, A_j)} \times^\top \left(A_j - A_i^0\right) \left(I_{\mathbb{R}^3} \otimes 2W_{(B,q)}\right), \quad (4.36)$$

où $\otimes$ représente la simple juxtaposition des matrices.

Remarque : les expressions (4.33) ainsi que (4.34), qui fournissent l'expression de la dérivée des coordonnées atomiques et des distances interatomiques par rapport aux degrés de liberté de rotation, pourront servir dans un calcul du gradient de l'énergie lorsqu'on fera du *docking*.

4.2.2.2 Heuristiques de recherche

Nous disposons maintenant d'un critère précis à optimiser afin d'obtenir une superposition *intelligente* (i.e. chimique) de deux molécules. Il n'y a cette fois que six degrés de liberté et on ne cherche *a priori* que l'optimum global (bien que la connaissance des principales superpositions sous-optimales serait un plus). Ce qui épice la question cette fois, c'est qu'il faut cribler d'énormes bases de données[5] en quelques minutes maximum.

De plus, les coordonnées atomiques ne sont généralement pas disponibles. Bien que l'exercice de prédiction de la géométrie pour des molécules aussi simples soit plus facile que pour des molécules de tailles supérieures, l'absence des coordonnées allonge considérablement les temps de calculs.

C'est pourquoi les algorithmes classiques de recherche opérationnelle ne sont pas envisageables. Nous avons alors cherché à combiner les résultats concernant le calcul du RMSD avec des approches topologiques par *descripteurs*. Les descripteurs sont des indices calculés à partir de la molécule (types atomiques, graphe de liaisons et éventuellement coordonnées atomiques) qui permettent de les classifier et/ou de prédire certaines de leurs propriétés (propriétés élémentaires ou plus élaborées comme les temps de repliement ou la présence de minima secondaires sur les chemins de repliement, voir Chavez *et al.*, 2004). Ils sont dits *topologiques* lorsqu'aucune information sur la structure tridimensionnelle n'est utilisée.

Leur nombre et la possibilité de les calculer *off-line* font de ces descripteurs des alliés de choix. Une comparaison des molécules sur la base de leurs descripteurs est en effet beaucoup plus simple et rapide que celle de leurs structures topologiques. De

[5] Irwin et Shoichet (2005) reportent 5×10^6 composés dans la base de données ZINC

plus, leur non-injectivité est compensée par la variété des descripteurs imaginables. Par exemple, les trois descripteurs binaires D_1, D_2 et D_3 dans le tableau suivant permettent de discriminer entièrement les huit molécules :

molécule n°	D_1	D_2	D_3
1	0	0	0
2	1	0	0
3	0	1	0
4	1	1	0
5	0	0	1
6	1	0	1
7	0	1	1
8	1	1	1

TAB. 4.2: exemple de trois descripteurs permettant de discriminer huit molécules.

Nous présentons, ci-après, un type particulier de descripteurs, qui caractérise les motifs pharmacophoriques.

4.2.3 Les descripteurs de motifs pharmacophoriques

On essaye de mettre en relation les molécules sur la base de leurs pharmacophores, ou plus exactement en mettant en évidence la présence de *motifs* pharmacophoriques. Ainsi, les descripteurs dits à « 2 points » répertorient les paires de pharmacophores avec les distances qui les séparent ; les descripteurs 3 points répertorient les triplets (voir figure 4.6), etc. Plus les descripteurs sont d'un ordre important, plus ils captent d'informations, ainsi, les descripteurs 4 points peuvent saisir jusqu'à la chiralité des atomes. Malheureusement, s'il y a N_{atomes} descripteurs à 1 point, on compte $\binom{N_{\text{atomes}}}{n} \sim o(N_{\text{atomes}}^n)$ descripteurs n points ($N_{\text{atomes}}(N_{\text{atomes}}-1)/2$ descripteurs 2 points et $\binom{N_{\text{atomes}}}{3}$ descripteurs 3 points). De plus, la complexité apparaît également dans l'énumération des motifs possibles : un descripteur 2 points se caractérise par deux pharmacophores et une unique distance, un descripteur 3 points nécessite trois pharmacophores et trois distances (même si les triplets ne respectant pas l'inégalité triangulaire peuvent être écartés) et pour $n > 2$, il faut n pharmacophores et $3(n-2)$ distances pour caractériser un descripteur n points. L'ordre choisi des descripteurs est donc rapidement limité par les ressources informatiques.

NB : il est toujours possible d'utiliser comme distance, la distance topologique (i.e. le nombre de liaisons séparant deux atomes), de sorte que ces descripteurs ont tous une version géométrique (dite 3D) et une version topologique (dite 2D).

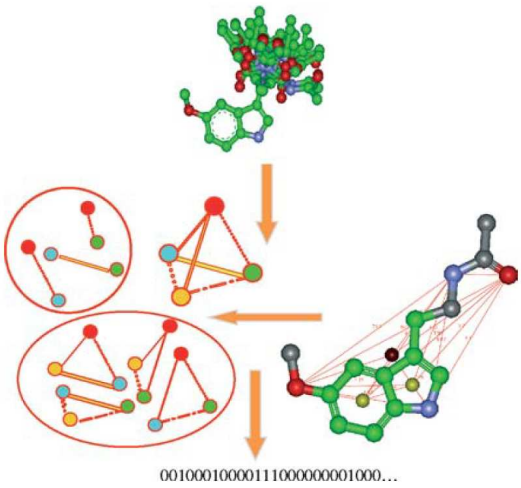


FIG. 4.6: caractérisation de la molécule par ses triplets pharmacophoriques, figure tirée de (Good *et al.*, 2004).

En se donnant un ensemble de polyèdres de *base* : $\mathfrak{B} = (\Delta_1, \Delta_2, \ldots, \Delta_n)$, on peut alors caractériser la molécule entrante par sa signature selon qu'elle contient ou non chacun des polyèdres Δ_i. On définit donc une fonction à valeurs dans $\mathbb{N}^n$ où la coordonnée i est égale au nombre de motifs Δ_i rencontrés dans la molécule.

L'apport de la logique floue. La difficulté, mise en évidence sur la figure 4.7, est que ces descripteurs ne sont pas *continus* au sens où deux motifs très proches peuvent être comptabilisés sur des motifs de base Δ_i différents. Pour pallier à ce défaut classique, l'utilisation de la logique floue est préconisée (Ross, 2004), voir figure 4.8. Dans ce formalisme, deux polyèdres proches contribuent sensiblement de la même façon sur chaque polyèdre de base.

Un autre apport de la logique floue, est qu'il est possible d'encoder des différences minimes (bien que le codage soit toujours discrétisé sur les entiers, la précision est un paramètre modifiable). Puisqu'on sait qu'il existe des composés similaires ayant des activités différentes, (« activity cliffs ») il est primordial que les descripteurs puissent capter ces différences. De plus, cela permet de réduire la taille de la base de triangles utilisée.

Nous avons donc opté pour des descripteurs topologiques flous à 3 points dont les avantages ont été mis en évidence dans l'article (Bonachera *et al.*, 2006). On forme alors la base $\mathfrak{B}$ en énumérant tous les triplets de types pharmacophoriques avec les distances possibles : $T_1^0 d_{2,3}^0 T_2^0 d_{3,1}^0 T_3^0 d_{1,2}^0$ où T_i^0 est le type pharmacophorique du sommet i et $d_{i,j}^0$ est la distance topologique (donc entière) entre les sommets i et

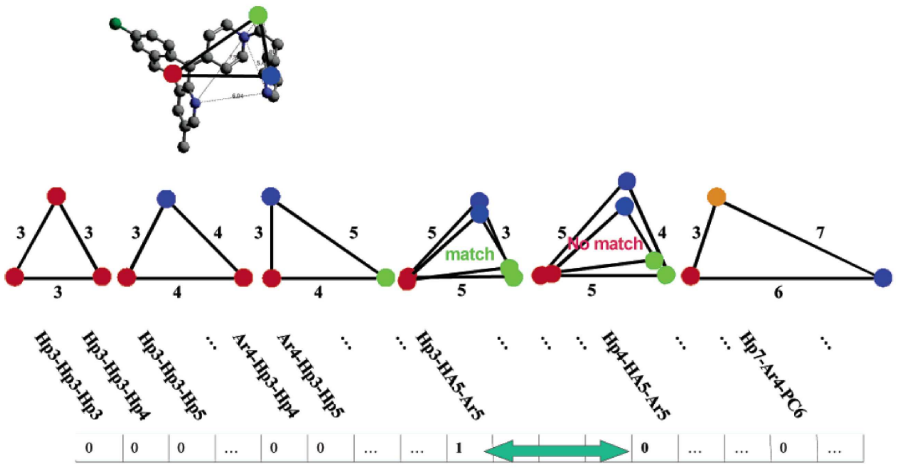


FIG. 4.7: deux molécules « proches » ne sont pas caractérisées par des descripteurs proches...

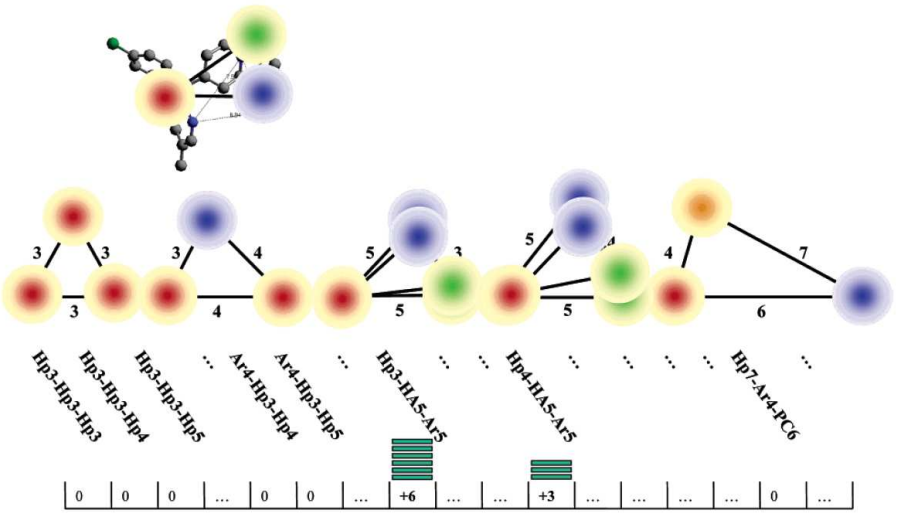


FIG. 4.8: l'utilisation de la logique floue préserve la continuité.

j, respectant l'inégalité triangulaire. Remarquons qu'il est possible de standardiser l'appellation des triplets afin d'éviter les redondances issues de transposition des sommets.

Pour une molécule donnée, on énumère chacun de ses triangles et on calcule la contribution de chaque triangle de type $T_1 d_{2,3} T_2 d_{3,1} T_3 d_{1,2}$ selon la composante $T_1^0 d_{2,3}^0 T_2^0 d_{3,1}^0 T_3^0 d_{1,2}^0$ grâce au critère $\mathfrak{C} = \sum_{T \in \mathfrak{T}\mathfrak{p}} \mathfrak{C}_T$ (équation (4.30)) de la manière suivante :

- consistance des types pharmacophoriques : il faut que $\omega(T_1^0, T_1)\omega(T_2^0, T_2)\omega(T_3^0, T_3) > 0$,
- prépositionnement des deux triangles selon l'algorithme de RMSD (§ 4.2.1 amont) utilisant un appariement dicté par les types pharmacophoriques (en cas de solutions multiples, celle qui donne le meilleur $\mathfrak{C}$ est retenue),
- une optimisation locale permet d'ajuster, dans le plan, la superposition des deux triplets

Remarque : le score $\mathfrak{C}$ prend également en compte la présence des différents états d'ionisation de la molécule au pH considéré et utilise alors une moyenne pondérée des sous-scores. Ceci est réalisé grâce à un outil ChemAxon[6].

La valeur finale de $\mathfrak{C}$ donne alors un score qui est mis à l'échelle pour couvrir l'intervalle $[0, 50]$ et qui donne la contribution recherchée.

La molécule est ainsi caractérisée par un vecteur de descripteurs pouvant être calculé *off-line*.

4.2.4 Résultats

Les principes de superposition pharmacophorique floue utilisant des descripteurs 3 points et l'algorithme de superposition fondé sur les quaternions ont été appliqués à une base de données de molécules commerciales (base BioPrint). Il a été montré que la distance entre les descripteurs introduits ci-dessus était plus revélatrice des « distances » entre les véritables activités chimiques des molécules. Ces résultats reposent en grande partie sur l'utilisation de la logique floue, mais également sur l'utilisation des différents états d'ionisation.

[6] http ://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (août 2007).

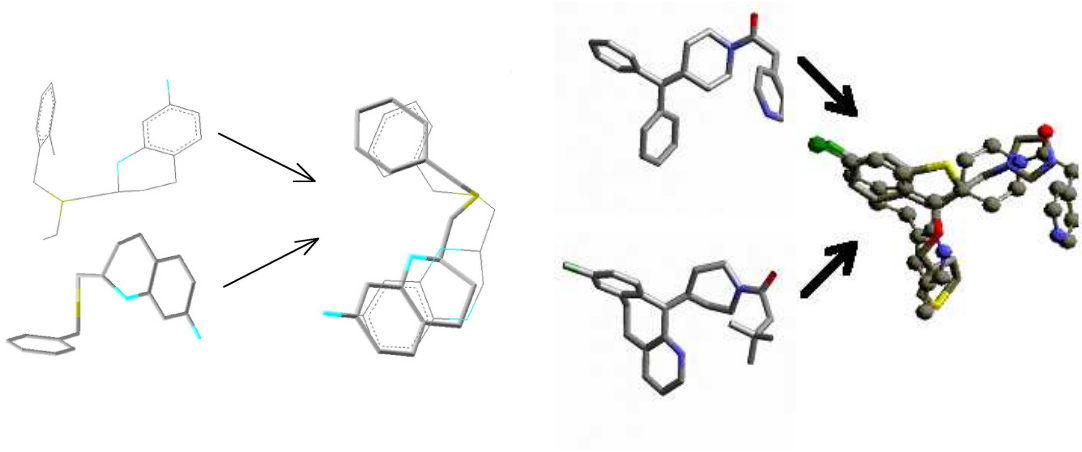


FIG. 4.9: superposition de composés différents sur la base de leurs pharmacophores.

FIG. 4.10: autre exemple de superposition pharmacophorique.

4.3 L'échantillonnage conformationnel de deux molécules

Ayant développé ces outils, nous allons revenir au problème initial du *docking* moléculaire. Si les approches QSAR peuvent servir à la recherche de cibles potentielles dans les bases de données, nous allons voir qu'elles vont aussi permettre de déterminer des sites possibles de fixation ainsi que des positions probables pour interagir.

Outre les résultats encourageants, l'intérêt de l'approche par triplets pharmacophoriques flous réside en effet dans le fait que la superposition et le calcul du score sont quasiment indépendants de la façon dont est fabriquée la base de triangles $\mathfrak{B}$: une généralisation à des descripteurs géométriques (plutôt que topologiques) sera donc implémentable. Le but est de mettre en évidence dans les deux molécules, des motifs pharmacophoriques complémentaires et, ainsi, de proposer des sites de fixation potentiels et des positionnements approximatifs éventuels. Ceci équivaut à l'étape de recherche d'invaginations et de complémentarité de formes développée par certains auteurs (Venkatachalam *et al.*, 2003).

4.3.1 Développements futurs

Dans le *docking*, on distingue trois niveaux de précision/complexité :

- le *docking* rigide, où l'on tente de mettre en évidence les possibilités d'accrochages des molécules sur la base de leur conformation préférentielle (et uniquement celle-ci), sans prise en compte des flexibilités des agents. C'est le cas de la recherche de complémentarités de formes (Venkatachalam *et al.*, 2003), mais aussi de toutes les approches par descripteurs géométriques ;
- le *docking* semi-flexible (Vieth *et al.*, 1998a; Klepeis *et al.*, 1998), prend de plus en compte la flexibilité du ligand ;
- enfin, le *docking* flexible prend en compte, à la fois la flexibilité du ligand, mais aussi celle du site actif (Najmanovich *et al.*, 2000; Hornak et Simmerling, 2007).

Remarque : pour des revues dans ce domaine, nous citons Mendez *et al.* (2003) qui ont analysé les résultats du concours de *docking* : CAPRI ; Bursulaya *et al.* (2003) qui ont comparé différents algorithmes (Autodock, DOCK, FlexX, GOLD, ICM) pour le *docking* ; et enfin Wang *et al.* (2003) qui ont comparé 11 fonctions de score pour le *docking* et l'estimation des affinités[7].

Nous proposons, comme développements à venir, d'adapter la méthode d'échantillonnage conformationnel par $C_S G_A$ à l'échantillonnage de deux molécules, prenant en plus en compte les degrés de liberté de la translation-rotation du ligand. Le cas considéré correspond au *docking* d'un ligand dans le site actif d'une protéine ou d'un compexe plus important.

Afin d'éviter les inconvénients du *docking* rigide, on peut envisager, non pas de positionner une conformation du ligand dans une conformation du site, mais de faire un *docking* rigide entre les familles de conformations obtenues après pré-échantillonnage de chacun des deux acteurs. Ces positionnements pourront être réalisés en utilisant des triplets pharmacophoriques flous géométriques.

Les géométries ainsi obtenues permettront alors d'initialiser des populations de solutions pour la deuxième étape de l'algorithme, consistant à échantillonner simultanément les deux molécules avec leurs degrés de liberté respectifs et leur positionnement relatif (*docking* complètement flexible).

Remarque : des stratégies complémentaires devront certainement être considérées et d'autres, adaptées. C'est le cas de l'optimisation par gradient qui va nécessiter

[7] pour cette étude, seule une moitié des heuristiques testées ont un taux de réussite supérieur à 66% pour la prédiction de la structure et seulement 4 sur les 11 obtiennent un coefficient de corrélation entre affinités prédite et expérimentale supérieur à 0,5...

la dérivée de l'énergie par rapport aux degrés de liberté supplémentaires de positionnement (translation rotation). La partie difficile de ce calcul réside dans la dérivation des positions atomiques par rapport au quaternion de la rotation ; pour cela, nous renvoyons le lecteur à l'équation (4.34), obtenue lors de la dérivation du critère de superposition.

Un schéma global du *docking* moléculaire est proposé figure 4.11 qui résume la stratégie.

Une implémentation parallèle et un déploiement sur grille de calcul sera possible grâce à la nature combinatoire de notre approche où nous explorons l'ensemble des assemblages possibles des conformations prééchantillonnées. Ces assemblages seront ensuite soumis à un AG utilisant le modèle planétaire.

4.3.2 Remarques sur la fonction score

Suite à l'étude et à l'optimisation des paramètres de champ de forces, nous espérons que cette fonction énergie sera un critère suffisant pour estimer l'affinité et la probabilité des différentes conformations du complexe. Ainsi, il ne devrait pas être fait appel à des fonctions scores complémentaires, comme c'est le cas habituellement (Bissantz *et al.*, 2000).

Étant donné que le *docking* est fait dans un site actif, où seuls quelques degrés de liberté seront autorisés (chaînes latérales et éventuelles boucles impliquées), il sera aussi possible de calculer *off-line* les contributions des atomes fixes sur un maillage de l'espace. Ces données seront alors reprises au cours de l'exécution de l'algorithme en interpolant les contributions des points du maillage aux coordonnées atomiques réelles.

Enfin, un autre point important sera peut-être l'utilisation de l'approche multi-critère qui permet de suivre l'évolution d'une fonction *vectorielle* de *fitness*. Le coût de calcul supplémentaire est négligeable et la quantité d'information récoltée est plus importante que si l'on traite une moyenne pondérée de tous les critères utiles[8]. De plus, cette approche permet de mettre en évidence les éventuels effets antagonistes à travers la forme des fronts *Pareto* (Zitzler *et al.*, 2003).

Certains auteurs ont par exemple distingué les contributions de valence des

[8]cette approche a été récemment utilisée pour l'échantillonnage conformationnel, voir Vainio et Johnson, 2007

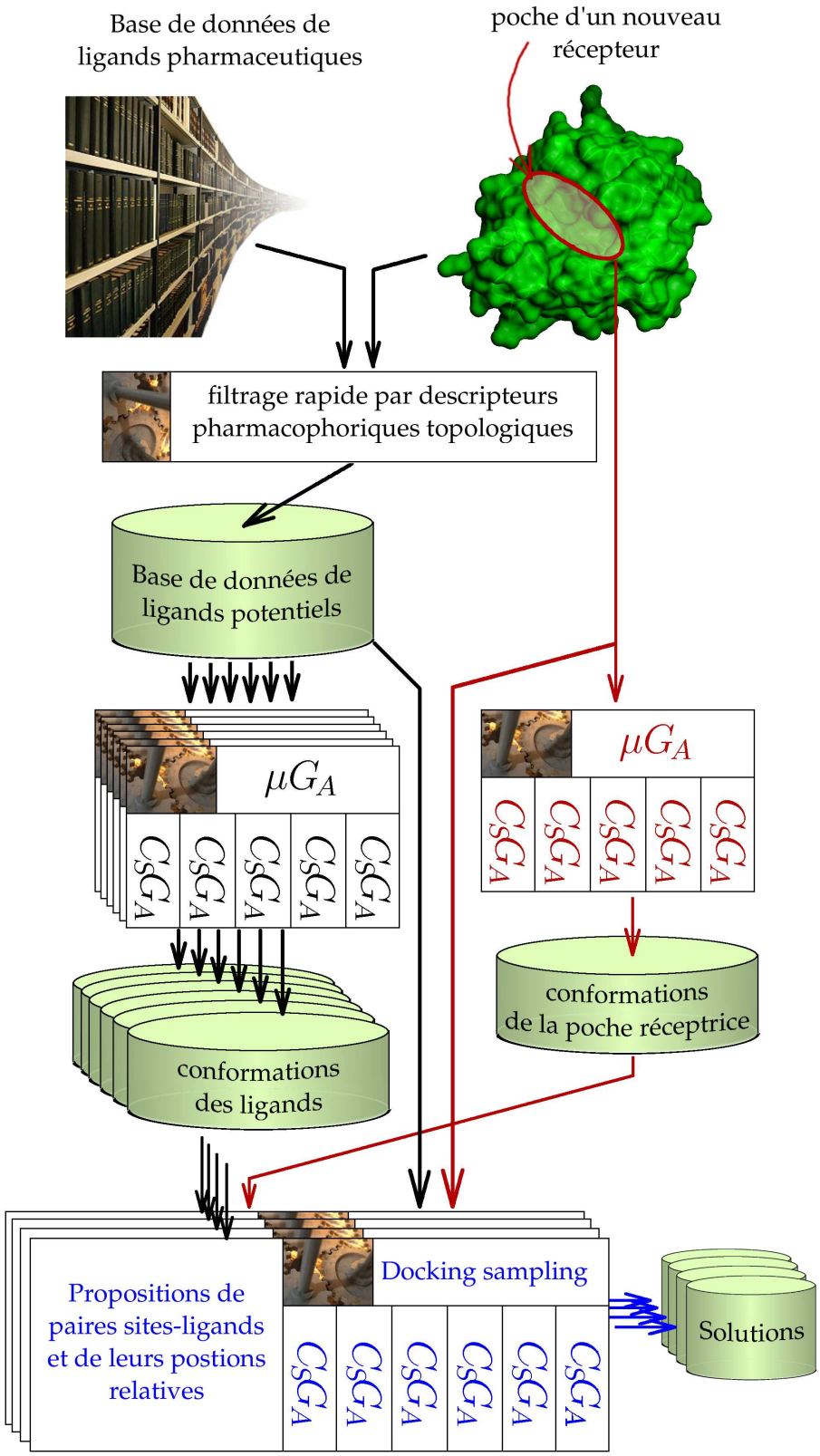


FIG. 4.11: ordonnancement des principales tâches pour le docking moléculaire.

contribution non-covalentes (Day *et al.*, 2002), mais, à notre sens, cette distinction n'est pas pertinente dans une description des degrés de liberté torsionnels.

À notre avis, il faut distinguer les points importants suivants :

1. les énergies *inter* et *intra*-moléculaires : effets antagonistes, déstabilisant les molécules individuelles pour stabiliser le complexe ;
2. l'entropie, comprise comme la *robustesse* d'une solution ;
3. l'enfouissement du ligand dans le complexe.

Pour l'entropie, on peut se rapporter au calcul page 37 où nous avions montré que S évoluait comme $\ln(V_D)$, V_D étant le volume du domaine D. Cependant, ce volume est difficile à estimer étant donnée l'extrême rugosité du paysage. En utilisant l'équation (1.15), on a également $S = \frac{E-G}{T}$ qui pourrait servir d'estimateur[9] dans notre approche par boîte ($\mathcal{R}$-conformations, voir section 3.7.3) : l'énergie interne E est donnée par la meilleure énergie dans la région échantillonnée, et l'énergie libre G est estimée grâce à la fonction de partition.

4.4 Conclusion

Dans ce chapitre, nous avons essentiellement traité le problème du positionnement relatif des acteurs. En commençant par la superposition de deux molécules identiques différant par leur conformation. Nous avons ensuite traité le cas plus général de le recherche de similarités chimiques entre deux acteurs, avançant ainsi progressivement vers le problème du positionnement du ligand dans la poche du site actif.

L'optimisation des temps de calculs par l'utilisation des quaternions a ouvert la porte au traitement à haut-débit de grandes bases de données pharmaceutiques pour lesquelles une stratégie de comparaison sur la base de triplets pharmacophoriques a été développée.

Concernant le *docking* proprement dit, les stratégies doivent encore être adaptées et/ou développées. Toutefois, nous avons présenté les étapes qui nous semblaient importantes de respecter et les critères qui pouvaient être utilisés.

Ceci achève la partie de notre travail concernant la « modélisation moléculaire ». Dans le chapitre suivant, nous nous intéressons encore aux interactions moléculaires, mais à une échelle beaucoup plus grossière, en considérant les *concentrations*

[9] nous gardons nos précautions vis-à-vis d'une telle définition de l'entropie, car nous n'avons pas, à notre disposition, le véritable ensemble de Boltzmann, mais un simple échantillonnage limité et épars.

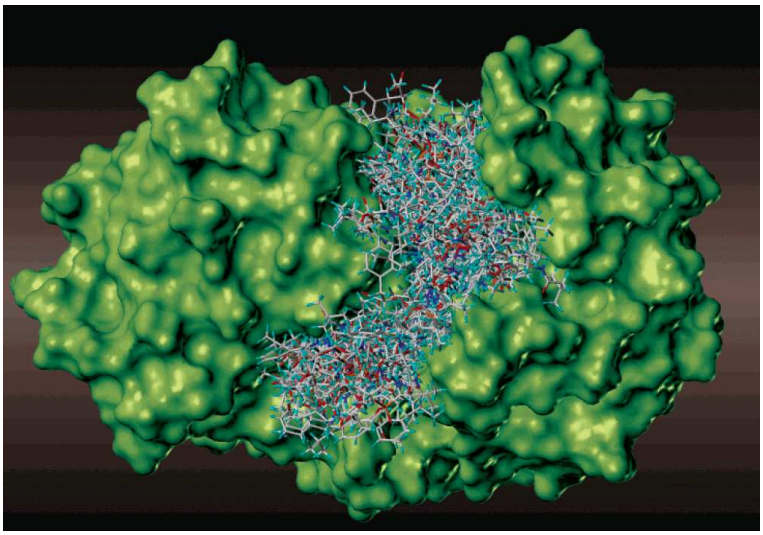


FIG. 4.12: figure extraite de Wang *et al.* (2003), montrant les différentes poses d'un ligand dans un site actif obtenues par Autodock (Morris *et al.*, 1998).

des acteurs. Nous verrons comment ces interactions influencent les dynamiques de réactions et, par suite, celle des réseaux de régulation de la cellule.

Deuxième partie

Les réseaux de régulation géniques

Chapitre 5

Modélisation des rythmes circadiens

5.1 Introduction

La complexité, déjà présente dans la structure géométrique des molécules et dans la prédiction de leurs interactions, explose à l'échelle de la cellule où des milliers d'acteurs interagissent en permanence. Ces acteurs activent ou répriment la production, la dégradation ou l'activité d'autres molécules, formant des réseaux d'interactions particulièrement compliqués. Toutes ces combinaisons possibles d'acteurs génèrent une variété extraordinaire de comportements différents, qui permettent à la cellule d'assurer ses fonctions vitales.

Nous aimerions savoir si une étude théorique pourrait permettre d'expliquer certains de ces comportements. Pour cela, nous avons considéré, en première approximation, les variables représentant la *concentration* des acteurs. En effet, si toutes les régulations reposent sur des interactions moléculaires (semblables à celles étudiées dans la première partie), la présence de milliards de molécules fait que l'on peut (dans une certaine mesure) abandonner la description individuelle de chaque acteur et de ses interactions, évitant ainsi de voir la cellule comme un assemblage combinatoire d'objets complexes. La façon dont ces concentrations sont modifiées au cours du temps dépend de certains mécanismes que nous rappelerons brièvement dans la section 5.2.

Bien entendu, cette approximation est sujette à caution et nous verrons, par la suite, ses limitations (section 5.3.5). Toutefois, elle autorise une première approche offrant quelques résultats (section 5.3.3).

Puisqu'il n'est pas possible d'envisager une modélisation de la cellule dans sa glo-

balité[1], nous nous sommes raccrochés à la notion de *module fonctionnel* : comme les molécules travaillent de concert pour élaborer des réponses aux stimuli, on a coutume de regrouper les gènes codant pour des protéines impliquées dans le même processus en modules. Une première protéine peut, par exemple, stimuler l'expression d'une deuxième, tandis que cette dernière inhibe la transcription de la première, formant ainsi une boucle de rétroaction négative. Le processus probablement le mieux caractérisé dans la cellule est le cycle de division cellulaire impliquant une dizaine de gènes (Novak et Pataki, 2000).

Nous nous sommes intéressés, pour notre part, à un autre exemple de cycle : celui des rythmes journaliers (dits *circadiens*). Ce projet, relié à l'Institut de Recherches Interdisciplinaires (IRI), a réuni des personnes de divers horizons (voir tableau 5.1). Ce groupe est en contact avec une équipe de chercheurs de l'Observatoire Océanographique de Banyuls sur Mer (OOB) qui étudient une algue verte appelée *Ostréococcus Tauri*. Dans le cadre d'un projet ANR « Biologie Systémique » commençant cette année[2], ces biologistes devraient fournir les données expérimentales nécessaires à l'élaboration d'hypothèses théoriques et les théoriciens, proposer de nouvelles expériences pour les valider. L'objectif ainsi poursuivi est de forcer un perpétuel aller et retour entre les deux disciplines.

De nombreux scientifiques ont déjà cherché à simuler des boucles de régulation géniques afin de générer des oscillations ; l'exercice consiste alors à trouver un jeu de paramètres permettant de reproduire les données expérimentales. Nous nous sommes intéressés à une autre thématique connexe qui est la recherche du module fonctionnel minimal — c'est-à-dire impliquant le plus petit nombre d'acteurs — permettant de créer des oscillations entretenues[3]. Cette recherche du modèle minimal traduit une volonté de comprendre en profondeur les mécanismes oscillants.

Par une approche formelle, nous avons pu mettre en évidence un mécanisme, utilisé depuis longtemps, mais dont l'impact est mal connu : les fonctions de dégradation non linéaires et, en particulier, les dégradations enzymatiques. Ainsi, nous montrerons à la section 5.3 qu'une dégradation linéaire des protéines ne permet pas

[1] bien qu'une équipe de chercheurs ait commencé à mettre en place une tentative d'intégration de toutes les connaissances actuelles dans un modèle global de la cellule (Takahashi *et al.*, 2002).

[2] incluant les équipes citées et celle d'Andrew Millar (Édimbourg).

[3] Le « record » est détenu par les systèmes à retards pour lesquels une unique équation suffit. En effet une protéine qui réprime sa propre expression grâce à un mécanisme modélisé par un delai temporel τ peut osciller : l'exemple de l'équation $dx/dt(t) = -x(t - \pi/2)$ avec la condition initiale fonctionnelle $x(t) = \sin(t), t \in [-\pi/2; 0]$ admet $x(t) = \sin(t)$ comme solution, cependant, ce type d'équations différentielles entre dans la catégorie des systèmes de dimension infinie (voir Richard, 2003, pour une revue sur les systèmes à retards).

Laboratoires et personnes impliquées dans le groupe de travail
Lille : partie modélisation
PhLAM : dynamiques non linéaires et chaos dans les systèmes physiques Marc Lefranc, Pierre-Emmanuel Morant, Constant Vandermoere, Quentin Thommen
LIFL : équipe de calcul formel François Boulier, François Lemaire, Asli Ürgüplü
LIFL : systèmes multi-agents Sébastien Picault
UGSF-LAGIS : automatique, analyse et commande des systèmes non linéaires Benjamin Parent
Banyuls/mer : partie expérimentale
OOB : études des couplages entre rythmes circadiens et rythmes de division cellulaire Florence Corellou, Christian Schwartz, Mickael Moulager, François-Yves Bouget

TAB. 5.1: personnes impliquées dans le groupe de travail sur les rythmes circadiens.

d'expliquer les oscillations du système différentiel ordinaire d'une protéine réprimant sa propre expression. La particularité de notre approche a été de traiter les équations sans donner de valeurs particulières aux paramètres. La connaissance de ces paramètres (typiquement, les constantes de réaction) est en effet un point sensible de la modélisation des réseaux de régulation géniques : souvent déterminées *in vitro*, parfois estimées *in vivo*, ces constantes dépendent généralement des conditions expérimentales et sont souvent sous le contrôle d'autres agents que les modèles ne peuvent pas prendre en compte.

5.2 Éléments de base pour la modélisation des réseaux géniques

> L'ordre du vivant ne réside pas dans la nature de ses composants élémentaires, mais dans leur organisation.
>
> François Jacob,
> La logique du vivant, une histoire de l'hérédité

5.2.1 Trois mécanismes de base

Nous ne ferons pas de présentation générale des mécanismes de la cellule (ce qui cadrerait plus avec les objectifs d'un livre de biologie[4]), bien qu'une grande partie de notre travail ait été de faire ces premiers pas dans la biologie tout en conservant un regard d'ingénieur automaticien.

Toutefois, afin de pouvoir élaborer un modèle des rythmes circadiens, nous devons donner un sens à la notion de *variation temporelle* des concentrations. Or, deux principaux phénomènes peuvent intervenir : la *production* de nouvelles protéines par les mécanismes de transcription et traduction et la *dégradation* des protéines par le protéasome. Nous tenterons, tant que possible, de quantifier les données afin de donner les ordres de grandeurs nécessaires lors de l'étape de modélisation.

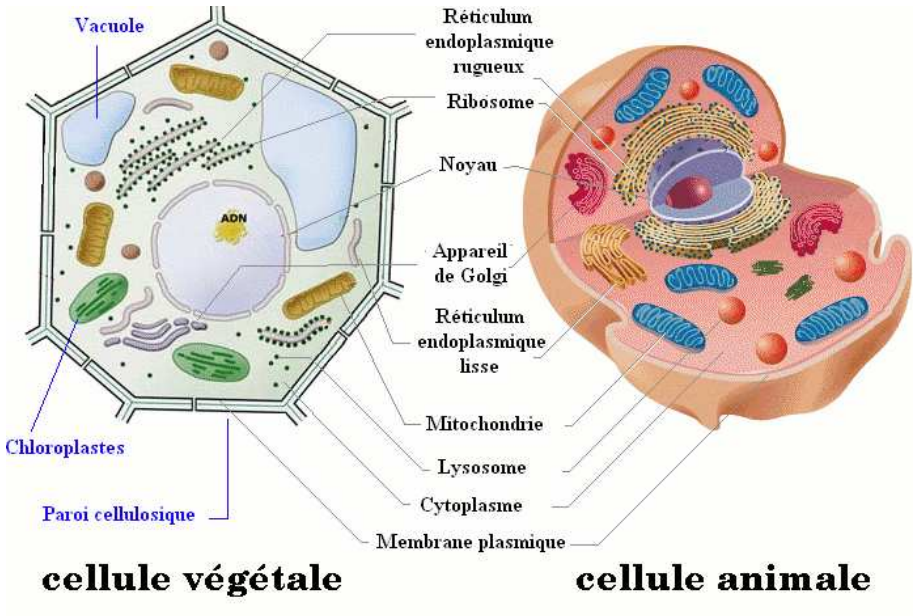


FIG. 5.1: structure et organites des cellules végétales et animales.

5.2.1.1 La transcription

Les signaux extérieurs (chimiques, lumineux, etc.) sont acheminés via des *voies de signalisation*, jusqu'aux chromosomes qui encodent l'information nécessaire à la production des protéines.

La première étape du décodage de cette information s'appelle la *transcription*, au cours de laquelle l'ADN est lu et *transcrit* en ARN messagers par des complexes

[4] voir par exemple « Molecular Biology of the Cell » (Alberts *et al.*, 2002).

moléculaires appelés *ARN-polymérases*. Ce processus est complexe et dépend du gène traité : de son initiation à la lecture des codons, jusqu'à son achèvement, il repose sur le recrutement d'agents moléculaires qui l'activent, le ralentissent ou l'inhibent complètement (voir figure 5.2). De plus, l'ADN est extrêmement compacté sur lui-même et offre un accès très limité, ce qui complique sa lecture (Li et Widom, 2004; Nagaich *et al.*, 2004).

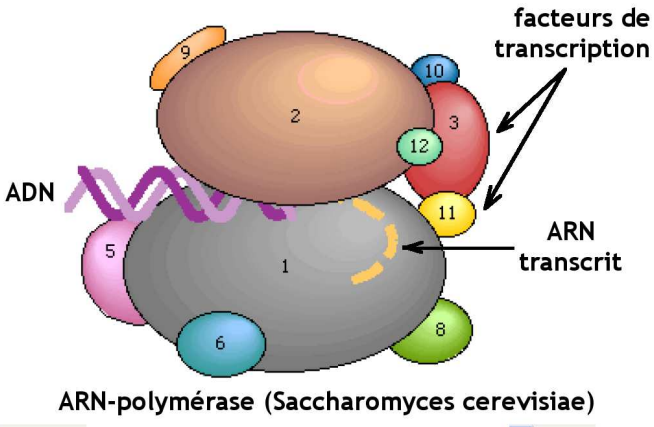


FIG. 5.2: ARN-polymérase et ses facteurs de transcription, en train de transcrire l'ADN en ARN. Figure extraite de Skhiri (2004).

De récentes études sur molécules uniques ont montré que la transcription n'était pas aussi linéaire dans le temps qu'on le croyait (Tolic-Norrelykke *et al.*, 2004) ; sont alors apparues les notions de « pauses » et de « salves » transcriptionnelles. Les échelles de temps pour décrire cette première étape ont donc été réévaluées : selon les gènes, l'ordre de grandeur pour les vitesses de transcription est environ de 5 à 50 paires de bases lues par seconde (chez *escherichia coli*), soit un temps caractéristique pouvant être de quelques minutes seulement. Paulsson (2005) s'est appliqué à mettre en évidence des causes théoriques possibles au phénomène de salves de transcription.

5.2.1.2 La traduction

Les ARN, synthétisés dans le noyau, traversent alors, quand elle existe, la membrane nuclaire et diffusent dans le cytoplasme. La deuxième étape — la *traduction* de l'ARN en protéine — peut alors avoir lieu grâce à de très gros complexes moléculaires, les *ribosomes*, qui lisent l'ARN, codon par codon et recrutent les acides aminés correspondants (voir figure 5.3).

Les ribosomes traduisent environ 1 à 3 résidus par seconde chez les eucaryotes[5]

[5]cellules à noyaux

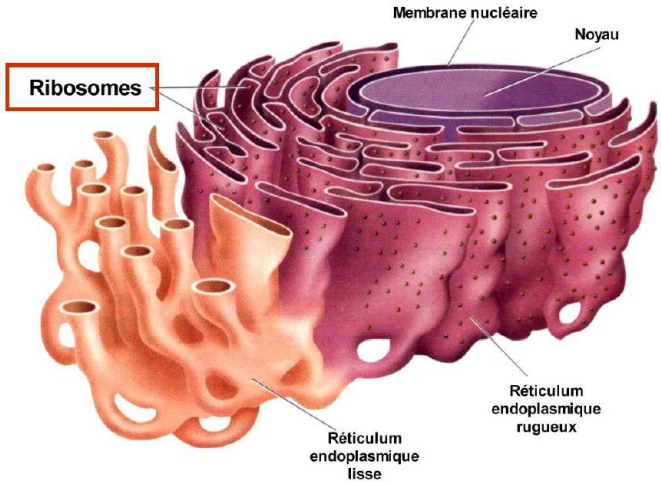


FIG. 5.3: la traduction des ARN messagers est assurée par les ribosomes au niveau du réticulum endoplasmique.

et jusqu'à 15 résidus par secondes chez les procaryotes[6], chez lesquels la traduction peut commencer dès la transcription puisqu'il n'y a pas de membrane séparatrice. Plusieurs ribosomes peuvent lire un même brin d'ARN simultanément, ce qui engendre des temps caractéristiques de l'ordre de quelques minutes également.

En réponse au stimulus, le système modifie donc le niveau de transcription/traduction de ses gènes et les molécules ainsi produites peuvent servir à rétablir l'équilibre (homéostasie), à contrer les agressions (anticorps par exemple), à propager de nouveaux signaux aux cellules voisines, à déclencher certaines phases de la vie d'une cellule (division, apoptose[7], etc.). Pour quantifier ces différences de niveaux, on parle de *taux d'expression relatifs* qui correspondent aux quantités de protéines produites par rapport à certaines quantités de référence (production moyenne, production au repos, etc.). Pour les taux d'expression absolus, il faut compter entre 50 et 10^6 protéines par cellule.

5.2.1.3 La dégradation

Enfin, les acteurs sont dégradés et recyclés. Ainsi, les ARN (moins stables que les protéines) sont généralement progressivement détruits par des *ARNases* ; la perte de leur fonction est alors retardée par l'existence d'une queue constituée de nombreuses bases d'adénines qui est attaquée avant que ne soient atteintes les bases codantes de l'ARN. Les temps de demi-vies caractéristiques de 4 661 ARN messagers chez la levure *Saccharomyces cerevisiae* ont été étudiés par Wang *et al.* (2002). Ils se situent

[6] cellules dépourvues de noyaux
[7] mort cellulaire programmée

entre 3 et 90 minutes avec une distribution (apparemment[8]) log-normale centrée sur 23 minutes.

La dégradation des protéines est beaucoup plus dépendante de leur état : sans marquage spécifique et correctement repliées, elles sont relativement stables. Leur dégradation se fait donc souvent de manière *active*, c'est-à-dire par des processus biologiques spécifiques (appelés protéasomes). La dégradation par des protéases peut, en particulier, nécessiter un marquage précis. Un autre mécanisme récurrent, est la dégradation par des enzymes très spécifiques, mais présentes en quantité restreinte ; ceci engendre des dynamiques de type Mickaëlis-Menten, où les protéases saturent rapidement.

Concernant la vitesse de dégradation des protéines : l'article de Belle *et al.* (2006) présente des études à l'échelle du génome (!) chez la levure *Saccharomyces cerevisiae* et exhibe une distribution bimodale des temps de demi-vie des protéines *in vivo* (donc avec le protéasome, voir figure 5.4). Le premier lobe suit une loi approximativement log-normale de moyenne 43 minutes, tandis que le deuxième correspond à 5% des protéines étudiées ayant un temps de demi-vie inférieur à 4 minutes[9] (voir l'article de Doherty et Beynon, 2006, pour une revue des dernières techniques permettant de mesurer les temps de vie dans la cellule, à l'échelle du protéome complet). Ces temps sont d'un ordre de grandeur compatible avec les temps caractéristiques des rythmes circadiens proches de 24 heures, c'est pourquoi nous pensons que leur influence mérite d'être étudiée.

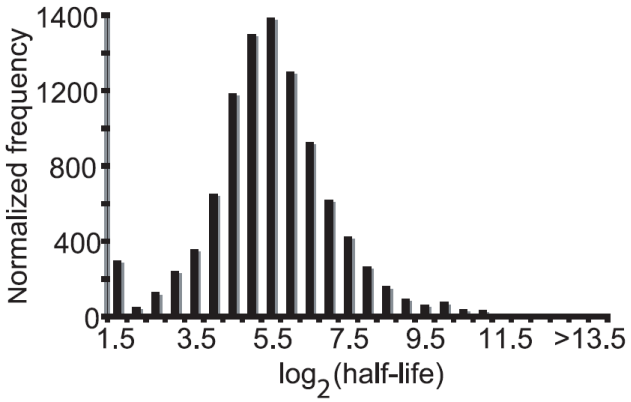


FIG. 5.4: distribution des temps de demi-vie *in vivo* des protéines chez Saccharomyces cerevisiae ; figure extraite de Belle *et al.*, 2006.

Remarque : un article, plus ancien (Pratt *et al.*, 2002), proposait des valeurs

[8] les auteurs ne le mentionnent pas, mais les données présentées l'évoquent

[9] un test de reproductibilité sur cette étude a montré que les données étaient fiables à un facteur multiplicatif 2 près.

plus importantes (en moyenne une trentaine d'heures, mais s'étalant de 6h jusqu'au delà de la limite mesurable), mais n'utilisait qu'une sélection d'une cinquantaine de protéines, ce qui explique probablement les différences.

Ayant rappelé certains principes généraux concernant la production et la dégradation des protéines et fourni quelques données permettant de se figurer les ordres de grandeurs, nous présentons maintenant brièvement la problématique à laquelle nous nous sommes attachés et un aperçu du paysage scientifique dans ce domaine.

5.2.2 Les rythmes circadiens

La cellule présente plusieurs rythmes : circadiens, division cellulaire, suivi du rythme des saisons, etc. Nous avons choisi d'analyser les rythmes circadiens, qui se caractérisent par :

1. des oscillations entretenues, même en conditions d'éclairement constantes (voir figure 5.5) avec une période propre proche de 24 heures,
2. une « compensation en température », c'est-à-dire une robustesse de la période vis-à-vis des variations de température,
3. la possibilité de réinitialiser le système par des impulsions lumineuses.

Ils présentent l'avantage d'être auto-entretenus : en particulier, il n'est pas nécessaire (dans un premier temps) de modéliser les entrées/sorties du module fonctionnel correspondant. Pour cela, nous nous sommes rapprochés de l'équipe de F.-Y. Bouget, qui les étudient chez l'algue verte *Ostréococcus Tauri*.

5.3 Étude complète de la répression autogène

Nous proposons ici l'étude précise des dynamiques de dégradation des protéines et leurs implications sur le comportement d'un modèle à un gène auto-régulé (dit *autogène*). Nous n'avons considéré initialement que trois variables, correspondant à un gène, son ARN associé et sa protéine, et nous nous sommes placés dans une description :

- continue : malgré le caractère discret des molécules, leur grand nombre permet d'utiliser la notion de *concentration* continue,

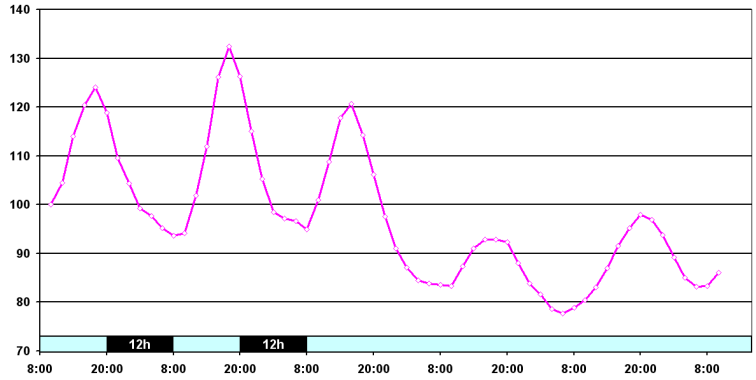


FIG. 5.5: taux d'expression d'une protéine suivi par fluorescence, en fonction du temps, les cadres bleus et noirs figurent respectivement les périodes d'éclairement et de pénombre. Les oscillations perdurent en conditions d'éclairement constant.

- uniforme : nous ne considérons ni les variables d'espace (gradients de concentrations, etc.), ni les compartiments,
- déterministe : pas de simulation stochastique,
- et sans retard.

5.3.1 Conception d'un modèle

Nous avons imaginé un modèle de boucle de rétroaction négative, la plus simple possible : un gène est transcrit en ARN, qui est lui même traduit en protéine. Cette protéine inhibe la transcription du gène et crée ainsi une boucle de rétroaction négative. Le modèle est résumé sur la figure 5.6.

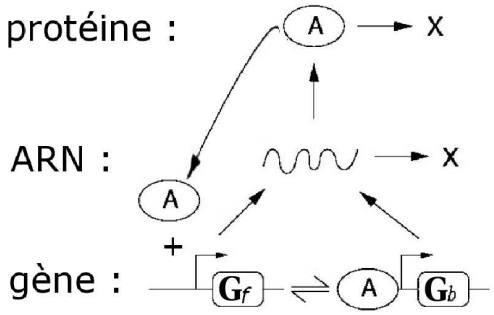


FIG. 5.6: modèle à un seul gène de boucle de rétroaction négative : la transcription est réprimée par la présence de protéines (symbolisée par l'ellipse A, liée au gène : G_b).

À l'exemple de François et Hakim (2004), nous n'avons pas utilisé une variable booléenne pour décrire l'état du gène (libre : f, ou liée : b) : nous avons considéré l'*activité* du gène comme une variable continue G_f, comprise entre 0 et une valeur

maximum G_T. Elle peut s'interpréter comme la *proportion* sur la population de cellules, de gènes sans répresseur[10]. À concentration de protéines fixée, P, cette activité converge vers une valeur dépendant de P.

La dégradation de l'ARN messager (noté M) est supposée linéaire car ces molécules sont peu stables et dégradées progressivement en commençant par leur queue poly-adénine. En revanche, aucune hypothèse n'est faite sur la dégradation des protéines, sinon que les agents qui les dégradent ne sont pas sous le contrôle circadien. Cette fonction de dégradation pourra ainsi être étudiée ultérieurement.

5.3.1.1 Les réactions

Le modèle de la figure 5.6 peut alors se réécrire sous la forme :

$$\begin{aligned}
G_f + P &\overset{k_b}{\underset{k_f}{\rightleftharpoons}} G_b, && (5.1)\\
G_f &\xrightarrow{\ell_f} G_f + M, && (5.2)\\
G_b &\xrightarrow{\ell_b} G_b + M, && (5.3)\\
M &\xrightarrow{\ell_M} M + P, && (5.4)\\
M &\xrightarrow{d_M} \emptyset, && (5.5)\\
P &\xrightarrow{\Phi} \emptyset. && (5.6)
\end{aligned}$$

La réaction (5.1) traduit la liaison de la protéine à l'ADN ; les réactions (5.2) et (5.3) concernent la transcription du gène en ARN messagers avec différents taux de transcription selon que la protéine est présente ou non. La réaction (5.4) correspond à l'étape de traduction, enfin, (5.5) et (5.6) indiquent la dégradation des acteurs (le symbole d'ensemble vide : $\emptyset$, indique la perte de la fonction de l'acteur).

Les lois de conservations de masse ne s'appliquent pas en général, puisque nous ne considérons pas le recrutement des bases et des acides aminés (pour former, respectivement, l'ARN messager et les protéines) ni leur recyclage lors de la dégradation. Cependant elles peuvent être utilisées dans la réaction (5.1) avec un sens légèrement modifié : la proportion de gènes libres et liés donne toujours le nombre de gènes total codant pour la protéine P :

$$G_f + G_b = G_T. \tag{5.7}$$

[10] G_T, le nombre de gènes par cellule, codant pour la même protéine, peut être supérieur à 1 quand il y a redondance.

5.3.1.2 Conditions requises

Toutes les constantes cinétiques (équations (5.1) à (5.6)) sont positives ou nulles. Nous imposons quelques conditions supplémentaires sur les variables et les constantes :

- régularité : les variables décrivant l'activité du gène et les concentrations sont toutes de classe C^1 par rapport au temps, c'est-à-dire continues et continûment dérivables (ce qui est nécessaire pour interpréter G_f, M et P comme solutions d'équations différentielles d'ordre 1) ;
- le modèle proposé fonctionne sur le principe d'une transcription différentielle lorsque la protéine intervient ou non ; comme il s'agit d'une répression, nous avons :

$$\ell_f > \ell_b \geq 0 \; ; \tag{5.8}$$

- de même, nous interdisons une traduction totalement inefficace : $\ell_M > 0$;
- remarquons enfin que $G_T > 0$, puisqu'au moins un gène code pour la protéine étudiée.

La fonction Φ, appelée *fonction de dégradation*, dépend de la concentration P. Elle est quelconque, cependant, nous émettons les hypothèses suivantes :

- continuité : si deux concentrations sont proches, les niveaux de dégradation sont nécessairement proches, Φ est donc continue ; en réalité, nous allons même supposer Φ lipschitzienne[11] ;
- monotonie : si la concentration augmente, le niveau de dégradation augmente également ; Φ est donc supposée monotone croissante ;
- positivité : s'agissant d'une dégradation, Φ doit être positive ; nous supposerons même Φ *strictement* positive sur $\mathbb{R}_+^*$ et nulle en 0.

Un des apports de notre travail sera de montrer l'intérêt d'employer une fonction de dégradation non linéaire plutôt que linéaire. Nous parlerons plus brièvement de « dégradation linéaire » et « dégradation non linéaire » plutôt que de « fonction de dégradation ».

Finalement, nous écartons les cas limites suivants :

- pas de dégradation de l'ARN : $d_M = 0$, car dans ce cas, M est monotone croissante (pas d'oscillations) ;
- pas d'équilibre entre G_f et G_b : $k_f = 0$ ou $k_b = 0$, car dans ce cas, c'est G_f qui est monotone et ne se stabilise qu'en 0 ou G_T.

[11] une fonction quelconque $F : \mathbb{R}^n \to \mathbb{R}^q$ est dite k-lipschitzienne sur un domaine D si, pour tout couple (x, y) de D, on a $\|F(y) - F(x)\| \leq k\|y - x\|$

5.3.1.3 Équations du système

Le modèle ci-dessus peut se traduire dans le système différentiel non linéaire suivant, où toutes les variables : $G_f = G_f(t)$, $M = M(t)$ et $P = P(t)$ sont exprimées à l'instant t (système ordinaire) :

$$\frac{dG_f}{dt} = k_f(G_T - G_f) - k_b G_f P, \tag{5.9}$$

$$\frac{dM}{dt} = \ell_f G_f + \ell_b(G_T - G_f) - d_M M, \tag{5.10}$$

$$\frac{dP}{dt} = k_f(G_T - G_f) - k_b G_f P + \ell_M M - \Phi(P). \tag{5.11}$$

Comme ce modèle est stationnaire, nous considèrerons, sans perte de généralités, que l'instant initial est $t_0 = 0$. Nous supposons alors les conditions initiales suivantes :

$$(G_f(0), M(0), P(0)) = (G_T, 0, 0). \tag{5.12}$$

Le système apparaît alors sous une forme « $\dot{x} = F(x)$ » où F est une fonction lipschitzienne, ce qui assure l'unicité de la trajectoire (problème de Cauchy).

5.3.2 Analyse du système

Étudions maintenant ce système.

5.3.2.1 Domaine invariant

Ce système fait partie des systèmes dits positifs (Mailleret, 2004) car les variables évoluent dans l'orthant positif : ${\mathbb{R}_+}^3$. Pour s'en convaincre, il suffit de vérifier que les frontières sont *infranchissables* : les dérivées temporelles des variables sont positives dès que la variable est nulle. Ainsi, $\forall t \in \mathbb{R}_+, (G_f(t), M(t), P(t)) \in {\mathbb{R}_+}^3$.

On peut même être plus précis et montrer, de la même manière, que $G_f(t) \leq G_T$ et $d_M M \leq \ell_f G_T$, puisqu'en $G_f = G_T$, la dérivée $\frac{dG_f}{dt} \leq 0$, et en $d_M M = \ell_f G_T$, on a bien $\frac{dM}{dt} \leq 0$ car $\ell_f > \ell_b$.

5.3.2.2 Étude des points d'équilibre

F étant lipschitzienne, les points d'équilibre (notés (G_0, M_0, P_0)), s'ils existent, doivent vérifier :

$$\left(\frac{d}{dt}(G_f, M, P)\right)_{\substack{G_f=G_0 \\ M=M_0 \\ P=P_0}} = (0, 0, 0), \tag{5.13}$$

soit :

$$G_0 = \frac{k_f G_T}{k_f + k_b P_0}, \tag{5.14}$$

$$M_0 = \frac{\ell_b G_T + (\ell_f - \ell_b) G_0}{d_M}, \tag{5.15}$$

$$\Phi(P_0) = \ell_M M_0, \tag{5.16}$$

où la non-nullité des dénominateurs est assurée par les précautions que nous avons prises au § 5.3.1.2.

D'après (5.14), G_0 est une fonction strictement décroissante de P_0 ; de même, comme $\ell_f - \ell_b > 0$ et, d'après (5.15), M_0 est strictement croissante en G_0 et donc strictement décroissante en P_0. Finalement, dans (5.16), P_0 apparaît comme le point d'intersection entre deux fonctions continues monotones contraires : Φ et $\ell_M M_0$ (dont une au moins est strictement monotone). S'il existe, le point d'équilibre est donc unique.

Pour qu'il existe, il faut s'assurer que les deux courbes se coupent. Or, les valeurs limites en $P_0 = 0$ et $P_0 = \infty$ sont regroupées dans le tableau 5.3.2.2.

P_0	0	∞
$G_0(P_0)$	G_T	0
$M_0(P_0)$	$\frac{\ell_f G_T}{d_M}$	$\frac{\ell_b G_T}{d_M}$
$\Phi(P_0)$	0	Φ_∞

En $P_0 = 0$, Φ est en dessous de $\ell_M M_0$, pour être sûrs de l'existence du point d'équilibre, nous imposons donc :

$$d_M \Phi_\infty > \ell_M \ell_b G_T. \tag{5.17}$$

Remarque : pour une dégradation linéaire, cette condition est toujours vérifiée.

Si cette condition n'est pas vérifiée, l'équilibre est *rejeté à l'infini* : $(G_0, M_0, P_0) = (0, \frac{\ell_b G_T}{d_M}, \infty)$, ce qui n'a pas de sens physique puisqu'en réalité, la production de protéines saturera. Néanmoins, dans ce cadre théorique, P est monotone croissant (au moins après un certain temps) et aucune oscillation entretenue ne peut avoir lieu.

La condition (5.17) peut s'interpréter physiquement : elle impose des dégradations suffisantes à très forte concentration de protéines. Or, pour P grand, le gène

est majoritairement dans l'état lié : $G_f \approx 0$, la production de nouvelles protéines est donc représentée par le terme $\ell_M \ell_b G_T$ tandis que le produit des dégradations est $d_M \Phi_\infty$.

5.3.2.3 Adimensionnement

Nous allons maintenant étudier le comportement du modèle au voisinage de son point d'équilibre, en fonction des paramètres. Or nous avons 7 paramètres : $(k_f, k_b, G_T, \ell_f, \ell_b, \ell_M, d_M)$ et une fonction inconnue : Φ. Pour simplifier notre étude, nous allons adimensionner le système par quelques changements de variables :

- plutôt que le paramètre de temps t, utilisons comme unité de temps, le temps de demi-vie de l'ARN : $\tau = d_M t$;
- considérons les nouvelles variables (g, m, p) définies par :

$$\left\{ \begin{array}{rcl} g & = & \frac{G_f}{G_T}, \\ m & = & M, \\ p & = & \frac{k_b}{k_f} P, \end{array} \right. \tag{5.18}$$

- et posons les constantes suivantes :

$$\left\{ \begin{array}{rcl} \theta & = & \frac{k_f}{d_M}, \\ \alpha & = & \frac{k_b G_T}{d_M}, \\ \delta & = & \frac{k_b \ell_M}{k_f d_M}, \\ \mu & = & \frac{\ell_b}{d_M} G_T, \\ \lambda & = & \frac{\ell_f - \ell_b}{d_M} G_T, \end{array} \right. \tag{5.19}$$

- enfin, posons la fonction f, telle que pour tout $u \geq 0$,

$$f(u) = \frac{1}{\ell_M} \Phi \left(\frac{k_f}{k_b} u \right). \tag{5.20}$$

Dans ces nouvelles coordonnées, seuls cinq paramètres subsistent et le système s'écrit :

$$\frac{dg}{d\tau} = \theta(1 - g - gp), \tag{5.21}$$

$$\frac{dm}{d\tau} = \mu + \lambda g - m, \tag{5.22}$$

$$\frac{dp}{d\tau} = \alpha(1 - g - gp) + \delta \left[m - f(p) \right]. \tag{5.23}$$

Les différentes conditions sur les paramètres deviennent :

– $(\theta, \alpha, \delta, \lambda)$ tous strictement positifs et μ positif ou nul ;

– f est de classe C^1, croissante, nulle en 0 et strictement positive ailleurs ;

– $\lim_{+\infty} f > \mu$.

Enfin, le point d'équilibre (g_0, m_0, p_0) vérifie

$$g_0 = \frac{1}{1+p_0}, \tag{5.24}$$

$$m_0 = \mu + \lambda g_0, \tag{5.25}$$

$$f(p_0) = \mu + \frac{\lambda}{1+p_0}. \tag{5.26}$$

p_0 est alors le point d'intersection entre deux courbes monotones contraires : voir figure 5.7

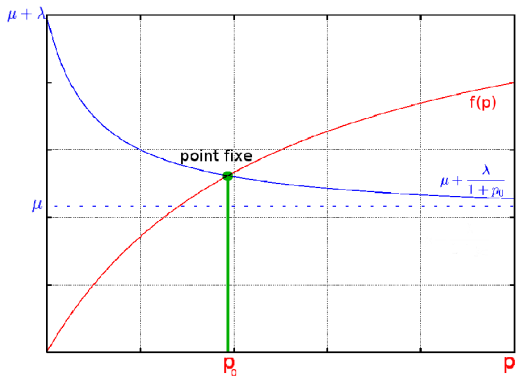


FIG. 5.7: s'il existe, le point d'équilibre est unique.

5.3.2.4 Étude locale autour du point d'équilibre

> If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is.
>
> von Neumann

Analysons la stabilité du système linéarisé autour de (g_0, m_0, p_0). Pour cela, nous allons poser $s_0 = \frac{df}{dp}(p_0)$. Comme f est croissante, s_0 est positif.

C'est ici que l'on met en évidence les non-linéarités de la fonction de dégradation : pour une fonction non linéaire, s_0 sera, en général, différent de $\frac{f(p_0)}{p_0}$.

Le linéarisé est de la forme :

$$\frac{d}{d\tau}\begin{pmatrix} g-g_0 \\ m-m_0 \\ p-p_0 \end{pmatrix} = J\begin{pmatrix} g-g_0 \\ m-m_0 \\ p-p_0 \end{pmatrix}, \tag{5.27}$$

$$\text{avec la jacobienne} \quad J = \begin{pmatrix} -\theta(1+p_0) & 0 & -\theta g_0 \\ \lambda & -1 & 0 \\ -\alpha(1+p_0) & \delta & -\alpha g_0 - \delta s_0 \end{pmatrix}. \tag{5.28}$$

Ce système possède trois pôles (valeurs propres de J que nous noterons $\sigma_{1,2,3}$, éventuellement complexes, éventuellement confondues) dont la position dans le plan complexe détermine le comportement local (voir figure 5.8).

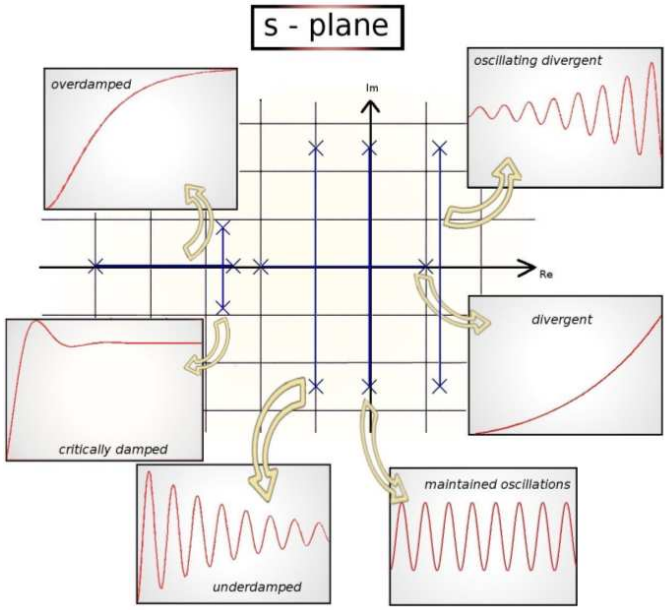


FIG. 5.8: comportement des systèmes linéaires en fonction de la position des pôles dans le plan complexe (figure extraite du poster présenté au « Gent-Lille workshop on computational biology », voir annexe I).

Le polynôme caractéristique de la jacobienne est donc de la forme :

$$Q_J(X) = X^3 - \left(\sum \sigma_i\right) X^2 + \left(\sum \sigma_i\sigma_j\right) X - \sigma_1\sigma_2\sigma_3, \tag{5.29}$$

$$= X^3 + aX^2 + bX + c. \tag{5.30}$$

Ce polynôme a au moins une racine réelle (que nous attribuerons à σ_1). Nous avons $\sigma_1\sigma_2\sigma_3 = \det(J) = -\delta\theta(\lambda g_0 + s_0 + s_0 p_0) < 0$ autrement dit, il y a un nombre impair de racines sur l'axe réel négatif :

- si il y en a trois, alors les trois pôles sont stables et le système linéarisé converge exponentiellement vers (g_0, m_0, p_0) sans oscillations. Dans ce cas, le système initial est localement asymptotiquement stable ;
- si il n'y en a qu'une et si les deux autres sont réelles positives, alors le système et son linéarisé sont tous deux instables mais ne présentent pas non plus d'oscillations entretenues ;
- enfin, si seule σ_1 est réelle (nécessairement négative), alors σ_2 et σ_3 sont complexes conjuguées et le système présentera des oscillations :

$$\mathrm{Re}(\sigma_2) = \mathrm{Re}(\sigma_3) \triangleq -\zeta \quad \text{donne l'amortissement des oscillations ;} \tag{5.31}$$

$$\mathrm{Im}(\sigma_2) = -\mathrm{Im}(\sigma_3) \triangleq \omega \quad \text{donne leur fréquence.} \tag{5.32}$$

Si ζ est positif, les pôles σ_2 et σ_3 sont stables et entraînent des oscillations amorties. Si ζ devient négatif (σ_2 et σ_3 franchissent l'axe imaginaire pur), le système linéarisé devient divergent et le système non linéaire présentera un point d'équilibre instable entouré par un cycle limite, c'est-à-dire des oscillations entretenues. À la frontière entre les deux domaines ($\zeta = 0$), il y a une bifurcation dite de Hopf (Richard, 2002) que nous allons caractériser en fonction des paramètres du système.

Physiquement, il est difficile de décider si le système biologique possède des oscillations entretenues ou des oscillations faiblement amorties : il est en effet difficile d'observer des cultures de cellules pendant très longtemps (les cultures sont en croissance exponentielle, à la lumière constante et leur désynchronisation atténue le signal). Toutefois, la figure suivante semble montrer que l'horloge redémarre après mise en condition d'éclairement constant (figure 5.9). De plus, nous pouvons chercher à rendre ζ le plus petit possible, indépendamment du fait qu'il soit négatif ou positif.

Pour étudier la position des pôles en fonction des paramètres, explicitons le polynôme caractéristique :

$$\begin{aligned} Q_J(X) &= X^3 + (1 + \alpha g_0 + \delta s_0 + \theta + \theta p_0)X^2 \\ &\quad + (\alpha g_0 + \delta s_0 + \theta + \theta p_0 + \theta\delta s_0 + \theta\delta s_0 p_0)X \\ &\quad + \delta\theta(\lambda g_0 + s_0 + s_0 p_0). \end{aligned} \tag{5.33}$$

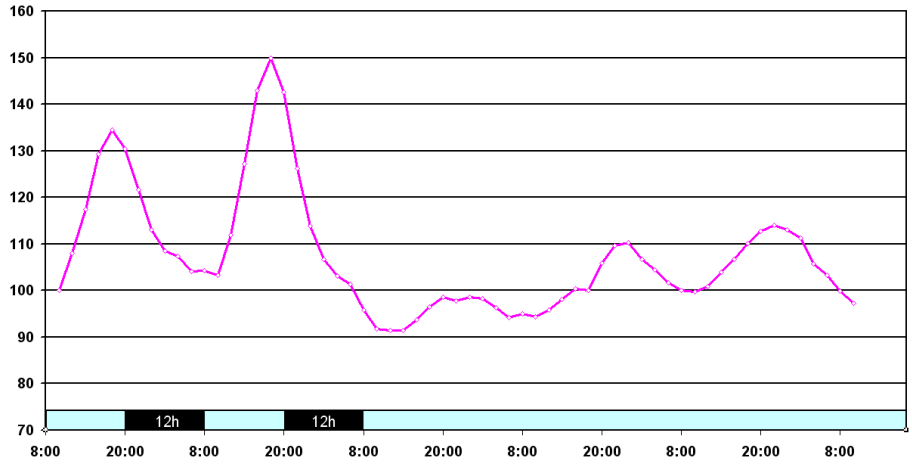


FIG. 5.9: niveau de fluorescence en fonction du temps, pendant entraînement en cycles jour-nuit, puis en condition d'éclairement constant. Le rythme, après s'être presque arrêté, semble retourner vers un cycle limite.

ζ pourrait s'exprimer comme une fonction implicite de a, b et c, puisque, d'après l'identité entre (5.29) et (5.30),

$$\begin{cases} a &= 2\zeta - \sigma_1 \\ b &= \omega^2 - \sigma_1^2 + (\sigma_1 - \zeta)^2 \\ c &= -\sigma_1(\zeta^2 + \omega^2) \end{cases}$$

ce qui fait apparaître ζ comme solution d'un polynôme de degré 3. Toutefois, le tableau de Routh (Borne *et al.*, 1990) nous donne une expression plus simple à rendre négative (tableau 5.2)

X^3	1	b
X^2	a	c
X	$\frac{ab-c}{a}$	0
1	c	0

TAB. 5.2: tableau de Routh pour un polynôme de degré trois.

Or, puisque $a = -\text{trace}(J) > 0$ et $c = -\det(J) > 0$, il ne reste donc plus que la condition sur $ab - c$ pour rendre le système localement instable autour du point d'équilibre. La bifurcation de Hopf a lieu sur la variété $ab = c$: avant $(ab - c > 0)$, le point fixe est localement asymptotiquement stable ; après $(ab - c < 0)$, le point fixe est instable, entouré d'un cycle limite.

Remarque : le critère $\mathfrak{R}_0 = ab - c$ fait apparaître ζ en facteur :

$$\begin{aligned} ab - c &= \sigma_1\sigma_2\sigma_3 - (\sigma_1 + \sigma_2 + \sigma_3)(\sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_1\sigma_3) \\ &= 2\zeta\left[\omega^2 + (\sigma_1 - \zeta)^2\right]. \end{aligned} \tag{5.34}$$

5.3.3 Étude du critère de Routh

En utilisant les expressions explicites de a, b et c dans l'équation (5.33), nous pouvons réécrire $\mathfrak{R}_0$ à l'aide des paramètres du modèle (en utilisant, au besoin, les équations du point d'équilibre). Comme cette expression reste très complexe et inexploitable, nous réduisons le nombre de paramètres en faisant les hypothèses suivantes :

- $\mu = 0$: inhibition totale de la transcription du gène par sa protéine ;
- la troisième équation du modèle (5.23), donnant l'évolution de p est dominée par les termes de traduction et de dégradation, tandis que la partie correspondant aux protéines qui se fixent sur l'ADN est négligeable. Ceci est obtenu de façon indirecte en prenant $\alpha \to 0$.

Alors, le critère devient :

$$\mathfrak{R}_0 = (\delta s_0 + 1)\big(1 + \theta + \theta p_0)(\delta s_0 + \theta + \theta p_0) - \lambda\delta\theta g_0 < 0. \tag{5.35}$$

5.3.3.1 Première conclusion

Cette expression suffit à démontrer ce que nous avons avancé dans l'introduction : si l'on considère une dégradation linéaire de la forme $f(p) = \pi p$, alors $s_0 = \pi$. En nous rappelant, d'après les équations du point d'équilibre, que $\lambda g_0 = f(p_0) = \pi p_0$, le seul terme négatif de $\mathfrak{R}_0$ s'annule et, bien qu'éventuellement oscillant, le système est nécessairement localement amorti. Dans nos conditions de modélisation, une approche linéaire ne permet donc pas de reproduire les phénomènes d'oscillations autoentretenues pourtant observés en pratique.

5.3.3.2 Interprétation

Nous venons donc de prouver que le système linéarisé ne pouvait pas osciller si la fonction de dégradation des protéines était linéaire. Nous allons continuer l'interprétation dans le cas général.

Tout d'abord, nous avons, avec l'équation (5.35), un critère de non-amortissement des oscillations du linéarisé de la forme $\lambda\delta > \mathcal{V}_{\min}$ où λ et δ sont les taux de trans-

cription et traduction et $\mathcal{V}_{\min}$ une valeur dépendante des autres paramètres[12]. Cela semble indiquer qu'un couplage minimal, du gène jusqu'à la protéine, est nécessaire pour entretenir les oscillations.

Par ailleurs, si, au lieu de s_0, nous utilisons la variable u, définie par :

$$u = \frac{\delta s_0}{\theta + g_0} g_0 \geq 0, \tag{5.36}$$

alors, le critère de Routh (5.35) devient :

$$\mathfrak{R}_1 = u^2 + u + \underbrace{\frac{\theta g_0}{(\theta + g_0)^2} - \lambda\delta \frac{\theta g_0^4}{(\theta + g_0)^3}}_{\Psi} < 0. \tag{5.37}$$

Ce polynôme, de coefficient dominant positif, doit être négatif en u et, par suite, doit donc avoir deux racines réelles. L'une d'elles est nécessairement négative puisque la somme des racines (opposé du coefficient du deuxième monôme) est negative. Pour que ce polynôme soit négatif en $u \geq 0$, il est donc *nécessaire* qu'il soit négatif en $u = 0$.

Il y a donc la condition nécessaire suivante :

$$\Psi \leq 0, \tag{5.38}$$

$$\text{c'est-à-dire } \lambda\delta \geq \frac{\theta + g_0}{g_0^3}. \tag{5.39}$$

Une fois cette condition vérifiée, nous savons que $\mathfrak{R}_1$ possède une racine à $u \geq 0$ (notée u_+) et nous imposons :

$$u \leq u_+ = -\frac{1}{2} + \sqrt{\frac{1}{4} + \lambda\delta \frac{\theta g_0^4}{(\theta + g_0)^3} - \frac{\theta g_0}{\theta + g_0}}, \tag{5.40}$$

que l'on peut facilement ramener à s, en utilisant (5.36).

Encore une fois, l'équation (5.39) réexprime que le couplage du gène à la protéine en passant par l'ARN doit être suffisamment important, tandis que l'inégalité (5.40) indique qu'un profil de dégradation saturé (s_0 inférieur à une valeur) facilite les oscillations.

[12] les autres apparitions du paramètre δ sont en facteur de s_0 c'est-à-dire uniquement parce que l'on a factorisé l'équation d'évolution en $\delta(m - f(p))$ afin de simplifier l'expression du point d'équilibre.

5.3.4 Cas particulier : les dégradations enzymatiques

Les résultats précédents semblent indiquer qu'une fonction de dégradation saturée, c'est-à-dire avec $\frac{df}{dp}(p_0) < \frac{f(p_0)}{p_0}$, est favorable aux oscillations. Nous avons donc étudié les dégradations enzymatiques de type Mickaëlis-Menten que nous rappelons ici.

5.3.4.1 Équation de Michaëlis-Menten

Ces fonctions sont issues d'une dégradation de type :

$$P + E \leftrightharpoons C \to E, \tag{5.41}$$

où C est un complexe intermédiaire entre la protéine P et l'enzyme E.

La première réaction est supposée très rapide. Les phénomènes transitoires sont donc omis et nous étudions la dynamique sur la variété où C est constant. Comme $C + E = E_T$ la quantité totale d'enzymes disponibles, cela revient à supposer $\frac{dE}{dt} \approx 0$. Nous avons alors :

$$\frac{dP}{dt} = -k_1 PE, \tag{5.42}$$

$$\frac{dE}{dt} = -k_1 PE + (k_2 + k_{-1})(E_T - E) \approx 0, \tag{5.43}$$

$$\text{soit} \quad E_0 = \frac{E_T}{1 + \frac{k_1}{k_{-1}+k_2} P} \tag{5.44}$$

$$\text{et} \quad \frac{dP}{dt} = -\frac{k_1 E_T P}{1 + \frac{k_1}{k_{-1}+k_2} P}. \tag{5.45}$$

La dégradation est donc de la forme :

$$f(p) = \frac{\chi p}{\kappa + p}, \qquad (\chi, \kappa) \in {\mathbb{R}_+}^2, \tag{5.46}$$

qui est bien une fonction non linéaire. En particulier, la pente s_0 de cette fonction en p_0 peut s'exprimer en fonction du rapport $\frac{f(p_0)}{p_0}$:

$$s_0 = \frac{\chi\kappa}{(\kappa + p_0)^2} = \frac{f(p_0)}{p_0} - \frac{f(p_0)}{\kappa + p_0}, \tag{5.47}$$

en particulier, on a bien $s_0 < \frac{f(p_0)}{p_0}$.

5.3.4.2 Analyse des résultats

Dans le cas particulier où f est de la forme $\frac{\chi p}{\kappa+p}$, il est possible d'exprimer le critère de Routh (équation (5.35)) plutôt comme un polynôme de θ de degré 2 de la forme :

$$\mathfrak{R}_2 = \theta^2 + \gamma\theta + 1 < 0. \tag{5.48}$$

Il faut donc imposer les trois conditions suivantes :

- $\gamma < 0$,
- le discriminant $\Delta = \gamma^2 - 4 > 0$ d'où $\gamma > -2$,
- $\theta_1 < \theta < \theta_2$

où (θ_1, θ_2) sont les racines du polynôme $\mathfrak{R}_2$.

Les domaines de paramètres se représentent mieux en posant π et ξ tels que

$$\pi = \frac{p_0}{\kappa}, \tag{5.49}$$

$$\xi = \frac{\chi}{(1+\pi)^2}, \tag{5.50}$$

ces conditions reviennent à prendre l'intersection (dans l'espace des trois paramètres (ξ, κ, π)) du domaine jaune ($\gamma < 0$, pour avoir deux racines à parties réelles positives : $(\theta_1, \theta_2) \in \mathcal{D}^+$), avec le complémentaire du domaine vert ($\gamma > -2$, pour avoir deux racines réelles), voir figure suivante (5.10).

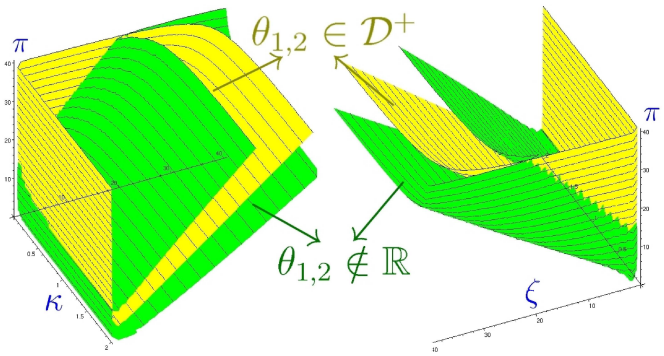


FIG. 5.10: espace des paramètres permettant d'obtenir des oscillations avec une dégradation de type enzymatique (figure réalisée avec Maple et, en particulier, la fonction implicitplot3D).

Nous avons simulé le système dans chacun des quatre domaines en prenant pour θ la valeur optimale (sommet de la parabole). La figure 5.11 présente les profils temporels obtenus, qui correspondent bien aux comportements prédits.

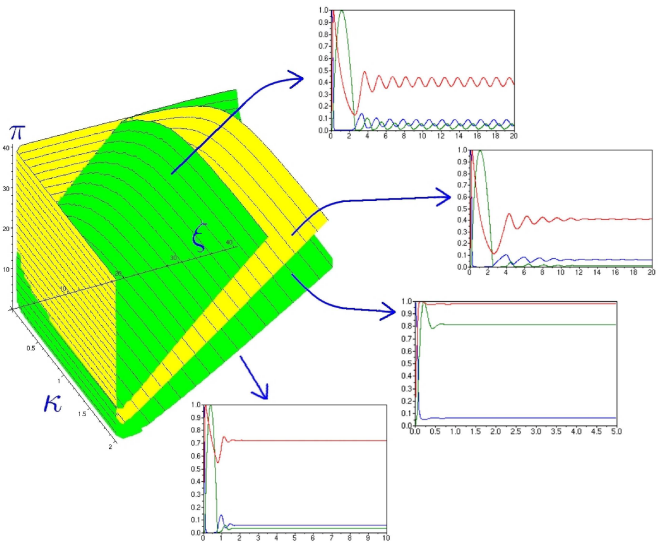


FIG. 5.11: simulations dans les différents domaines.

5.3.4.3 Conclusion

Nous avons réussi à obtenir des conditions sur les paramètres mettant en évidence les domaines dans lesquels les oscillations entretenues pouvaient avoir lieu. En particulier, nous avons montré que le profil de la fonction de dégradation influençait la possibilité de créer des oscillations : le domaine pour une dégradation linéaire est vide, tandis qu'en saturant la dégradation (dérivée inférieure à la pente moyenne), ces oscillations peuvent apparaître.

Ces résultats ont, en partie, été diffusés par l'intermédiaire de posters :

- l'un en juin 2006 (Lefranc *et al.*, 2006), au « gent-lille workshop on computational biology » (annexe I) ;
- l'autre en mars 2007 (Morant *et al.*, 2007), aux rencontres du non-linéaire de Paris, ayant donné lieu à une publication (annexe J).

5.3.5 Remarques sur nos choix pour la modélisation

Dans notre démarche, nous avons opté pour un modèle continu, homogène, déterministe et sans retards. Il faut cependant garder en mémoire un certain nombre de points importants

5.3.5.1 Les régulations

Même si un modèle à l'échelle 1 :1 n'existe pas, l'extrême complexité des mécanismes permet l'existence de nombreux *points de contrôle* et de *régulations* :

- accès limité à l'ADN par les ARN-polymérases,
- initiation, processus et terminaison de la transcription,
- processus éventuel d'épissage,
- traduction,
- transport parfois actif des acteurs et franchissement des membranes,
- contrôle des constantes de réaction,
- marquage pour la dégradation et processus de dégradation lui-même, etc.

De plus, les protéines ainsi produites sont encore sujettes à de nombreuses modifications qui permettent d'altérer leur fonctionnement. Ces modifications post-traductionnelles peuvent être de différentes natures :

1. *modifications covalentes*, lors de l'adjonction de certains groupements en des sites très spécifiques (méthylation, phosphorylation, etc.) ;

2. *polymérisation* avec d'autres partenaires (les ribosomes, les facteurs de transcription sont des exemples de complexes fonctionnels) ;

3. *modifications conformationnelles* lorsque la protéine ne se replie pas spontanément dans sa conformation native mais subit l'aide de protéines chaperones ou suite à des modifications comme dans les cas 1 et 2.

5.3.5.2 Les aspects spatiaux

Les acteurs moléculaires évoluent dans un espace à trois dimensions, en particulier il faut garder à l'esprit qu'il y a

- une compartimentation de la cellule,
- des gradients de concentrations (Hirata *et al.*, 2002),
- une colocalisation des acteurs (Huh *et al.*, 2003; Batada *et al.*, 2004),
- une dilution des composés dûe au grossissement des cellules[13] (Pratt *et al.*, 2002).

[13] voir également l'adresse suivante, consultée en août 2007 :
http ://genopole-toulouse.prd.fr/new_image/GenoToul2004_Presentation_M_Cocaign.pdf.

5.3.5.3 Les aspects stochastiques

Les cellules sont très sensibles aux conditions extérieures, ce qui leur permet de s'adapter et de répondre aux stimuli. Cependant, cette sensibilité peut apparaître désavantageuse : comment la cellule assure-t-elle un fonctionnement robuste malgré tant de variations de ses paramètres cinétiques ? Il s'agit d'un aspect particulièrement important, car les sources de bruits ne sont pas qu'extérieures : lorsque le fonctionnement d'un module repose sur quelques dizaines de molécules, la stochasticité de la diffusion et des interactions rendent la compréhension difficile.

Le rôle des bruits endogènes ainsi que la question de la robustesse du fonctionnement cellulaire sont des thématiques fortes qui ont récemment motivé de nombreuses recherches[14]. En effet, l'évolution et la pression de sélection de milliers de générations n'ont pas conduit à des systèmes parfaitement hermétiques au bruit, mais les êtres vivants semblent au contraire exploiter cette variabilité d'une façon et à des fins encore mal comprises.

C'est pourquoi certains modèles ont été développés afin de simuler explicitement cette variabilité : algorithme de Gillespie (1977), algorithme StochSim (Le Novere et Shimizu, 2001), π-calcul (Regev, 2002; R.Blossey *et al.*, 2006), etc.

5.3.5.4 Des mesures sur populations entières

Bien que les méthodes sur cellules uniques commencent à se répandre, elles posent des défis méthodologiques importants. Les données sur des populations de cellules[15], quant à elles, souffrent de la désynchronisation de ces cellules (Sako, 2006), surtout lors de l'étude des rythmes. Il faut donc garder à l'esprit que l'effet de moyenne peut expliquer certains comportements apparents (comme l'atténuation des signaux par exemple).

5.4 Discussion

L'étude théorique des modules fonctionnels a donné naissance à une nouvelle science qu'est la *biologie systémique* (Griffith, 1968a; Griffith, 1968b; De Jong, 2002; Thieffry et De Jong, 2002; Di Ventura *et al.*, 2006) — qui connaît d'ailleurs déjà, depuis plusieurs années, des applications thérapeutiques (Claude *et al.*, 2000).

[14] en témoigne l'école d'été « bruits et robustesse dans les réseaux de régulation transcriptionnelle » qui s'est tenue à Coquelles en septembre 2005.

[15] par exemple, issues des « micro-arrays »

Nous présentons ici un rapide aperçu des réalisations et des résultats théoriques obtenus dans la littérature. À ce titre, l'article de De Jong offre une revue dans ce domaine datant de 2002.

5.4.1 Les réseaux

La telle complexité des réseaux de régulation a tout d'abord poussé un certain nombre de scientifiques à étudier la topologie de ces réseaux (Watts et Strogatz, 1998; Strogatz, 2001; Maslov et Sneppen, 2002; Wuchty et Stadler, 2003; Lattner *et al.*, 2003; N.Przulj *et al.*, 2004; Koschützki et Schreiber, 2004). En particulier, la question de comment inférer cette topologie à partir de données expérimentales a été souvent abordée (Tavazoie *et al.*, 1999; Kim *et al.*, 2003; Kikuchi *et al.*, 2003; Gardner *et al.*, 2003; Sokhansanj *et al.*, 2004; Lok et Brent, 2005).

Certains chercheurs se sont restreints à l'étude de simples motifs correspondant à des sous-réseaux (Hartwell *et al.*, 1999; Struhl, 1999; Shen-Orr *et al.*, 2002; Milo *et al.*, 2002; François et Hakim, 2004), mettant ainsi en évidence le rôle des boucles de rétroaction négatives (Griffith, 1968a; Lema *et al.*, 2000; Roenneberg et Merrow, 2002; Hirata *et al.*, 2002; Monk, 2003) et positives (Griffith, 1968b; Mangan *et al.*, 2003). Kunz et Achermann (2003) ont étudié l'interfaçage de telles sous-unités entre plusieurs cellules ; Reppert et Weaver (2002) ont résumé les couplages entre oscillateurs chez les mamifères.

Inversement, d'autres scientifiques ont abordé la cellule dans sa globalité : c'est le cas du projet « E-cell » qui tente d'intégrer toutes les connaissances actuelles sur la cellule (Yugi et Tomita, 2004; Takahashi *et al.*, 2002).

Afin d'analyser ces réseaux, plusieurs méthodes — parfois originales — issues de l'ingénierie ont été appliquées et sont répertoriées dans l'article de Di Ventura *et al.* (2006). Citons par exemple — pour les plus exotiques — les approches logistiques par réseaux de Petri (Goss et Peccoud, 1998) ou utilisant des « statecharts » (Fisher *et al.*, 2005), les réseaux linéaires flous (Sokhansanj *et al.*, 2004), l'analyse « petit gain » des systèmes monotones (Angeli et Sontag, 2004; Leenheer *et al.*, 2004), etc.

5.4.2 Recherche de fonctions particulières

Ces modules permettent d'assurer certaines fonctions (François et Hakim, 2004), comme l'acheminement des signaux (Aldridge *et al.*, 2006), la bistabilité (Atkinson *et al.*, 2003; Lipshtat *et al.*, 2006), la régulation de certaines quantités (Struhl, 1999;

Alon, 2003), les oscillations[16] (Goldbeter, 1991; Goldbeter, 1995; Lewis, 2003; Naef, 2005; Xu *et al.*, 2007). L'étude de leur synchronisation reste une question ouverte (Gonze *et al.*, 2005) : pour des études expérimentales sur la synchronisation, voir Balsalobre *et al.* (1998) et Nagoshi *et al.* (2004).

Enfin, certains motifs ont été étudiés car ils font apparaître des comportements chaotiques (Leloup et Goldbeter, 1999) pouvant expliquer les arythmies pathologiques observées chez certains patients humains (Roenneberg et Merrow, 2002, 2003) ou après mutations chez la souris (Xu *et al.*, 2007).

5.4.3 Approches envisageables

De nombreuses démarches ont été suivies pour modéliser ces réseaux, l'article de Aldridge *et al.* (2006) donne quelques clefs pour débuter la modélisation d'un système.

En premier lieu, citons, les approches déterministes par équations différentielles ordinaires : EDO (Goldbeter, 1995; François et Hakim, 2004; Gonze *et al.*, 2004). L'attribution de valeurs aux paramètres cinétiques présents dans les équations pose alors un problème, car ils sont rarement connus *in vivo* et dépendent fortement des conditions environnementales. Ce problème peut être en partie traité[17] par l'étude de *diagrammes de bifurcations* présentant qualitativement le comportement du système dans les différents domaines de l'espace des paramètres (Arkin *et al.*, 1998; Atkinson *et al.*, 2003; Gonze *et al.*, 2005), voir figure 5.12.

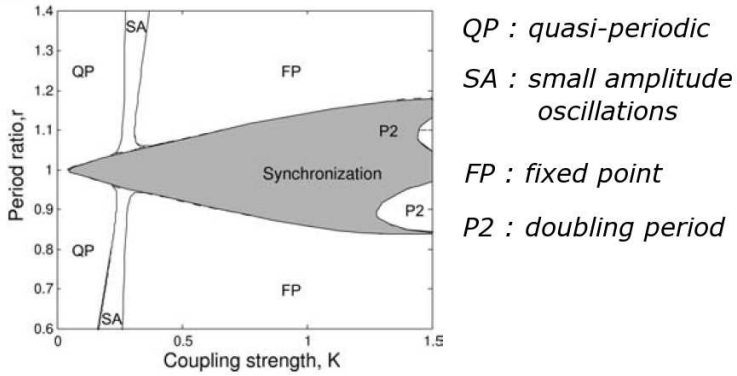


FIG. 5.12: diagramme de bifurcation donnant le comportement de deux oscillateurs couplés en fonction de deux paramètres : la constante de couplage et le rapport des deux périodes propres (extrait de Gonze *et al.*, 2005).

[16] pour une revue concernant les oscillations en biologie, voir Kruse et Jülicher (2005).
[17] quand le nombre de paramètres est restreint...

Une première généralisation des EDO consiste à autoriser l'existence de retards (Lema *et al.*, 2000; Lewis, 2003; Monk, 2003; Kerszberg, 2004), connus pour apporter une grande variété de comportements aux systèmes dynamiques[18] (déstabilisation, stabilisation, oscillations... voir, par exemple, le livre de Richard, 2002).

Citons également les méthodes prenant en compte la stochasticité moléculaire sous-jacente (Gillespie, 1977; McAdams et Arkin, 1997; Le Novere et Shimizu, 2001; El Samad *et al.*, 2005; Paulsson, 2005; R.Blossey *et al.*, 2006). Les modèles stochastiques permettent de mettre en évidence le rôle des bruits (Arkin *et al.*, 1998; Sasai et Wolynes, 2003; Locke *et al.*, 2005), ainsi que la robustesse du fonctionnement vis-à-vis de ces bruits (Vilar *et al.*, 2002; Kerszberg, 2004). Remarquons que Gonze *et al.* (2003 et 2004) ont comparé les comportements des modèles stochastiques et déterministes à de très faibles concentrations et en concluent la validité de l'approche déterministe.

classification des approches	
déterministe	stochastique
discrète	continue
spatiale	concentrations uniformes
avec retards	sans retard

TAB. 5.3: les différents types de description envisageables.

5.4.4 Littérature concernant la modélisation des rythmes biologiques

Depuis le début de la modélisation des rythmes biologiques (Goldbeter, 1991), il y a eu une « course » au plus petit module (en termes de nombre d'acteurs) permettant de produire des oscillations. Nous pouvons ainsi répertorier le repressilateur à 3 gènes (Elowitz et Leibler, 2000; R.Blossey *et al.*, 2006), un certain nombre de modèles à deux gènes formant une boucle négative : Leloup et Goldbeter (1999), Vilar *et al.* (2002), Guantes et Poyatos (2006) et même des systèmes n'impliquant qu'un seul gène dont la protéine réprime sa propre expression (Gonze *et al.*, 2004), introduisant généralement des retards[19] (Lema *et al.*, 2000; Lewis, 2003; Monk, 2003; Kerszberg, 2004). Pour justifier ces retards, les auteurs évoquent les pauses transcriptionnelles,

[18] en effet, on entre dans une classe de systèmes de dimension infinie.

[19] le système proposé par Lema *et al.* (2000), qui emploie une équation à retard, ne comporte concrètement qu'une seule variable dynamique.

les temps de diffusion, les franchissements de membranes et la *maturation* des acteurs (épissage, modifications post-traductionnelles).

Cette recherche de réduction de la taille des modèles traduit une volonté de comprendre en profondeur les mécanismes théoriques sous-jacents aux oscillateurs. C'est cette compréhension que nous avons voulu approfondir.

Pourtant, ce n'est pas uniquement le nombre de gènes qui détermine la possibilité de créer des oscillations : le nombre de variables et d'équations est, à note avis, aussi décisif. Ainsi, certains auteurs ont multiplié le nombre d'acteurs en considérant, comme nous, le niveau de transcription comme une variable à part entière. Ceci est maintenant clairement justifié par les dernières études listées à la section 5.2.1. D'autres auteurs ont utilisé les différents états de dimérisation des protéines (Tyson *et al.*, 1999; Vilar *et al.*, 2002) ou ont distingué les espèces selon qu'elles occupent le noyau ou le cytoplasme (Goldbeter, 1995; Locke *et al.*, 2005). Un autre élément important qui a été relevé, est les différences de phosphorylation de chaque protéine (Goldbeter, 1995; Gonze *et al.*, 2004) ; en effet, des expériences ont récemment mis en évidence des régulations de transcription et de dégradation basées sur la phosphorylation (Xu *et al.*, 2007). Un bel exemple étant celui d'un oscillateur *post-traductionnel*, uniquement basé sur la phosphorylation des acteurs pouvant même être observé *in vitro* (Nakajima *et al.*, 2005), voir aussi Iwasaki *et al.* (2002).

Un autre facteur qui a été utilisé est la *coopérativité* : on considère que la molécule n'est active que lorsqu'elle participe à un (homo-)multimère comportant N_h sous-unités. On utilise alors une fonction de Hill pour estimer la concentration en multimères à partir de la concentration en protéines et le coefficient de Hill est égal à N_h. Gonze *et al.* (2004) ainsi que Blossey *et al.* (2006) ont montré qu'une forte coopération améliorait la robustesse des oscillations dans les modèles stochastiques. De même, Griffith (1968a) a montré qu'une coopérativité minimale était nécessaire pour obtenir des oscillations entretenues.

Nous avons donc mis en évidence un dernier mécanisme, souvent utilisé pour simuler les systèmes (Goldbeter, 1995; Buchler *et al.*, 2005), mais dont l'impact reste mal connu : il s'agit des fonctions de dégradation non linéaires. Gonze *et al.* (2005) ont signalé qu'il était possible de réduire le coefficient de Hill (4 dans leur cas) s'il est fait usage d'une dégradation de type michaëlienne.

5.5 Conclusion

Dans ce chapitre, nous avons étudié un autre aspect des interactions moléculaires : par une approche purement théorique et formelle, nous avons pu mettre en évidence l'influence des interactions à l'échelle moléculaire sur le comportement global d'un module fonctionnel. Ainsi, les dégradations linéaires ne permettent pas d'expliquer, à elles seules, les oscillations d'un modèle minimal avec un seul triplet (gène ;ARN ;protéine). La littérature propose d'autres mécanismes tels que les retards purs, les aspects stochastiques, la compartimentation, etc. Cependant, les profils de dégradation de types enzymatiques sont souvent utilisés, sans que leur impact soit bien compris. Nous avons ici proposé une méthode permettant de caractériser l'espace des paramètres, en se basant sur un critère de stabilité de Routh appliqué au polynôme caractéristique du modèle linéarisé.

Pour pouvoir traiter les équations avec des paramètres quelconques, cette étude fait grandement appel au calcul formel. En particulier, l'utilisation du logiciel Maple nous a permis de manipuler et de factoriser des expressions souvent très volumineuses.

Comme développements futurs, nous voudrions étudier la dépendance et plus précisément la robustesse des oscillations en fonctions des constantes cinétiques. En particulier, il serait utile de savoir s'il est possible de reproduire, avec un modèle aussi simple, la compensation en température caractéristique des rythmes circadiens.

La deuxième étape (entamée dans notre article : Morant *et al.*, 2007, voir annexe J) est d'intégrer une donnée importante du système : la lumière. Pour cela, plusieurs points d'entrée ont été proposés dans la littérature : modification du niveau de dégradation des protéines, modification de la transcription, de la phosphorylation, etc. Ceci est nécessaire en vue de modéliser les données expérimentales.

Enfin, nous pensons également mettre en évidence d'autres mécanismes déstabilisants comme la diffusion des acteurs. En effet, les équations à retards reposent implicitement sur une équation de *propagation* des signaux, tandis qu'un modèle de *diffusion* semble mieux adapté.

Conclusion et perspectives

Conclusion et perspectives

Durant ce travail de thèse, nous avons étudié certains problèmes posés par la modélisation en biochimie autour d'un thème commun : les interactions moléculaires. Pour cela, nous avons parcouru différentes échelles en commençant par une description très détaillée de la molécule individuelle soumise aux potentiels de forces interatomiques. Pour prédire la conformation d'une molécule et, à l'avenir, la conformation et les affinités de complexes moléculaires, nous avons développé une stratégie adaptée à la caractérisation globale du paysage d'énergie potentielle, pourtant fortement multimodal et de grande dimension.

Cette stratégie repose sur l'heuristique des algorithmes génétiques qui, bien que gourmands en ressources de calculs, sont connus pour générer une bonne exploration de l'espace de recherche, indépendamment des barrières énergétiques éventuelles. Nous avons agrémenté cet algorithme central avec un certain nombre des idées récentes du domaine et hybridé l'ensemble avec des heuristiques complémentaires qui se sont révélées très précieuses pour améliorer les temps de calculs et la robustesse de la stratégie. L'originalité de notre approche est d'avoir laissé les paramètres de contrôle de ces algorithmes définissables par un procédé externe et de les avoir gérés par le biais d'une deuxième couche algorithmique (« *méta*-algorithme »). Afin d'optimiser ces paramètres et de mettre en évidence la meilleure stratégie d'hybridation des différentes heuristiques, nous avons proposé un critère d'évaluation d'une exécution particulière de l'algorithme génétique, ce qui nous a permis de valider l'ensemble des développements réalisés jusqu'alors.

Afin d'aborder des molécules de plus grandes tailles, la définition d'une stratégie de parallélisation des algorithmes sous forme de « planètes » représentant les nœuds de calcul a également été validée. Dans ce schéma, l'attribution des ressources à l'intensification par rapport à l'exploration est mise en évidence mais reste le point sensible car, comme nous l'avons montré, la balance optimale dépend de la molécule traitée.

Enfin, la capacité exploratrice de notre dispositif nous a permis de faire un retour critique sur le modèle de champ de forces utilisé pour estimer l'énergie et de revenir sur certains de ses paramètres. Plusieurs idées sont encore en cours de développements et des ouvertures envisageables ont été proposées. De plus, des applications à des cas concrets sont ou ont été étudiées.

Nous avons ensuite vu comment une description plus grossière des motifs pharmacophoriques pouvait être employée pour caractériser les molécules par des indices

topologiques. L'estimation de la similarité moléculaire repose alors sur un critère indépendant des translations rotations (déplacements) pour lesquels nous avons utilisé les quaternions qui nous ont permis de dériver des formules simples et peu coûteuses en temps de calculs. Ce travail a également été validé par une publication dans une revue internationale.

Enfin, dans la dernière partie, nous avons présenté des modèles globaux d'interactions à l'échelle des modules fonctionnels de la cellule. Pour ces derniers, des variables abstraites, représentant les concentrations des acteurs, permettent de masquer la complexité sous-jacente aux molécules individuelles. Toutefois, les réactions à l'échelle moléculaire engendrent des profils différents qui se répercutent sur la dynamique du module. C'est ce que nous avons montré sur un cas minimal en terme de nombre d'acteurs.

Entre la description atomique et la modélisation des modules fonctionnels, nous avons réalisé un formidable « zoom arrière », représentatif du fossé qui existe entre les données expérimentales sur les molécules et les informations qu'il est possible d'obtenir à l'échelle des cellules. Pour combler ce fossé, des méthodes expérimentales et computationnelles commencent à voir le jour. En particulier, les méthodes de microscopie sur molécules uniques commencent à offrir un aperçu de la variabilité et de la spatialité des acteurs. Inversement, des projets comme « e-cell » qui tente d'intégrer toutes les connaissances accumulées sur la cellule, représente, là encore, un premier pas pour joindre les extrêmes molécule-cellule. Nous reproduisons, à ce sujet, la figure de Sali *et al.* (2003) qui fait le point sur les méthodes les plus utilisées.

À notre sens — et ce travail de thèse en est l'illustration — les savoir-faire de l'Automatique sont applicables dans les deux approches : d'une part la conception de nouvelles méthodes et/ou l'ajustement de stratégies de recherche, usuellement appliquées à des problèmes d'ingénierie, peuvent être adaptés à des problèmes d'exploration d'espaces de phase ; d'autre part, la connaissance des outils de modélisation des systèmes dynamiques peut servir à une meilleure compréhension des mécanismes mis en place dans la cellule.

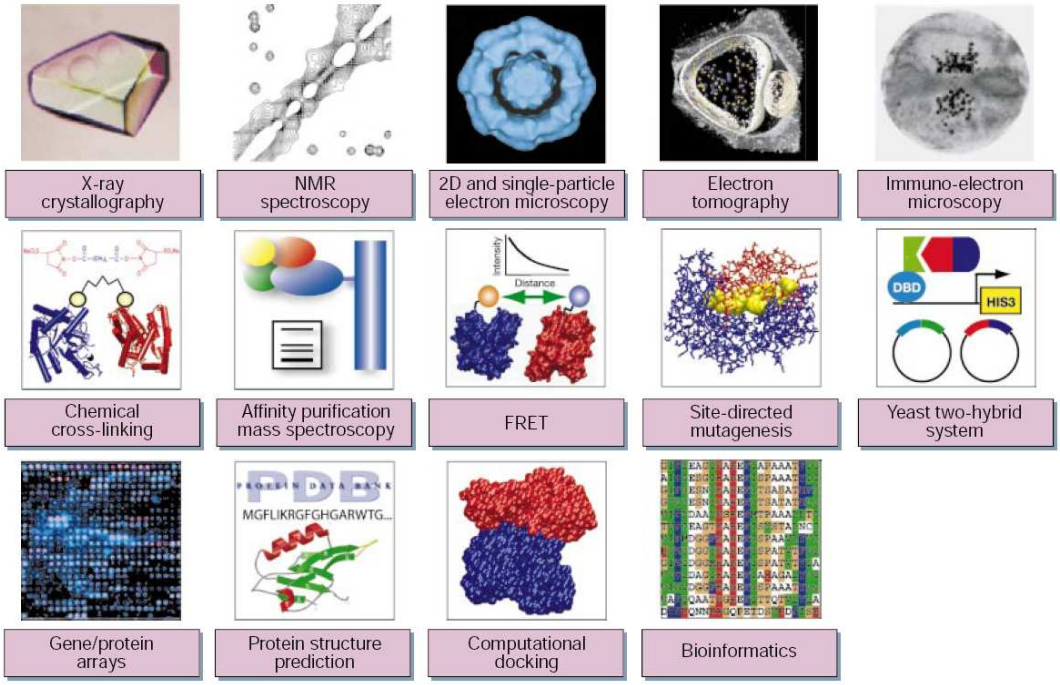


FIG. 5.13: différentes approches possibles afin d'acquérir les données nécessaires à une meilleure compréhension des mécanismes de la cellule ; figure extraite de Sali *et al.* (2003). Nous rajouterions volontiers une dernière case intitulée « modélisation par équations différentielles »...

Troisième partie

Annexes : compléments

Liste des abréviations

abréviation	Détails
1L2Y	Code PDB, Tryptophan Cage
1LE1	Code PDB, Tryptophan Zipper
1UAO	Code PDB, mini peptide formant un β-turn
ADN	Acide désoxyribonucléique
AG	Algorithme Génétique
AMBER	champ de forces et logiciel : Assisted Model Building with Energy Refinement
ARN	Acide ribonucléique
CASP	Critical Assessment of methods of protein Structure Prediction
CEA	Commissariat à l'Énergie Atomique
CFF	Consistent Force Field
CHARMM	Chemistry at HARvard Macromolecular Mechanics force field
CNRS	Centre National de la Recherche Scientifique française
CypB	Cyclophilline A
CRH	Conformationally Restrained Helix
CS	Conformational Sampling
CsA	Cyclosporin A
C_SG_A	Conformational Sampling Genetic Algorithm
CVFF	Consistent Valence Force Field
CypB	Cyclophilin B
ddl	degrés de liberté
ECEPP	Empirical Conformational Energy Program for Peptides
EEF1	Effective Energy Function 1
GB	Generalized Born models

abréviation	**Détails**
GNU	Gnu's Not Unix
GPL	Gnu General Public License
INRIA	Institut National des Recherches en Informatique et Automatique
IRI	Institut de Recherches Interdisciplinaires
LIFL	Laboratoire d'Informatique Fondamental de Lille
MC	Monte Carlo
MD	Molecular dynamics
μG_A	méta Algorithme Génétique
MM2/3/4	champ de forces : Molecular Modeling
MMFF	Merck Molecular Force Field
MW	Master-Worker
NCBI	National Center for Biotechnology Information
NOE	Nuclear Overhauser Effect
EDO	Équation Différentielle Ordinaire
OOB	Observatoire Océanographique de Banyuls sur mer
OPAC	Optimisation PArallèle Coopérative
ParadisEO	PARAllel DIStributed Evolving Objects
PDB	Protein Data Bank
PIN	Protein Interacting with Nima
PMF	Potential of Mean Force
PNAS	Proceedings of the National Academy of Sciences of the USA
QSAR	Quantitative Structure-Activity Relationship
RMN	Résonance Magnétique Nucléaire
RMSD	Root Mean Squared Deviation
UGSF	Unité de Glycobiologie Structurale et Fonctionnelle
UNRES	champ de forces : UNited RESidues
WWW	World Wide Web

Annexe A

Introduction et résultats utiles concernant les quaternions

A.1 Définition

$\mathbb{H}$: un $\mathbb{R}$-espace vectoriel

On appelle *quaternion* tout vecteur de $\mathbb{H} = \mathbb{R}^4$.

On munit alors $\mathbb{H}$ de la base canonique : (e, i, j, k) où $e = {}^t(1,0,0,0)$, $i = {}^t(0,1,0,0)$, $j = {}^t(0,0,1,0)$, $k = {}^t(0,0,0,1)$.

Tout quaternion Q se décompose de façon unique sur (e, i, j, k) et on note (q_0, q_1, q_2, q_3) ses composantes :

$$\forall Q \in \mathbb{H}, \quad \exists!(q_0, q_1, q_2, q_3) \in \mathbb{R}^4 \mid Q = q_0 e + q_1 i + q_2 j + q_3 k. \tag{A.1}$$

On appelle *partie réelle* la composante selon e et *partie imaginaire*, la composante selon (i, j, k). On notera $\mathrm{Re}(Q)$ la partie réelle ($\in \mathbb{R}$) et $\overrightarrow{Q}$ la partie imaginaire de Q ($\in \mathbb{R}^3$). Enfin, on notera $\mathcal{P}$ l'ensemble des quaternions imaginaires purs ; ils forment un sous-espace de $\mathbb{H}$ isomorphe à $\mathbb{R}^3$ de sorte que l'on identifiera $\overrightarrow{Q} = {}^t(q_1, q_2, q_3)$ à ${}^t(0, q_1, q_2, q_3)$ quand il n'y a pas d'ambiguité. On note alors (abusivement) :

$$\begin{aligned} Q &= \mathrm{Re}(Q)e + \overrightarrow{Q}, \\ \mathrm{Re}(Q) &\in \mathbb{R}, \\ \overrightarrow{Q} &\in \mathcal{P} = \mathrm{vect}(i, j, k). \end{aligned} \tag{A.2}$$

Conjuguaison Pour tout quaternion Q, on définit son *quaternion conjugué* : $\overline{Q}$ par

$$\overline{Q} = q_0 e - (q_1 i + q_2 j + q_3 k). \tag{A.3}$$

$\mathbb{H}$: une $\mathbb{R}$-algèbre

On définit maintenant le produit (interne) de deux quaternions par :

$$\begin{array}{rcccl} ex & = & x & & \forall x \in (i,j,k), \\ i^2 & = & j^2 & = & k^2 = -e, \\ ij & = & -ji & = & k, \\ jk & = & -kj & = & i, \\ ki & = & -ik & = & j. \end{array} \tag{A.4}$$

On reconnaît en e, l'élément neutre (qui sera noté en conséquence 1, lorsqu'il n'y a pas d'ambiguïté) et, pour (i,j,k), on a les formules habituelles du produit vectoriel de $\mathbb{R}^3$, à la différence qu'on a maintenant une partie réelle non-nulle en général.

Ce produit est associatif, distributif sur $+$, mais n'est pas commutatif.

L'expression du produit dans la base (e,i,j,k) est :

$$\begin{aligned} QQ' = \quad & (q_0 q'_0 - q_1 q'_1 - q_2 q'_2 - q_3 q'_3) e \\ & + (q_0 q'_1 + q_1 q'_0 + q_2 q'_3 - q_3 q'_2) i \\ & + (q_0 q'_2 + q_2 q'_0 + q_3 q'_1 - q_1 q'_3) j \\ & + (q_0 q'_3 + q_3 q'_0 + q_1 q'_2 - q_2 q'_1) k. \end{aligned} \tag{A.5}$$

En particulier, on a

$$Q\bar{Q} = q_0^2 + q_1^2 + q_2^2 + q_3^2 \triangleq |Q|^2, \tag{A.6}$$

qui est le quaternion réel égal, par définition, au module au carré de Q.

Autre expression du produit Nous disposons d'une autre expression pour le produit QQ' qui fait explicitement apparaître les parties réelle et imaginaire (utiliser (A.5)) :

$$QQ' = \left(\begin{array}{rcl} \mathrm{Re}(QQ') & = & q_0 q'_0 - \overrightarrow{Q}.\overrightarrow{Q'}, \\ \overrightarrow{QQ'} & = & q_0 \overrightarrow{Q'} + q'_0 \overrightarrow{Q} + \overrightarrow{Q} \wedge \overrightarrow{Q'} \end{array} \right), \tag{A.7}$$

où $\wedge$ represente le produit vectoriel habituel de $\mathbb{R}^3$ et $\overrightarrow{q}.\overrightarrow{q'}$ son produit scalaire.

Produit scalaire : ceci nous permet de définir une topologie, et même un produit scalaire dans $\mathbb{H}$:

$$\langle Q|Q'\rangle = \mathrm{Re}(Q.\bar{Q}') = \sum_{i=0}^{3} q_i q'_i, \tag{A.8}$$

$$|Q| = \sqrt{\langle Q|Q\rangle} = \sqrt{q_0^2+q_1^2+q_2^2+q_3^2}. \tag{A.9}$$

Dont on peut rapprocher l'expression de la définition du produit scalaire dans $\mathbb{R}^2$, isomorphe à $\mathbb{C}$:

$$\begin{aligned}\left\langle a\overrightarrow{i}+b\overrightarrow{j}\middle|a'\overrightarrow{i}+b'\overrightarrow{j}\right\rangle &= aa'+bb',\\ (a+ib).(a'-ib') &= aa'+bb' = \mathrm{Re}(z.\bar{z}').\end{aligned}$$

Lorsque les quaternions Q et Q' sont imaginaires purs ($q_0 = q'_0 = 0$), nous avons alors :

$$QQ' = -\langle Q|Q'\rangle\, e + Q\wedge Q'. \tag{A.10}$$

Avec (A.6) et (A.9), on voit que, dès que $Q \neq 0$, on a

$$Q^{-1} = \frac{\bar{Q}}{|Q|^2}. \tag{A.11}$$

En outre, $Q \perp Q' \Leftrightarrow q_0q'_0 + q_1q'_1 + q_2q'_2 + q_3q'_3 = 0 \Leftrightarrow Q\bar{Q}' \in \mathcal{P}$.

Finalement, $\mathbb{H}$ forme une $\mathbb{R}$-algèbre.

A.2 Interprétation géométrique dans $\mathbb{R}^3$

Les quaternions imaginaires purs forment un sous-espace vectoriel isomorphe à $\mathbb{R}^3$ que nous allons identifier à l'espace physique. Les quaternions de norme 1 vont alors encoder les isométries positives de $\mathbb{R}^3$, leur partie imaginaire correspondra à l'axe de rotation et leur partie réelle va nous permettre de stocker l'information « *angle de rotation* ».

Endomorphisme orthogonal Notons $\mathcal{S}$ la sphère unité de $\mathbb{H}$, c'est-à-dire l'ensemble des quaternions de module 1.

Pour tout $Q \in \mathcal{S}$, on définit également l'application

$$\begin{array}{rccc} f_Q & : & \mathbb{H} & \rightarrow & \mathbb{H} \\ & & p & \rightarrow & Qp\bar{Q}. \end{array} \tag{A.12}$$

Théorème : f_Q restreint à $\mathcal{P}$ (le sous-espace des quaternions imaginaires purs) est un endomorphisme orthogonal (une *isométrie* de $\mathcal{P}$).

Démonstration :

* La linéarité est évidente ;

* Vérifions la stabilité de $\mathcal{P}$ (elle ne tient pas au fait que $Q \in \mathcal{S}$) :

en utilisant (A.7), on a

$$Q = q_0 e + \overrightarrow{Q} \qquad ; \qquad p = \overrightarrow{p},$$

$$\begin{array}{rcl} \mathrm{Re}(Qp\bar{Q}) & = & \mathrm{Re}(Qp)\overbrace{\mathrm{Re}(\bar{Q})}^{q_0} - \mathrm{Im}(Qp)\overbrace{\mathrm{Im}(\bar{Q})}^{-\overrightarrow{Q}} \\ & = & (q_0 0 - \overrightarrow{Q}.\overrightarrow{p})q_0 + (q_0\overrightarrow{p} + 0 + \overrightarrow{Q} \wedge \overrightarrow{p}).\overrightarrow{Q} \\ & = & (\overrightarrow{Q} \wedge \overrightarrow{p}).\overrightarrow{Q} \\ & = & 0. \end{array}$$

* Enfin, montrons l'orthogonalité :

f_Q est orthogonal si et seulement si $|p| = |f_Q(p)|$ pour tout p imaginaire pur.
Or, $|f_Q(p)| = |Qp\bar{Q}| = |Q|.|p|.|\bar{Q}| = |p|$, puisque $Q \in \mathcal{S}$.

$\diamondsuit$

Puisque $Q \in \mathcal{S}$, $\mathrm{Re}(Q)^2 + \|\overrightarrow{Q}\|^2 = 1$, nous pouvons poser α et $\overrightarrow{u}$ tels que

$$\mathrm{Re}(Q) = \cos(\frac{\alpha}{2}), \tag{A.13}$$

$$\overrightarrow{Q} = \sin(\frac{\alpha}{2})\overrightarrow{u}, \tag{A.14}$$

$$\alpha \in [0, 2\pi], \tag{A.15}$$

$$\|\overrightarrow{u}\| = 1. \tag{A.16}$$

Le cas $\alpha = 2\pi$ ($Q = -1$) peut également être exclu puisque $\forall p \in \mathcal{P}, f_{-1}(p) = p$

autrement dit, $f_{-1} = f_1 =$ identité.

Cette décomposition est unique tant que $|\mathrm{Re}(Q)| \neq 1$ ($f_Q \neq$ identité), car dans ce cas le choix de $\overrightarrow{u}$ est arbitraire. Enfin, remarquons que l'axe porté par $\overrightarrow{u}$ est orienté car changer $\overrightarrow{u}$ en $(-\overrightarrow{u})$ revient à changer α en $(-\alpha)$, ce qui n'est pas possible d'après (A.15).

Si $\alpha = \pi$, Q est imaginaire pur et f_Q est la symétrie axiale d'axe $\overrightarrow{u}$.

Théorème : si $Q = \cos(\alpha/2)e + \sin(\alpha/2)\overrightarrow{u}$, $\overrightarrow{u}$ unitaire et $p \in \mathbb{R}^3$, alors $f_Q(p)$ est l'image du point p par la rotation d'angle α et d'axe la droite portée par le vecteur directeur $\overrightarrow{u}$.

Démonstration :

On notera les vecteurs avec des flêches ($\overrightarrow{u}$) pour mettre en évidence les produits scalaire et vectoriel de $\mathbb{R}^3$, tandis que les produits de quaternions seront non-fléchés.

$$\begin{aligned}
Q &= \cos(\frac{\alpha}{2}) + \sin(\frac{\alpha}{2})\vec{u}, \\
(\vec{u}, p) &\in \mathcal{P}^2.
\end{aligned}$$

Calculs préliminaires :

$$\begin{aligned}
\vec{u}p &= -\langle\vec{u}|\vec{p}\rangle + \vec{u}\wedge\vec{p}, \\
p\vec{u} &= -\langle\vec{u}|\vec{p}\rangle - \vec{u}\wedge\vec{p}, \\
\vec{u}p\vec{u} &= -\langle\vec{u}|\vec{p}\rangle\vec{u} + (\langle\vec{u}\wedge\vec{p}|\vec{u}\rangle + (\vec{u}\wedge\vec{p})\wedge\vec{u}).
\end{aligned}$$

$$\begin{aligned}
\text{Or } \langle\vec{u}\wedge\vec{p}|\vec{u}\rangle &= 0 \\
\text{et } (\vec{u}\wedge\vec{p})\wedge\vec{u} &= \langle\vec{u}|\vec{u}\rangle p - \langle\vec{u}|\vec{p}\rangle\vec{u} = p - \langle\vec{u}|\vec{p}\rangle\vec{u}, \\
\text{soit } \vec{u}p\vec{u} &= p - 2\langle\vec{u}|\vec{p}\rangle\vec{u}.
\end{aligned}$$

Calcul de $f_Q(p)$:

$$\begin{aligned}
Qp\bar{Q} &= \left[\cos(\frac{\alpha}{2}) + \sin(\frac{\alpha}{2})\vec{u}\right] p \left[\cos(\frac{\alpha}{2} - \sin(\frac{\alpha}{2})\vec{u}\right] \\
&= \cos^2(\frac{\alpha}{2})p + \cos(\frac{\alpha}{2})\sin(\frac{\alpha}{2})(\vec{u}p - p\vec{u}) - \sin^2(\frac{\alpha}{2})\vec{u}p\vec{u} \\
&= \cos(\alpha)p + \sin(\alpha)\vec{u}\wedge\vec{p} + 2\sin^2(\frac{\alpha}{2})\langle\vec{u}|\vec{p}\rangle\vec{u}.
\end{aligned}$$

En linéarisant le sinus, on a

$$\begin{aligned}
Qp\bar{Q} &= \langle\vec{u}|\vec{p}\rangle\vec{u} + \cos(\alpha)\,(p - \langle\vec{u}|\vec{p}\rangle\vec{u}) + \sin(\alpha)\vec{u}\wedge\vec{p}, \\
\text{qu'on peut écrire} \quad & \langle\vec{u}|\vec{p}\rangle\vec{u} + \cos(\alpha)\,(p - \langle\vec{u}|\vec{p}\rangle\vec{u}) + \sin(\alpha)\vec{u}\wedge(p - \langle\vec{u}|\vec{p}\rangle\vec{u})\,.
\end{aligned}$$

Le terme en $\langle\vec{u}|\vec{p}\rangle\vec{u}$ correspond à la composante de p selon l'axe de rotation $\vec{u}$ et est donc resté inchangé au cours de la rotation. Le vecteur $p_\perp \triangleq p - \langle\vec{u}|\vec{p}\rangle\vec{u}$ apparaissant dans les deux derniers termes est le projeté de p sur le plan orthogonal à $\vec{u}$; il est transformé en $\cos(\alpha)p_\perp + \sin(\alpha)\vec{u}\wedge p_\perp$ qui est bien la décomposition du vecteur image de $p_\perp$ par la rotation d'axe porté par $\overrightarrow{u}$ et d'angle α.

Enfin, il s'agit d'une rotation vectorielle, pour exprimer une rotation affine, il faut encore réaliser une translation :

$$r(p) = A + f_Q(p - A), \tag{A.17}$$

où A est un point quelconque de l'axe de rotation...

Lien avec les angles d'Euler. Voici pour finir, les équations liant les angles d'Euler avec les coefficients du quaternion correspondant :

$$\left\{\begin{array}{lcl} q_0 &=& \cos(\frac{\theta}{2})\cos(\frac{\psi+\phi}{2}) \\ q_1 &=& \sin(\frac{\theta}{2})\cos(\frac{\psi-\phi}{2}) \\ q_2 &=& \sin(\frac{\theta}{2})\sin(\frac{\psi-\phi}{2}) \\ q_3 &=& \cos(\frac{\theta}{2})\sin(\frac{\psi+\phi}{2}) \end{array}\right. \tag{A.18}$$

Et inversement :

$$\left\{\begin{array}{lcl} \theta &=& \arccos(q_0^2 + q_3^2 - q_1^2 - q_2^2) \\ \phi &=& \arctan_2(q_3, q_0) - \arctan_2(q_2, q_1) \\ \psi &=& \arctan_2(q_3, q_0) + \arctan_2(q_2, q_1) \end{array}\right. \tag{A.19}$$

A.3 Interprétation matricielle

L'interprétation matricielle, basée sur un isomorphisme entre $\mathbb{H}$ et l'espace $SO_4(\mathbb{R})$ des matrices réelles orthogonales (4×4) permet d'introduire plus naturellement la notion de produit et simplifie en outre certaines démonstrations. Par ailleurs, elle permet d'appréhender les quaternions comme une sous-algèbre d'un espace plus grand, plutôt que comme l'extension d'un espace plus petit...

On définit (e, i, j, k) de la manière suivante :

$$e = Id = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad , \qquad i = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix} \quad ,$$
$$j = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad , \qquad k = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \quad , \tag{A.20}$$

à rapprocher des complexes :

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad , \qquad i = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Autrement dit, on a l'isomorphisme caractérisé par la mise en bijection de

$$Q = q_0 e + q_1 i + q_2 j + q_3 k, \text{ avec } \mathcal{M}_Q = \begin{pmatrix} q_0 & -q_1 & -q_2 & q_3 \\ q_1 & q_0 & -q_3 & -q_2 \\ q_2 & q_3 & q_0 & q_1 \\ -q_3 & q_2 & -q_1 & q_0 \end{pmatrix}. \tag{A.21}$$

On vérifie aisément les propriétés suivantes :
- $\mathcal{M}_{\overline{Q}} = {}^t\mathcal{M}_Q$,
- $\det(\mathcal{M}_Q) = |Q|^4$, (on retrouve ainsi qu'il n'y a pas de diviseurs de zéro dans $\mathbb{H}$),
- $\overline{Qq} \equiv {}^t\mathcal{M}_{Qq} = {}^t(\mathcal{M}_Q\mathcal{M}_q) = {}^t\mathcal{M}_q\,{}^t\mathcal{M}_Q \equiv \overline{q}\,\overline{Q}$.

Ces matrices redonnent les formules des produits (A.4), ce qui montre que vect(e, i, j, k) (que l'on appellera $\mathbb{H}$) est bien stable par le produit des matrices.

Annexe B

Revue des principaux articles concernant 1LE1

B.1 Muñoz *et al.* 1997, Nature

Cet article ne concerne pas 1LE1 directement mais plutôt la formation des petites épingles β en général ; il marque le début de l'étude des structures β contrairement aux α-hélices, connues depuis plus longtemps. Les causes sont multiples :

- elles sont moins stables : l'auteur reporte un temps de repliement 30 fois plus long que pour les hélices soient 6μs environ,
- elles agrègent plus facilement.

Les auteurs confrontent leurs données expériementales de fluorescence à un modèle simpliste prenant en compte :

- ΔS : la perte entropique dûe au repliement,
- ΔH : gain énergétique dû à la formation de ponts hydrogènes (notés HB),
- ΔG : gain dû à la formation d'un cluster aromatique hydrophobe.

Le modèle consiste à compter les résidus « gelés », les ponts hydrogène et les contacts hydrophobes.

Le retour à l'équilibre après un saut de température a été suivi par fluorescence sur un tryptophane ; il suit une monoexponentielle permettant d'estimer les paramètres ΔS et ΔH (bien que le ratio de molécules en épingle ait diminué de 15%).

S'ensuit une discussion des différences de formation des hélices par rapport aux tournants β :

hélices α	structures β
L'apport d'un HB (énergétiquement bénéfique) se fait au prix du gèle d'un résidu (entropiquement défavorable).	Pour créer un HB, il faut bloquer deux résidus.
La formation d'une hélice peut commencer en plusieurs endroits simultanément.	La formation d'une épingle est quasiment séquentielle.

B.2 Cochran *et al.* 2001, PNAS

Premiers auteurs à concevoir la famille des « tryptophan zippers » (trpzips) et en particulier celui qui nous concerne : le trpzip 2, appelé alors 1HRX (futur 1LE1).

Dans la course à la plus petite épingle β stable n'utilisant pas de pont covalent (qu'on pensait limitée à 20-30 acides aminés), ils ont cherché à utiliser le tryptophane qui est connu pour faire des « stackings » stabilisant. Sachant que la modélisation des paires aromatiques est une question difficile, l'étude *in vitro* est justifiée.

Alors que les précédentes épingles connues n'étaient pas très stables (ΔG quasi nulle à 298K), l'utilisation d'un double stacking Trp-Trp stabilise grandement la structure. Les trpzips ont des énergies de repliement (par résidu) comparables à celles de protéines bien plus grandes ($\Delta G = 60 - 120$ cal.mol^{-1}.residu^{-1}). Ainsi l'étude par dichroïsme circulaire (DC) et résonance magnétique nucléaire (RMN) montre que *la dénaturation thermique est réversible*. L'entropie de dépliement est toutefois plus grande que celle des moyennes et grandes protéines avec un $\Delta S_{110°C} = 6.4 \pm 0.3$ cal.mol^{-1}.residu^{-1}.

Les structures ont été déterminées par « distance geometry » et recuit simulé, et les meilleures solutions ont été affinées par dynamique moléculaire avec Amber/Discover.

Les spectres CD indiquent des interactions entre chromophores aromatiques et attestent de la présence d'une structure tertiaire bien définie. Des expériences à différentes températures montrent également que *les trpzips ne dimérisent pas* aux concentrations considérées (entre mM et μM).

Les trpzips constituent donc des systèmes idéaux pour l'étude théorique et expérimentale du stacking aromatique.

B.3 Yang *et al.* 2004, Journal of Molecular Biology

Cet article présente une simulation « all atom » par « Replica Exchange Molecular Dynamics » et des expériences sur le repliement des trpzips...

Remarque : la structure initiale de trpzip2 (1HRX) a été revue à la lumière des dernières découvertes sur le stacking des cycles aromatiques, il devient 1LE1 ; les simulations en dynamique moléculaire avec un champ de force plus récent ainsi que des expériences de RMN ont plutôt montré une structuration des tryptophanes en forme de « T » (tranche contre face).

Le trpzip2 exhibé par Cochran *et al.* est extrèmement stable et monomérique, même à des concentrations élevées en dénaturant, et ce malgré le fait que le stacking aromatique n'est pas tout à fait isolé du solvant... À forte concentration de dénaturant (GuHCL), 1LE1 semble se replier suivant un modèle à deux états, tandis qu'en conditions normales, il exhibe plutôt une cinétique de repliement hétérogène avec de multiples minima. Les auteurs cherchent à mettre ceci en évidence par des simulations.

Au moins 3 régions de transitions ont été identifiées par dynamique moléculaire dont 2 sont observées expérimentalement en conditions normales et la dernière lorsque le dénaturant est ajouté. La dynamique moléculaire a également localisé 7 bassins d'attraction (structurellement distincts) à basse température, sans réelle barrière intermédiaire. Ceci accrédite la thèse du paysage d'énergie rugueux autour du natif. D'ailleurs, à *très* basse température, la simulation donne une unique structure proche des données RMN (voir figure B.1).

Afin de comparer les données du 1LE1, les auteurs disposent de « Carm5 » : un pentapeptide formé par l'un des deux feuillets d'1LE1 qui permet de mimer l'environnement des tryptophanes tout en empêchant leurs interactions croisées.

Les potentiels de champs moyens sont estimés sur la base de simulations de dynamique moléculaire (MD) avec « Replica exchange » (voir section 3.2.5). Les simulations ont été menées avec le champ de force de AMBER : parm96 et un modèle de solvant implicite (Generalized Born/Solvent Accessible Surface Area), les charges ont été fixées à pH 7. Les longueurs de liaisons sont restreintes à leurs dimensions nomiales par l'algorithme SHAKE.

Enfin, les auteurs font remarquer que dans leurs simulations, seulement 3 des 4 tryptophanes ont leurs angles dans la fourchette expérimentale.

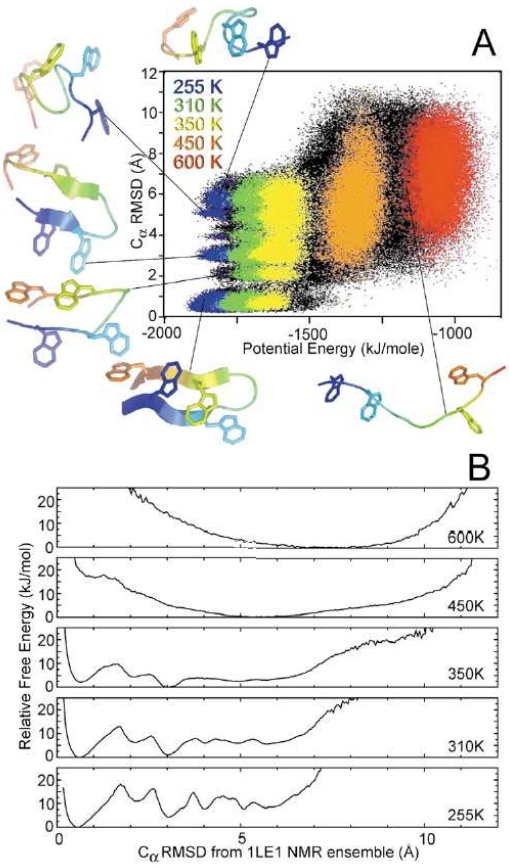


FIG. B.1: distribution de l'énergie potentielle en fonction de l'RMSD du squelette, et énergie libre en fonction de l'RMSD à différentes températures.

B.4 Snow *et al.* 2004, PNAS

Étude de la dynamique de repliement des trpzips 1, 2 et 3 par simulation de MD, par suivi expérimental de sauts de température par fluoresccence et par spectroscopie infrarouge.

La simulation, d'une durée totale (en faisant la somme) de 22ms, permet de mettre à jour les défauts du champ de forces « OPLS atomes unifiés » (qui prédit des minima non-natifs dans le paysage d'énergie libre), et de valider le champ de forces « OPLS all-atom » (OPLSaa) qui a bien reproduit les taux de repliement et les enthalpies de dépliement (bien que le trpzip 3 fût sensible aux conditions initiales).

Afin d'analyser les données statistiques sur l'ensemble de la simulation (démarée à partir d'une conformation expérimentale), les auteurs surveillent deux variables représentatives : le RMSD au natif et la somme, notée L, des distances correspondantes aux ponts hydrogène et ponts entre cycles aromatiques attendus (plus L est petit, plus les deux feuillets β seront ressérés). En chaque point du plan (RMSD,L), le potentiel de champ moyen est estimé par la formule de Boltzmann.

Énormément (dizaines de milliers) de simulations (de 10ns à plus de 1,5μs) ont été lancées en parallèle grâce à l'environnement de Folding@Home[1], pour différentes températures, différentes conformations initiales, différentes paramétrisations du champ de forces, etc. La formation de l'épingle a été reproduite plusieurs centaines de fois à température ambiante.

Les auteurs incitent à la validation des champs de forces par des approches expérimentales complémentaires (comme ici où les données de fluorescence sont parfois sensiblement différentes de celles de spectroscopie).

B.5 Guvench *et al.* 2005, Journal of the American Chemical Society

Cet article reporte la modélisation par dynamique moléculaire du peptide 1LE1 en « all-atom » afin de mieux comprendre le positionnement relatif des différents résidus tryptophanes.

Les auteurs rappellent quelques travaux antérieurs :

– des calculs quantiques *ab initio* sur les dimères de benzènes ont conclu que les conformations « Edge-to-Face » (EtF) et « Parallel Displaced » (PD) sont qua-

[1] http ://folding.stanford.edu

siment iso-énergétiques avec une faible barrière les séparant (~ 0.2kcal.mol^{-1}).
- Des statistiques sur les structures protéiques cristallines connues ont mis en évidence tout un continuum dans la répartition des angles entre plans aromatiques, avec un faible avantage énergétique pour les conformations PD (énergie libre inférieure à 1kcal.mol^{-1}).
- Enfin, des études de mécanique moléculaire dans le solvant ont montré que les stackings Phenyl-Phenyl, Phenyl-Tyrosine et Tyrosine-Tyrosine étaient légèrement plus stables en PD qu'en EtF contrairement aux dimères benzéniques, plutôt en EtF.

Les auteurs ont alors comparé les trajectoires de MD avec et sans la prise en compte des multipôles (notées respectivement +MP et -MP) afin de mettre en évidence leurs effets sur la conformation. Ceci est réalisé en lançant les simulations avec et sans les charges partielles, tout en conservant les groupements NH intacts, pour ne pas perturber les ponts hydrogènes. En utilisant le logiciel Charmm, le champ de forces Charmm22 (MacKerell *et al.*, 1998) et un solvant explicite, les auteurs ont constaté que

- dans les deux cas, le squelette fluctue peu (surtout au voisinage du *tournant* et des extrémités) et de manière similaire, protégeant ainsi les ponts hydrogènes du solvant ;
- par contre, les chaînes latérales des tryptophanes se comportent différemment : en -MP, elles fluctuent beaucoup plus entre EtF et PD, favorisant légèrement les conformations PD, tandis qu'en +MP, elles favorisent largement EtF.

Remarque : les simulations sont lancées à partir des structures RMN, qui sont toutes en EtF.

De plus amples simulations ont permis d'estimer les variations énergétiques entre EtF et PD (en séparant contributions électrostatiques et Van der Waals) dans les cas -MP et +MP. La principale différence tient aux interactions électrostatiques entre chaînes latérales aromatiques (en comparaison des interactions électrostatiques entre tryptophanes et solvant et des contributions Van der Waals avec le solvant ou entre tryptophanes).

Conclusion : les auteurs préconisent le développement d'un terme supplémentaire dans les champs de force, spécifique aux interactions aromatique-aromatique, qui évoluerait en $\frac{1}{d^n}$ où n est nécessairement supérieur à 1 puisqu'il ne s'agit pas de simples interactions coulombiennes (sous-jacentes ici) entre charges ponctuelles mais entre multipôles.

B.6 Wenzel *et al.* 2006, Europhysics Letters

Les auteurs ont précédemment développé un champ de forces « all-atom » pour l'estimation de l'énergie libre, dédié spécifiquement aux protéines hélicoïdales : PFF01 (Herges et Wenzel, 2004). Il a été testé sur des protéines de tailles 20-60 acides aminés, puis a été modifié pour accepter les protéines formant des épingles β. Ce champ de forces agit dans l'espace torsionel, il comprend les termes de Coulomb, de surface accessible au solvant, un potentiel de Lennard-Jones, les contributions des ponts hydrogène et un terme torsionel pour le squelette.

Alors que d'autres auteurs (Snow *et al.*, 2004) ont reproduit le repliement de 1LE1 (ainsi que deux autres trp zippers) par des simulations de 22ms (soient $O(10^{12})$ évaluations de l'énergie), on montre ici que la méthode de « Basin hopping technique » (BHT) permet de mettre à jour le repliement de 1LE1 avec $O(10^6)$ évaluations, de façon prédictive et reproductible.

Le principe de la BHT consiste à remplacer l'évaluation de l'énergie des conformations par celle du minimum le plus proche ; le paysage d'énergie potentielle ressemble alors à une succession de plateaux où les barrières ont disparues. Cette approche est utilisée ici de concert avec le recuit simulé.

Sur 10 simulations indépendantes, 4 ont convergé vers le minimum énergétique connu avec un RMSD (sur le squelette uniquement) inférieur à 2Å, une 5$^{\text{e}}$ a convergé en terme de RMSD mais avec $+3\text{kcal.mol}^{-1}$ par rapport au natif. Les 5 dernières se sont arrêtées dans la fourchette $[+4;+10]\text{kcal.mol}^{-1}$ et des RMSD supérieurs à 3Å. Les auteurs précisent également qu'ils reproduisent *correctement* le stacking des tryptophanes qui apparaît sur la figure comme étant en PD.

Statistiques sur les conformations échantillonnées. Afin de comprendre pourquoi le terme de formation de HB (2kcal.mol^{-1} par pont hydrogène, dès qu'un groupement CO est à moins de 3Å d'un NH) ne domine pas la dynamique — ce qui d'ailleurs favoriserait les conformations hélicoïdales — les auteurs ont analysé le terme de désolvatation qui le compense. Ces deux contributions semblent en effet antagonistes, et il apparaît que le repliement de 1LE1 repose sur la compétition entre désolvatation et formation de ponts hydrogène.

Lorsqu'on représente l'énergie libre selon les deux variables $\sharp s$ (short = nombre de ponts hydrogène entre résidus topologiquement proches) et $\sharp l$ (long = nombre de ponts hydrogène entre résidus topologiquement éloignés), on met en évidence les hélices et les épingles β. On voit alors apparaître un puits profond autour du natif,

mais également un deuxième minimum local dans la région hélicoïdale.

Lorsqu'on choisit les variables $\sharp HB$ (nombre de ponts hydrogène) et RMSD (sur le squelette), on voit que lorsque l'on s'éloigne de la conformation native, les 4 ou 5 ponts hydrogènes sont coupés pour ensuite en reformer 6 ou 7, voire plus dans une conformation en hélice.

Les auteurs mettent ces résultats en relation avec l'article de Yang *et al.* (2004) puisque le paysage d'énergie en conditions normale apparaît très complexe et rugueux. L'ajout de GuHCL comme dénaturant (connu pour stabiliser les tonneaux β) doit alors certainement déstabiliser les conformations hélicoïdales au profit de la structure native.

Quatrième partie

Publications personnelles, conférences et posters

Annexe C

Article 1 : Journal of Soft Computing, 2007

paru dans « Journal of Soft Computing » en janvier 2007, 11(1), p. 63-79

B. Parent, A. Kökösy et D. Horvath,

Optimized Evolutionnary Strategies in Conformational Sampling

Soft Comput (2006)
DOI 10.1007/s00500-006-0053-y

ORIGINAL PAPER

Benjamin Parent · Annemarie Kökösy
Dragos Horvath

Optimized evolutionary strategies in conformational sampling

Abstract Novel genetic algorithm (GA)-based strategies, specifically aimed at multimodal optimization problems, have been developed by hybridizing the GA with alternative optimization heuristics, and used for the search of a maximal number of minimum energy conformations (geometries) of complex molecules (conformational sampling). Intramolecular energy, the targeted function, describes a very complex nonlinear response hypersurface in the phase space of structural degrees of freedom. These are the torsional angles controlling the relative rotation of fragments connected by covalent bonds. The energy surface of cyclodextrine, a macrocyclic sugar molecule with $N = 65$ degrees of freedom served as model system for testing and tuning the herein proposed multimodal optimization strategies. The success of GAs is known to depend on the peculiar hypotheses used to simulate Darwinian evolution. Therefore, the conformational sampling GA (CSGA) was designed such as to allow an extensive control on the evolution process by means of tunable parameters, some being classical GA controls (population size, mutation frequency, etc.), while others control the herein designed population diversity management tools or the frequencies of calls to the alternative heuristics. They form a large set of operational parameters, and a (genetic) meta-optimization procedure was used to search for parameter configurations maximizing the efficiency of the CSGA process. The specific impact of disabling a given hybridizing heuristics was estimated relatively to the default sampling behavior (with all the implemented heuristics on). Optimal sampling performance was obtained with a GA featuring a built-in tabu search mechanism, a "Lamarckian" (gradient-based) optimization tool, and, most notably, a "directed mutations" engine (a torsional angle driving procedure generating chromosomes that radically differ from their parents but have good chances to be "fit", unlike offspring from spontaneous mutations). "Biasing" heuristics, implementing some more elaborated random draw distribution laws instead of the 'flat' default rule for torsional angle value picking, were at best unconvincing or outright harmful. Naive Bayesian analysis was employed in order to estimated the impact of the operational parameters on the CSGA success. The study emphasized the importance of proper tuning of the CSGA. The meta-optimization procedure implicitly ensures the management, in the context of an evolving operational parameterization, of the repeated GA runs that are absolutely mandatory for the reproducibility of the sampling of such vast phase spaces. Therefore, it should not be only seen as a tuning tool, but as the strategy for actual problem solving, essentially advocating a parallel exploration of problem space and parameter space.

Keywords Genetic algorithms · Multimodal optimization · Hybrid optimization techniques · Island model · Algorithm performance tuning · Molecular modeling, conformational sampling

Abbreviations GA Genetic algorithm · CSGA Conformational sampling GA · μGA Meta-GA (used for parameter setup optimization) · μF Meta-fitness score (target function of the μGA) a measure of success of conformational sampling

B. Parent
UMR 8117, Institut de Biologie de Lille, 1, rue Calmette
59019 Lille CEDEX, France

B. Parent
Institut Supérieur d'Electronique et du Numérique
41, Boulevard Vauban, 59000 Lille CEDEX, France

D. Horvath (✉)
UMR 8576 - CNRS, Université des Sciences
& Technologies de Lille, Cité Scientifique - Bât. C9
59655 Villeneuve d'Ascq, France
E-mail: dragos.horvath@univ-lille1.fr

A. Kökösy
LAGIS - UMR 8146, 59650 Villeneuve d'Ascq, France

1 Introduction

The study of complex (multi-dimensional and highly non-linear) functions, and, in particular, the search of their optima, has always been a major challenge in science and engineering. The study of such systems is, of course, directly

motivated by the fact that life itself is extraordinarily complex. Conformational sampling [14,24], e.g. predicting on hand of computational techniques how (bio)molecules "fold" [3,29,39] in a given solvent, is a problem of physical chemistry with a potentially high importance for theoretical biology and drug design. According to Boltzmann's distribution, the probability for a molecule to adopt a state of energy E, at a temperature T, is proportional to $\exp(-E/k_{\mathrm{B}}T)$ where k_{B} is Boltzmann's constant. A "state", in the above sense, would be fully defined by the set of $3\,\mathrm{N}_{\mathrm{atoms}}$ atomic coordinates. Here, however, the torsional angles around bonds that allow the free rotation of interconnected fragments are used as the actual degrees of freedom [37]. All the populated low-energy states, not only the absolute energy minimum, need to be discovered (multimodal optimization), as they are potential contributors to the experimentally measurable "average" molecular properties. Intramolecular potential energy is typically calculated according to some empirical molecular force field [16], based on an estimation of the different interactions between the atoms of the molecule.

Structure determination of biomolecules requires input of experimental constraints derived from measured nuclear Overhauser effects (NOE) or X-ray diffraction density maps [2]. The rugged energy landscape is thus turned into a funnel-like hypersurface with a clear-cut minimum representing conformers that fulfill these constraints. Other attempts to "ease" the problem solving involve the use rotamer libraries [33] enumerating the experimentally most often encountered torsional states.

This paper primarily focuses on the algorithmic aspects of exploring a molecular energy surface, like the one of cyclodextrine, chosen as benchmark in this work. The "success" of the sampling procedures will be assessed with respect to the deepness and number of independent minima of the energy surface found at given computational effort.

Different categories of stochastic algorithms inspired by statistical physics have already been used for conformational sampling, notably molecular dynamics [20] and simulated annealing [39]. However, their ability to visit relevant minima highly depends on the initial conditions, given the difficulty to cross the high potential barriers present in the energy landscapes. Other sampling heuristics deal with a pool of solutions called individuals or particles: sequential Monte Carlo sampling [5,8], and the "ant paradigm" [39] based on the recruitment of individuals ("ants") in interesting areas of the search space thanks to a temporary memory ("pheromones").

A powerful problem space exploration strategy, the genetic algorithm (GA) [1,19,26,32], simulates a Darwinian evolution process in order to achieve convergence of an initial random population of solutions towards an optimum of the response surface. Innovative strategies like elitism, parallelization, similarity filtering (to simulate food sharing) [40] have been added to the "core" GA [13]. GAs have already been used [4] for conformational sampling. However, the classical GA methodology suffers from a series of defaults with respect to certain peculiarities of the conformational sampling problem. A goal of this work is to suggest further improvements, mainly based on "hybridizations" of the classical GA with other optimization techniques, as follows:

- Adapted probability distributions for the random draw of torsion angle values: Classical GAs typically use "flat" random distributions to initialize the variables of the first, random population. In conformational sampling, each torsional angle value would therefore be equiprobably given a value between 0° and 359°. However, torsional angle values triggering extremely unfavorable local interactions (between the atoms directly bound to the heads of the torsional axis) are, except for highly strained ring systems, rarely seen in optimal molecular folds. Rather than waiting for the Darwinian selection process to eradicate such unfit genes from the "gene pool", two alternative "biasing" strategies of the torsion value random draw were assessed here: the "local strain" strategy favors the draw of values minimizing the local interaction strain, while the "tradition-based" approach prioritarily draws values observed in previously sampled, stable conformers.
- Tabu search: GAs were typically employed to quickly find a *reasonable* solution rather than the global optimum of a problem. Although GAs generate whole populations of solutions, they were rarely used for actual multimodal optimization, and their ability to find several different optima was not carefully assessed. Classical GAs may revisit previously found optima and therefore waste computational resources. In order to avoid this, the introduction of a "tabu" search mechanism [11,12] ensuring a self-avoiding walk in problem space has been attempted.
- Lamarckian optimization: Due to the peculiar nature of the potential energy function, including a "hard" atom–atom repulsion term depending on the inverse of the twelfth power of interatomic distance [16], a chromosome coding a near-optimal conformer with a slightly misplaced terminal fragment may score an energy largely above the level of typical "unfolded" structures. Waiting for a random mutation to "fix" the problematic detail is not a good strategy, as the "almost correct" solution may not pass the next selection step. The obvious choice is to let it glide to the closest energy optimum, following the gradient. To keep up the analogy with evolutionary theories, such a move may be viewed as a "Lamarckian" process, where the individual "learns" from its environment, improves itself and then "back-copies" the acquired knowledge into its genome.
- Directed mutations: Random mutations are a key element of natural evolution, although a notoriously ineffective one, as most such changes are highly detrimental. Likewise, a random change of a torsion in a stable conformer will rather lead to an impossible arrangement with overlapping atoms than to a more stable geometry. Rotations of fragments around their axes typically occur in a concerted manner, following the minimal resistance path between two local optima. It is therefore more realistic

to allow all other degrees of freedom to freely readjust while the "mutated" torsion is forced towards its newly imposed value. This is the classical principle of flexible "torsion angle driving" [9] in molecular mechanics. Its use in the context of a GA-driven approach as a source of high-fitness "mutants" is however original.

The central topic of this paper is thus the search of the best ways to combine or "hybridize" a GA-based approach with other optimization heuristics, in order to obtain a tool capable of efficient exploration of rugged energy landscapes of molecules. Conformational sampling has herein been used as a problem generator [6] for studying the behavior of the GA. The choice of the optimal modus operandi of this hybrid GA is not trivial, as all the previously introduced hybridization-related issues require some tuning, in addition to the choice of "classical" GA parameters (population size, mutation frequency, parallelization controls, chromosome migration frequency, etc.). As the tunable parameter space is vast, a meta-genetic algorithm (μGA) was used to explore it, in search of the optimal parameterization of the conformational sampling procedure. The "conformational sampling genetic algorithm" (CSGA), operates as a multimodal optimizer in torsion angle space, and its measure of "success" serves as fitness function for the μGA, mining the CSGA parameter space for optimal operational setups of the CSGA (Fig. 1).

The remainder of this paper is organized as follows: the first part of the Methods section depicts the implementation of the CSGA with a precise description of each parameter and each hybridizing heuristic, as well as the sampling success criterion used as "meta-Fitness" score by the μGA. The second part presents the setups of computational experiments aimed at assessing the specific impact of the key heuristics embedded in the CSGA, followed by Results, Discussions and Conclusions.

2 Methods

2.1 Description of the conformational sampling genetic algorithm

2.1.1 Data encoding

A chromosome encodes the list of the torsional angles of the molecule in degrees (as integers between 0 and 359). Torsional axis detection is automatic. Each torsional angle i (e.g. chromosome locus i) is assigned a weighing factor coding the expected impact of the rotation around that axis on the molecular conformation. Weighing factors w_i are thus chosen to linearly increase with the size of the moving fragment (for efficiency, the *smaller* end of each rotatable bond is submitted to a rotation procedure around the bond axis). They reach a maximum of 1.0 for all torsional axes coupled to fragments of size 50 atoms or more. In order to allow the sampling of cyclic conformers, the user needs to specify a ring edge to be formally "broken", allowing its ends to move away from each other upon rotation around other axes of the ring. Otherwise, a ring will appear as a rigid body to the torsion detection routine. Intracyclic torsional axes are assigned a weight of 1.0, since they control the proper closure of ring systems.

A chromosome will be "expressed" by a geometry buildup routine: using a "template" that can be any molecular geometry with correct bond length and valence angle values, the routine will, in turn, rotate the fragments around each axis i by an amount needed to set the corresponding torsional angle to the value θ_i at the locus i of the chromosome. This generates a set of $3N_{atoms}$ Cartesian coordinates completely characterizing the molecular fold (conformer) coded by a given chromosome.

The fitness of the individuals is defined as the opposite of the intramolecular energy E_{tot}: low-energy conformers are fittest. Energy is computed according to the consistent valence force field (CVFF) [16], completed with an implicit solvent effect term [21], as a sum of interatomic contributions that depend on the geometry returned at the "chromosome expression" step. The energy expression is detailed in Eqs. (1), (2), (3), (4) and (5), while graphically depicts the internal coordinates that correspond to each of the bond stretching $V_{bond}(l)$, angle bending $V_{ang}(\phi)$, torsional $V_{tors}(\theta)$ and non-bonded potentials $V_{nb}(d)$ (see Fig. 2). The internal coordinate values labeled by a "0" superscript stand for chemical context-dependent *parameters* (chosen in function of the nature of the atoms of each bond b, angle a or torsion t) and represent the "nominal" bond lengths, valence angle values, etc. Except for the point charges Q_i of the atoms i, intervening in the Coulomb and desolvation energies, the remaining variables are force field parameters controlling the intensity of the modeled interactions, most of them being dependent on the natures of the involved atoms. They will not be detailed here. The functional form in $1/d^2$ of the Coulomb potential is due to assuming a linear increase of the dielectric constant in function of the distance between the involved atoms.

$$E_{tot} = \sum_{\text{bonds } b} V_b(l_b) + \sum_{\text{angles } a} V_a(\phi_a) + \sum_{\text{torsions } t} V_t(\phi_t) + \sum_{\text{non-bonded atom pairs } i,j} V_{nb}(d_{ij}) \quad (1)$$

$$V_b(l_b) = K_b\left(l_b - l_b^0\right)^2 \quad (2)$$

$$V_a(\phi_a) = K_a\left(\phi_a - \phi_a^0\right)^2 \quad (3)$$

$$V_t(\theta_t) = K_t\left[1 - \cos\ n_t\left(\theta_t - \theta_t^0\right)\right] \quad (4)$$

$$V_{nb}(d_{ij}) = \frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} + K_{\text{Coulomb}}\frac{Q_i Q_j}{d_{ij}^2} + K_{\text{Desolv}}\frac{Q_i^2 + Q_j^2}{d_{ij}^4} \quad (5)$$

In torsion angle space, bond lengths and valence angles are constant and need not to be calculated except for user-defined bonds in cyclic systems, which need to be declared as "broken" in order to allow independent rotations of the intracyclic torsional axes. For these bonds, the harmonic V_b

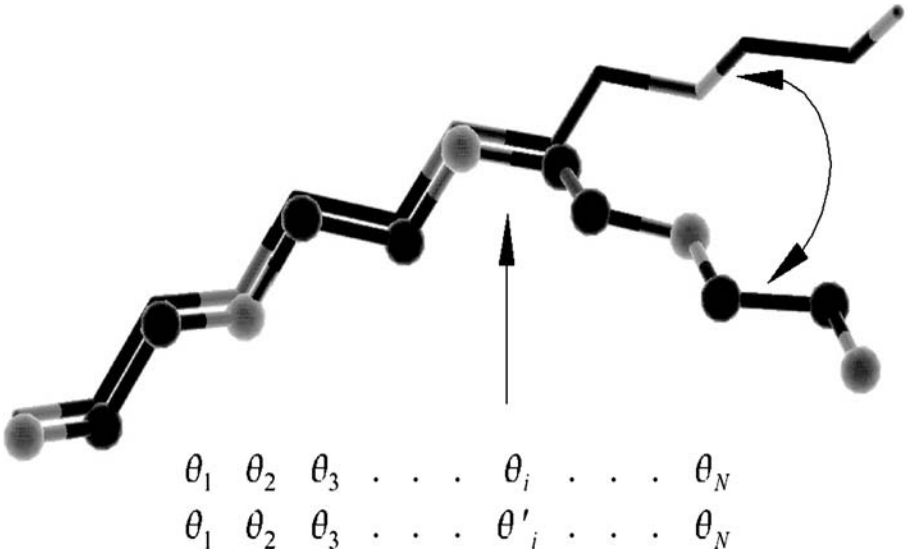


Fig. 1 Coding of the molecular structure as "chromosomes" in a GA: each chromosome locus contains a torsion angle value associated to a rotatable bond in the structure. The two structures correspond to two chromosomes differing with respect to a single locus i, which means that the corresponding molecular fragment is offset by a rotation of $|\theta_i - \theta'_i|$ around the pointed torsional axis.

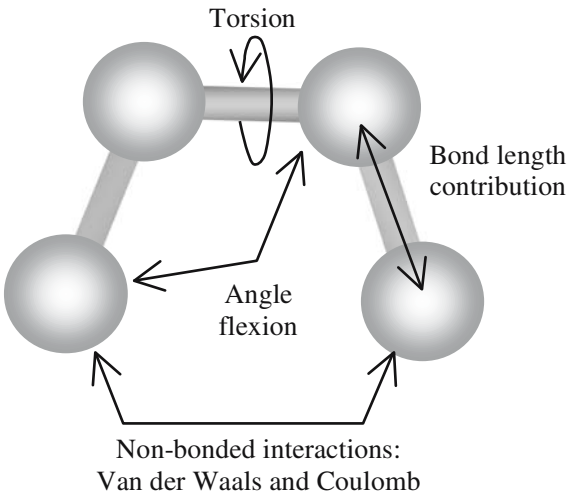


Fig. 2 Different types of energy contributions involved in the overall Hamiltonian of a conformer

terms as well as the V_a contributions of all the valence angles involving such bonds, must be included in the energy calculation in order to ensure that the ring will be closed such that the "loose" ends are set at a the expected distance l^0.

While the number of covalent bonds in a molecule scales linearly with size, the number of non-bonded atom pairs scales as $O(N^2)$. These interactions absorb most of the computer effort in energy evaluation. However, contributions of remote atom pairs are typically neglected: in the present work V_{nb} are explicitly estimated only if $d_{ij} < 10$ Å.

2.1.2 Population initialization

A GA starts from a random population of chromosomes, where the values assigned to each locus are drawn, according to a flat probability rule, out of the associated pool of options. Here, this "flat strategy" would amount to initialize each locus (torsion) with a random value between 0° and 359°. However, chemists know that torsional angles often adopt instances minimizing the *local* strain between the atoms directly bound to both ends of the torsional axis (that stereochemists call "staggered" conformers [23]). Of course, local strain is acceptable if it serves to relax the global tensions in the molecule. However, in practice – except for tensioned rings – strong local strain is rarely the price to pay in order to reach global stabilization. In the modeling community, rotamer libraries [33] are often used to cut down the size of search space by letting torsional angles only adopt values that were experimentally encountered in related compounds.

The herein introduced "local strain" biasing strategy uses the *calculated* local strain energy $-E^{\mathrm{loc}}(\theta_i)$, the sum of interactions between vicinal atoms directly bond to the torsion axis heads, to evaluate, at an empirical temperature T, the Boltzmann factor $\exp[-E^{\mathrm{loc}}(\theta_i)/k_{\mathrm{B}}T]$. If the molecular

Hamiltonian would consist of a simple sum of these local contributions, then the probability distribution of each torsional angle would be simply proportional to the corresponding Boltzmann factor. Using the Boltzmann distribution per se is not a good idea, because it might totally block higher local energy configurations from being drawn. Therefore, the following expression is used to calculate the "local strain" probability $p^{\mathrm{loc}}(\theta_i)$ of setting torsion i to a value θ_i:

$$p^{\mathrm{loc}}(\theta_i) = \frac{1 + N_{\mathrm{bias}} \exp\left[-E^{\mathrm{loc}}(\theta_i)/k_{\mathrm{B}}T\right]}{\sum\limits_{\text{all states } i} \left\{1 + N_{\mathrm{bias}} \exp\left[-E^{\mathrm{loc}}(\theta_i)/k_{\mathrm{B}}T\right]\right\}} \quad (6)$$

N_{bias} is a variable allowed to randomly change within the range (3,10) whenever the pace of progress towards fitter solutions decreases (see the "control" paragraph). When initializing a chromosome, there will be a three to tenfold increase in probability to "draw" a torsion angle value corresponding to minimal local strain than one causing strong local clashes.

An alternative strategy investigated here will be further on referred to as the "tradition-based" biasing strategy, relying on the analysis of the pool of conformers already generated at a given moment of the sampling process, in order to extract the torsion angle values that are preferentially adopted in the fittest solutions currently available. Assuming that, at a given moment of the sampling process, $j = 1, \ldots, N_{\mathrm{vis}}$ previously visited chromosomes χ_i^j of energies E^j are available. The "tradition-based" probabilities $p^{\mathrm{trad}}(\theta_i)$ of setting the torsion i at θ_i are related to the sum of the Boltzmann factors of all the previously generated conformers in which θ_i has been seen to occur:

$$p^{\mathrm{trad}}(\theta_i) = \frac{\sum_{j=1}^{N_{\mathrm{vis}}} \delta\left(\chi_i^j = \theta_i\right) \exp\left(-E^j/k_{\mathrm{B}}T\right)}{\sum_{j=1}^{N_{\mathrm{vis}}} \exp\left(-E^j/k_{\mathrm{B}}T\right)} \quad (7)$$

where the δ function in Eq. (7) returns 1 when its Boolean argument is true and 0 otherwise. Because of this risk of premature discarding of large zones of the problem space (torsional values not appearing in either of the most stable conformations will *never* be drawn), the strategy was always used in conjunction with the "local strain" technique and only within one of the parallel runs (islands; see Sect. 2.1.3 below). Obviously, initial CSGA runs that cannot benefit from the knowledge of any previously sampled conformers may not apply this strategy.

2.1.3 Population

Both the population size N_{pop} and the number N_{isl} of parallel runs (islands) to be launched are customizable parameters of a simulation. Currently, the initial population is formed by the N_{pop} fittest chromosomes out of a pool of 10^4 randomly generated individuals, according to the torsion probability distribution in use. It is worth noting that the current approach also supports the "seeding" of the initial, random population with chromosomes obtained from previous runs (details follow in Sect. 2.1.9).

Occasional migrations [35,40] of the momentarily fittest individuals are allowed, with a parameter N_{mig} controlling migration frequency. In CSGA, an island exports its fittest individual if the following conditions are simultaneously fulfilled:

- The fitness of this individual is strictly superior to the largest between the one of the previously exported "emigrant" and the one of the here so far best imported "immigrant". This directive ensures that an individual will be exported only once, thus avoiding the spread of multiple redundant copies of a same chromosome throughout various islands.
- At least N_{mig} generations have passed since the latest emigration event from this island.
- There is at least one of the active parallel runs for which there is no immigrant awaiting to be accepted (stored in a temporary file, an emigrant is waiting to be read by the run it has been addressed to, after which its file is deleted and the run gets ready to accept another).

Immigrant input in a CSGA run is immediately followed by reproduction, so that imported chromosomes that are unfit with respect to the host population and would not make it through the selection process have one chance to participate in crossovers with "indigenous" chromosomes.

2.1.4 Reproduction

This algorithm uses both crossovers and mutations in order to generate offspring. First, the N_{pop} members of a current population are regrouped into $N_{\mathrm{pairs}} \leq N_{\mathrm{pop}}/2$ parent couples. The fittest "free" individual (not yet assigned to a couple) randomly "picks" a partner out of the remaining unpaired chromosomes. Its "choice" may be rejected if the partner chromosome fails to display significant differences with respect to at least two loci coding important torsional angles (with assigned weights above 0.8). In case of rejection, a maximum of 20 other random picks are allowed until a valid couple is formed. Otherwise, the individual is discarded from sexual reproduction. Only a parameterizable crossover rate f_{mate} of the valid N_{pairs} couples are actually allowed to generate offspring. Crossovers are generated by randomly picking, for each couple, one out of the eligible crossover loci ensuring that offspring will be different form either of the parents. The decision to apply one- or two-point crossovers is random and the options are equiprobable. The tunable mutation rate f_{mut} controls the frequency of one-point mutations implying a random change of a single torsion value, according to the probability distributions currently in use for the selected torsion.

2.1.5 Selection mechanism

The extended population following the reproduction step is filtered according to two alternative selection mechanisms

- The default procedure sorts all individuals by decreasing fitness. Starting with the fittest, similarity filtering sets

the next individual of the set as a reference. Less fit conformers are discarded if they are "too similar" (according to a geometric fingerprint-based similarity score [22], not detailed here) to the reference (similarity score $> \sigma_{\max}$, an adaptive similarity threshold value). This feature simulates the process of "food sharing" [35]. The first N_{pop} non-redundant conformers kept by the procedure will form the next generation. If less than N_{pop} pass the similarity filtering, random chromosomes will be added. In this scenario, both parents and their children may pass to the next generation *if* they are dissimilar enough and fit enough.

- The "child-against-parent" competition specifically replaces the parents by their offspring if the fittest child outperforms the fittest parent. Similarity filtering proceeds as outlined before. As either children *or* parents make it into the next generation, this procedure favors solution diversity and slows down convergence. It is invoked instead of the default selection, once every (tunable) $N_{\text{c-p}}$ generations.

Since the interdiction of coexistence of related chromosomes may significantly slow down convergence, $\sigma_{\max}$ is steadily adapted to the current status of the population. In the beginning (random population), $\sigma_{\max}$ is set to a tunable, user-defined similarity control $S_{\max}$. As long as evolution proceeds at a reasonable pace (in the sense that the best-so-far energy is seen to decrease at least once every k generations), $\sigma_{\max}$ is kept at its current level. If, however, evolution appears to stall, the tolerated similarity is gradually increased, which may in turn relaunch the finding of fitter solutions. The number k of generations used to control the requested pace of evolution has been related to the parameter N_{nonew} controlling the overall tolerance of the process with respect to stalling evolution, as described further in Sect. 2.1.8 : $k = N_{\text{nonew}}/3$.

2.1.6 Tabu mechanism

A CSGA run maintains a "tabu list" featuring the chromosomes sampled by previous runs, and continuously updated with new ones generated by the run itself, as described in Sect. 2.1.8. Prior to fitness evaluation, the tabu list is checked for entries matching the current chromosome, if none of the important torsions (with weights above 0.9) differ by more than $\Delta_{\min}$ (tunable) degrees. If so, the procedure assigns an arbitrarily high energy to this redundant chromosome, triggering its demise.

2.1.7 Hybridization with deterministic optimization heuristics: Lamarckism and Directed Mutations (Explorers)

As already mentioned in Introduction section, two well-known problems encountered in force field-based molecular simulations were specifically addressed by adding the following heuristics to the GA engine:

- Lamarckism [27]: Whenever crossovers or mutations generate a new "best-so-far" chromosome, this may be further submitted, with a tunable probability p_{L}, to a conjugated gradient optimization in torsional angle space. The torsion values at the found local minimum replace (after folding back to the range [0, 359] and rounding to the closest integer) the ancient contents of the chromosome.
- Directed mutations ("Explorers"): An important constraint term $K(\theta_i - \theta_{\text{target}})^2$ is added to the molecular energy function, forcing the driven torsion $\theta$to evolve towards θ_{target}. A conjugated gradient optimization of this modified potential allows all the other degrees of freedom $j \neq i$ to find the optimal arrangement compatible with the constraint $\theta_i = \theta_{\text{target}}$. Once this point is found, the constraint term is removed and the structure reoptimized. If θ_{target} is very different from the ancient value of that torsion, it is unlikely that reoptimization will move back to the initial geometry. This approach is therefore a source of diversity, like random mutations, but the resulting conformers are much more likely to pass selection due to their low energy. However, the procedure is quite time consuming and would cause serious disruption of the evolutionary loop if run within the islands of the CSGA. Therefore, it has been programmed under the form of stand-alone "explorer" processes, that are started by a CSGA run, provided that no other such explorer is already running (there may be at most one "explorer" for N_{isl} CSGA islands at any time). The explorer process is provided with the chromosome of the momentarily fittest individual and a torsion to be driven, randomly picked within the list of important torsions (weight > 0.9). It proceeds in four cycles, "pushing" the driven torsion away from its initial value by 45°, 90°, 135° and 180°. At the end of each cycle, the resulting individual is transferred to any of the active CSGA islands by means of the migration mechanism.

2.1.8 Population management and convergence control

An "aging" parameter $A_{\max}$ specifies the maximal number of generations for which a chromosome may be kept in a population, to be thereafter replaced by a random chromosome (see aged genetic algorithm [25]). The progress of evolution is monitored in terms of decreasing energies of the top five ranked individuals. If evolution stagnates for a too long time (no fitness improvement among the top five during a parameterizable N_{nonew} generations), the whole population is removed and replaced by random chromosomes, while the fittest member of the population is added to the "tabu" list (see Sect 2.1.6) in order to avoid its rediscovery. In case of such a population reset, the adaptive similarity threshold $\sigma_{\max}$ is once again set to its extreme value $S_{\max}$. A parameterizable number N_{elit} of fittest individuals are preserved from deletion and aging (see elitism [40]). In the current implementation, N_{elit} may be either 0 or 1. However, these "immortal" individuals are always subjected to the "child-against-parent" selection rule: their direct offspring may not coexist with them in a same population, in order to avoid a premature convergence.

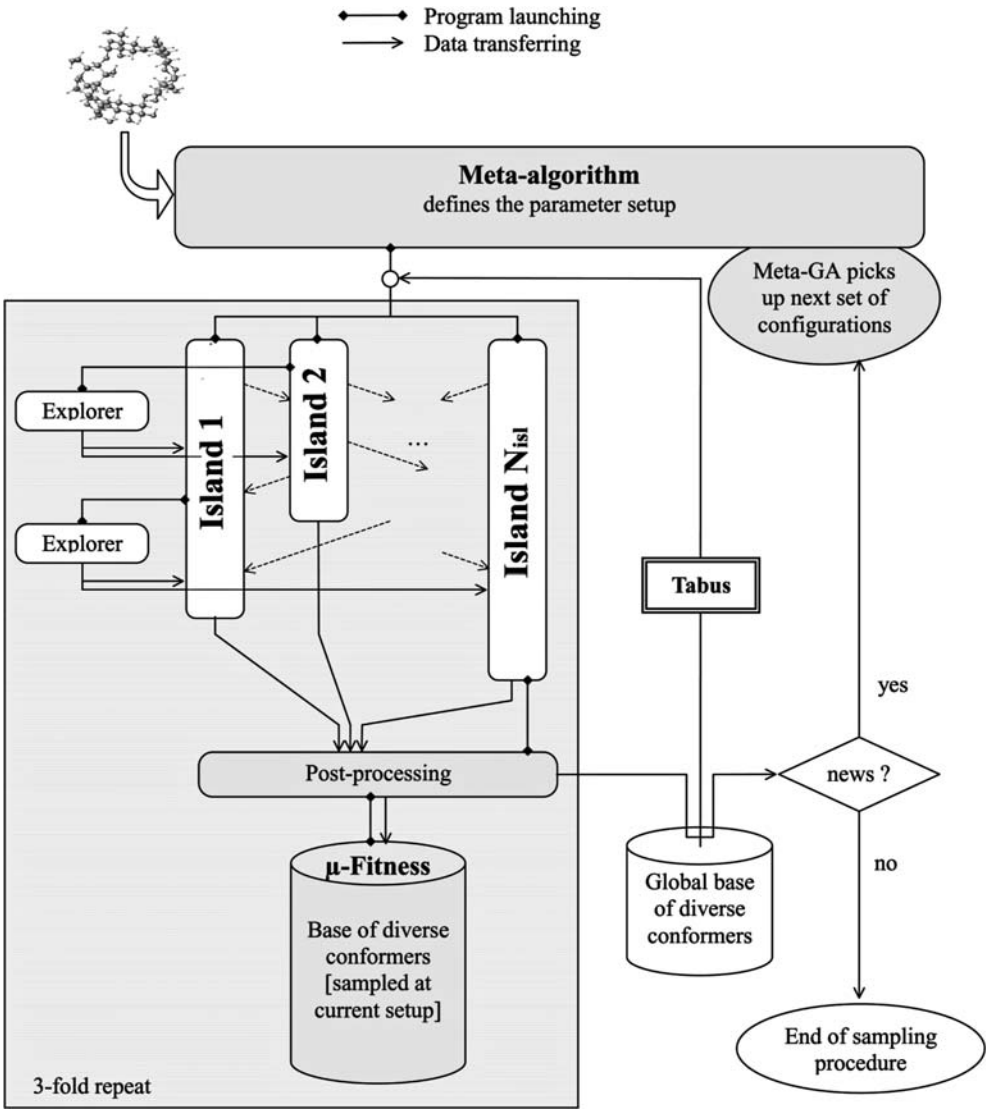


Fig. 3 Global conformational sampling scheme, featuring the triplicate CSGA runs embedded into the meta-optimization loop

Finally, the global ending condition for each island is double: either

- the total number of generations exceeds a global limit N_{gen}, or
- the best energy reached so far did not, in spite of several population reset attempts, progress by more than 0.5 kcal during the last N_{wait} generations.

In the current implementation, N_{gen} has been set to a very high value of 10^5 generations, so that the tunable N_{wait} parameter is actually controlling the ending of runs.

2.1.9 Triplicate runs: increasing the reproducibility of the CSGA

Given the stochastic nature of GAs, the final outcome of a sampling process (at given tunable parameter values) may strongly differ from run to run. In order to enhance reproducibility, runs are repeated thrice before proceeding with the analysis of the set of found conformers (Fig. 3). In this "block" of three successive runs, each run inherits "tabus" and "tradition" from the pool of previously sampled diverse solutions. After completion, the newly sampled chromosomes are post-processed, e.g. merged with the old set and subjected to diversity filtering. A same similarity threshold $S_{\max} = 0.8$ is used in post-process filtering, no matter what current value had been employed *during* the runs (two solution pools issued from differently parameterized runs may therefore be directly compared).

While tabu searching is expected to increase solution diversity, the steady increase of forbidden areas in the problem space may eventually impede on the convergence of the procedure. Therefore, the third run in the series "seeds" its initial population with the best chromosomes found by the two predecessors, and allows their further evolution in a tabu-*free* environment ($\Delta_{\min}$ is set to 0, overriding user choice). As this run is meant to ensure a complete

Table 1 Operational parameters of the CSGA and the pool of possible values defining the problem space of the μGA

Parameter	Possible values	Description
N_{isl}	2, 3, 4	Number of 'islands' (parallel runs)
N_{mig}	5, 10, 25, 50	Migration period
N_{gen}	99999	Maximum number of generations to go (constant)
N_{wait}	500, 800, 1000	Number of successive generations of stalled evolution triggering termination of the run
N_{nonew}	50, 75, 100	Number of generations without progress triggering population reset
N_{pop}	50, 100, 150, 200	Population size
N_{elit}	0, 1	Number of fittest individuals exempted from aging and population reset
A_{max}	10, 10^2, 10^3, 10^4	Maximum age of individuals (generations)
f_{mut}	1, 10%	Mutation rate
f_{mate}	40, 70, 100%	Crossover rate
$N_{\mathrm{c-p}}$	1, 2, 5, 10	"Child-against-parent" selection frequency (once every $N_{\mathrm{c-p}}$ generations)
S_{max}	75, 80, 85, 90%	Maximum similarity allowed throughout the population
Δ_{min}	20, 30, 40, 50, 60	Tabu avoidance threshold.
p_{L}	0.1, 0.3, 0.5	Probability of submitting a new "best-so-far" individual to "Lamarckian" conjugated gradient optimization

optimization of potentially suboptimal chromosomes, a strict termination criterion of $N_{\mathrm{wait}} = 2,000$ is set to override the user choice for this parameter.

Each island is a running copy of the CSGA executable in a dedicated directory, complied and executed on a Silicon Graphics 4-processor R12K at 360 MHz under IRIX 6.5. The CSGA and Explorer codes have been written in FORTRAN 77. A migrations directory serves as temporary storage for exchanged chromosome files, which are deleted after lecture by the target island. A layer of tcsh scripts is in charge of starting the runs after creating the execution directories. At termination, each CSGA island fires off a child post-processing script, which will die if other islands are still active. The child of the last active island will eventually proceed with the analysis, merging and diversity-filtering of the solutions files storing the chromosomes visited by each island. Then, the next triplicate run will be launched, or, if this had been the last of the three, control is passed back to the μGA loop.

2.2 Optimization of the tunable parameters of the CSGA: the Meta-GA Loop

GAs are known [15] to be very sensitive to the choice of their control parameters (Table 1). The best parameter setup could in principle be derived on hand of a purely analytical description of the GA (using Markov chains, or infinite population models) [7,31] and experimental analysis of its behavior [36]. This is however unlikely to succeed, given the complexity of the herein reported approach. The other option is to tackle this meta-optimization problem with appropriated methods for maximization of a noise-affected objective function, the "success score" of the CSGA run. Such methods may include auto-adaptation, fuzzy learning [18], or GAs [15]. The latter option, a μGA used to maximize the performance of the CSGA multimodal optimization tool has been adopted here. The success score of the CSGA in function of its operational parameters (the "meta-fitness" function μF) needs to embody three key aspects:

- The first one is the multimodal aspect of the task of the CSGA: finding as many as possible of the relevant minima of the energy landscape. The quality of a CSGA simulation thus cannot be measured by the classical '*best-so-far*' index [28,34]
- In order to reduce stochasticity, μF will be evaluated on hand of the conformer ensemble produced by a Triplicate run.
- Eventually, μF is also a matter of computer time: out of two CSGA runs yielding conformer samples of a same quality, the faster should be preferred.

The above demands are met by Eq. (8), which is a linear combination of the free energy $-k_{\mathrm{B}}T\ln(Z)$ of the set of n diverse conformers of energies E_i obtained by the current triplicate run and an empirical time penalty factor. The partition function Z of the conformer family is the sum of the conformer Boltzmann factors, at $T = 300\,\mathrm{K}$ and $k_{\mathrm{B}} = 2\,\mathrm{cal/(mol\,K)}$, with energies in kcal/mol.

$$\mu\mathrm{F} = -\left[-k_{\mathrm{B}}T\sum_{i=1}^{n}\exp\left(-\frac{E_i}{k_{\mathrm{B}}T}\right) - \alpha \times \mathrm{CPUtime}\right]. \quad (8)$$

The CPU time above is taken as the sum of run times of each processor, divided by $\sqrt{N_{\mathrm{isl}}}$ in order to favor setups with higher levels of parallelization. The mixing factor $\alpha = 1.4 \times 10^{-4}$ implies that a run that consumes two more "effective" hours is favored in terms of μF only if it succeeds to decrease the free energy of the conformer family by more than 1 kcal/mol.

Given the importance of the computer effort required for a single evaluation of μF (hours–days), meta-optimization is limited in terms of the total number of parameter configurations that can be explored. A basic μGA methodology has been used: starting from a set of ten random "meta-chromosomes" (complete sets of operational parameters), ten new individuals are generated, issued, in 15% of the cases, from single point mutations, and from cross-overs for the remaining 85% (here, cross-overs add a single "child" to the population, issued from two randomly selected parents). A history file of already visited parameterization schemes is kept, in order to ensure a self-avoiding walk. Selection is solely based on the μF score. The meta-optimization software consists of a series of tcsh (UNIX shell) scripts relying

on awk (pattern processing tool under UNIX) programs for the management of the parameter chromosomes.

2.3 The global conformational sampling scheme

Figure 3 shows the overall conformational sampling strategy, including the μGA-layer that fires off triplicate runs over the network, using the steadily evolving parameter sets coded by meta-chromosomes. The pool of conformers issued by a triplicate run is used to estimate the μF of the current operational parameter set, before being merged with the global conformer depository containing all the diverse ($S_{\max} = 0.8$) conformers within +20 kcal/mol of excess with respect to the global best-so-far energy. If four successive triplicate runs fail to add any new members to the global depository, the conformational sampling procedure of the molecule terminates. In order to avoid confusion, in the following the term "simulation" will be used to refer to the whole μGA-driven sampling scheme as described here.

2.4 Assessing the impact of the described strategies on the conformational sampling results

A rapid evaluation of the impact of meta-optimization has been done on hand of several small organic molecules, which were alternatively subjected to (a) ten different (triplicate) CSGA runs with randomly chosen operational parameters, then (b) subjected to the global μGA-driven simulations as outlined above and (c) resubmitted to ten triplicate CSGA runs using the top ten operational parameter setups found by the meta-optimizer. Individual CSGA runs performed at steps (a) and (c) were "ab initio" runs and were not provided with any information concerning previously sampled conformers, in order to ensure that their performances are comparable.

In order to understand the impact of the original strategies introduced here, a benchmark problem has been comparatively submitted to various CSGA versions, alternatively enabling and disabling each strategy under study. The chosen system was cyclodextrine (Fig. 4), a macrocyclic sugar composed of six glucose rings. All the rings were opened to sampling, which leads to a problem with 65 degrees of freedom. The algorithm needs to properly close each six-membered ring and the macrocycle formed by the latter.

The following series of simulations were performed (using a same random set of ten parameter sets as initial meta-population):

- "Default" simulations: the global sampling scheme (all strategies enabled).
- "No Tabu" simulations: the "tabu" strategy has been switched off.
- "No Explorer" simulations: "Explorer" processes were disabled.
- "No Tradition" simulations: disallow tradition-based bias (use only the "local strain" strategy to initialize random chromosomes).
- "Flat distribution" simulations: uses a flat probability density.

Four independent "default" simulations and three of each of the above noted variants have been performed.

2.5 Bayesian analysis of the choice of parameters on the performances of the CSGA

Bayesian learning [41] has been employed in order to discriminate, in the space of operational parameters, between the "good" and the "bad" CSGA runs. By estimating the probability of obtaining a "good" or a "bad" result upon setting a given parameter to a specified value, this approach provides a first estimation of the role of each CSGA control. The "Learn Good from Bad" toolbox of the Pipeline Pilot software [30] has been employed to mine for correlations between operational parameter values and the μF. For each strategy, the typically 90–120 parameter meta-chromosomes visited during the repeated simulations were sorted with respect to their μF, with the top 10% being considered "good" and the remaining "bad". A similar analysis has been conducted for the entire set of visited parameter chromosomes, all strategies confounded.

3 Results and discussion

It has been shown [17] that a combinatorial optimization problem over a broad class of functions is NP-hard. For the class of deterministic functions $f : \{0, 1\}^L \rightarrow Z$, that can be computed in polynomial time, the problem to know whether there exists a point p such that $f(p) < \lambda$ (at given λ) is NP-complete. The conclusion of this study is that the theoretical or experimental analysis of GA behavior cannot be performed regardless to the type of functions being optimized. Figure 5 illustrates the importance of searching for appropriate operational CSGA parameters. For each of the ten triplicate CSGA runs with random parameters (right side boxes) and the runs using the best ten setups visited by the μGA (left side boxes, respectively), free energies of the conformer sets issued from each run in the triplicates were calculated. The plots report the averages and variances of free energies over each triplicate CSGA run and clearly show that triplicates realized with randomly chosen setups may encounter serious difficulties with respect to both convergence and reproducibility. The tuning of CSGA setup is therefore of paramount importance and a GA is a well-suited tool for meta-optimization. Although other approaches, such as experimental design, might be well suited for such a task, the complexity of the problem is prohibiting an in-depth search for the best-suited meta-optimization tool.

Further results presented in this work are therefore restricted to the peculiar problem of the closure of the cyclodextrine ring system. This is a difficult problem for classical conformational sampling techniques such as molecular dynamics [20] because of the steepness of the potential wells due to the covalent ring closure constraints. Acyclic compounds, with extended low energy wells covering large phase space zones, allow for an easy discovery of many low-energy geometries, while raising a challenge of different nature: the

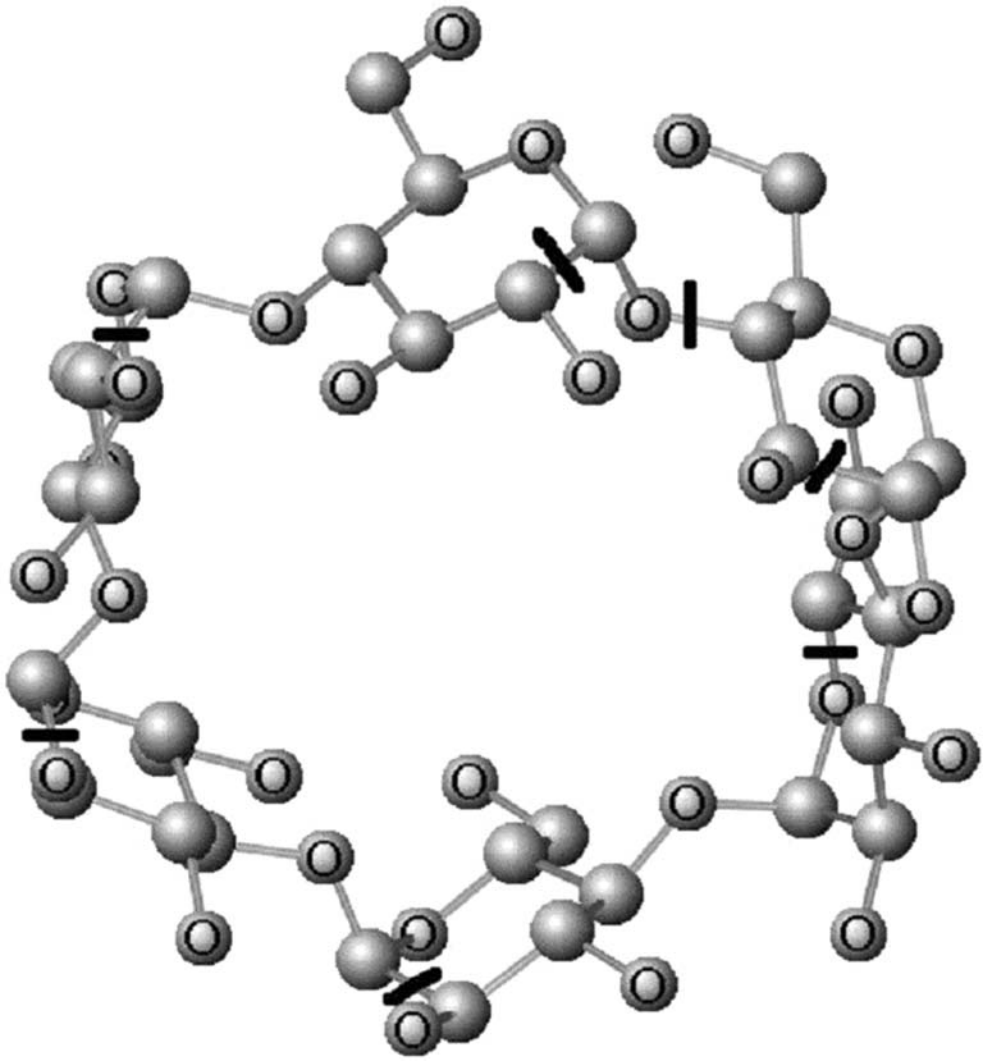


Fig. 4 The cyclodextrine molecule, shown without hydrogens. *Dashes mark* the bonds that were "broken" in order to open the ring systems for sampling

slightly deeper energy wells that are actually populated at room temperature may never be discovered within a reasonable simulation time. This work offers no insight about what the optimal parameter set for the sampling of such molecules may look like. Due to the expectedly huge number of low-energy conformers, the simulation of an acyclic compound similar in size to cyclodextrine would have taken much longer to complete and would have therefore been a poor benchmark problem.

3.1 General discussion of the success of the different strategies

In spite of repeated runs, results are affected by important fluctuations: A first observation based on Fig. 6, displaying the lowest energy levels versus the number of relevant minima obtained by each simulation, is the heavily stochastic nature of the results. The four different "default" simulations converged, in spite of triplicate repeats, to significantly different energy levels. The best minimum found by the less successful simulation is at +6 kcal/mol from the global best of this strategy. Moreover, two of the default simulations finished after having visited only four different local minima, while the two others managed to find 14 and 20, respectively. This is a consequence of the meta-optimization termination condition (four successive CSGA runs failing to enrich the pool of solutions with new, relevant visited minima). The probability of encountering such an "unlucky" series of "unproductive" CSGA simulations at early stages of meta-optimization appears to be intolerably high with the "default" strategy.

The best-found minima actually correspond to the experimentally determined structure of cyclodextrine. Each six-membered ring has been set in the proper "chair" conformation, and a strain-free closure of the macrocycle

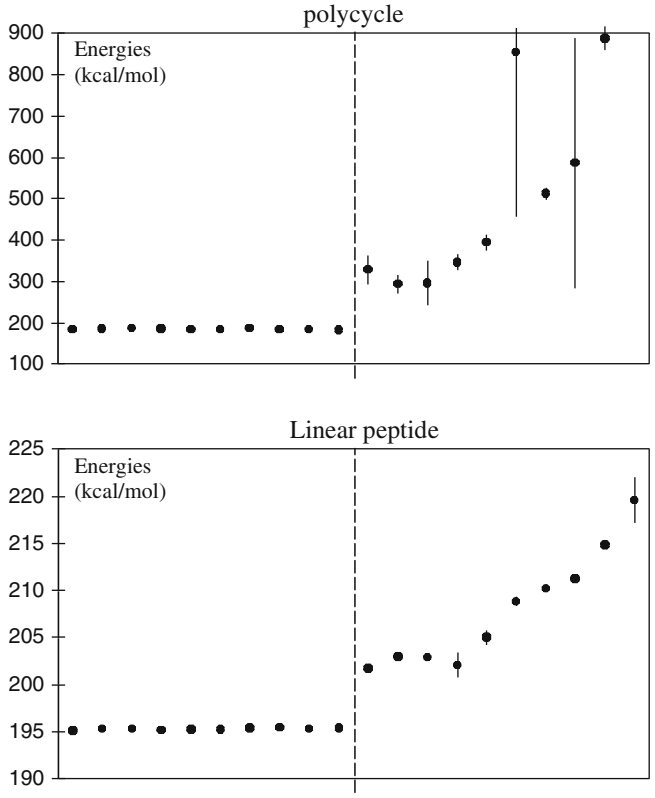


Fig. 5 Averages and variances of free energies for triplicate of CSGA runs with both a polycyclic molecule and a small linear peptide. The *right side boxes* are obtained with random parameterization whereas *left side boxes* show the same results with the ten best setups encountered so far. It can be seen from this that both convergence and reproducibility can be improved by the parameter choices

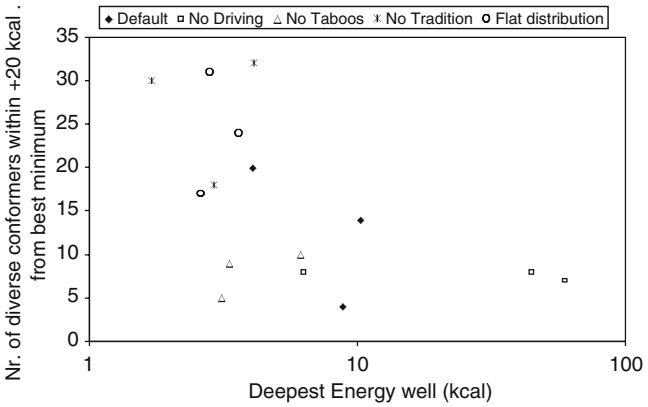


Fig. 6 Plot of the lowest energies reached by the different simulation strategies with respect to the number of found diverse minima

has been realized. The best minima found by each strategy all actually feature the correct ring geometry, they (and their energies) differ only because of different arrangements of the rotatable –OH and $–CH_2OH$ groups that "ornate" the ring system (and for which no experimental determination of their exact position is possible, since they are rapidly spinning in a molecule at room temperature).

"Explorers" are essential for effective conformational sampling: In absence of this directed mutation strategy, two of three simulations (squares in Fig. 6) failed to reach the bottom of the energy well by several tens of kilocalories per mole. Also, the total numbers of visited optima is limited in all three "No Explorer" runs. Directed mutations are therefore beneficial both in terms of energy decrease and population diversity increase. The implementation of torsional angle driving as an "intelligent" mutation strategy within a GA appears to be very useful. Its principle, a constraint-driven deterministic optimization of the objective function,

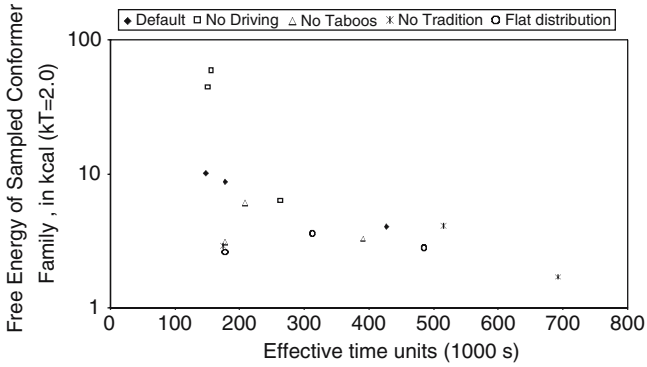


Fig. 7 Dependence of the quality of sampling (expressed as a free energy $-k_{\mathrm{B}}T \ln Z$) with respect to the total computer effort required by the strategy

may be generally applicable to other classes of problems outside the field of molecular modeling. In the current software, the effort-sharing between parent GA and child "explorer" processes is roughly controlled by the number of islands. As a single "explorer" may run at a time, the more GA islands are active, the (relatively) less computer effort is allocated for exploration. A search for more flexible management schemes of explorer processes has therefore been envisaged.

Setting of tabus increases population diversity, but slows down convergence: The three "No Tabus" simulations (plotted with triangles in Fig. 6) can be seen to lead to populations of few, but quite fit solutions. This is expectable within a fitness landscape with few sharp peaks. Simulations of flexible molecules with "flat" energy zones may need to be pursued for much longer until the risk of revisiting becomes tangible. The recurrent visiting of the same energy wells in the "No Tabus" strategies allowed for more chances to locally optimize the low-weight torsions controlling the arrangement of smaller molecular fragments. Tabus are imposed with respect to high-weight degrees of freedom controlling the overall molecular fold. For each fold, there are many possible arrangements of the side groups with respect to the central elements. There are however no guarantees that, between the first emergence of a fold and the adding of this fold to the tabu list, the algorithm had enough time to search through all these arrangements and find the optimal one (even though the third run of each triplicate is specifically dedicated to this purpose; see Methods). Once a tabu is set, it will effectively prohibit the algorithm to continue searching for better side group arrangements around a fold, since all conformations based on that fold are "forbidden". Therefore, the final conformer list in a tabu-based strategy may include geometries with suboptimal side group arrangements and higher energy.

The "tradition-based" strategy is the main trigger of premature convergence: Fig. 6 clearly shows that the two most successful strategies, returning a significant number of diverse minima and low energies, are the two approaches that do not rely on the torsion values in previously found solutions when defining the probability rules for the draw of random torsional angle values. Although only one of the N_{isl} islands applies "tradition-based" biasing, herein generated chromosomes are quite likely to be fitter than the ones of the other runs. The migration mechanism ensures their effective spread over the other islands, and the presence of "unnaturally" fit solutions at too early stages of evolution triggers long waiting times until the next improvement of the locally fittest individual, with the risk of premature fulfillment of stopping criteria. Tradition-based biasing may also clash with the tabu strategy: as the former encourages the reuse of previously seen torsional values, it implicitly increases the risk of regenerating tabu folds.

The herein performed simulations do not evidence any significant advantages of the "local strain-based" biasing strategy (depicted with stars) with respect to the "flat" strategy (circles). This is not surprising, since ring closure constraints, not taken into consideration by either of the biasing strategies, largely determine the torsional values that are allowed around intracyclic axes. Local strain-based biasing may still play a key role in modeling linear, flexible compounds.

The quality of the results of a simulation is roughly correlated with its total computer effort. As shown in Fig. 7, the free energies $-k_{\mathrm{B}}T \ln Z$ computed on hand of the final global set of diverse conformers generated by each simulation are roughly related to the sum of effective CPU times of all the triplicate runs performed within the simulation. Longer simulations tend to yield better results, applied strategies notwithstanding. With the notable exception of the two failed "No Explorer" simulations, the data points are slightly correlated ($R^2 = 0.31$). It can be concluded that none of the employed strategies has a direct impact on the rate at which the phase space of the problem is explored, nor on the expected number of generations needed to "discover" a fit solution, but rather control the risk of premature termination due to stagnation.

3.2 Statistical analysis of the operational parameters

Naive Bayesian learning is able to evidence loose dependencies between variables and observables even for noisy data sets, as is the case here. "Events" (e.g. a parameter p_i adopting a given value V_{ij} out of the $j = 1, \ldots, m_i$ eligible options) seen to occur within the subset of "good" examples with a frequency above the random expectation are considered to "favor" the obtaining of a good result (e.g. the value

adopted by the parameter was "correct"). Oppositely, values rarely seen to occur within the chromosomes of the top 10% best CSGA runs are "bad". The used software returns, for each event ($p_i = V_{ij}$) a positive or negative empirical "probability score" $P(p_i = V_{ij})$ stating how "correct" or how "wrong" the choice of V_{ij} has been for p_i. $P(p_i = V_{ij}) \approx 0$ means that setting p_i to V_{ij} neither improves nor decreases the chances of success of the CSGA.

It is important to note that the sample of data points $\mu F = \mu F(p_1, p_2, \ldots, p_i)$ submitted to the Bayesian analysis represent the output of an evolutionary program and are not randomly distributed in parameter phase space. Favorable phase space zones should be more densely populated, as the meta-optimization process selects offspring similar to the parents (unlike the CSGA, the μGA uses no dissimilarity enforcement). Convergence of the μGA towards a consensus zone in parameter space should trigger high probability scores associated to the corresponding parameter values. However, like in natural evolution, irrelevant features ("junk DNA") are also inherited, so that it cannot be excluded to see a fortuitous "pseudo-convergence" of irrelevant parameters towards a given value which gained the upper hand simply for been carried by a "winning" chromosome.

Also, the success of a triplicate CSGA run is, strictly speaking, not only a function of its operational parameters but also of the previously found solutions entering as tabus that block out whole conformational space regions and implicitly impact on the way in which the CSGA conducts the search for new optima. In other words, the μF landscape evolves as well during the meta-optimization process [18], which may further slow down the convergence of the optimal parameter search.

In spite of the potential bias of the above-cited phenomena on the observed parameter-μF correlations, many of the trends evidenced by the Bayesian analysis do make sense and will be discussed further on, after rescaling, within each of the comparative plots, the probability score of the most impacting event to ± 1.0.

Quick convergence of the meta-optimization process has been observed with the "No Explorers" and "No Tabus" strategies. Figure 8 locates the top 10% most successful CSGA runs of four different strategies, highlighted as triangles in the plane of the first two principal components (PC) [10] of the parameter space.

Within the "No Explorers" strategy, all successful runs are found in the vicinity of the x-axis (PC2 $\approx$ 0), with a marked cluster at the center of the plot, clearly evidencing a high degree of relatedness of the underlying operational parameter configurations. This is not surprising, as only one of the three simulations managed to find any low energy conformations: all the successful CSGA runs are indeed based on related parameter chromosomes issued from a same evolutionary process.

By contrast, the "No Tabus" successes represent runs from all the three simulations. The degree of interrelatedness of the underlying parameter configurations is less well marked than in the previous case, but nevertheless real: virtually all the points are grouped in the upper part of the plot (PC2 > 0). Different meta-runs of the "No Tabus" strategy convergently led to similar choices of operational parameters. The meta-optimization of the "No Tabus" CSGA appears to be the fastest to reproducibly converge. This may be related to the previously noted fact that the addition of tabus is actively modifying the μF landscape.

While the successes of the "Default" approach show some weak tendency towards higher PC1 values, the ones of the remaining strategies do not display any noticeable clustering behavior (as exemplified by the last of the four plots). It might therefore be concluded that the "No Explorers", "No Tabus" and to a lesser extent the "Default" strategies are more sensitive with respect to the parameter choice than the others. This conclusion is also supported by the fact that the latter strategies are also the ones for which the Bayesian learning tool consistently found quite strong correlations between parameter choices and success rate.

Bigger populations are a better guarantee of success, as can be seen from the Bayesian analysis of all parameter chromosomes, all strategies confounded, in Fig. 9. It is obvious to expect better sampling with larger populations; however, the required computer effort is seen to scale linearly with population size as well. Therefore, the choice of α in Eq. (8) eventually controls whether meta-evolution favors shorter, but less productive runs rather than longer ones, with better chances to find deeper energy wells.

The aging parameter $A_{\max}$ appears to play an important role within the "No Explorers" and "No Tabus" strategies only (Fig. 10). The former is the one with the most difficulties to converge and therefore tends to maintain the status-quo of the population rather than risking the insertion of new random and unfit chromosomes. Deleting chromosomes after ten generations is certainly a bad choice within this strategy. The apparent inappropriateness of the choice $A_{\max} = 1,000$ is puzzling. On the contrary, the "No Tabus" strategy would gain from often "refreshment" of chromosomes: low $A_{\max}$ values do indeed stand out as favorable.

A frequent use of Lamarckian optimization ($f_{\mathrm{L}} = 0.3$–0.5) is in general recommended, although this parameter plays a role only within the "No Explorers" and "No Tabus" strategies (Fig. 11). Lamarckian optimization is systematically used by the Explorer processes. When these are disabled, gradient-based optimization within the CSGA is expected to gain in importance, as the only source of fully optimized individuals. This is indeed being observed: success of the No Explorers protocol is significantly correlated with an often usage of the Lamarck optimizer. By contrast, extensive use of Lamarck optimization appears to be detrimental within the "No Tabus" strategy, probably because it favors revisiting minima (the deterministic optimizer acts as an attractor of diverse conformations towards a common local minimum).

Random mutations are being favored throughout all strategies: out of the two choices available for the random mutation frequency f_{mut}, 1 or 10%, the latter is being systematically preferred (plots not shown).

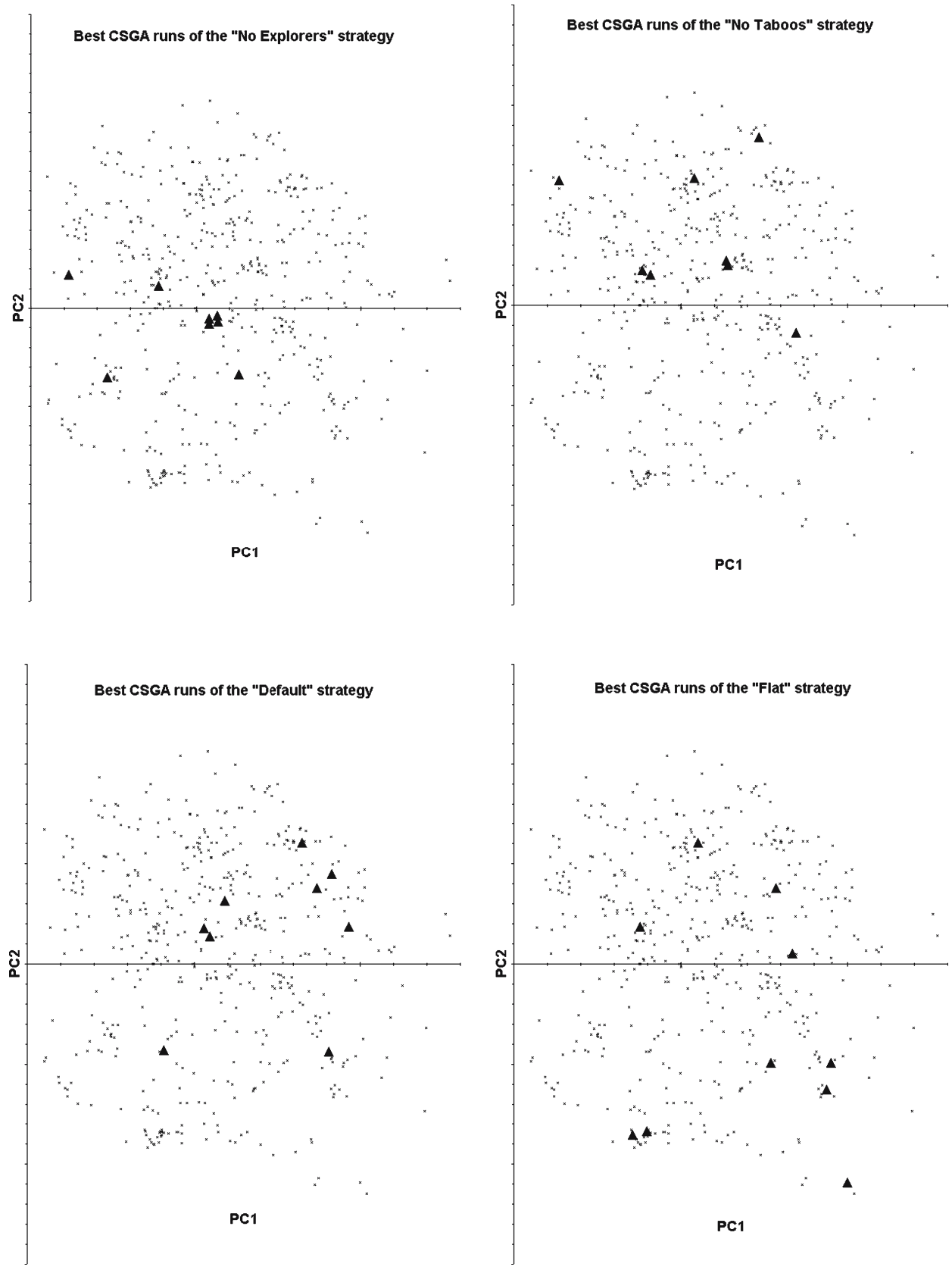


Fig. 8 Most successful CSGA runs of four strategies located in a principal component plot of parameter space

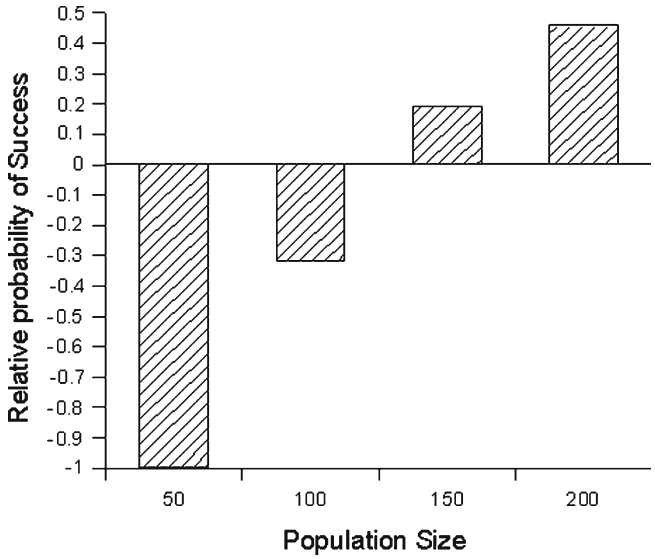


Fig. 9 Relative probability of success with respect to chosen population size, all strategies confounded

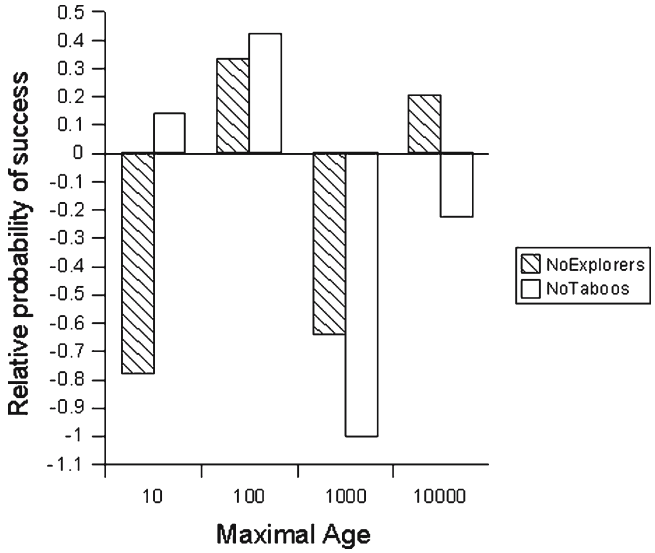


Fig. 10 Relative probability of success with respect to maximal age (in generations) within the "No Explorers" and "No Tabus" strategies

The tolerated stagnation of evolution before triggering a population reinitialization should not exceed 75 generations, in all the studied strategies. This tendency is, as expected, strongest within the "No Tabus" strategy, the most demanding for sources of population diversity.

Consensually, a high level of chromosome migration between islands appears to be optimal. Emigration of a new solution from its "native" island is permitted only once every N_{mig} generations: out of the four options of 5, 10, 25 and 50, $N_{\mathrm{mig}} = 10$ has been designed as the optimal choice, all strategies confounded.

The frequency of use of the "child-against-parent" selection rule only matters within the "No Explorers" and "No Tabus" strategies. In both latter cases, the Bayesian probability scores suggest that this selection rule should be completely abandoned. This is surprising in the "No Tabus" context, as the rule was supposed to enhance population diversity.

Imposing a strict similarity control parameter $S_{\max}$ within the populations is good policy. In virtually all strategies, the tolerated degree similarity between two conformers that are allowed to coexist in a population should be set below 75%, as this initial strict setup is being gradually relaxed in response to stalling evolution. The only exception is seen with the "No Explorers" strategy.

Eventually, a slight but consistent tendency in favor of elitism can be evidenced. No clear impact of the other tunable parameters of the CSGA could be established.

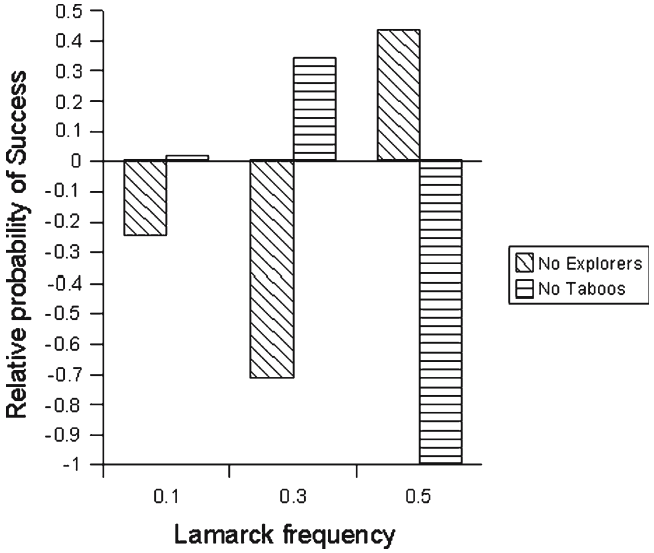


Fig. 11 Relative probability of success with respect to the frequency of use of Lamarckian optimization within the CSGAs in the "No Explorers" and "No Tabus" strategies

4 Conclusions

A GA-based conformational sampling procedure has been successfully used to search for relevant energy minima of a complex organic molecule, cyclodextrine. Specifically designed to handle multimodal optimization problems with about 100 degrees of freedom, the approach owns much of its success to its "hybridization" with other optimization strategies. Notably, the policy of "directed mutations (Explorers)" turned out to be extremely important for efficient discovery of low energy conformers. The mechanisms used to manage population diversity, and notably the "tabu search" employed in order to avoid revisiting of known optima appeared to be of paramount importance for ensuring the retrieval of various diverse local minima of the energy surface. Setting a "tabu" in the phase space neighborhood of a sampled conformation may involve the risk of blocking out some slightly deeper neighboring local minima corresponding to different arrangements of the small terminal moieties of the molecule. However, the benefit of the enforcement of non-redundant sampling is definitely more important than this drawback. In the specific molecule under study, replacement of the flat torsional value probability distribution with more sophisticated working hypotheses, aimed at returning the supposedly "correct" torsional values at higher rates, proved inconclusive. Biasing the random number generator in favor of torsional angle values that correspond to minimal local repulsions between vicinal atoms did not bring any clear advantage. The bias of torsional angle values in favor of values adopted in the previously sampled stable conformers proved to be, however, a cause for premature convergence of the sampling process and should be used with more restraint or fully abandoned.

Given the important number of operational parameter that control the CSGA, the genetic meta-optimization procedure proved extremely helpful in searching for reasonable parameter setup configurations. In a GA, a delicate balance needs to be kept between, on one hand, maintaining population diversity and, on the other, allowing for the convergence of this population towards a pool of related (sub)optimal chromosomes. For example, in the "No Tabus" strategy, which misses a key element acting in favor of population diversity, the fine-tuning provided by the meta-optimization procedure tried to compensate the "handicap" and empowered other diversity-enhancing mechanisms (lowering the maximal chromosome age, favoring population reinitialization by lowering the stagnation tolerance). This illustrates how important parameter tuning is for an effective use of genetic algorithms.

Due to the stochastic nature of genetic algorithms, the reproducibility of their results cannot be taken for granted, even if specific efforts were undertaken in this sense (triplicate rather than single runs being used as a basis for measuring the sampling success). The systematic repeat of triplicate runs triggered by the meta-optimization loop ensured that all the simulations eventually discovered the correct overall geometry of cyclodextrine, although the found solutions diverge with respect to the orientations predicted for the flexible rotatable substituents of the rings. However, flexible compounds with large "flat" energy wells in phase space may be much less easy to sample in a reproducible way.

As the optimal CSGA setups depend on the nature of the potential surface to be sampled, the specific conclusions and setups that were successful with cyclodextrine cannot be assumed to automatically apply to other molecules. In our opinion, the need to specifically tune a GA with respect to each new problem is general. Tuning cannot happen before the problem is solved, and therefore meta-optimization should not be regarded as a preliminary to problem-solving, but as the way to problem solving, that adjusts the tuning of the core GA on hand of the "experience" from previous trials.

References

1. Bäck T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, Oxford
2. Brunger AT, Clore GM, Gronenborn AM, Saffrich R, Nilges M (1993) Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science 261: 328–331
3. Calland PY (2003) On the structural complexity of a protein. Protein Eng 16:79–86
4. Damsbo M et al (2004) Application of evolutionary algorithm methods to polypeptidic folding: comparison with experimental results for unsolvated Ac-(Ala-Gly-Gly)5-LysH+. Proc Natl Acad Sci USA 101:7215–7222
5. Davy M, Del Moral P, Doucet A (2003) Méthodes Monte Carlo Séquentielles pour l'analyse Spectrale Bayésienne, Proceeding of the GRETSI Conference, Paris
6. De Jong KA, Potter MA, Spears WM (1997) Using a problem generator to explore the effects of epistasis. In: Proceedings of the 7th international conference on genetic algorithms. Morgan Kaufmann, San Fransisco, pp 338–345
7. De Jong KA, Spears WM, Gordon DF (1994) Using Markov chains to analyse GAFOs. In: Foundations of genetic algorithms 94, Morgan Kaufmann, San Fransisco, pp 115–137
8. Del Moral P, Doucet A (2002) Sequential Monte Carlo samplers, technical report 443, Cambridge University Press, Cambridge
9. Discover simulation package, Accelrys, San Diego, CA, http://www.accelrys.com/insight/discover.html
10. Glen WG, Dunn WJ, Scott DR (1989) Principal components analysis and partial least squares regressions. Tetrahedron Comput Technol 2:349–376
11. Glover F (1989) Tabu Search, Part I. ORSA J Comput 1(3):190–206
12. Glover F (1990) Tabu Search, Part II. ORSA J Comput 2(1):4–32
13. Goldberg DE (1989) Genetic algorithms in Search, optimization and machine learning. Addison-Wesley, Reading
14. Goto H, Osawa E (1993) An efficient algorithm for searching low-energy conformers of cyclic and acyclic molecules. J Chem Soc Perkin Trans 2:187–198
15. Grefenstette JJ (1986) Optimisation of control parameters for genetic algorithms. IEEE Trans SMC 16:122–128
16. Hagler AT, Huler E, Lifson S (1974) Energy functions for peptides and proteins: I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J Am Chem Soc 96: 5319–5327
17. Hart WE, Belew RK (1991) Optimizing an arbitrary function is hard for the genetic algorithm. In: Booker LB (ed) Proceedings of the 4th international conference on the genetic algorithms. Morgan Kaaufmann, San Mateo, pp 190–195
18. Herrera F, Lozano M (2001) Adaptative genetic operators based on coevolution with fuzzy behaviors. IEEE Trans Evol Comput 2:149–165
19. Heudin JC (1994) La vie artificielle. Hermès Editions, Paris
20. Hornak V, Simmerling C (2003) Generation of accurate protein loop conformations through low-barrier molecular dynamics. Proteins 51:577–590
21. Horvath D (1997) A virtual screening approach applied to the search of trypanothione reductase inhibitors. J Med Chem 15:2412–2423
22. Horvath D, Jeandenans C (2003) Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces – a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. J Chem Inf Comp Sci 43:680–690
23. Jarvis BB (2002) http://www.chem.umd.edu/courses/jarvis/chem 233spr04/Chapter04Notes.pdf
24. Kolossvary I, Guida WC (1996) Low mode search. An efficient, automated computational method for conformational analysis: Application to cyclic and acyclic alkanes and cyclic peptides. J Am Chem Soc 118:5011–5019
25. Kubota N, Fukuda T (1997) Genetic algorithms with age structure. Soft Comput 1:155–161
26. Michalewicz Z (1994) Genetic algorithms + data structure = evolution programs, 2nd edn. Springer, Berlin Heidelberg New York
27. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RE, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem 19:1639–1662
28. Ochoa G, Harvey J, Buxton H (1999) On recombination and Optimal Mutation Rates. In: Proceedings of genetic and evolutionary computation conference (GECCO-99), Morgan Kaufmann, San Francisco, pp 488–495
29. Packer MJ, Hunter CA (2001) Sequence-structure relationships in DNA oligomers: a computational approach. J Am Chem Soc 123:7399–7406
30. Pipeline Pilot version 3.0, available from SciTegic, Inc, at http://www.scitegic.com
31. Prebys EK (1999) The genetic algorithm in computer science. MIT Undergraduate J Math 1:165–170
32. Renders JM (1995) Algorithmes Génétiques et Réseaux de Neurones, Hermès Editions, Paris
33. Shetty RP, De Bakker PI, DePristo MA, Blundell TL (2003) Advantages of fine-grained side chain conformer libraries. Protein Eng 16:963–969
34. Spears WM (1992) Adapting crossover in a genetic algorithm, technical report AIC-92–025, Navy Center for Applied Research in AI, http://www.aic.nrl.navy.mil/∼spears/papers/adapt.crossover.pdf
35. Spears WM (1994) Simple subpopulation schemes. In: Proceedings of the third annual conference on evolutionary programming, Evolutionary Programming Society, San Diego, pp 296–307
36. Spears WM, De Jong KA (1996) Analysing GAs using Markov models with semantically ordered and lumped states. In: Foundations of genetic algorithms 96, Morgan Kaufmann, San Fransisco, pp 95–100
37. Stein EG, Rice LM, Brunger AT (1997) Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. J Magn Reson 124:154–164
38. Tai K (2004) Conformational sampling for the impatient. Biophys Chem 107:213–220
39. Teghem J (2003) Résolution de problèmes de RO par les métaheuristiques, Ed Hermès Sciences/Lavoisier, Paris
40. Vertanen K Genetic (1998) Adventures in parallel: towards a good island model under PVM. Oregon State University
41. Xia X, Maliski EG, Gallant P, Rogers D (2004) Classification of kinase inhibitors using a Bayesian model. J Med Chem 47:4463–4470

Annexe D

Article 2 : Journal of Chemical Informatic Models, 2006

paru dans « Journal of Chemical Informatic Models » en 2006, 46(6), p. 2457-2477

F. Bonachéra, B. Parent, Frédérique Barbosa, Nicolas Froloff et D. Horvath,

Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes.

Fuzzy Tricentric Pharmacophore Fingerprints. 1. Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes

Fanny Bonachéra,[†] Benjamin Parent,[†] Frédérique Barbosa,[‡] Nicolas Froloff,[‡] and Dragos Horvath*[,†]

Unite Mixte de Recherche 8576 Centre Nationale de la Recherche Scientifique − Unité de Glycobiologie Structurale & Fonctionnelle, Université des Sciences et Technologies de Lille, Bât. C9-59655 Villeneuve d'Ascq Cedex, France, and Cerep, Department of Molecular Modeling, 19 Avenue du Québec, 91951 Courtaboeuf Cedex, France

Received June 15, 2006

This paper introduces a novel molecular description—topological (2D) fuzzy pharmacophore triplets, 2D-FPT—using the number of interposed bonds as the measure of separation between the atoms representing pharmacophore types (hydrophobic, aromatic, hydrogen-bond donor and acceptor, cation, and anion). 2D-FPT features three key improvements with respect to the state-of-the-art pharmacophore fingerprints: (1) The first key novelty is fuzzy mapping of molecular triplets onto the basis set of pharmacophore triplets: unlike in the binary scheme where an atom triplet is set to highlight the bit of a single, best-matching basis triplet, the herein-defined fuzzy approach allows for gradual mapping of each atom triplet onto several related basis triplets, thus minimizing binary classification artifacts. (2) The second innovation is proteolytic equilibrium dependence, by explicitly considering all of the conjugated acids and bases (microspecies). 2D-FPTs are concentration-weighted (as predicted at pH = 7.4) averages of microspecies fingerprints. Therefore, small structural modifications, not affecting the overall pharmacophore pattern (in the sense of classical rule-based assignment), but nevertheless triggering a pK_a shift, will have a major impact on 2D-FPT. Pairs of almost identical compounds with significantly differing activities ("activity cliffs" in classical descriptor spaces) were in many cases predictable by 2D-FPT. (3) The third innovation is a new similarity scoring formula, acknowledging that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence. It displays excellent neighborhood behavior, outperforming 2D or 3D two-point pharmacophore descriptors or chemical fingerprints. The 2D-FPT calculator was developed using the chemoinformatics toolkit of ChemAxon (www.chemaxon.com).

1. INTRODUCTION

Rational drug design[1,2] largely relies on the paradigm of site−ligand shape and functional group complementarity in order to explain the affinity of a ligand for its macromolecular receptor. While molecular modeling may offer a deeper insight into ligand recognition mechanisms—molecular dynamics simulations[3] or free energy perturbation calculations[4] might, in principle, also account for the entropic effects at binding—it did not succeed to displace the more straightforward concept of binding pharmacophores[5−7] from the minds of medicinal chemists.

The idea that ligand-site affinity can be broken down into pairwise contributions from interacting functional groups is, after all, not all that far-fetched. Ligand binding is entropically penalizing—a ligand would not restrict its freedom of translation, rotation, and conformational flexibility by binding to a receptor unless this cost is compensated by enthalpic gains. The existence of at least one ligand pose making favorable contacts with the active site is a necessary, albeit not sufficient condition—but even so, a virtual filtering procedure, discarding all molecules failing to show enough complementarity to the site, might well score significant enrichment in actives. Complementarity, in the pharmacophoric sense, must be understood as the ability to form stabilizing interactions—hydrophobic contacts, hydrogen bonds, and salt bridges—between a ligand and a site. The exact chemical nature of the interacting functional groups can be dropped in favor of their pharmacophore type[8] T—hydrophobic (Hp) or aromatic (Ar), hydrogen-bond acceptor (HA) or donor (HD), and positively charged (PC) or negatively charged (NC) ions. Pharmacophorically equivalent functional groups are considered replaceable, ignoring the specific ways in which their chemical environment may modulate their properties (the hydrogen-bonding strengths, for example). Formally, pharmacophore-type information can be represented under the form of a binary pharmacophore flag matrix $F(a,T)$, with $F(a,T) = 1$ if atom a is of type T and $F(a,T) = 0$ otherwise.

While the pharmacophore paradigm had been introduced as a purely qualitative framework to explain ligand affinity and specificity for a given site, it has been recently taken over and used as a fundament for various chemoinformatics approaches—empirical algorithmic approaches for rational in silico compound selection, on the basis of some numeric descriptors[9,10] of the distribution pattern of pharmacophoric groups in the molecule. This overall pattern, mathematically represented by a fingerprint (vector) in which every component refers to a specific combination of types at given separations, accounts for the nature and relative position (in terms of topology or geometry) of all of the groups that are

* Corresponding author tel.: +333-20-43-49-97; fax: +333-20-43-65-55; e-mail: dragos.horvath@univ-lille1.fr, d.horvath@wanadoo.fr.
[†] Université des Sciences et Technologies de Lille.
[‡] Cerep.

10.1021/ci6002416 CCC: $33.50
Published on Web 10/21/2006 PAGE EST: 20.9

potentially involved in site−ligand interactions (the actually involved ones are not necessarily known at this stage). Pharmacophore fingerprints may be exploited in both similarity searches[11] and predictive quantitative structure−activity relationships (QSARs).[12] Similarity searches assume that molecules described by covariant fingerprints have similar overall pharmacophore patterns and, hence, a higher chance to share a common binding pharmacophore (and to bind to a same target) than any pair of randomly chosen compounds. In QSAR, model fitting may select[13] several key fingerprint components as arguments to enter an empirical (linear on nonlinear) function estimating the expected activities.

Despite their simplicity and potential pitfalls,[14] pharmacophore-based empirical models have been shown to be successful chemoinformatics tools. A key factor to success is the proper definition of underlying pharmacophore descriptors, with a minimal loss of chemically relevant information. One widely used approach is to monitor the numbers of pharmacophore group pairs[9,15] as a function of the pharmacophore-type combination they represent and the distance separating them. Distribution density plots of such pairs with respect to geometric or topological distance have been shown to display excellent neighborhood behavior (NB),[16] in the sense of selectively attributing high pharmacophore similarity scores to compound pairs with similar experimental properties. The use of fuzzy logics[17] at the descriptor buildup and similarity scoring stages appeared to be paramount in order to smooth out conformational sampling or categorization artifacts. Higher-order descriptors[18−20] monitor the triplets or quadruplets of pharmacophore types and, therefore, furnish a much more detailed description of the overall pharmacophore pattern but become more costly to evaluate and, more important, much more prone to categorization artifacts. This is the case of the binary three-dimensional three- and four-point fingerprints, which were found to show deceivingly low NB compared to their fuzzy two-point counterparts.[16] The main reason for this is the uncertainty of the assignment of a pharmacophore-type triplet or quadruplet to one of the predefined basis triangles or tetrahedra corresponding each to one of the fingerprint elements. In the context of a binary three-point fingerprint (see Figure 1), a basis triangle i is fully specified by a list of three pharmacophore types $T_j(i)$—each type T_j being associated with a corner $j = 1-3$ of the triangle—plus a set of three tolerance ranges $[d_{kj}^{\min}(i),d_{kj}^{\max}(i)]$ specifying constraints for triangle edge lengths. Basis triangles should thus be understood as the meshes of a grid onto which a molecule is being mapped. Considering an atom triplet $\{a_1, a_2, a_3\}$ in a molecule, this triplet is said to match a basis triangle i if (1) each atom a_j is of pharmacophore type $T_j(i)$, in other terms, $F[a_j,T_j(i)] > 0$ for each corner j and (2) the calculated—geometric or other—interatomic distances $\mathrm{dist}(a_j,a_k)$ each fall within the respective tolerance ranges: $d_{kj}^{\min}(i) \le \mathrm{dist}(a_j,a_k) < d_{kj}^{\max}(i)$.

If in a molecule M an atom triplet simultaneously fulfilling the above-mentioned conditions can be found, then the fingerprint of M will highlight the bit i corresponding to this basis triangle. The risk taken here is that in a very similar compound M'—or, if $\mathrm{dist}(a_j,a_k)$ are taken as geometric interatomic distances, in a slightly different conformation of the same molecule M—the equivalent atom triplet $\{a'_1,a'_2,a'_3\}$ may fail to match the basis triangle i. It is sufficient to have one of the three distances $\mathrm{dist}(a'_j,a'_k)$ exceeding by little one of the boundaries in order to highlight a completely different basis triangle i' in the fingerprint of M'. Basis triangles i' and i are similar, but this is ignored by a binary similarity scoring scheme failing to find either bit i or bit i' set in both compounds. In two-point descriptors, where elements standing for successive distance ranges are assigned successive indices $i' = i \pm 1$, the fingerprint scoring function could be trained to account for the covariance of neighboring bins. Such a straightforward fuzzy logics correction is no longer applicable here. There are, for example, three "successive" triangles of i {with the same $[d_{kj}^{\min}(i),d_{kj}^{\max}(i)]$ ranges for two of the edges and using the successive tolerance range for the third} but only one slot at position $i + 1$ of the fingerprint. The direct consequence is that relatively small differences in interatomic distances may trigger apparently random jumps (symbolized by the arrow of Figure 1, upper part) of the highlighted bits from one location in the fingerprint to another.

This paper shows that fuzzy tricentric pharmacophore descriptors can be successfully constructed and used. The current work reports the buildup of the topological fuzzy pharmacophore triplets (2D-FPT) using shortest-path topological distances as an indicator of pharmacophore group separation. The descriptor reports basis triangle population levels in a molecule instead of a binary presence/absence indicator. An atom triplet in the molecule will contribute to the population levels of all of the related basis triangles by an increment which is directly related to their fuzzy matching degree (Figure 1, below). In the fuzzy approach, it is sufficient to characterize basis triangles i by a set of three nominal edge lengths $d_{jk}(i)$ instead of the above-mentioned tolerance ranges. The fuzzy degree by which an atom triplet is said to match a basis triangle will be 100% if interatomic distances perfectly equal nominal edge lengths, $\mathrm{dist}(a_j,a_k) = d_{jk}(i)$, and smoothly decrease—according to a law to be detailed further on—as discrepancies between real and nominal distances become important.

While 2D-FPTs are obviously not subject to conformational sampling artifacts, fuzzy-logics-based descriptors nevertheless present essential advantages:

- Their tolerance with respect to the limited variability of topological distances between pharmacophore groups mimics the natural fuzziness of ligand recognition by active sites, which may tolerate the insertion or deletion of linker bonds in a series of analogues.
- Their size may be significantly reduced by an appropriate choice of the basis triangle set. In the fuzzy approach, it is, for example, possible to keep only basis triangles with edge sizes being multiples of 2, 3, or 4. Within the strict buildup procedure, any atom triplet featuring two atoms separated by an odd number of bonds would fail to highlight any of the basis triangles of even edge lengths—it would, in other words, slip between the meshes of the grid. A fine grid enumerating all basis triplets with all possible combinations of nominal distances must then be used—but many more of these will be required in order to cover the same global span in terms of possible distances.

A second element of originality introduced here is the pharmacophore-type assignment scheme for ionizable compounds. Classical rule-based pharmacophore typing ignores the mutual long-range influence of multiple ionizing groups,

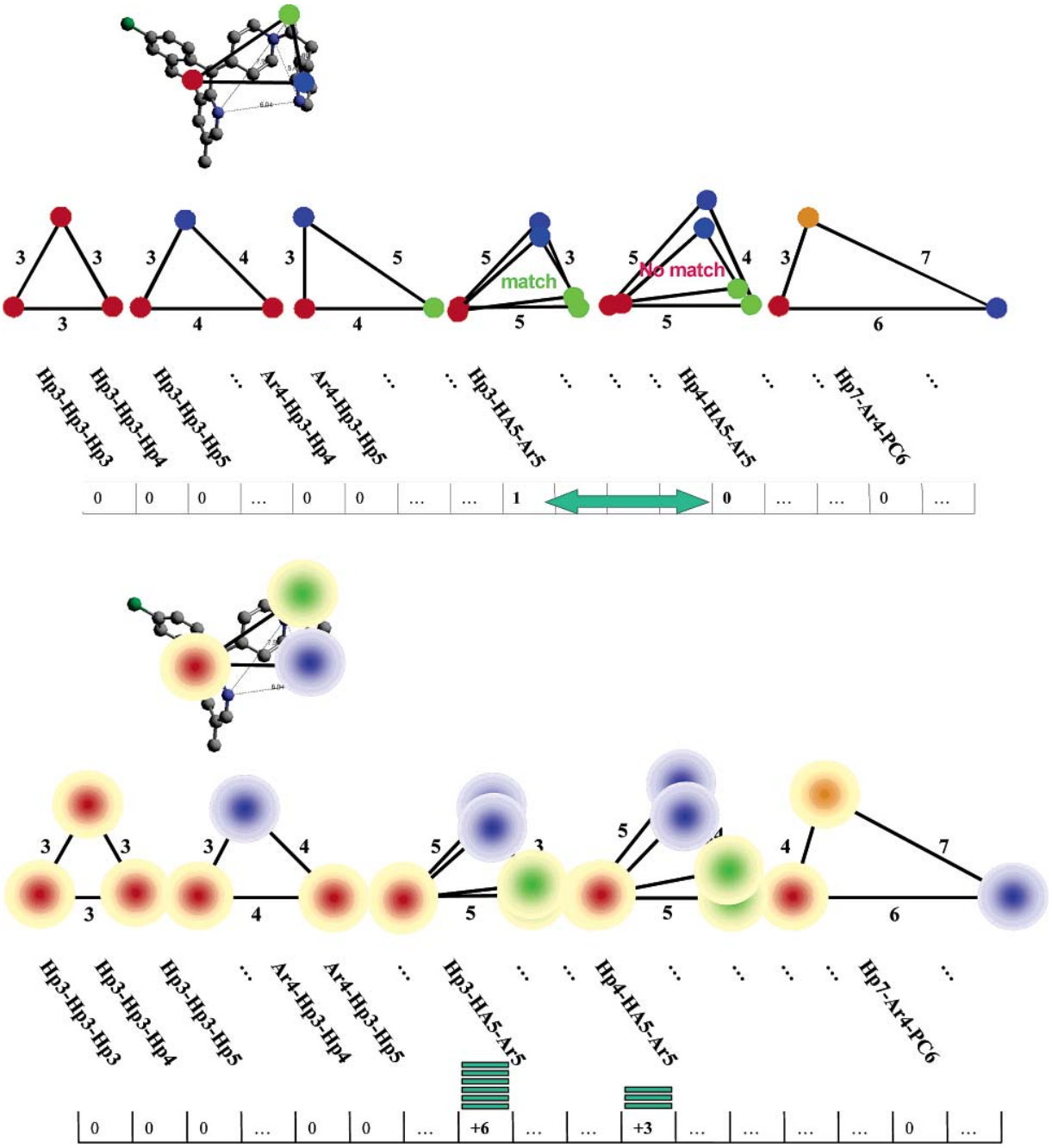


Figure 1. Buildup of a binary (above) and a fuzzy (below) pharmacophore triplet fingerprint, a vector in which every element stands for the presence (binary) or occurrence count (fuzzy) of given basis triplets. A triplet in a molecule (a) highlights a binary fingerprint component of the one best matching basis triangle or (b) increments the integer components of all of the matching basis triangles by amounts dependent on the match quality.

where each one of these is typed according to its protonation state of an isolated functional group at the considered pH. This leads to a typical overestimation of the occurrence of cation−cation or anion−anion pairs in polyamines and polyacids, respectively, and skews the molecular similarity measure upon the deletion of an ionizable group. Also, classical pharmacophore descriptors are not sensitive to electronic effects, being, for example, largely invariant upon the replacement of a methyl group (hydrophobe) by chlorine (another hydrophobe). This is acceptable unless, for example, the mentioned substitution prevents a neighboring amino group from accepting a proton in order to form a salt bridge at its binding site. To address these issues, 2D-FPT relies on the analysis of calculated[21] populations of all of the ionic or neutral forms involved in proton exchange equilibria—the "microspecies" μ, as they will be called throughout the paper—at a given pH. Each of these microspecies is mapped onto the basis triangle set, taking the actual anions and cations and donors and acceptors into account. The molecular fingerprint is rendered as the weighted average of microspecies fingerprints with respect to the predicted concentrations $c\%(\mu)$ of each microspecies μ at the considered pH of 7.4. In many cases, 2D-FPT-based analysis successfully proved that apparently near-identical compounds with puzzlingly different activities are not really as similar as they seem: the apparently minor (in the sense of classical rule-based

Table 1. Parameters Controlling 2D-FPT Buildup—Two Considered Setups

parameter	description	FPT-1	FPT-2
E_{min}	minimal edge length of basis triangles (number of bonds between two pharmacophore types)	2	4
E_{max}	maximal triangle edge length of basis triangles	12	15
E_{step}	edge length increment for enumeration of basis triangles	2	2
e	edge length excess parameter: in a molecule, triplets with edge length > $E_{max} + e$ are ignored	0	2
D	maximal edge length discrepancy tolerated when attempting to overlay a molecular triplet atop of a basis triangle	2	2
$\rho_{Hp} = \rho_{Ar}$	Gaussian fuzziness parameter for apolar (hydrophobic and aromatic) types	0.6	0.9
$\rho_{PC} = \rho_{NC}$	Gaussian fuzziness parameter for charged (positive and negative charge) types	0.6	0.8
$\rho_{HA} = \rho_{HD}$	Gaussian fuzziness parameter for polar (hydrogen bond donor and acceptor) types	0.6	0.7
l	aromatic−hydrophobic interchangeability level	0.6	0.5
	number of basis triplets at given setup	4494	7155

assignment) functional group substitutions actually had major impacts on ionization at the given pH. Many "activity cliffs" seen in classical descriptor spaces can be "leveled out" with pK_a-shift-sensitive 2D-FPT.

At last, the problem of appropriate similarity metrics to be used with 2D-FPT will be discussed, and an original scoring function, better adapted to such a high-dimensional descriptor, will be introduced. A plethora of various recipes have already been suggested[11] for comparing the descriptor sets (vectors) of two compounds m and M in order to determine a molecular dissimilarity score $\Sigma(m,M) = f[\vec{D}(M),\vec{D}(m)]$ (the distance in the structure space where each molecule is seen as a point localized by its vector of descriptors). 2D-FPT is, however, a large and potentially sparse fingerprint: out of the several thousands of basis triplets, only a few will be populated in simple molecules. Euclidean or Hamming distances may thus overemphasize the relative similarity of two simple molecules, while correlation coefficient-based metrics may be biased in favor of pairs of complex compounds. The original working hypothesis used here is to explicitly acknowledge that the simultaneous absence of a triplet in both molecules is a less-constraining indicator of similarity than its simultaneous presence, whereas its exclusive presence in only one of the compounds is a clear proof of dissimilarity. Specific partial distances are calculated with respect to the shared, exclusive, and null triplets in a fingerprint. A linear combination of these contributions leading to optimal neighborhood behavior was selected and used as the specific 2D-FPT similarity score.

For validation purposes, the NB of 2D-FPT was checked with respect to an activity profile featuring activity data (pIC_{50} values) of each molecule with respect to more that 150 targets, according to a previously outlined methodology.[22] Activity dissimilarity scores for $\sim 2.5 \times 10^6$ compound pairs were generated by Cerep, on the basis of the data in the BioPrint database[23,24] and according to a novel profile similarity scoring scheme. A second NB study has been carried out on publicly available data, by merging various QSAR data sets,[25−27] for different targets into an activity profile, assuming that each one of the molecules does not bind to any target except the one(s) for which pIC_{50} values above the micromolar threshold have been reported. Eventually, a validation study featuring virtual screening simulations will be presented. Virtual similarity screenings using 2D-FPT descriptors and metrics were performed by "seeding" a large commercially available compound collection (MayBridge) of 50 000 molecules with two sets of compounds (not used for 2D-FPT calibration) of known activities (featuring both actives and inactives) with respect to the dopamine receptor D2 and the tyrosine kinase c-Met, respectively. The ability of the 2D-FPT approach to retrieve the known actives and to avoid the selection of known inactives was benchmarked with respect to ChemAxon fuzzy pharmacophore fingerprints.[15]

2. METHODS

2.1. 2D-FPT Buildup: Fuzzy Mapping of Molecular Triplets onto Basis Triplets. Two prerequisite tasks must be completed prior to the actual construction of 2D-FPT.

Pharmacophore Flagging. This aspect will be detailed later on, because it is a central issue in ensuring the pK_a sensitivity of the fingerprints. At this time, the pharmacophore flag matrix $F_m(a,T)$, equaling 1 if atom a in the structure m is of type $T \in$ {"Hp", "Ar", "HA", "HD", "PC", "NC"} and zero otherwise, should be taken as granted. To account for the fact that aromatics and hydrophobes may, to some extent, interchangeably bind to the same binding pocket, in this work, aromatics are also flagged as lower-weight hydrophobes and vice versa. This requires the introduction of a fuzzy pharmacophore-type matrix $\Phi_m(a,T)$, identical to the binary flag matrix F for all of the polar types. For hydrophobes and aromatics, however, $\Phi_m(a,T) = \max[F_m(a,T), lF_m(a,T')]$ where T' stands for "aromatic" when T stands for "hydrophobic" and vice versa. $0 < l < 1$ is a tunable aromatic−hydrophobic compatibility parameter (Table 1). For example, an aromatic atom a has $F_m(a,\mathrm{Ar}) = \Phi_m(a,\mathrm{Ar}) = 1.0$, but $F_m(a,\mathrm{Hp}) = 0$ while $\Phi_m(a,\mathrm{Hp}) = l$.

Choice and Nonredundant Enumeration of the Basis Triplets Defining a Particular Version of 2D-FPT. The selection of a series of basis triplets to be monitored by the molecular fingerprint is essentially arbitrary and might be adapted to the specific problem for which 2D-FPTs are to be tailored. For the sake of concise specification, basis triplets are named $T_1d_{23}-T_2d_{13}-T_3d_{12}$, where T_i are the corner pharmacophore-type labels mentioned above and d_{ij} are the lengths of edges opposing each corner. For example, Ar4−Hp5−PC8 stands for a triangle in which the hydrophobe is four bonds away from the cation and eight bonds from the aromatic, while the aromatic and cation are five bonds apart. Basis triplets in this work were generated by systematic nonredundant enumeration, looping over each corner type, and respectively over each edge length from a user-defined minimal value E_{min} to a maximal E_{max}, with an integer step E_{step}. A pseudocode depiction of this procedure is given in Figure 2. Fingerprint element i hence monitors the population level of the basis triangle coded by the ith enumerated name in the list. The choice of E_{min}, E_{max}, and E_{step} (see Table 1) controls the coverage and graininess of the triplet basis set.

With these prerequisites, 2D-FPT buildup starts by the enumeration of all atom triplets $\{a_1, a_2, a_3\}$ in a molecule

```
for each T1 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {        #loop over type of corner1
   for each T2 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {     # ... corner 2
      for each T3 in ('Hp', 'Ar', 'HA', 'HD', 'PC', 'NC') {  #... and corner 3

        # Visit all the edge lengths from Emin to Emax with Estep
        for (d12=Emin,d12<=Emax,d12+=Estep) {

          #For 2nd edge, no need to loop over lengths below d12
          for (d13=d12,d13<=Emax,d13+=Estep) {

            # Only length combinations that may represent a triangle are enumerated
            # - third length may take only values verifying triangle inequalities
            dmin=max(Emin,|d12-d13|);
            dmax=min(Emax,d12+d13);
            for (d23=dmin,d23<dmax,d23+=Estep) {

              # Generate triangle corner labels Lk by concatenating types and
              # opposed edge lenght
              L1=T1d23;L2=T2d13;L3=T3d12;

              # Sort triangle corner label strings into a sorted list S.
              sort(L,S);

              # Final basis triplet name is obtained by concatenating corner labels in
              # their sorted alphabetical order
              NAME=S1'-'S2'-'S3;

              # Check whether this name had been generated previously;
              # if not add it to the list of basis triplets BLIST
              if !(BLIST.containsElement(NAME)) BLIST.add(NAME)

            } # end third edge length loop
          } # end second edge length loop
        } # end first edge length loop
      } # end third corner type loop
   } # end second corner type loop
} # end first corner type loop
```

Figure 2. Pseudocode rendering of the basis triplet enumeration procedure.

m, such that (1) the shortest topological distance between any two atoms equals or exceeds the minimal edge length E_{min} in basis triplets and (2) the longest one does not exceed the maximal edge length E_{max} by more than a tunable excess parameter e (Table 1).

To avoid confusion, in the following, the notation $t(a_k,a_j)$ to denote the (shortest-path) topological distance between two atoms will replace the generic interatomic distance dist(a_k,a_j) used in the introductory discussion on pharmacophore triplets. An atom triplet [note that the atoms of a triplet must be ordered such as to conveniently assign atoms to triangle corners; $\{a_1, a_2, a_3\}$ should not be understood as a list of three atoms taken according to their sequential ordering in the structure but the permuted list with the aromatic atom in position 1 if $T_1(i) =$ Ar etc.] is said to "potentially match" a basis triplet i if (1) each atom a_j features the pharmacophore type $T_j(i)$, in other terms, $\Phi_m[a_j,T_j(i)] > 0$ for each corner j, and (2) the topological distances $t(a_j,a_k)$ are close to the corresponding nominal edge lengths $d_{kj}(i)$, in the sense that $|t(a_j,a_k) - d_{kj}(i)| \leq \Delta$, the latter being a user-defined tolerance parameter (Table 1).

If a basis triangle is found to be a potential matcher of the triplet, their actual degree of similarity is calculated according to a simplified triangle overlay procedure related to the ComPharm[28] algorithm. Both the basis triplet i and the molecular triplet are represented as triangles of given (integer) edge lengths in the Euclidean plane. Each atom a_j in corner j is a source of a "pharmacophore field" $\psi_j(T,P)$ of type T. The intensity of such a pharmacophore field at any point P of space located at a distance d_{jP} from corner j is postulated to decrease according to a Gaussian function $\Phi(a_j,T) \exp(-\rho_{Tj}d_{jP}^2)$ of this distance, scaled by the extent $\Phi(a_j,T)$ to which atom a_j represents the pharmacophore type T. A 2D-superposition procedure translating and rotating the basis triangle with respect to the molecular triplet in order to achieve a relative alignment maximizing the covariance of these pharmacophore fields is launched after an initial triangle prealignment placing equivalent corners as closely together as possible. The fuzziness parameters ρ_T are treated as independent user-defined parameters of the method (Table 1).

Triplet-to-basis triangle overlay calculates a pharmacophore field covariance score ranging (in principle) between 0 (no match at all) and 1 (congruence). This score $O(i,\{a_k\})$ is an implicit function of the present pharmacophore types (and their intrinsic fuzziness parameters ρ_T), the nominal edge lengths of the basis triangle, and the actual topological distances within the atom triplet. In reality, covariance scores of 0 are never obtained, because the overlaid objects are filtered potential matchers. Actually, triangles sharing a common edge are guaranteed to score at least 0.67 (two conserved features out of three), no matter how far their third corners fall apart. Therefore, only covariance scores above the 2/3 threshold are considered:

$$O^*(i,\{a_k\}) = \max[0.0, O(i,\{a_k\}) - 2/3] \qquad (1)$$

The increment of the basis triplet population level due to the presence of a given atom triplet in m is proportional to $O^*(i,\{a_k\})$. Given the potentially large 2D-FPT fingerprint size, it is more practical to operate with integer rather than real population-level values. A scale-up factor of O^* has been introduced such that a basis triplet represented in a molecule by a single, perfectly congruent triplet reaches an arbitrary population level of 50. The ith 2D-FPT element $D_i(m)$, representing the total population level of a basis triplet i in species m, becomes

$$D_i(m) = \mathrm{int}[150 \times \sum_{\mathrm{atomtriplets}\{a_k\}\mathrm{in}m} O^*(i,\{a_k\})] \qquad (2)$$

2.2. Proteolytic Equilibrium-Dependent Fingerprint Buildup. The 2D-FPT generator uses ChemAxon's molecular reader classes[29] to input compounds in various formats and to standardize[30] connectivity and bond-order tables of compounds admitting several equivalent representations. Standardization rules were formally defined as chemical reactions in an XML configuration file read by the ChemAxon standardizer object (setup file in the Supporting Information).

On the basis of the standardized internal representations, the pharmacophore-type assignment procedure begins by submitting the current molecule to the ChemAxon pK_a plug-in.[31] This plug-in first predicts pK_a values for the ionizable groups of the molecule, then generates all of the possible conjugated acids and bases—the microspecies μ—together with their expected concentration $c\%(\mu)$, in percent, at the given pH (equal to 7.4 throughout this work). Next, the ChemAxon pharmacophore mapper tool (PMapper[15]) is used to flag the pharmacophore types within every microspecies. Specific pharmacophore flag matrices $F_\mu(a,T)$ and $\Phi_\mu(a,T)$ will be generated for each microspecies μ. PMapper is controlled by an XML file specifying flagging rules. A set of relevant substructures is specified as SMARTS[32] with labeled key atoms. Functional groups matching such sub-

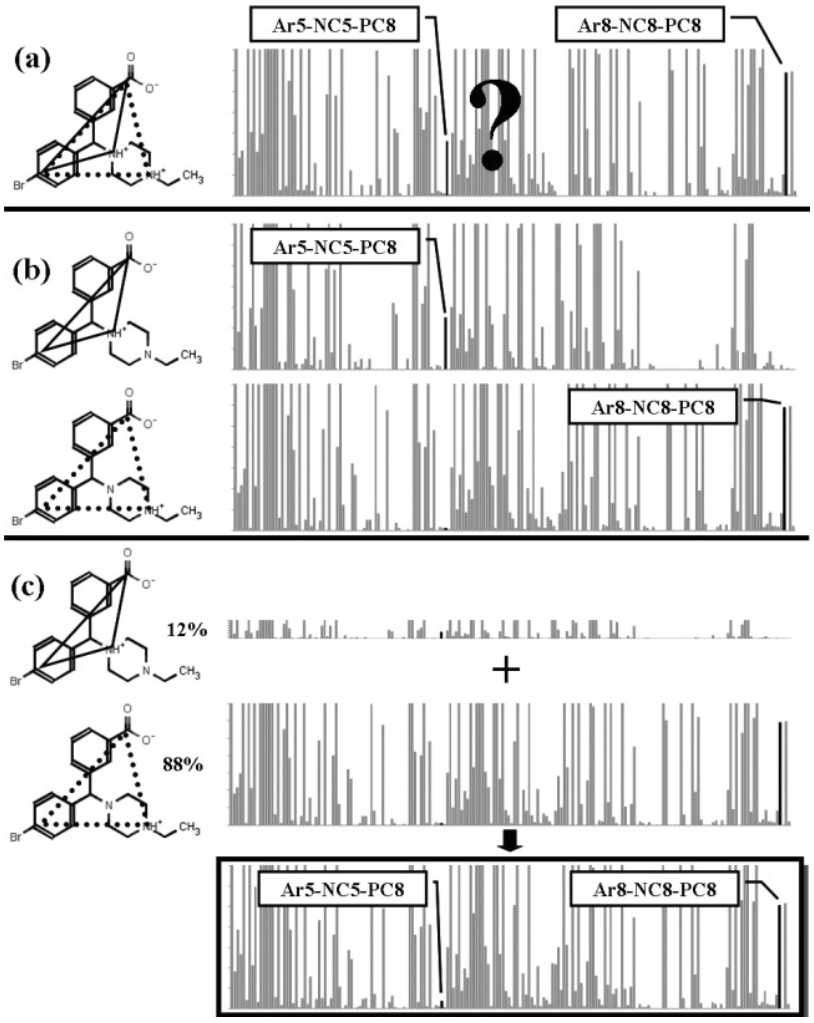


Figure 3. Graphical example of the principle of the construction of pK_a-sensitive 2D-FPT fingerprints: (a) rule-based pharmacophore flagging would assume three charged types in the molecule. Two triplets, both populated according to rule-based flagging, are localized in the sample fingerprint shown (bar sizes display population levels D_i, while the x axis enumerates the basis triplet counter i). Atom triplets that respectively contributed to each of the highlighted D_i's are marked in the structure. (b) The molecule actually appears at pH = 7 under the form of these two zwitterions. Each form carries only one of the triplets exemplified above. (c) The actual molecular fingerprint is obtained by weighed averaging of the microspecies fingerprints and, therefore, will resemble more the one of the zwitterionic forms predicted to occur at a concentration of 88% at equilibrium.

structures and the corresponding key atoms are detected in the molecule. An atom is assigned a given pharmacophore flag if it matches a certain substructure but not others. However, because formal charges are rigorously set in each microspecies, the assignment of PC and NC flags directly relies thereon. Any atom a carrying a positive formal charge (matching SMARTS "[*+]")—except for the nitrogen in nitro groups or nitrogen oxides—in the current microspecies μ will be assigned a flag $F_\mu(a,\mathrm{PC}) = 1$. By contrast, a classical flagging scheme would rely on the recognition of protonable group SMARTS and detect a potential cation even if it was not represented as such in the input molecule. Hydrogen-bond donor and acceptor flags are also set on the basis of specific rules pertaining to the microspecies. For example, a formally protonable N with a free electron pair, but not actually protonated in the current microspecies, will not be assigned an acceptor flag unless its pK_a value exceeds 5. Therefore, amide nitrogens will never be labeled as acceptors, but aniline nitrogens will unless they are strongly deactivated by electron-withdrawing groups. Oxygens always count as acceptors and −OH groups as donors. The recognition of aromatics is directly provided by ChemAxon's tools, while hydrophobes are defined as any carbon or halogen that is not aromatic and not charged.

The molecular fingerprint is thus obtained as a weighed average of microspecies fingerprints:

$$D_i(M) = \mathrm{int}\left[\sum_{\mathrm{microspecies}\,\mu\,\mathrm{of}\,M} \frac{c\%(\mu)}{100} D_i(\mu)\right] \qquad (3)$$

where $D_i(\mu)$'s are obtained for each microspecies μ, according to eq 2 using the specific pharmacophore flag matrix of the current microspecies for the estimation of the overlay score. The principle of proteolytic equilibrium-sensitive 2D-FPT buildup is illustrated in Figure 3. In the following, the notation D_i will, unless otherwise noted, implicitly refer to molecular average 2D-FPTs calculated according to eq 3.

2.3. FPT Similarity Scores. The appropriate choice of the similarity score $\Sigma(m,M) = f[\vec{D}(M),D(m)]$ comparing the 2D-FPT vectors of two molecules m and M is critical in order to ensure good NB. With classical metrics, such as the

Euclidean or Dice formulas, a first question is whether descriptors should be used as defined in eq 3 or after average/variance rescaling, leading to the set of normalized $\mathscr{D}_k(M)$: where $\alpha(D_k) = \langle D_k(m)\rangle_{\text{all } m}$ stands for the average of the

$$\mathscr{D}_k(M) = \frac{D_k(M) - \langle D_k(m)\rangle_{\text{all}m}}{\sqrt{\langle {D_k}^2(m)\rangle_{\text{all}m} - \langle D_k(m)\rangle_{\text{all}m}{}^2}} = \frac{D_k(M) - \alpha(D_k)}{\sigma(D_k)} \tag{4}$$

population level of triplet k over the BioPrint drugs and reference compounds[24] and $\sigma(D_k)$ stands for the corresponding variance. A further choice consisted in introducing a weighting scheme to specific triplets that are significantly populated in relatively few classes of compounds and absent from all of the others. These may be subject to an up to 10-fold increase of their relative importance with respect to ubiquitously present ones:

$$W_k = \min\left[10.0, \frac{\langle D_k(m)\rangle_{m\text{with}D_k(m)>0}}{\alpha(D_k)}\right] \tag{5}$$

Throughout this paper, structural dissimilarity metrics used with 2D-FPT will be denoted by the symbol Σ superscripted by the type of the metric, with an index informing on the use of normalized descriptors (N) as given in eq 4 or the weighting scheme (W) defined in eq 5. For example, the weighed Dice dissimilarity score using normalized descriptors is defined below, with N_T being the total number of basis triplets of the given 2D-FPT setup:

$$\Sigma^{\text{Dice}}_{N,W}(m,M) = 1 - \frac{2\sum_{k=1}^{N_T} W_{kk}(m)\mathscr{D}_k(M)}{\sum_{k=1}^{N_T} W_k{\mathscr{D}_k}^2(m) + \sum_{k=1}^{N_T} W_k{\mathscr{D}_k}^2(M)} \tag{6}$$

Indices N and W are omitted unless the metric explicitly relies on normalization and weighting and in cases of specific metrics (see below) or metrics from third-party software, whenever normalization and weighting options are no longer available.

The third, main, original contribution of this paper is the introduction of Σ^{FPT}, a specific metric of the dissimilarity of fuzzy pharmacophore triplets. Classical similarity scores, however, are generic metrics, applicable in arbitrary vector spaces, for example, independent of the actual nature of molecular descriptors associated with the degrees of freedom of the structure space. As this work will show, the specific design of a similarity scoring scheme based on an actual interpretation of the information in the fingerprint may significantly improve NB.

Concretely, the knowledge that $D_i(M)$ represents population levels of basis triplets, and that the simultaneous absence of a triplet in two molecules is a less-constraining indicator of similarity than its simultaneous presence, will be actively exploited. A first prerequisite in this sense is the introduction of a measure of the significance $S_k(M)$ of a triplet k for a molecule M, with respect to the observed averages and variances of each triplet population level:

$$S_k(M) = \begin{cases} 0 & \text{if } D_k(M) < 0.7\alpha(D_k) \\ 1 & \text{if } D_k(M) > 0.7\alpha(D_k) + \sigma(D_k) \\ \dfrac{D_k(M) - 0.7\alpha(D_k)}{\sigma(D_k)} & \text{otherwise} \end{cases} \tag{7}$$

A triplet k in a pair of molecules (m,M) may fall into one of the following categories: shared (++), for example, significant—in the above-mentioned sense—for both m and M, null (− −), for example, not significant for either, and exclusive (+ −), for example, significant for either m or M but not for both.

Rather than assigning it to one and only one of these, its fuzzy levels τ of association to each of the categories are defined in order to always sum up to 1:

$$\tau_k^{++}(m,M) = \frac{S_k(M)\, S_k(m)}{\text{norm}}$$

$$\tau_k^{-\,-}(m,M) = \frac{[1 - S_k(M)][1 - S_k(m)]}{\text{norm}}$$

$$\tau_k^{+\,-}(m,M) = \frac{|S_k(m) - S_k(M)|}{\text{norm}}$$

$$\text{norm} = S_k(M)\, S_k(m) + [1 - S_k(M)][1 - S_k(m)] + |S_k(m) - S_k(M)| \tag{8}$$

The fraction of triplets in a category c therefore becomes

$$f^c(M,m) = \frac{1}{N_T}\sum_{k=1}^{N_T}\tau_k^c(M,m) \tag{9}$$

Classical distance functions are typically calculated on the basis of the differences observed for each component k of the molecular descriptors $\delta_k(m,M) = |\mathscr{D}_k(m) - \mathscr{D}_k(M)|$. The herein introduced originality consists of a separate monitoring of these contributions for the shared, exclusive, and null triplets. Rather than simply summing up all $\delta_k(m,M)$ contributions (leading to a Hamming-type dissimilarity score), weighed partial distances $\Pi^c(m,M)$ are estimated in order to monitor how much of the difference stems from triplets in each category:

$$\Pi^c_{W,N}(m,M) = \frac{\sum_{k=1}^{N_T} W_k\tau_k^c(m,M)\,\delta_k(m,M)}{\sum_{k=1}^{N_T} W_k} \tag{10}$$

The working hypothesis adopted here was that a meaningful dissimilarity score can be expressed as some linear combination involving certain of the three fractions defined in eq 9 as well as the three partial distances (eq 10). Successive trials monitoring the NB of the resulting metric with respect to a subset of the entire learning set (see the following section) led to the following expression:

$$\Sigma^{\text{FPT}}(m,M) = 0.1323{\Pi_{W,N}}^{+-}(m,M) + 0.6357{\Pi_{W,N}}^{++}(m,M) + 0.2795[1 - f^{++}(m,M)] \tag{11}$$

The NB of the herein proposed scoring scheme was benchmarked with respect to classical dissimilarity metrics in various validation studies.

2.4. Experimental Data and Validation Studies. The performance of 2D-FPT in similarity searches has been assessed and compared to that of other 2D and 3D pharmacophore descriptors, following the previously published methodology[16] for monitoring the NB of in silico similarity scores. In the current work, activity profiles of 2275 nonproprietary (commercial drugs and drug precursors) molecules from the BioPrint database of Cerep were used to calculate the activity dissimilarity scores $\Lambda(m,M) = f[\vec{p}(M),\vec{p}(m)]$ expressing the amount of difference between the response patterns of the two molecules with respect to the considered battery of targets. Profiles $p_t(m)$ report measured $pIC_{50} = -\log IC_{50}$ (mol/l) values of every molecule m against each of $N_{targets} = 154$ different biological targets t (enzymes, receptors). $p_t(m) = 9/6/3$ means that molecule m is a nano-/micro-/millimolar binder of t, respectively. The actual algorithm used for estimating the activity profile dissimilarity score $\Lambda(M,m)$ is outlined in Appendix A.

An alternative NB study has been conducted on the basis of an activity profile compiled from publicly available data sets[25−27] (see the Supporting Information). Unlike the highly diverse BioPrint data, this study features a compilation of 112 compounds tested on the angiotensin converting enzyme (ACE), 111 on acetylcholine esterase (AchE), 163 on the benzodiazepine receptor (BzR), 321 on cyclooxygenase-II (Cox2), 641 on dihydrofolate reductase (DHFR), 66 on glycogen phosphorylase B, 67 on thermolysin, and 88 on thrombin (THR)—a total of 1569 molecules from eight activity classes. Each activity class is represented by a typical QSAR set, featuring variations of one or a few central scaffolds and including both actives ($pIC_{50} > 6$) and inactives in roughly equal proportions. The actual compilation of 1569 compounds has been realized by standardizing[30] the structures of molecules from the cited sources, then merging the sets and discarding duplicate compounds with conflicting activity data (associated activity values for a same target differing by more than one pIC_{50} log). In the absence of experimental data about the affinity of a compound m with respect to a target t, inactivity was assumed and $pIC_{50}(m,t)$ set to 3.5 in order to fill up the structure−activity profile matrix. Under this assumption, activity dissimilarity scores $\Lambda(M,m)$ were calculated according to Appendix A, with the conversion function ψ in equation A6 modified so as to return 1.0 only if its argument exceeds 12.5% of the number of targets in the profile (that is, one difference with respect to eight targets—the 5% threshold used with the much larger BioPrint profile makes no sense when $N_{targets} = 8$). With these specifications, an active compound M appears as equally distanced—at $\Lambda(M,m) = 1$—from any confirmed inactive of its own class, as well as from all of the molecules belonging to different classes. $\Lambda(M,m) = 0$ only if m and M are both actives within the same class. An inactive is set at $\Lambda(M,m) = 0.1$ from any other inactive, within its own series or not, but such pairs were consistently discarded, like in the BioPrint study case.

In the comparative NB studies, the experimental activity dissimilarity $\Lambda(M,m)$ is confronted to various calculated molecular dissimilarity scores $\Sigma(M,m)$. The purpose of such a benchmark is assessing in how far molecules (m,M) that are predicted to be neighbors in a given "structure space"—low $\Sigma(M,m)$—are systematically found to also be neighbors in "activity space"—low $\Lambda(M,m)$. The statistical formalism used to quantitatively evaluate NB is briefly revisited in Appendix B. NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. For simplicity, the plots were truncated at $\chi = 0.4$—displaying only the high-consistency range. Therefore, the characteristic U shape of $\Omega-\chi$ plots[16] may not always be apparent, but this is of little relevance for the discussion: the rule of thumb for the interpretation of the obtained graphs is that low Ω at high χ signals good neighborhood behavior.

2.4.1. Benchmarked Descriptors and Metrics. The NB of the 2D-FPT has been compared to the ones of different two-point pharmacophore descriptors, including fuzzy bipolar pharmacophore autocorrellograms (FBPA),[9] a 3D descriptor, and ChemAxon's topological fuzzy pharmacophore fingerprints.[15] The latter were calculated using both the recommended standard configuration (PF) and employing the "-R/--ignore-rotamers" (PFR) option of the ChemAxon descriptor generation tool. This option suppresses the default hypothesis according to which more fuzziness is applied when generating descriptor elements corresponding to more distanced atom pairs, as these have more options to experience important relative movements in the real molecule subjected to thermal agitation. ChemAxon's Chemical Fingerprints[33] (CF) were also used for benchmarking, as a representative of fragment-based fingerprints. To explicitly monitor the benefit of the novel-type flagging technique used with 2D-FPT, an alternative FPT relying on the same rule-based procedures used with PF/PFR has been generated. Molecular dissimilarity scores based on third-party descriptors were calculated according to the metrics best adapted for each—the Tanimoto score with ChemAxon's PF and CF and the fuzzy FBPA metric, respectively. XML setup files used for PF and CF descriptor and dissimilarity score calculations (PF.xml and CF.xml respectively) are included in the Supporting Information.

2.4.2. Virtual Screening of Seeded Compound Collections. A set of 50 000 random compounds—excluding organometallic derivatives and compounds of molecular mass above 1000 g/mol—from the MayBridge[34] vendor catalog were used as a reference chemical space to which molecules of known activities were added: (1) 194 compounds with reported c-Met tyrosine kinase activities from the literature,[35−37] including 72 actives with $IC_{50} \leq 10^{-7}$ M and (2) 460 molecules that were tested against the dopamine D2 receptor[38] (219 with $IC_{50} \leq 10^{-7}$ M). Both sets covered activity ranges from nanomolar to low millimolar values of IC_{50}. For each, the pharmacophorically most diverse three representatives were picked out of the respective subsets of very potent inhibitors ($IC_{50} < 10^{-8}$ M) and used as lead compounds for virtual screening according to both the 2D-FPT (FPT-2) and the PF-based Tanimoto metrics. The numbers of both confirmed actives ($IC_{50} \leq 10^{-7}$ M) *and* confirmed inactives ($IC_{50} > 10^{-7}$ M) were monitored within the sets of 200 nearest neighbors from the seeded chemical space found by each metric around each of these six leads.

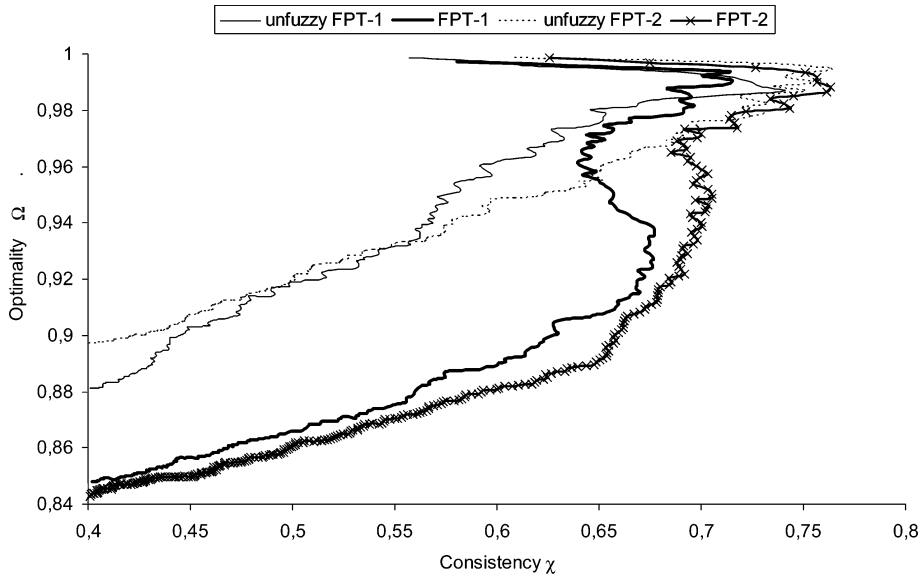


Figure 4. Comparative $\Omega-\chi$ plots illustrating the improvement of NB upon enabling the fuzzy mapping of atom triplets onto basis triplets, for both fingerprint versions FPT-1 and FPT-2, using the 2D-FPT specific similarity score Σ^{FPT} (BioPrint data set).

3. RESULTS AND DISCUSSIONS

3.1. The Importance of Fuzzy Mapping. To explicitly quantify the importance of fuzzy atom triplet mapping onto the basis triangles, the fuzziness factors ρ of considered FPT versions from Table 1 were temporarily set to 5.0 in order to generate comparative $\Omega-\chi$ plots for the corresponding unfuzzy fingerprints (the specific Σ^{FPT} score was used in all cases). At such high values of ρ, atom triplets will strictly highlight basis triplets of identical edge lengths. They will fail to highlight any basis triplet if the given combination of interatomic separations is not represented in the basis set. The corresponding curves in Figure 4 differ very little at their origins, where the selected pairs mostly include analogues with the same molecular scaffold and therefore are made of almost exactly the same atom triplets. However, the use of fuzzy logics is essential for extending the selection beyond these very first close analogues, to encompass pairs of compounds for which the underlying pharmacophore pattern similarity is not necessarily backed by a skeleton similarity. With fuzzy logics, many more activity-related compound pairs can be successfully picked without allowing pairs of different activities to enter the selection. Ω is observing a significant decrease without a loss of consistency, which is not seen when fuzzy mapping is turned off.

3.2. Importance of the pK_a-Dependent Fingerprint Buildup Strategy. The introduction of pK_a-dependent pharmacophore-type weights is expected to significantly contribute to the chemical meaningfulness of FPT. For example, a rule-based "educated guess" typically used to recognize potentially ionized groups in organic compounds would rely on the axiom that aliphatic amines are protonated, for example, must be flagged as cations and donors. Accordingly, N-alkylpiperazine-containing organic compounds will be assumed to harbor a cation−cation pair (see example in Figure 3). However, at pH = 7, only one of the two nitrogens is likely to carry a proton, its charge preventing the second one to do so. The cation−cation pair hence only appears in a minority of molecules, and its weight in the overall pharmacophore pattern should be adjusted accordingly. Piperazine may in reality be closer related to cyclohexylamine or morpholine than the rule-based pharmacophore pattern matching would suggest. Of course, rules can be tentatively optimized to avoid these kind of pitfalls: for example, the ChemAxon default pharmacophore mapping rules do not include tertiary amines into the cation category. This makes sense in medicinal chemistry, where the majority of amino groups in drugs are tertiary. The undue hypothesis of polycation patterns in the pharmacophore motif may hence be avoided, though at the cost of failing to perceive the similarity between secondary and tertiary amines.

An accurate prediction of the ionization status of protonable groups is a prerequisite for the success of the herein advocated flagging strategy. The NB of the fingerprints relying on ChemAxon's pK_a prediction plug-in outperforms the strategy of rule-based protonation state setup (Figure 5). This is thus an indirect proof of the accuracy of the pK_a prediction tool, offering an accurate estimation of expected protonation states. The rules used to build the alternative 2D-FPT (all other setup parameters being equal to FPT-1 values) were ChemAxon's default rules, the same used to construct the PF two-point pharmacophore fingerprints. A total of 59 pairs of compounds with identical activity profiles, ranking among the top 1000 most similar according to the pK_a-based approach, would lose their top-ranking positions and regress by more that 10000 ranks in the ordered pair list according to the rule-based method. Conversely, 50 activity-related pairs are perceived as similar by the rule-based metric, but not by the pK_a-based scoring scheme. The significant differences appear with respect to the distribution of activity-unrelated compound pairs. A total of 14 "violators" of the pK_a-based scheme (pairs with $\Lambda = 1$ but nevertheless ranked among the top 1000) are correctly reranked among the structurally dissimilar by the rule-based procedure. By contrast, 100 of the rule-based violators are successfully eliminated by the pK_a-based approach. Four typical examples of these latter ones are given in Figure 6. The similarity of compound pair a is clearly overstated by

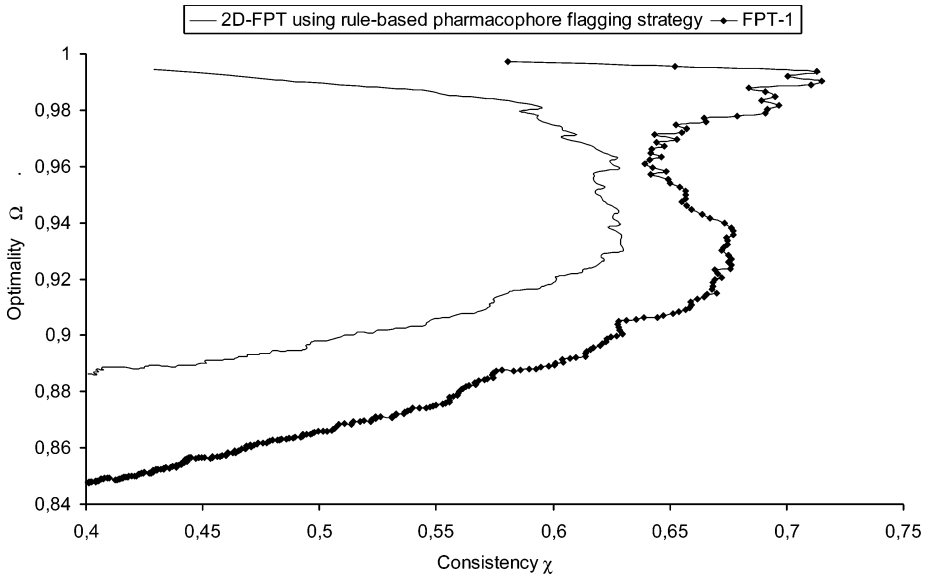


Figure 5. Standard rule-based flagging strategy of ionizable groups outperformed by the herein introduced pK_a-dependent fuzzy-type assignment procedure.

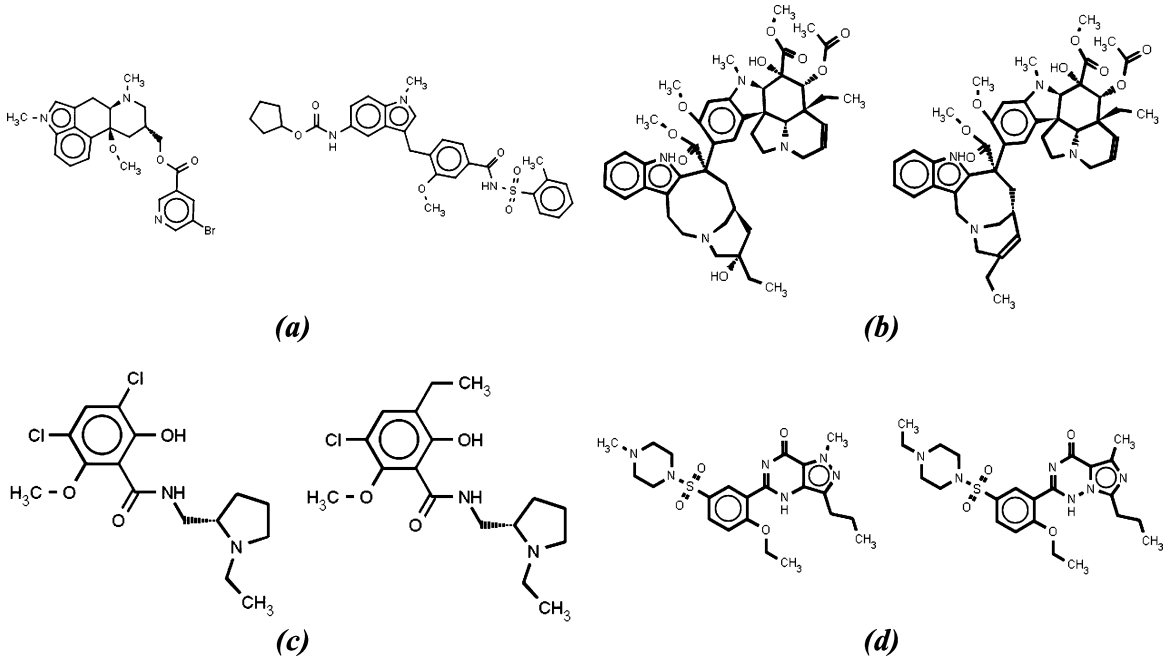


Figure 6. Examples of BioPrint compound pairs that look similar and are ranked among the top 1000 structurally closest pairs by the rule-based pharmacophore flagging scheme but, in reality, display radically different activity profiles and are correctly perceived as structurally different by the pK_a-based pharmacophore flagging scheme.

the rule-based scoring scheme, which regards both molecules as neutral species—acylsulfonamides are not declared as potential anions, and tertiary amines are not declared as cations in the ChemAxon default setup file pharma-frag.xml. Pair a stands thus for the numerous examples of activity-unrelated violator pairs that might have been avoided by redefining some of the flagging rules. In cases b, c, and d, however, pharmacophore dissimilarity cannot be accounted whatsoever by detailed flagging rule definitions: subtle substitution effects are seen to trigger relatively small pK_a shifts, but with dramatic impacts on the overall populations at proteolytic equilibrium. In compound pair c, the dissimilarity stems from the much more important ionization of the dichlorophenol compared to the monochlorophenol. While the left-hand compound mainly appears (according to the ChemAxon pK_a tool) under its zwitterionic form at pH = 7.4, the right-hand counterpart is predominantly positively charged. Even more dramatically, in example d, the addition of a simple methyl group enhances the protonation of the tertiary amine (70% cation at pH = 7.4 compared to 40% only in the left-hand molecule). Unless this effect is explicitly accounted for, a pharmacophore dissimilarity metric might never be able to explain the important activity differences observed upon the addition or deletion of a single hydrophobic center. Of course, the success of the approach relies on the precise pK_a estimation, or else the overestimated equilibrium population shifts that fortuitously explain observed activity differences might as well prevent the metric from recognizing the real pharmacophore similarity of activity-related pairs. As many com-

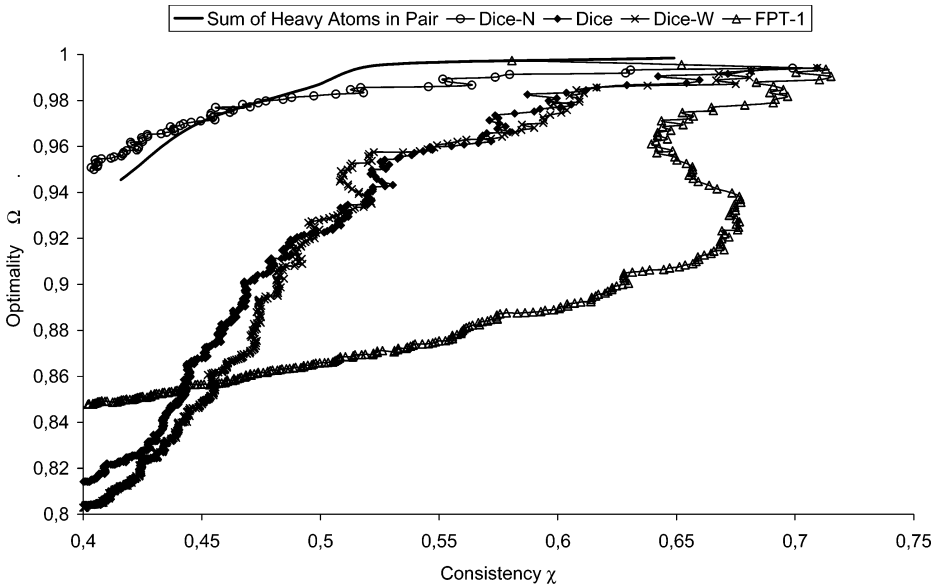


Figure 7. Comparative $\Omega-\chi$ plots of the NB (BioPrint data set) of various similarity scores with 2D-FPT (FPT-1 setup). Considered metrics are variants of the Dice formula: Σ^{Dice} ("Dice" in Figure legend), Σ_N^{Dice} ("Dice-N" in legend), and Σ_W^{Dice} ("Dice-W" in legend), as well as the 2D-FPT specific similarity score Σ^{FPT} ("FPT" in legend, eq 11).

pounds in this study are well-known drugs and reference molecules that are likely to have served for the pK_a tool calibration, further validation on the basis of original compound collections might be welcome. This notwithstanding, it can be concluded that one of the notorious limitations of pharmacophore-based similarity, the inability to explain activity shifts accompanying slight substitution pattern changes—a thorny issue raising fundamental questions about the validity of the neighborhood principle—might be successfully overcome in quite numerous cases of pK_a shift-related activity differences.

3.3. The Relative Performance of the Specific FPT Similarity Score. The NB of the various similarity scoring schemes using 2D-FPT (built according to setup 1 in Table 1) has been assessed, the results being shown in Figure 7.

The uppermost, solid curve represents the behavior of a fake dissimilarity score equaling the sum of heavy atoms in the molecule pair (m,M). It is nevertheless a well-shaped $\Omega-\chi$ plot, proving that activity-relatedness is statistically more likely to occur within subsets of small molecule pairs. This size effect is due to the fact that the smaller ($\sim$10 heavy atoms) of the employed molecules are unlikely to be strong binders to targets in the activity panel. Activity profiles of such compounds will be mostly empty, and their comparison returns low Λ scores (of about 0.1). Significant accumulation of such compound pairs at the top of the by-size sorted pair list ensures a significant consistency level of more than 60% within the top 20 lightest pairs (right-most point on the curve). Compound pairs with Λ scores of 0 (hitting common targets) are not contributing to these initial high consistency scores. The artifactual NB of size would have been even more marked if a bonus for binding to a same target would not have been included in Λ (results not shown).

Any rational pair selection strategy must therefore do better than (e.g., lay below) the size-driven NB curve. This is, unsurprisingly, not the case for the Dice metric based on normalized descriptors, which is quite sensitive to the complexity of the pharmacophore patterns of molecules, and implicitly to molecular size (see Figure 8). Small molecules with few populated triplets run an artificially high chance to be ranked as very similar: at $D_k(m) = 0$, $\mathscr{D}_k(m)$ simply relates to $-\alpha_k(m)$. The lesser the number of populated triplets is, the closer to the vector of average triplet populations—and the more correlated—the vectors $\mathscr{D}_k(m)$ and $\mathscr{D}_k(M)$ will be.

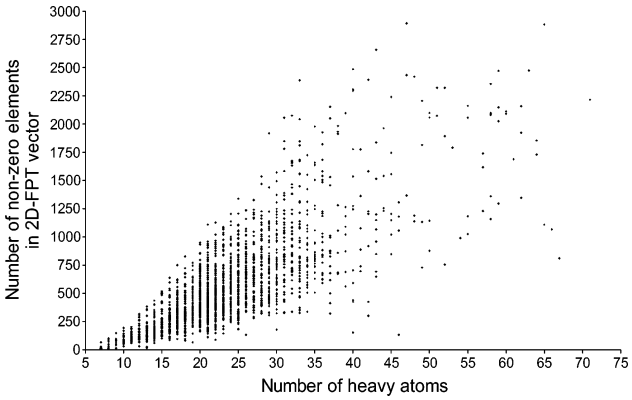


Figure 8. Dependence of the number of populated triplets on molecule size.

The same effect can be noticed with Euclidean scores (not shown). When $D_k(m) > 0$ and $D_k(M) > 0$, the chances that $D_k(m) = D_k(M)$ are quite small. Molecule pairs with a significant common set of populated basis triplets will, because of the summation of small but numerous residuals $\delta_k(m,M)$, typically end up at a higher Euclidean dissimilarity than pairs of small molecules with $D_k(m) = D_k(M) = 0$ for an overwhelming majority of triplets k. For example, the introduction of a methyl group in a large molecule M would trigger changes in the population levels of many more triplets k than the introduction of the same $-CH_3$ in a small compound m. Therefore, the calculated Euclidean distance score for a methyl/normethyl compound pair would counterintuitively increase with molecule size.

The Dice scores with or without the weighting of rare pharmacophore triplets can be successfully used to compare brute 2D-FPT, although they are clearly outperformed by the spe-

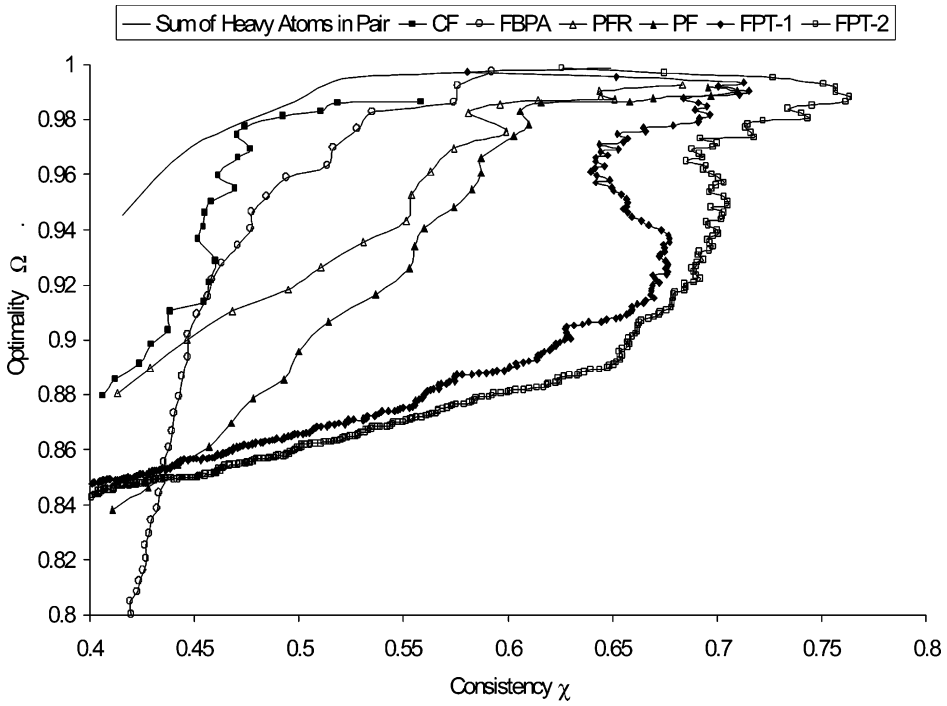


Figure 9. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (both setups, using the specific Σ^{FPT}) with respect to other descriptors and associated metrics (BioPrint data set).

cific FPT metric. In the Dice formula using 2D-FPT without any further norming or rescaling, the main criterion controlling dissimilarity is the number of common nonzero descriptor elements, as these are the only contributing to the sum of $D_k(m)D_k(M)$. Any molecules having no nonzero D_k values in common will be considered 100% dissimilar. However, two large molecules with less-sparse 2D-FPT vectors are much more likely to achieve some fortuitous overlap of their fingerprints than two small molecules. Even if an overwhelming number of exclusively populated D_k's exist, having $D_k(m)D_k(M) > 0$ for at least one k automatically ensures that such a molecule pair will nevertheless be ranked as more similar than any pair of small molecules with no shared triplets at all.

A general problem in molecular similarity scoring—be it molecular descriptor comparison or activity profile matching—appears to be the appropriate handling of the uncertain "null" situations describing the absence of an item (pharmacophore triplet, affinity with respect to a target) from both molecules. On one hand, it may be argued that the two compounds share the absence of an item, which makes them more similar. On the other, sharing the presence is clearly a stronger proof of similarity than sharing the absence, and the question is, how much stronger? Also, how can shared presence and shared absence be counterbalanced against the number of differences observed in the fingerprint, to achieve a meaningful final score?

The excellent NB of the dedicated dissimilarity score defined in eq 11 suggests an appropriate balancing of the contributions for the specific case of 2D-FPT. The dissimilarity score Σ^{FPT} is seen to increase in response to (a) observed differences between population levels of exclusively populated basis triplets and (b) observed differences between population levels of shared triplets. The coefficient of the latter is more important—however, it is the former that statistically contributes the most to the dissimilarity scores because situation a occurs more often.

Furthermore, Σ^{FPT} decreases as the total fraction of shared triplets increases—with the effect that $\Sigma^{FPT}(M,M)$ will decrease with molecule size: larger molecules (with richer pharmacophore patterns, strictly speaking) are "more similar to themselves" than smaller ones. This is not paradoxical if we give up considering Σ^{FPT} as a similarity metric, but consider it as a substitution score not unlike the ones used for sequence matching in bioinformatics:[39] the conservation of the rarer, larger, and functionally specific tryptophane in two sequences is seen as more significant and given a larger bonus than the conservation of a ubiquitous alanine.

3.4. Neighborhood Behavior of 2D-FPT, Compared to the Other Descriptors. Figure 9 compares the NB of 2D-FPT using Σ^{FPT} to that of other descriptor spaces and metrics. In can be seen that CF chemical fingerprints, which are tailored for (sub)structure recognition, do not fare better than size-driven artifacts. All of the pharmacophore descriptors, however, perform better than cumulated size. At low selection sizes (large Ω), PF outperform the fuzzy three-dimensional FBPA. However, although the latter metric tends to be too permissive (allowing compound pairs with different activities among its top-scoring pairs), it is nevertheless able to retrieve a maximum of existing activity-related pairs while maintaining a reasonable consistency of the selection (deep Ω minimum). Interestingly, applying higher fuzziness levels for more distant pharmacophore point pairs (default behavior in ChemAxon's pharmacophore fingerprint calculator) seems counterproductive in this benchmarking test: better results (PFR) are obtained when this approach is switched off.

It is remarkable that the 2D-FPT curves and notably the one obtained with the smaller triangle basis set (FPT-1) originate at relatively low consistency levels. As the selection is extended, the fraction of activity-related among the co-

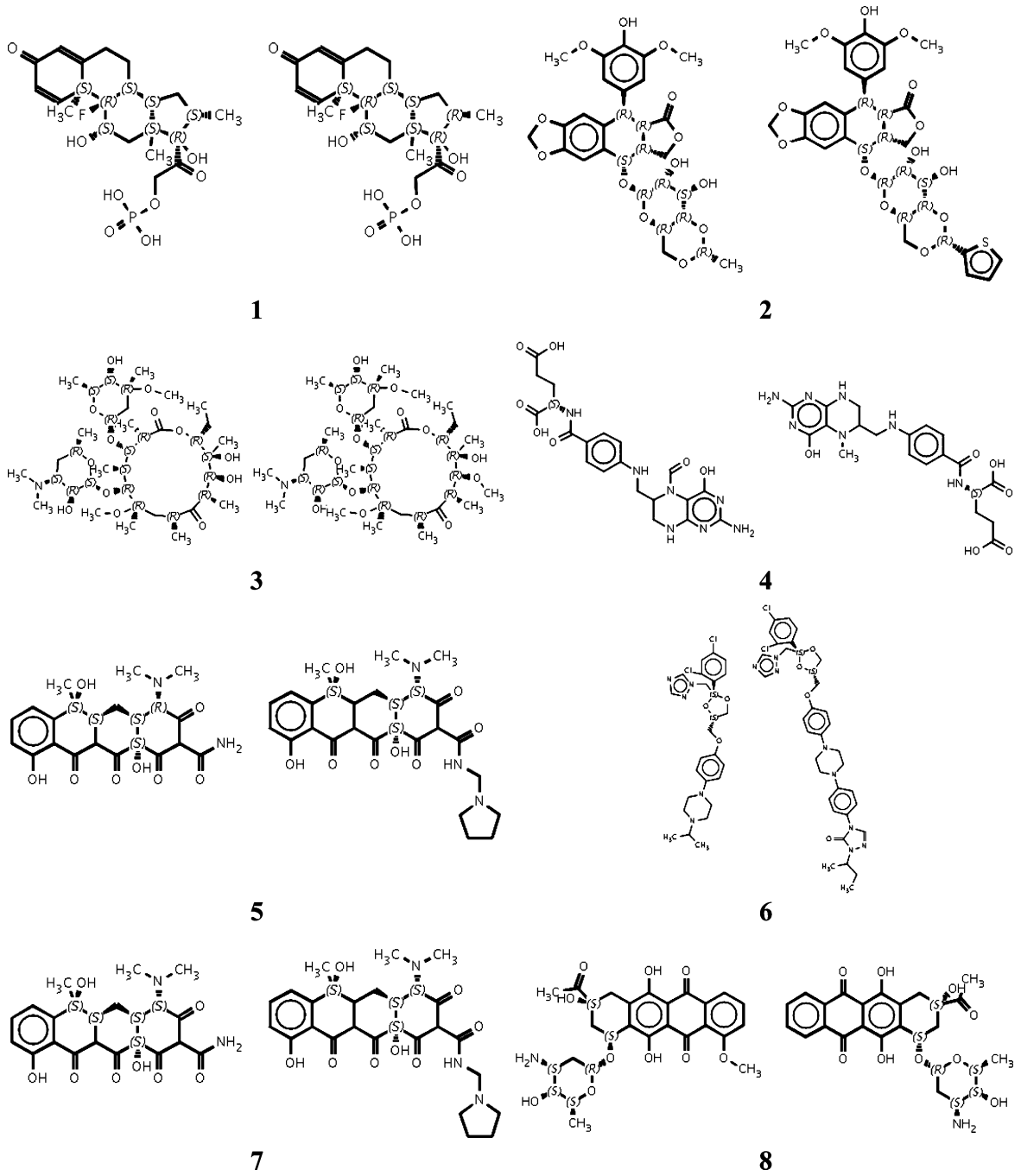


Figure 10. The eight pairs with highly dissimilar activity profiles found among the 50 most similar pairs according to 2D-FPT similarity scoring (FPT-1 setup).

opted pairs becomes much larger than that seen within the first top scorers. At high consistency values (0.5−0.7), significantly more activity-related compound pairs are retrieved by 2D-FPT than by any of the other scoring schemes.

Such behavior might be expected with topological descriptors such as 2D-FPT, because pairs of diastereomers (M,M*) score as much as a compound scores with respect to itself: $\Sigma^{FPT}(M,M^*) = \Sigma^{FPT}(M,M)$. The hypothesis that the initial inconsistency is due to the accumulation of activity-unrelated diastereomer and enantiomer pairs at the top of the similarity-sorted pair list must however be discarded. PFs, for example, are also topological distance-based and use a classical Tanimoto-based scoring scheme, so that $\Sigma^{PF}(M,M^*) = \Sigma^{PF}(M,M) = 0$ and diastereomers are always top scorers. However, the very high consistency of the right-most data point of the PFR curve proves that the 105 compound pairs with $0.00 \leq \Sigma^{PFR} < 0.01$, the herein included pairs of diastereomers, are not overwhelmingly activity-unrelated.

Actually, Σ^{FPT} no longer guarantees diastereomer pairs to rank among top scorers. $\Sigma^{FPT}(M,M) > 0$ decreases with the complexity of M, and pairs of slightly differently substituted analogues (M,M') sharing a highly complex pharmacophore pattern may score better than pairs of less complex molecules (m,m*) with identical fingerprints. Although $\Pi^{+-}(m,m^*) = \Pi^{++}(m,m^*) = 0$, having $f^{++}(M,M') > f^{++}(m,m^*)$ may eventually let the pair of close analogues score lower Σ^{FPT} values than the pair of diastereomers. The consistency inversion observed with 2D-FPT is, unexpectedly, not a

consequence of ignoring stereochemical information but actually stems from pairs of closely related analogues of very high molecular complexity. Among the best-ranked 100 pairs of compounds according to the FPT-1 setup of 2D-FPT scoring scheme, 66 have $\Lambda > 0.2$, 30 have $\Lambda > 0.5$, and 15 have $\Lambda > 0.8$. By contrast, in the pair subset ranked from 100 to 200, there are only 21 at $\Lambda > 0.2$, 13 at $\Lambda > 0.5$, and 6 at $\Lambda > 0.8$, for example, less than half as many NB violators than in the first 100 pairs. Violator pairs are, beyond doubt, chemically similar (to the point that finding the difference when looking at the structures is not always easy; Figure 10, except for examples 6 and 7, where substitution differences involve the introduction of a heterocycle and a cationic group, respectively). It is difficult to "blame" the 2D-FPT metric for having selected them. However, such "me-too" close analogue pairs are always among the top scorers of all of the similarity metrics, including PF and FBPA, but they are not seen to distort either of the herein-obtained NB curves. It can be safely assumed that, statistically speaking, closely related analogues differing in terms of either the stereochemistry or minor substituent changes tend to have similar biological activities, the exceptions to this rule being relatively rare (but widely publicized[40]). The previous section showed that 2D-FPTs are able to successfully explain some of these "activity cliffs" on the basis of predicted pK_a shifts. It appears however that they also tend to specifically pinpoint another subset of activity cliffs, pertaining to a specific series of close analogues that tend to score better than the ubiquitous activity-related "me-too" pairs. The 2D-FPT score-driven ranking of the BioPrint compound pairs evidenced a top-ranking subset of highly complex and very similar compound pairs with an increased propensity to form activity cliffs versus that of "typical me-too" pairs. At this point, it is however unclear whether this finding may be generalized to suggest that more-complex molecules are more likely to have their biological properties strongly affected by small chemical alterations. This is certainly not true with respect to overall physicochemical properties: methylation of a macrocycle like the third example in Figure 10 would hardly affect properties such as the octanol−water partition coefficient; by contrast, the methylation of methanol leads to the physicochemically different dimethyl ether. It is however important to remark that most of the compound pairs in Figure 10 are natural compounds or derivatives of natural compounds, optimized by Darwinian evolution to be perfect binders to a given target. From this viewpoint, it seems understandable that any small chemical alteration on the natural ligands may have dramatic changes in affinity. Synthetic drug molecules appear to be much less well-adapted to their targets and therefore, statistically spoken, much more tolerant to structural variations. 2D-FPT might provide a very useful metric for molecular complexity and implicit lead-likeness or drug-likeness—issues[41] that will be explored elsewhere.

The second parametrization attempt FPT-2 turned out to be more successful, but although the subsets of top scorers are significantly less marked by the accumulation of activity-unrelated pairs, the previously discussed consistency inversion does not vanish. Its better performance can be mainly ascribed to the shift of the minimal and maximal topological edge lengths from 2 to 4 and from 12 to 15, respectively. Monitoring triplets including directly bound, geminal or vicinal atoms does not enhance NB. This makes sense: binding pharmacophores typically include anchoring points from different parts of the ligand. Triplets involving, for example, both the carbonyl =O and the hydroxyl −OH in a hydroxamic acid RC(=O)−NH−OH are not accounted for in any of the versions—a specific fitting for metal enzyme inhibitors might prove necessary under these circumstances. The coverage of long-range molecular triplets seems to be very important: it also seems a good idea to extend the size of actually considered molecular triplets by $e = 2$ more bonds beyond E_{max}.

The initial choice of a grid of basis triplets having a mesh size (edge increment E_{step}) of 2 appears to be the good compromise. An E_{step} of 3 would have reduced the basis set size dramatically—however, molecular triangles with edge size values not appearing in the basis triplets would have been at risk to fall through the grid meshes, in failing to match any one of the basis triplets. Successful 2D-FPT setups with $E_{step} = 3$ may exist but must be actively searched for in the setup parameter space. $E_{step} = 1$ would, on the contrary, engender much larger grid sizes, thus causing significantly more practical problems with the handling of the resulting descriptors. Given the excellent behavior at $E_{step} = 2$, potential benefits of denser basis sets are unlikely to outweigh the descriptor size-related inconveniences.

A first key observation in Figure 11, monitoring the NB of various metrics with respect to the public data set obtained by merging eight independent QSAR series, is the much lower Ω values compared to what had been seen within the BioPrint set. Unsurprisingly, detecting structurally similar pairs of related activities is a much harder problem within the diverse set of drugs than within an artificially constructed set of series of analogues around a limited number of scaffolds. In this latter case, a simple discrimination between structural families—telling benzodiazepine-like chemotypes apart from acetycholine-like ligands and so forth—is sufficient to ensure significant NB. There are, for example, 65 active and 47 inactive ACE binders in the set; for example, 65/1569 = 4.14% of ACE actives in the entire set. Any metric that would consistently score lower dissimilarity between any two ACE set members than between an ACE and a non-ACE compound pair effectively discriminates between the ACE set and the rest of compounds. Within the ACE set, the rate of actives is however 65/112 = 58%, which represents a 58/4.14 = 14-fold enrichment in actives. Under these circumstances, dissimilarity scoring based on chemical fingerprints does display a significant NB, in sharp contrast to the observations made on the BioPrint set. The discrimination between the various chemical families that make up the public data set is readily achievable by all three metrics monitored in Figure 11: all of them avoided ranking any of the pairs of compounds from different series within the top 550 pairs corresponding to the checkpoints highlighted on the plots. All NB violators—in the sense of $\Lambda(m,M) > 0.5$—encountered at these checkpoints are intraseries activity cliffs regrouping an active and a structurally very close inactive. Within the top 550 pairs selected by the CF metric, the 128 observed NB violation instances break down into 15 ACE, 27 AchE, 5 BzR, 20 Cox2, 43 DHFR, and 18 THR compound pairs. Pharmacophore-based metrics should go beyond activity class recognition and successfully tell apart actives and inactives on the basis of a common scaffold. This

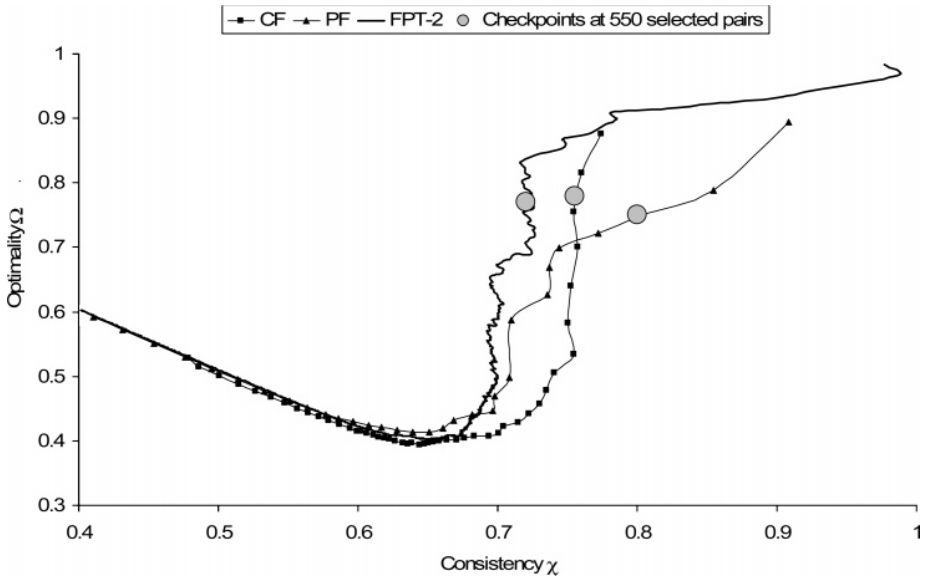


Figure 11. Comparative $\Omega-\chi$ plots illustrating the NB of 2D-FPT (setup FPT-2, using Σ^{FPT}) with respect to ChemAxon chemical and pharmacophore descriptors and associated metrics (public data set regrouping 1569 compounds from eight QSAR series).

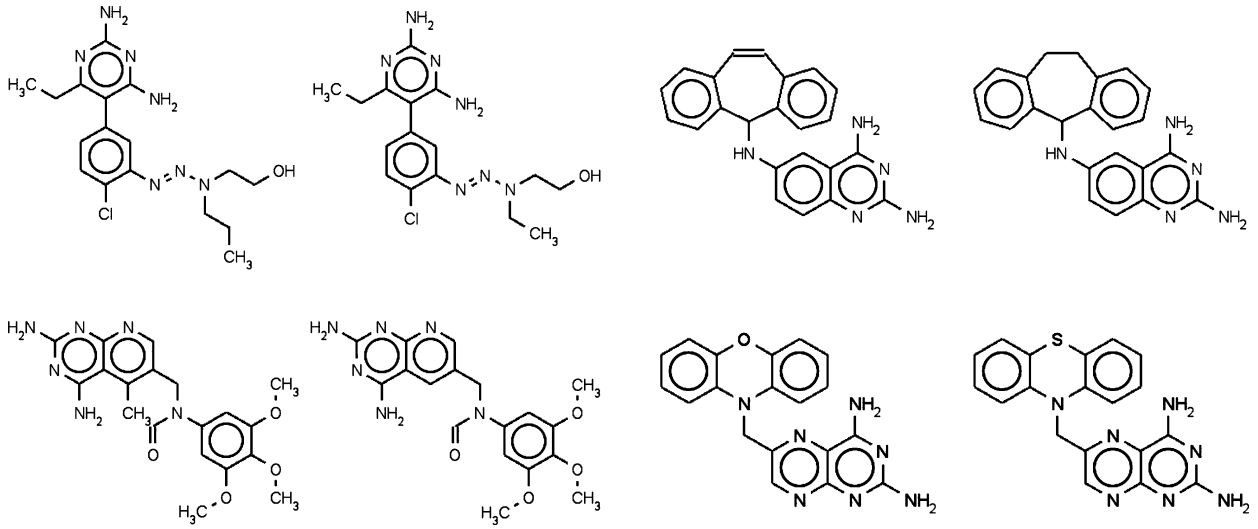


Figure 12. Typical "activity cliffs" of dihydrofolate reductase—very similar compound pairs with significantly differing DHFR activities ($\Lambda > 0.5$). Such compound pairs are consistently perceived as similar by all metrics—however, only the Σ^{FPT} formalism ranks these relatively complex compound pairs among the top 550.

is indeed observed with both PF and FPT metrics: both of these and particularly the latter reach out into higher consistency domains, not accessible to the CF approach. Unlike in the BioPrint study case, PF-driven NB reaches relatively better optimality scores at a same consistency or relatively higher consistencies at the same selection size (0.8 instead of 0.7 for the top 550 selected pairs, see checkpoints). An analysis of NB violators reveals that PF retrieved 92 such pairs within the top 550: 7ACE, 4 AchE, 3 BzR, 59 Cox2, and 19 DHFR, whereas FPT retrieved 138: 5 ACE, 48 Cox2, 83 DHFR, and 2 THR. The FPT approach thus experiences a sharp decrease of its NB criteria because of a local accumulation of DHFR activity cliffs, some typical examples of which are depicted in Figure 12. These are clearly structurally highly related compounds scoring very low dissimilarity values within both FPT and PF formalisms. However, only the former score includes a bonus for pharmacophore complexity, or it can be seen that DHFR ligands are among the most complex compounds in this set. DHFR pairs are therefore relatively better ranked than other intraset pairs when using FPT. Unfortunately, DHFR appears to display a rugged structure−activity landscape ridden by activity cliffs that cannot be conveniently explained by any of the herein explored metrics. This may be an illustration—but still no definite proof—of the possible correlation between ligand complexity and the propensity for activity cliffs, previously cited as an envisageable explanation for the observed consistency inversion of the FPT metric within the BioPrint set.

3.5. Virtual Screening Results of Seeded Compound Collections. Such simulations directly address the ability of the metrics to discover actives from databases but are less well-suited for rigorous benchmarking than the general NB analysis reported previously, insofar as the following are concerned:

• While a retrieval of a maximum of hidden actives among the top neighbors of each lead compound is desirable, it is not clear how many of the hidden actives are genuinely

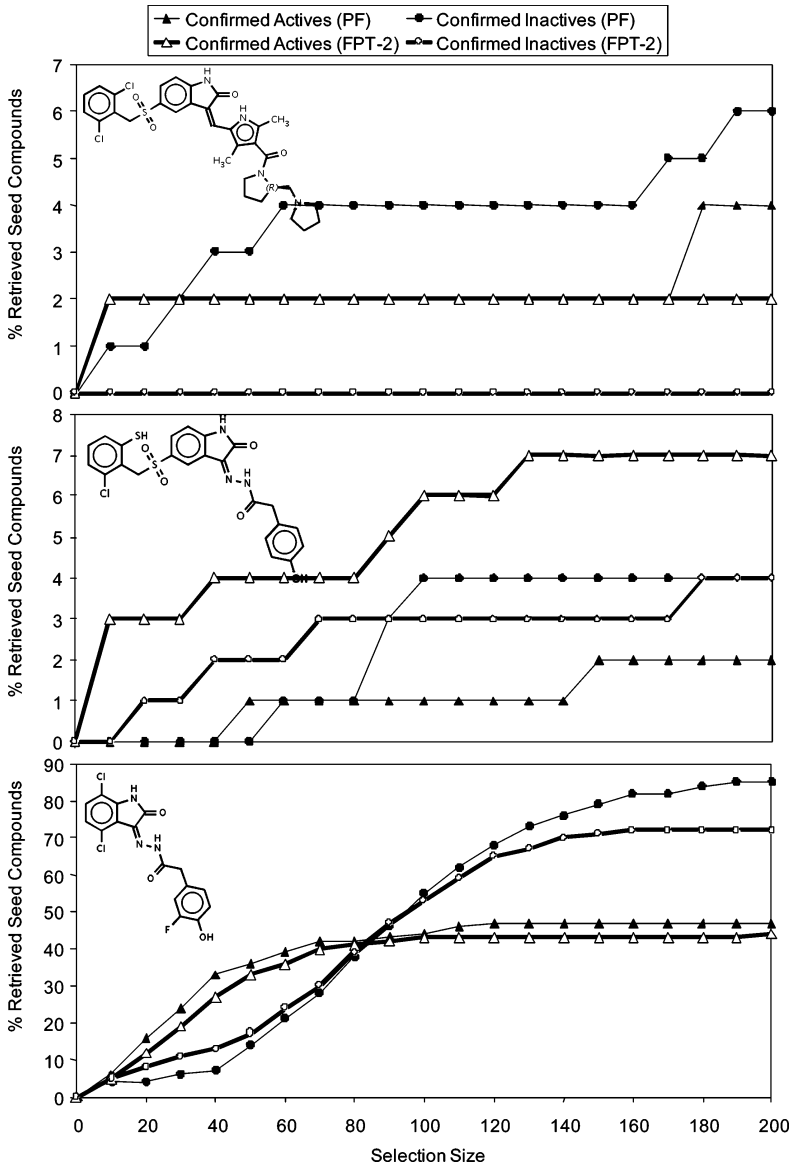


Figure 13. Results of virtual screening, probing each of the shown references against the MayBridge collection, seeded with compounds of known c-Met affinity (including actives with $pIC_{50} \geq 7$). Plots report the number of known actives and known inactives within subsets of nearest neighbors (subset size on the *x* axis) retrieved by the 2D-FPT (FPT-2 setup) and PF metrics, respectively.

similar to the lead and therefore eligible to be a virtual hit. Similarity to an active lead may be a sufficient but is clearly not a necessary condition. Unlike in virtual screening approaches based on QSAR or docking scores, successful similarity scoring is not expected to systematically score all of the actual active "ligands" better than the inactive "decoys"—if the set to be screened includes actives that are genuinely dissimilar to the reference, this subset of ligands might actually systematically score worse than decoys. The distributions of active ligands with respect to their similarity scores might actually be bi- or multimodal, complicating even more the statistical assessment of its robustness.[42] The selection criterion being the match of overall pharmacophore patterns—including those parts in which variability is not detrimental to binding—a search around a single lead may be too narrow.[43] In the present work, searches around single leads were performed with two different metrics (FPT and PF) and will be discussed in terms of relative retrieval rates.

- The key uncertainty in exploiting these results is the unknown activity status of the compounds from the bulk collection. The total number of actives present within the top neighbors is unknown, unless those compounds are ordered and tested against the target under study. Therefore, this study used both known actives and inactives for seeding. Selective enrichment in known actives, all while keeping the known inactives (often closely related analogues from the same series) out of the top neighbor set, is a strong indication of an increased probability to discover real actives among the hits from the bulk collection.

In the c-Met tyrosine kinase study case, the first two out of three lead compounds appear to be located at the rims of the cluster of the literature compounds of known activities. Both the PF and 2D-FPT-based metrics agree on the fact that the first lead (top plot in Figure 13) appears to have only two other known actives in its immediate neighborhood, with PF finding two more within the (arbitrary) limit of 200

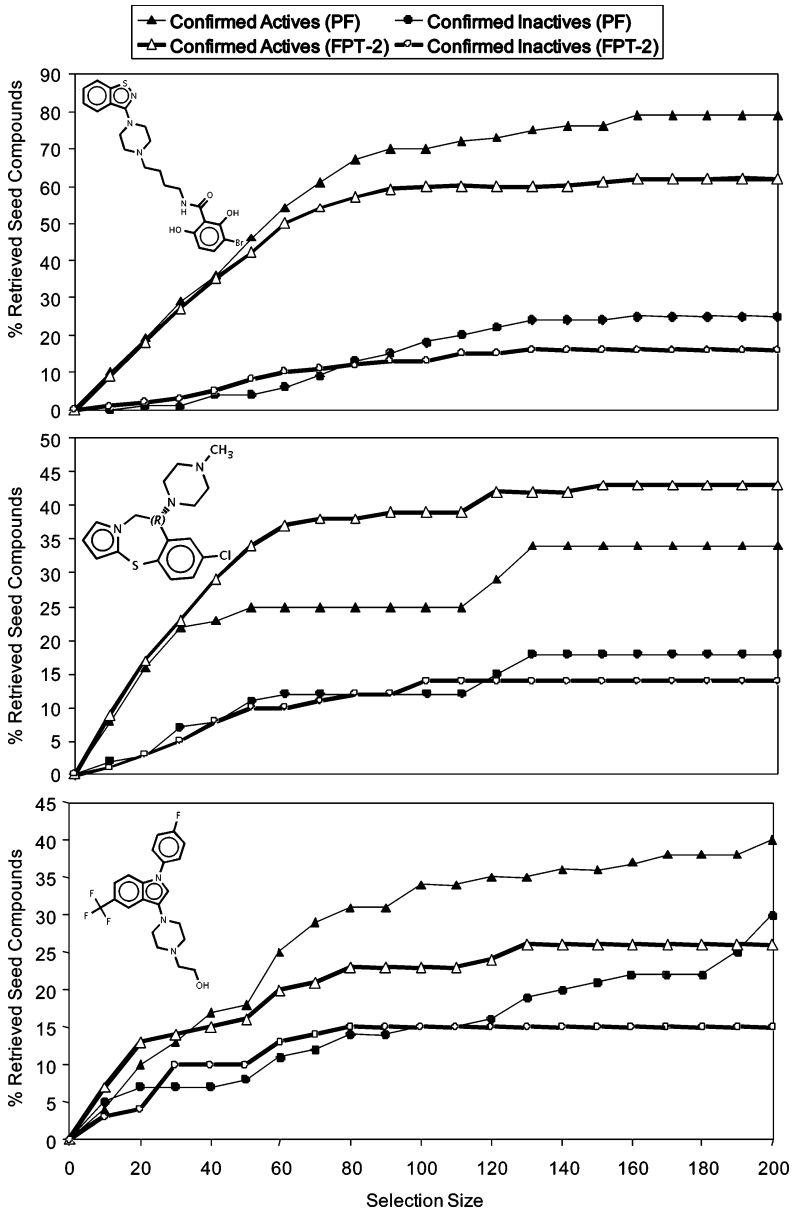


Figure 14. Virtual screening results for the D2 ligand study case (see legend of Figure 10 for details).

selected neighbors. However, the PF approach also co-opts four to six known inactives, which 2D-FPT successfully avoids. The results around the second lead compound are also clearly better with 2D-FPT, which recognizes roughly three times more known actives at basically equal numbers of co-opted inactives. The third c-Met lead appears, according to both metrics, to lay at the center of the c-Met compound cluster. Within the top 120 neighbors, retrieval levels closely match each other—with a slight advantage in favor of the PF approach, while at bigger selection sizes, the number of inactives co-opted by the PF significantly increases.

The study cases involving dopaminergic D2 compounds (Figure 14) showed that in all three situations lead molecules were well-surrounded by neighbors within the series. The first experiment may be considered a success of the PF approach—although it is still co-opting more inactives, it does better in known active retrieval by a clear margin. 2D-FPT clearly wins the second screening round, by simultaneously maximizing actives and minimizing co-opted inactives. The third experiment, eventually, is less clear-cut as the PF approach manages to retrieve more actives but only at the price of co-opting many more inactives than 2D-FPT.

Overall, the 2D-FPT-driven virtual screening appears to be more consistent—with respect to known actives and inactives—in the sense that higher active retrieval rates by PF are always accompanied by higher inactive retrieval rates as well. 2D-FPT systematically keeps the inactive retrieval rate equal or lower while nevertheless managing to improve the active retrieval rate in certain examples.

4. CONCLUSIONS

The insofar proven success of 2D-FPT-based similarity scoring compared to other fuzzy 2D and 3D pharmacophore descriptors is not surprising, as the three key innovations introduced here with respect to classical state-of-the-art descriptors and metrics are straightforward, chemically meaningful, and therefore expected to trigger improvements:

(1) The fuzzy mapping of molecular triplets on basis triplets is beneficial even in the context of topological distances (and assumed essential in a 3D context prone to conformational artifacts). It allows to accommodate the natural tolerance of receptors with respect to the number of bonds separating two binding groups and, from a practical point of view, allows a significant reduction of the descriptor dimension to a few thousands compared to >50 000 in binary fingerprints.

(2) The pK_a-dependent pharmacophore-type weighting scheme is able to correct many of the unavoidable inconsistencies that are introduced by rule-based flagging. Furthermore, local substituent swaps that, per se, would not translate to any significant pharmacophore pattern change as far as rule-based flagging is concerned may cause pK_a values to drift across the pH threshold and therefore trigger dramatic changes in the equilibrium population (and compound activity). Some of the "activity cliffs" in the structure—activity landscape of classical descriptor spaces are thus proven to be artifacts due to the failure of the latter to account for proteolytic equilibrium shifts. In the 2D-FPT space—for the first time, to our knowledge—this particular cause of landscape ruggedness has been successfully dealt with (insofar as the pK_a prediction tool is accurate, which appears to well be the case of the ChemAxon pK_a calculator employed in this work).

(3) The original similarity scoring scheme developed here recalls the simple truism that similarity due to the fact that a type is absent from both molecules is weaker than similarity due to the fact that both molecules contain the same type. As, in our hands, none of the classical scoring schemes managed to find the appropriate balance between contributions from shared, null, or exclusive triplets, such an optimal balance has been actively searched for—and found.

FPT as well as other pharmacophore-based descriptors have shown significant NB with respect to both diverse compound sets (BioPrint) and sets composed of several series of analogues. It is generally speaking much easier to demonstrate NB with respect to the latter situation, where simple discrimination between the main chemotypes at the basis of the various analogue series may suffice. The conclusions drawn on the basis of such studies may however be subject to different sources of bias due to relative size, chemical complexity, and other peculiarities of the considered analogue series. Mining for the underlying pharmacophore similarity in series with few representatives for each represented scaffold is much more challenging but successfully achieved by the FPT methodology. An interesting and recurring observation made in this work, requiring further investigation, is the possible correlation between the average pharmacophore complexity of the ligands of a target and its propensity for activity cliffs.

ACKNOWLEDGMENT

Special thanks to the ChemAxon (www.chemaxon.com) team, for allowing academics to freely use their software and for quick and effective hotline help. Sunset Molecular Inc. (http://sunsetmolecular.com/) and Tudor Oprea are acknowledged for providing the dopamine D2 data set. Nicole Dupont and Alexandre Barras (Institut de Biologie de Lille) are acknowledged for gathering the c-Met activity data from the literature. Thanks to Dr. Guy Lippens (University of Lille 1) for careful reading and important suggestions. ACCAMBA project members (http://accamba.imag.fr/) are acknowledged for encouraging this work.

APPENDIX A: THE ACTIVITY DISSIMILARITY SCORE

Similarity is an empirical concept, and there are no fundamental laws determining whether the activity profiles of two bioactive organic molecules are intrinsically similar or not. Like in the case of structural similarity, activity dissimilarity awaits for empirical definitions to be tried, validated, or discarded with respect to their usefulness in quantitative NB studies. Neighborhood behavior is necessarily a boot-strapping problem: its key assessment—that neighbors in a first (calculated) property space are likely to also be neighbors in a second (activity) property space—relies on two independent definitions of what "neighborhood" is supposed to mean in each one of the spaces.

For the above-mentioned reasons, this work postulates an activity dissimilarity score on the basis of plain medicinal chemistry common sense. Examples in which classical metrics (Euclidean, vector dot product, etc.) return counterintuitive dissimilarity measures will be discussed in order to highlight the need for a novel scoring scheme. Its implicit validation however comes from the fact that this definition of closeness in activity space respects the NB principle with respect to various molecular similarity metrics in structure space. In the following, the working hypotheses and parameters adopted in order to estimate the similarity of two activity profiles will be briefly outlined.

Profile similarity is determined by the behavior of a molecule pair (M,m) with respect to each target t. The target-specific response difference $\Delta_t(M,m)$ is defined as

$$\Delta_t(M,m) = \begin{cases} 0 & \text{if } |p_t(M) - p_t(m) \le 0.5 \\ 1 & \text{if } |p_t(M) - p_t(m) \ge 2.0 \\ \dfrac{|p_t(M) - p_t(m)| - 0.5}{1.5} & \text{otherwise} \end{cases} \quad \text{(A1)}$$

$\Delta_t(M,m)$ expresses a typical medicinal chemist's approach to activity comparison: two compounds with pIC_{50} values within 0.5 log units are said to have roughly the same activity; if however the pIC_{50} difference exceeds two log units, the molecules are beyond any doubt of different activity. In many situations, two log units is used as a landmark for selectivity: more than 2 orders of magnitude of affinity difference may not make any practical difference.

The activity index $\alpha_t(m)$ of a molecule m with respect to a target t is defined as a step function of the actual pIC_{50} value, such that compounds with affinities better than or equal to 1 μM count as active. A micromolar landmark for activity is widely used, especially in early stages of lead discovery.

$$\alpha_t(m) = \begin{cases} 0 \text{ if } p_t(m) < 6.0 \\ 1 \text{ otherwise} \end{cases} \quad \text{(A2)}$$

On the basis of definitions A1 and A2, $N_{diff}(m,M)$ and f_{diff}-(m,M)—the index and respectivel fraction of significant differences in the profiles of molecules M and m are defined

as

$$N_{\text{diff}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} [\alpha_t(m) + \alpha_t(M) - 2\alpha_t(m)\,\alpha_t(M)]\,\Delta_t(m,M)$$

$$f_{\text{diff}}(m,M) = \frac{N_{\text{diff}}(m,M)}{N_{\text{targets}}} \quad \text{(A3)}$$

In the N_{diff} index, the first factor plays the role of logical exclusive or it equals 1 if and only if either $\alpha_t(m) = 1$ or $\alpha_t(M) = 1$. If so, N_{diff} is incremented by the amount of the target-specific response difference $\Delta_t(M,m)$: a pair (M,m) of approximately micromolar affinities on opposite sides of the 1 μM threshold will not contribute. Intuitively, N_{diff} is a fuzzy counter of the obvious activity differences in the profile.

The index and respective fraction of similarities $N_{\text{sim}}(m,M)$ and $f_{\text{sim}}(m,M)$ observed in the activity profiles of the two molecules are defined as

$$N_{\text{sim}}(m,M) = \sum_{t=1}^{N_{\text{targets}}} \alpha_t(m)\,\alpha_t(M) \times [1 - \Delta_t(m,M)]$$

$$f_{\text{sim}}(m,M) = \frac{N_{\text{sim}}(m,M)}{N_{\text{targets}}} \quad \text{(A4)}$$

N_{sim} is the fuzzy counter of targets with respect to the two compounds having both strong $[\alpha_t(m) = \alpha_t(M) = 1]$ and similar $[\Delta_t(M,m) < 1]$ activities. Positive N_{sim} signals that the two compounds both interact with the same active site(s) and are therefore likely to include some common pharmacophore elements—insofar as most receptors tend to display a set of key interaction points that are always used in ligand binding, next to less important specific anchoring groups that form specific interactions with specific ligands. It is important to note that N_{diff} and N_{sim} do however not sum up to the total number N_{targets}. With respect to a pair of molecules, the set of targets making up the activity profile can be split into three domains: similarity, difference, and uncertainty, of sizes N_{sim}, N_{diff}, and $N_{\text{targets}} - N_{\text{diff}} - N_{\text{sim}}$, respectively. The uncertainty domain regroups targets for which molecules m and M display neither clear-cut different nor obviously similar behaviors. These include the (few) cases when compounds display significant potency differences despite both being active and the (ubiquitous) targets with respect to which m and M similarly fail to bind. A mutual lack of activity brings little information: molecules may be both inactive because of their similarity, or they may be each inactive in their own way.

The final activity dissimilarity score $\Lambda(m,M)$ associated with the activity profiles of molecules m and M is defined according to the following equation:

$$\Lambda(m,M) = \psi[f_{\text{diff}}(m,M) - \lambda \times f_{\text{sim}}(m,M)] \quad \text{(A5)}$$

with the conversion function $\psi(x)$ defined below:

$$\psi(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \le 0.05 \\ 0.1 + 18x & \text{if } 0 \le x < 0.05 \end{cases} \quad \text{(A6)}$$

In our opinion, this piecewise context-depending similarity scoring scheme returns a calculated profile activity score in agreement with medicinal chemistry and pharmaceutical know-how. Λ is a compromise between the sizes of the difference and similarity domains, with an empirical $\lambda = 5$ empirically chosen to emphasize the importance of observing actual similarities. The role of the conversion function $\psi(x)$ is to ensure the following:

- Only compound pairs sharing at least one significant (better than 1 μM) common hit in the profile may qualify to score top profile similarity (e.g., minimal $\Lambda = 0$), provided that the number of observed differences is low enough.
- If difference compensates for similarity, or if neither differences nor similarities could be evidenced (fully "uncertain" profiles, in the above-mentioned sense), a compromise score of 0.1 is returned. This value was chosen such as to signal that such profiles are clearly not different but should nevertheless not be allowed to compete in ranking with doubtlessly similar profiles at $\Lambda = 0$.
- Clearly different profiles, with $N_{\text{diff}} > \lambda N_{\text{sim}}$ score Λ values above 0.1, reach an upper limit of 1.0 if the excess differences make up more than 5% of the total number of targets in the profile.

It must be noted that Λ is not, strictly speaking, a metric: $\Lambda(M,M) = 0$ only if M binds at least to one target, with more than 1 μM of affinity. It is important to note that the conception of the Λ score ensures, unlike Euclidean or block distance metrics, a context-dependent activity difference interpretation. For example, the situation $p(m,t) = 5.0$ and $p(M,t) = 7.0$ marks an important difference between m and M, in the sense that selecting m from a database by means of a similarity screening experiment with respect to M might count as a failure. However, if $p(m,t) = 7.0$ and $p(M,t) = 9.0$, the discovery of m starting from M typically goes as a success, although the same 2 orders of magnitude of activity were lost. In the former case, target t contributes +1 to $N_{\text{diff}}(m,M)$, while in the latter, t contributes zero to both N_{diff} and N_{sim}. Eventually, if $p(m,t) > 7.0$ and $p(M,t) = 9.0$, target t becomes a contributor to N_{sim}. The Λ score therefore ranks a compound pair of activities (8,9) as more similar than a pair of activities (7,9) with respect to the target in question—like any Euclidean or Hamming score. Unlike these latter, however, Λ also meaningfully prioritizes the (7,9) pair over the (5,7) pair.

The failure of classical similarity metrics to respond differently to compound pairs that are both active and respectively both inactive often leads to an inappropriate, counterintuitive estimation of activity dissimilarity, as exemplified in Figure 15. The two bar plots represent comparative activity profiles—biological targets are aligned along the x axis, while the empty and filled bars respectively represent the pIC_{50} values of the compared molecules with respect to each target. Practically, IC_{50} values are only measurable starting from a certain activity threshold of the ligand—for compounds that are not active enough, a baseline pIC_{50} value of 3.0 is assumed (this also applies to BioPrint data). The left-hand graph displays a pair of molecules which have measurable pIC_{50} values with respect to a single target in the profile, and only one of them binds strongly enough to qualify as a potential hit or lead. A significant activity difference of three log units can be observed—obviously,

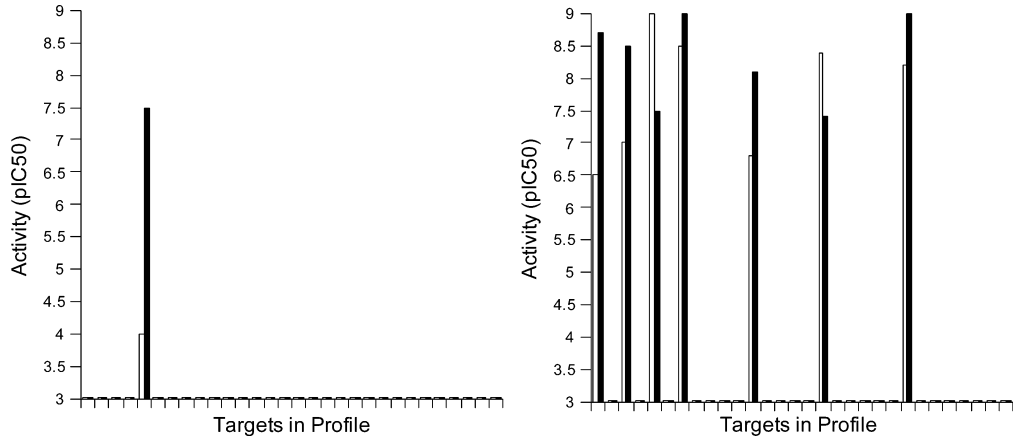


Figure 15. Two bar plots representing comparative activity profiles.

these molecules have different activity profiles. No other targets contribute to the Euclidean activity dissimilarity score, which therefore equals 3. The right-hand plot displays, by contrast, a pair of molecules with almost ideally covariant activities: they bind to the same targets, with comparable and significant—although not identical—affinities. However, every such target, rather than counting as a bonus in the profile similarity scoring, actually contributes some increment to the Euclidean profile dissimilarity score, which exceeds the dissimilarity level of the left-hand "different" compound pair and reaches 3.68. It is highly unlikely to expect identical activity values from binders to a same target, but it is guaranteed to get identical entries in the profile vector if none of the compounds have measurable pIC_{50} values—therefore, compound pairs with low hit rates in the profile will be spuriously favored by Euclidean scoring. A vector dot-product-based scoring metric would hardly perform better—as, in the left-hand plot, the only signals above the basis level stem from the same target; scores close to 1.0 (maximum similarity) are expected no matter what precise formula is used to calculate the profile correlation coefficient.

APPENDIX B: NEIGHBORHOOD BEHAVIOR CRITERIA.

NB analysis relies on monitoring activity dissimilarity within the subset $P(s)$ of molecule pairs (m,M) having calculated structural dissimilarity scores $\Sigma(M,m)$ below a variable dissimilarity threshold s. Let $N(s)$ represent the number of pairs retrieved by the selection $P(s)$ and which represent a fraction $f(s) = N(s)/N_{all}$ out of the total number of molecule pairs in the study. The consistency score $\chi(s)$ is defined in eq B1 by situating the average activity dissimilarity $\langle\Lambda(m,M)\rangle_{P(s)}$ of the $N(s)$ pairs in the actual selection at threshold s, in the context of (1) its upper baseline, the global average $\langle\Lambda(m,M)\rangle_{all}$ of all of the pairs in the study, which $\langle\Lambda(m,M)\rangle_{P(s)}$ approaches if selection at threshold s leads to a subset $P(s)$ as poor in activity-related pairs as a randomly picked one, and (2) its lower, ideal baseline, representing $\langle\Lambda(m,M)\rangle_{N(s)}^{MIN}$, the average Λ of the $N(s)$ compound pairs with the lowest Λ among the given N_{all} pairs.

$$\chi(s) = \frac{\langle\Lambda(m,M)\rangle_{all} - \langle\Lambda(m,M)\rangle_{P(s)}}{\langle\Lambda(m,M)\rangle_{all} - \langle\Lambda(m,M)\rangle_{N(s)}^{MIN}} \qquad \text{(B1)}$$

The overall optimality criterion $\Omega(s)$ renders a weighted account of two molecule pair counts in the actual selection of pairs $P(s)$ and randomly picked pairs:

- The first is the number of false similar pairs N_{FS} [structurally similar pairs with dissimilar activity profiles: $\Sigma(M,m) \le s$ and $\Lambda(M,m) > \kappa$]. A scaling factor $K > 1$ is applied to N_{FS} in order to take into account that, in virtual screening applied to drug discovery, the selection of pairs with diverging activity profiles is more penalizing than a failure to select all of the activity-related pairs (see below). In this work, $K = 100$.
- The second is the number of potentially false dissimilar pairs N_{PFD} [activity-related molecule pairs, apparently not structurally similar enough to be selected: $\Sigma(M,m) > s$ and $\Lambda(M,m) \le \kappa$].

The determination of N_{FS} and N_{PFD} requires in principle[16] a choice of the tolerated activity dissimilarity threshold κ—in the current context, however, every selected molecule pair (M,m) in $P(s)$ is fuzzily contributing an increment of $\Lambda(m,M)$ to N_{FS} and $1 - \Lambda(M,m)$ to N_{PFD}. In a random selection process, a set of size $N(s)$ would include activity-related and activity-unrelated pairs in a proportion equal to their overall occurrence in the total pair set and therefore

$$\Omega(s) = \frac{KN_{FS} + N_{PFD}}{KN_{FS}^{rand} + N_{PFD}^{rand}} = \frac{K\sum_{P(s)}\Lambda(M,m) + \sum_{All-P(s)}[1 - \Lambda(m,M)]}{K\frac{N(s)}{N_{all}}\sum_{all}\Lambda(m,M) + \left[1 - \frac{N(s)}{N_{all}}\right]\sum_{all}[1 - \Lambda(M,m)]} \qquad \text{(B2)}$$

NB can be graphically assessed by plotting the optimality criterion Ω against the consistency χ at various structural similarity thresholds s. Low Ω at high χ signals good neighborhood behavior.

Supporting Information Available: The public data set complied from eight QSAR series, including calculated FPT descriptors (FPT-2) and the .xml setup files controlling compound standardization and generation of ChemAxon PF and CF descriptors. This material is available free of charge via the Internet at http://pubs.acs.org. Activity dissimilarity $\Lambda(M,m)$ and FPT dissimilarity scores $\Sigma^{FTP}(M,m)$—not shared via

pubs.acs.org for technical reasons (files too large)—are available upon request (dragos.horvath@univ-lille1.fr).

REFERENCES AND NOTES

(1) Adam, M. Integrating Research and Development: The Emergence of Rational Drug Design in the Pharmaceutical Industry. *Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 513−37.

(2) Geney, R.; Sun, L.; Pera, P.; Bernacki, R. J.; Xia, S.; Horwitz, S. B.; Simmerling, C. L.; Ojima, I. Use of the Tubulin Bound Paclitaxel Conformation for Structure-Based Rational Drug Design. *Chem. Biol.* **2005**, *12*, 339−48.

(3) Ivanov, A. A.; Baskin, I. I.; Palyulin, V. A.; Piccagli, L.; Baraldi, P. G.; Zefirov, N. S. Molecular Modeling and Molecular Dynamics Simulation of the Human A2B Adenosine Receptor. The Study of the Possible Binding Modes of the A2B Receptor Antagonists. *J. Med. Chem.* **2005**, *48*, 6813−20.

(4) Bernacki, K.; Kalyanaraman, C.; Jacobson, M. P. Virtual Ligand Screening against Escherichia coli Dihydrofolate Reductase: Improving Docking Enrichment Using Physics-Based Methods. *J. Biomol. Screening* **2005**, *10*, 675−81.

(5) Barreca, M. L.; Ferro, S.; Rao, A.; De Luca, L.; Zappala, M.; Monforte, A. M.; Debyser, Z.; Witvrouw, M.; Chimirri, A. Pharmacophore-Based Design of HIV-1 Integrase Strand-Transfer Inhibitors. *J. Med. Chem.* **2005**, *48*, 7084−8.

(6) Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 6997−7004.

(7) Low, C. M.; Buck, I. M.; Cooke, T.; Cushnir, J. R.; Kalindjian, S. B.; Kotecha, A.; Pether, M. J.; Shankley, N. P.; Vinter, J. G.; Wright, L. Scaffold Hopping with Molecular Field Points: Identification of a Cholecystokinin-2 (CCK2) Receptor Pharmacophore and Its Use in the Design of a Prototypical Series of Pyrrole- and Imidazole-Based CCK2 Antagonists. *J. Med. Chem.* **2005**, *48*, 6790−802.

(8) Güner, O. F. *Pharmacophore Perception, Use and Development in Drug Design*; International University Line: La Jolla, CA, 2000.

(9) Horvath, D. High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and its Role in the Drug Discovery Laboratory. In *Combinatorial Library Design and Evaluation. Principles, Software Tools, and Applications in Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, 2001; pp 429−472.

(10) Makara, M. G. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(12) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of Validated QSAR Models of d(1) Dopaminergic Antagonists for Database Mining. *J. Med. Chem.* **2005**, *48*, 7322−32.

(13) Rolland, C.; Gozalbes, R.; Nicolai, E.; Paugam, M. F.; Coussy, L.; Barbosa, F.; Horvath, D.; Revah, F. G-Protein-Coupled Receptor Affinity Prediction Based on the Use of a Profiling Dataset: QSAR Design, Synthesis, and Experimental Validation. *J. Med. Chem.* **2005**, *48*, 6563−74.

(14) Horvath, D.; Mao, B.; Gozalbes, R.; Barbosa, F.; Rogalski, S. L. Strengths and Limitations of Pharmacophore-Based Virtual Screening. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2004.

(15) For details on the two-point topological pharmacophore descriptors developed by ChemAxon, see http://www.chemaxon.com/jchem/index.html?content=doc/user/Screen.html (accessed Sept 2006).

(16) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with respect to In Vitro Activity Spaces − A Benchmark for Neighborhood Behavior Assessment of Different in Silico Similarity Metrics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 691−698.

(17) Horvath, D.; Mao, B. Neighborhood Behavior − Fuzzy Molecular Descriptors and their Influence on the Relationship between Structural Similarity and Property Similarity. *QSAR Comb. Sci.* **2003**, *22*, 498−509; special issue "Machine Learning Methods in QSAR Modeling".

(18) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−23.

(19) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1995**, *38*, 144−150.

(20) Menard, J. P.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−13.

(21) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of Distribution Coefficient from Structure. 1. Estimation Method. *J. Pharm. Sci.* **1997**, *86*, 865−71.

(22) Horvath, D.; Jeandenans, C. Neighborhood Behavior of in Silico Structural Spaces with Respect to in Vitro Activity Spaces − A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 680−690.

(23) Krejsa, C. M.; Horvath, D.; Rogalski, S. L.; Penzotti, J. E.; Mao, B.; Barbosa, F.; Migeon J. C. Predicting ADME Properties and Side Effects: The BioPrint Approach. *Curr. Opin. Drug Discovery Dev.* **2003**, *6*, 470−80.

(24) http://www.cerep.fr/cerep/users/pages/Collaborations/Bioprint.asp (accessed Sept 2006).

(25) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(26) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541−5554.

(27) The above-mentioned data sets are also available via http://www.cheminformatics.org/ (accessed Sept 2006).

(28) Horvath, D. ComPharm − Automated Comparative Analysis of Pharmacophoric Patterns and Derived QSAR Approaches, Novel Tools in High Throughput Drug Discovery. A Proof of Concept Study Applied to Farnesyl Protein Transferase Inhibitor Design. In *QSPR/QSAR Studies by Molecular Descriptors*; Diudea, M., Ed.; Nova Science Publishers: New York, 2001; pp 395−439.

(29) http://www.chemaxon.com/jchem/doc/api/ (accessed Sept 2006).

(30) http://www.chemaxon.com/jchem/index.html?content=doc/user/Standardizer.html (accessed Sept 2006).

(31) http://www.chemaxon.com/marvin/chemaxon/marvin/help/calculator-plugins.html#pka (accessed Sept 2006).

(32) http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Sept 2006).

(33) http://www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed Sept 2006).

(34) http://www.maybridge.com/ (accessed Sept 2006).

(35) Christensena, J. G.; Burrows, J.; Salgiab R. c-Met as a Target for Human Cancer and Characterization of Inhibitors for Therapeutic Intervention. *Cancer Lett.* **2005**, *225*, 1−26.

(36) Vojkovsky, T.; Koenig, M.; Zhang, F.-J.; Cui, J. Tetracyclic Compounds as c-Met inhibitors. Patent WO2005004808, 2005.

(37) Koenig, M. Indolinonehydrazides as c-Met Inhibitors. Patent WO200500-5378, 2005.

(38) Compounds and activity data taken from the WOMBAT database of Sunset Molecular, Inc. (http://sunsetmolecular.com/products/?id=4) courtesy of Tudor I. Oprea, 2005.

(39) Altschul, S. F. Amino Acid Substitution Matrices from an Information Theoretic Perspective. *J. Mol. Biol.* **1991**, *219*, 555−65.

(40) Kubiny, H. Structure-Based Design of Enzyme Inhibitors and Receptor Ligands. Second European Workshop in Drug Design, Certosa di Pontignano, May 17−24, 1998; oral presentation.

(41) Hann, M. M.; Oprea, T. I. Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−63.

(42) Seifert, M. H. J. Assessing the Discriminatory Power of Scoring Functions for Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 1456−1465.

(43) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049−54.

CI6002416

Annexe E

Article 3 : Future Generation Computer Systems, 2007

paru dans « Future Generation Computer Systems » en mars 2007, 23(3), p. 398-409

A.-A. Tantar, N. Melab, E.-G. Talbi, B. Parent et D. Horvath,

A parallel hybrid genetic algorithm for protein structure prediction on the computational grid.

Available online at www.sciencedirect.com
ScienceDirect
Future Generation Computer Systems 23 (2007) 398–409
FGCS FUTURE GENERATION COMPUTER SYSTEMS
www.elsevier.com/locate/fgcs

A parallel hybrid genetic algorithm for protein structure prediction on the computational grid☆

A.-A. Tantar[a], N. Melab[a,*], E.-G. Talbi[a], B. Parent[b], D. Horvath[b]

[a] *Laboratoire d'Informatique Fondamentale de Lille, LIFL/CNRS UMR 8022, DOLPHIN Project - INRIA Futurs, Cité Scientifique, 59655 - Villeneuve d'Ascq Cedex, France*

[b] *CNRS UMR8576, Université des Sciences et Technologies de Lille, Bâtiment C9, Cité Scientifique 59655 - Villeneuve d'Ascq Cedex, France*

Received 2 February 2006; received in revised form 5 August 2006; accepted 7 September 2006
Available online 1 November 2006

Abstract

Solving the structure prediction problem for complex proteins is difficult and computationally expensive. In this paper, we propose a bicriterion parallel hybrid genetic algorithm (GA) in order to efficiently deal with the problem using the computational grid. The use of a near-optimal metaheuristic, such as a GA, allows a significant reduction in the number of explored potential structures. However, the complexity of the problem remains prohibitive as far as large proteins are concerned, making the use of parallel computing on the computational grid essential for its efficient resolution. A conjugated gradient-based Hill Climbing local search is combined with the GA in order to intensify the search in the neighborhood of its provided configurations. In this paper we consider two molecular complexes: the *tryptophan-cage* protein (Brookhaven Protein Data Bank ID 1L2Y) and *α-cyclodextrin*. The experimentation results obtained on a computational grid show the effectiveness of the approach.

Keywords: Protein structure prediction; Genetic algorithm; Hill climbing; Parallel computing; Grid computing

1. Introduction

Nowadays, grid computing is admitted as a powerful way to achieve high performance on computational-intensive applications. The protein structure prediction problem, further referred to as PSP, is one of the particularly interesting challenges of parallel computing on the computational grid. The problem consists in determining the ground-state conformation of a specified protein, given its amino acid sequence—the *primary structure*. In this context, the ground-state conformation term designates the associated tridimensional native form, known as zero energy *tertiary structure*. Addressing the mathematical model, paradigms based on quantum mechanics and the Schrödinger equation were developed in the literature, as well as empirical techniques based on classical dynamics—to be further discussed in the following sections.

Although there exist laboratory methods addressing the herein described problem, prohibitory costs and the long experimentation time required make them unfeasible for large scale application. As a consequence, computational protein structure prediction represents an interesting alternative, though complexity matters impose strong limitations. For a reduced size molecule composed of 40 residues, a number of 10^{40} conformations must be taken into account when considering, on average, 10 conformations per residue. Furthermore, if a number of 10^{14} conformations per second is explored, a time of more than 10^{18} years is needed for finding the native-state conformation. For example, for the *[met]-enkephalin* pentapeptide, composed of 75 atoms and having five amino acids, *Tyr-Gly-Gly-Phe-Met*, and 22 variable backbone dihedral angles, a number of 10^{11} local optima is estimated. Detailed

☆ The current article is developed as part of the **Conformational Sampling and Docking on Grids** project, supported by ANR (Agence Nationale de la Recherche—http://www.gip-anr.fr), under the coordination of Prof. El-Ghazali Talbi and reuniting LIFL (USTL-CNRS-INRIA), IBL (CNRS-INSERM) and CEA DSV/DRDC.

* Corresponding address: Université de Lille 1 - Cité Scientifique, CNRS/LIFL - INRIA DOLPHIN, Bâtiment M3 - Extension, 59655 Villeneuve d'Ascq, France.

E-mail addresses: tantar@lifl.fr (A.-A. Tantar), melab@lifl.fr (N. Melab), talbi@lifl.fr (E.-G. Talbi), benjamin.parent@ibl.fr (B. Parent), dragos.horvath@univ-lille1.fr (D. Horvath).

0167-739X/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2006.09.001

aspects concerning complexity matters were discussed in [20, 21], leading to the mention of the *Levinthal's paradox* [6] which states that, despite numerous pathways, *in vivo* molecular folding, for example, has a time scale magnitude of several milliseconds.

Notes on molecular structure prediction complexity may be found in [19]. Although it may not be possible to construct a general mathematical model for describing molecular structures, it may be inferred that no polynomial time resolution is possible if no or less *a priori* knowledge is employed. As a consequence, no simulation or resolution is possible unless extensive computational power is applied. Thus, a distributed grid approach is required.

Genetic algorithms are population-based metaheuristics that allow a powerful exploration of the conformational space. However, they have limited search intensification capabilities, which are essential for neighborhood-based improvement (the neighborhood of a solution refers to part of the problem's landscape). Therefore, the GA is combined with a conjugated gradient-based Hill Climbing local search method, in order to improve both the exploration and the intensification capabilities of the two techniques. In addition, the GA is parallelized in a hierarchical manner. Firstly, several GAs cooperate by exchanging their genetic material (parallel island model [3]). Secondly, as the fitness function of each GA is time-intensive the fitness evaluation phase of the GA is parallelized (parallel evaluation of the population model [3]). These two models are provided in a transparent way through the ParadisEO-CMW framework [1], dedicated to the reusable design of parallel hybrid metaheuristics on computational grids.

The interest in multicriterion structure prediction resides in result optimality and problem simplification. It can be argued that the native structure of a molecule should not be described through one unique conformation but through an ensemble of conformations, as in statistical mechanics [8]. As per environment interactions and the non-rigidity of a molecule's conformation, structural description may be performed by using a set of potentially transitory conformations. In this case, the transitory conformations are distributed at the base of a funnel-like energy landscape. As a consequence, relating to mesoscopic and macroscopic realm aspects, multicriterion analytical and computational models are extremely important for the complete *in silico* characterization of molecular systems.

The latter argument, concerning problem simplification, refers to molecular processes complexity, in terms of number of local optima—as mentioned above, a number of 10^{11} local optima is estimated for the *[met]-enkephalin* pentapeptide. The reduction of the number of local optima may be attained by transforming a monocriterion optimization problem into a multicriterion problem, experimental results in this respect being furnished in [7]. It should be mentioned that, at this time, the existing approaches focus on monocriterion definition terms for problem resolution.

The importance of the PSP problem is reinforced by the ubiquitousness of proteins in the living organisms, applications of computational protein structure prediction directed to computer assisted drug design and computer assisted molecular design. From a structural point of view, proteins are complex organic compounds composed of amino acid residues chains joined by peptide bonds. Proteins are involved in immune response mechanisms, enzymatic activity, signal transduction etc. Due to the intrinsic relation between the structure of a molecule and its functionality, the problem implies important consequences in medicine and biology related fields.

An extended referential resource for protein structural data may be accessed through the Brookhaven Protein Data Bank[1] [26]. For a comprehensive introductory article on protein structure, consult [9]. Also, for a glossary of terms, see [29].

In this paper, we propose a bicriterion genetic algorithm (GA), based on Newton's classical mechanics for performing molecular energy calculations. The proposed approach has been applied for two molecular complexes: the *tryptophan-cage* protein (Brookhaven Protein Data Bank ID 1L2Y) and *α-cyclodextrin*. The experimental results obtained on a computational grid demonstrate the effectiveness of the approach.

The remainder of the paper is organized as follows: a brief review on the related work is proposed in Section 2 indicating the main directions for solving the problem. Section 3 presents the basis for constructing the parallel GA approach—elementary theoretical elements are also presented. In Section 4, the ParadisEO-CMW framework is described, along with the subsidiary underlying middleware, Condor-MW, the final part of the corresponding section sketching the general implementation aspects. In Section 5, experimentation results are given with an introductory presentation of the GRID5000 computational grid. Section 6 comprises the conclusions.

2. Related work for the protein structure prediction problem (PSP)

In order to address the PSP problem, by analytical and computational means, a mathematical model that describes inter-atomic interactions must be constructed. The interactions to be considered are a resultant of electrostatic forces, entropy, hydrophobic characteristics, hydrogen bonding, etc. The interactions are quantified in terms of energy levels, relating to the internal energy of the molecule. Precise energy determination also relies on the solvent effect enclosed in the dielectric constant ϵ and in a continuum model based term.

A trade-off is accepted, opposing accuracy against the approximation level, varying from exact, physically correct mathematical formalisms to purely empirical approaches. The main categories to be mentioned are *de novo, ab initio* electronic structure calculations, semi-empirical methods and molecular mechanics based models. Hybrid and layered approaches were also designed, in order to reduce the amount of performed calculus to the detriment of accuracy.

The mathematical model describing molecular systems is formulated upon the Schrödinger equation, which makes use of molecular wavefunctions for modeling the spatio-temporal

[1] http://www.rcsb.org—Brookhaven Protein Data Bank; offers geometrical structural data for a large number of proteins.

probability distribution of constituent particles [10]. It should be noted that, though offering the most accurate approximation, the *Schrödinger* equation cannot be accurately solved for more than two interacting particles. For resolution related aspects, please consult [27,28]. Extended explanations for the herein exposed directions are available *via* [10–12,9].

Ab initio (first principles) calculations rely on quantum mechanics for determining different molecular characteristics, comprising no approximations and with no *a priori* required experimental data. Molecular orbital methods make use of *basis functions* for solving the *Schrödinger* equation. The high computational complexity of the formalism restricts their application area to systems composed of tens of atoms.

Semi-empirical methods substitute computationally expensive segments by approximating *ab initio* techniques. A decrease in the time required for calculus is obtained by employing simplified models for electron–electron interactions: *extended Hückel model*, *neglect of differential overlap*, *neglect of diatomic differential overlap*, etc.

Empirical methods rely upon molecular dynamics (*classical mechanics based methods*), and were introduced by Alder and Wainwright [16,17]. After more than a decade protein simulations were initiated on bovine pancreatic trypsin inhibitor—BPTI [18]. Empirical methods often represent the only applicable methods for large molecular systems, namely, proteins and polymers. Empirical methods do not make use of the quantum mechanics formalism, relying solely upon classical Newtonian mechanics, *i.e.* Newton's second law—the equation of motion. As to the basis of the considered approach, we should mention that, according to recent results [22,23], empirical methods exceed *ab initio* methods. Conceptually, molecular dynamics models do not dissociate atoms into electrons and nuclei but regard them as indivisible entities. The following list offers a few examples of molecular mechanics force fields:

- AMBER—*Assisted Model Building with Energy Refinement*;
- CHARMM—*Chemistry at HARvard Molecular Mechanics*;
- OPLS—*Optimized Potentials for Liquid Simulations*.

Also, hybrid and layered methods exist [13–15], connecting several methods through various computing architectures, in an attempt to obtain accurate results at low computational costs, and, consequently, in a reduced period of time.

3. A parallel hybrid metaheuristic for solving PSP

3.1. Multicriterion evolutionary algorithm basis

Evolutionary algorithms are stochastic search iterative techniques, with a large area of appliance—epistatic, multimodal, multicriterion and highly constrained problems [1]. Stochastic operators are applied for evolving the initial randomly generated population, in an iterative manner. Each generation undergoes a selection process, the individuals being evaluated by employing a problem specific fitness function.

Algorithm 3.1. EA pseudo-code.

```
Generate(P(0));
t := 0;
while not Termination_Criterion(P(t)) do
  Evaluate(P(t));
  P'(t)     := Selection(P(t));
  P'(t)     := Apply_Reproduction_Ops(P'(t));
  P(t + 1)  := Replace(P(t), P'(t));
  t := t + 1;
endwhile
```

The pseudo-code above shows the generic components of an EA. The main subclasses of EAs are the genetic algorithms, evolutionary programming, evolution strategies, etc.

Due to the nontriviality of the addressed problems, requiring extensive processing time, different approaches were designed in order to reduce the computational costs. Complexity is also addressed by developing specialized operators or hybrid and parallel algorithms. We have to note that the parallel affinity of the EAs represents a feature determined by their intrinsic population-based nature. More precisely, the main parallel models are the island synchronous cooperative model, the parallel evaluation of the population and the distributed evaluation of a single solution. For a complete overview on parallel and grid specific metaheuristcs refer to [1–4].

3.2. Multicriterion optimization context

A basic introduction to multicriterion theoretical tools is now presented. A succinct overview of existing research directions in multicriterion optimization may be found in [30].

The solution of a multicriterion optimization problem is represented by a multitude of individual feasible solutions—a *Pareto-optimal front*, to be defined in the following lines. A solution, identified as a composing point of a Pareto front, is designated as a *Pareto* point.

Definition 1. Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in A$ be two feasible solutions for a multicriterion problem $\mathcal{P}$, and $f : A \to B$, a cost function. We say that solution $\boldsymbol{x}_1$ dominates solution $\boldsymbol{x}_2$, denoted as $\boldsymbol{x}_1 < \boldsymbol{x}_2$, if the following are simultaneously true:

$$\forall i \in [1, \ldots, t],\ f_i(x_1) \leq f_i(x_2);$$
$$\exists i \in [1, \ldots, t],\ f_i(x_1) < f_i(x_2).$$

The solutions $\boldsymbol{x}_1, \boldsymbol{x}_2$ are said to be non-dominated with respect to each other if neither of the $\boldsymbol{x}_1 < \boldsymbol{x}_2, \boldsymbol{x}_2 < \boldsymbol{x}_1$ relations are true, *i.e.* neither solution dominates the other.

Definition 2. Let F be a set of solutions for a multicriterion problem $\mathcal{P}$, $F \subseteq A$. It is said that F is a *Pareto-optimal* set (or front) if $\forall x \in F$ and $\forall x' \in A - F, x < x'$.

Examples of domination relations may be found in Fig. 1, while Fig. 2 illustrates a Pareto front example.

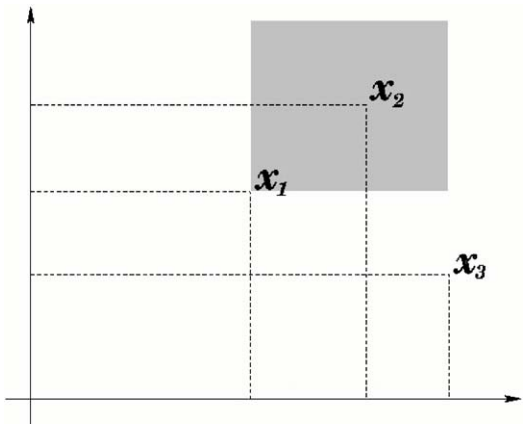


Fig. 1. x_1 dominates x_2; x_1 non-dominated with x_3 and x_2 non-dominated with x_3.

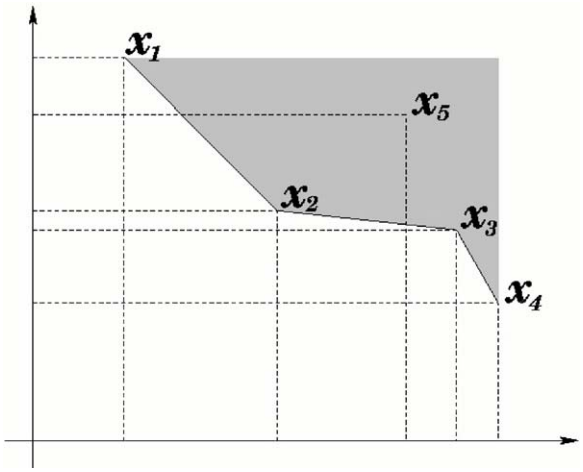


Fig. 2. Pareto front formed of: x_1, x_2, x_3, x_4; supported points (points located on the convex hull enclosing the entire set of solutions): x_1, x_2, x_4; non-supported point (point at the interior of the convex hull): x_3; dominated point: x_5.

3.3. Problem formulation and encoding

The algorithmic resolution of the PSP, in heuristic context, is directed through the exploration of the molecular energy surface. The sampling process is performed by altering the backbone structure in order to obtain different structural conformations.

Different encoding approaches were considered in the literature, the trivial approach considering the direct coding of atomic Cartesian coordinates [24]. The main disadvantage of direct coding is the fact that it requires filtering and correcting mechanisms, inducing non-negligible affected times. Moreover, by using amino acid based codings [25], hydrophobic/hydrophilic models were developed. In addition, several variations exist, making use of all-heavy-atom coordinates, C_α coordinates or backbone atom coordinates, where amino acids are approximated by their centroids.

For the herein described method, an indirect, less error-prone, torsional angle based representation was preferred, knowing that, for a given molecule, there exists an associated

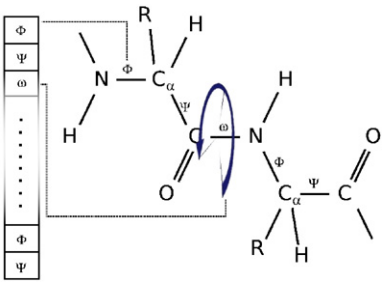


Fig. 3. Chromosome encoding based on specifying the backbone torsional angles.

sequence of atoms. More specifically, each individual is coded as a vector of torsion angle values— Fig. 3.

The defined number of torsion angles represents the degree of flexibility. Apart from torsion angles which move less than a specified parameter, all torsions are rotatable. Rotations are performed in integer increments, energy quantification of covalent bonds and non-bonded atom interactions being used as optimality evaluation criterion.

3.4. A parallel genetic algorithm for solving PSP

Genetic Algorithms (GAs) represent Darwinian-evolution inspired methods, a random population of individuals evolving in generations through different strategies in order for convergence to be achieved, with respect to optimality criteria. The *genotype* represents the raw encoding of individuals while the *phenotype* encloses the coded features. For each generation, individuals are selected on a fitness basis, genotype alteration being performed by means of crossover and mutation operators. Applying the genetic operators has as a result the modification of the population's structure as to intensify exploration inside a delimited segment or for diversification purposes.

The herein described algorithm comes as the result of a meta optimization process [5], experiments being performed for identifying optimal parametrization. A parallel design is considered, the general sustaining architecture of the developed algorithm conceptually following the generic parallel metaheuristic sketch, previously presented.

The granularity of the problem, as a counterpart for the computationally expensive fitness evaluations, biased the resolution pattern towards a parallel, island-model approach. As a consequence, several populations evolve on a master machine, fitness function estimations being distributed on remotely available computing units. We have to note that the evaluation of the fitness function consists of several stages, including the calculation of Cartesian atomic coordinates, inter-atomic distances determination etc. A distributed fitness calculation does not represent an option, incurring a significant synchronization overhead. Common one-point and two-point crossover and mutation operators were used.

3.5. Fitness function

The function to be optimized, under the bicriterion auspices, is computed by making use of bonded atom energy and non-bonded atom energy, as distinct entities. The result obtained is

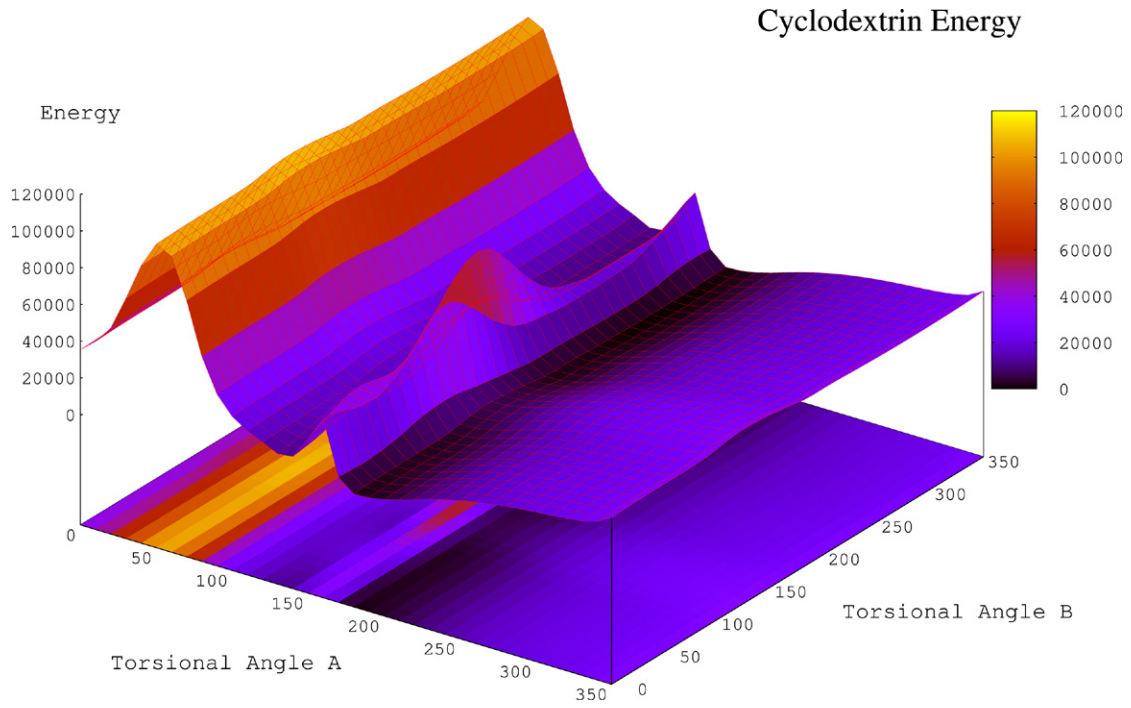


Fig. 4. Energy surface for *α-cyclodextrin*. High energy points are depicted in light colors, the low energy points being identified by the dark areas. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

compared with a Pareto front of solutions, the feasibility of a given individual being related to the dominance concept.

An intuitive reasoning leads to the fact that the bonded and the non-bonded energy terms are antagonist (verified through the performed experiments), although *no formal* demonstration exists in the literature. Hence, it may be stated that the problem qualifies for multicriterion optimization. The quantification of energy is performed by using empirical molecular mechanics, under the *CHARMM* realm as follows:

$$E_{\text{bonded}} = \sum_{\text{bonds}} K_b(b-b_0)^2 + \sum_{\text{bondangle}} K_\theta(\theta-\theta_0)^2 + \sum_{\text{torsion}} K_\phi(1-\cos n(\phi-\phi_0))$$

$$E_{\text{nbonded}} = \sum_{\text{Van der Waals}} \frac{K^a_{ij}}{d^{12}_{ij}} - \frac{K^b_{ij}}{d^6_{ij}} + \sum_{\text{Coulomb}} \frac{q_i q_j}{4\pi\varepsilon d_{ij}} + \sum_{\text{desolvation}} \frac{K q_i^2 V_j + q_j^2 V_i}{d^4_{ij}}$$

where E_{bondend} and E_{nbonded} represent the energy of the bonded and non-bonded contributions respectively.

The involved factors model oscillating entities, the interatomic forces being conceptually simulated by considering interconnecting springs between atoms. At this point, a specific constant is associated with each type of interaction, notationally denoted by K_{inter}. An optimal value for the considered entity (bond, angle, torsion) is introduced in the equation as reference for the variance magnitude—$(T - T_0)$. T stands for the experimentation value, while T_0 specifies the natural, experimentally observed value, when the entity is pulled out of its context.

In more specific terms, b represents the bond length, θ the bond angle, ϕ the torsion angle and q_a, d_{ij} and V_p the electrostatic charge associated with a given atom, the distance between the i and the j atoms and a volumetric measure for the p atom respectively.

An example of *α-cyclodextrin* energy surface is given in Fig. 4. The set of corresponding molecular conformations was obtained by modifying a random initial conformation. More specifically, an arbitrary conformation has been generated, subsequently, two torsional angles being chosen at random. For each of the two torsional angles, values between 0 and 360 have been considered, in 10° increments, all the other torsional angles being maintained rigid. Thus, 1225 conformations were obtained—the lighter areas on the obtained surface correspond to high-energy conformations. Furthermore, an energy-map representation is given, in the XY-plane—only the dark regions are meaningful.

Although smooth, the obtained surface is the result of only two torsional angles variation. The hyper-surface, generated by varying the entire set of torsional angles, has an extremely rough landscape, with a large number of local optima.

Fig. 5 depicts the bonded and non-bonded atom derived energies, corresponding to the previous energy surface, shown in 4. The energy surfaces are computed as given by the previously exposed force field.

As can be seen from the figure, the non-bonded atom derived energy component has large values in comparison with the bonded atom derived energy component. The high-energy values for the non-bonded component are determined by the large number of non-bonded interactions, as pairs of atoms are considered.

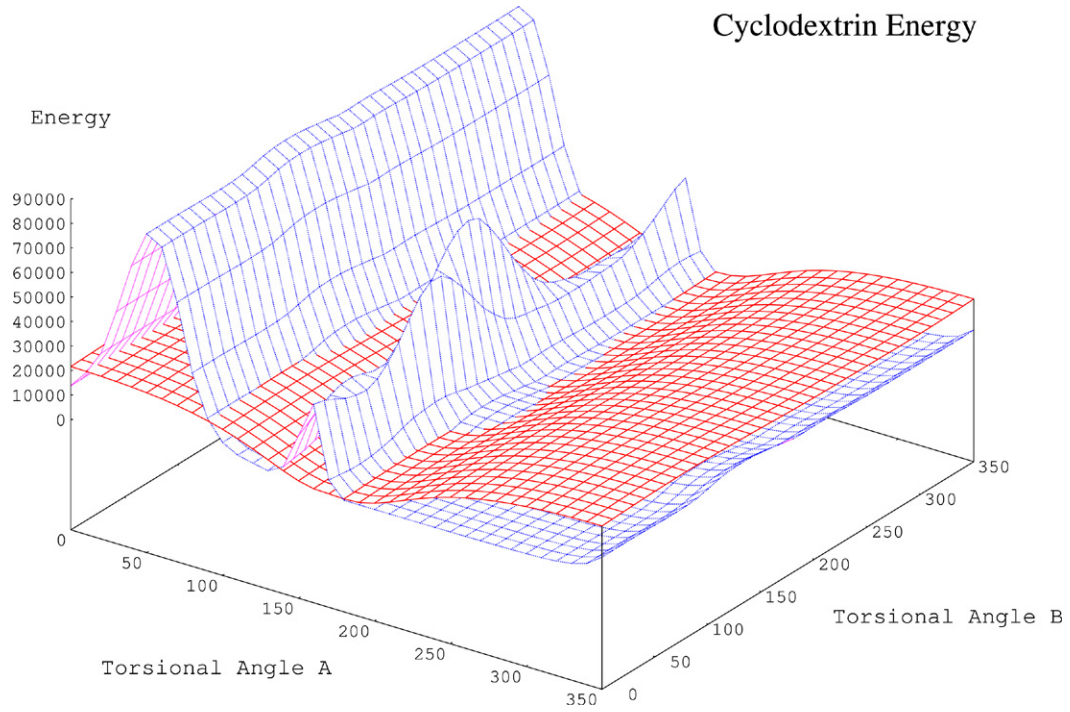


Fig. 5. The bonded atom derived energy component is represented by the blue grid. The non-bonded atom derived energy component is given by the smoother surface, with red grid lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.6. Hybridization with a Hill Climbing local search

The developed method has as backbone structure a hybrid architecture, combining a genetic algorithm with a conjugated gradient-based Hill Climbing local search method—a *Lamarckian* optimization technique.

The exploration and the intensification capabilities of the genetic algorithm do not suffice as paradigm, when addressing rough molecular energy function landscapes. Small variations of the torsion angle values may generate extremely different individuals, with respect to the fitness function. As a consequence, a nearly optimal configuration, considering the torsion angle values, may have a very high energy value, and thus it may not be taken into account for the next generations.

In order to correct the above exposed problem, a conjugated-gradient based method is applied for local search, alleviating the drawbacks determined by the conformation of the landscape. Fig. 6 was obtained by applying the local search technique for each of the conformations that were used for the *α-cyclodextrin* energy surface in Fig. 4.

Although reducing both energies, for the bonded and non-bonded type interactions, the non-bonded energy component still represents the major part of the total energy, as can be seen in Fig. 7.

4. ParadisEO-CMW based implementation

4.1. The ParadisEO framework

The ParadisEO[2] framework is dedicated to the reusable design of parallel hybrid meta-heuristics by providing a broad range of features, including EAs, local search methods, parallel and distributed models, different hybridization mechanisms, etc. The rich content and utility of ParadisEO increases its usefulness.

ParadisEO is a C++ LGPL white-box open source framework, based on a clear conceptual separation of the meta-heuristics from the problems they are intended to solve. This separation, and the large variety of implemented optimization features, allow a maximum code and design reuse. Changing existing components and adding new ones can be easily done, without impacting the rest of the application.

ParadisEO is one of the rare frameworks that provide the most common parallel and distributed models, portable on distributed-memory machines and shared-memory multi-processors, as they are implemented using standard libraries such as MPI, PVM and PThreads. The models can be exploited in a transparent way—one has just to instantiate its associated ParadisEO components. The user has the possibility of choosing, by a simple instantiation, the MPI or the PVM for the communication layer. The models have been validated on academic and industrial problems, and the experimental results demonstrate their efficiency [4].

4.2. The ParadisEO-CMW framework

The ParadisEO-CMW framework targets non-dedicated environments, having as sustaining structure the ParadisEO framework and the Condor-MW middleware.

The Condor[3] system [33,34] is a high-throughput computing (HTC) system that deals with heterogeneous computing

[2] http://www.lifl.fr/~cahon/paradisEO/common.

[3] http://www.cs.wisc.edu/condor/condorg.

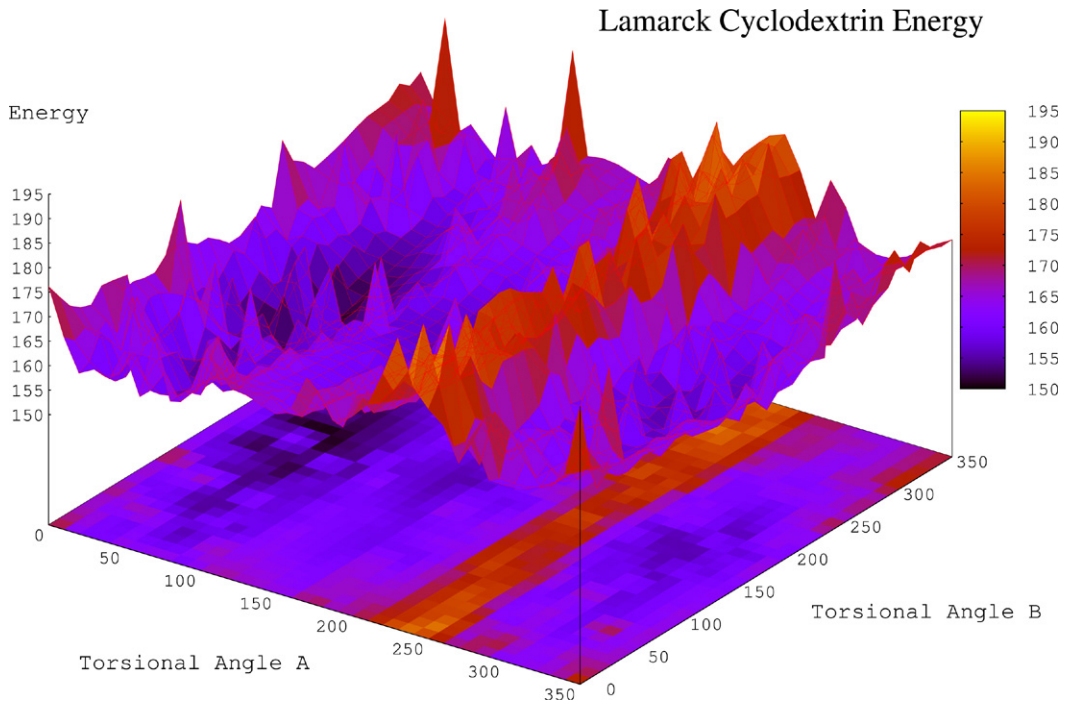


Fig. 6. Energy surface obtained after applying a Lamarck local search on the initial set of conformations.

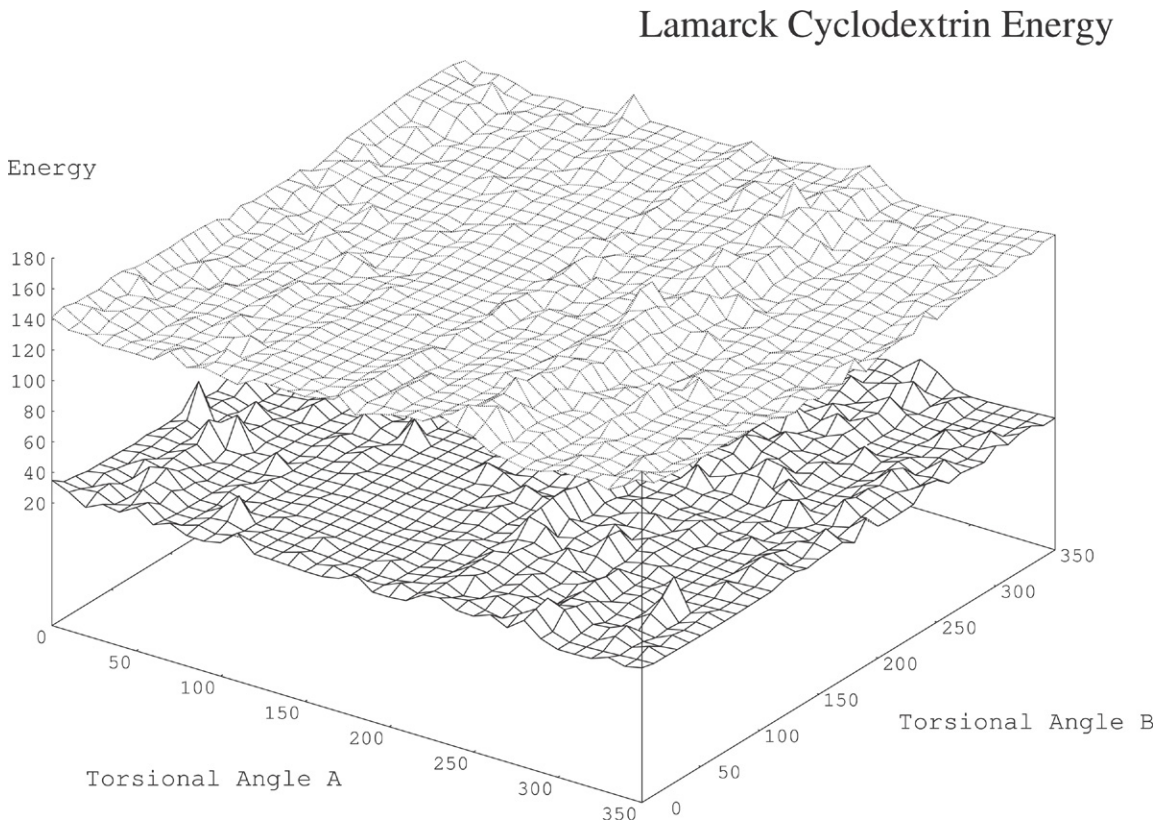


Fig. 7. The two components of the energy surface for the conformations obtained after applying the Lamarck local search. The upper and the lower surface correspond to the non-bonded atom derived energy, and, to the bonded atoms derived energy, respectively.

resources and multiple users. It allows the management of non-dedicated and volatile resources, by deciding their *availability*, using both the average CPU load and the information about the recent use of some peripherals, like the keyboard and the mouse. An environment including such resources is said to be adaptive, since tasks are scheduled among idle resources, and dynamically migrated when some resources get used or failed. In addition, Condor-PVM uses some sophisticated

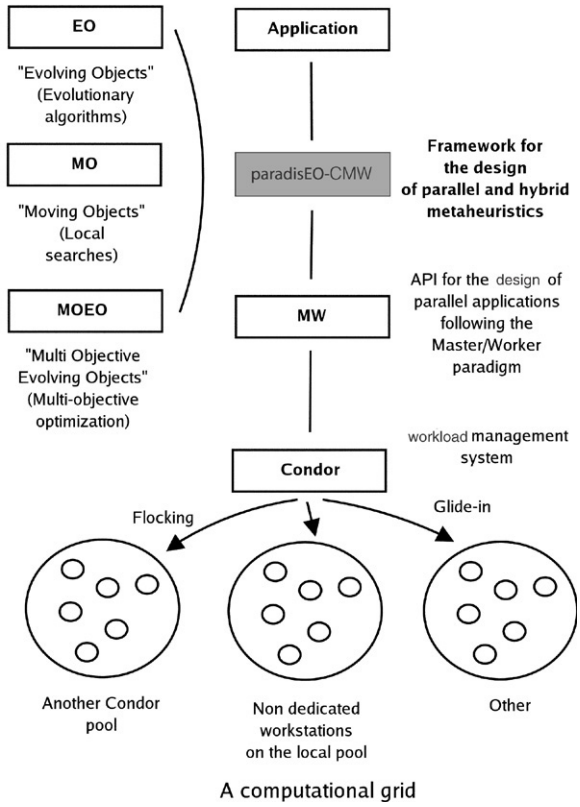


Fig. 8. A layered architecture of ParadisEO-CMW.

techniques [31] like matchmaking and checkpointing. These allow us, respectively, to associate job requirements and policies on resources owners, and to periodically save/restart the state of/from running jobs.

MW [32] is a software framework allowing an easy development of Master–Worker applications for computational grids. MW is a set of C++ abstract classes including interfaces to application programmers and Grid-infrastructure developers. Grid-enabling an application with MW, or porting MW to a new grid software toolkit, consists in re-implementing a small number of virtual functions. In MW, the infrastructure interface provides access to communication and resource management. The communication is performed between the master and the workers. The resource management encompasses: available resource request and detection, infrastructure querying to get information about resources, fault detection, and remote execution. These basic resource management services are provided by Condor-PVM.

One of the major design goals of MW is to ensure a maximum programmability, meaning that the users should easily be able to interface an existing code with the system. Therefore, porting ParadisEO to Condor-MW can be easily done through the use of the infrastructure and application programming interfaces provided by MW. Moreover, the coupling is facilitated by the fact that the two frameworks are written in C++. The architecture of ParadisEO-CMW is layered as is illustrated in Fig. 8.

From a top-down view, the first level supplies the optimization problems to be solved using the framework. The second level represents the ParadisEO framework, including optimization solvers, embedding single and multicriterion meta-heuristics (evolutionary algorithms and local searches). The third level provides interfaces for Grid-enabled programming and for access to the Condor infrastructure. The fourth and lowest level supplies communication and resource management services.

An important issue to deal with in Grid computing is the fault-tolerance. MW automatically reschedules unfinished tasks as they were running on processors that failed. This cannot be applied to the master process that launches and controls tasks on worker nodes. Nevertheless, a couple of primitives are provided to fold up or unfold the whole application, enabling the user to save/restart the state to/from a file stream. Dealing with meta-heuristics, these functionalities are easily investigated. Checkpointing most of the meta-heuristics is straightforward. It consists at least in saving the current solution(s), the best one found since the beginning of the search, the continuation criterion (*e.g.* the current iteration for a generational counter) and then some additional parameters controlling the behavior of the heuristic. In ParadisEO-CMW, default checkpoint policies are initially associated to the deployed meta-heuristics.

4.3. Implementation

The implementation relies on invariant elements provided by the ParadisEO-CMW framework, providing support for the insular model approach, as well as for distributed and parallel aspects concerning the parallel population evaluation. In this context, deployment related aspects are transparent, the focus being oriented on the application-specific elements.

The main steps to be performed, in order to configure the environment and to deploy the algorithm, consist in specifying the individual's encoding, the specific operators and the fitness function. Furthermore, elements concerning selection mechanisms and replacement strategies must be specified, along with configuration parameters (number of individuals, number of generations etc).

5. Experiments and results

For the developed application, the deployment has been performed on a layered framework design, the composing elements being the following: **Condor**, **MW—Master–Worker**, **ParadisEO-CMW**.

The underlying support for performing the experiments was GRID5000, a French nationwide experimental grid, connecting several sites which host clusters of PCs interconnected by RENATER[4] (the French academic network). The GRID5000 is promoted by CNRS, INRIA and several universities.[5]

By the end of 2006 the GRID should gather 2500 processors with 2.5 TB of cumulated memory and 100 TB of non-volatile storage capacity. Inter-connections sustain communications of

[4] Réseau National de Télécommunications pour la Technologie, l'Enseignement et la Recherche—http://www.renater.fr.

[5] CNRS—http://www.cnrs.fr/index.html; INRIA—http://www.inria.fr.

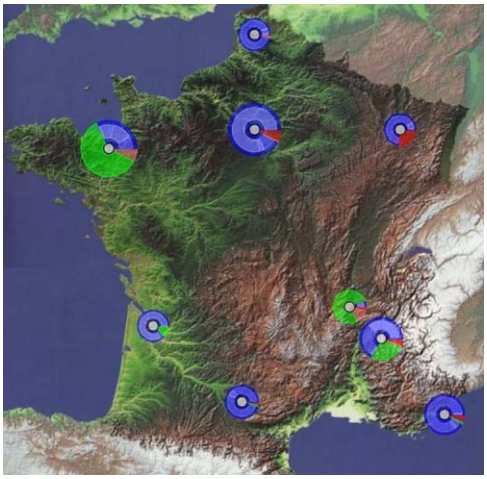


Fig. 9. GRID5000 centers are marked in grey, the colored disks around them offering a visual feedback regarding the status of their afferent workstations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.5 Gbps (10 Gbps soon). The GRID5000 infrastructure offers several tools for controlling, manipulating and supervising activities, Fig. 9 representing a real-time snapshot of the GRID.

The target point to be achieved is a marker-stone of 5000 processors for 2007—at this moment there are almost 2000 processors at this time being, regrouping nine centers: Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Sophia-Antipolis, Toulouse. The following results were obtained by performing deployments on the Lille cluster of GRID5000.

The addressed molecular complexes for grid deployment tests were *tryptophan-cage* (Protein Data Bank ID 1L2Y) and *α-cyclodextrin*. The *trp-cage* miniproteins present particular fast folding characteristics, while *cyclodextrins*, in α, β or γ conformations, are important for drug-stability applications, being used as drug protectors against micro-environment interactions or as homogeneous distribution stabilizers etc.

Structural profile of the *tryptophan-cage* protein: an α-helical N-terminal region, a short helix and a *polyproline* II helix at the C-terminus wrapping around for packing the *Trp* residue within a compact hydrophobic core [35]. *Cyclodextrins*, as non-reducing macrocyclic *oligosaccharides*, are constituted as *D-glucopyranosyl* units interconnected through $\alpha - (1, 4)$ glycosidic links. The ensemble builds as a toroidal structure with hydrophobic interior.

Table 1 offers information regarding the number of active elements used when executing the algorithm—determining the degree of flexibility considered for each of the molecules and, consequently, the dimension of the conformational space. The complexity of the model augments in concordance with the number of active elements—the table lists the considered active elements for each of the molecules under study. The last two lines offer the initial, respectively the final, number of interactions between non-bonded atoms. A cut-off is performed in order to reduce complexity, having as basis inter-atomic distances (interactions between atoms too far apart are ignored,

Table 1
Active elements for the performed experiments

	Tryptophan-cage	α-cyclodextrin
Active bonds	0	7
Active angles	0	40
Active torsions	524	336
Initial non-bonded inter.	44 369	7119
Final non-bonded inter.	44 223	7119

Table 2
Execution times for the performed experiments

No. of CPUs	Tryptophan-cage	α-cyclodextrin
80	79.380 s	46.600 s
60	87.060 s	48.340 s
30	162.550 s	79.370 s
10	459.880 s	270.420 s
5	1018.940 s	464.560 s
2	3069.830 s	1416.570 s

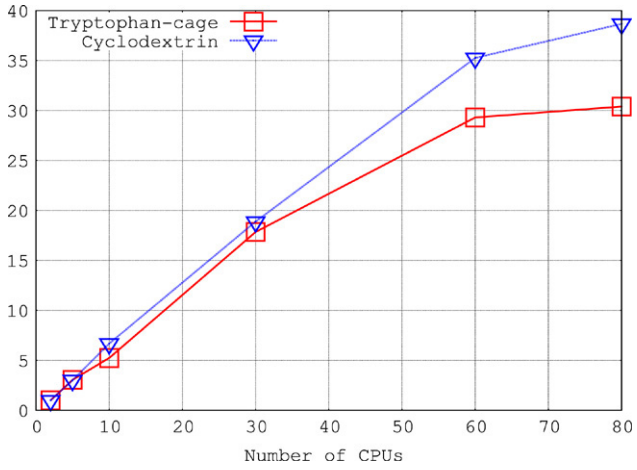


Fig. 10. Speed-up for the tryptophan-cage protein—marked with red rectangles—and α-cyclodextrin—blue triangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

as they cannot contribute significantly, in energy terms). We have to note that energy calculations for the non-bonded atoms set represent the main computational factor, as pairs of non-bonded atoms must be considered. In conjunction with the initial discussion on computational complexity, present in the introduction, the presented data confirm once more the need for a massive parallel computing environment.

In the followings lines, preliminary results are given, execution times for several performed tests being listed in Table 2. For each deployment, identical biprocessor machines were used, the number of computing units being listed on the left outer column. At the same time, the speed-up is depicted in Fig. 10—we are to remember that biprocessor machines were used, the enclosed data relating to distribution aspects.

Figs. 11 and 12 graphically represent the obtained Pareto fronts for the two above mentioned molecular systems—the Pareto points are marked by the blue triangles.

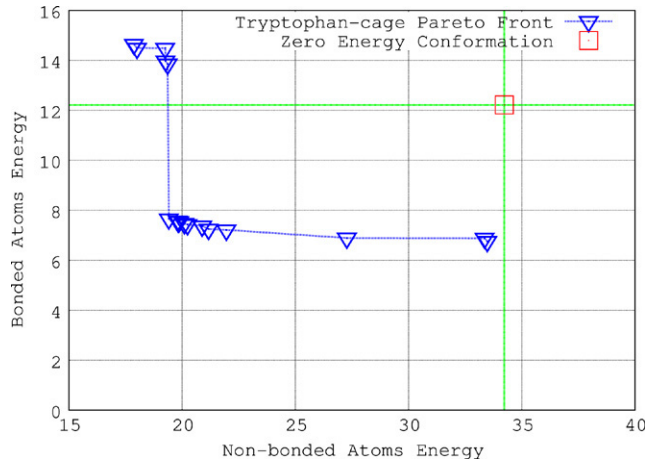


Fig. 11. 1L2Y Pareto front. Zero-energy conformation: 46.446 (non-bonded energy: 34.230, bonded energy: 12.216).

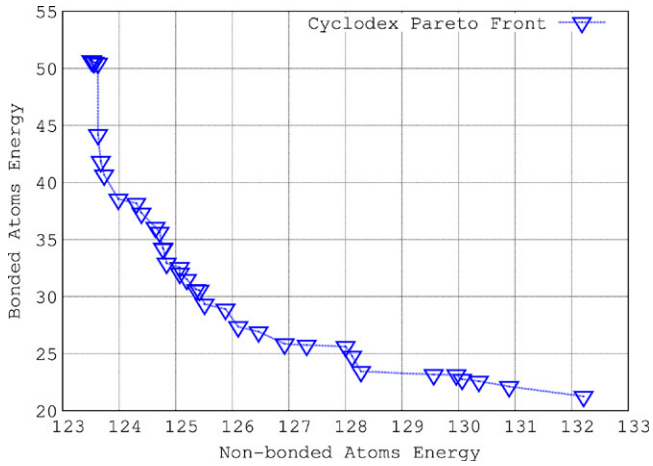


Fig. 12. α-cyclodextrin Pareto front. Zero-energy conformation: 242.157 (non-bonded energy: 216.579, bonded energy: 25.578).

Note—configuration for each of the machines: AMD Opteron(tm) Processor, 2193.504 MHz, 1024 kB *of cache and* 4 GB *of memory*.

In this context, the Pareto points correspond to metastable conformations, given that, at the end of its evolution, the algorithm significantly approaches a low energy level, close to the ground-state energy. Transitions may occur among close low-level energy metastable conformations, determined by the total energy of the molecule, driving to stability. Improvements may be effected by conducting further research on specialized operators capable of leading the search process towards regions of the search space corresponding to metastable conformations, combining efficient sampling with fast local search techniques.

As for the structure of the obtained Pareto fronts, there are several cases that deserve further research and which are worth discussing. The sparse structure is the combined result of the conformational sampling mechanism and of the energy-landscape structure. Thus, considering neighbor conformations with almost identical structure, a set of intermediary conformations might exist, though the sampling mechanism might have missed their associated region of space. The previous effect is driven by the granularity of the conformational sampling mechanism.

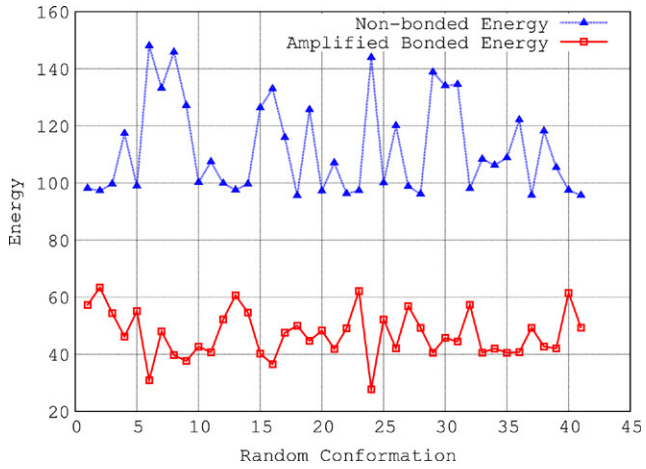


Fig. 13. Improvements in the value of a function generally attract a degradation in the value of the other function.

At the opposite extreme, as can be seen from Fig. 13, there are cases in which the degradation of one energy function does not incur improvements in the complementary energy function. In such cases, both energy functions undergo a degradation, in energy level terms. As a consequence, several neighboring conformations, having almost identical structure, might be separated by high potential barriers, in which case, no Pareto solutions exist between the conformations. This latter case is also determined by how close the mentioned conformations are to local optima, and the granularity considered when representing the torsional angles, with respect to energy variations.

6. Conclusions and future work

Multicriterion problems in general, and protein structure prediction under bicriterion aspects in particular, remain an open research field due to complexity matters, and of extreme importance in multiple domains. Mesoscopic and macroscopic characteristics represent the product of statistical interaction of an ensemble of near-optimal molecular conformations, a more complete description being achievable by defining not only the ground-state energy conformation of a molecule but also the ensemble of potential low-energy conformations.

The reported grid-enabled method offers a proof of feasibility, distributed techniques sustaining complex simulations. Multicriterion approaches, though potentially inducing augmented complexity, provide more accurate solutions for real-life problems, overcoming in particular cases the limitations of monocriterion resolution patterns. At this moment, experimentation and research are underway for specialized operators, exploiting directed mutation operators and approximative models as well as novel force fields. We also plan to tackle larger molecular complexes using parallel hybrid GAs on a larger computational grid. In this case, the exploitation of the two

parallel models of GAs in a hierarchical way requires several thousands of processors.

References

[1] S. Cahon, N. Melab, E.-G. Talbi, An enabling framework for parallel optimization on the computational grid, in: Proc. 5th IEEE/ACM Intl. Symposium on Cluster Computing and the Grid, CCGRID'2005, Cardiff, UK, 9–12 May, 2005.

[2] E.-G. Talbi, A taxonomy of hybrid metaheuristics, Journal of Heuristics 8 (2002) 541–564.

[3] E. Alba, G. Luque, E.-G. Talbi, N. Melab, in: E. Alba (Ed.), Metaheuristics and Parallelism, John Wiley and Sons, 2005.

[4] S. Cahon, N. Melab, E.-G. Talbi, ParadisEO: A framework for the reusable design of parallel and distributed metaheuristics, Journal of Heuristics 10 (2004) 357–380.

[5] B. Parent, A. Kökösy, D. Horvath, Optimized evolutionary strategies in conformational sampling, Journal of Soft Computing (2006).

[6] C. Levinthal, How to fold graciously, in: J.T.P. DeBrunner, E. Munck (Eds.), Mossbauer Spectroscopy in Biological Systems (Proceedings of a Meeting Held at Allerton House, Monticello, Illinois), University of Illinois Press, 1969, pp. 22–24.

[7] J.D. Knowles, D.W. Corne, Reducing local optima in single-objective problems by multi-objectivization, in: E. Zitzler, et al. (Eds.), Proc. First International Conference on Evolutionary Multi-criterion Optimization, EMO'01, Springer, Berlin, 2001, pp. 269–283.

[8] B. Ma, S. Kumar, C.-J. Tsai, R. Nussinov, Folding funnels and binding mechanisms, Protein Engineering 12, 713–720.

[9] A. Neumaier, Molecular modelling of proteins and mathematical prediction of protein structure, SIAM Review 39 (1997) 407–460.

[10] H. Dorsett, A. White, Overview of molecular modelling and ab initio molecular orbital methods suitable for use with energetic materials, Department of Defense, Weapons Systems Division, Aeronautical and Maritime Research Laboratory, DSTO-GD-0253, Salisbury South Australia, September 2000.

[11] A. White, F.J. Zerilli, H.D. Jones, Ab initio calculation of intermolecular potential parameters for gaseous decomposition products of energetic materials, Department of Defense, Energetic Materials Research and Technology Department, Naval Surface Warfare Center, DSTO-TR-1016, Melbourne Victoria 3001 Australia, August 2000.

[12] P. Sherwood, Hybrid quantum mechanics/molecular mechanics approaches, in: J. Grotendorst (Ed.), Modern Methods and Algorithms of Quantum Chemistry, Proceedings, 2nd edition, in: NIC Series, vol. 3, John von Neumann Institute for Computing, Jülich, ISBN: 3-00-005834-6, 2000, pp. 285–305.

[13] T. Vreven, K. Morokuma, Ö. Farkas, H.B. Schlegel, M.J. Frisch, Geometry optimization with QM/MM, ONIOM, and other combined methods. I. Microiterations and constraints, Journal of Computational Chemistry 24 (2003) 760–769.

[14] H. Kikuchi, R.K. Kalia, A. Nakano, P. Vashishta, H. Iyetomi, S. Ogata, T. Kouno, F. Shimojo, K. Tsuruta, S. Saini, Collaborative Simulation Grid: Multiscale Quantum-Mechanical/Classical Atomistic Simulations on Distributed PC Clusters in the US and Japan, IEEE, 2002.

[15] A. Nakano, R.K. Kalia, P. Vashishta, T.J. Campbell, S. Ogata, F. Shimojo, S. Saini, Scalable atomistic simulation algorithms for materials research, SC2001 November 2001, Denver (c) 2001 ACM.

[16] B.J. Alder, T.E. Wainwright, Journal of Chemical Physics 27 (1957) 1208.

[17] B.J. Alder, T.E. Wainwright, Journal of Chemical Physics 31 (1959) 459.

[18] J.A. McCammon, B.R. Gelin, M. Karplus, Nature 267 (1977) 585.

[19] J. Thomas Ngo, J. Marks, Computational complexity of a problem in molecular-structure prediction, Protein Engineering 5 (4) (1992) 313–321.

[20] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, M. Yannakakis, On the Complexity of Protein Folding.

[21] P.-Y. Calland, On the structural complexity of a protein, Protein Engineering 16 (2) (2003) 79–86.

[22] E.E. Lattman, CASP4, Proteins 44 (2001) 399.

[23] R. Bonneau, J. Tsui, I. Ruczinski, D. Chivian, C.M.E. Strauss, D. Baker Rosetta, CASP4: Progress in ab-initio protein structure prediction, Proteins 45 (2001) 119–126.

[24] A. Rabow, H. Scheraga, Protein Science 5 (1996) 1800–1815.

[25] N. Krasnogor, W. Hart, J. Smith, D. Pelta, Protein structure prediction problem with evolutionary algorithms, in: Proc. of the Genetic and Evolutionary Computation Conference, 1999.

[26] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E. Meyer, M.D. Bryce, J.R. Rogers, O. Kennard, T. Shikanouchi, M. Tasumi, The protein data bank: a computer-based archival file for macromolecular structures, Journal of Molecular Biology 112 (1977) 535–542.

[27] A.L. Islas, C.M. Schober, Multi-symplectic integration methods for generalized Schrödinger equations, Future Generation Computer Systems 19 (2003) 403–413.

[28] B.E. Moore, S. Reich, Multi-symplectic integration methods for Hamiltonian PDEs, Future Generation Computer Systems 19 (2003) 395–402.

[29] H. Van de Waterbeemd, R.E. Carter, G. Grassy, H. Kubinyi, Y.C. Martin, M.S. Tute, P. Willett, Glossary of terms used in computational drug design, Pure and Applied Chemistry 69 (5) (1997) 1137–1152.

[30] J.L. Cohon, in: J.S. Gero (Ed.), Multicriteria Programming: Brief Review and Application, Journal of Design Optimization (1985).

[31] M. Livny, J. Basney, R. Raman, T. Tannenbaum, Mechanisms for high throughput computing, Speedup Journal 11 (1) (1997).

[32] J. Linderoth, S. Kulkarni, J.P. Goux, M. Yoder, An enabling framework for master–worker applications on the computational grid, in: Proc. of the 9th IEEE Symposium on High Performance Distributed Computing, HPDC9, Pittsburgh, PA, August, 2000, pp. 43–50.

[33] D. Thain, T. Tannenbaum, M. Livny, Condor and the Grid, in: Grid Computing: Making the Global Infrastructure a Reality, John Wiley & Sons, December 2002.

[34] D. Thain, T. Tannenbaum, M. Livny, Distributed computing in practice: the condor experience, Concurrency and Computation: Practice & Experience (2004).

[35] L. Qiu, S.J. Hagen, Internal friction in the ultrafast folding of the tryptophan cage, Chemical Physics 312 (2005) 327–333.

A.-A. Tantar received the Master's degree from the Faculty of Computer Science, "A.I. Cuza" University of Iasi, Romania. He is currently a Ph.D. student within the OPAC team at Laboratoire d'Informatique Fondamentale de Lille (LIFL, Université de Lille 1). He is involved in the DOLPHIN project of INRIA Futurs. His major research interests include parallel and grid computing, and combinatorial optimization algorithms and applications.

N. Melab received the Master's, Ph.D. and HDR degrees in computer science, both from the Laboratoire d'Informatique Fondamentale de Lille (LIFL, Université de Lille 1). He is a Professor at Université de Lille 1 and a member of the OPAC team at LIFL. He is involved in the DOLPHIN project of INRIA Futurs. He is particularly a member of the Steering Committee of the French Nation-Wide project Grid5000. His major research interests include parallel and grid computing, combinatorial optimization algorithms and applications and software frameworks.

E.-G. Talbi received the Master's and Ph.D. degrees in computer science, both from the Institut National Polytechnique de Grenoble. He is presently Professor in computer science at Polytech'Lille (Université de Lille 1), and researcher in Laboratoire d'Informatique Fondamentale de Lille. He is the leader of OPAC team at LIFL and the DOLPHIN project of INRIA Futurs. He took part to several CEC Esprit and national research projects. His current research interests are mainly parallel and grid computing, combinatorial optimization algorithms and applications and software frameworks.

B. Parent is an engineer from the "Institut Supérieur d'Electronique et du Numerique" (Lille) and got his Master's degree in cybernetics and computer science from the "Ecole Centrale de Lille". Currently doing his Ph.D. in Biology and Biophysics, his main research interests involve the study and development of analysis and optimization algorithms for highly-dimensional, non-linear problems.

D. Horvath—Chemical engineer (Univ. Babes-Bolyai Cluj) 1991, Master & Ph.D. (Joint European Lab Pasteur Institute Lille—Free University of Brussels) 1996, Head of Chemoinformatics at Cerep (1997–2003), currently CNRS scientist. Development of methodology in chemoinformatics (molecular descriptors, similarity metrics, QSAR models) and molecular modeling (conformational sampling, docking). Virtual Screening applications in medicinal chemistry & drug design.

Annexe F

Article 4 : Journal of Biological Chemistry

paru dans « Journal of Biological Chemistry » en novembre 2007

Xavier Hanoulle, Aurélie Melchior, Nathalie Sibille, Benjamin Parent, Agnès Denys, Jean-Michel Wieruszeski, Dragos Horvath, Fabrice Allain, Guy Lippens et Isabelle Landrieu.

Structural and functional characterisation of the interaction between cyclophilin B and a heparin derived oligosaccharide.

à paraître...

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 47, pp. 34148–34158, November 23, 2007

Structural and Functional Characterization of the Interaction between Cyclophilin B and a Heparin-derived Oligosaccharide*[S]

Received for publication, August 1, 2007, and in revised form, September 11, 2007 Published, JBC Papers in Press, September 12, 2007, DOI 10.1074/jbc.M706353200

Xavier Hanoulle, Aurélie Melchior, Nathalie Sibille, Benjamin Parent, Agnès Denys, Jean-Michel Wieruszeski, Dragos Horvath, Fabrice Allain, Guy Lippens[1], and Isabelle Landrieu[2]

From the Structural and Functional Glycobiology Unit, UMR8576 CNRS, University of Sciences and Technologies of Lille, 59655 Villeneuve d'Ascq, France

The chemotaxis and integrin-mediated adhesion of T lymphocytes triggered by secreted cyclophilin B (CypB) depend on interactions with both cell surface heparan sulfate proteoglycans (HSPG) and the extracellular domain of the CD147 membrane receptor. Here, we use NMR spectroscopy to characterize the interaction of CypB with heparin-derived oligosaccharides. Chemical shift perturbation experiments allowed the precise definition of the heparan sulfate (HS) binding site of CypB. The N-terminal extremity of CypB, which contains a consensus sequence for heparin-binding proteins was modeled on the basis of our experimental NMR data. Because the HS binding site extends toward the CypB catalytic pocket, we measured its peptidyl-prolyl *cis-trans* isomerase (PPIase) activity in the absence or presence of a HS oligosaccharide toward a CD147-derived peptide. We report the first direct evidence that CypB is enzymatically active on CD147, as it is able to accelerate the *cis/trans* isomerization of the Asp^{179}-Pro^{180} bond in a CD147-derived peptide. However, HS binding has no significant influence on this PPIase activity. We thus conclude that the glycanic moiety of HSPG serves as anchor for CypB at the cell surface, and that the signal could be transduced by CypB via its PPIase activity toward CD147.

First characterized as the molecular targets of the immunosuppressive drug cyclosporin A (CsA),[3] cyclophilins (Cyps) constitute one class of the prolyl *cis*/*trans* isomerases that catalyze the *cis*/*trans* interconversion of the peptide bond preceding a proline (1, 2). Members of this class such as the predominantly cytoplasmic CypA, the secreted CypB, and the mitochondrial CypD are small ubiquitous proteins sharing a high sequence homology (65% identity between human CypA and CypB), that translates into a closely related three-dimensional fold. Indeed, the NMR and crystal structures of CypA free and in complex with CsA (3–6), as well as the crystal structure of CypB in complex with a cyclosporine analogue (7) all show the same core structure composed of eight antiparallel β-strands forming a β-barrel surrounded by α-helices and loops. Whereas the nearly identical active site and CsA binding pocket further underscore their close relationship, both proteins do differ in their N and C termini, CypB containing two peptides of some 10 residues long that are lacking in CypA.

CypA and CypB act in the progression of inflammatory diseases such as rheumatoid arthritis and psoriasis, but are equally involved in the first steps of certain viral infections (8–10). Their inflammatory activity is conditioned by their interaction with heparan sulfate proteoglycans (HSPGs) and the membrane receptor CD147, two binding partners at the cell surface of T cell lymphocytes, granulocytes and macrophages (11–14). Significantly, both molecular partners have equally been described as co-receptors for the HIV-1 virus (10, 12, 15).

Both intact prolyl *cis*/*trans* activity of the cyclophilins and the presence of the Pro^{180} residue of CD147, located on one of the two extracellular immunoglobulin-like domains, are required for its chemotactic activity, raising the possibility that isomerization of the accessible Asp^{179}–Pro^{180} peptide bond might be the molecular signal that translates ultimately in chemotactic activity (14, 16). Mutations in the catalytic site, with residues such as Trp^{129}, Phe^{67}, and Arg^{62} (17) negatively interfere with the signal transduction. Such a cyclophilin-dependent mechanism of regulation has already been demonstrated for the tyrosine kinase Itk (18), where CypA catalyzes the isomerization of the Asn^{286}–Pro^{287} peptide bond. According to the isomerization state in *trans* or *cis* of this peptide bond, the Itk SH2 domain interacts with either its natural phosphotyrosine substrate or with its own SH3 domain (19).

Both CypA and CypB *in vitro* induce extracellular signal-regulated kinase (Erk) 1/2 phosphorylation, calcium flux generation and chemotaxis of responsive cells, although CypB is a more potent agonist and uniquely triggers integrin-mediated adhesion of T lymphocytes to fibronectin (11, 13, 14). Tight

* This work was supported by the Région Nord-Pas de Calais (France), the CNRS, the Universities of Lille 1 and Lille 2, and the Institut Pasteur de Lille. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental data and Figs. S1–S3.

[1] To whom correspondence may be addressed. E-mail: guy.lippens@univ-lille1.fr.

[2] To whom correspondence may be addressed: Unité de glycobiologie Structurale et Fonctionnelle, UMR 8576 CNRS, IFR 147, Université des Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France. Tel.: 33-0-3-20-33-72-41; Fax: 33-0-3-20-43-65-55; E-mail: isabelle.landrieu@univ-lille1.fr.

[3] The abbreviations used are: CsA, cyclosporin A; Cyp, cyclophilin; HSPG, heparan sulfate proteoglycans; HS, heparan sulfate; NOESY, nuclear Overhauser effect spectroscopy; EXSY, exchange spectroscopy; dp, degree of polymerization; PPIase, peptidyl-prolyl *cis-trans* isomerase; RDCs, residual dipolar couplings; r.m.s. deviation, root mean square deviation; HIV, human immunodeficiency virus; HSQC, heteronuclear single quantrum coherence.

binding of CypB to HS moieties of proteoglycans is one source for this increased potency, as mutations in the N-terminal ^{3}KKK^{6} and ^{15}YFD^{17} tripeptide motifs not only affect HS binding (16) but equally reduce CypB chemotaxis and abolish integrin-mediated adhesion (11, 17). Mutants deprived of enzymatic activity still bind to the cell surface of T lymphocytes, but are unable to induce biological responses, indicating that CypB has to interact simultaneously with both CD147 and HSPGs. Very recently, the interaction of CypB with the HS moieties of syndecan-1 was shown to promote and/or stabilize the complex between syndecan-1 and CD147, resulting in mitogen-activated protein kinase activation and subsequent pro-adhesive activity (20).

The minimal motif of HS interacting with CypB was mapped to an octasaccharide (21). However, the length is not the sole parameter defining the complexity of the sugar chains of HSPGs, as the exact sulfation pattern and the conformation of the glycanic moieties equally may contribute to the specificity of the interaction (22–26). Altogether, these data suggest that the high-affinity binding of CypB to specialized HS moieties stabilizes the interaction with its substrate or directly modulates its PPIase activity, resulting in an enhanced intracellular signaling via CD147.

We examine here by NMR spectroscopy the interaction between heparin-derived oligosaccharides and CypB. Whereas we confirm the direct implication of the N-terminal extension that distinguishes CypB from CypA in the HS binding, NMR chemical shift mapping and NOE data indicate a binding site of heparin directed toward the catalytic site rather than to the N-terminal β-strand containing the ^{15}YFD^{17} motif. This novel identification of a HS binding patch close to the active site raises the possibility of a functional coupling between HS binding and prolyl *cis*/*trans* isomerase activity. We use EXSY spectroscopy in the absence or presence of an oligosaccharide to quantify the CypB isomerization efficiency toward the Asp^{179}–Pro^{180} bond in a CD147-derived peptide. Finally, the N-terminal peptide responsible for the CypB-specific induction of T-lymphocyte adhesion to the extracellular matrix being absent from the x-ray structure due to proteolytic cleavages during the CypB purification procedure (7), we derive its structure based on NMR parameters, and investigate whether the heparin binding consensus sequence (^{3}EKKKGPKV^{10} in CypB) adopts any regular heparin binding structure (23).

EXPERIMENTAL PROCEDURES

Expression and Purification of Cyclophilin B—A recombinant plasmid, *pET15b-CypB*, was constructed to increase the production of recombinant human CypB. The sequence coding CypB was amplified from the previously described plasmid PCGF (27), using the following forward primer 3′-acttccatggccgatgagaagaag-5′ and the reverse primer 5′-acaaggatcctactcctggcgat-3′ and then inserted in a *pET15b* plasmid (Merck-NOVAGEN, Darmstadt) between restriction enzyme sites NcoI and BamHI. The 24 first amino acids corresponding to the signal sequence were not included in recombinant CypB. Recombinant CypB starts with Ala^{1} and ends with Glu^{184}. Our numbering is as in the x-ray Protein Data Bank file 1CYN (7). The *pET15b-CypB* plasmid was introduced in *Escherichia coli* BL21(DE3) *pLysS* cells (NOVAGEN), and a ^{15}N-^{13}C-labeled sample was prepared by growing cells in M9 minimal medium with $^{15}NH_4Cl$ and ^{13}C glucose as sole nitrogen and carbon sources, respectively. The ^{15}N-^{2}H-labeled sample was prepared by growing cells in a semi-rich deuterated medium (M9 medium in 99.5% D_2O with $^{15}NH_4Cl$, $^{2}H_7$-glucose (2 g/liter) and 20% of deuterated ^{15}N-rich medium (*v*/*v*) (Isogro, Cambridge Isotopes Laboratories). The cells were grown at 37 °C to reach an A_{600} = 0.8 and expression was induced at 20 °C with 0.4 mM isopropyl 1-thio-β-D-galactopyranoside. The cells were harvested after overnight induction and disrupted in lysis buffer (20 mM NaH_2PO_4/Na_2HPO_4, pH 6.8, 10 mM EDTA, Proteases inhibitor mixture (Roche), DNase I, RNase A) by sonication. Cell debris was removed by centrifugation at 20,000 × *g* for 30 min, then DNA was precipitated with streptomycin sulfate. After centrifugation at 15,000 × *g* for 30 min the supernatant was dialyzed (6–8 kDa cut-off) overnight against 20 mM NaH_2PO_4/Na_2HPO_4, pH 6.85. The recombinant CypB was sequentially purified by ion exchange (SP Sepharose Fast Flow) and gel filtration (Superose 12 Prep Grade) chromatography. Finally the protein was dialyzed against 50 mM NaH_2PO_4/Na_2HPO_4, pH 6.3, 40 mM NaCl, 1 mM EDTA, 1 mM dithiothreitol and concentrated by ultrafiltration (cut-off 10 kDa). Recombinant CypB was filtered (0.2μ) and stored at −20 °C.

Preparation of Heparin-derived Oligosaccharides—The heparin-derived oligosaccharides were prepared as previously described (21). Briefly, heparin was enzymatically digested with heparinase I at 30 °C. The resulting digestion mixture was desalted on a Sephadex G-10 column (GE Healthcare), then fractionated by gel filtration chromatography on Bio-Gel P-6 (Bio-Rad) in 0.2 M NH_4Cl, pH 3.5. The fractions corresponding to increasing dp oligosaccharides were desalted and then freeze dried. The heparin-derived oligosaccharide fractions were kept at −20 °C until used.

Peptide from CD147—A 15-amino acid long peptide of CD147 centered around Pro^{180} (sequence ^{173}NLNMEAD-**P**GQYRCNG^{187}) was synthesized by classical solid phase chemistry (Neosystems, Strasbourg, France), and purified to homogeneity by high performance liquid chromatography. Upon dissolving this peptide in a phosphate buffer to a 1 mM concentration, some precipitate was observed. Comparison with NMR spectra of soluble peptides allowed an estimation of the concentration of the soluble fraction at 0.5 mM.

NMR Spectroscopy—All spectra were recorded on either a Bruker Avance 800 MHz spectrometer with standard triple resonance probe or a Bruker Avance 600 MHz equipped with a cryogenic triple resonance probe head, at 25 °C (Bruker, Karlsruhe, Germany). The proton chemical shifts were referenced using the methyl signal of TMSP (sodium 3-trimethyl-silyl-[2,2,3,3-d.4]propionate) at 0 ppm. The spectra were processed with the Bruker TOPSPIN software package and in-house routines with the SNARF program (van Hoesel FHJ, 2000 SNARF version 0.8.9, University of Groningen, The Netherlands). Resonance assignment of the CypB protein residues was performed by using the classical strategy of paired triple resonance experiments (28) on a ^{15}N/^{13}C CypB sample at 0.25 mM in a 50 mM NaH_2PO_4/Na_2HPO_4, pH 6.3, 40 mM NaCl, 1 mM EDTA, 1 mM

dithiothreitol buffer using standard Bruker pulse programs. HNCACB/CBCAcoNH spectra were recorded with 512/52/71 complex points for $^{1}H/^{15}N/^{13}C$ windows of 13.9/36/70 ppm centered at 4.8/118/37.4 ppm, respectively. HNCO and HN(CA)CO spectra were recorded with 512/52/24 complex points for $^{1}H/^{15}N/^{13}C$ windows of 13.9/36/20 ppm centered at 4.8/118/172.5 ppm, respectively. A three-dimensional NOESY-(^{1}H-^{15}N HSQC) spectrum with a mixing time of 400 ms on a 350 μM sample of ^{2}H-^{15}N CypB in the presence or absence of dp12 was acquired with 512/32/148 complex points in the $^{1}H/^{15}N/^{1}H$ dimensions. All spectra were zero filled to 1k/256/256 complex points and multiplied by a shifted square sine bell function prior to Fourier transformation.

The heteronuclear NOE effect was measured with standard refocused HSQC pulse sequence in the presence or absence of proton decoupling during the 5-s relaxation delay, on a 250 μM sample of ^{15}N-CypB in the absence or presence of dp12. Hetero-NOE values were derived from the intensity ratios of the cross-peak with and without proton decoupling.

Residual dipolar couplings (RDCs) were collected on CypB and CypB-dp12 complex at 0.2 mM in 95% H_2O, 5% D_2O, 50 mM NaH_2PO_4/Na_2HPO_4, pH 6.3, 2.5 mM EDTA, 5 mM dithiothreitol, 85 μM TMSP-D4 (trimethyl-silyl propionate). RDCs were acquired on these uniformly ^{15}N-labeled samples suspended in a liquid crystalline medium consisting of 5% (*w*/*v*) polyoxyethylene 5-lauryl ether ($C_{12}E_5$) and 1-hexanol (Sigma) with a molar ratio of 0.85 (29). $1D_{NH}$ dipolar couplings were measured at 600 MHz and obtained using two-dimensional TROSY-type experiments (30, 31). Quadrature detection in the indirect dimensions of the multidimensional experiments was achieved by the echo/antiecho detection scheme for ^{15}N, and by the TPPI States method for ^{1}H. 64 scans were recorded per (t_1, t_2) increment. Data processing and peak picking were performed using the software SNARF (van Hoesel FHJ, 2000 SNARF version 0.8.9. University of Groningen, The Netherlands). Because the complex was partially precipitated, RDC values on the isolated CypB were of better quality, and were used for the refinement of the core region (see below).

The PPIase activity of CypB on the CD147 peptide was assessed on a sample of 0.5 mM CD147 peptide and 25 μM CypB, in the absence or presence of dp14. EXSY spectra were acquired at 800 MHz with mixing times of 50, 100, 200, 300, and 400 ms, and 2k/256 complex points in the direct and indirect proton dimension, and Fourier transformed to 4k/1k complex frequency points after zero filling. Spectra at 100, 200, and 400 ms were repeated on an independent sample to evaluate the error margins. Because the exchange cross-peaks are close to the diagonal, the maximal peak intensity rather than the peak integral was measured for the Asp^{179} *cis*/*trans* cross-peaks, and normalized to the corresponding diagonal peak intensity. The exchange rate k_{exch} (s^{-1}) was calculated by fitting the theoretical curve given by Equation 1 (32) to the experimental data, where %[*cis* → *trans*], expressed as the intensity of the exchange cross-peak to the diagonal peak, corresponds to the fraction of molecules that undergoes a transition from *cis* to *trans* conformation during the mixing time, and 1/*a* is the excess of *trans* over *cis* forms, determined on the basis of the one-dimensional spectra of Fig. 1*C*.

$$\%[cis \rightarrow trans] = a\frac{1 - \exp(-(1 + 1/a)k_{exch}) \times MT}{1 + \exp(-(1 + 1/a)k_{exch}) \times MT} \quad \text{(Eq. 1)}$$

Modeling of CypB Structure in Its Complex with dp12—The peptide ADEKKK was manually constructed and added at the N terminus of the x-ray structure of CypB (PDB code 1CYN). This completed structure formed the starting point for the refinement procedure. Briefly, the core region (residues 15–173) was first refined using the RDC values obtained on the isolated CypB as input for the XPLOR-NIH program (33, 34). Using the program MODULE (35) and the RDC values obtained on the CypB-dp12 complex, we calculated the alignment tensor for the complex. This tensor was then fixed in a second refinement step for the full structure. Input data were back-calculated NOEs, backbone dihedral angles, and hydrogen bonds for the core region, and the experimental NOEs, dihedral angle constraints from the ^{13}C chemical shifts and RDC values for the N and C termini. A total of 250 structures was calculated, of which we analyzed in detail the 20 structures of lowest energy. Further details of the refinement steps can be found in supplemental materials. The PyMol software was used for molecular graphics (DeLano, W. L., The PyMOL Molecular Graphics System (www.pymol.org)).

RESULTS

Molecular Characterization of the Partners—Based on its high isoelectrical point, the recombinant human cyclophilin B, 184 amino acids residues, was purified in one step by ion exchange chromatography to above 95% based on SDS-PAGE. The protein eluted from gel filtration chromatography as a single peak with an elution volume corresponding to a monomer of 20 kDa, and the good dispersion of the methyl groups in the one-dimensional NMR spectrum indicated a globular tertiary folding (Fig. 1*A*). A doubly labeled ^{15}N-^{13}C CypB was used for the NMR assignment strategy and all backbone resonances (except for Ala^{1} and Lys^{52}) and Cβ carbons were fully assigned.[4] To observe potential NOE contacts with heparin (see below), a deuterated ^{2}H-^{15}N CypB was prepared. From the one-dimensional spectrum (Fig. 1*A*), the deuteration level was estimated to be around 95%. Even after 1 week in aqueous buffer, several amide functions from the core of the protein still did not exchange with protons from the solvent, thereby defining the rigid central core of the protein.

Previous gel mobility shift assays studies had determined an octasaccharide as the minimal length required for efficient binding of heparan sulfate to CypB (21). Therefore, we only considered oligosaccharides with a higher degree of polymerization in this work, and present the results with dp12 or dp14 oligosaccharides. As these molecules come from enzymatic digestion of heparin with heparinase I, there are several sources of heterogeneity, at the level of the sequence and the sulfation pattern. To minimize these heterogeneities for the NMR experiments, we selected for those dp12 oligosaccharides species that interact most tightly with CypB by mixing an excess of dp12 with CypB followed by purification of the complex by gel filtration chromatography. An even

[4] I. Landrieu, F. Bonnachera, N. Sibille, X. Hanoulle, G. Vugniauk, A. Sillen, A. Melchior, B. Parent, J.-M. Wieruszeski, A. Denys, A. Hamel, F. Allain, D. Horvath, and G. Lippens, manuscript in preparation.

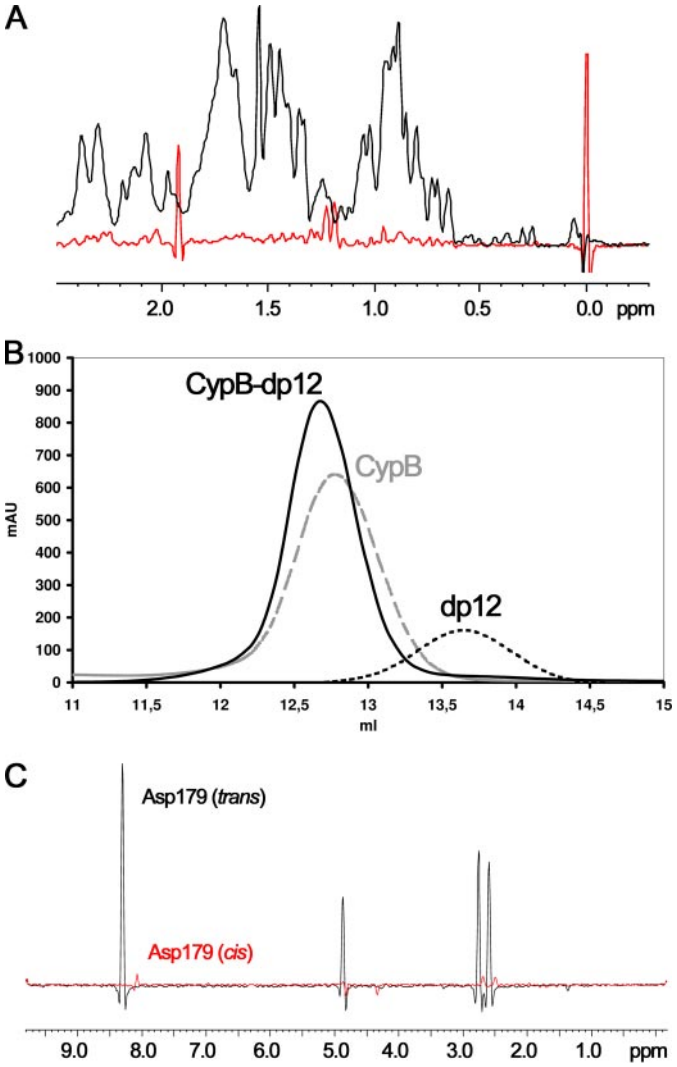


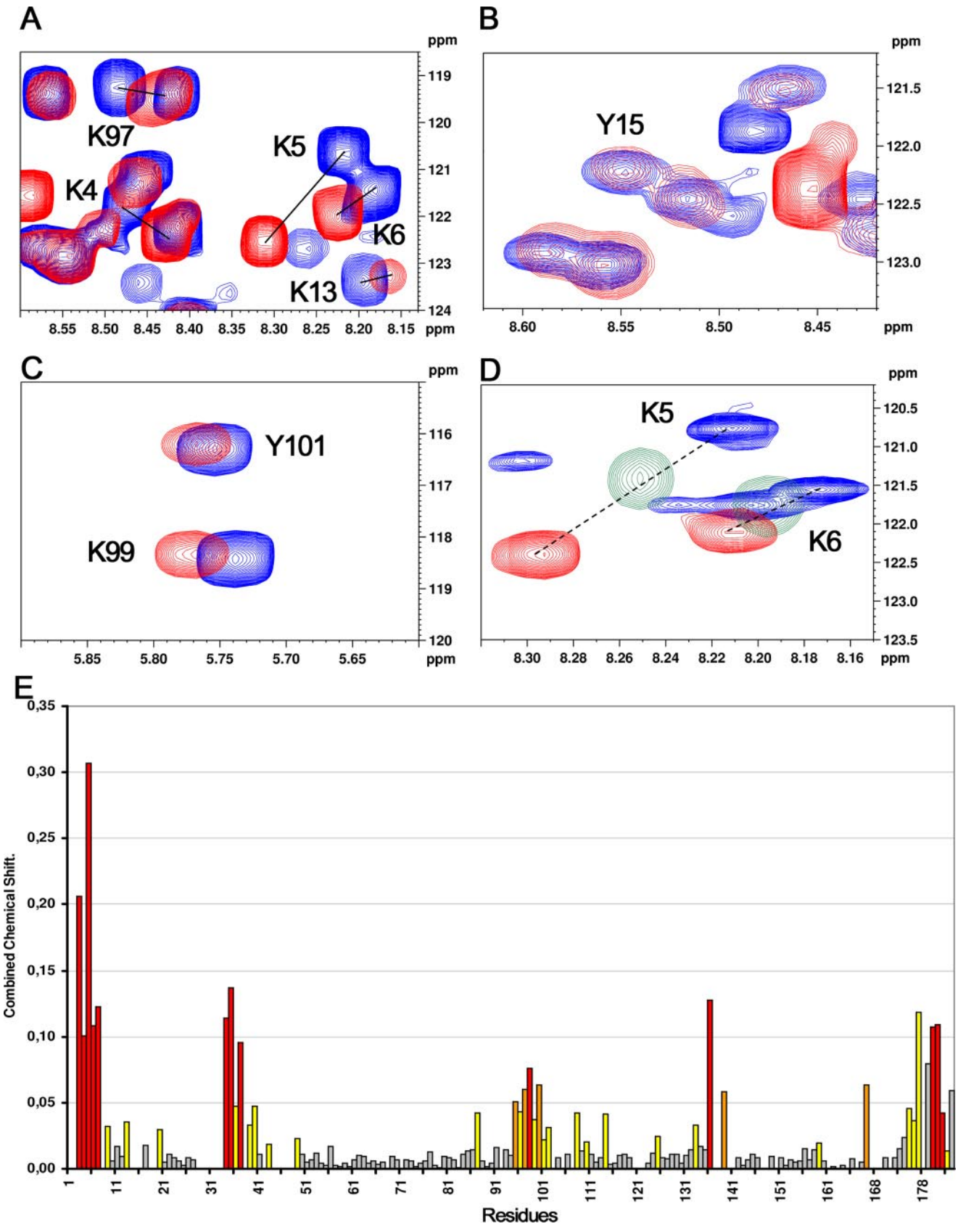
FIGURE 1. *A*, methyl region of the one-dimensional ^{1}H NMR spectra of ^{15}N CypB (in *black*) and ^{2}H-^{15}N CypB (in *red*). *B*, size exclusion chromatography elution profiles of the CypB-dp12 complex (*solid line*), free CypB (*dashed line*), and free dp12 heparin-derived oligosaccharides (*dotted line*). Protein profiles are at 280 nm, whereas the dp12 profile was recorded at 215 nm. *C*, traces from the ^{1}H TOCSY NMR spectrum at the amide proton frequency of the *cis* (in *red*) and *trans* (trans) forms of Asp^{179} in the CD147-derived peptide ^{173}NLNMEADPGQYRCNG^{187}.

more stringent selection criterion was introduced by incubation of the oligosaccharide/protein mixture in 400 mM NaCl before and during the gel filtration (Fig. 1*B*). As the interaction between CypB and heparan sulfates is expected to be essentially driven by electrostatic forces, the high ionic strength should preclude binding of oligosaccharides species that weakly interact with CypB. Following the gel filtration, the buffer was exchanged to reduce the salt concentration to 40 mM. Despite this stringent procedure, we cannot exclude that our heparin oligosaccharides still contain some degree of structural heterogeneity.

To assess a potential enzymatic activity of CypB toward CD147 and quantify the modulation of this activity by the heparin oligosaccharides, we used a synthetic peptide of 15 amino acids, ^{173}NLNMEAD**P**GQYRCNG^{187}, centered around the Pro^{180} CD147 residue. This CD147 peptide was characterized by homonuclear NMR spectroscopy, and both the absence of long range NOE contacts and $J_{HN\text{-}H\alpha}$ constants around 6 Hz indicate the absence of stable secondary structure. Both the *trans* and *cis* forms of the central Pro^{180} could be identified, and from the intensity of the *cis* and *trans* forms of the Asp^{179} amide proton, we estimate a *cis*/*trans* ratio of 6% (Fig. 1*C*). The dual proline conformation shows up not only as distinct resonance frequencies of the flanking amide protons, but also as far as the Tyr^{184} amide group (supplemental Fig. S1).

Definition of the CypB Zone in Interaction with Oligosaccharide dp12—A first complex between ^{15}N-labeled CypB and dp12 was obtained by mixing the two molecules in 400 mM NaCl to a molar ratio of 1:10, to ensure the ligand saturation of CypB (21). This complex was then purified by gel filtration chromatography using the same high ionic strength buffer (400 mM NaCl) to preferentially keep those dp12 species that strongly interact with CypB. To define the CypB residues involved in the interaction with heparin-derived dp12 oligosaccharides, we compared ^{1}H-^{15}N HSQC spectra of CypB alone and CypB in complex with dp12. Only a limited subset of CypB residues were affected by interaction with the dp12 oligosaccharide, excluding major conformational changes upon complex formation (Fig. 2). Previously, two CypB motifs, ^{4}KKK^{6} and ^{15}YFD^{17} (16), had been proposed to be directly involved in the interaction with heparan sulfates. Mapping the chemical shift changes along the sequence using our sequence-specific assignment confirmed the N-terminal ^{4}KKK^{6} motif as an effective part of the heparin binding site. The H_N resonances of these three lysine residues undergo the most important shift upon heparin binding (Fig. 2*A*). However, the binding of dp12 to CypB had no influence on the NMR signals corresponding to the residues of the ^{15}YFD^{17} motif (Fig. 2*B*), despite the fact that a ^{15}YFD^{17} deletion mutant was previously found unable to bind efficiently heparan sulfates (16, 21). We did several additional chemical shift mapping experiments with CypB and different heparin-derived oligosaccharides (dp8, dp12, and dp14), but were unable to detect any perturbation of these YFD motif resonances. These data suggest an indirect participation of the YFD motif in the binding of the heparan sulfates, probably through destabilization of the N-terminal β sheet. Beyond the N-terminal KKK motif, three additional regions of the protein had their amide chemical shift affected upon binding of dp12. These regions correspond to the C-terminal strand, the 34–43 region, and the 95–102 region (Figs. 2*E* and 3*D*). The backbone amide proton from lysine 97, lysine 99, and furthermore, the Hϵ-1 from the side chain from tryptophan 129 shifted in the presence of dp12, extending the interaction zone toward the active site of CypB (Fig. 3*D*). The latter one is known to play a dual role in the binding of CypB to cyclosporin A and CD147. In conclusion, whereas the previous mutational analysis had positioned the heparan sulfate binding site and the substrate binding site of CypB at opposite sites of the protein, we show here that these two sites are contiguous. Our identification of 12 lysines of a total of 25 (but no arginine) in the full interaction zone confirms that the complex formation is mainly driven by ionic interactions between lysines side chains and sulfate groups of HS.

The gel filtration experiment should ideally yield a 1:1 complex, with selection for those oligosaccharides that contain an optimal binding pattern. However, going through this procedure precludes a simple titration experiment to derive an affinity constant. Therefore, to estimate the order of magnitude of the affinity in solution



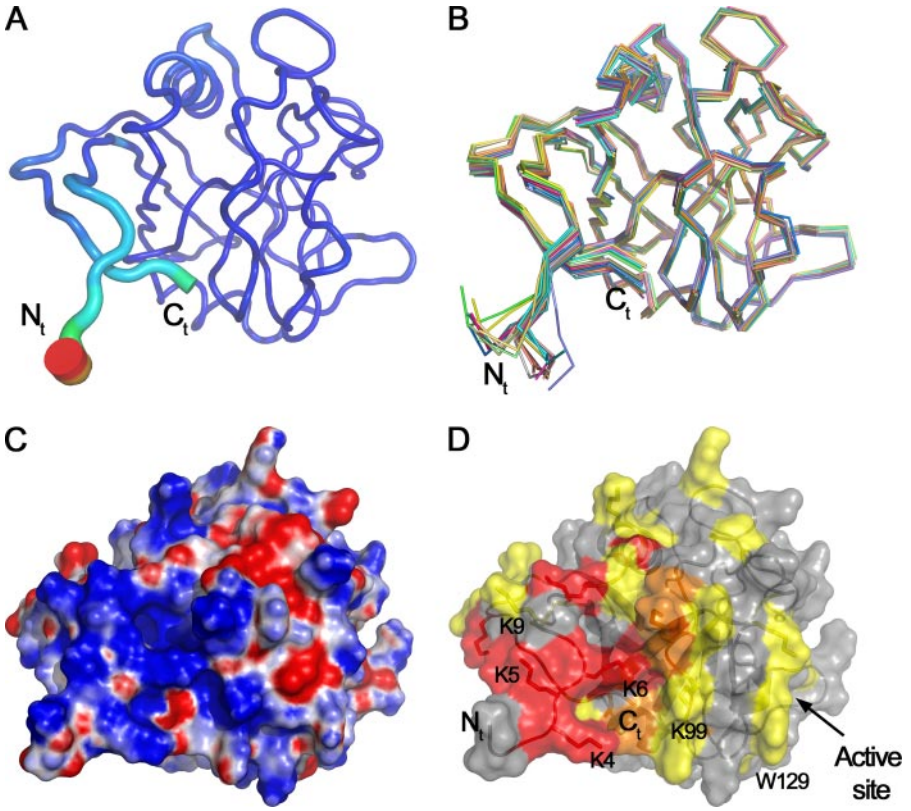


FIGURE 3. **Structure of CypB in the CypB-dp12 complex.** *A*, superimposition of the 20 lowest energy conformers of CypB in the CypB-dp12 complex. The structures, shown as Cα traces, were fitted on Cα and have a r.m.s. deviation of 0.87 Å. *B*, mean of the structures in *A*. The diameter of the sausage representation is representative of the r.m.s. deviation at each Cα position and the color goes from *marine blue* for the lowest Cα r.m.s. deviation (0.14Å) to *red* for the highest Cα r.m.s. deviation (4.77 Å). *C*, electrostatic potential surface of the CypB model in the same orientation as in *A*. The scale is from +10 kT/*e*, in *blue*, to −10 kT/*e*, in *red*. *D*, representation of the chemical shift NMR perturbations recorded on CypB upon dp12 binding on the molecular surface of the CypB model.

between CypB and a dp12 heparin-derived oligosaccharide, we did a reverse titration experiment. On the sample with the complex CypB-dp12, purified by gel filtration chromatography, we removed half of the sample and replaced it by an equivalent volume of CypB alone at the same concentration, and then recorded a new ^{1}H-^{15}N HSQC spectrum. Resonances that were previously affected upon addition of dp12 shifted back to an intermediate position between free CypB and CypB bound to dp12 (Fig. 2*D*). This suggests that on the NMR time scale, the heparin fragment exchanges rapidly between bound and free states, corresponding to an interaction of CypB and dp12 with a dissociation constant in the micromolar range or even weaker.

NMR Characterization of the CypB/dp12 Complex—Our results, together with previous studies, point out an important role of the ^{4}KKK^{6} motif from the CypB N terminus in the binding of heparan sulfates molecules. This N terminus is lacking in the x-ray structure of CypB (PDB code 1CYN) and structural data are only available starting from Gly^{7}, as the first 6 residues had undergone proteolytic cleavages during the purification process (7). In the absence of proteolysis of the N terminus during our purification, we recorded a three-dimensional NOESY-HSQC NMR experiment on the ^{2}H-^{15}N CypB alone or in complex with dp12 and compared the H_N-H_N NOE patterns of both spectra. The near identity of the NOE patterns involving residues from the core region in both spectra confirms that dp12 binding does not induce major conformational changes in the CypB structure. Moreover, most of the NOEs observed could be predicted from HN-HN distances derived from the x-ray structure, suggesting that this x-ray structure is a reasonably good starting point for the structure of the CypB bound to dp12. As for the C terminus, we detected NOEs between the Glu^{184} and Tyr^{101}, Gly^{102}, Trp^{105} side chain (Hε-1) and between Lys^{183} and the Trp^{105} side chain (Fig. 4*A*). The distance between the H_N of Glu^{184} and the Hε-1 proton of Trp^{105} in the crystal structure being 9.2 Å, the observation of a clear NOE contact between both protons suggests that the C terminus of CypB in solution is closer to the core of the protein than in the x-ray structure. However, these structural differences for the C terminus are not induced by dp12 binding, as we did observe the same NOEs with comparable intensity in the NOESY-HSQC spectrum of the free protein. Finally, several NOEs were observed between residues in the 7–10 region and the 179–183 region. These observations correlate with the x-ray structure where these regions of CypB form a small β-sheet.

The absence of NOEs between the 3 lysines in the N terminus and the rest of the protein suggests that this motif is highly flexible, which might be a determining character for it being the initial and preferential binding site for HS. We measured this dynamical aspect of the CypB backbone by heteronuclear NOE experiments in the presence or absence of dp12 (Fig. 5). A significant increase of heteronuclear NOEs was observed for the

FIGURE 2. **Chemical shift perturbations experiment.** *A–D*, superimposition of ^{1}H-^{15}N HSQC of free CypB (in *blue*) and CypB bound to dp12 heparin-derived oligosaccharides (in *red*). *A*, region of the spectra centered on the ^{4}KKK^{6} CypB N-terminal motif. *B*, region of the spectra centered on the residue Tyr^{15} from the ^{15}YFD^{17} CypB tripeptide. *C*, region of the spectra centered on residues Lys^{99} and Tyr^{101}, which are close to the CypB active site. *D*, reverse titration experiment where half of the CypB-dp12 sample was removed and replaced by an equivalent volume of CypB alone at the same concentration. A new ^{1}H-^{15}N HSQC spectrum was recorded (in *green*) and compared with those of free CypB (in *blue*) and CypB-dp12 complex (in *red*). Resonances from the ^{4}KKK^{6} motif that were previously affected upon addition of dp12 (*A*) shifted back to an intermediate position between CypB free and CypB bound to dp12. *E*, plot of the combined ^{1}H and ^{15}N chemical shift perturbations along the CypB sequence. The values were calculated with the following equation: Combined Chemical Shift ($\delta\Delta$) = $(\delta\Delta_{HN}^{2} + (\delta\Delta_{15N}/6.51)^{2})^{0.5}$. The $\delta\Delta$ values in the interval 0.018–0.05 ppm are colored *yellow*, those with $\delta\Delta$ values in the interval 0.05–0.07 ppm colored *orange*, and those with $\delta\Delta$ values >0.07 colored *red*.

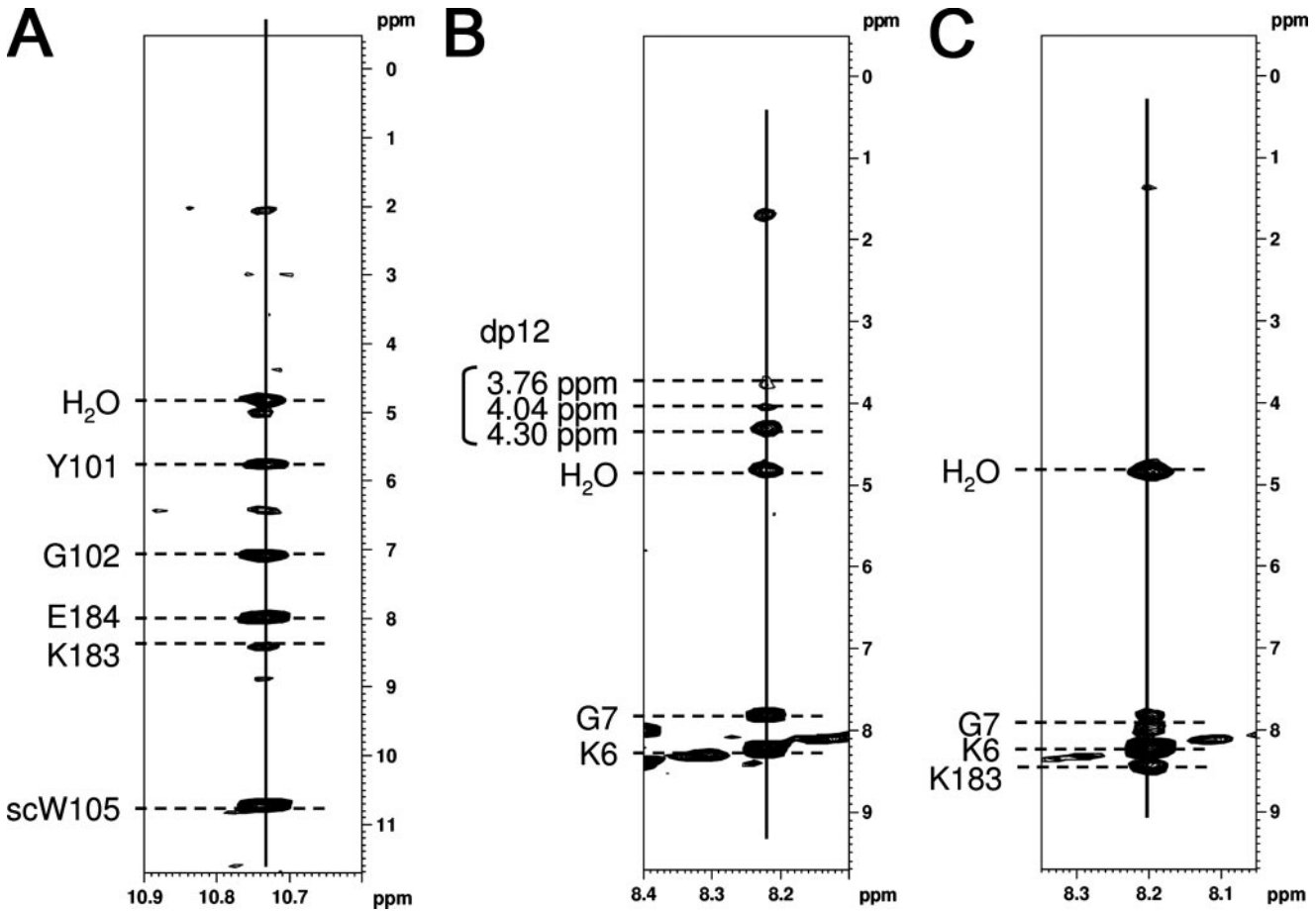


FIGURE 4. **Intra and intermolecular NOEs in the CypB-dp12 complex.** *A*, ^{1}H-^{1}H plane from the three-dimensional NOESY-^{1}H-^{15}N HSQC at the ^{15}N frequency of Hϵ-1 Trp^{105}. *B* and *C*, planes through the K6 amide resonance of ^{2}H-^{15}N CypB complexed to dp12 (*B*) or free in solution (*C*). The additional resonances come from the dp12 sugar moieties.

first 10 CypB residues upon binding of dp12, whereas a slight decrease of the heteronuclear NOE values was observed for residues 180–183. These observations suggest a direct interaction of the N-terminal lysine residues in the consensus sequence with the heparin-derived oligosaccharide, conferring a more rigid character upon binding. It further suggests that the chemical shift perturbations observed in the C terminus result from indirect effects rather than from a direct interaction with the oligosaccharide.

As the interaction between heparan sulfates and its binding partners involves negatively charged sulfate groups of HS and positively charged lysine side chains, backbone amide protons are seldom closer than 5 Å from the sugar protons, and ^{1}H-^{1}H intermolecular NOE correlations are not easily obtained (36–38). However, the use of a highly deuterated CypB limits spin diffusion (39), and moreover avoids confusion between heparin protons and aliphatic side chain protein resonances. We indeed detected some intermolecular NOEs between dp12 and Lys^{5}/Lys^{6} (Fig. 4*B*). These NOEs, involving protons in the range of 3.8–4.3 ppm, are absent in the control experiment on the same protein preparation without dp12 (Fig. 4*C*), and probably correspond to protons from the carbohydrate rings of dp12. Due to the severe overlap of heparin protons and the additional molecular heterogeneity of heparin-derived oligosaccharide dp12, these signals could, however, not be assigned without ambiguity, but they do confirm the direct physical interaction between the ^{4}KKK^{6} CypB motif and the dp12 molecule.

Because of the limited information that could be extracted from the NOEs involving residues in the N-terminal region, we assigned the ^{13}C chemical shifts in the absence and presence of dp12, and obtained RDC values on a partially oriented sample of the isolated and dp12 complexed protein. These data yield constraints on the dihedral angles for the former, and long range orientational constraints for the latter. All experimental constraints were used in a refinement protocol aimed at completing the structure of CypB in its complex with dp12.

Modeling of CypB in the CypB/dp12 Complex—A multistep protocol starting from the x-ray structure completed with coordinates for the lacking N terminus (^{1}ADEKKK^{6}) was used to obtain a family of structures compatible with all experimental data. The 20 structures of lowest energy (Fig. 3*A*) well conserve the typical cyclophilin fold, and when superimposed on all Cα atoms, give an overall r.m.s. deviation of 0.87 ± 0.27 Å. When we superimpose the core regions of these structures, from residue 15 to 173, and calculate the r.m.s. deviation values for the isolated N- and C-terminal extensions, we find values of 2.39 ± 1.05 and 1.14 ± 0.48 Å, respectively, indicating still a reasonable definition of these fragments.

The different CypB regions involved in dp12 binding as defined by the chemical shift perturbation mapping are close in

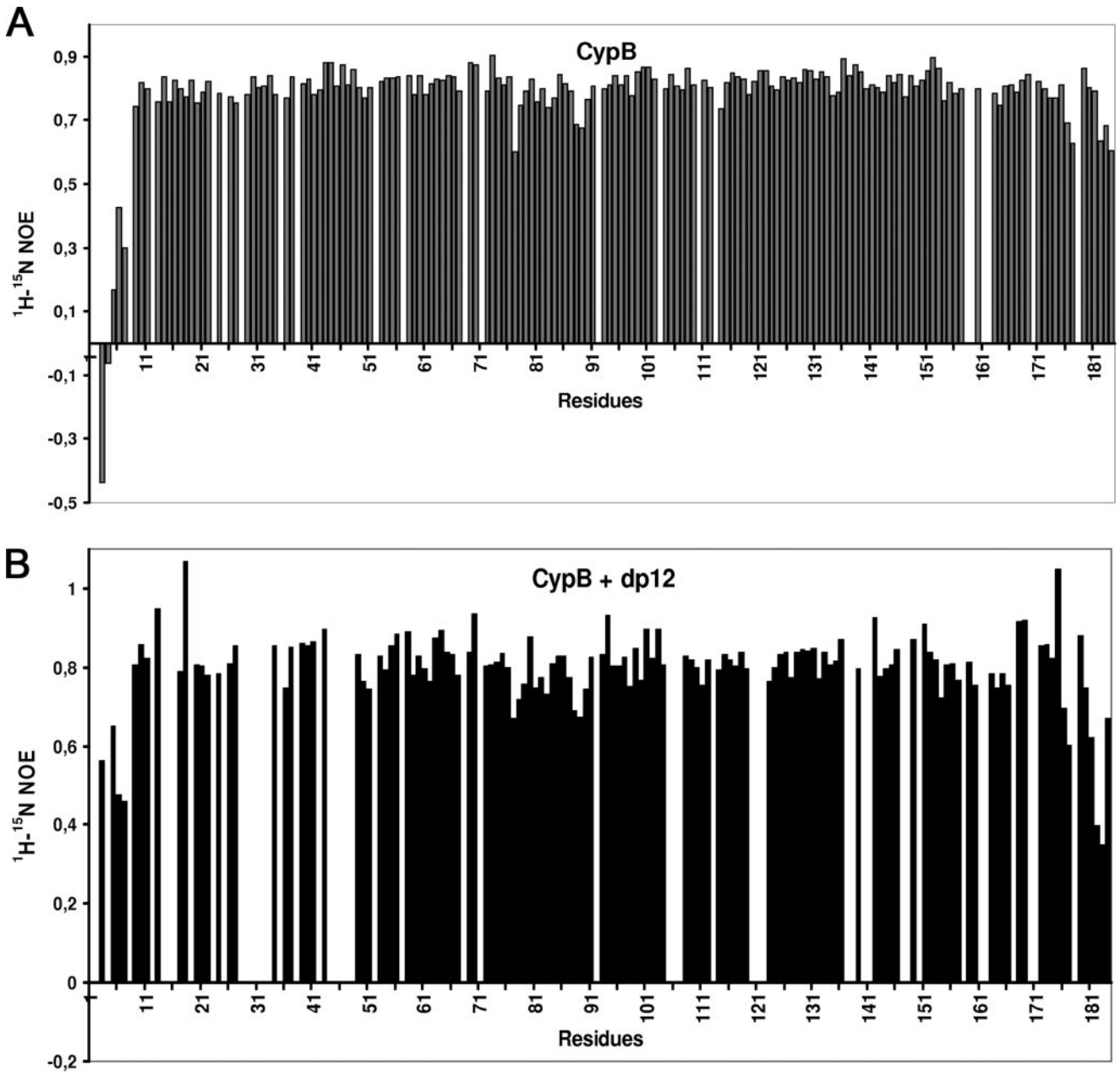


FIGURE 5. **Heteronuclear NOE data.** Heteronuclear NOE values of: (*A*) free CypB and (*B*) CypB in complex with dp12. The N terminus residues lose to a large extent their initial flexibility, whereas the extreme C terminus gains some flexibility.

space and form a well defined heparan sulfate binding site (Fig. 3*D*), which equally corresponds to the most electropositive area of CypB, containing 12 lysine residues (Fig. 3*C*). Closer examination of the N terminus in the CypB model showed that the structure of the consensus sequence *X*BBB*X*B*X* (^{3}EKKKGPK^9) may lead to the suitable orientation of the lysine side chains for promoting interaction with HS, although experimental data to define these side chains lack. However, our resulting structures indicate the absence of the canonical α-helix or β-strand structures that would project the basic side chains into the same direction (23). The strong NH_i-NH_{i+1} contacts that would characterize such a helical conformation were indeed not observed in the three-dimensional spectrum of CypB/dp12. Moreover the absence of regular secondary structural elements in the N terminus was confirmed by the ^{13}C chemical shift index method (40). Finally, our model shows that the N terminus is more surface accessible than the partially buried C terminus, in agreement with the experimental relaxation data.

Enzymatic Activity of CypB on a CD147-derived Peptide—Chemical shift mapping suggested that the heparan sulfate binding site extends to the edge of the active site of CypB. Our NOE data further support this result, as dp12 binding on CypB affects the Trp^{129} residue, which is part of the active site of CypB and plays a crucial role in the binding of cyclosporin A or the cell surface receptor CD147 (16). Indeed, the NOE patterns of the Hϵ-1 Trp^{129} side chain in the presence or absence of dp12 are not identical. In the absence of dp12, no NOE correlations were detected, whereas in the presence of dp12 the Hϵ-1 of Trp^{129} side chain correlates with the amide proton of the same residue and of two neighboring residues, Leu^{130} and Asp^{131}

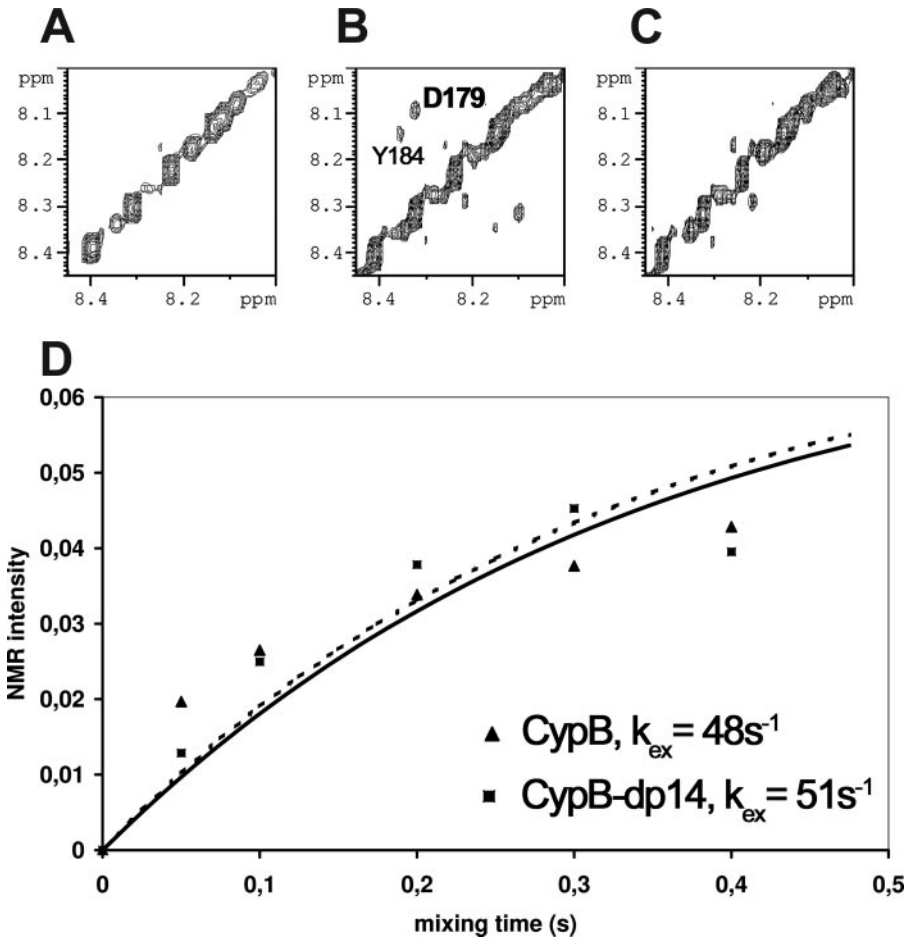


FIGURE 6. **EXSY data with the CD147 peptide.** The H_N-H_N region of a 400-ms EXSY spectrum is shown for the isolated CD147 peptide (*A*), the peptide in presence of a catalytic amount of CypB (*B*), and the peptide with CypB/CsA (*C*). The exchange rate in the free peptide is too slow to give observable exchange cross-peaks. When catalyzed by CypB, we do observe such peaks, but these disappear upon addition of CsA. *D*, normalized intensity of the exchange cross-peaks as a function of mixing time, for CypB (*triangles*, *solid line*) or CypB/dp14 (*squares*, *dotted line*).

(supplemental Fig. S2). Upon binding of dp12, the Trp^{129} side chain could at least transiently be reoriented, opening up the possibility of a functional coupling between the binding of heparin-derived oligosaccharides and the enzymatic PPIase activity of CypB. To further assess this possibility, the enzymatic activity of CypB toward a CD147-derived peptide, centered on the CD147 Pro^{180} residue (in bold), ^{173}NLNMEAD-**P**GQYRCNG^{187}, was characterized using EXSY NMR spectroscopy. The distinct *trans* and *cis* signals of the Asp^{179} amide protons were used to quantify the exchange process. Without CypB, both conformers are in very slow exchange, and no cross-peak between isoforms could be detected for mixing times up to 400 ms (Fig. 6*A* and supplemental Fig. S3*A*). However, when adding CypB in catalytic amounts to the same peptide sample, additional cross-peaks connecting the *cis* and *trans* amide protons of Asp^{179} are observed (Fig. 6*B* and supplemental Fig. S3*B*), confirming experimentally that CypB is able to catalyze the isomerization of Pro^{180} in CD147. Similar exchange peaks equally connected the *cis* and *trans* forms of Gly^{181} and Tyr^{184} by varying the mixing time of the EXSY spectra, an exchange rate $k_{ex} = 51\ s^{-1}$ was found (Fig. 6*D*). Addition of CsA to the sample did abolish the exchange cross-peaks, confirming the overlap between the prolyl *cis*/*trans* isomerase active site and the CsA binding site (Fig. 6*C* and supplemental Fig. S3*C*). As the dp12 binding site extends to the active site of CypB, and as on the cell surface, the heparan sulfate chains of proteoglycans are longer than a dp12 oligosaccharide, we used the longest heparin oligosaccharide, a dp14 molecule, to investigate any functional implications. The same EXSY spectra were thus run with CypB in the presence of dp14. Prolyl *cis*/*trans* isomerization was still present, and quantification of the rate led to a similar exchange rate of 48 s^{-1} (Fig. 6*D*). We therefore conclude that heparin binding extends up to the active site of CypB, but does not influence directly its enzymatic prolyl *cis*/*trans* isomerase activity.

DISCUSSION

Cyclophilins are proteins involved in several inflammatory diseases such as rheumatoid arthritis, and play a role in the HIV-1 viral infection process. It has been shown that both extracellular cyclophilin A and B are able to induce chemotaxis (13, 14, 41, 42) but that only CypB triggers T lymphocyte adhesion to fibronectin in the extracellular matrix (11). Although homologous, with more than 50% sequence identity for both full-length proteins, the main difference is in the N- and C-terminal extensions that characterize CypB but are lacking in CypA (43). The biological cyclophilin-mediated response requires both the cell surface HSPG and the cell surface receptor CD147. CypB would bind to one or more cell surface HS moiety of syndecan I and subsequently promote the syndecan I-CD147 association, resulting in an activation of p44/42 mitogen-activated kinases and a subsequent integrin-mediated induction of extracellular matrix adhesion (20). Precise molecular details of this activation mechanisms are as yet not available, be it for the early interaction steps or for the ensuing signal transduction.

Here, we use heparin-derived oligosaccharides to reproduce the physiological interaction between CypB and the glycan moiety of cell surface HSPG. The sulfated regions of heparin are similar to HS of HSPG. The use of oligosaccharides has already been validated as a good model to replace longer HS chains (44). Indeed, many structural and biological studies using oligosaccharides were consistent with the *in vivo* biological data (36, 45). The minimal binding unit for CypB has previously been shown to be an octasaccharide (dp8). Here, we used dp12 and dp14 oligosaccharides to study the interaction with CypB.

Their enzymatic preparation from heparin followed by size exclusion chromatography leads to length-defined compounds with heterogeneous sulfation patterns. To minimize this heterogeneity, we purified the CypB-dp12 complex in high salt conditions, which should result in the selection for the stronger interacting species. The 0.4 M salt concentration used is lower than the 0.6 M concentration needed to elute CypB from a heparin-Sepharose column (46), and our gel filtration data show that we indeed form a complex (Fig. 1*B*).

The chemical shift perturbation strategy showed that only a defined subset of CypB residues are involved in the dp12 binding. Even though located in 4 different regions of the linear sequence of CypB, they are spatially close and form a well defined HS binding site. NOE data confirm that at least the N-terminal ^{4}KKK^{6} motif is involved in direct physical interaction with the sugar moiety, explaining why their triple mutation into ^{4}AAA^{6} renders the protein unable to bind HSPG on the cell surface or to trigger the CypB T-lymphocytes adhesion to extracellular matrix. Heteronuclear relaxation data indicate that this interaction results in a loss of mobility for the N-terminal peptide, be it without the establishment of a regular secondary structure element as expected for heparin binding peptides (23).

If only intermolecular NOEs were observed for Lys^{5} and Lys^{6}, the complete binding site for dp12 is larger than simply this ^{3}KKK^{6} motif. Upon dp12 binding, 44 amide resonances were perturbed. The corresponding residues constitute a well defined binding site on the CypB molecular surface, including next to the N- and C-terminal extensions two loops (37–40 and 95–102) from the CypB core. The length of the defined binding site fits rather well with the length of a heparin-derived dp12 oligosaccharide in a helical conformation (PDB 1HPN). The binding site is characterized by a groove flanked by lysine ladders on each side. These lysine side chains constitute a positively charged patch on CypB that probably interacts with the bulky negatively charged sulfate groups of dp12 (22). Chemical shifts in the ^{15}YFD^{17} peptide, previously identified by site-directed mutagenesis as important for the interaction of CypB with HS, did not change upon interaction with dp12. However, only the side chains of Tyr^{15} and Asp^{17} are solvent accessible, whereas the side chain of Phe^{16} is buried into the hydrophobic core of CypB. A plausible explanation for the fact that the CypB mutant deleted of ^{15}YFD^{17} does not directly bind to the heparan sulfate is that this deletion induces a destabilization of the first β-strand (Thr^{11}–Arg^{19}) and thus potentially disrupts the location of the N-terminal anchoring patch with respect to the rest of the binding site.

A final interaction zone is defined by the loop of residues 125–133 surrounding the active site. We show specifically that the side chain of Trp^{129} not only undergoes chemical shift perturbations through the addition of dp12, but that equally its orientation could be modified as witnessed by differential NOEs. This suggests that the HPSG might exert a dual role in the biological function of CypB. They first might serve to anchor CypB in the close vicinity of the cell surface receptor CD147, and might in a second stage modify its prolyl *cis*/*trans* isomerase activity toward this same receptor. Using a synthetic peptide centered on the critical Pro^{180} and NMR exchange spectroscopy, we here directly demonstrate that CypB is catalytically active on an extracellular region of the membrane receptor CD147, and that this interaction is blocked by cyclopsporin A. Despite the dp12 binding site extending to the close vicinity of the CypB active site, the enzymatic activity of CypB on the CD147 peptide is not affected upon dp12 binding. This observation agrees with the previous finding that cyclosporin A, which bind in the CypB active site, does not influence the binding of cell surface HS (47, 48).

An affinity in the micromolar range between CypB and the dp12 oligosaccharide was inferred from the gradually shifting correlation peaks in our NMR reverse titration experiment. Such values are plausible if we consider the fact that CypB elutes from a heparin affinity column at 0.6 M NaCl (46) and that the HIV-1 Tat transduction domain, which elutes from the same column at 1.6 M, has a K_d of 0.37 μM for heparin (49). The micromolar range observed is also comparable with the affinity observed for CD44/HA (50) and several fibroblast growth factor-heparin complexes (0.5–85 μM) (51). Strikingly, Allain *et al.* (48) measured a much lower K_d around 10 nm between CypB and the full-length HS on the surface of T-lymphocytes (16). The discrepancy with our micromolar value could arise from different points. First of all, we used heparin-derived oligosaccharides dp12, whereas the cell-based assay used full-length cell surface HSPG. The flexibility of HS glycanic chains are probably length dependent, and this may play a crucial role in the affinity (37). Second, at the cell surface the HS chains are linked to the core of HSPG, leading to a crowded environment. Finally, the methods to assess the affinity were not the same. Here, we use in solution NMR spectroscopy on a molecular complex, whereas the binding on T-lymphocytes has been evaluated on a surface by competition experiments with radioiodinated and cold CypB.

CypB only shows weak transient interactions with CD147 peptide, compatible with an enzyme/substrate relationship. Our data thus validate the model proposed by Allain *et al.* (11), where during the inflammation response CypB interacts with the HS chains of cell surface HSPG and is subsequently locally concentrated in the surrounding of the membrane receptor CD147 (20). Without a direct influence of the HS on its enzymatic activity, CypB can isomerize the Asp^{179}–Pro^{180} peptidyl proline bond of the CD147 extracellular domain, which then triggers in an unknown fashion intracellular signaling events. Finally, we further validate the interaction of CypB with cell surface heparan sulfate as a potential therapeutic target to modulate the cyclophilin-mediated inflammation process.

Acknowledgments—We thank Drs. A. Hamel and G. Vugniaux from DebioPharm (Lausanne, Switzerland) for a generous gift of the CD147 peptide and cyclosporin A.

REFERENCES

1. Handschumacher, R. E., Harding, M. W., Rice, J., Drugge, R. J., and Speicher, D. W. (1984) *Science* **226,** 544–547
2. Schreiber, S. L. (1991) *Science* **251,** 283–287
3. Ke, H. (1992) *J. Mol. Biol.* **228,** 539–550
4. Mikol, V., Kallen, J., Pflugl, G., and Walkinshaw, M. D. (1993) *J. Mol. Biol.* **234,** 1119–1130

5. Ottiger, M., Zerbe, O., Guntert, P., and Wuthrich, K. (1997) *J. Mol. Biol.* **272,** 64–81
6. Spitzfaden, C., Braun, W., Wider, G., Widmer, H., and Wuthrich, K. (1994) *J. Biomol. NMR* **4,** 463–482
7. Mikol, V., Kallen, J., and Walkinshaw, M. D. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5183–5186
8. Bukrinsky, M. I. (2002) *Trends Immunol.* **23,** 323–325
9. Yurchenko, V., Constant, S., and Bukrinsky, M. (2006) *Immunology* **117,** 301–309
10. Sokolskaja, E., and Luban, J. (2006) *Curr. Opin. Microbiol.* **9,** 404–408
11. Allain, F., Vanpouille, C., Carpentier, M., Slomianny, M. C., Durieux, S., and Spik, G. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 2714–2719
12. Pushkarsky, T., Zybarth, G., Dubrovsky, L., Yurchenko, V., Tang, H., Guo, H., Toole, B., Sherry, B., and Bukrinsky, M. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 6360–6365
13. Yurchenko, V., O'Connor, M., Dai, W. W., Guo, H., Toole, B., Sherry, B., and Bukrinsky, M. (2001) *Biochem. Biophys. Res. Commun.* **288,** 786–788
14. Yurchenko, V., Zybarth, G., O'Connor, M., Dai, W. W., Franchin, G., Hao, T., Guo, H., Hung, H. C., Toole, B., Gallay, P., Sherry, B., and Bukrinsky, M. (2002) *J. Biol. Chem.* **277,** 22959–22965
15. Saphire, A. C., Bobardt, M. D., Zhang, Z., David, G., and Gallay, P. A. (2001) *J. Virol.* **75,** 9187–9200
16. Carpentier, M., Allain, F., Haendler, B., Denys, A., Mariller, C., Benaissa, M., and Spik, G. (1999) *J. Biol. Chem.* **274,** 10990–10998
17. Carpentier, M., Allain, F., Slomianny, M. C., Durieux, S., Vanpouille, C., Haendler, B., and Spik, G. (2002) *Biochemistry* **41,** 5222–5229
18. Andreotti, A. H. (2003) *Biochemistry* **42,** 9515–9524
19. Brazin, K. N., Mallis, R. J., Fulton, D. B., and Andreotti, A. H. (2002) *Proc. Natl. Acad. Sci. U. S. A.* **99,** 1899–1904
20. Pakula, R., Melchior, A., Denys, A., Vanpouille, C., Mazurier, J., and Allain, F. (2007) *Glycobiology* **17,** 492–503
21. Vanpouille, C., Denys, A., Carpentier, M., Pakula, R., Mazurier, J., and Allain, F. (2004) *Biochem. J.* **382,** 733–740
22. Vanpouille, C., Deligny, A., Delehedde, M., Denys, A., Melchior, A., Lienard, X., Lyon, M., Mazurier, J., Fernig, D. G., and Allain, F. (2007) *J. Biol. Chem.*
23. Capila, I., and Linhardt, R. J. (2002) *Angew. Chemie* **41,** 391–412
24. Delehedde, M., Allain, F., Payne, S. J., Borgo, R., Vanpouille, C., Fernig, D. G., and Deudon, E. (2002) *Curr. Med. Chem.* **1,** 89–102
25. Gama, C. I., and Hsieh-Wilson, L. C. (2005) *Curr. Opin. Chem. Biol.* **9,** 609–619
26. Rapraeger, A. C., and Ott, V. L. (1998) *Curr. Opin. Cell Biol.* **10,** 620–628
27. Spik, G., Haendler, B., Delmas, O., Mariller, C., Chamoux, M., Maes, P., Tartar, A., Montreuil, J., Stedman, K., and Kocher, H. P. (1991) *J. Biol. Chem.* **266,** 10735–10738
28. Grzesiek, S., Bax, A., Hu, J. S., Kaufman, J., Palmer, I., Stahl, S. J., Tjandra, N., and Wingfield, P. T. (1997) *Protein Sci.* **6,** 1248–1263
29. Otting, G., Ruckert, M., Levitt, M. H., and Moshref, A. (2000) *J. Biomol. NMR* **16,** 343–346
30. Pervushin, K., Riek, R., Wider, G., and Wuthrich, K. (1997) *Proc. Natl. Acad. Sci. U. S. A.* **94,** 12366–12371
31. Rance, M., Loria, J. P., and Palmer, A. G., III. (1999) *J. Magn. Reson.* **136,** 92–101
32. Kaplan, J. L., and Fraenkel, G. (1980) *NMR Chemically Exchanging Systems*, Academic Press, New York
33. Schwieters, C. D., Kuszewski, J. J., and Clore, G. M. (2006) *Prog. NMR Spectrosc.* **48,** 47–62
34. Schwieters, C. D., Kuszewski, J. J., Tjandra, N., and Clore, G. M. (2003) *J. Magn. Reson.* **160,** 65–73
35. Dosset, P., Hus, J. C., Marion, D., and Blackledge, M. (2001) *J. Biomol. NMR* **20,** 223–231
36. Canales, A., Lozano, R., Lopez-Mendez, B., Angulo, J., Ojeda, R., Nieto, P. M., Martin-Lomas, M., Gimenez-Gallego, G., and Jimenez-Barbero, J. (2006) *FEBS J.* **273,** 4716–4727
37. Powell, A. K., Yates, E. A., Fernig, D. G., and Turnbull, J. E. (2004) *Glycobiology* **14,** 17R–30R
38. Vanhaverbeke, C., Simorre, J. P., Sadir, R., Gans, P., and Lortat-Jacob, H. (2004) *Biochem. J.* **384,** 93–99
39. Koharudin, L. M., Bonvin, A. M., Kaptein, R., and Boelens, R. (2003) *J. Magn. Reson.* **163,** 228–235
40. Wishart, D. S., and Sykes, B. D. (1994) *J. Biomol. NMR* **4,** 171–180
41. Arora, K., Gwinn, W. M., Bower, M. A., Watson, A., Okwumabua, I., MacDonald, H. R., Bukrinsky, M. I., and Constant, S. L. (2005) *J. Immunol.* **175,** 517–522
42. Xu, Q., Leiva, M. C., Fischkoff, S. A., Handschumacher, R. E., and Lyttle, C. R. (1992) *J. Biol. Chem.* **267,** 11968–11971
43. Galat, A. (1999) *Arch. Biochem. Biophys.* **371,** 149–162
44. Angulo, J., Hricovini, M., Gairi, M., Guerrini, M., de Paz, J. L., Ojeda, R., Martin-Lomas, M., and Nieto, P. M. (2005) *Glycobiology* **15,** 1008–1015
45. Angulo, J., Ojeda, R., de Paz, J. L., Lucas, R., Nieto, P. M., Lozano, R. M., Redondo-Horcajo, M., Gimenez-Gallego, G., and Martin-Lomas, M. (2004) *ChemBioChem* **5,** 55–61
46. Denys, A., Allain, F., Carpentier, M., and Spik, G. (1998) *Biochem. J.* **336,** 689–697
47. Allain, F., Denys, A., and Spik, G. (1994) *J. Biol. Chem.* **269,** 16537–16540
48. Allain, F., Denys, A., and Spik, G. (1996) *Biochem. J.* **317,** 565–570
49. Hakansson, S., and Caffrey, M. (2003) *Biochemistry* **42,** 8999–9006
50. Takeda, M., Terasawa, H., Sakakura, M., Yamaguchi, Y., Kajiwara, M., Kawashima, H., Miyasaka, M., and Shimada, I. (2003) *J. Biol. Chem.* **278,** 43550–43555
51. Conrad, H. (1998) *Heparin Binding Proteins*, Academic Press, San Diego, CA

Annexe G

Conférence 1 : Congress on Evolutionary Computation, Singapour, 2007

« Congress on Evolutionary Computation », septembre 2007, Singapour.

B. Parent, Alexandru Tantar, Nouredine Melab, El-Ghazali Talbi, Dragos Horvath

Grid-based Evolutionary Strategies Applied to the Conformational Sampling Problem.

Grid-based Evolutionary Strategies Applied to the Conformational Sampling Problem.

Benjamin Parent, Alexandru Tantar, Nouredine Melab, El-Ghazali Talbi, Dragos Horvath

***Abstract*— Computational simulations of conformational sampling in general, and of macromolecular folding in particular represent one of the most important and yet one of the most challenging applications of computer science in biology and medicinal chemistry. The advent of GRID computing may trigger some major progress in this field. This paper presents our first attempts to design GRID-based conformational sampling strategies, exploring the extremely rugged energy response surface in function of molecular geometry, in search of low energy zones through phase spaces of hundreds of degrees of freedom. We have generalized the classical island model deployment of Genetic Algorithms ($\mathcal{GA}$) to a "planetary" model where each node of the grid is assimilated to a "planet" harboring quasi-independent multi-island simulations based on a hybrid GA-driven sampling approach. Although different "planets" do not communicate to each other — thus minimizing inter-CPU exchanges on the GRID — each new simulation will benefit from the preliminary knowledge extracted from the centralized pool of already visited geometries, located on the dispatcher machine, and which is disseminated to any new "planet". This "panspermic" strategy allows new simulations to be conducted such as to either be attracted towards an apparently promising phase space zone (biasing strategies, intensification procedures) or to avoid already in-depth sampled (tabu) areas. Successful folding of mini-proteins typically used in benchmarks for all-atoms protein simulations has been observed, although the reproducibility of these highly stochastic simulations in huge problem spaces is still in need of improvement. Work on two structured peptides (the "tryptophane cage" 1L2Y and the "tryptophane zipper" 1LE1) used as benchmarks for all-atom protein folding simulations has shown that the planetary model is able to reproducibly sample conformers from the neighborhood of the native geometries. However, within these neighborhoods (within ensembles of conformers similar to models published on hand of experimental geometry determinations), the energy landscapes are still extremely rugged. Therefore, simulations in general produce "correct" geometries (similar enough to experimental model for any practical purposes) which sometimes unfortunately correspond to relatively high energy levels and therefore are less stable than the most stable among misfolded conformers. The method thus reproducibly visits the native phase space zone, but fails to reproducibly hit the bottom of its rugged energy well. Intensifications of local sampling may in principle solve this problematic behavior, but is limited by computational ressources. The quest for the optimal time point at which a phase space zone should stop being intensively searched and declared tabu, a very difficult problem, is still awaiting for a practically useful solution.**

I. Introduction

The prediction of three-dimensional shapes of molecules on hand of their connectivity (the so-called *Conformational Sampling* task or simply $\mathcal{CS}$) is a widely addressed, central problem in structural biology and drug design [1]. There are yet no general approaches able to enumerate, for an arbitrary (macro)molecule, the most stable molecular geometries adopted in solution. Several proofs of the NP-completeness of such a problem have been proposed on hand of different models [2], [3] that frustrate computationalists and illustrate the Levinthal paradox [4]. The reformulation in terms of an energy landscape [5] where the energy, expressed as a function of geometry, is to be minimized, enables to attack the problem in the framework of function optimization. The energy minima then correspond to the populated geometries of the molecule; however entropic effects embedded in the widths of the wells, and which play an important role in determining the *free energy* are very difficult to estimate.

The huge problem size (hundreds of degrees of freedom), is actually not the major challenge: the extreme ruggedness of the response hypersurface (molecular energy as a function of internal coordinates: dihedral angles around the considered rotatable bonds, in this case) causes any deterministic optimization attempt to get stuck in local, most likely irrelevant optima and imposes the use of stochastic sampling procedures. However, the probability of discovering the very narrow low energy zones of phase space by randomly drawing the correct coordinates is virtually null.

A. Conformational sampling task in all-atom description

The estimation (according to a classical force field) of the internal energy of a given structure, in function of the relative positions of the atoms, offers an objective score, allowing to reformulate the question in terms of optimization theory: Boltzmann's equation (1) provides the population level of each state.

$$\Pr(\text{system in state of energy } E) \propto \exp\left(-\frac{E}{k_B T}\right) \quad (1)$$

where T is the absolute temperature and k_B, the Boltzmann constant.

This equation stresses that, no matter how numerous, all the low-energy minima within a few $k_B T$ from the absolute bottom of the energy hypersurface will be populated and are, therefore, important. Every conformational sampling algorithm must therefore address the (highly) multimodal aspect of the optimization.

Since the herein described software is aimed at docking problems and affinity estimation of small ligands with protein binding sites, an all-atom level of description is required. The empirical force field used to estimate the molecular energy as a function of geometry has been derived from the Consistent Valence Force Field [6], [7] (CVFF), enhanced by the addition of a continuum solvent model [8]. Although intrinsically

inaccurate, the force field-based energy estimation allows a far simpler, Newtonian, description of the problem compared to the correct quantum mechanical formalism.

Whereas molecular dynamics and/or Monte Carlo simulations, proceeding by small perturbations of a local geometry, may successfully avoid visiting the ubiquitous high-energy regions of phase space (provided a low-energy starting geometry is available!), they tend to spend too much time in exploring the local neighborhoods rather than pushing forward to yet uncharted phase space regions. The $\mathcal{GA}$ ability to deal with a set of solutions while deriving profit of both an intrinsic stochastic behavior in addition to the recombination principle, made them, in our opinion, the most suited tool for challenging highly multimodal / highly dimensional problems [9]. Our previous experience [10] showed that hybrid genetic algorithms, relying on the synergy between random exploration, selection and local calls to specific optimization procedures (tailor-made to respond to the peculiarities of the molecular energy landscape), have the ability to successfully cope with the challenges of conformational sampling. Nevertheless, this software would require weeks to month on a typical two-processor workstation in order to complete the successful folding (discovery of the experimentally known energy minimum) of peptides typically used in all-atom folding simulations (tryptophane cage, pdb code 1L2Y [11], 20 aminoacids; tryptophane zipper, pdb code 1LE1 [12], 13 aminoacids; the PIN1 WW domain, 34 aminoacids [13], etc.). The high computational costs, on one hand, and the straightforwardness of parallel deployment strategies for genetic algorithms, on the other, make this problem an ideal candidate for GRID computing.

Here we report, after a short introduction of the hybrid island model, a first successful deployment strategy on the parallel GRID[1] context. This "planetary" model was so dubbed as it represents a generalization of the classical island strategy, where each node of the grid represents a "planet" on which an island model will be started. It enables the controlled sharing of computational effort between global Darwinian exploration (some "planets" will be charged with the search for novel, different, low energy folds) and intensification (others perform local searches for the absolute energy minimum within the neighborhoods of newly discovered, "raw" geometries, to fine tune structural details - with potentially dramatic decreases in molecular energies).

II. $\mathcal{GA}$ IMPLEMENTATION

A. Genetic Algorithms

The hybrid GA deployed on the "planets" of the GRID operates on the degrees of freedom associated to the rotations around interatomic single bonds (figure 1), so that a chromosome actually represents the list, or vector of torsional angles associated to each of the considered rotatable bonds: $\overrightarrow{\Theta} = (\Theta_i, i = 1...N_{\text{rotBonds}})$.

[1]supported by the French GRID5000 initiative (www.grid5000.fr) and the Agence Nationale de la Recherche

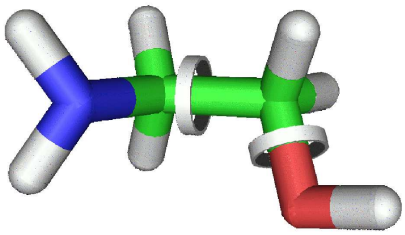


Fig. 1. Torsional angle coding.

Certain peculiarities of the sampling problem may ask for hybridizations of the genetic algorithm with other optimization procedures (conducting "Lamarckian" local optimizations to repair local clashes in what would otherwise represent stable conformers, allow for "directed" mutations, permitting the other degrees of freedom to adjust in response to the random shift applied to the mutated chromosome locus, introduce population diversity management and "tabu" criteria to block revisiting already sampled phase space zones, bias random distributions for each degree of freedom in order to enhance the probability of drawing values seen to occur in stable conformers, etc. — see below). Moreover, the control parameters inherent to the genetic algorithms (population size, mutation and crossover rates, maximal age, ending condition etc.) have a dramatic impact on the sampling performance. An additional layer of optimization, in search of the optimal operational regime of the $\mathcal{GA}$ for a current sampling problem, was therefore implemented as part of a global sampling strategy involving many successive and/or parallel $\mathcal{GA}$ runs.

B. Hybridizations

a) Parallelism: An island model [14] allows parallel implementations of the core $\mathcal{GA}$ to run independently, but with occasional inter-island migrations of solutions. This basic parallelization scheme favors exploration since each island may in principle harbor a distinct population which may nevertheless be challenged by fitter migrants if it fails to evolve as fast as competing islands. Care should be taken while designing the migration mechanism, to prevent genetic material from spreading to more than one island.

b) Non-uniform probability laws: while $\mathcal{GA}$s usually make use of flat distribution of probability to draw random values for each locus of the chromosome, introducing any knowledge and biasing the search towards peculiar regions of the phase space is possible by modifying these probability laws. The 'knowledge-based' biasing strategy relies on a local energy strain estimation, such that locally more stable staggered conformations will be favored over eclipsed ones. The other, 'tradition-based', strategy exploited here relies on statistics about the preferentially adopted torsional values in the fittest solutions currently available. This latest paradigm

suffers from its self-consistency and it has been shown that extreme caution should be taken to ensure that a sufficiently diverse and relevant pool of precursor solutions is at hand before actively favoring herein encountered torsion angle values. With this reserve, these biasing mechanisms have proven to speed up the overall progression of the populations.

c) Deterministic optimizations: in addition to an occasionally applied conjugated gradient relaxation of individuals (or 'Lamarckian optimization', [15]), a new heuristic has been implemented, taking advantage of both deterministic optimization and stochastic mutations. This search strategy, which actually relies on the 'Torsional Angle Driving' procedures [16], forces one randomly chosen degree of freedom towards a randomly determined target value, by means of an artificial harmonic constraint term added to the energy function to be minimized. A conjugated gradient optimization then allows the torsions to relax in a concerted manner, according to this new fitness landscape, towards the desired torsional value, avoiding the clashes that would have probably arisen if rigid fragments would have been rotated around the given axis (as is the case in classical random mutation). As this deterministic optimization procedure is quite time consuming and would cause serious disruption of the evolutionary loop if run within the islands; it has therefore been programmed under the form of stand-alone 'explorer' processes, started by a $\mathcal{GA}$ run.

III. Meta optimization

The performance of the Conformational Sampling $\mathcal{GA}$ ($\mathcal{CSGA}$) being quite sensitive with respect to the choice of the control parameter values, this choice has been addressed by means of a meta layer of optimization, favoring parameters sets that enhance the search procedure.

The '$\mathcal{CSGA}$ success' optimality criterion (equation 2), took into account both computational time and the so-called '*free energy*' of the sampled conformer ensemble (implicitly accounting for multimodality) at the current operational setup.

$$\begin{aligned} \mu Fitness &= -k_B T \times \ln \left[\sum_{\substack{i \in \text{ found} \\ \text{conformers}}} \exp\left(-\frac{E_i}{k_B T}\right) \right] \\ &+ \alpha \times Time \end{aligned} \quad (2)$$

The importances of the meta optimization procedure and the hybridizations was analysed in details elsewhere [10]. This optimized and hybridized tool was able to process bigger molecules (up to a hundred degrees of freedom) at the atomic level in acceptable computing times ($\sim$ one week).

IV. Massively parallel deployment — Planetary model

The above described hybrid Darwinian process is started simultaneously on an arbitrary, user-defined number of planets (nodes): a dispatcher script attempts to deploy island models on as many nodes as requested, if it can find the ressources on the GRID. There is no 'interplanetary' communication at all: fit solutions may only be swapped between islands. Once an island model is completed according to the locally specified termination criteria, or the generic reservation time of that node is about to expire, the pilot script in charge of running the island model will, before termination, send the locally sampled results back to the dispatcher, which will join them to the 'Universal' pool of solutions. Liberation of a node will prompt the dispatcher to restart an island model there, until a total (user-specified) number of sets of results were successfully retrieved, or until the latest (user-defined) N retrieved results failed to contain any fitter solutions. The exact behavior of the starting island model is controlled by a set of operational parameters dictated by the dispatcher, which actively tries to optimize these in order to achieve better sampling capacity of the further runs.

Like in the workstation version, the meta-optimization of the operational parameters is performed by learning from previous runs, though a simple genetic algorithm, which runs asynchronously in the planetary model (upon termination of a node, its sampling success is brought in relation to the operational parameters it had used, and this knowledge is stored in a database serving to pick a new operational parameter configuration whenever the next node is due to start).

A. Panspermia

A key element of our deployment strategy is 'panspermia', so entitled after the hypothesis that life on Earth might have been seeded by microorganisms from space: the dispatcher may randomly pick a subset of the already visited solutions from the 'Universal' pool and 'seed' any newly started planet. The latter may use the provided sample to specify these as 'tabu' zones [17] — forcing the exploration of other phase space zones — or to replace the random initialization of chromosomes by cross-over products of these 'ancestors', thus allowing an in-depth exploration of promising phase space regions.

B. Intensification

Although the sampling procedure may rapidly generate structures in the neighborhood of the 'native' (experimentally determined) geometries, the extreme ruggedness of the response surface is such that important energy fluctuations depending on geometry details are certain to occur even within this minimum energy well. As a consequence, many structures that may be regarded as 'correct' according to geometric criteria may nevertheless display high energies and fail to rank among the populated states. In other words, the discovery of the lowest point of the rugged energy well harboring the populated geometries is far from being a trivial problem and may require important intensification efforts. A specific setup scheme for the $\mathcal{GA}$, for fine exploration of limited phase space zones has been designed. It does not start with a random set of chromosomes, but from

previously sampled geometries representing a same global fold, in search for states of similar overall geometry but lower energy. Obviously, intensification runs compete for ressources with the default exploratory runs.

C. Tabu zones

Heavily visited phase space zones where it is 'believed' (see details below) that the deepest local optimum within the zone has already been sampled should be declared tabu areas. This amounts to (i.) eliminating the concerned chromosomes from the pool of 'ancestors' used for intensification and (ii.) defining an exclusion zone around each such chromosome. Any solution close, according to a to-be-defined similarity metric and similarity cut-off, to any tabu chromosome, and of higher energy than the tabu chromosome, will be assigned an abnormally low fitness score in order to force its demise at the next Darwinian selection step. If the new solution is fitter than the tabu chromosome, it will replace the latter. The choice of the similarity metric and cut-off is paramount: a too small cut-off discards only almost-identical pairs of solutions and unnecessarily spare redundant ones. On the opposite, too broad taboo areas may 'block' the access to unexplored deeper local minima in the neighborhood. In the present work we used a weighted block distance score in torsion angle space as a similarity metric of the two torsion angle vectors $\overrightarrow{\Theta}$, $\overrightarrow{\Theta}^{\text{tabu}}$:

$$\text{DISSIM}(\overrightarrow{\Theta}, \overrightarrow{\Theta}^{\text{tabu}}) = \sum_{i=1}^{N} w_i \times \Delta(\Theta_i, \Theta_i^{tabu}) \qquad (3)$$

where w_i is a weighting factor depending on fragment sizes, in order to tolerate larger variations with respect to terminal torsions, and Δ is the minimal positive rotation angle required to move from one torsional state to the other (e.g. 2 degrees to go from $\Theta_1 = 1$ degree to $\Theta_1^{tabu} = 359$ degrees, for example). Both the way in which torsional weighting factors are calculated with respect to the moving fragment sizes ($w_i = 0$ if fragment size $< MIN_{\text{FRAGSIZE}}$; $w_i = 1$ above MAX_{FRAGSIZE}; linear interpolation between these extremes) and the imposed tabu cut-off MIN_{DISSIM} are key control factors of the shape of the 'ellipsoidal' tabu zone around the tabu chromosome — several working hypotheses have been explored. In particular, all conformers differing only in terms of degrees of freedom associated to terminal fragments of MIN_{FRAGSIZE} and less become tabu.

As soon as regular diversification runs led to the discovery of a tunable minimal number of related geometries (regrouped according to a clustering procedure in torsional space, based on a chromosome dissimilarity score related to equation 3), the next planet will be dedicated to intensification within the phase space zone they populate. The key challenge of an optimal panspermia strategy is to decide at which point a cluster used as attractor in intensification searches has been sufficiently well sampled, in order to declare tabu the area around its cluster 'head' (its representative, most stable of its members). A too early decision in this sense may prematurely block the discovery of deep energy wells, while a too late one will translate in wasted computational time, at a scale proportional to the total number of independent solution clusters (of the order of $10^5 \ldots 10^6$ for a mini-protein like 1LE1 or 1L2Y). Common sense might suggest that intensification should be applied only to clusters of reasonably low energies, but in reality the ruggedness of the energy landscape is such that the energies of the first 'raw' conformers found by the diversification simulations that discovered the new clusters are completely uncorrelated with the final energies of fine-tuned geometries found by intensification in the immediate neighborhood. Restricting intensification to 'promising' solution clusters only is thus risky. The number N_{intens} of maximally tolerated intensification attempts of a cluster (set to 5, by default) is thus a key parameter of the panspermia strategy. Furthermore, the considered clusters are dynamic entities: when the newly added member is more stable than the current cluster head, it will replace the latter and recenter the cluster around the new head. Steadily evolving clusters will not become tabu — the number of maximally tolerated intensification attempts only applies if the cluster head remained unchallenged by the results of these biased searches (details not shown).

V. Results, discussion, prospect

Up-to-date attempts to use the planetary model led to successful folding experiments of the Tryptophane cage (α-helix) and Tryptophane zipper (β-sheet), as well as of key β-sheets and loops of the PIN1 WW domain in a matter of few days, using only a small subset (20-30 nodes) of GRID5000. Simulation results for the two first benchmark molecules will be discussed here.

The tryptophane cage contains an alpha-helical moiety stacked against an extended sequence to which it connects through a loop formed by 4 aminoacids (73 degrees of freedom, including both torsional axes of the protein backbone — except for the rigid peptidic bonds — and sidechains). α-helices are structural elements that fold quickly in solution, being stabilized by local, energetically favorable hydrogen bonds involving a residue and its 3rd successive neighbor. This situation is well suited for GA-based sampling: a helix turn is controlled by 6 degrees of freedom only, i.e. may quite easily emerge by hazard in a chromosome (and perhaps benefit from refinement by "Lamarckian" gradient optimization). Being stabilized by internal hydrogen bonds, this structural element may readily be inherited by the successors until a favorable cross-over may couple two spontaneously emerged helix loops together. Accordingly, the planetary model has successfully and reproducibly discovered geometries as shown in figure 2 that are very close to the native 1L2Y fold reported in literature (white — native geometry; red — typical folded structure). Furthermore, the most stable of all sampled conformers was systematically found to be one of the correctly folded structures.

By contrast, although the tryptophane zipper consists only 53 degrees of freedom, it is nevertheless more difficult to

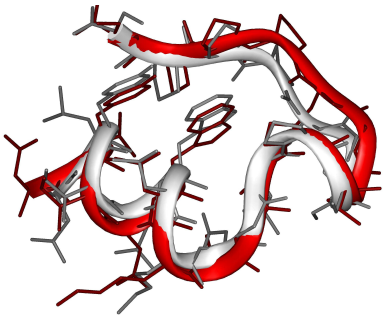


Fig. 2. Native state of 1L2Y, ranked as first among output conformers.

fold computationally than 1L2Y. The main reason is the β-hairpin structure it adopts, where stabilizing hydrogen bonds stem from topologically remote pairs of aminoacids. The β sheet "zipper" is a cooperative element: it gains stability only when fully structured: chromosomes displaying partly folded sheets will not benefit from stabilization, i.e. do not have any obvious evolutionary advantage. This notwithstanding, β-hairpin structures (correctly folded protein backbones) have been reproducibly obtained by planetary model-based simulations. In rare cases (2 out of several tens), the simulation actually returned a perfect replica of the experimental fold, both in terms of backbone and side chain orientations (figure 3), with the native geometry shown in white. This calculated geometry was also shown to be the most stable of all the ever visited 1LE1 conformers.

Typical simulations, however, will return geometries like in figure 4, where the backbone is correctly folded but sidechains are misplaced (are predicted to interact differently with each other). Furthermore, the alternative side chain interactions proposed by the model do make physico-chemical sense: they are aromatic stacking interactions of a same nature as the one seen in the native geometries. The differences between the two structures are subtle, the second is not obviously wrong and it may actually correspond to some less populated species which does exist in solution but escapes detection by state-of-the-art experimental methods. However, the energy of such a conformer is significantly higher than the one of the native state and, unfortunately, also higher than the one of misfolded structures like in figure 5. In that simulation, the almost correct fold 4 was ranked as 79^{th} most stable geometry out of several hundreds of thousands. If the geometry of 1LE1 would not have been known, this simulation would have erroneously predicted the misfolded geometry 5 instead of the almost correct fold 4.

Evolving the latter into the properly folded 3 may require a quite lengthy intensification simulation. An exhaustive search for an optimal 'panspermia' approach (guaranteeing the reproducible discovery of a 'native' geometry at the lowest energy level among the sampled conformers) does however not appear to be feasible: it would require the tuning of at least four parameters (N_{intens}, MIN_{FRAGSIZE}, MAX_{FRAGSIZE} and MIN_{DISSIM}, not mentioning the ones controlling cluster definition). Multiple simulations (of $20 \ldots 50$ hours each $\times 20 \ldots 30$ nodes or more for problems larger than 1LE1 or 1L2Y) would be required for due assessment of the reproducibility at each parameter combination. The termination criteria of the method should also be subject to scrutiny - would more important simulation efforts ensure the desired reproducibility? If so, which parameter should be first increased: the number of allocated planets or the total physical time? The obtained results show that reproducibility is not solely a matter of allocated ressources: note that the correctly folded 3 differs from the almost correctly folded 4 only by the placement of some low-weight side chains. Depending on the choice of MIN_{FRAGSIZE} and MAX_{FRAGSIZE}, the weighting factors from equation 3 may be such that the correct fold 3 actually falls within the tabu zone instated after the discovery of a structure like 4. If so, it will never be found, no matter for how long time the simulation continues. Renouncing the tabu strategy altogether is not an option, however: the simulations showed — and it makes perfect physical sense — that stable misfolded geometries, representing broader local optima than the native state, are reproducibly the first to be visited during the simulation. This would therefore systematically return to these same attraction pools each time a new run is started, unless tabu zones are declared. The native state owns its stability to more favorable intramolecular contacts. Or, a more compact packing of the protein chain is needed to enable more favorable contacts. This also means that any misplaced terminal fragment is likely to cause heavily penalizing intermolecular clashes, whereas in unfolded geometries side chains are free to move around in solvent. Protein folding amounts to an 'all-or-nothing' situation: the most stable states are achieved if either all degrees of freedom adopt their native values, or none of them do (i.e. all adopt random coil values corresponding to an unstructured peptide chain in solution). Situations in which most of the degrees of freedom are properly set, but a few of them are not, are likely to correspond to highly unfavorable energies due to clashes. The native state is a narrow but deep local minimum surrounded by an 'activation energy' barrier. As mentioned before, 1LE1 expectedly displays a much more marked 'all-or-nothing' behavior intrinsic to β-sheet folds. Therefore, optimal setup of the panspermia strategy is problem-dependent.

An alternative way to address the conformational problem is currently being considered: a thorough search of the maximal phase space volume that may be reproducibly sampled by local intensification procedures will be conducted, using diverse randomly picked phase space zones of different compounds. Phase space will be then divided into cells, optimally defined according to this study, and the overall conformational search will be conducted in this "discretized" problem space, where the fitness score of each phase space cell will be given by the free energy score returned by the local intensification simulation. In a broader perspective, novel deployment strategies using the PARADISEO[2] [18]

[2]http://paradiseo.gforge.inria.fr

core library for genetic algorithm deployment on the GRID will also be explored and compared to the planetary strategy, in search of a procedure optimally exploiting the potential of GRID5000 for solving molecular modeling problems.

Fig. 3. The almost correct geometry is found among more stable misfolds.

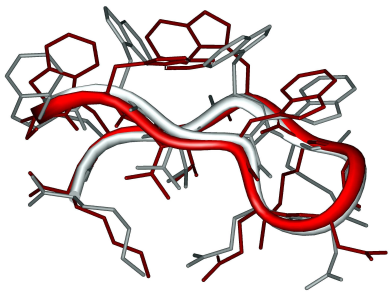


Fig. 4. Almost correctly folded geometry with correctly folded main chain but misplaced side chains, ranked only 79th in terms of stability.

Fig. 5. Top ranked misfolded geometry.

REFERENCES

[1] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, pp. 70–75, 2004.

[2] P. Crescenzi, D. Goldman, C. H. Papadimitriou, A. Piccolboni, and M. Yannakakis, "On the complexity of protein folding," *Journal of Computational Biology*, vol. 5, no. 3, pp. 423–466, 1998.

[3] R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," *Journal of Molecular Biology*, vol. 231, no. 1, pp. 75–81, may 1993.

[4] C. Levinthal, "How to fold graciously," in *Mossbauer Spectroscopy in Biological Systems*. University of Illinois Press: Proceedings of a meeting held at Allerton House, Monticello, Illinois, 1969, pp. 22–24.

[5] D. J. Wales and T. V. Bogdan, "Potential energy and free energy landscapes," *J. Phys. Chem.*, vol. 110, no. 42, pp. 20 765—20 776, 2006.

[6] A. T. Hagler, E. Huler, and S. Lifson, "Energy functions for peptides and proteins. i. derivation of a consistent force field including the hydrogen bond from amide crystals," *Journal of American Chemical Society*, vol. 96, no. 17, pp. 5319–5327, aug 1974.

[7] A. T. Hagler and S. Lifson, "Energy functions for peptides and proteins. ii. the amide hydrogen bond and calculation of amide crystal properties," *Journal of American Chemical Society*, vol. 96, no. 17, pp. 5327–5335, aug 1974.

[8] D. Horvath, "A virtual screening approach applied to the search for trypanothione reductase inhibitors," *Journal of Medicinal Chemistry*, vol. 40, no. 15, pp. 2412–2423, 1997.

[9] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, University of Michigan Press, 1975.

[10] B. Parent, A. Kökösy, and D. Horvath, "Optimized evolutionnary strategies in conformational sampling," *Journal of Soft Computing*, vol. 11, no. 1, jan 2007.

[11] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nature Structural Biology*, vol. 9, pp. 452–430, apr 2002.

[12] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, "Tryptophan zippers: Stable, monomeric β-hairpins," *Proc Natl Acad Sci USA*, vol. 98, no. 10, pp. 5578–5583, may 2001.

[13] H. Nguyen, M. J. M, J. Kelly, and M. Gruebele, "Engineering a beta-sheet protein toward the folding speed limit," *The Journal of Physical Chemistry B Condens Matter Mater Surf Interfaces Biophys.*, vol. 109, no. 32, pp. 15 182–15 186, aug 2005.

[14] K. Vertanen, "Genetic adventures in parallel: Towards a good island model under pvm," *Oregon State University*, 1998.

[15] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, "Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function," *Journal of Computational Chemistry*, vol. 19, no. 14, pp. 1639–1662, jun 1998.

[16] Accelery, "Accelerys discover simulation package." [Online]. Available: http://www.accelerys.com/insight/discover.html

[17] F. Glover, J. P. Kelly, and M. Laguna, "Genetic algorithms and tabu search: hybrids for optimization," *Computers and Operations Research*, vol. 22, no. 1, pp. 111–134, 1995.

[18] S. Cahon, N. Melab, and E.-G. Talbi, "Paradiseo: A framework for the reusable design of parallel and distributed metaheuristics," *Journal of Heuristics*, vol. 10, no. 3, pp. 357—380, 2004.

Annexe H

Affiche 1 : Gordon Conference, Suisse, 2006

Affiche lors de la « Computational Chemistry Gordon Research Conference » du 8 au 13 octobre 2006.

Benjamin Parent, Guy Lippens, Dragos Horvath

Steps towards an Ensemble-Based Force Field Fitting Procedure.

Steps towards an Ensemble-Based Force Field Fitting Procedure

Benjamin Parent, Guy Lippens & Dragos Horvath

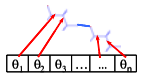
UMR 8576 CNRS/Université des Sciences & Technologies de Lille, Bât C9, 59655 Villeneuve d'Ascq, France

The problem: The classical force fields (FFs) used in molecular mechanics and dynamics were typically parameterized with respect to structural and energy barrier data of small molecules. Protein FFs are calibrated such as to guarantee that native folds actually correspond to an energy minimum of the structure-energy response surface, in the sense that a typical dynamics simulation at 300K of a stable protein is not expected to leave the neighborhood of the native fold. It is however unclear if the current force fields provide an accurate description over the entire phase space of (macro)molecules, or whether exhaustive conformational sampling methods which may easily tunnel through energy barriers might discover spurious, deeper energy minima corresponding to non-native folds. The existence of alternative minima is of little relevance for classical MD simulations gravitating around the native fold, but may become an issue in ab-initio all-atom protein folding simulations, which were only recently rendered feasible by the use of massively parallel computational resources. However, determining whether such alternative deeper minima are indeed spurious requires an in-depth analysis in terms of conformational free energies (deep but narrow alternative minima, not populated for entropic reasons, are tolerable). Estimating the conformational free energy directly from calculated partition functions is practically impossible in an all-atom explicit-solvent simulation including high-frequency bond stretching and angle-bending vibrational terms. Therefore, little is known about the overall accuracy of classical FFs throughout the entire phase space of a folding problem.
The goal: calibration of an empirical molecular FF for conformational sampling and docking. Generally applicable to proteins, sugars, organic ligands, tailor-made for use with torsional degrees of freedom only and using a continuum solvent model, it should be ***consistent***, in the sense that docking affinities & folding propensities should be directly linked to computed force field energies of sampled ensembles. *A posteriori* rescoring of docking poses should no longer be required, e.g. the docking problem should be reduced to a simultaneous conformational sampling of several molecules.

The approach: Our group (Parent et. al, *Soft Computing*, DOI: 10.1007/S00500-006-0053-y, **2006**) has recently developed a hybrid Genetic-algorithm-based conformational sampling method for problems of 100-200 torsional degrees of freedom – please refer to slides GA-1…5 below. It ignores bond stretching and angle bending, is based on CVFF (Hagler et. al., *J Am Chem Soc*, 96,5319-27, **1974**) van der Waals terms and includes a simple continuum desolvation model (Horvath, *J.Med.Chem* 15, 2412-23, **1997**). The method proved able to detect lower non-native energy minima corresponding to the initial FF setup – or to sample the native fold, whenever this coincided with the lowest energies found. Instead of an accurate estimation of conformational free energies, we rely on a semiquantitave criterion to decide whether the alternative minima are spurious: on hand of extensive torsional Monte Carlo sampling of the neighborhood of the native fold, a free energy index of the native state is determined from the herein obtained partition function. Alternative minima found by the evolutionary algorithm, however, are represented by a single state, e.g. their "free energy" may not be lower than their energy. If the energy of an alternative minimum lays below the free energy of the native state – in spite of artificially favoring the latter from an entropic point of view – then this is a clear proof of a force field failure and force field parameters need to be adjusted in order to appropriately reposition the relative energy levels of native and non-native states.
On the basis of a learning set including small structured proteins - the Tryptophane cage (1L2Y) the Tryptophane zipper (1LE1) , the WW domain of PIN1, etc., sugars – cyclodextrine, or chemically modified peptides, our first goal is to find a force field setup void of any "spurious" minima in the above-mentioned sense. This is a necessary, but not yet sufficient condition for FF accuracy. Furthermore, the question could be raised whether a "self-consistent" FF may be found, in the sense that experimental conformational free energies can be reproduced on hand of calculated partition functions from visited geometries – meaning that FF parameters must be chosen such as to compensate for the artifacts introduced by the ignoring of stretch-bend contributions and for the typical artifacts due to the inherent incompleteness of the sampling process itself.

(GA-1) Genetic Algorithm-based Conformational Sampling Tool

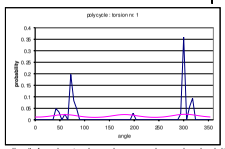
- Conformers are coded as "chromosomes" in which each locus stands for a torsional angle value.



- The efficacy of the computationally simulated Darwinian Evolution process (**offspring generation** by **cross-over** and **mutation** followed by the **selection** of the fittest), was enhanced by
  - hybridization with various optimization heuristics
  - Fine-tuning of the parameters controlling the evolutionary strategy

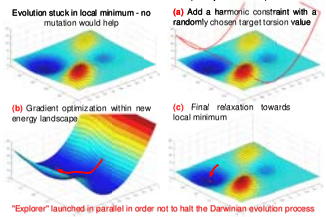
(GA-2) Hybrid Heuristics: (1)-Targeted torsion angle choice

- Biasing the probabilities to draw a given value for a given angle (according to a temperature parameter):
  - ➢ **Knowledge-based bias: favoring locally stable torsions…**



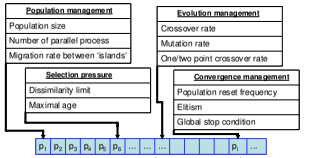
  - ➢ "Traditionalism": favoring torsion values seen in previously visited samples

(GA-3) Hybrid Heuristics (2) – Directed Mutants (Explorers)



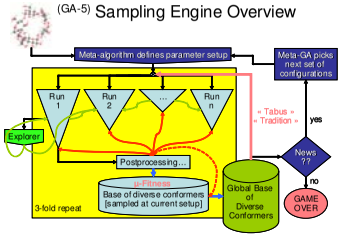
"Explorer" launched in parallel in order not to halt the Darwinian evolution process

(GA-4) Meta-Optimization: Search for Optimal Operational Parameters



$$\mu\_Fitness(p_1, p_2, ..., p_i, ...) = -k_b T . \ln\left(\sum_{found\ minima} \exp\left(-\frac{E_i}{k_b T} \right) \right) + \alpha . CPUtime$$

(GA-5) Sampling Engine Overview



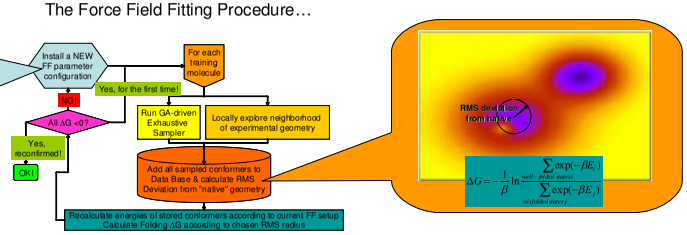
Up-to-date Results: After nine steps in force field parameter space (see right-hand schema), ab initio folding simulations with the latest set of parameters show an overall improved propensity to (a) sample native states and (b) to rank the native states among the most stable of the obtained conformer lists. Opposite images illustrate, for several of the training set molecules, the overlay between native and the closest-to-native of the GA-sampled geometries, using each of the N most recent visited force field setups. Both RMS deviation from native geometry and the rank number of the closest-to-native conformer in the energy-ranked conformer set are given (ideally, RMS<<1Å and rank# =1). More recent force field parameter sets can be seen to be more successful

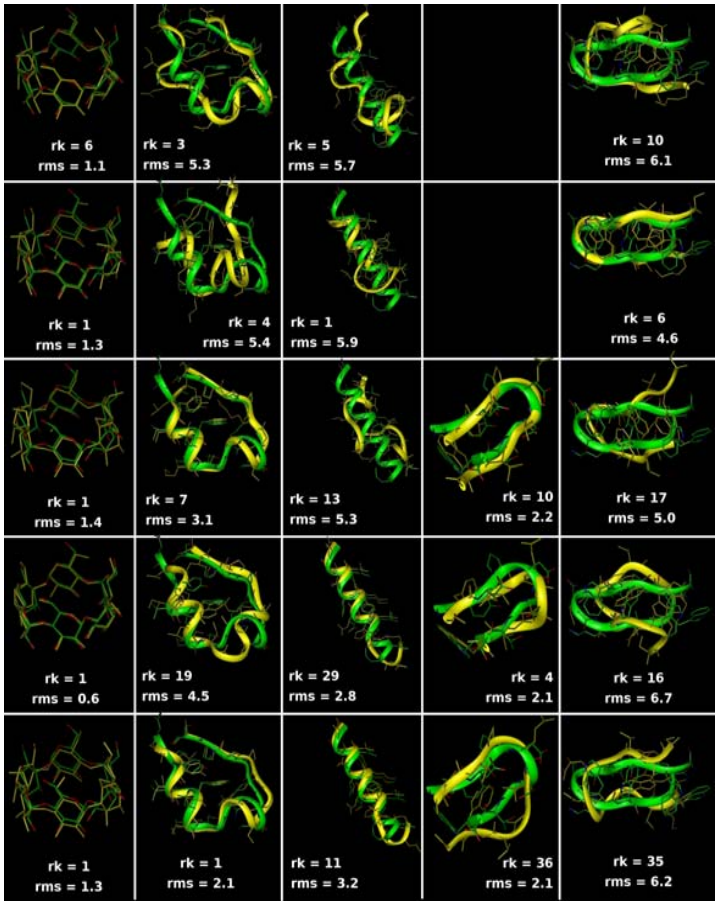
The Force Field Fitting Procedure…

- *Customized* CVFF force field, including a continuum solvent model:

$$E_{Solv} = k_{solv} \frac{Q_i^2 V_j + Q_j^2 V_i}{d_{i,j}^4} - k_b \delta^{hphob}(i,j) \qquad E_{coulomb} = \frac{\delta_i \delta_j}{4\pi\varepsilon_{dd} * d_{ij}^2}$$

- Current list of fittable force field parameters:
  - Distance-dependent dielectric constant
  - Weighing factor of the desolvation penalty
  - Weighing factor of the hydrophobic contacts
  - Weighing factor of repulsive van der Waals
  - Attractive & repulsive van der Waals coefficients of the following type: 'tcf' (carbonyl C), 'o' (ether-type O), 'hr (aliphatic H), 'cp' (aromatic C), 'tcf' (carbonyl O)





Annexe I

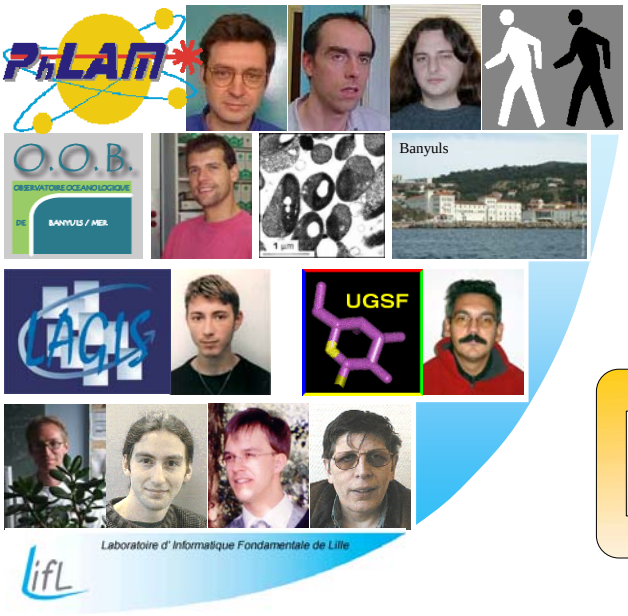
Affiche 2 : Computational Biology, Lille, 2006

Affiche lors du « Gent-Lille workshop on computational biology » du 20 juin 2006.

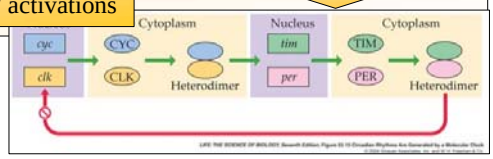
M. Lefranc, S. Bielwsky, F.-Y. Bouget, F. Boulier, F. Lemaire, S. Picault, M. Petitot, D. Horvath, Q. Thommen, P.-E. Morant, C. Vandermoere et Benjamin Parent

Studying, modeling and simulating circadian oscillations in regulatory networks

Studying, modeling & simulating circadian oscillations in regulatory networks

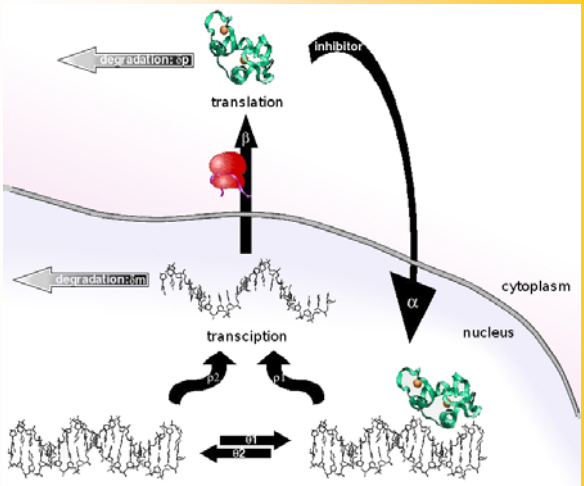


Regulatory networks → Mechanisms: Transcription / translation, Degradation / dimerization, Mutual inhibitons / activations → feed-{back / forward} loops



M. Lefranc, S. Bielawski, F-Y. Bouget, F. Boulier, F. Lemaire, S. Picault
M. Petitot, D. Horvath, Q. Thommen, P-E Morant, C. Vandermoere, B. Parent

Building minimal block allowing oscillations



Different approaches

- ✓ Deterministic non-linear differential equations (without & with delays)
- ✓ Stochastic multi-agent: (spatial diffusion & behavior specifications)
- ✓ Hybrid methods (stochastic / deterministic)
- ✓ Formal approaches

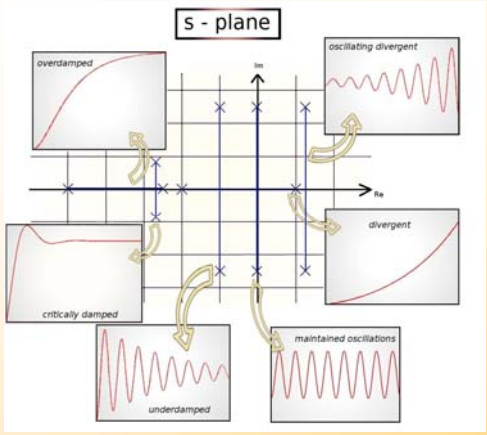
Protein degradation mechanism may influence the “near-equilibrium” behavior

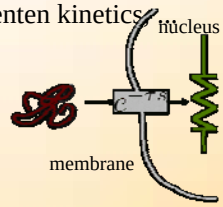
$$\frac{dP}{dt} = \cdots - f(P)$$

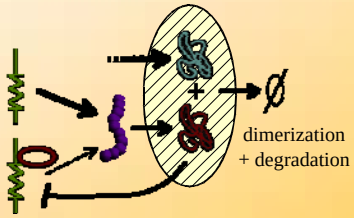
Destabilizing mechanisms

Delays, dimerizations, Michaëlis-Menten kinetics

Oscillation quest in parameter space



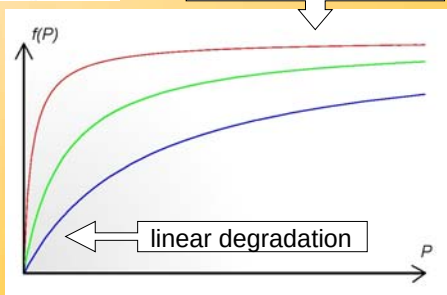


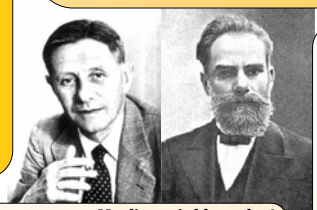


Michaëlis - Menten kinetics appears to be a key while engineering oscillators and trying to destabilize overdamped systems. (considering Hopf criterion)

$$\text{degradation} = f(P) = \frac{V_M P}{\kappa + P}$$

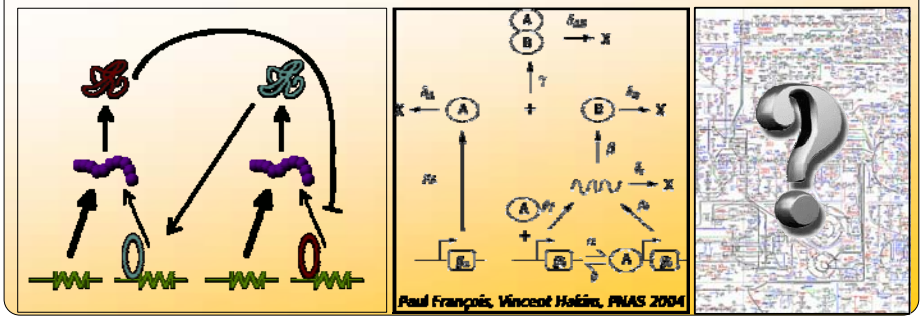
constant degradation





Nonlinear is kharacho !

Towards more complex models



Experimental (counter?)-part

Study of circadian and cell division cycles in Ostreococcus Tauri algae:
- identify components
- evidence coupling between cycles

Annexe J

Article relatif à l'affiche 3 : Rencontres du Non-Linéaire, Paris, 2007

« Rencontre du Non-Linéaire », 15 et 16 mars 2007, Paris.

Pierre-Emmanuel Morant, Constant Vandermoere, Quentin Thommen, Benjamin Parent, François Lemaire, Florence Corellou, Christian Schwartz, François-Yves Bouget, Marc Lefranc

Oscillateurs génétiques simples. Application à l'horloge circadienne d'une algue unicellulaire.

Oscillateurs génétiques simples. Application à l'horloge circadienne d'une algue unicellulaire

Pierre-Emmanuel Morant[1], Constant Vandermoere[1], Quentin Thommen[1], Benjamin Parent[2], François Lemaire[3], Florence Corellou[4], Christian Schwartz[4], François-Yves Bouget[4], Marc Lefranc[1]

[1] Laboratoire de Physique des Lasers, Atomes, Molécules, UMR CNRS 8523, UFR de Physique, Bât. P5, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France.
[2] Unité de Glycobiologie Structurale et Fonctionnelle, UMR CNRS 8576, Bât. C9, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France.
[3] Laboratoire d'Informatique Fondamentale de Lille, UMR CNRS 8022, Bât. M3, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France.
[4] Laboratoire Modèles en Biologie Cellulaire et Evolutive, UMR CNRS-Paris 6 7628, Observatoire Océanologique de Banyuls sur mer, BP44, 66651 Banyuls sur Mer Cedex, France.
marc.lefranc@univ-lille1.fr

Résumé. Un gène réprimé par l'expression de sa propre protéine constitue l'exemple le plus simple de circuit génétique à boucle de rétroaction négative, et l'apparition d'oscillations dans ce système est un problème classique de la biologie théorique. Nous nous intéressons ici au cas où le taux de transcription ne suit pas instantanément la concentration en protéine, mais se comporte comme une variable dynamique indépendante. Nous observons que l'existence d'une dynamique transcriptionnelle favorise les oscillations, et que ces dernières apparaissent de manière systématique dans la limite où les dégradations de l'ARN et de la protéine sont totalement saturées. Nous considérons également la généralisation la plus directe du gène auto-régulé : une boucle à deux gènes, l'un activateur, l'autre répresseur, se régulant réciproquement, et nous comparons ses prédictions auux données expérimentales concernant les oscillations circadiennes d'une algue unicellulaire verte.

Abstract. A gene which is repressed by its own protein is the simplest example of a genetic circuit with a negative feedback, and the appearance of oscillations in this system is a classical problem in theoretical biology. Here we study the case where the transcription rate does not react instantaneously to changes in protein concentration but is an independent dynamical variable. We observe that the transcriptional dynamics favors oscillations, and that periodic regimes appear unconditionnaly in the limit where enzymatic degradations of ARN and protein are completely saturated. We also consider the simplest generalization of this oscillatior, a circuit with two genes, an activator and a repressor, regulating each other, and compare its predictions to experimental data about circadian oscillations in a unicellular green alga.

1 Introduction

Les dizaines de milliers de gènes que porte la molécule d'ADN au coeur de chaque cellule contiennent l'information nécessaire à la synthèse des briques de la machinerie moléculaire de la Vie, les protéines. Cette synthèse s'effectue en deux étapes : "transcription" de la séquence codante en une molécule d'ARN messager, puis "traduction" de cet ARN en une séquence d'acides aminés, c'est-à-dire une protéine. Or, les taux de production des ARN ne sont pas constants : l'activité des gènes est en effet régulée par des protéines produites par d'autres gènes, au travers de réseaux complexes. L'ensemble constitue donc un système dynamique fortement non linéaire, susceptible de présenter toute une gamme de comportements bien connus : bistabilité, mais aussi oscillations, comme par exemple celles intervenant dans la segmentation des somites lors de l'embryogénèse [1], ou dans les horloges circadiennes [2]. Ces dernières fournissent à un grand nombre d'organismes une mesure interne du temps leur permettant de faire varier de nombreuses grandeurs physiologiques sur une période de 24 heures, et de s'adapter ainsi à l'alternance jour-nuit. Leur caractère autonome est démontré par le fait qu'elles persistent en éclairement constant, avec une période naturelle légèrement différente de 24 heures.

L'oscillateur génétique le plus simple est a priori celui constitué d'un gène réprimé par la protéine qu'il produit, comme sans doute le gène *hes1* dans la segmentation des somites [1]. Il s'agit d'un problème ancien [3,4], pour lequel il est admis qu'on ne peut observer des oscillations que si on introduit soit une étape cinétique intermédiaire, par exemple une phosphorylation de la protéine [5] ou un transport entre cytoplasme et noyau [6], soit un terme explicite de délai dans les équations [7,8,9]. Nous avons revisité ce problème en tenant compte de deux effets complémentaires.

D'une part, des expériences récentes ont montré que le processus de transcription se caractérise par une cinétique complexe, et notamment par l'existence de "salves de transcription" [10] modulant l'activité transcriptionnelle sur des durées allant jusqu'à quelques dizaines de minutes. Comme François et Hakim [11], nous considérons donc le taux de transcription comme une variable dynamique à part entière, contrairement à l'immense majorité des études où on suppose qu'il réagit instantanément à la concentration en protéine. D'autre part, les analyses théoriques postulent généralement que les acteurs moléculaires sont dégradés par des mécanismes génériques, par exemple dégradation spontanée ou dirigée par une enzyme, avec une cinétique de type Michaelis-Menten. Or, l'importance de la cinétique de dégradation, et le pouvoir déstabilisant d'effets non linéaires, tels que la stabilisation de la forme dimère d'une protéine [12], ont été récemment soulignés [13].

Dans le cas du gène auto-régulé, nous avons constaté que l'existence d'une dynamique transcriptionnelle peut élargir considérablement le domaine de paramètres dans lequel un mécanisme de dégradation non linéaire induit des oscillations. Celles-ci sont observées de manière systématique dans la limite où les dégradations de l'ARN et de la protéine sont saturées mais peuvent apparaître bien avant.

2 Oscillations d'un gène réprimé par sa propre protéine

Comme François et Hakim [11], nous décrivons la dynamique transcriptionnelle par une simple équation cinétique décrivant des processus élémentaires d'association/dissociation entre la protéine et l'ADN, mais une modélisation plus complexe pourrait être envisagée. Dans ces conditions, la dynamique du circuit à un gène auto-régulé peut être modélisée par les trois équations adimensionnées suivantes :

$$\dot{g} = \theta\,[1 - g(1 + p^n)] \tag{1a}$$

$$\dot{p} = n\alpha\,[1 - g(1 + p^n)] + \delta[m - f(p)] \tag{1b}$$

$$\dot{m} = \mu + \lambda g - h(m) \tag{1c}$$

où g,p et m représentent respectivement l'activité du gène, et les quantités de protéines et d'ARN. L'entier n indique la coopérativité de la régulation, c'est-à-dire le nombre de protéines contenues dans le complexe protéique modulant l'activité du gène. L'unité de temps est le temps de demi-vie de l'ARN. Les coefficients θ, α contrôlent les échelles de temps des processus de dissociation et d'association à l'ADN, tandis que $1/\delta$ est le temps de demi-vie de la protéine. Les paramètres μ et λ déterminent l'activité du gène selon que celui-ci est libre et actif ($g = 1$) ou lié et réprimé ($g = 0$). Les fonctions $f(p)$ et $h(m)$, qu'on suppose de pente unité à l'origine, décrivent respectivement les mécanismes de dégradation de la protéine et de l'ARN.

Pour étudier l'apparition d'oscillations dans ce système, nous n'envisagerons ici que la déstabilisation de l'état stationnaire des équations (1) via une bifurcation de Hopf menant à des oscillations périodiques. L'analyse de stabilité linéaire du système (1) montre que deux valeurs propres de la matrice jacobienne traversent l'axe imaginaire et acquièrent une partie réelle positive quand l'expression $\mathcal{H}$ ci-dessous passe par zéro pour devenir négative (critère de Routh-Hurwitz) :

$$\begin{aligned}\mathcal{H} = {} & u{h_0}^2\,(\alpha h_0 + \delta s\lambda)\,(\delta s\lambda + u\lambda + \alpha h_0) \\ & + \lambda^2 h_0\left[h_0(-\delta h_0 + 2u\alpha + \alpha\delta s) + \lambda(u + \delta s)^2\right]\theta \\ & + \lambda^4\,(u + \delta s)\,\theta^2\end{aligned} \tag{2}$$

où s et u sont les pentes des fonctions de dégradation $f(p)$ et $h(m)$ au point fixe et h_0 est la valeur prise par la fonction de dégradation h(m) en ce point.

On voit facilement que lorsque $T = u + \delta s \leq 0$, l'expression (2) est négative pour toutes valeurs des constantes cinétiques θ et α. Cela indique que l'on observe alors systématiquement des oscillations, même pour des dynamiques transcriptionnelles extrêmement rapides, et en particulier dans le cas $u, s \to 0$ où les dégradations enzymatiques sur l'ARN et la protéine sont saturées, un facteur d'instabilité bien connu [14]. Cela n'a rien de surprenant, car $-T$ est la trace du jacobien du modèle à deux variables où l'activité du gène g est supposée être asservie à la concentration en protéine p, et l'on sait que pour un système à deux variables, la positivité de cette trace est synonyme d'instabilité [12,15].

L'expression (2) est plus intéressante si on adopte le point de vue que les constantes θ et α ne sont pas très grandes, comme on le suppose généralement, mais qu'elles doivent correspondre aux échelles de temps des "salves de transcription" observées expérimentalement. Ces dernières se caractérisent par des temps d'extinction allant jusqu'à quelques dizaines de minutes [10], soit $\theta = O(1)$. Nous avons observé que des oscillations peuvent alors apparaître pour des valeurs de $T = u + \delta s$ nettement positives, ce qui correspond à des dégradations nettement moins saturées que lorsque $\theta, \alpha \to \infty$. La figure 1 montre ainsi des oscillations observées pour $\theta \sim 0.25$, ce qui correspond à des temps d'extinction d'environ 40 minutes pour une demi-vie de l'ARN de 10 minutes. Les pentes des fonctions de dégradation au point fixe sont alors $u = 0.14$ et $s = 0.56$, à comparer à une valeur unité à faible concentration. On voit si la dégradation de la protéine est relativement saturée, celle de l'ARN ne l'est que modérément.

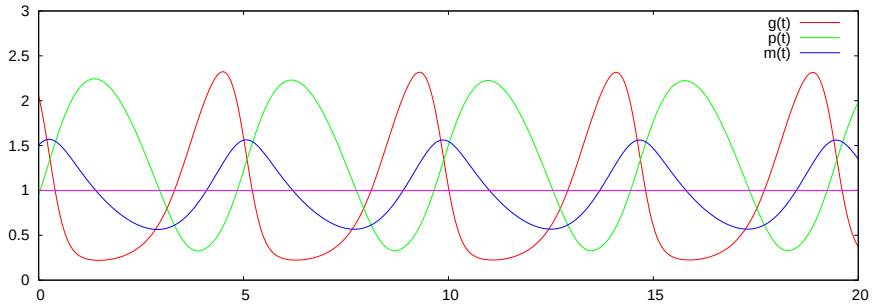


Fig.1. Oscillations du modèle (1) pour les valeurs des paramètres suivantes : $\theta = 0.25$, $\alpha = 8 \times 10^{-4}$, $\delta = 0.76$, $\lambda = 55.7$, $\mu = 0.6$, $n = 2$. Les variables g, p, m sont normalisées par rapport à leur valeur au point fixe. La protéine est supposée être dégradée par une enzyme allostérique avec une cinétique d'ordre 2, tandis que la dégradatation de l'ARN suit une cinétique de Michaelis-Menten classique. L'unité de temps est la demi-vie de l'ARN.

On peut légitimement se poser la question de la validité du modèle déterministe (1) si l'activité du gène g doit être considérée non comme une variable continue mais comme une variable stochastique alternant entre 0 et 1, et si les temps de commutation ne sont pas petits devant les temps d'évolution. A cela on peut répondre que les oscillations du modèle déterministe doivent se refléter de manière mesurable dans les propriétés statistisques du modèle stochastique, et entraîner par exemple une dispersion beaucoup moins importante des temps de commutation. D'autre part, il n'est pas exclu qu'une prise en compte plus fine des mécanismes de transcription montre la nécessité d'introduire certaines variables continues dans les description de ces mécanismes.

3 La boucle à deux gènes

Une généralisation naturelle du circuit à un gène auto-régulé est celui formé par une boucle de deux gènes, l'un activant le deuxième, le deuxième réprimant le premier. Nous utilisons dans ce qui suit un modèle semblable à (1), excepté que nous négligeons la dynamique transcriptionnelle. Nous nous intéressons ici à ce système en ce qu'il constitue un modèle minimal de l'horloge circadienne d'*Ostreococcus tauri*, une algue verte unicellulaire dont la physiologie et l'appareil génétique se caractérisent par

une compacité extrême, mais qui présente néanmoins de nombreux points communs avec les végétaux supérieurs. Deux gènes *TOC1* et *CCA1*, homologues de deux gènes centraux de l'horloge d'*Arabidopsis thaliana*, le modèle des végétaux supérieurs, ont pour l'instant été identifiés comme faisant partie de l'horloge circadienne de cette algue, qui est étudiée à l'Observatoire Océanologique de Banyuls.

En supposant des mécanismes de dégradation de type Michaelis-Menten, les équations réduites gouvernant la dynamique de la boucle à deux gènes peuvent s'écrire :

$$\frac{dm_T}{d\tau} = \mu_T + \frac{\lambda_T}{1+p_C^{n_C}} - \delta\,\frac{\kappa_{m_T}\,m_T}{\kappa_{m_T}+m_T} \tag{3a}$$

$$\frac{dp_T}{d\tau} = \delta_{p_T}(\,m_T - \frac{\kappa_{p_T}\,p_T}{\kappa_{p_T}+p_T}) \tag{3b}$$

$$\frac{dm_C}{d\tau} = \mu_C + \frac{\lambda_C\,p_T^{n_T}}{1+p_T^{n_T}} - \frac{\kappa_{m_C}\,m_C}{\kappa_{m_C}+m_C} \tag{3c}$$

$$\frac{dp_C}{d\tau} = \delta_{p_C}(\,m_C - \frac{\kappa_{p_C}\,p_C}{\kappa_{p_C}+p_C}). \tag{3d}$$

où m_T et p_T (m_C et p_C) représentent les quantités d'ARN et de protéine du gène *TOC1* (*CCA1*). Les paramètres $n_{T,C}$, $\lambda_{T,C}$, $\mu_{T,C}$ et δ, δ_{p_T,p_C} ont la même signifcation que dans (1). Les coefficients κ_i caractérisent la saturabilité des dégradations enzymatiques des différentes molécules en présence.

De même que pour le modèle (1), l'apparition d'oscillations dans le modèle (3) dépend de manière cruciale des mécanismes de dégradation. Plus précisément, il faut qu'au moins un certain nombre des quatre molécules impliquées dans la boucle soient dégradées de manière enzymatique, et que cette dégradation soit suffisamment saturée (à un moindre degré cependant que pour le circuit à un gène). Il est intéressant de noter au passage que le système (3) peut se ramener dans une certaine limite à la variante du célèbre oscillateur de Goodwin [3] donnée par Bliss *et al.* [17].

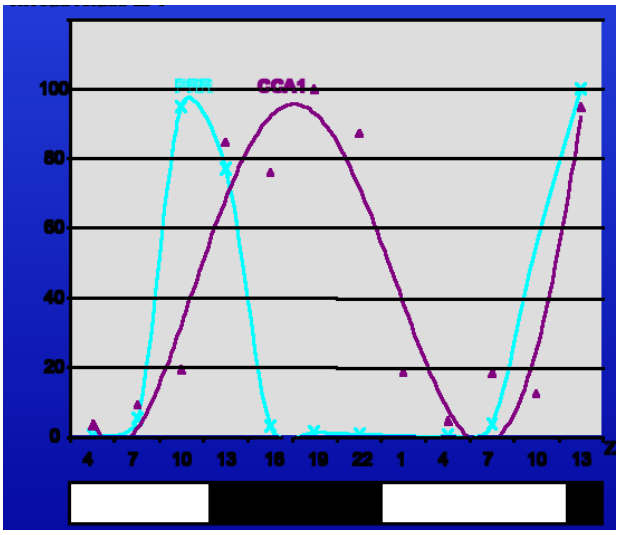


Fig.2. Niveaux d'expression en alternance jour/nuit des gènes *TOC1* (alias *PRR*) et *CCA1* d'O. tauri en fonction du temps circadien (CT : "circadian time"), CT0 correspondant au début du jour. Données expérimentales du groupe Horloge circadienne et cycle cellulaire de l'observatoire océanologique de Banyuls/mer. Malgré les incertitudes de mesure, on peut caractériser les deux courbes par des grandeurs relativement reproductibles. Ainsi, la quantité d'ARN de TOC1 est maximale vers CT10.5, avec une largeur à mi-hauteur d'environ 6 heures, et un long passage à zéro entre CT17 et CT7. En ce qui concerne CCA1, la présence de l'ARN est beaucoup plus étalée dans le temps, avec un pic vers CT17, une largeur à mi-hauteur d'environ 12 heures et un point bas aux alentours de CT7.

Notre but est de comparer les prédictions du modèle (3) aux données expérimentales concernant les variations dans le temps des ARN et des protéines de l'horloge. Cette comparaison est d'autant plus intéressante que la boucle TOC1/CCA1 a été un temps évoquée comme modèle pour l'horloge d'Arabidopsis [16,18] avant d'être délaissée au profit de circuits plus sophistiqués à plusieurs boucles de rétroaction [19]. Or, comme on le voit sur la figure 1, qui montre les variations dans le temps des niveaux d'ARN des gènes TOC1 et CCA1 en alternance jour/nuit, l'horloge d'*Ostreococcus* présente une différence importante avec celle d'*Arabidopsis* : CCA1 est à son maximum d'expression en début de nuit plutôt qu'au petit matin. Etant donné qu'Ostreococcus se caractérise généralement par une relative simplicité, il était donc important de déterminer si la boucle à deux gènes pourrait être un meilleur modèle pour cette algue que pour *Arabidopsis*. Dans un premier temps, nous nous sommes attachés à reproduire les régimes en alternance jour/nuit, généralement plus reproductibles que les régimes en éclairement constant.

Le modèle (3) décrit la régulation réciproque des gènes TOC1 et CCA1, mais ne précise pas le mécanisme d'action de la lumière sur la boucle. En l'absence d'informations précises, il nous faut donc envisager plusieurs scénarios différents, associés à des modulations différentes des paramètres. L'horloge pourrait être ainsi entraînée et synchronisée au cycle jour/nuit par une dégradation accélérée d'une protéine ou d'une autre, et ce le jour ou plutôt la nuit, ou encore par une réduction de l'activité transcriptionnelle d'une des deux protéines dans l'une des deux périodes. On peut évidemment espérer que les tests de ces différents mécanismes nous fournissent des pistes sur le couplage effectivement présent.

La figure 2 montre ainsi deux simulations préliminaires du modèle (3). Ces profils temporels ont été obtenus en cherchant des jeux de paramètres pour lesquels ils se rapprochaient le plus des données expérimentales (fig. 2). On constate sur la partie gauche de la figure que l'hypothèse d'une dégradation accélérée de la protéine TOC1 la nuit permet au modèle à deux gènes d'ajuster relativement bien les données expérimentales : les caractéristiques des profils expérimentaux et théoriques coïncident avec une très bonne précision, si ce n'est un pic de CCA1 un peu en avance. On note toutefois sur la figure 2 qu'il n'est pas exlu que ce pic arrive en fait plus tôt que ne l'indique la ligne tracée pour guider l'oeil.

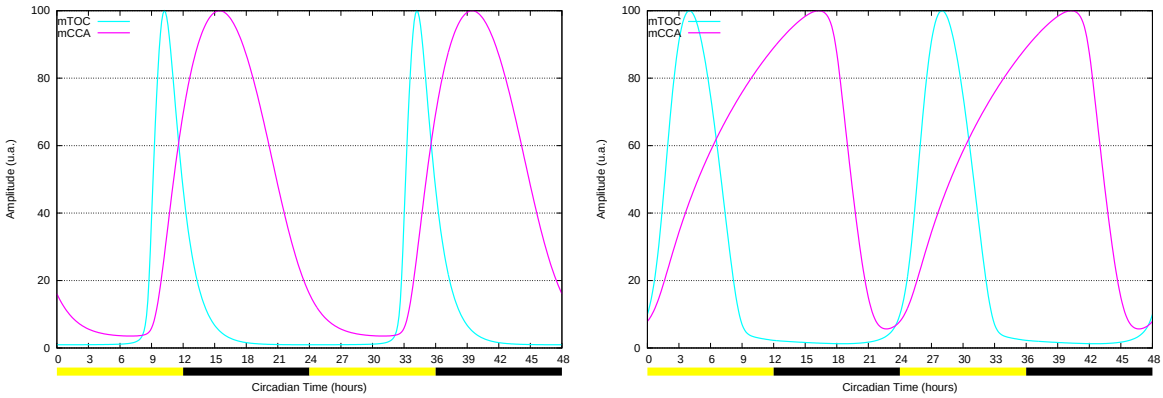


Fig.3. Simulations numériques du modèle (3) avec deux hypothèses différentes de coopérativité et de couplage de la lumière externe à la boucle génétique. Dans les deux cas, on teste un grand nombre de jeux de paramètres différents, et celui pour lequel les solutions s'approchent le plus de des courbes expérimentales de la figure 2 est retenue. (gauche) Dégradatation accélérée de la protéine TOC1 la nuit, régulation par un monomère de TOC1 et un dimère de CCA1 ; (droite) Dégradation accélérée de la protéine CCA1 la nuit, régulation par des monomères de TOC1 et de CCA1. On constate que l'hypothèse de gauche est nettement plus vraisemblable que celle de droite.

Evidemment, des comparaisons plus précises impliquant également les profils temporels des protéines ainsi que les données en éclairement constant seront nécessaires avant de se prononcer définitivement sur la pertinence du système (3) en tant que modèle de l'horloge circadienne d'*Ostreococcus*. Les résultats préliminaires présentés ici sont cependant étonnamment encourageants.

4 Conclusion

Nous avons observé que la prise en compte d'une dynamique transcriptionnelle élargit les zones de paramètres où des mécanismes de dégradation non linéaires peuvent induire des oscillations dans l'expression d'un gène réprimé par sa propre protéine. Ces mécanismes de dégradation sont également importants pour comprendre l'apparition d'oscillations dans la boucle à deux gènes, qui est par ailleurs un modèle hypothétique de l'horloge circadienne de l'algue unicellulaire *Ostreococcus tauri*. Des calculs préliminaires montrent qu'à condition de supposer certains modes d'action de la lumière sur les acteurs moléculaires, ce système semble bien reproduire les observations expérimentales.

Références

1. H. Hirata *et al.*, Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop, *Science* **298**, 840–843 (2002).
2. C. A. Strayer & S. A. Kay, The ins and outs of circadian regulated gene expression, *Curr. Opin. Plant. Biol.* **2**, 114 (1999).
3. B. C. Goodwin, Oscillatory behavior of enzymatic control processes, *Adv. Enzyme Regul.* **3**, 425-439 (1965).
4. J. S. Griffith, Mathematics of cellular control processes I. Negative feedback to one gene, *J. Theor. Biol.* **20**, 202 (1968).
5. A. Goldbeter, A model for circadian oscillations in the Drosophila period protein (PER), *Proc. R. Soc. Lond. B* **261**, 319 (1995).
6. J.-C. Leloup, D. Gonze, and A. Goldbeter, Limit cycle models for circadian rhythms based on transcriptional regulation in Drosophila and Neurospora, *J. Biol. Rhythms* **14**, 433 (1999).
7. M. H. Jensen, K. Sneppen & G. Tiana, Sustained oscillations and time delays in gene expression of protein Hes1, *FEBS Lett.* **541**, 176-177 (2003).
8. N. A. M. Monk, Oscilatory expression of Hes1, p53 and NK-κB driven by transcriptional time delays, *Curr. Biol.* **13**, 1409 (2003).
9. J. Lewis, Autoinhibition with transcriptional delay : a simple mechanism for the zebrafish somitogenesis oscillator, *Curr. Biol.* **13**, 1398 (2003).
10. I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Real-time kinetics of gene activity in individual bacteria, *Cell* **123**, 1025 (2005).
11. P. François & V. Hakim, Core genetic module : the mixed feedback loop, *Phys. Rev. E* **72**, 031908 (2005).
12. J. J. Tyson, C. .I. Hong, D. Thron and B. Novak, A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM, *Biophus. J.* **77**, 2411 (1999).
13. N. E. Buchler, U. Gerland, and T. Hwa, Nonlinear protein degradation and the function of genetic circuits, *Proc. Natl. Acad. Sci. USA* **102**, 9559 (2005).
14. A. Goldbeter, *Biochemical Oscillations and Cellular Rhythms : The molecular bases of periodic and chaotic behaviour* (Cambridge University Press, Cambridge, 1996).
15. C. P. Fall, E. S. Marland, J. M. Wagner, and J. J. Tyson, *Computational Cell Biology* (Springer, New York, 2002).
16. D. Alabadi, T. Oyama, M. J. Yanovsky, F. G. Harmon, P. Mas, S. A. Kay, Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock, *Science* **293**, 880 (2001).
17. R. D. Bliss, P. R. Painter, and A. G. Marr, Role of feedback inhibition in stabilizing the classical operon, *J. Theor. Biol.* **97**, 177 (1982).
18. J. C. W Locke, A. J. Millar, and M. S. Turner, Modelling genetic networks with noisy and varied experimental data : the circadian clock in *Arabidopsis thaliana*, *J. Theor. Biol.* **234**, 383 (2005).
19. J. C. W Locke, M. M. Southern, L. Kozma-Bognar, V. Hibberd, P. E. Brown, M. S. Turner, and A. J. Millar, Extension of a genetic network model by iterative experimentation and analysis, Mol. Systems Biol., doi :10.138/msb4100018.

Bibliographie

Accelerys (2005). Accelerys discover simulation package. San Diego, CA.

Aksimentiev, A., Balabin, I. A., Fillingame, R. H., et Schulten, K. (2004). Insights into the Molecular Mechanism of Rotation in the Fo Sector of ATP Synthase. *Biophys. J.*, 86(3) : 1332–1344.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., et Watson, J. D. (2002). *Molecular Biology of the Cell*. Garland, 4 edition.

Aldridge, B., Burke, J., Lauffenburger, D., et Sorger, P. (2006). Physicochemical modelling of cell signalling pathways. *Nature Cell Biology*, 8(11) : 1195–1203.

Alon, U. (2003). Biological networks : The tinkerer as an engineer. *Science*, 301 : 1866–1867.

Andersen, H. C. (1983). Rattle : a « velocity » version of the shake algorithm for molecular dynamics calculation. *Journal of Computational Physics*, 52(1) : 24–34.

Androulakis, I. P., Maranas, C. D., et Floudas, C. A. (1995). ffbb : a global optimization method for general constrained nonconvex problems. *Journal of Global Optimization*, 7 : 337–363.

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, 181(96) : 223–230.

Angeli, D. et Sontag, E. (2004). An analysis of a circadian model using the small-gain approach to monotone systems. Dans Publications, I., editeur, *Proceedings of the IEEE Conference Decision and Control*, pp. 575–578, Bahamas.

Antes, I., Merkwirth, C., et Lengauer, T. (2005). Poem : Parameter optimization using ensemble methods : Application to target specific scoring functions. *Journal of Chemical Information and Modeling*, 45(5) : 1291–1302.

Arkin, A., Ross, J., et McAdams, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected escherichia coli cells. *Genetics*, 149(4) : 1633–1648.

Atkinson, M. R., Savageau, M. A., Myers, J. T., et Ninfa, A. J. (2003). Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in escherichia coli. *Cell*, 113 : 597–607.

Audit, B., Vaillant, C., Arneodo, A., d'Aubenton Carafa, Y., et Thermes, C. (2002). Long-range correlations between dna bending sites : Relation to the structure and dynamics of nucleosomes. *Journal of Molecular Biology*, 316 : 903–918.

Auger, A., Schoenauer, M., et Vanhaecke, N. (2004). *Parallel Problem Solving from Nature - PPSN VIII*, chapter LS-CMA-ES : A Second-Order Algorithm for Covariance Matrix Adaptation, pp. 182–191. Lecture Notes in Computer Science. Springer Berlin / Heidelberg.

Baldwin, R. L. et Rose, G. D. (1999). Is protein folding hierarchic ? ii. folding intermediates and transition states. *Trends in Biochemical Sciences*, 24(2) : 77–83.

Balsalobre, A., Damiola, F., et Schibler, U. (1998). A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell*, 93 : 929–937.

Batada, N., Shepp, L., et Siegmund, D. (2004). Stochastic model of protein-protein interaction : why signaling proteins need to be colocalized. *PNAS*, 101(17) : 6445–6449.

Belle, A., Tanay, A., Bitincka, L., Shamir, R., et O'Shea, E. K. (2006). Quantification of protein half-lives in the budding yeast proteome. *PNAS*, 103(35) : 13004–13009.

Bissantz, C., Folkers, G., , et Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 43(25) : 4759–4767.

Blumenthal, L. M. et Menger, K. (1970). *Studies in Geometry*. W. H. Freeman & Co Ltd.

Bonachera, F., Parent, B., Barbosa, F., Froloff, N., et Horvath, D. (2006). Fuzzy tricentric pharmacophore fingerprints. 1. topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *Journal of Chemical Informatic Models*, 46(6) : 2457–2477.

Borel, E. (1913). Mécanique statistique et irréversibilité. *J. Phys.*, 3(5) : 189–196.

Borne, P., Dauphin-Tanguy, G., Richard, J.-P., Rotella, F., et Zambettakis, I. (1990). *Commande et optimisation des processus*. Technip, Paris, FRANCE, technip edition.

Braden, K. (2002). A simple approach to protein structure prediction using genetic algorithms. http ://www.genetic-programming.org/sp2002/Braden.pdf.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., et Karplus, M. (1983). Charmm : a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2) : 187–217.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., et Wolynes, P. G. (2004). Funnels, pathways, and the energy landscape of protein folding : A synthesis. *Proteins : Structure, Function, and Genetics*, 21(3) : 167–195.

Buchler, N. E., Gerland, U., et Hwa, T. (2005). Nonlinear protein degradation and the function of genetic circuits. *PNAS*, 102(27) : 9559–9564.

Bursulaya, B. D., Totrov, M., Abagyan, R., et Brooks, C. L. (2003). Comparative study of several algorithms for flexible ligand docking. *Journal of Computer-Aided Molecular Design*, 17(11) : 755–763.

Bussi, G., Gervasio, F. L., Laio, A., et Parrinello, M. (2006). Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *Journal of American Chemical Society*, 128(41) : 13435–13441.

Bäck, T. (1996). *Evolutionnary algorithms in Theory and Practice*. Oxford University Press.

Cahon, S., Melab, N., et Talbi, E.-G. (2004). Paradiseo : A framework for the reusable design of parallel and distributed metaheuristics. *Journal of Heuristics*, 10(3) : 357—380.

Calland, P.-Y. (2003). On the structural complexity of a protein. *Protein Engineering*, 16(2) : 76–86.

Canutescu, A. A., Shelenkov, A. A., et Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12 : 2001–2014.

Carugo, O. et Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Science*, 10(7) : 1470–1473.

Chavez, L. L., Onuchic, J. N., et Clementi, C. (2004). Quantifying the roughness on the free energy landscape : Entropic bottlenecks and protein folding rates. *Journal of American Chemical Society*, 126(27) : 8426–8432.

Claude, D., Clairambault, J., et Lévi, F. (2000). Rythmes biologiques et chronothérapeutique : comparaison entre des schémas d'administration théoriques et des thérapeutiques appliquées en cancérologie. *ESAIM proceedings*, 9 : 119–137.

Clore, G. M., Brunger, A. T., Karplus, M., et Gronenborn, A. M. (1986). Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. a model study of crambin. *Journal of Molecular Biology*, 191(3) : 523–551.

Cochran, A. G., Skelton, N. J., et Starovasnik, M. A. (2001). Tryptophan zippers : Stable, monomeric β-hairpins. *Proc Natl Acad Sci USA*, 98(10) : 5578–5583.

Coleman, T. F. et Wu, Z. (1996). Parallel continuation-based global optimization for molecular conformation and protein folding. *Journal of Global Optimization*, 8(1) : 49–65.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., et Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of American Chemical Society*, 117(9) : 5179–5197.

Coutsias, E. A., Seok, C., et Dill, K. A. (2004). Using quaternions to calculate rmsd. *Journal of Computational Chemistry*, 25(15) : 1849–1857.

Crescenzi, P., Goldman, D., Papadimitriou, C. H., Piccolboni, A., et Yannakakis, M. (1998). On the complexity of protein folding. *Journal of Computational Biology*, 5(3) : 423–466.

Crippen, G. M. et Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*. research studies press ltd.

Cui, G. et Simmerling, C. (2002). Conformational heterogeneity observed in simulations of a pyrene-substituted dna. *Journal of American Chemical Society*, 124(41) : 12154–12164.

Damsbo, M., Kinnear, B. S., Hartings, M. R., Ruhoff, P. T., Jarrold, M. F., et Ratner, M. A. (2004). Application of evolutionary algorithm methods to polypeptide folding : comparison with experimental results for unsolvated ac-(ala-gly-gly)$_5$-lysh+. *PNAS*, 101(19) : 7215–7222.

Dandekar, T. et Argos, P. (1997). Applying experimental data to protein fold prediction with the genetic algorithm. *Protein Eng.*, 10(8) : 877–893.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection*. Alfred Knoner Verlag, Stuttgard (German). Harvard University Press, 1995.

Davis, L. (1991). *Handbook of Genetic algorithms*. Van Nostrand Reinhold, New York.

Davy, M., Del Moral, P., et Doucet, A. (2003). méthodes monte carlo séquentielles pour l'analyse spectrale bayésienne. Dans *Proceedings of GRETSI Conference*.

Day, R. O., Zydallis, J. B., Lamont, G. B., et Pachter, R. (2002). Solving the protein structure prediction problem through a multiobjective genetic algorithm. Dans *Technical Proceedings of the 2002 International Conference on Computational Nanoscience and Nanotechnology*, volume 2, pp. 32 – 35, Air Force Institute of Technology, USA.

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems : A literature review. *Journal of Computational Biology*, 9(1) : 67–103.

De Jong, K. A. (1993). Genetic algorithms are not function optimizers. *Foundations of Genetic Algorithms*, 2.

De Jong, K. A., Potter, M. A., et Spears, W. M. (1997). Using problem generator to explore the effects of epistasis. Dans *Proceedings of The Seventh International Conference on Genetic Algorithms*, pp. 1–8, Michigan State University.

De Jong, K. A., Spears, W. M., et F., G. D. (1994). Using markov chains to analyse gafos. *Foundations of Genetic Algorithms*, 3 : 115–137.

Del Moral, P. et Doucet, A. (2002). Sequential monte carlo samplers. Rapport Technique 443, Cambridge University.

Di Ventura, B., Lemerle, C., Michalodimitrakis, K., et Serrano, L. (2006). From in vivo to in silico biology and back. *Nature*, 443 : 527–533.

Dill, K., Phillips, A., et Rosen, J. (1996). Molecular structure prediction by global optimization.

Dill, K. A. et Chan, H. S. (1997). From levinthal to pathways to funnels. *Nature Structural & Molecular Biology*, 4(1) : 10–19.

Djurdjevic, D. P. et Biggs, M. J. (2006). *Ab initio* protein fold prediction using evolutionary algorithms : Influence of design and control parameters on performance. *Journal of Computational Chemistry*, 27(11) : 1177–1195.

Dobson, C. (2003). Protein folding and misfolding. *Nature*, 426(6968) : 884–890.

Dobson, C. M., Sali, A., et Karplus, M. (1998). Protein folding : A perspective from theory and experiment. *Angewandte Chemie International Edition*, 37(7) : 868–893.

Doherty, M. K. et Beynon, R. J. (2006). Protein turnover on the scale of the proteome. *expert review of proteomics*, 3(1) : 97–110.

Dublanche, Y., Michalodimitrakis, K., Kümmerer, N., Foglierini, M., et Serrano, L. (2006). Noise in transcription negative feedback loops : simulation and experimental analysis. *Molecular Systems Biology*, 2(41) : 1–12.

El Samad, H., Khammash, M., Petzold, L., et Gillespie, D. (2005). Stochastic modeling of gene regulatory networks. *international journal of robust and nonlinear control*, 15(15) : 691–711.

Elowitz, M. B. et Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403 : 335–338.

Elston, T., Wang, H., et Oster, G. (1998). Energy transduction in atp synthase. *Nature*, 391(6666) : 510–513.

Fisher, J., Piterman, N., Hubbard, E. J. A., Stern, M. J., et Harel, D. (2005). Computational insights into caenorhabditis elegans vulval development. *PNAS*, 102(6) : 1951–1956.

Fraenkel, A. (1993). Complexity of protein folding. *Bull. Math. Biol.*, 55 : 1199.

François, P. et Hakim, V. (2004). Design of genetic networks with specified functions by evolution in silico. *PNAS*, 101(2) : 580–585.

Frauenfelder, H. et Leeson, D. T. (1998). The energy landscape in non-biological and biological molecules. *Nature structural & molecular biology*, 5 : 757–759.

Garcia, A. et Onuchic, J. (2003). Folding a protein in a computer : an atomic description of the folding/unfolding of protein a. *Journal of American Chemical Society*, 100(24) : 13898–13903.

Gardner, T. S., di Bernardo, D., Lorenz, D., et Collins, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301 : 102–105.

Gathen, J. V. Z. et Gerhard, J. (2003). *Modern Computer Algebra*. Cambridge University Press, New York, NY, USA.

Gfeller, D., Rios, P. D. L., Caflisch, A., et Rao, F. (2007). Complex network analysis of free-energy landscapes. *PNAS*, 104(6) : 1817–1822.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25) : 2340–2361.

Given, J. A. et Gilson, M. K. (1998). A hierarchical method for generating low-energy conformers of a protein-ligand complex. *Proteins : Structure, Function and Genetics*, 33(4) : 475–495.

Glover, F. (1989). Tabu search – part i. *ORSA Journal on Computing*, 1(3) : 190–206. Operations Research Society of America.

Glover, F. (1990). Tabu search – part ii. *ORSA Journal on Computing*, 2 : 4–32. Operations Research Society of America.

Glover, F. (1997). A template for scatter search and path relinking. *Lecture Notes in Computer Science*, 1363 : 13–54.

Glover, F., Kelly, J. P., et Laguna, M. (1995). Genetic algorithms and tabu search : hybrids for optimization. *Computers and Operations Research*, 22(1) : 111–134.

Goldberg, D. E. (1989). *Genetic algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

Goldbeter, A. (1991). A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *PNAS*, 88(20) : 9107–9111.

Goldbeter, A. (1995). A model for circadian oscillations in the drosophila period protein (per). *Proceedings of the Royal Society B, Biological Sciences*, 261(1362) : 319–324.

Gonze, D., Bernard, S., Waltermann, C., Kramer, A., et Herzel, H. (2005). Spontaneous synchronization of coupled circadian oscillators. *Biophysics Journal*, 89 : 120–129.

Gonze, D., Halloy, J., et Goldbeter, A. (2003). Deterministic and stochastic models for circadian rythms. *Pathologie Biologie*, 51(4) : 227–230.

Gonze, D., Halloy, J., et Goldbeter, A. (2004). Stochastic models for circadian oscillations : Emergence of a biological rhythm. *International Journal of Quantum Chemistry*, 98 : 228–238.

Good, A. C., Cho, S.-J., et Mason, J. S. (2004). Descriptors you can count on ? normalized and filtered pharmacophore descriptors for virtual screening. *Journal of Computer-Aided Molecular Design*, 18(7) : 523–527.

Goss, P. J. E. et Peccoud, J. (1998). Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. *PNAS*, 95(12) : 6750–6755.

Goto, H. et Osawa, E. (1989). Corner flapping : A simple and fast algorithm for exhaustive generation of ring conformations. *Journal of American Chemical Society*, 111 : 8950–8951.

Goto, H. et Osawa, E. (1992). Further developments in the algorithm for generating cyclic conformers. test with cycloheptadecane. *Tetrahedron Letters*, 33 : 1343–1346.

Goto, H. et Osawa, E. (1993). An efficient algorithm for searching low-energy conformers of cyclic and acyclic molecules. *Journal of Chemical Society*, 2 : 187–198.

Govindarajan, S. et Goldstein, R. A. (1998). On the thermodynamic hypothesis of protein folding. *PNAS*, 95(10) : 5545–5549.

Grassberger, P. (2004). Sequential monte carlo methods for protein folding. Dans Wolf, D., MĄünster, G., et Kremer, M., editeurs, *NIC Symposium 2004*, volume 20, pp. 1–10.

Grefenstette, J. J. (1986). Optimisation of control parameters for genetic algorithms. *IEEE Transaction on Systems, Man and Cybernetics*, 16(1) : 122–128.

Griffith, J. S. (1968a). Mathematics of cellular control processes, i. negative feedback to one gene. *Journal of Theoretical Biology*, 20(2) : 202–208.

Griffith, J. S. (1968b). Mathematics of cellular control processes, ii. positive feedback to one gene. *Journal of Theoretical Biology*, 20(2) : 209–216.

Guantes, R. et Poyatos, J. F. (2006). Dynamical principles of two-component genetic oscillators. *PLoS Comput Biol*, 2(e30) : 0188–0197.

Guvench, O. et Brooks, C. L. (2005). Tryptophan side chain electrostatic interactions determine edge-to-face vs parallel-displaced tryptophan side chain geometries in the designed beta-hairpin "trpzip2". *Journal of American Chemical Society*, 127 : 4668–4674.

Günter, R. (1992). Parallel approaches to stochastic global optimization. Dans *Proceedings of the European Workshop on Parallel Computing*, pp. 236–247. Barcelona, Spain.

Hagler, A. T., Huler, E., et Lifson, S. (1974). Energy functions for peptides and proteins. i. derivation of a consistent force field including the hydrogen bond from amide crystals. *Journal of American Chemical Society*, 96(17) : 5319–5327.

Hagler, A. T. et Lifson, S. (1974). Energy functions for peptides and proteins. ii. the amide hydrogen bond and calculation of amide crystal properties. *Journal of American Chemical Society*, 96(17) : 5327–5335.

Halgren, T. A. (1996). Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5) : 490–519.

Hanoulle, X., Melchior, A., Sibille, N., Parent, B., Denys, A., Wieruszeski, J.-M., Horvath, D., Allain, F., Lippens, G., et Landrieu, I. (2007). Structural and functional characterisation of the interaction between cyclophilin b and a heparin derived oligosaccharide. *Journal of Biological Chemistry*, 282(47) : 34148–34158.

Hansen, N. et Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies : The covariance matrix adaptation. Dans *Proceedings of the 1996 IEEE Intern. Conf. on Evolutionary Comutation (ICEC'96)*, pp. 312–317.

Hansen, N. et Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2) : 159–195.

Hart, W. E. et Belew, R. K. (1991). Optimizing an arbitrary function is hard for the genetic algorithm. Dans Belew, R. et L.B.Booker, editeurs, *Proceedings of the Fourth International Conference on the Genetic Algorithms*, pp. 190–195. L. Darrell Whitley. San Mateo CA : Morgan Kaufmann.

Hart, W. E. et Istrail, S. (1995). Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal. *Twenty-seventh Annual ACM Symp. on Theory of Computing (STOC95)*, pp. 157–168.

Hartwell, L. H., Hopfield, J. J., Leibler, S., et Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402 : C47–C52.

Herges, T. et Wenzel, W. (2004). An all-atom forcefield for teriary structure prediction of helical proteins. *Biophysical Journal*, 87 : 1–22.

Herrera, F. et Lozano, M. (2001). Adaptative genetic algorithms based on coevolution with fuzzy behaviors. *Evolutionary Computation, IEEE Transactions on*, 5(2) : 149–165.

Herrera, F. et Lozano, M. (2003). Fuzzy adaptive genetic algorithms : design, taxonomy, and future directions. *Soft Computing*, 7(8) : 545–562.

Herrera, F., Lozano, M., et Sánchez, A. M. (2003). A taxonomy for the crossover operator for real-coded genetic algorithms : An experimental study. *International Journal of Intelligent Systems*, 18 : 309–338.

Hirata, H., Yoshiura, S., Ohtsuka, T., Bessho, Y., Harada, T., Yoshikawa, K., et Kageyama, R. (2002). Oscillatory expression of the bhlh factor hes1 regulated by a negative feedback loop. *Science*, 298(5594) : 840–843.

Hobza, P., KabelĄác, M., Sponer, J., MejzlĄík, P., et VondrĄácek, J. (1998). Performance of empirical potentials (amber, cff95, cvff, charmm, opls, poltev), se-

miempirical quantum chemical methods (am1, mndo/m, pm3), and ab initio hartree-fock method for interaction of dna bases : Comparison with nonempirical beyond hartree-fock results. *Journal of Computational Chemistry*, 18(9) : 1136–1150.

Hoffmann, D. et Knapp, E. W. (1996). Polypeptide folding with off-lattice monte carlo dynamics : the method. *Eur. Biophysics J.*, 24(6) : 387–404.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, University of Michigan Press.

Honeycutt, J. D. et Thirumalai, D. (1990). Metastability of the folded states of globular proteins. *Proc Natl Acad Sci USA*, 87(9) : 3526–3529.

Honig, B. et Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268(5214) : 1144–1149.

Horn, B. K. P. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4) : 629–642.

Hornak, V. et Simmerling, C. (2003). Generation of accurate protein loop conformations through low-barrier molecular dynamics. *Proteins*, 51(4) : 577–590.

Hornak, V. et Simmerling, C. (2007). Targeting structural flexibility in hiv-1 protease inhibitor binding. *Drug Discovery Today*, 12(3–4) : 132–138.

Horvath, D. (1997). A virtual screening approach applied to the search for trypanothione reductase inhibitors. *Journal of Medicinal Chemistry*, 40(15) : 2412–2423.

Horvath, D. et Jeandenans, C. (2003). Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces-a novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *Journal of Chemical Information and Computer Science*, 43 : 680–690.

Huang, E. S., Subbiah, S., et Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *Journal of Molecular Biology*, 252(5) : 709–720.

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., et O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959) : 686–691.

Iftimie, R., Minary, P., et Tuckerman, M. E. (2005). Chemical theory and computation special feature : Ab initio molecular dynamics : Concepts, recent developments, and future trends. *PNAS*, 102(19) : 6654–6659.

Irwin, J. J. et Shoichet, B. K. (2005). Zinc : A free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1) : 177–182.

Ishwaran, H. (1999). Applications of hybrid monte carlo to bayesian generalized linear models : Quasicomplete separation and neural networks. *Journal of Computational and Graphical Statistics*, 8(4) : 779.

Iwasaki, H., Nishiwaki, T., Kitayama, Y., Nakajima, M., et Kondo, T. (2002). Kaia-stimulated kaic phosphorylation in circadian timing loops in cyanobacteria. *PNAS*, 99(24) : 15788–15793.

Jin, A. Y., Leung, F. Y., et Weaver, D. F. (1999). Three variations of genetic algorithm for searching biomolecular conformation space : Comparison of gap 1.0, 2.0, and 3.0. *Journal of Computational Chemistry*, 20(13) : 1329–1342.

Jin, L. et Harrison, S. (2002). Crystal structure of human calcineurin complexed with cyclosporin a and human cyclophilin. *PNAS*, 99(21) : 13522–13526.

Jorgensen, W. L. et Tirado-Rives, J. (2005). Chemical theory and computation special feature : Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *PNAS*, 102(19) : 6665–6670.

Jäger, M., Zhang, Y., Bieschke, J., Nguyen, H., Dendle, M., Bowman, M. E., Noel, J. P., Gruebele, M., et Kelly, J. W. (2006). Structure—function—folding relationship in a ww domain. *PNAS*, 103(28) : 10648–10653.

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5) : 922–923.

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 34(5) : 827–828.

Kamiya, N. et Higo, J. (2001). Repeated-annealing sampling combined with multicanonical algorithm for conformational sampling of bio-molecules. *Journal of Computational Chemistry*, 22(10) : 1098–1106.

Karplus, M. et Kuriyan, J. (2005). Chemical theory and computation special feature : Molecular dynamics and protein function. *PNAS*, 102(19) : 6679–6685.

Karplus, M. et Shakhnovich, E. (1992). *Protein Folding*, chapter Protein Folding : Theoretical Studies of Thermodynamics and Dynamics. W.H. Freeman, New York.

Kennedy, J. et Spears, W. M. (1998). Matching algorithms to problems : An experimental test of the particle swarm and some genetic algorithms on the multimo-

dal problem generator. Dans *Proceedings of the IEEE International Conference on Evolutionary Computation, Anchorage, Alaska, USA*.

Kerszberg, M. (2004). Noise, delays, robustness, canalization and all that. *Current Opinion in Genetics & Development*, 14(4) : 440–445.

Khimasia, M. M. et Coveney, P. V. (1997). Protein structure prediction as a hard optimization problem : the genetic algorithm approach. *Physics*, pp. 1–12.

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., et Tomita, M. (2003). Dynamic modeling of genetic networks using genetic algorithm and s-system. *Bioinformatics*, 19(5) : 643–650.

Kim, J. G., Fukunishi, Y., et Nakamura, H. (2004). Multicanonical molecular dynamics algorithm employing an adaptive force-biased iteration scheme. *Physical Review E*, 70(057103) : 1–4.

Kim, S., Weinstein, J. N., et Grefenstette, J. J. (2003). Inference of large-scale topology of gene regulation networks by neural nets. Dans *IEEE International Conference on Systems, Man & Cybernetics*, pp. 3969–3975.

Kirkpatrick, S., Gelatt, C., et Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220(4598) : 671–680.

Klepeis, J. L. et Floudas, C. A. (2001). *Advances in Convex Analysis and Global Optimization*, chapter Deterministic global optimization for protein structure prediction, pp. 31–74. Kluwer Academic Publishers.

Klepeis, J. L., Ierapetritou, M. G., et Floudas, C. A. (1998). Protein folding and peptide docking : A molecular modeling and global optimization approach. *Computers and Chemical Engineering*, 22 : S3–S10.

Kneller, G. R. (2005). Comment on "using quaternions to calculate rmsd" [j. comp. chem. 25, 1849 (2004)]. *Journal of Computational Chemistry*, 26(15) : 1660–1662.

Koehl, P. et Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Current Opinion in Structural Biology*, 6(2) : 222–226.

Kolossvary, I. et Guida, W. C. (1996). Low mode search. an efficient, automated computational method for conformational analysis : Application to cyclic and acyclic alkanes and cyclic peptides. *Journal of American Chemical Society*, 118(21) : 5011–5019.

König, R. et Dandekar, T. (1999). Improving genetic algorithms for protein folding simulations by systematic crossover. *BioSystems*, 50(1) : 17–25.

Koretke, K. K., Luthey-Schulten, Z., et Wolynes, P. G. (1998). Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *PNAS*, 95(6) : 2932–2937.

Koschützki, D. et Schreiber, F. (2004). Comparison of centralities for biological networks. Dans *Proceedings of German Conference on Bioinformatics (GCB'04)*, volume 53, pp. 199–206.

Kosinsky, Y. A., Volynsky, P. E., Lagant, P., Vergoten, G., Suzuki, E.-I., Arseniev, A. S., et Efremov, R. G. (2004). Development of the force field parameters for phosphoimidazole and phosphohistidine. *Journal of Computational Chemistry*, 25(11) : 1313–1321.

Krivov, S. V. et Karplus, M. (2004). Hidden complexity of free energy surfaces for peptide (protein) folding. *PNAS*, 101(41) : 14766–14770.

Kruse, K. et Jülicher, F. (2005). Oscillations in cell biology. *Current Opinion in Cell Biology*, 17(20) : 20–26.

Kubelka, J., Hofrichter, J., et Eaton, W. A. (2004). The protein folding 'speed limit'. *Current Opinion in Structural Biology*, 14 : 76–88.

Kubota, N. et Fukuda, T. (1997). Genetic algorithms with age structure. *Soft Computing*, 1 : 155–161.

Kunz, H. et Achermann, P. (2003). Simulation of circadian rhythm generation in the suprachiasmatic nucleus with locally coupled self-sustained oscillators. *Journal of Theoretical Biology*, 224(1) : 63–78.

Kutzner, C., Spoel, D. V. D., Fechner, M., Lindahl, E., Schmitt, U. W., Groot, B. L. D., et Grubmüller, H. (2007). Speeding up parallel gromacs on high-latency networks. *Journal of Computational Chemistry*, 28(12) : 2075 – 2084.

Lathrop, R. (1994). The protein threading problem with sequence amino acid interaction preferences is np-complete. *Protein Engineering*, 7(9) : 1059–1068.

Lattner, A. D., Kim, S., Cervone, G., et Grefenstette, J. J. (2003). Experimental comparison of symbolic learning programs. *In : FGML 2003 Workshop, Annual Meeting of the GI Working Group "Machine Learning, Knowledge Discovery, Data Mining" (FGML) : 2003 ; Karlsruhe, Germany ; 2003.*

Lauria, A., Diana, P., Barraja, P., Montalbano, A., Dattolo, G., Cirrincione, G., et Almerico, A. M. (2004). Docking of indolo- and pyrrolo-pyrimidines to dna. new dnainteractive polycycles from amino-indoles/pyrroles and bmma. *Arkivoc*, 5 : 263–271.

Lavelle, C. et Benecke, A. (2006). Chromatin physics : Replacing multiple, representation-centered descriptions at discrete scales by a continuous, function-dependent self-scaled model. *European Physical Journal E*, 19 : 379–384.

Lazaridis, T. et Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins : Structure, Function, and Genetics*, 35(2) : 133–152.

Le Novere, N. et Shimizu, T. S. (2001). Stochsim : modelling of stochastic biomolecular processes. *Bioinformatics*, 17(6) : 575–576.

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry : a review. *Journal of Chemometrics*, 15(7) : 559–569.

Leenheer, P. D., Angeli, D., et Sontag, E. D. (2004). A tutorial on monotone systems with an application to chemical reaction networks. Dans *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2004)*.

Lefranc, M., Bielwsky, S., Bouget, F.-Y., Boulier, F., Lemaire, F., Picault, S., Petitot, M., Horvath, D., Thommen, Q., Morant, P.-E., Vandermoere, C., et Parent, B. (2006). Studying, modeling & simulating circadian oscillations in regulatory networks. « Gent-Lille Workshop on Computational Biology ».

Leloup, J.-C. et Goldbeter, A. (1999). Chaos and birhytmicity in a model for circadian oscillations of the per and tim proteins in drosophila. *Journal of Theoretical Biology*, 198 : 445–459.

Lema, M. A., Golombek, D. A., et Echave, J. (2000). Delay model of the circadian pacemaker. *Journal of theoretical Biology*, 204 : 565–573.

Levinthal, C. (1969). How to fold graciously. Dans *Conference on Mossbauer Spectroscopy in Biological Systems*, pp. 22–24, University of Illinois Press. Proceedings of a meeting held at Allerton House, Monticello, Illinois.

Lewis, J. (2003). Autoinhibition with transcriptional delay : A simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, 13(16) : 1398–1408.

Li, G. et Widom, J. (2004). Nucleosomes facilitate their own invasion. *Nature*, 11(8) : 763–769.

Lin, G., Yao, X., Macleod, I., Kang, L., et Chen, Y. (1996). Parallel genetic algorithm on pvm. Dans *Proceedings of the International Conference on Parallel Algorithms (ICPA'95)*.

Lipshtat, A., Loinger, A., Balaban, N. Q., et Biham, O. (2006). Genetic toggle switch without cooperative binding. *Physical Review Letters*, 96(18) : 1–4.

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., et Scheraga, H. A. (1999). Protein structure prediction by global optimization of a potential energy function. *PNAS*, 96(10) : 5482–5485.

Locke, J. C. W., Millar, A. J., et Turnera, M. S. (2005). Modelling genetic networks with noisy and varied experimental data : the circadian clock in arabidopsis thaliana. *Journal of Theoretical Biology*, 234 : 383–393.

Lok, L. et Brent, R. (2005). Automatic generation of cellular reaction networks with moleculizer 1.0. *Computational Biology*, 23(1) : 131–136.

Loncharich, R. J. et Brooks, B. R. (1989). The effects of truncating long-range forces on protein dynamics. *Proteins : Structure, Function, and Genetics*, 6(1) : 32–45.

MacKerell, A., Bashford, D., Bellott, M., Dunbrack, R., Evanseck, J., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., et Karplus, M. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102(18) : 3586–3616.

Mackerell, A. D. (2004). Empirical force fields for biological macromolecules : Overview and issues. *Journal of Computational Chemistry*, 25(13) : 1584–1604.

Mailleret, L. (2004). *Stabilisation Globale de Systèmes Dynamiques Positifs Mal Connus. Applications en Biologie.* Thèse de Doctorat, Université de Nice Sophia-Antipolis.

Mangan, S., Zaslaver, A., et Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, 334 : 197–204.

Maple, J. R., Hwang, M.-J., Stockfish, T. P., Dinur, U., Waldman, M., Ewig, C., et Hagler, A. (1994). Derivation of class ii force fields. i. methodology and quantum force field for the alkyl functional group and alkane molecules. *Journal of Computational Chemistry*, 15(2) : 161–182.

Maslov, S. et Sneppen, K. (2002). Specificity and stability in topology of protein network. *Science*, 296 : 910–913.

Mathews, D. H. et Turner, D. H. (2006). Prediction of rna secondary structure by free energy minimization. *Current opinion in Structural Biology*, 16(3) : 270–278.

McAdams, H. H. et Arkin, A. (1997). Stochastic mechanisms in gene expression. *PNAS*, 94 : 814–819.

McLachlan, A. (1982). Rapid comparison of protein structures. *Acta Cristallography*, A38 : 871–873.

Michalewicz, Z. (1994). *Genetic Algorithms + Data structures = Evolution Programs*. Springer-Verlag, Berlin, second edition.

Millar, J. et Kollman, P. (1997). Theoretical studies of an exceptionally stable rna tetraloop : Observation of convergence from an incorrect nmr structure to the correct one using unrestrained molecular dynamics. *Journal of Molecular Biology*, 270(3) : 436–450.

Miller, W. H. (2005). Chemical theory and computation special feature : Quantum dynamics of complex molecular systems. *PNAS*, 102(19) : 6660–6664.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., et Alon, U. (2002). Network motifs : Simple building blocks of complex networks. *Science*, 298 : 824–827.

Mok, K. H., Kuhn, L. T., Goez, M., Day, I. J., Lin, J. C., Andersen, N. H., et Hore, P. J. (2007). A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. *Nature*, 447 : 106–109.

Momany, F., McGuire, R., Burgess, A., et Scheraga, H. (1975). Energy parameters in polypeptides. vii. geometric parameters, partial atomic charges, nonbounded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *Journal of Physical Chemistry*, 79 : 2361–2381.

Monk, N. A. M. (2003). Oscillatory expression of hes1, p53, and nf-κb driven by transcriptional time delays. *Current Biology*, 13(16) : 1409–Ąŋ1413.

Morant, P.-E., Vandermoere, C., Thommen, Q., Parent, B., Lemaire, F., Corellou, F., Schwartz, C., Bouget, F.-Y., et Lefranc, M. (2007). Oscillateurs génétiques simples. application à l'horloge circadienne d'une algue unicellulaire. Dans Lefranc, M., Letellier, C., et Pasteur, L., editeurs, *Compte-rendus de la 10^e Rencontre du Non-Linéaire*, volume 1, pp. 131–136, Paris. Institut Henri Poincaré, Non-linéaire publications. Orsay (Université de Paris-Sud, 91405).

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., et Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14) : 1639–1662.

Mu, Y., Nordenskiöld, L., et Tam, J. P. (2006). Folding, misfolding, and amyloid protofibril formation of ww domain fbp28. *Biophysical Journal*, 90 : 3983–3992.

Muñoz, V., Thompson, P. A., Hofrichter, J., et Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature*, 390(6656) : 196–199.

Méndez, R., Leplae, R., Maria, L. D., et Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions : Current status of docking methods. *Proteins : Structure, Function, and Genetics*, 52(1) : 51–67.

Mühlenbein, H. (1992). Asynchronous parallel search by the parallel genetic algorithm. *Third IEEE Symposium on Parallel and Distributed Products*, pp. 526–533.

Naef, F. (2005). Circadian clock go in vitro : purely post-translational oscillators in cyanobacteria. *Molecular System Biology*, 1(1) : E1–E5.

Nagaich, A. K., Walker, D. A., Wolford, R., et Hager, G. L. (2004). Rapid periodic binding and displacement of the glucocorticoid receptor during chromatin remodeling. *Molecular Cell*, 14 : 163–174.

Nagoshi, E., Saini, C., Bauer, C., Laroche, T., Naef, F., et Schibler, U. (2004). Circadian gene expression in individual fibroblasts : Cell-autonomous and self-sustained oscillators pass time to daughter cells. *Cell*, 119 : 693–705.

Najmanovich, R., Kuttner, J., Sobolev, V., et Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins : Structure, Function, and Genetics*, 39(3) : 261–268.

Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., et Kondo, T. (2005). Reconstruction of circadian oscillation of cyanobacterial kaic phosphorylation in vitro. *Science*, 308(5720) : 414–415.

Nayeem, A., Vila, J., et Scheraga, H. A. (1991). A comparative study of the simulated-annealing and monte carlo-with-minimization approaches to the minimum-energy structures of polypeptides : [met]-enkephalin. *Journal of Computational Chemistry*, 12(5) : 594–605.

Neidigh, J. W., Fesinmeyer, R. M., et Andersen, N. H. (2002). Designing a 20-residue protein. *Nature Structural Biology*, 9 : 430–452.

Neumaier, A. (1997). Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Revue*, 39(3) : 407–460.

Neumaier, A. (2004). *Acta Numerica 2004*, chapter Complete Search in Continuous

Global Optimization and Constraint Satisfaction, pp. 271–369. A. Iserles. Cambridge University Press.

Ngo, J. T. et Marks, J. (1992). Computational complexity of a problem in molecular structure prediction. *Prot. Eng.*, 5 : 313.

Nguyen, H., M, M. J., Kelly, J., et Gruebele, M. (2005). Engineering a beta-sheet protein toward the folding speed limit. *The Journal of Physical Chemistry B Condens Matter Mater Surf Interfaces Biophys.*, 109(32) : 15182–15186.

Nikitopoulos, T. G. et Emiris, I. Z. (2001). Molecular conformation search by matrix perturbations.

Nix, A. E. et Vose, M. D. (1992). Modeling genetic algorithms with markov chains. *Annals of Mathematics and Artificial Intelligence*, 5 : 79–88.

Novak, B. et Pataki, Z. (2000). Mathematical model of the cell division cycle of fission yeast. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 11(1) : 277–286.

N.Przulj, Wigle, D., et Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics*, 20(3) : 340–348.

Nùnez-Letamendia, L. (2003). Fitting the control parameters of a genetic algorithm to optimise technical trading rules. http ://207.36.165.114/Denver/Papers/FMA_2003_LETAMENDIA.pdf.

Ochoa, G., Harvey, I., et Buxton, H. (1999). On recombination and optimal mutation rates. Dans *Proceedings of Genetic and Evolutionary Computation Conference (GECCO'99)*.

Okur, A., Strockbine, B., Hornak, V., et Simmerling, C. (2003). Using pc clusters to evaluate the transferability of molecular mechanics force fields for proteins. *Journal of Computational Chemistry*, 24(1) : 21–31.

Onuchic, J. N., Socci, N. D., et Zaida Luthey-Schulten, P. G. W. (1996). Protein folding funnels : the nature of the transition state ensemble. *Folding and Design*, 1(6) : 441–450.

Oprea, T. I. (2005). *Chemoinformatics in Drug Discovery*, volume 23. Wiley-VCH, Weinheim, 1 edition.

Paci, E., Vendruscolo, M., et Karplus, M. (2002). Native and non-native interactions along protein folding and unfolding pathways. *Proteins : Structure, Function, and Genetics*, 47(3) : 379–392.

Packer, M. J. et Hunter, C. A. (2001). Sequence-structure relationships in dna oligomers : A computational approach. *Journal of American Chemical Society*, 123(30) : 7399–7406.

Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J., et Zagrovic, B. (2003). Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68(1) : 91–109.

Parent, B., Kökösy, A., et Horvath, D. (2007a). Optimized evolutionnary strategies in conformational sampling. *Journal of Soft Computing*, 11(1) : 63–79.

Parent, B., Lippens, G., et Horvath, D. (2006). Steps towards an ensemble-based force field fitting procedure. Computational Chemistry Gordon Research Conference.

Parent, B., Tantar, A., Melab, N., Talbi, E.-G., et Horvath, D. (2007b). Grid-based evolutionary strategies applied to the conformational sampling problem. Congress on Evolutionary Computation.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2(2) : 157–175.

Pillardy, J., Czaplewski, C., Liwo, A., Lee, J., Ripoll, D. R., Kazmierkiewicz, R., Oldziej, S., Wedemeyer, W. J., Gibson, K. D., Arnautova, Y. A., Saunders, J., Ye, Y.-J., et Scheraga, H. A. (2001). Recent improvements in prediction of protein structure by global optimization of a potential energy function. *PNAS*, 98(5) : 2329–2333.

Pratt, J. M., Petty, J., Riba-Garcia, I., Robertson, D. H. L., Gaskell, S. J., Oliver, S. G., et Beynon, R. J. (2002). Dynamics of protein turnover, a missing dimension in proteomics. *Mol Cell Proteomics*, 1(8) : 579–591.

Prebys, E. K. (1999). The genetic algorithm in computer science. *MIT Undergraduate Journal of Mathematics*, 1 : 165–170.

Ramachandran, G. et Sasisekhan, V. (1968). Conformation of polypeptides and proteins. *Advan. Prot. Chem.*, 23 : 283–438.

R.Blossey, L.Cardelli, et Phillips, A. (2006). Compositionality, stochasticity and cooperativity in dynamic models of gene regulation. *Quantitative Biology*, pp. 1–5.

Regev, A. (2002). *Computational Systems Biology : A Calculus for Biomolecular knowledge*. Thèse de Doctorat, Tel Aviv University.

Renders, J.-M. (1995). *Algorithmes génétiques et Réseaux de neurones*. Hermès, Paris.

Reppert, S. M. et Weaver, D. R. (2002). Coordination of circadian timing in mammals. *Nature*, 418 : 935–941.

Richard, J.-P. (2002). *Mathematiques pour les Systemes Dynamiques*. Hermes Science Publications, hermès science publications edition.

Richard, J.-P. (2003). Time-delay systems : an overview of some recent advances and open problems. *Automatica*, 39(10) : 1667–1694.

Roenneberg, T. et Merrow, M. (2002). Life before the clock : Modeling circadian evolution. *Journal of Biological Rhythms*, 17(6) : 495–505.

Roitberg, A. E., Okur, A., et Simmerling, C. (2007). Coupling of replica exchange simulations to a non-boltzmann structure reservoir. *Journal of Physical Chemistry*, 111(10) : 2415–2418.

Ross, T. J. (2004). *Fuzzy Logic With Engineering Applications*. John Wiley & Sons Inc, 2 edition.

Ruscio, J. et Onufriev, A. (2006). A computational study of nucleosomal dna flexibility. *Biophysical Journal*, 91(11) : 4121–4132.

Rylance, G. J., Johnston, R. L., Matsunaga, Y., Li, C.-B., Baba, A., et Komatsuzaki, T. (2006). Topographical complexity of multidimensional energy landscapes. *PNAS*, 103(49) : 18551–18555.

Sako, Y. (2006). Imaging single molecules in living cells for systems biology. *Molecular Systems Biology*, 2(56) : 1–6.

Sali, A., Glaeser, R., Earnest, T., et Baumeister, W. (2003). From words to literature in structural proteomics. *Nature*, 422 : 216–225.

Sasai, M. et Wolynes, P. G. (2003). Stochastic gene expression as a many-body problem. *PNAS*, 100(5) : 2374–2379.

Sawai, H. et Adachi, S. (2002). A comparative study of gene-duplicated gas based on pfga and ssga. Dans *Proceedings of GECCO-2000*, volume 1, pp. 74–81, Las Vegas.

Schug, A., Herges, T., Verma, A., Lee, K. H., et Wenzel, W. (2005a). Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. *ChemPhysChem*, 6(12) : 2640 – 2646.

Schug, A., Herges, T., et Wenzel, W. (2004). All-atom folding of the trp-cage protein with an adpative parallel tempering method. *European Physical Letter*, 67 : 307–313.

Schug, A. et Wenzel, W. (2004). Predictive in silico all-atom folding of a four-helix protein with a free-energy model. *Journal of American Chemical Society*, 126 : 16736–16737.

Schug, A., Wenzel, W., et Hansmann, U. H. E. (2005b). Energy landscape paving simulations of the trp-cage protein. *Journal of Chemical Physics*, 122(194711) : 1–7.

Schulze-Kremer, S. (1995). Biocomputing for everyone ! pages web.

Schulze-Kremer, S. et Tiedemann, U. (1994). Parameterizing genetic algorithms for protein folding simulation. Dans *HICSS (5)*, pp. 345–354.

Scitegic (2005). Scitegic pipeline pilot version 3.0. disponible depuis Scitegic, Inc à http ://www.scitegic.com.

Shen-Orr, S. S., Milo, R., Mangan, S., et Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature*, 31 : 64–68.

Shetty, R. P., de Bakker, P. I., DePristo, M. A., et Blundell, T. L. (2003). Advantages of fine-grained side chain conformer libraries. *Protein Engineering design & selection*, 16(12) : 963–969.

Shmygelska, A. et Hoos, H. (2005). An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem. *BMC Bioinformatics*, 6(1) : 30.

Shmygelska, A. et Hoos, H. H. (2003). An improved ant colony optimisation algorithm for the 2d hp protein folding problem.

Shoemaker, B. A., Wang, J., et Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals : the transition state ensemble. *Journal of Molecular Biology*, 287(3) : 675–694.

Skhiri, S. (2004). Interrogation des bases de données biochimiques : Conception d'un visualisateur de voies métaboliques et de transduction de signal. Mémoire de diplôme d'etudes approfondies en informatique, Université Libre de Bruxelles, Brussels, Belgium.

Snow, C., Sorin, E., Rhee, Y., et Pande, V. (2005). How well can simulation predict protein folding kinetics and thermodynamics ? *Biophysics Program*, 34 : 43–69.

Snow, C. D., Qiu, L., Du, D., Gai, F., Hagen, S. J., et Pande, V. S. (2004). Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy. *Proc Natl Acad Sci USA*, 101(12) : 4077–4082.

Sokhansanj, B. A., Fitch, J. P., Quong, J. N., et Quong, A. A. (2004). Linear fuzzy gene network models obtained from microarray data by exhaustive search. *BMC Bioinformatics*, 5(108) : 1–12.

Sommer, I., Rahnenführer, J., Domingues, F., de Lichtenberg, U., et Lengauer, T. (2004). Predicting protein structure classes from function predictions. *Bioinformatics*, 20(5) : 770–776.

Spears, W. M. (1992). Adapting crossover in a genetic algorithm. Rapport Technique AIC-92-025, Navy Center for Applied Research in AI.

Spears, W. M. (1994). Simple subpopulation schemes. Dans *Evolutionary Programming Society, Proceedings of the Third Annual Conference on Evolutionary Programming*, pp. 196–307. San Diego, CA.

Spears, W. M. et De Jong, K. A. (1996). Analysing gas using markov models with semantically ordered and lumped states. *Foundations of Genetic Algorithms*, 4 : 95–100.

Steipe, B. (2002). A revised proof of the metric properties of optimally superimposed vector sets. *Acta Crystallographica Section A*, 58(5) : 506.

Still, W., Tempczyk, A. C., Ronald, C. H., et Hendrickson, T. (1990). Semi-analytical treatment of solvation for molecular mechanics and dynamics. *JACS*, 112 : 6127–6129.

Strizhev, A., Abrahamian, E. J., Choi, S., Leonard, J. M., Wolohan, P. R. N., et Clark, R. D. (2006). The effects of biasing torsional mutations in a conformational ga. *Journal of Chemical Informatic Models*, 46(4) : 1862–1870.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410 : 268–276.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1) : 1–4.

Sun, J., Zhang, Q., et Schlick, T. (2005). electrostatic mechanism of nucleosomal array folding revealed by computer simulation. *PNAS*, 102(23) : 8180–8185.

Takahashi, K., Ishikawa, N., Sadamoto, Y., Sasamoto, H., Ohta, S., Shiozawa, A., Miyoshi, F., Naito, Y., Nakayama, Y., et Tomita, M. (2003). E-cell 2 : Multi-platform e-cell simulation system. *Bioinformatics*, 19(13) : 1727 – 1729.

Takahashi, K., Kaizu, K., Hu, B., et Tomita, M. (2004). A multi-algorithm, multi-timescale method for cell simulation. *Bioinformatics*, 20(4) : 538–546.

Takahashi, K., Yugi, K., Hashimoto, K., Yamada, Y., Pickett, C. J. F., et Tomita, M. (2002). Computational challenges in cell simulation : A software engineering approach. *IEEE Intelligent Systems in Biology*, 17(5) : 64–71.

Takahashi, O., Kita, H., et Kobayashi, S. (1999). Protein folding by a hierarchical genetic algorithm. Dans *Proceedings of the Fourth International Symposium on Artificial Life and Robotics (AROB 4th'99)*, pp. 334–339.

Taketomi, H., Ueda, Y., et Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represented by specific inter unit interactions. *International Journal of Peptide and Protein Research*, 7(6) : 445–459.

Tantar, A.-A., Melab, N., Talbi, E.-G., Parent, B., et Horvath, D. (2007). A parallel hybrid genetic algorithm for protein structure prediction on the computational grid. *Future Generation Computer Systems*, 23(3) : 398–409.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., et Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics (Letters)*, 22 : 281–285.

Teghem, J. (2003). *Résolution de problèmes de RO par les métaheuristiques*. Hermès Sciences/Lavoisier, Paris.

Thieffry, D. et De Jong, H. (2002). Modélisation, analyse et simulation des réseaux génétiques. *Médecine/sciences*, 18 : 492–502.

Thomsen, R. (2003). Flexible ligand docking using evolutionary algorithms : investigating the effects of variation operators and local search hybrids. *Biosystems*, 72(1) : 57–73.

Tolic-Norrelykke, S. F., Engh, A. M., Landick, R., et Gelles, J. (2004). Diversity in the rates of transcript elongation by single rna polymerase molecules. *Journal of Biological Chemistry*, 279(5) : 3292–3299.

Tsui, V. et Case, D. (2000). Molecular dynamics simulations of nucleic acids with a generalized born solvation model. *JACS*, 122(11) : 2489–2498.

Tupper, P. F. (2005). Ergodicity and the numerical simulation of hamiltonian systems. *SIAM Journal on Applied Dynamical Systems*, 4(3) : 563–587.

Tyson, J., Hong, C., Thron, C., et Novak, B. (1999). A simple model of circadian

rhythms based on dimerization and proteolysis of per and tim. *Biophysical Journal*, 77(5) : 2411–2417.

Ultsch, A. (2003). Pareto density estimation : Probability density estimation for knowledge discovery. *Innovations in Classification, Data Science, and Information Systems*, pp. 91–102.

Unger, R. et Moult, J. (1993a). Finding the lowest free energy conformation of a protein is an np-hard problem : Proof and implications. *Bulletin of Mathematical Biology*, 55(6) : 1183–1198.

Unger, R. et Moult, J. (1993b). Genetic algorithms for protein folding simulations. *Journal of Molecular Biology*, 231(1) : 75–81.

Vaillant, C., Audit, B., et Arnéodo, A. (2005). Thermodynamics of dna loops with long-range correlated structural disorder. *Physical Review Letters*, 95(6).

Vainio, M. J. et Johnson, M. S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of Chemical Informatic Models*, pp. A–M.

Van-Gunsteren, W. F. et Berendsen, H. J. C. (1977). Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics*, 34(5) : 1311–1327.

VandeVen, F. J. M. (1995). *Multidimensional NMR in Liquid*. VCH Publishers.

Varma, C. K. (2001). Molecular mechanical force fields. *Biochemistry*, 218 : 1–11.

Vengadesan, K. et Gautham, N. (2003). Enhanced sampling of the molecular potential energy surface using mutually orthogonal latin squares : Application to peptide structures. *Biophysical Journal*, 84(5) : 2897–2906.

Venkatachalam, C. M., Jiang, X., Oldfield, T., et Waldman, M. (2003). Ligandfit : a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling*, 21(4) : 289–307.

Vertanen, K. (1998). Genetic adventures in parallel : Towards a good island model under pvm. *Oregon State University*.

Vieth, M., Hirst, J. D., Dominy, B. N., Daigler, H., et III, C. L. B. (1998a). Assessing search strategies for flexible docking. *Journal of Computational Chemistry*, 19(14) : 1623–1631.

Vieth, M., Hirst, J. D., Kolinski, A., et III, C. L. B. (1998b). Assessing energy functions for flexible docking. *Journal of Computational Chemistry*, 19(14) : 1612–1622.

Vilar, J. M. G., Kueh, H. Y., Barkai, N., et Leibler, S. (2002). Mechanisms of noise-resistance in genetic oscillators. *PNAS*, 99(9) : 5988–5992.

Vinga, S. et Almeida, J. (2003). Alignment-free sequence comparison : a review. *Bioinformatics*, 19(4) : 513–523.

Volpon, L. et Lancelin, J. M. (2000). solution nmr structures of the polyene macrolide antibiotic filipin iii. *FEBS Letter*, 478 : 137–140.

Vose, M. D. (1999). Random heuristic search. *Theoretical Computer Science*, 229(1, 2) : 103–142.

Wang, R., Lu, Y., et Wang, S. (2003). Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry*, 46(12) : 2287–2303.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., et Brown, P. O. (2002). Precision and functional specificity in mrna decay. *PNAS*, 99(9) : 5860–5865.

Watson, J. D. et Crick, F. H. C. (1953). Molecular structure of nucleic acids : A structure for deoxyribose nucleic acid. *Nature*, 171(4356) : 737.

Watts, D. J. et Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393 : 440–442.

Wehrens, R., Pretsch, E., et Buydens, L. M. C. (1998). Quality criteria of genetic algorithms for structure optimization. *Journal of Chemical Information and Computer Sciences*, 38(2) : 151–157.

Wenzel, W. (2006). Predictive folding of a β-hairpin protein in an all-atom free-energy model. *Europhysics Letters*, 76 : 156–162.

Westhead, D. R., Clark, D. E., et Murray, C. W. (1997). A comparison of heuristic search algorithms for molecular docking. *Journal of Computer-Aided Molecular Design*, 11(3) : 209–228.

Whitley, D., Rana, S., et Heckendorn, R. B. (1999). The island model genetic algorithm : On separability, population size and convergence. *Journal of Computing and Information Technology*, 7(1) : 33–47.

Williams, D. J. et Hall, K. B. (1999). Unrestrained stochastic dynamics simulations of the uucg tetraloop using an implicit solvation model. *Biophysical Journal*, 76(6) : 3192–3205.

Wuchty, S. et Stadler, P. F. (2003). Centers of complex networks. *Journal of Theoretical Biology*, 223(1) : 45–53.

Xu, Y., Toh, K., Jones, C., Shin, J.-Y., Fu, Y.-H., et Ptáček, L. (2007). Modeling of a human circadian mutation yields insights into clock regulation by per2. *Cell*, 128(1) : 59–70.

Yang, A.-S. et Honig, B. (1995a). Free energy determinants of secondary structure formation : I. α-helices. *Journal of Molecular Biology*, 252(3) : 351–365.

Yang, A.-S. et Honig, B. (1995b). Free energy determinants of secondary structure formation : Ii. antiparallel β-sheets. *Journal of Molecular Biology*, 252(3) : 366–376.

Yang, S., Onuchic, J., et Levine, H. (2006). Effective stochastic dynamics on a protein folding energy landscape. *Journal of Chemical Physics*, 125(5) : 054910.

Yang, W. Y., Pitera, J. W., Swope, W. C., et Gruebele, M. (2004). Heterogeneous folding of the trpzip hairpin : full atom simulation and experiment. *Journal of Molecular Biology*, 336(1) : 241–251.

Yugi, K. et Tomita, M. (2004). A general computational model of mitochondrial metabolism in a whole organelle scale. *Bioinformatics*, 20(11) : 1795–1796.

Zhou, R. (2003). Free energy landscape of protein folding in water : Explicit vs. implicit solvent. *Proteins : Structure, Function, and Genetics*, 53(2) : 148 – 161.

Zhou, R. et Berne, B. J. (2002). Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water ? *PNAS*, 99(20) : 12777–12782.

Zhou, Y. et Karplus, M. (1999). Interpreting the folding kinetics of helical proteins. *Nature*, 401 : 400–403.

Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., et da Fonseca, V. (2003). Performance assessment of multiobjective optimizers : an analysis and review. *Evolutionary Computation, IEEE Transactions on*, 7(2) : 117–132.

Zwanzig, R., Szabo, A., et Bagchi, B. (1992). Levinthal's paradox. *PNAS*, 89(1) : 20–22.

Résumé

La complexité du vivant est omniprésente à toutes les échelles : des interactions entre molécules individuelles aux réseaux d'interactions permettant à la cellule d'assurer ses fonctions vitales et de répondre aux stimuli. Cette thèse se veut être une application des outils de l'Automatique et de l'Informatique à certaines questions de la Biologie et Biochimie.

Pour cela, nous avons abordé le problème *via* deux aspects : le premier concerne la modélisation des interactions moléculaires en vue de prédire les modes de fixation et les affinités entre molécules. Puisque ces estimations nécessitent de considérer la flexibilité des acteurs, nous avons abordé, en premier lieu, la prédiction des conformations moléculaires qui reste un challenge majeur, caractérisé par ses aspects multimodal et de grandes dimensions. Nous avons alors développé une suite d'heuristiques autour d'un algorithme génétique central. Les paramètres de contrôle et les stratégies d'hybridation sont pilotés par un *méta*-algorithme permettant d'optimiser la recherche. En outre, des stratégies innovantes de parallélisation sur grilles d'ordinateurs ont été validées afin de réduire les temps de calculs. Enfin, pour entreprendre l'étude des conformations de plusieurs molécules, nous avons développé des algorithmes de criblage rapides basés sur la comparaison d'indices topologiques.

Nous avons également étudié un autre aspect en modélisant formellement certains graphes d'interactions, ceci à une toute autre échelle : celle des concentrations des molécules. Nous avons alors mis en évidence l'impact des modes d'interactions moléculaires sur la dynamique globale.

Mots-clefs : Algorithmes génétiques ; optimisation multimodale ; problèmes de grandes dimensions ; stratégies d'hybridation ; optimisation automatique des paramètres de contrôle ; modélisation moléculaire ; échantillonnage conformationnel ; biologie systémique.

Abstract

The complexity of Life is present at every level of its study : from individual molecular interactions to "interaction networks", enabling the cell to achieve its functions and to elaborate responses to external stimuli. During this thesis, tools derived from Control and Computer Sciences were used to address questions originated from Biology and Biochemistry.

Two aspects were considered : firstly the atomic description of intermolecular interactions, allowing predictions of the docking poses and the affinities between the actors. However, since these estimations depend on the flexibility of these actors, the problem of the prediction of molecular conformations, characterized by multimodal and high dimensional aspects — still a major challenge — was first addressed. Therefore, we have developped a set of heuristics around a core genetic algorithm. The operational parameters and the hybridizing strategies are under control of an algorithmic meta-layer enabling the optimization of the search. Moreover, innovative strategies for parallel deployment of the algorithms have been validated in order to reduce computing times. Finally, while undertaking the study of multiple molecules conformations, we have developed fast screening algorithms based on topological indexes.

The second aspect studied here was a formal modeling of some interaction sub-networks at the scale of the *concentrations* of the molecules. In particular, we have shown how molecular interaction modes can alter the overall dynamic of the system.

Keywords : Genetic algorithm ; multimodal optimization ; high-dimensional problems ; hybridizing strategies ; automatic optimization of operational parameters ; molecular modeling ; conformational sampling ; system Biology.