Algorithmes Haute-Performance pour la Reconnaissance de Formes Moléculaires

David Ritchie

Nancy-Université

École doctorale IAEM Lorraine

UFR STMIA, Université Henri Poincaré

Mémoire Scientifique

Algorithmes Haute-Performance pour la Reconnaissance de Formes Moléculaires

Présenté publiquement à 14:00 le 5 avril 2011

dans la salle A08 au LORIA par

David W. Ritchie

pour l'obtention de l'

Habilitation à Diriger des Recherches de l'Université Henri Poincaré

(spécialité informatique)

Composition du jury

Rapporteurs	Gilles Bernot, professeur, Université Nice Sophia Antipolis
	Frederic Cazals, DR INRIA, INRIA Sophia Antipolis – Méditerranée
	Alexandre Varnek, professeur, Université de Strasbourg
Examinateurs	Bernard Girau, professeur, Université Henri Poincaré
	Bruno Lévy, DR INRIA, INRIA Nancy – Grand Est
	Paul Zimmermann, DR INRIA, INRIA Nancy – Grand Est

Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR 7503

Remerciements

“If we knew what we were doing, it would not be called research, would it?” (Albert Einstein).

Je ne le savais pas à l’époque, mais ma carrière de recherche a débuté il y a treize ans de cela à Aberdeen, quand j’ai commencé mes études de doctorat sous la direction de Graham Kemp et John Fothergill. Donc, je tiens à remercier en premier lieu Graham et John pour leur soutien chaleureux et leur enthousiasme durant ces premières années. Par la suite, étant devenu chargé de cours à Aberdeen, l’enseignement et les tâches administratives ne m’ont pas laissé beaucoup de temps pour faire de la recherche. Il fut donc toujours un plaisir de travailler avec mes jeunes (et parfois moins jeunes) collègues de recherche Alessandra Fano, Antonis Koussounadis, Lazaros Mavridis, Diana Mustard et Violeta Pérez-Nueno. Leurs projets figurent dans ce mémoire en grande partie et quoique je n’en aie pas eu conscience à l’époque, leur travaux ont contribué à m’amener en France. Donc, je tiens à remercier Alessandra, Antonis, Lazaros, Diana et Violeta pour m’avoir aidé à garder le contact avec la science et pour être des gens avec qui il est sympathique de travailler. Je souhaiterais également remercier d’autres amis et collègues d’Aberdeen avec qui j’ai partagé des discussions profitables et des moments agréables : Peter Gray, Frank Guerin, Judith Masthof, Chris Mellish, Chris Secombes et Wamberto Vasconcelos. Aberdeen va me manquer!

D’autre part, c’est toujours bon de rencontrer de nouvelles personnes (et parfois de vieux amis dans de nouveaux endroits) et de travailler sur de nouveaux projets. Donc, je tiens á remercier Yasmine Asses, Zainab Assaghir, Thomas Bourquard, Matthieu Chavent, Emmanuelle Deschamps, Marie-Dominique Devignes, Léo Ghemtio, Anisah Ghoorah, Stéphane Gégout, Bernard Maigret, Lazaros Mavridis, Amedeo Napoli, Violeta Pérez-Nueno, Malika Smaïl-Tabbone, Michel Souchet et Vishwesh Venkatraman pour aider á faire ma transition en France comme une expérience très agréable. En particulier, je suis vraiment reconnaissant á Yasmine Asses, Matthieu Chavent, Anisah Ghoorah, Bernard et Françoise Maigret et Malika Smaïl-Tabbone pour leur aide avec la version française de ce mémoire. Il aurait été impossible sans eux! Merci à tous et toutes!

Dave Ritchie

Nancy, janvier 2011.

Table des matières

Abréviations

1D :	unidimensionnel (one-dimensional).
2D :	bidimensionnel (two-dimensional).
3D :	tridimensionnel (three-dimensional).
5D :	cinq dimensions (five-dimensional).
6D :	six dimensions (six-dimensional).
ANR :	Agence Nationale de la Recherche.
AIR :	contrainte de distance ambiguë (ambiguous interaction restraint).
AUC :	aire sous la courbe (area under the curve).
CATH :	classe, architecture, topologie, homologie (class, architecture, topology, homology).
CAPRI :	Critical Assessment of PRedicted Interactions.
CoG :	centre de gravité (centre of gravity).
CoH :	origine de coordonnée harmonique (centre of harmonics).
CPU :	processeur central (central processing unit).
CUDA :	Common Unified Device Architecture.
DCED :	dynamique essentielle distance-contrainte (distance constrained essential dynamics).
DNA :	acide désoxyribonucléique (deoxyribonucleic acid).
ED :	dynamique essentielle (essential dynamics).
EF :	facteur d'enrichissement (enrichment factor).
EM :	microscopie électron (electron microscopy).
ETO :	orbitale de type exponentiel (exponential type orbital).
EVA :	analyse de vecteur propre (eigenvector analysis).
FG :	Fourier-Gegenbauer.
FFT :	transformée rapide de Fourier (fast Fourier transform).
FN :	faux négatif (false negative).
FP :	faux positif (false positive).
FPR :	taux de faux positifs (false positive rate).

GL : Gauss-Laguerre.

GMP : GNU Multiple Precision.

GPU : processeur graphique (graphics processor unit).

GTO : orbitale de type gaussien (Gaussian type orbital).

HIV : virus de l'immunodéficience humaine (human immuno-deficiency virus).

HPC : calculs haut-débit (high performance computing).

HMM : modèles de Markov cachés (hidden Markov models).

HTVS : criblage virtuel haute performance (high throughput virtual screening).

KDD : extraction de connaissances à partir de bases de données (knowledge discovery in databases).

LORIA : Laboratoire Lorrain de Recherche en Informatique et ses Applications.

MD : dynamique moléculaire (molecular dynamics).

MIF : champ interactif moléculaire (molecular interaction field).

MLR : moyenne de logarithme du rang (mean log rank).

NMA : analyse des modes normaux (normal mode analysis).

NMR : résonance magnétique nucléaire (nuclear magnetic resonance).

NPC : complexe pore nucléaire (nuclear pore complex).

PC : ordinateur individuel (personal computer).

PC : physico-chimique (physico-chemical).

PC : composant principal (principal component).

PCA : analyse en composante principales (principal component analysis).

PDB : banque de données de protéines (protein data bank).

PPI : interaction de protéine-protéine (protein-protein interaction).

QM : mécanique quantique (quantum mechanics).

RDM : fouille de données relationnelles (relational data mining).

RIF : empreinte invariante à la rotation (rotationally invariant fingerprint).

RMS : racine de la moyenne carré (root mean squared).

RMSD : déviation de la moyenne carré de racine (root mean squared deviation).

RNA : acide ribonuclétique (ribonucleic acid).

ROC : caractéristique d'opérateur-récepteur (receiver-operator characteristic).

ROT : fonction de score de rotation (rotational scoring function).

SAS : surface accessible au solvant (solvent accessible surface).

SH : harmonique sphérique (spherical harmonic).

SPF : polaire sphérique de Fourier (spherical polar Fourier).

TAP-MS : purification d'affinité en tandem par spectroscopie de masse (tandem affinity purification by mass spectroscopy).

TN : vrai négatif (true negative).

TP : vrai positif (true positive).

TPR : taux de vrai positifs (true positive rate).

VS : criblage virtuel (virtual screening).

VSM-G : Virtual Screening Manager – Grids.

VDW : van der Waals.

Y2H : double-hybride en levure (yeast two-hybrid).

Chapitre 1

Introduction

1.1 Contexte et motivation

1.1.1 Identification de formes moléculaires

Le sujet central de ce mémoire est le développement des techniques informatiques pour représenter et comparer les structures et les propriétés tridimensionnelles (3D) des molécules. Pour des macromolécules telles que les protéines, ceci implique de comparer et de classifier leurs formes afin d'étudier les liens entre la structure et la fonction de celle-ci. Ceci implique également de prévoir comment les paires de protéines peuvent se réarranger (ou "s'amarrer") pour former un complexe biomoléculaire. Dans le cas de petites molécules organiques, il est nécessaire de comparer et de classifier les formes des molécules pouvant potentiellement devenir des médicaments afin de prévoir comment celles-ci pourraient alors se lier aux cibles protéiques spécifiques. Par conséquent, le contenu de ce mémoire se trouve à l'interface entre la biologie informatique et le chemoinformatique.

Selon Wikipedia,[1] la biologie informatique est décrite comme *"... un champ interdisciplinaire où s'appliquent les techniques de l'informatique, des mathématiques appliquées et des statistiques pour résoudre des problèmes biologiques. L'objectif principal de cette méthode est de développer des outils mathématiques et des techniques informatiques de simulation. Grâce à ces moyens, il est possible de répondre aux questions théoriques et expérimentales sans avoir besoin de réaliser des expérimentations en laboratoire..."* D'un autre côté, le terme chemoinformatique a été employé pour la première fois par Frank Brown (1998) pour décrire *"... l'utilisation des techniques de calcul et d'informatique en chimie afin de transformer des données en information puis de l'information en connaissance dans le but de prendre des décisions plus adaptées et plus rapides dans le secteur de l'identification et de l'optimisation de médicaments."*

Historiquement, la biologie informatique et la chemoinformatique ont souvent été traitées en tant que disciplines séparées. Cependant, il est important de relier ces deux matières d'un point de

[1] http://en.wikipedia.org/wiki/Computational_biology.

vue scientifique puisque les structures tridimensionnelles de macromolécules et de petites molécules de médicaments sont souvent indissociables de leur fonction (Bourne & Weissig, 2003; Pevsner, 2003; Petsko & Ringe, 2004) et parce que les molécules médicamenteuses servent à moduler le comportement de leurs cibles biologiques (Larson, 2006). En outre, je crois que, d'un point de vue informatique, il est important de considérer de la même façon tous les domaines scientifiques dans le but de détecter les techniques informatiques ou les algorithmes qui sont bien connus dans un domaine mais qui pourraient être prolongés et transférés à un autre domaine.

1.1.2 Importance de la structure moléculaire

En biologie, on considère que l'ordre des acides aminés au sein d'une protéine détermine sa structure 3D moléculaire, et que la structure tridimensionnelle d'une protéine détermine sa fonction spécifique. Cependant, en raison de l'énorme volume et de la complexité considérable des données biologiques, il est nécessaire d'employer des techniques informatiques et de visualisation avancées dans le but de stocker et de manipuler celles-ci facilement. C'est dans cette *ère de post-génomique*, où de plus en plus de séquences complètes d'organismes sont déterminées de manière routinière, que l'attention des scientifiques se tourne plus particulièrement vers la transformation de cette information de base en connaissances structurales et par conséquent fonctionnelles. De ce fait, la capacité de représenter et manipuler les structures moléculaires *in silico* deviendra un aspect de plus en plus important de la biologie informatique. De plus, il existe déjà des techniques informatiques utilisées dans de nombreux secteurs des sciences de la vie pour aider à comprendre et à exploiter la quantité importante de séquences et de données structurales déjà disponibles. Par exemple, les biologistes réalisent souvent des alignements de séquences et des superpositions structurales de protéines afin d'améliorer la compréhension qu'ils ont de la fonction biologique de celles-ci. De plus, ils emploient également des programmes "d'amarrage" (ou "docking") de protéines pour essayer de prédire comment ces partenaires biologiques peuvent s'assembler au niveau moléculaire. De même, les chimistes médicinaux utilisent souvent des programmes d'amarrage entre protéine et petite molécule (ou ligand) afin d'identifier les molécules susceptibles de se lier à une cible protéique donnée et, par conséquent, d'en moduler le fonctionnement dans un but thérapeutique (Jensen, 1999; Dean, 1995; Petsko & Ringe, 2004). En effet, l'utilisation de la structure tridimensionnelle de protéines est une stratégie de plus en plus utilisée dans la découverte de médicaments (Richards, 2002; Congreve *et al.*, 2005).

1.1.3 Goulots d'étranglement expérimentaux et informatiques

Au niveau expérimental, la cristallographie aux rayons X et la résonance magnétique nucléaire (NMR)[2] sont souvent considérées comme des techniques "étalon or" pour déterminer les structures haute résolution de protéines ou d'autres macromolécules. Cependant, la résolution de la structure 3D d'une protéine est considérablement plus difficile que la détermination de sa séquence. Par exemple, bien qu'environ 12.000 structures distinctes de protéines aient été déposées à la banque de données de protéines (ou "protein databank," PDB; Sussman *et al.*, 1998) et que de nouvelles structures soient ajoutées régulièrement à la PDB (avec un taux de 100 nouvelles structures par semaine), ce nombre ne représente seulement qu'une proportion très restreinte de l'ensemble des séquences connues de protéines. Par conséquent, il y a un besoin de pouvoir créer les modèles 3D structuraux des protéines. En outre, seulement une petite proportion des structures 3D déposées dans le PDB correspondent aux complexes de protéine-protéine, et moins de 2% de toutes les structures connues comportent des hétéro-complexes de protéine-protéine. De plus, en raison d'un certain nombre de difficultés pratiques, il semble peu probable, dans un futur proche, de déterminer les structures de complexes protéiques en utilisant des techniques génomiques structurales haut-débit (Russell *et al.*, 2004). Par conséquent, l'utilisation de techniques informatiques telles que la modélisation par homologie et l'amarrage de protéines deviendront des protocoles de plus en plus utilisés pour aider à comprendre les mécanismes moléculaires des systèmes biologiques (Aloy *et al.*, 2004; Aloy & Russell, 2006).

Si l'on considère les biomolécules d'un point de vue structural, il est important de se rappeler que celles-ci et de nombreux petits ligands sont des entités intrinsèquement dynamiques dans les conditions physiologiques. Par exemple, nous pouvons noter que les différentes positions des atomes au sein d'une protéine fluctuent rapidement de façon continue à cause du mouvement brownien. À plus long terme, les conformations[3] des acides aminés dans une protéine peuvent passer d'un minimum local à un autre. À encore plus long terme, des sous-unités structurales de plus grande taille, comme les hèlices-α et les feuillets-β, peuvent subir des mouvements substantiels qui peuvent être très difficiles à prédire en utilisant des techniques informatiques.

Bien que les forces fondamentales régissant les interactions moléculaires soient presque entièrement comprises à un niveau théorique, simuler et manipuler *in silico* de grandes biomolécules telles

[2] Pour la compatibilité avec la littérature scientifique, j'emploierai généralement des acronymes anglais bien connus pour la plupart des abréviations. Ainsi, résonance magnétique nucléaire devient NMR (nuclear magnetic resonance) et acide désoxyribonucléique devient DNA (deoxyribonucleic acid), par exemple.

[3] En chimie, le terme conformation est employé pour décrire les positions relatives des atomes dans une molécule. Les molécules qui ont le même nombre et type d'atomes et les mêmes liaisons entre ceux-ci peuvent exister sous différentes conformations 3D. Une molécule dans une conformation particulière peut évoluer vers une autre conformation grâce à une ou plusieurs rotations des liaisons interatomiques. Il existe souvent une barrière d'énergie liée à de telles rotations, mais le mouvement thermique fournit le plus souvent l'énergie nécessaire pour surmonter celles-ci. Néanmoins, au-delà d'un temps suffisamment long, les protéines adoptent normalement la conformation de plus basse énergie.

que des protéines dans l'espace tridimensionnel restent des actions coûteuses en puissance informatique, ce qui limite celles-ci. Par exemple, effectuer une simulation de dynamique moléculaire (MD) sur un seul domaine protéique peut impliquer plusieurs jours voire plusieurs semaines de calculs. Bien que des progrès considérables aient été accomplis pour réduire les temps de calcul, modéliser de façon fiable la maniére dans deux protéines interagissent l'une avec l'autre au niveau moléculaire demeure un défi informatique considérable. Si l'on considère explicitement les protéines comme des structures flexibles, le temps nécessaire pour modéliser leurs interactions peut prendre environ 50 jours par processeur et par assemblage. Même le fait de superposer de façon optimale les structures de protéines similaires, demeure un problème informatique non trivial (Sippl & Wiederstein, 2008). De même, la découverte de médicaments basée sur la structure tridimensionnelle des molécules peut être très coûteuse, en temps de calculs, et ceci en particulier à cause de la taille démesurée des bases de données chimiques à examiner. Il est donc nécessaire de développer de nouvelles techniques informatiques permettant de représenter et de manipuler de manière efficace les structures 3D des protéines, ou d'autres molécules, pour simuler les interactions entre protéines ou entre protéines et ligands.

1.1.4 Interactions protéine-protéine

Si le DNA représente le plan d'architecte de la vie, alors les protéines sont définies comme les machines moléculaires accomplissant ce plan. Les protéines réalisent souvent leurs fonctions en agissant avec d'autres protéines pour former des complexes protéiques. Ces complexes peuvent exister en tant qu'associations transitoires de courte durée (comme, par exemple, dans le cas de la catalyse enzymatique), de systèmes multimériques (tels que le ribosome), les facteurs de transcription, les protéines à la surface des cellules, ou encore les protéines formant les canaux ioniques. Cependant, même si nous connaissons le plan de montage – c.-à-d. la séquence DNA – nous ne savons que très peu de choses sur le fonctionnement des protéines au niveau moléculaire. En effet, les études portant sur l'ensemble du génome (Uetz *et al.*, 2000; Ito *et al.*, 2001; Gavin *et al.*, 2002; Ho *et al.*, 2002) fournissent une liste croissante d'interactions protéine-protéine (PPI) potentielles; mais comprendre le fonctionnement de ces interactions exige davantage d'analyses biochimiques et structurales. Par exemple, la levure est l'un des organismes les plus étudiés et est connue pour avoir à peu près 6.000 protéines, impliquant environ entre 38.000 et 75.000 PPI. Environ 50% de ces PPI ont été observées expérimentalement. D'autre part, le génome humain code approximativement 30.000 protéines, donnant de 154.000 à 370.000 PPI, dont seulement $\sim$10% sont connus à ce jour (Aloy & Russell, 2004; Hart *et al.*, 2006). Par conséquent, le développement des approches automatisées permettant de fouiller et d'extrapoler les interactions de la levure vers l'humain sera une stratégie importante pour compléter le réseau d'interaction protéique de l'homme (Bork *et al.*, 2004). Comprendre comment les protéines interagissent est essentiel à la compréhension des mécanismes moléculaires provoquant

des maladies. Par exemple, les médicaments fonctionnent souvent en modulant ou en bloquant les PPI, et donc ces PPI représentent une classe importante de cibles médicamenteuses (Arkin & Wells, 2004; González-Ruiz & Gohlke, 2006).

1.1.5 Amarrage de protéines

L'amarrage de protéines consiste à modéliser la structure 3D d'un complexe de protéines à partir des structures individuelles de chaque protéine dans leur état non lié. Comme tous les bons problèmes scientifiques, l'amarrage de protéines est un problème facile à énoncer mais beaucoup plus délicat à résoudre. Puisque les structures des protéines sont intrinsèquement dynamiques, elles peuvent souvent changer de conformation lors de leur complexation. Ceci revient donc à essayer d'assembler les morceaux d'un puzzle 3D particulièrement complexe dans lequel les pièces indiquées ne s'adaptent pas parfaitement ensemble. Par conséquent, les temps de calculs nécessaires à une telle tâche sont particulièrement élevés. Afin d'effectuer le calcul en un temps acceptable, la plupart des algorithmes d'amarrage de protéines commencent en supposant que les structures à assembler sont rigides; on parle alors d'amarrage corps rigide. Ceci permet de ramener le problème à un espace de rotation-translation à six dimensions (6D) de recherche. Néanmoins, cette approximation peut également produire un grand nombre d'orientations d'amarrage incorrectes – ou faux-positifs. Par conséquent, le but général des algorithmes d'amarrage corps rigide est de trouver un nombre raisonnablement petit d'orientations faisables pour une paire de structures 3D qui peuvent plus tard être raffinées et reclassées en utilisant des techniques plus conventionnelles mais comportant plus de calculs. Pour plus d'informations sur le sujet, vous pouvez vous référer à la revue Ritchie (2008) et aux références qu'elle contient.

Actuellement, de nombreux algorithmes utilisent des techniques de corrélation de la transformée rapide de Fourier (ou "fast Fourier transform," FFT) afin d'obtenir des orientations de départ lors de la phase d'amarrage corps rigide. Cette approche a été présentée pour la première fois par Katchalski-Katzir *et al.* (1992) pour calculer rapidement la complémentarité de formes entre partenaires protéiques dans une grille 3D cartésienne. Cette méthode a ensuite été étendue pour inclure des termes additionnels représentant des interactions électrostatiques (Gabb *et al.*, 1997; Mandell *et al.*, 2001), voire un ensemble de termes comme les contributions électrostatiques et de désolvatation (Chen & Weng, 2003). Cependant, parce que cette approche se fonde sur une grille cartésienne, elle permet de calculer uniquement des corrélations de translation, et de nouvelles grilles FFT doivent être calculées pour chaque incrément lors de la rotation de la molécule. À cause de cette limitation, il est donc difficile d'incorporer des connaissances au sujet d'un complexe afin de focaliser le calcul autour d'un site de fixation connu ou présumé. Puisque couvrir entièrement l'espace de recherche exige des milliers d'échantillons de rotations, les algorithmes cartésiens d'amarrage nécessitent généralement plusieurs heures pour accomplir un calcul; et l'efficacité de ces approches diminue lorsque l'on

augmente la complexité du potentiel. Afin de simuler la flexibilité des protéines pendant les calculs d'amarrage, plusieurs groupes de recherche emploient des techniques de FFT pour assembler des *ensembles* de structures rigides (Grünberg *et al.*, 2004; Mustard & Ritchie, 2005; Smith *et al.*, 2005). Ceci augmente considérablement le coût de calcul de ces approches. L'utilisation croissante de tels potentiels basés sur la connaissance et de techniques d'amarrage d'ensemble entraîne donc un besoin important de développer des approches de FFT plus sophistiquées et plus souples.

1.1.6 Corrélations polaires sphériques de Fourier pour l'amarrage macromoléculaire

Beaucoup d'algorithmes actuels d'amarrage de protéines emploient des méthodes de FFT pour calculer les corrélations de translation (Eisenstein & Katchalski-Katzir, 2004). En d'autres termes, ils placent les protéines à associer dans une grille 3D cartésienne puis utilisent une bibliothèque existante de FFT 3D pour accélérer le calcul de la corrélation de translation d'une protéine mobile autour du partenaire fixe. La partie FFT de ce type de calculs d'amarrage peut être calculée rapidement, mais ceci doit se répéter pour les milliers d'incréments composant la rotation. En outre, plusieurs des pas de translation peuvent correspondre à des phénomènes d'interpénétration entre les deux protéines qui sont peu réalistes. De plus, si l'on garde en mémoire que deux des degrés de liberté de rotation sont redondants lorsque la composante de translation est nulle, ceci implique que de nombreuses orientations échantillonnées lors des FFTs de translation seront quasiment redondantes lorsque de telles FFTs seront appliquées aux multiples orientations de rotation.

La thématique principale de ce mémoire est donc que les problèmes d'amarrage de protéines et de reconnaissance moléculaire de formes sont, en soi, des problèmes *de rotation* et, de ce fait, devraient être décrits en utilisant des systèmes de coordonnées angulaires de sorte qu'ils puissent être plus naturellement adaptés à des FFT *de rotation*. Comme indiqué ci-dessus, les algorithmes conventionnels d'amarrage de FFT basés sur une grille divisent l'espace de recherche 6D en trois degrés de liberté de rotation et trois de translation. Cependant, ceci permet seulement à trois des degrés de liberté (de translation) d'être accélérés par la FFT. À l'inverse, un système de coordonnées sphériques polaires possède une dimension de translation et cinq de rotation. Ceci permet donc d'employer des techniques de FFT accélérant le calcul dans au moins cinq et potentiellement six degrés de liberté.

Ma contribution principale dans le domaine de l'amarrage de protéines a été d'explorer et de démontrer l'utilité de la méthode utilisant les coordonnées sphérique polaires. J'ai nommé cette approche: *polaire sphérique de Fourier* (ou "spherical polar Fourier," SPF). L'idée fondamentale est de représenter la forme et les propriétés électrostatiques des protéines (ou d'autres biomolécules telles que le DNA et le RNA) en tant qu'expansions de Fourier d'ordre supérieur des fonctions orthonormales sphériques d'harmonique (SH) et de base du Gauss-Laguerre (GL). Une grande partie de la théorie mathématique de cette approche est "bien connue" dans le sens où elle découle de la des-

cription des orbitales électroniques d'un atome en mécanique quantique (QM). Néanmoins, je fus le premier à employer cette approche pour représenter les formes des macromolécules entières, et j'ai montré que cette représentation est exceptionnellement bien adaptée pour calculer le recouvrement entre des paires de fonctions 3D (c.-à-d. les propriétés moléculaires) de manière très rapide en utilisant les techniques de FFT.

Ces idées ont été implémentées dans le programme d'amarrage *Hex* (Ritchie & Kemp, 2000). Pendant un calcul d'amarrage, le programme *Hex* peut évaluer plusieurs millions d'orientations par seconde sur un ordinateur individuel ordinaire (ou "personal computer," PC). Ce logiciel a été employé avec succès dans l'expérience CAPRI – expérience d'amarrage en aveugle - (Janin *et al.*, 2003; Méndez *et al.*, 2003; Méndez & Wodak, 2005), et est donc reconnu internationalement. Avec plus de 12.000 téléchargements, *Hex* est utilisé à travers le monde aussi bien dans le milieu universitaire qu'en industrie et a été cité dans plus de 200 publications scientifiques.

Récemment, Garzon *et al.* (2009) ont décrit une approche de corrélation d'amarrage (FRODOCK: "fast rotational docking," ou amarrage avec rotation rapide) dans laquelle la fonction de score est basée sur la forme, l'électrostatique et la désolvatation et les trois degrés de liberté de rotation sont accélérés par une FFT 3D. On rapporte que FRODOCK est presque aussi rapide que *Hex*. Cependant, comme FRODOCK ne possède pas les fonctions radiales spéciales utilisées dans *Hex*, il doit employer un jeu de sphères concentriques pour représenter les formes des protéines et doit exécuter la partie 3D de translation, lors de la recherche d'amarrage, par reéchantillonnage explicite de la fonction potentielle du partenaire mobile sur un grand nombre d'échantillons de translation. Par conséquent, je pense que mon approche de SPF continue à définir le "state-of-the-art" dans le domaine de l'amarrage entre protéines basé sur la FFT.

1.1.7 Criblage virtuel de médicaments

Une des activités principales dans le protocole de développement de médicaments est d'employer des techniques informatiques pour identifier ou prédire comment de petites molécules pourraient se lier à une cible protéique donnée pour en modifier la fonction. Cette méthode informatique est souvent appelée le criblage virtuel (ou "virtual screening," VS). Les initiatives de génomique structurale produisent des structures de protéines à une vitesse de plus en plus importante, et chaque nouvelle protéine caractérisée pourrait servir de nouvelle cible potentielle pour de futurs médicaments.

Par conséquent, il y a de plus en plus d'occasions d'exploiter cette connaissance structurale naissante dans des buts thérapeutiques. De plus, les sociétés pharmaceutiques ont maintenant des bases de données de composés qui contiennent des informations sur les structures chimiques pour, littéralement, des millions de molécules. Améliorer les méthodes permettant de fouiller dans de telles bases de données pourrait amener à un développement plus rapide et plus efficace de nouveaux médicaments. Cependant, bien que les outils courants d'amarrage protéine-ligand, tels

que Autodock, puissent examiner avec succès un nombre modeste (c.-à-d. de l'ordre de quelques milliers) de ligands sur une cible protéique donnée (Park *et al.*, 2006), ils restent encore trop lents pour le criblage virtuel haut débit (ou "high throughput virtual screening," HTVS) de bases de données chimiques contenant des millions de composés (Khodade *et al.*, 2007). Par conséquent, il y a un besoin croissant de développer des techniques avancées de recherche au sein des bases de données et d'amarrage protéine-ligand, ceci dans le domaine de la découverte de nouveaux médicaments basée sur la structure.

Globalement, il existe deux approches principales du criblage virtuel. Dans des approches basées sur le récepteur, la structure de la cible protéique est connue ou a été modélisée, et le but est de trouver les ligands appropriés qui, par exemple, pourraient se lier près de l'emplacement du site actif du récepteur et, par conséquent, moduler voir bloquer la fonction initiale du récepteur. Dans des approches basées sur le ligand, la structure de la cible protéique n'est généralement pas connue, et le but est de trouver de nouveaux ligands aux propriétés structurales semblables aux molécules antagonistes connues. Certains médicaments agissent comme des agonistes (c.-à-d. ils activent ou augmentent la fonction originale du récepteur); dans ce cas, les principes de criblage sont essentiellement les mêmes que pour des antagonistes. En raison du coût informatique des approches basées uniquement sur la structure du récepteur (c.-à-d. l'amarrage protéine-ligand), les campagnes de HTVS utilisent le plus souvent une combinaison des deux approches, dans laquelle des critères de similitude basés sur le ligand sont employés comme premier filtre. Les candidats sélectionnés durant ce premier filtre sont ensuite associés à la cible protéique. L'approche VSM-G ("virtual screening manager for grids") est un bon exemple de ce principe de filtrage (Beautrait *et al.*, 2008). Si les ressources informatiques suffisantes sont disponibles, il est possible de passer outre le filtre basé sur la structure du ligand et d'effectuer directement le criblage basé sur le récepteur. Par exemple, dans le profil "élevé" du projet d'économiseur d'écran THINK, plus de 1.000.000 PC autour du monde ont été rendus disponibles pour fournir 80.000 ans de temps-CPU afin de tester une base de données virtuelle d'environ 3.5 milliards de composés sur 12 cibles protéiques impliquées dans divers cancers (Davies *et al.*, 2002). Cependant, seules les plus grandes sociétés privées ou les organisations gouvernementales peuvent avoir les moyens d'utiliser des ressources de calcul comparables impliquant des calculs haute performance (ou "high performance computing", HPC) pour l'HTVS.

1.1.8 Utilisation des harmoniques sphériques pour le criblage virtuel

À mon avis, l'état de l'art actuel pour la comparaison moléculaire efficace de la forme 3D est basé sur les représentations gaussiennes de la forme moléculaire (Grant *et al.*, 1996) et sur l'approche plus récente de représentations de l'enveloppe en utilisant les SH, qui a été développée indépendamment par moi-même à Aberdeen (Ritchie & Kemp, 1999; Mavridis *et al.*, 2007), Bernard Maigret à Nancy (Cai *et al.*, 2002; Yamagishi *et al.*, 2006), et Tim Clark à Erlangen (Lin & Clark, 2005). À Aberdeen,

mon but était d'exploiter les propriétés de rotation des fonctions SH pour développer une manière très rapide de superposer et de comparer quantitativement les formes 3D des surfaces moléculaires. À Nancy, le travail du Dr Maigret concernait l'utilisation les représentations SH pour fournir un filtre rapide basé sur le récepteur pour le criblage virtuel (Beautrait *et al.*, 2008). À Erlangen, le professeur Clark a développé le programme ParaSurf pour représenter les principales propriétés des surfaces moléculaires en QM grâce à la représentation SH. Pour compléter le travail d'Erlangen, j'ai développé le programme ParaFit qui peut superposer les structures 3D moléculaires calculées par le programme ParaSurf à un taux allant jusqu'à cent molécules par seconde sur un seul processeur. ParaFit et ParaSurf sont actuellement commercialisés par la société Cepos Insilico Ltd.

1.2 Résumé et structure du document

Ce document présente un résumé de mes contributions dans le domaine de la représentation informatique de la forme moléculaire et de l'amarrage entre protéines en utilisant une transformation de Fourier efficace pour représenter celles-ci. Comme cité ci-dessus, une grande partie de la théorie mathématique utilisée ici est bien connue des chimistes théoriciens et des physiciens nucléaires, mais n'est généralement pas bien connue au delà de ces champs spécifiques. Néanmoins, les preuves mathématiques formelles ne seront généralement pas abordées dans ce mémoire. Le lecteur intéressé peut trouver des informations complémentaires dans de nombreux ouvrages de référence ayant pour sujet les fonctions spéciales et la QM (Talman, 1968; Luke, 1969; Hochstadt, 1971; Biedenharn & Louck, 1981; Sakurai, 1994; Bransden & Joachain, 1997). Au lieu de cela, la démarche adoptée dans les chapitres suivants est de présenter toute formule ou tout résultat mathématique nécessaire comme des axiomes, et d'employer ces derniers en tant que blocs fonctionnels de base à partir desquels seront développées les représentations informatiques 2D et 3D ainsi que les algorithmes de corrélation, ceci en utilisant seulement des techniques de calcul relativement simples.

Notons que, ces dernières années, les représentations de surfaces utilisant les SH sont de plus en plus appliquées à une large gamme d'identifications et d'enregistrements d'objets dans les secteurs allant, par exemple, de l'anatomie (McPeek *et al.*, 2008; McPeek *et al.*, 2009) au génie civil (Garboczi, 2002; Grigoriu *et al.*, 2006), en passant par des domaines comme la microscopie cryo-électronique (Kovacs & Wriggers, 2002) ou la formation d'images médicales (Frank, 2002b; Edvardson & Smedby, 2003; Huang *et al.*, 2005), l'infographie ou l'Internet (Kautz *et al.*, 2002; Funkhouser *et al.*, 2003; Kazhdan *et al.*, 2003; Novotni & Klein, 2003; Shen *et al.*, 2009b; Shen *et al.*, 2009a). Cependant, à ma connaissance, je reste la seule personne à avoir étendu les fonctions SH angulaires avec des fonctions radiales orthonormales pour construire des fonctions polaires sphériques en base 3D et à avoir employé cette représentation avec succès afin d'accélérer le calcul des corrélations de rotation et de translation dans l'espace 3D. Par conséquent, les approches présentées ici sont novatrices dans le domaine de l'amarrage protéine-protéine et pour l'identification de formes 3D en général.

Le reste de ce document est structuré comme suit. Le chapitre 2 récapitule quelques propriétés mathématiques utiles des fonctions spéciales et démontre comment elles peuvent être employées pour construire des fonctions orthogonales de base pour la représentation des formes 2D et 3D moléculaires ainsi que d'autres propriétés. Le chapitre 3 développe cela en discutant l'action de la rotation et des opérateurs de translation dans les espaces de Hilbert, en utilisant pour cela les matrices de rotation de Wigner pour les SH et en appliquant une transformation sphérique de Bessel afin de calculer des expressions de forme close pour les éléments correspondant aux matrices de translation. Par conséquent, ces chapitres décrivent les mécanismes de base nécessaires pour exécuter les corrélations de Fourier en six dimensions.

Le chapitre 4 montre comment l'approche globale peut être appliquée à la superposition, la comparaison et la classification des structures de protéines connues en utilisant des représentations de forme-densité de la SPF 3D, et comment ces représentations peuvent être exploitées pour exécuter des calculs efficaces d'amarrage de protéines en utilisant des corrélations de rotation FFT 1D, 3D, et 5D. Ce chapitre décrit également comment une approche 2D, de fonction SH simple, faisant correspondre les surfaces moléculaires, peut être employée pour le criblage virtuel de médicaments haut débit sur de grandes bases de données chimiques. En conclusion, le chapitre 5 présente une vue d'ensemble des projets en cours et de futurs objectifs. L'annexe énumère plusieurs articles de journaux à comité de lecture dans lesquels le travail présenté dans ce mémoire a été édité. Sauf indication contraire, toutes les figures de ce document représentant des surfaces moléculaires ont été produites à partir du programme *Hex* que j'ai développé.

Chapitre 2

Mathématiques de base

Le but principal de ce chapitre est d'introduire les concepts et les techniques mathématiques nécessaires pour représenter et manipuler des formes moléculaires de manière utile et efficace en computation. Une grande partie de ce document décrit les fonctions classiques spéciales des mathématiques qui sont essentiellement utilisées ici pour fournir des fonctions orthonormées de base pour des expansions de Fourier dans l'espace 3D (Luke, 1969; Lebedev, 1972). Puisque ces expansions impliqueront des polynômes d'ordre relativement supérieur, il est important de développer des méthodes de calcul efficaces qui ne sacrifient pas l'exactitude numérique. Il est aussi important d'utiliser des techniques d'échantillonnage spatiales fiables et efficaces. Par conséquent, ce chapitre décrit aussi en bref la tessellation icosaèdrique de la sphère.

2.1 Les fonctions spéciales

Les fonctions spéciales jouent un rôle important dans des applications de physique et d'ingénierie parce qu'elles apparaissent souvent comme les solutions pour certaines équations différentielles et intégrales (Lebedev, 1972). En conséquence, elles peuvent être utilisées pour représenter et modéliser plusieurs phénomènes naturels, allant de l'électrostatique et l'électromagnétisme aux théories quantiques de la matière. Les fonctions spéciales sont des polynômes analytiques lisses, dont les termes présentent souvent des motifs simples ou "spéciaux". Par exemple, toutes les fonctions spéciales peuvent être calculées par des formules de récursivité de trois-termes (Erdélyi *et al.*, 1953a). De même, beaucoup d'intégrales qui impliquent des produits des fonctions spéciales ont souvent des solutions relativement concises. Avant l'époque des ordinateurs modernes, ces propriétés ont considérablement facilité des calculs difficiles faits à la main. Bien sûr, de nos jours, il est possible de programmer et de calculer des fonctions d'ordre encore supérieur qui ne pourraient jamais être faits à la main. Néanmoins, il est encore utile d'étudier les fonctions spéciales et leurs propriétés parce que l'utilisation d'une formule de récursivité ou d'une intégrale transformée bien-choisie peut engendrer

un gain de performance spectaculaire en informatique.

2.1.1 Les fonctions analytiques

En mathématiques, les fonctions analytiques sont des fonctions lisses infiniment différentiables d'une ou de plusieurs variables. Un exemple fondamental d'une fonction analytique est la série entière

$$f(x) = \sum_{l=0}^{\infty} a_l x^l, \tag{2.1}$$

où les valeurs particulières des coefficients, a_l, distinguent une fonction des autres. Une telle série entière peut être différenciée un nombre quelconque de fois, m, pour donner

$$\frac{\mathrm{d}^m}{\mathrm{d}x^m} f(x) \equiv f^{(m)}(x) = \sum_{l=0}^{\infty} l(l-1)(l-2)...(l-m+1) a_l x^{l-m}. \tag{2.2}$$

En notant que chaque coefficient a_l peut être isolé en évaluant la dérivée l-ième de $f(x)$ à $x = 0$,

$$a_l = \frac{f^{(l)}(0)}{l!}, \tag{2.3}$$

la somme ci-dessus peut être écrite sous la forme d'une série de Taylor

$$f(x) = \sum_{l=0}^{\infty} \frac{f^{(l)}(0)}{l!} x^l. \tag{2.4}$$

En d'autres termes, si la forme de $f(x)$ est déjà connue et est facilement différentiable, les coefficients de ses séries entières peuvent être calculés en utilisant Eq 2.3. Par exemple, $\cos x$ est donné par

$$\begin{aligned} \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + ... \\ &= \sum_{l=0}^{\infty} \frac{(-1)^l x^{2l}}{(2l)!}. \end{aligned} \tag{2.5}$$

De nos jours, on pense qu'il est normal que les langages modernes de programmation nous fournissent des routines intégrées pour calculer des fonctions classiques de trigonométrie et de transcendantales, en utilisant normalement la série entière qui est semblable à la précédente. En revanche, quand la forme de $f(x)$ n'est pas connue à l'avance, comme c'est souvent le cas dans les problèmes d'ajustage ou d'analyse de données, il n'est souvent pas simple d'isoler et de calculer chaque coefficient. Au lieu de cela, les coefficients doivent être calculés "simultanément" par des techniques des moindres carrés par exemple et cette technique peut être coûteuse en calculs et sujette aux erreurs.

2.1.2 Les polynômes 3D homogènes

Ici, le but est de représenter des formes moléculaires et d'autres propriétés moléculaires comme des expansions de polynômes d'ordre supérieur dans l'espace tridimensionnel (3D). Par exemple, nous cherchons une représentation de la forme

$$f(\underline{x}) = \sum_{l=0}^{\infty}(a_l x + b_l y + c_l z)^l, \tag{2.6}$$

avec les coefficients a_l, b_l et c_l et où $\underline{x} = (x, y, z)$ sont les coordonnées 3D cartésiennes usuelles. Eq 2.6 est parfois appelée fonction harmonique solide (Hobson, 1931) parce qu'on peut la changer en des coordonnées polaires sphériques $\underline{r} = (r, \theta, \phi)$ en faisant les substitutions

$$\begin{aligned} x &= r\sin\theta\cos\phi, \\ y &= r\sin\theta\sin\phi, \\ z &= r\cos\theta, \end{aligned} \tag{2.7}$$

pour obtenir

$$f(\underline{r}) \equiv f(\underline{x}) = \sum_{l=0}^{\infty} r^l(a_l\sin\theta\cos\phi + b_l\sin\theta\sin\phi + c_l\cos\theta)^l, \tag{2.8}$$

qui est évidemment une somme de toutes les combinaisons possibles des fréquences harmoniques ou des puissances trigonométriques. Le rapport entre les coordonnées 3D cartésiennes et polaires sphériques est montré dans la figure 2.1.

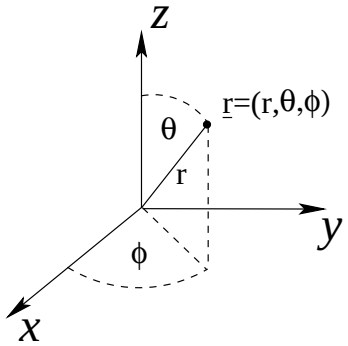

Figure 2.1: Le rapport entre les coordonnées cartésiennes (x, y, z) et polaires sphériques (r, θ, ϕ).

Malheureusement, les équations 2.6 et 2.8 ne sont pas très utiles dans des buts pratiques parce que quelques puissances et par conséquent quelques coefficients d'expansion, ne sont pas linéairement indépendants. Par exemple, en considérant les puissances de deux, on trouve que

$$(ax + by + cz)^2 = a^2x^2 + b^2y^2 + c^2z^2 + 2abxy + 2acxz + 2bcyz. \tag{2.9}$$

Mais à partir de l'Eq 2.7, il est clair que

$$x^2 + y^2 + z^2 = r^2, \tag{2.10}$$

ainsi x^2, y^2 et z^2 sont linéairement dépendants. D'où, seulement cinq des six puissances cartésiennes de deuxième ordre sont effectivement linéairement indépendantes. De la même manière, il peut être montré que seulement sept des dix puissances cartésiennes harmoniques de troisième ordre sont linéairement indépendantes et ainsi de suite. En général une façon fiable d'énumérer toutes les puissances polynômiales linéairement indépendantes dans l'espace 3D est d'écrire

$$f(\underline{r}) = \sum_{l=0}^{\infty} r^l \left(c_l P_{l0}(\cos\theta) + \sum_{m=1}^{l} \big(a_{lm} \cos m\phi + b_{lm} \sin m\phi \big) P_{lm}(\cos\theta) \right), \tag{2.11}$$

où $P_{lm}(\cos\theta)$ sont les polynômes de Legendre. Cependant, avant de considérer les polynômes de Legendre plus en détails, il est d'abord utile de réexaminer quelques unes des propriétés de base des fonctions trigonométrique ou circulaires plus simples.

2.1.3 Les fonctions circulaires

Les fonctions trigonométriques sinus et cosinus sont souvent appelées fonctions circulaires parce qu'elles décrivent le rapport entre une coordonnée angulaire ϕ et les coordonnées cartésiennes x et y le long du chemin d'un cercle

$$\begin{aligned} x &= \cos\phi, \\ y &= \sin\phi. \end{aligned} \tag{2.12}$$

Si un cercle avec un rayon d'une unité est tracé à l'origine dans le plan complexe, ses coordonnées peuvent être représentées de manière compacte comme un seul nombre complexe.[1]

$$w = \cos\phi + i \sin\phi, \tag{2.13}$$

où $i = \sqrt{-1}$ est l'unité imaginaire et où les parties réelles et complexes de w correspondent aux coordonnées x et y respectivement. Un des principaux résultats de l'analyse complexe est que les fonctions trigonométriques peuvent être liées à la fonction exponentielle par la formule d'Euler

$$e^{i\phi} = \cos\phi + i \sin\phi. \tag{2.14}$$

Plusieurs autres relations découlent de ceci. Par exemple, la formule de De Moivre

$$(\cos\phi + i \sin\phi)^m = \cos m\phi + i \sin m\phi \tag{2.15}$$

[1] Beaucoup de manuels utilise z pour représenter une variable complexe, mais ici j'utiliserai w pour laisser x, y et z correspondre aux coordonnées 3D cartésiennes habituelles sans aucune confusion.

découle directement du fait que

$$(e^{i\phi})^m = e^{i(m\phi)}. \tag{2.16}$$

Si on met $\phi = 2\pi/n$, les m nombres complexes $e^{i2\pi m/n}$, pour $0 \leq m < n$, comportant les sommets d'un polygone régulier dans le plan complexe qui sont parfois appelés les racines n-ème de l'unité. Ces points spéciaux occupent une place importante dans la théorie de la transformée de Fourier (FFT) rapide (Kammler, 2000). L'utilisation de FFT dans mon travail de recherche sera décrite dans le chapitre 4.

Quand les fonctions circulaires doivent être calculées pour de multiples intervalles réguliers, elles peuvent être calculées efficacement en utilisant les formules de récursivité. Par exemple, en écrivant

$$\cos m\phi = \cos(\phi + (m-1)\phi), \tag{2.17}$$

$$\sin m\phi = \sin(\phi + (m-1)\phi), \tag{2.18}$$

et en appliquant les identités (qui peuvent être dérivées de la formule de De Moivre)

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta, \tag{2.19}$$

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta, \tag{2.20}$$

on obtient les formules de récursivité stables

$$\cos m\phi = 2\cos\phi\cos(m-1)\phi - \cos(m-2)\phi, \tag{2.21}$$

$$\sin m\phi = 2\cos\phi\sin(m-1)\phi - \sin(m-2)\phi. \tag{2.22}$$

2.1.4 Les fonctions orthogonales et espaces d'Hilbert

Les fonctions circulaires sont orthogonales dans le sens que

$$\int_0^{2\pi} e^{im\phi}e^{-ij\phi}\mathrm{d}\phi = 2\pi\delta_{mj}, \tag{2.23}$$

où δ_{mj} est le symbole de Kronecker

$$\delta_{mj} = \begin{cases} 1 \text{ si } m = j \\ 0 \text{ autrement.} \end{cases} \tag{2.24}$$

En d'autres termes, le *recouvrement* total entre une quelconque paire distincte de fonctions de base est zéro. Les fonctions orthogonales jouent un rôle important dans la théorie des espaces de Hilbert. Un espace de Hilbert est essentiellement une extension algébrique de la notion d'un espace euclidien ordinaire, dans lequel l'espace est défini par une liste infinie de fonctions de base orthogonales (correspondant aux axes d'un espace euclidien) et où n'importe quel point, dans cet espace, peut

être décrit comme une combinaison linéaire de fonctions de base (correspondant à un vecteur de coordonnées dans l'espace euclidien) (Debnath & Mikusinski, 1999). Une application pratique d'un espace de Hilbert est de représenter une fonction arbitraire, $f(\phi)$, dans le domaine $0 \leq \phi < 2\pi$ comme une série de Fourier

$$f(\phi) = \sum_{m=0}^{\infty} a_m e^{im\phi}, \tag{2.25}$$

où les fonctions $e^{im\phi}$ servent d'ensemble de fonctions de base orthogonales et les coefficients d'expansion a_m servent de "coordonnées" par rapport à l'ensemble de base. En utilisant la propriété d'orthogonalité, le m-ème coefficient peut être déterminé en multipliant chaque côté de l'Eq 2.25 par $e^{-im\phi}$ et en intégrant pour obtenir

$$a_m = \frac{1}{2\pi} \int_0^{2\pi} f(\phi) e^{-im\phi} \mathrm{d}\phi. \tag{2.26}$$

À condition que $f(\phi)$ satisfasse quelques conditions de base sur la continuité et l'aspect lisse (ce qui est normalement le cas pour la plupart des problèmes physiques), il peut être montré que des expansions comme l'Eq 2.25 convergent monotoniquement dans le sens que

$$\lim_{N\to\infty} \left| f(\phi) - \sum_{m=0}^{N} a_m e^{im\phi} \right| = 0. \tag{2.27}$$

En conséquence, pour des raisons pratiques, on peut souvent représenter une fonction compliquée avec une exactitude suffisante en faisant un choix approprié pour l'ordre d'expansion, *N*.

Concernant la notation, il est souvent utile d'utiliser des fonctions de base orthogonales normalisées, ou *orthonormées,* comme

$$\psi_m(\phi) = \frac{1}{\sqrt{2\pi}} e^{im\phi} \mathrm{d}\phi \tag{2.28}$$

de sorte que la propriété d'orthogonalité puisse être écrite de façon concise comme

$$\int_0^{2\pi} \psi_m(\phi) \psi_j(\phi)^* = \delta_{mj} \tag{2.29}$$

où $\psi_j(\phi)^*$ dénote le conjugué du complexe $\psi_j(\phi)$ (c.-à-d. changer i en $-i$). Avec cette convention, l'expansion de Fourier d'une certaine fonction, $f(\phi)$, devient

$$f(\phi) = \sum_{m=0}^{\infty} a_m \psi_m(\phi) \tag{2.30}$$

et les coefficients d'expansion sont déterminés en utilisant

$$a_m = \int_0^{2\pi} f(\phi) \psi_m(\phi)^* \mathrm{d}\phi. \tag{2.31}$$

En algèbre, l'action de multiplier les deux côtés de l'équation par une certaine fonction et d'intégrer est souvent appelée une *transformée intégrale.* En travaillant avec des espaces de Hilbert ou des fonctions spéciales (Debnath & Mikusinski, 1999), il arrive souvent qu'une application appropriée d'une transformée intégrale puisse être une façon utile de procéder (Debnath & Bhatta, 2007).

2.1.5 La fonction gamma et les factorielles associées

La fonction gamma d'Euler, $\Gamma(w)$, peut être définie par l'intégrale

$$\int_0^\infty e^{-t}t^w \mathrm{d}t = \Gamma(w+1). \tag{2.32}$$

La fonction gamma peut être considérée comme une fonction factorielle généralisée dans le sens que

$$\begin{aligned}\Gamma(w+1) &= w\Gamma(w) \\ &= w(w-1)\Gamma(w-1) \\ &= w(w-1)(w-2)\Gamma(w-2) \\ &\text{... etc.}\end{aligned} \tag{2.33}$$

Quand $w = 0$, on peut voir à partir de l'Eq 2.32 que $\Gamma(0) = 1$. Quand w est un nombre entier réel, la fonction gamma se réduit à une factorielle ordinaire

$$\begin{aligned}\Gamma(n+1) &= n(n-1)(n-2)...\Gamma(0) \\ &= n!\end{aligned} \tag{2.34}$$

Pour le cas spécial de $w = 1/2$, l'Eq 2.32 peut être utilisée une fois encore pour montrer que

$$\Gamma(1/2) = \sqrt{\pi}. \tag{2.35}$$

Pour d'autres valeurs de w, la fonction gamma peut être estimée ou même calculée exactement en utilisant la formule de Lanczos (1964). Ici, ce qui nous interesse surtout c'est d'évaluer des fonctions gamma ou des produits de fonctions gamma ayant un argument entier ou demi-entier jusqu'à environ w=128. Par conséquent, quelques précautions sont requises afin d'éviter le dépassement numérique et de préserver une précision arithmétique élevée. Il est donc approprié d'introduire quelques notations supplémentaires qui faciliteront l'annulation des facteurs communs dans des expressions complexes. Spécifiquement, la factorielle descendante, $[w]_k$, peut être définie comme

$$[w]_k = w(w-1)(w-2)...(w-k+1). \tag{2.36}$$

De même, la factorielle ascendante, $(w)_k$, peut être définie comme

$$(w)_k = w(w+1)(w+2)...(w+k-1). \tag{2.37}$$

Pour des arguments de nombre entier, ces factorielles deviennent

$$[n]_k = n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!} \tag{2.38}$$

et

$$(n)_k = n(n+1)(n+2)...(n+k-1) = \frac{(n+k-1)!}{(n-1)!}. \tag{2.39}$$

En particulier, $\Gamma(n+1/2)$ peut être calculée en utilisant l'identité

$$\Gamma(n+1/2) = \sqrt{\pi}(1/2)_n \tag{2.40}$$

et le fait que

$$\begin{aligned}(1/2)_n &= \left(\frac{1}{2}\right)\left(\frac{3}{2}\right)\left(\frac{5}{2}\right)...\left(\frac{2n-1}{2}\right) \\ &= (1.3.5.7...(2n-1))\left(\frac{1}{2}\right)^n.\end{aligned} \tag{2.41}$$

Parfois il est aussi utile de définir un coefficient binomial comme

$$\binom{\alpha}{m} = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1-m)m!}. \tag{2.42}$$

2.1.6 La simplification symbolique de factorielles

Plusieurs des expressions utilisées ici impliquent la multiplication et la division des factorielles d'ordre relativement supérieur. Cependant, puisque ces expressions contiennent souvent plusieurs termes qui peuvent être annulés, il est ainsi utile d'appliquer une méthode symbolique afin de les simplifier le plus possible avant d'exécuter une quelconque arithmétique. Par exemple, un coefficient binomial peut être calculé comme

$$\binom{n}{m} = \frac{n!}{(n-m)!m!} = \frac{(0^1 \times 1^1 \times 2^1 ... n^1)}{(0^1 \times 1^1 \times 2^1 ... (n-m)^1)(0^1 \times 1^1 \times 2^1 ... m^1)}. \tag{2.43}$$

Évidemment, des combinaisons arbitraires des factorielles peuvent être accumulées en ajoutant et en soustrayant des puissances de nombres entiers des facteurs dans le numérateur et le dénominateur respectivement. Par exemple,

$$\binom{6}{2} = 0^{-1} \times 1^{-1} \times 2^{-1} \times 3^0 \times 4^0 \times 5^1 \times 6^1 = \frac{5 \times 6}{2}. \tag{2.44}$$

En outre, une telle expression peut être symboliquement réduite en produits de puissances de nombres premiers. Par exemple,

$$\frac{5 \times 6}{2} = 2^0 \times 3^1 \times 5^1. \tag{2.45}$$

Un petit ensemble de fonctions utiles a été implémenté dans le langage C pour effectuer automatiquement de telles manipulations pour les factorielles de nombres entiers et les puissances ci-dessus, et pour exécuter toute arithmétique restante en utilisant la bibliothèque mathématique de précision élevée GMP.[2] Par exemple, en utilisant ces utilitaires, une expression comme

$$x = \frac{n!}{\Gamma(n+1/2)} \tag{2.46}$$

peut être évaluée en utilisant le fragment de code C donné dans la figure 2.2.

[2] http://gmplib.org/.

```
#include <math.h>
double x;
sp_init();                              // initialise working memory
sp_fac(n, +1);                          // set numerator = n!
sp_rise2(n, -1);                        // set divisor   = (1/2)_n
x = sp_ans() / sqrt(M_PI);              // simplify and supply result
```

Figure 2.2: Un exemple de programmation C pour illustrer la simplification symbolique et l'évaluation des expressions impliquant des produits de factorielles.

2.1.7 Les polynômes de Jacobi

Les polynômes de Jacobi, $P_k^{(\alpha,\beta)}(x)$, peuvent être définis comme (Erdélyi *et al.*, 1953a)

$$P_k^{(\alpha,\beta)}(x) = \frac{1}{2^k}\sum_{j=0}^{k}\binom{k+\alpha}{j}\binom{k+\beta}{k-j}(x+1)^j(x-1)^{k-j}. \tag{2.47}$$

Dans quelques applications, il est approprié d'utiliser l'expansion décalée (Keister & Polyzou, 1997)

$$P_k^{(\alpha,\beta)}(x) = \frac{\Gamma(k+\alpha+1)}{k!\Gamma(k+\lambda)}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{\Gamma(k+j+\lambda)}{\Gamma(j+\alpha+1)}\left(\frac{1-x}{2}\right)^j, \tag{2.48}$$

où $\lambda = \alpha+\beta+1$. L'expansion inverse est donnée par (Erdélyi *et al.*, 1953b)

$$(1-x)^k = 2^k k!\Gamma(k+\alpha+1)\sum_{j=0}^{k}(-1)^j\frac{(2j+\lambda)\Gamma(j+\lambda)}{(k-j)!\Gamma(k+j+\lambda+1)\Gamma(j+\alpha+1)}P_j^{(\alpha,\beta)}(x). \tag{2.49}$$

Les polynômes de Jacobi sont orthogonaux par rapport à un facteur de pondération $(1-x)^\alpha(1+x)^\beta$ dans le sens que

$$\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta P_k^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(x)\mathrm{d}x = \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{k!\Gamma(k+\lambda)}\frac{2^\lambda}{2k+\lambda}\delta_{kn}. \tag{2.50}$$

Lorsque $k \geq 2$, les polynômes de Jacobi peuvent être calculés via la formule stable de récursivité

$$\begin{aligned}2(k+1)(k+\alpha+\beta+1)(2k+\alpha+\beta)P_{k+1}^{(\alpha,\beta)}(x) = \\ (2k+\alpha+\beta+1)\big[(2k+\alpha+\beta)(2k+\alpha+\beta+2)x+\alpha^2-\beta^2\big]P_k^{(\alpha,\beta)}(x) - \\ 2(k+\alpha)(k+\beta)(2k+\alpha+\beta+2)P_{k-1}^{(\alpha,\beta)}(x).\end{aligned} \tag{2.51}$$

Les polynômes de Jacobi sont le type de fonction orthogonale le plus général dans l'intervalle $[-1,1]$. Les polynômes de Gegenbauer (ou ultra-sphérique) apparaissent quand $\alpha = \beta$. Les polynômes de Legendre apparaissent quand $\alpha = \beta = 0$ et les polynômes de Chebychev correspondent au cas spécial de $\alpha = \beta = -1/2$.

2.1.8 Les polynômes de Legendre

Les polynômes de Legendre, $P_l(\mu)$, de l'ordre $l \geq 0$ peuvent être définis par la formule de Rodrigues

$$P_l(\mu) = \frac{1}{2^l l!} \frac{\mathrm{d}^l}{\mathrm{d}\mu^l} (\mu^2 - 1)^l. \tag{2.52}$$

Plus généralement, les polynômes associés de Legendre, $P_{lm}(\mu)$, de l'ordre l et de degré $m \leq l$ peuvent être définis comme

$$P_{lm}(\mu) = (-1)^m \frac{(1-\mu^2)^{m/2}}{2^l l!} \frac{\mathrm{d}^{l+m}}{\mathrm{d}\mu^{l+m}} (\mu^2 - 1)^l. \tag{2.53}$$

Dorénavant, le terme "polynôme de Legendre" sera pris pour signifier le polynôme général, $P_{lm}(\mu)$, où $|m| \leq l$. Les polynômes de Legendre dans lesquels m est négatif peuvent être calculés en utilisant l'identité (Hobson, 1931)

$$P_{l\overline{m}}(\mu) = (-1)^m P_{l|m|}(\mu). \tag{2.54}$$

Le domaine naturel des polynômes de Legendre est $-1 \leq \mu < 1$, ou avec le changement de variable, $\mu = \cos\theta$, le domaine devient $0 \leq \theta < \pi$. Ainsi ces polynômes peuvent également être définis comme

$$P_{lm}(\theta) = (-1)^m \frac{(\sin\theta)^m}{2^l l!} \frac{\mathrm{d}^{l+m}}{\mathrm{d}\mu^{l+m}} (\cos\theta)^l. \tag{2.55}$$

Les polynômes de Legendre sont orthogonaux dans le sens que

$$\int_{-1}^{1} P_{km}(\mu) P_{lm}(\mu) \mathrm{d}\mu = \frac{2}{(2l+1)} \frac{(l+m)!}{(l-m)!} \delta_{kl}. \tag{2.56}$$

Une expression explicite de série entière pour les polynômes de Legendre peut être obtenue en développant $(\mu^2 - 1)^l$ comme une série binomiale et en la différenciant $l + m$ fois :

$$P_{lm}(\mu) = (1-\mu^2)^{m/2} \sum_{k=\frac{l+m+1}{2}}^{l} \frac{(-1)^{k+l+m}}{2^l l!} \binom{l}{k} \frac{(2k)!}{(2k-l-m)!} \mu^{2k-l-m}, \tag{2.57}$$

où la sommation borne inférieure est prise en utilisant la troncature d'entier.

Souvent, les polynômes de Legendre sont calculés en utilisant des relations de récursivité. Par exemple, la formule conventionnelle de récursivité, modifiée pour inclure le facteur de phase de Condon-Shortley (Condon & Odabasi, 1980), commence avec

$$P_{ll}(\theta) = (-1)^l \frac{(2l)!}{l!} \left(\frac{\sin\theta}{2} \right)^l \tag{2.58}$$

et continue jusqu'à $m = 0$ avec (Hobson, 1931)

$$P_{lm}(\theta) = -2(m+1)\cot(\theta) P_{l,m+1}(\theta) - \frac{P_{l,m+2}(\theta)}{(l-m)(l+m+1)}, \tag{2.59}$$

où $P_{lm}(\theta) = 0$ quand $m > l$. Ceci a une stabilité numérique très bonne (Wiggins & Saito, 1971; Libbrecht, 1985).

2.1.9 Les harmoniques sphériques

Les fonctions régulières de SH solide, $Y_{lm}(\underline{r})$, sont généralement représentées comme des fonctions complexes des coordonnées polaires sphériques (Hobson, 1931) :

$$Y_{lm}(\underline{r}) = r^l Y_{lm}(\theta, \phi), \tag{2.60}$$

où $-l \leq m \leq +l$. Ici, nous sommes surtout intéressé par les fonctions harmoniques de surface, $Y_{lm}(\theta, \phi)$, parfois appelées harmoniques tessérales, obtenues en mettant $r = 1$. Les SH sont séparables

$$Y_{lm}(\theta, \phi) = \vartheta_{lm}(\theta)\psi_m(\phi), \tag{2.61}$$

où $\vartheta_{lm}(\theta)$ sont les polynômes normalisés de Legendre

$$\vartheta_{lm}(\theta) = \left[\frac{(2l+1)}{2}\frac{(l-m)!}{(l+m)!}\right]^{1/2} P_{lm}(\cos\theta), \tag{2.62}$$

et où $\psi_m(\phi)$ sont les fonctions circulaires normalisées

$$\psi_m(\phi) = \frac{1}{\sqrt{2\pi}} e^{im\phi}. \tag{2.63}$$

Grâce à l'orthogonalité des fonctions circulaires et de Legendre, les SH sont orthonormées dans le sens que

$$\int_0^{2\pi}\int_0^{\pi} Y_{lm}(\theta, \phi) Y_{l'm'}(\theta, \phi)^* \sin\theta \mathrm{d}\theta \mathrm{d}\phi = \delta_{ll'}\delta_{mm'}. \tag{2.64}$$

Les SH sont importantes dans de nombreux domaines de physique parce qu'elles semblent être des solutions pour l'équation de Laplace. Beaucoup de phénomènes naturels tels que par exemple la gravitation, l'électrostatique et l'écoulement de fluide et de chaleur peuvent être décrits par des potentiels conservatifs (c.-à-d. ceux qui dépendent seulement de la position spatiale d'une particule). L'équation de Laplace déclare que, pour un potentiel conservatif, $\psi(\underline{r})$:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\psi(\underline{r}) = 0. \tag{2.65}$$

Ceci est souvent écrit en utilisant l'opérateur "laplacien," ∇^2 :

$$\nabla^2\psi(\underline{r}) = 0. \tag{2.66}$$

En utilisant des techniques standard de calcul pour changer des variables, le laplacien peut être écrit en coordonnées polaires comme :

$$\nabla^2 = \frac{1}{r}\left(\frac{\partial^2}{\partial r^2}\right)r + \frac{1}{r^2}\Lambda^2 \tag{2.67}$$

où Λ^2 est l'opérateur "legendrien" :

$$\Lambda^2 = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}. \tag{2.68}$$

La partie angulaire de l'équation de Laplace a comme solutions :

$$\Lambda^2 Y_{lm}(\theta,\phi) = -l(l+1)Y_{lm}(\theta,\phi). \tag{2.69}$$

En d'autres termes, les SH sont des fonctions propres de l'opérateur legendrien. On trouve alors que la partie radiale de l'équation de Laplace donne deux solutions de la forme r^l et $r^{-(l+1)}$, qui correspondent aux harmoniques solides régulières et irrégulières respectivement.

Les SH solides régulières peuvent être écrites en termes de coordonnées cartésiennes en substituant l'identité

$$z^m = (r\cos\theta)^m \tag{2.70}$$

dans la définition de séries entières des polynômes normalisés de Legendre, Eqs 2.56 et 2.57, et en notant d'une façon semblable que

$$\begin{aligned}(x+iy)^m &= (r\sin\theta(\cos\phi + i\sin\phi))^m \\ &= (r\sin\theta)^m e^{im\phi},\end{aligned} \tag{2.71}$$

pour obtenir

$$Y_{lm}(\underline{x}) = \left[\frac{2l+1}{4\pi}\frac{(l-m)!}{(l+m)!}\right]^{1/2} \sum_{k=\frac{l+m+1}{2}}^{l} \binom{l}{k} \frac{(-1)^{k+l+m}(2k)!}{2^l l!(2k-l-m)!}\frac{(x+iy)^m z^{2k-l-m}}{r^{2k}}. \tag{2.72}$$

2.1.10 Les coefficients d'accouplement des harmoniques sphériques

Pour calculer certaines intégrales impliquant des fonctions SH, il est souvent utile de développer un produit des SH comme une combinaison linéaire (Biedenharn & Louck, 1981)

$$Y_{lm}(\theta,\phi)Y_{l'm'}(\theta,\phi) = \sum_{kj}\left[\frac{(2l+1)(2l'+1)}{4\pi(2k+1)}\right]^{1/2} C_{000}^{l\,l'\,k} C_{mm'\,j}^{l\,l'\,k} Y_{kj}(\theta,\phi), \tag{2.73}$$

où $C_{mm'\,j}^{l\,l'\,k}$ est le coefficient d'accouplement de Clebsh-Gordan. La formule de Wigner pour le coefficient de Clebsh-Gordan est donnée par

$$\begin{aligned}C_{m_1\,m_2\,m}^{j_1\,j_2\,j} = \delta_{m_1+m_2,m}&\left[(2j+1)\frac{(j+j_1-j_2)!(j-j_1+j_2)!(j_1+j_2-j)!}{(j_1+j_2+j+1)!}\right]^{1/2}\times \\ &\left[\frac{(j+m)!(j-m)!}{(j_1+m_1)!(j_1-m_1)!(j_2+m_2)!(j_2-m_2)!}\right]^{1/2}\times \\ &\sum_k (-1)^{j_2+m_2+k}\frac{(j_2+j+m_1-k)!(j_1-m_1+k)!}{(j-j_1+j_2-k)!(j+m-k)!(j_1-j_2-m+k)!k!},\end{aligned} \tag{2.74}$$

où l'addition s'étend sur toutes les valeurs de k dont les factorielles sont bien définies. Lorsque $m_1 = m_2 = m = 0$, le coefficient d'accouplement disparaît à moins que $j_1 + j_2 + j$ soit un chiffre pair; dans ce cas, l'expression peut être réduite à

$$C^{j_1\, j_2\, j}_{0\, 0\, 0} = \left[(2j+1)\frac{(j_1+j_2-j)!(j_1-j_2+j)!(j_2-j_1+j)!}{(j_1+j_2+j+1)!}\right]^{1/2} \frac{(-1)^{l-j}l!}{(l-j_1)!(l-j_2)!(l-j)!}, \tag{2.75}$$

où $l = (j_1 + j_2 + j)/2$. À partir des symétries de permutation des coefficients de Clebsch-Gordan,

$$\begin{aligned} C^{j_1\, j_2\, j}_{m_1\, m_2\, m} &= (-1)^{j_1+j_2-j} C^{j_2\, j_1\, j}_{m_2\, m_1\, m} \\ &= (-1)^{j_1+j_2-j} C^{j_1\, j_2\, j}_{\overline{m}_1\, \overline{m}_2\, \overline{m}}, \end{aligned} \tag{2.76}$$

il s'ensuit que

$$C^{j_1\, j_2\, j}_{m\, \overline{m}\, 0} = (-1)^{j_1-j_2-j} C^{j_1\, j_2\, j}_{\overline{m}\, m\, 0} \tag{2.77}$$

et

$$C^{j_1\, j_2\, j}_{m\, \overline{m}\, 0} = C^{j_2\, j_1\, j}_{m\, \overline{m}\, 0}. \tag{2.78}$$

Le symbole *3-j* de Wigner, qui apparaîtra dans le prochain chapitre, est étroitement lié aux coefficients de Clebsch-Gordan et est donné par

$$\begin{pmatrix} j_1 & j_2 & j \\ m_1 & m_2 & \overline{m} \end{pmatrix} = \frac{(-1)^{m+j_1-j_2}}{\sqrt{2j+1}} C^{j_1\, j_2\, j}_{m_1\, m_2\, m}. \tag{2.79}$$

Des symboles *3-j* de l'ordre supérieur peuvent être calculés efficacement par récursivité et en utilisant la formule de Wigner seulement pour initialiser la récursivité. Par exemple, les relations de récurrence de Sakurai (1994) pour les j_1, j_2 et j fixes sont

$$\begin{aligned} \big[(j+m)(j-m+1)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1, m_2, m-1} =& \big[(j_1+m_1+1)(j_1-m_1)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1+1, m_2, m} \\ &- \big[(j_2+m_2+1)(j_2-m_2)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1, m_2+1, m} \end{aligned} \tag{2.80}$$

et

$$\begin{aligned} \big[(j-m)(j+m+1)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1, m_2, m+1} =& \big[(j_1-m_1+1)(j_1+m_1)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1-1, m_2, m} \\ &- \big[(j_2-m_2+1)(j_2+m_2)\big]^{1/2} C^{j_1\, j_2\, j}_{m_1, m_2-1, m}. \end{aligned} \tag{2.81}$$

Les coefficients nécessaires $C^{j_1\, j_2\, j}_{m\, \overline{m}\, 0}$ peuvent être calculés en évaluant alternativement les équations 2.80 et 2.81 dans un chemin en "zig-zag" dans le plan m_1/m_2 (Sakurai, 1994). Puisque ces formules de récursivité ne sont pas particulièrement stables, des erreurs numériques peuvent être réduites en utilisant les deux types de récursivité : ascendante de $C^{j_1\, j_2\, j}_{000}$, et descendante de $C^{j_1\, j_2\, j}_{m, \overline{m}, 0}$. L'exactitude

du calcul peut être évaluée à partir de diverses relations d'orthogonalité des coefficients de Clebsch-Gordan (Biedenharn & Louck, 1981). Bien qu'une analyse numérique détaillée n'ait pas été réalisée, je trouve que l'arrondissement des erreurs devient significatif une fois que j_1 et j_2 atteignent environ 20, même si l'on utilise l'arithmétique de quadruple précision de matériel. Pour de grands nombres quantiques, une approche réussie consiste à utiliser la formule de Wigner (Eq 2.74) et à réduire symboliquement toutes les factorielles aux produits de nombres premiers (voir la section 2.1.6) avant d'accomplir le calcul dans l'arithmétique de 256 bits en utilisant la bibliothèque GMP.

2.1.11 Les harmoniques sphériques réelles

Les SH réelles, $y_{lm}(\theta, \phi)$, peuvent être trouvées en faisant des combinaisons linéaires des fonctions complexes suivantes

$$y_{lm}(\theta,\phi) = \begin{cases} (Y_{lm}(\theta,\phi) + Y_{lm}(\theta,\phi)^*)/\sqrt{2} & \text{si } m > 0 \\ Y_{l0}(\theta,\phi) & \text{si } m = 0 \\ -i(Y_{l\overline{m}}(\theta,\phi) - Y_{l\overline{m}}(\theta,\phi)^*)/\sqrt{2} & \text{si } m < 0. \end{cases} \tag{2.82}$$

Par conséquent les SH réelles sont aussi des fonctions propres de l'équation de Laplace. En développant les divers termes et en utilisant l'Eq 2.54, ceci donne

$$y_{lm}(\theta,\phi) = \begin{cases} \vartheta_{lm}(\theta)(\cos m\phi)/\sqrt{\pi} & \text{si } m > 0 \\ \vartheta_{lm}(\theta)/\sqrt{2\pi} & \text{si } m = 0 \\ \vartheta_{l\overline{m}}(\theta)(\sin \overline{m}\phi)/\sqrt{\pi} & \text{si } m < 0, \text{ c.-à-d. } \overline{m} > 0. \end{cases} \tag{2.83}$$

Ainsi les fonctions SH réelles peuvent être écrites comme

$$y_{lm}(\theta,\phi) = \vartheta_{l|m|}(\theta)\varphi_m(\phi) \tag{2.84}$$

où

$$\varphi_m(\phi) = \begin{cases} \cos m\phi/\sqrt{\pi} & \text{si } m > 0 \\ 1/\sqrt{2\pi} & \text{si } m = 0 \\ \sin \overline{m}\phi/\sqrt{\pi} & \text{si } m < 0. \end{cases} \tag{2.85}$$

La figure 2.3 montre les formes des SH réelles jusqu'à l=2.

Il est souvent considérablement plus efficace de représenter des quantités réelles en utilisant les SH réelles parce que ceci permet d'éviter toute arithmétique complexe. Cependant, comme nous l'avons montré dans le chapitre 4, lorsqu'on utilise des FFT pour accélérer les calculs, il est aussi utile de pouvoir commuter entre les bases complexes et réelles. Par conséquent, il est utile d'écrire des combinaisons linéaires comme l'Eq 2.82 sous la forme matricielle de

$$y_{lm}(\theta,\phi) = \sum_{m'=-l}^{l} U^{(l)}_{mm'} Y_{lm}(\theta,\phi), \tag{2.86}$$

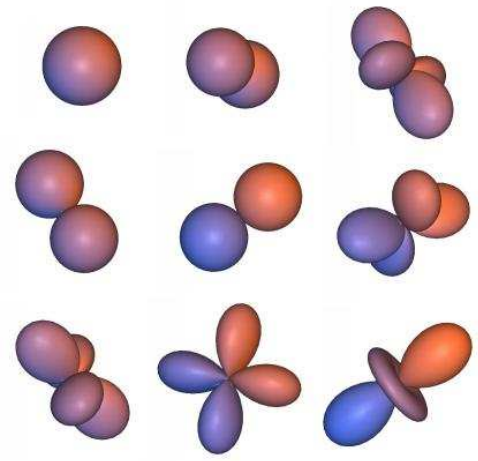

Figure 2.3: Les formes des fonctions SH réelles jusqu'à l'ordre l=2. Commençant à partir de la sphère, y_{00}, en haut à gauche de cette figure, les trois fonctions adjacentes ont l=1 et les cinq fonctions dans la rangée et la colonne finales ont l=2. Les trois fonctions sur la diagonale principale ont m=0; les fonctions avec des valeurs positives de m sont montrées au-dessus de la diagonale principale et celles avec des indices négatifs m sont montrées au-dessous de la diagonale principale.

où $U^{(l)}$ est une matrice unitaire (c.-à-d. le transposé conjugué complexe de $U^{(l)}$ est la matrice inverse; Biedenharn et Louck, 1981). En notant que tous les éléments non-diagonaux de $U^{(l)}$ sont zéro, le changement des équations de base peut être montré plus explicitement sous la forme matricielle de

$$\begin{pmatrix} y_{ll} \\ y_{lm} \\ y_{l0} \\ y_{l\overline{m}} \\ y_{l\bar{l}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{(-1)^l}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 & \frac{(-1)^m}{\sqrt{2}} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{i(-1)^m}{\sqrt{2}} & 0 & \frac{-i}{\sqrt{2}} & 0 \\ \frac{i(-1)^l}{\sqrt{2}} & 0 & 0 & 0 & \frac{-i}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} Y_{ll} \\ Y_{lm} \\ Y_{l0} \\ Y_{l\overline{m}} \\ Y_{l\bar{l}} \end{pmatrix}. \tag{2.87}$$

2.1.12 Les polynômes de Laguerre

Les polynômes généralisés de Laguerre, $L_k^{(\alpha)}(t)$, peuvent être définis par la formule de Rodrigue

$$L_k^{(\alpha)}(t) = \frac{t^{-\alpha}e^t}{k!}\frac{\mathrm{d}^k}{\mathrm{d}t^k}\left(e^{-t}e^{k+\alpha}\right). \tag{2.88}$$

Une expansion binomiale équivalente est donnée par (Erdélyi *et al.*, 1953b)

$$L_k^{(\alpha)}(t) = \sum_{j=0}^{k}\binom{k+\alpha}{k-j}\frac{(-t)^j}{j!}. \tag{2.89}$$

Les polynômes de Laguerre ont des racines positives réelles dans l'intervalle $[0, k+\alpha+(k-1)\sqrt{(k+\alpha)}]$. Des polynômes d'ordre supérieur peuvent être calculés efficacement en utilisant la

récursivité stable

$$(k+1)L_{k+1}^{(\alpha)}(t) = (2k+\alpha+1-t)L_k^{(\alpha)}(t) - (k+\alpha)L_{k-1}^{(\alpha)}(t), \tag{2.90}$$

avec les identités

$$L_0^{(\alpha)}(t) = 1 \tag{2.91}$$

et

$$L_1^{(\alpha)}(t) = \alpha + 1 - t. \tag{2.92}$$

Les polynômes de Laguerre sont orthogonaux par rapport à un facteur de pondération, $e^{-t}t^{\alpha}$, dans le sens que

$$\int_0^{\infty} e^{-t}t^{\alpha}L_k^{(\alpha)}(t)L_{k'}^{(\alpha)}(t)\mathrm{d}t = \frac{\Gamma(k+\alpha+1)}{k!}\delta_{kk'}. \tag{2.93}$$

2.1.13 Les fonctions radiales de base de GTO et d'ETO

Bien que les harmoniques solides soient des candidates normales pour représenter des potentiels conservatifs lisses, afin de pouvoir représenter des formes 3D moléculaires arbitraires, il serait souhaitable d'utiliser des fonctions radiales telles que les polynômes de Laguerre qui présentent des noeuds radials ou des zéros semblables aux zéros angulaires des SH. En prenant en compte la condition d'orthogonalité pour les polynômes de Laguerre dans la section précédente, il semble donc raisonnable de considérer des fonctions radiales de la forme :

$$R_k^{(\alpha)}(r) = N_k^{(\alpha)}e^{-t/2}t^{\alpha/2}L_k^{(\alpha)}(t) \tag{2.94}$$

où $N_k^{(\alpha)}$ est un facteur de normalisation et où t est une distance radiale bien proportionnée. Puisque l'élément de volume 3D est donné par

$$\mathrm{d}V = r^2\mathrm{d}r\sin\theta\mathrm{d}\theta\mathrm{d}\phi, \tag{2.95}$$

la façon dans laquelle la distance radiale r est proportionnée sur le paramètre formel t fixera alors la valeur de α. En choisissant expressément un facteur de balance λ et en faisant un changement de variable

$$t = r^2/\lambda \tag{2.96}$$

on obtient

$$r^2\mathrm{d}r = \frac{\lambda^{3/2}}{2}t^{1/2}\mathrm{d}t \tag{2.97}$$

et ceci implique de mettre $\alpha = l+1/2$ afin de maintenir la forme d'orthogonalité, Eq 2.93. Puisque ces fonctions seront utilisées avec les SH, l'affectation $k = n - l - 1$ est faite ci-dessous pour s'assurer que les produits de telles fonctions radiales avec les SH énuméreront des combinaisons distinctes de produits de puissances de $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$ et $z = r\cos\theta$. Sans donner une preuve formelle, ceci assurera que l'ensemble final de fonctions orthogonales de base sera *complet*. Quelques travaux additionnels donnent alors

$$R_{nl}(r) = \left[\frac{2}{\lambda^{3/2}}\frac{(n-l-1)!}{\Gamma(n+1/2)}\right]^{1/2} e^{-\rho^2/2}\rho^l L_{n-l-1}^{(l+1/2)}(\rho^2), \tag{2.98}$$

où maintenant $\rho^2 = r^2/\lambda$.

Ces fonctions de Gauss-Laguerre (GL) sont orthonormées dans le sens que

$$\int_0^\infty R_{nl}(r)R_{n'l}(r)r^2\mathrm{d}r = \delta_{nn'}. \tag{2.99}$$

Puisque ces fonctions ont un pré-facteur gaussien, les fonctions de base 3D GL et SH sont souvent appelées les orbitales de type gaussien (GTO) dans la littérature de la chimie quantique. Le facteur gaussien assure que ces fonctions tendent vers zéro pour de grandes valeurs de paramètre radiales. En effet, de la section 2.1.12, les zéros de ces fonctions sont liés dans la marge $[0, n-1/2+(n-l-2)\sqrt{(n-1/2)}]$. Pour l'amarrage des domaines typiques de protéine, utiliser un facteur de balance $\lambda = 20$ donne de bons résultats pour des domaines globulaires avec un rayon moyen allant jusqu'à environ 30 Å (Ritchie, 1998). La figure 2.4 montre les formes de quelques fonctions de base de GTO.

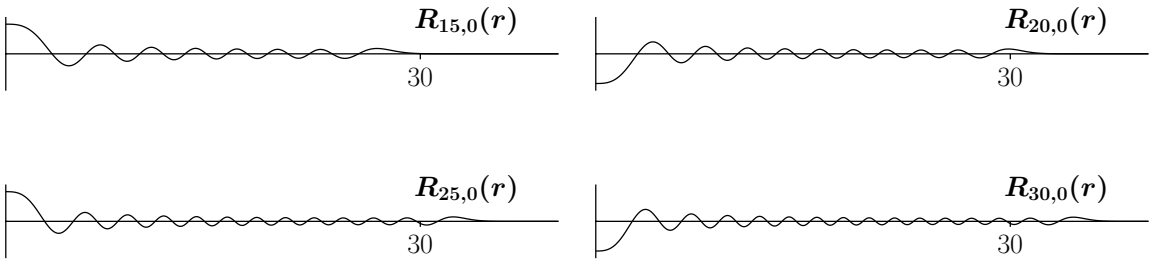

Figure 2.4: Quelques exemples des formes de fonctions radiales de base de GTO, proportionées à 30 Å.

Il est aussi possible de faire un changement linéaire de variable

$$t = \rho = \Lambda r, \tag{2.100}$$

avec un facteur de balance Λ. Ceci implique de mettre $\alpha = 2l + 2$ pour satisfaire la condition d'orthogonalité. Quelques travaux supplémentaires donnent alors les fonctions orthonormées

$$S_{nl}(r) = \left[(2\Lambda)^3\frac{(n-l-1)!}{(n+l+1)!}\right]^{1/2} e^{-\rho/2}\rho^l L_{n-l-1}^{(2l+2)}(\rho). \tag{2.101}$$

En chimie quantique, ces fonctions radiales sont souvent appelées orbitales de type exponentiel (ETO). J'utilise des ETO pour représenter les propriétés électrostatiques des protéines.

J'aimerais mentionner que, pendant mon travail pour ma thèse de doctorat, quand j'étudiais les fonctions radiales pour en choisir une appropriée (Ritchie, 1998), j'ai été inspiré par l'équation de Schrödinger pour l'atome d'hydrogène qui a la forme

$$\psi_{nlm}(\underline{r}) = N_{nl} e^{-\rho/2} \rho^l L_{n-l-1}^{(2l+1)}(\rho) Y_{lm}(\theta, \phi), \tag{2.102}$$

où N_{nl} est un facteur de normalisation et ρ est une distance proportionnée. Dans l'équation de Schrödinger, l'index n correspond au nombre quantique principal de l'atome d'hydrogène qui, par convention, se compte à partir de l'unité. J'ai suivi la même convention de numérotation dans mes publications précédentes. Ceci tend à rendre les formules ci-dessous un peu plus complexes que nécessaires, mais je continue à l'utiliser pour être cohérent.

2.1.14 Les fonctions de Bessel

Finalement, les fonctions de Bessel sont introduites ici parce qu'elles donnent une manière analytique de décrire le rapport entre les différents systèmes de coordonnées en utilisant des coordonnées polaires. En d'autres termes, elles donnent les mécanismes analytiques nécessaires pour calculer des translations des expansions polaires de Fourier. La fonction générale de Bessel, $J_\nu(w)$, de degré ν et d'argument complexe w peut être définie comme (Hochstadt, 1971)

$$J_\nu(w) = \left(\frac{w}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k (w/2)^{2k}}{\Gamma(\nu + k + 1) k!}. \tag{2.103}$$

La fonction *sphérique de Bessel*, $j_l(w)$, de degré entier l est liée à $J_\nu(w)$ par

$$j_l(w) = \sqrt{\frac{\pi}{2w}} J_{l+1/2}(w). \tag{2.104}$$

En utilisant

$$(\alpha)_n = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)}, \tag{2.105}$$

dans l'Eq 2.103, il est simple de montrer que $j_l(w)$ peut être calculée comme

$$j_l(w) = \frac{1}{2} \sum_{k=0}^{\infty} C_k^{(l)} \left(\frac{w}{2}\right)^{2k+l} \tag{2.106}$$

dans laquelle les coefficients, $C_k^{(l)}$, sont donnés par

$$C_k^{(l)} = \frac{-2C_{k-1}^{(l)}}{k(2k + 2l + 1)}, \tag{2.107}$$

et où

$$C_0^{(l)} = \frac{1}{(1/2)_{l+1}}. \tag{2.108}$$

Pour $w \leq 2$, l'Eq 2.106 converge rapidement et l'addition peut être terminée une fois qu'on obtient le niveau désiré d'exactitude. Pour de grandes valeurs de w (c.-à-d. jusqu'à environ $w \simeq 100$), les fonctions sphériques de Bessel peuvent être calculées en utilisant la relation de récursivité

$$j_l(w) = \frac{(2l-1)}{w} j_{l-1}(w) - j_{l-2}(w) \tag{2.109}$$

avec

$$j_0(w) = \frac{\sin w}{w} \tag{2.110}$$

et

$$j_1(w) = \frac{\sin w}{w^2} - \frac{\cos w}{w}. \tag{2.111}$$

Ainsi, on voit que les fonctions sphériques de Bessel ont une forme sinusoïdale qui se décompose selon une puissance inverse de la distance à l'origine.

Il peut être montré que les fonctions sphériques de Bessel sont orthogonales dans le sens que (Gottfried, 1966)

$$\int_0^\infty j_l(\beta r) j_l(\beta r')\, \beta^2 \mathrm{d}\beta = \frac{\pi}{2r} \delta(r - r'), \tag{2.112}$$

où

$$\delta(x) = \begin{cases} 1 \text{ si } x = 0, \\ 0 \text{ autrement.} \end{cases} \tag{2.113}$$

Si la transformée sphérique de Bessel d'une fonction $f(r)$ est définie comme

$$\tilde{f}_l(\beta) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(r)\, j_l(\beta r)\, r^2 \mathrm{d}r, \tag{2.114}$$

alors, en utilisant l'Eq 2.112, il peut être montré que la transformée inverse est donnée par

$$f(r) = \sqrt{\frac{2}{\pi}} \int_0^\infty \tilde{f}_l(\beta)\, j_l(\beta r)\, \beta^2 \mathrm{d}\beta. \tag{2.115}$$

Ainsi on voit que la transformée sphérique de Bessel est son propre inverse. De même, il est simple de montrer que la transformée sphérique de Bessel des fonctions orthogonales est elle-même orthogonale. En d'autres termes, si par exemple

$$\int_0^\infty R_{kl}(r) R_{nl}(r) r^2 \mathrm{d}r = \delta_{kn}, \tag{2.116}$$

alors

$$\int_0^\infty \tilde{R}_{kl}(\beta) \tilde{R}_{nl}(\beta) \beta^2 \mathrm{d}\beta = \delta_{kn}. \tag{2.117}$$

Ces propriétés seront évoquées dans la section 3.4 pour développer des expressions analytiques de translation pour des expansions SPF 3D.

2.2 Les représentations de forme-densité 3D des molécules

En mécanique quantique, les molécules sont souvent traitées comme des arrangements fixes de noyaux atomiques entourés de nuages d'électrons. Mathématiquement, ceci peut être représenté comme une superposition des fonctions d'ondes électroniques centrées sur les coordonnées nucléaires, qui définissent ensemble un modèle probabiliste sur la façon dont les électrons sont distribués dans l'espace entier. Cependant, les grandes molécules de protéines se composent normalement de centaines d'acides aminés, de milliers d'atomes et de dizaines de milliers d'électrons. Par conséquent, même avec les plus grands superordinateurs, il est clairement impossible de représenter et de calculer les propriétés de grandes protéines en utilisant des fonctions *ab initio* d'onde électronique. Même les petites molécules organiques se composent souvent de dizaines d'atomes et de centaines d'électrons et il devient très coûteux d'appliquer les techniques *ab initio* à de grands nombres de petites molécules. Par conséquent, des représentations approximatives sont souvent utilisées actuellement pour décrire les petites molécules ainsi que les grandes. Une façon commune et simple de montrer les structures des molécules en utilisant l'infographie est de dessiner simplement chaque atome de la molécule comme une sphère d'un certain rayon. L'utilisateur voit ainsi un espace rempli qui est en fait une union de toutes les sphères atomiques (voir la figure 2.5). Généralement, à chaque type d'atome est assigné un rayon de van der Waals (VDW), calculé à partir des données des assemblages cristallographiques. Cependant, cette représentation "sphère dure" ne modélise pas d'une manière réaliste la nature fondamentalement lisse de la densité moléculaire d'électron, et elle ne fournit pas une méthode facile pour calculer exactement la superficie entière ou le volume d'une molécule, par exemple. Cependant, Grant et Pickup (1995) ont montré qu'en assignant une seule fonction de densité gaussienne à chaque position atomique, on donne une façon remarquablement efficace de décrire la densité globale de matière de petites molécules. Par exemple, en écrivant

$$\rho_i(\underline{r}) = \alpha e^{-\beta(r/r_i)^2}, \tag{2.118}$$

où $\rho_i(\underline{r})$ est la fonction de densité pour le i-ème atome, r_i est son rayon de VDW et où α et β sont des paramètres ajustables, la densité globale de matière d'une molécule de N_A atomes est alors donnée par la somme des densités atomiques

$$\rho(\underline{r}) = \sum_{i=1}^{N_A} \rho_i(\underline{r}). \tag{2.119}$$

De plus, le recouvrement entre les paires de ces fonctions gaussiennes a une forme particulièrement simple (Boys, 1950). Grant et Pickup (1995) ont exploité cette propriété pour développer ROCS ("rapid overlay of chemical structures", ou superposition rapide des structures chimiques), le programme très efficace d'appariement de formes de petites molécules (Grant *et al.*, 1996). Suivant le travail de Grant et Pickup (1995), je mets $\alpha = 2.70$ et $\beta = 2.3442$. Par conséquent, à une distance r_i du centre

d'atome i, la densité prend la valeur constante de 2.7 $e^{-2.3442}$ = 0.259. Ainsi, une bonne estimation de la surface de VDW peut être calculée en additionnant les contributions de densité d'atome à chaque noeud dans une grille 3D et en contournant la grille en utilisant un seuil de densité de ρ = 0.259. La figure 2.5 montre la représentation contournée de densité gaussienne de lorazépam, une petite molécule de médicament. Le contournage est exécuté en utilisant ma propre implémentation de l'algorithme de "marching tetrahedra" (Guéziec & Hummel, 1995).

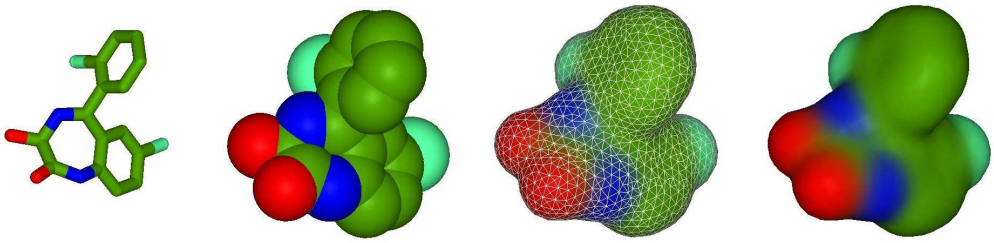

Figure 2.5: Des illustrations d'une petite molécule de médicament, le lorazépam, dessinée en utilisant (de gauche à droite) les "licorice sticks" pour représenter les liaisons covalentes entre les atomes, les sphères de VDW, une surface gaussienne contournée avec les triangles de surface traçés en blanc, et la surface gaussienne contournée sans aucun trace. Dans toutes les représentations, les atomes sont colorés par le type d'atome: carbone en vert; oxygène en rouge; azote en bleu; chlore en bleu-vert. Dans les surfaces gaussiennes, les triangles de surface sont colorés en utilisant une règle de mélange de couleur distance-pondérée qui donne une gradation lisse des nuances à travers la surface.

En plus de calculer une surface lisse de VDW, il est souvent aussi utile de pouvoir calculer la surface accessible au solvant (ou "solvent accessible surface," SAS). Si une molécule sonde sphérique d'un rayon donné, en général 1.4 Å, est roulée au-dessus de la surface de VDW d'une molécule sans jamais la pénétrer, alors la surface balayée par le centre de la sonde définit la SAS. La figure 2.6 montre un croquis de SAS et des surfaces de VDW.

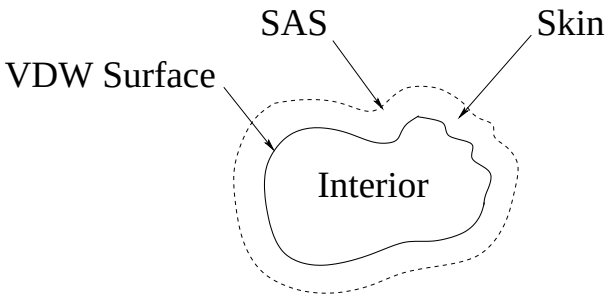

Figure 2.6: Une illustration schématique de la surface moléculaire de VDW, la SAS et les volumes de *"surface skin"* et d'intérieur.

En utilisant l'approche gaussienne atomique pour représenter les volumes moléculaires, la SAS peut être calculée en contournant une surface gaussienne dans laquelle le rayon de chaque atome

est étendu par celle du rayon de sonde. La figure 2.7 montre la SAS pour le lorazépam. Dans le chapitre 4 il est montré que le volume limité par la SAS et les surfaces de VDW, que j'appelle "skin volume," jouent un rôle essentiel pour calculer la complémentarité de forme des protéines pendant les calculs d'amarrage.

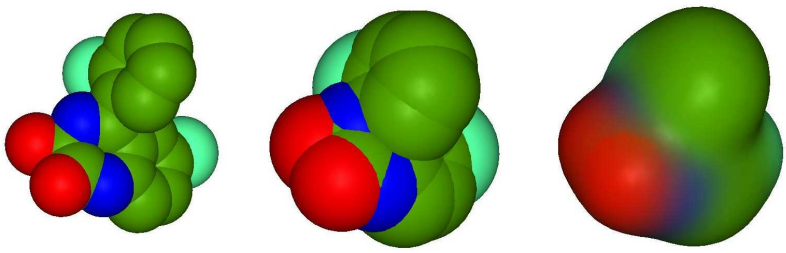

Figure 2.7: Une comparaison (de gauche à droite) de la sphère dure de VDW, de la sphère dure étendue de VDW et des surfaces gaussiennes de SAS de lorazépam.

2.3 Les tessellations icosaèdriques de la sphère

Si l'on considère une grille sphérique conventionnelle, comme celle constituée par les lignes de latitude et de longitude sur une carte du monde, il est clair que les lignes de grille deviennent très concentrées vers les pôles nord et sud. Une manière beaucoup plus juste de diviser la surface d'une sphère est de construire une tessellation sphérique en utilisant les faces d'un icosaèdre régulier, comme illustré dans la figure 2.8. Ici, des tessellations icosaèdriques sont calculées en construisant des triangles sphériques des faces icosaèdriques et en utilisant des courbes géodésiques pour subdiviser la surface de chaque triangle sphérique. Ceci permet à chaque face icosaèdrique d'être divisées en nombre entier de subdivisions. J'utilise la distribution presque régulière des sommets d'une tessellation icosaèdrique pour échantillonner l'espace de rotation de façon homogène et pour donner une méthode rapide d'estimer les intégrales au-dessus de la sphère, qui surgissent dans le calcul des coefficients d'expansion SH de surface moléculaire, comme décrit ci-dessous.

2.4 Les harmoniques sphériques 2D des surfaces moléculaires

Les SH peuvent être utilisées comme des "building blocks" orthogonaux avec lesquels on construit des fonctions paramétrisées par les coordonnées sphériques, (θ, ϕ). Par exemple, la distance radiale d'un domaine globulaire de protéine peut être encodée comme une somme de SH réelles de l'ordre

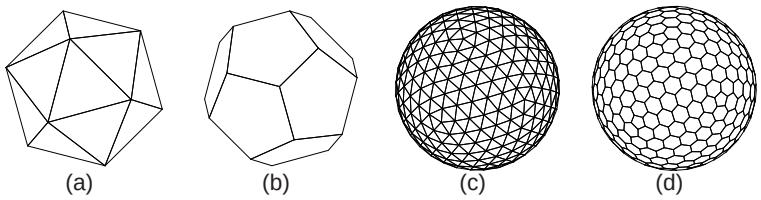

Figure 2.8: Des illustrations d'un icosaèdre (a) et son double, le dodécaèdre (b). La subdivision des patches triangulaires sphériques de l'icosaèdre donne la tessellation montrée dans (c). La tessellation double (d) est obtenue en reliant les centres des faces triangulaires de (c). Dans cet exemple, 6 subdivisions sont faites le long de chaque bord d'icosaèdre pour donner une tessellation icosaèdrique de 362 sommets et de 720 faces. La tessellation double a 720 sommets qui définissent les 12 faces hexagonales et les 350 faces pentagonales.

L en utilisant

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi), \tag{2.120}$$

où a_{lm} sont les coefficients d'expansion. En multipliant chaque côté de l'Eq 2.120 par $y_{kj}(\theta,\phi)$ et en intégrant sur la sphère, ceci donne

$$\int_0^{2\pi} \int_0^{\pi} r(\theta,\phi) y_{kj}(\theta,\phi) \sin\theta d\theta d\phi = a_{lm}\delta_{kl}\delta_{jm}. \tag{2.121}$$

Grâce à l'orthogonalité des fonctions de base, ceci se réduit à

$$a_{lm} = \int_0^{2\pi} \int_0^{\pi} r(\theta,\phi) y_{lm}(\theta,\phi) \sin\theta \mathrm{d}\theta \mathrm{d}\phi. \tag{2.122}$$

En d'autres termes, chaque coefficient est uniquement déterminé par le degré de recouvrement avec sa fonction de base correspondante. En utilisant la tessellation icosaèdrique de la sphère, cette intégrale peut être estimée comme

$$a_{lm} = \sum_{i=1}^{N_V} r(\theta_i,\phi_i) y_{lm}(\theta_i,\phi_i) A_i, \tag{2.123}$$

où la somme s'étend sur les N_V sommets de tessellation, (θ_i,ϕ_i) sont les coordonnées angulaires du sommet de la i-ème tessellation et A_i est l'aire de la face correspondante dans la double grille. Puisque les éléments finis de l'aire ne s'additionneront pas exactement pour égaler 4π, une façon un peu plus exacte de calculer les coefficients est d'utiliser :

$$a_{lm} = \left(\frac{4\pi}{\sum_{i=1}^{N_V} A_i}\right) \sum_{i=1}^{N_V} r(\theta_i,\phi_i) y_{lm}(\theta_i,\phi_i) A_i. \tag{2.124}$$

Ainsi, avec l'aide d'une tessellation icosaèdrique de la sphère, calculer une surface SH se réduit en gros à échantillonner la surface à chaque sommet de tessellation en utilisant l'Eq 2.124 pour calculer les coefficients d'expansion SH.

La figure 2.9 montre la surface de VDW de lorazépam échantillonnée sur une tessellation icosaèdrique de la sphère, avec la surface lisse SH résultante. La figure 2.10 montre les surfaces SH de lorazépam reconstruites à partir de divers ordres d'expansion. La figure 2.11 montre les surfaces SH avec *L=16* des deux molécules de protéines, un anticorps et un lysozyme, pris du complexe de HyHel-5/lysozyme (code PDB 3HFL). En comparant ces figures, on peut voir que les expansions SH d'ordre inférieur peuvent capturer correctement les formes de petites molécules globulaires. En utilisant des expansions plus grandes que *L=8*, ceci donne une très petite différence apparente au niveau de la résolution. En revanche, les formes globales des molécules de protéines plus grandes sont clairement visibles avec les expansions de *L=16*, mais une grande partie des détails atomiques de ces grandes molécules est manquante. De plus, puisque les représentations SH doivent être de valeur unique, ou "star-like", par rapport aux rayons radiaux de l'origine choisie, les formes détaillées des cavités ou des poches dans une protéine ne peuvent pas être représentées. Néanmoins, le chapitre 4 montre que les représentations des surfaces de l'ordre inférieur donnent une manière efficace de chercher dans de grandes bases de données de petites molécules.

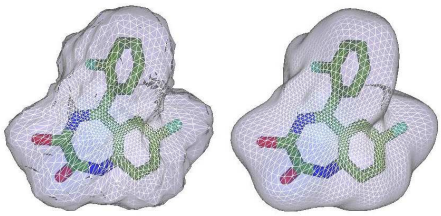

Figure 2.9: La surface de VDW de lorazépam échantillonnée sur une tessellation icosaèdrique (gauche) et la surface SH correspondante *L=16* (droite) calculée en utilisant l'Eq 2.124 et reconstruite des coefficients d'expansion SH. À noter que l'apparence un peu "cassée" de l'image à gauche est un artefact dû à l'utilisation d'un algorithme simple de tri en profondeur des triangles, utilisé pour réaliser l'effet de transparence. Le même algorithme de tri en profondeur est utilisé dans l'image à droite, mais peu de profondeurs incorrectes de surfaces de triangles sont calculées car la surface SH est plus lisse.

2.5 Les expansions polaires sphériques 3D de Fourier

Afin de capturer les formes détaillées de grandes molécules de protéines avec suffisamment de précision pour pouvoir exécuter des calculs d'amarrage, il est nécessaire d'augmenter les fonctions de base SH avec des fonctions radiales de base orthonormées appropriées. Ceci implique essentiellement d'abandonner la notion des surfaces tangibles et d'adopter un modèle de masse-densité de forme de protéines. Par conséquent, pour les représentations 3D, chaque propriété scalaire d'intérêt,

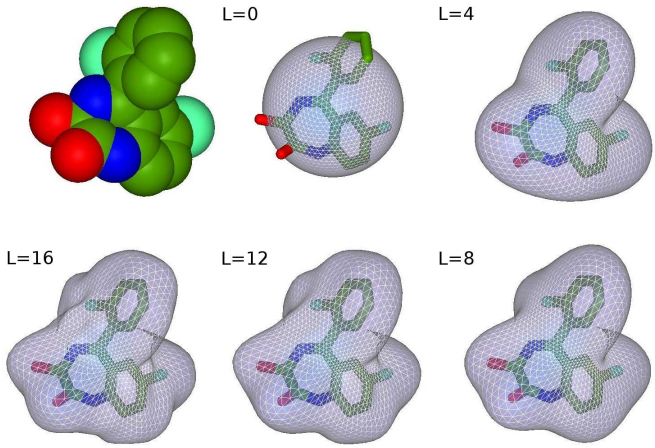

Figure 2.10: Les surfaces SH de lorazépam à divers ordres d'expansion.

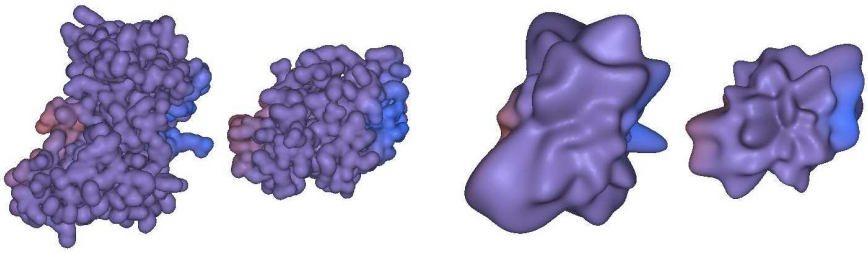

Figure 2.11: Une comparaison d'une paire de surfaces moléculaires SH 2D à l'ordre d'expansion *L=16* (à droite) avec la représentation originale de densité gaussienne atomique (à gauche) du domaine Fv anticorps HyHel-5 (grand domaine de protéine) et du lysozyme d'oeuf de poule (petit domaine). Les deux domaines sont séparés par une distance de 15 Å pour plus de clarté.

$A(\underline{r})$, est encodée comme une expansion de type Fourier des fonctions de base SH et radiales orthonormées jusqu'à un ordre donné N comme

$$A(\underline{r}) = \sum_{n=1}^{N} \sum_{l=0}^{n-1} \sum_{m=-l}^{l} a_{nlm} R_{nl}(r) y_{lm}(\theta, \phi). \tag{2.125}$$

J'utilise les GTO pour représenter la forme stérique et les ETO plus diffuses pour représenter les propriétés électrostatiques. La notation utilisée ici suit la convention de la chimie quantique dans laquelle l'index radial n, ou le nombre quantique principal, compte à partir de l'unité. Par conséquent, l'ordre harmonique le plus élevé et la puissance polynômiale la plus élevée dans une coordonnée individuelle quelconque est *L=N-1*.

Il vaut la peine de noter que ce n'est pas le seul type de fonction radiale qui pourrait être utilisé pour la représentation 3D de la forme. Par exemple, Mak *et al.* et Sael *et al.* ont récemment décrit

l'utilisation des polynômes de Zernike (Novotni & Klein, 2003) avec les SH pour construire les descripteurs rotation-invariables avec lesquels on peut comparer des formes de protéines (Mak *et al.*, 2008; Sael *et al.*, 2008). Cependant, des représentations rotation-invariables ne peuvent pas être utilisées pour superposer ou amarrer des protéines parce que toutes les informations sur l'orientation sont détruites. La question de savoir s'il pourrait y avoir des fonctions radiales meilleures que les fonctions de GL est considérée plus en détail dans la section 5.9.

2.5.1 Calcul de fonctions de densité de forme 3D

Suivant une approche semblable au cas de surface 2D, des représentations de forme-densité 3D de volume de l'intérieur moléculaire de VDW (τ) et de la surface skin (σ) peuvent être définies comme des fonctions de densité :

$$\tau(\underline{r}) = \begin{cases} 1 & \text{si } \underline{r} \in \text{atome de protéine} \\ 0 & \text{autrement,} \end{cases} \tag{2.126}$$

et

$$\sigma(\underline{r}) = \begin{cases} 1 & \text{si } \underline{r} \in \text{surface skin} \\ 0 & \text{autrement.} \end{cases} \tag{2.127}$$

En écrivant ces fonctions comme des expansions SPF à l'ordre N, ceci donne, par exemple,

$$\tau(\underline{r}) = \sum_{nlm} a^{\tau}_{nlm} R_{nl}(r)\, y_{lm}(\theta, \phi). \tag{2.128}$$

En échantillonnant les formes de la protéine sur une grille cartésienne régulière, les coefficients d'expansion peuvent être déterminés en additionnant les cellules de la grille non nulles :

$$\begin{aligned} a^{\tau}_{nlm} &= \int \tau(\underline{r}) R_{nl}(r)\, y_{lm}(\theta, \phi) \mathrm{d}V \\ &\simeq \sum_{k} R_{nl}(r_k)\, y_{lm}(\theta_k, \phi_k) \Delta V, \end{aligned} \tag{2.129}$$

où l'addition s'étend sur les cellules de grille non nulles k et où ΔV est le volume de chaque cellule de grille, et (r_k, θ_k, ϕ_k) sont les coordonnées polaires du centre de la k-ème cellule. J'utilise une grille d'échantillonnage cartésienne ayant généralement des cellules de 0.6 Å^3 de volume. La figure 2.12 montre quelques exemples de représentation de SPF du complexe entre l'anticorps HyHel-5 et le lysozyme d'oeuf de poule (code PDB 3HFL), calculés à partir des coefficients d'expansion de GTO à divers ordres *L=N-1*. Cette figure montre que différents atomes individuels de protéines commencent à être résolus clairement avec des expansions de l'ordre supérieur d'environ *L=28*.

D'une manière semblable, le volume de la surface skin peut être échantillonné sur une grille cartésienne en utilisant les vecteurs normaux de SAS calculés à partir de l'algorithme contourné

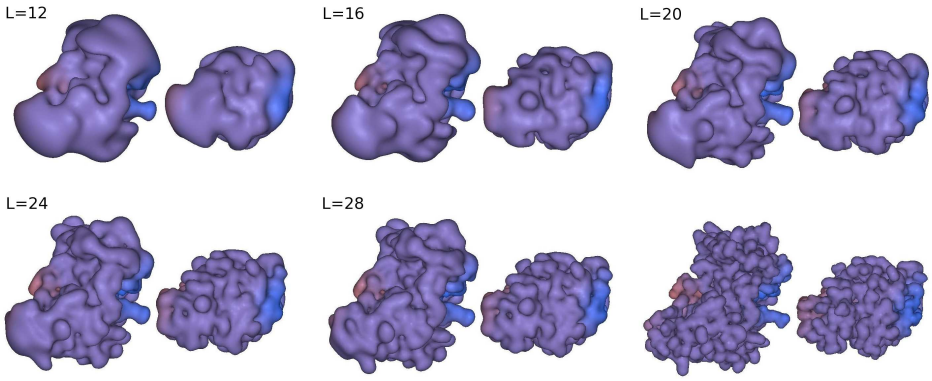

Figure 2.12: Les isosurfaces stériques de densité SPF de diverses expansions 3D de GTO pour le complexe entre le domaine de Fv de l'anticorps HyHel-5 (à gauche) et le lysozyme d'oeuf de poule (à droite). Les sous-unités sont séparées par une distance de 15 Å pour plus de clarté. La paire au bas à gauche montre les représentations gaussiennes atomiques des surfaces de VDW desquelles les expansions de SPF sont dérivées.

tétraédral pour remplir le volume entre la surface de SAS et de VDW avec un grand nombre de sphères échantillonnées. La figure 2.13 illustre cette idée. Les coefficients de surface skin, a^{σ}_{nlm}, peuvent alors être calculés en additionnant les voxels de grille non nuls comme auparavant.

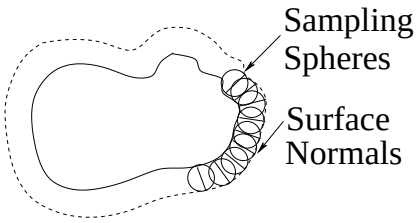

Figure 2.13: Une illustration schématique d'échantillonnage du volume de surface skin en utilisant de petites sphères échantillonnées placées tangentiellement à l'intérieur de SAS contournée et triangulée.

2.5.2 Calcul des propriétés électrostatiques de protéines

L'énergie électrostatique d'une distribution de charge, $\rho(\underline{r})$, sous l'influence d'un potentiel, $\phi(\underline{r})$, est donnée par (Jackson, 1975)

$$E = \frac{1}{2}\int \rho(\underline{r})\phi(\underline{r})\mathrm{d}V. \tag{2.130}$$

Dans la simulation MD conventionnelle, les paramètres nécessaires pour évaluer des interactions électrostatiques sont souvent représentés comme des point-charges (ou "point-charges") pour chaque type d'atome dans une molécule. Par conséquent, il est commode d'utiliser ces point-charges pour

calculer des expansions SPF 3D. Cependant, puisque ces modèles de charge assument en général la présence des atomes d'hydrogène, et puisque des atomes d'hydrogène ne sont pas normalement résolus dans des structures protéiques de rayon X, il est d'abord nécessaire d'ajouter à la protéine des hydrogènes polaires. Ceci est fait automatiquement dans le programme *Hex* en utilisant des géométries standard d'acides aminés pour déduire ou deviner les positions des hydrogènes. Des charges d'atome sont alors assignées à partir de l'ensemble de paramètres AMBER (Weiner *et al.*, 1984). Les coefficients de densité de charge de SPF peuvent être calculés en égalisant une expansion d'ETO pour $\rho(\underline{r})$ à une expression classique de densité de charge due à un ensemble de point-charges, q_i, aux positions $\underline{x}_i \equiv \underline{r}_i$, en utilisant

$$\rho(\underline{r}) = \sum_i q_i \, \delta(\underline{x} - \underline{x}_i) = \sum_{n'=1}^{N} \sum_{l'=0}^{n'-1} \sum_{m'=-l'}^{l'} a^{\rho}_{n'l'm'} S_{n'l'}(r) y_{l'm'}(\theta, \phi), \tag{2.131}$$

où $\delta(\underline{x})$ est le delta 3D de Dirac

$$\delta(\underline{x}) = \begin{cases} 1 \text{ si } \underline{x} = (0,0,0), \\ 0 \text{ autrement.} \end{cases} \tag{2.132}$$

Puis en multipliant les deux côtés de l'Eq 2.131 par $S_{nl}(r) y_{lm}(\theta, \phi)$ et en intégrant immédiatement on obtient le résultat

$$a^{\rho}_{nlm} = \sum_i q_i \, S_{nl}(r_i) y_{lm}(\theta_i, \phi_i). \tag{2.133}$$

Les coefficients d'expansion pour le potentiel *in vacuo* peuvent être calculés à partir de la densité de charge en résolvant l'équation de Poisson

$$\nabla^2 \phi(\underline{r}) = -4\pi \rho(\underline{r}). \tag{2.134}$$

En substituant l'expansion de série de chaque côté, en appliquant ∇^2 aux fonctions de base, en multipliant les deux côtés du résultat par $S_{n'l'}(r) y_{l'm'}(\theta, \phi)$ et en intégrant, ceci donne

$$\sum_{n=l+1}^{N} a^{\phi}_{nlm} \int_0^{\infty} \left(S''_{nl}(r) + 2S'_{nl}(r)/r - l(l+1) S_{nl}(r)/r^2 \right) S_{n'l}(r) r^2 \mathrm{d}r = -4\pi a^{\rho}_{n'lm}, \tag{2.135}$$

où S' dénote $\partial S / \partial r$ etc. Puis, en intégrant par parties le terme $S'_{nl}(r)$ ceci donne

$$\sum_{n=l+1}^{N} a^{\phi}_{nlm} G^{(l)}_{nn'} = -4\pi a^{\rho}_{n'lm}, \tag{2.136}$$

où chaque élément de $G^{(l)}$ a la forme symétrique

$$G^{(l)}_{nn'} = -\int_0^{\infty} \left(S'_{nl}(r) S'_{n'l}(r) r^2 + l(l+1) S_{nl}(r) S_{n'l}(r) \right) \mathrm{d}r. \tag{2.137}$$

On peut voir que pour chaque l et m, l'Eq 2.136 représente un ensemble d'équations simultanées avec des coefficients, $a^{\phi}_{n'lm}$, qui peuvent être déterminés en inversant chaque matrice $G^{(l)}$. Les éléments de $G^{(l)}$ peuvent être calculés par la manipulation directe de l'expansion de série pour les polynômes de Laguerre. Par exemple, en écrivant

$$S_{nl}(r) = \sum_{k=0}^{n-l-1} D_{nlk}\, e^{-\rho/2} \rho^{l+k}, \tag{2.138}$$

où $\rho = 2\Lambda r$ et

$$D_{nlk} = (-1)^k \frac{\big((n-l-1)!(n+l+1)!\big)^{1/2}}{(n-l-k-1)!(k+2l+2)!k!}, \tag{2.139}$$

permet à l'Eq 2.137 d'être simplifiée terme à terme pour donner l'expression symétrique

$$G^{(l)}_{nn'} = \frac{1}{4} \sum_{k=0}^{n-l-1} \sum_{k'=0}^{n'-l-1} D_{nlk} D_{n'lk'} (2l+k+k')![(k-k')^2 - (k+k') - 2(2l+1)(l+1)]. \tag{2.140}$$

Chaque matrice $G^{(l)}$ peut alors être inversée en utilisant des techniques numériques standard. Grâce à cette approche, les expansions de charge-densité et de potentiel à l'ordre *N=30* peuvent être calculées pour une protéine typique en moins d'une seconde. Il vaut la peine de noter que l'équation de Poisson est normalement résolue en utilisant l'intégration numérique sur une grille 3D qui peut être coûteuse, et sujette à erreur. Si les point-charges ne sont pas trop éloignés de l'origine, la solution analytique présentée ici est efficacement exacte jusqu'à l'ordre d'expansion choisi. La figure 2.14 montre le potentiel électrostatique d'ETO sur la surface de la protéine de lysozyme d'oeuf de poule (Code PDB 1LZA) calculée à l'ordre *N=30*.

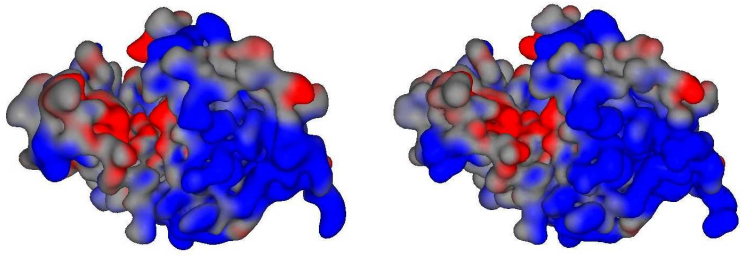

Figure 2.14: Le potentiel électrostatique d'ETO calculé à l'ordre *N=30* pour le lysozome (code PDB 1LZA). L'image sur la gauche montre le potentiel électrostatique sur l'isosurface contournée de densité de SPF *N=30*. L'image sur la droite montre le même potentiel sur la densité gaussienne atomique contournée originale. Les couleurs rouges et bleues représentent des potentiels négatifs et positifs respectivement. Dans ces images, le lysozyme est orienté de manière à montrer son site catalytique (rouge) dans la région gauche de la molécule.

Chapitre 3

Corrélations polaires sphériques de Fourier

3.1 Notation d'un opérateur et opérations de coordonnées 3D

Dans les applications infographiques, par exemple, dans lesquelles des objets graphiques complexes sont souvent représentés par des listes de polygones connectés, il est relativement simple de localiser de tels objets dans une scène en multipliant les coordonnées des sommets du composant polygone par des matrices de rotation et de translation homogènes 4×4 bien choisies. D'autre part, afin de comparer une paire de formes moléculaires semblables ou d'amarrer une paire de molécules complémentaires, il est nécessaire de tourner et translater l'une ou les deux molécules afin de trouver l'orientation relative qui donnera la meilleure superposition ou contraposition respectivement. Cependant, même si les molécules sont représentées comme des expansions équivalentes aux expansions de Fourier et que le but est de transformer de telles expansions directement, il n'est pourtant pas trivial d'en déduire comment les opérations correspondantes des coordonnées pourraient être représentées et implémentées.

Par conséquent, il me semble utile de commencer par définir des opérateurs de coordonnées abstraits, $\hat{R}$ et $\hat{T}$, qui, respectivement, tourneront et translateront des objets ou des représentations de l'espace euclidien ou de Hilbert de manière appropriée. Par exemple, si $\hat{R}_z(\alpha)$ représente un opérateur permettant à une molécule (objet) d'effectuer une rotation de α autour de l'axe z et si la molécule est représentée par une expansion SPF avec les coefficients a_{nlm}, alors nous souhaitons trouver les coefficients de rotation a'_{nlm} appropriés de sorte que la molécule tournée puisse être représentée comme

$$\hat{R}_z(\alpha)\sigma(\underline{r}) = \sum_{n=1}^{N}\sum_{l=0}^{n-1}\sum_{m=-l}^{l} a'_{nlm}R_{nl}(r)y_{lm}(\theta,\phi). \tag{3.1}$$

Nous pouvons noter qu'il y a une analogie directe entre les rotations et les translations dans un espace

de Hilbert et l'opération correspondante sur les coordonnées dans un espace euclidien. Par exemple, une rotation "active" positive des coefficients d'expansion (objet) est complètement équivalente à une rotation opposée "passive" des fonctions de base (axes de coordonnées). En d'autres termes, Eq 3.1 peut également être écrite comme

$$\hat{R}_z(\alpha)\sigma(\underline{r}) \equiv \sigma(\hat{R}_z(\alpha)^{-1}\underline{r}) = \sum_{k=1}^{N}\sum_{j=0}^{k-1}\sum_{p=-j}^{j} a_{kjp}R_{kj}(r)y_{jp}(\theta,\phi-\alpha), \tag{3.2}$$

où les indices sur la droite ont été re-nommés pour faciliter la prochaine étape. Si nécessaire, la forme de l'opérateur de rotation sur la gauche peut être instanciée comme une matrice 3×3

$$\hat{R}_z(\alpha)^{-1} = \underline{R}_z(-\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.3}$$

Maintenant, en égalisant les additions dans les équations 3.1 et 3.2, en multipliant les deux côtés par $R_{nl}(r)y_{lm}(\theta,\phi)$ et en intégrant, on obtient

$$a'_{nlm} = a_{nlp}\int_0^{2\pi} \psi_m(\phi)\psi_p(\phi-\alpha). \tag{3.4}$$

En d'autres mots, en considérant l'*intégrale de recouvrement* entre les fonctions de base originales et transformées, on peut calculer (au moins en principe) les coefficients d'expansion transformés. En pratique, ceci est simple pour des rotations-z pures, mais beaucoup plus difficile pour des rotations et des translations générales. Néanmoins, ceci est essentiellement le point de départ utilisé dans la section 3.4 pour calculer les translations. Toutefois, il est d'abord nécessaire d'étudier plus en détails les rotations.

Puisqu'une rotation implique toujours un axe et un angle de rotation, trois paramètres angulaires sont nécessaires pour décrire une rotation générale dans l'espace 3D (c-à-d. deux angles fixant l'orientation de l'axe et le troisième angle indiquant la valeur de rotation sur cet axe). En travaillant avec les SH, il est normal de suivre la convention "z-y-z" d'Euler pour des rotations dans l'espace 3D, dans lequel une rotation active d'un objet est décrite par des rotations successives de γ autour de l'axe z, de β autour de l'axe y et finalement de α (encore) autour de l'axe z. Ceci peut être représenté comme

$$\hat{R}(\alpha,\beta,\gamma) = \hat{R}_z(\alpha)\hat{R}_y(\beta)\hat{R}_z(\gamma) \tag{3.5}$$

où l'opérateur à l'extrême droite est appliqué en premier. Comme il est montré ci-dessous, des rotations autour de l'axe z sont simples à implémenter pour des expansions de SH. Par conséquent, la convention z-y-z d'Euler est un choix naturel pour représenter les rotations 3D. Lorsqu'il est appliqué

aux coordonnées cartésiennes, l'opérateur général de rotation Euler peut être instancié comme

$$\underline{R}(\alpha,\beta,\gamma) = \begin{pmatrix} \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma & -\cos\alpha\cos\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta \\ \sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\cos\beta\sin\gamma & \sin\alpha\sin\beta \\ -\sin\beta\cos\gamma & \sin\beta\sin\gamma & \cos\beta \end{pmatrix}. \tag{3.6}$$

Inversement, la forme de l'Eq 3.6 nous indique qu'une matrice générale de rotation, $\underline{R}$, peut être décomposée en trois angles de rotation d'Euler selon

$$\begin{aligned} \beta &= \cos^{-1}(R_{22}) \\ \gamma &= \cos^{-1}(-R_{20}/\sin\beta) \\ \alpha &= \cos^{-1}(R_{02}/\sin\beta). \end{aligned} \tag{3.7}$$

Les cas spéciaux de $R_{22} = 1$ et $R_{22} = -1$ peuvent être résolus en mettant $\gamma = 0$ et $\alpha = \cos^{-1}(R_{00})$ ou $\alpha = \cos^{-1}(R_{11})$ respectivement.

Bien qu'un certain système puisse demander l'application d'une transformation arbitraire de coordonnées à une expansion polaire de Fourier, une translation générale est beaucoup plus difficile à calculer qu'une rotation générale dans une base SH. Par conséquent, il est avantageux de calculer des translations seulement par rapport à l'axe de z et de définir un mouvement général comme une composition d'une ou plusieurs rotations et d'une seule translation en z pure. Par exemple, si $\underline{T}$ représente une matrice générale de transformation homogène 4×4,

$$\underline{T} = \begin{pmatrix} R_{00} & R_{01} & R_{02} & T_x \\ R_{10} & R_{11} & R_{12} & T_y \\ R_{20} & R_{21} & R_{22} & T_z \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{3.8}$$

et si

$$\begin{aligned} r &= \sqrt{T_x^2 + T_y^2 + T_z^2} \\ \theta &= \cos^{-1}(T_z/r) \\ \phi &= \cos^{-1}(T_x/r\sin\theta) \end{aligned} \tag{3.9}$$

(où $\phi = 0$ si $\theta = 0$), alors $\underline{T}$ peut être décomposé en

$$\underline{T} = \underline{R}_2\,\underline{T}_z(r)\,\underline{R}_1, \tag{3.10}$$

où $\underline{T}_z(r)$ représente une translation pure de r le long de l'axe positif de z, et où $\underline{R}_1$ et $\underline{R}_2$ sont des rotations pures. Il est alors simple de voir que

$$\underline{R}_2 = \underline{R}_z(\phi)\,\underline{R}_y(\theta) \tag{3.11}$$

et

$$\underline{R}_1 = \underline{T}_z(-r)\,\underline{R}_y(-\theta)\,\underline{R}_z(-\phi)\,\underline{T}. \tag{3.12}$$

Puisque $\underline{R}_1$ se résoudra normalement grâce aux trois angles distincts de rotation d'Euler, on peut voir qu'une transformation générale 3D peut être caractérisée par un paramètre de translation et cinq paramètres angulaires.

3.2 Théorèmes d'addition et corrélations

Un théorème d'addition est une relation algébrique entre les paramètres d'une fonction telle que $f(a+b)$ en termes de paramètres individuels, $f(a)$ et $f(b)$. Un exemple simple d'un théorème d'addition est

$$e^{i(\alpha+\beta)} = e^{i\alpha}e^{i\beta}. \tag{3.13}$$

En termes d'opérateurs, l'action d'appliquer deux rotations consécutives de α et β autour d'un axe dans l'espace 3D peut aussi être considérée comme une sorte de théorème d'addition. Par exemple, en considérant Eq 3.13, il est normal de s'attendre à pouvoir écrire

$$\hat{R}_z(\alpha+\beta) = \hat{R}_z(\alpha)\hat{R}_z(\beta), \tag{3.14}$$

et, effectivement, en utilisant la formule d'Euler (Eq 2.14) il est simple de montrer que la représentation de la matrice correspondante 3×3 d'Eq 3.14 est donnée par

$$\begin{pmatrix} \cos(\alpha+\beta) & \sin(\alpha+\beta) & 0 \\ -\sin(\alpha+\beta) & \cos(\alpha+\beta) & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{3.15}$$

Cette équation de matrice applique simultanément les théorèmes classiques d'addition géométrique :

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \tag{3.16}$$

et

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta. \tag{3.17}$$

Des théorèmes d'addition plus sophistiqués ont été établis, qui impliquent des fonctions spéciales. Par exemple, un théorème d'addition de translation pour des SH solides régulières peut être représenté comme (Biedenharn & Louck, 1981) :

$$Y_{lm}(\underline{x}+\underline{T}) = \sum_{kj}\left[\frac{4\pi(2l+1)}{(2l-2k+1)(2k+1)}\binom{l+m}{k+j}\binom{l-m}{k-j}\right]^{1/2} Y_{kj}(\underline{x})Y_{l-k,m-j}(\underline{T}), \tag{3.18}$$

où l'addition comprend toutes les valeurs de k et j dont les factoriels sont bien définis.

Plus généralement, deux systèmes de coordonnées $\underline{r} = (r, \theta, \phi)$ et $\underline{r}' = \underline{r} - \underline{T} = (r', \theta', \phi)$, comme ceux illustrés dans la figure 3.1, peuvent être fonctionnellement reliés en utilisant le théorème d'addition de l'onde plane de Raleigh (Bransden & Joachain, 1997). Par exemple, en multipliant l'équation de vecteur

$$\underline{r} = \underline{T} + \underline{r}' \tag{3.19}$$

par un vecteur complexe arbitraire $i\underline{k}$ et en passant chaque côté de l'égalité en exponentielle, ceci donne :

$$e^{i\underline{k}.\underline{r}} = e^{i\underline{k}.\underline{T}}\, e.^{i\underline{k}.\underline{r}'} \tag{3.20}$$

En écrivant les vecteurs directeurs en coordonnées polaires sphériques, $\underline{k} = (\beta, \Theta, \Phi)$, $\underline{r} = (r, \theta, \phi)$, $\underline{r}' = (r', \theta', \phi')$, et $\underline{T} = (R, \gamma, \delta)$, le théorème d'addition de Raleigh relie les deux systèmes de coordonnées selon :

$$e^{i\underline{k}.\underline{r}} = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(\beta r) Y_{lm}(\Theta, \Phi)^* Y_{lm}(\theta, \phi). \tag{3.21}$$

Quand la translation est limitée à la direction positive de z, il peut être montré que l'équation de Raleigh peut être simplifiée pour obtenir un théorème d'addition sphérique de Bessel :

$$j_l(\beta r) Y_{lm}(\theta, \phi) = \sum_{l'=0}^{\infty} \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} j_k(\beta R) j_{l'}(\beta r') Y_{l'm}(\theta', \phi), \tag{3.22}$$

où le coefficient angulaire $A_k^{(ll'|m|)}$ est donné par

$$A_k^{(ll'|m|)} = (-1)^{(k+l'-l)/2+m}(2k+1)\big[(2l+1)(2l'+1)\big]^{1/2} \begin{pmatrix} l & l' & k \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} l & l' & k \\ m & \overline{m} & 0 \end{pmatrix}. \tag{3.23}$$

Les symétries de permutation du deuxième symbole *3-j* montrent que la partie droite est indépendante du signe de m, justifiant ainsi l'utilisation de $|m|$ pour marquer les éléments de la matrice. Du fait de la disparation du premier symbole *3-j* quand $l + l' + k$ est impair, on peut voir que les coefficients restants sont toujours réels et que l'addition sur k doit seulement être calculée pour des incréments pairs avec $k = |l - l'|, |l - l'| + 2, ..., l + l'$. Pour une translation négative, il peut être montré qu'un facteur additionnel de $(-1)^{l-l'}$ apparaît dans l'expression ci-dessus. Pour des détails complets, vous pouvez vous référer par exemple, à l'annexe A de Ritchie (2005).

Les théorèmes d'addition sont souvent très utiles pour calculer des corrélations. Dans l'analyse classique de Fourier, la *corrélation* entre une paire de fonctions, $f(\phi)$ et $g(\phi)$, est conventionnellement définie comme

$$(f * g)(\alpha) \equiv \int_0^{2\pi} f(\phi)^* g(\phi + \alpha) \mathrm{d}\phi. \tag{3.24}$$

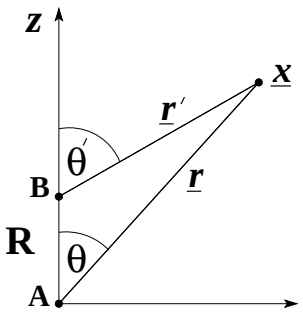

Figure 3.1: Une illustration des systèmes de coordonnées utilisés pour représenter des translations. La position d'un point $\underline{x}$ par rapport à un système de coordonnées où l'origine est à la position A et $\underline{x} = \underline{r} = (r, \theta, \phi)$, alors que le même point a des coordonnées $\underline{x} = \underline{r}' = (r', \theta', \phi)$ par rapport à un système de coordonnées où l'origine est à la position B. Les deux systèmes de coordonnées, A et B, sont reliés par une translation R dans la direction de z.

En d'autres termes, une corrélation est le degré de recouvrement entre une fonction et une version décalée de l'autre fonction. Bien sûr, dans les cas d'appariement moléculaire et d'amarrage, l'objectif est de calculer des intégrales de recouvrement dans l'espace 3D dans lequel l'une ou les deux fonctions ont été tournées ou translatées pour des valeurs spécifiques. Cependant, la notation concise utilisée dans Eq 3.24 pour des corrélations unidimensionnelles ne peut pas être facilement développée pour décrire des corrélations multidimensionnelles sans ambiguïté. Par conséquent, je préfère utiliser la notation d'opérateur définie ci-dessus. Donc, par exemple, Eq 3.24 serait représentée par

$$(f * g)(\alpha) \equiv \int_0^{2\pi} f(\phi)^* [g(\hat{R}(\alpha)\phi)] \mathrm{d}\phi \equiv \int_0^{2\pi} f(\phi)^* [\hat{R}(-\alpha) g(\phi)] \mathrm{d}\phi. \tag{3.25}$$

3.3 Rotation des expansions polaires sphériques de Fourier

3.3.1 Les matrices de rotation de Wigner

Il peut être montré que les fonctions de SH de chaque ordre l, sont transformées en une rotation général d'Euler selon (Rose, 1957; Biedenharn & Louck, 1981)

$$\hat{R}(\alpha, \beta, \gamma) Y_{lm}(\theta, \phi) = \sum_{m'} Y_{lm'}(\theta, \phi) D^{(l)}_{m'm}(\alpha, \beta, \gamma) \tag{3.26}$$

où $\hat{R}(\alpha, \beta, \gamma)$ représente un opérateur de rotation exprimé en termes de paramétrisation d'angle d'Euler (α, β, γ). C'est à Wigner (1939) que l'on doit les matrices unitaires de rotation $D^{(l)}(\alpha, \beta, \gamma)$. Cependant, il semble qu'il n'ait jamais publié une dérivation complète. La façon la plus élégante et la plus directe pour calculer des rotations générales de SH est probablement d'utiliser une technique d'opérateur de Boson (Sakurai, 1994). Le résultat est (Biedenharn & Louck, 1981)

$$D^{(l)}_{m'm}(\alpha, \beta, \gamma) = e^{-im'\alpha} d^{(l)}_{m'm}(\beta) e^{-im\gamma}, \tag{3.27}$$

où

$$d^{(l)}_{m'm}(\beta) = \left[(l+m')!(l-m')!(l+m)!(l-m)!\right]^{1/2} \times \sum_{k=\max(0,m-m')}^{\min(l-m',l+m)} \frac{(-1)^{k+m'-m}(\cos\beta/2)^{2l+m-m'-2k}(\sin\beta/2)^{2k+m'-m}}{(l+m-k)!k!(m'-m+k)!(l-m'-k)!}. \quad (3.28)$$

Les éléments de la matrice $d^{(l)}$ ont les symétries utiles (Biedenharn & Louck, 1981)

$$\begin{aligned} d^{(l)}_{m'm}(\beta) &= (-1)^{m'-m} d^{(l)}_{mm'}(\beta) \\ &= (-1)^{m'-m} d^{(l)}_{\overline{m}'\overline{m}}(\beta). \end{aligned} \quad (3.29)$$

De point de vue de la programmation, ceci signifie qu'à peu près trois quarts des éléments de chaque matrice $d^{(l)}$ peuvent être calculés de façon triviale. Lorsque $|m'| = l$ ou $|m| = l$, la somme dans Eq 3.28 se réduit en un seul terme et on obtient, par exemple,

$$d^{(l)}_{m'l} = \left[\frac{(2l)!}{(l-m')!(l+m')!}\right]^{1/2} (\cos\beta/2)^{l+m'}(\sin\beta/2)^{l-m'}. \quad (3.30)$$

Comme les autres fonctions spéciales, il y a quelques formules de récursivité de trois-termes qui peuvent être utilisées pour faire des calculs efficaces. Par exemple, sauf aux pôles, la formule

$$\frac{2(m\cos\beta - m')}{\sin\beta} d^{(l)}_{m'm} = \left[(l+m+1)(l-m)\right]^{1/2} d^{(l)}_{m',m+1} + \left[(l-m+1)(l+m)\right]^{1/2} d^{(l)}_{m',m'-1} \quad (3.31)$$

peut être utilisée pour calculer efficacement les éléments consécutifs de la matrice à partir d'Eq 3.30.

3.3.2 Les matrices réelles de rotation de Wigner

Des combinaisons linéaires de SH complexes comme Eq 2.82 préservent la symétrie de rotation et les fonctions réelles de SH se transforment également en une rotation. Ce fonctionnement peut être représenté comme

$$\hat{R}(\alpha,\beta,\gamma) y_{lm}(\theta,\phi) = \sum_{m'} y_{lm'}(\theta,\phi) R^{(l)}_{m'm}(\alpha,\beta,\gamma). \quad (3.32)$$

où $R^{(l)}$ est une matrice réelle de rotation. Si Eq 2.82 est écrite comme une somme,

$$y_{lm}(\theta,\phi) = \sum_{m'} U^{(l)}_{mm'} Y_{lm'}(\theta,\phi), \quad (3.33)$$

il peut alors être montré que, pour chaque ordre l, la matrice réelle de rotation, $\underline{R}$, est donnée par l'équation de matrice

$$\underline{R}^{(l)} = \underline{U}^{(l)} \underline{D}^{(l)} \underline{U}^{(l)\dagger}, \quad (3.34)$$

où $\underline{U}^\dagger$ est la matrice complexe transposée conjuguée de $\underline{U}$. En reconnaissant que tous les éléments non-diagonaux de $\underline{U}$ sont à zéro, il est relativement simple, mais fastidieux, de simplifier symboliquement Eq 3.34. Le résultat est

$$R^{(l)}_{m'm} = \begin{cases} d^{(l)}_{m'm}(\beta)\cos(m\gamma + m'\alpha) + (-1)^{m'} d^{(l)}_{\overline{m}'m}(\beta)\cos(m\gamma - m'\alpha) & ;\ m' > 0, m > 0 \\ d^{(l)}_{0m}(\beta)\sqrt{2}\cos(m\gamma) & ;\ m' = 0, m > 0 \\ (-1)^{m'+1} d^{(l)}_{m'm}(\beta)\sin(m\gamma + m'\alpha) + d^{(l)}_{\overline{m}'m}(\beta)\sin(m\gamma - m'\alpha) & ;\ m' < 0, m > 0 \\ d^{(l)}_{m'0}(\beta)\sqrt{2}\cos(m'\alpha) & ;\ m' > 0, m = 0 \\ d^{(l)}_{00}(\beta) & ;\ m' = 0, m = 0 \\ (-1)^{m'+1} d^{(l)}_{m'0}(\beta)\sqrt{2}\sin(m'\alpha) & ;\ m' < 0, m = 0 \\ (-1)^{m} d^{(l)}_{m'm}(\beta)\sin(m\gamma + m'\alpha) + (-1)^{m+m'} d^{(l)}_{\overline{m}'m}(\beta)\sin(m\gamma - m'\alpha) & ;\ m' > 0, m < 0 \\ (-1)^{m} d^{(l)}_{0m}(\beta)\sqrt{2}\sin(m\gamma) & ;\ m' = 0, m < 0 \\ (-1)^{m+m'} d^{(l)}_{m'm}(\beta)\cos(m\gamma + m'\alpha) + (-1)^{m+1} d^{(l)}_{\overline{m}'m}(\beta)\cos(m\gamma - m'\alpha) & ;\ m' < 0, m < 0. \end{cases} \tag{3.35}$$

Dans le cas où $\beta = \gamma = 0$, ceci engendre une seule rotation α, autour de l'axe z qui a une forme particulièrement simple. Puisque

$$\hat{R}(\alpha, 0, 0)\ y_{lm}(\theta, \phi) = y_{lm}(\theta, \phi - \alpha), \tag{3.36}$$

les identités

$$\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta \tag{3.37}$$

et

$$\sin(\alpha \pm \beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \tag{3.38}$$

peuvent être utilisées pour montrer que

$$\hat{R}(\alpha, 0, 0)\ y_{lm}(\theta, \phi) = y_{lm}(\theta, \phi)\cos m\alpha + y_{l\overline{m}}(\theta, \phi)\sin m\alpha. \tag{3.39}$$

Avec un peu plus de travail, l'identité

$$d^{(l)}_{m'm}(\beta = 0) = \delta_{m'm} \tag{3.40}$$

peut être substituée par des représentations d'éléments de matrice de rotation (Eq 3.35) et Eq 3.32 peut être simplifiée pour obtenir le même résultat. Pour une rotation générale, dans laquelle $\beta \neq 0$, tous les éléments de la matrice de rotation doivent être calculés mais le nombre de calculs peut être réduit considérablement en utilisant des symétries des matrices $d^{(l)}$ (Eq 3.29).

3.4 Translation des expansions polaires sphériques de Fourier

Une grande partie de la QM consiste à calculer des intégrales des produits de fonctions d'onde électronique centrées sur différentes coordonnées nucléaires. Cependant, bien que ces intégrales dépendent des distances entre les centres nucléaires, il y a peu de littérature sur la façon d'associater de telles intégrales aux actions d'un opérateur de translation. Au contraire, il semble que la pratique habituelle dans QM est de recalculer toutes les intégrales de recouvrement nécessaires chaque fois que les coordonnées nucléaires changent. Ceci est sans doute la façon correcte de procéder dans QM où la tâche est d'évaluer aussi exactement que possible un très grand nombre d'intégrales de multi-centres d'ordre relativement inférieur. Cependant, pour l'amarrage protéique et l'appariement moléculaire basé sur la forme, nous devons calculer l'intégrale de recouvrement entre les fonctions de base d'ordre considérablement supérieur de seulement deux centres d'expansion. De plus, nous devons potentiellement répéter ces calculs sur un grand nombre d'orientations de rotation distinctes par rapport à une ou deux molécules. Par conséquent, il est impératif de développer une représentation explicite de matrice de l'opérateur de translation afin d'éviter de recalculer inutilement un grand nombre d'intégrales de recouvrement d'ordre supérieur pendant une recherche de corrélation.

3.4.1 Intégrales de recouvrement comme éléments de matrice de translation

Une méthode simple pour trouver la forme générale des matrices de translation est de considérer en premier le recouvrement entre un "corps" fixe A ou une fonction scalaire $A(\underline{r})$ et un corps mobile B ou une fonction $B(\underline{r})$, sous une translation active de B par $\underline{T} = (R, 0, 0)$ le long de l'axe positif z, comme il est illustré dans la figure 3.1. Symboliquement, ceci peut être représenté par

$$C(\underline{T}) = \int A(\underline{r}) B(\underline{T}^{-1}\underline{r}) \, \mathrm{d}V. \tag{3.41}$$

En substituant les expansions de $A(\underline{r})$ et $B(\underline{r})$, ceci donne

$$C(\underline{T}) = \sum_{nlm} \sum_{n'l'm'} a_{nlm} b_{n'l'm'} \int R_{nl}(r) y_{lm}(\theta, \phi) R_{n'l'}(r') y_{l'm'}(\theta', \phi') \, \mathrm{d}V \tag{3.42}$$

où la notation abrégée $\sum_{nlm}$, etc., est utilisée pour indiquer la somme sur les plages des indices données dans Eq 2.125, et où $\underline{r} = (r, \theta, \phi)$ et $\underline{r}' = \underline{r} - \underline{T} = (r', \theta', \phi')$. Dans ce cas où ϕ et ϕ' demeurent coïncidentes, les fonctions circulaires, $\varphi_m(\phi)$, peuvent être intégrées et Eq 3.42 se réduit à une somme sur les intégrales 2D dans la plan (r, θ). Puisque la valeur de chacune de ces intégrales dépend seulement de la distance R et qu'elle est indépendante du signe de m (voir ci-dessous), nous pouvons écrire :

$$T^{(|m|)}_{nl,n'l'}(R) = \int R_{nl}(r) \vartheta_{lm}(\theta) R_{n'l'}(r') \vartheta_{l'm}(\theta') r^2 \mathrm{d}r \sin\theta \mathrm{d}\theta, \tag{3.43}$$

et

$$C(R) = \sum_{nlm} \sum_{n'l'm'} a_{nlm} b_{n'l'm'} T^{(|m|)}_{nl,n'l'}(R) \delta_{mm'}, \tag{3.44}$$

et interpréter chaque $T^{(|m|)}_{nl,n'l'}(R)$ comme un élément de matrice de l'opérateur de translation. Par exemple, à partir d'Eq 3.44 on peut voir que les deux additions

$$b^R_{nlm} = \sum_{n'l'} T^{(|m|)}_{nl,n'l'}(R) b_{n'l'm} \tag{3.45}$$

et, après avoir renommé les indices,

$$a^R_{nlm} = \sum_{n'l'} T^{(|m|)}_{n'l',nl}(R) a_{n'l'm} \tag{3.46}$$

représentent une translation positive du corps B ou, d'une façon équivalente, une translation négative du corps A respectivement. Les matrices de translation sont évidemment des quantités à cinq dimensions. Cependant, parce qu'elles ne dépendent pas du signe de m, il est utile de considérer que chaque matrice se compose de rangées bidimensionnelles $\sum_{m=0}^{N-1} = N$, chacune étant classée par $\sum_{nl} = N(N+1)/2$ valeurs possibles pour chaque paire d'indices nl. Les éléments de matrice disparaissent trivialement lorsque $|m| > l$. En outre, la notation utilisée ici est destinée à être en accord avec la convention habituelle pour des éléments SH complexes et réels de matrices de rotation, $D^{(l)}_{m'm}(\alpha, \beta, \gamma)$ et $R^{(l)}_{m'm}(\alpha, \beta, \gamma)$, respectivement, dans le sens où une translation positive en z pure des fonctions de base est représentée comme

$$R_{nl}(r') y_{lm}(\theta', \phi) = \sum_{n'=1}^{\infty} \sum_{l'=0}^{n'-1} T^{(|m|)}_{n'l',nl}(R) R_{n'l'}(r) y_{l'm}(\theta, \phi). \tag{3.47}$$

Dans le reste de cette section, il sera montré que les éléments de matrice de translation pour des fonctions de base de SPF peuvent être calculés comme des sommes sur les transformations sphériques inverses 1D de Bessel. Premièrement, de la section 2.1.14, si la transformation sphérique de Bessel est définie comme

$$\tilde{R}_{nl}(\beta) = (2/\pi)^{1/2} \int_0^{\infty} R_{nl}(r) j_l(\beta r) r^2 \mathrm{d}r, \tag{3.48}$$

alors la transformation inverse est donnée par

$$R_{nl}(r) = (2/\pi)^{1/2} \int_0^{\infty} \tilde{R}_{nl}(\beta) j_l(\beta r) \beta^2 \mathrm{d}\beta. \tag{3.49}$$

Ensuite, en multipliant les deux côtés du théorème sphérique d'addition de Bessel, Eq 3.22, par $(2/\pi)^{1/2} \tilde{R}_{nl}(\beta) \beta^2$ et en intégrant par rapport à β (c-à-d. en appliquant la transformation inverse ci-dessus), on obtient :

$$R_{nl}(r) Y_{lm}(\theta, \phi) = \left(\frac{2}{\pi}\right)^{1/2} \sum_{l'=0}^{\infty} \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \int_0^{\infty} \tilde{R}_{nl}(\beta) j_k(\beta R) j_{l'}(\beta r') \beta^2 \mathrm{d}\beta \; Y_{l'm}(\theta', \phi).$$

(3.50)

Puis, en multipliant chaque côté par $R_{n'j'}(r')Y_{j'm'}(\theta', \phi)$ et en intégrant sur tout l'espace des variables correspondantes, ceci donne

$$T^{(|m|)}_{n'j',nl}(R) = \left(\frac{2}{\pi}\right)^{1/2} \sum_{l'=0}^{\infty} \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \int_0^\infty \int_0^\infty \delta_{j'l'} R_{n'j'}(r') j_{l'}(\beta r')\, \tilde{R}_{nl}(\beta) j_k(\beta R) \beta^2 \mathrm{d}\beta\, r'^2 \mathrm{d}r'. \tag{3.51}$$

Pour finir, en reconnaissant l'intégrale dans r' comme une transformation sphérique de Bessel de $R_{n'l'}(r')$, le résultat se réduit à une somme finie de termes :

$$T^{(|m|)}_{n'l',nl}(R) = \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \int_0^\infty \tilde{R}_{n'l'}(\beta) \tilde{R}_{nl}(\beta) j_k(\beta R) \beta^2 \mathrm{d}\beta. \tag{3.52}$$

Ceci généralise l'expression donnée par Danos et Maximon (1965) pour translater des expansions multipolaires au cas plus général pour des fonctions radiales orthonormales arbitraires, $R_{nl}(r)$. Avec des arguments similaires, il peut aussi être montré que

$$T^{(|m|)}_{nl,n'l'}(R) = T^{(|m|)}_{n'l',nl}(-R) = (-1)^{l'-l} T^{(|m|)}_{n'l',nl}(R). \tag{3.53}$$

En conséquence, pratiquement la moitié de tous les éléments de matrice peut être trouvée par symétrie. Étant donné que les fonctions originales de base forment un ensemble orthonormé complet, il est simple de montrer que les matrices de translation sont aussi orthonormées dans le sens que

$$\sum_{n'=1}^{\infty} \sum_{l'=0}^{n'-1} T^{(|m|)}_{n'l',nl}(R) T^{(|m|)}_{n'l',n''l''}(R) = \delta_{nn''}\delta_{ll''}. \tag{3.54}$$

L'évaluation de cette équation donne une méthode convenable afin de vérifier l'exactitude des calculs suivants.

Les sections suivantes développeront des expressions explicites de formes closes pour les éléments de matrice de translation pour des fonctions de base de GTO et d'ETO.

3.4.2 Les éléments de matrice de translation de GTO

À partir d'Eq 2.98, les fonctions radiales normalisées de GTO sont données par

$$R_{nl}(r) = \left[\frac{2}{\lambda^{3/2}\pi^{1/2}} \frac{(n-l-1)!}{(1/2)_n}\right]^{1/2} e^{-\rho^2/2} \rho^l L_{n-l-1}^{(l+1/2)}(\rho^2)\ ; \quad \rho^2 = r^2/\lambda. \tag{3.55}$$

Les fonctions de GTO sont des fonctions propres de la transformation sphérique de Bessel (dérivées de (Erdélyi *et al.*, 1953c), p42, Eq 3)

$$\tilde{R}_{nl}(\beta) = (-1)^{n-l-1} \left[\frac{2\lambda^{3/2}}{\pi^{1/2}} \frac{(n-l-1)!}{(1/2)_n}\right]^{1/2} e^{-x^2/2} x^l L_{n-l-1}^{(l+1/2)}(x^2), \tag{3.56}$$

où $x^2 = \lambda\beta^2$. Ici, il est convenable d'utiliser Eq 2.89 pour développer Eq 3.56 comme une série entière

$$\tilde{R}_{nl}(\beta) = \left[\frac{4\lambda^{3/2}}{\pi^{1/2}}\right]^{1/2} \sum_j X_{nlj} e^{-x^2/2} x^{2j+l}, \tag{3.57}$$

où $\sum_j$ sert de notation abrégée pour $\sum_{j=0}^{n-l-1}$ et où les coefficients X_{nlj} sont donnés par

$$X_{nlj} = \left[\frac{(n-l-1)!(1/2)_n}{2}\right]^{1/2} \frac{(-1)^{n-l-j-1}}{j!(n-l-j-1)!(1/2)_{l+j+1}}. \tag{3.58}$$

En substituant Eq 3.57 deux fois dans Eq 3.52 et rassemblant les coefficients de x^{2k} en utilisant

$$C_k^{(nl,n'l')} = \sum_j \sum_{j'} \delta_{k,j+j'} X_{nlj} X_{n'l'j'} \tag{3.59}$$

nous obtenons, pour les éléments de la matrice de translation de GTO,

$$T_{n'l',nl}^{(|m|)}(R) = \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \sum_{j=0}^{n-l+n'-l'-2} C_j^{(nl,n'l')} \frac{4}{\pi^{1/2}} \int_0^\infty x^{2j+l+l'} j_k(xR/\lambda^{1/2}) x^2 \mathrm{d}x. \tag{3.60}$$

En applicant la relation (de (Erdélyi *et al.*, 1953c), p30, Eq 13)

$$\frac{4}{\pi^{1/2}} \int_0^\infty e^{-x^2} x^{2m+k} j_k(xy) x^2 \mathrm{d}x = m! e^{-y^2/4} (y^2/4)^{k/2} L_m^{(k+1/2)}(y^2/4), \tag{3.61}$$

l'équation donne alors le résultat analytique final

$$T_{n'l',nl}^{(|m|)}(R) = \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \sum_{j=0}^{n-l+n'-l'-2} C_j^{(nl,n'l')} M! e^{-R^2/4\lambda} (R^2/4\lambda)^{k/2} L_M^{(k+1/2)}(R^2/4\lambda), \tag{3.62}$$

où $M = j + (l + l' - k)/2$.

3.4.3 Les éléments de matrice de translation d'ETO

À partir d'Eq 2.101, les fonctions radiales normalisées d'ETO sont données par

$$S_{nl}(r) = \left[(2\Lambda)^3 \frac{(n-l-1)!}{(n+l+1)!}\right]^{1/2} e^{-\rho/2} \rho^l L_{n-l-1}^{(2l+2)}(\rho), \tag{3.63}$$

où $\rho = 2\Lambda r$ avec un facteur d'échelle de Λ. J'ai défini $\Lambda = 1/2$ pour des calculs électrostatiques entre protéines (Ritchie & Kemp, 2000). En utilisant un argument basé sur l'orthogonalité, Keister et Polyzou (1997) ont récemment montré que la transformation sphérique de Bessel de ces fonctions peut être écrite en termes de polynômes de Jacobi, $P_k^{(\alpha,\beta)}(t)$

$$\tilde{S}_{nl}(\beta) = \frac{2}{(1/2)_n} \left[\frac{(n-l-1)!(n+l+1)!}{\pi\Lambda^3}\right]^{1/2} \frac{s^l}{(s^2+1)^{l+2}} P_{n-l-1}^{(l+3/2,l+1/2)}\left(\frac{s^2-1}{s^2+1}\right), \tag{3.64}$$

où $s = \beta/\Lambda$. À la suite d'un traitement semblable au cas de GTO, l'expansion décalée de la série (Keister & Polyzou, 1997)

$$P_k^{(\mu,\nu)}(t) = \frac{1}{k!}\frac{\Gamma(k+\mu+1)}{\Gamma(k+\mu+\nu+1)}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\frac{\Gamma(k+j+\mu+\nu+1)}{\Gamma(j+\mu+1)}\left(\frac{1-t}{2}\right)^j \tag{3.65}$$

peut être utilisée pour rassembler des facteurs de $1/(s^2+1) = (1-t)/2$ pour écrire Eq 3.64 comme une série entière

$$\tilde{S}_{nl}(\beta) = \left[\frac{2}{\pi\Lambda^3}\right]^{1/2}\sum_j Y_{nlj}\frac{s^l}{(s^2+1)^{l+j+2}} \tag{3.66}$$

où

$$Y_{nlj} = \left[\frac{1}{2}\frac{(n-l-1)!}{(n+l+1)!}\right]^{1/2}\frac{(-1)^j(2n+1)(n+l+j+1)!}{j!(n-l-j-1)!(1/2)_{l+j+2}}. \tag{3.67}$$

En substituant Eq 3.66 deux fois dans Eq 3.52 et en rassemblant les coefficients de $1/(s^2+1)^k$ en utilisant

$$D_k^{(nl,n'l')} = \sum_j\sum_{j'}\delta_{k,j+j'}Y_{nlj}Y_{n'l'j'} \tag{3.68}$$

nous obtenons, pour les éléments de la matrice de translation d'ETO,

$$U_{n'l',nl}^{(|m|)}(R) = \sum_{k=|l-l'|}^{l+l'}A_k^{(ll'|m|)}\sum_{j=0}^{n-l+n'-l'-2}D_j^{(nl,n'l')}\frac{2}{\pi}\int_0^\infty\frac{s^{2M+k}}{(s^2+1)^{J+2}}j_k(s\Lambda R)s^2\mathrm{d}s, \tag{3.69}$$

où $M = (l+l'-k)/2$ et $J = j+l+l'+2$. Il est montré dans l'annexe B de Ritchie (2005) que l'intégrale restante peut être calculée sous la forme finie comme

$$\frac{2}{\pi}\int_0^\infty\frac{s^{2M+k}}{(s^2+1)^{J+2}}j_k(s\Lambda R)s^2ds = \sum_{q=0}^{M}\binom{M}{q}\frac{(-1)^{M+q}}{2^{J+1-q}(J+1-q)!}(\Lambda R)^k\hat{k}_{J-k-q+1/2}(\Lambda R), \tag{3.70}$$

où $\hat{k}_\sigma(z)$ est une fonction réduite de Bessel du deuxième genre. Pour les nombres entiers et demi, ces fonctions peuvent être calculées en utilisant les relations de récurrence (Weniger & Steinborn, 1983)

$$\begin{aligned}\hat{k}_{1/2}(z) &= e^{-z},\\ \hat{k}_{3/2}(z) &= (1+z)e^{-z},\\ \hat{k}_{n+3/2}(z) &= (2n+1)\hat{k}_{n+1/2}(z) + z^2\hat{k}_{n-1/2}(z).\end{aligned} \tag{3.71}$$

Ainsi, les éléments de matrice de translation d'ETO peuvent aussi être calculés analytiquement, bien qu'en comparant au cas de GTO, une somme interne additionnelle soit nécessaire.

3.4.4 Les matrices non-orthogonales de translation

Des translations d'expansions de SPF dans les deux bases de GTO et d'ETO peuvent être calculées plus économiquement en éliminant la somme intérieure sur l'indice j dans Eq.s 3.62 et 3.69. Ceci revient à calculer les intégrales de recouvrement qui correspondent aux expansions des fonctions radiales de base non-orthogonales. Par exemple, en substituant Eq 3.57 dans Eq 3.52 et en appliquant Eq 3.61 on obtient directement la factorisation

$$T^{(|m|)}_{n'l',nl}(R) = \sum_{j'}\sum_{j} X_{n'l'j'}\bar{T}^{(|m|)}_{j'l',jl}(R)X_{nlj}, \tag{3.72}$$

où chaque $\bar{T}^{(|m|)}_{j'l',jl}(R)$ est une intégrale de recouvrement dans une base non-orthogonale

$$\bar{T}^{(|m|)}_{j'l',jl}(R) = \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} M! e^{-R^2/4\lambda}(R^2/4\lambda)^{k/2} L_M^{(k+1/2)}(R^2/4\lambda) \tag{3.73}$$

maintenant avec $M = j + j' + (l + l' - k)/2$. Ceci revient à développer $R_{nl}(r)$ comme une somme de fonctions non-orthogonales, $\bar{R}_{nl}(r)$

$$R_{nl}(r) = (-1)^{n-l-1}\sum_{j} X_{nlj}\bar{R}_{jl}(r). \tag{3.74}$$

Il peut être montré que ces fonctions correspondent à la base non-orthogonale de GL proposée par (Chiu & Moharerrzadeh, 1999). Avec cette factorisation, les coefficients d'expansion translatés, a^R_{nlm}, dans la base orthogonale originale peuvent être calculés en utilisant la séquence

$$\bar{a}_{jlm} = \sum_{n} X_{nlj}a_{nlm}, \tag{3.75}$$

$$\bar{a}^R_{jlm} = \sum_{j'l'} \bar{T}^{(|m|)}_{jl,j'l'}(R)\bar{a}_{j'l'm}, \tag{3.76}$$

$$a^R_{nlm} = \sum_{j} X_{nlj}\bar{a}^R_{jlm}. \tag{3.77}$$

D'une façon semblable, des translations d'ETO peuvent être calculées dans une base non-orthogonale en substituant Eq 3.66 dans Eq 3.52 et en rassemblant les puissances de $1/(s^2+1)$ pour donner directement

$$U^{(|m|)}_{n'l',nl}(R) = \sum_{j'}\sum_{j} Y_{n'l'j'}\bar{U}^{(|m|)}_{j'l',jl}(R)Y_{nlj} \tag{3.78}$$

où les éléments non-orthogonaux de matrice, $\bar{U}^{(|m|)}_{j'l',jl}(R)$, sont donnés par

$$\bar{U}^{(|m|)}_{j'l',jl}(R) = \sum_{k=|l-l'|}^{l+l'} A_k^{(ll'|m|)} \sum_{q=0}^{M} \binom{M}{q} \frac{(-1)^{M+q}}{2^{J+1-q}(J+1-q)!}(\Lambda R)^k \hat{k}_{J-k-q+1/2}(\Lambda R) \tag{3.79}$$

avec $M = (l + l' - k)/2$ et maintenant $J = j + j' + l + l' + 2$.

3.4.5 Résultats numériques

En utilisant la propriété d'orthogonalité de matrice de rotation et à partir d'une inspection visuelle des surfaces tournées 2D et 3D moléculaires, il est relativement simple de vérifier si les expressions de rotation SH sont correctes et qu'une rotation arbitraire d'Euler préserve la longueur du vecteur des coefficients d'expansion quand on travaille avec l'arithmétique de "double précision" du langage C.

Afin de vérifier l'implémentation des calculs de matrice de translation, des résultats numériques des expressions analytiques ont été comparés à ceux d'une intégration numérique 2D d'Eq 3.42 en utilisant une grille régulière de 200×200 dans la plan $(r, \cos\theta)$ et aussi comparés avec une intégration 1D d'Eq 3.52 en utilisant 200 étapes dans β en employant un schéma log-numérique. Des détails complets sont donnés dans (Ritchie, 2005). Cette comparaison a montré qu'afin de réaliser une précision numérique satisfaisante pour tous les éléments de matrice jusqu'à *N=32* pour des calculs subséquents utilisant des instructions de double précision (15 chiffres décimaux) en C, il est nécessaire d'exécuter ces calculs en utilisant la bibliothèque mathématique de précision étendue GMP, en employant l'arithmétique de 160 bits pour les matrices de translation de GTO et en utilisant l'arithmétique de 192 bits pour les matrices de translation d'ETO. Les calculs des expansions non-orthogonales se sont avérés deux fois plus rapides qu'en utilisant des expansions orthogonales (Ritchie, 2005). Cependant, quand les matrices de translation sont nécessaires pour des intervalles réguliers fixes, il est plus efficace de les calculer une seule fois dans des bases orthogonales et de les sauvegarder sur disque pour les usages suivants. D'autre part, les expansions non-orthogonales seraient plus appropriées pour calculer des translations d'intervalles irrégulières ou imprévisibles pendant un calcul de minimisation, par exemple. De toute façon, parce que les espaces de recherche d'appariement de forme et d'amarrage peuvent être partitionnés en cinq degrés de liberté de rotation et un degré de liberté de translation, les expressions analytiques pour les matrices de translation développées ici offrent maintenant les mécanismes nécessaires pour calculer des corrélations analytiques SPF d'ordre supérieur jusqu'à six dimensions.

Chapitre 4

Applications de la biologie structurale des systèmes

4.1 Reconnaissance de formes moléculaires

Le terme “reconnaissance” est souvent utilisé pour décrire l’identification des régions aussi bien similaires que complémentaires d’une paire de protéines ou d’autres molécules (Cherfils & Janin, 1993; Dean & Callow, 1987; Katchalski-Katzir *et al.*, 1992; Vakser & Aflalo, 1994; Jones *et al.*, 1995). Afin qu’une paire de molécules soit considérée comme similaire, leur forme ainsi que leurs caractéristiques physico-chimiques (par exemple, les propriétés électrostatiques, l’hydrophobicité et les propensions des liaisons hydrogèniques) doivent s’apparier au niveau des régions d’intérêt. À l’inverse, ces propriétés doivent être contraposées de façon appropriée pour donner un arrangement complémentaire. Ainsi les tâches de rotation et de translation d’une paire de molécules, pour trouver des régions de surface moléculaire similaires ou complémentaires, peuvent être interprétées comme deux facettes étroitement liées d’un *problème de reconnaissance*. Si l’on suppose que les molécules sont rigides, alors les deux tâches sont caractérisées par un espace de recherche à six dimensions (6D), comme il est illustré dans la figure 4.1. Cependant, dans le cas de la similarité, si les deux molécules sont initialement situées à l’origine, la composante de translation sera souvent petite et deux des angles de rotation, par exemple, (β_A, γ_A), seront presque redondants.

4.1.1 Superposition de SPF des formes protéiques

Afin de démontrer la technique de corrélation de SPF, cette section décrit la superposition d’une paire de protéines similaires où chaque protéine est représentée comme une seule expansion de densité 3D de forme de GTO (Eq 2.125). En laissant $A(\underline{r})$ et $B(\underline{r})$ représenter les formes des protéines A et B respectivement et en plaçant chaque protéine initialement à l’origine, il est approprié de représenter

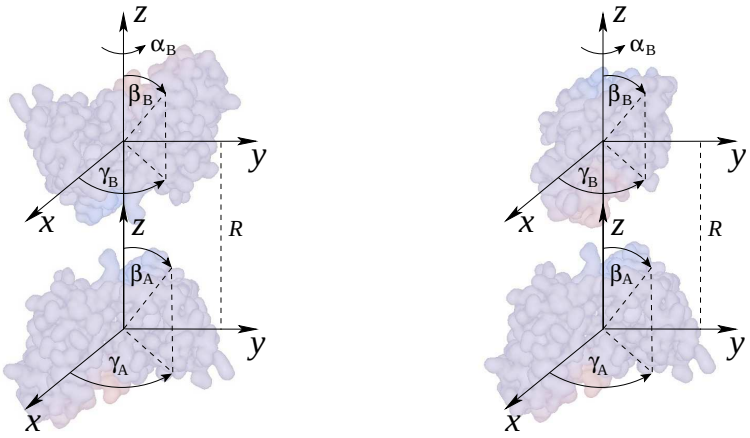

Figure 4.1: Une illustration schématique des espaces de recherche en corps rigide 6D d'appariement de forme (à gauche) et d'amarrage (à droite) en termes d'une coordonnée de translation, R, et de cinq coordonnées de rotation d'Euler, (β_A, γ_A) et, $(\alpha_B, \beta_B, \gamma_B)$, assignées au récepteur et au ligand, respectivement. La distance entre les centres moléculaires sera petite et deux des angles de rotation seront presque redondants.

la corrélation comme

$$E \equiv E(\beta_A, \gamma_A, R, \alpha_B, \beta_B, \gamma_B) = \int [\hat{T}_z(-R)\hat{R}(0, \beta_A, \gamma_A)A(\underline{r})][\hat{R}(\alpha_B, \beta_B, \gamma_B)B(\underline{r})]\,\mathrm{d}V, \quad (4.1)$$

où $\hat{T}_z(-R)$ translate l'$A(\underline{r})$ tournée par R le long de l'axe négatif z. La substitution de chaque terme dans la représentation de SPF donne

$$E = \sum_{nln'l'mm'm''}^{N} T^{(|m'|)}_{n'l',nl}(R) R^{(l')}_{mm'}(0, \beta_A, \gamma_A) a_{n'l'm'} R^{(l)}_{mm''}(\alpha_B, \beta_B, \gamma_B) b_{nlm''}. \quad (4.2)$$

Par conséquent, le but est de trouver les paramètres de rotation et de translation qui maximisent la somme ci-dessus. Dans cette section et les suivantes, la notation d'addition abrégée utilisée ci-dessus sera employée pour des raisons de brièveté, lorsque l'addition s'étendra sur les plages permises de tous les indices, par exemple $|m| \leq l < N$, etc.

Évidemment, le coût du calcul direct d'Eq 4.2 est de l'ordre de $O(N^7)$ opérations pour chaque orientation testée et il serait prohibitivement coûteux de calculer cette somme à plusieurs reprises dans une recherche 6D exhaustive. Cependant, une stratégie beaucoup plus efficace est de calculer la somme en plusieurs étapes en utilisant des éléments pré-calculés des matrices de rotation et de translation. Par exemple, en supposant que tous les éléments de la matrice de rotation ont été pré-calculés pour une certaine rotation (β, γ), un vecteur de $O(N^3)$ coefficients pour la molécule A peut être tourné en $O(N^3) \times O(N) = O(N^4)$ opérations en utilisant

$$a_{nlm}(\beta_A, \gamma_A) = \sum_{m'} R^{(l)}_{mm'}(0, \beta_A, \gamma_A) a_{nlm'}. \quad (4.3)$$

En supposant que tous les éléments de la matrice de translation ont aussi été pré-calculés, un vecteur des coefficients représentant une certaine orientation (β, γ, R) peut être calculé en $O(N^3) \times (O(N) + O(N^2)) = O(N^5)$ opérations en utilisant

$$a_{nlm}(R, \beta_A, \gamma_A) = \sum_{n'l'} T^{(|m|)}_{nl,n'l'}(-R) a_{n'l'm}(\beta_A, \gamma_A). \tag{4.4}$$

De même, les vecteurs des instances tournées de la molécule B peuvent être calculés en $O(N^3) \times O(N) = O(N^4)$ opérations par orientation en utilisant

$$b_{nlm}(\beta_B, \gamma_B) = \sum_{m'} R^{(l)}_{mm'}(0, \beta_B, \gamma_B) b_{nlm'}. \tag{4.5}$$

Le degré de liberté final est une rotation autour de l'axe z. Ceci pourrait être inclus dans la formule ci-dessus, mais puisque cette rotation peut être écrite comme

$$b_{nlm}(\alpha_B, \beta_B, \gamma_B) = \sum_{m'} R^{(l)}_{mm'}(\alpha_B, 0, 0) b_{nlm'}(\beta_B, \gamma_B) \tag{4.6}$$

$$= \sum_{m'} b_{nlm'}(\beta_B, \gamma_B) \cos m'\alpha_B + b_{nl\overline{m}'}(\beta_B, \gamma_B) \sin \overline{m}'\alpha_B, \tag{4.7}$$

il est plus efficace de calculer les quantités intermédiaires

$$P_m = \sum_{nl} a_{nlm}(\beta_A, \gamma_A) b_{nlm}(\beta_B, \gamma_B), \tag{4.8}$$

et

$$Q_m = \sum_{nl} a_{nlm}(\beta_A, \gamma_A) b_{nl\overline{m}}(\beta_B, \gamma_B), \tag{4.9}$$

et pour exécuter la corrélation en réitérant sur toutes les combinaisons d'orientation de (R, β_A, γ_A) et (β_B, γ_B) afin de calculer les rotations α_B en utilisant une série réelle de Fourier :

$$E = \sum_{m} P_m \cos m\alpha_B + Q_m \sin \overline{m}\alpha_B. \tag{4.10}$$

Pour des expansions d'ordre supérieur, le calcul d'Eq 4.10 sur de multiples échantillons angulaires, M, peut être exécuté en $O(M \log M)$ opérations en utilisant un FFT 1D. Cependant, quand *N<16*, il est plus rapide de calculer explicitement Eq 4.10 pour chaque valeur de α_B dans un temps de $O(MN)$. Quoiqu'il en soit, l'algorithme d'appariement de forme 6D peut en être implémenté comme une séquence imbriquée de transformations, ainsi qu'il est illustré dans la figure 4.2. En dépit du coût relativement élevé de rotation et de translation des différents vecteurs de coefficients 3D, chaque orientation distincte (R, β_A, γ_A) de la molécule A et chaque orientation distincte (β_B, γ_B) de la molécule B est calculée une fois seulement. Ainsi, le coût principal de l'algorithme de recherche de

begin 6D Superposition
  foreach (β_A, γ_A)
    calculate $\underline{a}(\beta_A, \gamma_A)$ using $a_{nlm'}(\beta_A, \gamma_A) = \sum_{m'} R^{(l)}_{mm'}(0, \beta_A, \gamma_A) a_{nlm'}$
  endfor
  foreach (β_B, γ_B)
    calculate $\underline{b}(\beta_B, \gamma_B)$ using $b_{nlm}(\beta_B, \gamma_B) = \sum_{m'} R^{(l)}_{mm'}(0, \beta_B, \gamma_B) b_{nlm'}$
  endfor
  foreach R
    load $T(R)$ from disc
    foreach (β_A, γ_A)
      calculate $\underline{a'}$ using $a'_{nlm} = \sum_{n'l'} T^{(|m)|}_{nl,n'l'}(R) a_{nlm}(\beta_A, \gamma_A)$
      foreach (β_B, γ_B)
        calculate $\underline{P}$ using $P_m = \sum_{m'} a'_{nlm'} b_{nlm'}(\beta_B, \gamma_B)$
        calculate $\underline{Q}$ using $Q_m = \sum_{m'} a'_{nlm'} b_{nl\overline{m'}}(\beta_B, \gamma_B)$
        foreach α_B
          calculate $c[\alpha_B] = \sum_m (P_m \cos m\alpha_B + Q_m \sin \overline{m}\alpha_B)$
        endfor
        save $C[R, \beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B] = \max_{\alpha_B}(c[\alpha_B])$
      endfor
    endfor
  endfor
end

Figure 4.2: Un résumé du pseudo-code de l'algorithme de recherche de superposition 6D. Des tessellations icosaèdriques de la sphère sont utilisées pour produire un modèle presque régulier des échantillons de rotation (β, γ) pour le récepteur (A) et le ligand (B), respectivement. Des vecteurs des coefficients tournés sont calculés et sauvegardés une seule fois pour chaque échantillon de rotation. Les vecteurs de coefficients de récepteur sont alors translatés pour chaque étape de recherche translationnelle et une itération interne sur l'angle de torsion est effectuée pour chaque paire de vecteurs de récepteur et de ligand pour compléter la recherche 6D. Les tableaux de $\cos m\alpha_B$ et $\sin m\alpha_B$ sont aussi pré-calculés en dehors de la boucle principale. Puisque la recherche itère sur un nombre discret d'étapes de rotation et de translation, les corrélations finales peuvent être sauvegardées en utilisant un seul nombre entier comme identifiant (détails non montrés).

superposition est une itération combinatoire de l'ordre $O(N^2)$ pour comparer des paires de vecteurs transformés de A et de B, où chacun a un cycle interne autour de l'angle de torsion, α_B.

La figure 4.3 montre quelques représentations GTO de densité stériques pour une paire de domaines globulaires de protéines, le superantigène exotoxine A1 de *Streptococcique* pyrogène (Code PDB 1B1Z) (Papageorgiou *et al.*, 1995) et l'exotoxine SEC3 de *Staphylococcus* aureus (Code PDB 1JCK) (Fields *et al.*, 1996). Ces protéines globulaires ont une identité de séquence relativement basse de 46% mais partagent un repliement très similaire. Par conséquent, elles peuvent être bien superposées par la méthode conventionnelle d'ajustement par les moindres carrés des coordonnées des atomes conservés de C_α. Cependant, dans cette illustration, la superposition a été exécutée en

maximisant le recouvrement entre les expansions de densité stériques de GTO respectives avec des corrélations de *N=6*. Une superposition presque-identique (non montrée) a été également réalisée en corrélant des densités de charge électrostatique dans la base d'ETO. À partir de la figure 4.3, on peut voir que les expansions de GTO d'ordre supérieur capturent remarquablement bien la forme détaillée de chaque protéine, bien que les expansions d'ordre inférieur encodent toujours les informations suffisantes pour permettre de calculer une très bonne superposition globale. La superposition montrée a été calculée en cherchant parmi quelques 21×10^6 essais d'orientation produits à partir de 162 échantillons (β, γ) angulaires icosaèdriques pour chaque protéine, 128 échantillons de torsion dans α_B et 40 pas de distance de 0.25Å. Tous les calculs ont pris environ 8 secondes sur un processeur Pentium Xeon de 2GHz.

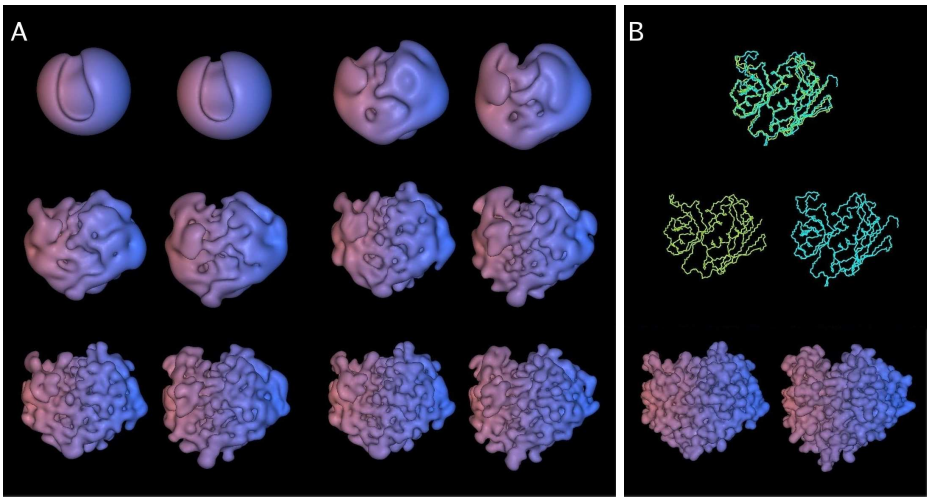

Figure 4.3: Une illustration de la représentation de la forme GTO et de la superposition d'une paire de protéines globulaires, les superantigènes SpeA et SEC3. (A) Du haut à gauche au bas à droite : les fonctions stériques de densité de SpeA et de SEC3 montrées aux ordres d'expansion de *N=6, 12, 16, 20, 25* et *30*. Chaque paire est dans la même orientation superposée, séparée horizontalement pour plus de clarté, avec SpeA sur la gauche et SEC3 sur la droite. Pour la visualisation, chaque forme de surface moléculaire a été entourée par la fonction de densité 3D. Les expansions jusqu'à *N=32* sont visuellement quasi indifférenciables des expansions *N=30* et ne sont donc pas montrées ici. (B) Les traces correspondantes du squelette de la superposition avec SpeA en jaune et SEC3 en cyan (haut), les squelettes séparés avec SpeA sur la gauche et SEC3 sur la droite (centre) et les surfaces moléculaires originales desquelles les représentations de densité de GTO ont été dérivées (bas).

4.1.2 Clustering des structures protéiques des super-familles de CATH

Le logiciel d'alignement de séquences, BLAST (Altschul *et al.*, 1990), est probablement connu de tous les biologistes comme un outil standard pour faire des requêtes dans les bases de données génomiques de séquences de nucléotides ou d'acides aminés. Les biologistes effectuent souvent des alignements de séquences comme une première étape pour déterminer la fonction et l'origine

évolutionnaire d'une protéine ou d'un gène inconnu. Cependant, ceci exige normalement qu'il y ait au moins une similarité de 20% entre une séquence connue et la séquence en question[1]. Néanmoins, les protéines peuvent avoir des similarités de séquence inférieures alors que leurs structures partagent globalement le même pli. En d'autres mots, dans la nature, les structures de protéine sont plus conservés que les séquences de ces protéines. Par conséquent, aligner et réaliser un clustering des structures de protéines constitue une façon utile d'analyser les relations fonctionnelles et évolutives entre les protéines, même lorsqu'elles ont une similarité de séquence très basse (Kolodny *et al.*, 2005). Toutefois, tandis que les techniques symboliques, telles que les algorithmes de Smith-Waterman et de Needleman-Wunsch, pour aligner les séquences de protéines sont devenus des outils standard en bioinformatique, la meilleure façon d'aligner les structures 3D des protéines reste une question ouverte (Sippl & Wiederstein, 2008). La plupart des algorithmes d'alignement structuraux existants sont basés sur des comparaisons des traces de squelettes C_α ou des vecteurs formés par les atomes C_α-C_β de chaque acide aminé à l'exception de la glycine, par exemple. Cependant, ces approches ont un coût significativement plus élevé que les techniques symboliques parce qu'elles impliquent des calculs de multiples matrices de rotation en moindres carrés qui sont coûteux dans le contexte des algorithmes d'alignement en programmation dynamique. Par exemple, les algorithmes courants d'alignement structuraux tels que SSM (Taylor & Orengo, 1989), DALI (Holm & Sander, 1991), SAP (Taylor, 1999), CE (Shindyalov & Bourne, 1998) et VAST (Madej *et al.*, 1995) prennent normalement plusieurs secondes de temps CPU par paire d'alignements (Kolodny *et al.*, 2005). Par conséquent, il y a un besoin de développer des méthodes rapides indépendantes de la séquence pour comparer les structures de protéines.

Dans le cadre de ma convention ANR Chaire d'Excellence au LORIA, Lazaros Mavridis est actuellement employé comme assistant post-doctoral sur le projet "3D-Blast", dont l'objectif est d'appliquer des techniques de SPF 3D pour accomplir la tâche précédente. Pour une évaluation préliminaire de notre approche, nous avons effectué quelques expériences sur l'alignement et le clustering de plusieurs protéines sélectionnées parmi la base de données de structures protéiques, CATH (Orengo *et al.*, 1997; Cuff *et al.*, 2008). Lazaros a récemment présenté ce travail lors de la rencontre satellite 3DSIG de la conférence internationale sur les systèmes intelligents dans la biologie moléculaire (ISMB2009-3DSIG), et un article a été accepté au colloque du pacific symposium de bio-informatique (PSB-2010). Le reste de cette section et la prochaine section résument les résultats obtenus jusqu'à présent.

Bien que la section 4.1.1 décrive comment superposer une paire de protéines en maximisant le recouvrement entre leurs volumes de van der Waals, il est approprié d'utiliser une fonction de score normalisée pour comparer plusieurs protéines. Par conséquent, les calculs suivants utilisent un score

[1]On dit que parfois des protéines avec des identités de séquence entre 20% and 35% sont dans la zone crépusculaire, ou "twilight zone," entre les structures semblables et non-semblables (Rost, 1999)

de similarité de Carbo (Carbo *et al.*, 1980) calculé comme :

$$S = \frac{\underline{a}.\underline{b}'}{|\underline{a}|.|\underline{b}|}, \tag{4.11}$$

où $\underline{b}'$ dénote le vecteur des coefficients forme-densité tournés de la molécule B, $|\underline{b}|$ dénote la norme de ce vecteur et la notation abrégée du produit scalaire est

$$\underline{a}.\underline{b}' = \sum_{nlm}^{N} a_{nlm} b'_{nlm}, \tag{4.12}$$

Cette fonction de score de type cosinus donne les valeurs qui s'étendent de zéro (aucune similarité) à un (deux protéines identiques s'alignant parfaitement).

Dans la classification CATH, des structures de protéines sont assignées à une super-famille selon la classe, l'architecture, la topologie et l'homologie de leur pli. Ceci est essentiellement un schéma hiérarchique, dont la classe supérieure se compose de quatre types possibles de plis : All-α (i.e. la structure est composée entièrement d'éléments hélices-α de la structure secondaire), All-β (la structure est composée entièrement d'éléments feuillets-β de la structure secondaire), α+β (la structure contient à la fois des hélices-α et des feuillets-β) et "irregular" (aucun élément de structure secondaire identifiable). Chacun des quatre niveaux de la hiérarchie CATH est identifié par un code numérique. De plus, la classification CATH nomme chaque protéine selon les quatre lettres de son code PDB, la lettre de sa chaîne et le nombre de domaines qu'elle contient (par exemple, 1IOMA02). Pour chaque expérience de clustering, cinq ou six super-familles avec la même architecture ont été sélectionnées de façon à donner environ 30 structures de protéines pour chaque classe de plis dans CATH. Par conséquent, le but de ces expériences est d'évaluer l'efficacité de notre approche pour identifier les protéines qui ont la même topologie et sont homologues au sein d'une certaine architecture de plis. Les détails des super-familles de CATH utilisées ici sont donnés dans le tableau 4.1.

Pour la classe All-α, cinq super-familles de CATH ont été sélectionnées et énumérées dans le tableau 4.1. Pour chaque paire de protéines dans cet ensemble, une recherche de corrélation a été exécutée pour trouver l'orientation qui donne des scores maxima de similarité de Carbo (Eq 4.11). Le clustering agglomératif de Ward (1963) a été ensuite appliqué aux scores de similarité deux à deux pour donner un total de cinq clusters. Les résultats du clustering dans la figure 4.4 montrent que les super-familles 1.10.230.10 et 1.10.167.10 ont été correctement réparties en deux clusters distincts. Bien qu'il existe une certaine différence dans les clusters produits pour les autres super-familles, on peut voir qu'il y a toujours un très bon accord entre les clusters de SPF calculés et la hiérarchie CATH. L'exception la plus notable est le fait que suposin (Code PDB 1N69) se soit groupé avec la super-famille 1.10.30.10. Une inspection visuelle de la figure 4.4 montre que le pli global de suposin ressemble plus à la super-famille 1.10.30.10 que 1.10.225.10, attribuée par CATH. Ceci suggère que la classification automatique de SPF pourrait potentiellement aider les curateurs de CATH à résoudre les cas anormaux ou ambigus.

Table 4.1: Les 23 super-familles de CATH utilisées pour nos expériences de clustering de protéines.

Class + Architecture	Topology + Homology	Protein Name and Function	Representative Structure
All-α Orthogonal Bundle (1.10)	230.10	Citrate synthase	1iomA02
	120.10	Trypsin/Alpha-Amylase Inhibitor	1beaA00
	225.10	NK – Lysin	1l9lA00
	167.10	G-Protein Signalling Regulator	1dk8A02
	30.10	HMG DNA Binding Domain	1qrvA00
All-β Ribbon (2.10)	109.10	LexA repressor	1jhfB00
	150.10	Urease	1ejxB00
	110.10	LIM domain PINCH protein	1g47A00
	77.10	Hemagglutinin	1jsdA02
	160.10	Endothelial Growth Factor 165	1kmxA00
	10.10	PDC-109	1h8pA02
α+β Roll (3.10)	130.10	Ribonuclease A	1dy5A00
	170.10	Elastase	1u4gA01
	150.10	DNA Polymerase	1ok7A01
	110.10	Ubiquitin Conjugating Enzyme	2grrA00
	120.10	Flavocytochrome B2	1cyoA00
Irregular (4.10)	280.10	MYOD Helix-Loop-Helix Domain	1nlwE00
	290.10	Bacteriorhodopsin Fragment	1bctA00
	410.10	Factor Xa Inhibitor	1g6xA00
	490.10	HiPIP	1iuaA00
	400.10	LDLR	2fcwB02
	320.10	Dihydrolipoamide Transferase	1w85I00

Pour la classe All-β, six super-familles ont été groupées. Comme on le voit dans la figure 4.5, le clustering SPF distingue correctement les six groupes, mais deux protéines sont mal placées selon la classification de CATH. Ce sont le domaine de la carboxy-terminal LIM (Code PDB 1CTL) et l'hémagglutinine du virus influenza (Code PDB 2VIR) qui sont groupés avec le domaine singleton héparine-liant (Code PDB 1KMX). Ceci semble se produire parce que 1CTL et 2VIR sont moins similaires, selon les scores calculés, aux autres membres de leurs super-familles de CATH respectives. En outre, ils sont groupés avec 1KMX en grande partie parce que les trois protéines ont un volume stérique semblable.

Pour la classe α+β, cinq super-familles de CATH ont été groupées. La figure 4.6 montre que les super-familles 3.10.130.10 et 3.10.120.10 sont correctement réparties en deux groupes. Les trois super-familles restantes (3.10.110.10, 3.10.150.10 et 3.10.170.10), présentent un cas semblable aux résultats All-β, dans lequel un groupe super-famille (3.10.110.10) est divisé en deux et deux groupes super-famille (3.10.170.10 et 3.10.150.10) sont fusionnés en un. Néanmoins, en dépit de ces différences, la cohérence globale du clustering SPF avec la hiérarchie CATH est évidemment très bonne.

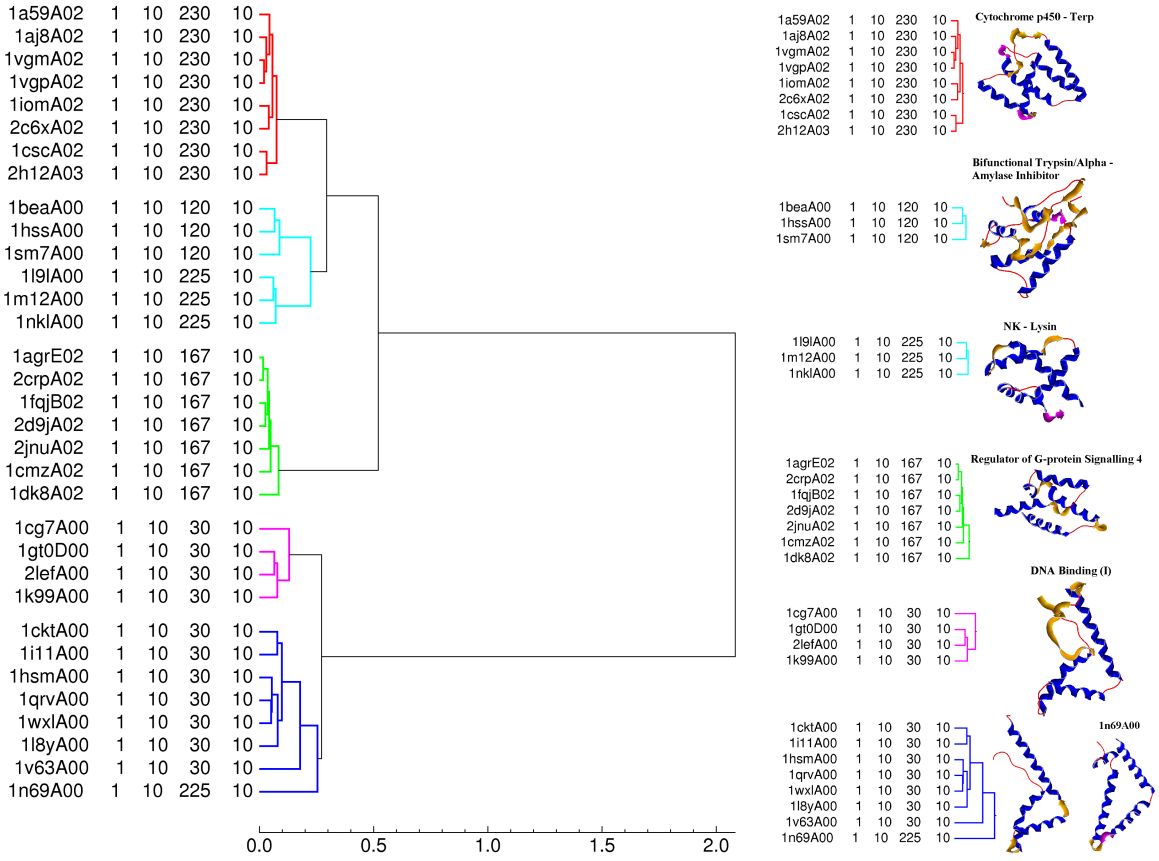

Figure 4.4: Les résultats du clustering SPF de la classe All-α, utilisant *N=6* avec cinq clusters.

Pour la classe Irregular, six super-familles ont été groupées. Comme l'exemple de All-β, le clustering SPF est complètement cohérent avec la hiérarchie CATH. Cependant, deux protéines sont mal placées par rapport à la classification CATH, en particulier la proteine bikunin du complexe inter-alpha-inhibiteur (Code PDB 1BIK) et la protéine d'anticoagulant tique (Code PDB 1D0D), qui sont groupées avec la super-famille 4.10.490.10. Ceci semble être dû à la différence de taille entre ces protéines et le reste de la super-famille des inhibiteurs de facteur Xa. Par exemple, la figure 4.7 montre que le bikunin a une répétition du même motif que celui des autres inhibiteurs de facteur Xa. Par conséquent, il est stériquement trop grand pour être groupé avec les autres inhibiteurs de Xa et il est plutôt placé avec les protéines plus grandes de la super-famille 4.10.490.10.

4.1.3 Recherche dans la base de structures protéiques de CATH

En guise de deuxième test d'utilité de notre approche, la base entière de données CATH, qui contient environ 12.000 structures, a été interrogée en utilisant des fonctions de densité de SPF comme

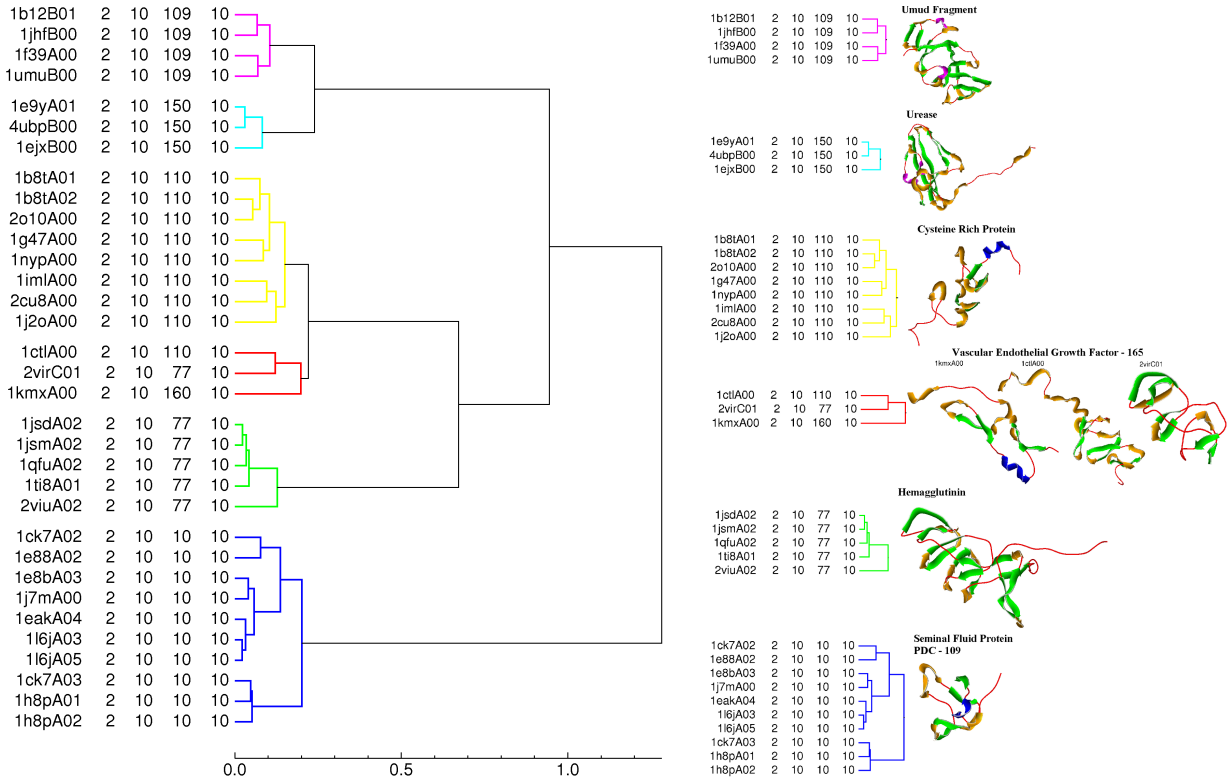

Figure 4.5: Les résultats du clustering SPF de la classe All-β, utilisant *N=6* avec six clusters.

requêtes. Lorsqu'on fait une requête pour trouver des structures de protéines qui sont similaires à une structure-requête dans une si grande base de données, il serait souhaitable de pouvoir éliminer rapidement plusieurs protéines qui ont des formes fondamentalement différentes de la requête et qui ne peuvent pas avoir une bonne superposition quelle que soit l'orientation. En notant que les coefficients d'expansion ayant les mêmes valeurs de l'index m se transforment sous une rotation, il est normal d'utiliser l'interprétation de vecteurs de coefficients SH pour construire des empreintes invariantes à la rotation ("rotationally invariant fingerprints," RIFs) comme :

$$A_n = \sum_{l=0}^{n-1} \sum_{m=-l}^{m=l} a_{nlm}^2. \tag{4.13}$$

Si les coefficients a_{nlm} définissent la densité de forme d'une protéine, les descripteurs invariants à la rotation A_n encodent ensemble la distribution de masse radiale de la protéine. Par analogie à Eq 4.11, le score de similarité de RIF est écrit comme :

$$S_{RIF} = \frac{\sum_{n=1}^{N} A_n B_n}{\left(\sum_{n=1}^{N} A_n^2\right)^{1/2} \left(\sum_{n=1}^{N} B_n^2\right)^{1/2}}. \tag{4.14}$$

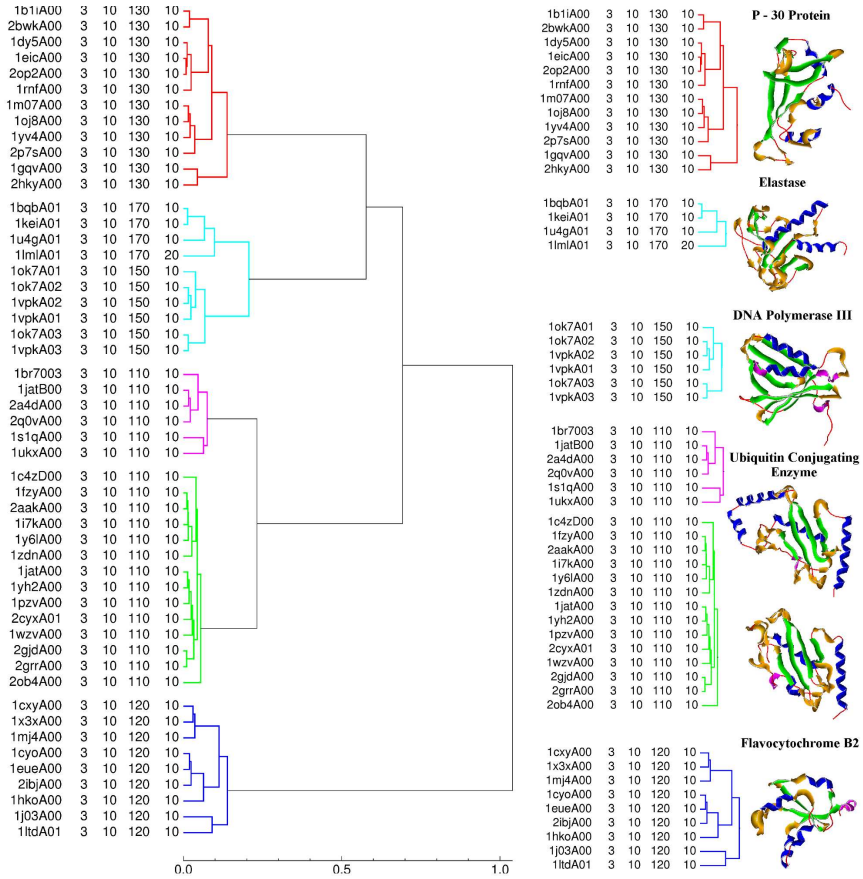

Figure 4.6: Les résultats du clustering SPF de la classe α+β, utilisant *N=6* avec cinq clusters.

Pour l'expérience initiale de recherche dans la base de données, l'asparagine synthétase (Code PDB 12AS, CATH super-famille 3.30.930.10) a été sélectionnée comme structure-requête. La super-famille 3.30.930.10 a 27 membres, qui sont traités en tant que "positifs" tandis que les protéines restantes dans la base de données sont traitées en tant que "négatives" par rapport à la requête. Si une fonction de score pouvait reproduire exactement la classification CATH, les 27 positifs apparaîtraient parmi les premiers dans la liste classée. Cependant un tel résultat parfait est rarement observé en pratique. Par conséquent, les courbes de caractéristique d'opérateur-récepteur (ROC) (Egan, 1975; Fawcett, 2006) sont utilisées pour analyser objectivement les caractéristiques de précision/rappel des fonctions de score. Dans une analyse ROC, chaque élément de la liste classée est considéré à son tour et le nombre de positifs et de négatifs dans les sous-listes de chaque côté de l'élément en question est compté. Ici, la sous-liste ayant une similarité élevée se nomme la "hit list". Un vrai positif ("true positive," TP) est assigné quand un élément dans la hit list contient un positif original ou un faux positif ("false positive," FP) est assigné si cet élément contient un négatif. À l'inverse,

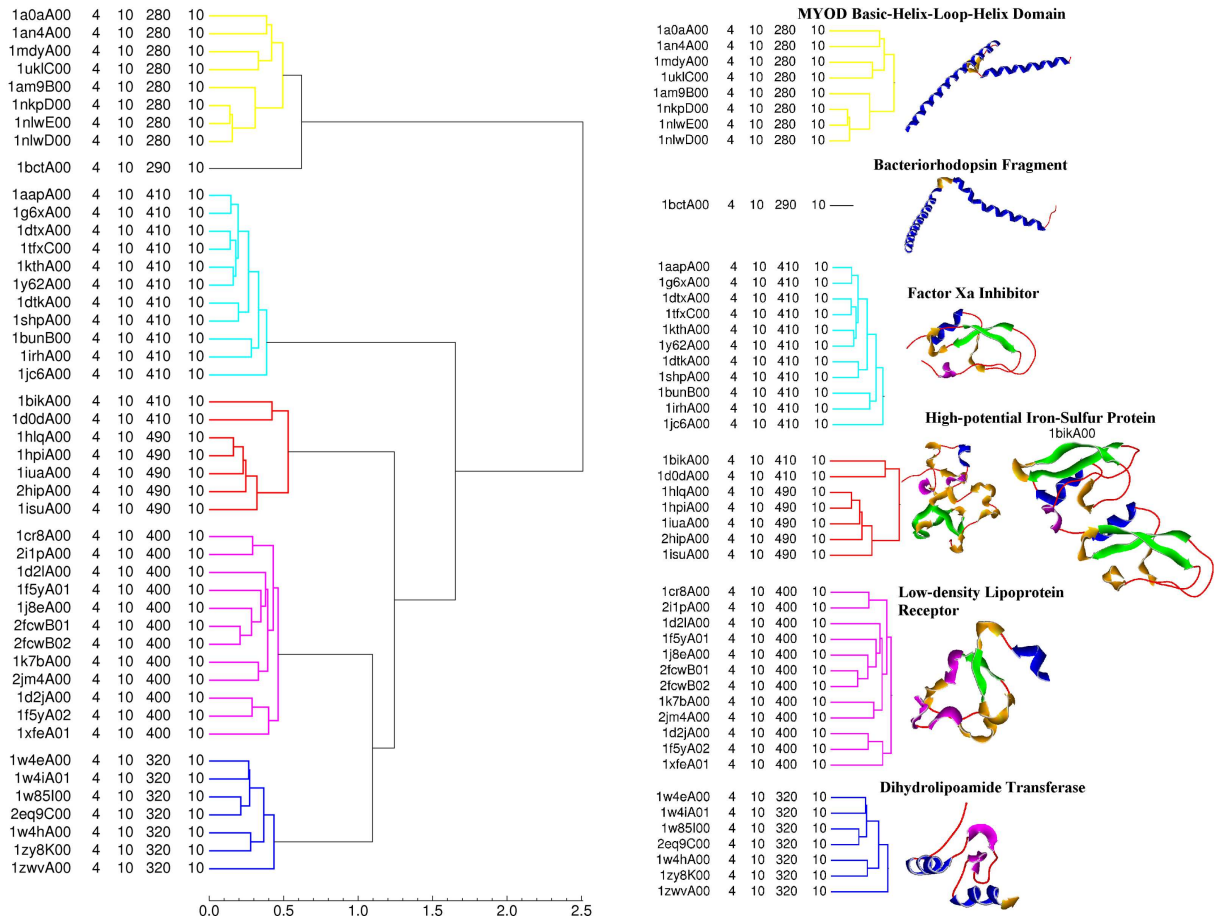

Figure 4.7: Les résultats du clustering SPF de la classe Irregular, utilisant *N=6* avec six clusters.

un vrai négatif (“true negative,” TN) est assigné quand un négatif original se trouve à l’extérieur de la hit list et un faux négatif (“false negative,” FN) est assigné si cette position est occupée par un membre positif. Les courbes ROC sont produites en traçant le taux de vrais positifs (“true positive rate,” TPR) sur l’axe y contre le taux de faux positifs (“false positive rate,” FPR) sur l’axe x, où TPR et FPR sont donnés par :

$$TPR = \frac{TP}{TP + FN} \tag{4.15}$$

et

$$FPR = \frac{FP}{FP + TN}. \tag{4.16}$$

La qualité d’une fonction de score peut être rapidement évaluée par la forme d’une courbe ROC. Par exemple, une fonction de score aléatoire donnerait une ligne diagonale (TPR=FPR), tandis qu’une

fonction de score parfaite donnerait un trait horizontal (TPR=1) avec une valeur maximum pour l'aire sous la courbe (ou "area under the curve," AUC=1).

Puisqu'il est relativement lent de calculer des corrélations d'ordre supérieur, différents paramètres de rotation et invariants à la rotation ont été testés pour explorer la mesure pour laquelle des requêtes peuvent être exécutées plus rapidement sans sacrifier l'exactitude. Dans chaque cas, la structure de l'asparagine synthétase (Code PDB 12AS, CATH super-famille 3.30.930.10) fut superposée sur chaque protéine dans la base de données en utilisant des recherches de corrélation de SPF *N=6* et ces structures de base de données étaient classées par similarité par rapport à la requête. La figure 4.8 montre les courbes résultantes de ROC obtenues pour une plage des ordres d'expansion en interrogeant la base de données CATH. Cette figure montre que l'approche globale donne une précision et un rappel très bons pour tous les ordres d'expansion au-dessus de *N=2* et qu'il y a un avantage très minime à utiliser des ordres d'expansion de *N=6* ou plus. La figure 4.9 montre les 27 membres de cette super-famille qui sont traités comme des TP par rapport à la requête. Pour une analyse poussée des résultats, les TP sont groupés en 5 groupes. La requête appartient au groupe 1 et tous les s groupes 2 et 3 ont rents des hélices-α. 20% de la base de d iers 30%. Pour finir, pplémentaire.

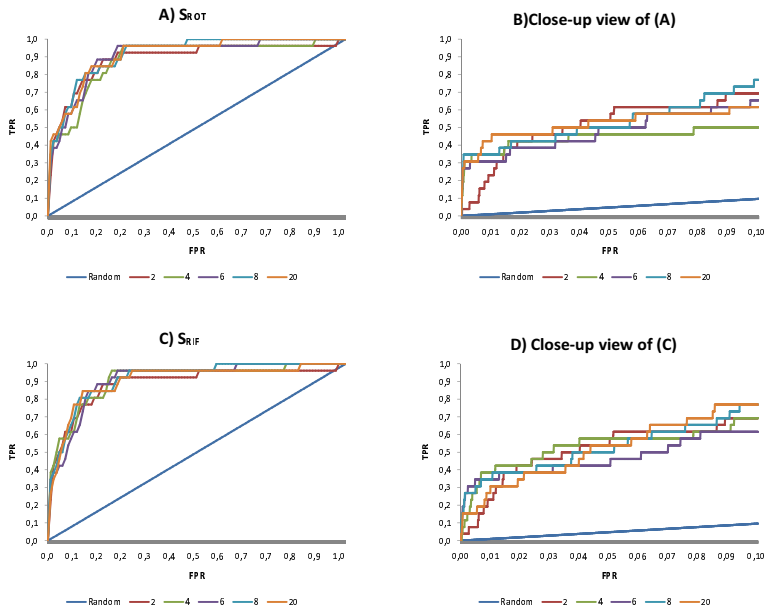

Figure 4.8: Les analyses des courbes ROC obtenues en interrogeant la base entière de données CATH avec la structure 12asA00. A : (ROT) fonction de score rotation-dépendante, Eq 4.11; B : vue en gros plan de (A); C : fonction de score RIF, Eq 4.14; D : vue en gros plan de (C).

La figure 4.8 montre aussi les résultats pour la fonction de score RIF. Comparée à la fonction de score rotation-dépendante, la fonction RIF présente des performances générales remarquables. Cependant les deux fonctions se comportent plutôt différemment sur les premiers pourcentages de

la base de données. Par exemple, les recherches de rotation donnent un TPR d'environ 40% sur le premier 0.1% de la base de données, tandis que les recherches RIF donnent un TPR d'environ 10% seulement. Par conséquent, la fonction RIF n'est pas suffisamment sensible pour être utilisée toute seule mais elle pourrait constituer un pré-filtre utile et rapide sur la base de données de sorte que la fonction rotation-dépendante, plus coûteuse, puisse être appliquée seulement aux candidats les plus prometteurs.

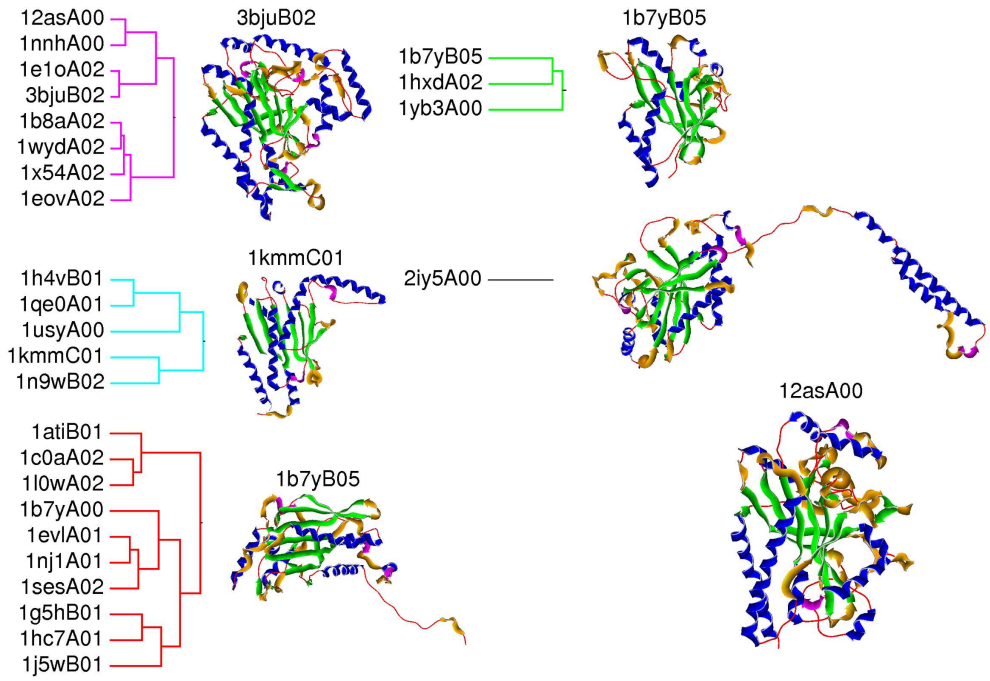

Figure 4.9: Le clustering de la super-famille 3.30.930.10 en cinq groupes. Le membre représentatif de chaque groupe est montré avec la protéine-requête 12asA00.

Afin de tester la notion ci-dessus, des recherches dans la base de données de CATH ont été faites en utilisant les fonctions de score de RIF, de rotation et les deux en tandem en utilisant plusieurs structures de protéines comme requêtes : l'asparagine synthétase, la protéine ALF4-activated Giα1 (Code PDB 1AGR), la protéine cystéine-riche de la poule (Code PDB 1B8T), transférase d'acétyle de dihydrolipoic-lysine-résidu (Code PDB 1W4E), et UbcH7 (Code PDB 1C4Z). En utilisant un pré-filtre RIF avec un seuil de similarité de 0.99, qui choisit parmi les 2% à 15% de la base de données pour le re-classement de rotation, chaque recherche en tandem prend moins de 10 minutes comparées à 75 minutes pour des recherches complètes de rotation sur un processeur Pentium Xeon de 2.3GHz. La figure 4.10 montre les AUC pour les 1% meilleurs de la base de données pour les recherches de rotation, de RIF et les deux en tandem. Cette figure montre que les recherches en tandem réalisent le même niveau de performance élevé que les recherches de rotation. Le tableau 4.2 présente les AUC

globaux obtenus pour ces recherches. Les valeurs très élevées dans ce tableau confirment l'utilité des fonctions de score SPF.

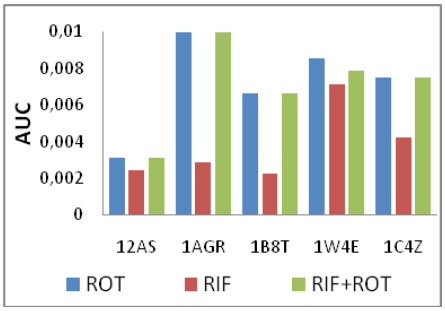

Figure 4.10: Les AUC des courbes ROC obtenues en interrogeant la base entière de données CATH en utilisant les cinq protéines sélectionnées comme requêtes, mais en considérant seulement les meilleurs 1% des résultats trouvés par les fonctions de score de rotation (ROT), RIF et les deux en tandem (RIF+ROT).

Table 4.2: Un résumé des valeurs des AUC obtenues en interrogeant la base entière de données CATH.

Query	RIF	ROT	RIF+ROT
12AS	0.944	0.907	0.929
1AGR	0.960	1.000	1.000
1B8T	0.964	0.983	0.997
1W4E	0.995	0.999	0.997
1C4Z	0.968	0.995	0.995

Ce tableau présente les valeurs des AUC obtenues en interrogeant la base entière de données CATH avec les structures de protéines données (énumérées par leur code PDB) comme requêtes pour les recherches de RIF, de rotation (ROT) et les deux en tandem (RIF+ROT).

En général, les résultats ci-dessus montrent que les expansions SPF à basse résolution fournissent une méthode séquence-indépendante fiable et rapide pour superposer et comparer des structures de protéines. Nous croyons que l'approche SPF pourrait fournir une méthode automatique et objective pour améliorer la qualité des classifications de structures de protéines et nous travaillons afin de créer une interface web pour interroger des structures de protéines en temps réel.

4.2 Corrélations SPF pour l'amarrage protéique

Afin de représenter la *complémentarité de forme* de protéines d'une manière appropriée pour calculer des corrélations SPF d'amarrage, il est utile de définir une fonction de "skin surface" $\sigma(\underline{r})$ comme une fonction de forme-densité qui décrit le volume englobé par la surface de VDW et la surface de SAS

de chaque protéine. En d'autres termes,

$$\sigma(\underline{r}) = \begin{cases} 1 \text{ si } \underline{r} \in \text{ skin surface,} \\ 0 \text{ autrement.} \end{cases} \tag{4.17}$$

Cette fonction de densité peut être calculée numériquement sur une grille de manière similaire au calcul de la fonction de densité de VDW, $\tau(\underline{r})$. L'utilisation d'une représentation de skin surface pour modéliser la complémentarité de forme de protéines est justifiée par la figure 4.11. Cette figure suggère qu'une bonne stratégie pour trouver des orientations complémentaires entre une paire de protéines est de maximiser le recouvrement entre la densité intérieure d'une des protéines avec la skin extérieure de l'autre. Des chocs (ou "clashes") stériques peuvent être pénalisés avec un terme de pénalité du recouvrement de forme-densité intérieure-intérieure. En suivant ces idées, le score de complémentarité de forme, E, pour les protéines A et B peut être écrit comme

$$E = \int \sigma_A \tau_B \mathrm{d}V + \int \tau_A \sigma_B \mathrm{d}V - \int Q \tau_A \tau_B \mathrm{d}V, \tag{4.18}$$

où $\sigma_A \equiv \sigma_A(\underline{r}_A)$ etc. et où Q est un facteur de pénalité positif d'intérieure-intérieure. J'utilise actuellement $Q = 11$. Quand la SAS est calculée avec une sonde de rayon 1.4Å les deux premiers termes donnent une expression pour le volume d'eau expulsée des surfaces de protéines pendant leur liaison (voir la figure 4.11). Avec un facteur d'échelle approprié, le volume expulsé peut être utilisé comme une approximation de premier ordre pour l'énergie libre d'association hydrophobe (Richmond, 1984). En mettant $Q_B = \sigma_B - Q\tau_B$, Eq 4.18 peut être écrite comme une expression de pseudo-énergie de deux termes

$$E_{\text{SHAPE}} = K \int \big(\sigma_A \tau_B + \tau_A Q_B\big) \mathrm{d}V, \tag{4.19}$$

où K est une constante négative qui donne des scores négatifs pour des orientations favorables.

4.2.1 Corrélations FFT 1D pour l'amarrage protéique

D'après la section 2.5.2, l'énergie d'interaction électrostatique *in vacuo* d'une paire de protéines avec les densités de charge $\rho_A(\underline{r})$ et $\rho_B(\underline{r})$ et avec des potentiels électrostatiques $\phi_A(\underline{r})$ et $\phi_B(\underline{r})$, respectivement, est donnée par

$$E_{\text{ELEC}} = \frac{1}{2} \int \big(\rho_A(\underline{r})\phi_B(\underline{r}) + \rho_B(\underline{r})\phi_A(\underline{r})\big) \mathrm{d}V. \tag{4.20}$$

Donc, en représentant chaque fonction comme une expansion sphérique polaire d'ETO à l'origine avec les fonctions radiales $S(r)$, Eq 2.101 donne immédiatement l'expression de corrélation électrostatique

$$E_{\text{ELEC}}(R, \beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B) = \frac{1}{2} \sum_{nlm}^{N} \big(a'^{\rho}_{nlm} b'^{\phi}_{n'l'm'} + a'^{\phi}_{nlm} b'^{\rho}_{n'l'm'}\big), \tag{4.21}$$

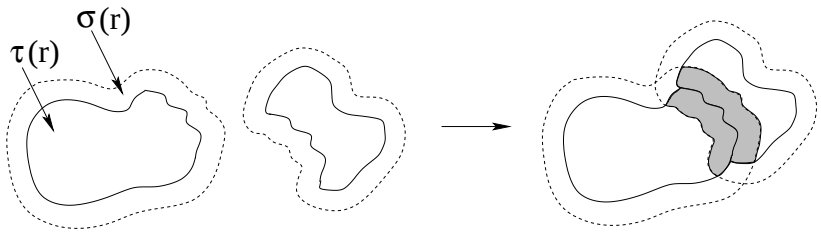

Figure 4.11: Une illustration schématique de la complémentarité de forme en utilisant des fonctions de densité 3D. Ici, les lignes continues représentent les surfaces de VDW et les lignes pointillées représentent les SAS. La fonction $\tau(\underline{r})$ est définie comme l'unité dans la surface de VDW et zéro partout ailleurs; $\sigma(\underline{r})$ est définie comme une unité dans le volume englobé par la SAS et la surface de VDW et zéro partout ailleurs. Quand une paire de protéines est assemblée dans un arrangement complémentaire, il y a un grand recouvrement (région dégradée) entre $\sigma(\underline{r})$ de l'une des protéines et $\tau(\underline{r})$ de l'autre et *vice-versa*.

où a'^{ρ}_{nlm} et a'^{ϕ}_{nlm} dénotent les coefficients d'expansion tournés et translatés de charge-densité et de potentiel électrostatique de la protéine A, etc., selon les équations 4.3–4.6.

De même, à partir d'Eq 4.18, la pseudo-énergie de la forme-densité peut être développée comme

$$E_{\text{SHAPE}}(R,\beta_A,\gamma_A,\alpha_B,\beta_B,\gamma_B) = K\sum_{nlm}^{N}\left(a'^{\sigma}_{nlm}b'^{\tau}_{n'l'm'} + a'^{\tau}_{nlm}b'^{Q}_{n'l'm'}\right). \tag{4.22}$$

Donc une pseudo-énergie globale de forme et d'électrostatique pour le système peut être calculée comme

$$E_{\text{TOTAL}}(R,\beta_A,\gamma_A,\alpha_B,\beta_B,\gamma_B) = E_{\text{ELEC}}(R,\beta_A,\gamma_A,\alpha_B,\beta_B,\gamma_B)+E_{\text{SHAPE}}(R,\beta_A,\gamma_A,\alpha_B,\beta_B,\gamma_B). \tag{4.23}$$

Cette expression de corrélation d'amarrage a été implémentée comme une série imbriquée d'une FFT 1D, ainsi qu'il est décrit dans la section 4.1.1. Pour obtenir des résultats d'amarrage satisfaisants, il est nécessaire d'utiliser des incréments de recherche angulaire (β,γ) d'environ 7.5° produits à partir des tessellations icosaèdriques de 812 sommets et d'utiliser des expansions SPF d'ordre d'au moins *N=25*. Cependant, puisque l'ordre de l'expansion peut être varié indépendamment de la taille de l'incrément de recherche angulaire, le calcul peut être accéléré de manière significative en effectuant une recherche de "forme-seulement" de basse résolution dans l'espace de recherche en utilisant *N=16* et en re-classant seulement les 20.000 meilleures orientations avec des corrélations de formes et d'électrostatiques d'ordre élevé avec *N=25*, par exemple. Cette approche a été utilisée sur un certain nombre de complexes protéiques et les résultats obtenus ont été publiés (Ritchie & Kemp, 2000). Quand les interactions électrostatiques sont importantes dans un complexe particulier, le terme électrostatique peut souvent aider à améliorer le score des orientations qui ressemblent à la solution correcte. Cependant, dans certains cas, la contribution électrostatique détériore la qualité des

prédictions. Malheureusement, il n'est pas toujours évident de savoir quand l'électrostatique devrait être incluse pour chaque cas particulier. Dans presque tous les cas étudiés, la complémentarité de forme s'avère nettement importante que l'électrostatique. Dans Eq 4.18, j'ai mis K =-0.6 KJ/mol/Å^3, ce qui rend la composante de forme environ cinq fois plus grande que la contribution électrostatique dans le score global. Néanmoins, l'approche ci-dessus a été utilisée avec succès sur plusieurs des cibles protéiques de l'expérience CAPRI - amarrage en aveugle (Ritchie, 2003; Mustard & Ritchie, 2005). La figure 4.12 montre les meilleures orientations proposées par *Hex* pour deux des cibles protéiques dans les premiers rounds de CAPRI.

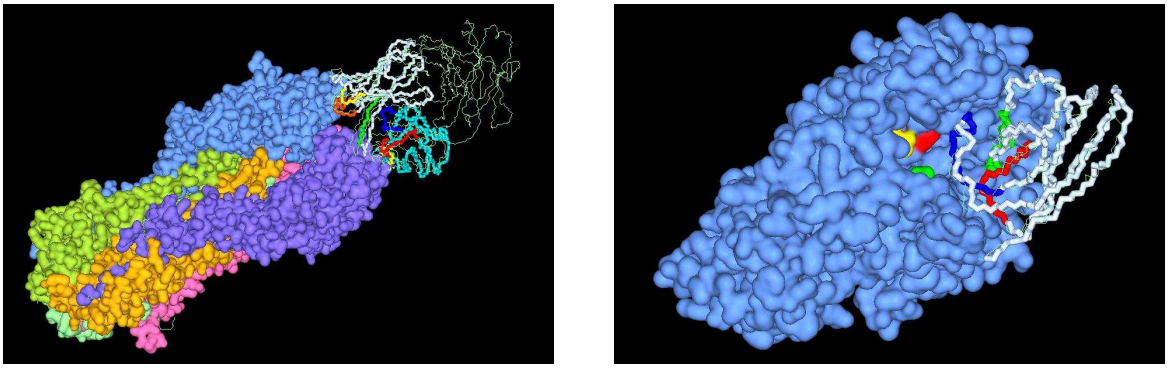

Figure 4.12: Les résultats de la cible 3 de CAPRI, HA/HC63 (à gauche) et 6, α-Amylase/AMD9 (à droite). La meilleure solution d'amarrage obtenue pour le complexe HA/HC63 était la quatrième solution soumise à CAPRI. Cette solution a 43/63 contacts corrects de résidus. La déviation entre les coordonnées du fragment d'anticorps Fv et celles de la structure cristallographique du complexe est 7.43Å RMS. Le fragment de HC63 Fv est coloré comme suit : VH en blanc, VL en cyan, H1 en orange, H2 en jaune, H3 en vert, L1 en rouge, L2 en jaune et L3 en bleu. L'orientation cristallographique du Fv est montrée en vert léger. Les chaînes de HA sont colorées comme suit : A en bleu léger, C en rose, E en bleu-foncé et F en orange. La solution d'amarrage obtenue pour le complexe AMB9/α-amylase (à droite) était la cinquième solution soumise. Cette prédiction a 53/65 contacts corrects de résidus et une déviation de 2.16Å RMS entre les coordonnées AMB9 prédites et vraies. L'α-amylase est en bleu et le domaine AMB9 VHH est en blanc. Les résidus du site actif d'amylase sont colorés comme suit – ASP-197 en rouge, GLU-233 en jaune et ASP-300 en vert. Les boucles de VHH CDR sont colorées comme suit : CRD1 en rouge, CDR2 en vert et CDR3 en bleu. L'orientation cristallographique du VHH est montrée en vert léger.

4.2.2 Guider les corrélations d'amarrage

Comparée aux approches FFT conventionnelles de grille cartésienne, l'approche SPF nous permet de façon relativement simple d'utiliser la connaissance antérieure (ou "prior knowledge") sur une interaction protéique connue pour guider et accélérer le calcul d'amarrage. Par exemple, la connaissance d'un résidu impliqué dans l'interaction suffit pour établir une orientation initiale d'amarrage et pour limiter ou guider la recherche d'amarrage autour de l'interface initiale. Ceci est illustré dans la figure 4.13. Utiliser cette approche pour contraindre l'espace de recherche peut, de manière signi-

ficative, augmenter la qualité des prédictions d'amarrage (Ritchie *et al.*, 2008).

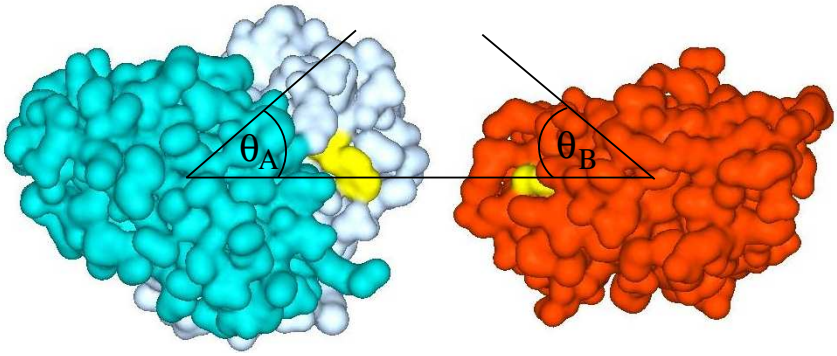

Figure 4.13: Une illustration d'une recherche d'amarrage guidée par la connaissance antérieure – un résidu interactif sur chacun des associés d'amarrage. Cette figure montre l'orientation initiale pour le complexe HyHel-5 anticorps/lysozyme, dans laquelle des résidus de l'anticorps H-33:TRP (à gauche) et le lysozyme Y-53:TYR (à droite) ont été accentués en jaune. L'orientation montrée a été initialisée automatiquement dans *Hex* en tournant l'atome C_α de chaque résidu accentué sur l'axe intermoléculaire z. La corrélation d'amarrage peut alors être contrainte de rechercher autour de l'orientation initiale en limitant les plages permises pour les angles de recherche θ_A et θ_B.

4.2.3 Amarrage de très grandes protéines

Puisque les fonctions radiales de base décroissent rapidement au-delà d'environ 30Å de l'origine choisie, l'approche ci-dessus n'est pas directement appropriée pour amarrer de très grandes protéines, telles que les cibles 2 et 3 de CAPRI. Cependant, il n'est pas nécessaire de s'appuyer sur une seule origine de coordonnées. Par exemple, la surface de la plus grande protéine considérée comme "le récepteur," peut être couverte de multiples copies de la plus petite protéine "le ligand." Ensuite, une recherche guidée d'amarrage peut être exécutée autour de chaque position initiale du ligand. La figure 4.14 illustre cette approche pour la cible 2 de CAPRI, un complexe entre la grande protéine de surface virale VP6 et un domaine d'anticorps Fv.

L'algorithme couverture d'amarrage contient quatre étapes. Premièrement, si la connaissance sur la surface de liaison du ligand est disponible, la molécule de ligand est orientée le long de l'axe négatif z pour faire face au récepteur. Deuxièmement, une surface sphérique harmonique de basse résolution avec *L=5* est calculée pour le récepteur en échantillonnant sa surface sur une tessellation icosaèdrique de la sphère, comme il est montré dans la figure 4.14(B). Pour chaque facette triangulaire de la surface, un vecteur normal est calculé et une sphère de rayon 15Å est centrée sur chaque facette normale de l'extérieur et tangentielle à la surface. Ceci couvre la surface avec des sphères. Dans la troisième étape, les sphères de surface sont échantillonnées en identifiant itérativement et rejetant la sphère qui a le plus grand recouvrement de volume avec ses voisins. Cette procédure est répétée jusqu'à ce qu'aucun recouvrement de volume n'excède 5Å^3. Ceci donne une distribution as-

sez régulière des sphères restantes sur la surface du récepteur. Finalement, chaque sphère restante (vecteur normal) est utilisée pour définir un axe intermoléculaire local pour l'amarrage, avec une orientation initiale ligand (axe) transférée sur la normale extérieure, et une origine de coordonnées locale pour le récepteur définie à une distance égale le long de la normale intérieure. La figure 4.14(C) montre le résultat de l'application de l'algorithme de surface de sphères limité à la chaîne C du trimère VP6 et la figure 4.14(D) montre le trimère couvert de 23 orientations initiales produites pour le MCV Fv.

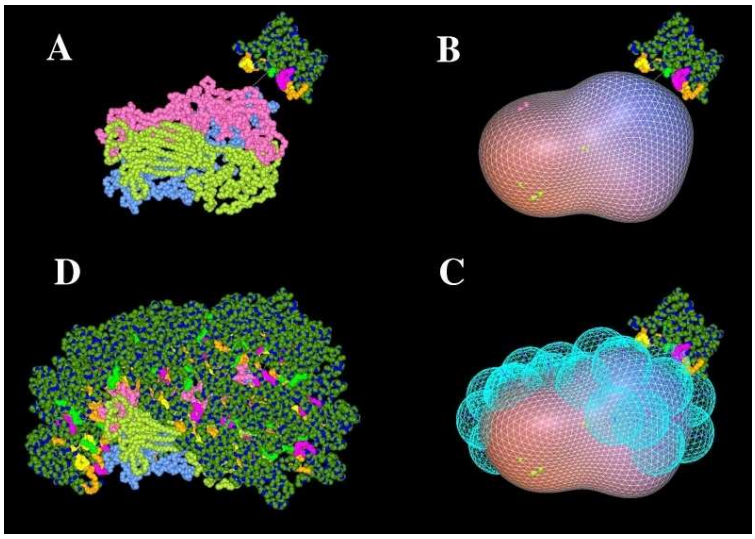

Figure 4.14: Les quatre étapes de l'algorithme d'échantillonnage de surface macromoléculaire, illustrées schématiquement pour le complexe de l'anticorps MCV/VP6 (cible 2 de CAPRI). (A) Les boucles hypervariables du fragment MCV Fv (le "ligand") sont initialement orientées pour faire face au trimère VP6 (le "récepteur"). Les chaînes VP6 sont colorées comme suit : A en bleu, B en jaune et C en rose. (B) Une surface SH à basse résolution avec *L=5* est calculée pour le récepteur (2.252 triangles de surface). (C) La surface SH du récepteur après l'application de l'algorithme de couverture de sphères à la chaîne C du récepteur. (D) Des multiples orientations initiales d'amarrage pour le ligand sont produites à partir des centres de sphères. Cet exemple montre les 23 fragments de MCV Fv répartis sur la chaîne C de VP6.

4.2.4 Clustering des solutions d'amarrage

Puisque la procédure de couverture sphérique de surface macromoléculaire tente de sur-échantillonner l'espace de recherche d'orientation, toutes les solutions de basse énergie sont groupées afin d'identifier des orientations distinctes. L'algorithme de clustering classe en premier les solutions d'amarrage par leur énergie et assigne la solution de plus basse énergie comme le membre initial ou "seed" du premier cluster. La liste des solutions restantes est alors parcourue pour trouver des entrées non-assignées et toutes orientations pour lesquelles les atomes C_α du ligand qui sont distants de 2Å de RMS des atomes correspondants du membre "seed" sont assignées au cluster. La liste est

alors re-parcourue pour trouver la prochaine solution non-assignée ayant la plus basse énergie qui deviendra le seed du prochain cluster et la procédure est répétée jusqu'à ce que toutes les solutions aient été assignées à un cluster.

Même lorsqu'il n'est pas nécessaire d'utiliser de multiples orientations initiales pour le ligand, cet algorithme de clustering donne une façon utile de réduire le nombre de "faux-positifs" produits par une recherche d'amarrage. Par exemple, dans une recherche exhaustive comme la nôtre, plusieurs orientations semblables mais néanmoins distinctes peuvent être trouvées et celles-ci tendraient "à pousser une bonne solution vers le bas de la liste" si un clustering n'a pas été utilisé. Le clustering est aussi utile quand on utilise la minimisation d'énergie parce que les multiples solutions d'amarrage peuvent être fusionnées en une seule orientation minimisée.

4.2.5 Sondes potentielles sélectionnées par la PCA pour l'amarrage protéique

Une des difficultés que présente l'amarrage macromoléculaire est de concevoir une fonction de score fiable basée sur l'énergie avec laquelle on évaluera les orientations échantillonnées d'amarrage. La complémentarité de forme est très efficace comme premier filtre, mais il serait souhaitable d'incorporer des interactions chimiques dans le schéma de score pour aider à distinguer le vrai complexe des nombreux faux-positifs produits par une recherche de corrélation d'amarrage. Cependant, on ne peut pas déduire directement lesquels des multiples types d'interactions intermoléculaires (électrostatiques, liaisons d'hydrogène, désolvatation, ponts salins, forces de dispersion, etc.) donnent la force motrice pour lier deux molécules dans un quelconque cas particulier. Par conséquent, dans le cadre de projet de thèse de doctorat d'Alessandra Fano, nous avons étudié comment l'analyse de composantes principales (ou "principal component analysis," PCA) des champs interactifs moléculaires (ou "molecular interaction fields," MIFs) peut être utilisée afin de choisir les types d'interactions chimiques les plus significatifs pour une certaine cible d'amarrage. Pour vérifier l'efficacité de cette approche, nous avons essayé d'amarrer les composants non liés du complexe entre la subtilisine de streptomyces (SUP) et son inhibiteur naturel (SSI). Ce complexe, qui est un cas difficile à amarrer avec les techniques courantes, présente ainsi un test parfait.

En premier, nous avons utilisé le programme GRID (Goodford, 1985) afin de produire des cartes 3D d'énergie potentielle de plusieurs types d'atomes de sonde placés dans une grille pour les surfaces de chaque protéine. Seuls les points avec des valeurs potentielles au delà d'un seuil donné sont maintenus; les autres sont rejetés. La figure 4.15 illustre cette approche pour la protéine SSI. Un calcul semblable a été effectué pour la SUP. Bien que le but soit d'apparier les distributions de tels potentiels avec des distributions complémentaires des potentiels de l'associé d'amarrage, il serait impraticable d'utiliser toutes les cartes d'énergie potentielle possibles pendant un calcul d'amarrage. Donc nous avons utilisé la PCA pour choisir les types de sondes les plus pertinents et les plus significatifs pour guider la recherche d'amarrage. La PCA est une approche standard de chemoinformatique

pour extraire des informations à partir d'ensembles de données très grands. En général, la PCA est utilisée dans des études de conception de médicaments, en particulier pour l'association quantitative entre la structure 3D et l'activité (ou "3D quantitative structure-activity relationship, 3D-QSAR; Pastor *et al.*, 1995) et elle est aussi utilisée pour relier les sites de liaison de récepteur-ligand (Matter & Schwab, 1999). Cependant, l'approche d'Alessandra représente une première application de PCA sur des potentielles des sondes pour classer les solutions de l'amarrage macromoléculaire. La raison principale d'effectuer une PCA sur les MIF est le fait que ces sondes qui contribuent le plus à la variance dans les cartes d'énergie sont supposées d'être les meilleures indicatrices de complémentarité pour le système sous considération.

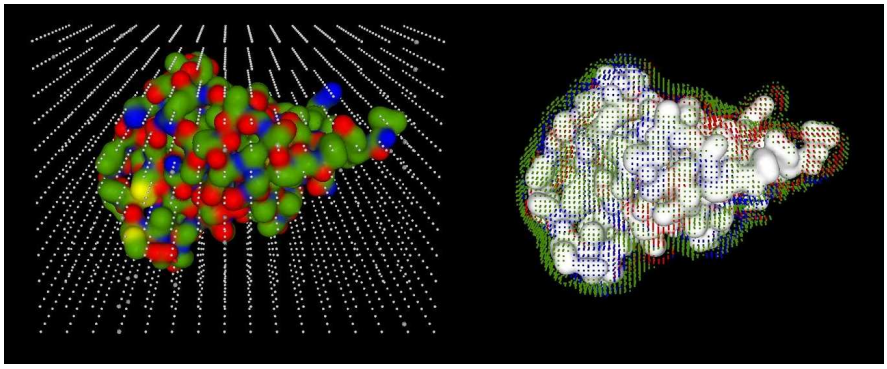

Figure 4.15: À gauche : la protéine SSI placée dans une grille 3D. À droite : positions des sondes de surface codées en couleur par le potentiel de GRID – donneur de protons en rouge, accepteur de protons en bleu et hydrophobe en vert.

Dans une PCA, la matrice de sondes potentielles est décomposée en deux matrices plus petites des chargements et scores. Les chargements mesurent la pondération des variables originales dans l'analyse et les scores donnent une réprésentation simplifiée des objets (les sondes dans notre cas) en termes d'un petit nombre de nouvelles variables non-corrélatives (les composantes principales, ou PCs). Le graphe des scores des objets contre les PC nous permet d'identifier les objets (ou clusters d'objets) qui expliquent en grande partie la variance. Les graphes de PCA montrés sur la figure 4.16 montrent la distribution des six sondes les plus significatives (C sp3, NH amide, N+ sp3, O carbonyle, O- carboxyle et Sèche, ou "Dry") dans les deux premiers composants de l'espace de chimiométrie pour SUP et SSI. Il est intéressant de noter la symétrie de miroir dans le plan horizontal. Ceci suggère qu'il y a une bonne complémentarité chimique entre les protéines. Ces graphes montrent que seulement trois sondes (N+, O- et Dry) sont suffisantes pour expliquer en grande partie la variance dans les cartes de potentiels. Par conséquent, seul ces trois sondes sont utilisées dans les prochains calculs d'amarrage.

Pour un certain type de sonde, les sondes potentielles sélectionnées par la PCA peuvent être transformées en fonctions continues lisses pour l'amarrage dans *Hex* de la même façon que les

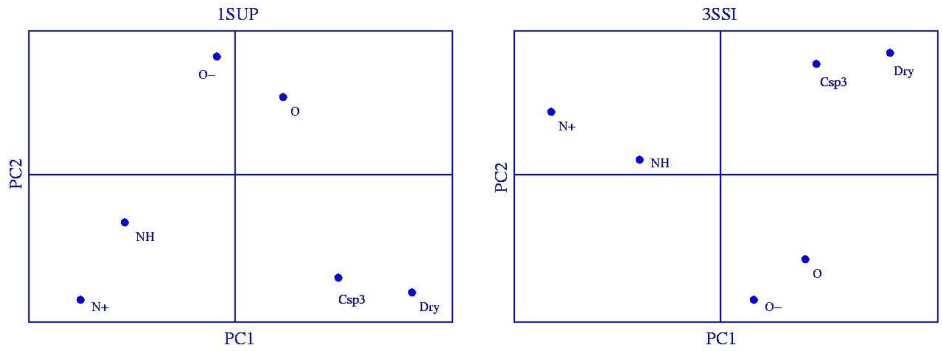

Figure 4.16: À gauche : Le graphe de PCA pour SUP montrant les contributions des six types de sondes potentielles les plus significatifs (C sp3, NH amide, N+ sp3, O carbonyle, O- carboxyle et Dry) aux deux premières composantes principales, PC1 et PC2. PC1 et PC2 expliquent 69.3% et 21.5% de la variance totale, respectivement. À droite : pareillement, le graphe de PCA pour SSI. PC1 et PC2 expliquent 74.3% et 16.0% de la variance totale, respectivement.

charges des atomes points (ou “point-charges”) sont transformées pour donner une fonction de charge-densité (voir la section 2.5.2). Par exemple, en traitant les positions de sonde N+ pour la protéine A comme une somme sur les points potentiels, $\phi^{N+}(\underline{x}_i)$, on peut écrire la somme de potentiels comme :

$$\phi^{N+}(\underline{x}) = \sum_i \phi^{N+}(\underline{x}_i)\delta(\underline{x} - \underline{x}_i). \tag{4.24}$$

Ceci peut alors être représenté comme une série coupée de SPF dans la façon usuelle

$$\phi^{N+}(\underline{r}) \simeq \sum_{nlm} a^{N+}_{nlm} R_{nl}(r) y_{lm}(\theta, \phi), \tag{4.25}$$

où les coefficients d’expansion sont calculés en utilisant (c.f. Eq 2.133) :

$$a^{N+}_{nlm} = \sum_i \phi^{N+}(\underline{r}_i) R_{nl}(r_i) y_{lm}(\theta_i, \phi_i). \tag{4.26}$$

Des expressions similaires peuvent être écrites pour les autres types potentiels. L’énergie globale d’interaction peut alors être estimée comme

$$E_{GRID} = \frac{1}{2} \int \left[\tau_A(\underline{r})\left(\phi^{\mathrm{N+}}_B(\underline{r}) + \phi^{\mathrm{O-}}_B(\underline{r}) + \phi^{\mathrm{Dry}}_B(\underline{r})\right) + \tau_B(\underline{r})\left(\phi^{\mathrm{N+}}_A(\underline{r}) + \phi^{\mathrm{O-}}_A(\underline{r}) + \phi^{\mathrm{Dry}}_A(\underline{r})\right)\right] \mathrm{d}V. \tag{4.27}$$

Il est à noter que cette expression ne favorise pas spécifiquement des paires complémentaires de types de sondes individuelles, ni ne pénalise des paires non-favorables. Cependant, par conception, cela devrait donner un minimum profond quand les protéines sont contraposées dans une orientation bien ajustée. Par conséquent, ce terme d’énergie devrait augmenter les scores des orientations

d'amarrage basées sur la complémentarité de forme. La figure 4.17 montre le potentiel N+ calculé par SPF sur la SAS de la protéine SSI, avec le potentiel correspondant O- sur la protéine SUP.

Après comparaison avec les calculs d'amarrage basés sur la forme-seulement et la forme-plus-électrostatique décrites précédemment, nous avons trouvé qu'en ajoutant le potentiel de grille dans la fonction de corrélation de forme améliore le rang de la première solution presque-native par au moins un facteur de 2 (spécifiquement, la meilleure solution trouvée basée sur la forme-seulement était classée 13-ème, la forme-plus-électrostatique 10-ème et la forme-plus-grille 5-ème). Malheureusement, puisque notre approche exige plusieurs étapes manuelles pour effectuer les analyses PCA et pour importer les données de GRID dans *Hex*, il n'a pas été possible de faire des tests plus approfondis pendant la période où Alessandra était à Aberdeen. Cependant, ce travail a démontré pour la première fois l'utilité des corrélations des potentiels chimiques dans des calculs d'amarrage. L'approche globale a été par la suite utilisée pour produire quelques modèles utiles d'amarrage pour la protéine co-réceptrice de surface CCR5 et des protéines chemokines MIP-1β et RANTES (Fano *et al.*, 2006).

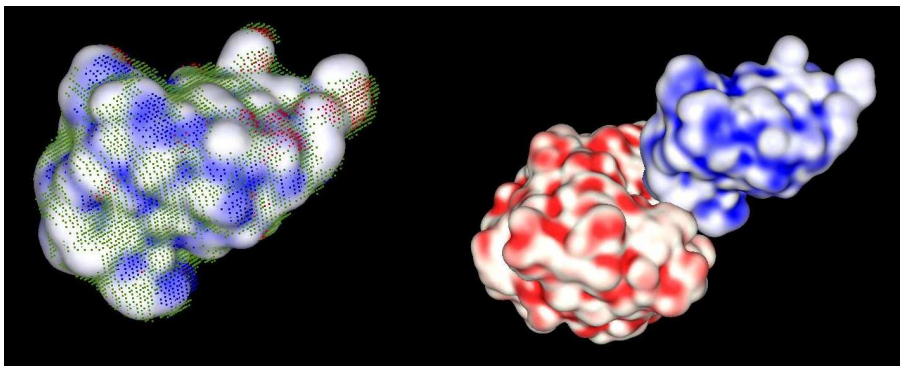

Figure 4.17: À gauche : le potentiel N+ (régions bleues) calculé à partir de la SAS (montrée en blanc) de SSI. Les points bleus montrent les positions des points originaux de N+ de GRID, les points rouges montrent les positions O- et les points verts montrent les points hydrophobes. À droite : les points chauds de sonde N+ (bleu) et O- (rouge) sur les surfaces SAS de SSI et de SUP, respectivement. Cette image montre l'orientation de liaison du complexe mais avec les deux protéines légèrement séparées pour une meilleure vue.

4.2.6 FFT multi-dimensionnelles pour l'amarrage protéique

Puisque la FFT permet à un problème qui requiert formellement $O(N^2)$ opérations d'être calculé en $O(N \log N)$ étapes, il est raisonnable de s'attendre à ce qu'une plus grande vitesse informatique soit atteinte quand la FFT est appliquée à autant de degrés de liberté que possible. Cette section décrit comment l'approche SPF peut être utilisée pour développer des expressions de corrélation tridimensionnelles (3D) et cinq-dimensionnelles (5D) d'amarrage. Pour réaliser ceci, il est commode d'utiliser des fonctions SH réelles et complexes, où les deux types de fonction sont reliés par la matrice

unitaire de transformation, $U^{(l)}$ (voir la section 2.1.11),

$$y_{lm}(\theta,\phi) = \sum_{m'} U^{(l)}_{mm'} Y_{lm'}(\theta,\phi). \tag{4.28}$$

Si une propriété 3D particulière est initialement échantillonnée comme une expansion réelle, elle peut être représentée dans la forme complexe en utilisant

$$A(\underline{r}) = \sum_{nlm}^{N} A_{nlm} R_{nl}(r) Y_{lm}(\theta,\phi) \tag{4.29}$$

où les coefficients complexes d'expansion, A_{nlm}, sont liés aux coefficients réels originaux par

$$A_{nlm} = \sum_{m'} U^{(l)}_{m'm} a_{nlm'}. \tag{4.30}$$

La corrélation entre une paire de propriétés complexes, $A(\underline{r})$ et $B(\underline{r})$, peut alors être écrite comme

$$E = \int (\hat{T}(-R)\hat{R}(0,\beta_A,\gamma_A)A(\underline{r}))^* (\hat{R}(\alpha_B,\beta_B,\gamma_B)B(\underline{r}))\mathrm{d}\underline{r}, \tag{4.31}$$

où l'astérisque dénote le conjugé complexe. En substituant les expansions SPF, on obtient alors la version complexe d'Eq 4.1

$$E = \sum_{kjsmnlv} D^{(j)*}_{ms}(0,\beta_A,\gamma_A) A^*_{kjs} T^{(|m|)}_{kj,nl}(R) D^{(l)}_{mv}(\alpha_B,\beta_B,\gamma_B) B_{nlv}. \tag{4.32}$$

Puis, en sommant sur les indices radiaux k et n, ceci donne

$$E = \sum_{jsmlv} D^{(j)*}_{ms}(0,\beta_A,\gamma_A) S^{(|m|)}_{js,lv}(R) D^{(l)}_{mv}(\alpha_B,\beta_B,\gamma_B), \tag{4.33}$$

où $S^{(|m|)}_{js,lv}(R)$ sont les éléments d'une matrice réduite de translation/recouvrement qui est donnée par

$$S^{(|m|)}_{js,lv}(R) = \sum_{kn} A^*_{kjs} T^{(|m|)}_{kj,nl}(R) B_{nlv}. \tag{4.34}$$

Eq 4.33 peut être représentée en forme exponentielle en notant qu'une rotation de β autour de l'axe y peut être toujours calculée comme une rotation de β autour d'un axe z tourné en utilisant (Edmonds, 1957)

$$\hat{R}_y(\beta) \equiv \hat{R}_z(-\pi/2)\hat{R}_y(-\pi/2)\hat{R}_z(\beta)\hat{R}_y(\pi/2)\hat{R}_z(\pi/2), \tag{4.35}$$

et en re-écrivant les matrices $d^l_{mm'}(\beta)$ d'une rotation générale de Wigner

$$D^{(l)}_{mm'}(\alpha,\beta,\gamma) = e^{-im\alpha} d^l_{mm'}(\beta) e^{-im'\gamma}, \tag{4.36}$$

comme le produit correspondant des exponentiels complexes

$$d^l_{mm'}(\beta) = \sum_t e^{im\pi/2} d^l_{mt}(-\pi/2) e^{-it\beta} d^l_{tm'}(\pi/2) e^{-im'\pi/2}. \tag{4.37}$$

Puis, en écrivant

$$\begin{aligned} \Delta^l_{tm} &= d^l_{tm}(\pi/2) \\ &= d^l_{mt}(-\pi/2), \end{aligned} \tag{4.38}$$

et en rassemblant les coefficients constants

$$\begin{aligned} \Gamma^{tm}_{lm'} &= e^{i(m-m')\pi/2} \Delta^l_{tm} \Delta^l_{tm'} \\ &= i^{m-m'} \Delta^l_{tm} \Delta^l_{tm'} \end{aligned} \tag{4.39}$$

permet aux éléments de matrice de rotation de Wigner d'être écrits sous une forme complètement exponentielle

$$D^{(l)}_{mm'}(\alpha, \beta, \gamma) = \sum_t \Gamma^{tm}_{lm'} e^{-im\alpha} e^{-it\beta} e^{-im'\gamma}. \tag{4.40}$$

En substituant Eq 4.40 deux fois dans Eq 4.33 on obtient alors le résultat entièrement factorisé

$$E = \sum_{jsmlvrt} \Gamma^{rm}_{js} S^{(|m|)}_{js,lv}(R) \Gamma^{tm}_{lv} e^{-i(r\beta_A - s\gamma_A + m\alpha_B + t\beta_B + v\gamma_B)}, \tag{4.41}$$

où l'addition s'étend sur toutes les valeurs des indices qui satisfont $|r| \leq j, |s| \leq j, |t| \leq l, |v| \leq l$ et $|m| \leq \min(l, j) \leq L$. Dans cette équation, r et t énumèrent les composants de fréquence azimutaux et s, v, et m énumèrent les fréquences circulaires. J'appelle Eq 4.41 l'équation principale de corrélation d'amarrage.

L'équation 4.41 a la forme évidente d'une série complexe cinq-dimensionnelle de Fourier. Par conséquent, elle peut être calculée en utilisant une FFT multi-dimensionnelle. Cependant, puisque les angles de rotation d'Euler ont les plages $0 \leq \alpha, \gamma < 360^{\text{o}}$ et $0 \leq \beta < 180^{\text{o}}$, il est utile de changer le signe de rotation γ_A et de proportionner les angles de rotation β de sorte que toutes les coordonnées de rotation se mappent à la phase et la période normales de la FFT. Si ceci n'est pas fait, le calcul FFT sur-échantillonnera les coordonnées β pour donner des solutions dupliquées, chacune ayant la moitié de la résolution désirée. En proportionnant les coordonnées β, on élimine cet effet et l'on peut utiliser une plus petite grille FFT afin de réduire de moitié la quantité de mémoire requise pour chaque dimension β et accélérer le calcul FFT.

Travailler avec le signe de γ_A est simple. Par exemple, en mettant $\gamma'_A = -\gamma_A$ et écrivant

$$e^{is\gamma_A} = \sum_q \eta_{sq} e^{-iq\gamma'_A}, \tag{4.42}$$

et en utilisant l'orthogonalité des exponentiels pour résoudre les coefficients, η_{sq}, donne

$$\eta_{sq} = \delta_{s\bar{q}}. \tag{4.43}$$

De même, les rotations β peuvent être proportionnées en mettant $\beta' = 2\beta$ et en écrivant

$$e^{-it\beta} = \sum_u \lambda_{tu} e^{-iu\beta'}, \tag{4.44}$$

et en utilisant une fois de plus l'orthogonalité des exponentiels pour résoudre les coefficients λ_{tu}. Dans ce cas, il peut être montré, en utilisant des relations trigonométriques de base, que les coefficients sont donnés par

$$\lambda_{tu} = \begin{cases} 2i/\pi(2u-t) & \text{si } t \text{ est un chiffre impair,} \\ 1 & \text{si } t = 2u, \\ 0 & \text{autrement.} \end{cases} \tag{4.45}$$

En d'autres termes, il existe des solutions exactes quand t est un chiffre pair, et des solutions convergentes de série entière quand t est un chiffre impair. Cependant, pour les besoins actuels, les coefficients λ_{tu} peuvent être déterminés pour reproduire *exactement* le même ensemble fini d'échantillons de rotation M_β en traitant Eq 4.44 comme une équation discrète d'analyse transformée de Fourier

$$\lambda_{tu} = \frac{1}{M_\beta} \sum_{n=0}^{M_\beta - 1} e^{-\pi i t n / M_\beta} e^{2\pi i u n / M_\beta}. \tag{4.46}$$

D'autres plages angulaires (e.g. quand on exécute une recherche guidée d'amarrage) peuvent être proportionnées sur la période normale de FFT d'une façon semblable. Substituer les changements de variable mentionnés ci-dessus dans Eq 4.41 et appliquer une transformée inverse de Fourier au résultat donne

$$E[p,q,m,u,v;R] = \sum_{rt} \sum_{jl} \Gamma^{rm}_{j\bar{q}} S^{(|m|)}_{j\bar{q},lv}(R) \Gamma^{tm}_{lv} \lambda_{rp} \lambda_{tu}. \tag{4.47}$$

Rassemblant les coefficients comme

$$\Lambda^{um}_{lv} = \sum_t \Gamma^{tm}_{lv} \lambda_{tu} \tag{4.48}$$

donne finalement

$$E[p,q,m,u,v;R] = \sum_{jl} \Lambda^{pm}_{j\bar{q}} S^{(|m|)}_{j\bar{q},lv}(R) \Lambda^{um}_{lv}. \tag{4.49}$$

Cette équation peut être considérée comme une recette analytique pour calculer les éléments d'une grille FFT 5D. Appliquer une transformée en avant de Fourier aux éléments de ce tableau produira un tableau 5D de valeurs de fonction $E(\beta_A, \gamma_A, \alpha_B, \beta_B, \gamma_B, R)$ pour d'*uniques* combinaisons d'angles de rotation d'Euler. Par conséquent, Eq 4.49 peut être interprétée comme une fonction analytique génératrice (ou "generating function," GF) pour des corrélations FFT 5D d'amarrage.

4.2.7 FFT multi-dimensionnelles

Bien qu'il soit satisfaisant de pouvoir obtenir une formule très compacte pour produire des corrélations FFT 5D, il ne s'ensuit pas automatiquement qu'une GF d'ordre supérieur comme Eq 4.49 serait la plus efficace pour calculer en pratique. Par exemple, dans Eq 4.49 on peut voir que la double somme sur les indices jl signifie que le coût d'initialiser chaque cellule de grille FFT 5D est de l'ordre de $O(N^2)$ et donc le coût global d'initialiser une grille complète FFT 5D est de l'ordre de $O(N^7)$. Par conséquent, il est avantageux de calculer Eq 4.49

$$W_{lv}^{pqm}(R) = \sum_{j} \Lambda_{j\bar{q}}^{pm} S_{j\bar{q},lv}^{(|m|)}(R) \tag{4.50}$$

et

$$E[p,q,m,u,v;R] = \sum_{l} W_{lv}^{pqm}(R) \Lambda_{lv}^{um}. \tag{4.51}$$

Ainsi, en utilisant un tableau temporaire, W, l'initialisation d'une FFT 5D au coût de $O(N^7)$ peut être calculée pratiquement en deux étapes de $O(N^6)$. La double somme dans l'expression de la matrice réduite de recouvrement, Eq 4.34, peut être calculée efficacement d'une manière semblable. Néanmoins, utiliser un grand tableau intermédiaire requiert une mémoire additionnelle significative. Une façon de réduire la quantité de mémoire nécessaire est de mettre $\gamma_A = 0$ dans l'expression de corrélation et de tourner explicitement les coefficients d'expansion du récepteur pour donner

$$S_{jq,lv}^{(|m|)}(R,\gamma_A) = \sum_{kn}^{N} A_{kjq}^{*}(\gamma_A) T_{kj,nl}^{(|m|)}(R) B_{nlv} \tag{4.52}$$

où $A_{kjq}(\gamma_A)$ représente un coefficient tourné d'expansion. Les éléments de tableau FFT 4D peuvent alors être calculés à partir de la GF 4D

$$E[p,m,u,v;R,\gamma_A] = \sum_{jql} \Lambda_{j\bar{q}}^{pm} S_{j\bar{q},lv}^{(|m|)}(R,\gamma_A) \Lambda_{lv}^{um}. \tag{4.53}$$

Ainsi, en principe, une recherche d'amarrage 6D pourrait être exécutée en réitérant sur des paires d'échantillons (R,γ_A) et, pour chaque paire, en calculant un FFT 4D sur les angles restants de rotation. Cependant, on peut voir immédiatement que cette approche est impraticable parce que la triple somme dans Eq 4.53 indique que le coût d'initialisation d'une grille FFT 4D est toujours de l'ordre de $O(N^7)$. En revanche, la complexité de la GF tombe de manière significative si l'angle de rotation β_A est annulé de la FFT. Par exemple, en transformant explicitement les coefficients d'expansion réels du récepteur avec les équations 4.3 et 4.4 et puis en calculant

$$A_{nlm}(R,\beta_A,\gamma_A) = \sum_{m'} U_{m'm}^{(l)} a_{nlm'}(R,\beta_A,\gamma_A), \tag{4.54}$$

la GF est alors

$$E[m,u,v;R,\beta_A,\gamma_A] = \sum_l S^m_{lv}(R,\beta_A,\gamma_A)\Lambda^{um}_{lv} \tag{4.55}$$

où

$$S^m_{lv}(R,\beta_A,\gamma_A) = \sum_n A^*_{nlm}(R,\beta_A,\gamma_A)B_{nlv}. \tag{4.56}$$

Par conséquent, en tenant compte des précédents vecteurs pré-calculés de coefficient du récepteur, on peut voir que le coût d'initialisation d'une FFT 3D de rotation est essentiellement de l'ordre de $O(N^2)$ pour chaque cellule de grille FFT. Afin d'être exhaustif, la GF 2D a la même complexité de structure et d'initialisation que ci-dessus et elle peut être représentée comme

$$E[m,u;R,\beta_A,\gamma_A,\gamma_B] = \sum_{lv} S^m_{lv}(R,\beta_A,\gamma_A,\gamma_B)\Lambda^{um}_{lv} \tag{4.57}$$

où

$$S^m_{lv}(R,\beta_A,\gamma_A,\gamma_B) = \sum_n A^*_{nlm}(R,\beta_A,\gamma_A)B_{nlv}(\gamma_B). \tag{4.58}$$

Par conséquent, comme le cas de 4D, les corrélations 2D peuvent être annulées comme étant impraticables en termes de calcul. La GF 1D (initialisation FFT de complexité $O(N^3)$ pour chaque angle α_B de torsion de recherche) a été implémentée précédemment dans la forme réelle (voir les sections 4.2 et 4.1.1) et est donnée par

$$E[m;R,\beta_A,\gamma_A,\beta_B,\gamma_B] = \sum_{nl} A^*_{nlm}(R,\beta_A,\gamma_A)B_{nlm}(\beta_B,\gamma_B). \tag{4.59}$$

4.2.8 FFT de multi-propriétés

Il est bien connu que la corrélation entre deux paires de propriétés réelles peut être calculée simultanément en utilisant une FFT complexe. Par exemple, si la charge-densité et le potentiel d'électrostatique *in vacuo* d'un système de deux protéines, A et B, sont écrits comme

$$\begin{aligned}\phi(\underline{r}) &= \phi_A(\underline{r}) + \phi_B(\underline{r})\\ \rho(\underline{r}) &= \rho_A(\underline{r}) + \rho_B(\underline{r})\end{aligned} \tag{4.60}$$

et si les combinaisons linéaires des expansions SPF sont formées comme

$$\begin{aligned}\underline{A} &= \underline{U}^T(\underline{a}^\phi + i\underline{a}^\rho)\\ \underline{B} &= \underline{U}^T(\underline{b}^\rho + i\underline{b}^\phi),\end{aligned} \tag{4.61}$$

où $\underline{U}^T$ est le transposé de la matrice unitaire de transformation complexe-à-réelle $\underline{U}$ (c.f. équations 4.28 et 4.30), alors l'énergie électrostatique d'interaction pour une orientation paire peut être calculée comme

$$E = Re(\underline{A}^* \underline{B}). \tag{4.62}$$

De même, éliminer les indices d'addition et utiliser la notation de matrice pour l'énergie électrostatique d'interaction 6D de GF (Eq 4.49) donne

$$E[p,q,m,u,v;R] = \underline{\Lambda}^{pqm} \underline{S}^{qmv}(R) \underline{\Lambda}^{uvm}. \tag{4.63}$$

Cependant, grâce à la linéarité de cette expression, il s'ensuit que de multiples corrélations d'énergie d'interaction $e = 0, 1, 2, ...$ peuvent être calculées simultanément en additionnant en premier la partie de distance-dépendante de chaque interaction de potentiel/densité

$$\big(\underline{S}_e^{qmv}(R)\big)_{jl} = \sum_{kn} A_{kjq}^{e*} T_{kj,nl}^{(|m|)}(R) B_{nlv}^e, \tag{4.64}$$

pour donner

$$E[p,q,m,u,v;R] = \underline{\Lambda}^{pqm} \big(\sum_e \underline{S}_e^{qmv}(R) \big) \underline{\Lambda}^{uvm}. \tag{4.65}$$

Ainsi, des combinaisons arbitraires de corrélations, *incluant* celles qui utilisent différentes fonctions radiales, peuvent être évaluées ensemble dans une seule FFT 5D avec un coût additionnel très petit.

4.2.9 FFT de multi-résolutions

Il vaut la peine de noter qu'il n'y a aucune condition requise pour que les dimensions de grille FFT correspondent exactement à l'ordre polynomial des fonctions de base de SPF. Par exemple, une GF d'ordre inférieur peut être évaluée sur une grille FFT d'ordre supérieur et *vice-versa*. Ceci consiste à remplir la grille FFT avec des zéros ou exclure les composants de fréquence qui excèdent les limites de la grille, respectivement. Par conséquent, il est important de considérer attentivement l'ordre polynomial de l'expansion et les dimensions de la grille FFT, parce que chacun peut influencer de manière significative la performance globale. Il a été montré précédemment (Ritchie & Kemp, 2000; Ritchie, 2003) que l'utilisation des ordres polynômaux d'expansion dans la plage *L=24* à *30* est souvent suffisante pour donner une résolution satisfaisante pour l'amarrage des domaines globulaires de protéines. Selon la théorie d'échantillonnage de Shannon, ceci implique qu'une dimension angulaire de grille FFT d'au moins *M=2L=48* devrait être utilisée pour un échantillonnage complet de rotation. Ceci consiste à utiliser un incrément de recherche angulaire de $360^o/48 = 7.5^o$, qui est en quelque sorte plus correct que les tailles des incréments de rotation utilisés par convention dans des algorithmes cartésiens FFT. Néanmoins, puisque deux des cinq degrés de liberté de rotation peuvent être

décrits en utilisant les angles d'Euler qui s'étendent de 0 à 180°, il est évident qu'une grille FFT 5D de $48^3 \times 24^2$ cellules, par exemple, peut être accommodée dans au moins un gigaoctet (Gb) de mémoire si les valeurs de la grille sont sauvegardées comme des nombres complexes de virgule flottante de simple précision (deux fois 4 bytes par cellule de grille).

4.2.10 Comparaison de la performance de FFT

En guise d'un premier test d'utilité de l'approche FFT multidimensionnelle, le complexe HyHel-5/lysozyme (figure 4.13) a été amarré pour plusieurs valeurs d'ordre d'expansion, L, en utilisant la conformation liée du fragment d'anticorps Fv et la conformation non liée du lysozyme. Tableau 4.3 présente une comparaison de l'exactitude et des temps d'exécution des corrélations de forme-seulement et de forme-plus-électrostatique pour cet exemple. Tous les calculs ont échantillonné 53 étapes de translation de $\pm$0.75Å de l'orientation initiale du complexe. Afin de faciliter la comparaison des corrélations FFT 3D et 5D avec la FFT 1D radix-2 existant implémentée dans *Hex*, $M_\alpha = 64$ a été utilisé pour la dimension de l'angle de torsion. Les grilles 3D et 5D ont chacune utilisée $M_\gamma = 48$ et $M_\beta = 24$ pour donner des incréments (β, γ) de 7.5^o. Les degrés de liberté de rotation restants dans les cas de 3D et 1D ont respectivement utilisé une et deux tessellations icosaèdriques de la sphère, chacune de 812 sommets, pour produire des échantillons de rotation avec une séparation angulaire de moyenne d'environ 7.7^o. Considérant que les grilles d'Euler tendent de sur-échantillonner près des pôles, ce schéma donne pratiquement les équivalentes des densités échantillonnées avec environ 1.7, 2.5 et 3.5 milliard orientations d'amarrage pour les cas 1D, 3D et 5D, respectivement.

Table 4.3: Une comparaison des temps de corrélation d'amarrage de forme-seulement ("Shape-Only") et de forme-plus-électrostatique ("Shape+Electro").

	1D Shape-Only		1D Shape+Electro		3D Shape-Only		3D Shape+Electro		5D Shape-Only		5D Shape+Electro	
L	Rank (RMS)	Time/m	Rank (RMS)	Time/m	Rank (RMS)	Time/m	Rank (RMS)	Time/m	Rank (RMS)	Time/m	Rank (RMS)	Time/m
16	646 (6.8)	28.7	428 (8.0)	52.0	864 (7.1)	15.1	254 (8.2)	18.1	–	37.5	669 (6.0)	40.3
20	336 (1.2)	52.7	20 (1.3)	102.7	410 (1.2)	23.5	17 (1.3)	29.2	336 (7.9)	39.3	29 (1.3)	46.5
24	417 (1.2)	92.4	52 (1.2)	184.2	501 (1.2)	33.2	53 (1.2)	51.2	833 (1.2)	53.0	82 (1.2)	56.2
26	49 (1.2)	123.3	15 (1.2)	243.1	48 (1.2)	43.5	15 (1.6)	69.0	45 (1.2)	58.7	13 (1.6)	63.1
28	54 (1.5)	158.1	8 (1.2)	315.6	22 (5.2)	54.2	11 (1.3)	92.2	19 (5.5)	64.5	13 (1.2)	71.7
30	113 (2.2)	203.5	43 (1.3)	403.0	47 (1.6)	69.8	20 (1.6)	122.5	61 (1.6)	74.3	19 (1.6)	108.0

Dans ce tableau, L est l'ordre polynomial d'expansion, Rank est le rang de la première orientation trouvée dans laquelle le ligand est dans 10Å RMS (montré entre parenthèses) de la structure cristallographique après le clustering par défaut de *Hex*. Un tiret indique qu'aucune orientation presque-native n'a été trouvée dans les 2.000 premières solutions. Time est le temps total de calcul en minutes sur un seul processeur Xeon Pentium de 1.8GHz. Les calculs FFT de 3D et 5D ont utilisé la bibliothèque de KissFFT.[2] Pour ces calculs, la quantité de temps nécessaire pour la bibliothèque de FFT est essentiellement constante à 13.1 et 34.3 minutes, respectivement. Tous les temps excluent le calcul des éléments de matrice de translation.

Comme prévu, le tableau 4.3 montre que les expansions d'ordre supérieur assignent générale-

ment aux orientations presque-natives un meilleur rang que les expansions d'ordre inférieur, mais cette tendance n'est pas nécessairement monotonique. La meilleure combinaison d'un bon rang et d'une déviation basse de RMS de ligand du complexe est souvent obtenue avec *L=28* ou *L=30*. Cette table montre aussi que les calculs des FFT 3D de forme-seulement sont environ trois fois plus rapides que ceux de 1D et, ce qui est très surprenant, ils sont aussi en général plus rapides que les FFT 5D. Cependant, en raison de la linéarité de la GF, le coût d'inclusion de l'électrostatique dans les corrélations 3D et 5D est bas, comparé au coût de calcul des FFT 1D de forme-plus-électrostatique. En effet, les FFT 5D de forme-plus-électrostatique sont plus rapides que les FFT 3D quand $L \geq 26$. Ces différences deviendraient plus prononcées si plus de potentiels étaient inclus dans le calcul.

4.2.11 Résultats d'amarrages protéiques pré-établis

L'approche ci-dessus a été appliquée aux 84 complexes tirés de jeux de données tests pré-établis pour l'amarrage protéique (ou "Docking Benchmark" version 2.0; Mintseris *et al.*, 2005) en utilisant les corrélations de forme-seulement et de forme-plus-électrostatique (Ritchie *et al.*, 2008). Comme il est suggéré par la table 4.3, un protocole de recherche de deux étapes utilisant des balayages FFT de rotation 3D de forme-seulement avec *L=20* suivis par un reclassement de forme 1D-plus-électrostatique avec *L=30*, a été utilisé pour obtenir un bon compromis entre la vitesse et l'exactitude.

Pour donner une orientation initiale pseudo-aléatoire cohérente, toutes les protéines ont été initialement orientées par l'ajustage de moindres carrés au complexe et une petite rotation hors grille, $\hat{R}(\alpha, \beta, \gamma)$=$\hat{R}(11°, 9°, 0)$ a été ensuite appliquée au ligand. Les orientations calculées dans chaque passage d'amarrage ont été groupées en utilisant un algorithme glouton avec un seuil de clustering 9Å (Kozakov *et al.*, 2005) et le membre avec la plus basse énergie de chaque cluster a été sélectionné comme la "solution" pour ce cluster. Tous les autres membres de chaque cluster ont été rejetés.

Sept passages différents d'amarrage ont été exécutés pour chaque complexe afin d'évaluer les composants de forme et d'électrostatique de la fonction de score et d'étudier la différence entre l'amarrage en aveugle et l'utilisation des connaissances antérieures de l'un ou des deux sites de liaisons. Les résultats sont montrés dans le tableau 4.4. Le premier ensemble de figures dans ce tableau donne les résultats pour l'amarrage en aveugle de forme-seulement des sous-unités liées, présentés comme le rang et les déviations de la première solution trouvée dans une déviation de 10 Å RMS du complexe (ici appelé un "hit") avec le nombre de tels hits trouvés dans les 2000 premières solutions. Ce seuil correspond grosso modo à la définition d'une prédiction "acceptable" selon les critères d'évaluation de CAPRI (Méndez *et al.*, 2003). Bien que l'objectif final soit d'amarrer les sous-unités non liées, la considération des résultats d'amarrage liés donne une façon pratique d'identifier les complexes qui *a priori* sont difficiles á prévoir à amarrer acceptablement dans le cas non lié. Des solutions acceptables ont été trouvées dans les 10 premiers des 33 cas et dans les 20 premiers des 37 cas, ce qui est encourageant. Ceci montre que la fonction de score de *Hex* basée sur la forme peut souvent

identifier des orientations cristallographiques presque-natives.

Cependant, ces résultats montrent aussi le fait que *Hex* n'arrive pas à trouver une solution acceptable pour les 22 complexes lié-lié des exemples pré-établis. Une inspection visuelle de ces complexes montre que plusieurs (1AK4, 1GHQ, 1KTZ, 1BJ1, 1QFW, 2QFW et 1ATN) ont en particulier de petites aires d'interface, qui sont donc difficiles à identifier par n'importe quel algorithme d'amarrage basé sur la forme. En outre, plusieurs des autres complexes non-réussis incluent au moins un grand domaine protéique (e.g. 1KLU, 1ML0, 1KKL, 1HE8, 1N2C, 1DE4, 1H1V et 2HMI) qui ne peut pas être encodé exactement dans la fonction radiale standard de *Hex*. Ces cas seront aussi des cas difficiles pour la fonction de score de *Hex*. Parmi les complexes non-réussis restants, plusieurs sont des complexes d'anticorps/antigène (e.g. 1DQJ, 1E6J, 1WEJ, 2VIS) et il n'est pas nécessaire en général d'exécuter des calculs d'amarrage totalement en aveugle sur de tels systèmes qui sont bien compris.

Le reste du tableau 4.4 présente les résultats des amarrages des structures non liées. Comme il était prévu, le rang de la meilleure solution d'amarrage en aveugle de forme-seulement est souvent considérablement plus mauvais que ceux des composants liés d'amarrage, avec seulement 6 complexes dans les 20 premiers. En revanche, inclure le terme électrostatique d'interaction d'ETO dans la corrélation améliore souvent le rang de la meilleure solution, donnant 16 complexes parmi les 20 premiers. Cependant, utiliser des corrélations électrostatiques peut empirer la prédiction dans certains cas, mais il n'est pas clair comment on peut prédire *ab initio* ces cas là.

Néanmoins, en pratique, il devient de plus en plus rare qu'un amarrage en aveugle total soit nécessaire parce que, par exemple, pour les familles d'anticorps, les connaissances biochimiques ou biophysiques sont souvent disponibles pour identifier les résidus clés des interactions. Par conséquent, quatre autres passages contraints d'amarrage ont été exécutés pour chaque complexe afin de simuler de tels scénarios d'amarrage incorporant des données. Ici, les plages des recherches FFT ont été contraintes avec la restriction $\beta_A \leq 45°$ pour simuler en utilisant la connaissance du site de liaison du récepteur (présenté comme "One Constraint" dans le tableau) et de la même façon avec $\beta_A \leq 45°$ et $\beta_B \leq 45°$ pour simuler en utilisant les connaissances des sites de liaison du récepteur et du ligand ("Two Constraints"). Ces contraintes réduisent la taille de l'espace de recherche et des dimensions de la grille FFT correspondante par un facteur d'environ quatre et accélèrent le balayage FFT en conséquence. Ainsi, pour les passages d'amarrage contraints, les temps totaux de calcul de seulement quelques minutes résultent en grande partie de l'étape de reclassement de *L=30*. Indiquer une contrainte de $\beta_A = 45°$ sur le récepteur correspondrait physiquement à tourner un antigène au-dessus de la région hypervariable de boucle de l'anticorps dans un complexe d'anticorps/antigène, comme il est illustré dans la figure 4.1, par exemple. En général, *Hex* permet à un certain résidu de récepteur et de ligand d'être tourné sur l'axe z avant chaque passage d'amarrage. Par conséquent, en mettant de petites valeurs pour les plages angulaires β_A et β_B, il est simple de guider un calcul d'amarrage autour d'une paire de résidus donnée pour une interface protéique connue ou présumée.

Comme on peut le voir dans le tableau 4.4, les contraintes généreuses ci-dessus sont souvent

Table 4.4: Les résultats de *Hex* pour les amarrages de “Docking Benchmark” (version 2).

	B-B Shape-Only Blind Search		U-U Shape-Only Blind Search		U-U Shape+Elec Blind Search		U-U Shape-Only One Constraint		U-U Shape+Elec One Constraint		U-U Shape-Only Two Constraints		U-U Shape+Elec Two Constraints	
Code	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits
Rigid-Body (63)														
1AVX	46 (4.8)	20	108 (8.9)	7	111 (8.9)	4	40 (8.9)	12	75 (9.0)	14	18 (9.0)	43	12 (9.0)	45
1AY7	40 (8.9)	16	645 (9.9)	4	–	–	99 (3.5)	20	234 (9.8)	1	17 (6.7)	39	17 (9.7)	18
1BVN	1 (1.1)	29	63 (9.1)	20	389 (9.6)	7	29 (9.6)	35	3 (6.6)	36	4 (5.1)	49	2 (9.6)	39
1CGI	1 (0.7)	24	42 (9.4)	17	47 (4.6)	9	20 (9.4)	14	42 (9.8)	11	4 (9.4)	31	4 (4.6)	24
1D6R	273 (1.3)	24	447 (7.7)	1	119 (7.6)	4	49 (7.7)	8	31 (7.7)	8	8 (7.7)	37	5 (7.7)	31
1DFJ	167 (4.2)	14	17 (9.5)	14	1 (4.2)	30	3 (9.5)	24	1 (4.2)	30	2 (9.5)	32	1 (4.2)	35
1E6E	1 (2.1)	14	109 (5.6)	10	5 (2.2)	24	24 (5.6)	19	3 (1.5)	29	5 (5.6)	38	1 (7.7)	49
1EAW	1 (1.0)	17	9 (5.0)	20	1 (4.0)	37	7 (5.0)	25	1 (4.0)	35	1 (5.0)	42	1 (4.0)	42
1EWY	19 (7.7)	16	76 (9.1)	12	24 (9.7)	14	114 (8.1)	12	103 (6.8)	7	9 (8.1)	37	9 (7.6)	23
1EZU	2 (0.9)	13	–	–	–	–	–	–	–	–	86 (6.7)	10	287 (6.2)	4
1F34	1 (1.4)	25	124 (6.7)	11	–	–	48 (7.1)	15	–	–	11 (5.4)	22	26 (6.5)	11
1HIA	3 (1.2)	30	51 (8.7)	6	8 (8.9)	15	72 (8.7)	21	15 (9.9)	22	15 (6.7)	33	6 (8.3)	32
1MAH	1 (0.9)	16	2 (1.2)	20	1 (1.1)	28	1 (1.2)	27	1 (1.2)	30	1 (1.2)	33	1 (1.2)	30
1PPE	1 (1.0)	42	2 (9.7)	47	4 (3.0)	31	1 (9.7)	49	1 (3.0)	46	1 (3.0)	43	1 (3.0)	45
1TMQ	1 (2.1)	19	356 (5.9)	9	427 (6.0)	6	45 (5.9)	21	264 (2.3)	7	7 (5.9)	39	10 (6.6)	38
1UDI	1 (1.6)	17	8 (6.2)	9	20 (6.2)	10	4 (6.2)	22	7 (6.2)	25	1 (6.2)	32	5 (6.2)	37
2MTA	11 (1.4)	18	136 (9.0)	4	79 (9.8)	20	38 (9.0)	17	12 (8.4)	24	15 (7.7)	33	15 (8.7)	31
2PCC	1007 (9.1)	1	–	–	18 (6.9)	33	14 (9.3)	20	12 (5.1)	31	5 (9.3)	37	14 (6.3)	44
2SIC	3 (0.7)	10	57 (8.8)	8	–	–	21 (8.9)	10	44 (1.0)	9	4 (8.9)	31	4 (1.0)	35
2SNI	1 (1.5)	18	256 (9.6)	7	101 (9.6)	6	39 (7.1)	15	40 (4.4)	11	5 (7.1)	31	5 (4.4)	25
7CEI	5 (1.3)	17	61 (8.7)	5	4 (8.4)	19	11 (8.7)	17	3 (8.4)	22	2 (8.7)	29	1 (8.4)	35
1AHW	6 (1.9)	10	234 (8.0)	3	7 (8.0)	12	31 (8.0)	12	5 (8.0)	40	3 (8.0)	42	5 (8.0)	38
1BVK	44 (1.5)	6	–	–	508 (6.7)	7	134 (9.4)	7	184 (6.8)	10	71 (9.9)	23	22 (6.8)	24
1DQJ	–	–	–	–	–	–	216 (8.6)	6	440 (9.9)	2	22 (8.6)	24	73 (8.1)	11
1E6J	–	–	–	–	–	–	26 (8.9)	12	16 (8.4)	22	2 (8.9)	37	4 (8.4)	41
1JPS	24 (1.3)	5	–	–	36 (8.8)	11	170 (6.6)	9	14 (6.6)	27	15 (6.6)	29	1 (8.8)	30
1MLC	62 (1.2)	5	408 (3.6)	2	–	–	25 (3.6)	13	22 (3.7)	28	3 (3.6)	29	2 (3.7)	23
1VFB	23 (1.1)	3	–	–	–	–	97 (9.1)	14	51 (7.1)	10	14 (9.1)	36	12 (7.1)	35
1WEJ	–	–	–	–	–	–	26 (1.7)	13	2 (1.7)	20	8 (1.7)	29	1 (1.7)	37
2VIS	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1A2K	29 (5.4)	12	–	–	–	–	–	–	–	–	186 (9.3)	5	274 (9.1)	4
1AK4	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1AKJ	30 (8.4)	25	209 (9.6)	10	17 (9.4)	27	110 (6.3)	15	23 (2.7)	35	23 (9.6)	36	5 (9.6)	48
1B6C	3 (1.8)	19	593 (9.0)	2	755 (8.9)	2	88 (9.0)	5	133 (8.5)	5	19 (9.0)	27	7 (9.7)	36
1BUH	28 (1.0)	9	743 (7.7)	2	289 (7.8)	4	52 (7.7)	14	19 (7.7)	13	28 (7.7)	19	8 (7.7)	18
1E96	133 (1.1)	5	–	–	302 (8.6)	2	246 (9.4)	6	119 (8.6)	8	37 (9.7)	13	43 (8.5)	20
1F51	3 (1.4)	21	371 (9.6)	5	–	–	149 (9.6)	12	58 (9.3)	3	9 (7.6)	19	8 (7.5)	27
1FC2	605 (6.5)	2	–	–	–	–	–	–	–	–	–	–	297 (7.7)	10
1FQJ	7 (1.0)	14	41 (8.0)	12	7 (7.9)	14	14 (8.0)	21	7 (7.7)	28	5 (7.8)	31	4 (7.7)	41
1GCQ	1 (1.0)	16	–	–	–	–	–	–	–	–	92 (6.2)	6	–	–
1GHQ	–	–	–	–	–	–	828 (8.9)	2	–	–	30 (8.9)	13	175 (6.7)	6
1HE1	1 (1.5)	24	37 (6.4)	18	88 (6.3)	15	10 (6.4)	26	28 (7.2)	25	2 (7.6)	39	9 (7.2)	39
1I4D	31 (1.5)	19	–	–	–	–	–	–	–	–	505 (8.1)	1	481 (9.4)	1
1KAC	36 (1.2)	7	687 (8.7)	1	271 (8.9)	5	7 (4.4)	19	4 (4.4)	26	4 (4.4)	33	2 (4.4)	32
1KLU	–	–	–	–	–	–	–	–	–	–	591 (9.7)	2	–	–
1KTZ	–	–	–	–	–	–	–	–	–	–	238 (9.4)	4	25 (6.0)	10
1KXP	1 (1.1)	22	36 (9.4)	13	1 (7.5)	13	15 (9.4)	19	1 (6.9)	30	7 (9.4)	24	1 (6.9)	29
1ML0	–	–	–	–	–	–	7 (9.1)	8	33 (7.0)	11	1 (9.1)	22	3 (5.6)	27
1QA9	86 (5.9)	7	–	–	161 (9.9)	3	587 (7.5)	8	481 (6.8)	4	25 (5.3)	28	23 (4.5)	28
1RLB	409 (8.8)	2	–	–	–	–	–	–	–	–	305 (6.3)	7	384 (6.3)	6
1SBB	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2BTF	5 (0.8)	8	–	–	–	–	133 (8.6)	13	16 (6.7)	22	32 (8.6)	19	4 (6.7)	34
1BJ1	–	–	–	–	–	–	–	–	–	–	7 (6.7)	13	10 (6.9)	10
1FSK	10 (1.3)	16	5 (1.8)	16	6 (1.4)	10	1 (1.8)	31	1 (1.8)	31	1 (1.8)	43	1 (1.8)	46

(suite)

Table 4.2: (continue).

	B-B Shape-Only Blind Search		U-U Shape-Only Blind Search		U-U Shape+Elec Blind Search		U-U Shape-Only One Constraint		U-U Shape+Elec One Constraint		U-U Shape-Only Two Constraints		U-U Shape+Elec Two Constraints	
Code	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits	Rank (RMS)	Hits
1I9R	5 (5.7)	14	82 (2.1)	8	4 (2.1)	15	23 (2.1)	19	13 (2.1)	26	7 (2.1)	29	5 (2.1)	26
1IQD	42 (0.7)	8	–	–	760 (1.4)	3	276 (6.1)	7	5 (6.1)	16	5 (9.4)	27	3 (6.1)	29
1K4C	24 (0.7)	4	21 (9.6)	1	–	–	4 (9.6)	3	311 (9.6)	2	2 (9.6)	17	46 (9.6)	19
1KXQ	6 (5.5)	10	488 (7.1)	5	35 (6.3)	12	48 (7.1)	16	27 (7.1)	15	27 (7.1)	18	24 (7.1)	16
1NCA	1 (1.1)	11	116 (1.2)	5	139 (1.9)	3	20 (1.2)	13	8 (0.9)	16	2 (9.9)	22	3 (0.9)	30
1NSN	11 (1.7)	8	142 (1.5)	6	–	–	18 (1.5)	19	14 (1.5)	12	6 (1.5)	22	3 (1.5)	23
1QFW	–	–	–	–	–	–	–	–	–	–	333 (6.3)	3	37 (6.3)	6
2QFW	–	–	–	–	–	–	–	–	–	–	522 (9.7)	1	–	–
2JEL	10 (1.1)	10	164 (6.0)	3	–	–	7 (6.0)	27	4 (5.6)	29	6 (6.0)	39	2 (6.0)	38
Mean	**25 (4.1)**	**11**	**242 (8.4)**	**5**	**156 (8.1)**	**7**	**66 (7.6)**	**13**	**46 (7.0)**	**14**	**15 (7.3)**	**25**	**13 (6.7)**	**25**
Medium Difficulty (13)														
1ACB	36 (0.9)	8	694 (8.3)	3	674 (8.5)	2	156 (8.3)	7	163 (8.3)	1	10 (8.3)	33	88 (8.4)	14
1KKL	–	–	–	–	–	–	48 (8.6)	18	94 (8.4)	10	8 (8.7)	40	14 (8.0)	31
1BGX	1 (3.0)	3	–	–	–	–	–	–	–	–	–	–	–	–
1GP2	–	–	–	–	419 (6.9)	5	–	–	137 (7.1)	8	113 (5.6)	12	68 (7.1)	17
1GRN	1 (1.3)	13	914 (9.1)	2	586 (2.5)	5	661 (7.1)	4	27 (6.3)	23	14 (7.4)	31	20 (6.3)	29
1HE8	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1I2M	1 (1.8)	17	–	–	29 (5.4)	24	754 (8.5)	3	15 (8.5)	24	107 (6.7)	14	21 (8.5)	24
1IB1	10 (5.0)	13	–	–	–	–	–	–	–	–	14 (9.8)	13	22 (9.9)	7
1IJK	189 (3.0)	10	1012 (8.7)	3	–	–	145 (8.7)	5	383 (8.7)	1	14 (8.7)	18	70 (8.7)	5
1K5D	406 (5.9)	4	–	–	146 (7.6)	3	–	–	128 (9.1)	5	377 (7.6)	4	21 (9.7)	17
1M10	429 (9.1)	4	514 (9.5)	2	48 (9.2)	4	130 (9.5)	4	46 (9.3)	6	13 (9.5)	8	124 (8.4)	12
1N2C	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1WQ1	1 (1.5)	26	125 (7.1)	10	16 (7.2)	17	34 (7.1)	14	13 (7.1)	20	6 (7.1)	27	3 (7.1)	33
Mean	**50 (5.5)**	**8**	**782 (9.5)**	**1**	**329 (8.2)**	**5**	**306 (8.8)**	**5**	**153 (8.7)**	**8**	**58 (8.4)**	**15**	**66 (8.6)**	**15**
Difficult (8)														
1ATN	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1DE4	–	–	946 (8.6)	1	15 (8.4)	3	164 (8.6)	3	–	–	184 (8.5)	8	35 (9.9)	8
1EER	1 (4.0)	25	609 (9.2)	8	43 (9.2)	16	106 (7.6)	18	30 (7.7)	18	34 (7.6)	23	39 (7.7)	13
1FAK	–	–	–	–	–	–	–	–	–	–	768 (7.0)	2	221 (7.0)	8
1FQ1	162 (5.6)	5	–	–	–	–	469 (8.4)	2	–	–	82 (8.4)	5	508 (8.4)	3
1H1V	–	–	–	–	–	–	–	–	–	–	–	–	–	–
1IBR	4 (3.0)	27	–	–	–	–	–	–	–	–	314 (8.8)	4	68 (8.4)	6
2HMI	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Mean	**168 (7.8)**	**7**	**933 (9.7)**	**1**	**399 (9.7)**	**2**	**549 (9.3)**	**3**	**359 (9.3)**	**3**	**325 (8.8)**	**5**	**238 (8.9)**	**5**

Dans cette table, B-B dénote les amarrages de lié-lié (ou “bound-bound”) et U-U dénote ceux de non lié-non lié (ou “unbound-unbound”). Un tiret dénote le fait qu’aucune solution acceptable n’a été obtenue dans les premières 2.000 solutions et, dans ce cas, une valeur de 10Å est utilisée pour calculer la moyenne de la déviation de RMS. Les moyennes des rangs (dénotées par “Mean”) sont calculées en utilisant la formule de MLR, Eq 4.66. Pour les complexes d’anticorps/antigène (1AHW, 1BVK, 1DQJ, 1E6J, 1DQJ, 1JPS, 1MLC, 1VFB, 1WEJ, 2VIS, 1BJ1, 1FSK, 1I9R, 1IQD, 1K4C, 1KXQ, 1NCA, 1NSN, 1QFW, 2QFW, 2JEL, 1BGX, 2HMI), les coordonnées Cα du résidu 37 de la chaîne lourde ont été utilisées comme l’origine de coordonnées de l’anticorps. Pour toutes les autres structures, le centre de masse a été utilisé comme l’origine des coordonnées. Il est à noter que les complexes pré-établis incluent plusieurs complexes d’anticorps (1BJ1, 1FSK, 1I9R, 1IQD, 1K4C, 1KXQ, 1NCA, 1NSN, 1QFW, 2QFW, 2JEL, 2HMI) pour lesquels seules les coordonnées *liées* de l’anticorps de Fab sont disponibles.

suffisantes pour améliorer considérablement le rang des solutions presque-natives. Par exemple, utiliser seulement la contrainte de récepteur est suffisant pour augmenter le taux de solutions acceptables de 6 à 17 parmi les 20 premières. Ajouter le terme électrostatique de corrélation de *Hex*

amplifie cette amélioration à 28 parmi les 20 premières. Appliquer une contrainte semblable de ligand améliore le taux de succès à 48 dans les 20 premières et à 35 dans les 10 premières pour des corrélations de forme-seulement, ou à 45 dans les 20 premières et à 37 dans les 10 premières pour la forme-plus-électrostatique. En d'autres termes, le composant électrostatique aide de manière significative à identifier l'orientation générale du mode de liaison, et il peut aussi aider à distinguer une orientation presque-native parmi des orientations bien plaçées grâce à la complémentarité de forme, bien que l'amélioration avec cette dernière soit moins impressionnante. Il vaut la peine de noter que utiliser l'amarrage contraint améliore aussi les résultats de plusieurs complexes que les calculs d'amarrage de type corps-rigide ont indiqués comme difficiles (en particulier 1GHQ, 1KTZ, 1ML0, 1BJ1, 1QFW, 1KKL et 1DE4).

Afin de comparer de telles tendances plus objectivement, le tableau 4.4 présente les moyennes des résultats globaux pour chaque ensemble de calculs. Ici, on utilise la moyenne de logarithme du rang (ou "mean log rank," MLR) pour calculer la moyenne du rang de chaque premier hit acceptable selon :

$$\mathrm{MLR} = \exp\{\frac{1}{N_C}\sum_{i=1}^{N_C}\ln(\min(\mathrm{Rank}_i, 1000))\}, \tag{4.66}$$

où N_C est le nombre de complexes dans chaque catégorie des exemples pré-établis. En réduisant le nombre de mauvais résultats à 1000 dans cette formule, cela peut aider à empêcher les résultats aberrants d'influencer le score global. Ainsi les scores de MLR s'étendent de 1 (des hits de rang 1 pour tous les complexes) à 1000 (pas de hits pour aucun complexe). Les chiffres de MLR dans le tableau 4.4 montrent facilement l'avantage d'utiliser seulement une ou de préférence deux, contraintes vagues pour enrichir le nombre de prédictions ayant un classement élevé dans chaque catégorie des exemples pré-établis. Cet avantage est le plus impressionnant dans la catégorie de "Rigid-Body," ou corps-rigide, bien qu'utiliser deux contraintes améliore aussi de manière significative les résultats des catégories "Medium Difficulty" et "Difficult."

En général, les corrélations d'amarrage en aveugle de forme-seulement trouvent les solutions acceptables dans les 20 premières dans 6 cas, tandis qu'inclure l'électrostatique dans le calcul donne 16 solutions dans les 20 premières. L'application d'une seule contrainte angulaire vague pour focaliser le calcul autour du site de liaison de récepteur est suffisante pour produire des solutions acceptables dans les 20 premiers des 28 cas. Des contraintes plus poussées de recherche sur le site de liaison de ligand d'une manière semblable donnent jusqu'à 48 solutions parmi les 20 premiers.

En termes de vitesses de traitement brut, les FFT de forme-seulement et de forme-plus-électrostatique s'avèrent environ trois fois plus rapides que la FFT 1D initialement implémentée dans *Hex* (voir la section 4.1.1) mais, fait très surprenant, les FFT 3D sont aussi souvent plus rapides que les FFT 5D. D'autre part, grâce aux linéarités de la FFT, les multiples propriétés peuvent être corrélées simultanément dans la FFT 5D et l'on s'attend à ce que celle-ci soit particulièrement avantageuse pour les calculs de corrélations d'ordre supérieur des potentiels d'interaction de protéine-protéine à

multitermes basés sur la connaissance.

4.2.12 Simulation de la flexibilité de protéine pendant l'amarrage

Il devrait être souligné que l'amarrage corps-rigide est en général seulement la première étape d'un calcul d'amarrage. Si un complexe de protéines se compose de N atomes, il y a *3N-6* degrés de liberté internes disponibles aux atomes constitutifs. Puisque les protéines se composent souvent de quelques centaines de résidus d'acide aminé ou, d'une manière équivalente, de plusieurs milliers d'atomes, la dimensionnalité de cet espace est vraiment vaste. Cependant, les coordonnées atomiques internes ne sont pas complètement indépendantes mais sont, en fait, fortement contraintes par les liens covalents entre les atomes. Néanmoins, effectuer de la MD pour simuler les fluctuations atomiques et les mouvements internes dans de grandes protéines est très coûteus en calcul. En revanche, on connaît à partir des évidences expérimentales et de simulation MD que, souvent, seul un nombre relativement petit d'angles de torsion change de manière significative quand deux protéines forment un complexe. Le challenge non résolu, cependant, est d'identifier et de n'incorporer dans le calcul que ces régions flexibles d'une protéine qui doivent être modélisées pendant l'amarrage.

Actuellement, l'approche la plus prometteuse pour réduire la dimensionnalité du problème de flexibilité d'amarrage semble être l'utilisation du "mode-lent" ou des techniques de l'analyse des modes normaux (NMA; Hinsel *et al.* 1999). Ces approches sont dérivées de l'utilisation des techniques de diagonalisation de matrice pour analyser les fluctuations atomiques de grande échelle dans une protéine pendant les simulations MD. On observe souvent que seul un petit nombre de vecteurs propres les plus significatifs, ou les PC de la matrice, de fluctuation peuvent expliquer les mouvements importants dans une protéine. En d'autres termes, il y a seulement quelques (en général pas plus de 20 environ) degrés de liberté mutuellement indépendants (Amadei *et al.*, 1993). Plus récemment, des techniques ont été développées pour calculer rapidement et approximativement les composantes principales tout en évitant les dépenses de calcul pour effectuer une simulation MD complète (Tirion, 1996; de Groot *et al.*, 1997).

Dans le cadre de son projet de doctorat à Aberdeen pour simuler la flexibilité de protéine pendant les simulations d'amarrage de protéine-protéine, Diana Mustard a étudié l'utilisation d'une approche basée sur l'analyse des vecteurs propres pour échantillonner l'espace de conformation de protéine et, par conséquent, produire de multiples conformations possibles de protéines pour l'amarrage corps-rigide dans *Hex*. L'idée globale est que la flexibilité de protéines serait simulée en amarrant de multiples conformations corps rigide produites à partir des structures initiales de la protéine. Cette approche a été appliquée à neuf des cibles complexes protéiques (cibles T8-T14, T18 et T19) de l'expérience CAPRI (Mustard & Ritchie, 2005) et a été présentée lors de la deuxième réunion d'évaluation de CAPRI (2004) à Gaeta en Italie.

Afin de produire un grand nombre de conformations 3D initiales de protéines, on a utilisé les

programmes CONCOORD et DISCO basés sur les contraintes de distance (de Groot *et al.*, 1997) pour créer un ensemble de structures 3D pseudo-aléatoires. Cet ensemble de structures peut être considéré comme des points permis échantillonnés dans l'espace multidimensionnel de conformation de la protéine. La figure 4.18 montre quelques conformations de la protéine Laminin (cible T8 de CAPRI) produites par CONCOORD. Pour capturer les fluctuations les plus significatives dans cet espace, l'approche de la dynamique essentielle (ED) construit une matrice carrée de covariance $\underline{C}$ des moyennes des déviations des coordonnées x_i de chaque atome de sa position initiale non liée u_i :

$$C_{ij} = < (x_i - u_i)(x_j - u_j) >, \tag{4.67}$$

où les indices $i, j = 1...3N$ marquent les composantes des coordonnées cartésiennes des N atomes sous considération et où les parenthèses angulaires dénotent la moyenne sur toutes les structures pseudo-aléatoires échantillonnées. Par conséquent, au moins *3N* échantillons de conformation sont requis pour une analyse d'ED. Puisque la matrice de covariance est carrée-symétrique, elle peut être factorisée selon :

$$\underline{C} = \underline{T}.\underline{\Lambda}.\underline{T}^T \tag{4.68}$$

où $\underline{T}$ est la matrice des vecteurs propres $\underline{e}_k$ et où $\underline{\Lambda}$ est une matrice diagonale des valeurs propres, λ_k. Les vecteurs propres et les valeurs propres représentent les composantes principales et les normes carrées des fluctuations des coordonnées dans $\underline{C}$, respectivement. Si les vecteurs propres sont considérés par ordre de taille décroissante de leurs valeurs propres correspondantes, la majeure partie de la fluctuation est trouvée dans les premiers vecteurs propres. Ainsi la technique d'ED de contrainte de distance (DCED) capture une grande partie du mouvement interne d'une protéine tout en évitant les dépenses de calculs pour effectuer une simulation MD complète (Amadei *et al.*, 1993). La figure 4.19 montre que les premiers vecteurs propres capturent en grande partie les mouvements internes dans la protéine Laminin.

Ici, on a adapté l'approche ci-dessus pour produire des conformations possibles pour l'amarrage corps-rigide dans *Hex*, en exécutant en premier une analyse DCED sur la conformation de la structure initiale. Puisque les vecteurs propres sont orthonormés et couvrent l'espace de conformation de la protéine, une quelconque conformation 3D peut être, en principe, construite à partir d'une combinaison appropriée des vecteurs propres. Par exemple, en dénotant des vecteurs de coordonnées de protéine liée et non-liée par $\underline{B}$ et $\underline{U}$ (avec $\underline{U} = \{u_i; i = 1...3N\}$, etc.), on peut écrire

$$\underline{B} = \underline{U} + \sum_k \alpha_k \underline{e}_k. \tag{4.69}$$

Les coefficients α_k représentent les pondérations avec lesquelles les vecteurs propres devraient être combinés afin d'obtenir la conformation liée à partir des coordonnées de la structure non liée. Quand

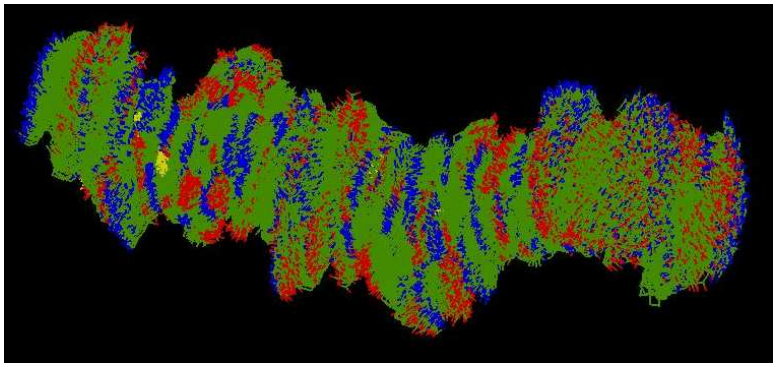

Figure 4.18: Les multiples conformations corps-rigide de la protéine Laminin (cible T8 de CAPRI) produites par CONCOORD.

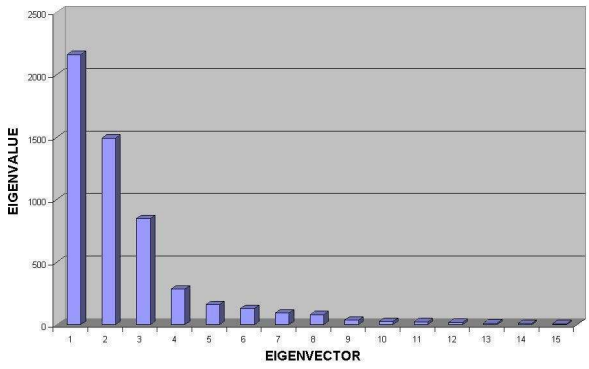

Figure 4.19: Les normes des plus grands vecteurs propres pour la protéine Laminin (cible T8 de CAPRI). Cette figure montre que les premiers vecteurs propres décrivent en grande partie les mouvements internes.

les coordonnées non-liées et liées sont connues, on appelle la quantité

$$\underline{V} = \underline{B} - \underline{U} \tag{4.70}$$

le “vecteur d’amarrage.” Si les coordonnées des deux structures liée et non-liée sont disponibles, les pondérations α_k peuvent être résolues exactement en utilisant une projection :

$$\alpha_k = \underline{V}.\underline{e}_k. \tag{4.71}$$

Ceci donne une façon utile d’évaluer l’approche en utilisant les structures des complexes de protéines connus. La figure 4.20 montre la déviation de RMS de C_α entre les conformations liées calculées et réelles en fonction du nombre de vecteurs propres utilisés dans Eq 4.69. Ceci montre que les premiers vecteurs propres peuvent expliquer en grande partie le changement de conformation du squelette qui a lieu pendant la liaison. Bien sûr, dans l’amarrage prédictif, la conformation liée désirée n’est pas

disponible et, en effet, la conformation non liée initiale a pu être obtenue par modélisation. Néanmoins, il est raisonnable de supposer que, suite à une analyse d'ED de la structure initiale, il existera une certaine combinaison des vecteurs propres les plus significatifs qui transformeront la structure initiale en une autre qui ressemble plus à la forme liée. En d'autres termes, nous supposons que chaque conformation est accessible à l'autre par les modes de fluctuation inclus dans sa structure. La tâche se réduit alors à calculer les pondérations avec lesquelles un nombre suffisant de vecteurs propres devra être combiné.

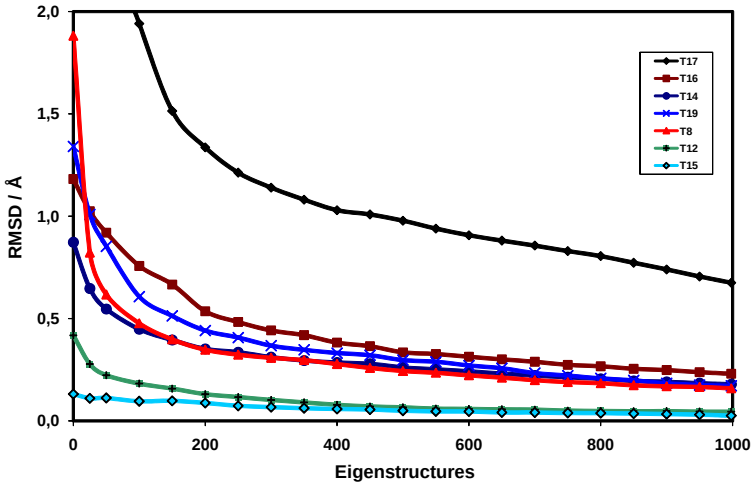

Figure 4.20: La déviation de RMS entre les conformations liées calculées et réelles des structures de ligand de CAPRI en fonction du nombre de vecteurs propres utilisés dans Eq 4.69. Des vecteurs propres sont combinés en utilisant des pondérations α_k calculées à partir d'une projection (Eq 4.71) de vecteur d'amarrage non lié-lié (Eq 4.70). Les graphes montrent que la majeure partie du changement de conformation qui se produit pendant la liaison peut être expliquée par les quelques premiers vecteurs propres.

Dans notre approche, l'analyse DCED est appliquée aux atomes lourds C_α, C, O, N et C_β (si présent). Puis, chaque conformation candidate de squelette $\underline{B}_{nj}$ est construite comme :

$$\underline{B}_{nj} = \underline{U} + \sum_{k=1}^{n} \alpha_{kj} \underline{e}_k \tag{4.72}$$

où l'indice j énumère les échantillons le long du k-ème vecteur propre $\alpha_{kj} = \pm\delta, \pm 2\delta$, etc. On a utilisé $\delta = 0.25$Å. Ainsi, chaque conformation candidate dévie de toutes les autres par un multiple intégral de 0.25Å RMS. Cependant, plusieurs de ces conformations auront des longueurs de liaison covalentes et des angles impossibles. Par conséquent, on a défini arbitrairement une des liaisons covalentes qui diffère de plus de 1% de la structure originale comme une "mauvaise liaison" et l'on a rejeté toute conformation avec plus de cinq mauvaises liaisons. La figure 4.21 montre la distribution de mauvaises liaisons pour les structures propres de la protéine Laminin. Dans notre implémentation

actuelle, jusqu'à n=8 vecteurs propres sont échantillonnés sous la contrainte que $|\alpha_{kj}| \leq \sqrt{\lambda_k}$, et ceci peut produire jusqu'à environ 10^5 conformations candidates de squelette. Appliquer notre filtre simple sur la longueur de liaison réduit souvent ce nombre à moins de 100. Puisque ces structures candidates ont des géometries de squelette très similaires à la conformation initiale, les coordonnées des atomes des chaînes latérales sont transférées directement à partir de la structure initiale. Nous appelons les structures 3D protéiques résultantes : "structures propres". Dans la présente étude, des structures propres ont été produites pour seulement une des protéines dans chaque complexe (typiquement la composante non-liée ou la plus petite) pour éviter les dépenses de calcul de multiples amarrages-croisés.

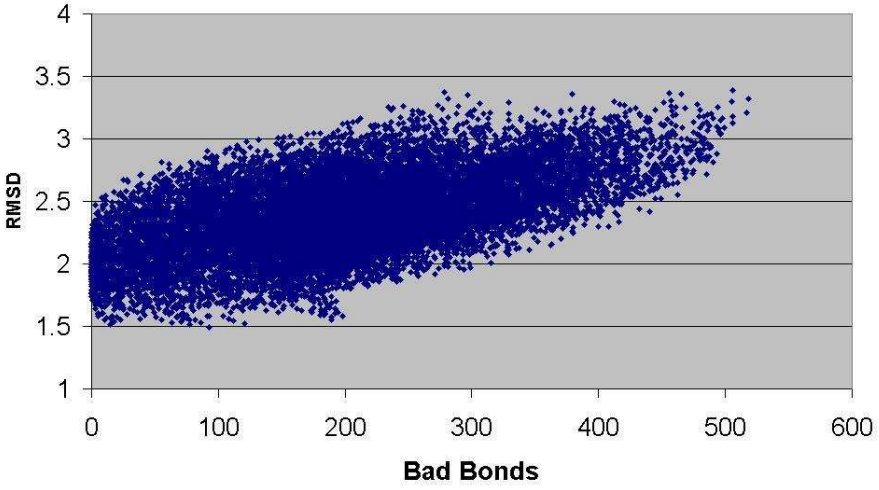

Figure 4.21: Les violations des longueurs de liaison dans les 8 premiers vecteurs propres pour la protéine Laminin.

Le tableau 4.3 résume le nombre de modèles de conformation produits et filtrés pour les cibles de CAPRI en utilisant la procédure ci-dessus. Cette table omet la structure très flexible de T9 LiCT, et T18 est aussi omis parce que le ligand TAXI dans cette cible n'a pas un squelette contigué, selon les exigences de CONCOORD. Les trois dernières colonnes de déviation de RMS montrent que, dans tous les cas, des conformations peuvent être produites qui ont une déviation inférieure C_α de la forme liée que la structure initiale non liée. Par exemple, quand on utilise une projection sur les huit premiers vecteurs propres, la déviation de C_α non lié-lié pour T8 est réduite par 0.58Å RMS de 1.88 à 1.30Å RMS. Pour T17, la réduction correspondante est 0.70Å RMS. Avec l'échantillonnage en aveugle du vecteur propre en utilisant un incrément fixe, la réduction disponible dans les déviations de RMS est inférieure à l'optimum théorique, mais elle peut encore être significative dans des cas favorables (par exemple, 0.26 et 0.21Å RMS, pour T8 et T17, respectivement). Par conséquent, le tableau 4.3 montre

que des conformations améliorées du squelette peuvent être produites avec relativement peu d'efforts avec l'approche DCED.

Table 4.3: Un résumé des structures propres produites pour les cibles de rounds 3-5 de CAPRI.

Target	Ligand	AA	EV	CS	ES	B/UB	B/ES(opt)	B/ES(δ)
T8	Laminin	162	8	405,405	624	1.88	1.30	1.62
T10	TBEV-B	395	6	8,505	49	11.33	10.90	11.13
T12	Dockerin	138	5	1,215	19	0.44	0.37	0.42
T13	SAG1	129	8	54,675	52	0.96	0.92	0.94
T14	PP1	294	8	59,535	229	0.88	0.83	0.85
T15	ImmD	87	5	2,025	6	0.15	0.11	0.15
T16	GH10	257	8	120,285	49	1.19	1.12	1.16
T17	GH11	188	8	54,675	33	5.09	4.39	4.88
T19	PrP	121	7	28,431	54	1.59	1.26	1.47

Cette table résume le nombre de modèles de conformation produits et filtrés pour les cibles (ou "Targets") T8-T19 de rounds 3-5 de CAPRI. Les colonnes sont marquées comme suit – AA: le nombre d'acides aminés dans la structure du ligand; EV: le nombre de vecteurs propres utilisé; CS: le nombre de structures candidates produites des vecteurs propres en utilisant un incrément $\delta = 0.275$Å; ES: le nombre de structures propres restantes après l'application du filtre de longueur de liaison; B/UB: la déviation de RMS de C_α entre les structures liées (ou "Bound") et non-liées (ou "Unbound"); B/ES (OPT): la meilleure déviation de RMS entre la conformation liée et la structure propre (ou "eigenstructure," ES) optimale calculée à partir des projections des 8 premiers vecteurs propres; B/ES (δ): la plus basse déviation de RMS entre la conformation liée et la meilleure structure propre trouvée en utilisant un incrément (δ) de recherche le long des 8 premiers vecteurs propres.

Afin d'étudier l'utilité de notre approche d'amarrage de structure propre, nous avons rétrospectivement amarré les structures propres produites par ED pour plusieurs des cibles de CAPRI qui étaient faciles à analyser par l'approche de DCED. Puisque les structures cristallographiques des complexes avaient été révélées, nous avons choisi de commencer chaque passage d'amarrage avec les structures propres initialement superposées sur les atomes C_α du complexe. Tous les passages d'amarrage dans ce test ont utilisé des corrélations de forme-seulement avec *N=30* et 45 degrés angulaires de contrainte de recherche sur chaque protéine. Dans certains cas, les paramètres de l'incrément de vecteur propre et de filtre de longueur de liaison ont été un peu modifiés (par exemple, pour T8 et T14) afin d'obtenir une liste plus raisonnable de structures propres ($<$100) à amarrer. Chaque liste de structures propres a été alors initialisée avec les conformations des structures non-liées et liées de ligand afin de faciliter la comparaison de l'amarrage de trois types de structures. En raison du grand nombre d'échantillons d'orientations produit, toutes les solutions d'amarrage ont été triées et clusterisées comme il a été décrit précédemment (Ritchie, 2003). Le coût total des calculs était autour de 12 heures par complexe sur un processeur AMD Athlon de 1.8GHz.

Table 4.4: Les résultats d'amarrage de structures propres pour les cibles de rounds 3-5 de CAPRI.

Target	Docked ES	Bound	RMS	Unbound	RMS	ES	RMS
T8	94	84(1/2)	9.71	30(40/94)	8.80	30(1/94)	8.24
T11	37	19(1/5)	5.52	2(29/183)	9.55	2(1/183)	9.20
T12	40	1(1/90)	0.64	1(23/90)	1.53	1(6/90)	1.53
T13	52	5(1/9)	1.17	1(32/306)	0.96	1(1/306)	6.24
T14	60	16(1/3)	9.95	10(10/177)	8.81	10(1/177)	8.81
T15	39	20(6/17)	7.80	8(20/77)	3.47	8(1/77)	4.94
T17	33	3(1/8)	1.56	–	–	12(1/43)	8.64
T19	40	1(1/12)	0.95	13(46/66)	7.70	13(1/66)	5.28

Ce tableau présente le nombre de structures propres de ligand amarrées pour chaque cible (ou "Target") suivi par le rang du cluster de la première solution avec une déviation C_α de 10Å RMS ou moins, obtenue en amarrant les structures propres liées, non liées et produites par DCED, respectivement. Les figures entre parenthèses (n/m) donnent le rang de l'orientation (n) dans le cluster d'une taille donnée (m). Un tiret dénote dénote qu'aucune solution de basse RMS n'a été trouvée dans les 512 premiers clusters.

En général, nos études initiales sur des structures propres de DCED semblent très prometteuses. Nous avons montré que les premiers vecteurs propres encodent intrinsèquement une grande partie de la flexibilité de conformation de squelette observée pendant la liaison. Les résultats avec les cibles de CAPRI montrent que l'amarrage des multiples structures propres produites des huit premiers vecteurs propres est suffisant pour donner de meilleures prédictions d'amarrage que l'amarrage seul des structures initiales non-liées ou modélisées. Cependant, il est évidemment possible d'améliorer la qualité des conformations produites. Par exemple, l'augmentation du nombre de vecteurs propres échantillonnés et l'utilisation d'un incrément variable pour rechercher le long de chaque vecteur propre donneront une meilleure couverture de l'espace de conformation accessible à chaque protéine. En plus, utiliser une méthode plus sensible pour estimer et possiblement minimiser les énergies internes des structures propres produites devrait donner des structures physiquement plus réalistes dans cet espace. Cependant, il sera aussi nécessaire de faire en sorte que la fonction de score d'amarrage soit plus sélective afin d'identifier une conformation presque-native du complexe d'un grand répertoire de leurres physiquement réalistes.

4.3 Criblage virtuel de petites molécules

En général, il y a deux approches principales du criblage virtuel. Dans des approches basées sur le récepteur, la structure de la protéine cible est connue ou a été modélisée et le but est de trouver des ligands appropriés qui se lieront près du site actif du récepteur et, par conséquent, inhiberont (bloquent) la fonction native, par exemple. Dans des approches basées sur le ligand, la structure de la protéine cible n'est en général pas connue et le but est de trouver de nouveaux ligands qui soient similaires aux antagonistes connus. Certaines molécules à vertus thérapeutiques agissent

comme des agonistes (c.-à-d. qu'elles activent ou améliorent la fonction native du récepteur) mais les principes du criblage s'appliquent de la même façon aussi bien pour les agonistes que pour les antagonistes. En raison des dépenses informatiques des approches basées sur le récepteur (c.-à-d. des approches d'amarrage), les campagnes de criblage virtuel de haut-débit (ou "high-throughput virtual screening", HTVS) utilisent souvent une combinaison d'approches, dans laquelle des critères basés sur la similarité du ligand sont utilisés comme un premier filtre et les composés candidats qui ressortent de ce filtre sont ensuite amarrés à la protéine cible.

Actuellement, les techniques les plus utilisées pour faire des recherches rapides sur de grandes bases de données chimiques pour le criblage virtuel basé sur le ligand, utilisent des représentations de "bit-string" des propriétés et topologies moléculaires telles que les empreintes (ou "fingerprints") de Daylight, d'UNITY et de MACCS. Cependant, par leur nature, ces représentations ont tendance à trouver des analogues chimiques proches de la requête donnée, qui peuvent ne pas être suffisamment nouvelles pour justifier un programme de développement de médicaments. D'un autre côté, l'action pharmacologique de la plupart des molécules de médicaments est régie par leur mode de liaison avec leurs cibles biologiques *via* la liaison ligand-récepteur. Par conséquent, on pourrait raisonnablement prévoir que des molécules ayant des formes globales similaires se lient à une protéine de façon semblable. Cependant, la comparaison des formes 3D moléculaires est considérablement plus coûteuse en calculs que la comparaison des bit-strings (Lemmen & Lengauer, 2000).

À mon avis, les approches "state-of-the-art" actuelles pour faire de la comparaison efficace des formes 3D moléculaires sont basées sur les représentations gaussiennes de la forme moléculaire (Grant *et al.*, 1996) et, récemment, les SH de l'enveloppe de surface. Cette approche dernière a été développée indépendamment par moi-même à Aberdeen (Ritchie & Kemp, 1999; Mavridis *et al.*, 2007), Tim Clark à Erlangen (Lin & Clark, 2005) et Bernard Maigret à Nancy (Cai *et al.*, 2002; Yamagishi *et al.*, 2006). À Aberdeen, l'objectif était d'exploiter les propriétés de rotation des fonctions SH pour développer une méthode très rapide de superposition et de comparaison quantitative des formes 3D des surfaces moléculaires. À Erlangen, le but était de représenter les propriétés clés des surfaces moléculaires de QM en utilisant la représentation SH. À Nancy, l'objectif principal était d'employer les représentations SH pour fournir un filtre rapide basé sur la forme pour l'amarrage protéine-ligand. Cette approche a été récemment incorporée dans le projet VSM-G (manager de grilles pour le criblage virtuel, ou "virtual screening manager for grids") (Beautrait *et al.*, 2008). Pour compléter le travail d'Erlangen, j'ai récemment développé le programme ParaFit qui peut superposer des structures 3D moléculaires calculées par ParaSurf à un taux pouvant aller jusqu'à cent molécules par seconde sur un seul processeur. ParaFit et ParaSurf sont actuellement commercialisés par Cepos Insilico Ltd.

4.3.1 Le programme ParaFit

ParaFit superpose et compare des molécules en utilisant des expansions SH des surfaces moléculaires et des propriétés moléculaires locales de surface calculées par ParaSurf (Lin & Clark, 2005). En exploitant les propriétés de rotation spéciales des fonctions SH de base (Wigner, 1939; Rose, 1957), les temps de calculs peuvent être réduits par plusieurs ordres de grandeur comparés aux algorithmes conventionnels d'appariement de formes (Ritchie & Kemp, 1999; Ritchie & Kemp, 2000). ParaFit fournit trois modes principaux de calcul. Dans le mode par défaut d'appariement, ParaFit superpose une ou plusieurs molécules mobiles sur une seule molécule fixe de référence. Le programme peut aussi exécuter des "toute-contre-toute" superpositions dans lesquelles chaque molécule est superposée alternativement sur toutes les autres. En ce mode de "matrice", un tableau de scores de distance est écrit dans un format approprié pour des analyses subséquentes de clustering, par exemple. ParaFit peut aussi être utilisé pour aligner des molécules sur les axes de coordonnées afin de les placer dans une orientation standard ou "canonique." Ceci est souvent une première étape utile dans des études de QSAR. ParaFit peut aussi appliquer des transformations arbitraires de coordonnées à une liste donnée de SDFs (fichiers de description des structures) créée par ParaSurf. Ces transformations pourraient être fournies en tant qu'élément d'un pipeline de traitement par d'autres programmes de superposition qui ne peuvent pas tourner les propriétés complexes de QM tels que des moments quadrupolaires et octupolaires et des éléments orbitaux atomiques de matrice de charge-densité. La capacité de ParaFit de considerer toutes les informations de QM dépendante de l'orientation dans un SDF élimine la nécessité de recalculer de grandes quantités de QM pour de nouvelles orientations moléculaires.

4.3.2 Similarité de forme de surface harmonique sphérique

Des formes de surface moléculaires SH sont représentées comme des expansions radiales bidimensionnelles de la forme

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\phi), \tag{4.73}$$

où $y_{lm}(\theta,\phi)$ sont les fonctions SH réelles normalisées et a_{lm} sont les coefficients d'expansion. Les coordonnées sont définies par rapport à l'origine de coordonnée harmonique (CoH), qui est en général équivalente au centre de gravité moléculaire (CoG). Afin de calculer une superposition entre une paire de molécules, ParaFit translate le CoH de la molécule mobile (B) à celle de la molécule fixe de référence (A) et puis il cherche la rotation qui minimise la "distance" entre les paires correspondantes d'expansions SH :

$$D_{\text{Euclidean}} = \int_0^{\pi} \int_0^{2\pi} \left(r_A(\theta,\phi) - \hat{R}(\alpha,\beta,\gamma) r_B(\theta,\phi)\right)^2 \sin\theta \mathrm{d}\theta \mathrm{d}\phi. \tag{4.74}$$

En utilisant les matrices réelles de rotation de Wigner (Eq 3.35) pour tourner les coefficients d'expansion et en exploitant l'orthonormalité des fonctions de base, cette expression se réduit à

$$D_{\text{Euclidean}} = |\underline{a}|^2 + |\underline{b}|^2 - 2\underline{a}.\underline{b}'. \tag{4.75}$$

où $\underline{b}'$ représente le vecteur des coefficients d'expansion tournés SH de la molécule mobile, etc. Cette fonction a des unités de Å^2 et dépend clairement de la taille relative des molécules comparées. Ceci s'appelle une fonction de distance euclidienne à cause de son analogie avec les distances euclidiennes dans l'espace 3D ordinaire. Cependant, en comparant de multiples molécules, il est souvent commode d'utiliser les fonctions normalisées de similarité dans lesquelles les molécules identiques donnent un score d'unité. Par exemple, diviser par la somme des grandeurs des vecteurs de forme SH donne le score de similarité de Hodgkin :

$$S_{\text{Hodgkin}} = \frac{2\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2} = 1 - \frac{D_{\text{Euclidean}}}{|\underline{a}|^2 + |\underline{b}|^2}. \tag{4.76}$$

ParaFit implémente aussi des fonctions de similarité de Carbo et de Tanimoto :

$$S_{\text{CARBO}} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|^2.|\underline{b}|^2}. \tag{4.77}$$

et

$$S_{\text{TANIMOTO}} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}'}. \tag{4.78}$$

ParaFit utilise la fonction de Tanimoto comme sa fonction de similarité par défaut. Il n'est généralement pas évident de déterminer quelle fonction de score ci-dessus doit être privilégiée. D'après mon expérience, elles donnent toutes de bonnes superpositions de paires avec un ordre d'expansion par défaut *L=6*.

ParaFit superpose des molécules en utilisant une recherche de rotation force brute sur les trois angles de rotation d'Euler. Conceptuellement, chaque molécule mobile est tournée par rapport à la molécule fixe de référence et la rotation d'Euler qui donne le plus grand score de similarité (ou la plus petite distance) est notée. Ceci est essentiellement une recherche de corrélation de Fourier en des coordonnées d'angle d'Euler. Cependant, puisque de bonnes superpositions peuvent être réalisées en utilisant seulement des expansions harmoniques d'ordre inférieur, il n'est pas nécessaire d'employer des techniques de la transformée rapide de Fourier (FFT) pour accélérer le calcul sauf *L*$\geq$*16* (voir la section 4.1.1).

En plus d'utiliser des recherches de corrélation d'ordre inférieur, les calculs de superposition de ParaFit sont accélérés de deux autres manières. La première technique exploite le fait que les expansions harmoniques d'ordre L ne peuvent pas avoir plus de L^2 éléments maxima locaux. Par conséquent, ParaFit utilise initialement des incréments de recherche angulaires relativement grands d'environ 8° pour couvrir l'espace de recherche. Afin d'échantillonner l'espace angulaire de façon

régulière et efficace, ces échantillons angulaires sont produits à partir des sommets d'une tessellation icosaèdrique de la sphère (section 2.3). Pour un incrément angulaire donné, ceci donne environ 30% de moins de points échantillonnés qu'une grille naïve équi-angulaire (Ritchie & Kemp, 1999). Une fois que la position approximative de similarité maximum a été identifiée, elle est alors raffinée en utilisant une recherche localisée de grille par des incréments de 2°. Les deux incréments angulaires peuvent être ajustés par l'utilisateur.

La deuxième technique d'accélération est employée pour comparer de multiples molécules. Plutôt que de tourner séparément chacune des molécules mobiles alternativement, il est beaucoup plus efficace de tourner les expansions SH de la molécule de référence seulement et de comparer ces dernières avec chacune des molécules mobiles. Ainsi, des rotations SH relativement coûteuses sont appliquées seulement à une au lieu de N molécules. Une fois que les rotations optimales ont été trouvées, les molécules mobiles sont tournées en utilisant l'inverse des rotations correspondantes de référence. En utilisant ces techniques, une paire de molécules peut être superposée dans environ 0.05 secondes sur un processeur Xeon Pentium de 1.8GHz et les temps de calcul peuvent être réduits de plus par un facteur pouvant aller jusqu'à cinq lorsque de multiples molécules sont comparées dans un seul passage de ParaFit.

4.3.3 Empreintes rotation-invariables et orientations canoniques

Bien que les paires de molécules puissent être superposées et comparées très rapidement en utilisant les techniques mentionnées ci-dessus, il est nécessaire de développer des techniques de comparaison encore plus rapides afin de faire des requêtes dans de très grandes bases de données structurales 3D (par exemple $> 10^6$ molécules) pour le HTVS. Il est donc normal d'utiliser l'interprétation de vecteur des coefficients SH pour construire une empreinte rotation-invariable (ou "rotation-invariant fingerprint", RIF) pour chaque forme moléculaire SH. En notant que des coefficients d'expansion avec la même valeur de l se transforment sous une rotation, les coefficients de RIF sont définis comme :

$$A_l = \left(\sum_{m=-l}^{m=l} a_{lm}^2 \right)^{1/2} \tag{4.79}$$

et

$$A_L = \left(\sum_{l=0}^{L} A_l^2 \right)^{1/2}. \tag{4.80}$$

Par analogie avec l'équation 4.75, les scores de distance de RIF entre une paire de molécules peuvent être calculés comme

$$D_{RIF} = A_L^2 + B_L^2 - 2 \sum_{l=0}^{L} A_l B_l. \tag{4.81}$$

Une étude de la performance de ces fonctions de score de rotation-dépendantes et de rotation-invariables a été publiée (Mavridis *et al.*, 2007). Cette étude a montré que de très bonnes superpositions peuvent être obtenues en utilisant des expansions avec *L=3* seulement et qu'utiliser des ordres d'expansion au delà de *L=6* donne peu ou pas d'amélioration dans la qualité des résultats si l'on compare avec les superpositions d'ordre supérieur *L=15* qui sont traitées comme les "étalon or".

Bien que des comparaisons de RIF puissent être calculées très rapidement, elles sont considérablement moins exactes que les recherches les plus coûteuses de corrélation de rotation. Par conséquent, nous avons étudié la possibilité de comparer des formes moléculaires dans des orientations standard, ou "canonicalisées," comme une façon de retenir une grande partie de la précision des comparaisons de rotation tout en évitant les dépenses informatiques d'une recherche de rotation complète. Ici, une orientation canonique est calculée en utilisant une expansion SH de forme avec *L=6* de sorte que sa plus grande étendue radiale ait été alignée avec l'axe global z et ensuite en appliquant une rotation-z pure pour placer l'étendue équatoriale maximale sur l'axe positif x. Cette procédure est en quelque sorte semblable à l'alignement des moments d'inertie avec les axes principaux (Lanzavecchia *et al.*, 2001), mais l'utilisation des expansions avec *L>2* élimine toute ambiguïté par rapport aux rotations ou "flips" de $180°$. La figure 4.22 illustre les orientations canoniques de quatre molécules de benzodiazépine.

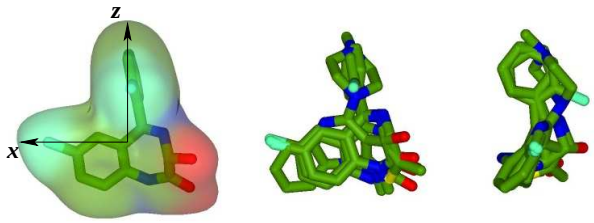

Figure 4.22: Une illustration de l'alignement canonique des quatre agonistes récepteurs de GABA, les benzodiazépines lorazépam, diazépam, temazépam et clonazépam. À gauche : la surface moléculaire SH avec *L=6* et l'orientation canonicalisée de lorazépam; Au milieu : les quatre benzodiazépines canonicalisées ensemble; À droite : les mêmes orientations tournées par $90°$ autour de l'axe z.

4.3.4 Grouper et classifier les données de *Drug* et d'*Odour*

Afin de comparer l'approche de comparaison de surface SH aux autres mesures traditionnelles de similarité moléculaire, des analyses de cluster des deux ensembles de données moléculaires, ici appelés *Drug* et *Odour*, ont été effectuées et les résultats ont été comparés, avec ceux du clustering physico-chimique (PC) de l'ensemble des données de *Drug*, aux résultats de clustering basés sur la fréquence vibratoire de l'ensemble de données *Odour* obtenu par Takane et Mitchell (2004). Ce travail a également été réalisé dans le cadre du projet de thèse de doctorat de Lazaros Mavridis.

L'ensemble de données *Drug* a été initialement classifié par mon collègue Dr Brian Hudson dans six larges catégories pharmacologiques et jusqu'à sept sous-groupes basés sur les mécanismes d'action pharmacologiques connus. Ceci a donné un total de 22 classes de médicament, comme il est montré dans le tableau 4.5. Il doit être noté que cette classification n'est pas unique parce que plusieurs de ces composés ont de multiples modes d'action et sont utilisés à diverses fins thérapeutiques. Néanmoins, la classification experte donne une indication de similarité pharmacologique avec laquelle des clusters calculés peuvent être comparés.

Pour le clustering PC, 11 descripteurs moléculaires (y compris l'aspect de polarité, le rayon de la giration, le poids moléculaire, le logP, etc.) ont été calculés pour l'ensemble de données *Drug* en utilisant Cerius-2[3] et ceux-ci ont été auto-mesurés et groupés en utilisant l'algorithme de clustering agglomératif de Ward (1963) pour produire un total de 22 clusters, comme il est montré dans la figure 4.23. Pour le clustering SH, une matrice de distance de forme-seulement pour le même groupe de molécules avec des expansions *L=6* a été calculée et l'algorithme de Ward a été appliqué directement pour produire 22 clusters. Ceux-ci sont également montrés dans la figure 4.23. Cette figure indique que les deux méthodes de clustering groupent souvent les classes similaires de médicaments dans le même cluster ou dans un cluster similaire. Par exemple, les deux approches de clustering PC et SH groupent les antibiotiques (AB) ensemble et les deux approches groupent plusieurs des médicaments de tranquillisants et d'antidépresseurs (système nerveux central; CN) étroitement ensemble. Ceci suggérerait que l'utilisation des surfaces SH pour classifier les molécules est au moins aussi bonne que des méthodes traditionnelles basées sur les propriétés globales moléculaires. En effet, une comparaison des deux dendrogrammes suggère que le clustering SH tend à placer plus de molécules pharmacologiquement apparentées dans les groupes plus étroitement liés que le clustering PC. Par exemple, le clustering SH place les médicaments gastro-intestinaux (GI) dans le même groupe, tandis que ces médicaments sont répartis sur quatre groupes distincts dans le clustering PC. Pour les composés de CN, un groupe contient les benzodiazépines clonazépam, lorazépam et diazépam; et un autre groupe contient essentiellement des composés apparentés à l'activité de GPCR tels que les inhibiteurs de la recapture de la sérotonine amitriptyline, nortriptyline, citalopram, fluoxétine et paroxétine aussi bien que l'antagoniste de récepteur de sérotonine olanzapine. Ces caractéristiques de l'analyse de cluster ne sont pas concluantes mais néanmoins encourageantes.

En guise d'un autre test de l'approche SH, les 46 molécules de l'ensemble de données d'*Odour* ont été groupées dans dix groupes à l'aide des descripteurs de forme SH avec *L=6* en utilisant des superpositions de rotation et en comparant les molécules dans leurs orientations canoniques préalignées. Takane et Mitchell (2004) ont, à l'origine, groupé cet ensemble de données dans dix groupes distincts en utilisant des descripteurs de valeur propre (ou "eigenvalue," EVA) dérivés des calculs de fréquences vibratoires de QM. Par conséquent, le même nombre de clusters a été utilisé dans la présente étude pour faciliter la comparaison avec les résultats de SH. La figure 4.25 montre les clus-

[3] http://www.accelrys.com/.

Table 4.5: La classification pharmacologique des 73 molécules de l'ensemble de données *Drug*.

Nom	Classe	Mots Clés du World Drug Index
MINOCYCLINE	AB 1	ANTIBIOTICS
DOXYCYCLINE	AB 1	ANTIBIOTICS
TETRACYCLINE	AB 1	ANTIBIOTICS
CEFPROZIL	AB 1	ANTIBIOTICS
CLINDAMYCIN	AB 1	ANTIBIOTICS
IBUPROFEN	AI 1	ANALGESICS; ANTIINFLAMMATORIES; ANTIPYRETICS
ASPIRIN	AI 1	ANALGESICS; ANTIINFLAMMATORIES; ANTIPYRETICS; ANTICOAGULANTS
DICLOFENAC	AI 1	ANALGESICS; ANTIINFLAMMATORIES; PROSTAGLANDIN-ANTAGONISTS
NAPROXEN	AI 1	ANALGESICS; ANTIINFLAMMATORIES; PROSTAGLANDIN-ANTAGONISTS; ANTIPYRETICS
CODEINE	AI 1	ANALGESICS; ANTITUSSIVES; NARCOTICS
CARISOPRODOL	AI 1	ANALGESICS; RELAXANTS
LORATADINE	AI 2	ANTIHISTAMINES-H1
CETIRIZINE	AI 2	ANTIHISTAMINES-H1
PROMETHAZINE	AI 2	ANTIHISTAMINES-H1; SEDATIVES
TRIAMCINOLONE	AI 3	CORTICOSTEROIDS
METHYLPREDNISOLONE	AI 3	CORTICOSTEROIDS
BUDESONIDE	AI 3	CORTICOSTEROIDS
PREDNISONE	AI 3	CORTICOSTEROIDS
CLONAZEPAM	CN 1	ANTICONVULSANTS
GABAPENTIN	CN 1	ANTICONVULSANTS
PHENYTOIN	CN 1	ANTICONVULSANTS
TOPIRAMATE	CN 1	ANTICONVULSANTS
SERTRALINE	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
FLUOXETINE	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
NORTRIPTYLINE	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
AMITRIPTYLINE	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
PAROXETINE	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
CITALOPRAM	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
BUPROPION	CN 2	ANTIDEPRESSANTS; PSYCHOSTIMULANTS
OLANZAPINE	CN 3	PSYCHOSEDATIVES; DOPAMINE-ANTAGONISTS; NEUROLEPTICS
RISPERIDONE	CN 3	PSYCHOSEDATIVES; NEUROLEPTICS; ANTISEROTONINS; DOPAMINE-ANTAGONISTS
LORAZEPAM	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS
BUSPIRONE	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS
DIAZEPAM	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS
TEMAZEPAM	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS; ANTICONVULSANTS
TRAZODONE	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS; PSYCHOSTIMULANTS; ANTIDEPRESSANTS
CYCLOBENZAPRINE	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS; RELAXANTS
ZOLPIDEM	CN 3	PSYCHOSEDATIVES; TRANQUILIZERS; SEDATIVES
FENOFIBRATE	CV 1	ANTIARTERIOSCLEROTICS
GEMFIBROZIL	CV 1	ANTIARTERIOSCLEROTICS
SIMVASTATIN	CV 1	ANTIARTERIOSCLEROTICS; HMG-COA-REDUCTASE-INHIBITORS
PRAVASTATIN	CV 1	ANTIARTERIOSCLEROTICS; HMG-COA-REDUCTASE-INHIBITORS
WARFARIN	CV 2	ANTICOAGULANTS
NIFEDIPINE	CV 3	CARDIANTS; CALCIUM-ANTAGONISTS
DILTIAZEM	CV 3	CARDIANTS; CALCIUM-ANTAGONISTS
VERAPAMIL	CV 3	CARDIANTS; CALCIUM-ANTAGONISTS; PROTEIN-KINASE-C-INHIBITORS
TRIAMTERENE	CV 4	DIURETICS
SPIRONOLACTONE	CV 4	DIURETICS; ALDOSTERONE-ANTAGONISTS
HYDROCHLOROTHIAZIDE	CV 4	DIURETICS; CARBONIC-ANHYDRASE-INHIBITORS; HYPOTENSIVES
FUROSEMIDE	CV 4	DIURETICS; PROTEIN-KINASE-C-INHIBITORS
VALSARTAN	CV 5	HYPOTENSIVES
TERAZOSIN	CV 5	HYPOTENSIVES
CAPTOPRIL	CV 5	HYPOTENSIVES; ANGIOTENSIN-ANTAGONISTS
FOSINOPRIL	CV 5	HYPOTENSIVES; ANGIOTENSIN-ANTAGONISTS
DOXAZOSIN	CV 5	HYPOTENSIVES; SYMPATHOLYTICS-ALPHA
BISOPROLOL	CV 5	HYPOTENSIVES; SYMPATHOLYTICS-BETA; ANTIARRHYTHMICS
CARVEDILOL	CV 5	HYPOTENSIVES; SYMPATHOLYTICS-BETA; VASODILATORS
CLONIDINE	CV 5	HYPOTENSIVES; SYMPATHOMIMETICS-ALPHA
ATENOLOL	CV 6	SYMPATHOLYTICS-BETA
METOPROLOL	CV 6	SYMPATHOLYTICS-BETA
TIMOLOL	CV 6	SYMPATHOLYTICS-BETA
FAMOTIDINE	GI 1	GASTRIC-SECRETION-INHIBITORS; ANTIHISTAMINES-H2
RANITIDINE	GI 1	GASTRIC-SECRETION-INHIBITORS; ANTIHISTAMINES-H2; ANTIULCERS
LANSOPRAZOLE	GI 2	GASTRIC-SECRETION-INHIBITORS; H-K-ATPASE-INHIBITORS
OMEPRAZOLE	GI 2	GASTRIC-SECRETION-INHIBITORS; H-K-ATPASE-INHIBITORS; ANTIULCERS
GLIPIZIDE	OT 1	ANTIDIABETICS
METFORMIN	OT 1	ANTIDIABETICS
METOCLOPRAMIDE	OT 2	ANTIEMETICS; DOPAMINE-ANTAGONISTS
ALLOPURINOL	OT 3	ANTIGOUTS; ANTIRHEUMATICS
CARBIDOPA	OT 4	ANTIPARKINSONIANS; DOPA-DECARBOXYLASE-INHIBITORS
ESTRADIOL	OT 5	ESTROGENS
FLUCONAZOLE	OT 6	FUNGICIDES
TAMOXIFEN	OT 7	PROTEIN-KINASE-C-INHIBITORS; ESTROGEN-ANTAGONISTS

Chaque molécule de médicament dans ce tableau a été assignée à un code à deux lettres par un expert en chemoinformatique selon sa classe pharmacologique principale comme suit, AB : antibiotique; AI : anti-inflammatoire; CN : système nerveux central; CV : cardio-vasculaire; GI : gastro-intestinal; OT : autre. Les chiffres numériques indiquent des sous-classes dans chacune des six classes pharmacologiques principales.

-ters résultants de SH avec les superpositions 3D correspondantes. Les deux méthodes de SH et d'EVA donnent largement des groupements similaires. Cependant, le clustering de SH distingue bien les molécules de camphre et d'amande amère comme deux groupes séparés, tandis que le clustering précédent d'EVA divise les molécules de camphre en deux sous-groupes, dont l'un inclut une odeur de jasmin et deux de rose (voir le tableau 3 de Takane et Mitchell, 2004). Le clustering d'EVA divise également les odeurs d'amande amère en trois groupes distincts, tandis que le clustering SH assigne correctement ces molécules à deux sous-groupes voisins. Le clustering de SH place également toutes sauf une des odeurs de rose et de jasmin dans deux sous-groupes étroitement liés. En général, la figure 4.25 montre une correspondance saisissante entre la classification SH basée sur la forme et les superpositions correspondantes de forme moléculaire.

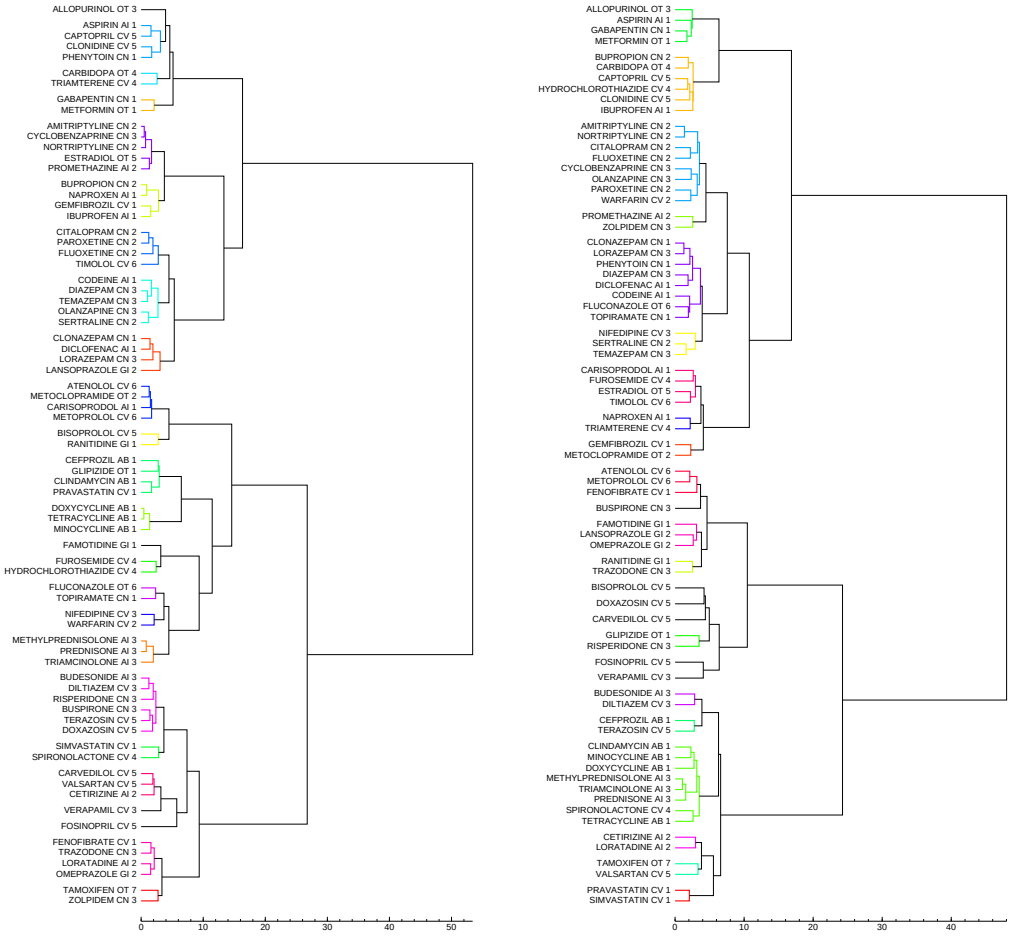

Figure 4.23: Des dendrogrammes des 73 molécules de l'ensemble de données de *Drug*, calculés en utilisant l'algorithme du clustering agglomératif de Ward pour donner 22 clusters. À gauche : le clustering conventionnel chimique utilisant 11 descripteurs macroscopiques auto-mesurés de PC. À droite : le clustering de formes de surface SH avec L=*6*.

Les résultats de clustering des ensembles de données de *Drug* et d'*Odour* montrent que les comparaisons de forme SH donnent souvent des groupements chimiquement signicatifs. Nos résultats pour l'ensemble de données d'*Odour* montrent une correspondance saisissante entre la classification basée sur la SH et les superpositions correspondantes de forme moléculaire. En effet, le clustering de forme SH réalise des clusters comparables ou meilleurs que le travail précédent basé sur l'analyse de fréquence vibratoire de QM qui est plus onéreuse en calculs. L'analyse de l'ensemble des données de *Drug* est également encourageante dans le sens qu'en dépit d'utiliser seulement des expansions de forme, il est possible d'identifier des groupements pharmacologiques similaires qui semblent être au moins aussi bons que ceux produits en utilisant des analyses traditionnelles de PC.

La figure 4.26 montre les résultats du clustering de l'ensemble de données *Odour* en utilisant les comparaisons de surface SH des molécules dans leurs orientations canoniques. En comparant ces résultats à la figure 4.25, on peut voir qu'une comparaison des molécules dans des orientations canoniques est presque aussi fiable que des comparaisons rotationnelles par paires. En outre, si les vecteurs de coefficient SH moléculaire sont comparés dans leurs orientations canonicalisées, le coût informatique de la comparaison n'est essentiellement pas plus élevé que celui d'une comparaison rotation-invariable.

D'autres expériences sur l'ensemble de données de *Drug*, qui a été augmenté par un grand nombre de molécules similaires de leurre, ont confirmé que les comparaisons canoniques sont beaucoup plus exactes que des comparaisons rotation-invariables (Mavridis *et al.*, 2007). On peut donc en conclure que, pour obtenir une meilleure performance sur une base de données en utilisant des représentations SH, les coefficients SH moléculaires devraient être pré-calculés et sauvegardés dans la base de données en utilisant des orientations canonicalisées.

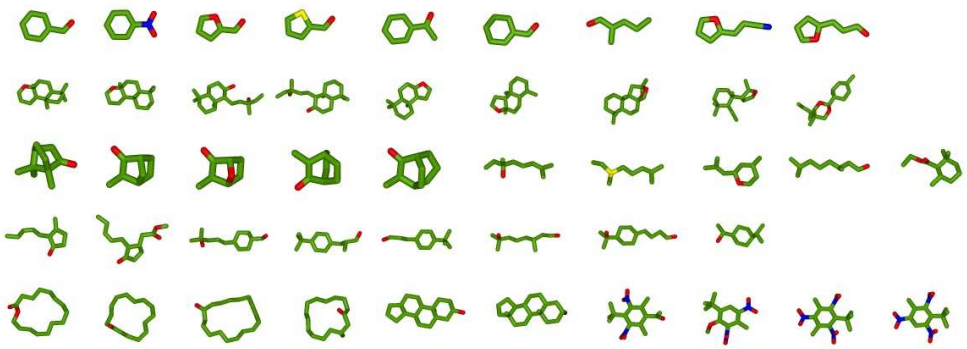

Figure 4.24: Les 46 molécules dans l'ensemble de données d'*Odour*. Ces molécules peuvent être classifiées dans 7 odeurs principales : amer, ambre gris, camphre, rose, jasmin, muguet et musc.

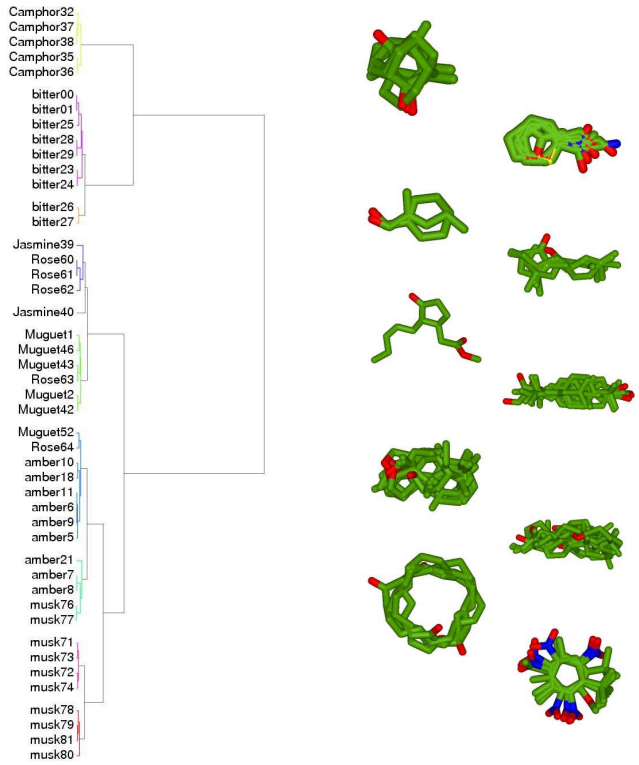

Figure 4.25: Le clustering de forme de surface SH de l'ensemble de données d'*Odour*. À gauche : le dendrogramme obtenu en utilisant des calculs de similarité de forme SH avec $L=6$. À droite : les superpositions correspondantes 3D moléculaires.

4.3.5 Criblage virtuel des molécules qui bloquent l'entrée du HIV

Selon l'organisation mondiale de la santé ("World Health Organisation", WHO), 33 millions de personnes environ vivent actuellement avec le syndrome d'immunodéficience acquise (AIDS).[4] La cause principale du AIDS est une infection par le virus de l'immunodéficience humaine (HIV) qui, au niveau moléculaire, commence par une liaison de la glycoprotéine virale de l'enveloppe gp120 au récepteur cellulaire de surface CD4 et à un des co-récepteurs de chémokine CXCR4 ou CCR5 qui, par la suite, entraîne une fusion de la capside virale avec la membrane cellulaire. Des thérapies antirétrovirales courantes (ARTs) contre le AIDS sont généralement basées sur les inhibiteurs renversés de transcriptase et les inhibiteurs de protéase. En dépit des progrès dans le développement de ces agents qui bloquent la transcription et l'assemblage d'HIV, il reste des problèmes concernant la résistance aux médicaments, les réservoirs viraux latents et les effets toxiques induits par les médicaments qui peuvent tous compromettre le contrôle effectif du virus. Par conséquent, il y a un intérêt considérable à développer de nouvelles classes de médicaments anti-HIV ayant différents modes d'action. Une

[4] http://www.who.int/hiv/en/.

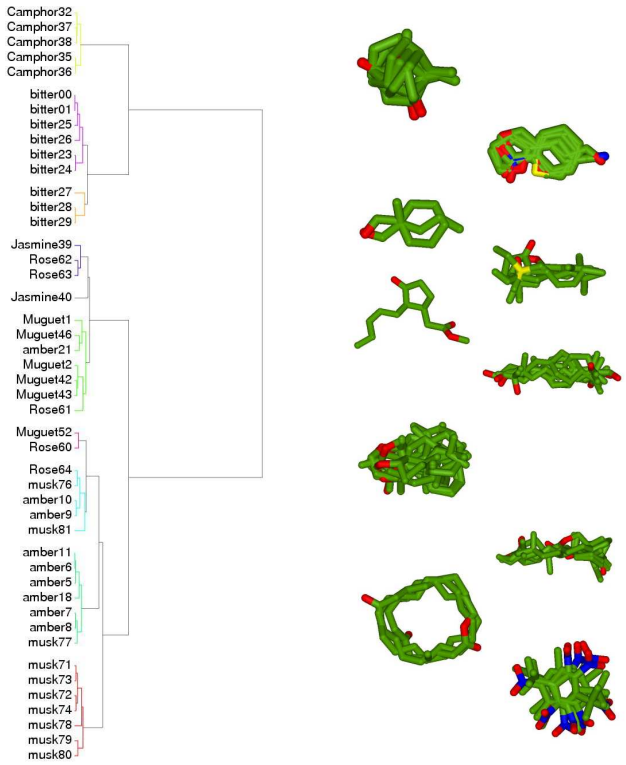

Figure 4.26: Le clustering de forme de surface canonique SH de l'ensemble de données d'*Odour*. À gauche : le dendrogramme obtenu avec dix clusters avec *L=6*; À droite : les superpositions correspondantes 3D moléculaires.

approche très prometteuse pour atteindre cet objectif a été le développement des médicaments qui bloquent l'entrée du virus HIV en interférant avec l'interaction initiale entre la protéine virale gp120 et les protéines réceptrices CXCR4 et CCR5 présentes à la surface de la cellule. Plusieurs petites molécules antagonistes de CXCR4 et CCR5 ont déjà été identifiées et la première molécule commercialisée qui bloque l'entrée du virus HIV, Maraviroc, a été récemment présentée. Toutefois, il reste un besoin continu de développer de nouvelles molécules plus puissantes qui puissent bloquer l'entrée du HIV (Carrieri *et al.*, 2009).

Dans le cadre de ma contribution pour atteindre l'objectif ci-dessus, au cours des cinq dernières années, j'ai développé des collaborations très enrichissantes et productives avec mes collègues de chimie informatique Professeur Antonio Carrieri de l'université de Bari (Italie) et Professeur Jordi Teixidó de l'Institut Químic de Sarriá de l'Universitat Ramon Llul à Barcelone (Espagne). Les deux collaborations ont impliqué des visites prolongées dans mon groupe à Aberdeen des doctorantes Alessandra Fano de Bari et Violeta Pérez-Nueno de Barcelone.

L'approche d'appariement de formes SH implémentée dans les programmes ParaFit et ParaSurf a été évaluée de manière approfondie par Violeta Pérez-Nueno dans le cadre de son projet de thèse

de doctorat sur le VS des inhibiteurs d'entrée du HIV. Afin d'effectuer un VS des ligands pour les protéines réceptrices CXCR4 et CCR5, Violeta a, tout d'abord, compilé un grand ensemble de données contenant 248 inhibiteurs de CXCR4 et 354 de CCR5 extraits de la littérature qui se composent principalement de 5 familles de composés pour les inhibiteurs de CXCR4 et 13 familles de composés pour les inhibiteurs de CCR5. Elle a également assemblé une base de données contenant 4696 composés présumés inactifs (ou leurres) ayant les propriétés physico-chimiques similaires aux composés actifs assemblés à partir de la collection de criblage de Maybridge.[5] Un VS rétrospectif des inhibiteurs de CXCR4 a été alors effectué en utilisant AMD3100 (un ligand connu de haute-affinité pour CXCR4) comme la molécule "requête" et en calculant la similarité entre elle et chacun des 248 ligands CXCR4 connus et les 4696 leurres. De même, les inhibiteurs de CCR5 ont été criblés en utilisant la molécule TAK779 de haute affinité comme requête, qui a été comparée à chacun des 354 ligands CCR5 connus et les 4696 leurres.

En chimie informatique, c'est une pratique courante d'évaluer des expériences de VS en utilisant les courbes de facteur d'enrichissement (EF) ou les courbes de ROC. Les deux approches sont largement équivalentes et consistent à tracer le rappel contre la précision comme dans des analyses conventionnelles de recherche de l'information. Cependant, les courbes de ROC deviennent de plus en plus populaires parce qu'elles fournissent une manière plus objective de comparer des expériences qui utilisent différents nombres d'exemples positifs et négatifs. Néanmoins, des courbes EF ont été initialement utilisées dans le cas actuel parce que des macros de tableurs étaient déjà disponibles pour exécuter les calculs. Pour les deux genres d'analyse, les composés sont classés dans une liste selon leur similarité avec la requête et la liste est alors analysée pour évaluer combien de composés actifs apparaissent parmi les premiers de la liste. Plus spécifiquement, dans une analyse de EF, à chaque position dans la liste, le ratio du nombre d'actifs connus (ou de hits), $\mathrm{Hits}_{\mathrm{sampled}}$, par rapport au nombre de composés échantillonnés jusqu'ici, $\mathrm{N}_{\mathrm{sampled}}$, est comparé au ratio du nombre total d'actifs, $\mathrm{Hits}_{\mathrm{total}}$, par rapport au nombre total de molécules dans la base de données, $\mathrm{N}_{\mathrm{total}}$. En d'autres mots, EF est calculée à chaque position de la liste classée en utilisant :

$$\mathrm{EF} = \frac{\mathrm{Hits}_{\mathrm{sampled}}/\mathrm{N}_{\mathrm{sampled}}}{\mathrm{Hits}_{\mathrm{total}}/\mathrm{N}_{\mathrm{total}}}. \tag{4.82}$$

La figure 4.27 montre des courbes d'enrichissement de VS obtenues en comparant la fonction d'appariement de formes SH 2D de Parafit et la fonction d'appariement de densité 3D de *Hex* avec les fonctions de score basées sur la forme uniquement (COMBO) et celles basées sur la forme en plus de paramètres chimiques (issues de l'industrie-standard ROCS). Cette figure montre que le score de Combo de ROCS donne les meilleures EF en utilisant AMD3100 et TAK779 comme requêtes. Cependant, on peut voir aussi que la forme Tanimoto 2D de ParaFit donne en général de meilleurs résultats que celle de ROCS et donne souvent des résultats comparables au score de Combo de ROCS pour les deux inhibiteurs de CXCR4 et CCR5. Les fonctions de forme Tanimoto 3D de *Hex*

[5] http://www.maybridge.com/

donnent de bons résultats pour les inhibiteurs de CXCR4 dans les premiers pourcentages de la base de données criblées, mais les EF sont considérablement inférieurs pour les inhibiteurs de CCR5. Pour les inhibiteurs de CXCR4, la forme Tanimoto de *Hex* et le score de Combo de ROCS donnent des EF comparables au maximum théorique (19.9%) dans le premier pourcentage de la base de données criblées. Pour les inhibiteurs de CCR5, le score de Combo de ROCS et celui de ParaFit donnent des EF comparables au maximum théorique (14.3%) dans le premier pourcentage de la base de données criblées. De plus, pour les inhibiteurs de CXCR4, les quatre fonctions de score d'appariement de formes fonctionnent bien dans les prochains pourcentages de la base de données criblées. Cependant, les EF pour les inhibiteurs de CCR5 ne sont pas aussi bons, en général, que les EF pour les inhibiteurs de CXCR4, bien que l'utilité relative des différentes fonctions de score soit similaire dans les deux cas. Les EF inférieurs obtenus pour le CCR5 semblent être ainsi parce que la conformation de la requête ne peut pas bien se superposer sur toutes les familles de ligand CCR5. Cette conformation se superpose bien sur les actifs ayant les mêmes familles structurales (qui ont été retrouvés en premier) mais elle ne peut pas bien se superposer sur celles qui possèdent une famille différente.

Cette étude a également comparé l'utilité du VS basé sur le ligand au VS basé sur le récepteur des deux protéines CXCR4 et CCR5 en utilisant plusieurs outils d'amarrage protéine-ligand. Il est à noter que, au moins pour ces récepteurs, les approches de criblage basées sur le ligand ont montré de meilleurs enrichissements de VS que n'importe quelle approche basée sur le récepteur (Pérez-Nueno *et al.*, 2008).

4.3.6 Grouper et classifier divers ligands CCR5

Plusieurs études précédentes sur l'amarrage ont indiqué que différents ligands CCR5 se lient de manière fondamentalement différente dans la poche extracellulaire de CCR5. En outre, en considérant qu'il est très difficile de superposer les différentes familles de composés actifs CCR5, il y a de bonne raisons pour soutenir l'hypothèse que les molécules connues liantes appartiennent à deux groupes ou plus et que les membres de chaque groupe se lient à la même région de la poche extracellulaire. Afin d'explorer cette hypothèse davantage, Violeta et moi avons développé une méthode simple pour calculer une représentation de "consensus" de forme (ou "consensus-forme") en prenant la moyenne des différentes formes de surface SH pour des groupes sélectionnés de N molécules :

$$\tilde{r}(\theta,\phi) = \frac{1}{N}\sum_{k=1}^{N}\sum_{l=0}^{L}\sum_{m=-l}^{l} a_{lm}^{k} y_{lm}(\theta,\phi). \tag{4.83}$$

Cependant, avant de calculer la moyenne, chaque molécule dans le consensus-forme doit en premier être tournée pour minimiser la distance entre elle et les *N-1* molécules restantes. Puisque ces rotations ne sont pas connues *a priori*, le consensus-forme est construit itérativement comme suit. En premier, des toute-contre-toute superpositions par paire de rotation sont calculées afin de trouver les

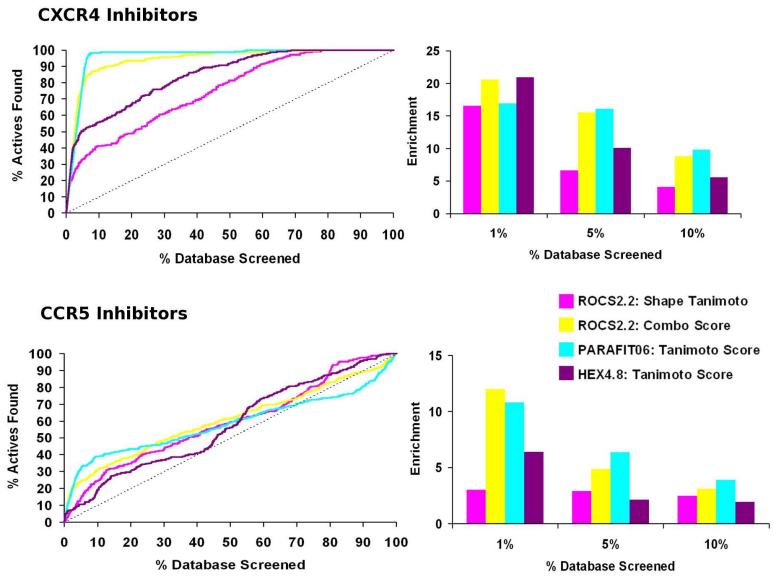

Figure 4.27: Des courbes d'enrichissement de VS pour les inhibiteurs de CXCR4 et CCR5. Les courbes d'enrichissement sur la gauche montrent une capacité relative des programmes ROCS, ParaFit et *Hex* d'identifier des inhibiteurs connus à partir d'une base de données de molécules inhibitrices et de leurre pour les protéines CXCR4 et CCR5. La ligne pointillée représente l'enrichissement qui aurait été obtenu si des inhibiteurs étaient retrouvés au hasard. Les diagrammes à barres sur la droite montre des EF pour les premiers, 1%, 5% et 10% des bases de données criblées.

deux formes de surface les plus similaires. Puis, la moyenne de ces deux formes est prise comme la forme initiale du consensus-forme ou "seed" et les *N-2* formes SH restantes sont tournées pour être superposées avec la forme seed. La moyenne globale de tous les coefficients SH est alors calculée pour donner la première estimation du consensus-forme. La forme moyenne obtenue est alors raffinée en la superposant sur les formes membres et en recalculant une nouvelle forme moyenne. Cette procédure est répétée jusqu'à convergence d'un recouvrement optimal entre chaque molécule. Et ainsi le consensus-forme est obtenu. Ceci requiert normalement seulement trois ou quatre cycles. Ainsi le calcul d'un consensus-forme est rapide.

La figure 4.28(a) montre le consensus-forme calculé à partir des trois composés les plus actifs des différentes familles classées selon leurs structures moléculaires dans la base de données des inhibiteurs de CXCR4. La figure 4.28(b) montre le consensus-forme calculé à partir de tous les inhibiteurs de CXCR4 dans la base de données. Une inspection visuelle de ces figures montre que la première consensus-forme capture plutôt bien la forme globale des trois inhibiteurs sélectionnés, tandis que le consensus-forme de "toute-molécule" a beaucoup moins de détails sur la surface locale; pourtant il maintient toujours, en gros, les caractéristiques brutes des formes membres. La figure 4.28(c) montre le consensus-forme calculé pour les trois composés les plus actifs de différentes familles dans la

base de données des inhibiteurs de CCR5. La figure 4.28(d) montre le consensus-forme de tous les inhibiteurs actifs de CCR5. Dans ce cas-ci, on peut voir qu'en utilisant tous les composés de la base de données pour construire la requête consensuelle, on obtient une forme moyenne beaucoup plus sphérique que les inhibiteurs de CXCR4, grâce à la plus grande importance en nombre et en diversité de composés CCR5 dans la base de données.

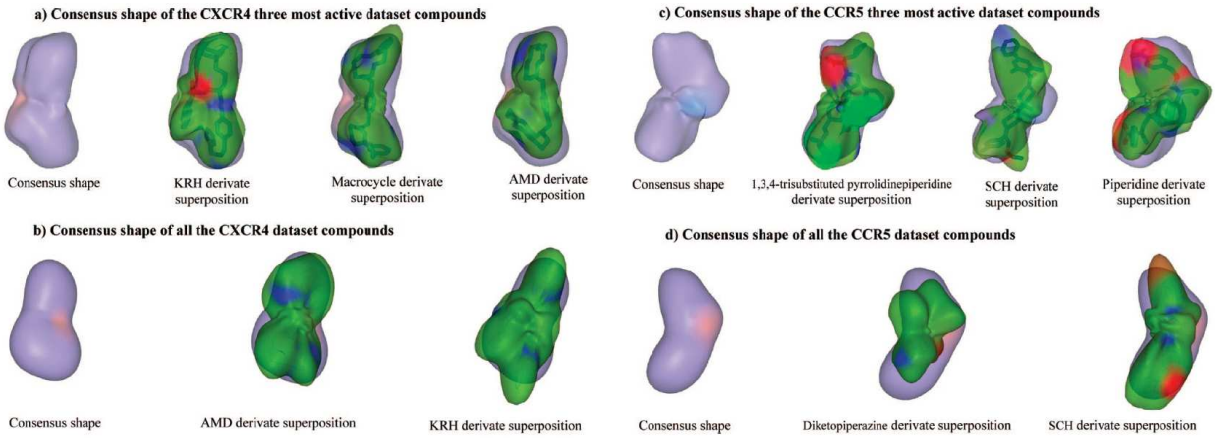

Figure 4.28: Les formes consensuelles des antagonistes de CXCR4 et CCR5. (a) L'image sur la gauche montre le consensus-forme calculé à partir des trois composés les plus actifs de différentes familles de structure moléculaire dans la base de données des inhibiteurs de CXCR4. Les trois images suivantes montrent les superpositions de ces composés sur le consensus-forme. (b) Le consensus-forme calculé à partir de tous les actifs de la base de données CXCR4 et des exemples de superposition sur le consensus-forme des deux composés sélectionnés aléatoirement. (c) Sur la gauche, le consensus-forme calculé à partir des trois composés les plus actifs de différentes familles structurales moléculaires de la base de données CCR5. Sur la droite, les superpositions de ces composés sur le consensus-forme. (d) Une forme consensuelle calculée à partir de tous les actifs CCR5 avec des exemples de superposition sur le consensus-forme de deux actifs sélectionnés aléatoirement.

La figure 4.29 montre la performance des requêtes consensuelles de VS de CXCR4 comparée au VS basé sur l'amarrage et la forme en utilisant un seul ligand de haute affinité (AMD3100). Cette figure montre que les requêtes de formes consensuelles donnent des AUC plus élevés que d'autres approches, bien que la requête de ParaFit avec un seul ligand donne également de bons résultats. Comme on pouvait s'y attendre en considérant la figure 4.28, le consensus-forme de trois ligands fonctionne considérablement mieux que le consensus-forme de tous les ligands, ce qui est dû au degré supérieur de l'aspect lisse et à la perte de détails de surface des formes de tous-ligands. D'un autre côté, considérer la très bonne performance de la requête d'un seul ligand de haute affinité et la performance marginalement supérieure de la requête de trois ligands suggère que ces trois derniers ligands partagent des formes fortement similaires (comme confirmé par la figure 4.28(a)) qui se lient tous probablement de manière similaire dans la poche de CXCR4.

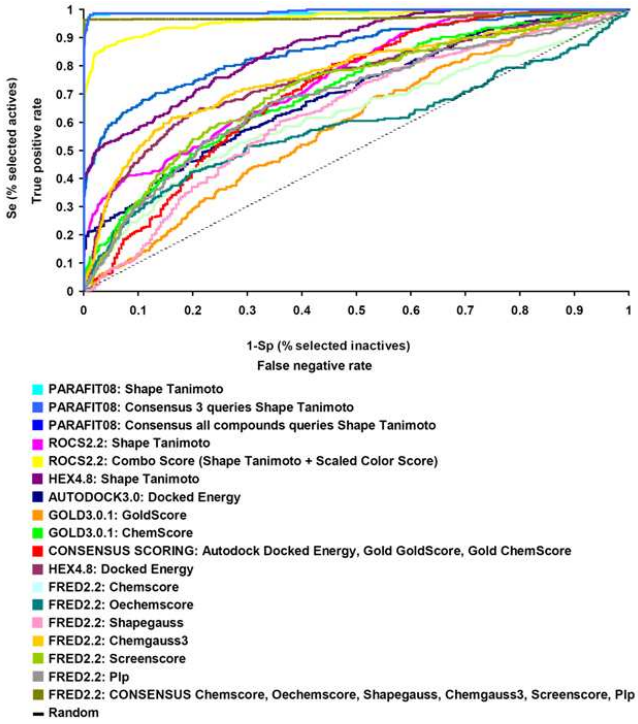

Figure 4.29: Les analyses des courbes de ROC de VS basé sur le ligand et le récepteur en utilisant AMD3100 et des requêtes de forme consensuelles des pseudo-molécules contre la base de données des inhibiteurs de CXCR4.

Concernant le problème plus difficile à comprendre des modes de liaison de divers ligands CCR5, nous avons supposé que, si tous les membres d'un groupe se lient à la même région de la poche réceptrice, ils devraient tous partager un degré significatif de similarité de forme et qu'il pourrait être possible de décrire ces similarités en construisant un consensus-forme de "pseudo-molécule". En effet, parce que la plupart des conformations de ligand avaient été calculées en utilisant le logiciel de modélisation moléculaire et que certaines de ces conformations pourraient donc être incorrectes, nous avons aussi supposé que calculer une forme consensuelle pourrait faire disparaître certaines de ces erreurs et probablement donner une meilleure requête de forme avec laquelle on pourrait effectuer du VS basé sur le ligand. *A priori*, l'appartenance de certains actifs à certains groupes, ainsi que le moyen de mieux superposer toutes les molécules dans un groupe consensuel hypothétique, restent vagues. Par conséquent, nous avons décidé de procéder itérativement en effectuant un clustering initial hiérarchique de Ward des descripteurs chimiques conventionnels (pour des détails complets, voir Pérez-Nueno *et al.* 2008) afin d'obtenir un total de dix clusters. Des formes consensuelles de surface SH ont été calculées pour chaque cluster et une *toute-contre-toute* comparaison SH de chaque surface consensuelle a été calculée en utilisant ParaFit. Les coefficients de similarité

de Tanimoto résultants par paires ont été ensuite utilisés dans un deuxième clustering hiérarchique de Ward. La figure 4.30 montre un dendrogramme des super-clusters résultants dans lesquels les dix formes consensuelles des surfaces initiales sont groupées pour donner quatre groupes principaux: A, B, C et D. Les membres de ces quatre groupes ont été re-alignés pour former quatre formes “super-consensuelles” (SC) de surface SH pseudo-moléculaire. La figure 4.31 montre les superpositions moléculaires 3D, les formes SH des dix clusters initiaux et les quatre formes SC calculées à partir du clustering des surfaces consensuelles.

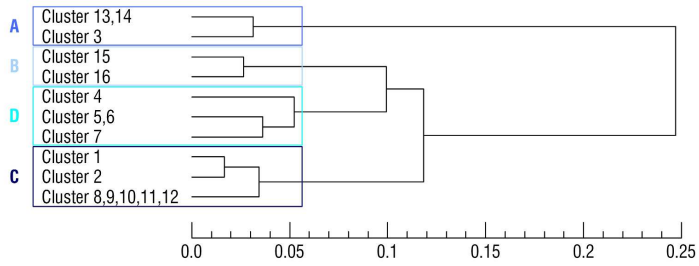

Figure 4.30: Un dendrogramme de dix groupes initiaux d’antagonistes de CCR5 groupés en utilisant l’algorithme de Ward des distances SH entre les consensus-formes de surface de chaque groupe. Quatre groupes super-consensuels principaux, marqués A, B, C et D, sont identifiés.

La figure 4.32 montre les résultats du VS obtenus en utilisant les quatre formes SC comme requêtes pour des recherches de bases de données. SC C donne la meilleure performance de VS avec un AUC de 0.91. Il n’est peut-être pas surprenant que cette requête SC fonctionne très bien parce qu’elle inclut les trois composés les plus actifs dans la base de données ainsi qu’un grand nombre d’autres actifs (c.-à-d. 184/424) avec des formes similaires aux dérivés de 4-pipéridine, de SCH, et de 1,3,4-trisubstituée pyrrolidinepiperidine. La requête SC A (87/424 actifs) fonctionne aussi plutôt bien avec un AUC de 0.79 et la requête SC D (84/424 actifs) s’exécute raisonnablement bien (AUC=0.63). Cependant, la courbe de ROC de SC B montre que cette requête a une bonne sensibilité et sélectivité dans les premiers pourcentages de la base de données criblées, mais l’AUC global est bas (0.41) parce que la base de données contient relativement peu de membres des deux familles de SC B (c.-à-d. un total seulement de 69/424). Cependant, si les membres des clusters B et D sont groupés ensemble pour former un seul SC, comme ceci pourrait être suggéré par le dendrogramme dans la figure 4.30, la performance du criblage devient essentiellement aléatoire (AUC=0.51). Ainsi, en dépit de petites populations de ces deux groupes, leurs membres ont des formes globales considérablement différentes et ils devraient être classifiés comme deux groupes structuraux distincts pour des buts de VS. Effectuer un exercice similaire avec d’autres combinaisons de clusters SC montre des réductions similaires mais moins impressionnantes dans des AUC comparés aux AUC des clusters

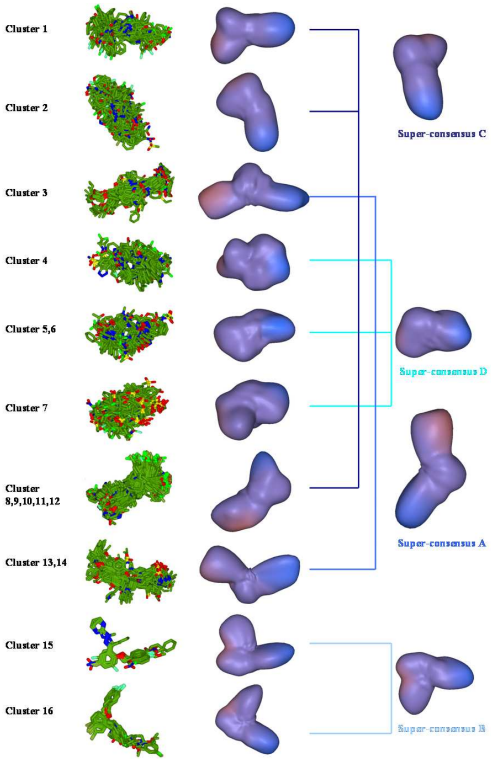

Figure 4.31: Des superpositions moléculaires et des formes consensuelles des dix clusters de Ward utilisées pour calculer les quatre formes super-consensuelles finales du CCR5.

non-fusionnés (détails non montrés). Ce comportement suggère que les familles des inhibiteurs de CCR5 peuvent être groupées dans pas moins de quatre groupes principaux.

Puisque *Hex* utilise principalement une représentation de forme-densité pour ses calculs d'amarrage protéine-protéine, il était immédiatement possible d'appliquer un amarrage SPF en corps-rigide pour trouver les quatre formes SC pseudo-moléculaires dans la poche extracellulaire de CCR5. Même si les calculs d'amarrage protéine-protéine produisent normalement de multiples orientations faux-positives, dans le cas présent fortement contraint, il y a très peu de possibilités dans lesquelles les "ligands" SC peuvent se rattacher d'une manière satisfaisante à la poche. En fait, il y a seulement trois sites de liaison possibles dans les régions des hélices de la transmembrane (TM) CCR5 qui pourraient accommoder les quatre formes SC trouvées. Celles-ci sont montrées dans la figure 4.33. Ces prédictions d'amarrage basées sur la SC sont cohérentes avec les données expérimentales existantes (voir Pérez-Nueno *et al.* 2008 et les références 11, 13, 14 et 16 qu'elle contient pour des détails complets).

Afin de confirmer que les requêtes SC sont correctement appariées avec leurs sites cibles prédits, les trois sites de liaison proposés ont été traités chacun comme s'ils étaient des cibles séparées

pour du VS basé sur l'amarrage en utilisant l'amarrage en corps-rigide des pseudo-molécules SC correspondantes. En d'autres termes, en amarrant au site 1, des composés appartenant à SC A sont traités comme actifs, et des composés appartenant à SC B, C et D sont traités comme inactifs. De même, en amarrant au site 2, des composés appartenant à SC C sont traités comme actifs, et des composés appartenant à SC A, B et D sont traités comme inactifs. De même pour le site 3, des composés appartenant à SC B et D sont traités comme actifs et des composés appartenant à SC A et C sont traités comme inactifs. La figure 4.34 montre la performance d'amarrage VS pour chacune des trois régions de liaison proposées de CCR5. En comparant les figures 4.34 et 4.32, on peut observer que l'amarrage VS sur les sites 1, 2 et 3 (AUC=0.83, 0.96 et 0.85, respectivement) améliore les AUC d'appariement de formes SC A, C et B/D (AUC=0.79, 0.91 et 0.41/0.63, respectivement). Étant donné que les SC A et C donnent déjà de bons enrichissements d'appariement de formes, re-assigner les membres de B et de D comme des inactifs améliore seulement marginalement les AUC correspondants. Cependant, le traitement du grand ensemble de membres de C et d'A comme inactifs pour le site 3 donne des AUC bien supérieurs pour les requêtes SC B et D, qui soutient clairement la notion que les antagonistes de CCR5 se lient au moins à trois sites principaux dans la poche extracellulaire.

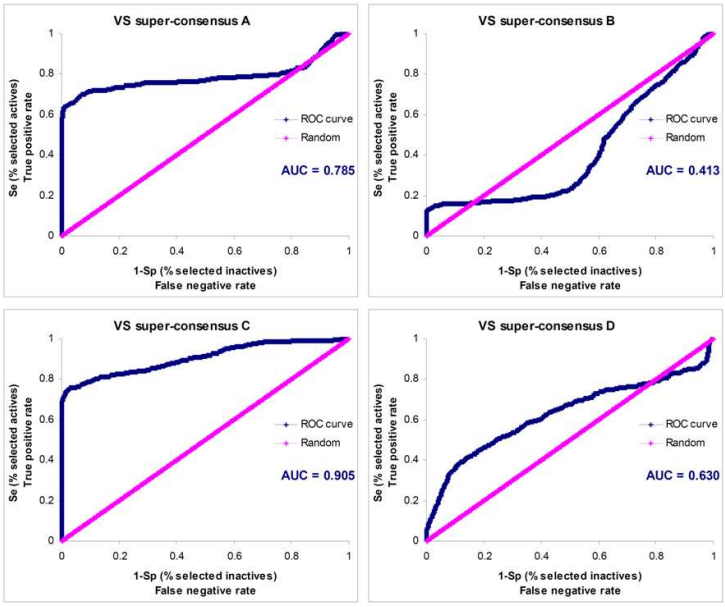

Figure 4.32: Des courbes de ROC pour valider les pseudo-molécules SC des inhibiteurs de CCR5.

Les résultats de cette étude ont montré que les formes consensuelles SH peuvent fournir des requêtes structurales 3D efficaces pour le VS basé sur la forme. Pour les ligands CXCR4 et CCR5 étudiés ici, nos résultats montrent que les requêtes de formes consensuelles bien choisies peuvent

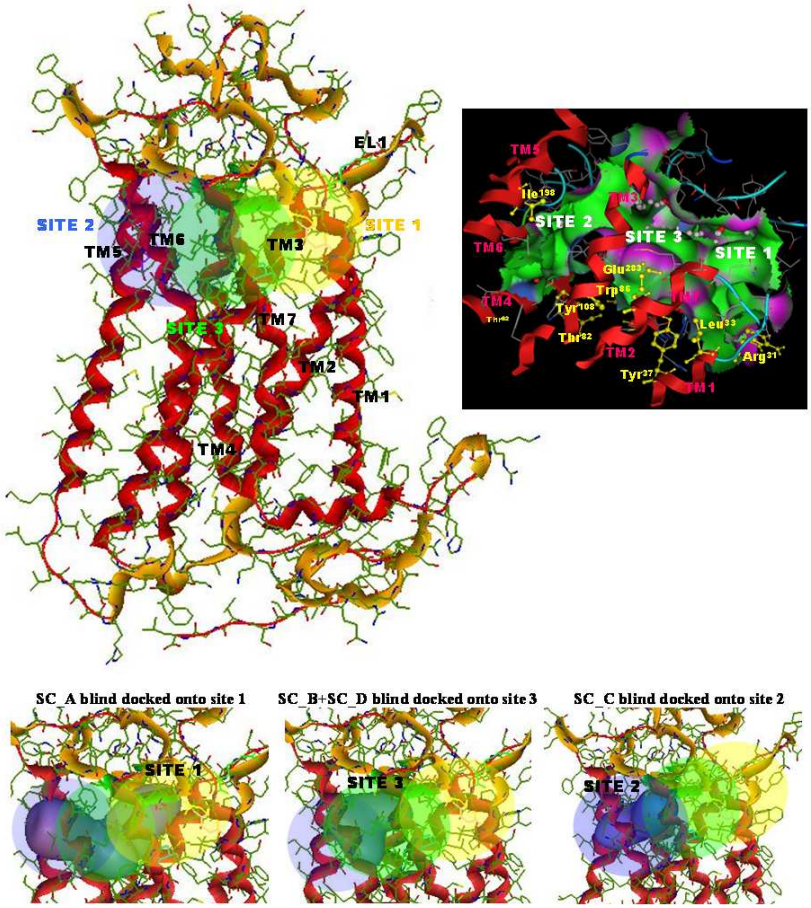

Figure 4.33: Des sous-sites de liaison de la poche CCR5 proposés par les résultats du VS consensuel et d'amarrage. La pseudo-molécule SC A est amarrée sur le site 1, la pseudo-molécule SC C est amarrée sur le site 2 et les pseudo-molécules SC B et D sont amarrées sur le site 3, qui recouvre les sous-sites de SC A et C.

donner de meilleurs (CXCR4) ou considérablement meilleurs (CCR5) enrichissements de criblage virtuel que les requêtes conventionnelles du VS d'une seule molécule. Cependant, pour CXCR4, ces résultats sont largement similaires à l'approche de base d'appariement de formes d'une molécule de ParaFit parce que les inhibiteurs de cette cible partagent plutôt des formes moléculaires similaires qui correspondent individuellement assez bien à la forme requête sélectionnée. Pour CCR5, qui a un ensemble beaucoup plus grand et plus divers de familles d'inhibiteurs, les requêtes SC de famille C et toute-famille donnent toutes deux une très bonne performance globale de VS. Cependant, ceci semble être ainsi, au moins en partie, parce qu'une proportion élevée de toutes les familles classées selon leur structure moléculaire se groupent en la famille super-consensuelle C. Par conséquent, par une construction, la forme consensuelle SH dérivée de ces membres de famille fournit une seule forme pseudo-moléculaire représentative qui identifie bien différents membres structuraux.

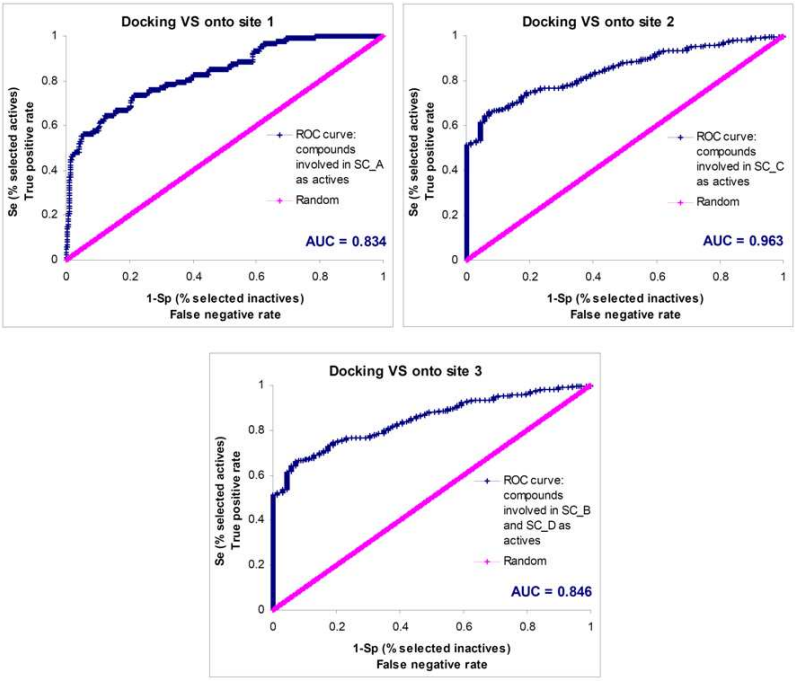

Figure 4.34: Des courbes de ROC pour valider l'amarrage VS sur les trois sous-sites identifiés sur le CCR5 pour les antagonistes de CCR5.

Concernant le challenge de comprendre comment tant de diverses familles d'inhibiteurs pourraient se lier dans la poche CCR5, notre approche de consensus basée sur la forme fournit une méthode simple d'identifier des clusters des familles d'inhibiteurs d'un grand ensemble d'actifs connus qui est largement cohérentes avec les précédents modèles et aux données expérimentales courantes. Néanmoins, la seule méthode complètement fiable pour vérifier la validité des prédictions basées sur l'amarrage est de comparer des structures cristallographiques connues. Malheureusement, de telles références d'étalon or ne sont pas disponibles pour CXCR4 et CCR5. Par conséquent, une quelconque comparaison avec des études précédentes d'amarrage peut, au mieux, servir seulement à ajouter un soutien supplémentaire à la prédiction originale. D'un autre côté, pour des raisons pratiques, une méthode impartiale et objective pour valider une prédiction informatique, même lorsqu'aucune structure cristallographique n'est disponible, est de tester son utilité dans le contexte de VS. Les résultats VS obtenus ici utilisent quatre requêtes SC basées sur la similarité et l'amarrage donnent de bien meilleurs enrichissements de VS comparés aux requêtes d'une seule-molécule. Ces résultats apportent un soutien fort aussi bien à la validité de la notion des structures SC, qu'à l'hypothèse que les membres de ces clusters SC se lient à au moins trois sites de liaison principaux dans la poche extracellulaire de CCR5.

Chapitre 5

Résumé et perspectives

5.1 Résumé

Les chapitres précédents ont présenté les formules mathématiques nécessaires au calcul des représentations SH et SPF pour les propriétés physiques et de formes des protéines et des petites molécules. On a montré que la représentation SPF présente une nouvelle méthode très efficace pour exécuter des calculs exhaustifs pour l'amarrage protéique, en utilisant des corrélations FFT de rotation d'ordre supérieur. Cette approche a été réalisée et explorée dans le programme d'amarrage *Hex*, qui est maintenant le véhicule principal et le banc d'essai pour la majeure partie de ma recherche. Les performances de *Hex* sont très honorables sur plusieurs exemples tirés de jeux de données tests pré-établis pour l'amarrage protéique. *Hex* a également donné de bons résultats pour plusieurs cibles d'amarrage protéique de CAPRI dont on ne connaissait pas la solution. *Hex* est plus rapide par au moins un ordre de grandeur que d'autres approches de corrélation conventionnelles basées sur des grilles cartésiennes. Il est, en ce moment, un des programmes les plus utilisés pour l'amarrage protéique dans les institutions académiques du monde entier. De même, on a montré qu'une représentation SH de la surface, mathématiquement plus simple, nous donne une méthode puissante pour effectuer du clustering et du VS très rapidement dans des bases de données de petites molécules. Le programme ParaFit est actuellement commercialisé par Cepos Insilico Ltd. comme un des programmes les plus rapides actuellement disponibles pour le VS.

5.2 Futurs challenges

Bien évidemment, améliorer la vitesse des calculs est toujours souhaitable, surtout pour des applications graphiques très interactives. D'autre part, en science, une grande exactitude est presque toujours requise et on devrait toujours être disposé à sacrifier un gain de vitesse contre des calculs plus sensibles et plus exacts. De plus, dans les sciences de la vie, il y a un besoin continu de

pouvoir appliquer des logiciels scientifiques à des systèmes biologiques de plus en plus grands. La croissance exponentielle récente de données structurales expérimentales génomiques et protéiques à de multiples niveaux de résolution ainsi qu'une expansion semblable dans les bases de données chimiques de petites molécules font qu'il y a maintenant des perspectives immenses pour exploiter les informations structurales protéiques et chimiques à des buts scientifiques et thérapeutiques. Mon objectif continuel est d'aider les biologistes et les chimistes à exploiter au maximum cette richesse croissante de données en continuant à développer de nouvelles techniques de calculs pour représenter et manipuler les informations complexes sur les structures moléculaires biologiques et chimiques.

Mais l'explosion récente des données signifie aussi que nous devons maintenant ajuster nos objectifs et viser des cibles et des challenges scientifiques beaucoup plus grands. Au lieu d'être satisfaits seulement avec les amarrages de paires de protéines, nous devrions maintenant viser à amarrer des centaines de paires de protéines à la fois afin de nous permettre d'assembler et de simuler de grands amarrages macromoléculaires, de l'ordre du nanomètre, comme pour le complexe du pore nucléaire transmembranaire (NPC) et les moteurs moléculaires tels que les flagelles bactériennes et les cils eukaryotes, par exemple. Nous devrions même viser à explorer le processus de reconnaissance protéine-protéine, extrêmement sensible et sélectif, par un amarrage-croisé de milliers de paires de protéines afin de simuler et vérifier le grand volume de données d'interactions produites par des méthodes haut-débit comme TAP-MS et Y2H. De même, nous devrions viser la recherche de nouvelles molécules dans de grandes bases de composés virtuels pour cibler de nouvelles molécules candidates potentielles contre de multiples structures de protéines afin de prédire, par exemple, des effets secondaires inattendus ou non désirés. Bien sûr, ce sont des objectifs très grands. Et bien entendu, je ne suis pas le seul à y avoir pensé. Mais, à mon avis, viser de grand objectifs est un bon moyen de réaliser au moins des résultats acceptables. Les sections suivantes décrivent brièvement certaines directions plus modestes dans lesquelles je continuerai mes recherches grâce à la subvention actuelle de recherche de l'ANR et quelques idées probablement plus ambitieuses que je voudrais développer à plus long terme.

5.3 Utiliser des potentiels basés sur la connaissance dans l'amarrage protéique

Je travaille de façon constante à améliorer l'état de l'art dans l'amarrage basé sur la FFT et pour rendre disponibles les nouveaux développements dans le programme d'amarrage *Hex*. La nouvelle méthode 5D polar Fourier, que j'ai récemment développée (section 4.2) et qui permet d'analyser l'espace d'amarrage en corps-rigide, est plus rapide par au moins un ordre de grandeur que les approches conventionnelles d'amarrage de corrélation basées sur des grilles cartésiennes et offre plusieurs possibilités de développements futurs. Par exemple, comme le nombre de complexes protéiques connus augmente, les potentiels basés sur la connaissance (ou "knowledge-based potentials,"

KBP) des interactions protéiques deviendront de plus en plus plus exacts et fiables (Ritchie, 2008). Cependant, parce que de tels potentiels sont en général représentés par la somme des contributions de multiples types d'atomes ou résidus (Kozakov *et al.*, 2006), leur calcul, en utilisant des techniques conventionnelles FFT 3D, s'avère coûteux. Néanmoins, en développant la représentation FFT 5D, j'ai montré que les contributions de chaque terme dans un KBP peuvent être additionnées *avant* de faire la recherche corps-rigide de FFT. En d'autres termes, le coût de calcul des potentiels multi-termes dans un FFT 5D sera marginalement supérieur au calcul de simples corrélations basées sur la forme. Je collabore avec Dima Kozakov et Sandor Vajda de l'université de Boston afin d'utiliser l'approche de FFT 5D pour calculer l'exact multi-terme correspondant au potentiel "Decoys as Reference State" (DARS) de Kozakov *et al.* (2006). Cette nouvelle approche permettra d'analyser l'énorme espace de recherche 6D beaucoup plus vite et de façon fiable en comparaison de ce qui est actuellement possible. La combinaison de nos deux approches devrait donner à nos deux groupes plusieurs années d'avance dans le domaine.

5.4 Modéliser la flexibilité protéique pendant l'amarrage

Le projet mené avec Diana Mustard pour simuler la flexibilité de la protéine pendant l'amarrage a employé une approche de dynamique fondamentale afin de produire de multiples structures propres d'une protéine pour représenter des instantanés de l'espace conformationnel possible (Mustard & Ritchie, 2005). Nous avons constaté que l'amarrage de tels instantanés conformationnels donne des résultats prometteurs mais sont coûteux en calculs. Plus récemment, Rueda *et al.* (2009) ont montré que l'utilisation des instantanés conformationnels de protéine de NMA améliore systématiquement la qualité de leurs simulations d'amarrage protéine-ligand. May et Zacharias (2008) ont publié des résultats très prometteurs sur l'amarrage protéique utilisant une approche basée sur GNM, qui modélise une protéine comme un réseau élastique de ressorts harmoniques reliant les C_α atomes du squelette. Un des avantages de GNMs est que toute information nécessaire pour l'analyse par vecteurs propres peut être dérivée directement de la matrice d'interaction de Hesse, en évitant ainsi la nécessité de produire de multiples conformations pseudo-aléaotoires. Néanmoins, on pourrait faire mieux. Par exemple, un inconvénient de l'approche actuelle de GNM (et des approches à base de vecteurs propres en général) est la nécessité de diagonaliser de larges matrices dont le coût est de l'ordre de $O(N^3)$. Par conséquent, les approches actuelles se sont jusqu'ici limitées à employer les coordonnées des atomes $C_\alpha$et ont dû exécuter une seule diagonalisation pour chaque protéine avant que l'amarrage puisse se faire.

Dans le projet Eigen-Hex, nous appliquerons une nouvelle analyse de GNM pour chaque pose putative de l'amarrage en corps-rigide. Ainsi, chaque analyse par vecteurs propres tiendra compte des fluctuations accessibles de chaque protéine dans le contexte des contraintes présentées par son partenaire d'amarrage. Bien que cette approche implique beaucoup de diagonalisations de matri-

ces, nous réduirons les coûts de calcul en utilisant l'approche "building block" de Tama *et al.* (2000) dans laquelle une protéine est subdivisée en petits blocs, comportant jusqu'à six acides aminés par bloc, chaque bloc étant consideré comme un bloc rigide par rapport à la rotation et à la translation (RTB). Tama *et al.* ont montré que les modes normaux obtenus à partir de cette approximation sont presque aussi exacts qu'en utilisant un acide aminé par bloc, et alors que le coût de calcul est considérablement réduit (Tama & Sanejouand, 2001). Si on uUtilise les vecteurs propres résultants pour produire de nouvelles poses conformationnelles putatives et minimiser rapidement les énergies de chaque nouvelle pose en utilisant des "mécaniques moléculaires souple" (Ritchie, 2003), alors chaque pose en corps-rigide sera orientée de façon flexible dans un mode d'assemblage localement optimal. Cependant, même en utilisant des vecteurs propres RTB efficaces (et peut-être aussi des techniques de diagonalisation de matrice creuse), cette approche nécessitera des calculs intensifs parce que chaque conformation prometteuse verra son score re-évalué en utilisant un champ de forces mécanique moléculaire complet. Par conséquent, les calculs seront répartis sur de multiples processeurs. J'étudie actuellement ces idées avec Vishwesh Venkatraman qui a récemment rejoint mon groupe comme post-doctorant et est financé par ma subvention de recherche ANR.

5.5 Automatiser le processus d'amarrage protéique incorporant des données expérimentales

Comme les initiatives génomiques structurales continuent à enrichir l'espace des structures 3D des protéines, l'utilisation de base de données structurales pour effectuer l'amarrage par l'homologie (qu'on pourrait décrire comme "amarrage à partir de cas") deviendra évidemment une approche de plus en plus puissante pour prédire les interactions protéiques. Bien qu'en ce moment la PDB contienne seulement une fraction très petite du nombre de complexes protéiques existants, plusieurs groupes ont créé récemment des bases de données d'interactions protéine-protéine (pour des publications de synthèse, voyez, par exemple, Mathivanan *et al.* 2006 et Ritchie 2008), et celles-ci constituent maintenant des atouts très importants pour prédire les structures de complexes protéiques. Le nombre et la couverture grandissants de ces bases de données nous offre maintenant la perspective de développer des méthodes automatiques pour le transfert des connaissances sur les interactions protéiques existantes vers des cas inconnus mais homologues. Même dans les cas où il n'y a pas de structure homologue directe avec la cible d'amarrage, il est tout de même extrêmement utile de pouvoir incorporer les connaissances sur des résidus clés pour l'interaction dans le protocole d'amarrage. En effet, même la connaissance d'un seul résidu participant à l'interaction peut aider à orienter le calcul d'amarrage et à améliorer significativement la qualité du résultat (Ritchie, 2008).

Si l'on considère les diverses approches utilisées par les différents participants de CAPRI, il est clair que plusieurs prédicteurs humains sont devenus habiles dans la recherche et l'assimilation de connaissances sur les cibles disponibles dans la littérature. Cependant, c'est une activité longue et

sujette aux erreurs et il serait souhaitable de développer des approches automatiques d'extraction de données qui puissent simuler les étapes d'acquisition de connaissances faites par les experts pour l'amarrage incorporant des données ("data-driven docking"). Je voudrais développer une méthode générique capable d'incorporer la connaissance à partir de sources de données externes dans les calculs d'amarrage sous forme de termes de contraintes de distance notées "ambiguous interaction restraints" ou AIRs (Nilges, 1995), par exemple, dans la fonction d'énergie de l'amarrage. Cependant, la manière de bien décrire ces contraintes obtenues faites par les experts humains en termes de règles simples, qui puissent être exécutées et appliquées depuis une base de données, n'est pas encore claire.

Le projet de thèse doctorale d'Anisah Ghoorah intitulé "KDD-Dock" représente une première étape vers une application formelle des techniques d'extraction de connaissances à partir de bases de données (ou "knowledge discovery in databases," KDD) pour extraire des informations à partir des bases de données existantes d'interactions protéine-protéine afin d'aider les calculs d'amarrage protéique. Ce projet, que je supervise en collaboration avec mes collègues d'Orpailleur Marie-Dominique Devignes et Malika Smaïl-Tabbone, est financé par l'ANR. Le KDD implique souvent de traiter des volumes énormes de données en utilisant une variété de techniques de fouille de données, afin d'extraire des règles ou des "unités de connaissances" utiles et ré-utilisables. Les techniques de fouille de données communes incluent la classification par des treillis, l'extraction de motifs fréquents et l'extraction de règles d'association (Napoli, 2005). Cependant, il est souvent nécessaire de rassembler en premier les données dans une seule grande table afin d'appliquer la fouille de données. Ceci peut être une tâche difficile si les données sont réparties sur plusieurs tables ou sources de données parce qu'on doit condenser toutes les données dans une seule table régulière (par exemple, en faisant une vue comportant la concaténation de plusiers tables de la base de données). D'un autre côté, les approches de fouille de données relationnelles (ou "relational data mining," RDM) permettent la fouille de données réparties dans des tables multiples, mais elles ne peuvent pas facilement être appliquées à de grands ensembles de données à cause de leur complexité algorithmique. Néanmoins, il est possible de combiner RDM avec d'autres approches conventionnelles plus efficaces (Helma *et al.*, 2004; Muggleton, 2005; Phuong & Ho, 2005). Par conséquent, l'objectif général de ce projet est d'appliquer les techniques de KDD sur les bases de données existantes d'interactions protéine-protéine pour découvrir des règles simples sur les interfaces de protéines qu'on pourrait transformer en AIR pour des calculs d'amarrage.

A plus long terme, je crois que l'amélioration de notre aptitude à identifier les informations issues de la littérature sur les PPI et l'utilisation des modèles de Markov cachés (ou "hidden Markov models," HMMs) pour détecter les sites d'interface des protéines, seront d'autres stratégies utiles pour contraindre et guider les calculs d'amarrage. C'est dans cet ordre d'idées que je voudrais collaborer avec mes collègues d'Orpailleur.

5.6 Améliorer l'alignement structural et classification des structures 3D des protéines

Bien que Lazaros Mavridis n'ait commencé le travail sur le projet "3D-Blast" qu'en mars 2009 – projet financé par ANR – les premiers résultats rapportés dans les Sections 4.1.2 et 4.1.3 sont extrêmement prometteurs. Nous avons déjà montré que l'approche SPF est suffisamment sensible pour superposer et clusteriser les structures des protéines. Les résultats sont en très bon accord avec la classification CATH, qui est l'une des sources de références les mieux acceptées pour la classification des structures des protéines. Nous avons également montré que ce serait bientôt possible de faire des requêtes structurales séquence-indépendante sur toute la base de données CATH en temps réel, en utilisant en tandem des recherches invariantes à la rotation et des recherches de corrélation rotationnelles. Bien que cela puisse être très intéressant de faire des expériences de clustering à grande échelle sur la base de données CATH entière, nous ne proposons pas que notre approche de clustering basé sur SPF remplace complètement celle de CATH. Nous considérons plutôt notre contribution comme une méthode indépendante et objective complémentaire à la classification experte mais lente actuellement utilisée dans CATH. Nous prévoyons que cela pourrait aider à résoudre des cas difficiles ou ambigus et donnerait une façon d'identifier les similarités inattendues ou nouvelles qui pourraient exister au delà de la zone "twilight" de l'alignement conventionnel de séquences. Développer notre approche pour permettre de chercher et d'apparier des sous-structures pourrait donner des outils supplémentaires pour effectuer des analyses plus détaillées sur la relation structure-fonction. Par exemple, dans le cadre de son projet de Master au LORIA, Emmanuel Bresso étudie actuellement comment cette approche pourrait être utilisée pour réaliser une étude à grande échelle des déterminants structuraux des sites de phosphorylation des surfaces protéiques.

5.7 Explorer le retour haptique pour orienter l'amarrage protéique

Actuellement, un des grands challenges dans l'amarrage protéique est de distinguer les modes d'assemblage presque natifs d'une liste de faux-positifs très plausibles (Ritchie, 2008). Bien que des progrès continuent d'être faits pour développer des approches d'amarrage automatiques, il y a eu remarquablement peu de recherche sur la façon dont le savoir-faire et les compétences humaines pourraient être exploités pour aider à améliorer les résultats d'amarrage. L'analyse des expériences de CAPRI suggère que la compétence humaine peut apporter une contribution très importante à la qualité d'un ensemble de prédictions d'amarrage. Beaucoup de participants de CAPRI et sans aucun doute la plupart des chercheurs de laboratoire, consacrent un temps considérable à visionner et analyser les solutions candidates d'amarrage en utilisant des outils graphiques moléculaires 3D. Généralement, la faisabilité de chaque prédiction d'amarrage est évaluée visuellement pour considérer la complémentarité de l'assemblage entre les surfaces interactives calculées. Cependant, au

delà du développement de meilleures fonctions de score à base d'énergie, l'automatisation de cette activité n'est pas résolue. Néanmoins, Gillet *et al.* (2005) ont démontré que la possibilité de sentir et manipuler des modèles moléculaires tangibles en plastique donne un niveau d'intuition exceptionnel et des informations sur les formes et les propriétés des macromolécules. Malheureusement, la création des modèles des molécules en 3D ("l'impression 3D") reste un processus lent et coûteux. D'autre part, un travail récent sur la réalité virtuelle pour modéliser des interactions protéine-ligand a montré que les forces interactives entre les protéines et les petits ligands peuvent être simulées et expérimentées physiquement en utilisant des techniques de retour haptique (Nagata *et al.*, 2002; Wollacott & Mertz, 2005). Par conséquent, il serait très utile d'explorer une telle approche dans le contexte de l'amarrage protéique.

De façon exclusive, le programme d'amarrage *Hex* intègre déjà la visualisation moléculaire stéréographique avec laquelle on peut visualiser les poses d'amarrage calculées. Mais, comme la plupart des programmes graphiques moléculaires conventionnels, il manque le retour haptique. Néanmoins, la nature interactive du programme existant montre que ce sera relativement simple d'ajouter cette possibilité et d'explorer son utilité. L'idée fondamentale est que l'utilisateur visualiserait et manoeuvrerait devant ses yeux les protéines comme s'il tenait une protéine dans chaque main. Les mouvements des poignets (les mouvements de tangage et roulis) seraient transformés en angles de rotation d'Euler avec lesquels on ferait tourner les protéines. Les actions de rapprocher les mains ou de les séparer commanderaient les positions cartésiennes des molécules. Cela fournirait une façon beaucoup plus naturelle de manoeuvrer et d'inspecter les poses d'amarrage que le contrôle 2D conventionnel offert par la souris, qui est actuellement implémenté dans *Hex*, par exemple. Du retour visuel additionnel serait donné par le codage-couleur des surfaces protéiques selon les énergies d'amarrage et les désaccords stériques, par exemple. Si des résidus participant à l'interaction ou faisant l'objet de contraintes ont été identifiés grâce aux données expérimentales, ceux-ci seront accentués graphiquement pendant l'amarrage. L'exploration de cette façon de l'amarrage protéique interactif irait de pair avec les objectifs mentionnés ci-dessus concernant l'amarrage incorporant des données et le développement des fonctions de score et d'énergie rapides. Par exemple, des techniques de gradient peuvent être utilisées afin de calculer les forces nécessaires pour le retour haptique inertiel et ce serait très intéressant d'explorer des méthodes pour visualiser interactivement les entonnoirs d'énergies autour d'une pose d'amarrage putative. Il est à noter que le LORIA dispose d'excellentes facilités dans le domaine de recherche de la réalité virtuelle et pluisiers equipes travaillent dans des secteurs relatifs. Par conséquent, d'autres synergies et innovations pourraient émerger à travers ce genre de projet.

5.8 Développer des consensus-formes SH pour le criblage virtuel

Comprendre comment les protéines interagissent est d'une importance cruciale pour comprendre les mécanismes moléculaires d'une maladie. Par exemple, les molécules thérapeutiques fonctionnent souvent en modulant ou en bloquant des PPI et donc les PPI représentent une classe importante de cibles (González-Ruiz & Gohlke, 2006). Par conséquent, chercher dans des bases de données chimiques pour trouver des petites molécules ou ligands qui pourraient se lier à des protéines spécifiques et modéliser ces interactions deviendront des stratégies de plus en plus importantes pour la découverte structurale de médicaments (Richards, 2002; Congreve *et al.*, 2005). Bien qu'il soit clairement souhaitable de connaître la structure tri-dimensionnelle de la protéine cible, ce n'est pas souvent possible, en particulier pour les grandes protéines transmembranaires qui sont très difficiles à cristalliser par exemple.

Néanmoins, pour le VS basé sur le ligand, la structure de la protéine cible n'est généralement pas connue, et l'approche générale est d'utiliser la connaissance sur les ligands connus ayant une affinité avec la cible afin de trouver ou concevoir d'autres molécules semblables qui pourraient se lier à la cible d'une façon semblable. En effet, plusieurs études menées pour comparer les approches de VS ont montré que les méthodes basées sur l'appariement selon la forme du ligand font au moins aussi bien ou même mieux que l'amarrage des ligands sur la cible (Hawkins *et al.*, 2007). Cependant, en dépit du succès relatif du VS basé sur le ligand, il reste des problèmes difficiles relatifs aux choix des composés initiaux pour faire la requête et des conformations de ces composés (Hawkins *et al.*, 2007). De plus, les sociétés pharmaceutiques ont déjà souvent développé de multiples ligands pour une certaine cible. Cependant, les approches traditionnelles procédant par appariement de formes utilisent normalement une seule conformation d'un composé comme requête. Mais on ne connaît pas *a priori* si c'est la bonne requête à utiliser pour cribler une base de données entière. Par exemple, d'autres familles de composés pourraient également être actives pour la même cible mais elles pourraient être retrouvées dans la base de données seulement si une conformation-requête différente est utilisée. En d'autres termes, le VS conventionnel suppose qu'il y a un seul mode de liaison pour une certaine protéine cible. Ceci peut être correct pour quelques cibles, mais certainement pas dans tous les cas. Plusieurs études récentes ont montré que quelques protéines cibles se lient avec différents ligands de différentes manières, par exemple, CCR5 (Kellenberger *et al.*, 2007), CXCR4 (Wong *et al.*, 2008), CDK2 (Wong *et al.*, 2006), HIVRT (Lewis *et al.*, 2003), FXA (Taha *et al.*, 2005), et LXR (Williams *et al.*, 2003). Par conséquent, il y a un besoin de développer de nouvelles approches de criblage qui puissent détecter de tels cas et qui puissent associer des sous-ensembles de ligands spécifiques au site de liaison du récepteur correspondant.

Comme il a été résumé dans la Section 4.3.5, le travail avec Violeta Pérez-Nueno sur le VS SH a démontré que la représentation SH fournit une méthode efficace et rapide pour identifier des ligands qui sont globalement semblables à une certaine molécule. L'approche SH implémentée dans Para-

Surf et ParaFit donne des résultats comparables avec ceux du programme de superposition ”ROCS Gaussian” qui est le standard industriel (Grant *et al.*, 1996). Nous avons également montré qu'en utilisant une représentation SH, il est simple de construire la forme moyenne, ou “consensus-forme” d'un groupe de molécules en calculant la moyenne de leurs vecteurs de coefficients et que le clustering de telles formes consensus indique que les ligands CCR5 peuvent être classifiés dans quatre familles principales, qui semblent se lier dans trois sous-sites principaux dans la poche extracellulaire de CCR5. Puisque nous avons eu de très bons résultats pour les systèmes CCR5 et CXCR4, nous souhaitons développer l'approche clustering de consensus et l'appliquer à d'autres protéines cibles. Par conséquent, en août 2009, Violeta et moi avons déposé une demande de bourse Marie-Curie Intra-European, qui lui permettrait de développer ces idées au LORIA.

5.9 Implémenter le VS protéine-ligand basé sur la représentation FG

Les représentations SH ont aussi bien été utilisées pour représenter les formes de petites molécules que pour caractériser avec succès les formes des surfaces des poches liées de la protéine (Cai *et al.*, 2002; Morris *et al.*, 2005) et pour exécuter du VS à haut-débit basé sur le récepteur (Yamagishi *et al.*, 2006). Cependant, il n'est pas encore clair si les représentations SH des surfaces permettent exactement d'apparier les formes de ligands avec des poches protéiques car, mathématiquement, la représentation de l'enveloppe de la surface ne permet pas de résoudre la partie translationnelle du problème d'appariement. À mon avis, on doit utiliser des représentations SPF 3D plus sophistiquées, semblables à celles utilisées dans *Hex*, afin de donner un niveau de sensibilité plus élevé dans un filtre HTVS tout en évitant le recours à des facilités HPC trop coûteuses. En outre, parce qu'il a été suggéré que l'appariement 3D basé sur la forme du ligand peut être, en fait, supérieur à celui d'amarrage dans un contexte de VS (Hawkins *et al.*, 2007), il y a une forte motivation et l'espoir de retombées importantes dans le développement des approches SPF pour les filtres HTVS de ligands.

Bien que les fonctions GL radiales de base décrites ici fonctionnent très bien pour l'amarrage protéique, je crois que la méthode calculatoire la plus utile pour représenter et comparer les formes 3D de petites molécules est d'employer les expansions SPF dans lesquelles les fonctions GL radiales de base existantes sont remplacées par les polynômes Gegenbauer. Il y a deux raisons principales pour lesquelles je crois que les polynômes Gegenbauer donneront un ensemble de base plus approprié avec lequel on puisse développer de nouvelles techniques HTVS 3D récepteur *et* ligand. Premièrement, ces polynômes n'ont pas un facteur de désintégration exponentiel comme les polynômes GL, mais leur domaine normal est l'unité hypersphère. Cela signifie qu'avec le choix approprié d'un facteur d'échelle, les polynômes Gegenbauer permettront d'encoder un certain niveau de détails en utilisant des expansions de l'ordre inférieur de manière plus compacte qu'avec les fonctions GL. Deuxièmement, parce qu'un théorème d'addition existe pour les polynômes Gegenbauer (Srinivasan *et al.*, 2005), il devrait être possible de calculer les effets de translation et ceux de rotation en transformant

seulement les coefficients d'expansion originaux, comme cela est le cas actuellement pour les fonctions GL de base existantes. À ma connaissance, il n'existe pas un théorème de translation pour les polynômes Zernike. Par conséquent, en dépit des résultats prometteurs récents avec la représentation de forme invariable à la rotation de Zernike (Mak *et al.*, 2008; Sael *et al.*, 2008; Venkatraman *et al.*, 2009), la base Zernike semblerait être moins attrayante que les bases GL ou Gegenbauer. Il est important de noter que les descripteurs prétendus invariables à la rotation peuvent être obtenus trivialement à partir de toutes ces expansions de base orthonormées simplement en additionnant les valeurs des coefficients d'expansion (Mavridis & Ritchie, 2009).

De toute façon, les considérations ci-dessus sont importantes parce que le coût de calcul des coefficients d'expansion SPF est de l'ordre de $O(N^3)$, et le coût de rotation et translation de ces représentations est pratiquement de l'ordre de $O(N^4)$ et $O(N^5)$, respectivement, où N est l'ordre du polynôme de l'expansion. Par conséquent, en choisissant judicieusement le type d'expansion radial et le facteur d'échelle, on peut réduire l'ordre d'expansion et ainsi accélérer de manière significative les calculs de l'appariement de formes et d'amarrage sans sacrifier la résolution.

On envisage que le nouveau programme "3D-Snap" pourra être utilisé en mode ligand ou récepteur, ou les deux, pour le criblage. En d'autres termes, si l'entrée est une petite molécule, le programme cherchera dans sa base de données des molécules semblables en utilisant des corrélations rotationnelles de superposition 3D. Si l'entrée est une protéine cible, le programme cherchera des molécules qui peuvent se lier dans un site de liaison spécifié de la protéine cible. Dans ce cas, le site de liaison sera indiqué en donnant les coordonnées et le rayon d'un atome factice sur la protéine, par exemple. Si l'entrée est un complexe protéine-ligand existant, le programme cherchera dans sa base de données des ligands semblables et il les liera alors dans un site indiqué par la position du ligand. Il sera possible de chercher dans la base de données en utilisant des modes de filtrage différents. Par exemple, pour une certaine requête de forme, des comparaisons ultra-rapides invariables à la rotation seront utilisées pour éliminer presque instantanément une grande proportion de la base de données. Un appariement plus sensible sera alors effectué en utilisant des corrélations rotationnelles explicites FFT. Quand cela sera approprié, des corrélations d'amarrage récepteur-ligand seront appliquées seulement aux ligands candidats qui auront passé les filtres initiaux de forme. Par conséquent, il sera nécessaire d'exécuter des calculs d'amarrage complets seulement sur un sous-ensemble relativement petit de la base de données.

Avec un certain site de liaison d'un ligand, la version courante de *Hex* peut effectuer l'amarrage en corps-rigide en environ une minute parce que le calcul se réduit en grande partie à une corrélation rotationnelle dans laquelle le ligand tourne à l'intérieur du site de liaison et seule une recherche de translation limitée est nécessaire pour accomplir la manoeuvre d'amarrage. Cependant, *Hex* n'est pas conçu pour accéder à une base de données ou pour travailler avec un processus de filtrage, et il serait inapproprié de surcharger sa structure. Par conséquent, on propose le nouveau programme 3D-Snap. Néanmoins, 3D-Snap sera largement implémenté en adaptant et en réutilisant une grosse partie du

code source de *Hex*. En utilisant de nouvelles représentations de forme FG du ligand et seulement une zone relativement petite de la protéine centrée sur le site de liaison, il sera possible d'exécuter des corrélations d'appariement ou d'amarrage très rapides mais exactes. L'approche globale est illustrée schématiquement dans la figure 5.1. D'après les expériences précédentes, on s'attend à ce que chaque corrélation d'amarrage ne requière que quelques secondes de processeur.

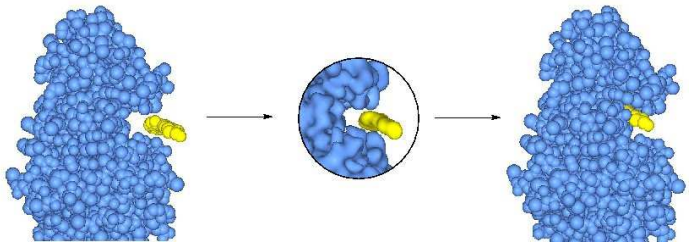

Figure 5.1: Une illustration schématique du calcul de corrélation d'amarrage protéine-ligand FG proposé. À gauche : un domaine du récepteur de fibroblast growth factor (FGFRK) avec un ligand ATP analogue près du site de liaison du ligand (code PDB 1AGW). Centre : Représentations FG du ligand et de la zone locale autour du site de liaison du ligand FGFRK. À droite : le complexe FGFRK-ligand calculé. Puisque le site de liaison du ligand est normalement connu *a priori*, il n'est pas nécessaire d'analyser la surface entière de la protéine pendant l'amarrage. Autrement, une corrélation SPF rapide peut être faite en tournant et en translatant le ligand à l'intérieur d'une petite zone sphérique locale au site de liaison (centre). Cette illustration est faite en utilisant des fonctions GL au lieu des fonctions hypersphériques de base FG proposées.

Bien que j'aie beaucoup d'expérience en calcul et en théorie des fonctions spéciales, un travail non trivial est à prévoir pour développer les expressions de translation nécessaires pour exploiter au maximum l'approche FG proposée. Et il est difficile de prédire combien de temps cela pourrait prendre. Je n'ai pas prévu qu'un post-doctorant puisse entreprendre cette partie du projet. Néanmoins, il convient de noter que ni 3D-Blast ni 3D-Snap ne dépendent absolument des fonctions de base FG pour leur succès. Ainsi, le développement et le test du programme de ces deux projets peuvent s'accomplir en utilisant les fonctions de base GL existantes dans *Hex*.

Cette approche sera développée et évaluée en utilisant des données HTVS sur les protéines cibles et les ligands, comme celles fournis par les données publiques de DUD ("directory of useful decoys") et de ZINC (Irwin & Shoichet, 2005; Huang & Shoichet, 2006). Bien que les calculs proposés pour l'amarrage protéine-ligand ne soient pas aussi exacts qu'AutoDock, cette approche fournirait un filtre additionnel rapide qui pourrait précéder l'étape conventionnelle de l'amarrage protéine-ligand basée sur le champ de forces. Je crois qu'en construisant un processus de filtrage et d'amarrage utilisant des représentations FG 3D et des techniques de corrélation SPF 3D et 5D rapides, il sera possible d'analyser des millions de ligands avec une exactitude comparable à celle d'AutoDock en environ 24 heures sur du matériel GPU moderne très abordable (détaillé dans la section suivante). Si l'on atteint avec succès ce niveau de performance, cela révolutionnerait la découverte structurale de médicaments.

5.10 Exploiter des processeurs graphiques de pointe

Les progrès récents dans la technologie du processeur graphique (“graphics processing unit,” GPU) ont donné lieu à une puissance énorme de calculs sur un ordinateur personnel à un prix abordable. Des développements dans la technologie GPU ont été initialement incités par les demandes des industries de jeu, mais beaucoup de calculs scientifiques ont été, depuis, portés sur des GPU (Owens *et al.*, 2007). En effet, il est maintenant possible d’exécuter sur un ordinateur personnel calculs la plupart des scientifiques rêvaient il y a seulement quelques années. Par exemple, le GPU de pointe actuel nVidia contient 240 unités de traitement arithmétique qui peuvent fournir ensemble presque 1000 milliards d’opérations en virgule flottante par seconde (c.-à-d. 1 Teraflop). Ceci correspond à plus de 100 fois la puissance de calcul d’un ordinateur personnel conventionnel. Jusqu’à récemment, il fallait avoir une compétence considérable et une connaissance poussée du matériel pour écrire des programmes sur des GPU. Cependant, avec l’arrivée du modèle de hardware SIMT (“simultaneous instructions on multiple threads”) et avec de nouveaux outils de développement de logiciels tels que Brook (Buck *et al.*, 2004) et CUDA[1], il est maintenant beaucoup plus facile de déployer des programmes scientifiques sur des GPU. Par exemple, pour des simulations de la dynamique moléculaire (MD), Buck *et al.* (2004) ont atteint presque 10 fois la vitesse pour le programme Gromacs en utilisant Brook; et Stone *et al.* (2007) ont rapporté une amélioration de la vitesse allant jusqu’à 36 fois pour NAMD en utilisant CUDA. Bien qu’il reste une différence significative entre les modèles de programmation Brook et CUDA, il s’avérerait que les prochaines spécifications OpenCL[2] deviendront bientôt le modèle de programmation standard pour le traitement parallèle multi-plateforme. Naturellement, nous suivrons tout nouveau développement tel qu’OpenCL.

J’ai récemment implémenté les corrélations d’amarrage 1D et 3D de *Hex* en CUDA et ceci donne une vitesse d’au moins 45 fois meilleure si l’on compare aux mêmes calculs sur un seul processeur conventionnel (Ritchie & Venkatraman, 2010; Macindoe *et al.*, 2010). En outre, dans le cadre du projet de HPASSB, nous avons amélioré *Hex* pour être en mesure d’utiliser jusqu’à huit processeurs et deux processeurs graphiques simultanément sur un seul ordinateur personnel. Ceci maintenant permet d’exécuter des calculs exhaustifs sur l’amarrage de forme en quelques secondes, permettant ainsi à des biologistes d’effectuer un vrai amarrage “orienté” interactif sur leur ordinateur personnel. Il permettra d’incorporer dans des calculs des modèles de flexibilité de protéines beaucoup plus sophistiqués en échangeant une plus grande vitesse pour plus d’exactitude. Cependant, en dépit de l’utilisation de techniques efficaces, les projets “Eigen-Hex” et “3D-Snap” seront coûteux en calculs. Effectuer de l’amarrage flexible dans “Eigen-Hex” impliquera un amarrage-croisé et une minimisation de l’énergie pour les multiples conformations de protéines. De même, effectuer du criblage virtuel sur de grandes bases de composés impliquera de comparer et de filtrer des millions de composés. Par conséquent,

[1] http://www.nvidia.com/.

[2] http://www.khronos.org/opencl/.

développer des algorithmes efficaces en utilisant des modèles de programmation parallèle modernes pour exploiter le matériel de pointe, constituera un aspect important de développements futurs pour la biologie structurale.

5.11 Modéliser les assemblages macromoléculaires

Développer des algorithmes de haute-performance sera aussi très important dans des projets à grande échelle sur la modélisation macromoléculaire. La figure 5.2 montre les structures cristallographiques du complexe de sept composants Arp2/3 (Robinson *et al.*, 2001) qui est responsable de l'initiation de la polymérisation actine des cellules eukaryotes et du complexe de dix composants de RNA polymérase II de levure (Gnatt *et al.*, 2001) qui est impliqué dans la transcription du DNA au RNA messager. Curieusement, Inbar *et al.* (2005) ont démontré qu'il est possible de construire ces structures de multiples composants à partir des protéines composantes *non-liées* grâce à une technique d'assemblage combinatoire utilisant seulement des résultats d'amarrage deux à deux. Il semble que l'approche combinatoire d'assemblage fonctionne très bien parce que chaque paire d'arrangements putative donne des contraintes stériques substantielles sur la façon dont des structures suivantes pourraient être ajoutées au complexe émergeant. En d'autres termes, il y a suffisamment de contraintes stériques mutuelles pour permettre à une seule solution presque-native d'être identifiée. Ce travail révolutionnaire a démontré la faisabilité d'assembler des structures très grandes à partir de composants multiples. Cependant, des recherches supplémentaires sont nécessaires pour développer des techniques de recherche plus pratiques qui ne demandent pas une recherche combinatoire exhaustive.

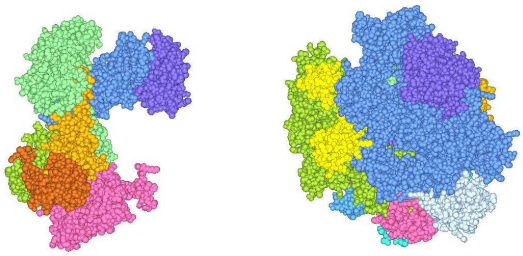

Figure 5.2: Deux exemples d'assemblages crystallographiques multimoléculaires. La structure sur la gauche est le complexe de sept composants Arp2/3 (code PDB 1K8K) qui initie la polymérisation d'actine dans les cellules eukaryotes. La structure sur la droite est le complexe de dix composants de RNA polymérase II de la levure (code PDB 1I6H) qui est responsable de la transcription du DNA au RNA messager.

Bien que la cristallographie de rayons X soit sans doute la technique standard pour déterminer une structure, la résolution disponible dans la microscopie cryo-électron (cryo-EM) s'est améliorée considérablement au cours des années et commence à s'approcher de celle de la cristallographie (Stowell

et al., 1998). Un avantage de cryo-EM est qu'il peut fournir des structures à basse résolution pour des assemblages macromoléculaires très grands qui sont peut être difficiles ou impossibles à résoudre en utilisant des techniques cristallographiques conventionnelles (Frank, 2002a). En d'autres termes, cryo-EM peut fournir les formes globales d'assemblages très grands mais il est toujours nécessaire de compléter le niveau atomique de telles structures. Par exemple, la structure du grand moteur ATPase vacuole transmembranaire (voir la figure 5.3), a été récemment résolue par cette méthode (Muench *et al.*, 2009). Des techniques de corrélation FFT sont utilisées de plus en plus pour adapter les structures de protéines de rayon X de haute résolution à des cartes de densité cryo-EM de basse résolution (Wriggers *et al.*, 1999; Roseman, 2000; Kovacs *et al.*, 2003). De telles techniques d'ajustage ou "d'amarrage intérieur" sont susceptibles de rester des approches très puissantes pour déterminer les structures de grands complexes avec une résolution atomique qui sont peu susceptibles d'être résolues en utilisant des techniques cristallographiques (Rossmann, 2000; Rossmann *et al.*, 2007).

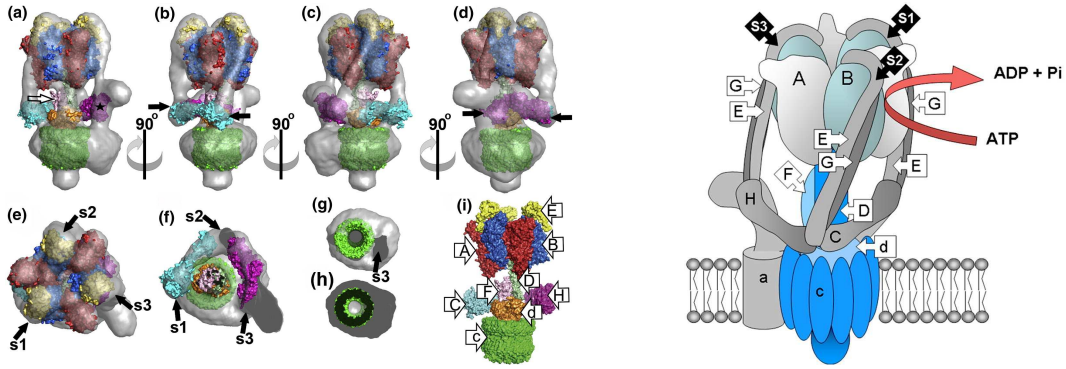

Figure 5.3: À gauche : une illustration de la structure moléculaire de moteur ATPase montrant les domaines composants de la protéine arrimés dans une carte de densité cryo-EM de basse-résolution (en gris). À droite : un diagramme schématique montrant le modèle d'arrangement des stators (en gris) et rotors (en bleu). Ces images sont reprises de Muench *et al.* (2009).

Sur une plus grande échelle, Alber *et al.* ont récemment montré que l'architecture globale du très grand complexe pore nucléaire (NPC) comportant un total de 456 protéines, peut être modélisée en combinant diverses données de multi-résolution sur les structures des protéines composantes et leurs interactions (Alber *et al.*, 2007a; Alber *et al.*, 2007b). La figure 5.4 montre les positions prédites des sous-structures principales dans ce modèle remarquable. Ce modèle est une bonne illustration du besoin futur de pouvoir utiliser des approches hybrides qui puissent intégrer des sources de diverses données expérimentales et des connaissances sur les PPI pour combler l'écart entre la résolution des différentes techniques de collecte de données structurales expérimentales (Elad *et al.*, 2009; Lindert *et al.*, 2009).

Le moteur d'ATPase et les NPC sont simplement deux exemples de machines moléculaires. L'amélioration de notre compétence à modéliser la biologie structurale de telles machines molécu-

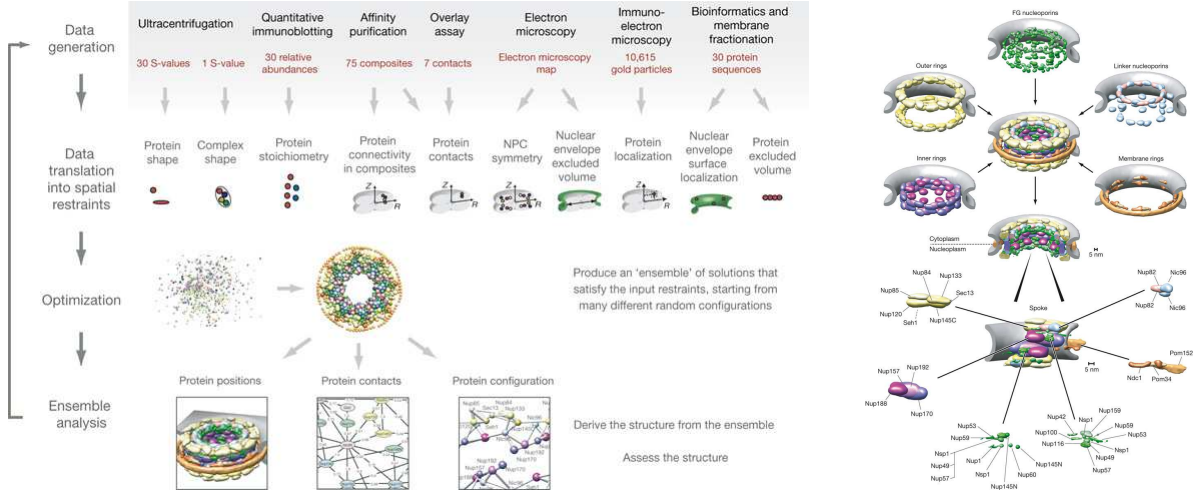

Figure 5.4: À gauche : une présentation schématique du processus de modélisation de NPC d'Alber *et al.* nécessitant l'intégration des diverses données expérimentales et des contraintes spatiales. À droite : la localisation des sous-structures principales dans le modèle final. Ces images sont reprises d'Alber *et al.* (2009a et 2009b).

laires apporterait une énorme contribution à une compréhension des processus dynamiques cellulaires (Stein *et al.*, 2007). Aider à la compréhension des mécanismes fondamentaux de la vie et permettre de modéliser la biologie structurale des virus tels que le HIV (Ganser-Pornillos *et al.*, 2008) ainsi que la façon dont ils s'assemblent, pourrait mener au développement de nouvelles molécules thérapeutiques inhibitrices d'assemblages (Sticht *et al.*, 2005). Je suis convaincu que les algorithmes et les techniques SPF présentés ici auront un rôle utile à jouer dans de nombreux défis futurs fascinants de la biologie structurale.

Bibliographie

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P., & Sali, A. (2007a). Determining the architectures of macromolecular assemblies. *Nature,* **450**, 683–694.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A., & Rout, M. P. (2007b). The molecular architecture of the nuclear pore complex. *Nature,* **450**, 695–701.

Aloy, P., Pichaud, M., & Russell, R. B. (2004). Protein complexes: structure predicition challenges for the 21st century. *Curr. Op. Struct. Biol.* **15**, 15–22.

Aloy, P. & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nat. Biotech.* **22**, 1317–1321.

Aloy, P. & Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nature Rev. Mol. Cell. Biol.* **7**, 188–197.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Amadei, A., Linssen, A. B. M., & Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Struct. Func. Genet.* **17**, 412–425.

Arkin, M. R. & Wells, J. A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discov.* **3**, 301–317.

Beautrait, A., Leroux, V., Chavent, M., Ghemtio, L., Devignes, M. D., Smaïl-Tabbone, M., Cai, W., Shao, X., Moreau, G., Bladon, P., Yao, J., & Maigret, B. (2008). Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment. *J. Mol. Model.* **14**, 135–148.

Biedenharn, L. C. & Louck, J. C. (1981). *Angular Momentum in Quantum Physics*. Reading, MA: Addison-Wesley.

Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., & Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Curr. Op. Struct. Biol.* **14**, 292–299.

Bourne, P. E. & Weissig, H. (2003). *Structural Bioinformatics*. New York: Wiley.

Boys, S. F. (1950). Electronic wave functions I. A general method of calculation for the stationary states of any molecular system. *Proc. Roy. Soc.* **A200**, 542–554.

Bransden, B. H. & Joachain, C. J. (1997). *Introduction to Quantum Mechanics*. Harlow, UK: Addison Wesley Longman.

Brown, F. K. (1998). Chemoinformatics: What is it and how does it impact drug discovery? *Annual Reports in Med. Chem.* **33**, 375.

Buck, I., Foley, T., Horn, D., Sugerman, J., Fatahalian, K., & Hanrahan, M. H. P. (2004). Brook for GPUs: stream computing for graphics hardware. *ACM Trans. Graph.* **23**, 777–786.

Cai, W., Shao, X., & Maigret, M. (2002). Protein-ligand recognition using spherical harmonic molecular surfaces: towards a fast and efficient filter for large virtual throughput screening. *J. Mol. Graph.* **20**, 313–328.

Carbo, R., Leyda, L., & Arnau, M. (1980). An electron density measure of the similarity between two compounds. *Int. J. Quant. Chem.* **17**, 1185–1189.

Carrieri, A., Pérez-Nueno, V. I., Fano, A., Pistone, C., Ritchie, D. W., & Teixidó, J. (2009). Biological profiling of anti-HIV agents and insights into CCR5 antagonist binding using in silico techniques. *ChemMedChem,* **4**, 1153–1163.

Chen, R. & Weng, Z. (2003). A novel shape complementarity scoring function for protein-protein docking. *Proteins: Struct. Func. Genet.* **51**, 397–408.

Cherfils, J. & Janin, J. (1993). Protein docking algorithms: simulating molecular recognition. *Curr. Op. Struct. Biol.* **3**, 265–269.

Chiu, L. Y. C. & Moharerrzadeh, M. (1999). Fourier transform of spherical Laguerre Gaussian functions and its application in molecular integrals. *Int. J. Quant. Chem.* **73**, 265–273.

Condon, E. U. & Odabasi, H. (1980). *Atomic Spectra*. Cambridge: Cambridge University Press.

Congreve, M., Murray, C. W., & Blundell, T. L. (2005). Structural biology and drug discovery. *Drug Discovery Today,* **10**, 895–907.

Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O., amd J. Thornton, R. G., & Orengo, C. A. (2008). The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res.* **37**, D310–D314.

Danos, M. & Maximon, L. C. (1965). Multipole matrix elements of the translation operator. *J. Math. Phys.* **6** (1), 766–778.

Davies, E. K., Glick, M., Harrison, K. N., & Richards, W. G. (2002). Pattern recognition and massively distributed computing. *J. Comp. Chem.* **23**, 1544–1550.

de Groot, B. L., van Aalten, D. M. F., Scheek, R. M., Amadei, A., Vriend, G., & Berendsen, H. J. C. (1997). Prediction of protein conformational freedom from distance constraints. *Proteins: Struct. Func. Genet.* **29**, 240–251.

Dean, P. M. (1995). *Molecular Similarity in Drug Design*. London: Blackie Academic & Professional.

Dean, P. M. & Callow, P. (1987). Molecular recognition: identification of local minima for matching in rotational 3-space by cluster analysis. *J. Mol. Graph.* **5** (3), 159–164.

Debnath, L. & Bhatta, D. (2007). *Integral Transforms and their Applications*. London: Chapman Hall.

Debnath, L. & Mikusinski, P. (1999). *Introduction to Hilbert Spaces with Applications*. London: Academic Press.

Edmonds, A. R. (1957). *Angular Momentum in Quantum Physics*. New Jersey: Princeton University Press.

Edvardson, H. & Smedby, O. (2003). Compact and efficient 3D shape description through radial function approximation. *Computer Methods and Programs in Biomedicine,* **72**, 89–97.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. New York: Academic Press.

Eisenstein, M. & Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *Comptes Rendus Biologies,* **327**, 409–420.

Elad, N., Maimon, T., Frenkiel-Krispin, D., Lim, R. Y. H., & Medalia, O. (2009). Structural analysis of the nuclear pore complex by integrated approaches. *Curr. Op. Struct. Biol.* **19**, 226–232.

Erdélyi, A., Magnus, W., Oberhettinger, F., & Tricomi, F. G. (1953a). *Higher Transcendental Functions*. New York: McGraw-Hill.

Erdélyi, A., Magnus, W., Oberhettinger, F., & Tricomi, F. G. (1953b). *Higher Transcendental Functions Vol 2*. New York: McGraw-Hill.

Erdélyi, A., Magnus, W., Oberhettinger, F., & Tricomi, F. G. (1953c). *Tables of Integral Transforms Vol 2*. New York: McGraw-Hill.

Fano, A., Ritchie, D. W., & Carrieri, A. (2006). Modelling the structural basis of human CCR5 chemokine receptor function: from homology model-building and molecular dynamics validation to agonist and antagonist docking. *J. Chem. Inf. Model.* **46** (3), 1223–1235.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* **7**, 861–874.

Fields, B. A., Malchiodi, E. L., Li, H., Ysern, X., Stauffacher, C. V., Schlievert, P. M., Karjalainen, K., & Mariuzza, R. A. (1996). Crystal structure of a T-cell receptor β-chain complexed with a superantigen. *Nature,* **384**, 188–192.

Frank, J. (2002a). Single-particle imaging of macromolecules by cryo-electron microscopy. *Ann. Rev. Biophys. Biolmol. Struct.* **31**, 303–319.

Frank, L. R. (2002b). Characterization of anisotropy in high angular resolution diffusion-weighted MRI. *Magnet. Reson. Med.* **47**, 1083–1099.

Funkhouser, T., Min, P., Kazhdan, M., Chen, D. Y., Halderman, A., & Dobkin, D. (2003). A search engine for 3D models. *ACM Transactions on Graphics,* **22**, 83–105.

Gabb, H. A., Jackson, R. M., & Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272** (1), 106–120.

Ganser-Pornillos, B. K., Yeager, M., & Sundquist, W. I. (2008). The structural biology of hiv assembly. *Curr. Op. Struct. Biol.* **18**, 203–217.

Garboczi, E. (2002). Three-dimensional mathematical analysis of particle shape using X-ray tomography and spherical harmonics: Application to aggregates used in concrete. *Cement and Concrete Research,* **32**, 1621–1638.

Garzón, J. I., Lopéz-Blanco, J. R., Pons, C., Kovacs, J., Abagyan, R., & Chacón, P. (2009). FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics,* , Advanced Access, 20 July 2009: doi:10.1093/bioinformatics/btp447.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., & Cruciat, C. M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141 – 147.

Gillet, A., Sanner, M., Stoffler, D., & Olson, A. (2005). Tangible interfaces for structural molecular biology. *Structure,* **13**, 483–491.

Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., & Kornberg, R. D. (2001). Structural basis of transcription: An RNA polymerase II elongation complex at 3.3å resolution. *Science,* **292**, 1876–1882.

González-Ruiz, D. & Gohlke, H. (2006). Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. *Curr. Med. Chem.* **13**, 2607–2625.

Goodford, P. J. (1985). A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857.

Gottfried, K. (1966). *Quantum mechanics*. New York: Benjamin.

Grant, J. A., Gallardo, M. A., & Pickup, B. T. (1996). A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comp. Chem.* **17** (14), 1653–1666.

Grant, J. A. & Pickup, B. T. (1995). A Gaussian description of molecular shape. *J. Phys. Chem.* **99**, 3503–3510.

Grigoriu, M., Garboczi, E., & Kafali, C. (2006). Spherical harmonic-based random fields for aggregates used in concrete. *Powder Technology,* **166**, 123–138.

Grünberg, R., Leckner, J., & Nilges, M. (2004). Complementarity of structure ensembles in protein-protein docking. *Structure,* **12**, 2125–2136.

Guéziec, A. & Hummel, R. (1995). Exploiting triangulated surface extraction using tetrahedral decomposition. *IEEE Trans. Vis. Comp. Graph.* **1** (4), 328–342.

Hart, G. T., Ramani, A. K., & Marcotte, E. M. (2006). How complete are current yeast and human protein interaction networks? *Genome Biol.* **7**, 120.

Hawkins, P. C. D., Skilllman, A. G., & Nicholls, A. (2007). Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74–82.

Helma, C., Cramer, T., Kramer, S., & De Raedt, L. (2004). Data mining and machine learning techniques for the identification of mutagenicity inducing substrates and structure activity relationships of noncongeneric compounds. *J. Chem. Inf. Comput. Sci.* **44**, 1402–1411.

Hinsen, K., Thomas, A., & Field, M. J. (1999). Analysis of domain motions in large proteins. *Proteins: Struct. Func. Genet.* **34**, 369–382.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., & Boutilier, K. (2002). Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature,* **415**, 180 – 183.

Hobson, E. W. (1931). *The Theory of Spherical and Ellipsoidal Harmonics*. London: Cambridge University Press.

Hochstadt, H. (1971). *The Functions of Mathematical Physics*. New York: Wiley.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C_α trace. Application to model building and detection of co-ordinate errors. *J. Mol. Biol.* **218**, 183–194.

Huang, H., Shen, L., Zhang, R., Makedon, F., Hettleman, B., & Perlman, J. (2005). Surface alignment of 3D spherical harmonic models: Application to cardiac MRI analysis. *LNCS 3749 – Medical Image Computing and Computer-Assisted Intervention,* **8** (1), 67–74.

Huang, N. & Shoichet, B. K. (2006). Benchmarking sets for molecular docking. *J. Med. Chem.* **49** (23), 6789–6801.

Inbar, Y., Schneidman-Duhovny, D., Oron, A., Nussinov, R., & Wolfson, H. J. (2005). Approaching the CAPRI challenge with an efficient geometry-based docking. *Proteins: Struct. Func. Bioinf.* **60**, 217–223.

Irwin, J. J. & Shoichet, B. K. (2005). ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* **98**, 4569–4574.

Jackson, J. D. (1975). *Classical Electrodynamics*. New York: Wiley.

Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M. J. E., Vajda, S., Vakser, I., & Wodak, S. J. (2003). CAPRI: a critical assessment of predicted interactions. *Proteins: Struct. Func. Genet.* **52**, 2–9.

Jensen, F. (1999). *Introduction to Computational Chemistry*. New York: Wiley.

Jones, G., Willett, P., & Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53.

Kammler, D. (2000). *A First Course in Fourier Analysis*. Upper Saddle River, NJ: Prentice-Hall.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.* **89**, 2195–2199.

Kautz, J., Sloan, P. P., & Snyder, J. (2002). Fourier method for large-scale surface modeling and registration. *Proceedings of 13th Eurographics workshop on rendering,* **28**, 291–296.

Kazhdan, M., Funkhouser, T., & Rusinkiewcz, S. (2003). Rotation invariant spherical harmonic representation of 3D shape descriptors. *Proceedings of 2003 Eurographics/ACM SIGGRAPH symposium on geometry processing,* **43**, 156–164.

Keister, B. D. & Polyzou, W. N. (1997). Useful bases for problems in nuclear and particle physics. *J. Comp. Phys.* **134**, 231–235.

Kellenberger, E., Springael, J. Y., Parmentier, M., Hachet-Haas, M., Galzi, J. L., & Rognan, D. (2007). Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J. Med. Chem.* **50**, 1294–1303.

Khodade, P., Prabhu, R., Chandra, N., Raha, S., & Govindrajan, R. (2007). Parallel implementation of autodock. *J. Appl. Cryst.* **40**, 598–599.

Kolodny, R., Koehl, P., & Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures. *J. Mol. Biol.* **346**, 1173–1188.

Kovacs, J. A., Chacon, P., Cong, Y., Metwally, E., & Wriggers, W. (2003). Fast rotation matching of rigid bodies by fast Fourier transform acceleration of five degrees of freedom. *Acta Cryst.* **D59**, 1371–1376.

Kovacs, J. A. & Wriggers, W. (2002). Fast rotation matching. *Acta Cryst.* **D58**, 1282–1286.

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Struct. Func. Bioinf.* **65**, 392–406.

Kozakov, D., Clodfelter, K. H., Vajda, S., & Camacho, C. J. (2005). Optimal clustering for detecting near-native conformations in protein docking. *Biophys. J.* **89**, 867–875.

Lanczos, C. (1964). A precision approximation of the gamma function. *J. SIAM Numer. Anal.* **B1**, 86–96.

Lanzavecchia, S., Cantele, F., & Bellon, P. L. (2001). Alignment of 3D structures of macromolecular assemblies. *Bioinformatics,* **17** (1), 58–62.

Larson, R. S. (2006). *Bioinformatics and Drug Discovery*. New Jersey: Humana Press.

Lebedev, N. N. (1972). *Special Functions and Their Applications*. New York: Dover.

Lemmen, C. & Lengauer, T. (2000). Computational methods for the structural alignment of molecules. *J. Comput.-Aid. Mol. Des.* **14**, 215–232.

Lewis, P., De Jonge, M., Daeyaert, F., Koymans, L., Vinkers, M., Heeres, J., Janssen, P. A. J., Arnold, E., Das, K., Clark, A. D., Hughes, S. H., Boyer, P. L., De Béthune, M. P., Pauwels, R., Andries,

K., Kukla, M., Ludovici, D., De Corte, B., Cavas, R., & Ho, C. (2003). On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J. Comput.-Aid. Mol. Des.* **17**, 129–134.

Libbrecht, K. G. (1985). Practical considerations for the generation of large-order spherical harmonics. *Solar Physics,* **99** (1-2), 371–373.

Lin, J. & Clark, T. (2005). An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *J. Chem. Inf. Model.* **45**, 1010–1016.

Lindert, S., Stewart, P. L., & Meiler, J. (2009). Hybrid approaches: applying computational methods in cryo-electron microscopy. *Curr. Op. Struct. Biol.* **19**, 218–225.

Luke, Y. L. (1969). *The Special Functions and their Approximation*. New York: Academic Press.

Macindoe, G., Mavridis, L., Venkatraman, V., Devignes, M.-D., & Ritchie, D. W. (2010). HexServer: an FFT-based protein docking server powered by graphics processors. *Nucleic Acids Res.* **38**, W445–W449.

Madej, T., Gibrat, J.-F., & Bryant, S. H. (1995). Threading a database of protein cores. *Proteins: Struct. Func. Bioinf.* **23** (3), 356–369.

Mak, L., Grandison, S., & Morris, R. J. (2008). An extension of spherical harmonics to region-based rotationally invariant descriptors for molecular shape description and comparison. *J. Mol. Graph. Model.* **26**, 1035–1045.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., & Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng.* **14** (2), 105–113.

Mathivavnan, S., Periaswamy, B., Gandhi, T. K. B., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y. L., & Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics,* **7**, S19.

Matter, H. & Schwab, W. (1999). Affinity and selectivity of matrix metalloproteinase inhibitors: a chemometrical study from the perspective of ligands and proteins. *J. Med. Chem.* **42**, 4506–4523.

Mavridis, L., Hudson, B. D., & Ritchie, D. W. (2007). Toward high throughput screening using spherical harmonic surface representations. *J. Chem. Inf. Model.* **47** (5), 1878–1796.

Mavridis, L. & Ritchie, D. W. (2009). 3D-Blast: protein protein structure alignment, comparison, and classification using spherical polar fourier correlations. *Pacific Symposium on Biocomputing,* **2010**, 281–292.

May, A. & Zacharias, M. (2008). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins: Struct. Func. Bioinf.* **70**, 794–809.

McPeek, M. A., Shen, L., & Farid, H. (2009). The correlated evolution of three-dimensional reproductive structures between male and female damsel flies. *Evolution,* **63**, 73–83.

McPeek, M. A., Shen, L., & Torrey, J. Z. (2008). The tempo and mode of three-dimensional morphological evolution in male reproductive structures. *Americal Naturalist,* **171**, E158–E178.

Méndez, R., Leplae, R., De Maria, L., & Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins: Struct. Func. Genet.* **52**, 51–67.

Méndez, R. & Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Struct. Func. Bioinf.* **60**, 150–169.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., & Weng, Z. (2005). Protein-protein docking benchmark 2.0: An update. *Proteins: Struct. Func. Bioinf.* **60**, 214–216.

Morris, R. J., Najmanovich, R. J., Kahraman, A., & Thornton, J. (2005). Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics,* **21**, 2347–2355.

Muench, S. P., Huss, M., Song, C. F., Phillips, C., Wieczorek, H., Trinick, J., & Harrison, M. A. (2009). Cryo-electron microscopy of the vacuolar ATPase motor reveals its mechanical and regulatory complexity. *J. Mol. Biol.* **386**, 989–999.

Muggleton, S. (2005). Machine learning for systems biology. In: *15th International Conference on Inductive Logic Programming*, (Kramer, S. & Pfahringer, B., eds) pp. 416–423, Bonn: Springer LNAI 3625.

Mustard, D. & Ritchie, D. W. (2005). Docking essential dynamics eigenstructures. *Proteins: Struct. Func. Bioinf.* **60**, 269–274.

Nagata, A., Mizushima, H., & Tanaka, H. (2002). Concept and prototype of protein-ligand docking simulator with force feedback technology. *Bioinformatics,* **18**, 140–146.

Napoli, A. (2005). A smooth introduction to symbolic methods for knowledge discovery. In: *Handbook of Categorization in Cognitive Science*, (Cohen, H. & Lefebvre, C., eds) pp. 813–933, Amsterdam: Elsevier.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**, 645–660.

Novotni, M. & Klein, R. (2003). 3d zernike descriptors for content based shape retrieval. *Proceedings of the eighth ACM symposium on Solid modeling and applications,* **SPM08**, 216–225.

Orengo, C. A., Michine, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH - A hierarchic classification of protein domain structures. *Structure,* **5** (8), 1093–1108.

Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., & Purcell, T. J. (2007). A survey of general-purpose computation on graphics hardware. *Comp. Graph. Forum,* **26**, 80–113.

Papageorgiou, A. C., Collins, C. M., Gutman, D. M., Kline, J. B., O'Brien, S. M., Tranter, H. S., & Acharya, K. R. (1995). Structural basis for the recognition of superantigen streptococcal pyrogenic exotoxin A (SpeA1) by MHC class II molecules and T-cell receptors. *EMBO J.* **18**, 9–21.

Park, H., Lee, J., & Lee, S. (2006). Critical assessment of the automated autodock as a new docking tool for virtual screening. *Proteins: Struct. Func. Genet.* **65**, 594–554.

Pastor, M. & Cruciani, G. (1995). A novel strategy for improving ligand selectivity in receptor-based drug design. *J. Med. Chem.* **38**, 4637–4647.

Pérez-Nueno, V. I., Ritchie, D. W., Rabal, O., Pascual, R., Borrell, J. I., & Teixidó, J. (2008). Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. *J. Chem. Inf. Model.* **48** (3), 509–533.

Petsko, G. A. & Ringe, D. (2004). *Protein Structure and Function*. London: New Science Press.

Pevsner, J. (2003). *Bioinformatics and Functional Genomics*. New York: Wiley.

Phuong, T. & Ho, T. B. (2005). Prediction of domain-domain interactions using inductive logic programming from multiple genome databases. In: *Ninth International Conference on Discovery Science*, (Lavrac, N., Todorovski, L., Jantke, K., & Klaus, P., eds) pp. 185–196, Bonn: Springer LNAI 4256.

Richards, W. G. (2002). Virtual screening using grid computing: the screensaver project. *Nature Rev. Drug. Disc.* **1**, 551–555.

Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.* **178**, 63–89.

Ritchie, D. W. (1998). *Parametric Protein Shape Recognition*. PhD thesis University of Aberdeen U.K.

Ritchie, D. W. (2003). Evaluation of protein docking predictions using *Hex* 3.1 in CAPRI rounds 1 and 2. *Proteins: Struct. Func. Genet.* **52** (1), 98–106.

Ritchie, D. W. (2005). High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. *J. Appl. Cryst.* **38**, 808–818.

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. *Curr. Prot. Pep. Sci.* **9** (1), 1–15.

Ritchie, D. W. & Kemp, G. J. L. (1999). Fast computation, rotation and comparison of low resolution spherical harmonic molecular surfaces. *J. Comp. Chem.* **20** (4), 383–395.

Ritchie, D. W. & Kemp, G. J. L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins: Struct. Func. Genet.* **39** (2), 178–194.

Ritchie, D. W., Kozakov, D., & Vajda, S. (2008). Accelerating protein-protein docking correlations using a six-dimensional analytic FFT generating function. *Bioinformatics,* **24** (4), 810–823.

Ritchie, D. W. & Venkatraman, V. (2010). Ultra-fast FFT protein docking on graphics processors. *Bioinformatics,* **26**, 2398–2405.

Robinson, R. C., Turbedsky, K., Kaiser, D. A., Marchand, J. B., Higgs, H. N., Choe, S., & Pollard, T. D. (2001). Crystal structure of arp2/3 complex. *Science,* **294**, 1679–1684.

Rose, M. E. (1957). *Elementary Theory of Angular Momentum*. New York: Wiley.

Roseman, A. M. (2000). Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Cryst.* **D56**, 1332–1340.

Rossmann, M. G. (2000). Fitting atomic models into electron-microscopy maps. *Acta Cryst.* **D56**, 1341–1349.

Rossmann, M. G., Arisaka, F., Battisti, A. J., Bowman, V. D., Chipman, P. R., Fokine, A., Halfstein, S., Kanamura, S., Kostyuchemko, V. A., Mesyanzhinov, V. V., Schneider, M. M., Morais, M. C., Leiman, P. G., Palermo, L. M., Parrish, C. R., & Xiao, C. (2007). From structure of the complex to understanding of the biology. *Acta Cryst.* **D63**, 9–16.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.

Rueda, M., Bottegoni, G., & Abagyan, R. (2009). Consistent improvement of cross-docking results using binding site ensembles generated with elastic network normal modes. *J. Chem. Inf. Model.* **49**, 716–725.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., & Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr. Op. Struct. Biol.* **14**, 313–324.

Sael, L., , La, D., Fang, Y., Ramani, K., R.Rustamov, & Kihara, D. (2008). Fast protein tertiary structure retrieval based on global surface shape similarity. *Proteins: Struct. Func. Bioinf.* **72**, 1259–1273.

Sakurai, J. J. (1994). *Modern Quantum Mechanics*. Reading, MA: Addison-Wesley.

Shen, L., Farid, H., & McPeek, M. A. (2009a). Modeling three-dimensional morphological structures using spherical harmonics. *Evolution,* **63**, 1003–1016.

Shen, L., Kim, S., & Saykin, A. J. (2009b). Fourier method for large-scale surface modeling and registration. *Computers & Graphics,* **33**, 299–311.

Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.

Sippl, M. J. & Wiederstein, M. (2008). A note on difficult structure alignment problems. *Bioinformatics,* **24**, 426–427.

Smith, G. R., Sternberg, M. J. E., & Bates, P. A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.* **347**, 1077–1101.

Srinivasan, K., Mahawar, H., & Sarin, V. (2005). A multipole based treecode using spherical harmonics for potentials of the form $r^{-\lambda}$. *Lect. Notes Comp. Sci.* **3514**, 107–104.

Stein, M., Gabdoulline, R. R., & Wade, R. C. (2007). Bridging from molecular simulations to biochemical networks. *Curr. Op. Struct. Biol.* **17**, 166–172.

Sticht, J., Humbert, M., Findlow, S., Bodem, J., Müller, B., Deitrich, U., Werner, J., & Kr"ausslich, H. G. (2005). A peptide inhibitor of HIV-1 assembly *in vitro*. *Nature Struct. Biol.* **12** (8), 671–677.

Stowell, M. H. B., Miyazawa, A., & Unwin, N. (1998). Macromolecular structure determination by electron microscopy: new advances and recent results. *Curr. Op. Struct. Biol.* **8**, 606–611.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola., E. E. (1998). Protein data bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Cryst.* **D54**, 1078–1084.

Taha, M. O., Qandil, A. M., Zaki, D. D., & Aldaben, M. A. (2005). Ligand-based assessment of factor Xa binding site flexibility via elaborate pharmacophore exploration and genetic algorithm-based QSAR modeling. *Eur. J. Med. Chem.* **40**, 701–727.

Takana, S. Y. & Mitchell, J. B. O. (2004). A structure-odour relationship study using EVA descriptors and hierarchical clustering. *Organic Biomol. Chem.* **22** (2), 3250–3255.

Talman, J. D. (1968). *Special Functions: A Group Theoretical Approach*. New York: W. A. Benjamin Inc.

Tama, F., Gadea, F. X., Marques, O., & Sanejouand, Y. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins: Struct. Func. Genet.* **41**, 1–7.

Tama, F. & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14** (1), 1–6.

Taylor, W. R. (1999). Protein structure comparison using iterated double dynamic programming. *Protein Sci.* **8**, 654–665.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

Tirion, M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M. J., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *saccaromyces cerevisiae*. *Nature,* **403**, 623–671.

Vakser, I. A. & Aflalo, C. (1994). Hydrophobic docking: A proposed enhancement to molecular recognition techniques. *Proteins: Struct. Func. Genet.* **20**, 320–329.

Venkatraman, V., Sael, L., & Kihara, D. (2009). Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. *Cell Biochem. Biophys,* **54**, 23–32.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association,* **58**, 236–244.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta Jr., S., & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784.

Weniger, E. J. & Steinborn, E. O. (1983). The Fourier transforms of some exponential-type basis functions and their relevance to multicenter problems. *J. Chem. Phys.* **78**, 6121–6132.

Wiggins, R. A. & Saito, M. (1971). Evaluation of computational algorithms for the associated Legendre polynomials by interval analysis. *Bull Seismol. Soc. Am.* **61** (2), 375–381.

Wigner, E. P. (1939). On the unitary representations of the inhomogeneous repesentation of the Lorentz group. *Annals of Mathematics,* **40** (1), 149–204.

Williams, S., Bledsoe, R. K., Collins, J. L., Boggs, S., Lambert, M. H., Miller, A. B., Moore, J., McKee, D. D., Moore, L., Nichols, J., Parks, D., Watson, M., Wisely, B., & Willson, T. M. (2003). X-ray crystal structure of the liver X receptor β ligand binding domain: Regulation by a histidine-tryptophan switch. *J. Biol. Chem.* **278**, 27138–27143.

Wollacott, A. M. & Mertz, K. M. (2005). Haptic applications for molecular structure manipulation. *J. Mol. Graph. Model.* **25**, 801–805.

Wong, R. S., Bodart, V., Metz, M., Labrecque, J., Bridger, G., & Fricker, S. P. (2006). Prediction of multiple binding modes of the CDK2 inhibitors, anilinopyrazoles, using the automated docking programs GOLD, FlexX, and LigandFit: an evaluation of performance. *J. Chem. Inf. Model.* **46**, 2552–2562.

Wong, R. S., Bodart, V., Metz, M., Labrecque, J., Bridger, G., & Fricker, S. P. (2008). Comparison of the potential multiple binding modes of bicyclam, monocylam, and noncyclam small-molecule CXC chemokine receptor 4 inhibitors. *Mol. Pharmacol.* **74**, 1485–1495.

Wriggers, W., Milligan, R. A., & McCammon, J. A. (1999). Situs: A package for docking crystal structures into low-resolution maps from electron density. *J. Struct. Biol.* **125**, 185–195.

Yamagishi, M. E. B., Martins, N. F., Neshich, G., Cai, W., Shao, X., Beautrait, A., & Maigret, B. (2006). A fast surface-matching procedure for protein-ligand docking. *J. Mol. Model.* **12** (2), 965–972.

Annexe A

Publications appropriées

Cette annexe contient des copies des articles dans les revues internationales à comité de lecture :

- Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. D.W. **Ritchie**, G.J.L. Kemp. *J. Comp. Chem.* **20**(4), 383–395 (1999).
- Protein docking using spherical polar Fourier correlations. D.W. **Ritchie**, G.J.L. Kemp. *Proteins: Struct. Func. Genet.* **39**, 178–194 (2000).
- Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. D.W. **Ritchie**. *Proteins: Struct. Func. Genet.* **52**, 98–106 (2003).
- Docking essential dynamics eigenstructures. D. Mustard and D.W. **Ritchie**. *Proteins: Struct. Funct. Bioinf.* **60**, 269–274 (2005).
- High order analytic translation matrix elements for real space six-dimensional polar Fourier correlations. D.W. **Ritchie**. *J. Appl. Cryst.* **38** 808–818 (2005).
- Modelling the structural basis of human CCR5 chemokine receptor function: from homology model-building and molecular dynamics validation to agonist and antagonist docking. A. Fano, D.W. **Ritchie**, A. Carrieri. *J. Chem. Inf. Model.* **46**(3) 1223–1235 (2006).
- Toward high throughput screening using spherical harmonic surface representations. L. Mavridis, B.D. Hudson, D.W. **Ritchie**. *J. Chem. Inf. Model.* **47**(5) 1787–1796 (2007).
- Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking. V.I. Pérez-Nueno, D.W. **Ritchie**, O. Rabal, R. Pascual, J.I. Borel, J. Teixidó. *J. Chem. Inf. Model.* **48**(3) 509–533 (2008).

- Recent progress and future directions in protein-protein docking. D.W. **Ritchie**. *Curr. Prot. Pep. Sci.* **9**(1) 1–15 (2008).
- Accelerating protein-protein docking correlations using a six-dimensional analytic FFT generating function. D.W. **Ritchie**, D. Kozakov, and S. Vajda. *Bioinformatics* **24**(4) 810–823 (2008).
- Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket. V.I. Pérez-Nueno, D.W. **Ritchie**, J.I. Borrell, and J. Teixidó. *J. Chem. Inf. Model.* **48**(11) 2146–2165 (2008).
- 3D-Blast: 3D protein structure alignment, comparison, and classification using spherical polar Fourier correlations. L. Mavridis and D.W. **Ritchie**. *Pacific Symposium on Biocomputing (PSB 2010),* 281–292.
- SHREC-10 Track: Protein Models. L. Mavridis, V. Venkatraman, D. W. Ritchie, N. Morikawa, R. Andonov, A. Cornu, N. Malod-Dognin, J. Nicolas, M. Temerinac-Ott, M. Reisert, H. Burkhardt, A. Axenopoulos (2010). 3DOR: Eurographics Workshop on 3D Object Retrieval (2010), 117–124.
- Ultra-Fast Protein Docking on Graphics Processors. D.W. **Ritchie**, V. Venkatraman, (2010). *Bioinformatics*. **26**, 2398–2405.
- HexServer: an FFT-based protein docking server powered by graphics processors. G. Macindoe, L. Mavridis, V. Venkatraman, M.-D. Devignes, D.W. **Ritchie** (2010). *Nucleic Acids Research*, **38**, W445–W449.