Détermination de la structure de protéines à l'aide de données faiblement résolues

Marc Piuzzi

Thèse de doctorat
de l'Université Pierre et Marie Curie

Spécialité Bioinformatique
École doctorale iViv

Présentée par :

Marc Piuzzi

Pour obtenir le grade de

Docteur de l'Université Pierre et Marie Curie

Sujet de la thèse :

Détermination de la structure de protéines à l'aide de données faiblement résolues

Soutenue le 3 Décembre 2010 devant le jury composé de :

M. François BONTEMS, Directeur de thèse

Mme Anne POUPON, Rapporteur

M. Christian ROUMESTAND, Rapporteur

M. Bernard GILQUIN, Examinateur

M. Olivier LEQUIN, Examinateur

M. Javier PEREZ, Examinateur

Mlle Christina SIZUN

REMERCIEMENTS

Le travail présenté dans cette thèse a été effectué au sein de l'Institut de Chimie des Substances Naturelles qui en a assuré le financement. Je tiens à remercier Jean-Yves Lallemand d'avoir accepté de financer ce travail qui n'est pas dans le cœur de métier de l'institut. La direction de la thèse a été assurée par François Bontems qui a permis d'engager ce travail pluridisciplinaire.

Cette thèse a commencé en 2007 en collaboration avec Fabien Mareuil dont j'ai en partie repris le travail. Grâce à son accueil et sa motivation, le début de la thèse en a été grandement facilité et je l'en remercie pour cela. De même, je remercie Christina Sizun pour son aide précieuse lors du travail expérimental et pour ses commentaires toujours productifs durant le reste de l'étude. Malgré mes passages souvent rapides, l'équipe du 23B m'a toujours offert un excellent accueil et des discussions intéressantes lors des repas. Eric Guittet a assuré les derniers mois de financement de mon travail et je lui en suis reconnaissant.

Je remercie aussi Pierre, Alain, Philippe pour les diverses discussions sur des sujets aussi intéressants que variés. Enfin, je ne peux conclure ces remerciements sans aborder le soutien inconditionnel et essentiel d'Hannah qui m'a souvent permis de prendre du recul sur ce que j'étais en train de faire et qui a toujours trouvé les bons mots dans les moments un peu difficiles. Je tiens à remercier mes parents qui m'ont encouragé tout au long de ces trois années.

RÉSUMÉ

La connaissance des structures tridimensionnelles des macromolécules biologiques est indispensable pour mieux comprendre leur rôle et pour la conception de nouvelles molécules thérapeutiques. Les techniques utilisées actuellement offrent une grande variété d'approches qui utilisent à la fois des informations spécifiques à la protéine étudiée et des informations génériques communes à l'ensemble des protéines. Il est possible de classer ces méthodes en fonction de la quantité d'information utilisée dans chacune de ces deux catégories avec d'un côté des méthodes utilisant le plus possible de données spécifiques à la protéine étudiée et de l'autre les méthodes utilisant le plus possibles de données génériques présentes dans les bases de données.

Le travail présenté dans cette thèse aborde deux utilisations de techniques mixtes, présentant une autre combinaison entre données spécifiques et données génériques. En particulier nous avons cherché à obtenir la structure de protéines composée d'un ou deux domaines en ne disposant que d'un nombre restreint de données spécifiques. Pour déterminer la structure d'une protéine de grande taille composée de deux domaines à l'aide de données de diffusion des rayons X et de modèles obtenus par de la modélisation par homologie, nous avons adapté puis optimisé un programme récemment développé au laboratoire. Nous avons ensuite modélisé la structure d'un domaine d'une protéine de virus en incorporant un faible nombre de contraintes issues des données obtenues par RMN dans une méthode de prédiction de structure « *ab initio* ». Enfin, nous avons étudié l'intérêt d'intégrer les courants de cycle, une composante du déplacement chimique, dans un programme d'arrimage moléculaire pour la résolution de complexes protéine-ADN.

ABSTRACT

The knowledge of the tridimensional structure of biological macromolecules is essential to fully comprehend their role and to design new pharmaceutical molecules. Commonly used methods offer a broad spectrum of techniques which uses data that are specific to the studied protein and generic through the use of databases. The work of this thesis present various mixed techniques which uses a limited set of specific data combined with the use of databases. In particular, we tried to generate models of a bi-domain protein using low-resolution data. This has been done using small angle X-ray Scattering (SAXS) and models of the individual domains obtained by homology modeling. Models of the C-terminal domain of protein P from PPRV were determined using Rosetta and sparse NMR experimental data (NOE). Eventually, we discuss the use of ring current effects to drive protein-nucleic acids docking.

1 LA DÉTERMINATION DE LA STRUCTURE DES PROTÉINES

Le travail présenté dans cette thèse aborde différents aspects de la biologie structurale, domaine portant sur l'étude de la structure des protéines et des macromolécules biologiques. La problématique développée ici est la détermination de la structure de protéines ou d'assemblage de protéines par des techniques intermédiaires entre les méthodes expérimentales classiques utilisant des données spécifiques à la protéine étudiée et les méthodes de prédiction « *ab initio* » utilisant le plus possible de données génériques communes à l'ensemble des protéines. Pourquoi s'intéresser à cette problématique ? La machinerie cellulaire repose en grande partie sur les protéines qui peuvent interagir avec d'autres protéines, une petite molécule ou encore un acide nucléique dans un processus biologique. Connaître le repliement d'une protéine, c'est à dire l'agencement dans l'espace tridimensionnel des acides aminés formant sa séquence, est essentiel pour mieux comprendre son rôle, ses possibles interactions et d'un niveau plus global le fonctionnement d'un mécanisme biologique au niveau moléculaire. La connaissance du repliement peut aussi être utile pour des applications à but thérapeutique direct comme la compréhension des effets d'une mutation de la séquence des acides aminés ou la recherche de nouvelles molécules thérapeutiques de synthèse, le « drug design ». Cette recherche porte sur la conception de petites molécules ciblant de manière très précise un mécanisme biochimique particulier et nécessite donc une bonne connaissance de la structure des molécules impliquées.

2 LES MÉTHODES DITES « EXPÉRIMENTALES »

Depuis la création des projets de génomique haut-débit, de plus en plus de séquences de protéines sont disponibles pour la communauté scientifique. Pour un grand nombre de ces séquences, on ne connaît ni le repliement associé ni nécessairement la fonction. Jusque dans les années 1990, il était indispensable de disposer d'un grand nombre de données spécifiques obtenues à l'aide de techniques expérimentales pour pouvoir déterminer des structures de protéines. Les deux techniques les plus répandues sont la

radio-cristallographie aux rayons X et la résonnance magnétique nucléaire (RMN) dont voici une brève présentation.

La radio-cristallographie aux rayons X nécessite d'obtenir la protéine sous la forme d'un cristal ordonné. Le motif de diffraction des rayons X après passage dans le cristal est utilisé pour déterminer l'empreinte atomique de la structure en trois dimensions. À l'aide de cette empreinte et de données génériques communes à l'ensemble des protéines (structures des acides aminés, angles dihédriques, …), il est possible de déterminer un modèle de la protéine. Ce motif résulte de la diffusion de la radiation électromagnétique par les électrons et dépend de leur distribution et de leur densité. Le principal désavantage de cette technique (en dehors de la nécessité d'avoir accès à une source de rayons X) est la nécessité de former des cristaux ordonnés de protéine. En effet, les conditions de formation du cristal varient en fonction de chaque protéine et il est souvent nécessaire d'en tester un grand nombre avant d'obtenir les premiers cristaux. L'avantage de cette technique est sa très bonne résolution spatiale, proche du rayon de van der Waals d'un atome d'hydrogène dans le meilleur de cas.

En comparaison, la résonance magnétique nucléaire permet d'obtenir un ensemble de contraintes de distances et d'angles sur les acides aminés de la protéine. En combinant ces contraintes spécifiques à la protéine avec des données génériques, il est possible d'obtenir un ensemble de modèles satisfaisant ces contraintes. L'avantage de cette technique expérimentale est de pouvoir étudier la protéine en solution et donc de pouvoir étudier sa structure dans un milieu dynamique et non pas figée dans un cristal. Comme la protéine n'est pas immobile en solution, la fluctuation du repliement va engendrer un ensemble de contraintes reflétant cette mobilité. C'est pour cette raison que la RMN, contrairement à la radio-cristallographie, qui propose un modèle unique, propose un ensemble de modèles.

Il est important de noter, par conséquent, que les techniques expérimentales « pures » n'existent pas car, dans tous les cas, il est nécessaire d'introduire de l'information *a priori*. Pour les méthodes qualifiées d'expérimentales, cette information ne concerne que la géométrie covalente et pas l'organisation spatiale. En général, il y a toujours la tentation d'introduire de l'information ne provenant pas directement des expériences réalisées. Par

exemple, en RMN, Clore et Gronenborn (Kuszewski et al. 1996) ont proposé de contraindre le diagramme de Ramachandran.

3 LES MÉTHODES DITES « IN SILICO »

À l'opposition de ces techniques utilisant le plus possibles de données spécifiques à la protéine étudiée, des méthodes de prédiction de structures à l'aide de données génériques se sont développées. Ces méthodes de prédiction ne sont pas uniquement « *in silico* » puisque la plupart d'entre elles reposent sur la conversion de données statistiques, issues d'expériences, en potentiels de contrainte ou de tri. L'approche proposée pour résoudre le problème de la prédiction de structure des protéines est d'utiliser la structure d'une ou plusieurs protéines dont la séquence présente une grande similarité. Cette solution est appelée modélisation comparative, aussi dénommée modélisation par homologie. Il a été estimé que la résolution d'au moins 16000 nouvelles structures (soigneusement sélectionnées) permettrait de couvrir, grâce à cette approche, 90% des familles de domaines des protéines. Au rythme actuel, ce résultat devrait être atteint dans les dix prochaines années (Vitkup et al. 2001).

Aidée par l'augmentation de la puissance des unités de calcul, ces nouvelles méthodes pour la détermination de modèles de protéines se sont rapidement développées. Néanmoins, les modèles obtenus ne sont pas toujours parfaitement résolus et leur champ d'application dépend de la résolution de la modélisation estimée (Figure 1). Le développement rapide de ces méthodes a été poussé à la fois par le besoin grandissant de l'automatisation et par la difficulté d'obtenir suffisamment de données expérimentales pour la détermination des structures.

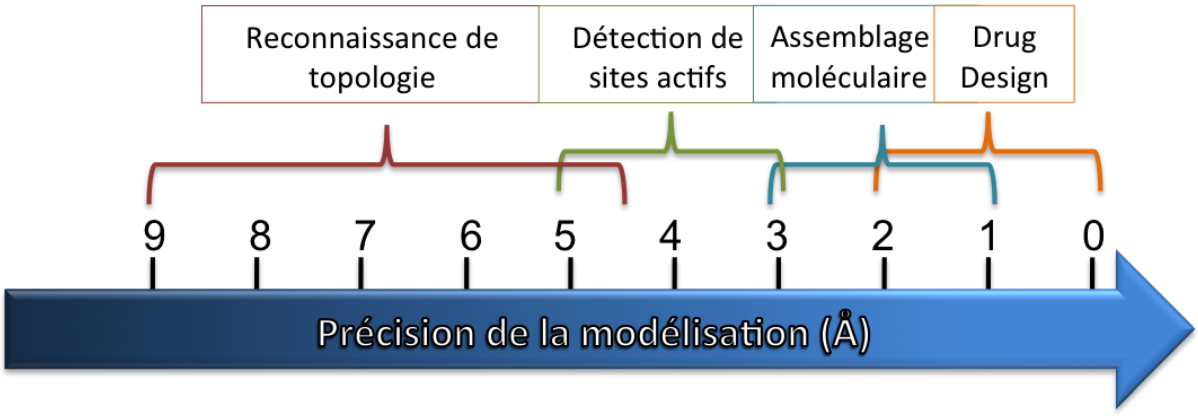


Figure 1 : Utilisation possible des modèles en fonction de la précision de la modélisation

3.1 Problématique générale

Le problème général posé par le repliement des protéines est un problème de minimisation d'énergie. Si il est admis que la forme la plus stable d'une protéine correspond à la structure de plus basse énergie adoptée par les acides aminés (qui peut varier en fonction du solvant et de la température), trouver cette forme est loin d'être trivial car il s'agit de trouver une méthode permettant de s'assurer de trouver le minimum global dans une fonction d'énergie qui admet un grand nombre de minima locaux (Figure 2).

Pour faire une analogie simple, on peut représenter l'ensemble des conformations possibles comme la surface d'une nouvelle planète et chaque coordonnée correspond à un repliement de la protéine. Dans ce cas, la recherche du repliement de plus basse énergie devient la recherche du lieu de plus basse élévation. Comme l'atmosphère de la planète est opaque, on ne peut connaître l'élévation qu'en étant très près de la surface et donc sans avoir une vue d'ensemble de toute la surface de la planète. Le moyen le plus simple de trouver le point de plus basse élévation est d'explorer toute la surface mais cela prend beaucoup trop de temps. On cherche donc des méthodes permettant d'isoler les régions contenant les lieux de basse élévation tout en s'assurant de ne pas écarter le lieu de plus basse élévation.

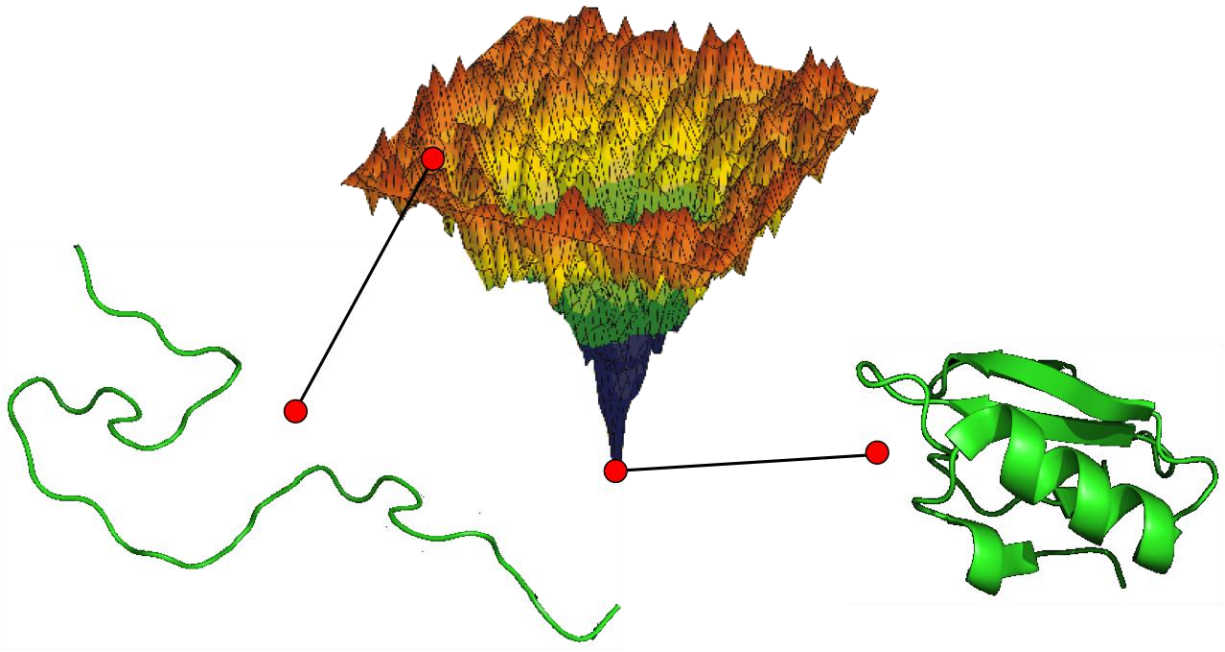


Figure 2 : Représentation schématique de l'état de repliement d'une protéine en fonction de son niveau d'énergie.

3.2 Les différentes catégories de méthodes

Les méthodes qui n'utilisent pas de données spécifiques à la protéine étudiée (à l'exception de la séquence en acides aminés) peuvent être classées en trois catégories, suivant les données qu'elles utilisent dans leur fonctionnement : la modélisation par homologie, la reconnaissance de repliement et la prédiction de structure *ab initio*. Cette catégorisation est issue du Critical Assessment of Protein Structure Prediction (CASP), un concours qui permet de comparer les résultats des méthodes de ces différentes catégories (Barton & Russell 1993; Benner et al. 1992) dont voici une brève présentation.

La modélisation par homologie utilise les outils d'alignement de séquences pour trouver des protéines dont la séquence est proche de la protéine que l'on cherche à modéliser. La qualité des modèles produits est généralement bonne jusqu'à 25% d'identité de séquence. C'est une méthode facile à utiliser avec des protocoles qui sont aujourd'hui parvenus à maturité. Les méthodes de reconnaissance de repliement cherchent à trouver quel type de repliement la nouvelle protéine peut adopter quand on ne dispose que de sa séquence. Comme le nombre de repliements est limité, il est possible de retrouver les mêmes repliements dans des protéines très éloignées. À ce jour, ce sont les deux méthodes présentant les meilleurs résultats (Moult 1999) mais elles reposent sur le fait que l'on dispose de séquences homologues dans les bases de données ce qui n'est fréquemment pas le cas. Dans le cas où l'on ne dispose pas de séquences homologues, des méthodes efficaces de prédiction de structure secondaire et de motif locaux ont vu le jour (Rost et al. 1994; King et al. 1997; Bystroff & Baker 1998; Jones 1999; Karplus et al. 1999). Cependant, ces méthodes ne permettent pas de modéliser une protéine dans son intégralité et c'est dans ce but qu'ont émergé les méthodes dites « *ab initio* ». L'objectif des méthodes *ab initio* (couramment appelées *de novo*) est de trouver la structure d'une protéine à partir de sa séquence en acide aminé, de potentiels statistiques et de bases de données. Si elles n'utilisent pas de séquences homologues de façon directe, il est important de noter qu'elles utilisent l'intégralité des bases de données à leur disposition. Ainsi, si l'on cherche à évaluer les résultats obtenus par une méthode de ce type il est important de retirer de la base de données utilisée les structures dont les séquences sont proche de celles du jeu de test.

Le point commun des méthodes de ces trois catégories de modélisation (homologie, reconnaissance de repliement et *ab initio*) est de produire un grand nombre de modèles

minimisant une fonction de score interne. On peut déjà noter une différence dans les objectifs des techniques utilisant des données expérimentales spécifiques à l'objet étudiée et les méthodes utilisant des données génériques. Dans le premier cas, l'objectif est de trouver « la » structure, ce qui est réalisable en jouant sur la quantité et la qualité des données spécifiques. Schématiquement, l'objectif (trouver la structure) nécessite d'acquérir assez de données spécifiques. Dans le deuxième, le problème est inverse car le fait de n'utiliser que des contraintes génériques pose une condition sur l'objectif : trouver toutes les structures possibles compatibles avec la connaissance du système. Cela pose alors le problème de pouvoir identifier parmi ces structures les structures les plus proches de la « vraie » structure. C'est un problème récurrent rencontré par tous les programmes de modélisation.

En résumé, quelle que soit la méthode utilisée lorsque l'on détermine un modèle de protéine, on mélange des données issues de l'expérience concernant l'objet particulier auquel on s'intéresse (la protéine) et de la classe d'objet auquel il appartient (les protéines). Par exemple, un cliché de diffraction ou une carte de RMN appartient manifestement au premier type de données. A l'inverse les contraintes sur les potentiels de la longueur des liaisons, de respect du diagramme de Ramachandran ou sur les résidus hydrophobes appartiennent manifestement au second type. On peut remarquer que les différentes techniques vont varier entre des situations avec le moins possible de données génériques (techniques expérimentales) et des situations avec le moins possible de données spécifiques (techniques « *ab initio* »). Au milieu, on trouvera des techniques intermédiaires, comme celles qui seront présentés dans la prochaine section et dont certaines ont été utilisées au cours des trois prochains chapitres.

4 LES MÉTHODES INTERMÉDIAIRES

Si l'utilisation des méthodes utilisant le plus possible de données génériques est plus aisée, il faut noter que, par définition, elles ne seront pas capables de produire des structures possédant un repliement très différent des structures déjà connues. Dans ce cas, l'utilisation d'un faible nombre de données spécifiques permet de mieux « guider » la modélisation. L'émergence des techniques expérimentales à basse résolution permet d'obtenir des données spécifiques rapidement, même dans des cas où les techniques à

haute résolution, telles que la RMN et la cristallographie aux rayons X, ne peuvent s'appliquer. Néanmoins, cela se fait au dépend d'une information ambigüe qui à elle seule ne suffit pas à déterminer la structure détaillée de la protéine. L'intérêt de l'ajout de ces techniques dans des programmes de modélisation est aussi important car il peut permettre d'améliorer la fonction de score pour qu'elle permette de mieux discriminer parmi les structures produites. Parmi ces méthodes qualifiées de basse résolution les plus utilisées sont la diffusion des rayons X ou des neutrons aux petits angles en solution (SAXS/SANS), des données issues de la RMN (couplages dipolaires résiduels et effets NOE) ainsi que la microscopie électronique. Comme ces données seules n'apportent pas assez de contraintes pour produire des modèles détaillés des protéines, de nombreuses techniques sont apparues depuis les années 2000 pour les combiner. Un travail récent (Das et al. 2009) confirme l'intérêt de l'ajout de données expérimentales pour la discrimination des structures en sortie du programme. Parmi les techniques expérimentales, on peut noter un intérêt particulier pour la combinaison SAXS et les couplages dipolaires résiduels obtenus par RMN qui sera détaillée ici.

Le SAXS et les couplages dipolaires résiduels ont connu de grand progrès ces dernières années (Prestegard et al. 2000; Bax 2003; Blackledge 2005; Svergun & Koch 2002; Koch et al. 2003; Neylon 2008). Indirectement (le processus est détaillé dans le chapitre 1), le SAXS permet d'avoir plusieurs informations sur la géométrie de la protéine comme le rayon de giration ou encore les distances inter-domaines. Comme les contraintes d'acquisitions sont relativement faibles, le SAXS est rapidement devenu un outil de choix pour l'analyse de la forme globale des macromolécules biologiques. Cependant, dans le cas d'une protéine à plusieurs domaines, si on a une idée de la forme globale de la protéine il est rarement possible de déterminer l'orientation relative des domaines. L'information apportée par les couplages dipolaires résiduels permet de pallier à ce défaut, ce qui en fait deux techniques parfaitement complémentaires. Avant la sortie de la première version du programme Dadimodo spécialement développé au laboratoire, la combinaison de ces techniques a été utilisée dans la résolution de la structure de protéines à plusieurs domaines (Mattinen et al. 2002; Yuzawa 2004; Bernado 2005; Grishaev et al. 2005; Gabel et al. 2006) et d'un enzyme (Grishaev et al. 2007). À cette période, aucun programme ou protocole n'était disponible à la communauté pour pouvoir exploiter simultanément ces deux données expérimentales. A l'heure actuelle, il existe

une version du programme de modélisation par homologie MODELLER capable de gérer ces deux types de données expérimentales comme contraintes sur les structures (Förster et al. 2008) et un protocole utilisable avec le programme de modélisation moléculaire CNS (Grishaev et al. 2007). Cette combinaison a même été très récemment utilisée pour la détermination de complexes de protéines (Madl et al. 2010).

Durant cette thèse, différentes méthodes utilisant une combinaison de données spécifiques et génériques ont été utilisées pour déterminer des modèles de protéines. Dans le chapitre 1, je présente l'application d'un programme de détermination de structure de protéines à plusieurs domaines développé au laboratoire. Cette application a pour but de déterminer des modèles d'une protéine de grande taille composée de deux domaines pour laquelle on ne dispose que d'une donnée expérimentale faiblement résolue (courbe de diffusion des rayons X en solution) et a nécessité plusieurs développements spécifiques du programme. Le chapitre 2 illustre la difficulté d'obtenir suffisamment de données expérimentales pour déterminer de manière classique la structure d'une protéine, même si celle-ci paraît être un cas relativement favorable. Je présente alors ce qu'il est possible de faire lorsque l'on ne dispose que d'un ensemble partiel de données expérimentales obtenues par RMN. Le dernier chapitre aborde le problème de l'assemblage de complexes de protéines, en particulier avec les acides nucléiques. Le but est d'isoler une donnée issue des expériences RMN capable de mieux guider les programmes d'assemblage moléculaire en indiquant non seulement les résidus interagissant dans l'assemblage (information qualitative) mais aussi dans quelle mesure ils sont impliqués (information quantitative).

1 INTRODUCTION

Dans ce chapitre, sera présentée la détermination de la structure de la protéine Penicillin Binding Protein 1b (PBP1b) à l'aide de données de diffusion des rayons X aux petits angles (Small Angle X-ray Scattering ou SAXS) et de modèles de chacun des domaines. Les modèles obtenus par le programme spécifiquement développé au laboratoire seront comparés à ceux obtenus par un logiciel couramment utilisé par la communauté.

1.1 LE SAXS, UN OUTIL DE CHOIX POUR LA BIOLOGIE STRUCTURALE

La diffusion des rayons X aux petits angles est une méthode puissante pour l'analyse de macromolécules biologiques en solution. L'intérêt de l'utilisation de cette technique réside dans sa relative simplicité, à la fois pour l'acquisition des données et pour la caractérisation de l'échantillon. En terme de préparation de l'échantillon expérimental, les pré-requis (concentration, solvant, etc…) sont similaires à la résonnance magnétique nucléaire (RMN) ou à la radio-cristallographie aux rayons X, sans nécessiter la formation de cristaux. Les solutions de macromolécules étant isotropes du fait de l'agitation thermique (mouvement brownien), seule la moyenne sphérique des intensités diffusées est accessible expérimentalement. Les données expérimentales sont représentées sous la forme d'une courbe représentant l'évolution de l'intensité de diffusion I(Q) en fonction du transfert de moment $Q=4\pi\sin\theta/\lambda$ $Å^{-1}$ où 2θ est l'angle de diffusion et λ (Å) la longueur d'onde des rayons-X. Comme la fonction I(Q) est obtenue par transformée de Fourier de p(r), la fonction de distribution des distances au sein de l'objet, contraindre I(Q) revient à contraindre p(r).

Des méthodes de détermination de modèles tridimensionnels à basse résolution n'utilisant que les données SAXS ont été développées pour déterminer la forme globale des macromolécules. Ces méthodes sont basées sur la comparaison de la courbe de

diffusion calculée à partir de la forme géométrique à la courbe de diffusion expérimentale. Le programme le plus communément utilisé est DAMMIN (D.I. Svergun 1999) qui combine une modélisation par un ensemble de pseudo-atomes sur un réseau tridimensionnel avec un recuit simulé (voir encart ci-dessous).

DAMMIN

DAMMIN est un programme couramment utilisé pour la détermination de la forme globale des protéines à partir de leur courbe de diffusion obtenue en solution.

Principe du programme :

Le programme commence par générer une sphère dont le diamètre est déterminé par la valeur du D_{max} pour s'assurer que la protéine soit bien contenue dans ce volume. Cette sphère est ensuite remplie de pseudo-atomes dont le rayon $r_0 << D_{max}$ est dépendant de la taille de la sphère et de leur nombre (fixé par l'utilisateur). Chaque pseudo-atome se voit ensuite attribuer un index en fonction indiquant la phase dans laquelle il se trouve (0 pour le solvant, 1 pour la protéine), déterminant ainsi le modèle (Figure 3).

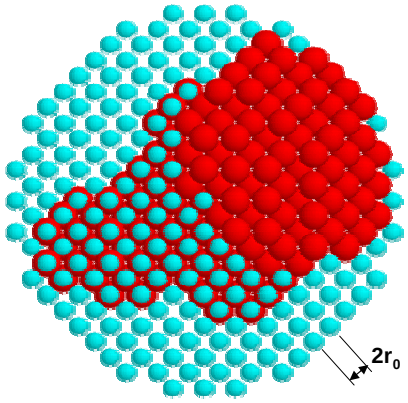


Figure 3 : Représentation de la sphère générée par dammin et de ses pseudo-atomes. Les pseudo-atomes cyan correspondent aux pseudo-atomes se trouvant dans le solvant, les rouges ceux se trouvant dans la protéine. L'espace séparant le centre de deux pseudo-atomes voisins est de $2r_0$.

La courbe de diffusion correspondante est alors calculée à l'aide de CRYSOL puis comparée à la courbe expérimentale par l'opération suivante :

$$\chi^2\left[I(S_j), I_{calc}(S_j)\right] = \frac{1}{N-1}\sum_j \left[\frac{I(S_j) - I_{calc}(S_j)}{\sigma(S_j)}\right]^2$$

où N est le nombre de points sur la courbe expérimentale, $\sigma(S_j)$ l'écart type au j-ième point.

De cette façon, en utilisant une méthode de recuit simulé, le programme cherche une configuration de pseudo-atomes compacte X minimisant la fonction :

$$f(X) = \chi^2\left[I(S_j), I_{calc}(S_j)\right] + \alpha P(X)$$

où P(X) est la fonction de pénalité permettant de s'assurer de la compacité et de la connectivité du modèle et α son poids associé.

Si des modèles à haute-résolution des domaines sont disponibles, le SAXS peut aussi être utilisé pour étudier des complexes de macromolécules comme par exemple des protéines à plusieurs domaines. Des techniques de modélisation en corps rigides qui utilisent les données SAXS existent, comme le programme SASREF (Petoukhov & Svergun 2005) qui utilise CRYSOL (Svergun et al. 1995) pour le calcul des courbes de diffusion à partir des structures. La recherche du meilleur accord entre la courbe calculée et la courbe expérimentale se fait en effectuant des rotations et des translations des sous-unités. Ce programme permet d'obtenir rapidement des complexes approchant de façon correcte les données expérimentales. Cependant, plusieurs problèmes se posent : la nécessité de disposer de modèles fiables du/des domaine(s) de départ, l'exploration de l'espace des conformations qui peut ne pas être suffisante et la validité des structures du point de vue biochimique. Le manque d'exploration de l'espace des conformations est dû au fonctionnement du programme car il n'est pas capable de générer des structures avec des domaines dont l'orientation a été fortement modifiée. De plus, il est nécessaire d'affiner les structures dans un autre programme, CNS par exemple, pour reconstruire les peptides de liaisons et minimiser les chaines latérales.

Dans le but de résoudre ces problèmes, un programme a été développé au laboratoire. Dadimodo, pour DAta DrIven MODular Orientation, est un programme d'assemblage de structures de domaines protéiques conçu à l'origine pour utiliser conjointement deux types de données expérimentales : des couplages dipolaires résiduels, obtenus par des mesures de RMN réalisées en milieu faiblement orientant, et des courbes de diffusion des rayons X en solution (SAXS). L'idée générale du programme est de modifier la position et

l'orientation relative des domaines constituant une protéine pour que sa forme globale et l'orientation des domaines soient compatibles avec les données expérimentales. Initialement développé par Fabien Mareuil au cours de sa thèse pour l'étude de la structure de fragments d'une grande protéine (S1) composée de plusieurs domaines (Mareuil et al. 2007), Dadimodo a connu depuis une série d'optimisations qui seront abordées plus loin dans ce chapitre.

1.2 La protéine Penicillin Binding Protein 1b (PBP1b)

1.2.1 Rôle biologique

La bactérie *Streptoccus pneumoniae* est responsable de nombreuses infections telles que la pneumonie, la méningite et l'otite aigüe chez les nourrissons et les personnes âgées. À partir des années 50, ces infections ont été traitées par des antibiotiques tels que la pénicilline ou les céphalosporines mais, depuis la fin des années 1980, des germes résistants sont apparus.

Dans le cadre de la recherche de nouvelles cibles thérapeutiques pour la conception d'un nouvel antibiotique, les études se sont portées sur les penicillin-binding proteins (PBP) qui sont les cibles des antibiotiques précédemment cités. Les PBP sont impliquées dans la biosynthèse de la paroi bactérienne (Goffin & Ghuysen 1998) avec deux activités catalytiques menées par deux domaines différents. La première activité est la polymérisation des sucres par le domaine glycosyltransférase et la deuxième activité est la liaison entre les chaînes de sucres par des liaisons peptidiques réalisée par le domaine transpeptidase. Face aux nouvelles résistances développées par les bactéries, de nombreux travaux visent à l'obtention d'une molécule provoquant l'inhibition de l'une ou l'autre des fonctions de cette protéine, conduisant à la mort de la bactérie. L'objectif étant, à terme, de remplacer les pénicillines qui ne sont plus efficaces dans ce rôle d'inhibiteur

1.2.2 Détermination de modèles de PBP1b

La structure de la protéine PBP1b, dont la séquence est présentée dans la Figure 4, est composée de deux domaines séparés par deux peptides de liaisons et d'une extrémité N-terminale correspondant à un morceau de l'ancre se fixant dans la membrane lipidique. La structure du domaine le plus grand, le domaine transpeptidase) en bleu sur la Figure 4, a été déterminée par radio-cristallographie. La structure du deuxième domaine, le domaine glycosyltransférase, n'a pas encore été déterminée. Cependant, comme il existe pour ce domaine Glycosyltransférase deux structures homologues dans la PDB, il a été possible d'utiliser le programme MODELLER en utilisant ces structures pour réaliser un premier modèle complet de la molécule. Les structures utilisées sont indiquées dans le Tableau 1. PBP1b* correspond une structure tronquée de PBP1b, contenant un peptide composé de 25 résidus du domaine Glycosyltransférase, un peptide de liaison inter domaines et le domaine Transpeptidase entier. Le modèle proposé par MODELLER servira ensuite de modèle de départ pour les programmes de détermination de structures Dadimodo et SASREF.

Tableau 1 : Structures utilisées pour la génération de modèles par MODELLER. Le chiffre entre parenthèses correspond au pourcentage d'identité avec le domaine équivalent de la protéine PBP1b.

Nom	Type domaine Glycosyltransférase (code PDB)	Type domaine Transpeptidase (code PDB)
PBP1a	PBP1a (2oqo 30%)	-
PBP1b*	-	PBP1b (2xd1)
Structure chimérique	PBP2 (2olu 31%)	PBP1b (2xd1)

```
                  70        80        90       100       110
-----------GSGIALGYGVALFDKVRVPQTEELVNQVKDISSISEITYSDGTVIASIE

                 130       140       150       160       170
SDMLRTSISSEQISENLKKAIIATEDEHFKEHKGVVPNAVIPATLGKFVGLGSSSGGSTL

                 190       200       210       220       230
TQQLIKQQVVGDAPTLARKAAEIVDALALERAMNKDEILTTYLNVAPFGRNNKGQNIAGA

                 250       260       270       280       290
RQAAEGIFGVDASQLTVPQAAFLAGLPQSPITYSPYENTGELKSDEDLEIGLRRAKAVLY

                 310       320       330       340       350
SMYRTGALSKDEYSQYKDYDLKQDFLPSGTVTGISQDYLYFTTLAEAQERMYDYLAQRDN

                 370       380       390       400       410
VSAKELKNEATQKFYRDLAAKEIENGGYKITTTIDQKIHSAMQSAVADYGYLLDDGTGRV

                 430       440       450       460       470
EVGNVLMDNQTGAILGFVGGRNYQENQNNHAFDTKRSPASTTKPLLAYGIAIDQGLMGSE

                 490       500       510       520       530
TILSNYPTNFANGNPIMYANSKGTGMMTLGEALNYSWNIPAYWTYRMLRENGVDVKGYME

                 550       560       570       580       590
KMGYEIPEYGIESLPMGGGIEVTVAQHTNGYQTLANNGVYHQKHVISKIEAADGRVVYEY

                 610       620       630       640       650
QDKPVQVYSKATATIMQGLLREVLSSRVTTTFKSNLTSLNPTLANADWIGKTGTTNQDEN

                 670       680       690       700       710
MWLMLSTPRLTLGGWIGHDDNHSLSQQAGYSNNSNYMAHLVNAIQQASPSIWGNERFALD

                 730       740       750       760       770
PSVVKSEVLKSTGQKPGKVSVEGKEVEVTGSTVTSYWANKSGAPATSYRFAIGGSDADYQ

NAWSSIVGSLP
```

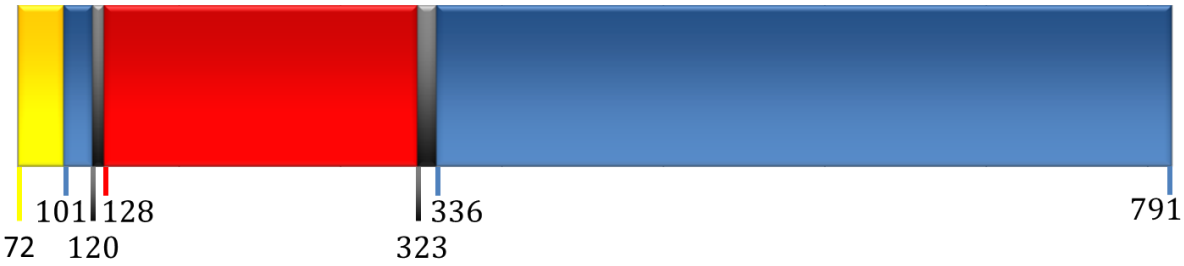


Figure 4 Haut : Séquence de PBP1b, en jaune l'extrémité N-terminale, en bleu plein le domaine Transpeptidase, en bleu souligné le premier peptide de liaison séparant les domaines Transpeptidase et Glycosyltransférase, en rouge plein le domaine Glycosyltransférase et en rouge souligné le deuxième peptide de liaison. Bas : Schéma de la séquence.

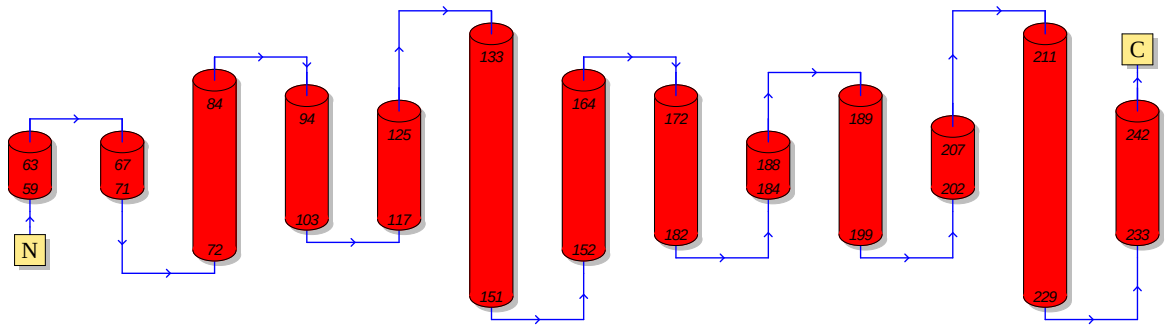


Figure 5 : Schéma de topologie du domaine Glycosyltransférase de la structure 2oqo. Elle présente 30% d'identité avec la séquence de ce même domaine de PBP1b.

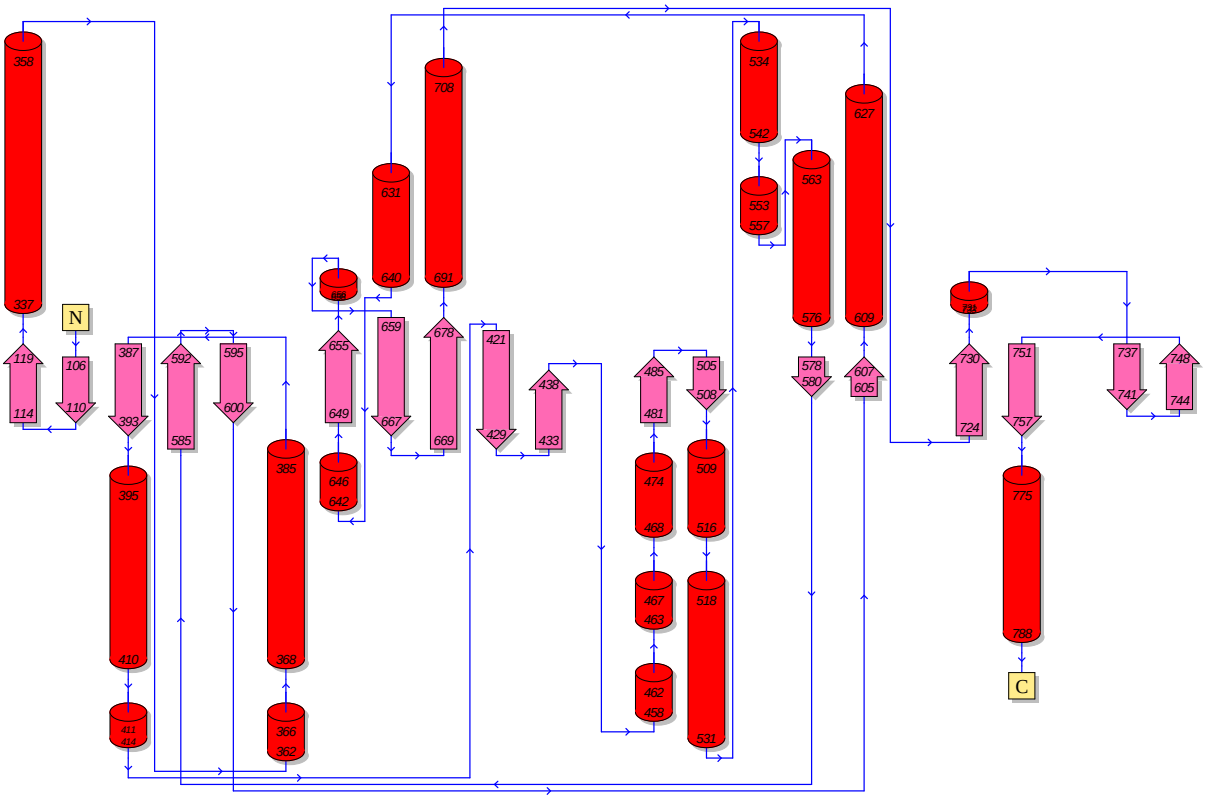


Figure 6 : Schéma de topologie du domaine Transpeptidase de la structure 2xd1 qui correspond au domaine Transpeptidase isolé de PBP1b.

2 PRÉSENTATION DU PROGRAMME DADIMODO

2.1 ORIGINE DU PROGRAMME

À l'origine de Dadimodo, il y avait deux réflexions. La première est que la biologie structurale est de plus en plus souvent confrontée à l'étude d'objets complexes comme les protéines formées de plusieurs domaines (c'était le cas de la protéine S1 sur laquelle le laboratoire travaillait à l'époque) ou les complexes formés de plusieurs macromolécules, en particulier des assemblages de protéines et éventuellement d'acides nucléiques. La seconde réflexion est que s'il semblait assez illusoire que l'on puisse déterminer expérimentalement la structure de tous les objets complexes biologiquement pertinents on peut espérer disposer, dans un avenir proche, d'un nombre suffisant de structures de domaines pour pouvoir calculer celles de tous les autres par homologie. Dans ces conditions le laboratoire, confronté à un problème de cette nature avec l'étude de la structure de la protéine S1, a cherché à mettre au point une technique permettant de reconstruire l'architecture de protéines modulaires (formées de plusieurs domaines) à partir de modèles par homologie (c'est à dire approximatifs, donc devant pouvoir être déformés) des domaines et utilisant des données expérimentales « simples », à savoir des courbes de SAXS, des mesures de couplages dipolaires résiduels et un peu plus tard des données de cartographie de déplacement chimique. Cette technique a été développée en collaboration avec Javier Perez au synchrotron Soleil. Ces travaux ont conduit à la mise au point d'une version initiale de Dadimodo (Mareuil et al, 2007) puis d'une première version distribuable du programme (disponible sur le site du synchrotron) après un travail d'optimisation que j'ai effectué à mon arrivée au laboratoire. Cette version a servi de base pour réaliser les développements qui seront exposés dans ce chapitre. Le développement du programme a maintenant été repris par Guillaume Evrard et Javier Perez qui se chargent de la réécriture complète du programme de façon à le rendre plus performant et convivial.

L'utilisation combinée de ces trois types de données expérimentales n'est pas le fruit du hasard, le SAXS permet d'obtenir une information sur la distribution des distances au sein de la protéine alors que les RDC apportent une information sur l'orientation des

structures et la cartographie de déplacement chimique permet de caractériser les interfaces entre les domaines (cette dernière mesure nécessitant, cependant, plus de travail expérimental, dans la mesure où elle impose, dans le cas d'une protéine à domaine, de disposer d'enregistrer des spectres sur la protéine et sur les domaines séparés).

Dès son origine, Dadimodo a été conçu pour être un programme simple, facilement adaptable à différents types de protéines et de données expérimentales. C'est dans cette optique que certains choix techniques ont été faits : un algorithme génétique pour le corps principal, l'utilisation de XPLOR pour la gestion des structures et l'utilisation de programmes externes pour gérer les données expérimentales. Ayant repris le développement du logiciel, j'ai dans un premier temps intégré plusieurs optimisations d'ordre technique : la réécriture du code de l'algorithme génétique en Python puis l'utilisation de CNS pour remplacer XPLOR pour pouvoir gérer des protéines de grande taille. Les autres optimisations, spécifiques à la protéine étudiée, seront abordées plus loin dans ce chapitre.

2.2 Présentation générale des algorithmes génétiques

Les algorithmes génétiques sont des algorithmes de type stochastique tirant leur nom du principe de sélection naturelle dont ils s'inspirent pour déterminer des solutions satisfaisant un ensemble de conditions données. La théorie de la sélection naturelle permet d'expliquer comment les organismes évoluent face à leur environnement. En résumant, les traits (phénotypes) favorisant la survie et la reproduction sont sélectionnés, ce qui entraîne l'augmentation, d'une génération à l'autre, des génotypes associés dans la population. Mais cette « contamination » de la population par un (ou un ensemble de) génotype donné est en permanence contrebalancée par l'introduction de nouvelles mutations dans le génome, ce qui permet de maintenir une diversité dans la population. Il est possible d'assimiler ce mécanisme à la recherche de l'adaptation optimale face à un ensemble de conditions environnementales.

Introduit dans les années 70 (Holland 1992), les algorithmes génétiques ont été adaptés dans le domaine de la biologie structurale dans les années 90 pour la prédiction de la structure de protéines et d'ARN (Shapiro & Navetta 1994; Gultyaev et al. 1995). Depuis, ces d'algorithmes sont devenus un outil de choix pour résoudre des problèmes

d'optimisation en relation avec la structure des protéines. S'ils ne donnent pas l'assurance de trouver la solution optimale, ils permettent en général de trouver des solutions approchées qui vont satisfaire de manière correcte l'ensemble des conditions que l'on aura fixé au départ, même lorsque le problème est très complexe (en particulier lorsque les données sont fortement dégénérées).

L'idée de base des algorithmes génétiques est de réaliser, sur une population d'individus représentant un ensemble de solutions potentielles à un problème, une série de cycle de mutations consistant en la perturbation d'un paramètre, et de sélection en fonction de l'accord de la solution potentielle avec les données du problème. Les processus de mutation et de sélection sont ajustés finement, de façon à ce que la population s'enrichisse en individus satisfaisant de mieux en mieux les données du problème tout en gardant le plus longtemps possible une grande variabilité. Le schéma classique d'une génération d'un algorithme génétique est présenté en Figure 7.

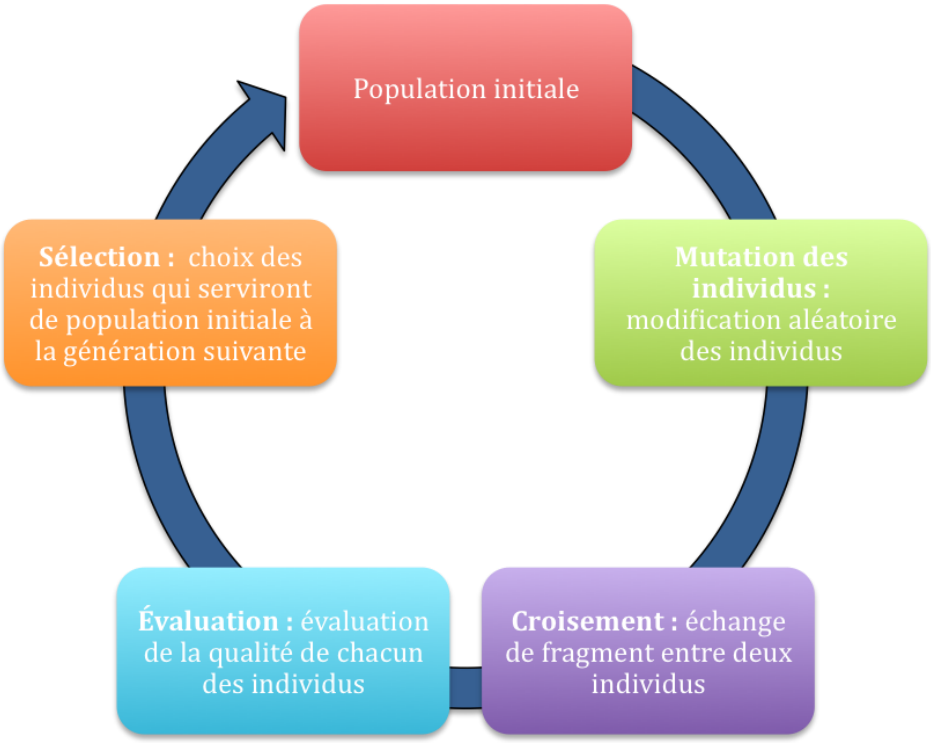


Figure 7 : Schéma d'une génération d'un algorithme génétique. Dans la première étape les individus de la population initiale sont répliqués (amplification) puis mutés aléatoirement. L'étape de croisement permet d'échanger des morceaux entre deux individus tirés au hasard. L'étape d'évaluation permet d'établir l'accord des individus avec les contraintes, la qualité de cet accord étant utile pour l'étape de sélection. Cette dernière étape sélectionne les individus qui constitueront la population initiale de la génération i+1

2.3 Mécanique de l'algorithme génétique dans Dadimodo

2.3.1 Mutation des individus

L'étape de mutation consiste à modifier les individus, mais c'est aussi une étape d'amplification de la population : chaque individu est répliqué plusieurs fois et chaque copie est ensuite modifiée. L'objectif de Dadimodo étant de positionner les différents domaines d'une protéine les uns par rapport aux autres, il aurait été possible de considérer chaque domaine comme un corps rigide et de définir les mutations comme des variations élémentaires de leurs positions relatives. Mais, nous voulions aussi être capables de déformer la structure interne des domaines. Le choix a donc été fait de représenter toute la protéine sous la forme d'une chaîne continue et de définir les mutations à partir de variations des angles phi et psi du squelette. Par ailleurs, comme nous souhaitions pouvoir utiliser des données expérimentales issues de la RMN, il était indispensable de représenter tous les atomes de façon explicite.

Avec cette méthodologie, il était nécessaire de distinguer deux types de modifications. Celles-ci sont, d'une part, des modifications « globales » qui changent la disposition relative des deux régions de la chaine situées en amont et en aval du point de modification et, d'autre part, des modifications locales qui changent la géométrie d'un fragment plus ou moins long au sein de la chaine sans affecter les parties de part et d'autre de ce fragment. Typiquement, les modifications globales peuvent servir à déplacer un domaine par rapport à un autre tandis que les modifications locales peuvent servir à modifier la conformation d'une boucle au sein d'un domaine sans perturber celui-ci, ni son positionnement relatif dans l'ensemble de la protéine.

Ces modifications effectuées sur la structure sont appelées « mutations » et correspondent donc à une modification de la conformation de la protéine en affectant la valeur des angles Ψ et Φ d'un résidu choisi au hasard. Le choix du type de mutation dépend de la valeur de l'identifiant de région que l'on aura préalablement placé à la place du facteur B dans le fichier PDB. Ce fonctionnement permet d'une part de choisir quels résidus vont être impactés par les mutations mais aussi de choisir quel type de mutation

sera possible en fonction de la région de la protéine (boucle, structure secondaire, peptides de liaison, etc.). L'ampleur de la mutation est déterminée par les formules de l'Équation 1. Les angles Φ_n et Ψ_n sont les nouvelles valeurs, les angles Φ_a et Ψ_a les anciennes. $\Delta\Psi$ et $\Delta\Phi$ sont des variables aléatoires définies par une gaussienne centrée en 0 et dont l'écart type a été défini à 5,36 degrés multiplié par un coefficient décrivant l'accord local entre les données expérimentales et la structure. Ce coefficient n'est utilisé que si les données propres aux RDC sont présentes, sinon il est égal à 1. Cette valeur de 5,36 degrés a été définie pour que la perturbation des angles soit suffisante pour générer des conformations véritablement différentes sans être trop forte pour éviter de générer un trop grand nombre de conformations incorrectes d'un point de vue énergétique.

$$\Psi_n = \Psi_a + \Delta\Psi$$
$$\Phi_n = \Phi_a + \Delta\Phi$$

Équation 1 : Formules de modification des angles Φ et Ψ du résidu modifié

2.3.1.1 Mutation globale

Pour modifier le positionnement relatif des domaines, on modifie la valeur des angles Φ et ψ d'un résidu situé dans le peptide de liaison. Cette mutation affecte toute la structure en aval du point de mutation. Une fois la modification effectuée, la conformation des chaînes latérales fait l'objet de plusieurs cycles de minimisation d'énergie en bloquant les atomes du squelette.

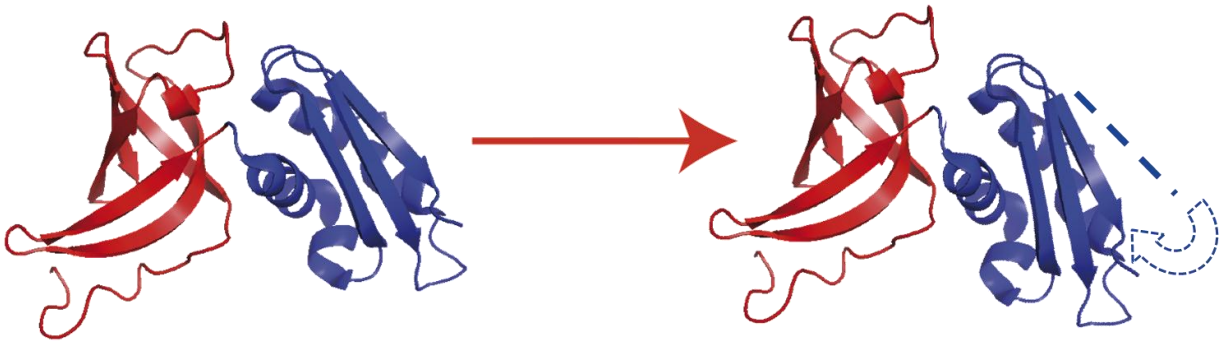


Figure 8 : Effet d'une mutation globale : les angles phi et psi de l'acide aminé à la jonction des deux domaines de la protéine NusA (en rouge le domaine S1, en bleu le domaine KH) ont été modifiés. Le résultat est un déplacement global du domaine KH par rapport au domaine S1.

2.3.1.2 Mutation locale

Pour modifier localement la structure, les angles Φ et ψ du résidu choisi sont modifiés de la même manière, mais la perturbation n'est propagée que sur quelques résidus en aval, la structure restant inchangée au-delà. La connectivité est rétablie par une minimisation de l'énergie de toute la région perturbée.

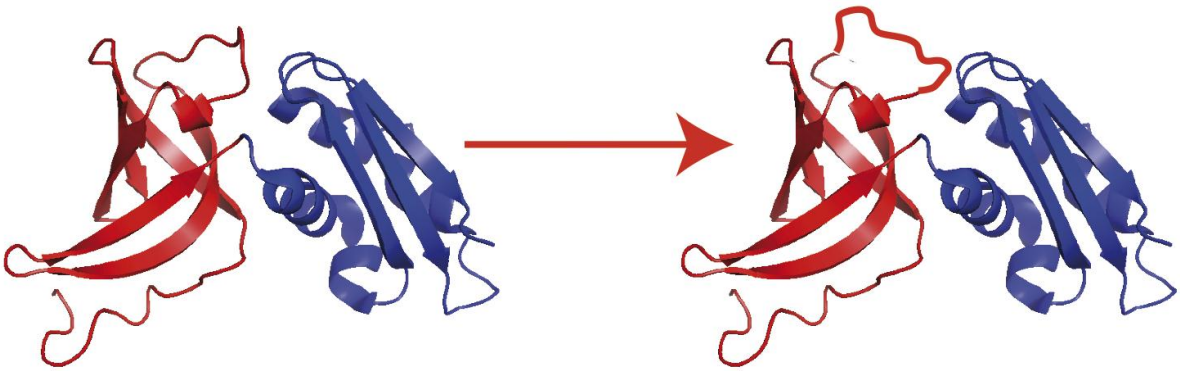


Figure 9 : Représentation de l'effet d'une mutation locale au niveau d'une boucle du domaine S1 : les angles phi et psi du premier résidu de la boucle ont été modifiés aléatoirement, la connectivité a été rétablie à l'autre extrémité de la boucle par une minimisation d'énergie appliquée à toute la boucle. Cette minimisation a pour effet de rétablir l'intégrité de la structure est d'adapter la boucle modifiée au contexte du reste de la protéine.

Après chaque mutation, l'énergie des chaines latérales de la protéine est minimisée. L'énergie de la protéine est évaluée. La structure n'est conservée que si cette valeur est inférieure à un seuil défini par l'utilisateur, de façon à s'assurer que toutes les structures conservées sont vraisemblables du point de vue de leur géométrie covalente.

2.3.1.3 Remarques

L'avantage de ce système est qu'il est simple à mettre en œuvre. De plus, comme CNS est un programme couramment utilisé par la communauté, les scripts peuvent être modifiés par les utilisateurs. Le principal désavantage est le temps nécessaire à CNS pour réaliser ces opérations, en partie causée par la nécessité de passer par des

lectures/écritures de fichiers à chaque opération. Avec sa configuration actuelle, le programme génère à cette étape 125 structures, ce qui implique un temps de calcul est assez élevé.

Comme indiqué précédemment, Dadimodo a été conçu pour être facilement adaptable à différentes protéines. Le cas d'étude présenté ici, la protéine PBP1b, a nécessité un certain nombre d'adaptations spécifiques. En effet, la structure de cette protéine est particulière car composée de deux domaines reliés par deux peptides de liaisons (Figure 4 Bas). Du fait de la présence de ce double peptide de liaison, il n'est pas possible de se contenter de la mutation globale pour changer le positionnement relatif des domaines. Après une mutation globale, en effet, la structure du domaine Transpeptidase sera altérée à cause du déplacement du feuillet β à l'interface des deux domaines (les deux feuillets juste après le N-terminal sur la Figure 6). La superposition de deux structures similaires ayant subi des mutations globales du peptide de liaison inter domaine (en rouge) et une structure de départ (en bleu) illustre ce déplacement (Figure 10). Ce feuillet appartenant au domaine Transpeptidase, sa position doit rester fixe. Il a donc été nécessaire d'introduire un nouveau type de mutation, qui sera présenté dans la section 3.1.

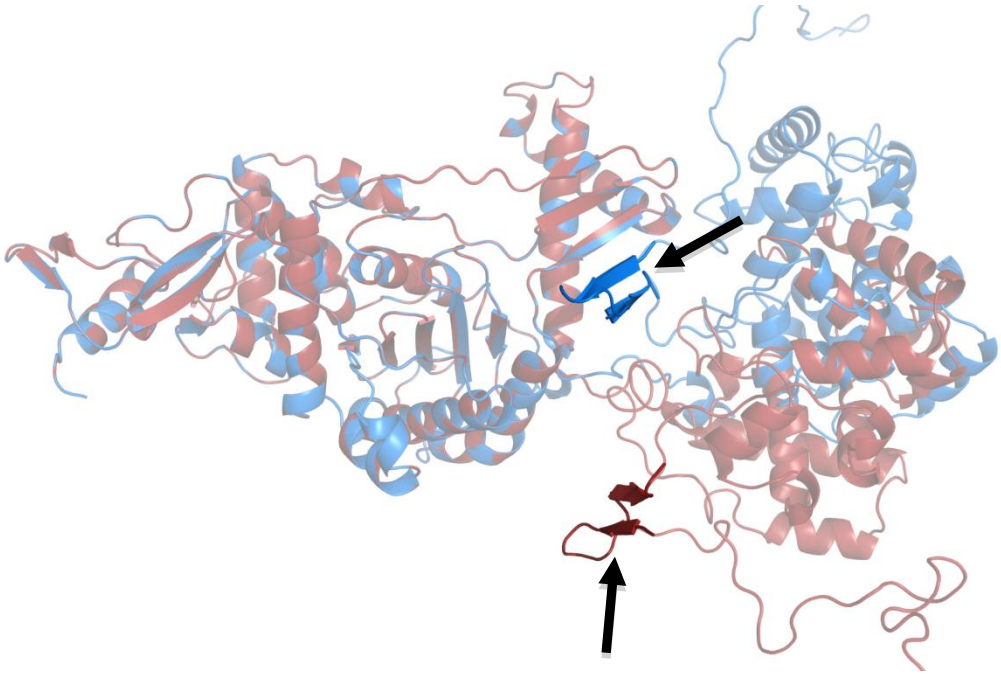


Figure 10 : Superposition d'une structure de départ (en bleu) et d'une structure ayant subi des mutations globales au niveau du peptide de liaison inter-domaines (en rouge). Les flèches indiquent les feuillets β du domaine Transpeptidase à l'interface avec le domaine Glycosyltransférase. Ce feuillet fait partie du domaine Transpeptidase et sa position ne doit pas être modifiée

2.3.2 Évaluation des individus

Après l'étape de modification/réplication des individus, vient la phase d'évaluation. Il s'agit ici de déterminer la qualité des individus par rapport à un ensemble de conditions définies. Dans le cas présenté ici, il s'agit de déterminer l'accord des structures avec les données de diffusion des rayons X en solution.

L'évaluation est réalisée en comparant la courbe expérimentale à la courbe calculée pour chaque structure par le programme CRYSOL. Cette comparaison est quantifiée en calculant le χ^2 entre les deux courbes. Il serait possible de réaliser cette opération en utilisant aussi CRYSOL, mais nous avons préféré utiliser notre propre routine de calcul car au cours du développement du programme nous avons été conduit à réaliser des tests sur des courbes « parfaites » (sans incertitude expérimentale), or CRYSOL ne sait pas faire le calcul dans ces conditions. Par ailleurs, le calcul de l'accord entre deux courbes par CRYSOL pose un certain nombre de problèmes qui seront abordés dans la section 3.2.

2.3.3 Sélection des individus d'une génération à l'autre

Après l'étape de mutation, la population est passée de 25 à 150 individus, 125 structures étant engendrées à cette étape. Il est donc nécessaire « d'élaguer » cette population pour n'en garder que 25, en conservant une bonne diversité et en améliorant globalement la qualité des individus. Dans cette optique, c'est la méthode du tournoi (David E. Goldberg 1991) qui a été utilisée. Le principe est représenté par la Figure 11.

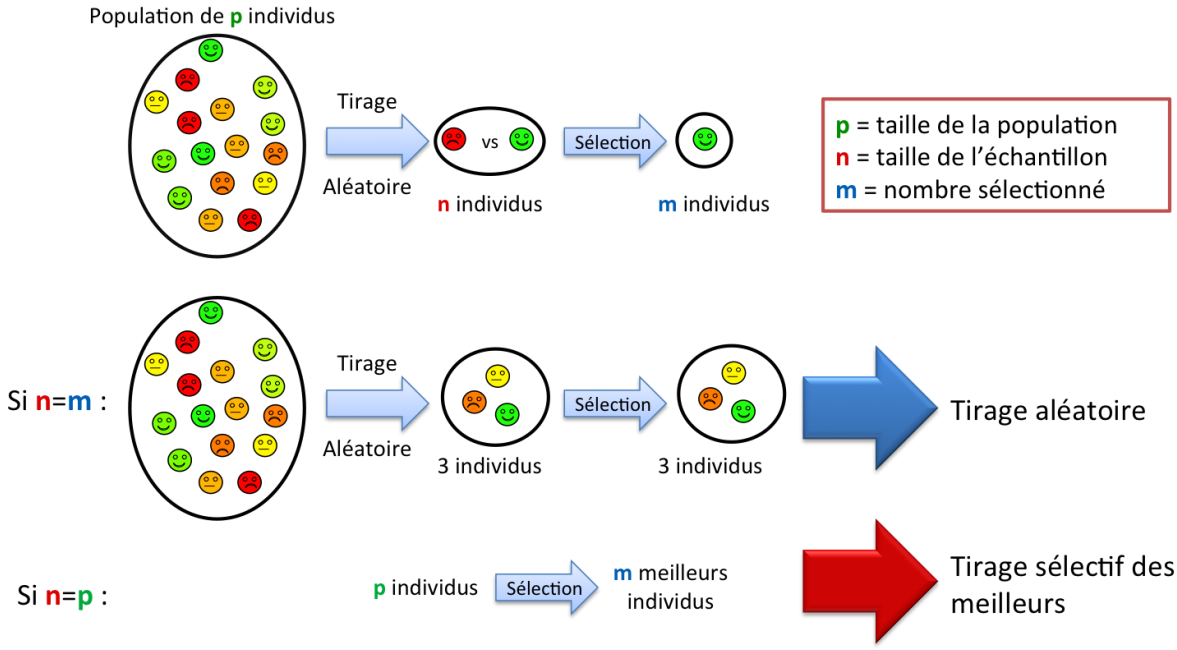


Figure 11 : Fonctionnement d'une sélection par méthode de tournoi. Un sous ensemble de n individus est tiré au sort au sein de la population initiale composée de p individus. Les n individus sont classés en fonction de leur accord aux données expérimentales et les m meilleurs sont conservés pour le tour suivant. On peut faire varier la pression de sélection en jouant sur p et m. Dans le premier cas présenté (seconde ligne), n=m. Dans ces conditions, tous les individus sélectionnés lors de la première phase sont retenus. Le processus s'apparente à un pur tirage aléatoire et la pression de sélection aléatoire est nulle. Dans le second cas (troisième ligne), n=p. Dans ces conditions, le classement des meilleurs se fait au sein de la population initiale complète : seuls les meilleurs individus sont transmis à la génération suivante et la pression de sélection est maximale.

Dans la version développée par F. Mareuil, le programme fonctionnait en utilisant deux types de tirages : un premier semi-aléatoire (*m* très proche de *n*) pendant 200 générations, puis un deuxième très sélectif (*n*=*p*) pendant 100 générations. Les courbes de la Figure 12 montrent l'évolution de la qualité des individus au cours du processus. Elle mettent en évidence une très grande variabilité des structures et une faible amélioration du score jusqu'à la génération 200, puis une nette amélioration du score et une forte baisse de la variabilité dès les premiers cycles suivant l'augmentation de la pression de sélection. Durant ces cycles, le score n'évolue que très peu et, pour le cycle en forte pression de sélection, il est probable que le système est trop contraint à ce moment-là pour sortir du minimum dans lequel il se retrouve piégé. Pour régler ce problème, un système de sélection optimisé a été développé, il sera abordé dans la section 3.3.

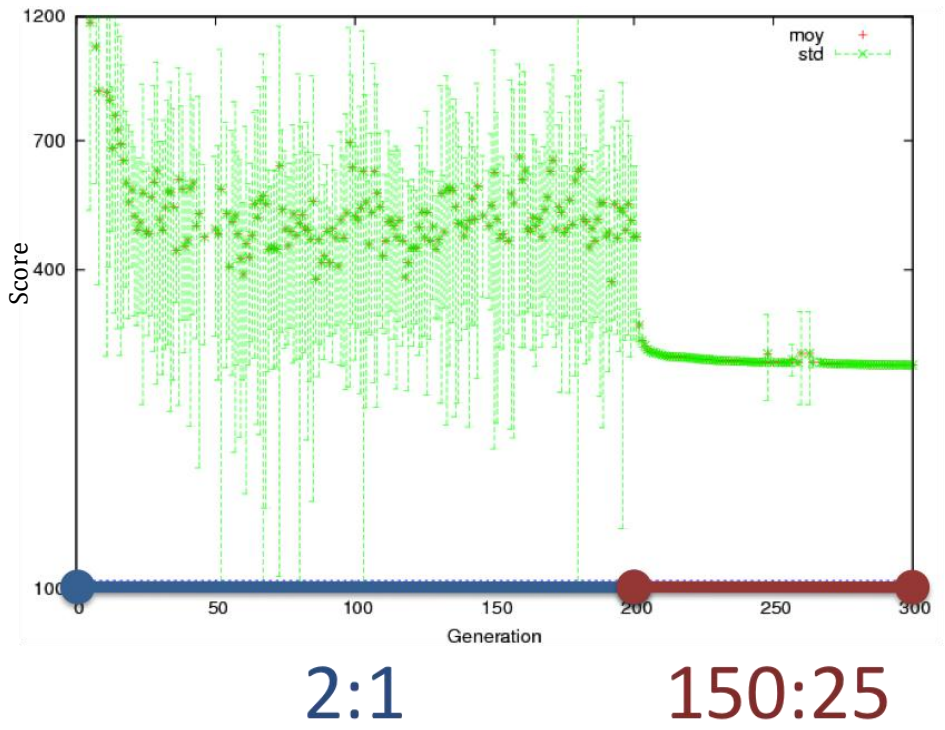


Figure 12 : Évolution du score moyen d'une famille de 25 structure au cours d'un test réalisé avec Dadimodo sur la protéine NusA en présence de données de SAXS et de mesures de couplages dipolaires résiduels (exprimé en une unité arbitraire, correspondant à une combinaison linéaire des écarts mesurées vis-à-vis des deux grandeurs expérimentales). Les points verts représentent le score moyen, les barres représentent l'enveloppe de toutes les valeurs de la population. L'abscisse correspond au nombre de génération. Le calcul a été réalisé en appliquant deux pressions de sélection très différentes. Lors des 200 premières générations, la pression est très faible (p=150, n=2, m=1). La variabilité au sein de la population reste grande, mais après une première phase d'optimisation (lors de 25 premières générations) la qualité de la population n'évolue pratiquement plus. Lors des 100 générations suivantes, la pression de sélection est maximale (p=150, n=150, m=25). La population s'améliore fortement, mais au prix d'une perte de variabilité qui interdit toute évolution ultérieure.

3 OPTIMISATION ET DÉVELOPPEMENTS SPÉCIFIQUES POUR PBP1B

L'application de Dadimodo à PBP1b a demandé la résolution de trois problèmes distincts (Tableau 2). Le premier découle de l'existence d'un double peptide de liaison entre les deux domaines de la protéine et a requis la mise en place d'un nouvel opérateur de mutation. Un second problème a concerné le mode d'évaluation des structures vis-à-vis des données de SAXS. Plus précisément, il a été nécessaire de vérifier la cohérence de nos résultats avec ceux donnés par le programme CRYSOL et de comprendre d'où pouvaient provenir d'éventuelles différences. Cette cohérence est importante car CRYSOL est le programme généralement utilisé par la communauté scientifique. Le troisième problème est un problème plus général d'optimisation et a nécessité l'implémentation d'un optimiseur spécifique pour améliorer le fonctionnement du programme. Dans le cadre de cette étude, la fonction de croisement n'est pas utilisée car elle est principalement utile dans le cas où l'on cherche à échanger des boucles. Les structures des domaines ayant été considérées comme rigides dans le cadre de cette application, nous ne l'avons pas utilisé.

Tableau 2 : Étapes du cycle de calcul que nous avons modifié à l'occasion de notre étude de PBP1b. La présence de deux peptides de liaison nous a imposé de développer un nouvel opérateur de mutation (dit local-étendu). Cela donne la possibilité de gérer tous les cas où un domaine est inséré dans une boucle. Par ailleurs, nous avons amélioré la convergence de l'algorithme en introduisant la possibilité de faire varier la pression de sélection au cours du processus et nous avons comparé la façon dont nous évaluons nos structures vis-à-vis d'une courbe de SAXS avec le programme CRYSOL, très utilisé par les gens qui utilisent le SAXS.

Étape	Problème rencontré
Mutation	Présence d'un double peptide de liaison
Évaluation	Évaluation de l'accord entre la courbe calculée et expérimentale par CRYSOL
Sélection	Espace des conformations exploré susceptible d'être insuffisant

3.1 Problème de la mutation

Comme nous l'avons indiqué précédemment, la difficulté de l'étude de la PBP1b réside dans le fait que le domaine Glycosyltransférase est inséré après une petite épingle en feuillet β du domaine Transpeptidase. Il s'agissait donc de construire un opérateur de mutation capable de déplacer le domaine Glycosyltransférase sans rompre l'intégrité du domaine Transpeptidase. Pour se faire, nous avons combiné les propriétés des opérateurs de mutation globale et locale. Comme dans le cas de l'opérateur global, nous avons appliqué une perturbation ponctuelle des angles phi et psi à un acide aminé du premier peptide de liaison, cette perturbation étant répercutée sur la totalité du domaine Glycosyltransférase jusqu'à un acide aminé du second peptide de liaison. La connectivité du second peptide de liaison est ensuite restaurée par l'application d'une minimisation sur quelques résidus de part et d'autre.

L'existence de deux peptides de liaison entre les deux domaines pose un second problème : partant d'une situation donnée, une rotation de 180° ou plus d'un des domaines par rapport à l'autre autour de l'axe les joignant entraîne un croisement des peptides de liaison. Pratiquement, ce problème a été géré en procédant à la rectification de la géométrie des deux peptides de liaison en deux temps. Dans un premier temps, une minimisation de leur énergie est appliquée en supprimant le terme de répulsion entre les atomes des deux peptides ce qui leur permet de « se croiser » si besoin est (Figure 13 en bas à droite). Dans un second temps, une seconde minimisation est réalisée après rétablissement de la fonction d'énergie complète.

Il est important de préciser que pour pouvoir traiter des protéines de la taille de PBP1b il a été nécessaire de passer du programme X-PLOR au programme CNS. La conversion des scripts de mutation a nécessité de nombreux ajustements, notamment au niveau du système de gestion des coordonnées atomiques. Dans X-PLOR, on utilisait des allers-retours entre coordonnées cartésiennes et coordonnées internes pour effectuer les modifications spatiales alors que dans CNS, ce sont des matrices de rotation qui sont appliquées à des sélections d'atomes.

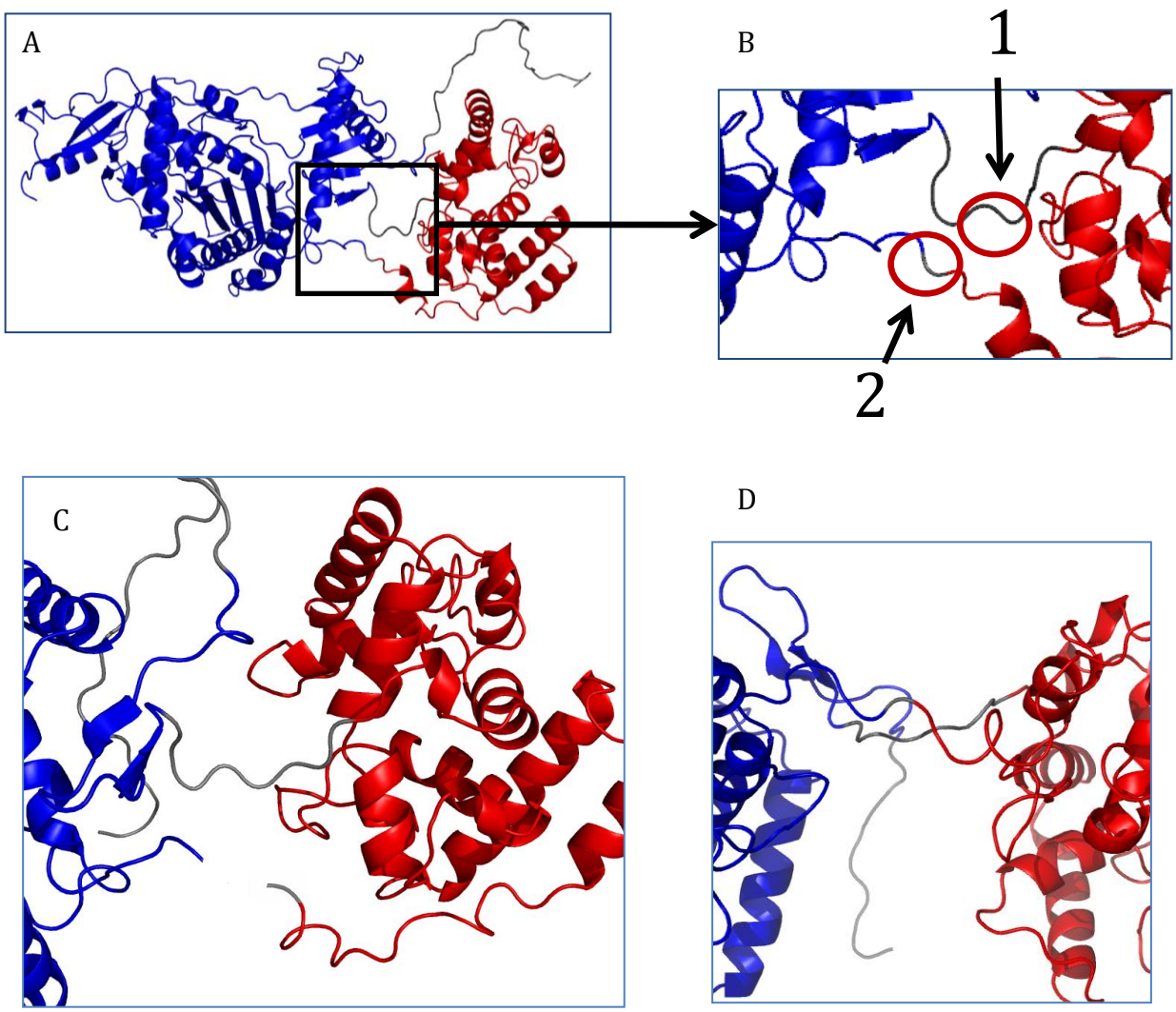


Figure 13 : Fonctionnement de l'opérateur de mutation locale-étendue. Les deux domaines de la protéine sont représentés en bleu (domaine Transpeptidase) et en rouge (domaine Glycosyl-transférase). La vue A met en évidence les deux peptides de liaison existant entre les deux domaines. Sur la vue B, nous avons indiqué les deux régions des peptides de liaison limitant le domaine affecté par la mutation. Un acide aminé est choisi aléatoirement dans la région 1 et ses angles phi et psi sont modifiés aléatoirement. L'effet de la mutation est calculé jusqu'au milieu de la région 2, cette région subissant ensuite une minimisation d'énergie de façon à rétablir la connectivité de la chaîne. En C est présentée la structure résultant de l'application de la mutation avant la minimisation. En D est présenté un exemple de structure après mutation dans le cas de peptide de liaison croisés.

3.2 Problème de l'évaluation

Comme indiqué précédemment, le programme CRYSOL est le programme de référence pour calculer une courbe SAXS à partir d'une structure de protéine. Ce programme peut aussi être utilisé pour vérifier l'accord entre la courbe calculée et la courbe expérimentale par un calcul de χ. Dans sa version actuelle, CRYSOL calcule le χ en plusieurs fois en ajustant à chaque itération les paramètres de contraste de la couche d'eau, de rayon atomique et de volume exclu. Le contraste de la couche d'eau exprime le contraste de densité de l'eau d'hydratation autour de la protéine. Dans le programme, elle est considérée comme une monocouche dont l'épaisseur correspond à la valeur du paramètre. Le rayon atomique moyen est, comme son nom l'indique, le rayon atomique moyen des atomes de la protéine. Le volume exclu correspond à la somme des volumes des atomes de la molécule et est donc dépendant de la valeur du rayon atomique. Lors du calcul de l'accord, CRYSOL les découple pour avoir un plus grand nombre de paramètres à modifier. Un dernier paramètre que CRYSOL ajuste pour améliorer l'accord avec les données expérimentales est le facteur d'échelle. Ce facteur permet de mettre les courbes sur la même échelle avant de les comparer et CRYSOL l'ajuste sur différentes parties de la courbe pour assurer le meilleur accord possible sans jamais l'indiquer explicitement à l'utilisateur.

Pour pouvoir maitriser la valeur de ces paramètres, nous avons décidé de procéder au calcul en deux étapes. Nous calculons la courbe de SAXS correspondant à chacun des individus (structures de la protéine) en utilisant CRYSOL avec un jeu de paramètres imposant des valeurs fixes de la couche d'eau, du rayon atomique et du volume exclu (respectivement 0,03 $Å^{-3}$; 1,615 Å et 96950 $Å^3$). Chaque courbe est ensuite comparée à la courbe expérimentale en calculant un χ^2 utilisant un facteur d'échelle déterminé par le rapport des sommes des intensités expérimentales et calculées. La formule de ce χ^2 est représentée dans l'Équation 2.

$$\chi^2 = \sum_i \left(\frac{R_i - \frac{R}{S} S_i}{Inc(R_i)} \right)^2 \quad \text{avec} \quad R = \sum_i R_i \quad \text{et} \quad S = \sum_i S_i$$

Équation 2 : Formule du χ^2 utilisée dans Dadimodo. Ri et Si désignent les valeurs du i^{ème} point des courbes expérimentales (Ri) et calculées sur la structure (Si). Inc(Ri) est l'incertitude expérimentale associée au point Ri. Les deux courbes sont recalées l'une par rapport à l'autre en utilisant un facteur d'échelle défini comme le rapport de la somme des points de la courbe expérimentale sur la somme des points de la courbe calculée (il s'agit du facteur minimisant l'écart entre les deux courbes lorsqu'elles sont considérées dans leur totalité).

Dans ces conditions, il était important de savoir dans quelle mesure notre fonction d'évaluation était en accord avec les résultats donnés par CRYSOL. Nous voulions en particulier savoir si l'utilisation de cette formule risquait d'introduire un biais dans les résultats de nos optimisations lors de comparaisons avec des techniques standard (comme BUNCH ou SASREF). Lorsque CRYSOL est utilisé pour comparer la courbe calculée à la courbe expérimentale, il donne le résultat sous la forme d'un χ (désigné par la suite par χ_{CRYSOL}). Pour faciliter la comparaison des résultats, nous avons donc converti le χ^2 Dadimodo en χ (désigné par la suite comme étant le $\chi_{Dadimodo}=\sqrt{\chi^2}$). Les résultats d'évaluation des structures obtenues par le programme Dadimodo sont regroupés dans la Figure 14, classés par ordre décroissant de χ CRYSOL. Les 9 meilleures structures présentent un χ inférieur à 1, tandis que les 7 plus mauvaises ont des valeurs supérieures à 1,5. La majorité des structures (27) présentent une valeur entre 1 et 1,5. La moyenne des χ est de 1,21 et l'écart-type de 0,25. La distribution des scores Dadimodo est légèrement plus étendue, avec une moyenne de 1,41 et un écart-type de 0,4.

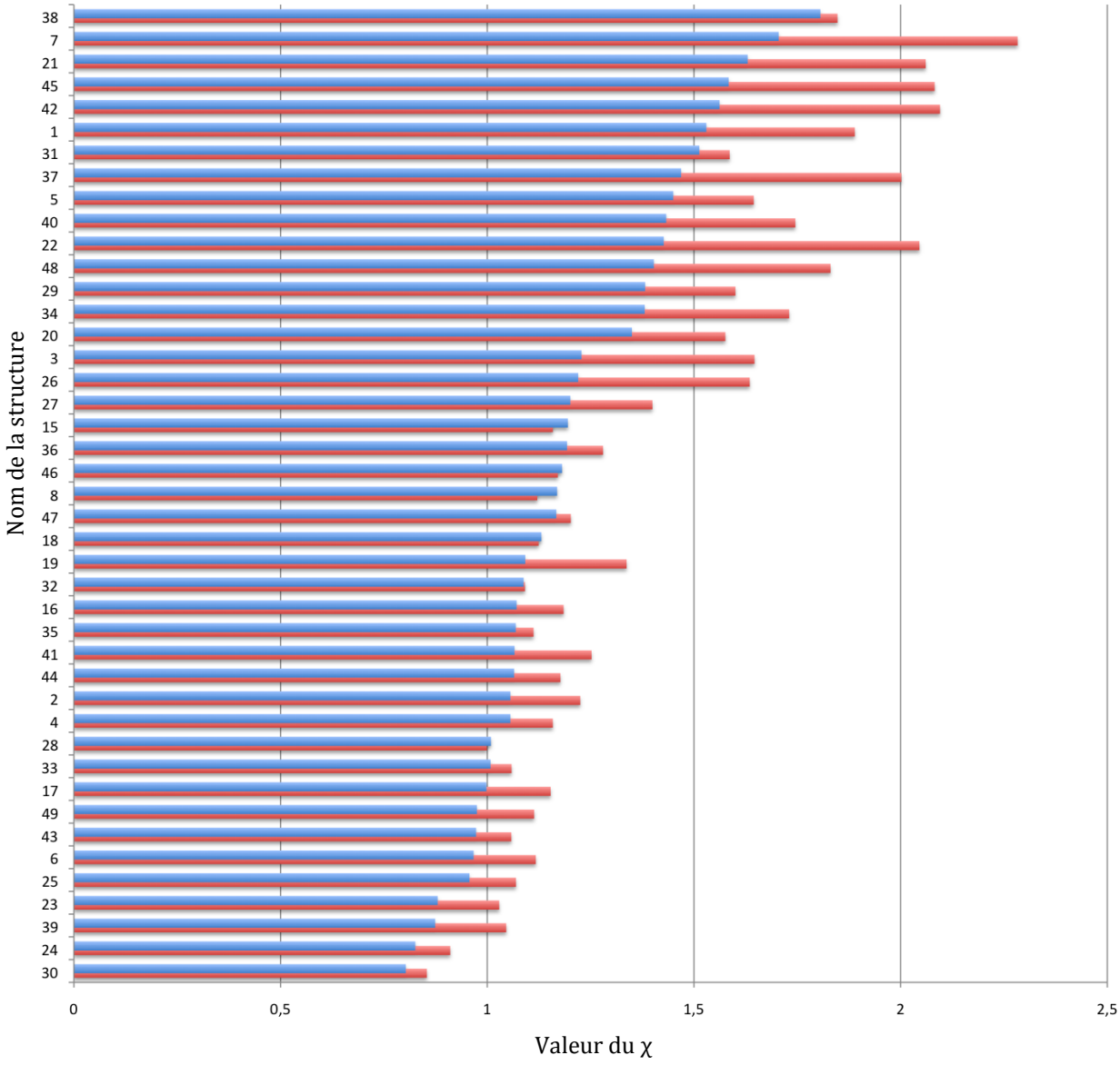


Figure 14 : Valeur des χ_{CRYSOL} (en bleu) et $\chi_{Dadimodo}$ (en rouge) des 43 structures finales obtenues par Dadimodo triées par valeur croissante du χ_{CRYSOL}.

Les deux scores semblent corrélés, plus le χ_{CRYSOL} est élevé, plus le score Dadimodo augmente. L'évolution globale du score Dadimodo est similaire à l'évolution du χ_{CRYSOL}, comme l'indique la Figure 15. En fait, la hiérarchie est identique jusqu'à une valeur de χ d'environ 1,2. Passé cette valeur le classement commence à différer. Cette différence est due à la manière dont la courbe est calculée par CRYSOL, comme abordé précédemment. Dans le cadre de Dadimodo, les paramètres de couche d'eau, de volume exclu et de rayon atomique sont fixés alors que CRYSOL calcule le χ en ajustant ces paramètres au cas par cas (deux modèles d'une même protéine pouvant avoir des valeurs de paramètres

différents). La différence de valeur entre ces paramètres explique la différence observée des tendances et des valeurs. De même, cela explique pourquoi le χ_{CRYSOL} évolue moins vite car CRYSOL compense le désaccord entre la courbe expérimentale et la courbe calculée par l'ajustement de ces paramètres dont la valeur n'est pas toujours indiquée à l'utilisateur. Comme ce sont les structures qui présentent le meilleur accord qui sont les plus intéressantes, on peut dire que le χ^2 Dadimodo est un bon indicateur de la qualité de l'accord avec les données expérimentales puisqu'on retrouve la même hiérarchie que CRYSOL pour ces structures.

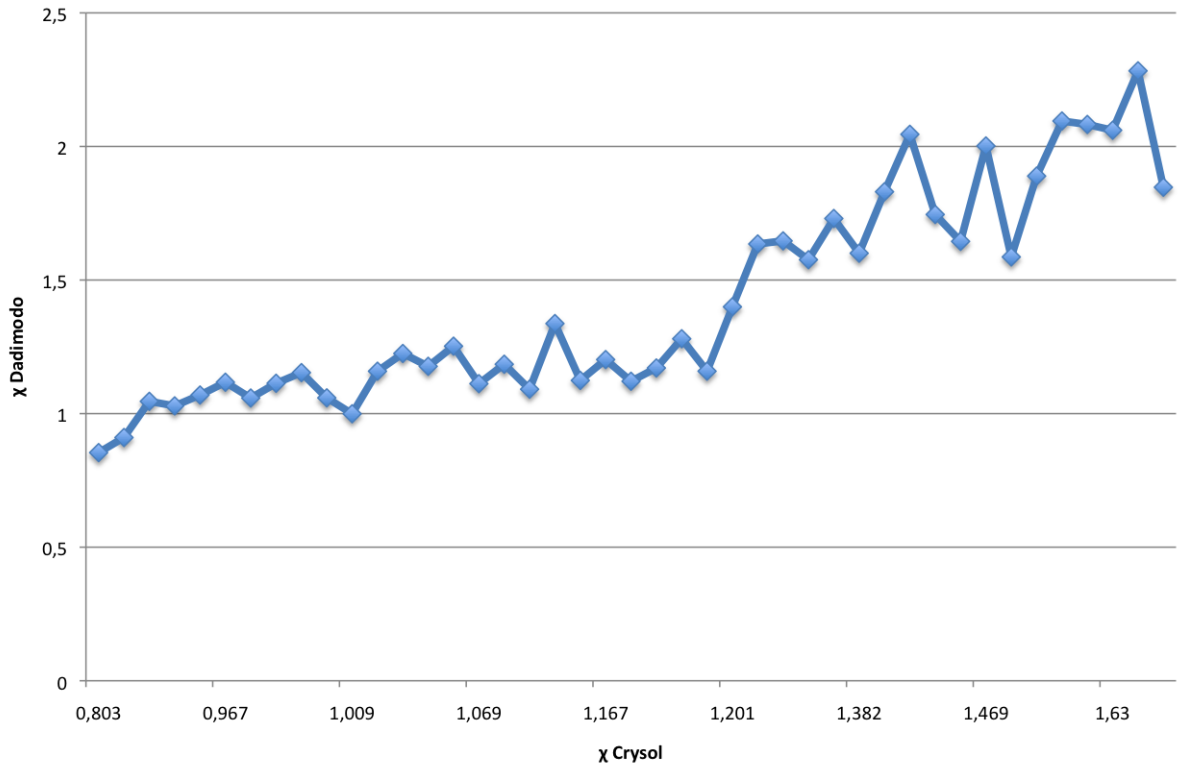


Figure 15 : Valeur du $\chi_{Dadimodo}$ en fonction du χ_{CRYSOL}. On remarque que CRYSOL compense les écarts les plus importants en ajustant les paramètres.

3.3 Problème de sélection

Pour optimiser le temps de calcul et éviter de converger trop vite vers un optimum local, un système de pression de sélection progressive a été implémenté. Ce système permet de définir pour chaque génération les valeurs *n* et *m* à utiliser dans la méthode de sélection. Avec cette méthode similaire au recuit simulé, il est possible d'augmenter au fur et à mesure la pression de sélection (Équation 3).

$$P_S = \frac{n - m}{p - n}$$

Équation 3 : Formule nous ayant servi à définir la pression de sélection associée à chaque tournoi. p est la taille de la population à la fin du processus de mutation, n est la taille de la sous-population choisie aléatoirement au sein de laquelle les individus sont mis en compétition, m est le nombre d'individus conservés au sein des n à la fin de chaque tournoi. Avec cette définition, lorsque m = n (tous les individus choisi aléatoirement pour un tournoi sont conservé à la fin de ce tournoi, ce qui revient à faire un simple tirage aléatoire), la pression de sélection est nulle. De même lorsque n = p (le tournois implique tous les individus, ce qui revient à sélectionner les m meilleurs de la population p), la pression devient infinie.

La pression sera dite « faible » si *n* est proche de *m* (et nulle si *n*=*m*) et dans ce cas on aura tendance à conserver des « mauvais » individus dans la population. Au contraire, plus *m* est petit devant *n*, plus la tendance à sélectionner les « meilleurs » individus sera élevée ce qui correspond à une pression de sélection « forte ». Pour pouvoir explorer un maximum de conformations et éviter de converger trop rapidement, il est donc nécessaire de trouver une bonne combinaison entre les différentes pressions de sélection.

L'analyse des courbes obtenues avec cette nouvelle méthode de sélection (Figure 16) permet de se rendre compte de l'intérêt de cette amélioration. En particulier, la baisse progressive des écarts types indiquent que l'algorithme explore un plus grand nombre de conformations avant de tomber dans un minimum global.

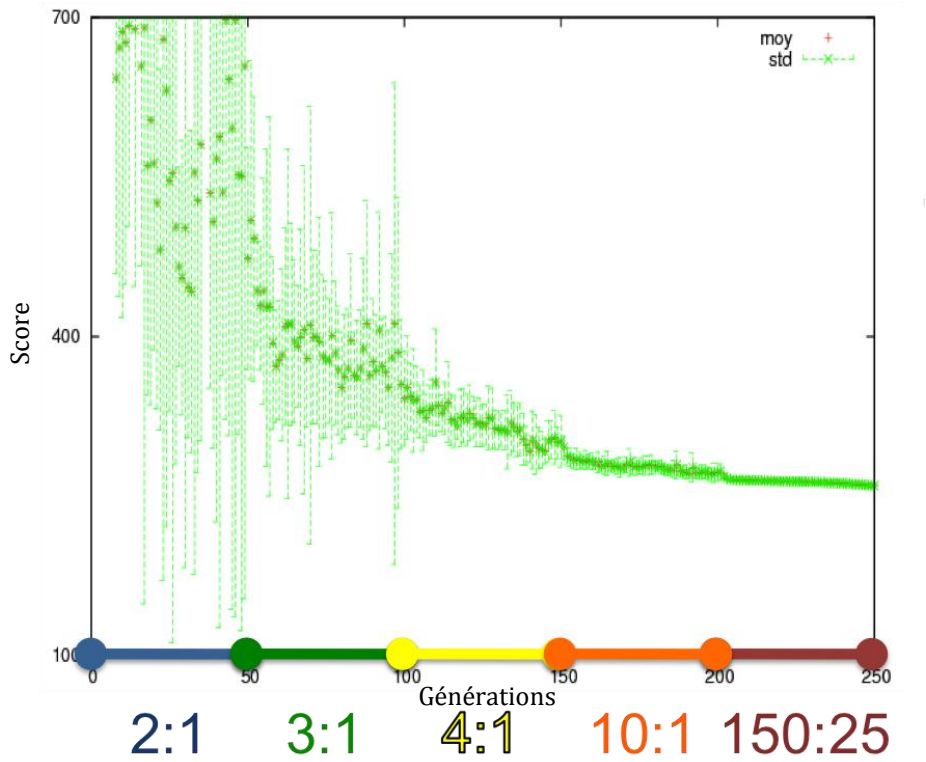


Figure 16 : Évolution du score moyen d'une famille de 25 structures générées par Dadimodo au cours d'un test réalisé dans les mêmes conditions que celui de la Figure 12 mais en utilisant une stratégie d'optimisation impliquant une variation régulière de la pression de sélection. Comme précédemment, les points verts représentent le score moyen et les traits l'enveloppe de toutes les valeurs présentent dans la population. Les nombres indiqués au niveau de l'axe des abscisses représentent pour chaque cycle la valeur de pression de sélection utilisée sous la forme « taille échantillon:nombre d'individus sélectionnés ». On constate une diminution beaucoup plus régulière de la valeur moyenne et de la variabilité associées à la population au cours des générations. Le résultat est une amélioration significative et reproductible du score à la fin du calcul.

Dans le cadre de l'application de Dadimodo à PBP1b, l'algorithme a été lancé sur cinquante générations avec trois valeurs de pressions de sélection. Ce nombre de génération relativement faible permet de diminuer le temps nécessaire aux calculs car, la protéine étant de très grande taille, les calculs de modification de structure et CRYSOL sont relativement longs. Pour compenser ce nombre réduit de générations, nous avons généré un ensemble de structures de départ présentant une forte variabilité. De la génération 1 à 20, une pression de sélection faible est utilisée (2 :1), de la génération 21 à la génération 35 une pression de sélection intermédiaire (5 :1) et de la génération 36 à la fin, seules les meilleures sont sélectionnées (125 :25). L'évolution du score moyen des huit premières séries de calcul est représentée par le Figure 17.

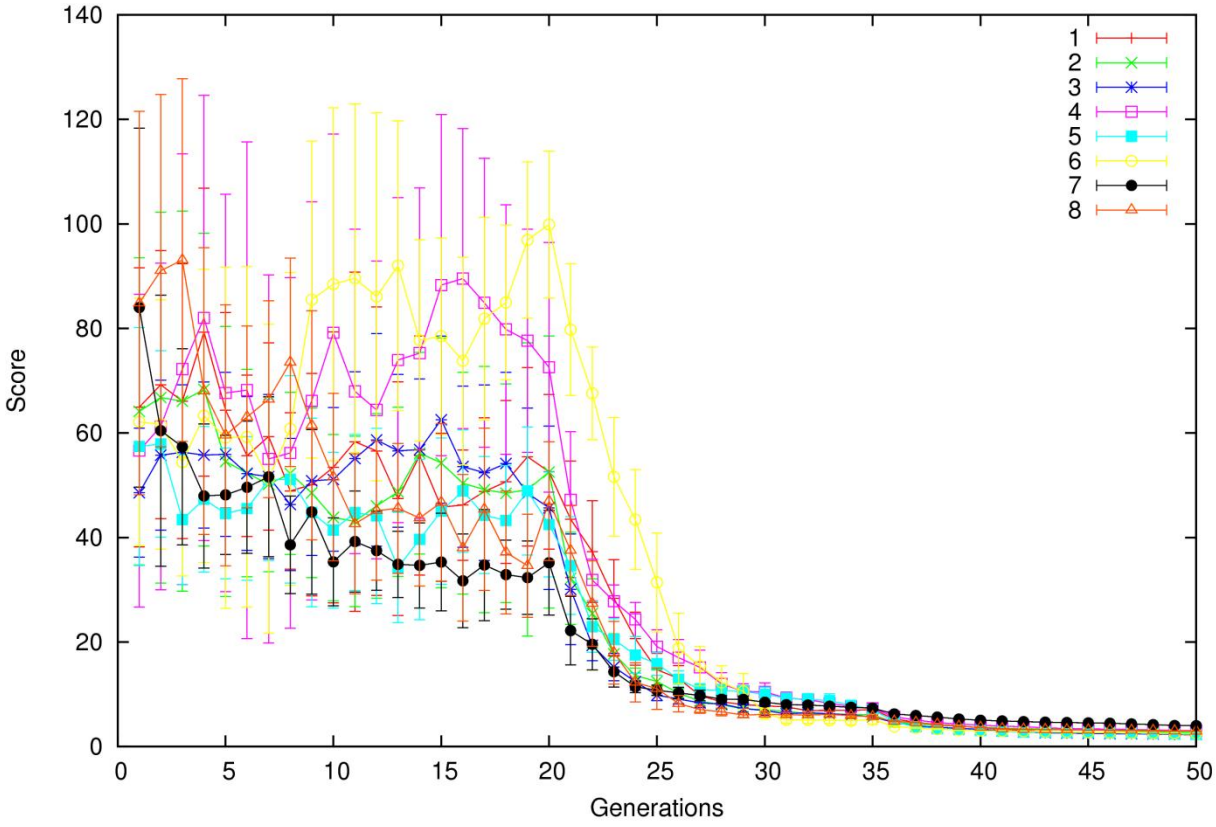


Figure 17 : Évolution du score moyen Dadimodo en fonction des générations pour les 8 premières familles de calcul. La taille des écarts-types représente la dispersion du score dans les familles, elle diminue notablement à partir de la génération 20 (passage à une pression de sélection plus importante) puis devient négligeable lors du passage à la pression de sélection infinie à partir de la génération 35.

4 APPLICATION À PBP1B

4.1 STRUCTURES DE DÉPART

4.1.1 IDENTIFICATION DES RÉGIONS À MODIFIER

Comme nous l'avons signalé précédemment, l'un des points forts de Dadimodo est qu'il est possible de moduler le type de mutation (et sa probabilité) dans les différentes régions de la protéine. Cela permet, par exemple, de maintenir le cœur d'une structure invariante, tout en faisant varier la structure des boucles (par des mutations locales) ou du peptide de liaison entre deux domaines (par des mutations globales). Dans le cas de PBP1b, nous avons distingué cinq régions : l'extrémité N-terminale (résidus 72 à 100), le premier et le deuxième peptide de liaison reliant les deux domaines (résidus 120 à 127 et 323 à 335, respectivement) et les deux domaines. Les deux domaines ont été définis comme des régions ne devant subir aucune mutation (ce qui revient à les traiter comme des corps rigides), l'extrémité N-terminale comme soumise à des mutations globales et les deux peptides de liaison comme les points de départ et d'arrivée de mutations locales étendues. L'encodage du type de mutation autorisée dans chaque région est réalisé au niveau de la colonne du facteur B du fichier PDB. Les différentes valeurs ainsi que les types de mutation associés sont indiqués dans le Tableau 3.

Tableau 3 : Tableau d'encodage des types de région pour les mutations

Région	Résidus	Valeur du facteur B	Type de mutation
N-ter	72-100	-1	Mutation **globale**
Peptide de liaison 1	120-127	-2	Mutation **locale étendue**
Peptide de liaison 2	323-335	-2,50	Résidus de raccord pour la Mutation **locale étendue**
Domaine Glycosyltransférase	128-322	1	Pas de mutations
Domaine Transpeptidase	101-119 et 336-791	2	Pas de mutations

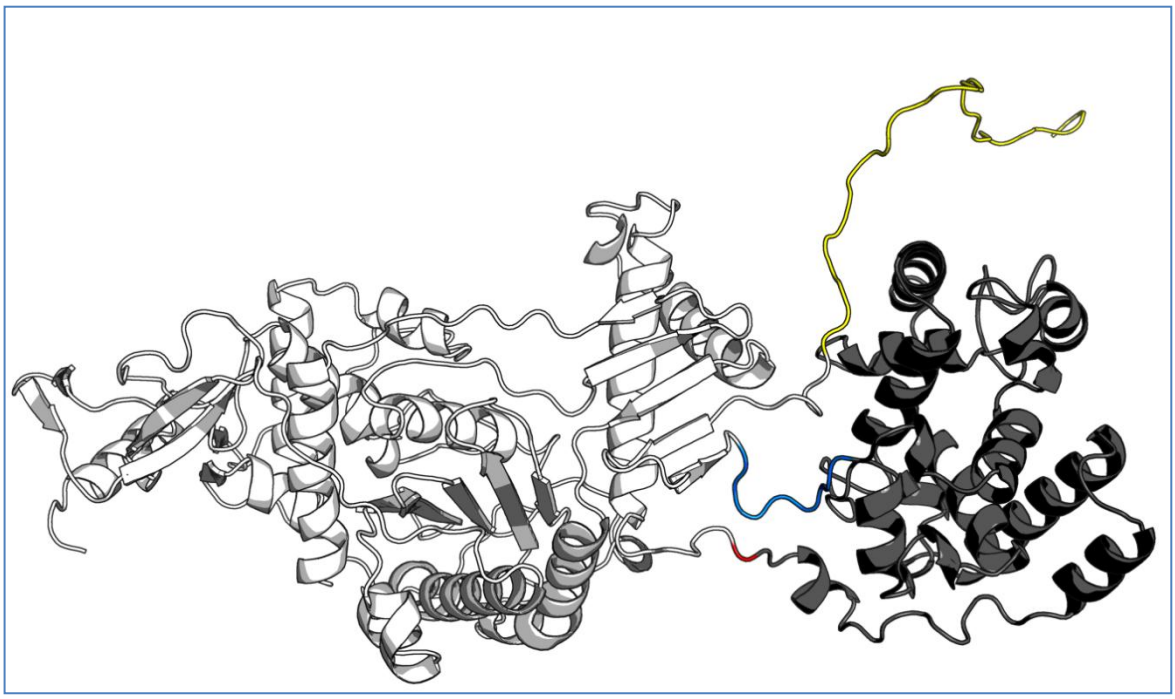


Figure 18 : Représentation, sur un modèle de la structure de PBP1b, des différentes régions pouvant être mutées au cours du processus. Les deux régions ne pouvant être mutées sont représentées en blanc (domaine Transpeptidase) et en noir (domaine Glycosyltransférase). L'extrémité N-terrminale, subissant des mutations globales (donc considéré comme complètement déformable) est en jaune. Les deux peptides de liaisons, entre lesquels il peut y avaoir des mutations locales étendues, sont en bleu et en rouge.

4.1.2 Génération des structures de départ

La taille de la protéine (750 acides aminés) augmentant fortement le temps de calcul associé à chaque cycle, nous avons été obligé de réduire la phase du processus réalisé avec une pression de sélection basse (équivalent à la phase d'exploration dans un recuit simulé). Pour pallier cela, nous sommes partis d'un jeu de structures initiales présentant la plus grande variabilité possible. Pour ce faire, Dadimodo a été lancé « à vide » (c'est à dire sans pression de sélection), à partir de la structure fournie par notre collaborateur pendant 200 générations avec des paramètres de mutations élevées : trois mutations globales et deux mutations locales étendues par génération avec une gaussienne élargie (pour augmenter la modification des valeurs des angles Φ/Ψ). L'ensemble des 5000 structures obtenues (25 par génération) est mélangé et distribué aléatoirement à travers les 43 familles de départ (25 structures par familles de départ).

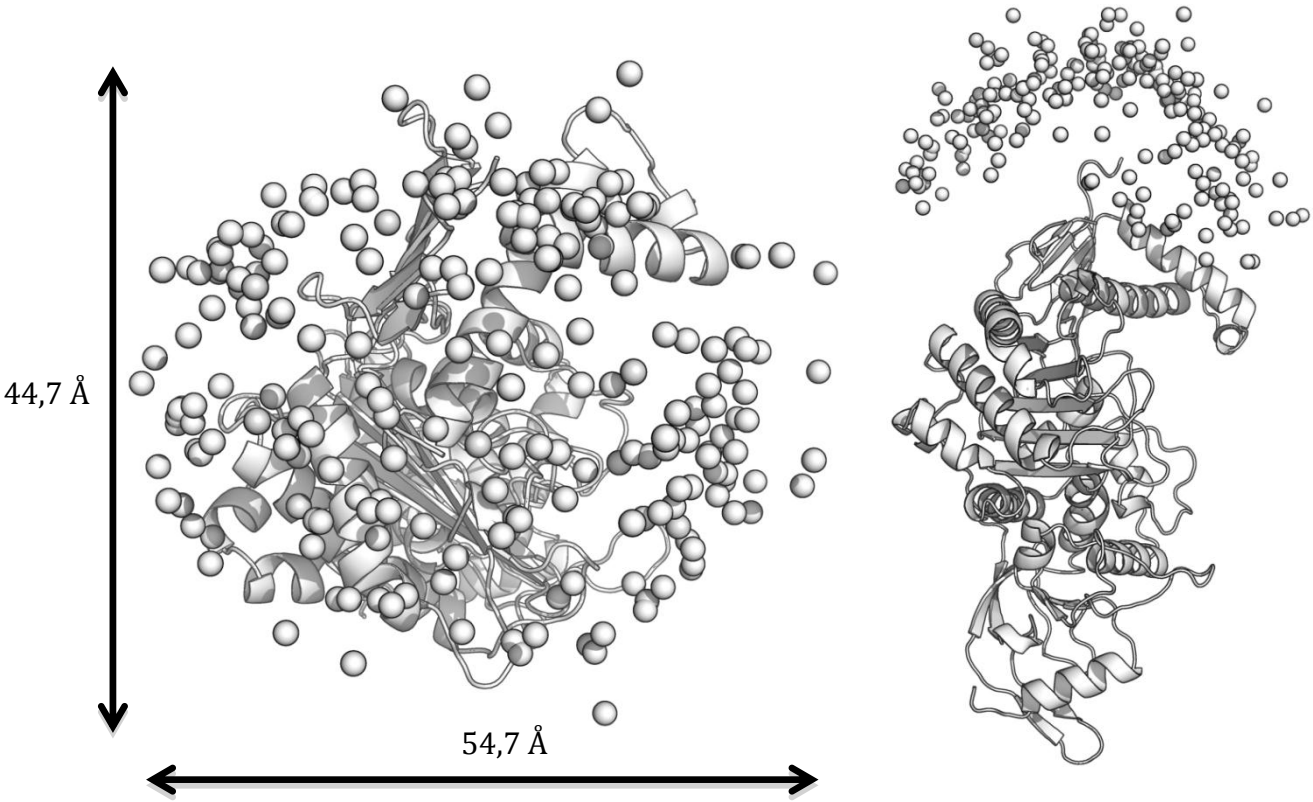


Figure 19 : Distribution des structures utilisées en entrée du processus d'optimisation et engendrées en faisant fonctionner Dadimodo sans pression de sélection en vue de dessus (à gauche) et de face (à droite). Pour simplifier la présentation, nous avons superposé tous les domaines Transpeptidase (représenté en ruban) et indiqué les centres de gravité des domaines Glycosyltransférase sous forme de sphères. Les deux flèches servent à donner l'échelle.

4.2 ÉCHANTILLONNAGE DE LA COURBE DE DIFFUSION DES RAYONS X

Le fichier contenant les données expérimentales qui nous a été donné est un fichier contenant une courbe de diffusion définie par 714 points de Q=0,0142 $Å^{-1}$ à Q=0,534 $Å^{-1}$. À chacun des points est associée une incertitude expérimentale dépendante du matériel et de la technique d'estimation utilisée qui est propre à chaque équipe. Ces point ne sont pas uniformément répartis sur toute la courbe, l'écart entre deux points est de $3,42^{E}$-04 $Å^{-1}$ du premier point à Q=0,113 $Å^{-1}$, puis de $9,92^{E}$-04 $Å^{-1}$ de Q=0,113 $Å^{-1}$ jusqu'à la fin. Cette différence résulte de la combinaison de données enregistrées à deux distances différentes entre l'échantillon et le détecteur. Afin de pouvoir comparer les courbes calculées par CRYSOL à la courbe expérimentale, il est nécessaire de ré-échantillonner la courbe expérimentale pour obtenir le même nombre de points. En effet, les courbes générées par CRYSOL à partir des structures sont d'une longueur définie par l'utilisateur et à intervalle fixe. Au sein de Dadimodo, les courbes générées par CRYSOL contiennent 101 points uniformément répartis de Q=0 à Q=0,5 $Å^{-1}$. L'échantillonnage est réalisé de la manière suivante :

- Détermination de la valeur des 101 Q à écarts constants entre 0 et 0,5 $Å^{-1}$.
- Recherche de sept points consécutifs de la courbe expérimentale dont la moyenne des Q est égale au point recherché dans la courbe interpolée.
- Détermination de la valeur de I(Q) et de l'incertitude en faisant la moyenne des I(Q) et des incertitudes des sept valeurs associées sur la courbe expérimentale.

Il est important de noter que l'interpolation de la courbe expérimentale est une étape importante dans l'utilisation du programme et n'est pas à négliger. Une mauvaise estimation du bruit peut en effet aboutir à des structures de mauvaise qualité.

5 REGROUPEMENT DES STRUCTURES

De part le système de pression de sélection utilisé, chaque calcul de Dadimodo aboutit à une structure unique. Chacune de ces structures correspond à un minimum local de la fonction d'évaluation trouvé par l'algorithme. Pour obtenir un grand nombre de solutions, il est donc nécessaire de lancer l'algorithme un grand nombre de fois avec un maximum de variabilité dans les structures de départ. Ici, 43 calculs ont été lancés en parallèle sur un cluster, pour une durée moyenne de 3 jours.

La Figure 20 représente les positions des centres de gravité des domaines Glycosyltransférase, après superposition des domaines Transpeptidase, dans le cas des structures initiales (en blanc) et finales (en rouge). On constate une évolution nette du système. Les centres de gravité des domaines Glycosyltransférase des structures initiales sont fortement dispersés et couvrent une grande variété d'angles entre les domaines Transpeptidase et Glycosyltransférase. En revanche, dans le cas des structures finales, les centres de gravités sont pour l'essentiel rassemblés le long d'un axe x en face du domaine Transpeptidase avec une faible dispersion sur l'axe y (passage d'une dispersion sur 44,7 Å à 17,7 Å). On peut aussi remarquer sur la figure de droite que les centres de gravité des structures finales se retrouvent préférentiellement en haut, c'est à dire relativement éloignés du domaine Transpeptidase. Aucune de ces structures finales ne présente d'interaction forte entre les deux domaines.

Pour aller plus loin dans l'analyse des 43 structures obtenues, il nous a semblé pratique de les regrouper. Nous voulions en particulier savoir si un ensemble (ou un petit nombre d'ensembles de structures) émergerait ou si les données étaient compatibles avec une distribution large et uniforme de solutions

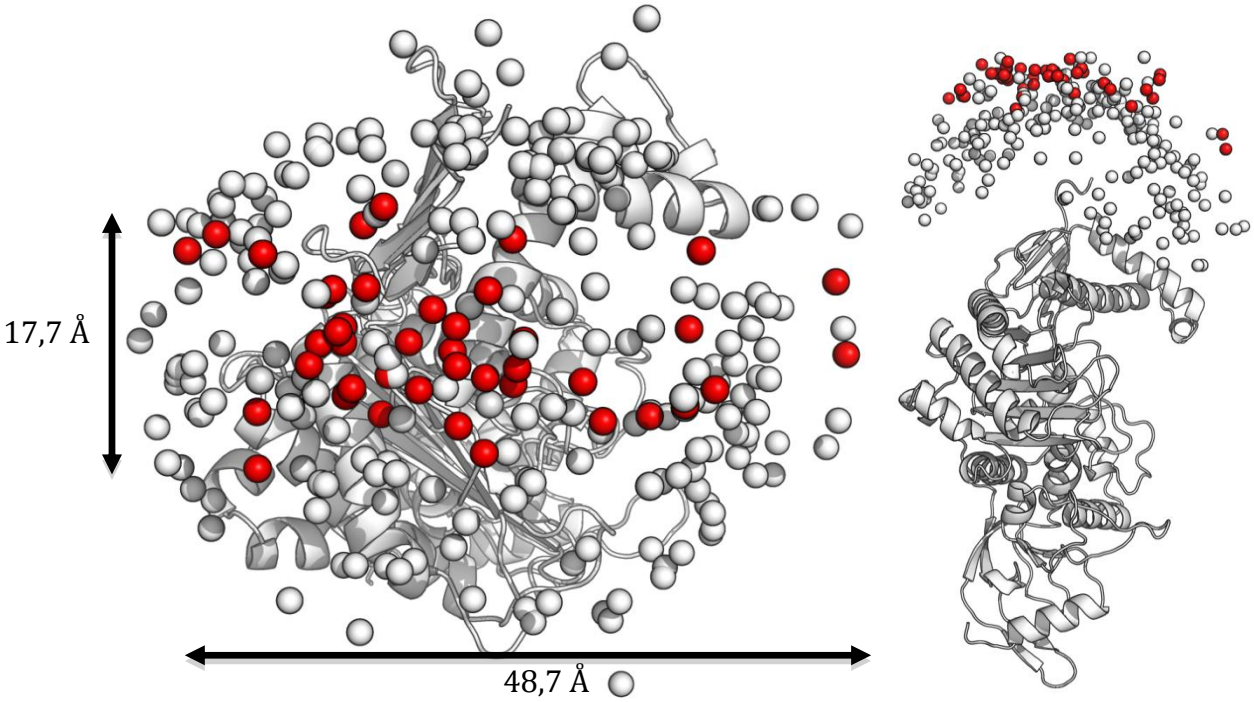


Figure 20 : Comparaison de la variabilité au sein des structures initiales et finales. La représentation adoptée est la même que dans la figure précédente. Les domaines Transpeptidase (représenté en ruban) de toutes les structures ont été superposés. Les centres de gravités des domaines Glycosyltransférase sont représentés sous formes de sphères blanches (pour les structures de départ) et rouges (pour les structures finales). Sur la vue de gauche, on note le resserrement des positions de ces centres de gravités le long de l'axe x (passage d'une dispersion d'environ 50 Å à une dispersion d'environ 20 Å le long de y). De même, on constate sur la vue de droite que les centres de gravités des domaines Glycosyltransférase tendent à être plus éloignés du domaine Transpeptidase dans les structures finales que dans les structures initiales.

5.1 Regroupement par RMSD

Le regroupement par RMSD est la méthode la plus classique pour classer des ensembles de structures. Le regroupement a été réalisé par deux scripts distincts développés spécialement pour cette application. Le premier utilise PyMOL pour déterminer une matrice de distance des RMSD des Cα entre les domaines Glycosyltransférase. Dans un premier temps, le script importe toutes les structures dans le programme PyMOL et les aligne sur le domaine Transpeptidase. Dans un deuxième temps, la fonction rms_cur de PyMOL est utilisée pour déterminer le RMSD, cette fonction permettant de calculer le RMSD entre deux sélections sans alignement préalable. Les

valeurs de RMSD sont ensuite écrites dans un fichier texte sous forme de matrice. Le deuxième script utilise ensuite cette matrice pour effectuer le regroupement à proprement parler avec la méthode illustrée par la Figure 21 en fixant le seuil de regroupement à 10 Å.

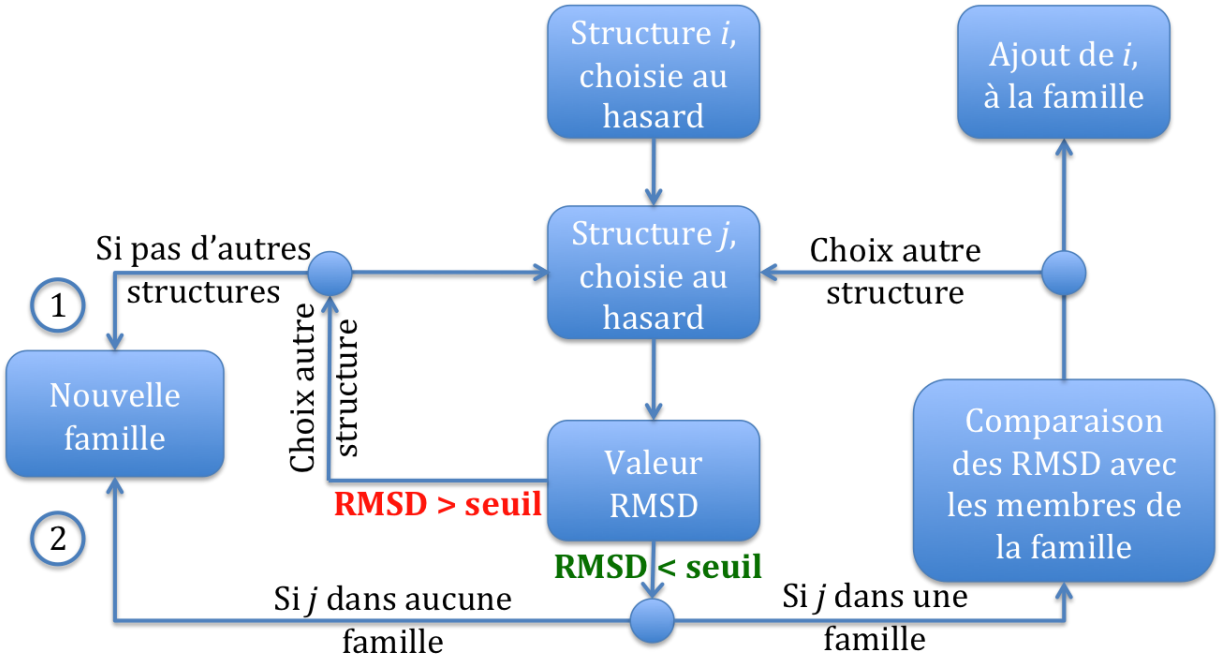


Figure 21 : Fonctionnement du script de regroupement. Deux cas sont possibles pour créer une nouvelle famille : dans le cas 1 la nouvelle famille créée ne contient que *i*, dans le cas 2 elle contient *i* et *j*. Cette méthode de permet de contraindre le nombre de famille créées et plusieurs peuvent exécutions donner des résultats différents. Chaque regroupement a donc été fait 10 fois et seule l'exécution présentant le plus faible nombre de regroupements a été conservée.

La grande variabilité des valeurs de RMSD n'a pas permis d'identifier de grands regroupements. En ajoutant les structures les plus proches des plus grands regroupements en augmentant un peu le seuil (de 10 à 14 Å), il a été possible de réduire sensiblement leur nombre. Néanmoins, il en reste encore ne contenant qu'une ou deux structures.

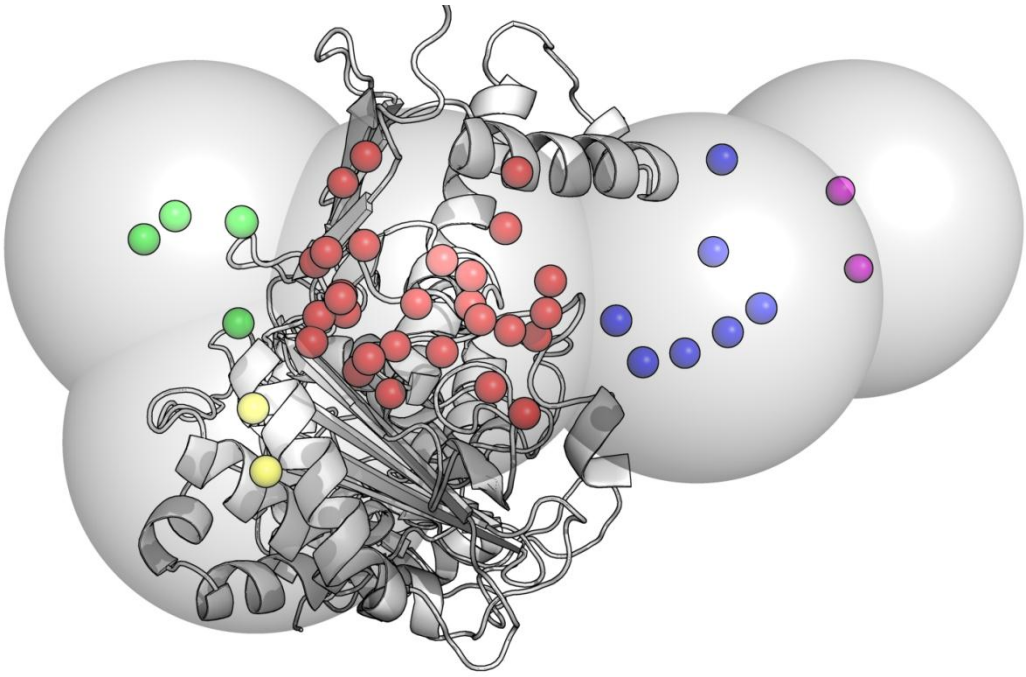
5.2 Regroupement par centre de gravité

Pour obtenir des regroupements de taille significative, les RMSD entre les domaines ont été remplacés par les distances entre les centres de gravités. Dans le cas du RMSD, la

valeur est extrêmement dépendante de la rotation et de la translation entre les domaines. En utilisant les centres de gravités, l'information sur la rotation n'est pas conservée et seules les translations interviennent dans le classement des structures.

La méthode de regroupement est similaire, seul le premier script responsable du calcul des RMSD est remplacé par un calcul des distances des centres de gravité. Les coordonnées des centres de gravité peuvent être calculées par PyMOL à l'aide d'un module externe prenant en compte le poids de chaque atome. Ainsi, il a été possible de déterminer les coordonnées du centre de gravité de chaque domaine Glycosyltransférase (en conservant la superposition des domaines Transpeptidase) et de calculer la matrice des distances de ces centres. Le regroupement est ensuite effectué par le même script que précédemment en utilisant un seuil à 10 Å.

Après ajustement du seuil à 12 Å pour regrouper des structures isolées et par l'analyse manuelle de la matrice des distances, il a été possible de diminuer le nombre de regroupements à 5, représentés sur la Figure 22.



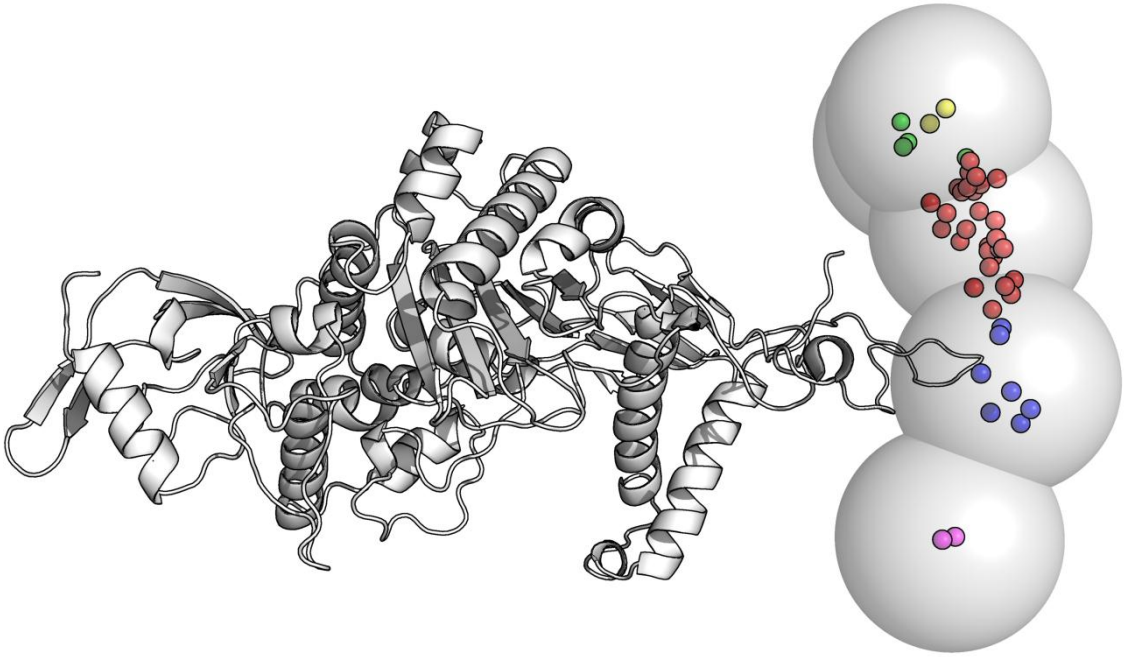


Figure 22 : Représentation schématique des regroupements obtenus à la fin de la procédure explicitée dans le texte (en haut, vue le long du grand axe de la protéine, en bas, vue perpendiculaire au grand axe). Comme précédemment, les domaines Transpeptidase des différentes structures obtenues à la fin du calcul ont été superposés et les positions des centres de gravité des domaines Glycosyltransférase ont été indiquées par sphères de 1 Å de rayon (sphères colorées). Les sphères de mêmes couleurs appartiennent à un même regroupement. L'enveloppe (approximative) de chaque regroupement est matérialisée par une sphère blanche d'un rayon de 12 Å (distance seuil pour le regroupement par RMSD) centrée sur les coordonnées du barycentre du regroupement.

5.3 Regroupement par volume commun

Une fois effectué l'ajustement manuel des regroupements obtenus en utilisant les centres de gravités, nous avons eu l'intention de mieux utiliser l'information contenue dans la courbe de diffusion. Avec cette donnée il semble naturel d'utiliser, pour comparer deux positions différentes du domaine Glycosyltransférase, le volume commun à ces deux domaines. Cette grandeur est difficile à déterminer directement, aussi avons-nous utilisé comme approximation le nombre d'atomes communs (en fait voisins) aux deux domaines.

Le nombre d'atomes communs entre deux domaines Glycosyltransférase des structures *i* et *j* correspond au nombre d'atomes du domaine Glycosyltransférase de la structure *j* se trouvant dans le volume défini par l'enveloppe du domaine de la structure *i*

plus une distance seuil. Cette opération est effectuée grâce à la commande «sel2 within x of sel1 » de PyMOL qui permet de créer une sélection contenant tous les atomes de la sélection 2 à moins de *x* Å d'au moins un atome de la sélection 1. Les sélections 1 et 2 correspondent ici aux atomes Cα, C et N des domaines Glycosyltransférase des deux structures que l'on cherche à comparer et on utilise une valeur de seuil de 5 Å (Figure 23 Haut). Il reste ensuite à compter le nombre d'atomes de cette nouvelle sélection pour évaluer la superposition des volumes. Le regroupement est ensuite effectué en regroupant les structures présentant au moins 50% d'atomes en commun dans le volume du domaine Glycosyltransférase ainsi défini. L'application de cette méthode permet de valider les résultats obtenus en utilisant les centres de gravités après ajustements. Seules deux structures sont groupées dans un regroupement différent, mais ce sont des structures situées à la frontière des regroupements rouge et vert (Figure 24).

Dans le cadre du regroupement de structures obtenues par l'utilisation de courbe SAXS, la méthode de regroupement la plus intéressante est celle du volume commun car elle permet d'obtenir directement des regroupements de bonne taille. Cette technique a aussi l'avantage d'utiliser de façon directe une des caractéristiques de la structure obtenue par le SAXS.

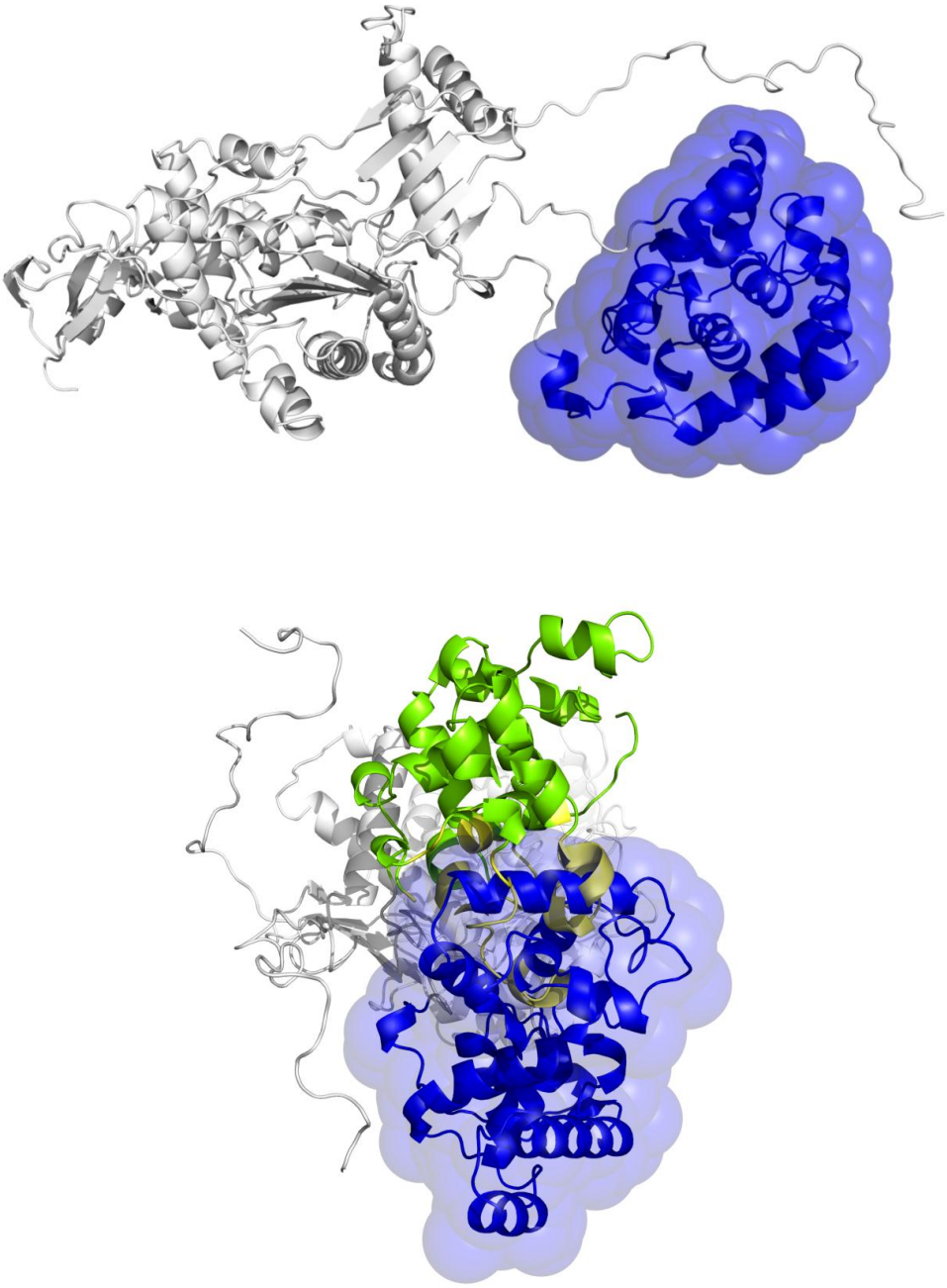


Figure 23 : Illustration du processus d'évaluation de l'espace commun entre deux solutions. En haut, ne solution obtenue par Dadimodo a été représenté en ruban (domaine Transpetidase en blanc, domaine Glycosyltransférase en bleu). L'enveloppe du volume occupé par le domaine Glycosyltransférase a aussi été indiquée en bleu. En bas, deux solutions ont été superposées au niveau de leur domaine Transpetidase. Le domaine Glycosyltransférase et le volume occupé correspondant de l'une des deux solutions sont représentés en bleu. Le domaine Glycosyltransférase de la seconde est représenté en vert et jaune, le jaune correspondant aux acides aminés dont l'un des atomes du squelette au moins est contenu dans le volume de la première solution. Dans ce cas précis, le volume commun est faible (106 atomes Cα, C ou N sur 573 soit ~18,5%).

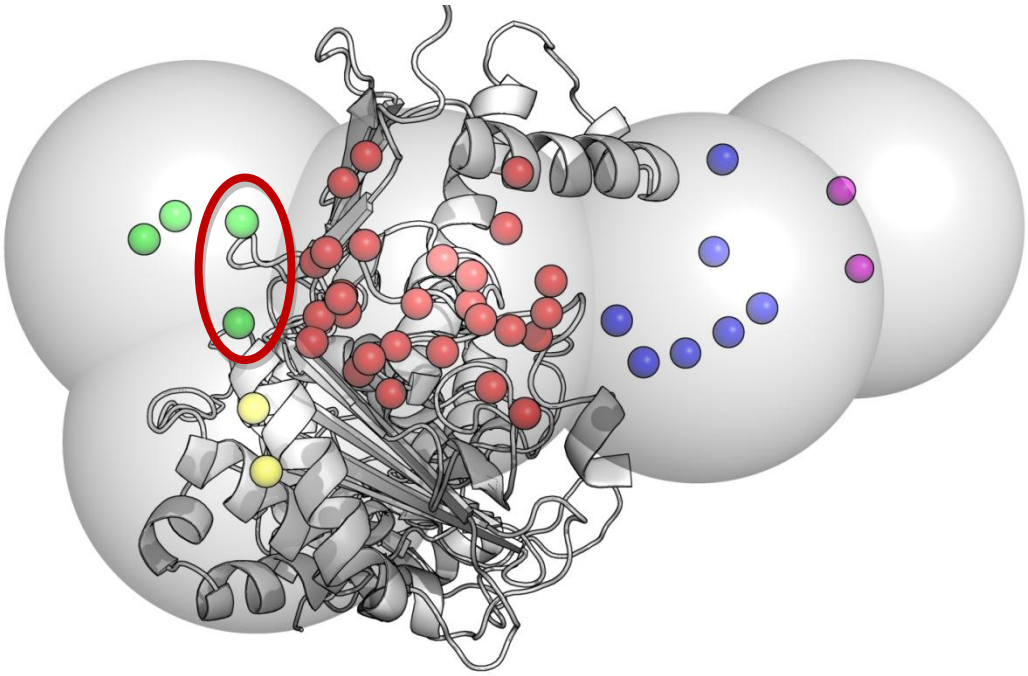


Figure 24 : Différence entre les résultats obtenus par la méthode des RMSD et des volumes occupés : Les regroupements obtenus par la méthode des RMSD ont été représenté de la même façon que dans la figure 22. Les résultats obtenus en utilisant la mesuré de l'espace communs sont identiques, à l'exception des deux solutions « vertes » entourées de l'ellipse rouge qui sont maintenant agrégées à l'ensemble des solutions « rouges ».

6 RÉSULTATS

6.1 ÉVALUATION DE LA QUALITÉ DE L'ACCORD

Le calcul du χ^2 permet d'évaluer la qualité globale de l'accord des structures calculées avec les données expérimentales. Cependant, il est aussi important de s'assurer de la qualité de l'accord pour chacun des points pour pouvoir repérer un très mauvais accord local qui serait « gommé » par un très bon accord global. Dans ce but, la distribution des résidus normalisés (correspondant à la différence, pour chaque point, entre la courbe calculée et la courbe expérimentale rapportée à la valeur de l'incertitude expérimentale) est calculée pour chaque modèle entre Q=0,02 $Å^{-1}$ et Q=0,35 $Å^{-1}$. Cette distribution correspond aux écarts entre intensité calculée et intensité expérimentale (mise à l'échelle) divisés par l'incertitude expérimentale correspondante. Dans l'hypothèse d'une distribution gaussienne des erreurs, les résidus normalisés sont compris à 95% entre -2 et 2 pour un modèle rendant compte des données.

D'après la courbe de la Figure 25, on observe systématiquement deux pics à Q=0,075$Å^{-1}$ et Q=0,13$Å^{-1}$ d'amplitude variable et pouvant atteindre des valeurs supérieures à 3. Ces pics correspondent à des valeurs localement plus basses des erreurs expérimentales. En particulier, on observe une discontinuité de la valeur des I(Q) au point Q=0,13 $Å^{-1}$ sur la courbe rééchantillonée (

Figure 26). La valeur au niveau de ce point sur la courbe rééchantillonée est localement trop faible, ce qui cause une augmentation importante de la valeur du résidu normalisé en ce point. En dehors de ces deux pics, la majorité des structures générées par Dadimodo présentent des valeurs de résidus normalisés comprises entre 2 et -2 jusqu'à 0,27-0,3 $Å^{-1}$. Au-delà de 0,3, les résidus normalisés dépassent 2 pour l'ensemble des structures.

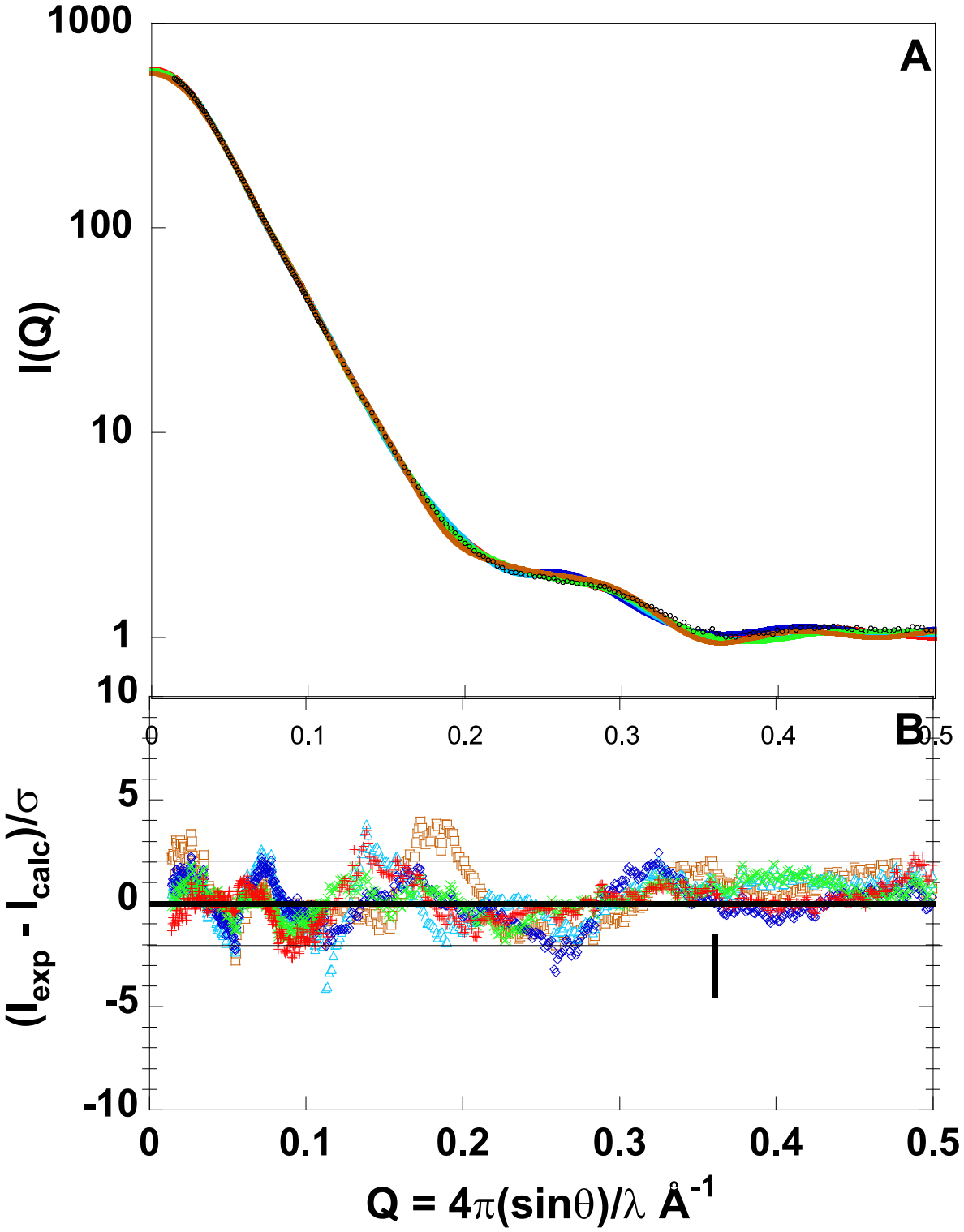


Figure 25 : Courbes de diffusion et valeur des résidus normalisés pour les structures Dadimodo. A : Courbes de diffusion des structures (une courbe par regroupement), rouge : regroupement 1 (le plus gros), bleu : regroupement 2 (2e plus gros), vert : regroupement 3, marron : regroupement 4, cyan : regroupement 5.

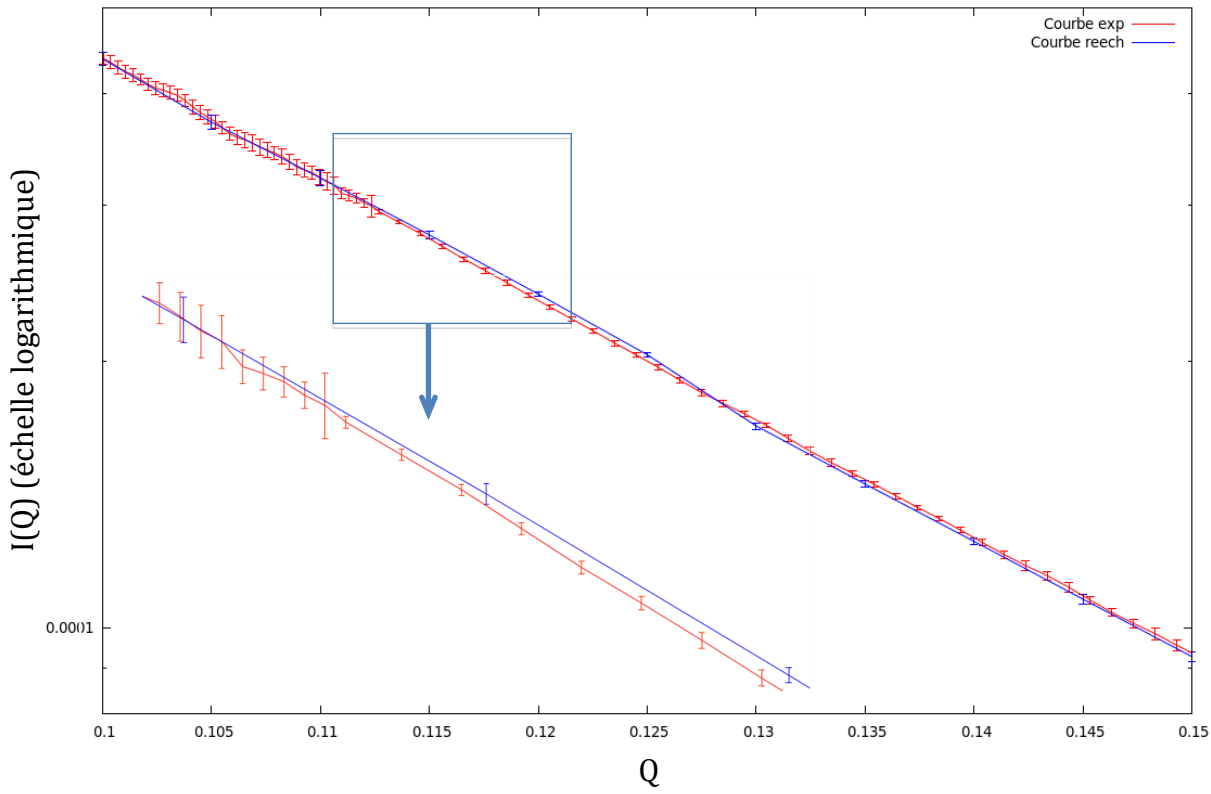


Figure 26 : Comparaison entre la courbe expérimentale (en vert) et la courbe rééchantillonée (en bleu) dans la région Q=0,105 Å^{-1} à Q= 0,155 Å^{-1}. Du fait du rééchantillonnage, on observe la présence de quelques points en dehors des incertitudes expérimentales dans la région de transition de la courbe (baisse brutale de la valeur des incertitudes). Il faut cependant noter que la valeur des incertitudes dans cette région est sous-évaluée.

L'accroissement des valeurs observé à partir de Q=0,25 Å^{-1} vient du fait que cette région de la courbe correspond aux petites distances. Comme on ne modifie pas la structure interne des domaines, Dadimodo n'est pas capable de satisfaire les contraintes pour ces types de distances. Néanmoins, les courbes obtenues par SASREF présentent des distributions de résiduelles réduites intégralement comprises entre -2 et 2 alors qu'il fonctionne en corps rigides. En fait ceci vient du paramétrage CRYSOL car lors du calcul Dadimodo, trois paramètres sont imposés : la couche d'eau à 0,030, le rayon atomique à 1.6112 (sa valeur par défaut) et le volume exclu à 96730 Å^3 (valeur par défaut). Lors des calculs SASREF, les paramètres sont libres. On peut néanmoins remarquer qu'en augmentant la valeur du volume exclu à 99500 Å^3 (soit une augmentation de 2,86%), la courbe des résiduelles de la meilleure structure Dadimodo est intégralement comprise entre les bornes. Les courbes de meilleur et plus mauvais χ CRYSOL sont représentées sur la Figure 27.

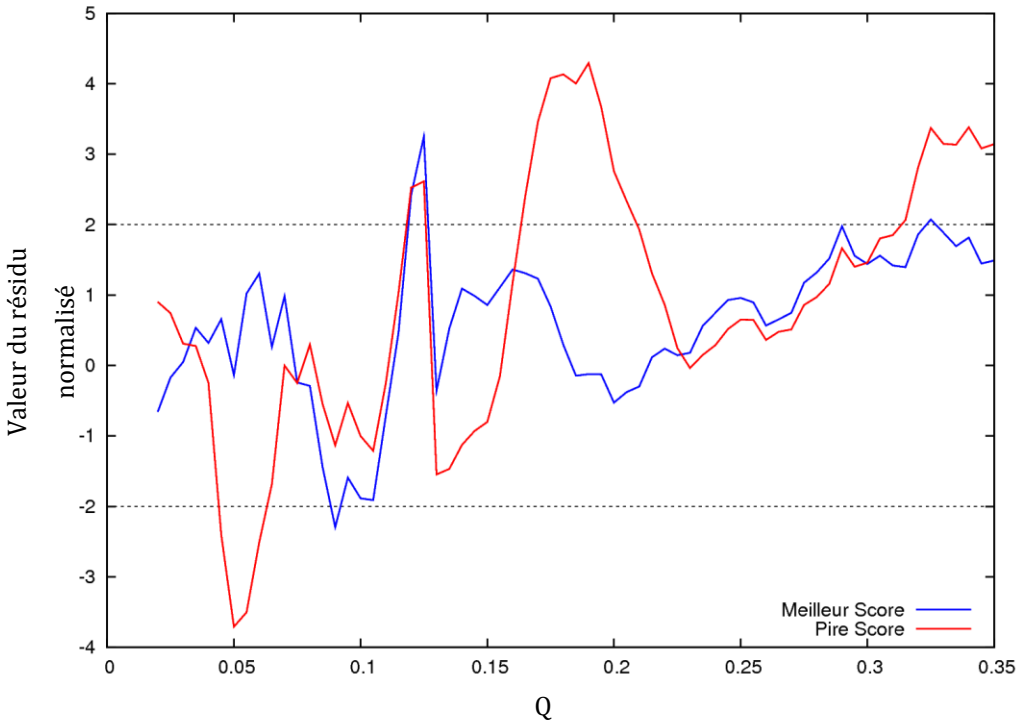


Figure 27 : Courbes des résiduelles réduites pour la structure de meilleur et de plus mauvais χ CRYSOL (respectivement en bleu et en rouge) après ajustement de la valeur du volume exclu à 99500 Å^3 au lieu de la valeur 96730 Å^3 utilisée pendant le calcul de Dadimodo.

6.2 Caractéristiques des structures

En examinant l'orientation du domaine Glycosyltransférase des 10 structures présentant les plus mauvais χ^2, on remarque que pour 7 d'entre elles les peptides de liaison séparant les domaines Glycosyltransférase et Transpeptidase se croisent. Ce croisement est dû à la rotation relative du domaine Glycosyltransférase par rapport au domaine Transpeptidase le long de l'axe de la molécule. Dans ces cas, les peptides de liaison sont très étirés et l'extrémité N-Terminale présente des conformations qui paraissent peu probables (compaction extrême, nœuds, etc.). Ces conformations « exotiques » semblent avoir été le meilleur compromis trouvé par l'algorithme pour approcher au mieux les données expérimentales, l'extrémité N-Terminale étant alors la seule partie de la structure modifiable sans dégrader le score. Parmi les sept structures aux peptides de liaison croisés, deux d'entre elles se trouvent dans le regroupement 1 qui regroupe huit des meilleures structures. En modifiant le seuil lors des regroupements, il n'est pas possible de les séparer de ce regroupement : ils ne sont pas donc pas à la bordure.

En comparant directement les structures 7 et 30 (Figure 28), il est possible de voir que malgré une superposition de volume de 75%, l'écart en termes de score et de χ est considérable. D'après la disposition de deux couples d'hélice, il est clair que l'orientation du domaine Glycosyltransférase de la structure 7 est à environ 180° de celui de la structure 30. Comme les volumes des domaines Glycosyltransférase se superposent très bien, l'élément qui varie le plus entre ces deux structures est l'extrémité N-Terminale. Pour vérifier que la position du N-Terminale de la structure 7 est bien optimale, une inversion des N-Terminale entre les structures 7 et 30 a été effectuée. Après cette inversion, le χ augmente de façon assez importante (Tableau 4).

Tableau 4 : χ avant et après échange des peptides de liaison N-Terminale

Structure	Origine N-Terminale	χ
30	**30**	0,803
30	7	1,521
7	30	2,404
7	**7**	1,705

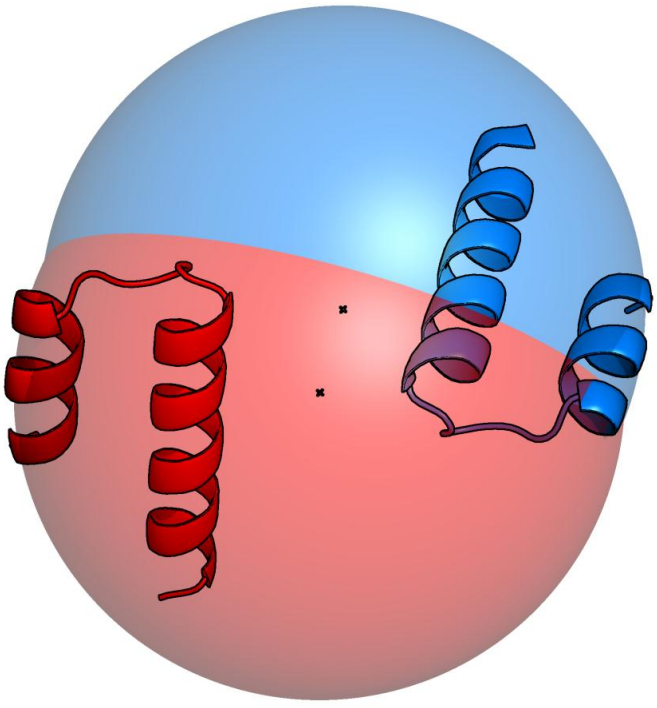


Figure 28 : Comparaison de deux solutions apparemment très proches en terme de position des centres de gravité et de recouvrement des volumes des domaines Glycosyltransférase mais présentant des valeurs de χ^2 très différentes. Les domaines Transpeptidase des deux solutions ont été superposés. Les positions des centres de gravité des deux domaines Glycosyltransférase sont matérialisées par les deux croix et les volumes des deux domaines par les deux sphères rouge et bleu. La figure montre clairement que les centres de gravités sont très proches l'un de l'autre et que les deux sphères se recouvrent largement. La même paire d'hélices est représentée pour les deux structures (représentation en ruban). La comparaison de la position de cette paire d'hélice dans les deux structures montre clairement que les deux solutions sont quasiment symétriques l'une de l'autre, résultant d'une rotation d'environ 180° du domaine Glycosyltransférase autour du grand axe de la protéine (perpendiculaire à la figure).

La dégradation importante de la valeur du χ est en grande partie due aux conformations des peptides de liaison séparant les domaines Glycosyltransférase et Transpeptidase. Pour basculer à 180°, le programme a généré des structures avec des peptides de liaison extrêmement tendus piégeant ainsi les structures dans un minimum local. La seule possibilité d'améliorer l'ajustement est alors de modifier l'extrémité N-terminale, et la conformation ainsi obtenue est optimale pour cette structure particulière. Le fait de changer l'extrémité N-terminale par celle d'une autre structure dégradera alors nécessairement la qualité de l'accord. Il est important de noter qu'en SAXS il est très difficile d'avoir une bonne approximation de la structure des parties flexibles car pendant le temps des différentes acquisitions, ces parties vont être en mouvement. Par conséquent, la structure de ces régions doit être considérée comme une moyenne (au sens de la contribution à la diffusion) des conformations explorées.

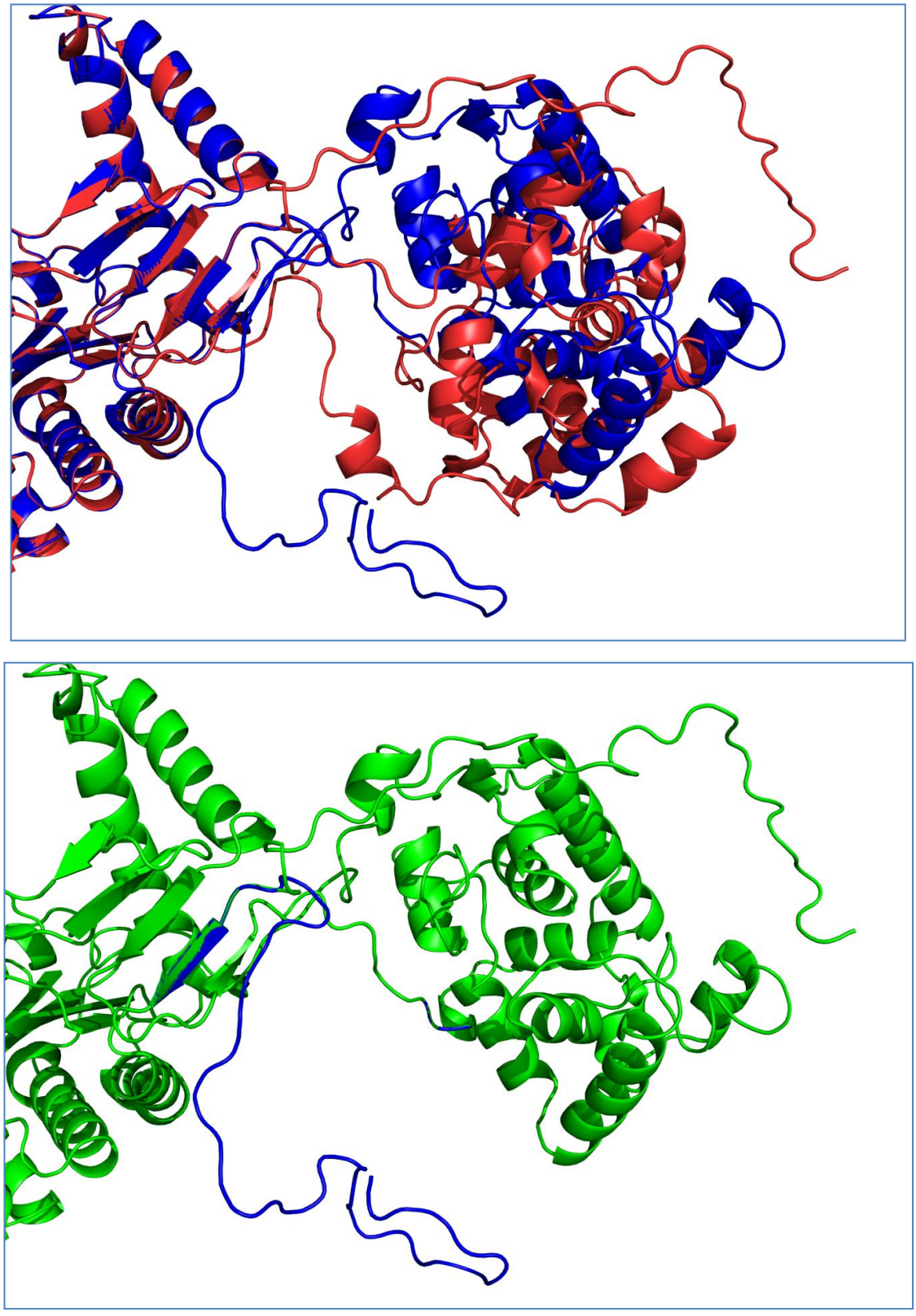


Figure 29 : L'analyse des structures suggère que la région N-terminale sert de variable d'ajustement permettant d'améliorer l'accord des courbes calculées et expérimentales. Nous avons vérifié cette hypothèse en construisant deux solutions chimères obtenues en échangeant les extrémités des deux solutions croisée et non-croisée. Dans la figure, nous avons présenté la structure de la solution non-croisée (en vert) et l'extrémité de la solution croisée (en bleu).

6.3 Position du domaine Glycosyltransférase

Il est maintenant intéressant de savoir s'il est possible de déterminer un positionnement du domaine Glycosyltransférase relativement au domaine Transpeptidase. La majorité des structures Dadimodo, représentées par le regroupement le plus important, placent le domaine Glycosyltransférase dans l'axe du domaine Transpeptidase (positions en rouge dans la Figure 30). Au niveau de l'orientation, les structures contenant des peptides de liaison croisés, et donc avec un domaine Glycosyltransférase à au moins 180° de son orientation initiale, sont nettement défavorisées en terme de χ (cf. 6.2). Ces orientations sont présentes dans près d'un quart des structures (10/43).

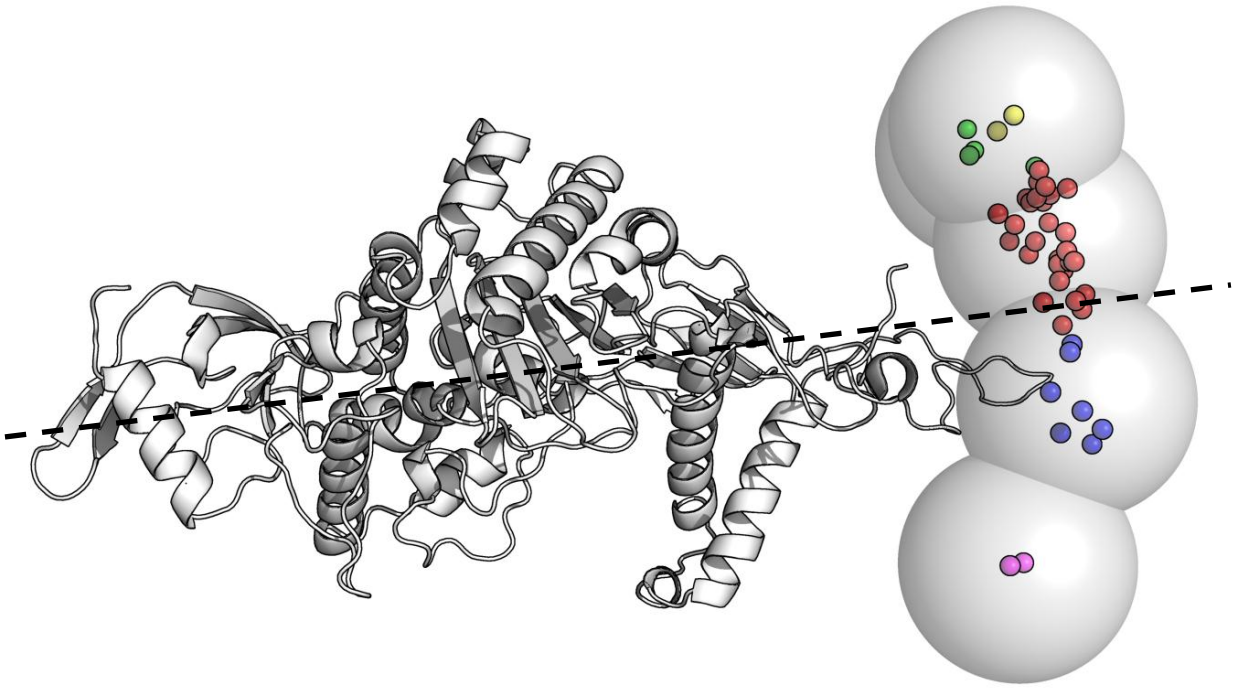


Figure 30 : Représentation de la distribution des centres de gravité des structures Dadimodo en vue de face. Le trait en pointillés représente l'axe du domaine Transpeptidase.

6.4 Comparaison avec les structures SASREF

Le programme SASREF a été utilisé pour construire des modèles en parallèle à l'utilisation de Dadimodo. La structure a été découpée en trois morceaux pour isoler le segment N-terminal et les domaines Glycosyltransférase et Transpeptidase. Le domaine Transpeptidase a été gardé dans un repère de référence tandis que le segment N-Terminal et le domaine Glycolsyltransférase ont été déplacés avec une contrainte de distance de 10, 15 et 20 Å entre les Cα de chaque côté de chaque coupure. Les trois fragments ont ensuite été ressoudés par ajustement des angles phi/psi à l'aide du logiciel Coot (Emsley & Cowtan 2004). Enfin, la position et la conformation du segment N-terminal ont été ajustées manuellement en utilisant les angles phi/psi.

6.4.1 Score

Les scores des structures SASREF et Dadimodo sont représentés par la Figure 32. Les structures SASREF portent le nom 3NFXX et 3CDXX. Les structures SASREF sont en grande majorité dans la 2e partie du classement (28 structures sur 33) et la meilleure structure est classée 12e. Il faut cependant noter qu'il est possible d'améliorer les scores des structures SASREF en modifiant manuellement la structure de l'extrémité N-terminale, SASREF ne pouvant le faire car fonctionnant en corps rigides. Après cette modification, les scores sont équivalents aux meilleures structures Dadimodo. Les courbes des résiduelles réduites (Figure 31) illustrent le très bon accord des meilleures structures à la fois pour SASREF que pour Dadimodo.

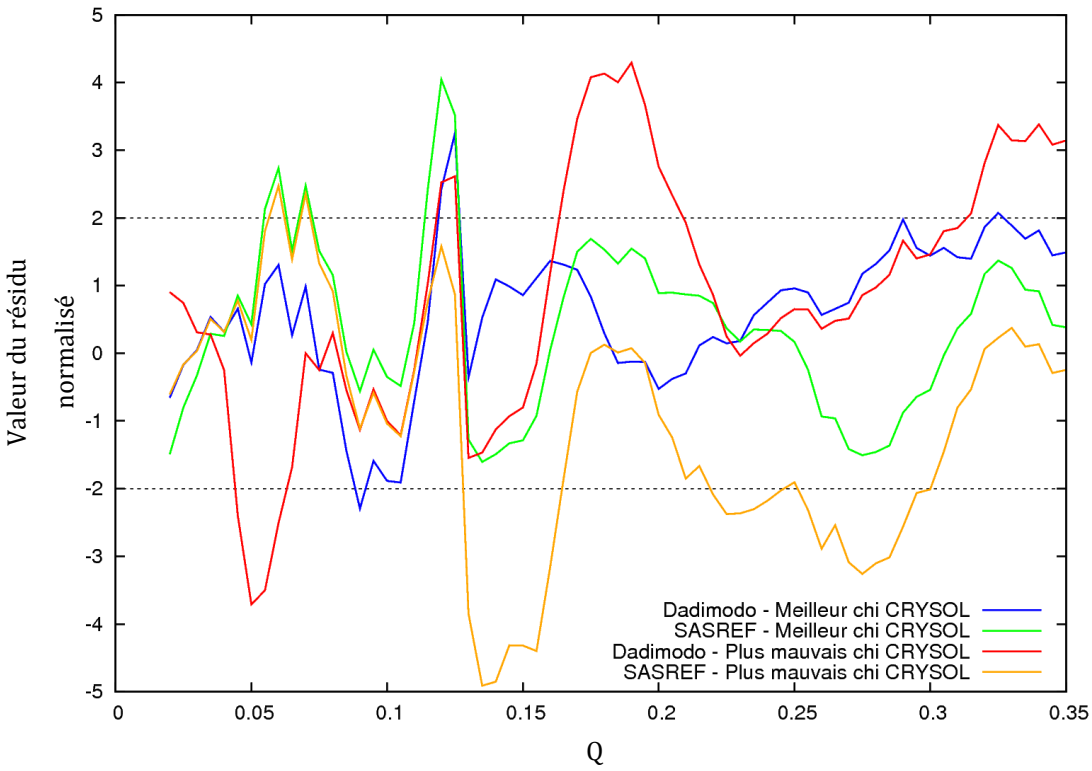


Figure 31 : Courbes des résiduelles réduites des structures de meilleur et de plus mauvais χ CRYSOL obtenues par SASREF et Dadimodo.

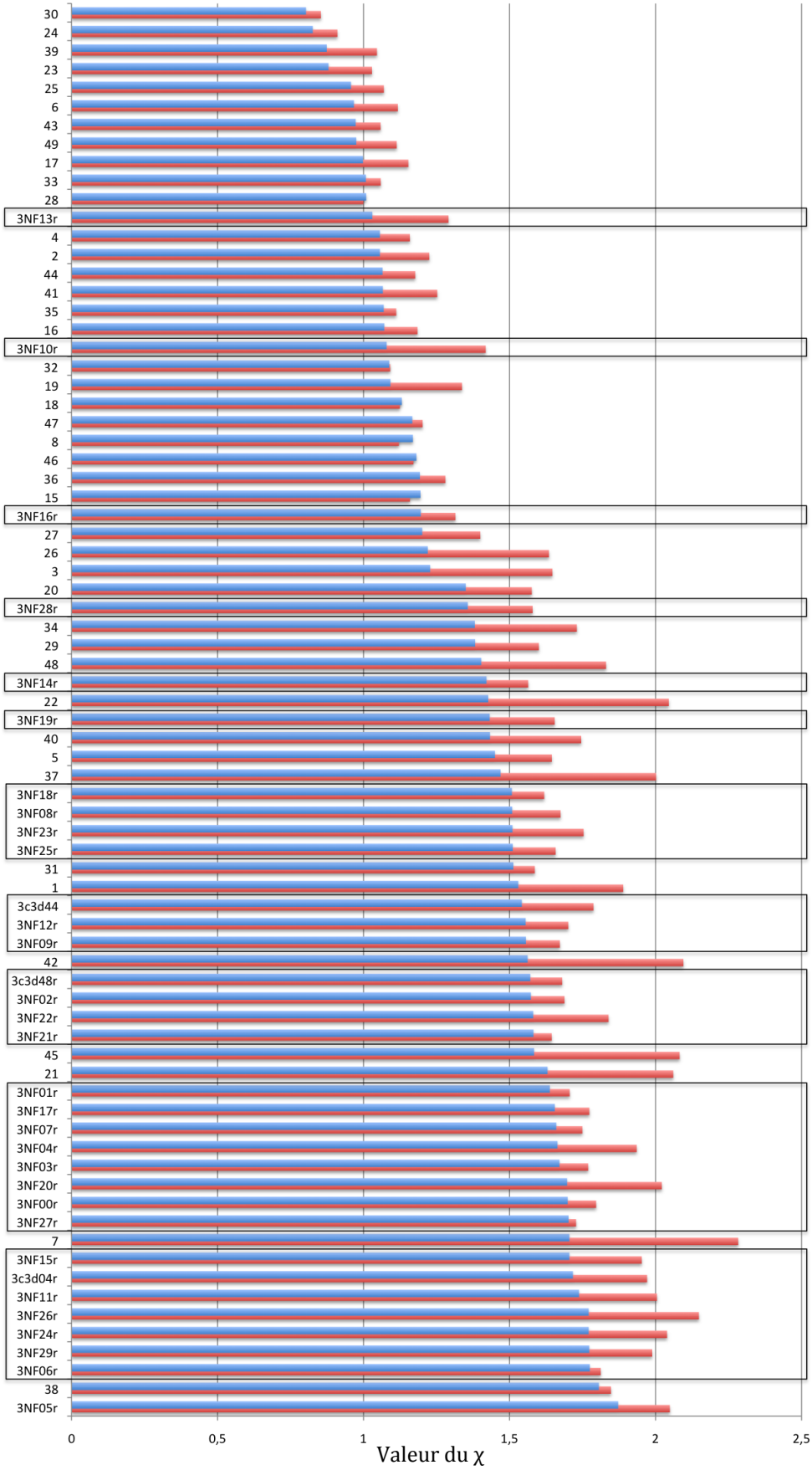


Figure 32 : χ CRYSOL (en bleu) et racine carrée du score Dadimodo (en rouge) des structures SASREF (dans les cadres noirs) et Dadimodo

6.4.2 Position du domaine Glycosyltransférase

La Figure 33 représente les regroupements des structures Dadimodo obtenus précédemment avec les regroupements des structures SASREF ; les sphères de couleur rouge correspondant aux structures SASREF. Les centres de gravité des structures SASREF ont des positions similaires aux structures des regroupements 1 et 2 (en rouge et en vert) et pour une structure avec le regroupement 4 (jaune). En revanche, les structures des regroupements 3 et 5 (en bleu et en violet) n'ont pas d'équivalents parmi les structures générées par SASREF.

La diversité des positions est inférieure à celle proposée par Dadimodo, tout l'espace occupé par les structures du regroupement violet et la moitié des bleus n'est pas exploré par SASREF. De plus, SASREF ne trouve pas de configurations avec des peptides de liaison croisés. Ces structures présentent ici un moins bon accord avec les données expérimentales mais 2 ou 3 résidus de plus dans les peptides de liaison auraient probablement permis l'adoption de conformations en bon accord avec les données.

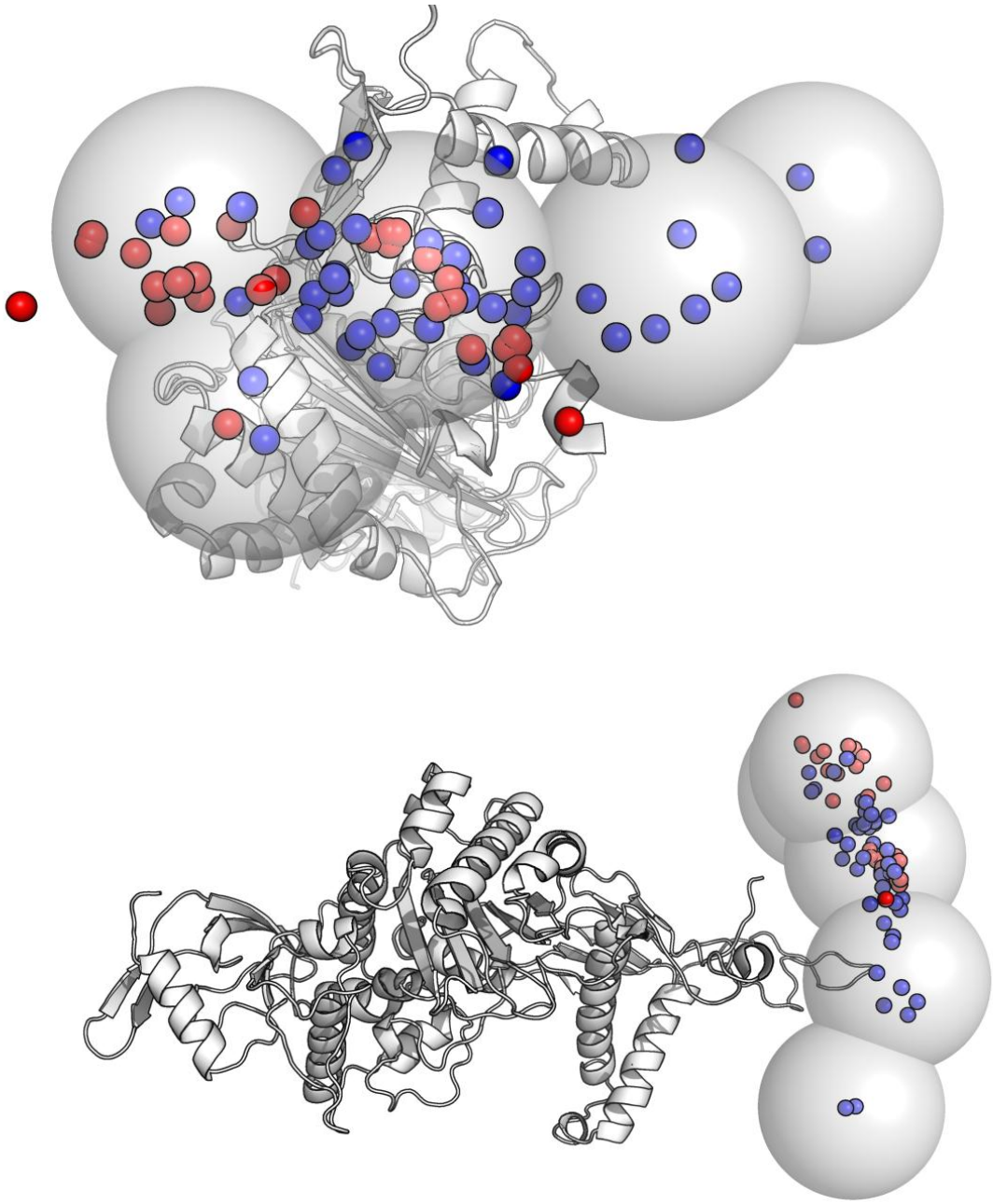


Figure 33 : Comparaison des solutions obtenues par Dadimodo et par SASREF en utilisant les mêmes données expérimentales. La représentation est identique à celle des figures précédentes. Les solutions obtenues par Dadimodo sont en bleu et celles avec SASREF en rouge. Les solutions obtenues avec Dadimodo ont été regroupées en utilisant la procédure du recouvrement de volume. A l'exception d'une solution (à l'extrême gauche de la figure du haut), toutes les solutions obtenues par SASREF sont à l'intérieur de l'espace exploré par Dadimodo. En revanche, les structures de Dadimodo à droite sur la figure du haut n'ont clairement pas de contrepartie dans l'espace exploré par SASREF.

7 CONCLUSION

7.1 COMPARAISON DES PROGRAMMES SASREF ET DADIMODO

L'approche proposée par Dadimodo permet d'obtenir un plus grand espace de solutions et donc d'une meilleure exploration de l'espace des conformations que le programme SASREF et ce d'une manière totalement automatisée. Il est important de noter qu'il a été possible d'obtenir des conformations avec des peptides de liaison croisés que SASREF n'est pas capable de proposer. Le principal défaut de Dadimodo est d'une part le temps de calcul nécessaire et le besoin de disposer d'un cluster pour pouvoir lancer les calculs massivement en parallèle. Le temps de calcul nécessaire varie en fonction de la taille de la protéine et dans le cas étudié dans ce chapitre chaque calcul a eu une durée moyenne de 3 jours. En comparaison, les calculs SASREF ne prennent que 2 à 3 heures mais nécessitent l'utilisation d'un deuxième programme pour affiner les structures et d'un ajustement manuel des structures pour améliorer l'accord avec les données expérimentales.

Pour revenir à Dadimodo, il faut aussi préciser qu'une connaissance approfondie du logiciel est nécessaire pour pouvoir l'appliquer à des cas d'études différents. En effet, chaque cas nécessite des paramètres particuliers (nombre de mutations, amplitudes des perturbations, …) voire le développement de nouvelles fonctions, comme le cas qui a été présenté dans ce chapitre. On peut aussi ajouter que la possibilité de combiner d'autres types de données expérimentales est un avantage majeur mais qui n'a pas pu être exploité dans ce cas précis. Par exemple, l'ajout des couplages dipolaires résiduels aurait permis de déterminer l'orientation du domaine Glycosyltransférase mais la grande taille de la molécule n'a pas permis cette étude complémentaire.

Il est aussi important de noter que l'obtention de modèles par l'une ou l'autre des méthodes n'est pas une fin en soi, leur analyse étant d'une importance capitale. Le développement des méthodes de regroupement par volume commun et l'utilisation de résidus normalisés pour l'évaluation de l'accord avec les données expérimentales sont deux méthodes qui ont été développées ici mais qui ne sont pas du tout généralisées.

7.2 INFORMATION BIOLOGIQUE

Les modèles en pseudo-atomes proposés par les programmes BUNCH et DAMMIN présentent une forme étendue, ce qui est compatible avec les paramètres structuraux obtenus par le SAXS. Cette forme étendue est confirmée par les approches SASREF et Dadimodo mais il n'est pas possible de définir l'orientation du domaine Glycosyltransférase car la distribution de la densité électronique change peu du fait de la forme relativement globulaire du domaine. Parmi l'ensemble des solutions proposées, aucune ne présente un contact entre les domaines Glycosyltransférase et Transpeptidase ce qui laisse la possibilité d'une mobilité mutuelle entre ces deux domaines. Cette remarque est compatible avec la grande flexibilité des domaines Glycosyltransférase et Transpeptidase observée dans les structures cristallographiques de PBP2 (Lovering et al. 2008). Dans cette étude, les structures sont en général pliées, avec un angle important entre les domaines Transpeptidase et Glycosyltransférase alors que les structures en solution sont plus étendues et les structures pliées sont minoritaires.

Le fait de trouver un ensemble de solutions n'est pas une preuve en soi de la mobilité des domaines, mais indique que les données expérimentales sont trop limitées pour restreindre l'espace des solutions. Cependant, comme les structures partagent certaines caractéristiques comme la forme étendue et l'absence de contacts entre les deux domaines il est possible de conclure qu'en solution et sans partenaire, PBP1b présente ces deux caractéristiques. La liberté de mouvement apportée par ces conformations étendues pourrait être un avantage dans l'environnement de PBP1b, permettant à chacun des domaines de se lier avec son partenaire plus facilement. L'étude présentée dans ce chapitre a été l'objet d'une publication (Macheboeuf et al. 2011).

DÉTERMINATION DE LA STRUCTURE D'UN DOMAINE PAR UNE MÉTHODE DE NOVO, APPLICATION À UN DOMAINE DE LA PROTÉINE P DU VIRUS PPRV

1 INTRODUCTION

Au chapitre précédent, Dadimodo a été utilisé dans un cas où l'un des domaines a une structure cristallographique connue tandis que pour obtenir un modèle de départ du deuxième domaine il a été nécessaire d'utiliser une technique de modélisation par homologie. Comme abordé au chapitre précédent, la modélisation par homologie permet d'obtenir des modèles de bonne qualité pour peu que l'on dispose de structures homologues et/ou de protéines de séquences fortement similaires. Mais comment faire lorsqu'on ne dispose pas de structures similaires dans les bases de données ? L'obtention de modèles fiables par des méthodes qui ne sont pas basées sur la similitude permettrait d'élargir le champ d'application de Dadimodo à des protéines dont la structure des domaines n'est pas modélisable par homologie.

1.1 MODÉLISATION DE STRUCTURE DE PROTÉINE À PARTIR DE LA SÉQUENCE

Les projets de génomique à grande échelle produisent une grande quantité de séquences qui, pour la plupart, codent pour des protéines dont la structure n'est pas connue. Être capable de prédire la structure de la protéine seulement à partir de sa séquence en acides aminés est donc un énorme défi car il peut permettre à terme d'automatiser tout le processus du séquençage à la détermination de la structure. Grâce aux avancées dans la connaissance du repliement des protéines, les méthodes de prédictions *de novo* ont vu le jour. Le terme « *de novo* », parfois nommé « nouveau repliement » (par opposition aux méthodes de reconnaissance de repliements) ou encore « *ab initio* », regroupe les méthodes de prédiction de structures qui ne se basent pas sur l'homologie entre la séquence cible et une séquence de la PDB pour la prédiction de la structure. Ce regroupement est issu de la classification des méthodes de prédiction de structures de protéine établie lors du Critical Assessment of Protein Structure Prediction

ou CASP (Moult et al. 2007) qui classe les méthodes en trois catégories : les méthodes par homologie, les méthodes par reconnaissance de repliement et les méthodes *de novo*.

Il est admis que l'état énergétique d'une protéine dans son état natif est à (ou est proche de) son minimum global. Anfinsen a suggéré, il y a près de 40 ans (Anfinsen 1973), que, comme le repliement des protéines se fait dans un temps fini, il doit exister un « chemin » menant de la protéine déstructurée à son état natif. Son hypothèse thermodynamique est que le repliement est un processus purement physique (et non pas biologique) qui ne dépend que de la séquence en acides aminés et du solvant. De nombreuses méthodes de prédiction *de novo* se sont basées sur cette hypothèse, mais depuis la fin des années 90, un nouveau courant fondé essentiellement sur l'utilisation de bases de données est apparu. C'est une méthode de ce courant qui sera présentée par la suite.

Par définition, les méthodes *de novo* doivent explorer un espace des conformations beaucoup plus grand par rapport aux méthodes de modélisation par homologie et de reconnaissance de repliements, qui limitent l'espace en explorant seulement les régions autour d'un ou de plusieurs modèles de départ. Explorer cet espace demande une grande puissance de calcul qui augmente de façon importante avec la taille de la séquence. Les méthodes actuelles sont limitées à des protéines et à des domaines de protéines de moins de 150 acides aminés environ. Une comparaison des différents programmes a été présentée par Helles (Helles 2008). Dans cette revue, l'auteur a regroupé les résultats obtenus par chacun des programmes sur leur jeu de test respectif. En conclusion, l'auteur retrouve à peu près le classement du CASP avec deux programmes présentant de bons résultats avec un RMSD inférieur à 4,6 Å sur un jeu de test de plus de 10 protéines : Rosetta (Rohl et al. 2004; Bradley et al. 2005) et I-TASSER (Wu et al. 2007). Par la suite, nous nous sommes particulièrement intéressés au programme Rosetta pour trois raisons : il possède une importante communauté d'utilisateurs, il est activement maintenu et ses auteurs cherchent à ajouter différents types de données expérimentales pour restreindre le nombre de structure différentes produites.

1.2 Combinaison avec des données expérimentales

L'incorporation d'un faible nombre de données expérimentales dans les méthodologies *de novo* permet d'obtenir rapidement une estimation du repliement global de la protéine (Bowers et al. 2000a; Shen. et al. 2009; Rohl et al. 2005). Dans ces travaux, ce sont des données issues de la RMN qui sont utilisées : les effets nucléaires Overhauser (NOE) et les déplacements chimiques. Les informations apportées par les corrélations NOE peuvent être assimilées à des contraintes de distance aussi utiles pour déterminer la forme globale que pour les affinements. Les déplacements chimiques apportent une information sur l'environnement immédiat de l'atome que l'on peut utiliser comme contraintes sur les angles dièdres. Ils seront abordés plus en détail dans le dernier chapitre de la thèse.

1.3 Cas d'étude avec des données RMN incomplètes

Lors de ma thèse, j'ai fait face à un problème de stabilité de protéine qui a empêché l'acquisition d'une grande partie des spectres RMN nécessaires à la résolution de la structure de manière classique. En effet, seuls quelques spectres ont pu être obtenus avec la protéine marquée 15N et par la suite nous n'avons pas réussi à obtenir de nouveau une protéine stable en solution. Aucun essai en double marquage 15N et 13C n'a pu abouti à cause de la dégradation rapide de la protéine dans le tube RMN. Comme nous ne disposons que de quelques NOE, cette protéine de petite taille (53 résidus) se présente comme un cas d'étude idéal pour tester une méthode *de novo* utilisant ces données expérimentales, appelée Rosetta. Parallèlement MODELLER a été utilisé pour générer des modèles par homologie de la protéine à partir de deux structures homologues. Dans ce chapitre seront présentées une étude de l'intérêt de l'ajout des contraintes expérimentales dans le cadre de l'utilisation d'une méthode de prédiction de structure *de novo* ainsi qu'une comparaison avec les modèles obtenus avec MODELLER. Nous mettrons ainsi en évidence l'intérêt de l'approche *de novo* avec contraintes.

2 ÉTUDE PAR RMN D'UN FRAGMENT DE PROTÉINE VIRALE

2.1 LE VIRUS DE LA PESTE DES PETITS RUMINANTS

Le virus de la peste des petits ruminants (PPRV) est un virus dont le génome est constitué d'une molécule d'ARN à polarité négative. Il appartient à la famille des Paramyxoviridés du genre Morbillivirus (Figure 34), tout comme les virus responsable de la peste bovine ou encore celui de la rougeole chez l'homme. Les espèces les plus touchées par le PPRV sont les petits ruminants, terme qui regroupe les ovins et les caprins. Le taux de morbidité est très élevé dans les populations à risque, principalement en Afrique subsaharienne et au Moyen-Orient. Le taux de mortalité associé est de 50 à 80% et, à l'heure actuelle, aucun vaccin ou molécule anti virale n'est disponible, contrairement aux virus de la peste bovine et de la rougeole.

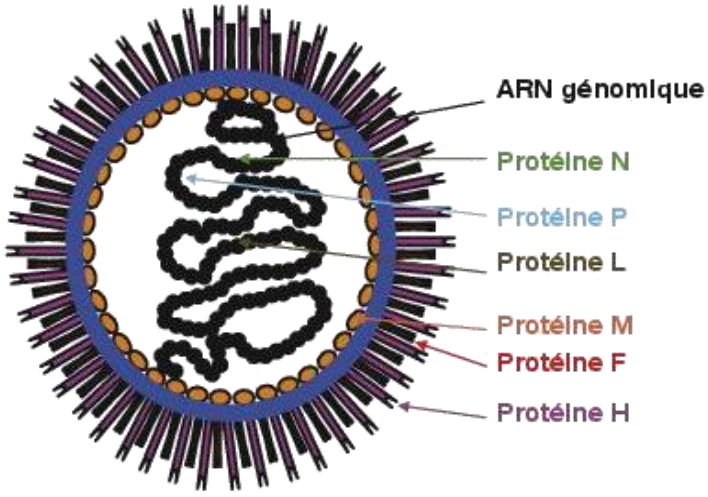


Figure 34 : Représentation schématique des différentes familles de protéines du virus de la peste des petits ruminants (PPRV)

Dans le cadre de la recherche d'un vaccin ou d'une molécule antivirale à moyen terme, les études se sont portées sur la phosphoprotéine P, un cofacteur essentiel de la polymérase du virus qui intervient à la fois dans le processus de réplication du virus et celui de la transcription de l'ARN génomique en ARN messager. La protéine P se lie d'une part à la polymérase L et d'autre part à la nucléocapside formée de l'ARN génomique encapsidé par la nucléoprotéine N (Figure 35). Dans ce chapitre, c'est le domaine C-

terminal de la protéine P, responsable de la liaison au complexe ARN-Protéine N, qui a été étudié (noté PPRV-PΔ459N ou P459N dans la suite).

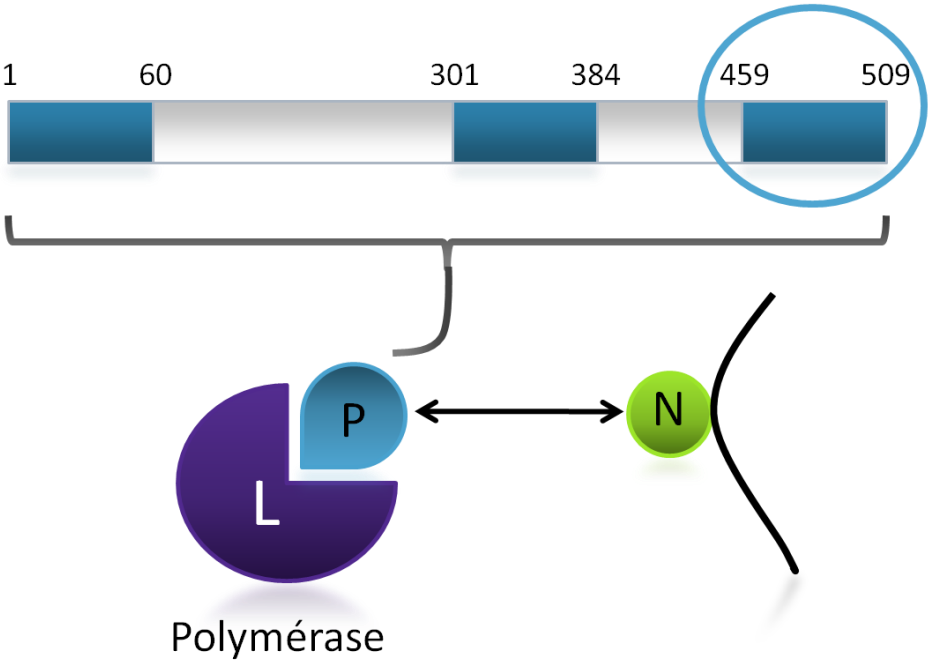


Figure 35 : Schéma de l'organisation structurale de la protéine P du PPRV (haut) et de ses interactions avec la polymérase (L) et le complexe ribonucléoprotéique ARN-Nucléoprotéine N (bas). Le cercle correspond au domaine C-terminal (résidus 459-509) étudié, impliqué dans la liaison avec la nucléocapside.
Les régions 1-60 et 301-384 correspondent respectivement au domaine de liaison putatif à la nucléoprotéine libre et un domaine d'oligomérisation.

Le domaine PPRV-PΔ459N possède une forte identité de séquence avec le domaine C-terminal de la protéine P du virus de la rougeole, appelé domaine XD et composé de trois hélices alpha. Ce domaine interagit avec la nucléocapside du virus de la rougeole et induit la structuration du domaine N-terminal de la protéine N. Deux structures obtenues par radiocristallographie sont déposées à la PDB (Figure 36) et ont été utilisées pour fabriquer des modèles par homologie avec MODELLER.

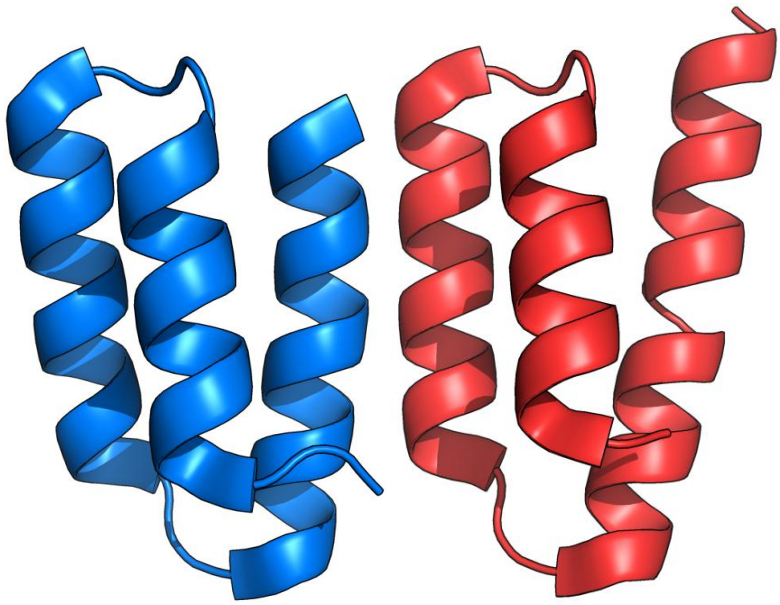


Figure 36 : Structures homologues au domaine C-terminal de la protéine P de PPRV : domaine XD de la protéine P du virus de la rougeole. En bleu, code PDB 1T6O (résidus 457-507, 61% d'identité de séquence). En rouge code PDB 1OKS, (résidus 459-507, 58% d'identité).

2.2 Préparation de l'échantillon de P459N

2.2.1 Production

Le détail du protocole expérimental se situe en annexe. Ici, seules les grandes étapes seront citées. Le plasmide de type pGEX (Novagen) avait été fourni par notre collaborateur. Ce plasmide contient la séquence codant pour le fragment C-terminal de la protéine P du PPRV (résidus 459-509, entouré en bleu sur la Figure 35), noté P459N, produit en fusion avec une étiquette glutathion S-transférase (GST). La séquence de la protéine est indiquée dans la Figure 37 avec le site de clivage de l'étiquette GST par la thrombine en couleur. La protéine GST-P459N a été obtenue par surexpression dans la souche BL21(DE3) d'*E. coli*. Le gel de contrôle SDS-PAGE post-production est satisfaisant, une bonne quantité de protéine GST-P459N ayant été produite. Le poids moléculaire de la protéine en fusion avec la GST est de 32,1 kDa et sa localisation sur le gel (Figure 38) est indiquée par une flèche.

```
          10         20         30         40         50         60
M SP ILGYWK IKGLVQ PTRLLLEY LEEKYEEH LY ERDEGDKWRNKK FELGLEFPNLPYY IDGDVKLTQ SMA

          80         90        100        110        120        130
I IRY IADKHNM LGGCPK ERAE ISM LEGAVLD IRYGVSR IAY SKDFETLKVD FLSKLP EM LKM FEDRLCHK

         150        160        170        180        190        200
TYLNGDHVTHPDFM LYDALDVVLYM DPMCLDAFPKLVC FKKR IEA IPQ IDKYLKSSKY IAWPLQGWQATF

         220        230        240        250        260        270
GGGDHPPKSDLVPRGSSSR SV IRS I IKSSKLN IDHKDYLLDLLNDVKGSKDLKEFHKM LTA ILAKQP
```

Figure 37 : Séquence de GST-P459N : le site de la coupure de la thrombine se situe au niveau de la séquence LVPR/GS (entre les régions en bleu et en rouge).

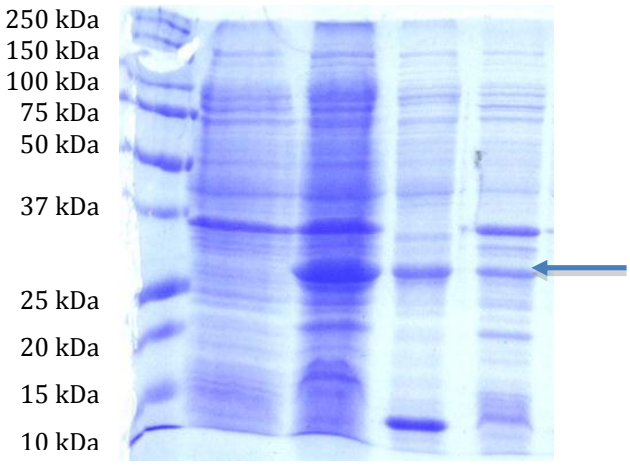


Figure 38 : Contrôle de la production de GST-P459N par SDS-PAGEà 12 %. Colonne 1 : marqueurs de poids moléculaire, colonne 2 : extrait total bactérien avant induction, colonne 3 : extrait total bactérien après induction, colonne 4 : surnageant après ultracentrifugation, colonne 5: culot d'ultracentrifugation. La flèche indique la bande correspondant au poids moléculaire de la protéine GST-P459N.

2.2.2 Purification

La purification s'est déroulée en trois étapes. La première est le dépôt sur une colonne contenant des groupements glutathions (GSTrap FF) avec un tampon composé de Tris-HCl 50 mM pH 8, de 150 mNaCl et de 2.5 mM $CaCl_2$. La deuxième étape est la coupure du tag GST qui est réalisée sur la colonne après injection de thrombine. On laisse incuber pendant au moins une nuit à température ambiante. La troisième étape est l'élution pour isoler la protéine P459N du tag GST ou de la protéine GST-P459N non clivée, qui restent accrochés sur la colonne. On élue avec le même tampon que précédemment, puis on lave la colonne avec une solution contenant du glutathion pour évacuer les protéines qui contiennent un tag

GST. La coupure est bien réalisée, comme l'indique le gel de contrôle en Figure 39. Les chromatogrammes sont disponibles en annexes.

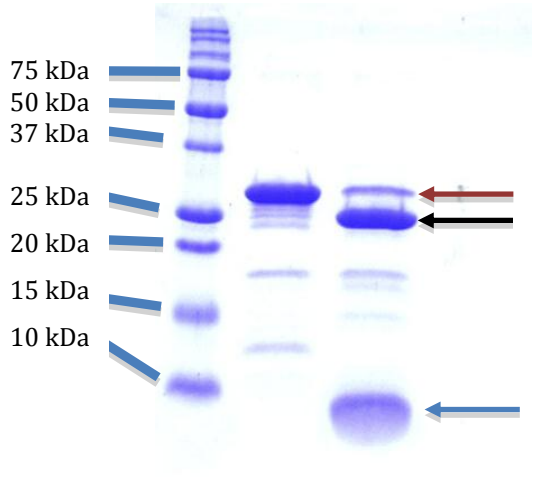


Figure 39 : Coupure de la protéine GST-P459N par la thrombine analysée par SDS-PAGE. La flèche bleue indique la protéine P459N (5,9 kDa), la flèche noire le tag GST seul (26,2 kDa) et la flèche rouge ce qui reste de GST-P459N (32,1 kDa).

2.2.3 Problème de stabilité

En dépit de rendements très variables, il a toujours été possible d'obtenir suffisamment de protéine pour fabriquer un échantillon pour la RMN. Le problème est qu'une fois dans le tube, la protéine se dégrade très rapidement (en moins d'une heure). Les analyses par spectrométrie de masse montrent que la protéine est à la fois dégradée en N-terminal et coupée au milieu de la séquence, probablement entre deux hélices. L'utilisation d'antiprotéases (Pefabloc, Roche) n'y a rien changé. Nous n'avons réussi à faire des mesures qu'à partir d'un seul échantillon marqué en 15N, tous les échantillons produits se dégradant trop rapidement. A ce jour, nous n'avons pu isoler la cause de ce problème. Il est cependant possible que le problème vienne de l'étape de la coupure du tag GST et que l'on introduise une protéase contaminante avec la thrombine. Une manière de s'en affranchir consisterait à produire P459N avec une étiquette plus petite (poly-histidine par exemple) que l'on n'aurait pas besoin de couper. Mais, ne disposant pas des compétences requises, je n'ai pas pu changer de vecteur d'expression.

2.3 ATTRIBUTION DES FRÉQUENCES DE RÉSONANCE

Trois types d'acquisitions ont pu être faites sur la protéine marquée 15N : HSQC-N15, une NOESY-HSQC-N15 et une TOCSY-HSQC-N15. Il n'a pas été possible d'obtenir un spectre NOE 2D. L'attribution de 86% des fréquences de résonance des protons amides de la protéine (43 résidus sur 50 résidus non prolines) a pu être effectuée, de ~80% de protons Hα et Hβ et ~70% de protons méthyles ainsi que des protons aromatiques Hδ de Tyr480 et Phe497.

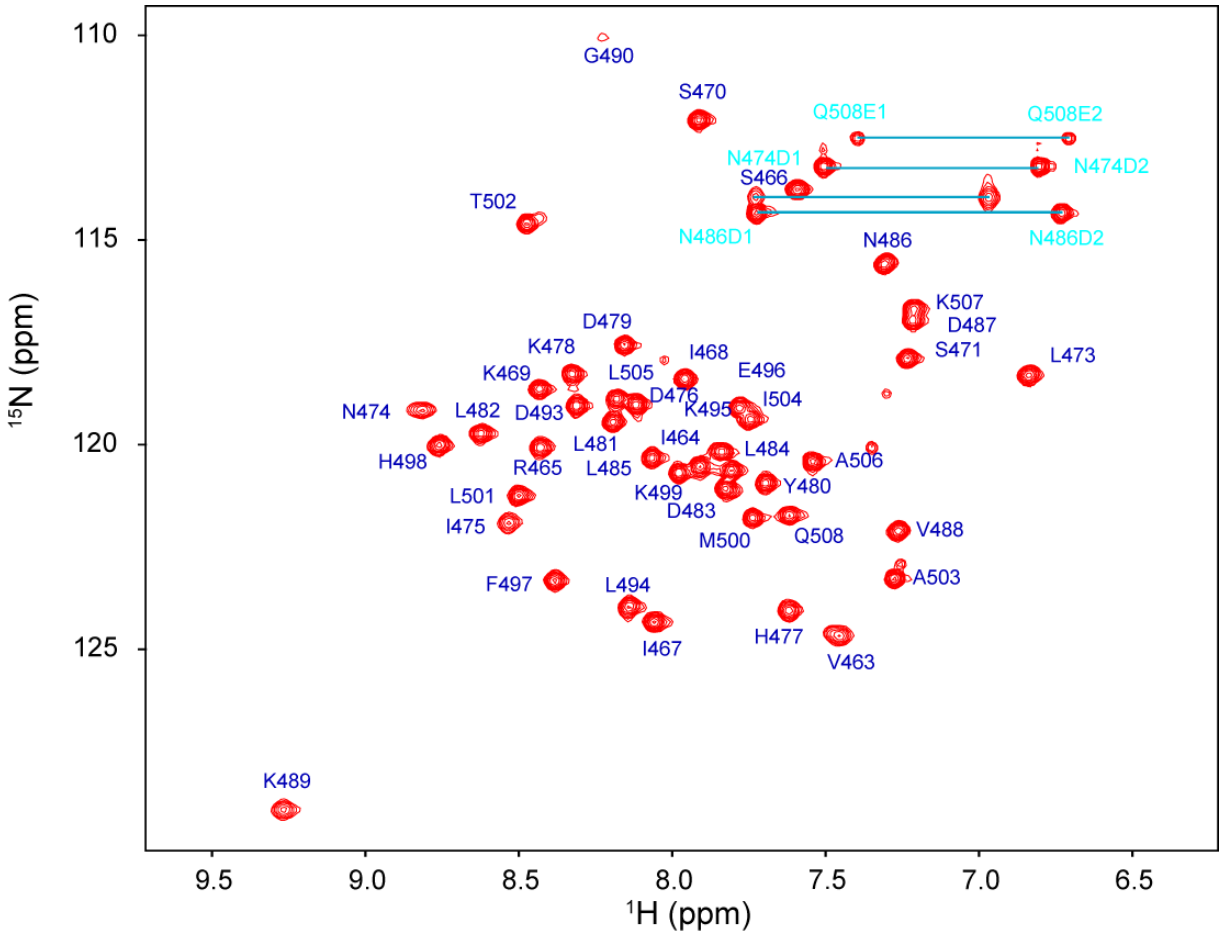


Figure 40 : Spectre HSQC-N15 du domaine C-terminal de la protéine P de PPRV (P459N) à 600 MHz et 293 K

A partir des spectres obtenus, il a été possible de définir, en plus des 149 NOE intra et inter résiduels, 8 NOE longues distances. Les NOE intra résiduels sont les NOE HN-HA ou HN-H de la chaîne latérale dans un même résidu et les inter-résiduels sont les NOE HN_i-HN_{i+1}, caractéristiques de la structuration en hélice alpha. Les effets NOE longues

distances déterminés impliquent des atomes de résidus différents et sont répertoriés dans le Tableau 5.

Tableau 5 : Effets NOE longues distances

Numéro du résidu i*	**Type du résidu**	**Type d'atome**	**Numéro du résidu j***	**Type du résidu**	**Type d'atome**	**Limite inférieure**	**Limite supérieure**
8 (464)	Ile	HD1 ou HG2	**42 (498)**	His	HN	2,44	3,20
17 (473)	Leu	HD*	**22 (478)**	Lys	HN	2,95	3,86
17 (473)	Leu	HD*	**52 (508)**	Gln	HE2*	2,98	3,82
29 (485)	Leu	HD*	**30 (486)**	Asn	HN	2,96	3,87
29 (485)	Leu	HD*	**9 (465)**	Arg	HN	2,45	3,21
41 (497)	Phe	HD1	**29 (485)**	Leu	HN	2,64	3,45
41 (497)	Phe	HZ	**44 (500)**	Met	HN	2,70	3,53
49 (505)	Leu	HD*	**15 (471)**	Ser	HN	2,70	3,53

*Le numéro de résidu est donné par rapport au premier résidu du fragment P459N. La numérotation dans la protéine P entière est donnée entre parenthèses.

3 PRÉSENTATION DE ROSETTA

Rosetta est un programme de détermination de modèles de protéines *de novo* qui utilise une méthode d'assemblage par fragments couplée à des potentiels statistiques (Rohl et al. 2004). Ce programme, conçu par l'équipe de David Baker à l'université de Washington, a déjà obtenu de bons résultats dans de nombreuses applications et a toujours été classé premier dans la catégorie *de novo* dans les expériences CASP (Critical Assessment of Techniques for Protein Structure Prediction, http://predictioncenter.org/) depuis CASP3. L'idée de départ derrière ce programme est de décomposer le problème du repliement des protéines en plusieurs parties et de les résoudre avec des méthodes différentes. Par conséquent, Rosetta n'est pas un algorithme au sens strict du terme, mais un mélange d'algorithmes, de bases de données et de potentiels statistiques. Le fonctionnement général du programme sera présenté dans les paragraphes suivants et est illustré par la

Figure 41. Son succès repose à la fois sur sa stratégie de balayage de l'espace conformationnel et sur l'étape finale d'affinement.

Depuis sa création, Rosetta a connu plusieurs adaptations et développements pour lui permettre de gérer d'autres cas que la prédiction de repliement. Les cas tels que l'arrimage moléculaire (docking), la détermination de structures à partir de données expérimentales incomplètes (Bowers et al. 2000), les interactions protéine-protéine (Kortemme et al. 2004) et protéine-ADN (Havranek et al. 2004) sont non seulement fonctionnels, mais donnent des résultats satisfaisants. Le groupe de D. Baker s'est aussi intéressé à la démarche inverse, le Protein Design. Celle-ci consiste à prédire une séquence de protéine qui adopterait un repliement (Kuhlman et al. 2003). Dans le cadre de ce sujet, c'est la procédure de détermination de structure à partir de données expérimentales incomplètes qui sera présentée et utilisée.

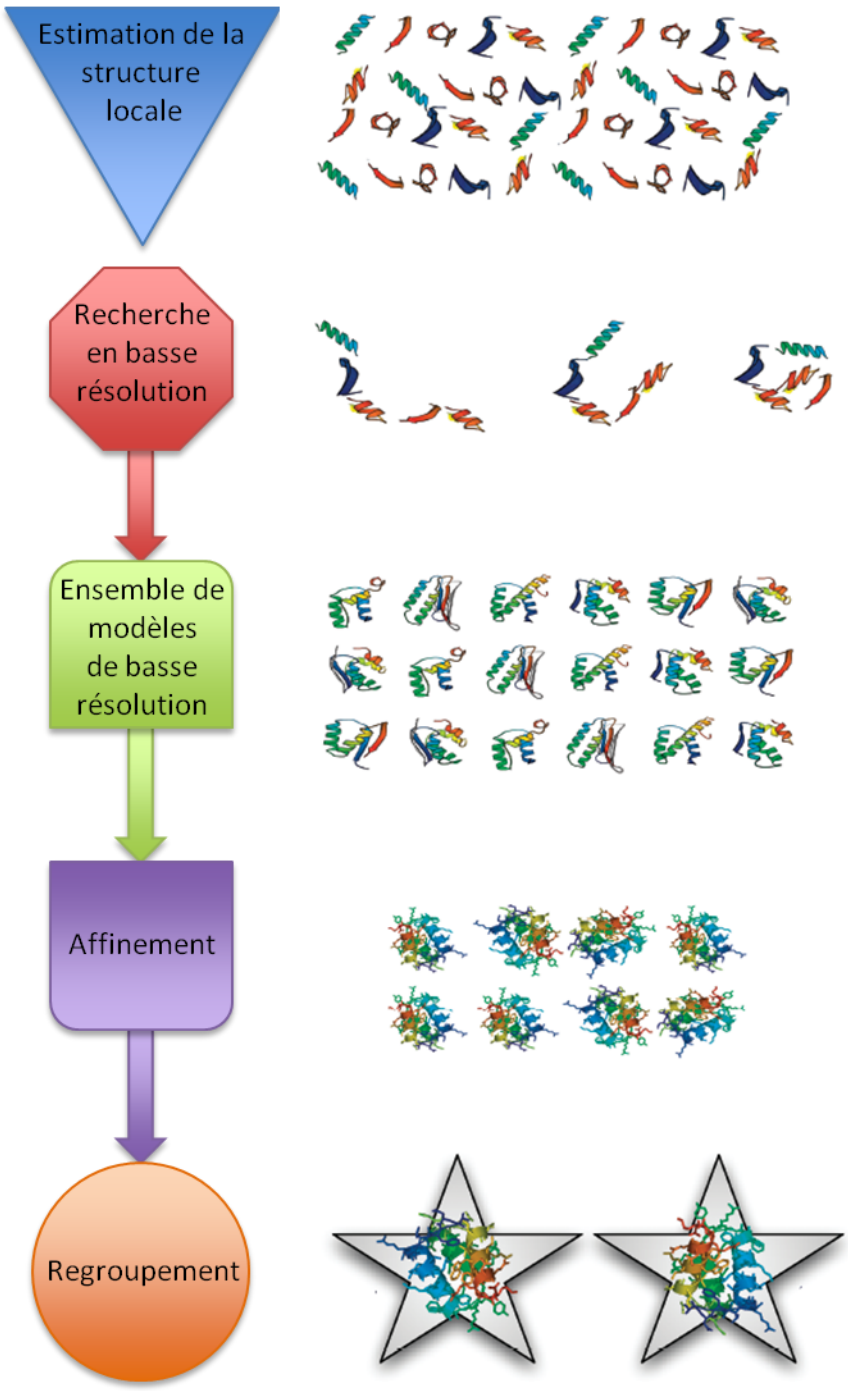


Figure 41 : Protocole Rosetta pour la prédiction de structure. La prédiction de structure Rosetta se déroule en deux phases, une phase dite de basse résolution (étapes 1-3) pour la recherche de la topologie globale en utilisant un assemblage par fragments et une fonction de score et une phase dite de haute résolution (étape 4) pour l'affinement des structures en utilisant des rotamères, de petits changements d'angles au sein du squelette et une fonction de score plus complète du point de vue physique.

3.1 Exploration de l'espace des conformations

L'exploration de l'espace des conformations est assurée par une méthode d'assemblage par fragments de 3 ou 9 résidus (Simons et al. 1997). La première partie de cette méthode consiste à découper de manière aléatoire la séquence de la protéine étudiée en segments de 3 à 9 résidus de longueur. Pour chacun de ces segments, le programme cherche un fragment de même séquence dans les structures contenues dans la PDB. L'intérêt de la constitution d'une librairie de fragments est de gagner un temps considérable en réduisant de manière drastique l'espace des conformations à explorer. Comme les fragments recherchés dans la PDB sont courts, la présence de protéines homologues et/ou à similarité élevée n'est pas indispensable.

La deuxième partie consiste en l'assemblage lui-même, testant toutes les combinaisons possibles et les évaluant avec la fonction de score. Les structures tertiaires sont générées en utilisant une recherche Monte-Carlo des combinaisons possibles des structures locales minimisant la fonction de score. Cette dernière prend en compte les paramètres comme la compacité et l'enfouissement des hydrophobes ou encore les interactions tels que les appariements de brins et l'électrostatique. Les modèles de la protéine ainsi générés sont des modèles pour lesquels les chaînes latérales sont remplacées par des centroïdes : on les considère donc de « basse résolution ». Ces choix se justifient au niveau du temps de calcul car grâce à une fonction de score « simple » et à la représentation par pseudo-atomes, le programme peut générer un grand nombre de modèles dans un temps restreint.

3.2 Affinement des modèles

Une fois l'ensemble de modèles généré par la méthode d'assemblage de fragment, le programme sélectionne les modèles présentant les meilleurs scores pour les affiner. L'affinement permet de passer d'une phase de structures « basse résolution » à une phase de détail atomique dans laquelle on introduit les atomes des chaînes latérales en blocs rigides. Chaque modèle de « bon » score sert de point de départ à la phase d'affinement. Dans cette phase, le programme joue en grande partie sur l'orientation des chaînes latérales et les modifications sur le squelette se limitent à quelques perturbations sur les angles dièdres. L'introduction du détail atomique est accompagnée par le changement de

la fonction de score afin de prendre en charge cette nouvelle représentation en tout atome. Les termes de solvatation et de liaison hydrogène sont entre autres présents dans cette nouvelle fonction de score.

Les modèles tout atomes obtenus sont ensuite regroupés par RMSD (entre elles) et par la valeur du deuxième score. Les centres des regroupements les plus gros sont ensuite choisis pour représenter les meilleures conformations obtenues. En effet, les plus gros regroupements peuvent être vus comme des minimums potentiels dans le paysage énergétique de la protéine car après minimisations le programme converge vers ces modèles.

3.3 Prise en charge des contraintes expérimentales

L'idée de départ pour l'étude de la structure de P459N était d'utiliser RosettaNMR (Rohl 2005), une variante du protocole Rosetta conçue spécifiquement pour utiliser les données issues des expériences RMN. Cependant, RosettaNMR n'est plus développé car il est basé sur la version 2.3 de Rosetta, aussi appelée Rosetta++. La nouvelle version de Rosetta, Rosetta 3.1 aussi appelé minirosetta, est issue d'une réécriture complète du code, nécessaire pour faciliter le développement des versions futures en le rendant plus modulable. Pour utiliser les données RMN, il est maintenant nécessaire d'utiliser CS-Rosetta pour les déplacements chimiques ou utiliser le système de contraintes de distances de Rosetta pour les corrélations NOE, ce qui a été fait ici.

Rosetta utilise les contraintes expérimentales lors de la construction de la banque de fragments, de leur assemblage ainsi que pour l'étape d'affinement. Il est cependant nécessaire d'ajouter que seules les données concernant les atomes du squelette (C, Cα, N, Hα, HN) sont prises en compte lors des deux premières étapes puisqu'elles font appel à une représentation simplifiée des résidus (sous forme de centroïdes). L'intérêt d'utiliser les contraintes lors de la constitution de la banque de fragments est de sélectionner les fragments les plus proches de la conformation attendue et ainsi gagner du temps lors de l'exploration de l'espace des conformations.

Les valeurs de distances sont stockées dans un fichier de contrainte Rosetta 3. Pour chaque contrainte, il faut définir les atomes impliqués, les distances et la fonction de score à utiliser. Dans le cadre de l'utilisation des NOE, c'est la fonction BOUNDED qui a été utilisée. Cette fonction utilise un potentiel « en casserole » dont le seuil, intervalle entre

les bornes supérieures et inférieures (« lb » et « ub » sur la Figure 42), correspond aux valeurs entre la borne inférieure et supérieure (Équation 4).

$$f(x) = \begin{cases} \left(\frac{x-lb}{sd}\right)^2 & \text{si} \quad x < lb \\ 0 & \text{si} \quad lb \leq x < ub \\ \left(\frac{x-ub}{sd}\right)^2 & \text{si} \quad ub < x \leq ub + rswitch \cdot sd \\ \frac{1}{sd}\left(x\left(ub + rswitch \cdot sd\right)\right) + \left(\frac{rswitch \cdot sd}{sd}\right)^2 & \text{si} \quad x > ub + rswitch \cdot sd \end{cases}$$

Équation 4 : Formule de la fonction de score associée aux contraintes de distance. Le paramètre rswitch est fixé à 0,5.

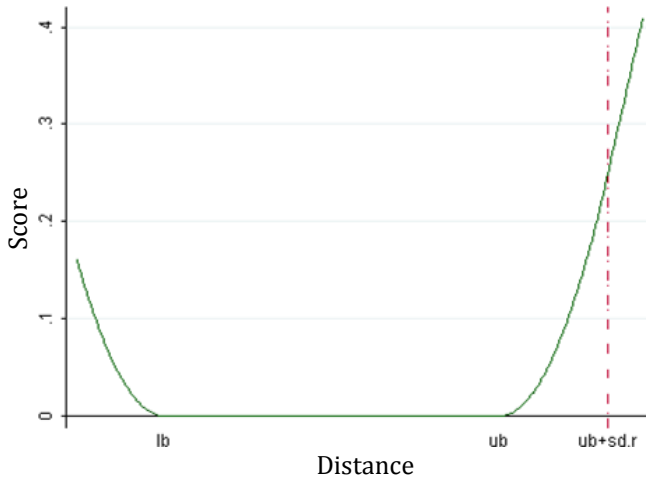


Figure 42 : Représentation du score en fonction de la distance. La valeur lb correspond à la borne inférieure, ub à la borne supérieure. ub+sd.r est la valeur de la borne supérieure plus l'écart type fixé à 0,5

3.4 Limitations du programme

En RMN, des protons ayant des déplacements chimiques identiques donnent lieu à des contraintes de distance ambiguës. La leucine de la Figure 43 illustre ce problème. Dans le spectre RMN, les fréquences de résonance des protons correspondant aux atomes 1HD1, 2HD1 et 3HD1 de ce résidu sont moyennées à cause de la libre rotation des groupements méthyles. Dans certains cas, l'environnement des méthyles HD1 et HD2 peut être tel que le déplacement chimique est même identique pour les six protons. Certains programmes, comme CNS, permettent l'utilisation d'opérateurs spécifiques pour gérer ce cas de figure en donnant la possibilité au programme de gérer la distance moyenne des protons, grâce au concept de pseudo-atome. Dans ce cas, le programme ne cherche pas spécifiquement à

satisfaire la contrainte 1HD1-X, mais fait en sorte que la moyenne des distances 3HD1-X, 2HD1-X, 1HD1-X soit compatible avec la contrainte. Comme Rosetta n'est pas capable de le faire, la contrainte de distance est modifiée pour porter sur l'atome lourd au lieu du proton et sa valeur est augmentée de 1,3 Å. Pour gérer le problème de stéréospécificité, la contrainte est appliquée systématiquement sur le premier atome lourd (atome CD1 sur la Figure 43). Cette décision est purement arbitraire, mais elle était la plus rapide à mettre en œuvre. Dans l'idéal il faudrait soit dupliquer la contrainte pour affecter les deux atomes lourds, soit introduire des pseudo-atomes. Il est maintenant possible de modifier la bibliothèque des acides aminés de Rosetta et donc d'appliquer la deuxième solution, mais je n'ai pu l'appliquer dans le temps imparti.

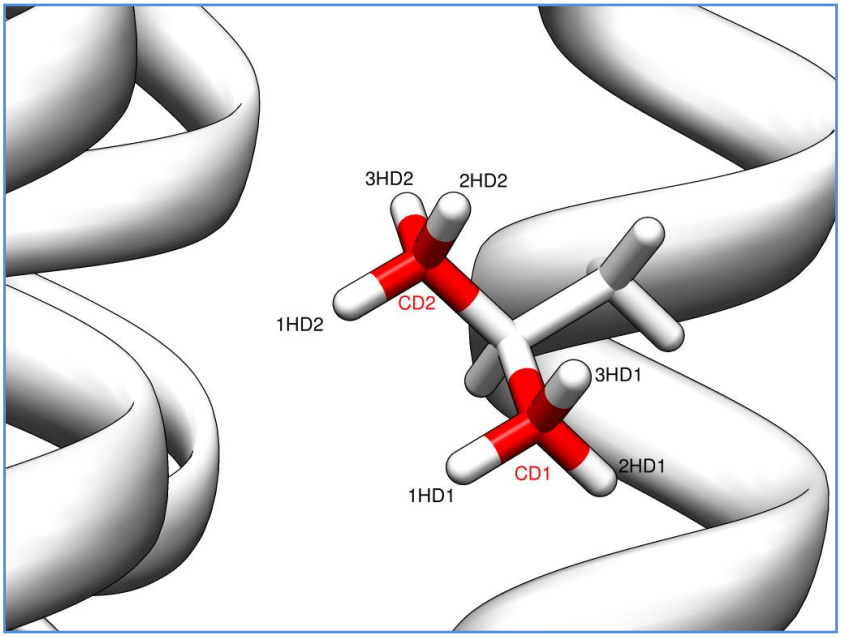


Figure 43 : Représentation d'atomes dont l'attribution stéréospécifique n'a pas été possible (atomes de carbone en rouge)

3.5 Du point de vue de l'utilisateur

Rosetta se présente sous deux formes : un serveur web appelé Robetta (http://robetta.bakerlab.org) automatisant les processus de création de la banque de fragments et de génération/regroupement des modèles et un regroupement de logiciels en téléchargement libre, appelé minirosetta, contenant les différents protocoles Rosetta (prédiction *de novo*, docking, …). Si Robetta est de loin la version la plus simple d'accès, les paramètres avancés du programme tels que le nombre de modèles à générer ou le

RMSD seuil pour le regroupement sont automatiquement ajustés sans laisser le moindre choix à l'utilisateur. Son utilisation est surtout utile pour évaluer rapidement les possibilités de Rosetta avant de s'investir davantage.

L'investissement en temps demandé par Rosetta est loin d'être négligeable, que ce soit pour son installation ou son usage. L'installation sur un ordinateur 64 bits sous Ubuntu 9.04 a nécessité la compilation des exécutables, cette opération a duré au moins une semaine pour régler les problèmes de dépendances des différentes librairies et les différences de version des compilateurs.

L'utilisation n'est pas non plus simple, même si on peut noter la présence d'un manuel et d'un wiki disponibles sur le site Rosetta Commons. Ces derniers, bien que décrivant l'utilisation de certaines fonctions, sont loin d'être exhaustifs et en règle générale ne proposent que des exemples d'utilisation pour des cas simples (pas de contraintes, pas d'ajustement de score, etc.). Le pire étant que les exemples de fichiers de configuration fournis ne contiennent aucun commentaire permettant de savoir à quoi correspondent les valeurs des différents paramètres. Heureusement, il est possible de trouver plusieurs revues (Kaufmann et al. 2010; Rohl et al. 2004) permettant d'accéder à des exemples plus complets et beaucoup mieux commentés.

Ce manque de documentation et la difficulté d'utilisation du programme est assez étonnant car l'équipe de D. Baker est aussi responsable de deux logiciels grand public : Rosetta@Home, un logiciel de calcul distribué pour la recherche du minimum global de structures complexes, et surtout FoldIt (http://fold.it/), un jeu permettant de s'initier à la problématique du repliement des protéines. Bien conçu et plutôt ergonomique, ce logiciel devrait être une bonne source d'inspiration pour Rosetta.

4 APPLICATION AU DOMAINE C-TERMINAL DE LA PROTÉINE P DE PPRV

4.1 CALIBRATION DES NOE

Les valeurs des NOE correspondant à des intensités, il est nécessaire de les convertir en distances pour les intégrer dans Rosetta 3 et MODELLER. Pour convertir les NOE en contraintes de distance, il est courant d'utiliser un facteur fixe. Cependant, cela introduit un biais lié à la dynamique du système. Dans le cas de la protéine P, comme nous savions qu'elle était essentiellement constituée d'hélices *a priori*, nous avons utilisé la distribution des distances HN-HA intra résiduels et HN_i-HN_{i+1} inter résiduels dans les hélices α pour servir de référence. Ces distributions sont déterminées par le programme STARS (Zheng & Yang 2005) qui permet de calculer, sur la base de 500 fichiers PDB, les distances moyennes et les écarts type de ces deux couples dans les hélices α. Ces valeurs moyennes permettent ensuite de déterminer la valeur du facteur multiplicatif selon l'Équation 5. Les bornes supérieures/inférieures sont ensuite définies à +/- 12,5%.

$$d = \frac{f}{\sqrt[6]{i}}$$

Équation 5 : Relation entre la distance et l'intensité du pic en RMN. d : distance entre les deux atomes, f : facteur multiplicateur fixe, i : intensité du pic

Avec cette méthode de calibration, la distribution de distance pour les NOE HN-HA (Figure 44 Haut) est compatible avec la distribution obtenue avec STARS (Figure 44 Bas). Si la moyenne des valeurs est identique à celle obtenue par STARS (2,835), l'écart type est différent, cette différence pouvant se justifier par la différence dans les tailles d'échantillon. Après avoir vérifié que la valeur du facteur donne des valeurs correctes pour les distances HN-HN, un intervalle à ± 12,5% a été défini de part et d'autre de la valeur du NOE. Cet intervalle permet d'assouplir les contraintes NOE en tenant en compte l'incertitude expérimentale.

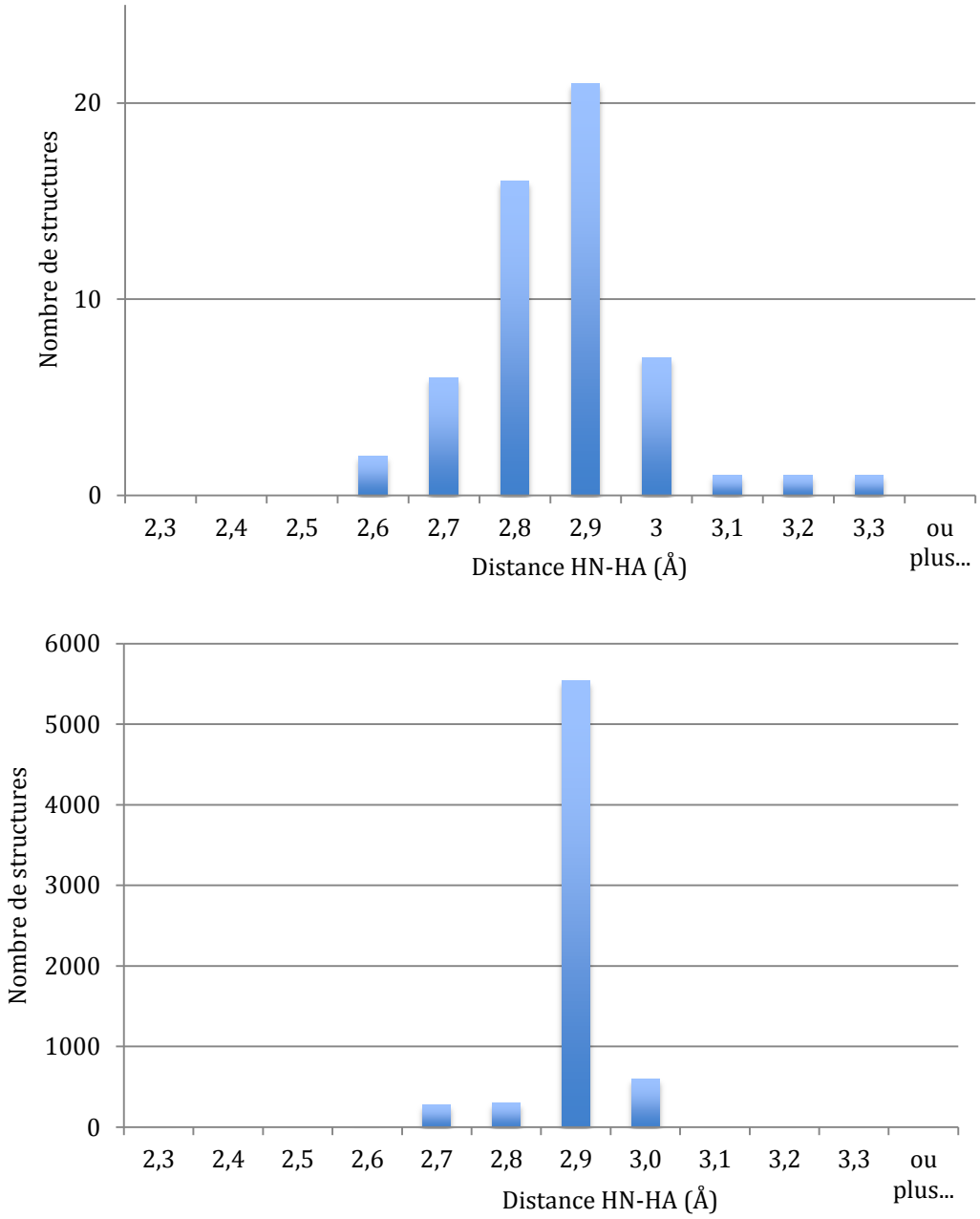


Figure 44 Haut : Distribution des valeurs de distances HN-HA intra-résiduels pour P459N après calibration. Bas : Statistique d'après le programme STARS pour les distances HN-HA intra-résiduels de résidus situés dans des hélices α (sur la base de 100 structures).

4.2 Configuration des calculs

Dans le cadre de cette étude, il est intéressant d'évaluer les résultats de Rosetta dans des conditions différentes : avec et sans contraintes et avec et sans prise en compte des structures homologues (cf. Tableau 6). Ainsi, il sera possible d'évaluer d'une part l'apport des données expérimentales et d'autre part l'importance de disposer de structures

homologues. Dans le cas d'étude présenté, on s'attend à un impact important des contraintes expérimentales longues distances car les contraintes courtes vont servir à définir la géométrie des hélices qui est déjà connue. En revanche, on ne connaît pas a priori l'apport des structures homologues. Chaque calcul a été configuré pour produire chacun 1000 structures.

Tableau 6 : Configuration des calculs Rosetta

Nom du calcul	Prise en charge des NOE	Utilisation des Homologues
Témoin 1	✗	✗
Témoin 2	✗	✓
Contraint 1	✓	✗
Contraint 2	✓	✓

4.3 Regroupement des structures

Chaque calcul aboutissant à la génération de 1000 structures, le programme a généré au total 4000 structures à la suite des quatre calculs. Il est maintenant intéressant de les regrouper indépendamment pour éliminer les structures trop similaires et/ou non informatives (d'énergie trop élevée). Dans ce but, il est possible d'utiliser l'outil de regroupement intégré à Rosetta 3 (Shortle et al. 1998) qui fonctionne en deux temps (Figure 45).

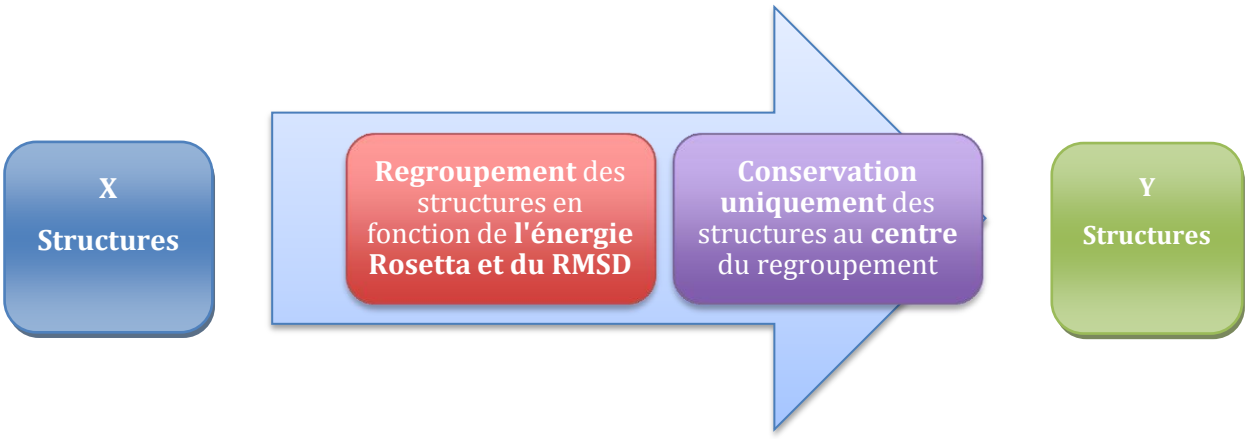


Figure 45 : Fonctionnement du programme de regroupement intégré à Rosetta (X>Y)

Ainsi, pour les regroupements les plus importants, Rosetta ne conservera que les structures les plus proches du centre et donc seule une partie des 1000 structures sera conservée (Figure 46). La taille des regroupements illustre d'une certaine manière l'exploration du paysage énergétique, plus un regroupement est important : plus l'exploration est importante. C'est pour cette raison qu'en général on ne traite que les regroupements de taille significative (ici, de taille supérieure à 20 individus). Le centre des regroupements est la région la plus peuplée, et représente donc une région de convergence. Dans la philosophie Rosetta, cette région de convergence correspond aux meilleures structures du regroupement et plus le regroupement est important, plus cette région a de chances de contenir la meilleure structure globale.

Il faut noter que, lors des regroupements, si le fichier de score des structures donné par Rosetta est présent dans le même répertoire que les structures, ce dernier sera utilisé pour éviter de recalculer l'énergie Rosetta. Or il n'est pas possible de comparer les énergies Rosetta des structures issues des calculs avec et sans contraintes car l'énergie Rosetta contient le terme d'accord avec les données expérimentales si celles-ci sont présentes. Pour regrouper ces structures « à l'aveugle », il est donc nécessaire de les copier dans un même répertoire en effaçant les fichiers de score.

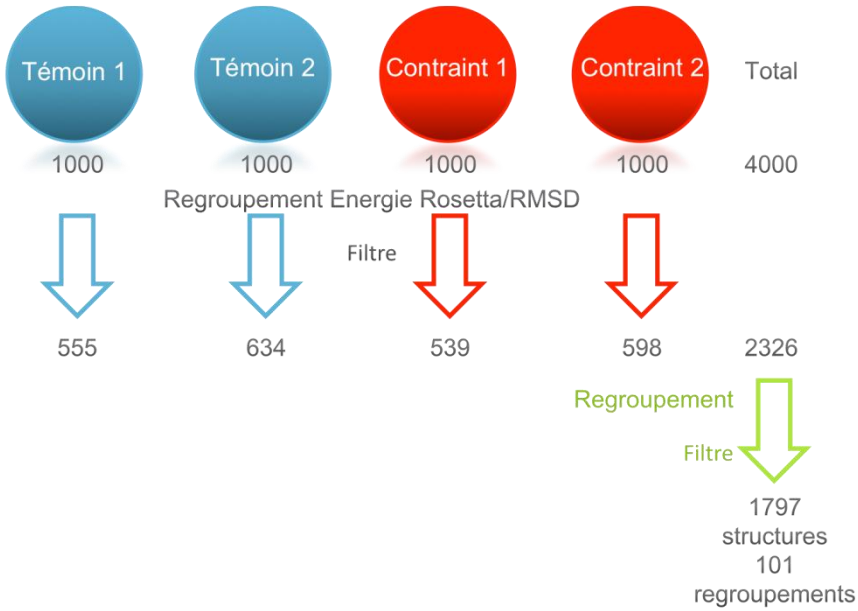


Figure 46 : Après chaque calcul, les structures ont été regroupées entre elles, leur nombre est indiqué en gris sur la deuxième ligne. L'ensemble de ces 2326 structures ont ensuite été regroupées entre elles dans un deuxième temps. Le nombre de structures obtenues au final (1797) est inférieur à la moitié du nombre de structures générées au total par Rosetta (4000).

Au final, sur les 4000 structures générées, 2326 ont été conservées. Ces structures ont ensuite été rassemblées dans un même répertoire avant de relancer le programme de regroupement de Rosetta. Cette étape a pour but de déterminer la distribution des origines des structures et de les regrouper sans prendre en compte l'accord avec les données expérimentales. En étudiant la distribution des structures dans les regroupements ainsi créés, il va être possible de déterminer l'impact des données expérimentales ainsi que l'apport des structures homologues. Par exemple, si les structures sont aléatoirement distribuées dans les regroupements, alors il n'y a pas d'effet des données expérimentales et des homologues. On peut s'attendre à retrouver d'un côté des regroupements composés de structures issues des calculs avec contraintes expérimentales et de l'autre des regroupements composés des structures provenant de calculs sans contraintes.

Le regroupement de ces 2326 structures aboutit à la constitution de 101 regroupements cumulant 1797 structures. La diminution du nombre de structures s'explique toujours par le fait que seules les structures les plus proches des centres des regroupements sont conservées. Les 10 regroupements les plus peuplés regroupent 40% du nombre total de structures. Passé les 20 premiers regroupements environ (Figure 47), les regroupements restants ne sont composés que d'une vingtaine de structures au maximum.

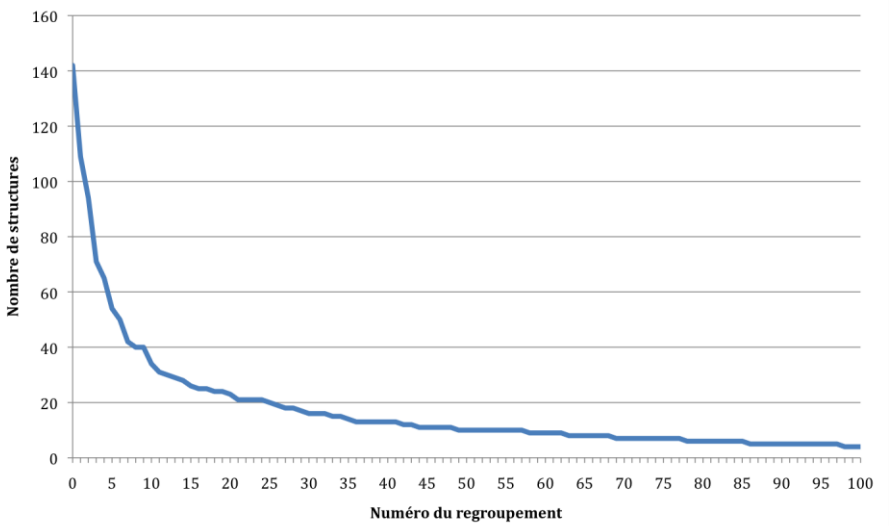


Figure 47 : Représentation du nombre de structures par regroupement. Les 10 plus gros regroupements regroupent 40% des structures, les 20 plus gros 53%. On considère que les regroupements les plus peuplés représentent les points de convergence les plus importants et, de fait, correspondent aux solutions les plus plausibles.

Afin d'évaluer leur accord avec les données expérimentales, la même fonction d'évaluation de respect des contraintes de Rosetta a été utilisée (Équation 4). La Figure 48 représente la corrélation entre le respect des contraintes et le terme d'énergie Rosetta pour les structures des 10 plus gros regroupements. En utilisant ces points et en lançant un grand nombre de fois la fonction de regroupement K-means implémentée dans le logiciel de statistiques R, il est possible de repérer des regroupements qui apparaissent de façon régulière (Figure 49). Le regroupement en bleu en bas à gauche correspond au regroupement des meilleures structures Rosetta.

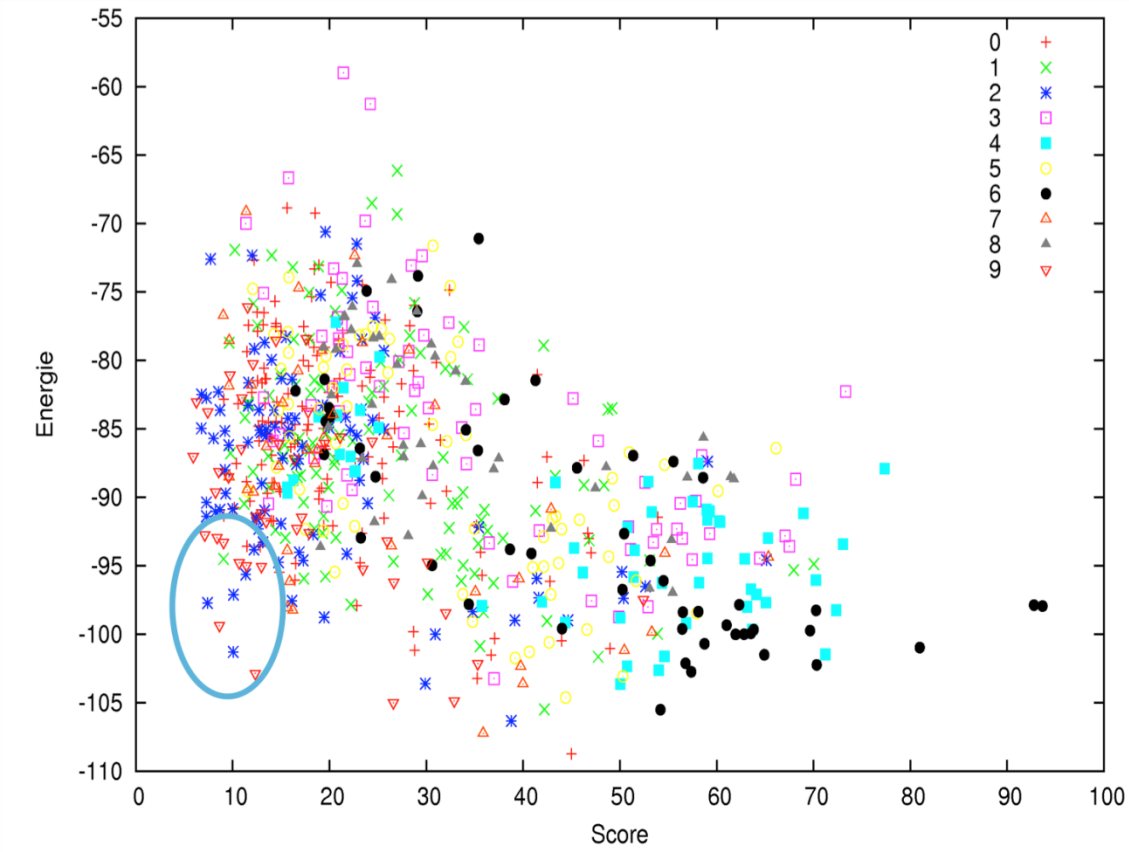


Figure 48 : Répartition des 707 structures des 10 plus gros clusters Rosetta. Plus le score est élevé, moins l'accord avec les données expérimentales est important. Plus l'énergie est élevée, moins la structure est correcte du point de vue de Rosetta. L'ellipse bleue correspond à la localisation des meilleures structures.

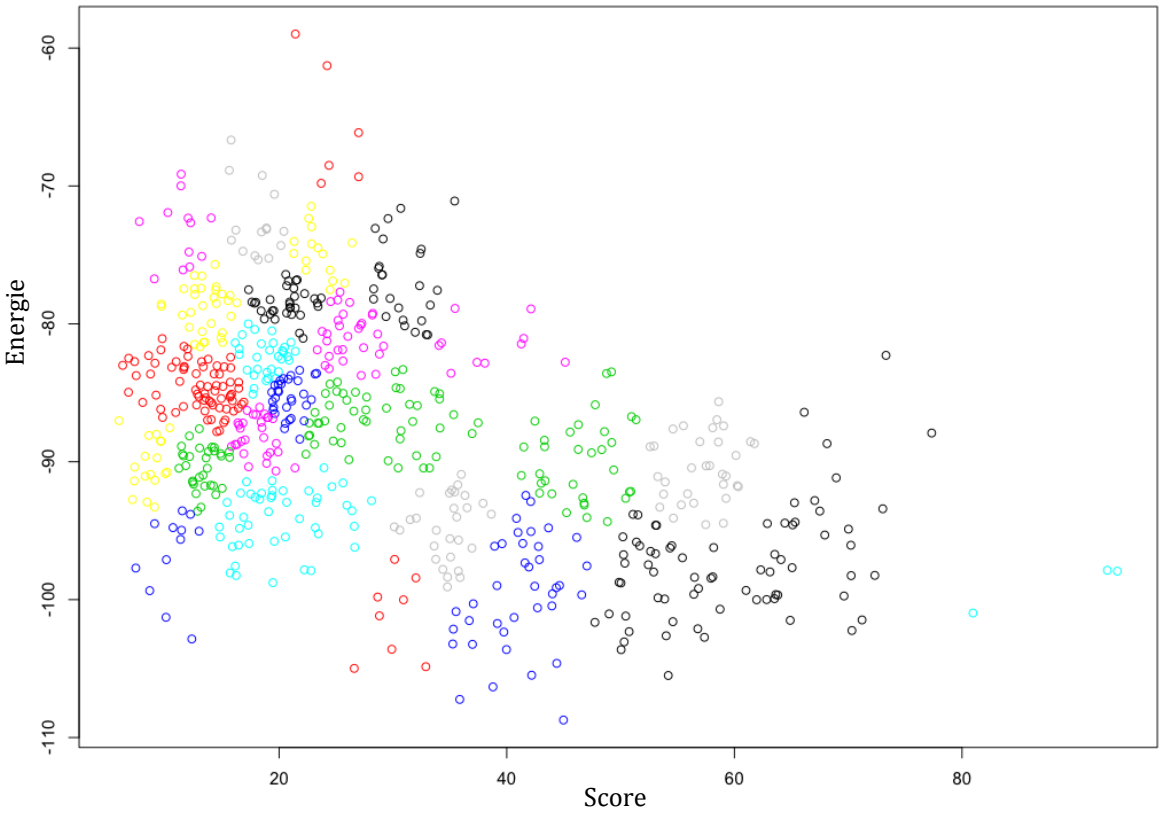


Figure 49 : Résultat du regroupement des 707 points par la méthode K-means de R, seuls les points voisins d'autres points de la même couleur appartiennent au même regroupement (2 points d'une même couleur n'appartiennent pas nécessairement au même regroupement).

Ces 10 structures, présentant à la fois un bon accord avec les données expérimentales et une énergie Rosetta basse, proviennent des regroupements 2 et 9 dont les meilleures structures sont représentées en Figure 50. Le RMSD entre ces structures varie dans l'intervalle 1,1<RMSD<3,03Å pour les Cα et 1,56<RMSD<3,35 Å en tout atome. Les 10 meilleures structures proviennent de calculs avec prise en charge des contraintes expérimentales et 6 de ces 10 structures sont issues des calculs réalisés sans incorporation des homologues structuraux.

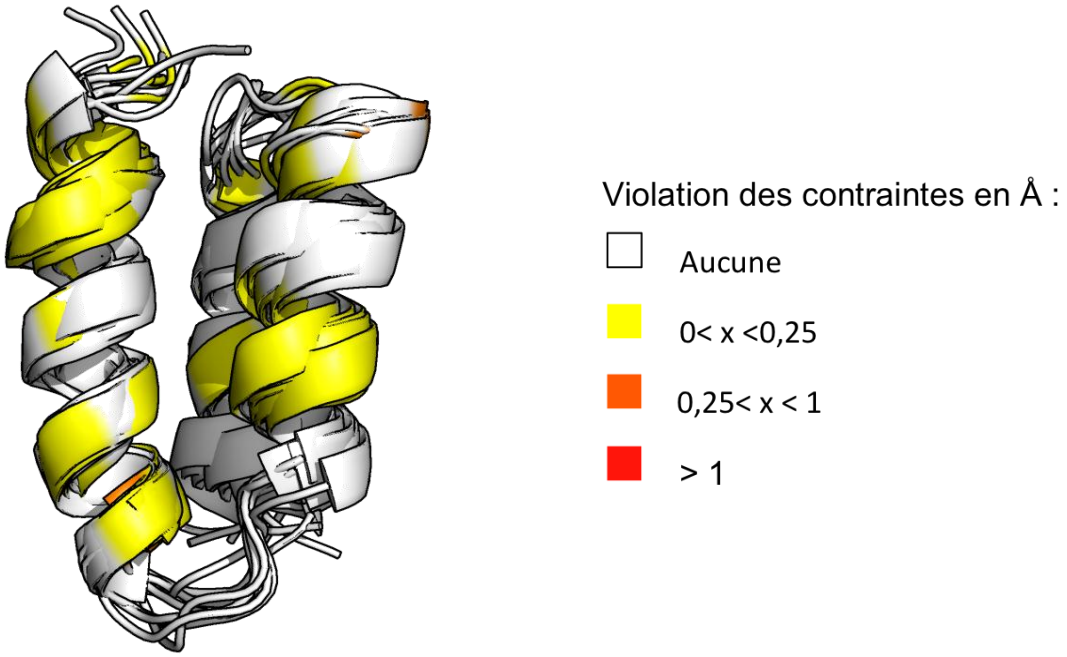


Figure 50 : Superposition des 10 meilleures structures obtenues par Rosetta. Les résidus colorés présentent au moins une violation de contrainte inter ou intra résiduelle.

Le respect des contraintes au niveau des intra et inter résiduels est plutôt bon avec seulement quelques résidus en bout d'hélice qui présentent une violation de contrainte entre 0,25 Å et 1 Å. La principale différence entre ces structures, hormis le positionnement des chaînes latérales, est la position du début et de la fin des hélices α.

Tableau 7 : Résidus de début et de fin des hélices α des meilleures structures Rosetta

Hélice	**Résidu de Début**	**Résidu de Fin**
1	3	14,15
2	19, 22	31
3	35, 37, 38	50, 51

Au niveau des contraintes longues distances, les résultats sont assez bons avec seulement une ou deux contraintes violées à plus d'un Å et quelques-unes à moins d'un Å. En particulier, la meilleure structure globale (Figure 52) présente seulement deux contraintes violées à plus de 0,25 Å dont une à plus d'un Å. Pour cette dernière, l'un des deux atomes impliqués dans cette contrainte est dans un cycle aromatique (contrainte CZ

de Phe41 relié au HN de la Met44) et pour lequel il peut y avoir un problème d'estimation du volume du pic (superposition du bruit, etc.).

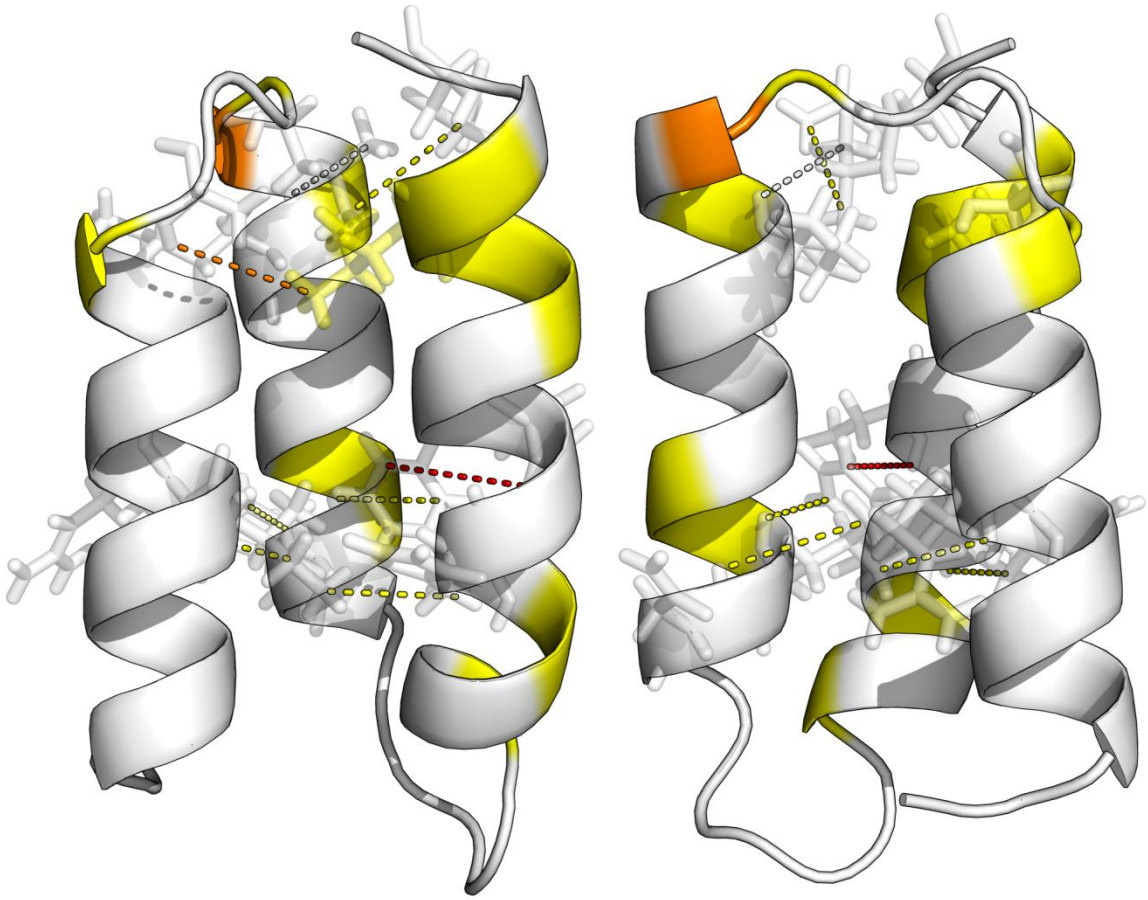


Figure 51 : Représentation de face et de derrière de la meilleure structure Rosetta. Les contraintes inter et intra résiduelles sont bien respectées, à l'exception d'un résidu au niveau de l'extrémité de la deuxième hélice (résidu en orange). De même, seule une liaison longue distance est significativement mauvaise (pointillés en rouge).

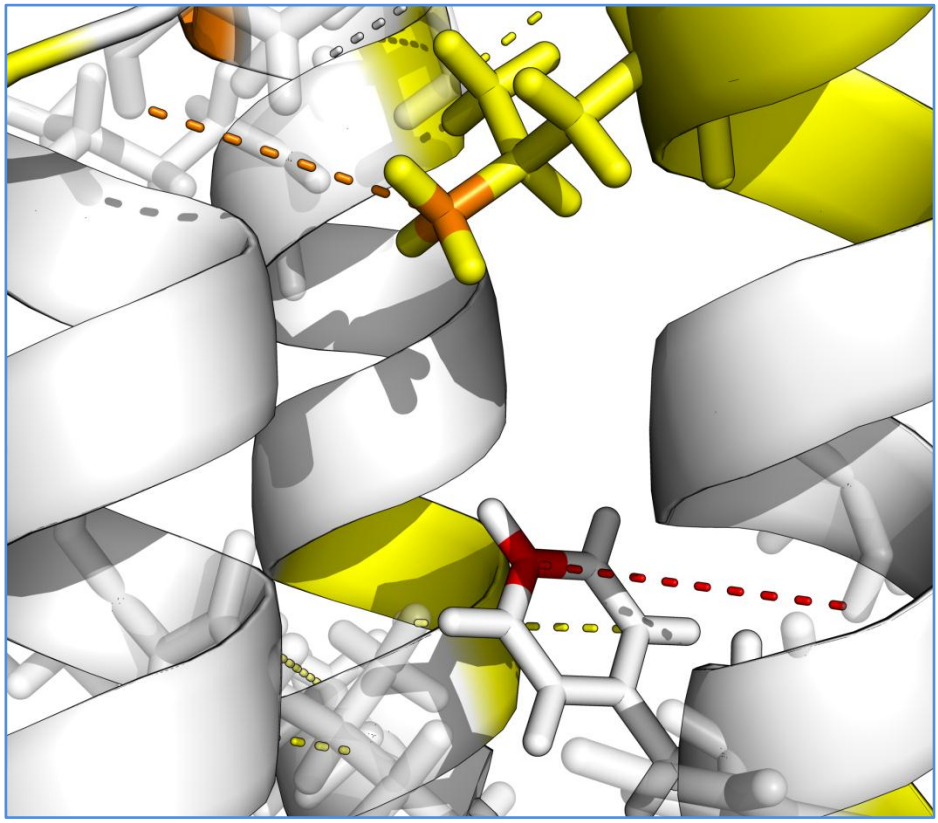


Figure 52 : Zoom sur les contraintes violées à plus de 0,25 Å (rouge et orange). La contrainte en rouge aurait pu être satisfaite par le choix d'un autre carbone du cycle aromatique.

4.4 Comparaison des structures Modeller et Rosetta

La comparaison avec la structure obtenue avec MODELLER (Figure 53) montre que les structures Rosetta présentent une meilleure adéquation avec les données expérimentales. Si ce résultat confirme l'intérêt des contraintes expérimentales, il faut remarquer que ce sont surtout les contraintes longues distances qui ont le plus d'effet. C'est sur les chaînes latérales des résidus impactés par les NOE longues distances que l'on observe le plus de différences.

Si la comparaison d'une méthode utilisant des contraintes expérimentales et une autre qui ne les utilise pas aboutit logiquement à l'avantage de la première, il est important d'indiquer que Rosetta n'a pas nécessairement besoin de structures homologues. L'utilisation de cet outil peut donc se faire dans des cas où l'on ne dispose pas de structures homologues et/ou à forte similarité, cas où MODELLER sera nettement plus limité.

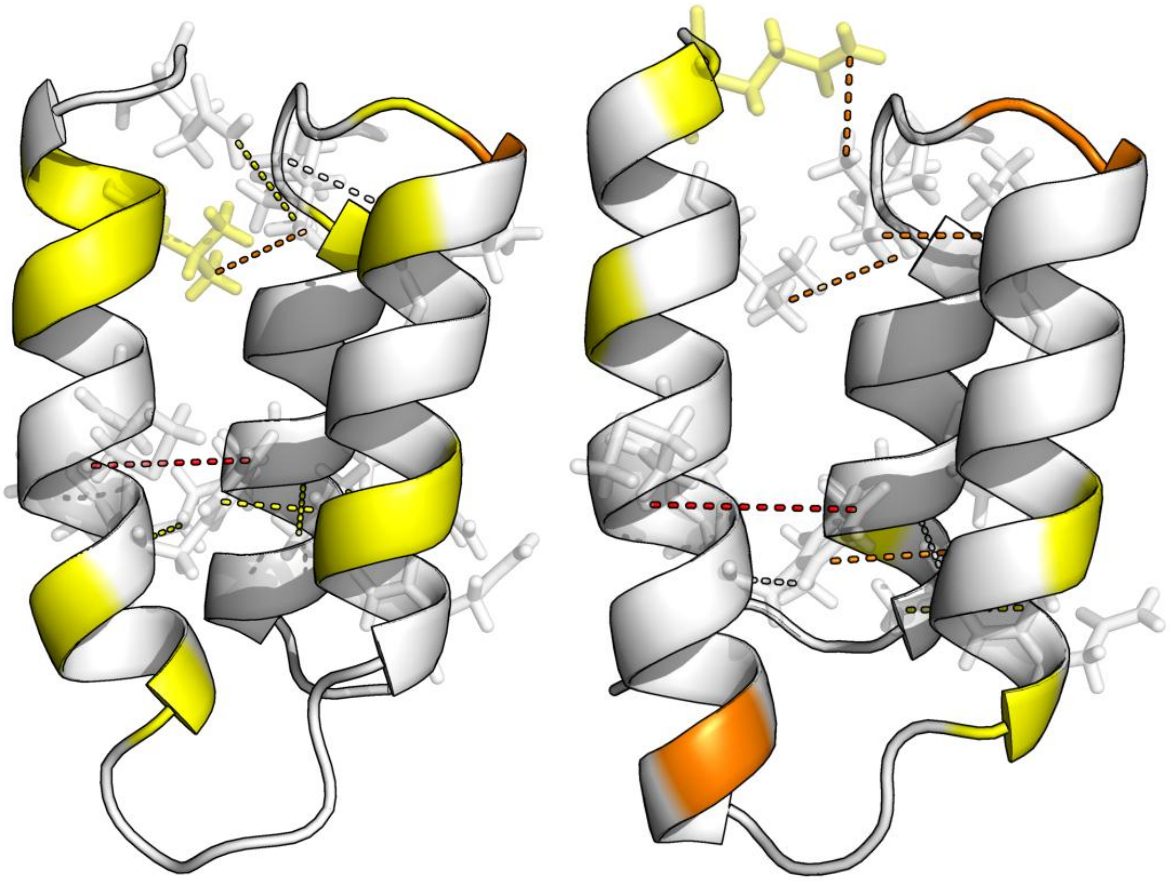


Figure 53 : Structure Rosetta (à gauche) et MODELLER (à droite) : les résidus colorés correspondent aux résidus présentant au moins un NOE inter ou intra résiduel violant la contrainte. Les pointillés correspondent aux NOE longues distances, leur couleur est associée aux mêmes valeurs de violation.

Pour compléter la comparaison, des calculs MODELLER avec les contraintes de distances ont été effectués. Étonnamment, les résultats ne sont pas très bons (Figure 54), les structures (le calcul ayant été fait plusieurs fois) respectant moins les contraintes intra/inter résiduelles que la structure issue du calcul sans contrainte. Les structures MODELLER et Rosetta ont ensuite été évaluées avec le potentiel DOPE (Colubri et al. 2006) intégré à MODELLER (Figure 55). Les structures Rosetta se classent très bien et sont généralement meilleures que les structures MODELLER. Ce potentiel reflète le respect des angles et de l'orientation de la chaîne latérale jusqu'au Cβ.

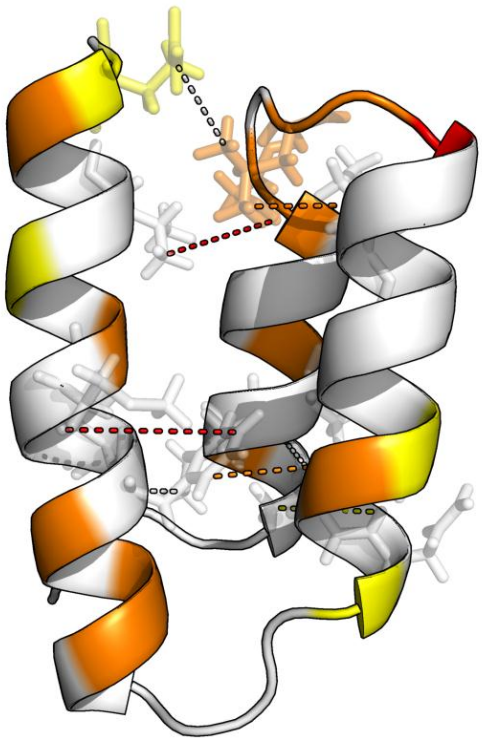


Figure 54 : Structure MODELLER issue du calcul avec prise en charge des contraintes de distance. Les contraintes inter/intra résiduelles et longues distances ne sont clairement pas satisfaites

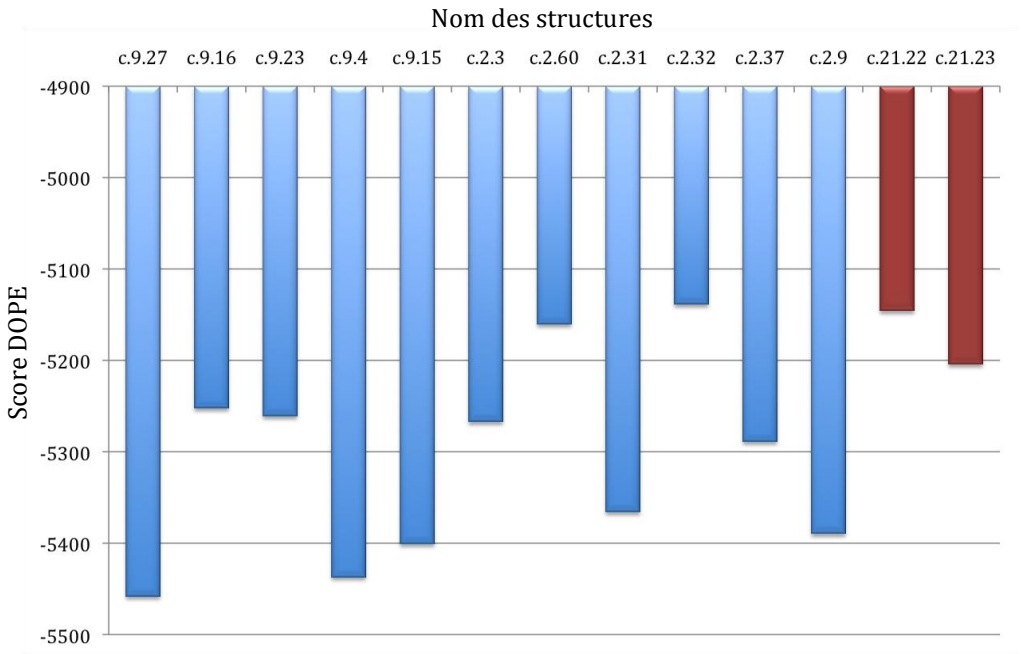


Figure 55 : Score DOPE des meilleures structures Rosetta (bleues) et des deux structures MODELLER (rouges). Les structures c.21.22 et c.21.23 correspondent aux structures MODELLER issues de calculs sans et avec contraintes (respectivement). L'énergie la plus basse est la meilleure.

5 CONCLUSION

Dans le premier article présentant l'utilisation de Rosetta avec des contraintes RMN (Bowers et al. 2000), les auteurs ont signalé les difficultés présentées par les structures formées uniquement d'hélices α. Un des cas tests présente une taille similaire à PPRV-P459N et est aussi uniquement composé d'hélices. Lorsqu'ils ont utilisé 84 contraintes courtes distances (inférieures ou égales à 5 résidus) et 5 contraintes longues distances (strictement supérieures à 5 résidus), le RMSD final entre la structure de meilleur score Rosetta et la structure de référence est de 3,01 Å. En utilisant 51 contraintes courtes distances et 7 contraintes longues distances, le RMSD est de 1,96 Å. Les structures en hélices α sont en général des cas difficiles à traiter avec des NOE seuls (1cmz, Figure 56), sauf si on dispose d'un nombre suffisant de NOE longues distances.

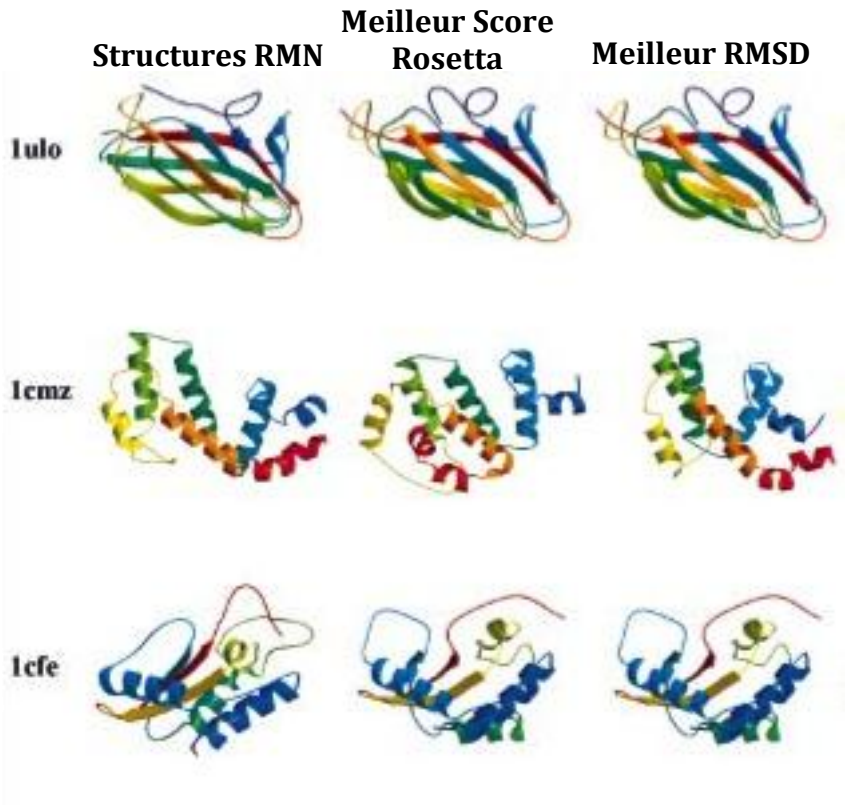


Figure 56 : Résultats obtenus par Rosetta avec prise en charge des NOE (figure provenant de l'article de Bowers et al. 2000). Les structures de la colonne « meilleur RMSD » sont les structures Rosetta présentant le RMSD le plus bas avec la structure RMN. Si les structures de meilleur RMSD et de meilleur score Rosetta diffèrent, cela indique que les données n'apportent pas assez de contraintes pour pouvoir identifier les meilleures structures.

Un deuxième article (Shen et al. 2008), présentant la première version de CS-Rosetta, montre les performances du protocole Rosetta couplé avec l'utilisation de déplacements chimiques incomplets. Avec ce nouveau protocole, les résultats obtenus sont très bons quel que soit le repliement global de la protéine. Ceci n'est pas étonnant dans le sens où il est plus facile d'incorporer des contraintes d'angles (apportées par les déplacements chimiques) que des contraintes de distances (apportées par les corrélations NOE). Une autre étude très récente (Raman et al. 2009) montre tout l'intérêt de l'incorporation de données RMN du squelette uniquement (déplacement chimique, couplages dipolaires résiduels et NOE des protons amides) dans le protocole Rosetta pour la détermination de structures à haute-résolution.

Les résultats obtenus dans ce chapitre montrent l'évolution du protocole Rosetta depuis ces 10 dernières années car, même dans un cas défavorable, les modèles obtenus sont au même niveau voire meilleurs que ceux obtenus par la modélisation par homologie et ce sans utiliser de structures homologues. L'effet des contraintes expérimentales a, au final, relativement peu d'impact sur le squelette et sur le repliement global de la protéine. Le principal apport de ces données se situe au niveau de l'orientation des chaînes latérales.

En conclusion, on ne peut que recommander l'utilisation de Rosetta dans la résolution de structures de petite taille avec peu de données expérimentales. Si les NOE peuvent être suffisants pour des protéines composées essentiellement en feuillets β, l'utilisation des déplacements chimiques est préférable dans les autres cas, y compris celui présenté dans ce chapitre, à moins de disposer d'un grand nombre de contraintes longues distances. MODELLER permet quand même d'avoir une bonne idée du repliement global de la protéine mais l'ajout de données expérimentales n'a pas permis d'améliorer la position des chaînes latérales.

ÉTUDE DE L'EFFET DES COURANTS DE CYCLE DANS LES COMPLEXES PROTÉINE-ADN

1 INTRODUCTION

La RMN permet de localiser les surfaces d'interaction entre macromolécules (ou entre une macromolécule et un petit ligand). L'une des méthodes les plus utilisées est la cartographie des déplacements chimiques (*chemical shift mapping*) dont un exemple est représenté en Figure 57. Une série de spectres HSQC d'une protéine (généralement marquée à l'azote ^{15}N) est enregistrée en présence de quantités croissantes de son partenaire (non marqué). Le déplacement chimique d'un atome étant très fortement sensible à son environnement (présence du solvant, implication dans des liaisons hydrogène, présence de courant de cycles, géométrie de la molécule, etc.), les fréquences de résonance des atomes impliqués dans l'interaction vont généralement être perturbées. Le repérage des atomes perturbés sur les spectres permet donc de définir la surface. Cependant, la relation entre atomes dont les fréquences sont perturbées et atomes localisée dans la surface d'interaction n'est pas absolue, des atomes au voisinage de la zone d'interaction peuvent être affectés indirectement tandis que des atomes de la surface peuvent ne pas être affectés.

Cette information peut ensuite être utilisée dans le cadre de l'utilisation d'outils de d'assemblage moléculaires comme Haddock (Dominguez et al. 2003) en réduisant fortement l'espace à explorer par le programme, ou encore dans le criblage de nouvelles molécules dans la recherche de nouvelles molécules actives pour l'industrie pharmaceutique. Il est difficile d'utiliser l'information contenue dans les expériences cartographie de déplacement chimique, car elle est très dégénérée. Si l'on connaît, une voire deux surfaces en interactions, on ne connaît pas le positionnement de ces deux surfaces l'une par rapport à l'autre. Par ailleurs, cette information est aussi fortement dégradée. Elle est très généralement utilisée de façon binaire (un atome est affecté ou pas) sans prendre en considération la valeur et le signe de la perturbation.

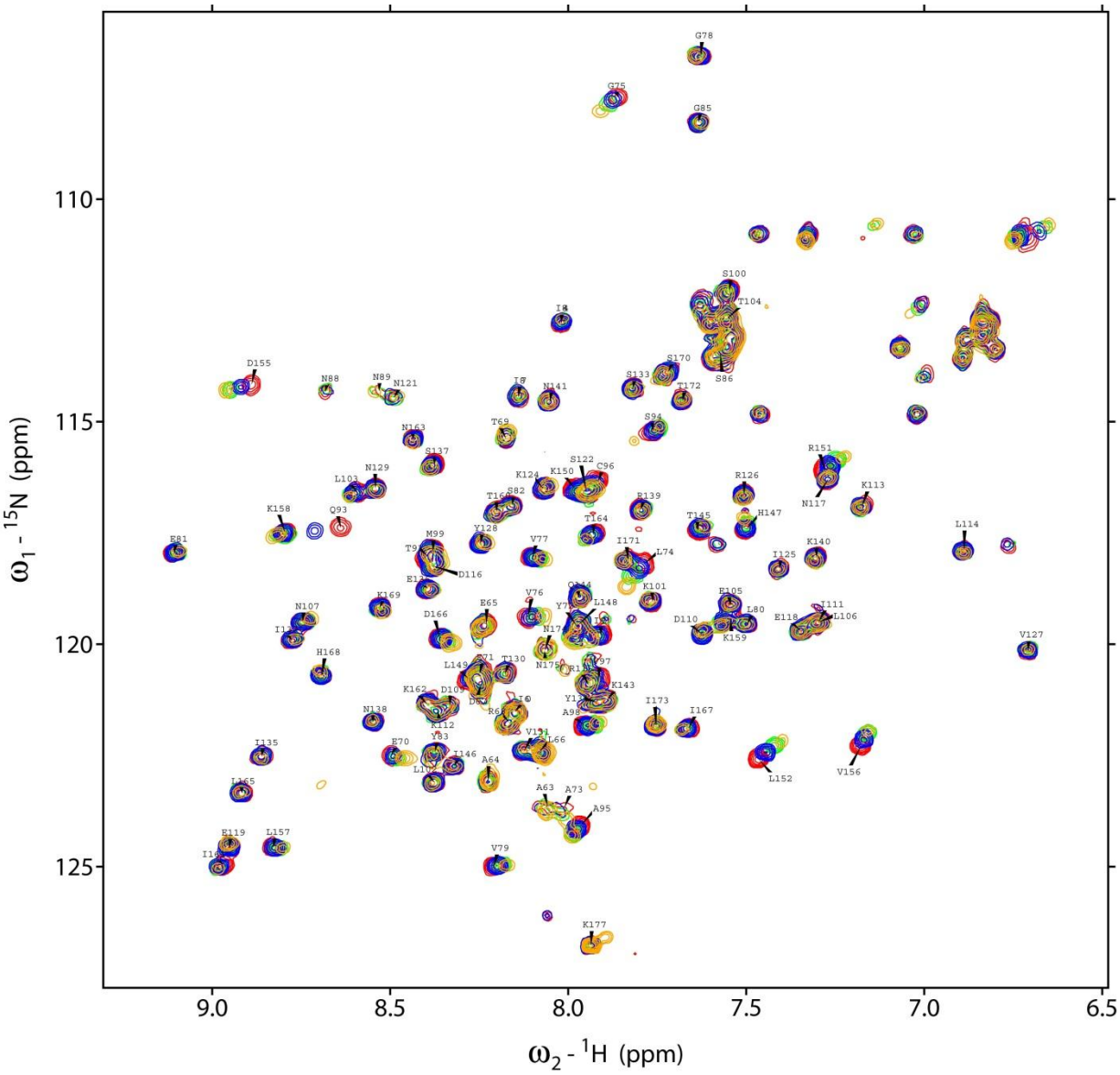


Figure 57 : Cartographie de déplacement chimique d'une protéine virale titrée avec de l'ARN, les différentes couleurs des pics correspondent à des concentrations différentes d'ARN. Les pics dont la position varie correspondent à des résidus affectés par le changement de concentration de l'ARN et donc potentiellement en interaction avec celui-ci.

1.1 Le problème de l'arrimage Protéine-ADN

Dans le cas de complexes protéine-ADN, nous avons cherché une information plus fine qui permettrait de mieux guider les programmes d'assemblage moléculaires. Nous nous sommes intéressés aux interactions protéine/acide nucléique en partant de l'idée que le courant de cycles était un élément intéressant à prendre en compte et que l'on pouvait attendre à ce que les perturbations engendrées par les interactions protéines/acides

nucléiques dépendent assez fortement de ce paramètre. Les interactions entre les protéines et les acides nucléiques sont aussi d'une très grande importance dans la machinerie cellulaire. La plupart des fonctions portées par les acides nucléiques sont réalisées à l'aide de protéines. En particulier, les interactions protéine-ADN sont à la base de fonction indispensables à l'organisme, telles que la réplication du génome, la régulation de la transcription ou encore le maintien de l'intégrité du génome. En effet, ce sont les protéines qui assurent la réparation des molécules d'ADN endommagées. Des maladies majeures telles que le cancer sont causées par l'altération de ces fonctions. Une meilleure connaissance de ces interactions au niveau atomique est indispensable pour la recherche de nouvelles molécules thérapeutiques ciblées.

L'arrimage protéine/acide nucléique est un domaine récent et il n'y a que peu d'exemples d'applications pratiques à grande échelle. Les premières applications d'arrimage moléculaire datent d'une quinzaine d'années avec la présentation du programme MONTY développé par Knegtel et al. Ils ont implémenté dans leur programme la flexibilité des chaines latérales de la protéine (Knegtel et al. 1994) puis ont ajouté celle de l'ADN (Knegtel et al. 1994b) mais le programme n'a été appliqué que sur un nombre restreint de complexes protéine-ADN. Quelques années plus tard, le programme FTDock (Aloy et al. 1998) a proposé un algorithme innovant de recherche en corps rigides pour prédire les interactions protéine-ADN pour un nombre plus important de complexes. La méthode a ensuite été modifiée pour prendre en compte la flexibilité des chaînes latérales de la protéine à l'interface (Sternberg et al. 1998). Plus récemment, Chen a obtenu de bons résultats pour les complexes protéines ARN (Chen et al. 2004), en particulier pour la discrimination des structures produites avec le programme RosettaDock (Gray et al. 2003). Enfin, un des exemples les plus avancés est l'utilisation de la méthode HADDOCK qui a permis de réaliser l'arrimage totalement flexible (chaines latérales et ADN) de trois complexes protéine-ADN (van Dijk et al. 2006) en partant de leur forme non liée. Depuis ce travail, l'équipe responsable de HADDOCK conserve un grand intérêt pour l'arrimage protéine-ADN avec la conception de 3D-DART, un serveur pour la modélisation de l'ADN (van Dijk & Bonvin 2009) et d'une revue présentant ce qu'il est possible de faire avec la dernière version du programme (van Dijk & Bonvin 2010).

Malgré ces travaux, il y a toujours de grandes limitations pratiques quant à la généralisation de ces méthodes. En particulier, seules quelques-unes d'entre elles

permettent de partir des formes libres de la protéine et de l'acide nucléique. La nature flexible de l'ADN et de l'ARN et l'importance de la reconnaissance indirecte font qu'il va être nécessaire d'ajouter une flexibilité moléculaire plus importante et de manière plus robuste que ce qui a été fait jusqu'à présent. C'est à ce niveau que l'incorporation de données expérimentales peut aider en permettant l'identification des régions flexibles à la surface des deux partenaires.

1.2 Le déplacement chimique

Comme nous l'avons dit précédemment, l'information apportée par les expériences de cartographie de déplacement chimique est généralement exploitée sous une forme très dégradée. On ne la considère que sous forme binaire (significativement affecté – non significativement affecté). Par ailleurs, les déplacements chimiques des atomes d'une protéine contiennent une très grande quantité d'information structurale (Shen et al. 2009) et, depuis les travaux de Wishart (Wishart et al. 1992), on sait qu'il existe une corrélation claire entre le déplacement chimique secondaire et la structure secondaire. Le déplacement chimique secondaire est la différence entre le déplacement chimique d'un atome d'un acide aminé donné dans la protéine considérée et le déplacement chimique du même atome du même acide aminé dans une structure aléatoire. Par exemple, on observe pour les protons un changement de -0,3 ppm dans les hélices α et de 0,5 ppm pour les feuillets β (Shen & Bax 2007; Wang & Jardetzky 2002). De façon plus fine, il est possible d'utiliser l'information contenue dans le déplacement chimique des atomes du « squelette » d'une protéine (Cα, Hα, C, N, HN et Cβ) pour déterminer assez précisément les angles phi-psi. Cette donnée étant maintenant utilisée de façon courante comme contrainte structurale dans les programmes de reconstruction comme TALOS (Cornilescu et al. 1999). Plus récemment, il a été montré que l'utilisation combinée de déplacements chimiques et de techniques de prédiction de structures permettent d'obtenir des structures de bonne résolution (Bowers et al. 2000; Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008; Berjanskii et al. 2009; Robustelli et al. 2010; Shen et al. 2009; Kohlhoff et al. 2009; Vila et al. 2007; Hunter et al. 2005). Ces nouvelles approches ont pu voir le jour grâce à l'apparition de méthodes de prédiction de déplacements chimiques. Ces programmes, dont les plus utilisés sont ShiftX (Neal et al. 2003) et SPARTA (Shen & Bax 2007), sont basés sur des méthodes utilisant une grande collection de déplacements

chimiques expérimentaux associés aux structures 3D correspondantes. En plus de la résolution de la structure de protéines seules, ces approches ont aussi été appliquées aux complexes protéine-protéine avec succès (Montalvao et al. 2008).

Toutes ces approches ne concernent que les protéines. À ce jour, il ne semble pas exister de programme permettant de calculer les déplacements chimiques de dans le cas des acides nucléiques, même si quelques méthodes ont vu le jour pour les formes canoniques de l'ADN (Altona et al. 2000). De plus, il n'existe pas de programmes permettant de calculer l'effet d'un acide nucléique sur une protéine dans le cas d'un complexe. Or, dans ce dernier cas, on est peut-être dans une situation favorable dans la mesure où l'on peut penser que la contribution des courants de cycle de l'acide nucléique à la perturbation du déplacement chimique des atomes de la protéine va être importante. Cette contribution étant simple à calculer, on disposerait alors d'une façon simple d'améliorer la qualité des structures de complexes protéine/acides-nucléiques calculées à partir des expériences de cartographie de déplacement chimique.

De ce point de vue, une analyse de la contribution des différents facteurs susceptible d'affecter les déplacements chimiques des atomes du squelette est intéressante (Tableau 8 d'après (Wishart & Case 2002). Elle montre qu'il n'y a pas ou peu de contribution des courants de cycle au déplacement chimique des atomes lourds (azote et carbone) et, à géométrie constante, la contribution majeur au déplacement chimique des H^N est de façon prévisible la présence de liaison hydrogène et les facteurs divers (dont l'exposition au solvant). On constate aussi, toujours à géométrie constante, que la contribution des courants de cycle est majoritaire sur les protons H^a. De plus, il faut remarquer que ces données ont été obtenues à partir de l'analyse de mesures réalisées sur des protéines seules (pour lesquelles les cycles aromatiques ne sont pas forcément très nombreux) et on peut donc s'attendre à ce que la situation soit encore plus favorable dans le cas de complexes protéines/acides nucléiques. Au vu de la faiblesse de l'effet constaté sur les atomes lourds, nous avons décidé de restreindre notre analyse aux protons dans la suite de ce travail.

Tableau 8 : Analyse des différents facteurs influençant le déplacement chimique des atomes du squelette des protéines (d'après Wishart et Case). Neuf contributions ont été distinguées par les auteurs. Random coil désigne la valeur du déplacement chimique de l'atome dans un peptide de structure aléatoire, elle mesure, de fait l'influence de la nature de l'acide aminé. Torsion phi/psi, phi/psi-1, chi, chi-1, désignent respectivement les variations dues aux changement des angles phi et psi autour de l'acide aminé considéré, du plan peptidique précédent l'acide aminé considéré, les changements des angles chi du résidu considéré et du résidu précédent. Ces quatre contributions mesurent donc l'influence de la géométrie locale sur le déplacement chimique d'un atome. Les trois suivantes correspondent à l'implication de l'atome dans des liaisons hydrogènes, à l'influence des courants de cycle dus aux groupements voisins et à l'influence des charges électriques des atomes voisins. Une analyse plus précise et plus détaillée peut être trouvée dans la publication de Neal et al. (Neal et al. 2003).

Contribution	HN (en %)	N (en %)	Hα (en %)	Cα (en %)	Cβ (en %)
Random Coil	0	50	25	50	75
Torsions ϕ/ψ	0	0	50	25	10
Torsions ϕ/ψ_{i-1}	25	25	0	0	0
Chaîne latérale χ	5	0	0	0	5
Chaîne latérale χ_{i-1}	5	5	0	0	0
Liaisons Hydrogène	25	5	5	5	0
Courant de cycle	10	0	10	10	5
Charges locales	10	0	0	0	0
Divers	20	15	10	10	5

1.3 Le courant de cycle

Le courant de cycle est une contribution qui provient de l'anisotropie de la susceptibilité magnétique des groupements aromatiques. Lorsqu'un cycle aromatique est placé dans un champ magnétique, les électrons vont circuler dans celui-ci et ainsi générer

un champ magnétique local comme le ferait un courant électrique circulant dans une boucle de fil conducteur (Figure 58). Le champ magnétique généré affecte un volume spatial assez grand pour influencer les protons au voisinage du cycle. Les zones propices au blindage (variation négative du déplacement chimique) ou de déblindage (variation positive) des protons affectés sont indiquées sur la Figure 58. La contribution du courant de cycle est intéressante car elle est relativement facile à calculer. De plus, la molécule d'ADN contient un grand nombre de cycles aromatiques qui pourraient affecter les protons de la région d'interaction de la protéine.

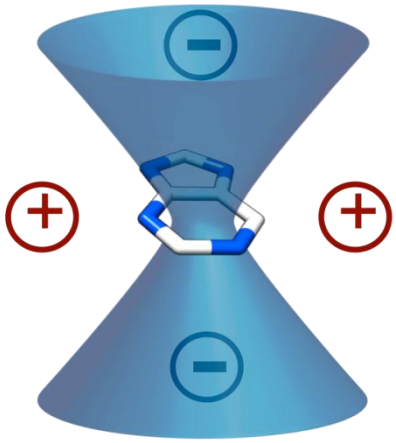


Figure 58 : Illustration de l'effet d'un courant de cycle. Le volume de signe négatif représente l'espace de blindage engendré par le courant de cycle. Les atomes compris dans l'espace en dehors de ce volume seront déblindés.

L'objectif de ce chapitre est de savoir si la contribution des courants de cycle aux variations de déplacement chimique constatées lors d'expériences de titrage protéine/acide nucléique est majoritaire. Ainsi, leur calcul en retour pourrait servir à améliorer de façon significative la qualité des structures de complexes obtenus à partir de l'analyse de ces expériences. Les données expérimentales directes étant peu nombreuses, comme nous le verrons par la suite, nous avons utilisé une approche indirecte. Nous avons simulé les effets des courants de cycle dus aux acides nucléiques sur des protéines en utilisant des structures connues de complexes et essayé, à partir de là, d'estimer leur importance. Nous avons aussi essayé de préciser les conditions dans lesquelles une expérience de RMN permettrait d'obtenir l'information la plus pertinente de ce point de vue.

1.4 Méthodes de calcul de l'effet des courants de cycle

La contribution de l'effet des courants de cycle dans le déplacement chimique peut être représentée sous la forme générale $\delta_{ring}(r) = iBG(r)$ où le facteur i est un facteur d'intensité spécifique au cycle, B est une constante liée à la susceptibilité magnétique et G(r) une fonction géométrique. Celle-ci peut être calculée par différentes méthodes dont trois sont particulièrement utilisées : Haigh et Mallion (Haigh & Mallion 1979), Johnson et Bovey (Johnson & Bovey 1958) et point-dipole (Pople 1958) classées par l'ordre de précision défini selon le travail de Perkins (Perkins & Dwek 1980).

1.4.1 Haigh-Mallion

La méthode Haigh-Mallion est une formulation dérivée de la mécanique quantique. Elle est entre autres utilisée dans le programme ShiftX (Neal et al. 2003) car les auteurs estiment que c'est la méthode la plus précise.

$$G(r) = \sum_{ij} S_{ij} \left(\frac{1}{r_i^3} + \frac{1}{r_j^3} \right)$$

Équation 6 : Formule de la méthode Haigh-Mallion

Dans cette équation, *r* est le vecteur définissant le point où est calculé le courant de cycle, r*i* et r*j* les distances de ce point à deux atomes (i et j) adjacents du cycle et S_{ij} l'aire signée du triangle formée par les deux atomes du cycle et la projection de *r* sur le plan du cycle. La somme étant réalisée sur toutes les paires d'atomes adjacents du cycle. Nous avons décidé d'utiliser cette méthode car en plus d'être précise, elle est relativement simple à coder.

Le programme de calcul des courants de cycle a été codé en Python. Il est nécessaire fixer la valeur de deux facteurs : le facteur B et le facteur *i*. La valeur du facteur B utilisée est la même que pour le programme shiftX et les valeurs du facteur *i* spécifiques aux cycles de l'ADN ont été définies par la publication de Case (Case 1995). Le programme de

calcul commence par définir tous les cycles des bases de l'acide nucléique puis détermine pour chacun les atomes à moins de 8 Å. Ce seuil est déterminé par la courbe de la fonction de calcul, après une distance de 8 Å l'effet des courants de cycle est très faible (Figure 59). Il est important de noter que, dans la plupart des cas, les protons affectés de la protéine seront proches du plan du cycle du fait de la conformation en double hélice de l'ADN. Le changement de signe que l'on observe au début de la courbe de rho=5,0 correspond au passage de l'extérieur à l'intérieur du cône d'effet du courant de cycle. Une fois que l'on dispose de la liste contenant chaque cycle de l'ADN, le programme détermine pour chacun une nouvelle liste contenant les protons de la protéine affectés par celui-ci. Le calcul de l'effet du courant de cycle lui-même est ensuite effectué sur chaque atome de la liste. Si un atome est affecté par plus d'un courant de cycle, alors les valeurs sont additionnées.

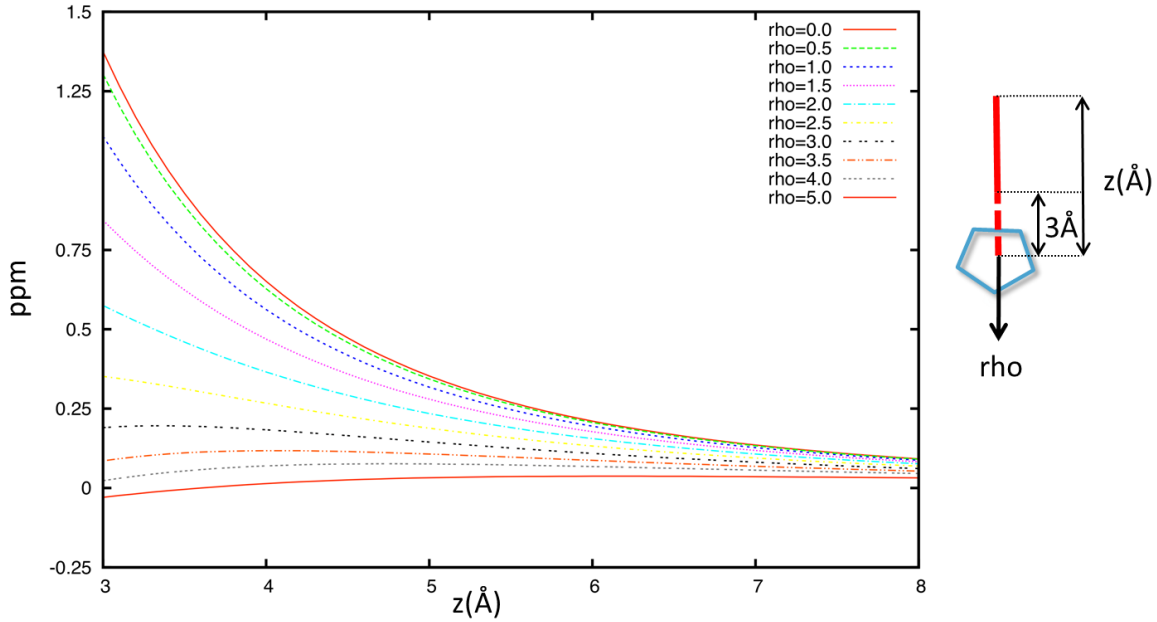


Figure 59 : Courbes calculées de l'effet des courants de cycle en fonction de la position d'un proton par rapport au cycle d'adénine à 5 carbone en utilisant la formule de Haigh et Mallion. z désigne l'élévation au-dessus du plan du cycle. Le calcul est réalisé à partir de z=3 Å pour ne pas tomber sur une incohérence due au fait que « l'épaisseur » du cycle peut être évaluée à 3 Å. rho désigne la distance au centre du cycle de la projection du point considéré sur le plan du cycle (coordonnées cylindriques).

1.4.2 JOHNSON BOVEY

Cette méthode est la description semi-classique de l'effet des courants de cycle déterminée par Johnson et Bovey.

$$G(\rho, z) = \frac{1}{\sqrt{(1+\rho)^2 + z^2}} \left[K(k) + \frac{1 - \rho^2 + z^2}{(1-\rho)^2 + z^2} E(k) \right]$$

Équation 7 : Formule de la méthode Johnson-Bovey

Les variables p et z sont les coordonnées cylindriques de l'atome affecté relativement au centre des orbitales électroniques jointes et E et K sont des intégrales elliptiques. Comme il y a deux orbitales, une au-dessus et une en dessous du cycle, il est nécessaire de résoudre cette équation deux fois par cycle.

1.4.3 POINT-DIPOLE

La méthode dite du point-dipole a l'avantage d'être plus rapide à calculer que les autres mais est moins précise. Il faut cependant noter qu'avec certains paramètres, elle n'est plus significativement inférieure (Moyna et al. 1998). C'est cette méthode qui a été implémentée dans CamShift (Kohlhoff et al. 2009).

$$G(r) = (1 - 3cos^2\theta)\frac{1}{r^3}$$

Équation 8 : Formule de la méthode Point-Dipole

2 Effet des courants de cycles dans les complexes Protéine-ADN

2.1 Étude de complexes avec des données expérimentales

La première idée qui vient à l'esprit, pour déterminer l'incidence des courants de cycle dans les perturbations des déplacements chimiques constatées lors de la formation de complexe protéine/acides nucléiques est bien évidemment de rechercher des structures de complexes pour lesquelles on disposerait des fréquences de la protéine seule dans le complexe. Dans ce but, nous avons utilisé la BioMagResBank (BMRB) qui est une base de données en libre accès proposant aux utilisateurs de télécharger et d'envoyer des données expérimentales obtenues par RMN. Ces données ne sont pas traitées, c'est-à-dire qu'elles se retrouvent dans l'état dans lequel l'auteur les a envoyées. Elles peuvent donc contenir des erreurs, que ce soit par rapport au référencement de l'atome ou de la valeur du déplacement chimique. Le groupe de Wishart est en train d'automatiser la correction de ces erreurs avec le programme ShiftX pour constituer une nouvelle base de données (Zhang et al. 2003) mais, au moment de la rédaction, le nombre de données corrigées n'était pas suffisant pour remplacer la base de données de la BMRB.

Après recherche approfondie, il a été possible d'identifier 37 complexes protéine-ADN pour lesquelles on dispose des déplacements chimiques de la protéine. Sur ces 37 complexes, seules les protéines HMG et Tn916 possèdent des déplacements chimiques pour les formes libres et liées de la protéine.

Tableau 9 : Données pour la protéine HMG

Type	Code PDB	Code BMRB	Nombre de résidus	Nombre de déplacements chimiques
Protéine libre	1HMA	4732	74	391
Protéine en complexe	1E7J	4734	74	235

Tableau 10 : Données pour la protéine Tn916

Type	Code PDB	Code BMRB	Nombre de résidus	Nombre de déplacements chimiques
Protéine libre	1BB8	4160	73	439
Protéine en complexe	1B69	4165	71	414

2.1.1 Protéine HMG

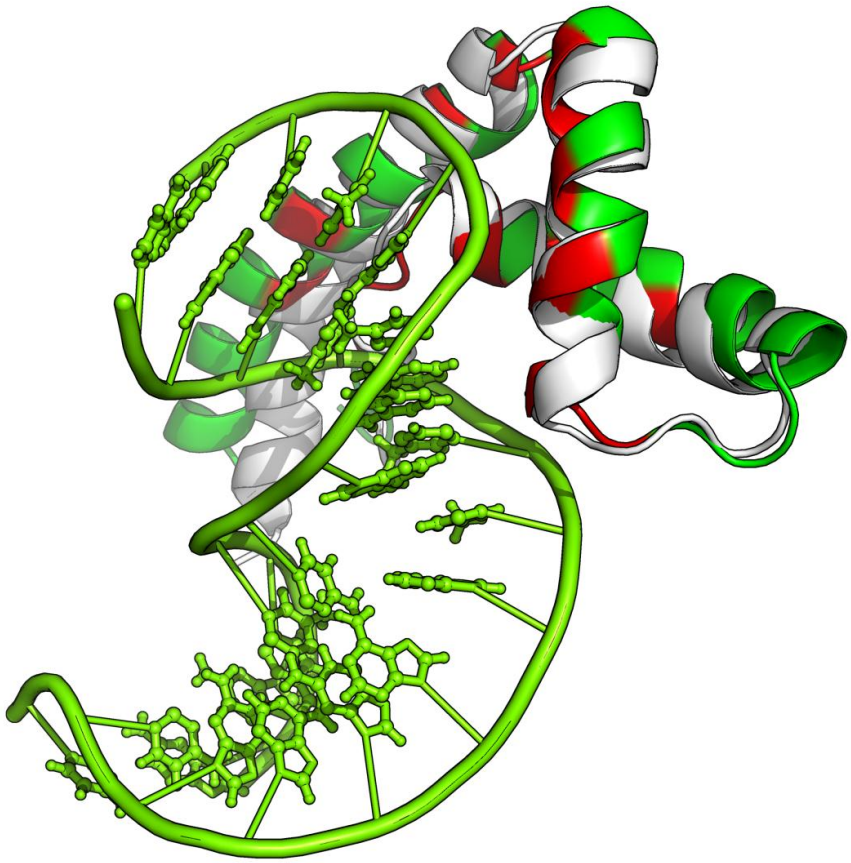


Figure 60 : Structure de la protéine HMG dans la forme libre (blanc) et la forme complexée avec l'ADN (vert). Les résidus en rouge correspondent aux résidus à l'interface. Le RMSD entre les deux structures est de 1,684 Å (990 atomes).

Ce complexe HMG/ADN déformé (l'ADN contient un bulge d'adénine) est en échange rapide. C'est un cas défavorable pour l'étude des déplacements chimiques car les résonances des résidus à l'interface sont relativement peu modifiées. Pour cette protéine, il est possible de déterminer la différence en déplacement chimique pour 199 atomes. Le calcul des courants de cycles permet d'identifier 12 atomes affectés, dont 3 pour lesquels on dispose des déplacements chimiques pour les formes libres et liées (Figure 61).

La variation observée du déplacement chimique suite à la liaison avec l'ADN est relativement faible. La majorité des pics de grandes valeurs sont associées à des protons HN ou Hα et cela pourrait s'expliquer en grande partie par la susceptibilité du déplacement chimique de ces atomes à la modification de la géométrie du squelette.

Très peu de courants de cycle affectent les protons de la protéine car la distance entre les deux partenaires est trop importante. On observe pour le résidu 33 une bonne corrélation entre la valeur de la variation du déplacement chimique et celle de l'effet des courants de cycle (Figure 62). Pour les deux autres protons affectés, il y a une corrélation entre le signe du courant de cycle et celui de la variation du déplacement chimique. Pour le résidu 10, les deux valeurs sont très proches alors que pour le résidu 33 la valeur du courant de cycle est très élevée par rapport à la variation du déplacement chimique. Dans ce cas, il est possible que d'autres contributions du déplacement chimique (comme les liaisons hydrogène) atténuent la valeur de la différence.

En conclusion pour la protéine HMG, il semble impossible d'utiliser l'information apportée par les courants de cycle car trop parcellaire et de valeur insuffisante (valeurs assez proches du bruit). La variation du déplacement chimique des chaînes latérales (Figure 61 Bas) est aussi relativement difficile à utiliser, mais on peut observer que les pics les plus importants sont pratiquement tous dans la région d'interaction. Il faut cependant noter que cette protéine est un cas plutôt défavorable car la protéine est relativement éloignée de la molécule d'ADN. Dans ce cas la contribution des courants de cycle est très faible.

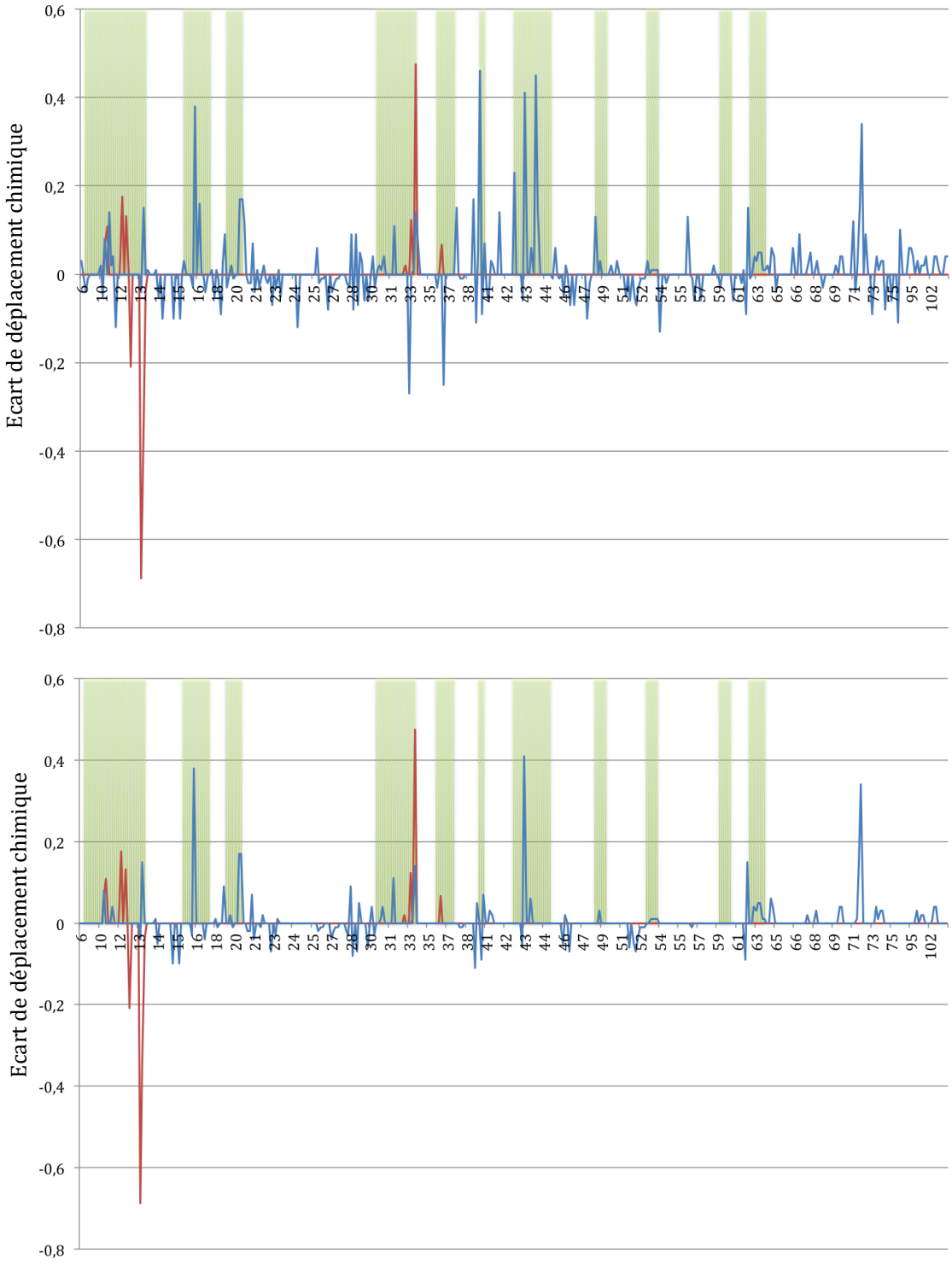


Figure 61 : Haut : Différence de déplacement chimique entre la forme complexée et libre de HMG (bleu) et valeur du courant de cycle calculé pour la forme complexée (rouge) pour chaque atome de chaque résidu (numéro du résidu sur l'axe des abscisses).

Bas : Idem que précédent sans les atomes HN et Hα. Comme on ne dispose pas du même nombre de déplacement chimique pour chaque résidu, l'axe des abscisses n'est pas linéaire. Les zones en vert correspondent aux régions d'interface avec l'ADN.

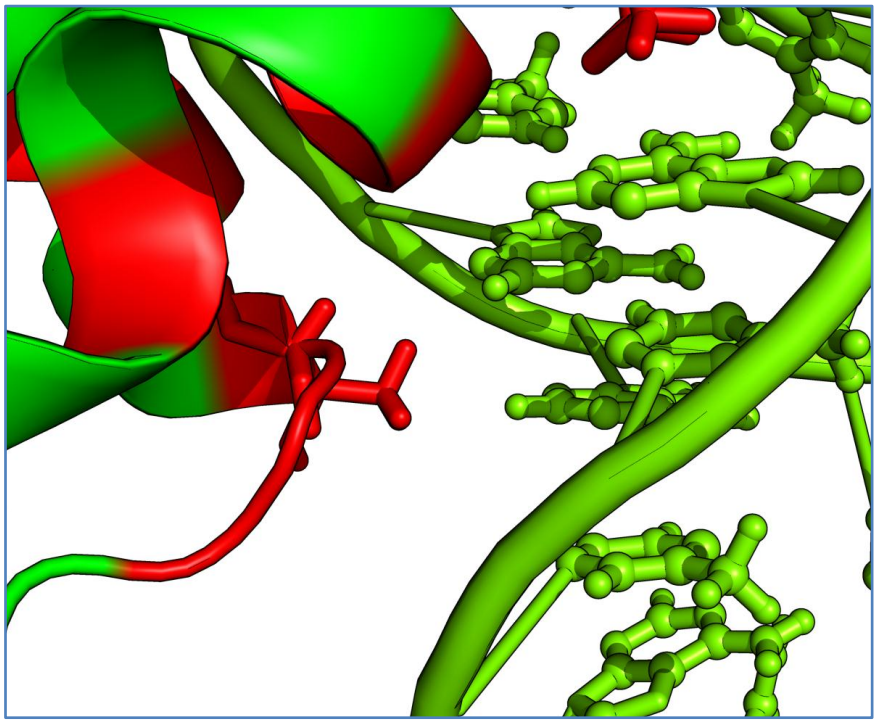


Figure 62 : Vue de détail de la position de la chaîne latérale au niveau du résidu 33 (représentée sous forme de bâtons, en rouge). Le squelette de la protéine est représenté sous forme d'un ruban en vert et rouge (pour les résidus dont au moins un proton est affecté par l'effet des courants de cycle) et l'ADN est en vert. On constate que le groupement méthyle est très proche de bases de l'ADN, ce qui se traduit par un très fort effet calculé des courants de cycle.

2.1.2 Protéine Tn916

Le complexe Tn916/ADN propose un mode de liaison particulier entre la protéine et l'ADN. En effet, l'interface est composée d'un feuillet β à trois brins s'insérant dans le grand sillon de l'ADN. La protéine Tn916 a une taille équivalente à la précédente mais l'on dispose d'un nombre plus important de déplacements chimiques (en particulier pour la forme complexée, voir Tableau 10). Pour cette protéine, on observe une variation du déplacement chimique entre la forme libre et la forme en complexe (Figure 65 Haut). Comme pour HMG, les pics les plus importants affectent des atomes HN ou Hα. On peut remarquer d'emblée que les atomes dont la variation du déplacement chimique est supérieure à 0,2 ppm appartiennent quasiment tous à des résidus à l'interface entre la protéine et l'ADN. Les exceptions sont des atomes HN de boucle, sauf pour deux atomes du résidu 51 dont un présente un pic inférieur à -1 ppm. Il est probable que ce pic soit dû

à une erreur d'attribution car ce résidu particulier ne montre pas de différence notable entre la forme libre et la forme liée (Figure 64), mais il est aussi possible que le pic soit causé par une autre contribution du déplacement chimique (comme un effet de charge par exemple).

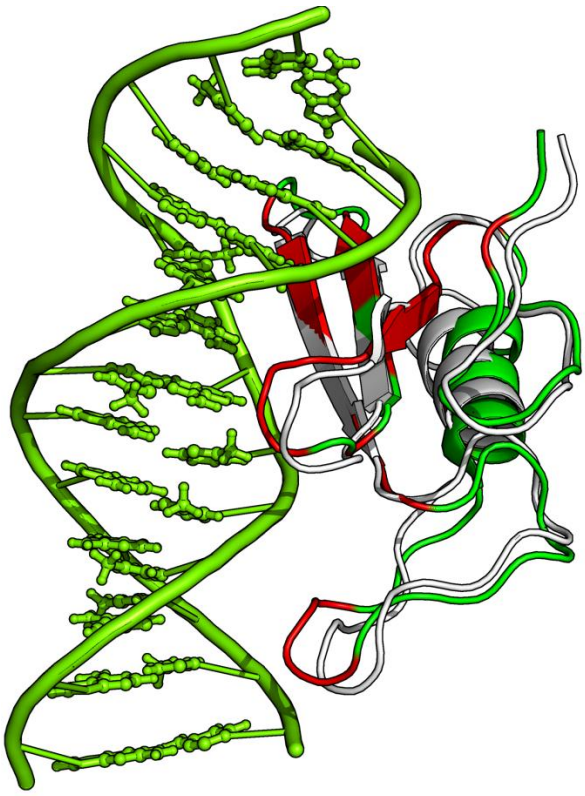


Figure 63 : Structure de la protéine Tn916 dans la forme libre (blanc) et la forme complexée avec l'ADN (vert). Les résidus en rouge correspondent aux résidus à l'interface. Le RMSD entre les deux structures est de 1,887 Å (906 atomes).

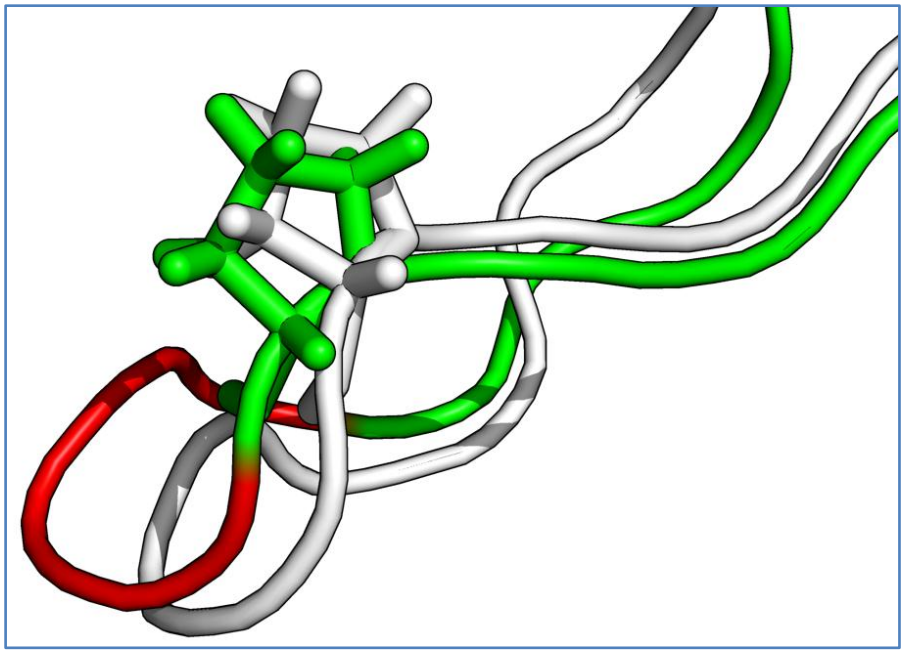


Figure 64 : Zoom sur le résidu 51 de la protéine Tn916. La position du résidu et de sa chaîne latérale ne varie que très peu entre la forme libre (en blanc) et la forme liée (en vert et rouge).

Pour mettre en valeur les atomes dont le déplacement chimique est plus affecté par l'interaction avec l'ADN que par le changement de géométrie, les atomes HN et Hα ont été retirés de l'analyse pour la Figure 65 Bas. En dehors de l'interface, il ne reste plus que les deux atomes du résidu dont le déplacement chimique change de plus de 0,2 ppm. L'étude de la variation du déplacement chimique des protons des chaînes latérales permet donc d'identifier la région d'interaction. Malheureusement, les programmes de prédiction de déplacement chimique ne sont pas capables à ce jour de calculer les valeurs pour ces atomes. De plus, ces programmes ne gèrent pas encore les effets des acides nucléiques sur les atomes de la protéine.

Comme il n'est pas possible de prédire les déplacements chimiques avec ces programmes, nous allons étudier la contribution du courant de cycle. Cette étude a pour but de déterminer s'il est possible d'identifier la région d'interface entre la protéine et l'ADN en calculant la valeur du courant de cycle engendré par les cycles des bases sur les protons de la protéine. Dans le complexe Tn916/ADN, le courant de cycle affecte 24 atomes sur 7 résidus, tous compris dans la région d'interface. En valeur absolue, la moyenne est de la contribution du courant de cycle est de 0,13 ppm. En particulier, pour les résidus 21, 36 et 38 le courant de cycle atteint à peine 0,10 ppm. Si on place un seuil à

cette valeur, le courant de cycle permet d'identifier 4 résidus sur les 26 de la région d'interface.

Il est maintenant intéressant de déterminer s'il existe une corrélation entre la variation du déplacement chimique et la valeur du courant de cycle. La Figure 66 représente pour chaque point affecté par un courant de cycle la valeur de celui-ci en fonction de la variation du déplacement chimique. Il n'y a pas de corrélation évidente si on traite l'ensemble de ces points (courbe rouge). On peut remarquer la présence de 4 points en dehors des quadrants +/+ et -/-. L'étude des atomes associés à ces points montre que l'opposition de signe entre la variation du déplacement chimique et du courant de cycle peut être due à un courant de cycle venant d'un résidu voisin de la protéine (non pris en compte ici) ou à une rotation de la chaîne latérale. Si l'on enlève ces 4, alors la corrélation est plus évidente (courbe noire).

Après étude du complexe Tn916 il est possible de conclure que l'étude de la variation du déplacement chimique des protons des chaînes latérales permet d'identifier les résidus de l'interface. L'étude des courants de cycle indique que le taux de couverture de la région d'interaction (nombre de résidu affecté par les courants de cycle/nombre de résidu de l'interface) est faible. Cependant, il semble qu'une corrélation existe entre la valeur du courant de cycle et la variation du déplacement chimique.

Figure 65 Haut : Différence de déplacement chimique entre la forme complexée et libre de Tn916 (bleu) et valeur du courant de cycle calculé pour la forme complexée (rouge) pour chaque atome de chaque résidu (numéro du résidu sur l'axe des abscisses).

Bas : Idem que précédent sans les atomes HN et Hα. Comme on ne dispose pas du même nombre de déplacement chimique pour chaque résidu pour ces deux figures, l'axe des abscisses n'est pas linéaire. Les zones en vert correspondent aux régions d'interface avec l'ADN.

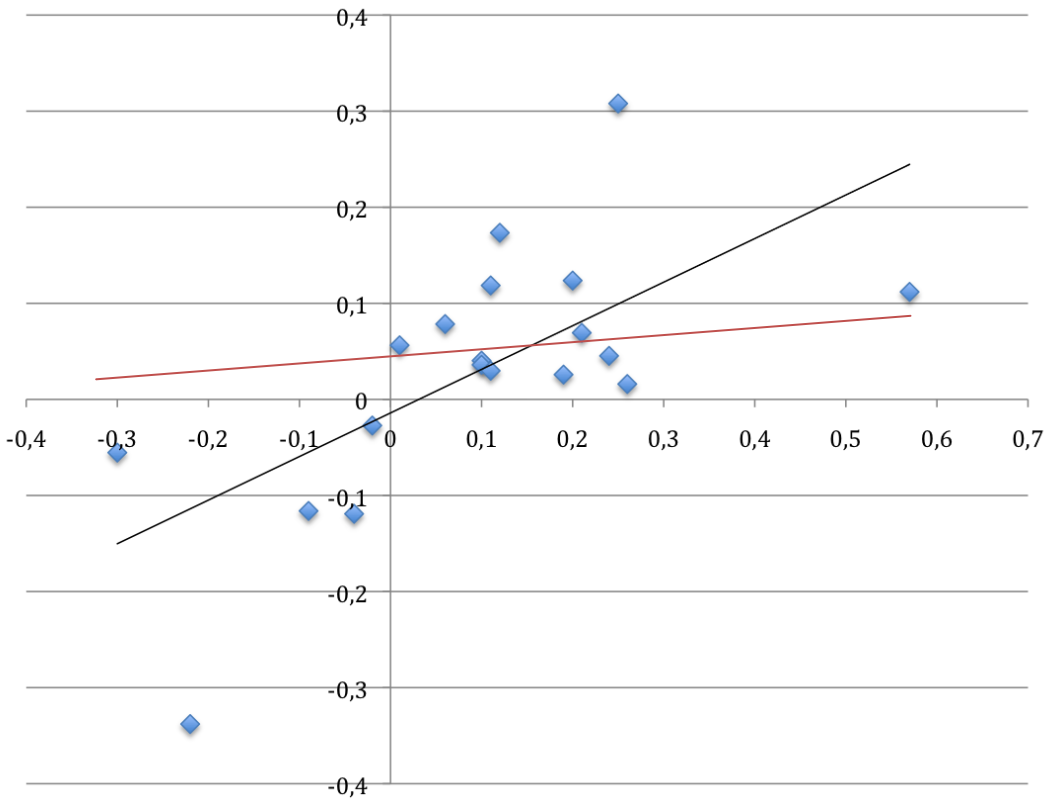


Figure 66 : Essai de corrélation entre la valeur du courant de cycle (en ordonnée) et la variation de déplacement chimique entre la forme libre et liée des protons de la protéine (en abscisse). À l'exception de quatre protons (discutés dans le texte), on constate que les courants de cycle et les variations mesurées de déplacement chimique sont de même signe. Lorsque l'on retire ces quatre protons, les données peuvent être approximées par une droite passant pratiquement par l'origine et de coefficient de corrélation de 0,43. L'échantillonnage est certainement trop faible pour être significatif mais l'ensemble de ces valeurs suggère que les courants de cycle contribuent pour environ la moitié de la variation observée.

2.2 Étude sur une banque de complexes

L'étude de deux cas « expérimentaux » ne permet pas de tirer des conclusions statistiquement valables. Par conséquent, nous avons choisi d'analyser la valeur de la contribution du courant de cycle dans un ensemble de complexes protéines/ADN extraits de la PDB. Comme précédemment, la première étape est de savoir si les courants de cycle affectent suffisamment de résidus pour pouvoir identifier la région de la protéine impliquée dans l'interaction avec l'acide nucléique. Cette étape permet de déterminer s'il

est possible de trouver suffisamment de points d'ancrage pour limiter le nombre de conformations compatibles avec les données expérimentales. La deuxième étape de cette étude est de déterminer de façon statistique si la perturbation du déplacement chimique causée par le courant de cycle est significative par rapport à la dispersion des déplacements chimiques observés.

2.2.1 Choix des structures

La PDB contient 1828 structures contenant à la fois une protéine et une molécule d'ADN. Cependant, nombre de ces structures sont trop proches en termes de structure et/ou de séquence ou sont composées de molécules de trop petite taille. Pour faciliter le choix des structures, la banque de complexes ADN/protéine 3d-footprint a été choisie. Composée par Contreras-Moreira (Contreras-Moreira 2009), cette banque regroupait au moment de son téléchargement 239 structures de complexes. Mise à jour de façon hebdomadaire, cette banque utilise les approches de contact cumulatif (Morozov & Siggia 2007) et de DNAPROT (Angarica et al. 2008) pour extraire les nouvelles structures de la PDB. Chaque structure ajoutée est corrigée si besoin (problèmes de coordonnées, notamment pour l'ADN) et si plusieurs modèles de complexes sont présents c'est celui présentant le plus de contacts ADN-Protéine qui sera choisi. Les structures ne sont ajoutées que si elles présentent une similarité de séquence inférieure à 70% avec les séquences des protéines déjà présentes. Les structures de la base de données sont dépourvues d'atomes d'hydrogène et comme leur présence est nécessaire pour le calcul des courants de cycle, ils ont été ajoutés à l'aide du programme REDUCE (Word et al. 1999).

2.2.2 Identification de la région d'interaction

En partant de la banque de données 3d-footprint, chaque structure a été copiée sous deux formes, une contenant le complexe protéine-ADN et l'autre ne contenant que la protéine. Pour chacune de ces formes, le coefficient d'accessibilité au solvant est calculé pour chaque atome à l'aide de l'algorithme de Shrake-Rupley (Shrake & Rupley 1973). Cet algorithme consiste à déplacer une sphère de rayon fixé à 1.4 Å (soit le rayon d'une molécule d'eau) sur la surface de la protéine et de déterminer pour chaque atome de la

surface l'espace qui a pu être exploré par la sphère (Figure 67). Une fois les surfaces accessibles de chacune des formes déterminées, l'identification de la région d'interaction revient à repérer les atomes pour lesquels la valeur de la surface accessible change.

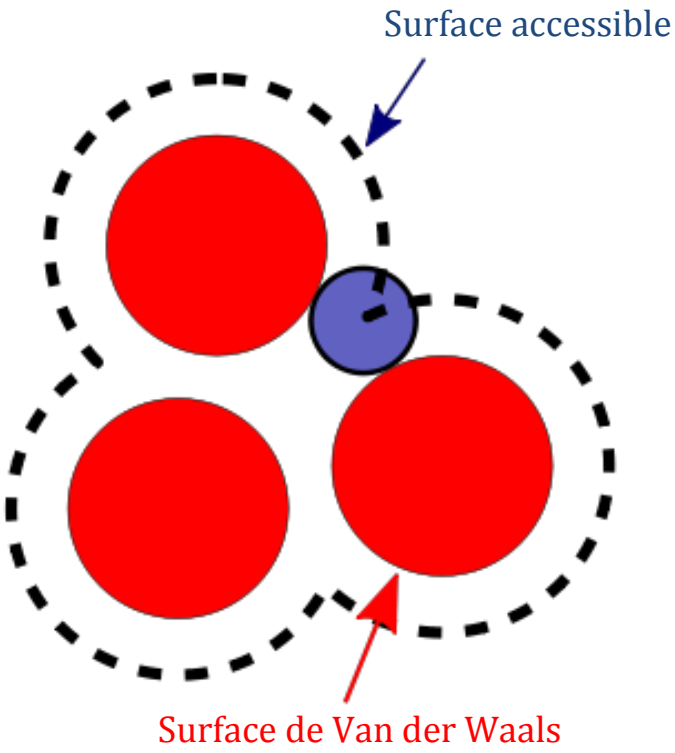


Figure 67 : Représentation schématique de la détermination de surface accessible. La sphère bleue correspond à une molécule d'eau

Une fois cette région d'interaction identifiée, on peut s'intéresser à la distribution des acides aminés dans cette interface. Dans un article de 1999 (S. Jones et al. 1999), les auteurs avaient étudié la tendance à trouver les acides aminés dans les interfaces protéine-protéine et protéine-ADN lorsqu'ils sont à la surface de la protéine. Ce travail avait été fait sur la base de 26 complexes protéine-ADN, ce qui est faible par rapport au nombre de structures de la banque de complexes (256). Nous avons refait cette analyse à partir de la banque de complexes 3d-footprint (Figure 68) car ces valeurs de tendance sont utiles pour déterminer quels résidus sont plus présents à l'interface qu'à la surface.

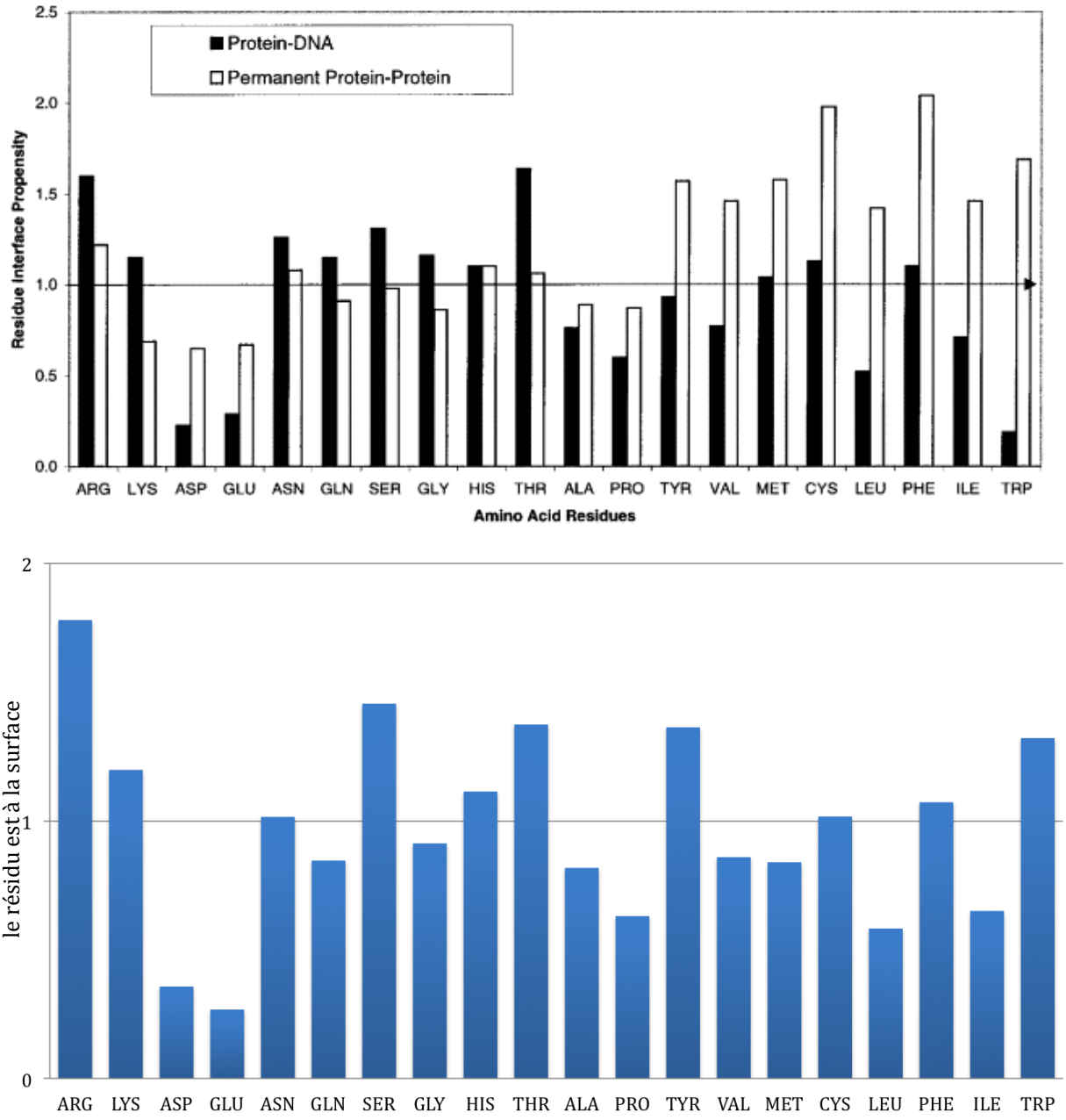


Figure 68 Haut : Tendance à trouver chaque acide aminé dans une interface protéine-ADN lorsqu'ils sont présents à la surface de la protéine d'après l'article de Jones et Thornton en 1999. Bas : À partir de la banque de complexes 3d-footprint en utilisant la même formule. Les acides aminés sont triés par ordre croissant d'hydrophobicité. Une tendance supérieure à 1 indique que l'on trouve plus fréquemment le résidu à l'interface qu'à la surface de la protéine.

Les deux distributions sont assez similaires, à l'exception de la thréonine et du tryptophane qui présentent une tendance beaucoup plus élevée dans notre analyse que dans celle de 1999. À titre de comparaison, la Figure 69 représente pour chaque type d'acide aminé le nombre de fois où il est présent dans une interface protéine-ADN. À

l'exception du tryptophane et de la tyrosine, on retrouve les résidus présentant les plus fortes tendances.

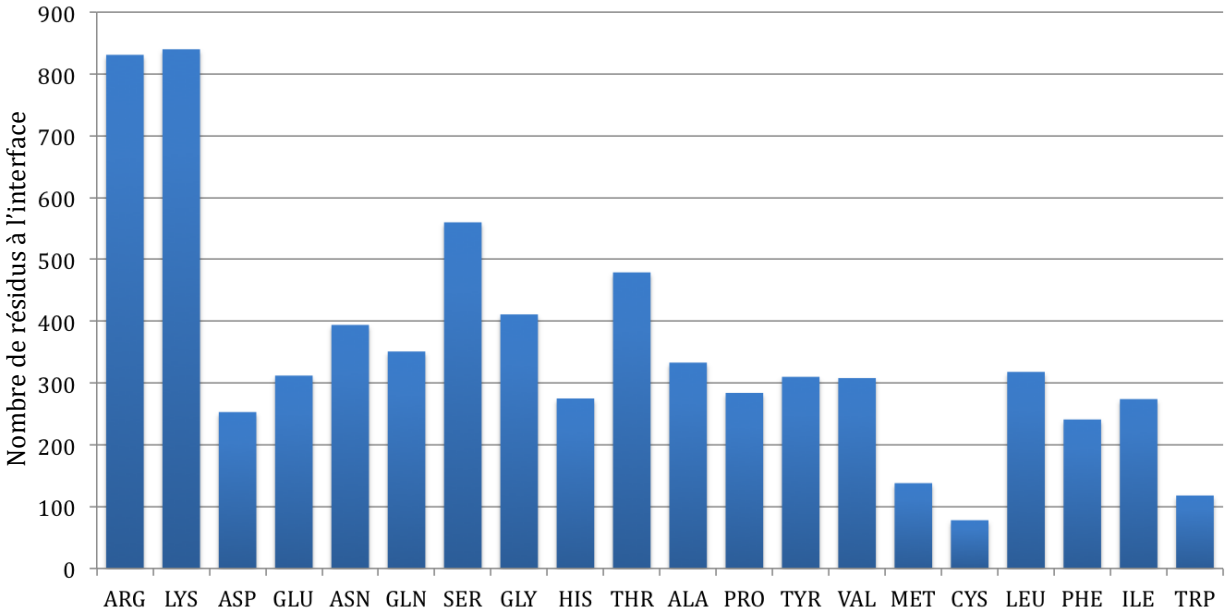


Figure 69 : Nombre de résidus impliqués dans l'interface protéine-ADN pour tous les complexes de la banque 3d-footprint.

2.3 TAUX DE COUVERTURE DES INTERFACES PAR LES COURANTS DE CYCLE

Dans un premier temps, nous nous sommes intéressés à comparer, pour chaque protéine, la surface d'interaction (définie comme la surface de la protéine masquée par l'acide nucléique) et la surface affectée par les courants de cycles. Les deux exemples décrits précédemment suggèrent que les taux de recouvrement peuvent être très variables et nous nous sommes demandé si cette observation est générale. Par ailleurs, la façon la plus classique de caractériser une interface est d'utiliser une HSQC proton-azote. Or l'hydrogène amide ne semble pas forcément être le meilleurs rapporteur, dans la mesure où, au vue du tableau de Wishart (Tableau 8), son déplacement chimique est essentiellement conditionné par la géométrie du squelette et relativement peu par les autres facteurs. Par ailleurs, toujours au vue des deux exemples décrits, ces protons semblent éloignés des bases de l'ADN mais nous voulions néanmoins tester cette hypothèse. Une autre façon de caractériser l'interface est d'utiliser une HSQC proton

carbone. Il s'agissait donc de savoir si certains protons aliphatiques constitueraient de meilleurs rapporteurs.

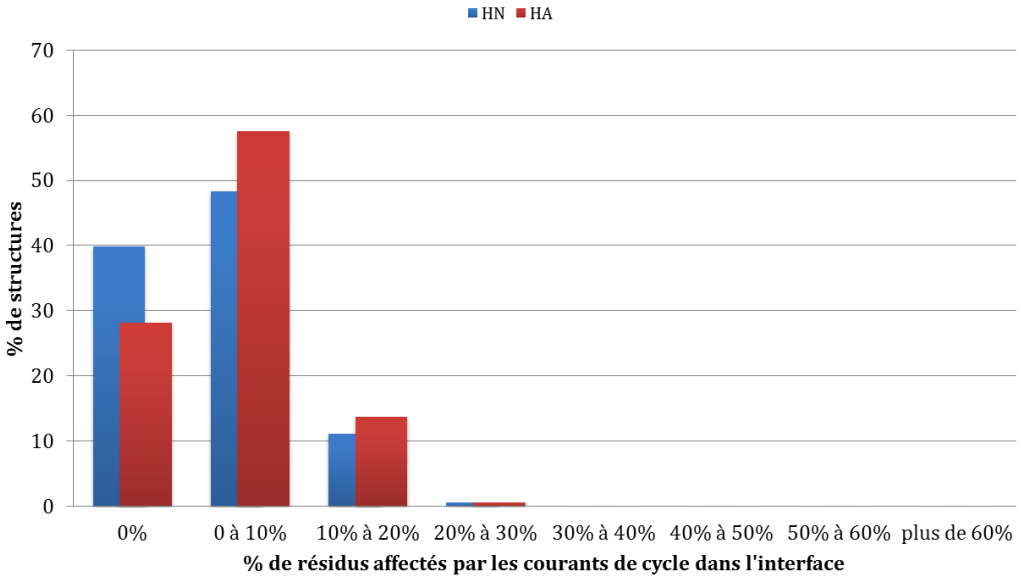


Figure 70 : Distribution du rapport du nombre d'acides aminés dont un proton HN (en bleu) ou HA (en rouge) est affecté par l'effet des courants de cycle au nombre d'acides aminés dans l'interface protéine-ADN au sein des 256 complexes de la banque de données 3d-footprint. On constate que dans 28% (HA) et 40% (HN) des complexes, aucun atome HA ou HN n'est affecté par l'effet des courants de cycle. Seuls 12 à 15% des complexes présentent un rapport supérieur à 10%. Les protons HN et HA semblent être en général trop éloignés des bases de l'ADN pour être significativement affectés.

D'après la Figure 71 et la Figure 72, il est clair que les HN ne constituent pas une bonne sonde des courants de cycle. Pour près de la moitié des structures, le taux de couverture est égal à 0 et pour le reste des structures ce taux est majoritairement compris entre 0 et 10%. Aucune structure ne présente un taux de couverture supérieur à 20%. Lorsque l'on regarde la valeur du courant de cycle, on observe des valeurs comprises entre 0 et 0,1 ppm pour plus de 50% d'entre elles. Il est aussi intéressant de remarquer qu'on peut observer une certaine corrélation entre le nombre de résidus pour lequel un HN ou HA est affecté et le type d'acide aminé (Figure 72). En effet, plus la chaîne de l'acide aminé est courte plus on observe de résidus de ce type dont un proton HA affecté (et dans une moindre mesure de proton HN). Par ailleurs, le nombre « élevé » de valeurs pour les résidus arginine, asparagine, lysine et thréonine peut s'expliquer par leur fréquence dans les interfaces (Figure 69).

Le faible nombre de valeurs, mis en relation avec le faible taux de couverture, peut s'expliquer par la distance proton HN/cycles des bases de l'ADN et probablement aussi par la sensibilité de ces protons aux déformations du squelette. Cette analyse est valable aussi pour les Hα.

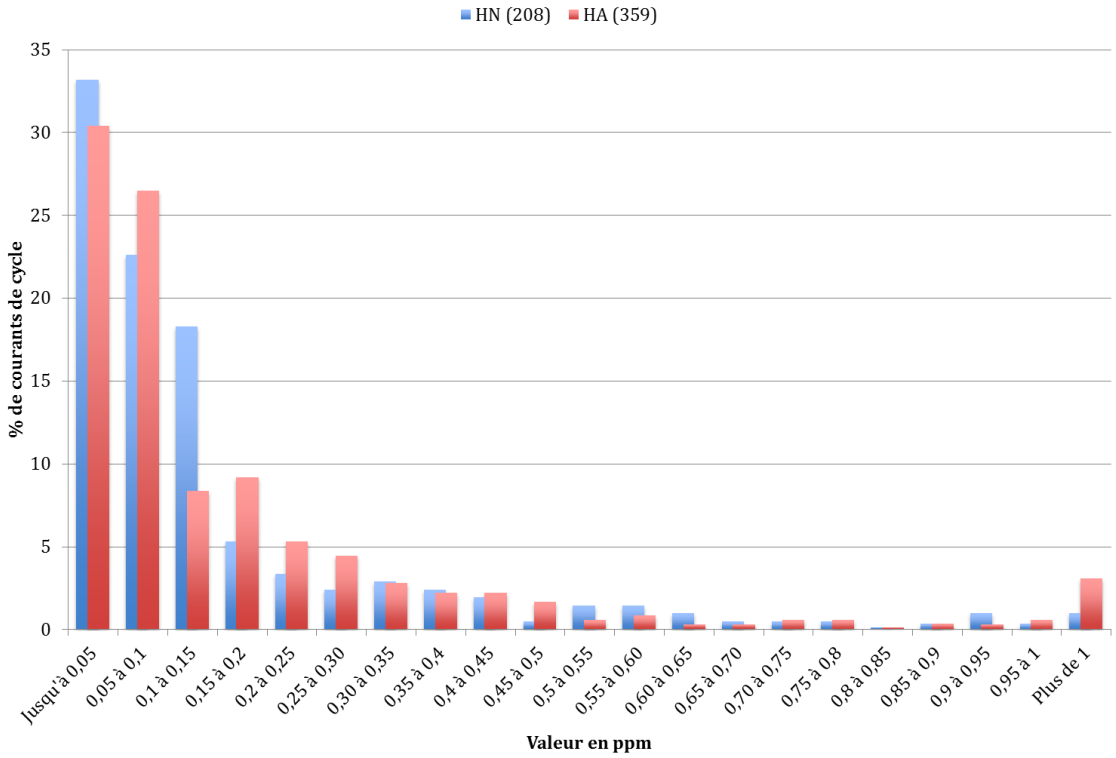


Figure 71 : Distribution du pourcentage de courants de cycle observé sur les protons classés par plages de valeurs. La première colonne regroupe les valeurs des effets de courants de cycle comprises entre 0 (non inclus) et 0,05 (inclus).

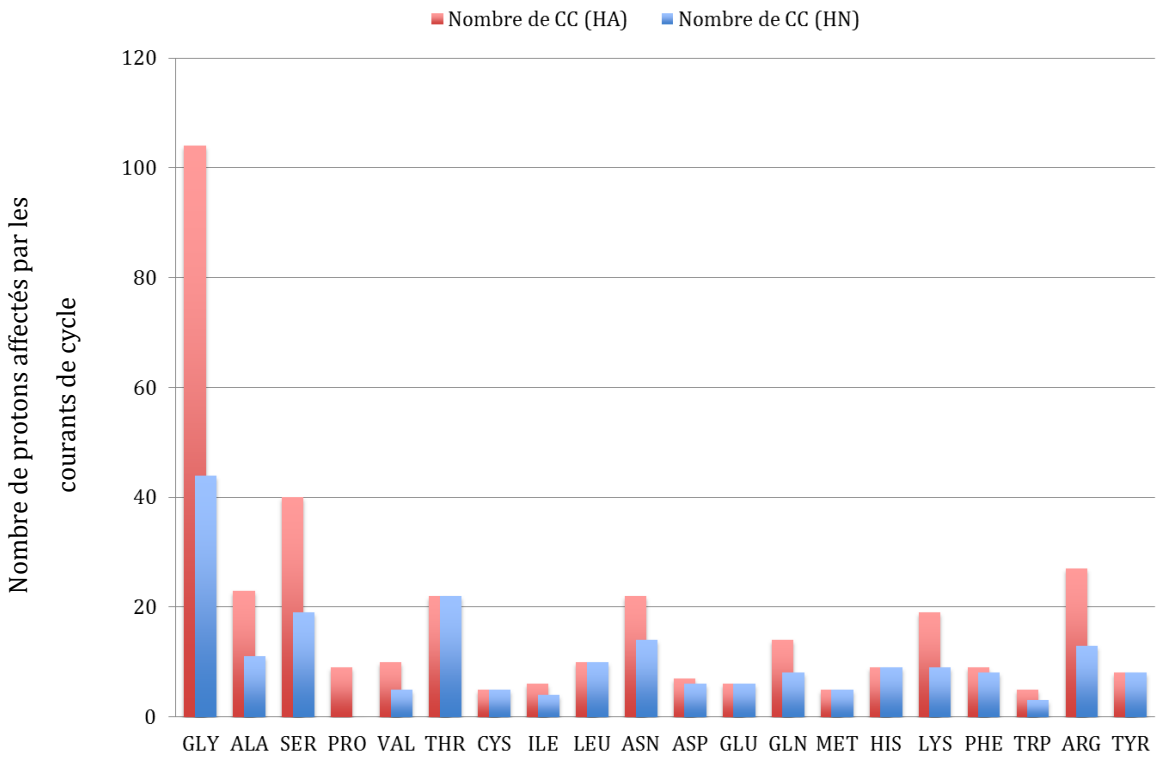


Figure 72 : Distribution du nombre de valeurs de courants de cycle par type d'acide aminé. Les acides aminés sont triés par longueur de chaîne latérale.

Le taux de couvertures a ensuite été calculé en prenant en compte l'ensemble des protons en faisant le rapport du nombre de protons affectés sur le nombre total de protons à l'interface. Dans ce cas, 70 % des structures présentent entre 20 à 40% des résidus de l'interface affecté par au moins un courant de cycle (en bleu sur la Figure 73). Cette distribution varie peu en plaçant le seuil à 0,02 (en vert sur la Figure 73). À titre indicatif, le taux de couverture baisse de façon importante si l'on place un seuil à 0,1 ppm pour le courant de cycle (en rouge sur la Figure 73). Dans ce cas, plus de 50% des structures présentent un taux de couverture inférieur à 20%. Au vu de ces résultats, il est possible de dire que le taux de couverture des résidus à l'interface par les courants de cycle est faible.

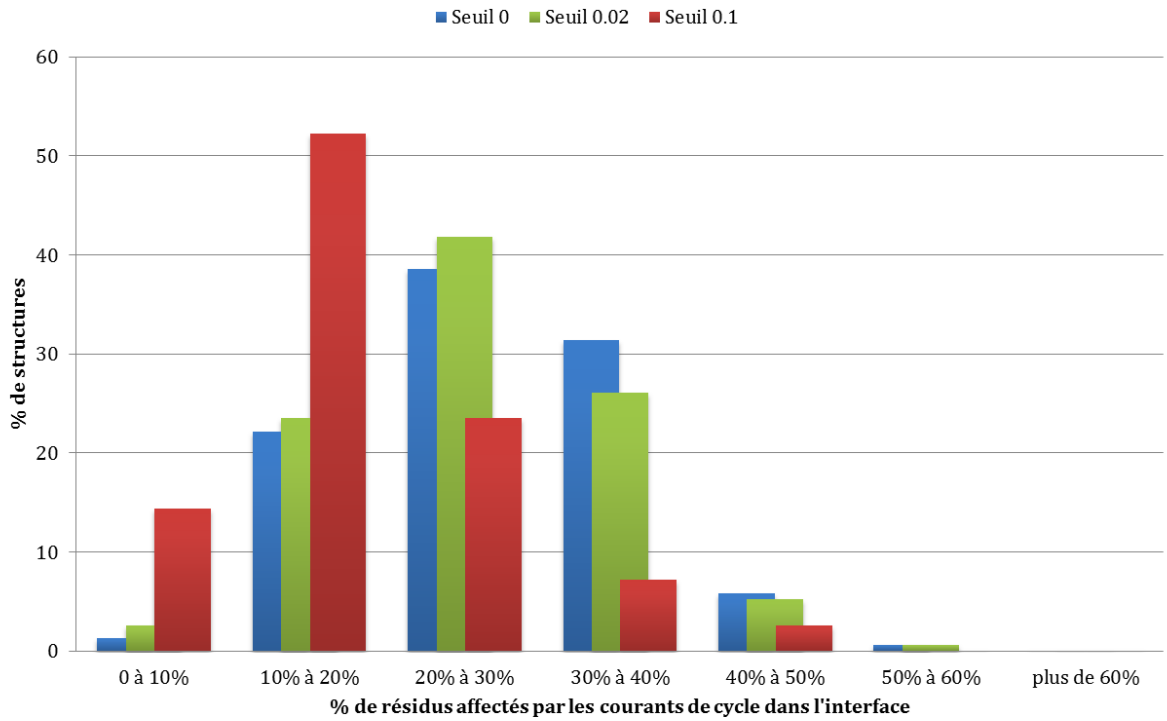


Figure 73 : Distribution du taux de couverture de l'interface de l'effet des courants de cycle. Trois seuils sont représentés selon la valeur de l'effet : différent de 0 (en bleu) ou de valeur absolue supérieure à 0,02 ppm (en vert) et 0,1 (en rouge).

Il existe néanmoins des cas favorables avec un taux de couverture supérieur à 30%. Ces structures présentent une hélice α bien insérée dans le grand sillon de l'ADN (Figure 74). C'est avec ce type de complexes qu'il sera éventuellement possible d'utiliser les courants de cycle pour repérer l'interface. À l'autre extrémité de la distribution, on observe toute une série de configuration conduisant à un faible taux de protons affectés par les courants de cycle (Figure 75). Dans ces différents cas, il semble bien que les interactions entre la protéine et l'ADN se font essentiellement au niveau du squelette ribose-phosphate ou à travers de petits éléments de structures insérés dans le grand sillon, ce qui conduit à un éloignement de la protéine des bases de l'ADN. Ces complexes sont relativement nombreux dans la base de données et correspondent à des protéines se fixant de manière non spécifique à l'ADN. Il est peu probable que l'on puisse utiliser les courants de cycle pour ces types de complexes.

Dans ces cas, les bases de la molécule d'ADN sont trop éloignées des protons des chaînes latérales pour que les courants de cycles les affectent. Ces complexes sont relativement nombreux dans la base de données et correspondent à des protéines se

fixant de manière non spécifique à l'ADN. Il est peu probable que l'on puisse utiliser les courants de cycle pour ces types de complexes.

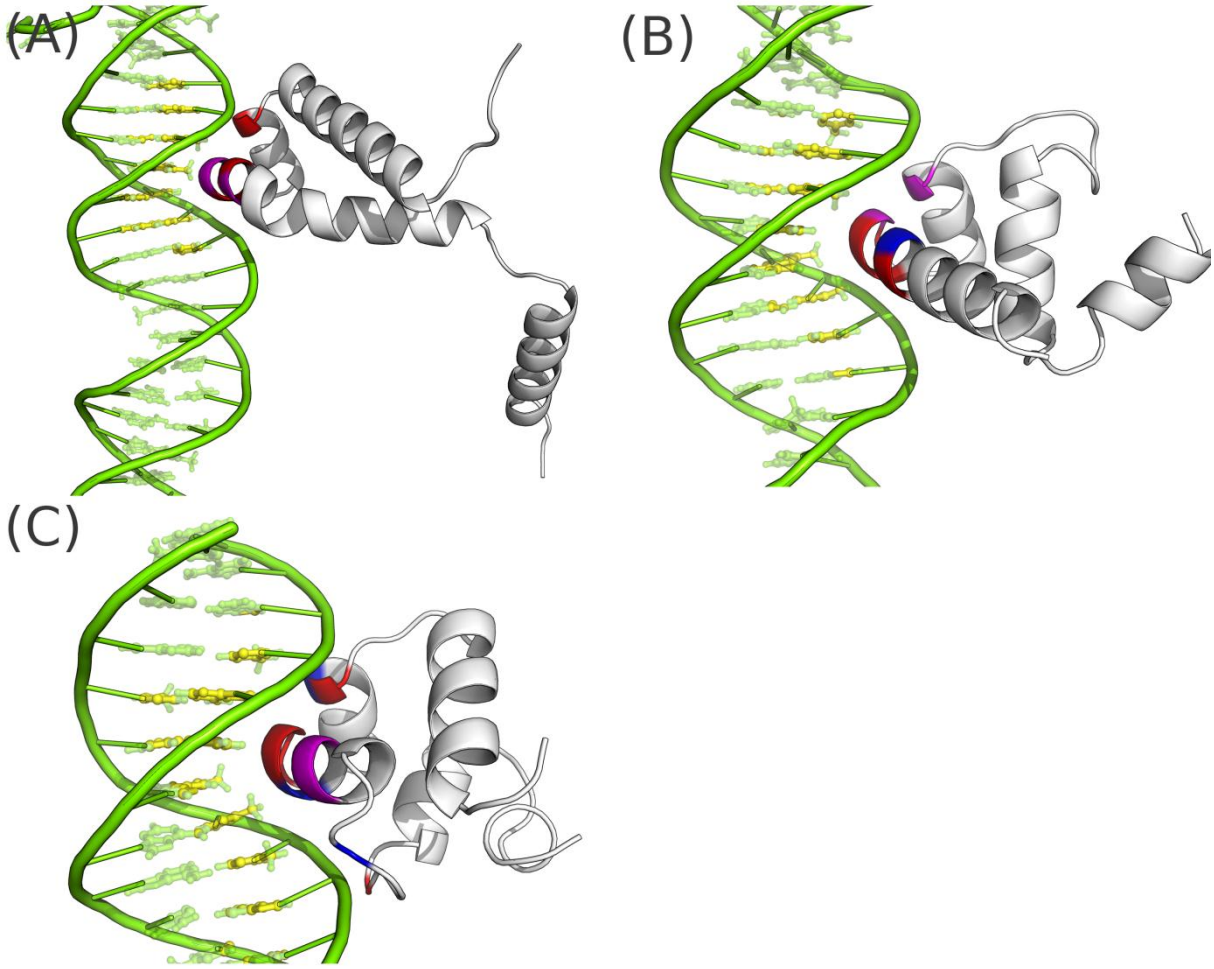


Figure 74 : Structures présentant une forte couverture de l'interface par les courants de cycle les résidus colorés correspondent aux résidus affectés par les courants de cycle causés par les bases de l'ADN en jaune. En bleu : résidus dont les protons sont affectés par un effet de courant de cycle de signe négatif, en rouge par un effet positif et en violet les deux à la fois.

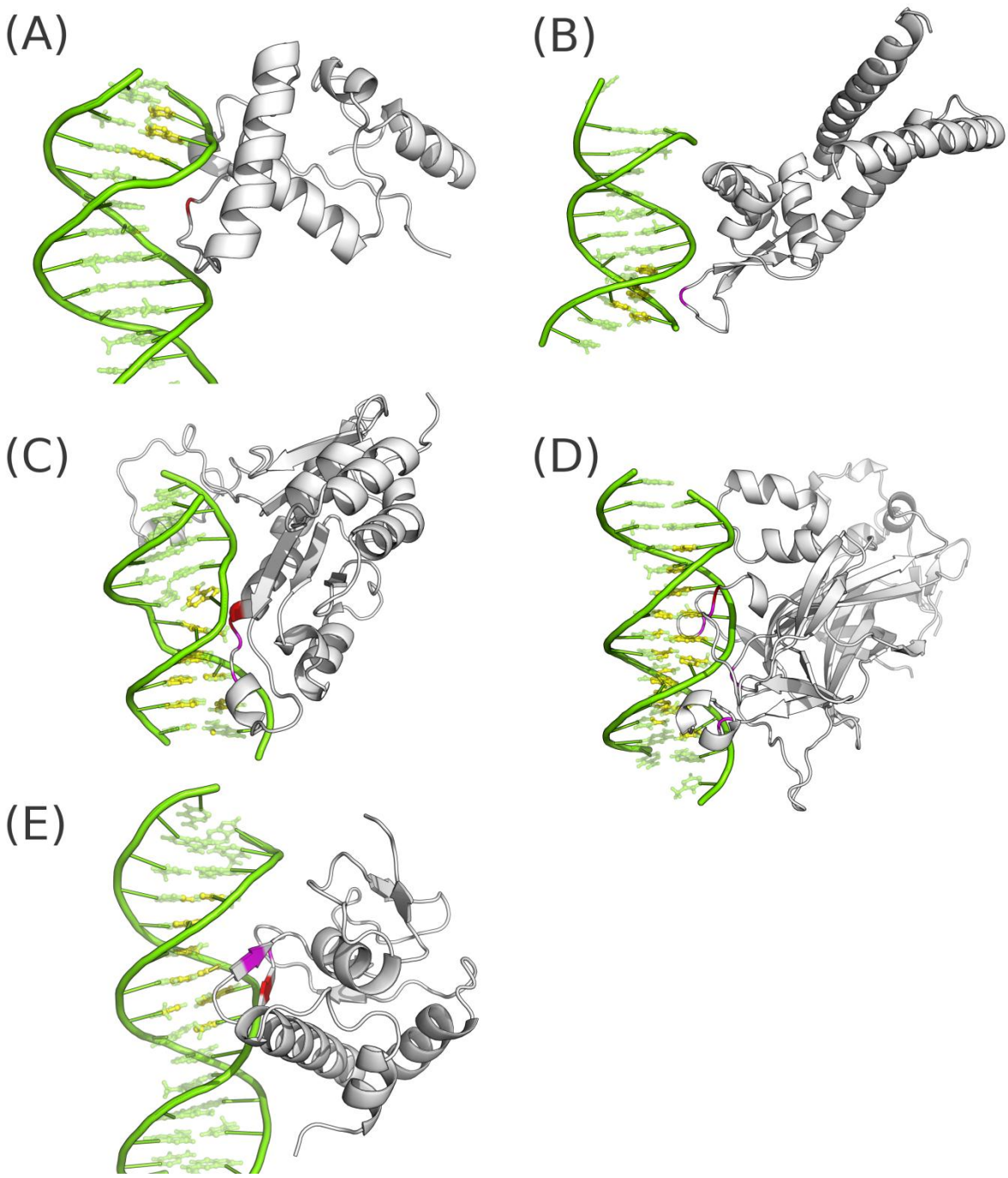


Figure 75 : Structures présentant une faible couverture de l'interface par les courants de cycle les résidus colorés correspondent aux résidus affectés par les courants de cycle causés par les bases de l'ADN en jaune. Descriptions des cas : ADN éloigné de la protéine (A et B), peptide de liaison inséré dans le sillon (C), hélice α mal insérée dans le sillon de l'ADN (D) et interface composée de feuillet(s) β (E)

3 Évaluation de la contribution du courant de cycle dans le déplacement chimique

Si l'on admet qu'un certain nombre de protons aliphatiques sont suffisamment proches de l'acide nucléique et suffisamment nombreux dans les interfaces pour présenter un intérêt en ce qui concerne notre étude, une seconde question se pose : comment déterminer la part des courants de cycles dans leurs variations éventuelles de déplacement chimique ? Une contribution importante indiquerait en effet qu'il y aurait bien un gain à attendre de l'introduction du calcul des courants de cycle dans un programme d'arrimage tandis qu'une contribution faible signifierait qu'il faudrait plutôt s'intéresser à d'autres contributions. Le deuxième exemple traité au début de cette étude suggère que la contribution est significative, dans la mesure où nous voyons qu'il y a pratiquement toujours accord entre le signe de la variation et le signe de l'effet du courant. Cette observation n'étant pas généralisable, fautes d'un nombre suffisant de complexes pour lesquels nous aurions les déplacements chimiques de la forme libre et de la forme liée à l'ADN, nous avons essayé de trouver un autre estimateur. La question est de comparer l'effet calculable des courants de cycles aux autres sources de variations possibles – les changements de géométrie, les changements de solvatations, les effets de la présence des charges. Une façon d'estimer ces changements est de comparer pour un groupement donné, l'effet du courant de cycle que nous pouvons calculer aux variations de déplacements chimiques constatés dans les protéines lorsque cette chaîne change de structure et d'environnement. Dans un premier temps, nous avons simplement cherché à estimer cette variation à travers l'écart-type de la distribution des déplacements chimiques reportés dans la BMRB pour chacun des types de protons. Cette approche sous-estime probablement certains effets (comme celui des charges) mais permet néanmoins de se faire une première idée.

Dans ce but, une fois les déplacements chimiques téléchargés, il a été possible d'associer à chaque identifiant BMRB, le type de structure auquel il appartient en utilisant la « Query Grid Interface ». Les différents types de structures répertoriés sont les protéines seules ou en complexe avec d'autres protéines, les complexes protéine-ADN, les complexes protéine-ARN et les complexes protéine-ligand. Ici, seuls les deux premiers

types nous ont intéressés : les protéines seules et les complexes protéine-ADN. Au moment de l'analyse, 37 entrées de complexes protéines/ADN et 4893 entrées contenant uniquement des protéines ont pu être exploitées. Pour ces deux catégories, chaque type d'atome et de résidu la moyenne, l'écart type, la valeur minimale et maximale du déplacement chimique sont calculés.

3.1 Pour les protons HN et Ha

Nous avons suggéré précédemment que les protons H^N et H^α n'étaient pas une bonne sonde pour les courants de cycle. Il est quand même intéressant de comparer l'écart type du déplacement chimique à la valeur moyenne observée pour le courant de cycle. Les tableaux ci-dessous confirment notre analyse. La variabilité du déplacement chimique de ces protons est grande et l'influence moyenne des courants de cycles dans les interfaces est faible. Dans le détail, la situation semble un peu plus favorable pour les protons H^a que pour les protons H^N (leur variabilité est sensiblement plus basse) et pour les acides aminés à chaîne courte que pour les acides aminés à chaîne longue (l'effet moyen du courant de cycle est plus important). Le calcul de la valeur moyenne du courant de cycle a été effectué en faisant la moyenne de la valeur absolue des courants de cycle des atomes affectés, c'est à dire présentant une valeur de courant de cycle différent de 0.

Tableau 11 : Statistiques des valeurs de déplacement chimique et de courant de cycle pour les protons HN par type de résidu. Les résidus sont triés par taille de chaîne latérale et par poids moléculaire. δ=Déplacement chimique, CC=Courant de Cycle.

Résidu	Moyenne δ	Ecart-type δ	Moyenne CC	Ecart-type CC	Nombre CC	Nombre total CC	% de CC
GLY	8.256	0.651	0.213	0.261	44	148	29.73%
ALA	8.187	0.58	0.11	0.125	11	122	9.02%
SER	8.4	0.652	0.075	0.05	19	330	5.76%
VAL	8.141	0.675	0.087	0.043	5	176	2.84%
THR	8.125	0.665	0.083	0.116	22	281	7.83%
CYS	8.437	0.567	0.115	0.113	5	35	14.29%

ILE	8.123	0.656	0.192	0.226	4	167	2.40%
LEU	8.087	0.714	0.199	0.149	10	252	3.97%
ASN	8.28	0.638	0.2	0.182	14	259	5.41%
ASP	8.338	0.587	0.143	0.169	6	61	9.84%
GLN	8.116	0.687	0.079	0.044	8	281	2.85%
GLU	8.341	0.693	0.098	0.067	6	106	5.66%
MET	8.303	0.584	0.371	0.534	5	110	4.55%
HIS	7.872	0.717	0.155	0.136	9	165	5.45%
LYS	8.083	0.532	0.139	0.167	9	686	1.31%
PHE	8.451	0.75	0.04	0.034	8	183	4.37%
TRP	8.47	0.729	0.101	0.108	3	90	3.33%
ARG	8.126	0.608	0.127	0.088	13	1385	0.94%
TYR	8.132	0.657	0.257	0.304	8	254	3.15%

Tableau 12 : Statistiques des valeurs de déplacement chimique et de l'effet des courants de cycle pour les protons Hα par type de résidu. Les résidus sont triés par taille de chaîne latérale et par poids moléculaire.

Résidu	Moyenne δ	Ecart-type δ	Moyenne CC	Ecart-type CC	Nombre CC	Nombre total CC	% de CC
GLY	3.907	0.313	0.261	0.382	104	148	70.27%
ALA	4.129	0.338	0.166	0.389	23	122	18.85%
SER	4.355	0.394	0.072	0.071	40	330	12.12%
PRO	4.354	0.33	0.189	0.167	9	135	6.67%
VAL	4.054	0.61	0.097	0.068	10	176	5.68%
THR	4.322	0.505	0.131	0.151	22	281	7.83%
CYS	4.455	0.503	0.14	0.113	5	35	14.29%

ILE	4.017	0.638	0.209	0.261	6	167	3.59%
LEU	4.15	0.371	0.335	0.308	10	252	3.97%
ASN	4.587	0.372	0.182	0.305	22	259	8.49%
ASP	4.54	0.245	0.101	0.136	7	61	11.48%
GLN	4.126	0.284	0.139	0.194	14	281	4.98%
GLU	4.144	0.312	0.071	0.065	6	106	5.66%
MET	4.286	0.482	0.126	0.058	5	110	4.55%
HIS	4.469	0.487	0.144	0.139	9	165	5.45%
LYS	4.187	0.373	0.181	0.235	19	686	2.77%
PHE	4.524	0.685	0.214	0.355	9	183	4.92%
TRP	4.448	0.489	0.152	0.104	5	90	5.56%
ARG	4.125	0.393	0.129	0.137	27	1385	1.95%

3.2 Pour les protons en bout de chaîne latérale

Dans cette partie de l'étude, nous nous sommes intéressés aux protons en bout de chaîne latérale et en particulier aux protons des groupes méthyles et amines en bout de chaîne. Nous avons choisi ces protons parce que d'une part ils sont au bout de la chaîne latérale et donc potentiellement le plus proche de l'ADN et d'autre part parce qu'ils sont facilement repérables par RMN. Pour chacun de ces protons, la valeur de l'écart type du déplacement chimique a été déterminée à partir des données des complexes protéines/ADN de la BMRB. Ensuite, un script détermine le nombre de courants de cycle supérieur à cet écart-type pour chaque couple acide aminé-proton.

Tableau 13 : Statistiques des valeurs de déplacement chimique et de courant de cycle pour les protons des méthyles par type de résidu. ET=Ecart-type, « Nombre affectés par les CC» correspond au nombre de résidu de ce type affectés par au moins un courant de cycle. « Nombre à l'interface » correspond au nombre de fois où l'on trouve ce résidu dans une interface protéine-ADN. Pour la leucine, on observe plus de courants de cycle supérieur à l'écart type que de résidus affectés par les courants de cycle car plusieurs protons d'un méthyle peuvent être affectés.

Résidu	Type de proton	ET δ	Moyenne CC	ET CC	CC>ET δ	Nombre Affectés par les CC	Nombre à l'interface
ALA	HB	0,222	0,157	0,224	16	45	333
THR	HG	0,907	0,145	0,257	5	85	479
VAL	HG	0,248	0,171	0,244	32	42	308
ILE	HD	0,322	0,192	0,307	8	33	274
LEU	HD	0,302	0,313	0,466	46	43	318
MET	HE	0,519	0,436	0,791	9	26	138

Tableau 14 : Statistiques des valeurs de déplacement chimique et de courant de cycle pour les protons des amines en bout de chaîne latérale, par type de résidu

Résidu	Type de proton	ET δ	Moyenne CC	EC CC	CC > ET δ	Nombre Affectés par les CC	Nombre à l'interface
ARG	HH	0.57	0.266	0.311	93	192	831
ASN	HD	0.523	0.251	0.297	24	79	394
GLN	HE	0.485	0.337	0.38	34	80	351
LYS	HZ	1.039	0.277	0.297	10	108	840

Tableau 15 : Statistiques des valeurs de déplacement chimique et de courant de cycle pour les protons des cycles aromatiques en bout de chaîne latérale, par type de résidu

Résidu	Type de proton	ET δ	Moyenne CC	EC CC	CC > ET δ	Nombre Affectés par les CC	Nombre à l'interface
HIS	HD	0,423	0,183	0,224	5	53	275
	HE	0,458	0,255	0,32	12	53	275
PHE	HD	0,441	0,302	0,401	9	39	241
	HE	0,363	0,562	0,654	24	39	241
	HZ	0,445	0,577	0,563	11	39	241
TRP	HE	1,849	0,221	0,338	0	21	118
	HZ	0,596	0,198	0,246	2	21	118
	HH	0,537	0,219	0,229	1	21	118
TYR	HD	0,298	0,234	0,44	12	66	310
	HE	0,246	0,186	0,259	17	66	310

On constate que la situation est, en effet, nettement plus favorable dans deux cas. Dans le cas de méthyles, à l'exception de celui de la thréonine situé en a d'un groupement hydroxyle et donc fortement susceptible d'être affecté indirectement, la variabilité des déplacements chimiques est faible tandis que l'effet du courant de cycle est relativement important. Ceci conduit, dans le cas des valines et des leucines en particulier, à un nombre important de méthyles pour lesquels l'effet du courant de cycle est supérieur à la variabilité du déplacement chimique.

Dans le cas des protons des amines et imines, la variabilité est, comme on pouvait s'y attendre, plus importante. Le nombre de situation, en particulier dans le cas des arginines, asparagines et glutamines pour laquelle l'effet du courant de cycle semble prédominant est grand.

4 CONCLUSION

Dans la littérature, les courants de cycle ont déjà été utilisés pour repérer l'interface entre un enzyme et son substrat (Kleinpeter & Klod 2004) et pour l'arrimage d'un ligand sur une protéine (Cioffi et al. 2009). Ces études ont montré les limites de l'utilisation des courants de cycle, en particulier au niveau du type de résidu affecté. Pour les complexes protéine-ADN, il était communément admis (Schmiedeskamp et al. 1997) que les variations de déplacement chimique les plus importantes observées sur les protons des résidus de la protéine sont dues aux effets des courants de cycle produits par les bases de l'ADN. L'étude présentée ici montre qu'il est nécessaire de nuancer un peu cette observation car on observe seulement quelques résidus susceptibles de contenir des protons dont les variations de déplacement chimique sont essentiellement liées aux courants de cycle.

L'utilisation des courants de cycle comme sonde pour repérer les résidus d'une protéine à l'interface avec l'ADN est envisageable principalement dans le cas de complexes présentant des résidus valine et leucine à la surface. Dans ce cas, le courant de cycle permet non seulement de repérer les résidus en interaction, mais permet en plus de les évaluer quantitativement. Cette information peut ensuite être utilisée par des logiciels d'arrimage pour affiner les régions d'interaction. En effet, il serait possible de calculer, pour chaque modèle généré par le programme, la valeur des courants de cycle sur les protons affectés de ces résidus et les comparer à la variation de déplacement chimique expérimentale. Ainsi, il est possible de définir un terme de score, dépendant de l'accord entre la valeur des courants de cycle calculés et de la variation du déplacement chimique expérimental observé.

L'étude a été réalisée ici sur des complexes protéine-ADN pour une raison de quantité de données car, à ce jour, il n'existe pas de librairie de structures spécifiques aux complexes protéines/ARN. Néanmoins, il serait intéressant d'étudier ce deuxième type de complexe protéine/acide nucléique car lorsque l'ARN est en forme simple brin, les cycles de l'ARN sont beaucoup plus proches des résidus de la protéine. Il serait alors possible que l'effet des courants de cycle soit plus important à l'interface.

CONCLUSION GÉNÉRALE

À la suite de ces différentes études, il est possible de mettre en valeur les différents apports des données faiblement résolues dans la résolution de la structure de protéines. Dans le cas de la protéine PBP1b, nous avons proposé une nouvelle approche pour l'utilisation des données de SAXS en tenant en compte la spécificité de l'objet d'étude. Les résultats obtenus ont permis d'identifier une plus grande variabilité dans la position relative des deux domaines ce qui confirme l'intérêt du programme Dadimodo. Ce programme, contrairement aux autres logiciels développés en interne dans les autres laboratoires, sera bientôt disponible à la communauté scientifique. Au moment de la rédaction de cette thèse, seuls les programmes MODELLER et XPLOR-NIH permettent d'évaluer l'accord des structures avec une courbe de diffusion. Dans MODELLER cette fonction est encore expérimentale et n'est pas utilisée pour discriminer les modèles les uns par rapport aux autres. Nous espérons que la diffusion de notre outil permettra de développer l'utilisation du SAXS dans des laboratoires de petite et/ou de moyenne taille qui n'ont pas la possibilité de développer des outils en interne.

L'utilisation des effets NOE dans la résolution de structure à l'aide du programme Rosetta a donné des résultats satisfaisants mais a nécessité une longue période d'apprentissage du fait du manque de documentation. Depuis 2000, il n'y a pas eu d'études présentant l'intérêt de l'ajout de ce type de contrainte malgré une nette amélioration du programme. À l'issue de ce travail, nous proposons un protocole constitué de différents fichiers de configuration Rosetta ainsi que d'un exemple de fichier de contrainte. Une méthodologie particulière de calibration des effets NOE a aussi été présentée. D'après les résultats obtenus, il est possible de conseiller l'utilisation de Rosetta dans le cadre d'un laboratoire où le programme est déjà installé pour faciliter la résolution de structures de protéines de petite taille (inférieures à 130 résidus) en permettant d'obtenir rapidement des premiers modèles fiables (même en l'absence de protéines homologues) qui peuvent être utilisés pour accélérer le processus d'attribution. Dans le cas particulier de la protéine PPRV-Δ459N, l'utilisation des données expérimentales a permis d'obtenir des modèles avec une bonne précision sur la localisation des hélices ainsi que des chaînes latérales. Le modèle obtenu pourra être utilisé pour comparaison avec d'autres protéines virales du même type.

Les effets des courants de cycle ont fait l'objet d'un grand nombre d'études mais leur application dans le cadre de problèmes d'arrimage n'est pas encore très répandue. L'étude des effets des courants de cycle dans le cadre de l'arrimage protéine-ADN que nous avons proposé a permis d'évaluer leur influence sur le déplacement chimique des atomes de la protéine à l'interface ce qui n'avait pas été réalisé précédemment à grande échelle. Les résultats de cette étude sont mitigés du fait des faibles valeurs des effets calculés. Il est possible, dans certaines conformations de complexes présentées dans le chapitre, d'observer suffisamment d'effets pour établir des points d'ancrage et les évaluer de manière quantitative. Il serait intéressant d'appliquer cette approche à des complexes protéine/ARN notamment lorsque l'ARN est en simple brin car la distance entre bases et acides aminés est plus faible. Pour faciliter cette étude, il est préférable d'attendre l'apparition de banques de structures spécifiques aux complexes protéines/ARN.

En conclusion, je tiens à signaler l'importante évolution des techniques de résolution de structure de protéines depuis ces dix dernières années. D'un côté, les techniques dites « expérimentales » bénéficient de l'automatisation d'un grand nombre d'étapes et, de l'autre, les techniques dites « *in silico* » se sont nettement améliorées, notamment grâce à la croissance importante des bases de données. Depuis l'apparition de techniques mixtes, il est possible d'obtenir des modèles de protéines avec une quantité limitée de données, même dans des cas plutôt défavorables aux techniques expérimentales classiques tels que ceux présentés dans cette thèse. Il est possible que l'on puisse utiliser, dans un futur proche, les techniques mixtes pour apporter des informations supplémentaires dans des simulations comme par exemple en dynamique moléculaire.

RÉFÉRENCES

1. Aloy, Patrick et al., 1998. Modelling repressor proteins docking to DNA. *Proteins: Structure, Function, and Genetics*, 33(4), pp.535-549.
2. Altona, C., Faber, D.H. & Hoekzema, A.J.A.W., 2000. Double-helical DNA1H chemical shifts: an accurate and balanced predictive empirical scheme. *Magnetic Resonance in Chemistry*, 38(2), p.95-107.
3. Anfinsen, C.B., 1973. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096), p.223-230
4. Angarica, V. et al., 2008. Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, 9(1), p.436.
5. Barton, G.J. & Russell, R.B., 1993. Protein structure prediction. *Nature*, 361(6412), pp.505-506.
6. Bax, A., 2003. Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Science*, 12(1), pp.1-16.
7. Benner, S.A., Cohen, M.A. & Gerloff, D., 1992. Correct structure prediction? *Nature*, 359(6398), pp.781-781.
8. Berjanskii, M. et al., 2009. GeNMR: a web server for rapid NMR-based protein structure determination. *Nucl. Acids Res.*, 37(suppl_2), pp.W670-677.
9. Bernado, P., 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences*, 102(47), pp.17002-17007.
10. Blackledge, M., 2005. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 46(1), pp.23-61.
11. Bowers, P.M., Strauss, Charlie E.M. & Baker, D., 2000. De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR*, 18(4), p.311-318.
12. Bradley, P., Misura, K.M.S. & Baker, D., 2005b. Toward High-Resolution de Novo Structure Prediction for Small Proteins. *Science*, 309(5742), pp.1868-1871.
13. Bystroff, C. & Baker, D., 1998b. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of Molecular Biology*, 281(3), pp.565-577.
14. Case, D.A., 1995. Calibration of ring-current effects in proteins and nucleic acids. *Journal of Biomolecular NMR*, 6(4), p.341-346.
15. Cavalli, A. et al., 2007. Protein structure determination from NMR chemical shifts. *Proceedings of the National Academy of Sciences*, 104(23), p.9615-9620.
16. Chen, Y. et al., 2004c. A new hydrogen-bonding potential for the design of protein–RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Research*, 32(17), pp.5147 -5162.
17. Cioffi, M. et al., 2009. Use of quantitative 1H NMR chemical shift changes for ligand docking into barnase. *Journal of Biomolecular NMR*, 43(1), p.11-19.
18. Colubri, A. et al., 2006. Minimalist Representations and the Importance of Nearest Neighbor Effects in Protein Folding Simulations. *Journal of Molecular Biology*, 363(4), p.835-857.
19. Contreras-Moreira, B., 2009. 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucl. Acids Res.*, p.gkp781.

20. Cornilescu, G., Delaglio, F. & Bax, Ad, 1999. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13(3), p.289-302.
21. Das, R. et al., 2009e. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences*, 106(45), pp.18978 - 18983.
22. David E. Goldberg, 1991. A comparative analysis of selection schemes used in genetic algorithms.
23. Dijk, M. van & Bonvin, A.M.J.J., 2009. 3D-DART: a DNA structure modelling server. *Nucl. Acids Res.*, p.gkp287.
24. Dijk, M. van & Bonvin, A.M.J.J., 2010. Pushing the limits of what is achievable in protein-DNA docking: benchmarking HADDOCK's performance. *Nucl. Acids Res.*, p.gkq222.
25. Dijk, M. van et al., 2006. Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Research*, 34(11), pp.3317 -3325.
26. Dominguez, C., Boelens, R. & Bonvin, A.M.J.J., 2003. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7), pp.1731-1737.
27. Emsley, P. & Cowtan, K., 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D Biological Crystallography*, 60(12), pp.2126-2132.
28. Förster, F. et al., 2008. Integration of Small-Angle X-Ray Scattering Data into Structural Modeling of Proteins and Their Assemblies. *Journal of Molecular Biology*, 382(4), pp.1089-1106.
29. Gabel, F., Simon, B. & Sattler, M., 2006. A target function for quaternary structural refinement from small angle scattering and NMR orientational restraints. *European Biophysics Journal*, 35(4), pp.313-327.
30. Goffin, C. & Ghuysen, J.-M., 1998. Multimodular Penicillin-Binding Proteins: An Enigmatic Family of Orthologs and Paralogs. *Microbiol. Mol. Biol. Rev.*, 62(4), p.1079-1093.
31. Gray, J.J. et al., 2003 Protein-Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. *Journal of Molecular Biology*, 331(1), pp.281-299.
32. Grishaev, A. et al., 2007. Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. *Journal of Biomolecular NMR*, 40(2), p.95-106.
33. Grishaev, A. et al., 2005. Refinement of Multidomain Protein Structures by Combination of Solution Small-Angle X-ray Scattering and NMR Data. *Journal of the American Chemical Society*, 127(47), p.16621-16628.
34. Gultyaev, A.P., van Batenburg, F.H.D. & Pleij, C.W.A., 1995. The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm. *Journal of Molecular Biology*, 250(1), p.37-51.
35. Haigh, C.W. & Mallion, R.B., 1979. Ring current theories in nuclear magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 13(4), p.303-344.
36. Havranek, J.J., Duarte, C.M. & Baker, D., 2004. A Simple Physical Model for the Prediction and Design of Protein-DNA Interactions. *Journal of Molecular Biology*, 344(1), p.59-70.
37. Helles, G., 2008. A comparative study of the reported performance of ab initio protein structure prediction algorithms. *Journal of The Royal Society Interface*, 5(21), pp.387 - 396.
38. Holland, J., 1992. *Adaptation in Natural and Artificial Systems*, MIT Press.

39. Hunter, C.A., Packer, M.J. & Zonta, C., 2005. From structure to chemical shift and vice-versa. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 47(1-2), pp.27-39.
40. Johnson, C.E. & Bovey, F.A., 1958. Calculation of Nuclear Magnetic Resonance Spectra of Aromatic Hydrocarbons. *The Journal of Chemical Physics*, 29(5), p.1012.
41. Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2), pp.195-202.
42. Jones, S. et al., 1999. Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology*, 287(5), p.877-896.
43. Karplus, K. et al., 1999. Predicting protein structure using only sequence information. *Proteins: Structure, Function, and Genetics*, 37(S3), pp.121-125.
44. Kaufmann, K.W. et al., 2010. Practically Useful: What the Rosetta Protein Modeling Suite Can Do for You. *Biochemistry*, 49(14), p.2987-2998.
45. King, R.D. et al., 1997. DSC: public domain protein secondary structure predication. *Computer Applications in the Biosciences: CABIOS*, 13(4), pp.473-474.
46. Kleinpeter, E. & Klod, S., 2004. Ab initio calculation of the anisotropic/ring current effects of amino acid residues to locate the position of substrates in the binding site of enzymes. *Journal of Molecular Structure*, 704(1-3), p.79-82.
47. Knegtel, Ronald M.A, Boelens, R. & Kaptein, R., 1994. Monte Carlo docking of protein-DNA complexes: incorporation of DNA flexibility and experimental data. *Protein Engineering*, 7(6), pp.761 -768.
48. Knegtel, Ronald M.A. et al., 1994. MONTY: a Monte Carlo approach to protein-DNA recognition. *Journal of Molecular Biology*, 235(1), pp.318-324.
49. Koch, M.H., Vachette, P. & Svergun, Dmitri I, 2003. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Quarterly Reviews of Biophysics*, 36(2), pp.147-227.
50. Kohlhoff, K.J. et al., 2009. Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *Journal of the American Chemical Society*, 0(0).
51. Kortemme, T. et al., 2004. Computational redesign of protein-protein interaction specificity. *Nat Struct Mol Biol*, 11(4), p.371-379.
52. Kuhlman, B. et al., 2003. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649), p.1364-1368.
53. Kuszewski, J., Gronenborn, A.M. & Clore, G.M., 1996. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science*, 5(6), p.1067-1080
54. Lovering, A.L., De Castro, L. & Strynadka, N.C.J., 2008. Identification of Dynamic Structural Motifs Involved in Peptidoglycan Glycosyltransfer. *Journal of Molecular Biology*, 383(1), p.167-177.
55. Macheboeuf, P. et al., 2011. Solution X-ray scattering study of a full-length class A penicillin-binding protein. *Biochemical and Biophysical Research Communications*
56. Madl, T., Gabel, F. & Sattler, M., 2010. NMR and small-angle scattering-based structural analysis of protein complexes in solution. *Journal of Structural Biology*, In Press
57. Mareuil, F. et al., 2007. A simple genetic algorithm for the optimization of multidomain protein homology models driven by NMR residual dipolar coupling and small angle X-ray scattering data. *European Biophysics Journal*, 37(1), p.95-104.
58. Mattinen, M.-L. et al., 2002. Quaternary Structure Built from Subunits Combining NMR and Small-Angle X-Ray Scattering Data. *Biophysical Journal*, 83(2), p.1177-1183.

59. Montalvao, R.W. et al., 2008b. Structure Determination of Protein−Protein Complexes Using NMR Chemical Shifts: Case of an Endonuclease Colicin−Immunity Protein Complex. *Journal of the American Chemical Society*, 130(47), p.15990-15996.
60. Morozov, A.V. & Siggia, E.D., 2007. Connecting protein structure with predictions of regulatory sites. *Proceedings of the National Academy of Sciences*, 104(17), p.7068-7073.
61. Moult, J., 1999. Predicting protein three-dimensional structure. *Current Opinion in Biotechnology*, 10(6), pp.583-588.
62. Moult, J. et al., 2007. Critical assessment of methods of protein structure prediction—Round VII. *Proteins*, 69(S8), p.3-9.
63. Moyna, G. et al., 1998. Comparison of Ring Current Methods for Use in Molecular Modeling Refinement of NMR Derived Three-Dimensional Structures. *Journal of Chemical Information and Computer Sciences*, 38(4), p.702-709.
64. Neal, S. et al., 2003. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *Journal of Biomolecular NMR*, 26(3), p.215-240.
65. Neylon, C., 2008. Small angle neutron and X-ray scattering in structural biology: recent examples from the literature. *European Biophysics Journal*, 37(5), p.531-541.
66. Perkins, S.J. & Dwek, R.A., 1980. Comparisons of ring-current shifts calculated from the crystal structure of egg white lysozyme of hen with the proton nuclear magnetic resonance spectrum of lysozyme in solution. *Biochemistry*, 19(2), p.245-258.
67. Petoukhov, M.V. & Svergun, Dmitri I., 2005. Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. *Biophysical Journal*, 89(2), p.1237-1250.
68. Pople, J.A., 1958. Molecular orbital theory of aromatic ring currents. *Molecular Physics: An International Journal at the Interface Between Chemistry and Physics*, 1(2), p.175.
69. Prestegard, J.H., al-Hashimi, H.M. & Tolman, J.R., 2000. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Quarterly Reviews of Biophysics*, 33(4), pp.371-424.
70. Raman, S. et al., 2009i. Accurate Automated Protein NMR Structure Determination Using Unassigned NOESY Data. *Journal of the American Chemical Society*, 0(0).
71. Robustelli, P. et al., 2010. Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure*, 18(8), pp.923-933.
72. Rohl, C.A., 2005. Protein Structure Estimation from Minimal Restraints Using Rosetta. In *Nuclear Magnetic Resonance of Biological Macromolecules*. Academic Press, pp. 244-260.
73. Rohl, C.A. et al., 2004. Protein Structure Prediction Using Rosetta. In *Numerical Computer Methods, Part D*. Academic Press, pp. 66-93.
74. Rost, B., Sander, C. & Schneider, R., 1994. PHD-an automatic mail server for protein secondary structure prediction. *Computer applications in the biosciences : CABIOS*, 10(1), pp.53 -60.
75. Schmiedeskamp, M., Rajagopal, P. & Klevit, R.E., 1997. NMR chemical shift perturbation mapping of dna binding by a zinc-finger domain from the yeast transcription factor ADR1. *Protein Science*, 6(9), p.1835-1848.
76. Shapiro, B.A. & Navetta, J., 1994. A massively parallel genetic algorithm for RNA secondary structure prediction. *The Journal of Supercomputing*, 8(3), p.195-207.
77. Shen, Y. & Bax, A., 2007a. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of Biomolecular NMR*, 38(4), p.289-302.

78. Shen, Y. et al., 2009. TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. *Journal of Biomolecular NMR*, 44(4), p.213-223.
79. Shen, Y. et al., 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12), p.4685-4690.
80. Shen, Y. et al., 2009. De novo protein structure generation from incomplete chemical shift assignments. *Journal of Biomolecular NMR*, 43(2), p.63-78.
81. Shortle, D., Simons, K.T. & Baker, D., 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 95(19), pp.11158 -11162.
82. Shrake, A. & Rupley, J.A., 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79(2), p.351-364.
83. Simons, K.T. et al., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1), pp.209-225.
84. Sternberg, M.J. et al., 1998. A computational system for modelling flexible protein-protein and protein-DNA docking. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6, pp.183-192.
85. Svergun, D., Barberato, C. & Koch, M.H.J., 1995. CRYSOL – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*, 28(6), p.768-773.
86. Svergun, D.I., 1999. Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. *Biophysical Journal*, 76(6), p.2879-2886.
87. Svergun, Dmitri I. & Koch, Michel H. J., 2002. Advances in structure analysis using small-angle scattering in solution. *Current Opinion in Structural Biology*, 12(5), pp.654-660.
88. Vila, J.A., Ripoll, D.R. & Scheraga, H.A., 2007. Use of 13Cα Chemical Shifts in Protein Structure Determination. *The Journal of Physical Chemistry B*, 111(23), pp.6577-6585.
89. Vitkup, D. et al., 2001. Completeness in structural genomics. *Nat Struct Mol Biol*, 8(6), p.559-566.
90. Wang, Y. & Jardetzky, O., 2002. Investigation of the Neighboring Residue Effects on Protein Chemical Shifts. *Journal of the American Chemical Society*, 124(47), p.14075-14084.
91. Wishart, D. S., Sykes, B.D. & Richards, F.M., 1992. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6), p.1647-1651.
92. Wishart, D.S. et al., 2008. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucl. Acids Res.*, p.gkn305.
93. Wishart, David S. & Case, D.A., 2002. Use of chemical shifts in macromolecular structure determination. In *Nuclear Magnetic Resonance of Biological Macromolecules Part A*. Academic Press, pp. 3-34.
94. Word, J.M. et al., 1999. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4), p.1735-1747.
95. Wu, S., Skolnick, J. & Zhang, Y., 2007. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology*, 5(1), p.17.

96. Yuzawa, S., 2004. Solution Structure of the Tandem Src Homology 3 Domains of p47phox in an Autoinhibited Form. *Journal of Biological Chemistry*, 279(28), pp.29752-29760.
97. Zhang, H., Neal, S. & Wishart, D.S., 2003. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25(3), p.173-195.
98. Zheng, Y. & Yang, D., 2005. STARS: statistics on inter-atomic distances and torsion angles in protein secondary structures. *Bioinformatics*, 21(12), p.2925-2926.

ANNEXE 1 : SCRIPT DE MUTATION CNS

```
set message = on end
set echo = on end

   inline @entete_cns.inp
   inline @option_cns.inp


{           Initialisation des varaibles generales}

{           Lecture des parametres, du PSF et des IC}


    structure @$psf end                          ! lecture de PSF

    set display = $disp end                                                 !
ouverture de fichier d'impression des resultats
    display   num_str  num_residu   type_res  type_muta  var_adapt var_phi
var_psi

    parameter                                              ! defini le
potentiel de vdW repulsif
       nbond repel=0.8 rcon=1 cutnb = 3.0 end
       @$par
    end


{           Algorithme genetique : Boucle de mutation}

    evaluate ($comp_str = 1)
    while ($comp_str <= $nombre_str) loop BOUCLE_STR           ! Boucle sur
les structures

      { gestion du nom de fichier }

      evaluate ($lecture  = $pdb_in  + encode($comp_str) + $suf_in)

      {lecture des coordonnees, copie dans le vecteur de reference,
remplissage de la table des IC}

      coordinate @@$lecture   ! end

      energy end

      do (XCOMP = X) (all)                                                  !
coor X vers XCOMP
      do (YCOMP = Y) (all)                                                  !
coor Y vers YCOMP
      do (ZCOMP = Z) (all)                                                  !
coor Z vers ZCOMP
      do (BCOMP = B) (all)                                                  !
facteur B vers BCOMP


      { determination du Q_seuil : Q, Q_max -> Q_seuil }
```

```
      show ave (Q) (attribute Q > -1 and (name CA or name C or name O or
name N))
      evaluate ($Q_seuil = min($Q_max,$RESULT))


      { boucle sur la decendance, jusqu'a nombre_fils }

      evaluate ($comp_essai_fils = 0)
      while ($comp_essai_fils <= 50) loop BOUCLE_ECHAP
        do (X = XCOMP) (all)
        do (Y = YCOMP) (all)
        do (Z = ZCOMP) (all)


        { Choix du residu a muter, dans la gamme 2 - (n-1) }
        { avec un biais vers les residus non conserves     }
      { premier_res, dernier_res, B -> res_mut, type_res_mute }

      evaluate ($nombre_essai = 0)                                  ! compteur
du nombre d'essais
        evaluate ($res_mut = 0)                                     ! numero de
l'aa a muter
      evaluate ($nombre_res = ($dernier_res - $premier_res) -2)   ! Nombre de
residus utiles (ni le 1 ni de dernier)

      while ($nombre_essai < $N_echap)  loop BOUCLE_RES

            evaluate ($res_mute_reel = $nombre_res*random())
      ! tirage d'un nombre reel 1-nombre_res
            evaluate ($res_mut = int($res_mute_reel) + $premier_res + 1)
      ! conversion du reel en entier
            show (B) (resid $res_mut and name CA)                   ! verifie
le type du residu lui meme
            evaluate ($type_res_mute = $RESULT)

            if ($Flag_link = 0) then
                if ($type_res_mute >= 1) then
                evaluate($Proba_choix = (1 - $C_restype))
              end if
              if ($type_res_mute <= 0) then
                evaluate($Proba_choix = $C_restype)
              end if
            end if
              if ($Flag_link = 1) then
              if ($type_res_mute = -1.00) then
                evaluate($Proba_choix = 1)
            end if
            if ($type_res_mute > -1.00) then
              evaluate($Proba_choix = -1)
            end if
            if ($type_res_mute = -2.00) then
              evaluate($Proba_choix = 1)
            end if
            end if

              evaluate ($tirage = random())
              if ($tirage <= $Proba_choix) then
                  exit BOUCLE_RES
              end if
            evaluate ($nombre_essai = $nombre_essai+1)
            if ($nombre_essai >= $N_echap) then
```

```
                exit BOUCLE_RES
              end if

        end loop BOUCLE_RES

        if ($nombre_essai >= $N_echap) then
          evaluate ($ecriture = $pdb_out + encode($comp_str) + $suf_out)
          display $comp_str  Aucun residu de ce type a muter
          evaluate ($comp_str = $comp_str + 1)
          write coordinate output = $ecriture end
          exit BOUCLE_ECHAP
        end if


        { Choix du type de mutation a realiser }
        { C_muttype -> M_type }

        evaluate ($M_type = "Globale")
        evaluate ($fin_segment = $res_mut + $N_res_locale)
        if ($fin_segment < $dernier_res) then
           evaluate ($tirage = random())
           if ($tirage <= $C_muttype) then
              evaluate ($M_type = "Locale")
           end if

        end if

        { valeur des mutations }
        { M_type, dernier_res, phi_old, psi_old Q Q_seul -> val_mut_phi,
val_mut_psi }
        { en utilisant un generateur aleatoire gaussien }

      if ($type_res_mute=-2) then
            evaluate($M_type = "LocaleEtendu")
            show ELEM (RESID) (ATTRIBUTE B=-2.50 and name CA)
            evaluate($res_fin=decode($RESULT))
      end if

        if ($M_type = "Locale") then
            evaluate ($fin_segment = $res_mut + $N_res_locale)
      end if
      if ($M_type = "Globale") then
            evaluate ($fin_segment = $dernier_res)
      end if
      if ($M_type = "LocaleEtendu") then
            evaluate ($fin_segment = $res_fin)
      end if



      if ($res_mut < $fin_segment) then
            show average (Q) (attribute Q > -1 and (resid
$res_mut:$fin_segment))
      end if
      if ($res_mut > $fin_segment) then
            show average (Q) (attribute Q > -1 and (resid
$fin_segment:$res_mut))
      end if
        evaluate ($var_adapt = $A_globale*$RESULT/$Q_seuil)        !
adaptation de la valeur de la mutation au Q
        evaluate ($var_phi = gauss($var_adapt))
```

```
        evaluate ($var_psi = gauss($var_adapt))

        { realisation des mutations }
        evaluate($res_mut_plus = $res_mut + 1)

      { mutation phi }
      show (RESN) (resid $res_mut and name CA)
      evaluate ($Residuname = $RESULT)

      if ($Residuname # "PRO") then
        show (X) (resid $res_mut and name CA)
        evaluate($Xres_mut1 = $RESULT)
        show (Y) (resid $res_mut and name CA)
        evaluate($Yres_mut1 = $RESULT)
        show (Z) (resid $res_mut and name CA)
        evaluate($Zres_mut1 = $RESULT)

        show (X) (resid $res_mut and name N)
        evaluate($Xres_mut2 = $RESULT)
        show (Y) (resid $res_mut and name N)
        evaluate($Yres_mut2 = $RESULT)
          show (Z) (resid $res_mut and name N)
        evaluate($Zres_mut2 = $RESULT)

        evaluate($Xaxis = $Xres_mut1-$Xres_mut2)
        evaluate($Yaxis = $Yres_mut1-$Yres_mut2)
        evaluate($Zaxis = $Zres_mut1-$Zres_mut2)

          coor rotate sele= (resid $res_mut_plus:$fin_segment or resid
$res_mut and not (name N or name HN))
                    center ($Xres_mut1, $Yres_mut1, $Zres_mut1)
                    axis ($Xaxis, $Yaxis, $Zaxis) $var_phi end
      end if

      { mutation psi }
      show (X) (resid $res_mut and name C)
      evaluate($Xres_mut1 = $RESULT)
      show (Y) (resid $res_mut and name C)
      evaluate($Yres_mut1 = $RESULT)
      show (Z) (resid $res_mut and name C)
      evaluate($Zres_mut1 = $RESULT)

      show (X) (resid $res_mut and name CA)
      evaluate($Xres_mut2 = $RESULT)
      show (Y) (resid $res_mut and name CA)
      evaluate($Yres_mut2 = $RESULT)
      show (Z) (resid $res_mut and name CA)
      evaluate($Zres_mut2 = $RESULT)

      evaluate($Xaxis = $Xres_mut1-$Xres_mut2)
      evaluate($Yaxis = $Yres_mut1-$Yres_mut2)
      evaluate($Zaxis = $Zres_mut1-$Zres_mut2)

      if ($res_mut < $fin_segment) then
            coor rotate sele= (resid $res_mut_plus:$fin_segment or resid
$res_mut and (name C or name O))
                  center ($Xres_mut1, $Yres_mut1, $Zres_mut1)
                  axis ($Xaxis, $Yaxis, $Zaxis) $var_psi end
      end if
      if ($res_mut > $fin_segment) then
```

```
            coor rotate sele= (resid $fin_segment:$res_mut_plus or resid
$res_mut and (name C or name O))
                  center ($Xres_mut1, $Yres_mut1, $Zres_mut1)
                  axis ($Xaxis, $Yaxis, $Zaxis) $var_psi end
      end if



           { rectification de la structure en cas de mutation locale }
           { M_type, res_mut, N_res_locale, dernier_res, Xcomp, Ycomp,
Z_comp, X,Y,Z -> X,Y,Z }

      if ($M_type = "Locale") then
           fix selection = (not resid $res_mut:$fin_segment) end
           flags exclude ELEC end
           minimize powell
              nstep 500
           end
        end if

        if ($M_type = "LocaleEtendu") then
         evaluate($debut_segment=$fin_segment-5)
         evaluate($fin_segment=$fin_segment+5)
           fix selection = (not resid $debut_segment:$fin_segment) end
           flags exclude ELEC end
           minimize powell
              nstep 500
           end
        end if

        { minimisation des chaines laterales }

        fix selection =  (name ca or name c or name n) end
        flags exclude ELEC  end


      minimize powell
           nstep = $N_minim
        end

        { evaluation}

        energy end

      { ecriture de la structure }

        if ($ENER < $E_seuil) then
             evaluate ($ecriture = $pdb_out + encode($comp_str) + $suf_out)
               display $comp_str  $res_mut  $type_res_mute $M_type
$var_adapt  $var_phi  $var_psi
             evaluate ($comp_str = $comp_str + 1)
               write coordinate output = $ecriture end
             exit BOUCLE_ECHAP!reinitialisation de la condition d'echapement
        end if

        evaluate ($comp_essai_fils = $comp_essai_fils + 1)
        if ($comp_essai_fils > 50) then
             do (X = XCOMP) (all)
               do (Y = YCOMP) (all)
               do (Z = ZCOMP) (all)
```

```
                evaluate ($ecriture = $pdb_out + encode($comp_str)+$suf_out)
                display $comp_str  Echappement
                evaluate ($comp_str = $comp_str + 1)
                write coordinate output = $ecriture end
              exit BOUCLE_ECH !réinitialisation de la condition d'échappement
          end if

        end loop BOUCLE_ECHAP

      end loop BOUCLE_STR

STOP
```

ANNEXE 2 : SCRIPT D'ESTIMATION DU VOLUME COMMUN POUR UN ENSEMBLE DE STRUCTURES

```
#!/usr/bin/env python

import __main__
__main__.pymol_argv = [ 'pymol', '-qc']
# Parametres pymol (pas de GUI et pas de messages)
import pymol
from pymol import cmd

import os,shutil
from msms_pymol import *

from progressbar import *

pymol.finish_launching()

path= "/Clusters/"
exp="035/"

outname="Volume"

distance=5

dom=1

rms={}
structures=[]
previous="none"

nbstructure=0

for j in sorted(os.listdir(path+exp)):
  if str.split(j,'.')[1]=="pdb":
    nbstructure+=1

widgets=[Bar('>'),Percentage()]
sum=0
maxval=nbstructure*(nbstructure-1)/2
pbar=ProgressBar(widgets=widgets,maxval=maxval).start()
comptiter=0

for j in sorted(os.listdir(path+exp)):
 if str.split(j,'.')[1]=="pdb":

   structname=str.split(j,'.')[0]
   cmd.load(path+exp+j,structname)

   structures.append(str.split(j,'.')[0])

   if dom==0:
     cmd.select("fit2",str.split(j,'.')[0]+" and resid 72:98 and name CA")

   if dom ==1:
     cmd.select("fit2",str.split(j,'.')[0]+" and resid 330:791 and name CA")

   if dom ==2:
     cmd.select("fit2",str.split(j,'.')[0]+" and resid 130:320 and name CA")
```

```
    if previous!="none":
      if dom==0:
        cmd.select("fit1",previous+" and resid 72:98 and name CA")

      if dom==1:
        cmd.select("fit1",previous+" and resid 330:791 and name CA")

      if dom==2:
        cmd.select("fit1",previous+" and resid 130:320 and name CA")

      cmd.align("fit2","fit1")

    previous=structname

if len(structures)==len(os.listdir(path+exp)):
     for struct in structures:
         rms[struct]={}
         for struct2 in structures:
             cmd.select("tmp1",struct+" and resid 130:320 and (n. CA n. CO n.
N n. C)")
             cmd.select("tmp2",struct2+" and resid 130:320 and (n. CA n. CO
n. N n. C)")
             nbatomtot=cmd.count_atoms("tmp1")
             if struct2 in rms and struct in rms[struct2] and
struct!=struct2:
               #rms[struct][struct2]=rms[struct2][struct]
               rms[struct][struct2]=""
               #pass
             elif struct==struct2:
               rms[struct][struct2]="%5.2f" % 0.0
             else:
               cmd.select("tmp3", "tmp2 within "+str(distance)+" of tmp1")
               rms[struct][struct2]="%5.2f" %
float(cmd.count_atoms("tmp3"))/nbatomtot
             comptiter+=1
               try:
                 pbar.update(comptiter)
               except AssertionError:
                 print comptiter

#Ecriture de la matrice
line1="%16s" % ""
frms=open(path+outname+'.matrice','w')
for i in sorted(rms):
  line1=line1+" %16s" % i
frms.write(line1+"\n")
for i in sorted(rms):
  line="%16s" % i
  for j in sorted(rms[i]):
    line=line+" %16s" % rms[i][j]
  frms.write(line+"\n")
frms.close()
pbar.finish()
frms=open(path+outname+'.disp','w')
for i in sorted(rms):
  for j in sorted(rms[i]):
    line="%5s %3s %3s" % (rms[i][j],i,j)
    frms.write(line+"\n")
frms.close()
```

ANNEXE 3 : SCRIPT DE REGROUPEMENT DES STRUCTURES PAR FAMILLE

```
#!/usr/bin/env python

#Script de clusterisation par volume commun
#Le fichier fichier_cluster est calcule par calcsimil.py

from random import choice
from operator import itemgetter

#Nb atomes du squelette du domaine Glycosyltransférase : 579
minvol=358 #Nb d'atomes minimum en commun entre les structures du meme
cluster
nbiter=10  #Nombre d'iterations du script

path="/Clusters/"
type_calc="Simil50p"
fichier_cluster=path+"Simil.disp"

#Format du fichier d'entree :
#col1 col2 col3
#RMSD  s1   s2

#Fonction de lecture des RMSD depuis le dictionnaire

def rms_calc(rms_index,s1,s2):
   if s2 in rms_index[s1].keys():
      rms_calc=rms_index[s1][s2]
   else:
      rms_calc=rms_index[s2][s1]
   return rms_calc

rmsindex={} #Dictionnaire des RMSD
struct=[] #Liste des structures
for line in open(fichier_cluster,'r'):
   line=str.split(line)
   if len(line)==3:
      if line[1] not in struct and "3NF" in line[1] or "3c3" in line[1]:
        struct.append(line[1])
      try:
        rmsindex[line[1]][line[2]]=float(line[0])
      except KeyError:
        rmsindex[line[1]]={}
        rmsindex[line[1]][line[2]]=float(line[0])


for iter in range(nbiter):
   name="clusters"+type_calc+"."+str(iter)
   out=open(path+name,'w')

   struct_temp=[]
   struct_temp.extend(struct) #Liste des structures restant a assigner
   fam=[0]*len(struct)

   indice=0
   while struct_temp:
      structure=choice(struct_temp) #Choix d'une structure au hasard
      fam[indice]=[structure]
      struct_temp.remove(structure)
```

```
      struct_temp2=[]
      struct_temp2.extend(struct_temp) #Liste des structures pour le 2e
passage
      while struct_temp2: #Choix de la 2e structure au hasard
        structure2=choice(struct_temp2)
        if structure!=structure2:
         #Comparaison d'une structure de la liste (structure2) avec
         structure1
         rms=rms_calc(rmsindex,structure,structure2)
         if rms>minvol:
           ok=0
           comptTest=0
           rms2=0
           for structurefam in fam[indice]:
             #Comparaison de l'ensemble de la famille de structure1 avec
             structure2
             if structurefam!=structure2 and structure2 in struct_temp:
               comptTest+=1
               rms2+=rms_calc(rmsindex,structure2,structurefam)
           if float(rms2)/comptTest>minvol:
              fam[indice].append(structure2)
              struct_temp.remove(structure2)
         struct_temp2.remove(structure2)
      indice+=1

   #Partie pour faire un tri inverse des familles par nb d'elements
   longueurs={}
   for indice in fam:
      if indice==0:
        fam.remove(indice)
      else:
         longueurs[fam.index(indice)]=len(indice)

   sfam=[]
   f=sorted(longueurs.items(), key=itemgetter(1))
   compt=len(f)-1
   for i in range(len(f)):
     sfam.append(fam[f[compt-i][0]])

   #Ecriture du fichier de sortie
   for indice in sfam:
         out.write("\nFamille %i\n" %(sfam.index(indice)))
         for stru in sorted(indice):
            out.write(stru+"\n")
   out.close()
```

ANNEXE 4 : PROTOCOLE DE PRODUCTION ET DE PURIFICATION POUR LA PROTÉINE PPRV-P459N

Culture

1. Pré-culture à 37°C pendant la nuit dans 25 mL de LB + 100 µg/mL d'ampicilline
2. Culture dans 1 L de M9 à 37°C jusqu'à DO = 0,6
3. Induction par IPTG (80 µg/mL) : 3 h à 37°C
4. Collecte des cellules par centrifugation : 6000 rpm, 20 min, 4°C (JLA 9100)

Lyse

1. Reprendre le culot avec 40 mL de tampon de lyse + 200 µg/mL lysozyme (= 8 mg) + cocktail antiprotéases (+ DNAse).
2. Incubation sur glace : 1 h
3. Sonication : Puissance 70%, durée 5 min, impulsions = 4 sec on et 6 sec off
4. Centrifugation : 10 000 rpm, 30 min, 4°C (70 Ti)

Purification *(utiliser des tampons filtrés et dégazés sur l'Akta !)*

Etape 1 : Dépôt sur résine GST

1. Equilibrer une colonne GSTrap FF (GE) en tampon "GST-A".
2. Injecter le surnageant sur la colonne GSTrap, soit en utilisant la boucle d'injection de 10 mL, soit en utilisant la pompe P950 (sample flow 0,8 mL/min et flow 0 mL/min).
3. Laver avec 5 à 10 volumes de colonne du tampon "GST-A".

Etape 2 : Coupure sur colonne

1. Préparer 1,1 mL de solution de thrombine : par exemple 100 µL de solution de thrombine (1 U/µL) avec 1 mL de tampon "GST-A".
2. Injecter avec une seringue équipée d'un filtre 0,22 µm sur la colonne GSTrap démontée.
3. Laisser incuber la colonne à température ambiante (22 à 25°C) pendant 2-16 h.

Etape 3 : Elution

1. Equilibrer une colonne benzamidine en tampon"GST-A".
2. Monter en série les colonne GSTrap et benzamidine .
3. Eluer en tampon "GST-A" et collecter.
4. Démonter la colonne benzamidine et laver la colonne GSTrap en tampon "GST-B".
5. Démonter la colonne GSTrap et monter la colonne benzamidine.
6. Laver la colonne benzamidine en tampon "benza".

Dialyse et concentration

Utiliser des membranes avec un cut-off de 3,5 kDa (PPRV-P459N fait 6 kDa)

Tampon RMN= tampon "PBS-RMN"

Tampons

Solutions-mères

$MgSO_4$	1 M
$CaCl_2$	1 M
EDTA pH 8	0,5 M
Tris-HCl pH 8	0,5 M
NaP pH 6,8	0,5 M

Tampon de lyse		**Préparation pour 1 L**
Tris-HCl pH 8	50 mM	100 mL Tris-HCl 0,5 M
NaCl	150 mM	8,76 g NaCl (FW 58,44)
EDTA	1 mM	2 mL EDTA pH 8 0,5 M
DTT	2 mM	0,3 g DTT (FW 154,25)
Triton X-100	0,5%	5 mL Triton X-100
$MgSO_4$	10 mM	10 mL $MgSO_4$ 1 M
$CaCl_2$	1 mM	1 mL $CaCl_2$ 1 M
		compléter à 1 L avec de l'eau MQ

Tampon "GST-A"		**Préparation pour 0,5 L**
Tris-HCl	50 mM pH 8	50 mL Tris-HCl 0,5 M
NaCl	150 mM	4,38 g NaCl
$CaCl_2$	2,5 mM	1,25 mL $CaCl_2$ 1 M
		compléter à 0,5 L avec de l'eau MQ

Tampon "GST-B"		**Préparation pour 200 mL**
Tris-HCl	50 mM pH 8	20 mL Tris-HCl 0,5 M
GSH	10 mM	0,62 g (FW 307)
		compléter à 0,2 L avec de l'eau MQ

Tampon "benza"		**Préparation pour 200 mL**
Tris-HCl	50 mM pH 8	20 mL Tris-HCl 0,5 M
NaCl	1,5 M	17,5 g NaCl
		compléter à 0,2 L avec de l'eau MQ

Tampon "PBS-RMN"		**Préparation pour 1 L**
NaP	25 mM pH 6,8	50 mL NaP 0,5 M
NaCl	150 mM	8,76 g NaCl
		compléter à 1 L avec de l'eau MQ

ANNEXE 5 : FICHIER DE CONFIGURATION ROSETTA 3

```
-abinitio
 -relax     # do a relax refinement after each abinitio folding
            # this does full-atom mode
 -fastrelax # using a reduced number of cycles in the relax refinement
            protocol

-in
 -file
  -fasta PPRV.fasta
  -frag3 fragments/aaPPRV_03_05.200_v1_3
  -frag9 fragments/aaPPRV_09_05.200_v1_3
 -path
  -database  /rosetta3_database/

-out
 #-path output    # path where PDB output files will be written to '.'

 -nstruct 1000    # the number of output decoys to produce
 #-pdb            # output PDB files as well as silent.out, default=false
 -prefix bb       # the beginning part of output filenames

 -file
  -silent folding_silent2.out # name the "silent file" output; mult.
                              structures in one file
  -fullatom
      # low-resolution or "centroid" folding mode is the default; you must
      specify fullatom output if you want to match -relax, above

# uncomment these commands to impose either centroid-based or full-atom
constraints
-constraints
 #-cst_file PPRVnoeV2wLD.cst
 #-cst_weight 1.0
 -cst_fa_file PPRVnoeV2wLD.cst
 -cst_fa_weight 1.0


-mute core.util.prof    # don't show timing info
#-no_prof_info_in_silentout         # or profiling info
```

ANNEXE 6 : SCRIPT DE CALCUL DES COURANTS DE CYCLE

```
import numpy as np
import re
import math
import os
import molecule

def sortDict(adict):
    items = adict.items()
    items.sort()
    return [value for key, value in items]


#Script de calcul des courants de cycles


rep="/Librairie_H/" #Repertoire des structures
rep2="/RC/" #Repertoire des fichiers de sortie
log=rep2+"RC_full.log"
log2=rep2+"RCFiltreStruct.rc"


listeStruct=[]
for line in open(rep2+"ListeStructureX.txt",'r'):
  listeStruct.append(str.split(line)[0])

for line in open(rep2+"ListeStructureY.txt",'r'):
  listeStruct.append(str.split(line)[0])



res=open(log,'w')
res2=open(log2,'w')
for data in sorted(os.listdir(rep)):
 if data in listeStruct:
   comptH=0
   res.write(data+'\n')
   res2.write(data+'\n')
   atomes=[]
   cycles={}
   nbarom=0
   prec_resnum=0
   maxdist=8 # distance maximale de l'atome par rapport au cycle
   purines="DA DG A G ADE GUA"
   pyrimidines="DC DT C T CYT THY"
   cycle_purine1=['N9','C8','N7','C5','C4']
   cycle_purine2=['C4','C5','C6','N1','C2','N3']
   cycle_pyrimidine=['N1','C2','N3','C4','C5','C6']

   atomes_calcules="H"
   HDNA=["H1","H1\'","H2","H2\'","H2\'\'","H3","H3\'","H4","H4\'",\
   "H5","H5\'","H5\'\'","H6","H8","H21","H22","H41","H42","H61","H62",\
   "H71","H72","H73","HO3\'","HO5\'"]
   Carbons=["CA","CB","CO"]
   facteurBH=5.13
   facteurBHN=7.06
   facteurBCA=1.5
   facteurBCBCO=1.0
```

```
    #facteurB prendra une de ces valeurs selon le type d'atome
    facteurB=0

    for line in open(rep+data,'r').readlines():
      if line.startswith('ATOM'):
        atome=molecule.AtomFromPdbLine(line)
        if atome.type.startswith('H') and not atome.type in HDNA:
          comptH+=1
        atomes.append(atome)
        if atome.res_type in purines:
        #Traitement des bases A et G (deux cycles)
        resnum=prec_resnum
        resnum=atome.res_num
        if prec_resnum != resnum:
            nbarom+=1
            cycles[atome.res_type+str(resnum)+"S"]=[0]*len(cycle_purine1)
            cycles[atome.res_type+str(resnum)+"B"]=[0]*len(cycle_purine2)
            prec_resnum=resnum
        if atome.type in cycle_purine1:

cycles[atome.res_type+str(resnum)+"S"][cycle_purine1.index(atome.type)]=atom
e
        if atome.type in cycle_purine2:

cycles[atome.res_type+str(resnum)+"B"][cycle_purine2.index(atome.type)]=atom
e

        if atome.res_type in pyrimidines:
        #Traitement des bases C et T (un cycle)
        resnum=prec_resnum
        resnum=atome.res_num
        if prec_resnum != resnum:
            nbarom+=1
            cycles[atome.res_type+str(resnum)]=[0]*len(cycle_pyrimidine)
            prec_resnum=resnum
        if atome.type in cycle_pyrimidine:

cycles[atome.res_type+str(resnum)][cycle_pyrimidine.index(atome.type)]=atome


    Gp={}

    #Calcul du barycentre : Gp[numero du residu]=[x,y,z]
    toRem=[]
    for cycle in cycles:
      sumx=0
      sumy=0
      sumz=0
      if 0 not in cycles[cycle]:
        for atomecycle in cycles[cycle]:
        sumx+=atomecycle.pos.x
        sumy+=atomecycle.pos.y
        sumz+=atomecycle.pos.z
        if atomecycle.type in cycle_purine1:
        Gp[cycle]=[sumx/5,sumy/5,sumz/5]
        else:
        Gp[cycle]=[sumx/6,sumy/6,sumz/6]
      else:
        toRem.append(cycle)

    for elem in toRem:
```

```
      del cycles[elem]

   #Choix des atomes a moins de X=amp A du barycentre situes dans un autre
residu

   closeList={}

   for cycle in Gp:
     closeList[cycle]=[]
     if Gp[cycle]!=[0,0,0]:
       for atome in atomes:
         if (atome.type.startswith('H') and not atome.type in HDNA) or
atome.type in Carbons:
           dist=math.sqrt((Gp[cycle][0]-atome.pos.x)**2+(Gp[cycle][1]-
atome.pos.y)**2+(Gp[cycle][2]-atome.pos.z)**2)
           if dist < float(maxdist):
             closeList[cycle].append(atome)

   #Nettoyage de la liste
   toRemove=[]
   for elem in closeList:
     if len(closeList[elem])==0:
       toRemove.append(elem)

   for elem in toRemove:
     del closeList[elem]


   #Calcul de la contribution du courant de cycle sur ces atomes
   norm=lambda x: np.sqrt(np.square(x).sum())



   for j in cycles:
   #Pour chaque cycle
     atomesCycle=[]
     for i in cycles[j]:
     #Ajout des atomes du cycle
       atomesCycle.append([i.pos.x,i.pos.y,i.pos.z])

     try:
       for k in closeList[j]:
      #Pour chaque atome proche, faire le calcul
       G=0
       Surface=[]
       P=np.array([k.pos.x,k.pos.y,k.pos.z])
       for i in range(len(atomesCycle)):
         A=np.array(atomesCycle[i])
         try:
           B=np.array(atomesCycle[i+1])

         except IndexError:
           B=np.array(atomesCycle[0])

         try:
           C=np.array(atomesCycle[i+2])

         except IndexError:
           if len(atomesCycle)==6:
             C=np.array(atomesCycle[i-4])
           if len(atomesCycle)==5:
```

```
            C=np.array(atomesCycle[i-3])


          BA=np.array([A[0]-B[0],A[1]-B[1],A[2]-B[2]])
          BC=np.array([C[0]-B[0],C[1]-B[1],C[2]-B[2]])

          N=np.array(np.cross(BA,BC))
          AP=np.array([P[0]-A[0],P[1]-A[1],P[2]-A[2]])
          BP=np.array([P[0]-B[0],P[1]-B[1],P[2]-B[2]])

          #Le vecteur BPproj verifie l'equation BPproj.N=0 car Pproj et B
sont dans le plan du cycle, d'ou
          #zPproj=(-((P[0]-B[0])*N[0]+(P[1]-B[1])*N[1])/N[2])+B[2]
          #Pproj=np.array([P[0],P[1],zPproj])

          #BPproj=np.array([Pproj[0]-B[0],Pproj[1]-B[1],Pproj[2]-B[2]])
          #APproj=np.array([Pproj[0]-A[0],Pproj[1]-A[1],Pproj[2]-A[2]])

          Axe=np.cross(BA,N)


Surface.append((1/(norm(BP)**3)+1/(norm(AP)**3))*(0.5*np.dot(BP,Axe/norm(BA)
)))

        #Determination du facteur I
          #ade
        if re.search('A\d+S',j) or re.search('DA\d+S',j):
         #facteur petit cycle
          facteuri=1.14

        if re.search('A\d+B',j) or re.search('DA\d+B',j):
         #facteur grand cycle
          facteuri=0.90

        #cyt
        if re.search('C\d+',j) or re.search('DC\d+',j):
          facteuri=0.37

        #thy
        if re.search('T\d+',j) or re.search('DT\d+',j):
          facteuri=0.35

        #gua
        if re.search('G\d+S',j) or re.search('DG\d+S',j):
         #facteur petit cycle
          facteuri=1.00

        if re.search('G\d+B',j) or re.search('DG\d+B',j):
         #facteur grand cycle
          facteuri=0.51

        #Determination du facteur B
        if k.type.startswith("H") and not k.type == "HN":
          facteurB=facteurBH
        if k.type == "HN":
          facteurB=facteurBHN

        S=sum(Surface)*facteuri*facteurB
        G+=sum(Surface)*facteuri*facteurB
```

```
        #Ajout du courant de cycle sur l'atome
        if k.rc != 0:
         k.rc += S
        else:
         k.rc = S

        #Ecriture pour le fichier de log complet
        res.write("%5s %3s %3i %7.4f %-7s\n" %
(k.type,k.res_type,k.res_num,float(S),j))

     except KeyError:pass

   #Ecriture pour le fichier de log normal
   for k in atomes:
     if k.rc != 0:
       res2.write("%5s %3s %3i %7.4f\n" %
(k.type,k.res_type,k.res_num,k.rc))

   print data+" "+str(comptH)

res.close()
res2.close()
```