UFR de Biologie et Sciences de la Nature
Université Paris Diderot - Paris 7
Case 7007
Tour 54 - 4ème étage - couloir 54/53
2, place Jussieu
75251 PARIS CEDEX 05

EBGM - INSERM U726
Université Paris Diderot - Paris 7
Case 7113
Tour 53 - 1er étage - couloir 53/54
2, place Jussieu
75251 PARIS CEDEX 05

Thèse de Doctorat de l'Université Paris 7 – Denis Diderot

Ecole Doctorale B2M – Biochimie et Biologie Moléculaire

Spécialité : *ANALYSE DE GENOMES ET MODELISATION MOLECULAIRE*

Présentée par

Julien Maupetit

Pour obtenir le titre de Docteur de l'Université Paris 7

Génération *ab initio* de modèles protéiques à partir de représentations discrètes des protéines et de critères d'énergie simplifiés.

Soutenue le 6 novembre 2007, devant le jury composé de :

Pr Catherine Etchebest, EBGM - Université Paris Diderot - Paris 7	Présidente
Dr Jacques Chomilier, IMPMC - CNRS	Rapporteur
Dr Gilles Labesse, CBS - CNRS	Rapporteur
Pr Philippe Derreumaux, IBPC - Université Paris Diderot - Paris 7	Examinateur
Dr Konrad Hinsen, CBM - CNRS	Examinateur
Dr Anne Poupon, IBBMC - CNRS	Examinatrice
Dr Joël Pothier, ABI - Université Paris 6	Examinateur
Dr Pierre Tufféry, EBGM - INSERM	Directeur

A ma doudoune.

Remerciements

Et voilà, trois ans ont passé, très vite, trop vite. Durant ces trois années de thèse, j'ai beaucoup appris à la fois sur la science, mais aussi sur les autres et sur moi même.

J'ai tout d'abord une pensée émue pour Serge Hazout qui a d'abord été mon Professeur, puis m'a accepté dans son laboratoire pour y faire ma thèse. Je ne l'oublierai pas.

Je tiens à remercier Pierre Tufféry pour ce qu'il a été durant des quatre dernières années. Merci de m'avoir accepté comme étudiant, d'avoir été présent pour moi, et de m'avoir formé à la recherche.

Merci à Jacques Chomilier et Gilles Labesse pour avoir accepté de juger mon travail. Leurs conseils ont notamment été d'une grande aide pour améliorer la fluidité de ce manuscrit. Merci à Cathy, Konrad Hinsen, Anne Poupon et Joël Potier d'avoir bien voulu examiner ce travail aussi. Merci à Philippe Derreumaux d'avoir travaillé avec Pierre et moi, et de m'avoir accepté dans son laboratoire pour cette année à venir.

Merci à l'equipe "HMM-SA" : Anne-Claude, Fred, Leslie et Juliette parce que j'ai vraiment apprécié de travailler avec eux. Merci aux filles d'à côté, pour leur bonne humeur constante ... enfin quand elles ne jouent pas les autistes ! Merci P@t pour tous tes tips de geek de la modélisation moléculaire et tous ces guitaristes que tu m'as fait découvrir, qui raisonnent encore à mes oreilles. J'ai hâte que l'on se fasse un concert de Tommy ensemble ! Merci Alex d'avoir toujours le mot pour rire : non, je ne m'appelle pas Emilie ! Sinon Aurélie et Eric, arrêtez de parler on n'entend que vous. Merci à Joëlle pour les nouvelles du matin, et ses astuces de super admin : vive fedora ! Bref merci à tous les membres de l'EBGM que je n'ai pas cité mais qui se reconnaitrons.

Ces trois ans ont aussi été les grandes années du bureau 5 pendant lesquelles, de vraies amitiés sont nées. Gaëlle, Yann et Anita merci de m'avoir supporté !

Merci à ma (belle) famille d'avoir toujours été présente, et à l'écoute pendant les périodes difficiles. Et merci aussi aux théatreux, ma seconde famille !

Et enfin, merci à ma poulette d'être toujours là pour moi. Maintenant, c'est ton tour ...

Bok !

Table des matières

Liste des tableaux

Table des figures

Liste des symboles

Q_n	Pourcentage de bonne prédiction à n états
Å	Angström
e.g.	*exempli gratia*
i.e.	*id est*
ART	*Activation Relaxation Technique*
BB	*Building Blocks*
BIC	*Bayesian Information Criterion*
BLAST	*Basic Local Alignment search Tool*
BS	Banque de Structures
Cα	Carbone alpha
CABS	*Cα-β and Side group*
CASP	*community wide experiment on the Critical Assessment of techniques for protein Structure Prediction*
CL	Chaîne Latérale
CP	Chaîne Principale
CP'	$CP - C\alpha$
CPU	*Central Process Unit*
cRMSd	*alpha carbon Root Mean Square deviation*
CSA	*Conformational Space Annealing*
DM	Dynamique moléculaire
DOPE	*Discrete Optimized Protein Energy Model*
EBGM	Equipe de Bioinformatique Génomique et Moléculaire
FAA	Filtre sur les Acides Aminés

FM	*Free Modeling*
FS	Fonction de Score
FU	*Folding Units*
HA-TBM	*High Accuracy-Template Based Modeling*
HCPM	*Hierarchical Clustering of Protein Models*
HMM	*Hidden Markov Model*
HMM-SA	*Hidden Markov Model Structural Alphabet*
JA	Jeux d'Apprentissage
JV	Jeu de Validation
kCal	kilocalorie
LS	*Learning Set*
MC	Monte-Carlo
Neq	Nombre équivalent (de lettres)
OPEP	*Optimized Potential for Efficient protein structure Prediction*
PB	*Protein Blocks*
PDB	*Protein Data Bank*
PSI-BLAST	*Position Specific Iterative BLAST*
PSN	Proche de la Structure Native
PSSM	*Position Specific Score Matrix*
RAM	*Random Access Memory*
REMD	*Replica Exchange Molecular Dynamics*
RMN	Résonance Magnétique Nucléaire
RMSd	*Root Mean Square Deviation*
RMSF	*Root Mean Square Fluctuation*
RX	Rayon X
SABBAC	*Structural Alphabet based protein Backbone Builder from Alpha Carbon trace*
SAFrAN	*Structural Alphabet candidate Fragments from AmiNo-acid sequence*
SFAA	Sans Filtre sur les Acides Aminés

SVM	*Support Vector Machines*
TBM	*Template Based Modeling*
TEF	*Tight End Fragments*

Première partie

Introduction

Les protéines sont des macromolécules biologiques essentielles à une majorité de processus biologiques. Elles jouent à la fois un rôle structural nécessaire à la survie des cellules (le cytosquelette en est la charpente), mais elles sont aussi impliquées dans des cascades de réactions chimiques complexes, soit en rentrant en interaction avec d'autres macromolécules biologiques, soit en jouant le rôle de catalyseur biologique de ces réactions chimiques. L'expression de ces macromolécules et la régulation de leur fonction sont des processus très finement régulés garantissant, dans des conditions normales, l'harmonie cellulaire.

Un concept important lié aux protéines est la relation étroite entre leur structure et leur fonction. Dès 1894, Emil Fischer avait émis l'hypothèse que la spécificité des enzymes était due à leur conformation spatiale par un modèle de clé-serrure (Fischer, 1894). En 1936, Mirsky et Pauling avaient suggéré que la conformation des protéines, c'est à dire l'organisation spatiale de leurs atomes, est responsable de leur activité (Mirsky and Pauling, 1936). Depuis, la détermination de nombreuses structures par des techniques expérimentales comme la radiocrystallographie des rayons X (RX), ou la résonance magnétique nucléaire (RMN), ont largement permis de confirmer cette hypothèse. Il est par exemple frappant de voir la proximité spatiale des acides aminés impliqués dans les sites catalytiques enzymatiques, ou encore, l'impact structural de certaines mutations sur la fonction des protéines. Cependant, un autre aspect important de la fonction des protéines est lié aux interactions fonctionnelles spécifiques de partenaires biologiques (protéines, acides nucleiques, lipides, ...), impliquées dans les différents chemins du métabolisme cellulaire, tissulaire, et plus généralement du vivant. Là encore, la compréhension de ces interactions, et leur perturbation à des fins d'ingénierie ou des fins médicales est grandement facilitée par la connaissance de la structure tridimensionnelle des protéines. Il est désormais possible de faire du criblage *in silico* de petits composés sur des modèles structuraux prédits par homologie (si l'identité de séquence entre le modèle et le patron moléculaire est supérieure ou égale à 50 %). En aval de la détermination des structures protéiques, des enjeux importants se trouvent donc également dans la détermination *in silico* d'effecteurs des protéines, à visée thérapeutique ou fondamentale.

Dans l'ère post-génomique où les projets de séquencages de génomes complets nourissent exponentiellement les bases de données de séquence (voir http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=Genome, plus de 420 génomes complets d'eucaryotes séquencés, plus de 1300 génomes bacteriens séquencés ou en cours de séquencage courant 2007) la détermination de la structure des protéines reste une préoccupation majeure. Malgré les progrès des techniques expérimentales, et l'arrivée annoncée de méthodes de résolution haut débit (Rupp, 2005), l'écart entre le nombre de séquences et le nombre de structures résolues demeure très important (environ 36000 structures déposées à la *Protein Data Bank* (PDB) (Berman et al., 2000b) à l'été 2007. La motivation pour développer des mé-

thodes *in silico* de prédiction de la structure des protéines demeure donc très forte. Un important effort au niveau mondial, que l'on peut estimer à plusieurs centaines d'équipes (voir les équipes participants aux différentes éditions de CASP (*community wide experiment on the Critical Assessment of techniques for protein Structure Prediction*) (Moult et al., 2007), a permis l'avènement de techniques fiables de modélisation par homologie. Dans ce cas, on génère un modèle protéique à partir de la connaissance de la structure d'une protéine de séquence similaire, sur la base de la constatation que des protéines de séquence similaire ont généralement des structures voisines. Un des premiers exemples de telle modélisation est sans doute, en 1968, la modélisation de la structure de l'élastase de la trypsine à partir de la chymotrypsine, à l'époque la seule protéase à sérine de structure connue (Hartley, 1970). Depuis, ce type d'approche a énormément progressé, et aujourd'hui des modèles sont générées automatiquement pour être ajoutés dans des banques (Pieper et al., 2006). Cependant, un certain nombre de protéines tombe encore hors du champ d'application de ces techniques. On estime entre 20 et 30 % le nombre de protéines dites "orphelines", qui ne présentent aucune homologie de séquence détectable avec aucune protéine de structure connue (Siew et al., 2004). Les efforts actuels se focalisent donc sur le développement de méthodes alternatives de modélisation dites *ab initio* ou *de novo*.

Mon travail de thèse se situe dans ce contexte. Il a porté sur différents aspects du développement d'une approche de modélisation de la structure des protéines basée sur le concept d'alphabet structural. Dans la suite de ce manuscrit, après un rappel sur l'état de l'art des méthodes de modélisation de la structure des protéines et des outils déjà en place au début de ma thèse, j'aborderai les différentes contributions que j'ai apporté à l'approche HMM-SA au cours de ma thèse.

Chapitre 1

De la prédiction de la structure des protéines

Malgré des progrès importants, la prédiction de la structure des protéines à partir de leur seule séquence en acides aminés reste encore un défi en bioinformatique structurale. A partir de la structure primaire d'une protéine, l'enjeu est de pouvoir déterminer, sans notion d'ordre, (i) les structures secondaires, *i.e.* les régions en hélices α ou en brins β, (ii) toutes les paires de brins β qui forment des feuillets (la topologie du feuillet β), (iii) les ponts disulfure si des cystéines sont présentes, (iv) les boucles qui connectent les structures secondaires, et enfin, (v) la structure tertiaire (repliement tridimensionnel) de la protéine.

Le problème de la prédiction de la structure des protéines a attiré un grand nombre de chercheurs venant de disciplines différentes, proposant des explications diverses à la question du repliement protéique. Selon l'hypothèse thermodynamique d'Anfinsen (1973), l'assemblage des protéines dans leur structure native n'est pas un processus biologique, mais purement physique, dépendant uniquement de la spécificité de sa séquence en acides aminés, et du solvant environnant. Cette hypothèse est aujourd'hui à nuancer, car nous savons d'une part qu'il existe un certain nombre de protéines chaperonnes aidant au repliement des protéines (Ellis, 2000, 2006) et, d'autre part, qu'il existe un certain nombre de protéines dites globalement désordonnées (Dyson and Wright, 2005), *i.e.* n'ayant pas de structures secondaires ni tertiaire bien déterminées, mais étant fonctionelles.

Cinq modèles principaux ont été proposés pour le repliement des protéines (Haspel et al., 2003a) :

i. le modèle de nucléation-condensation, dans lequel une première étape de nucléation est suivie par une propagation rapide de la structure (Zimm and Bragg, 1959; Wetlaufer, 1973).

ii. le modèle de diffusion-collision, ici la nucléation intervient en différent points de la chaîne polypeptidique, ces noyaux diffusent puis s'unissent pour former des micro-structures natives (Karplus and Weaver, 1994).

iii. le repliement séquentiel, pour lequel plusieurs segments de structure sont formés et assemblés à différents niveaux en suivant un chemin unique de repliement (Baldwin, 1975).

iv. le modèle d'effondrement hydrophobe, ce dernier implique que le repliement est directement guidé par les interactions hydrophobes formant le coeur protéique. Les structures secondaires sont formées dans un deuxième temps (Levitt and Warshel, 1975).

v. le modèle de repliement hiérarchique, pour lequel, le phénomène de nucléation est suivi par la formation des super-structures secondaires puis des domaines (Schulz, 1977; Lesk and Rose, 1981; Baldwin and Rose, 1999).

Chacun de ces modèles est corroboré par des analyses expérimentales (Tsai et al., 2002). Le repliement des protéines semble donc être un *continuum* entre ces différents modèles (Fersht and Daggett, 2002). Cependant, le modèle de repliement hiérarchique serait le plus communément admis (Haspel et al., 2003a; Floudas, 2007). Il implique que des éléments des structures secondaires se forment rapidement, suivis par un réarrangement tridimensionnel plus lent pour former la structure tertiaire.

Une vision classique dans le domaine est de découper les méthodes de prédiction de la structure des protéines en deux catégories principales : les méthodes de modélisation comparatives et les méthodes dites *ab initio*. Les méthodes de modélisation comparative nécessitent une ou des structures matrices identifiées par homologie entre la séquence de la structure à prédire, et un ensemble de séquences de structures expérimentalement résolues. Cette catégorie regroupe la modélisation par homologie, et les méthodes dites d'enfilage ou de reconnaissance de repliement. A l'opposé, les méthodes *ab initio* permettent de prédire une structure protéique pour des pourcentages d'identité de séquence très faibles. Ce sont donc potentiellement des méthodes permettant de générer de nouveaux repliements. Au sein des méthodes *ab initio*, nous pouvons faire la distinction entre les méthodes dites *ab initio* pures, basées uniquement sur des principes physiques, et les méthodes dites *de novo* qui, quant à elles, utilisent une batterie d'informations issues de bases de données.

Devant l'enjeu de ces méthodes de prédiction, a été mise en place une procédure d'évaluation objective de leur performance. Cette évaluation a lieu tous les deux ans lors de l'expérience CASP (Moult, 1999; Moult et al., 1997, 2001, 2003, 2005, 2007). Lors de cette expérience, les organisateurs gardent secrètes des structures protéiques récemment résolues expérimentalement, et ne fournissent aux équipes participantes que la séquence

en acides aminés de ces protéines. Chaque équipe doit donc appliquer sa méthode pour proposer des modèles. A la fin de l'expérience, les structures sont rendues publiques, et les différentes méthodes évaluées pour la pertinence de leurs prédictions. Cette expérience est ainsi un très bon outil d'analyse des points forts et des faiblesses d'une méthode. Dans la suite de notre manuscrit, nous détaillerons les résultats que nous avons obtenu lors de notre participation à la septième édition de cette compétition.

1.1 Les méthodes de modélisation comparative

1.1.1 La modélisation par homologie

Historiquement, les méthodes de modélisation par homologie sont les plus anciennes, et la technique n'a que peu évolué depuis ces vingts dernières années (Tramontano and Morea, 2003). Cette méthode se base sur le principe que des séquences, étant reliées du point de vue évolutif, possèdent un repliement tridimensionnel similaire (Holm and Sander, 1996). Ainsi, une similarité de séquence suggère une similarité de repliement. Partant de ce principe, la modélisation par homologie consiste en quatre étapes principales (Sanchez and Sali, 1997; Fiser and Sali, 2003) : (i) identification, par leur séquence, de structures connues pouvant servir de matrice, (ii) alignement de la séquence à modéliser avec la structure matrice, (iii) modélisation des régions conservées, en utilisant la matrice, et des boucles et chaînes latérales absentes de la matrice, et enfin (iv) raffinement et évaluation de la qualité du modèle produit.

La performance des méthodes de modélisation par homologie dépend directement du pourcentage d'identité de séquence que partagent la cible et la matrice. Si plus de 50% de ces deux séquences sont identiques, les prédictions sont de très haute qualité, et, il a été montré que ces modèles sont aussi pertinents que des modèles RX à faible résolution (Kopp and Schwede, 2004).
Entre 30 et 50% d'identité de séquence, plus de 80% des carbones alpha sont attendus comme étant à moins de 3,5 Å de leur position réelle (Kopp and Schwede, 2004), alors qu'à moins de 30% d'identité de séquence, il y a de fortes chances que le modèle contienne des erreurs importantes (Vitkup et al., 2001; Kopp and Schwede, 2004).

Une étude récente évaluant la pertinence des différentes méthodes de modélisation par homologie peut être trouvée dans Dalton and Jackson (2007).

Détection d'homologues et méthodes d'alignements

Classiquement, pour des pourcentages d'identité de séquence supérieurs à 30%, l'identification d'homologues structuraux est réalisée en comparant la séquence de la cible avec

l'ensemble des séquences des structures de la PDB (*Protein Data Bank*) (Berman et al., 2000b). L'alignement des séquences est réalisé par des algorithmes de type programmation dynamique (Needleman and Wunsch, 1970) et ses dérivés (Smith et al., 1981). Le programme le plus couramment utilisé dans ce cas est BLAST (*Basic Local Alignment search Tool*). Pour des pourcentages d'identité de séquence plus faibles (inférieurs à 30%), des méthodes alternatives ont du être développées. Ce sont les méthodes basées sur des profils, comme par exemple, les méthodes de recherche position-spécifiques tel que PSI-BLAST (*Position Specific Iterative BLAST*) (Altschul et al., 1997) ou des chaînes de Markov cachées (HMM, *Hidden Markov Models*) (Krogh et al., 1994). En améliorant la qualité des alignements produits, ces méthodes ont rendu possible la détection d'homologues plus distants.

Les méthodes basées sur des profils de séquences (comme PSI-BLAST) commencent par réaliser une recherche d'alignements deux à deux avec une base de données. Les alignements les plus significatifs sont concervés pour construire une matrice de score spécifique de chaque position (*PSSM : Position Specific Score Matrix*). Cette matrice remplace alors la séquence de la cible pour les prochaines itérations ; le processus est itéré jusqu'a ce qu'aucun nouvel alignement significatif ne puisse être trouvé.

Les méthodes de comparaison séquence-profil peuvent être améliorées en ajoutant des informations relatives à l'évolution de la séquence protéique pour, à la fois la séquence de la cible et les séquences de la base de données. Différentes méthodes d'alignement de profils ont été proposées en ce sens (Rychlewski et al., 2000; Yona and Levitt, 2002; Sadreyev and Grishin, 2003; Edgar and Sjolander, 2003; Pei et al., 2003). Ohlson et al. (2004) ont démontré que les méthodes profil-profil sont 30% plus performantes que les méthodes séquence-profil, à la fois pour leur capacité à reconnaître des protéines de la même super-famille, mais aussi pour la qualité des alignements produits.

Depuis CASP5 (Moult et al., 2003), il est apparu que les méta-méthodes, tirant parti de différentes approches, permettent de détecter des homologues structuraux plus lointains (Kinch et al., 2003). Deux serveurs se sont ainsi illustrés lors de CASP5 dans la catégorie reconnaissance de repliement. Il s'agit de @TOME (Douguet and Labesse, 2001) et 3D-Jury (Ginalski et al., 2003a). @TOME combine les résultats des programmes PDB Blast, 3D-PSSM (Kelley et al., 2000), mGenTHREADER (Jones, 1999a), FUGUE (Shi et al., 2001), SAM-T99 (Karplus et al., 1999), et JPRED2 (Cuff et al., 1998), tandis que 3D-Jury combine les 8 méthodes ORFeus (Ginalski et al., 2003b), SAM-T02 (Karplus et al., 2003), FFAS03 (Rychlewski et al., 2000), mGenTHREADER (Jones, 1999a), INBGU (Fischer, 2000), RAPTOR (Xu et al., 2003) FUGUE-2 (Shi et al., 2001) et 3D-PSSM (Kelley et al., 2000) pour détecter des candidats structuraux qui sont ensuite réordonnés selon un score de similarité entre les modèles basé sur l'outil MaxSub (Siew et al., 2000).

Génération de modèles

Etant donné un alignement entre une séquence cible et une matrice, trois types de méthodes peuvent être utilisées pour générer un modèle, dépendant de la manière dont l'information de structure est transférée à la séquence. Ces méthodes sont : l'assemblage en corps rigides, la correspondance de segments, et la satisfaction de contraintes spatiales.

Les premiers programmes de modélisation étaient basés sur des méthodes d'assemblage rigide, dans lesquelles un modèle est assemblé à partir d'un nombre restreint de corps rigides obtenus à partir du coeur des régions alignées (Blundell et al., 1987; Greer, 1990). L'assemblage consiste donc en un ajustement des corps rigides sur la matrice, puis de reconstruire les parties non conservées (*i.e.* les boucles et chaînes latérales). Les programmes les plus connus dans cette catégorie sont SWISS-MODEL (Schwede et al., 2003), NEST (Petrey et al., 2003), 3D-JIGSAW (Bates et al., 2001) et Builder (Koehl and Delarue, 1994, 1995). La principale différence entre ces méthodes réside dans la technique employée pour reconstruire les boucles et chaînes latérales. NEST par exemple utilise une approche séquentielle, appliquant un évenement évolutif à la fois, tandis que 3D-JIGSAW et Builder utilisent des méthodes de minimisation de champ moyen (Koehl and Delarue, 1996).

L'approche par correspondance de fragments utilise un sous-ensemble de fragments protéiques dérivés de l'alignement pour rechercher des fragments compatibles dans une base de données représentative de structures résolues (Jones and Thirup, 1986; Claessens et al., 1989). La base de données de recherche contient de courts fragments protéiques sélectionnés selon des critères énergétiques et/ou géométriques. SegMod/ENCAD (Levitt, 1992) dévelopée par Michael Levitt appartient à cette catégorie.

La dernière méthode utilisée en modélisation par homologie utilise des contraintes spatiales pour générer des modèles, contraintes dérivées de l'alignement de séquences toujours. Le modèle est ainsi obtenu en minimisant les violations de ces contraintes spatiales. Le programme le plus performant, et donc le plus couramment utilisé, dans cette catégorie est MODELLER (Sali and Blundell, 1993).

Il semble difficile de dire quelle méthode semble la plus pertinente étant donné que les trois programmes NEST, SegMod/ENCAD, et MODELLER, appartenant chacun à une catégorie distincte, sont aussi performants (Wallner and Elofsson, 2005).

1.1.2 Les méthodes dites “d’enfilage”

Les méthodes de reconnaissance de repliement reposent sur le principe que le nombre de repliements différents que les protéines peuvent adopter est bien moins conséquent que la vaste diversité des séquences générées par les projets génomes. Il a été démontré que la PDB contenait d’ores et déjà l’ensemble des repliements différents que peuvent adopter les protéines de taille allant jusqu’à 100 résidus (Kihara and Skolnick, 2003). Bien que les structures soient mieux conservées que les séquences, il semble nécessaire que la position de certains résidus spécifiques soit conservée durant le processus d’évolution pour garantir la stabilité et la fonction d’une protéine. La sensibilité des méthodes basées sur des profils a ainsi été augmentée en exploitant cette propriété. Pour ce faire, sont inclues dans les profils des informations relatives aux structures des protéines, telles que des alignements multiples de structures, l’environnement de certains résidus, la prédiction des structures secondaires ou l’accessibilité au solvant (Tang et al., 2003; Przybylski and Rost, 2004).

Les méthodes d’enfilage évaluent la pertinence d’enfiler une séquence protéique dans une structure connue issue d’une librairie de repliements. Xu and Xu (2000) ont développé un algorithme d’enfilage qui considère les paires de contacts entre les hélices α et les brins β, et permettant des *gaps* dans l’alignement au niveau des boucles. La méthode permet d’incorporer un certain nombre de contraintes à propos de la protéine cible, telle que des ponts disulfure ou des contraintes de distance. Dans une autre approche, le problème de la reconnaissance de repliement est considéré comme un problème d’optimisation globale d’une fonction d’énergie (Xu et al., 2003), résolu par programmation linéaire.

Il a été montré lors de l’expérience CASP6 que les méthodes de reconnaissance de repliements avaient fait des progrès notables par rapport aux expériences précédentes (Wang et al., 2005).

1.2 Les méthodes dites *ab initio*

1.2.1 Les méthodes *ab initio* pures

Les méthodes *ab initio* pures n’utilisent pas d’informations directement issues de bases de données. Elles tentent d’identifier, pour une protéine dans son environnement, la structure ayant l’énergie libre la plus basse, et ce en utilisant uniquement la séquence en acides aminés de cette dernière et les lois de la physique. Cette classe de prédiction de la structure des protéines peut *a priori* être utilisée pour n’importe quelle séquence protéique avec des potentiels ayant un sens physique et une représentation atomique des modèles. C’est de loin la catégorie la plus complexe, mais aussi la plus intéressante tant elle peut nous apprendre sur le repliement des protéines.

Rose et al. ont introduit une approche hiérarchique pour prédire la structure des protéines (LINUS) qui met l'accent sur le rôle des interactions stériques et de l'entropie conformationnelle (Srinivasan and Rose, 1995, 2002).
Scheraga et al. ont aussi introduit une approche hiérarchique en utilisant un champ de force simplifié pour les calculs initiaux, suivit d'une étape de raffinement avec un champ de force tous atomes (Lee et al., 2001; Liwo et al., 1997a,b, 2001, 2002; Pillardy et al., 2001). Ce champ de force gros grain UNRES (*UNited RESidue*), réduisant la représentation des acides aminés à seulement deux sites d'interactions, permet à l'algorithme de *Conformational Space Annealing* (CSA) d'identifier des structures de plus basse énergie (Lee et al., 1997, 1998; Lee and Scheraga, 1999; Lee et al., 2000). Les travaux récents ont porté sur une amélioration de la gestion des brins β par l'algorithme (Czaplewski et al., 2004a), une analyse détaillée du rôle des ponts disulfure dans la structure (Czaplewski et al., 2004b), et l'introduction d'un algorithme de Monte-Carlo basé sur l'échange de répliques avec la minimisation du champ de force UNRES (Nanias et al., 2005).

Floudas et al. ont développé une méthode originale de prédiction *ab initio* : ASTRO-FOLD (Klepeis and Floudas, 2003b). Cette approche suit aussi un modèle hiérarchique : dans un premier temps, les segments en hélices sont prédits (Klepeis and Floudas, 2002), puis les régions en brins β et la topologie du feuillet correspondant sont envisagées en maximisant le nombre d'interactions hydrophobes (Klepeis and Floudas, 2003a). Un ensemble de conformères est ensuite prédit pour chaques boucles *via* des calculs d'énergie libre (Klepeis and Floudas, 2005) couplés à un échantillonnage intensif et une procédure de classification (Monnigmann and Floudas, 2005). L'ensemble de ces prédictions isolées permettent de définir des contraintes utilisées pour la prédiction de la structure tertiaire par l'intermédiaire d'une nouvelle classe hybride d'optimisation globale (Klepeis et al., 2003a,b) similaire aux protocoles de raffinement des structures RMN (Klepeis and Floudas, 1999).

Dill et al. ont proposé un autre mécanisme potentiel pour le repliement protéique : la fermeture éclair hydrophobe (ou *Zipping and Assembly*) (Dill et al., 1993). Dans ce cas, la formation de structures secondaires s'effectue en même temps que l'effondrement hydrophobe : les structures secondaires commencent à se former localement à différents endroits indépendants de la chaîne, puis ces noyaux s'assemblent pour former une structure complète (Dill et al., 2007). Ce mode de repliement permettrait d'apporter une explication au paradoxe de Levinthal (1968), *i.e.* pourquoi le phénomène de repliement est-il aussi rapide ? Ainsi, Dill et al. ont appliqué ce mécanisme de recherche dans leur méthode de prédiction *ab initio*. Ils partent d'une structure totalement dépliée, sur laquelle s'applique un algorithme basé sur des simulations de dynamique moléculaire avec échange de ré-

pliques, guidé par le champ de force AMBER96 (Kollman et al., 1997) et un modèle de solvant implicite de type born généralisé (GBSA) (Tsui and Case, 2000). Les structures à prédire sont d'abord découpées en fragments de 8 à 12 résidus, puis simulées indépendamment jusqu'à ce que se forme un nombre suffisant de contacts hydrophobes. Leur algorithme fait ensuite croître le fragment en appliquant des contraintes sur les contacts déjà formés. Ces ensembles de simulations sont répétées sur plusieurs régions de la protéine à modéliser, jusqu'à ce qu'aucun contact supplémentaire ne puisse se former. Ils ont ensuite recours à une méthode d'assemblage de fragments pour générer une structure complète. Cette méthode semble donner de bons résultats pour des protéines de petite taille : une étude récente a montré qu'ils ont pu générer des modèles éloignés en moyenne de 2,2 Å de la structure native pour huit des neuf protéines de leur jeu de test ayant des tailles comprises entre 25 et 73 acides aminés (Ozkan et al., 2007).

Il s'avère qu'à l'heure actuelle, bien que ces méthodes donnent des résultats encourageants pour des protéines de petite taille, elles ne sont pas les plus performantes pour générer des modèles protéiques face aux méthodes *de novo*, car trop limitées par les ressources de calcul nécessaires.

1.2.2 Les méthodes dites *de novo*

Les méthodes dites *de novo* sont actuellement les plus performantes pour fournir des modèles à moyenne voir haute résolution (Floudas, 2007). Elles peuvent être découpées en deux catégories : (i) les méthodes d'assemblage de fragments ; (ii) les méthodes hybrides qui mêlent à la fois de l'assemblage de fragments et des simulations de repliement sur réseau.

Les méthodes d'assemblage de fragments

Les méthodes d'assemblage de fragments s'appuient sur le principe fondamental suivant : les interactions locales séquence-dépendantes conduisent la chaîne protéique à n'échantillonner qu'un sous-ensemble restreint de conformations, tandis que les interactions non locales préfèrent des conformères d'énergie libre compatibles avec le biais de conformères locaux (Floudas, 2007).

Dans ce type d'approche le principe est de proposer un ensemble de fragments candidats prédits à partir de la séquence en acides aminés et couvrant la totalité de cette dernière. Ces fragments sont ensuite assemblés pour former un modèle protéique complet.

La méthode la plus performante à l'heure actuelle, dans cette catégorie, est certaine-

ment Rosetta (Rohl et al., 2004; Bradley et al., 2005) et ses dérivées développées dans le groupe de David Baker. La méthode Rosetta initiale se déroule en deux étapes : (i) prédiction des fragments de 3 et 9 résidus pouvant décrire la séquence par des méthodes de comparaison de profils raffinées par différentes méthodes de prédiction des structures secondaires (dont PSIPRED (Jones, 1999b)), et (ii) génération de modèles par assemblage de fragments en utilisant un algorithme de type recuit simulé (Simons et al., 1997). Dans cette méthode d'assemblage, un grand nombre de simulations indépendantes est réalisé. Pour chaque simulation, le point de départ est une chaîne protéique totalement dépliée. A chaque pas du recuit simulé, une fenêtre sélectionnée au hasard est sujette à l'insertion d'un des meilleurs fragments de neufs résidus prédits à cette position. Cette insertion se fait en imposant les angles de torsion du fragment considéré. A chaque pas, la pertinence du modèle est évaluée par une fonction de score (Rohl et al., 2004) dont la formulation évolue au cours de la simulation (chacun des termes est associé jusqu'à la formulation complète à la fin de la simulation). A l'issue de la simulation, un raffinement du modèle est effectué par l'insertion des fragments prédits de 3 résidus avec la formulation complète de la fonction objectif. Dans son protocole standard, Rosetta utilise un modèle gros grain, dans lequel, la chaîne principale est représentée explicitement, ainsi que les carbones β, tandis que le reste de la chaîne latérale est représenté par une sphère localisée sur le centre de masse de cette dernière (pour la glycine le carbone α est choisi comme référence). Cependant, Rosetta est aussi capable de raffiner des modèles tous atomes avec un potentiel physique. Les modèles générés sont ensuite classifiés selon leur similarité structurale, et les modèles retenus sont en général les centres des classes (Bonneau et al., 2002).

SIMFOLD développée récemment par l'équipe de Shoji Takada (Chikenji et al., 2003) est une méthode proche de Rosetta qui a donné de bons résultats lors de la $6^{\text{ème}}$ expérience CASP. Dans cette dernière, sont utilisés des fragments de 4 et 9 résidus assemblés par un algorithme très évolué de type Monte Carlo, guidé par une fonction d'énergie basée sur des considérations physiques (Fujitsuka et al., 2004). La principale originalité de la méthode réside dans son mode d'insertion des fragments qui est réversible. A la différence de Rosetta qui n'utilise que le sous-ensemble de fragments prédits pour les insérer dans le modèle, SIMFOLD insère un nouveau fragment pour faire la jonction entre deux fragments, l'ancienne conformation de la région substituée est ajoutée à l'ensemble des fragments disponibles, et peut donc être insérée à nouveau.

PROFESY est une autre approche *de novo* basée sur un assemblage de fragments mis en place au sein du groupe de Jooyoung Lee (Lee et al., 2004). Comme SIMFOLD, et à la différence des algorithmes classiques de recuit simulé, PROFESY se base sur une méthode efficace d'échantillonage de l'espace conformationnel et utilise un potentiel physique plus que statistique. La minimisation globale de la fonction d'énergie est rendue possible par

l'algorithme de CSA développé au sein du groupe de Harold Scheraga (Lee et al., 1999).

FRAGFOLD, une méthode originale de David Jones, utilise une librairie de fragments constituée de super-structures secondaires (composées de deux ou trois structures secondaires consécutives) issues d'une collection de structures protéiques à haute résolution, ainsi que de petits fragments de 3 à 5 résidus de long (Jones, 1997; Jones and McGuffin, 2003). Les fragments compatibles prédits sont assignés à la séquence à prédire par une méthode d'enfilage directement inspirée de l'algorithme de GenThreader (Jones, 1999a). La structure globale est reconstruite par un algorithme génétique ou un recuit simulé, dans lesquels, la moitié des mouvements aléatoires correspondent à l'insertion de motifs de super-structures secondaires sélectionnés, et l'autre moitié à l'insertion de petits fragments. L'ensemble des modèles générés est ensuite évalué en terme de collisions stériques, de compacité et de nombre de liaisons hydrogène. Enfin, les modèles sont regroupés en classes de repliements les plus représentatifs.

Les méthodes hybrides

Les modèles sur réseau sont une alternative aux méthodes d'assemblage de fragments. Dans ces méthodes, l'espace conformationnel est limité à un ensemble de points que sont les points d'intersection d'une grille à trois dimensions (Skolnick and Kolinski, 1991; Hinds and Levitt, 1992). La résolution du système dépend de la taille de la maille. Ces méthodes permettent une exploration très rapide d'un grand nombre de conformations, mais souffrent d'une faible résolution et par conséquent, de la difficulté d'y implémenter des fonctions d'énergie ayant un sens physique. Récemment, dû au succès grandissant des méthodes d'assemblage de fragments, de nouvelles méthodes hybrides sont apparues combinant les deux approches.

La méthode la plus aboutie dans cette catégorie est TASSER (Zhang et al., 2005). Mise en place par Jeffrey Skolnick et Yang Zhang, TASSER et ses méthodes dérivées ont remporté un franc succès lors de CASP6 (Zhang et al., 2005) et CASP7 (Zhou et al., 2007). La première étape de TASSER est d'identifier à la fois un repliement consensus, mais aussi un ensemble de repliements matrice distincts par la méthode d'enfilage PROSPECTOR (Skolnick et al., 2004). De par les alignements obtenus par la méthode de reconnaissance de repliement, la chaîne protéique est découpée en fragments contigus alignés d'au moins 5 résidus et en regions non alignées. La conformation des régions alignées est copiée en l'état et n'est pas changée lors de la procédure d'assemblage, alors que les régions non alignées sont repliées *ab initio* sur un réseau cubique similaire à ceux développées par Skolnick, Kolinski et al. (Kolinski et al., 2001; Kihara et al., 2001). L'ensemble des modèles générés est alors classé par la méthode itérative SPICKER (Zhang and Skolnick,

2004a) identifiant les modèles selon la densité des classes générées. Deux améliorations majeures ont été récemment apportées à la méthode et testées avec succès à CASP7 : Wu et al. (2007) ont rendu la méthode itérative pour progressivement raffiner les modèles obtenus, et Zhou and Skolnick (2007) ont ajouté, à l'ensemble des fragments identifiés par reconnaissance de repliement, des super-structures secondaires repliées *ab initio* : les *chunks*.

Une autre approche hybride a été développée par Kolinski et Bujnicki, tirant partie de deux méthodes : FRankenstein's Monster (FRM) (Kosinski et al., 2003) et CABS (*Cα-β and Side group*) (Kolinski, 2004). Dans un premier temps, des modèles hybrides sont générés avec la méthode FRM de recombinaison de matrices, puis classés selon leur score Verify3D (Bowie et al., 1991; Luthy et al., 1992) pour identifier les fragments correctement repliés. Ces fragments ne sont pas utilisés directement, mais sont utilisés comme source de contraintes spatiales permettant de guider les simulations de Monte-Carlo avec échange de répliques du modèle CABS. L'ensemble des modèles générés est ensuite classé par la méthode HCPM (Gront and Kolinski, 2005) (*Hierarchical Clustering of Protein Models*) pour identifier les modèles finaux. Cette méthode performante a donné de bons résultats lors de la sixième édition de CASP (Kolinski and Bujnicki, 2005).

Si l'on fait un bilan des performances de l'ensemble des méthodes de prédiction de la structure des protéines, il apparait que les méthodes de modélisation comparative sont très efficaces depuis quelques années et ne semblent de ce fait que très peu évoluer. Les méthodes d'assemblage de fragments continuent de s'améliorer, mais il apparaît que des méthodes hybrides, telle TASSER, combinant à la fois de l'enfilage et une méthode d'assemblage de fragments, puisse rapidement progresser pour donner des résultats très pertinent dans un avenir proche. @TOME et 3D-Jury ont ouvert la porte aux méta-méthodes de prédiction de la structure des protéines. Un méta-serveur, combinant les approches d'assemblage de fragments les plus performantes à l'heure actuelle, serait bienvenue pour parfaire notre connaissance du repliement protéique, et fournir aux biologistes un outil de confiance.

Chapitre 2

Présentation de l'approche HMM-SA

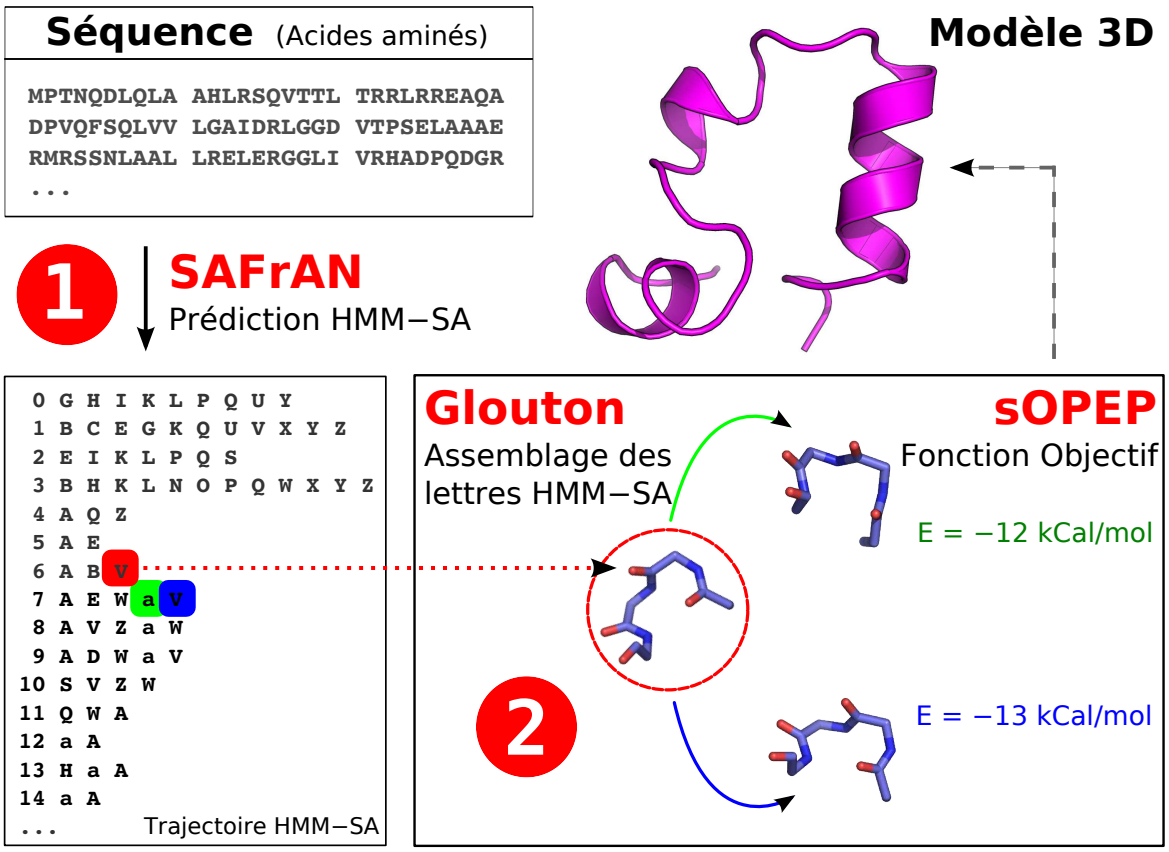


Fig. 2.1: Vue d'ensemble de la méthode HMM-SA. La méthode HMM-SA consiste en deux étapes essentielles que sont (i) prédire un ensemble de lettres de l'alphabet structural HMM-SA compatibles avec la structure à prédire, et (ii) assembler cet ensemble de fragments prédits par un algorithme glouton couplé au potentiel énergétique OPEP.

La méthode HMM-SA, développée au sein de l'EBGM, est une méthode de prédiction de la structure des protéines dite *de novo*. *De novo*, car, les paramètres de la méthode sont dérivés des propriétés de structures protéiques connues, et non uniquement des lois

physiques régissant le repliement des protéines. Cette méthode se base sur la notion d'alphabet structural, *i.e.* un ensemble restreint de conformations prototypes permettant de décrire l'ensemble des structures protéiques. L'alphabet structural HMM-SA (*Hidden Markov Model Structural Alphabet*) a été appris par l'intermédiaire de chaînes de Markov cachées (Camproux et al., 1999, 2004; Camproux and Tuffery, 2005). L'apprentissage de la dépendance entre les lettres de l'alphabet structural et les séquences en acides aminés qui les émettent, nous a permis de prédire l'ensemble des lettres de l'alphabet structural pouvant décrire une séquence protéique donnée.

Ainsi, la première étape de la méthode est de prédire l'ensemble des fragments à assembler pour pouvoir générer un modèle protéique complet. Pour cette étape, Tuffery et al. ont développé SAFrAN (*Structural Alphabet candidate Fragments from AmiNo-acid sequence*) (Manuscrit en cours de préparation), un algorithme original de recherche de fragments compatibles avec une séquence HMM-SA prédite sous contrainte de la prediction des structures secondaires par le logiciel PSIPRED (Jones, 1999b).

Lors de la deuxième étape, l'assemblage des fragments prédits est envisagé par un algorithme glouton (Tuffery et al., 2005; Tuffery and Derreumaux, 2005), guidé par une version simplifiée du potentiel énergétique OPEP (*Optimized Potential for Efficient protein structure Prediction*) (Maupetit et al., 2007).

Lors de ma thèse, il m'a été donné l'occasion de me pencher sur les trois aspects essentiels de la méthode que sont SAFrAN, l'algorithme glouton et le potentiel gros grain OPEP.

Les deux premières années de ma thèse ont porté sur l'optimisation du champ de force OPEP et son implémentation dans l'algorithme glouton développé par Tuffery et Derreumaux. Ce développement a été l'occasion de mettre en place SABBAC (*Structural Alphabet based protein Backbone Builder from Alpha Carbon trace*), un outil de reconstruction de structures protéiques complètes à partir d'une trace en carbones alpha. Je me suis ensuite penché sur l'implémentation de nouveaux opérateurs permettant d'améliorer l'algorithme glouton existant. L'été 2006 a été l'occasion de participer à la septième expérience internationale CASP, qui nous a permis d'apprendre beaucoup sur les limites de la méthode et les améliorations que nous pouvions y apporter. Lors de ma dernière année de thèse, j'ai donc eu l'occasion d'initier certaines de ces améliorations, concernant, la méthode d'assemblage de fragments et sa version simplifiée du potentiel OPEP. J'ai aussi participé à la validation et l'amélioration de la méthode de prédiction SAFrAN déjà en place au sein du laboratoire.

Le plan suivit pour ce manuscrit suit l'histoire d'une prédiction. Dans la suite de cette

introduction, je vais présenter l'ensemble des outils déjà en place au début de ma thèse, *i.e.* l'alphabet structural HMM-SA, ainsi que les méthodes développées à partir de ce dernier, l'algorithme glouton permettant de générer des modèles protéiques à partir de cette représentation discrète de la structure des protéines et la version originale du champ de force OPEP.

Le corps du manuscrit comporte trois parties pour les trois aspects de la méthode de prédiction de la structure des protéines que j'ai abordé : (i) SAFrAN, permettant de passer d'une séquence en acides aminés à un ensemble de fragments candidats décrivant la structure à prédire, (ii) l'optimisation du champ de force gros grain OPEP, son implémentation dans l'algorithme glouton pour guider le repliement des structures protéiques, et une première application qu'est SABBAC, une méthode de reconstruction de structures protéiques complètes à partir de leur trace en carbones α, et enfin, dans une dernière partie (iii), j'exposerai les quelques tests de l'approche que nous avons réalisé, à la fois pour la prédiction de structures protéiques complètes, et pour le repliement de peptides ou de fragments protéiques.

Chapitre 3

L'alphabet structural HMM-SA

Les protéines présentent des motifs récurrents à tous les niveaux de leur organisation structurale. La première description apparue est le modèle à trois états décrivant les protéines comme une suite de structures répétitives que sont les hélices α et les brins β, connectées entre elles par les boucles (Pauling and Corey, 1951). Ce modèle s'est ensuite raffiné pour y ajouter les hélices 3.10 (4% des résidus protéiques) (Sun and Doig, 1998) et les hélices π (plus de 0,2% des résidus), principalement situées dans les régions de transition avec les hélices α (Rajashankar and Ramakumar, 1996). Concernant les régions en boucles, un certain nombre de descriptions ont été entreprises pour les petits coudes β (Richardson, 1981; Wilmot and Thornton, 1988; Hutchinson and Thornton, 1994). Pour des régions plus longues, des motifs répétés ont été identifiés, comme les coudes γ (Rose et al., 1985; Milner-White, 1988), les bosses β (β-*bulges*) (Richardson et al., 1978; Chan et al., 1993), les épingles à cheveux β (Sibanda and Thornton, 1991), les boucles Ω (Fetrow, 1995) ou les coudes α (Pavone et al., 1996; Chou, 1997). Un grand nombre de classifications des zones de connection entre les fragments α et β ont été proposées (Wintjens et al., 1996; Kwasigroch et al., 1996; Boutonnet et al., 1998; Wojcik et al., 1999). Cependant, face à une grande variabilité conformationnelle, la classification des boucles est une tâche complexe même pour des tailles comprises entre 6 et 16 résidus (Leszczynski and Rose, 1986). Pour des longueurs importantes, le fait de combiner des conformations de régions plus courtes n'a pas conduit aux résultats escomptés (Ring et al., 1992). Néanmoins, une telle stratégie basée sur des structures connues, qui aurait pour objectif de caractériser de nouveaux motifs non encore identifiés, donnerait lieu à une tâche sans fin, et nous pourrions nous poser la question de la signification et de la représentativité de tels motifs.

Alternativement, beaucoup d'efforts ont été entrepris pour mettre en place des librairies de fragments, *i.e.* des collections de fragments protéiques de petite taille, capables de décrire l'ensemble des structures protéiques, sans connaissance *a priori* de ces dernières. Depuis les années 80, les travaux pionniers de Unger et al. (Unger et al., 1989; Unger and Sussman, 1993) ont été suivis par un grand nombre d'équipes (Rooman et al., 1990;

Prestrelski et al., 1992; Orengo and Taylor, 1993; Schuchhardt et al., 1996; Fetrow et al., 1997; Camproux et al., 1999; de Brevern et al., 2000) (liste non exhaustive).

Ces librairies de fragments peuvent poursuivre deux objectifs distincts. Dans un cas, elles visent à raffiner la description de la structure des protéines, et, dans l'autre cas, elles ont été mises en place pour permettre de reconstruire et modéliser les structures protéiques.

Les descripteurs de la chaîne polypeptidique utilisés dans ces librairies sont variés. Dans certains cas (Unger et al., 1989; Rooman et al., 1990), c'est la conformation de la trace en carbones α qui permet de caractériser les conformations de segments de 4 et 7 résidus de long. Prestrelski et al. (1992) ont utilisé quant à eux un jeu de descripteurs combinant à la fois les angles dièdres formés par 4 carbones α consécutifs, et la distance entre ces carbones α consécutifs pour des segments de taille 8. Schuchhardt et al. (1996) ainsi que de Brevern et al. (2000) ont utilisé une description classique de la chaîne principale par les couples d'angles dièdres (ϕ, ψ) pour des fragments de taille 9 et 5 respectivement. Zhang et al. (1993); Fetrow et al. (1997) ont utilisé les distances inter-carbones α (sans les distances inter-carbones consécutives), combinées avec une description des angles (ϕ, ψ) et des angles dièdres formés par quatre carbones α consécutifs, pour des fragments de taille 7 ; tandis que Camproux et al. (1999) n'ont utilisé que les distances inter-carbones α non consécutives pour décrire des segments protéiques de quatre résidus de long. Dans une optique de prédiction, Bystroff and Baker (1998) ont développé les I-sites, une librairie de fragments de tailles allant de 3 à 19 possédant une signature de séquence et de structure marquée.

Afin d'évaluer la pertinence des librairies de fragments pour décrire les structures protéiques, Kolodny et al. (2002); Kolodny and Levitt (2003) ont mis en place plusieurs librairies de fragments de taille allant de 4 à 7 résidus de long. Micheletti et al. (2000) ont quant à eux défini des librairies de fragments, décrits par leurs carbones α, de longueurs allant de 3 à 10, et permettant de décrire les structures protéiques avec une grande précision (moins de 1 Å).

En marge de ces librairies de fragments s'appliquant aux protéines globulaires, Martin et al. (2007) ont développé un alphabet spécifique des OMP (*Outer Membrane Proteins*, protéines membranaires formant des tonneaux β) permettant une description fine de la spécificité séquence-structure qui les caractérise.

Dans ce chapitre, nous allons brièvement présenter comment l'alphabet structural que nous utilisons a été mis en place, ses propriétés et les méthodes dérivées de ce dernier.

3.1 Le modèle markovien

Un des alphabets structuraux identifié au sein de l'EBGM, et que nous avons utilisé dans le cadre de l'approche HMM-SA, a été développé à partir de chaînes de Markov cachées (HMM) dont il a été montré qu'elles étaient pertinentes pour décrire les conformations locales des protéines (Camproux et al., 1999).

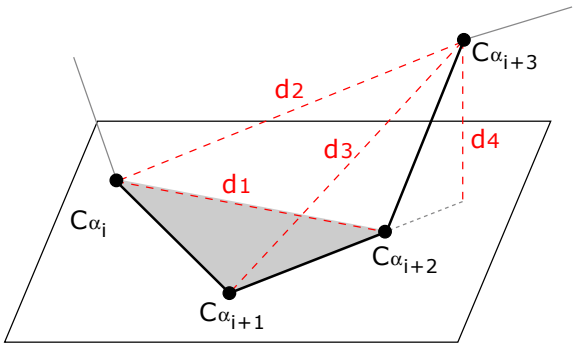


Fig. 3.1: Les descripteurs de l'alphabet structural HMM-SA. Chaque lettre de l'alphabet structural HMM-SA est la combinaison de valeurs spécifiques des descripteurs d_1 à d_4.

Chaque lettre de l'alphabet structural HMM-SA est un fragment protéique de quatre résidus de long, caractérisé par quatre descripteurs que sont les 3 distances inter-carbones *alpha* non consécutifs (d_1, d_2, d_3), et la projection du quatrième carbone *alpha* sur le plan formé par les trois premiers (d_4). Les lettres de l'alphabet se chevauchent sur trois résidus.

Les HMM permettent d'apprendre simultanément la géométrie de chaque état, mais aussi leur transition. Le processus d'apprentissage s'effectue en reconstruisant une séquence non observable d'états d'un système, à partir d'une séquence d'observation dépendant aléatoirement de la séquence d'états.

Il s'agit de discrétiser ou de classifier en k états la conformation de la chaîne peptidique des protéines. Ces k états ou blocs structuraux (S_1, S_2, ... , S_k) ne sont pas définis *a priori*, mais sont appris des fragments observés. Ils sont ainsi dits cachés. Si l'on suppose que la taille des blocs structuraux est de 4 résidus se chevauchant sur 3, alors une protéine de longueur N peut être décomposée en une série de $N-3$ fragments. Chaque fragment, à la position i, est alors décrit par un vecteur de 4 descripteurs $Y_i = (d_1^i, d_2^i, d_3^i, d_4^i)$ (voir la figure 3.1). Notre protéine est donc une composition de $N-3$ vecteurs descripteurs, notés $Y_1, Y_2, ..., Y_{N-3}$. Le but est ainsi d'obtenir la séquence optimale d'états ou de blocs cachés correspondante $X_1, X_2, ..., X_{N-3}$. Nous savons que la séquence optimale d'états $X_1, X_2, ..., X_{N-3}$ existe avec une certaine variabilité, qui est ici décrite par une loi normale de dimension 4. Pour prendre en compte la dépendance entre les blocs, nous supposons

que l'agencement de ces derniers est gouverné par une chaîne de Markov d'ordre 1. Ainsi, lorsqu'un état est choisi à une position i, seul un sous-ensemble d'états peut décrire la position $i + 1$.

Le modèle de Markov caché a donc pour but d'identifier des blocs structuraux, leurs règles logiques et de reconstruire la séquence d'états la plus probable étant donné un ensemble d'observations.

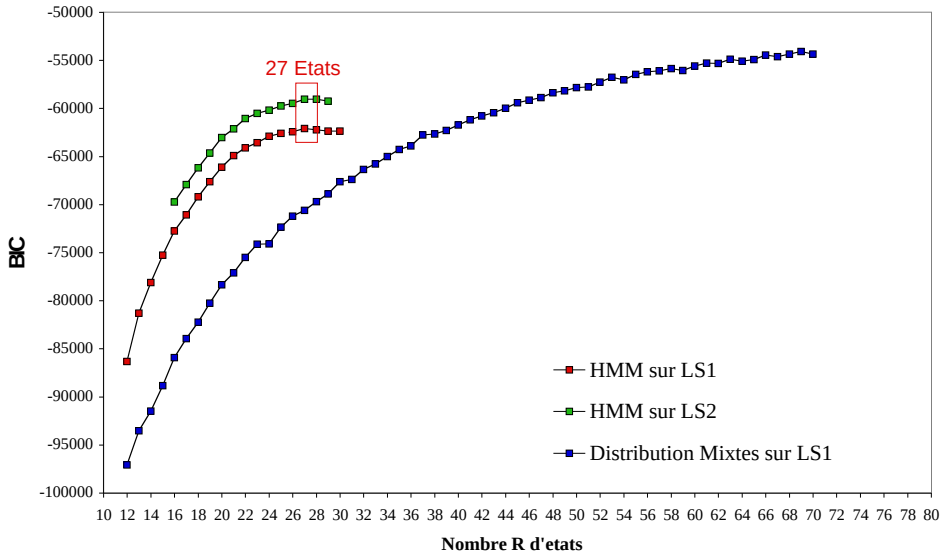


Fig. 3.2: Evaluation du nombre d'états optimaux par le critère BIC. L'évolution du critère BIC est ici estimée en fonction du nombre d'états pour deux modèles HMM et un modèle mixte. Figure extraite de Camproux et al. (2004).

L'apprentissage a été effectué sur deux échantillons LS_1 et LS_2 de 250 protéines, ce qui représente un ensemble de 56167 et 57544 résidus respectivement. Le nombre optimal d'états, du point de vue de la parcimonie du modèle, a été testé par l'intermédiaire du critère statistique BIC (*Bayesian Information Criterion*) (Schwartz, 1978). L'évolution de ce critère en fonction du nombre d'états de l'alphabet structural HMM-SA (voir la figure 3.2) a permis de déterminer que l'optimum est atteint pour 27 états (Camproux et al., 2004) sur les deux jeux de protéines LS_1 et LS_2. A titre de comparaison, un modèle mixte, correspondant à un mélange de lois normales aléatoires, sans prise en compte des transitions (chaîne de Markov d'ordre 0) est aussi présenté. Avec ce modèle, l'optimum n'est jamais atteint, même pour 70 états différents, mettant en évidence la pertinence du modèle HMM.

Les 27 lettres ou états de l'alphabet obtenus sont représentés dans la figure 3.3. Les valeurs des descripteurs associés, ainsi que les occurences des prototypes sont présentés dans la table 3.1. Nous pouvons noter que la variation géométrique des carbones *alpha*

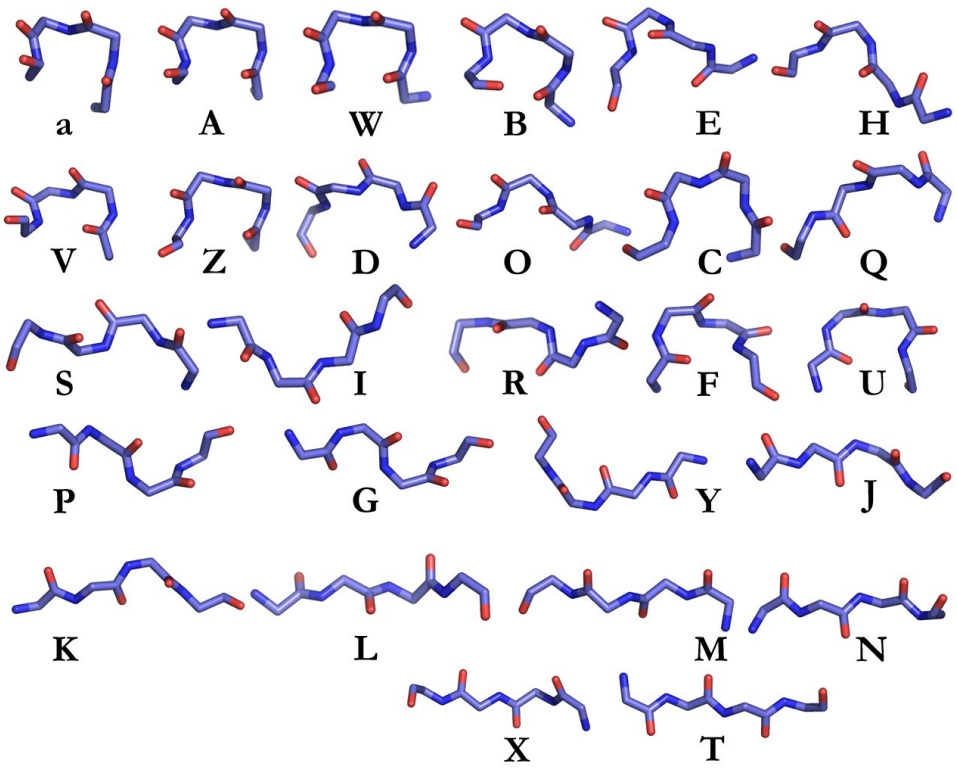


Fig. 3.3: Les lettres de l'alphabet structural HMM-SA. L'alphabet structural HMM-SA est composé de 27 lettres ou états. Ne sont représentés dans cette figure que les prototypes moyens correspondant à chaque état. Les hélices α sont représentées par les lettres `A` et `a`, tandis que les brins β sont représentés par les lettres `L`, `M`, `N`, `T` et `X`.

de chaque état (cRMSd, α *carbon Root Mean Square Deviation*) est faible : elle s'étend entre 0,08 et 0,91 Å pour une valeur moyenne de 0,23 $\pm$ 0,14 Å.

La correspondance des états aux structures secondaires canoniques a par ailleurs aussi été évaluée (voir la figure 3.4). La distribution des Z-scores des 27 blocs pour les 12 classes de structures secondaires STRIDE ont permis de mettre en évidence la sur- ou sous-représentation de certaines lettres HMM-SA dans ces dernières. Ainsi, les blocs `A` et `a` semblent spécifiques des hélices α, tandis que les blocs `L`, `M`, `N`, `T` et `X` sont plus spécifiques des brins β.

3.2 Encodage des structures protéiques

L'encodage des structures protéiques dans l'espace de l'alphabet structural HMM-SA permet de compresser l'information 3D en une seule dimension : la séquence HMM-SA. Un exemple de protéine encodée est présenté dans la figure 3.5. Cette discrétisation de la structure des protéines peut être réalisée soit par l'algorithme de Viterbi (Rabiner, 1989), soit par l'algorithme de *forward-backward* (Rabiner, 1989).

Etats	Occurences		Descripteurs (Å)				cRMSd Å	
	%	56167	d_1	d_2	d_3	d_4	w	dev
a	2,6	1460	5,39	5,09	5,38	2,92	0,13	0,15
A	12,6	7077	5,43	5,09	5,42	2,94	0,08	0,15
V	5,6	3145	5,41	5,23	5,61	2,86	0,09	0,15
W	5,3	2977	5,62	5,25	5,42	2,87	0,09	0,15
Z	4,5	2528	5,59	5,49	5,78	2,58	0,20	0,27
B	4,7	2640	5,40	5,58	5,42	3,39	0,13	0,27
C	1,8	1011	5,78	5,68	6,07	1,46	0,22	0,42
D	2,0	1123	5,55	7,74	5,60	-3,31	0,25	0,63
E	2,0	1123	5,60	6,71	5,50	3,69	0,27	0,53
O	1,5	843	5,69	8,09	5,67	3,09	0,25	0,64
S	3,2	1797	5,66	8,95	6,54	2,09	0,28	0,67
R	1,7	955	5,66	8,91	6,66	-1,46	0,49	0,67
Q	4,1	2303	5,66	8,07	6,70	2,96	0,31	0,69
I	2,9	1629	5,70	7,26	7,02	0,88	0,31	0,75
F	1,9	1067	6,03	6,85	5,64	-0,63	0,91	1,08
U	2,0	1123	6,47	5,92	5,56	0,53	0,38	0,54
P	4,4	2471	6,57	8,96	5,58	-2,19	0,29	0,53
H	2,7	1517	6,71	8,27	5,47	-3,56	0,22	0,53
G	3,4	1910	6,21	9,21	5,77	0,27	0,29	0,69
Y	2,0	1123	6,87	8,28	6,03	-3,44	0,29	0,53
J	2,0	1123	6,89	8,94	6,76	-0,48	0,55	0,83
K	4,1	2303	6,72	9,12	6,41	-3,31	0,23	0,49
L	5,1	2865	6,71	9,64	6,50	-2,60	0,24	0,49
N	4,9	2752	6,39	9,93	6,75	-1,07	0,22	0,43
M	5,3	2977	6,87	10,06	6,51	-1,41	0,23	0,46
T	3,0	1685	6,48	10,17	7,09	0,66	0,21	0,43
X	4,7	2640	6,80	10,35	6,85	-0,25	0,25	0,53

Tab. 3.1: Description des 27 états de HMM-SA. Pour chaque état sont données les occurences de ce dernier, ainsi que les valeurs moyennes des descripteurs. $cRMSd_w$ est la déviation moyenne du fragment par rapport au centroïde de l'état, et $cRMSd_{dev}$, la déviation minimum moyenne entre les fragments correspondant aux différents états. Table extraite de Camproux et al. (2004).

L'algorithme de *forward-backward* permet de calculer l'ensemble des trajectoires possibles à chaque position. Pour chaque position, nous pouvons donc retenir les N états les plus probables. Ce critère maximise localement le nombre d'états correctement prédits, mais peut proposer, comme solution finale, une succession d'états incorrecte, la dépendance étant prise en compte à chaque position.

L'algorithme de Viterbi (Rabiner, 1989; Baum et al., 1970) identifie la séquence de N états la plus probable parmi toutes les trajectoires possibles dans l'espace, à l'aide de l'algorithme de programmation dynamique prenant en compte la transition markovienne entre les états consécutifs. Contrairement à l'algorithme de *forward-backward*, l'algorithme de Viterbi ne propose qu'une solution unique à chaque position, sans indice de confiance local de l'état choisi. Quoi qu'il en soit, il a été observé une grande concordance entre les encodages produits par ces deux algorithmes.

Il est important de noter que cet encodage des structures protéiques résulte bien d'une compression de l'information tridimensionnelle. La perte associée à cette compression n'est

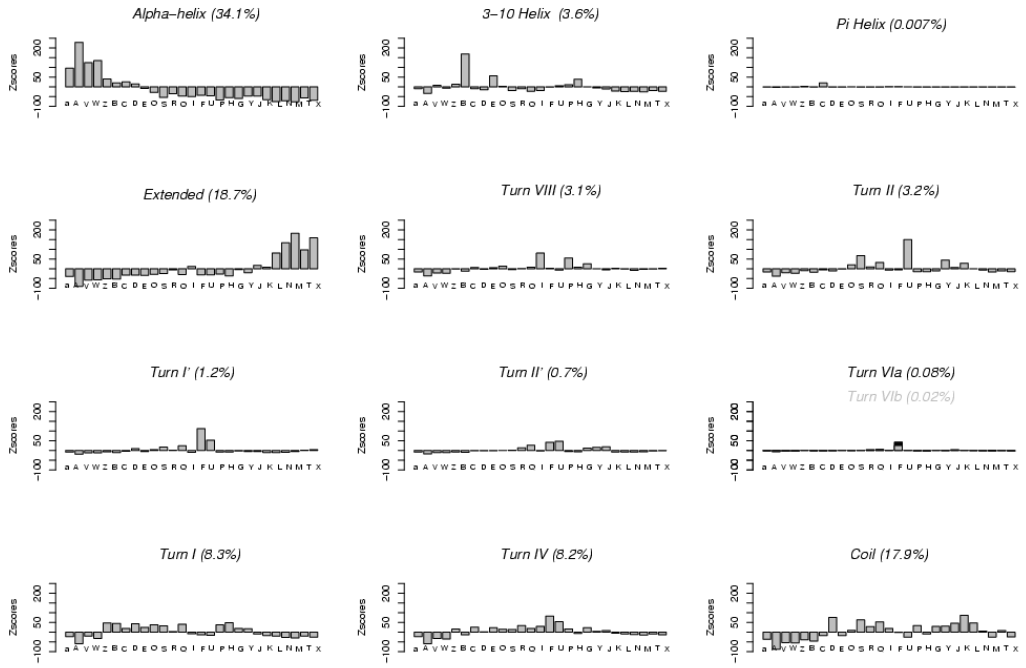


Fig. 3.4: Les états HMM-SA dans les structures secondaires. Pour chacune des 12 classes de structures secondaires identifiées par STRIDE (Frishman and Argos, 1995), la distribution des Z-scores des 27 lettres HMM-SA est présentée. La fréquence des structures est reportée entre parenthèses. Figure extraite de Camproux et al. (2004).

```
>Acides Aminés
KVYGRCELAAAMKRLGLDNYRGYSLGNWVCAAKFESNFNT
HATNRNTDGSTDYGILQINSRWWCNDGRTPGSKNLCNIPC
SALLSSDITASVNCAKKIASGGNGMNAWVAWRNRCKGTDV
HAWIRGCRL

>HMM-SA
NLHWAAAAAVWAVDOQUSUFSLHBBVWAAAVZZFFFSPBS
XTLNHZDSNLNJFZDRLPECCILGDEQLUGPRGBDSKHBB
BBQHEGOWAVWAAAVWABQHZRUEEEGWAAZCCQUQYGEB
BVSUSP
```

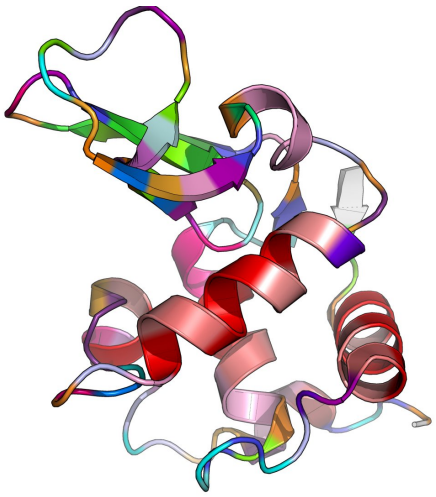


Fig. 3.5: Un exemple de structure encodée dans l'espace HMM-SA. Cette structure du lyzozyme (Code PDB : `1351`) a été encodée dans l'espace de l'alphabet structural par l'algorithme de Viterbi. Chaque lettre de l'alphabet structural est ici figurée par une couleur spécifique dans la séquence HMM-SA ; couleur que l'on retrouve dans sa structure à droite.

que minime, car si l'on adopte la démarche inverse, de décompresser les séquences HMM-SA, en reconstruisant un modèle tridimensionnel, nous savons que ce modèle sera en

moyenne éloigné de 1 Å de la structure native (Tuffery et al., 2005). Cette décompression est réalisée par l'intermédiaire d'un algorithme glouton que nous présenterons en détails dans la suite de cette introduction.

3.3 Les travaux dérivés de HMM-SA

Depuis sa création, l'alphabet structural HMM-SA est utilisé dans divers domaines parmi lesquels la recherche de similitude 3D et l'analyse fine des structures protéiques.

3.3.1 Recherche de similitude 3D

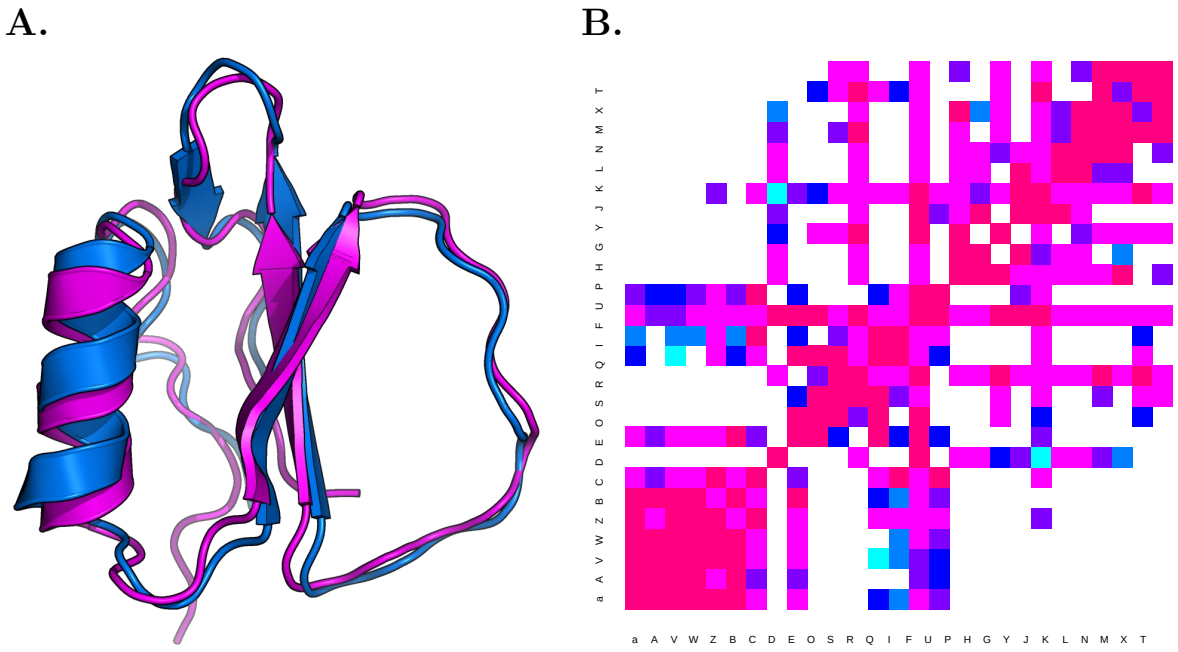


Fig. 3.6: SA-Search, un outil pour détecter des homologues structuraux. A. La structure requête 2ci2I (en magenta) superposée à la structure 1cseL (en bleu) un de ses homologues structuraux détecté par SA-Search. **B.** La matrice de substitution utilisée dans SA-Seach. Le gradient va du rouge (substitution favorable) au blanc (substitution interdite).

SA-Search est fondée sur le fait que des séquences de lettres identiques donnent des conformations 3D identiques (Camproux et al., 1999). Elle a alors été développée dans l'optique suivante : si l'on n'a pas des séquences de lettres strictement identiques, qu'en est il de la similitude structurale ?

SA-Search (Guyon et al., 2004) est un outil permettant de détecter des homologues structuraux à une protéine donnée. Partant d'un ensemble de coordonnées 3D encodées dans l'espace de l'alphabet structural, il est possible d'utiliser les techniques classiques d'alignement pour explorer une banque de recherche de protéines codées en une suite de

lettres HMM-SA. Ainsi, l'algorithme de Smith et Waterman (Smith and Waterman, 1981) est utilisé pour détecter rapidement des structures compatibles entre elles (voir la figure 3.6.A). Des matrices de *log-odds* ont été construites pour quantifier le "coût" associé à la substitution d'une lettre HMM-SA par une autre. Cela a été rendu possible par l'encodage d'un ensemble de structures par l'algorithme *forward-backward* et considérant la probabilité que chaque lettre HMM-SA puisse encoder chaque position de cet ensemble de structures. Une représentation de la matrice de substitution déduite est présentée dans la figure 3.6.B.

Cet outil a été mis à la disposition de la communauté à l'adresse suivante : http ://bioserv.rpbs.jussieu.fr/cgi-bin/SA-Search

Cette méthode a initié un certain nombre d'approches dérivées parmi lesquelles Tyagi et al. (2006); Yang and Tung (2006); Friedberg et al. (2007).

3.3.2 Les boucles protéiques

Les boucles protéiques sont classiquement décrites comme des régions connectant les structures secondaires. Elles présentent une grande diversité de séquence, et sont très variables en terme de structure, ce qui les rend très difficiles à prédire. Cependant, la mise en place de méthodes de classification des boucles (Wojcik et al., 1999; Fernandez-Fuentes et al., 2006a), ont permis de mettre en évidence des motifs structuraux spécifiques de certaines classes. Sachant que ces motifs ont très souvent une implication fonctionnelle, Regad et al. ont entrepris d'analyser ces motifs, dans l'espace de l'alphabet structural HMM-SA. Ainsi, il a été suggéré que 18 des 27 lettres de l'alphabet sont spécifiques des régions en boucles (Regad et al., In press). De plus, 97% des motifs présentant une faible variabilité structurale (cRMSd inter-boucle inférieur à 0,4 Å) sont sur-représentés, et ont une grande spécificité de séquence (Nuel et al., Submited).

3.3.3 Les chaînes latérales

Nous savons qu'il existe une forte dépendance entre la conformation de la chaîne principale et la celle des chaînes latérales. Cependant, étant donné le nombre conséquent de combinaisons ϕ,ψ envisageables, l'utilisation d'un alphabet structural décrivant un ensemble réduit de conformations types de la chaîne principale, semble être un outil de choix pour analyser cette dépendance. Ainsi, Gautier et al. (2004) ont consitué une librairie de rotamères dépendant de l'état de la chaîne principale décrite par l'alphabet structural HMM-SA.

La méthode de reconstruction et d'analyse qui en sont dérivées ont été mises à la disposition de la communauté à l'adresse suivante : http ://bioserv.rpbs.jussieu.fr/SCit

3.3.4 Les interactions protéiques

Il existe un certain nombre d'arguments en faveur de l'hypothèse de l'ajustement induit pour permettre l'interaction entre deux protéines. Une étude systématique de protéines, dont la structure a pu être résolue isolée et au sein d'un complexe protéique, a révélé que ce phénomène d'ajustement induit implique des changements d'états significatifs associés à certaines lettres de l'alphabet structural HMM-SA. Ces lettres sont souvent associées à des régions non structurées (Martin et al., Submited).

Plus généralement, il a aussi été suggéré que la spécificité des contacts, observables dans le cas d'une interaction protéine-protéine, peut aussi être décrite au sein même d'une protéine. Une analyse systématique des contacts intra-protéine, définis à partir de tesselations de Voronoï (Voronoi, 1907), a permis de mettre en évidence que l'alphabet structural permet de décrire une plus grande spécificité de paires de contacts que les acides aminés (Martin et al., en préparation).

Chapitre 4

L'algorithme glouton

Lors du développement d'une méthode de prédiction de la structure des protéines, deux problèmes majeurs se posent. La première difficulté réside dans l'obtention d'une fonction de score capable, sans information aucune sur la structure protéique à reconstruire, de retrouver la structure native comme étant un minimum global dans l'ensemble du paysage énergétique de conformations alternatives. Nous allons aborder ce point lors de l'optimisation de la fonction d'énergie OPEP. La deuxième difficulté consiste en une exploration efficace de l'espace conformationnel.

Beaucoup de travaux ont été entrepris durant ces dernières années pour tenter de résoudre ce problème d'exploration de l'espace conformationnel. Les méthodes d'échange de répliques (Hansmann, 1997) ou les dynamiques d'ensemble (Zagrovic et al., 2001) ont ouvert de nouvelles perspectives pour le repliement des protéines, mais nécessitent tout de même de grandes ressources de calcul. D'autres heuristiques ont ainsi été utilisées pour trouver une bonne solution, qui on le sait, peut ne pas être optimale. C'est le cas du recuit simulé (Kirkpatrick et al., 1983), de la recherche tabou (Glover, 1989), des algorithmes génétiques (Gunn, 1997), combinés à la recherche tabou (Jiang et al., 2003), des stratégies de croissance de chaîne (Yue and Dill, 2000), et de divers méthodes dérivées d'algorithmes de Monte-Carlo (MC) (Abagyan and Totrov, 1994; Wei et al., 2003; Hansmann and Okamoto, 1993).

L'algorithme que nous allons présenter, retenu dans l'approche HMM-SA, est un algorithme glouton inspiré des travaux de Park and Levitt (1995) et Kolodny et al. (2002).

Le choix d'un algorithme glouton, basé sur une stratégie locale de reconstruction, peut sembler étonnant, étant donné que nous savons que le repliement des protéines fait appel à des contacts de longue portée. Cependant, il était espéré que le modèle markovien, sous-jacent à l'alphabet structural, permettrait d'effectuer le passage du niveau local au niveau global. La pertinence d'un tel algorithme avait déjà été explorée par Vendruscolo et al. (1997) pour la reconstruction de structures protéiques à partir de cartes de contacts.

4.1 L'algorithme

4.1.1 Les fragments candidats à assembler

Chaque structure protéique peut être décrite par un ensemble de fragments candidats que nous appelerons la trajectoire de la protéine à reconstruire. Lorsque l'on utilise les lettres de l'alphabet structural HMM-SA, les fragments protéiques à assembler ont une longueur de 4 résidus. Les 27 lettres de l'alphabet structural correspondent à un ensemble de 155 prototypes distincts, chaque lettre étant représentée par un ensemble de sous-conformations allant de 1 à 23 centroïdes.

Complexité de la recherche

A chaque position i d'une trajectoire de longueur L (pour une protéine de $L+3$ acides aminés), est envisagé un ensemble de fragments candidats n_i. Ainsi le nombre de reconstructions possibles de la protéine complète est $\prod_{i=1}^{L} n_i$. En suivant la convention définie par Park et Levitt (Park and Levitt, 1995), la complexité de la recherche (*i.e.* le nombre moyen d'états par résidu) est $\sqrt[L+3]{\prod_{i=1}^{L} n_i}$. Prenons l'exemple, d'une protéine de sept résidus de long, décrite par quatre lettres de l'alphabet structural HMM-SA (une lettre par position) (*e.g.* `AXPJ`), si le nombre de prototypes associés à ces lettres sont respectivement 1, 2, 5 et 4, alors la complexité est : $\sqrt[7]{1*2*5*4} = 1,69$.
Considérons maintenant, une trajectoire plus floue dans laquelle, la position 1 est décrite par les lettres `[A,B,a]`, la position 2 par `[V,X]`, la position 3 par `[P,Q,R]` et la position 4 par `[J]`. Sachant que le nombre de prototypes associés aux lettres `A`, `B`, `a`, `V`, `X`, `P`, `Q`, `R` et `J` est respectivement 1, 1, 1, 3, 2, 5, 3, 6 et 4, le nombre de fragments possibles pour les quatre positions est maintenant : 3, 5, 14 et 4. La complexité qui en résulte est donc de $\sqrt[7]{3*5*14*4} = 2,61$. Il est important de noter que ces trajectoires floues augmentent l'espace conformationnel, mais elles ne conduisent par forcément l'algorithme vers la solution, produisant ainsi des topologies non natives.

Assemblage des fragments candidats

Si l'on considère que nous reconstruisons la structure de l'extrémité N vers C terminale, les fragments sont assemblés en superposant les trois premiers carbones *alpha* du nouveau fragment sur les trois derniers carbones *alpha* de la structure reconstruite de longueur l. La superposition est assurée par une méthode recherchant le meilleur ajustement entre des ensembles de coordonnées en utilisant les quaternions (Zuker and Somorjai, 1989). Cela résulte en une structure de longueur $l+1$ où le quatrième carbone *alpha* du nouveau fragment est inséré (voir la figure 4.1).
Afin de garantir la réversibilité de la procédure, permettant ainsi l'échange d'un fragment contre un autre, ou pour assembler les fragments de l'extrémité C vers N terminale, nous

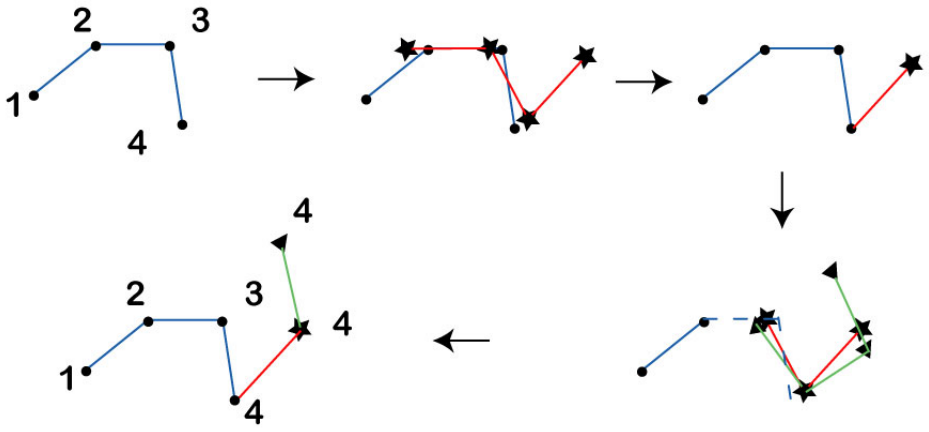


Fig. 4.1: Méthode de superposition des fragments candidats. Les fragments candidats sont assemblés en superposant les 3 derniers carbones *alpha* du fragment à la position i sur les 3 premiers carbones *alpha* du fragment à la position $i+1$. N'est conservé que le quatrième carbone *alpha* du fragment superposé.

conservons les coordonnées ayant servi à la procédure de superposition, et pas seulement celles du carbone *alpha* ajouté à la position $i+1$. Ainsi, si l'on change le fragment k à la position i par un fragment k', puis que l'on décide de revenir au fragment k, les coordonnées générées seront celles d'origine.

4.1.2 L'algorithme original

L'algorithme glouton reconstruit incrémentalement une structure depuis l'extrémité N vers C terminale. Le nombre possible de reconstructions pour une structure de longueur $L+3$ est clairement trop important pour être exploré exhaustivement. L'idée majeure des algorithmes gloutons est, à chaque position i, de reconstruire l'ensemble des solutions possibles ; de les trier selon un critère objectif, puis de ne retenir qu'un maximum de h (pour *heap*) conformations pour l'itération suivante. Ces h conformations constituent la pile. Cette procédure est répétée jusqu'à ce que soit atteinte l'extrémité C terminale. La performance de l'algorithme dépend directement de son seul paramètre h. Dans notre cas, nous utilisons une taille de pile de 3000, une valeur proche de celle utilisée dans d'autres études (Kolodny et al., 2002), et semblant être un bon compromis entre performance et rapidité. Ainsi, si à la position i, nous avons généré 12.000 conformations, ces 12.000 conformations seront triées selon un critère objectif (cRMSd, énergie, ...), et seules les 3.000 meilleures d'entre elles seront conservées pour l'itération $i+1$ (voir la figure 4.2). Une telle procédure ne garantit cependant pas que l'optimum global soit atteint. Ainsi, Tuffery et al. ont apporté un certain nombre d'améliorations que nous allons détailler dans la section suivante.

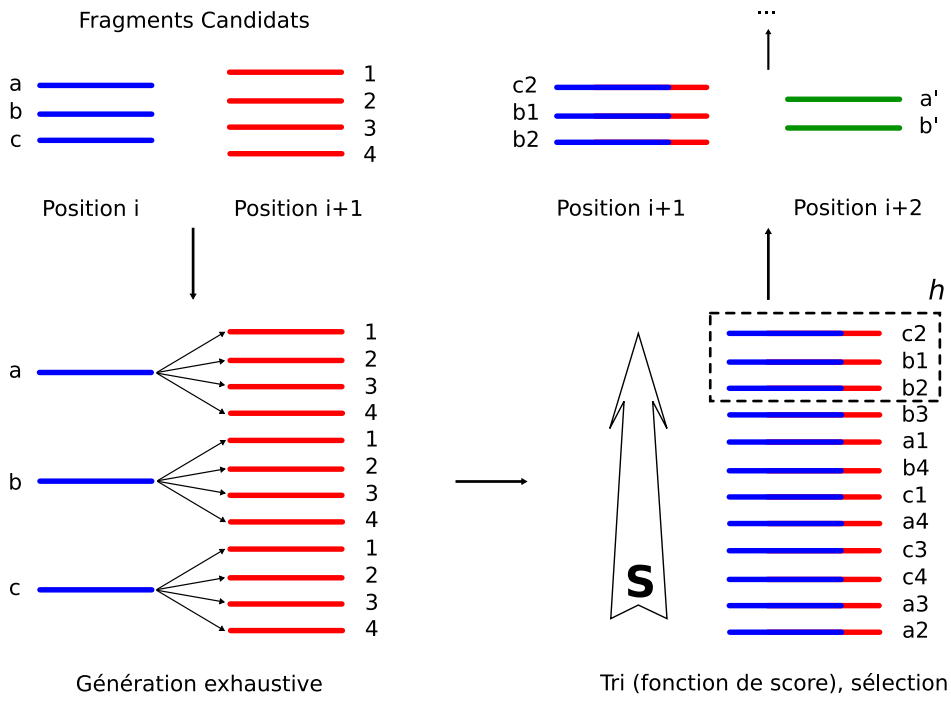


Fig. 4.2: Principe de l'algorithme glouton. Si l'on considère un ensemble de trois fragments à la position i et quatre à la position $i+1$, l'ensemble des combinaisons de fragments est généré, puis trié, et ne sont conservées que les h meilleures combinaisons, h étant la taille de la pile. Cette pile servira de point de départ pour la prochaine itération.

4.1.3 Améliorations de l'algorithme original

Stochastique

La performance d'un tel algorithme réside dans sa capacité à s'échapper des minima locaux. Une version stochastique de l'algorithme est utilisée dans la selection des h meilleures conformations. Si l'on considère un nombre courant de conformations C_i à la position i, les h_b meilleures conformations (au sens de la fonction objectif) sont conservées, et $h - h_b$ conformations sont sélectionnées aléatoirement dans le reste de la pile $(C_i - h_b)$. Par exemple, si l'on considère toutes les lettres de l'alphabet HMM-SA à la position i et une taille de pile $h = 3000$, nous avons donc $C_i = 465.000$ $(3000 * 155)$ conformations. Si $h_b = 1000$, il en résulte une sélection aléatoire de $h - h_b = 2000$ conformations parmi les $C_i - h_b = 464.000$ conformations restantes.

Itéré

Un des problèmes associé à une procédure incrémentale est que la sélection des conformations ne prend en compte la séquence entière que lorsque l'extrémité opposée est atteinte. Or, nous ne pouvons pas ignorer que, dans les structures protéiques, les interactions de longue portée peuvent conduire l'algorithme dans un autre bassin énergétique. Afin de pallier ce problème, a été introduite une procédure itérative dans l'algorithme glouton.

Toutes les itérations, sauf la première, sont donc réalisées en ayant généré une structure protéique complète. Dans ce processus, si l'on considère que l'on progresse de l'extrémité N vers C terminale, et que l'on substitue un fragment, l'extrémité C terminale résultante est ainsi repositionnée en conséquence. Ceci a été rendu possible grâce à la réversibilité de la procédure. D'une manière générale, le nombre d'itérations utilisées jusqu'alors a été fixé à deux.
L'algorithme peut partir de l'extrémité N (noté *forward*) ou C terminale (noté *backward*), ce sens de progression s'inversant à la fin de chaque itération.

Préfiltré

Les préfiltres sont des mini-simulations indépendantes réalisées avant la simulation complète et visant (i) à réduire la complexité de la recherche, et (ii) à inactiver certaines combinaisons de fragments dont les géométries sont incompatibles avec des structures cohérentes. Concrètement, nous considérons une fenêtre glissante de 20 acides aminés tous les 15 résidus. Pour chaque mini-simulation, une itération *forward* et une itération *backward* sont réalisées avec une taille de pile de 1000, la sélection stochastique étant fixée à 300.

4.1.4 La fonction objectif

cRMSd

Dans un premier temps, pour évaluer les performances de l'algorithme, la fonction objectif testée a été le critère cRMSd garantissant qu'un essemble de minima locaux pouvait conduire à minimum global. Ce critère nous permet ainsi de connaître la meilleure reconstruction que l'on peut atteindre avec une trajectoire donnée.

Potentiel de Go

Dans un deuxième temps, Tuffery et al. ont testé un critère de type Go (Udea et al., 1978) donnant une information plus floue sur la structure à reconstruire. L'expression de ce critère énergétique est la suivante :

$$E_{Go} = \sum_{j>i} \Delta_{ij} B_{ij} + E_{rep} \tag{4.1}$$

avec $B_{ij} = -1$ kCal/mol et $\Delta = 1$ quand les résidus i et j forment un contact natif, sinon $\Delta = 0$. Un contact natif est défini sur la structure expérimentale tous-atomes (y compris les atomes d'hydrogène) lorsque deux atomes quelconques appartenant à deux chaînes latérales non consécutives ($j >= i + 4$) sont séparés d'une distance inférieure à

3 Å. Dans les structures générées, un contact est considéré comme natif lorsque la distance $C_\alpha - C_\alpha$ est inférieure à λ fois sa valeur dans la structure expérimentale. λ est ici fixé à 1,3 (Clementi et al., 2001), mais il semblerait qu'une valeur de 1,2 n'affecte que peu les résultats (données non présentées). Le potentiel de volume exclu, E_{rep}, est de 2 kCal/mol pour tous les C_α séparés au moins de six acides aminés dans la séquence, dont la distance est inférieure à 3.5 Å.

Potentiel gros grain sOPEP

Dans un objectif *ab initio*, la fonction objectif utilisée maintenant en production est une version simplifiée du potentiel gros grains OPEP (*Optimized Potential for Efficient protein structure Prediction*) qui sera présenté en détails dans la suite de cette introduction.

4.2 Performances de l'algorithme

4.2.1 A partir d'une description floue de la structure

La trajectoire optimale d'un protéine correspond à l'encodage de sa structure par l'algorithme de Viterbi. Elle contient donc une seule lettre HMM-SA par position. Ce que nous appellerons une trajectoire floue dans le reste de ce manuscrit, correspond à l'encodage d'une structure protéique par l'algorithme *forward-backward*. Dans ce cas, toutes les lettres HMM-SA ayant une probabilité d'apparition supérieure à un seuil (ici $p \geq 10^{-6}$) sont sélectionnées. Ces trajectoires contiennent en général 2 à 4 lettres HMM-SA par position.

Si l'on considère comme critère objectif le cRMSd, pour reconstruire un ensemble de 16 structures protéiques de tailles et de classes protéiques diverses (voir la table 4.1 pour les classes et identifiants), les modèles générés sont en moyenne éloignés de 0,93 $\pm$ 0,2 Å (cRMSd) de la structure native en partant d'une trajectoire optimale, et de 0,55 $\pm$ 0,15 Å, avec une trajectoire floue. Ceci montre que la reconstruction complète d'une structure protéique, en ayant connaissance de la position des carbones alpha, est indépendant du nombre de degrés de liberté. De ce fait, une complexité accrue améliore les performances de reconstruction. Ce point a déjà été discuté sur de nombreux jeux de protéines, en utilisant différentes librairies de fragments et des fonctions objectifs basées sur le RMSd (Micheletti et al., 2000; Kolodny et al., 2002; Camproux et al., 2004). *A contrario*, une grande complexité devient problématique en utilisant une fonction moins stringente tel qu'un critère énergétique, car la recherche n'est pas restreinte à chaque position.
Dans notre cas, pour les trajectoires floues, en comparaison avec les trajectoires optimales, la complexité augmente en moyenne d'un facteur 1,5 à chaque position (voir la table 4.1), et le nombre total de conformations correspondant de $1,5^{L+3}$ pour une protéine de lon-

gueur $L + 3$.

PDB	**L**	**CATH**	C_o	C_f	E_n	E_m	**cRMSd**
1csp	67	β	4,37	6,28	-78	-75 (-63)	3,14 (7,70)
1ctf	68	α/β	2,48	4,45	-79	-78 (-46)	1,16 (9,69)
1pgb	56	α/β	2,88	4,54	-62	-62 (-55)	2,18 (3,24)
1shg	57	β	4,51	6,30	-73	-72 (-61)	3,21 (5,59)
1ubq	76	α/β	3,44	4,78	-95	-95 (-82)	1,86 (4,97)
2acy	98	α/β	3,44	6,28	-130	-117 (-79)	3,41 (11,14)
2bby	69	α	2,65	4,92	-65	-65 (-64)	2,04 (4,98)
2ci2	65	α/β	3,54	4,08	-70	-70 (-56)	1,51 (4,00)
2ife	91	α/β	5,17	5,82	-112	-111 (-96)	2,68 (5,90)
6pti	56	*few*	4,02	6,12	-53	-53 (-44)	1,93 (6,12)
1hlb	157	α	2,95	5,90	-154	-152 (-90)	4,70 (15,10)
1pdo	129	α/β	2,95	4,72	-166	-166 (-160)	2,45 (3,90)
1tfe	142	α/β	2,41	3,98	-183	-183 (-159)	2,62 (6,40)
1timA	247	α/β	3,65	6,75	-263	-243 (-190)	4,84 (12,5)
2lzm	164	α	2,59	4,57	-204	-204 (-148)	2,56 (10,2)
3chy	128	α/β	2,73	4,40	-171	-171 (-157)	2,97 (4,10)

Tab. 4.1: Performance de l'algorithme glouton sur des trajectoires floues avec un critère énergétique de type Go. Pour chaque cible (colonne PDB), sont présentées : sa taille (L), la classe CATH (Pearl et al., 2003) à laquelle elle appartient (CATH), la complexité de la trajectoire optimale (C_o) et de la trajectoire floue (C_f) associées, l'énergie de référence (E_n) calculée sur la structure native, l'énergie du meilleur modèle (E_m) et le cRMSd (cRMSd) du meilleur modèle généré par l'algorithme glouton par rapport à la structure native. Les valeurs entre parenthèses correspondent aux résultats obtenus avec la version basique de l'algorithme. Table extraite de Tuffery et al. (2005).

Les résultats obtenus avec la version améliorée de l'algorithme glouton, guidé par le potentiel de Go, avec des trajectoires floues pour les 16 cibles précédemment citées, sont présentés dans la table 4.1. Du point de vue des énergies du potentiel, les modèles ont des énergies (E_m) très proches de l'énergie de référence (E_n), *i.e.* de la structure native, montrant que l'algorithme a été capable d'identifier la majorité des contacts présents dans la structure de référence. Le cRMSd des modèles reconstruits s'échelonne entre 1,5 et 4.8 Å avec cette version de l'algorithme, et entre 3,24 et 15,1 Å avec la version basique. Ce résultat, clairement significatif, met en évidence le gain apporté par les différentes améliorations de l'algorithme de base (Tuffery et al., 2005), et l'avantage d'utiliser un alphabet structural dérivé de chaînes de Markov cachées. Par ailleurs, les cibles 2acy et 1timA sont des cibles très difficiles pour un tel algorithme guidé par un critère de type Go, car ces dernières sont stabilisées par des interactions de très longue portée. Dans le cas de 2acy, le brin β1 (L6-K16) est stabilisé au sein de la structure en interagissant avec les brins β3 (V47-P54) et 4 (H74-V85), séparés de 40 et 70 acides aminés respectivement. Pour 1timA, les deux extrémités séparées de plus de 200 acides aminés sont en contact.

4.2.2 A partir de la définition des structures secondaires

Les trajectoires floues précedemment décrites requièrent une connaissance de la structure native. Ainsi, dans une optique *ab initio*, si l'on va un peu plus loin dans la complexité, et que l'on ne fournit que la seule information des structures secondaires à l'algorithme glouton, guidé par le critère de type Go, est-il capable de générer des modèles de topologie native ?

PDB	**L**	**CATH**	C_{s2i}	C_{s2f}	C_{s2fM}	E_n	E_m	**cRMSd**
1csp	67	β	18,29	10,02	8,52	-78	-72 (-63)	3,07 (6,16)
1ctf	68	α/β	7,94	5,43	4,60	-79	-71 (-58)	3,34 (4,83)
1hlb	157	α	11,11	7,34	6,08	-154	-123 (n.d.)	6,04 (n.d.)
1lea	72	α	28,00	14,55	12,32	-91	-90 (-87)	4,77 (5,35)
1mba	146	α	5,60	4,68	4,11	-164	-142 (-125)	3,86 (5,15)
1pdo	129	α/β	8,21	6,07	5,28	-166	-155 (-134)	3,61 (4,87)
1shg	57	β	20,00	11,32	9,34	-73	-72 (-56)	2,72 (4,65)
1tfe	142	α/β	7,22	5,30	4,43	-183	-178 (n.d.)	4,44 (n.d.)
1tfi	50	β	26,10	12,78	10,47	-50	-47 (-43)	4,99 (5,32)
1ubq	76	α/β	11,51	8,20	7,16	-95	-88 (-78)	3,66 (5,00)
2acy	98	α/β	14,15	9,17	7,57	-130	-120 (-103)	4,33 (6,17)
2bby	69	α	8,26	6,29	5,49	-65	-63 (-64)	3,00 (3,00)
2ci2	65	α/β	20,25	10,54	8,78	-70	-67 (-49)	3,81 (4,79)
2ezh	65	α	6,34	5,26	4,54	-44	-44 (-44)	3,32 (3,78)
2ife	91	α/β	19,03	10,31	8,54	-112	-107 (-94)	4,27 (4,51)
2lzm	164	α	9,09	6,35	5,49	-152	-124 (n.d.)	5,47 (n.d.)
2mtaC	147	α	27,37	12,64	9,82	-204	-189 (-187)	5,50 (4,96)
2ptl	78	α/β	18,80	8,69	7,69	-64	-61 (-56)	5,33 (5,88)
3chy	128	α/β	8,59	6,65	5,51	-171	-169 (-168)	3,80 (3,25)
5fdl	106	α/β	31,32	12,80	10,67	-145	-137 (-130)	4,27 (5,04)

Tab. 4.2: Performance de l'algorithme glouton à partir de la définition des structures secondaires, avec un critère énergétique de type Go. Pour chaque cible (colonne PDB), sont présentées : sa taille (L), la classe CATH (Pearl et al., 2003) à laquelle elle appartient (CATH), la complexité de la trajectoire initiale définie par les structures secondaires (C_{s2i}), préfiltrée par Greedy (C_{s2f}), et, par Greedy et la matrice de transition markovienne (C_{s2fM}), l'énergie de référence (E_n) calculée sur la structure native, l'énergie du meilleur modèle (E_m) et le cRMSd (cRMSd) du meilleur modèle généré par l'algorithme glouton par rapport à la structure native. Ces valeurs ont été déterminées à partir des trajectoires $s2fM$. Les valeurs entre parenthèses correspondent aux résultats obtenus pour les trajectoires $s2f$. Les trois longues chaînes 1hlb, 1tfe et 2mtaC n'ont pas été déterminées (n.d.) en utilisant les trajectoires $s2f$. Table extraite de Tuffery and Derreumaux (2005).

C'est la question à laquelle ont voulu répondre Tuffery and Derreumaux (2005) dans la suite de leurs travaux sur l'algorithme glouton déjà présenté. Dans ce cas, les trajectoires des protéines du jeu de validation sont définies comme suit : (i) pour chaque acide aminé, encodé par une lettre décrivant une structure secondaire dans la trajectoire optimale

HMM-SA, sont sélectionnées toutes les lettres de l'alphabet structural correspondant à cet état (voir (Camproux et al., 2004)), et (ii) pour les autres acides aminés, les 27 lettres de l'alphabet sont retenues. Ainsi, pour chaque position décrite comme une hélice, les lettres `[a,A,B,V,W]` sont utilisées, et pour les brins β, les lettres `[L,M,N,T,X]` sont choisies. Ces trajectoires brutes, notées C_{s2i} dans la table 4.2, sont ensuite raffinées, par les préfiltres de l'algorithme glouton, puis par l'intermédiaire de la matrice de probabilité de transition markovienne permettant, ainsi, de réduire la complexité des trajectoires.

Comme nous pouvons le constater dans la table 4.2, les trajectoires initiales sont d'une complexité importante, lorsqu'on les compare aux trajectoires floues de la table 4.1. Cependant, cette complexité est grandement diminuée par l'utilisation des préfiltres de l'algorithme glouton d'une part, et par les transitions markoviennes d'autre part. Si l'on ne prend pas en compte les cibles possédant une longue chaîne (1hlb, 1tfe et 2mtaC), le cRMSd moyen des meilleurs modèles générés (en terme de cRMSd) est de 3,98 Å pour les trajectoires $s2fM$ contre 4,87 Å pour les trajectoires $s2f$. Par ailleurs, les modèles d'énergie la plus basse générés à partir des trajectoires $s2f$ pour sept des cibles (1csp, 1ctf, 1shg, 1tfi, 1ubq, 2acy et 2ptl) ont des topologies non natives (Tuffery and Derreumaux, 2005), alors que les modèles d'énergie la plus basse, générés à partir des trajectoires $s2fM$ ont tous des topologies natives.

Dans ces deux premères études, Tufféry et Derreumaux ont montré que l'algorithme glouton était capable de générer des modèles ayant une conformation proche de la structure native. Différentes fonctions objectifs ont été utilisées pour se détacher de plus en plus de connaissances *a priori* sur la structure à reconstruire. Le critère cRMSd par un ensemble d'approximations locales nous permet d'obtenir une bonne approximation globale des structures. Un critère de type potentiel de Go, nous permet de générer des modèles de topologie native, même à partir de la seule information des structures secondaires.

Cependant, dans une optique de prédiction *ab initio*, nous avions besoin d'un critère objectif ne nécessitant aucune connaissance *a priori* sur la structure à prédire, si ce n'est sa seule séquence en acides aminés. Ainsi, nous avons implémenté une version simplifiée du potentiel gros grain OPEP dans la méthode de reconstruction. Après une brève présentation des différents potentiels gros grains utilisés à l'heure actuelle, nous allons maintenant présenter la formulation originale de ce champ de force.

Chapitre 5

Le potentiel gros grain OPEP

Un des ingrédients essentiels pour prédire la structure tertiaire d'une protéine, à partir de sa seule séquence en acides aminés, est une fonction d'énergie capable d'identifier l'état natif d'une protéine comme étant le minimum global de l'énergie libre. Malheureusement après 30 ans d'efforts, une telle fonction n'a pas encore été trouvée (Buchete et al., 2004a). Les approches développées en ce sens sont de deux types : les potentiels physiques d'un côté, et les potentiels statistiques de l'autre.

Les méthodes physiques se basent sur des calculs de mécanique quantique décrivant les forces s'appliquant aux particules. Les modèles protéiques, dans ce type d'approches, sont détaillés au niveau atomique. De ce fait, ils contiennent des termes énergétiques associés aux longueurs de liaisons, aux angles de valence, aux angles de torsion (voir la figure 5.1), et aux interactions électrostatiques et de Van der Waals (Mackerell, 2004). Les champs de force les plus réputés dans cette catégorie sont CHARMM (Brooks et al., 1983; Mackerell et al., 1998), AMBER (Case et al., 2005), OPLS (Jorgensen and Tirado-Rives, 1988; Jorgensen et al., 1996) et GROMOS (Christen et al., 2005). Bien que les ressources de calculs aient grandement augmenté durant les vingt dernières années, elles sont toujours un facteur limitant pour avoir des temps de simulations supérieurs à la microseconde. Ainsi, même ces potentiels physiques ont recours à des modèles pour accroître ce temps de simulation. Par exemple, lorsque l'on simule le comportement d'une protéine dans une boîte d'eau, la majeure partie du temps de calcul est consacrée au solvant lui même et non à la protéine d'intérêt. Pour pallier cette contrainte, le solvant peut être représenté implicitement, par des approches de type Born généralisé (Feig et al., 2004; Pokala and Handel, 2005). Une autre astuce peut être de réduire le nombre d'atomes du système en fusionnant les groupements méthyl en une seule entité, nous parlons alors, dans ce cas précis, de champ de force *united atom* (Yang et al., 2006) et non tous atomes.

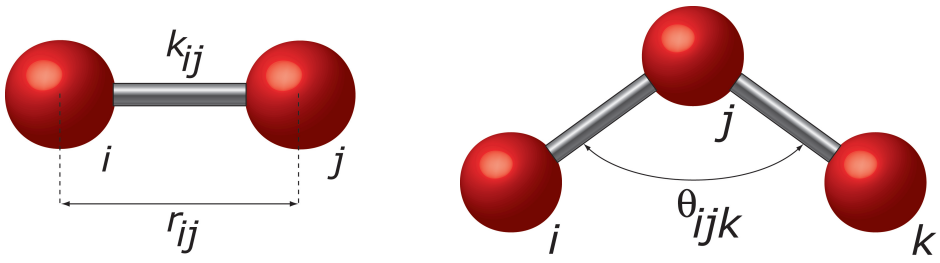
Malgré les approximations qui peuvent être faîtes pour augmenter les temps de simulations, de tels potentiels sont difficilement utilisables dans le cadre du repliement d'une

Potentiel de liaison

$$E_L = k_{ij}(r_{ij} - r_0)^2$$

Potentiel d'angle

$$E_A = k^H_{ijk}(\theta_{ijk} - \theta_0)^2$$



Potentiels de torsion

Angle dièdre propre

$$E_{DP} = k^P_{ijkl}(\theta^P_{ijkl} - \theta^P_0)^2$$

Angle dièdre impropre

$$E_{DI} = k^I_{ijkl}(\theta^I_{ijkl} - \theta^I_0)^2$$

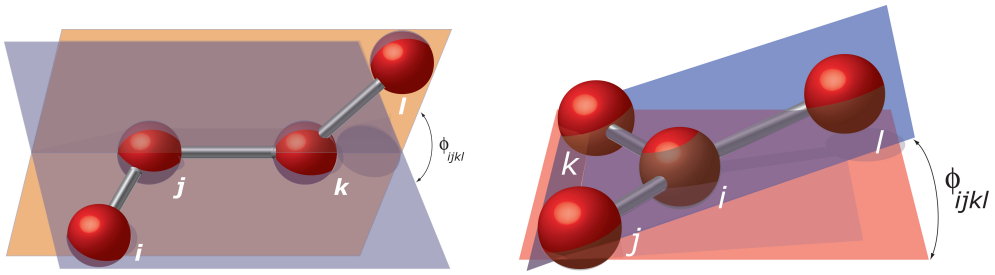


Fig. 5.1: Potentiels énergétiques liés à la géométrie des molécules. Voici quatre exemples de potentiels harmoniques que l'on peut retrouver dans les potentiels physiques, visant à rapprocher une mesure atomique (un angle ou une longueur de liaison) de sa valeur canonique.

protéine, processus de l'ordre de la milli-seconde à la seconde (Anfinsen and Scheraga, 1975). Il est donc nécessaire de lisser le paysage énergétique des protéines en simplifiant leur représentation (Kolinski and Skolnick, 1996; Kolinski, 2004) et donc, les modèles énergétiques qui en découlent (Buchete et al., 2004a; Tozzini, 2005; Skolnick, 2006).

Le nombre de structures résolues expérimentalement disponibles dans la *Protein Data Bank* (Berman et al., 2000b) couvrant la majeure partie des repliements théoriquement observables (Zhang and Skolnick, 2005a), un certain nombre de potentiels statistiques ont pu être mis en place. Ces potentiels visent à tirer un maximum d'informations pertinentes sur un sous-ensemble statistiquement représentatif de structures. Paradoxalement, ce sont ces champs de force empiriques qui semblent les plus performants pour ce qui est de la prédiction de la structure des protéines (Zhang et al., 2005; Bradley et al., 2005). La plupart de ces potentiels statistiques utilisent une représentation simplifiée des protéines, on parle de modèles gros grain ou *coarse grained*. Chacun des grains du modèle est basé sur une représentation de type *united atom* : une sphère représente ainsi un groupement

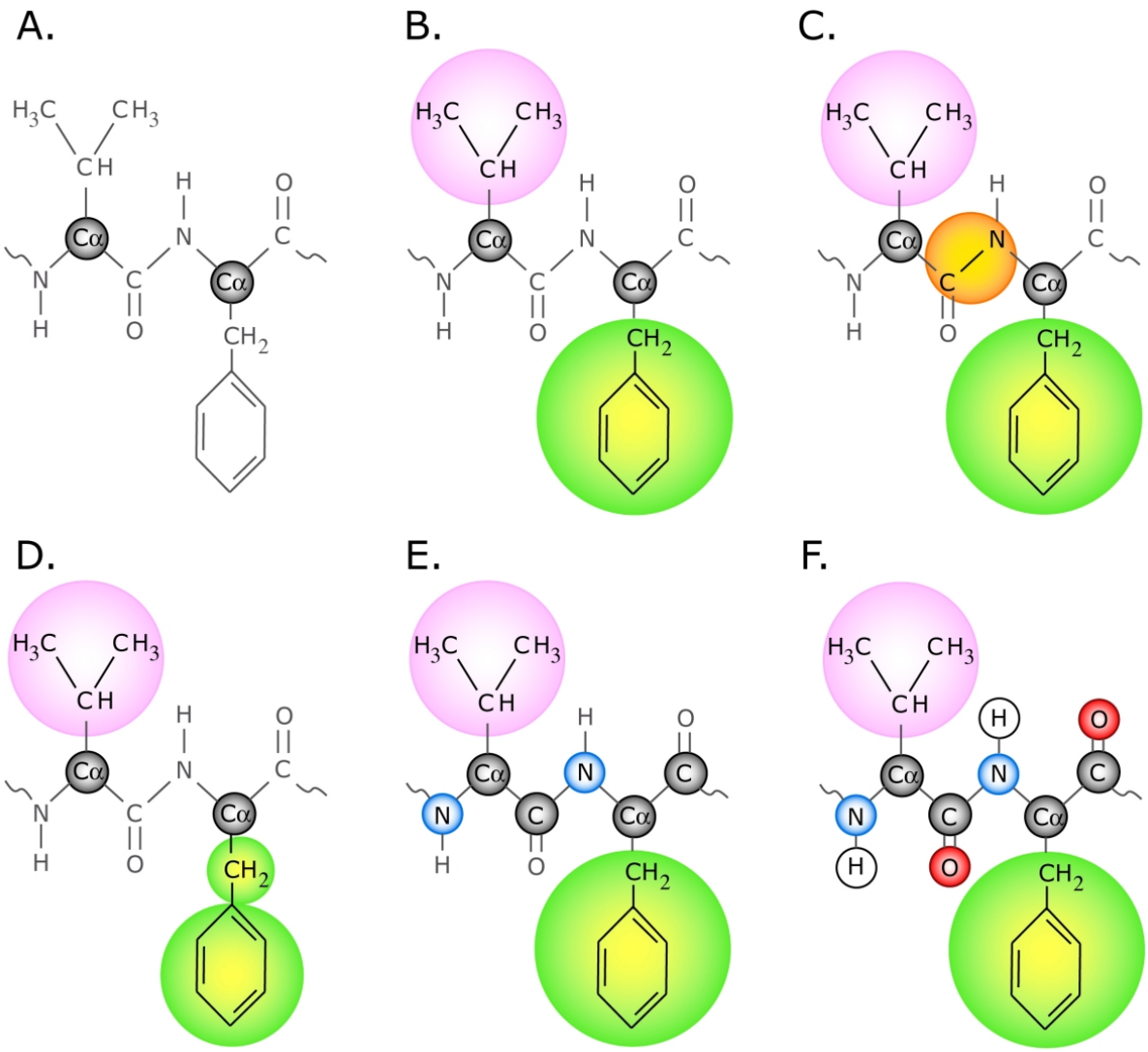


Fig. 5.2: Les différents modèles gros grain. Il exite différents types de modèles gros grains selon le nombre de sphères utilisées pour simplifier la structure des proteines. **A.** une sphère : C_α ; **B.** deux sphères : C_α et la chaîne latérale ; **C.** trois sphères : C_α, la chaîne latérale et la liaison peptidique ; **D.** trois sphères : C_α et la chaîne latérale à une ou deux sphères selon sa taille ; **E.** quatre sphères : N, C_α, C et la chaîne latérale ; **F.** six sphères : HN, N, C_α, C, O et la chaîne latérale.

d'atomes. Ces modèles peuvent être classés selon leur degré de "granularité" (allant de une à six sphères), directement proportionnel à la rugosité du paysage énergétique qu'ils décrivent (voir la figure 5.2).

Historiquement les premiers modèles gros grain ont fait leur apparition dans les années 70 avec les travaux de Go (Udea et al., 1978) et Levitt (Levitt, 1976). Dans le modèle de Go original, les protéines sont représentées comme un collier de perles, chaque perle étant un résidu. Le potentiel associé est définit par un ensemble d'interactions non liantes attractives ou répulsives entre les résidus, à partir de leur définition dans la structure

native. Ce potentiel a été développé pour simuler le repliement des protéines. Il donne des résultats satisfaisants, car le biais introduit façonne le paysage énergétique en entonnoir, la conformation de plus basse énergie étant la structure native. La plupart des modèles à une sphère (généralement le carbone α) sont des évolutions de ce modèle de Go, nécessitant une paramétrisation spécifique ou une connaissance *a priori* de la structure native (Brown et al., 2003). Ceci est dû à la difficulté de prendre en compte, dans un nombre limité de paramètres, l'effet de la taille de l'acide aminé, de sa géométrie et de sa conformation. Certains modèles fortement anisotropes sont cependant développés pour combler cette lacune (Bahar and Jernigan, 1997; Buchete et al., 2004b).

Dans les modèles gros grain à deux sphères, la deuxième sphère, qui représente en général la chaîne latérale, permet d'ajouter une spécificité aux interactions locales (Sippl, 1990; Miyazawa and Jernigan, 1985). Bahar and Jernigan (1997) ont développé un tel potentiel paramétré sur un ensemble de structures expérimentales (Reith et al., 2003), potentiel qui a permis d'évaluer l'affinité d'un ensemble de substrats de la protéase de HIV-1 (Kurt et al., 2003).

Deux types de modèles à trois sphères ont été développés. Dans le premier cas, la troisième sphère peut être ajoutée au modèle pour représenter le centre de masse de la liaison peptidique. Ce type de modèle a été utilisé dans des champs d'application divers tels que la reconnaissance de repliement (Buchete et al., 2004a), la simulation de repliements dans l'eau (Oldziej et al., 2005) ou dans un environnement lipidique (Shih et al., 2006), la liaison (Zhang et al., 2004) et l'assemblage (Bond and Sansom, 2006; Bond et al., 2007) de protéines membranaires. Pour le deuxième type de modèle gros grain à trois sphères, les longues chaînes latérales sont représentées par deux sphères au lieu d'une seule (Srinivasan et al., 2004; Wallqvist and Ullner, 1994; Zacharias, 2003), permettant d'introduire une certaine flexibilité des chaînes latérales, notamment pour des problématiques d'amarrage moléculaire.

Il existe une dernière catégorie de modèle gros grain décrit par quatre ou six sphères. Dans le modèle à quatre sphères, les atomes de la chaîne principale N, C_α, C' sont représentés explicitement, tandis que la chaîne latérale est représentée par une seule sphère, et, dans le modèle à six sphères, les atomes HN et O de la chaîne principale sont ajoutés. Ce type de modèle a l'avantage non négligeable de permettre une représentation explicite des liaisons hydrogène. Il a été utilisé dans le cadre de l'agrégation de peptides β-amyloïdes (Urbanc et al., 2004; Favrin et al., 2004; Santini et al., 2004b,a), mais aussi pour la prédiction de structure et le repliement de petites protéines (Derreumaux, 1999; Fujitsuka et al., 2004; Colubri, 2004). Devant un tel niveau de description, le terme gros grain peut sembler impropre, le seul gros grain étant finalement la chaîne latérale.

Dans le travail qui suit, nous avons mis à jour la fonction d'énergie OPEP. Ce potentiel gros grain utilise six sphères pour représenter un acide aminé, ce qui permet de décrire très précisément la structure de la chaîne principale, et a permis de discriminer des structures natives de peptides de modèles *ab initio* (Derreumaux, 1999). Couplé à des simulations de Monte-Carlo , OPEP a permis de prédire la structure native des 46 résidus du faisceau de trois hélices de la protéine A (Derreumaux, 2000), des épingles à cheveux d'hélices α (Forcellino and Derreumaux, 2001), et des 56 résidus du domaine B1 de la protéine G (Derreumaux, 2002), avec une déviation moyenne (cRMSd) de moins de 3 Å par rapport à la structure expérimentale. OPEP a aussi été grandement utilisé pour étudier le repliement de peptides et leur agrégation. En utilisant la technique d'activation-relaxation (ART) (Malek and Mousseau, 2000; Wei et al., 2002), les chemins de repliement explorés pour un modèle d'épingle à cheveux β ont été en accord avec les résultats obtenus par simulation de dynamique moléculaire (DM) ou de Monte-Carlo (MC) (Wei et al., 2004). De plus, les simulations ART-OPEP de l'agrégation du peptide amyloïde ont permis de mettre en évidence des mouvements de reptation de la chaîne, mouvements qui sont en accord avec des données expérimentales de spectroscopie infrarouge (Mousseau and Derreumaux, 2005).

Bien que cet ensemble de succès montre la pertinence du champ de force OPEP, deux types de données nous indiquent que les paramètres d'OPEP ne semblent pas optimaux. Premièrement, des simulations de MC ont suggéré que la super-structure sedondaire $\beta\alpha\beta$, dans laquelle deux brins β parallèles sont joints par une hélice α pouvait être une unité de repliement à elle seule (Derreumaux, 2000), mais ces données n'ont été confirmées ni par des expériences de dichroïsme circulaire, ni par des expériences de résonance magnétique nucléaire (RMN) (Coincon et al., 2005). Deuxièmement, les paramètres des chaînes latérales n'ont pas été appris sur des distributions réelles de probabilités de distances entre paires de résidus, et, par ailleurs, ces paires n'étaient alors que faiblement peuplées (Derreumaux, 1999).

5.1 Le modèle gros grain

Le champ de force OPEP n'a que peu évolué depuis sa version originale (Derreumaux, 1999). Les atomes de la chaîne principale (N, HN, $C\alpha$, C', O) sont représentés explicitement, tandis que les chaînes latérales sont assimilées à une sphère de position et de rayon variable selon le résidu considéré (voir figure 5.3).

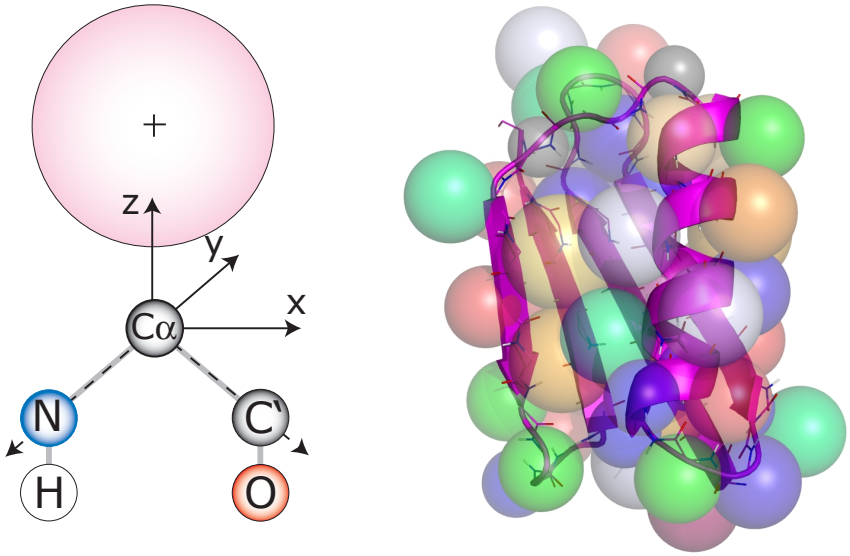


Fig. 5.3: OPEP : un modèle gros grain. Chaque résidu est représenté par six sphères : cinq sphères pour les atomes de la chaîne principale, et une sphère pour représenter la chaîne latérale. Dans la partie droite de la figure, est représentée une protéine complète dans le modèle OPEP (code PDB 1pgb).

5.2 La fonction d'énergie

La formulation des termes énergétiques d'OPEP pour modéliser les interactions de courte et de grande portée dans cette version 3 sont les mêmes que dans les versions 1 (Derreumaux, 1999, 2000), et 2 (Forcellino and Derreumaux, 2001). L'approche originale d'OPEP tient dans le traitement coopératif à quatre corps des liaisons hydrogène. Ce terme vise à reproduire la forte coopérativité des liaisons hydrogène de type amide, comme il a été démontré par des calculs de mécanique quantique (Guo and Salahub, 1998). Par ailleurs, dans ce terme, est pris en compte la propensité des résidus à se retrouver dans une structure secondaire de type α ou β, paramètre crucial dans le *protein design* ou la prédiction du taux d'agrégation des protéines formant des fibres amyloïdes (Chiti et al., 2003). Les formes analytiques des potentiels de liaisons hydrogène et de torsion de la chaîne principale ont été modifiées pour les rendre dérivables, et ont été utilisées à maintes reprises dans le cadre d'études de repliement et d'agrégation protéique (Wei et al., 2004; Mousseau and Derreumaux, 2005; Wei et al., 2003; Santini et al., 2003). Dans la version 3 qui est présentée ici, seuls les paramètres d'OPEP sont changés.

La formulation générale d'OPEP est exprimée comme la somme de l'énergie locale, non liée et associée aux liaisons hydrogène :

$$E = E_{locale} + E_{non-liante} + E_{liaisons-H} \tag{5.1}$$

5.2.1 L'énergie locale

Le calcul de l'énergie locale s'exprime comme :

$$\begin{aligned} E_{locale} &= w_l \sum_{liaison} K_b(r - r_{eq})^2 + w_a \sum_{angles} K_\alpha(\alpha - \alpha_{eq})^2 \\ &\quad + w_\Omega \sum_{torsions-imp} k_\Omega(\Omega - \Omega_{eq})^2 + \ w_{\phi,\psi} \ (\sum_{\phi} E_\phi + \sum_{\psi} E_\psi) \end{aligned} \tag{5.2}$$

Le terme E_{locale} contient des constantes de forces associées à un changement des longueurs et angles de liaison de toutes les particules ainsi que des constantes de force en relation avec un changement des torsions impropres des chaînes latérales et de la liaison peptidique. Les constantes de forces et les valeurs d'équilibre associées aux atomes de la chaîne principale sont issues d'AMBER (Case et al., 2005) ; les constantes de force associées aux chaînes latérales sont très similaires à celles d'AMBER. Les termes E_ϕ et E_ψ reproduisent le diagramme de Ramachandran dans le modèle gros grain, en accord avec les structures protéiques complètes ; ils sont exprimés par les polynômes quadratiques suivants :

$$E_\phi = k_{\phi\psi}(\phi - \phi_o)^2 \tag{5.3}$$

$$E_\psi = k_{\phi\psi}(\psi - \psi_o)^2 \tag{5.4}$$

où $\phi_0 = \phi$ dans l'interval $[\phi_{inf}, \phi sup]$, sinon, $\phi_0 = min(\phi - \phi_{inf}, \phi - \phi_{sup})$, avec $\phi_{inf} = -160^\circ$, et $\phi_{sup} = -60^\circ$. De la même manière, nous utilisons $\psi_{inf} = -60^\circ$, et $\psi_{sup} = -160^\circ$ dans l'équation 5.4. Il est important de noter qu'une telle formulation n'empêche pas d'explorer des conformations couvrant l'ensemble des valeurs de ϕ et de ψ.

5.2.2 Les interactions non liantes

L'énergie associée aux interactions non liantes s'exprime par :

$$\begin{aligned} E_{non-liante} &= w_{1,4} \sum_{1,4} E_{VdW} + w_{C\alpha,C\alpha} \sum_{C\alpha,C\alpha} E_{VdW} + \ w_{1>4} \sum_{CP',CP'} E_{VdW} \\ &\quad + w_{1>4} \sum_{CP',C\alpha} E_{VdW} + w_{1>4} \sum_{CP,CL} E_{VdW} + \sum_{CL,CL} w_{CL,CL} \ E_{VdW} \end{aligned} \tag{5.5}$$

avec $1,4$, les interactions de type 1-4 pour chaque degré de liberté le long des angles de torsion, CP' les atomes N, HN, C' et O le long de la chaîne principale ($CP = CP' + C\alpha$), et CL les chaînes latérales. Les interactions de courte portée sont séparées des interactions de longue portée ($j > i+4$), et le $C\alpha$ est traité séparément des autres atomes de la chaîne principale.

Les interactions de type Van der Walls (E_{VdW}) sont définies comme :

$$E_{VdW} = \begin{cases} -\epsilon_{ij}\left(\frac{r^0_{ij}}{r_{ij}}\right)^6, \text{ pour } \epsilon_{ij} < 0, \\ \epsilon_{ij}\left(\left(\frac{r^0_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r^0_{ij}}{r_{ij}}\right)^6\right) \text{ sinon.} \end{cases} \tag{5.6}$$

avec r_{ij} la distance entre les particules i et j, $r^0_{ij} = r^0_i + r^0_j$, où r^0_i est le rayon de Van Der Waals de la particule i. Les rayons utilisés pour les chaînes latérales sont donnés dans la table 8.1. Pour les atomes de la chaîne principale, les rayons suivants sont utilisés : $r^0_N = 1,75$ Å, $r^0_H = 1,00$ Å, $r^0_{C\alpha} = 2,385$ Å, $r^0_{C'} = 1,85$ Å et $r^0_O = 1,60$ Å. Toutes les interactions non liantes sont considérées, sans distance seuil (*cut-off*).
Dans cette 3ème formulation de OPEP, un potentiel de type 6-12 est utilisé pour toutes les interactions, sauf pour les chaînes latérales. Pour ces dernières, si l'interaction a un caractère hydrophobe ou résulte de deux résidus de charge opposée le potentiel 6-12 est utilisé ; sinon, seul le potentiel répulsif -6 est utilisé. Les valeurs initiales des ϵ_{ij} entre les couples de particules (CP', CP'), (CP, CL), $(CP', C\alpha)$, et $(C\alpha, CL)$ sont initialisées à 5.10^{-3} kCal/mol, et $0,4$ kCal/mol pour les interactions $(C\alpha, C\alpha)$. Les valeurs de ϵ_{ij} pour les chaînes latérales sont issues des travaux de Betancourt et Thirumalai (Betancourt and Thirumalai, 1999) qui ont raffiné la matrice de contact préalablement décrite par Skolnick et al. sur la base de 224 protéines homologues (Skolnick et al., 1997). Sur un total de 210 interactions, 93 paires sont gouvernées par un potentiel répulsif -6.

5.2.3 Les liaisons hydrogène

L'énergie associée aux liaisons hydrogène ($E_{liaisonsH}$) est composée de deux termes : un terme classique à deux corps (E_{LH1}), et un terme plus original à quatre corps (E_{LH2}). Le terme à deux corps est défini par :

$$E_{LH1} = w_{lh1-4} \sum_{ij,j=i+4} \varepsilon_{lh1-4}\ \mu(r_{ij})\nu(\alpha_{ij}) + w_{lh1>4} \sum_{ij,j>i+4} \varepsilon_{lh1>4}\ \mu(r_{ij})\nu(\alpha_{ij}) \tag{5.7}$$

où :

$$\mu(r_{ij}) = 5\ (\frac{\sigma}{r_{ij}})^{12} - 6\ (\frac{\sigma}{r_{ij}})^{10} \tag{5.8}$$

$$\nu(\alpha_{ij}) = \begin{cases} \cos^2\alpha_{ij}\ ,\ \alpha_{ij} > 90° \\ 0,\ sinon \end{cases} \tag{5.9}$$

La somme est effectuée sur tous les résidus i et j séparés par $j = i + 4$ et $j > i + 4$ (les hélices 3_{10} sont donc exclues), r_{ij} est la distance $O \cdots H$ (l'oxygène du groupement carbonyle et l'hydrogène du groupement amide), α_{ij} l'angle $\widehat{NHO}$ et σ (initialisé à $1,8$ Å), la valeur d'équilibre de la distance $O \cdots H$. Les paramètres ε_{lh1-4} et $\varepsilon_{lh1>4}$ sont initialisés à $1,0$ kCal/mol. Dans ce cas aussi, nous distinguons les énergies des liaisons hydrogène de courte (ε_{lh1-4}) et de longue portée ($\varepsilon_{lh1>4}$), car les distances d'équilibre $C\alpha \cdots C\alpha$ varient

pour les hélices α (6,1 Å pour les interactions 1-4), et les feuillets β (4,5 Å).

L'effet à quatre corps mimant l'effet coopératif entre les liaisons hydrogène ij et kl est défini par :

$$\begin{aligned} E_{LH2} &= \sum \varepsilon_{\alpha}^{coop} \exp(-(r_{ij}-\sigma)^2/2)\ \exp(-(r_{kl}-\sigma)^2/2)\ \Delta(ijkl) \\ &\quad + \sum \varepsilon_{\beta}^{coop} \exp(-(r_{ij}-\sigma)^2/2)\ \exp(-(r_{kl}-\sigma)^2/2)\ \Delta'(ijkl) \end{aligned} \tag{5.10}$$

Le paramètre $\Delta(ijkl)$ est initialisé à 1 si les résidus $(k,\, l) = (i+1,\, j+1)$ et $(j = i+4,\, l = k+4)$, sinon $\Delta(ijkl) = 0$. Ainsi, les hélices Π ne sont pas stabilisées. Le paramètre $\Delta'(ijkl) = 1$ si k et l satisfassent l'une des conditions suivantes : soit $(k,\, l) = (i+2,\, j-2)$, soit $(i+2,\, j+2)$; sinon $\Delta'(ijkl) = 0$. L'ensemble de ces conditions permet de stabiliser les hélices α et les feuillets β parallèles et anti-parallèles, indépendamment des angles dièdres ϕ et ψ, mais aussi n'importe quelle region satisfaisant les conditions imposées pour $ijkl$. Les paramètres $\varepsilon_{\alpha}^{coop}$ et $\varepsilon_{\beta}^{coop}$ sont exprimés par :

$$\varepsilon_{\alpha}^{coop} = w_{\alpha}^{coop} E_{\alpha}^{coop} + \sum_{R} w_{\alpha}^{R} E_{\alpha}^{R} \tag{5.11}$$

$$\varepsilon_{\beta}^{coop} = w_{\beta}^{coop} E_{\beta}^{coop} + \sum_{R} w_{\beta}^{R} E_{\beta}^{R} \tag{5.12}$$

E_{α}^{coop} et E_{β}^{coop} sont les énergies coopératives, indépendamment des acides aminés impliqués, et E_{α}^{R} et E_{β}^{R} sont les potentiels associés aux propensités du résidu R pour les hélices α et les feuillets β respectivement, comme il a été défini dans la version 1.0 d'OPEP (Derreumaux, 1999, 2000). La somme est sur 4 résidus pour la première liaison hydrogène au sein d'une hélice, puis sur un résidu pour chaque liaison hydrogène supplémentaire, et la somme est généralement sur 2 résidus impliqués dans une liaison hydrogène au sein d'un feuillet. Il a été vérifié que chaque type de résidu n'est compté qu'une fois au sein d'une longue hélice ou d'un long feuillet.

Vu dans son ensemble, le potentiel OPEP est exprimé comme une fonction de 261 poids : 213 poids pour $E_{non-liante}$, dont 210 pour $E_{CL,CL}$; 4 poids pour E_{locale} ; 4 pour les liaisons hydrogène et 40 pour les propensités des résidus à être impliqués dans des structures secondaires de type α ou β. Ce nombre de paramètres à optimiser est beaucoup plus important que pour la version 2 d'OPEP pour laquelle, il n'y avait que 47 paramètres. Dans la suite de ce manuscrit, la version 3.0 d'OPEP fait référence à la formulation courante d'OPEP non optimisée (tous les poids étant à 1,0), incluant les paramètres E_{α}^{R} et E_{β}^{R}. OPEP v3.1 est cette même version optimisée, tandis que la version 3.2 a ses paramètres E_{α}^{R} et E_{β}^{R} initialisés à zéro.

Deuxième partie

Prédiction de la structure locale des protéines

Chapitre 6

SAFrAN : une méthode de recherche de fragments candidats

Beaucoup d'efforts ont été investis pour la description de la structure locale des protéines et leur prédiction.

Nous pouvons considérer que les premières méthodes de prédiction de la structure locale des protéines sont les méthodes de prédiction des structures secondaires. Ces méthodes ont beaucoup progressé depuis les années 70 pour atteindre maintenant des valeurs de Q_3 supérieures à 75 %. Pour rappel, le Q_n représente la qualité de la prédiction (% de bonne prédiction) à n états. Parmi les méthodes les plus performantes, nous pouvons citer PHD (Rost and Sander, 1994), PSIPRED (Jones, 1999b) ou SSPRO (Baldi et al., 1999; Pollastri et al., 2002), toutes trois basées sur des réseaux de neurones, GOR (Garnier et al., 1978; Gibrat et al., 1987; Garnier et al., 1996) une méthode dérivée de la théorie de l'information, des méthodes basées sur des chaînes de Markov cachées (Thorne et al., 1996; Lio et al., 1998; Crooks and Brenner, 2004; Martin et al., 2006) et, plus récemment, des méthodes basées sur des SVM (*Support Vector Machine*) (Kim and Park, 2003; Ward et al., 2003) sont apparues.

En complément des structures secondaires, un certain nombre d'efforts ont été entrepris pour prédire la structure des boucles protéiques (Donate et al., 1996; Oliva et al., 1997; Wojcik et al., 1999; Michalsky et al., 2003; Espadaler et al., 2004; Fernandez-Fuentes et al., 2005, 2006b).

Pour échapper à la description classique à trois états, différentes méthodes génériques ont été décrites. Bystroff and Baker (1998) ont proposé les I-sites, un ensemble de conformations récurrentes au sein des structures protéiques. Ces fragments de 3 et 9 résidus de long possèdent une signature de séquence suffisante pour permettre la prédiction de conformations préférentielles de certaines régions. Pour améliorer les performances de prédiction, les I-Sites ont ensuite été inclus dans un modèle de prédiction basé sur des chaînes de Markov cachées : HMMSTR (Bystroff et al., 2000). Ce modèle de Markov permet de décrire les transitions compatibles entre les motifs séquence-structure constituant les I-

sites. A partir des 28 centroïdes de leur alphabet structural (Hunter and Subramaniam, 2003b), Hunter and Subramaniam (2003a) ont défini une approche de prédiction de type bayesienne après avoir évalué la probabilité d'apparition de chaque type d'acide aminé à chaque position de leur fragments canoniques. de Brevern et al. (2000) ont mis en place une méthode de prédiction baysienne de leur 16 *Protein Blocks* (PB) (de Brevern et al., 2000) dont la performance a été améliorée (i) en élargissant la banque d'apprentissage de la relation séquence-structure et en s'aidant de méthode de prédiction de la structure secondaire (Etchebest et al., 2005), et en prennant en compte la dépendance des transitions entre les PB (de Brevern et al., 2007). En parallèle, Benros et al. (2006) ont développé une méthode de prédiction de fragments longs basée sur le modèle de la protéine hybride (de Brevern and Hazout, 2003), une méthode de classification non supervisée. Yang and Wang (2003) ont défini une méthode consensus de prédiction de fragments de 9 résidus basée sur quatre états définis par des régions clés du diagramme de Ramachandran. Sander et al. (2006) ont quant à eux mis en place une stratégie hybride de classification prenant en compte à la fois l'information de séquence et de structure pendant le partitionnement.

Nous avons exploré une nouvelle approche sur la base de chaînes de Markov cachées. A partir d'une séquence en acides aminés, nous prédisons un ensemble de lettres HMM-SA servant de point de départ pour la recherche de fragments candidats dans une banque de structures. A la différence des approches précédentes, notre approche n'impose pas une taille fixe des fragments identifiés.

Les étapes de l'approche SAFrAN, décrites dans la figure 6.1, consistent à : (i) prédire la séquence de lettres HMM-SA à partir de la séquence en acides aminés, conditionnellement à la prédiction des structures secondaires par la méthode PSIPRED (Jones, 1999b), (ii) rechercher des fragments de taille minimum L, compatibles avec la séquence HMM-SA prédite, dans une banque de structures encodées dans l'espace de l'alphabet structural HMM-SA, et (iii) appliquer des filtres pour raffiner la sélection de fragments. Les étapes (ii) et (iii) sont itérées en diminuant la taille minimum des fragments identifiés à $L-1$ jusqu'à ce que la séquence soit couverte intégralement ou qu'aucun nouveau fragment ne puisse être détecté ou que L soit égal à deux. Une position de la séquence en acides aminés est considérée comme couverte, si nous avons trouvé au moins N fragments la recouvrant. N est ici fixé à 4. Les fragments résultants sont ensuite groupés par l'intermédiaire d'une procédure de classification hiérarchique.

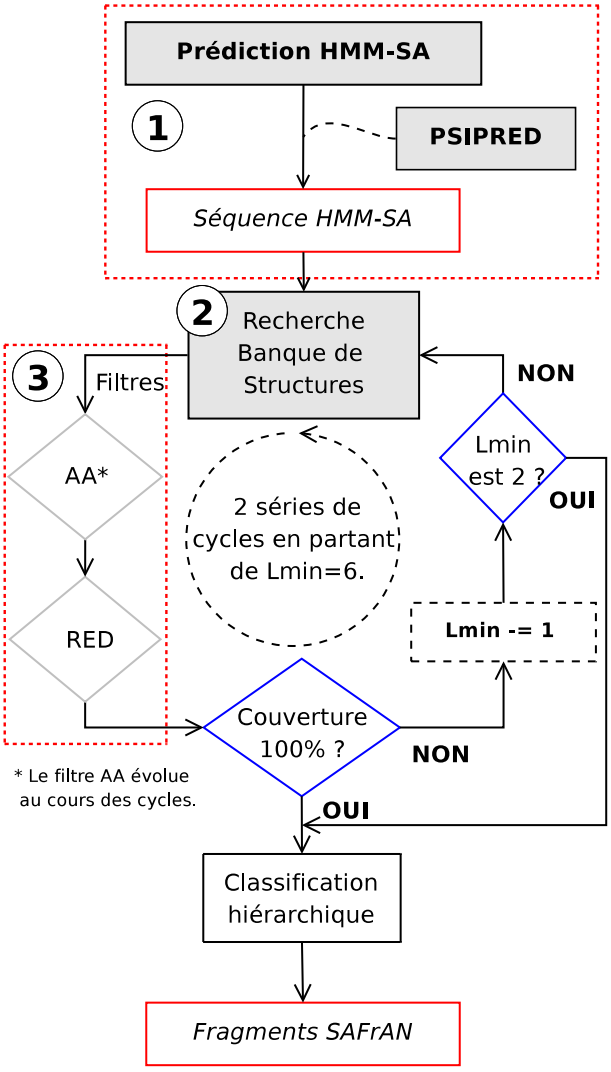


Fig. 6.1: La méthode SAFrAN. Etape 1 : prédiction de la séquence en lettres HMM-SA compatible avec la séquence en acides aminés. Etape 2 : recherche de fragments candidats de longueur minimum L_{min} compatibles avec la séquence HMM-SA prédite, dans une banque de structures encodées dans l'espace de l'alphabet structural. Etape 3 : filtrage des solutions par leur séquence en acides aminés (AA), et traitement de la redondance (RED). A chaque cycle la couverture est analysée et L_{min} diminue de 1 résidu. Si L_{min} est égale à 2 ou que la couverture est de 100% l'algorithme s'arrête. Les fragments résultants sont ensuite regroupés par une procédure de classification hiérarchique.

6.1 Matériels et méthodes

6.1.1 Les jeux de données

Le jeu d'apprentissage. Le jeu d'apprentissage (JA), datant de 2003, est contitué de 3439 protéines dont la structure a été résolue par rayons X à haute résolution (moins de 2,5 Å), n'ayant aucune rupture de chaîne et partageant moins de 50 % d'identité de séquence entre elles. Cet ensemble de structures, identifié à partir du site de la *Protein Data Bank* (Berman et al., 2000b) en utilisant un algorithme similaire au logiciel PISCES (Wang and Dunbrack, 2003), constitue le jeu d'apprentissage (JA), utilisé pour la prédiction à partir de la séquence en acides aminés (voir la section 6.1.2).

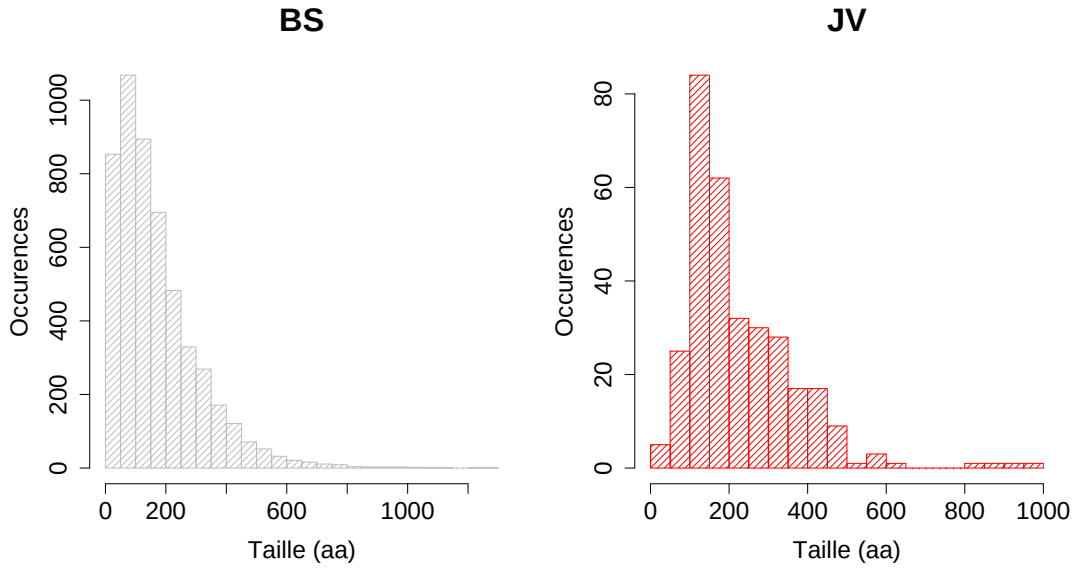


Fig. 6.2: SAFrAN, distribution des tailles des protéines des jeux de données. Les distributions des tailles, en acides aminés, des protéines de la banque de structures (BS), et du jeu de validation (JV) de nouvelles entrées PDB sont ici tracées.

La banque de structures. De la même manière, la banque de structures (BS), datant de 2006, est contituée de 3672 protéines dont la structure a été résolue par RX à moins de 2,5 Å, et partageant entre elles moins de 30% d'identité de séquence. Pour les chaînes incomplètes, chaque fragment est considéré comme une entrée à part entière dans la banque. La taille de la banque est ainsi de 5114 chaînes/fragments dont les tailles varient de 10 à 1264 résidus (voir la figure 6.2), 70 % des structures ayant une taille comprise entre 60 et 350 résidus.

Le jeu de validation. Afin de tester la capacité de SAFrAN à trouver des fragments pertinents pour de nouveaux repliements, nous avons constitué un jeu de validation (JV) uniquement composé d'entrées récentes de la PDB (postérieures à la date de constitution de la BS) similairement à la BS, et partageant moins de 30 % d'identité de séquence avec les protéines de cette dernière. Le JV contient ainsi 322 protéines. En terme de structuration secondaire, les protéines du JV se répartissent comme suit dans les différentes classes : 22,9 % α, 10,8 % β, 65,3 % α/β, et 1 % de petits peptides sans structure secondaire canonique. Ces protéines n'étant pour la plupart pas encore référencées dans les bases CATH (Orengo et al., 1997; Pearl et al., 2003, 2005) ou SCOP (Murzin et al., 1995), nous avons déterminé ces classes à partir des assignations des structures secondaires par le logiciel STRIDE (Frishman and Argos, 1995) et les critères définis par Michie et al. (1996).

6.1.2 Prédiction HMM-SA sous contrainte

La première étape de la méthode SAFrAN est de prédire les lettres HMM-SA pouvant décrire la structure à modéliser à partir de sa seule séquence en acides aminés.

Une étape clé pour vérifier la pertinence potentielle de la prédiction par HMM-SA est d'abord de s'assurer qu'il existe bien une dépendance significative entre la séquence primaire (acides aminés) et l'alphabet structural. Camproux and Tuffery (2005) ont ainsi montré qu'à l'optimum statistique des HMM (27 lettres structurales), l'ensemble des lettres structurales présentent des associations significatives avec un ensemble d'acides aminés. De plus, l'analyse des types d'acides aminés favorisés ou défavorisés dans les différentes lettres structurales apparaissent consistantes avec la littérature, particulièrement pour les lettres HMM-SA associées à une structure secondaire (Argos and Palau, 1982; Richardson and Richardson, 1988).

La démarche de prédiction à partir de la séquence en acides aminés, ne consite pas à se limiter à une prédiction bayésienne p(HMM-SA/AA), mais à intégrer la dépendance entre séquence en acides aminés et lettres HMM-SA dans le processus markovien préalablement décrit. Concrètement, cela revient à remplacer dans la chaîne de Markov cachée, la densité multi-normale de chaque état (correspondant à l'émission des fragments associés, soit les 4 descripteurs), par la probabilité que chaque état émette les 20 acides aminés, *i.e.* p(AA/HMM-SA).

Lors de la prédiction, il est possible de contraindre l'algorithme *forward-backward* à n'utiliser qu'un sous-ensemble des 27 lettres de l'alphabet structural HMM-SA à chaque position de la séquence. Nous avons choisi d'utiliser PSIPRED (Jones, 1999b), une des

méthodes les plus performantes à l'heure actuelle. PSIPRED évalue un indice de confiance de prédiction selon une échelle allant de 0 à 9, 0 étant le plus faible degré de confiance. Une étude de la pertinence de la prédiction en fonction du degré de confiance a été réalisée (données non présentées), et le seuil de 5 s'est révelé être le meilleur compromis entre contrainte et qualité de prédiction. Ainsi, pour les régions de la séquence prédites avec un degré de confiance supérieur à 5, nous appliquons les contraintes suivantes : (i) pour les hélices, nous ne considérons que les lettres `[a A V W Z B C D E]`, (ii) pour les brins, le jeu de lettres utilisé est `[L M N T X J K]`, (iii) les régions en *coil* sont prédites à partir du sous-ensemble `[B C D E F G H I J K L N O P Q R S T U Y Z]`, et (iv) si aucune contrainte n'est applicable, l'ensemble des 27 lettres est utilisé pour la prédiction.

6.1.3 Recherche de fragments candidats à taille minimale imposée

La recherche de fragments candidats est basée sur une approche similaire à SA-Search (Guyon et al., 2004).

Bien qu'il soit possible de produire des alignements des structures avec *gap*, nous n'avons pas considéré cette possibilité, notre but étant d'identifier des fragments structuraux qui correspondent exactement à des régions de la séquence considérée.

L'identification des fragments n'est pas basée sur la recherche du fragment ayant le meilleur score d'alignement. Au lieu d'utiliser un algorithme de programmation dynamique simple, nous utilisons un algorithme capable de trouver l'ensemble des mots ayant une taille minimale imposée et une valeur de score supérieure à un seuil.

Enfin, en raison de la discordance possible entre l'encodage de structures 3D et la séquence HMM-SA résultant de la prédiction à partir de la séquence en acides aminés, nous avons encodé l'ensemble des structures de la BS, avec la même méthode que celle employée pour la prédiction à partir de la séquence. La recherche de fragments candidats s'effectue donc sur cet ensemble de séquences HMM-SA.

6.1.4 Recherche itérative de fragments candidats

Les fragments SAFrAN sont détectés itérativement pour une taille minimum et un score seuil dont les valeurs ont été calibrées à partir de séries de tests (données non présentées). Une position est considérée comme couverte si au moins quatre fragments recouvrent cette position. Ce paramètre est très sensible à la taille de la BS.

6.1.5 Filtrage des solutions

Les fragments candidats issus d'une recherche SAFrAN sont filtrés pour limiter le nombre de candidats. Deux types de filtres ont été mis en place pour améliorer (i) la

compatiblité de séquence en acides aminés et (ii) pour éviter la redondance des fragments sélectionnés.

Compatibilité Acides Aminés : chacune des séquences des fragments candidats est alignée avec la séquence de la protéine d'intérêt par l'algorithme de Smith et Waterman. Si le score d'alignement est supérieur à un seuil donné (nous utilisons une matrice BLOSUM 62 (Henikoff and Henikoff, 1992)), et qu'il y a au maximum un gap dans l'alignement, le fragment est alors conservé.

Redondance 3D : afin de limiter la redondance tridimensionnelle, nous avons décidé d'élaguer les solutions par plusieurs mécanismes. Un seul exemplaire des mots strictement identiques, présents plusieurs fois, est conservé. Les séries de lettres correspondant à des conformations hélicales pures sont aussi traitées. Un seul exemplaire pouvant être présent pour une région donnée. Par ailleurs, nous avons aussi mis en place une classification hiérarchique des fragments candidats de même longueur à chaque position sur un critère cRMSd. Le seuil utilisé pour déterminer les classes est de 0.5 Å. Seuls les fragments centroïdes des classes sont conservés.

6.1.6 Reconstruction de structures protéiques complètes

Afin de tester la pertinence des fragments candidats détectés par SAFrAN, nous avons reconstruit l'ensemble des cibles du JV à partir d'une trajectoire issue de ces fragments prédits. Pour naviguer dans la combinatoire des fragments, nous avons utilisé l'algorithme glouton mis en place dans le laboratoire (Tuffery et al., 2005), en utilisant comme fonction objectif le cRMSd. Notre but est ici de déterminer si l'ensemble des fragments prédits par SAFrAN permet d'approximer la structure à prédire avec une bonne précision.

6.2 Résultats - Discussion

La sortie du programme SAFrAN est un alignement de fragments candidats par rapport à la séquence de recherche en acides aminés (voir la figure 6.3). La superposition de cet ensemble de fragments, sur la protéine à modéliser, nous montre que ces fragments permettent de décrire globalement la topologie de la structure (voir la figure 6.4). Nous allons maintenant détailler la pertinence de ces fragments (i) en terme d'approximation locale et (ii) en terme d'approximation globale des structures à prédire, validées sur un ensemble de 322 protéines du JV n'appartenant ni au JA, ni à la BS, et n'ayant pas d'homologues dans ces derniers.

```
NLKTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAEVPRVG << TARGET SEQUENCE
ZERFF-------------------------------------------------------------   1   5 2pf1   54  58
FDS---------------------------------------------------------------   1   3 1eazA  14  16
-ZQM--------------------------------------------------------------   2   4 1duyA 178 180
-SLTY-------------------------------------------------------------   2   5 1egyA  60  63
-SHOEG------------------------------------------------------------   2   6 1h0nA   9  13
--CRKP------------------------------------------------------------   3   6 1jiaA 107 110
--CPRP------------------------------------------------------------   3   6 1nxwA  41  44
---FFR------------------------------------------------------------   4   6 1usoA   5   7
---CRPQP----------------------------------------------------------   4   8 1pi1A  99 103
---CGEQX----------------------------------------------------------   4   8 1fr1A  91  95
-----ZCQXLT-------------------------------------------------------   6  11 1nbfA 114 119
-----PILLNT-------------------------------------------------------   6  11 1np2A 271 276
--------YQHBQ-----------------------------------------------------   9  13 1opmA 262 266
---------SXYWA----------------------------------------------------  10  14 1kkjA 366 370
---------SKHAA----------------------------------------------------  10  14 1lw6I   8  12
----------LHAVWAVZZ-----------------------------------------------  11  19 1qpwB   2  10
-----------YBAAAAVZW----------------------------------------------  12  20 1b79A   3  11
-----------GBBBBV-------------------------------------------------  12  17 1qvyA  26  31
--------------aaaaaaaaZ------------------------------------------  15  23 1fvaB 101 109
----------------AAVWAAV-----------------------------------------  17  23 1cjcA 404 410
--------------------VZGEB-----------------------------------------  22  26 1eaf   84  88
--------------------VQYH------------------------------------------  22  25 1o7kA 102 105
----------------------HAVQP-------------------------------------  24  28 1rwzA 176 180
------------------------YNMNX-----------------------------------  25  29 1h4iA 120 124
------------------------YNMN------------------------------------  25  28 1h4iA 120 123
------------------------SMNM------------------------------------  25  28 1n0lA 100 103
--------------------------MXKNUO----------------------------------  28  33 1aec  193 198
----------------------------LXKLUQF-------------------------------  30  36 1t0iA 152 158
-----------------------------KKKUSXY-------------------------------  31  37 1e96B 174 180
-------------------------------KLUQF-------------------------------  32  36 1t0iA 154 158
-------------------------------KKUSXY-------------------------------  32  37 1e96B 175 180
---------------------------------USMNM-------------------------------  34  38 1cm2A  56  60
---------------------------------JHIMM-------------------------------  34  38 1nbvH 139 143
-----------------------------------LKPB---------------------------  37  40 1aocA  20  23
-----------------------------------GBSP---------------------------  37  40 1i39A 187 190
-------------------------------------XLGY-------------------------  39  42 1tig   71  74
-------------------------------------PEQP-------------------------  39  42 1nu7D   7  10
-------------------------------------NPCS-------------------------  39  42 1i10A  16  19
-------------------------------------KPBS-------------------------  39  42 1cla  133 136
-------------------------------------HZRG-------------------------  39  42 1jo0A  87  90
---------------------------------------LTPVQ---------------------  41  45 1lw6I  39  43
---------------------------------------CSP----------------------  41  43 1vjcA  73  75
-----------------------------------------ZQG--------------------  43  45 1whsB  72  74
-------------------------------------------PIJLNK--------------  44  49 1dsn   52  57
-----------------------------------------------NXMMN-----------  48  52 1pot  227 231
-----------------------------------------------NLNTYKY---------  48  54 1gqvA 108 114
-----------------------------------------------MNMYADSYG------  49  57 1lw6I  47  55
---------------------------------------------------KYUS------  54  57 1ng5A 190 193
-------------------------------------------------------NPGIT  58  62 1bxbA  82  86
-------------------------------------------------------KLKKU  58  62 1qexA  76  80
```

Fig. 6.3: Un exemple de résultat de SAFrAN. La colonne 1 contient les mots HMM-SA trouvés par SAFrAN alignés sur la séquence de recherche en acides aminés (dont les positions sont renseignées dans les colonnes 2 et 3). La colonne 4 contient l'identifiant de la structure PDB dont est extrait le mot trouvé, les bornes d'extraction étant situées respectivement dans les colonnes 5 et 6. Notez que le résultat de la cible 2ci2I a ici volontairement été tronqué pour des besoins de présentation.

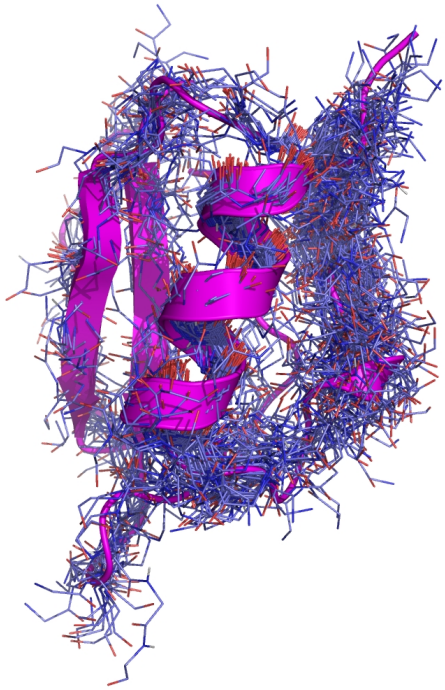


Fig. 6.4: Un exemple de fragments candidats prédits par SAFrAN. L'ensemble des fragments identifiés par SAFrAN pour la cible 2ci2I (mode *lines* en bleu) sont ici superposés à la structure native dont la séquence de recherche est issue (mode *cartoon* en magenta).

6.2.1 Approximation locale des structures protéiques

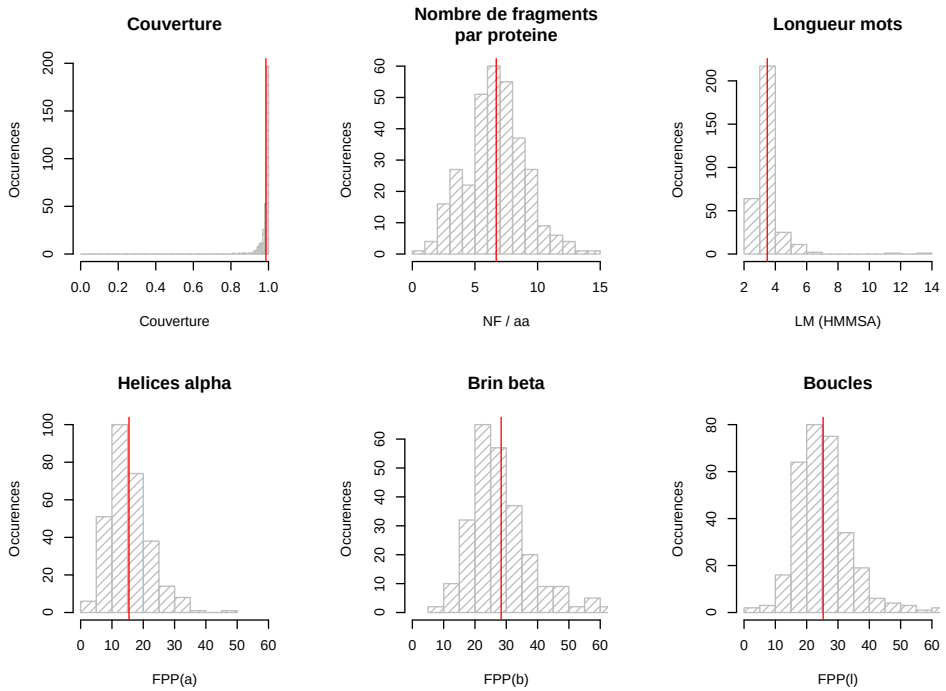
Propriétés des fragments identifiés

La figure 6.5 présente une analyse des fragments prédits par SAFrAN avec le Filtre sur les Acides Aminés (FAA) et Sans le Filtre sur les Acides Aminés (SFAA). On voit que, selon la calibration de la méthode, la longueur des fragments candidats varie de 8,3 lettres HMM-SA à 3,5. Le filtre sur les acides aminés semble tellement stringent qu'il force l'algorithme SAFrAN à aller chercher des fragments de petite taille pour satisfaire nos critères de redondance par position et de couverture. Sans ce filtre, SAFrAN détecte plus de fragments par protéine.

Concernant la répartition des fragments selon la structure secondaire des résidus couverts, SAFrAN détecte, en moyenne, 28 fragments par position (FPP) pour les brins β (41%), 15 FPP pour les hélices α (22%), et 25 FPP pour les boucles (37%). Les hélices α étant sur-représentées dans les résultats SAFrAN, nous les avons hautement filtrées, ce qui explique que l'on ait moins de fragments candidats pour ces zones. Les brins β sont quant à eux bien représentés, et il est intéressant de voir que SAFrAN ne détecte pas que des zones de structuration secondaire, mais propose un grand nombre de candidats pour les boucles. Sans le FAA, les résultats penchent vers une nette préférence pour les hélices α (deux fois plus que les boucles et presque trois fois plus que les brins β).

La couverture, *i.e.* la fraction de résidus pour lesquels SAFrAN propose des fragments

A. FAA



B. SFAA

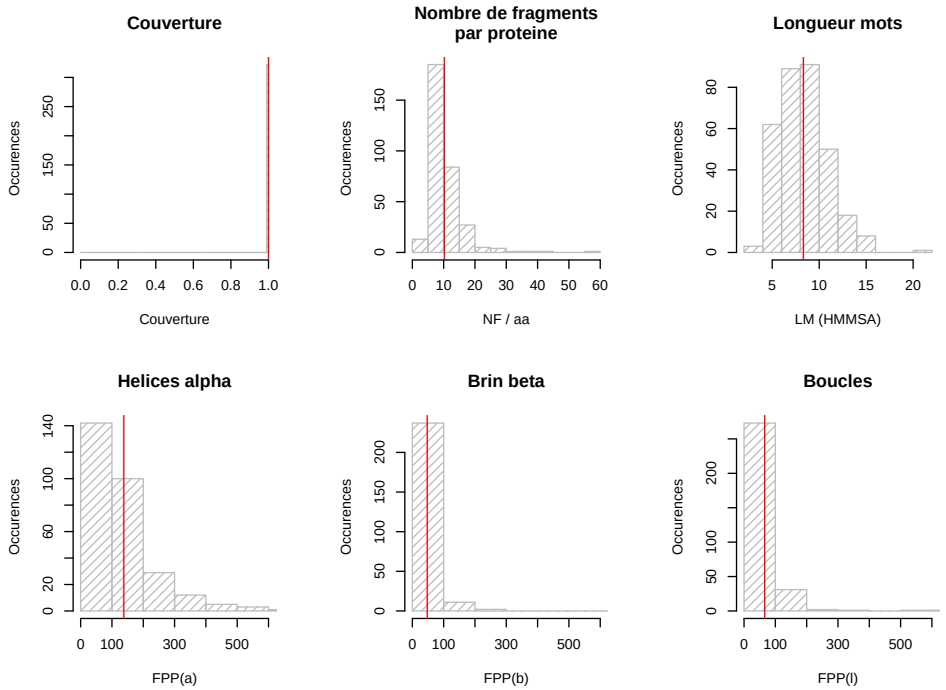


Fig. 6.5: Propriétés des fragments SAFrAN. Les fragments candidats détectés par SAFrAN sont ici analysés en terme de couverture de la séquence de recherche, du nombre de fragments trouvés pour chaque protéine, normalisé par la taille de cette dernière, de longueur des mots trouvés, et le nombre de fragments par position (FPP) assignée comme appartenant à une structure secondaire spécifique par le logiciel STRIDE. Les lignes rouges verticales représentent les valeurs moyennes des distributions. **A.** Résultats obtenus avec le Filtre sur les Acides Aminés (FAA) et **B.** Sans le Filtre sur les Acides Aminés (SFAA).

candidats, quant à elle, varie beaucoup moins : elle est de 98,6% avec le FAA et de 100% sans le FAA.

Qualité des fragments

Les distributions des cRMSds des fragments candidats détectés par SAFrAN, contre la structure native de la séquence de recherche, sont présentés dans la figure 6.6 A. A titre de comparaison, nous avons tracé le cRMSd moyen que l'on pourrait obtenir, pour chaque taille de fragment, si l'on sélectionnait, au hasard, un fragment dans la PDB. Ces valeurs ont été recalculées pour des tailles de fragments allant de 3 à 21 résidus. Les distributions des cRMSds obtenues sont strictement identiques à celles calculées par Micheletti et al. (2000) (fragments de taille 3 à 7 et 10 résidus).

Les valeurs de cRMSds présentées ont été calculées pour chaque fragment, en ignorant les 2 derniers acides aminés C-terminaux dont on sait qu'ils ne créent que du bruit de fond (données non présentées). Les plus petits fragments détectés par SAFrAN ont donc une longueur de 3 résidus.

Globalement, plus la taille des fragments détectés augmente, plus le nombre de fragments de faible cRMSd augmente, et donc, plus l'écart se creuse avec la valeur moyenne due au hasard. Cependant, pour les fragments de taille 4 à 7, les distributions semblent bimodales, suggérant qu'il existe un certain nombre de faux positifs. Ce phénomène semble s'atténuer pour les mots de taille supérieure à neuf pour lesquels nous avons un distribution idéale.

Pour évaluer le bruit associé à la recherche SAFrAN, nous avons retracé ces mêmes distributions, en ne considérant que le fragment de plus faible cRMSd, si à une position donnée nous avons plusieurs fragments de même taille. Ces résultats sont présentés dans la figure 6.6 B. Pour les fragments de taille 4 à 6, les distributions ont été déplacées vers un mode de plus faible cRMSd. Ce phénomène ne se retrouve pas pour des fragments plus longs.

Si l'on trace ces mêmes distributions des cRMSd des fragments obtenus par SAFrAN SFAA, il apparait que ce dernier n'a qu'une influence faible sur les fragments de petites tailles, mais devient pertinent pour les fragments de plus de 7 résidus de long (voir la figure 6.7). Autrement dit, l'information de séquence devient une information pertinente pouvant aider à la reconnaissance de repliement à partir de 7 résidus. Une des améliorations possibles de ce filtre serait donc de ne l'appliquer qu'aux fragments de taille supérieure à 7 résidus.

Une analyse plus détaillée des fragments en fonction de leur localisation au sein des structures secondaires, pour les mots de taille 5 et 7, nous a permis de mettre en évidence les populations de fragments en jeu dans la distribution bimodale observée (voir la figure 6.8). Nous voyons ici clairement que pour les mots de taille 5, il existe bien deux popu-

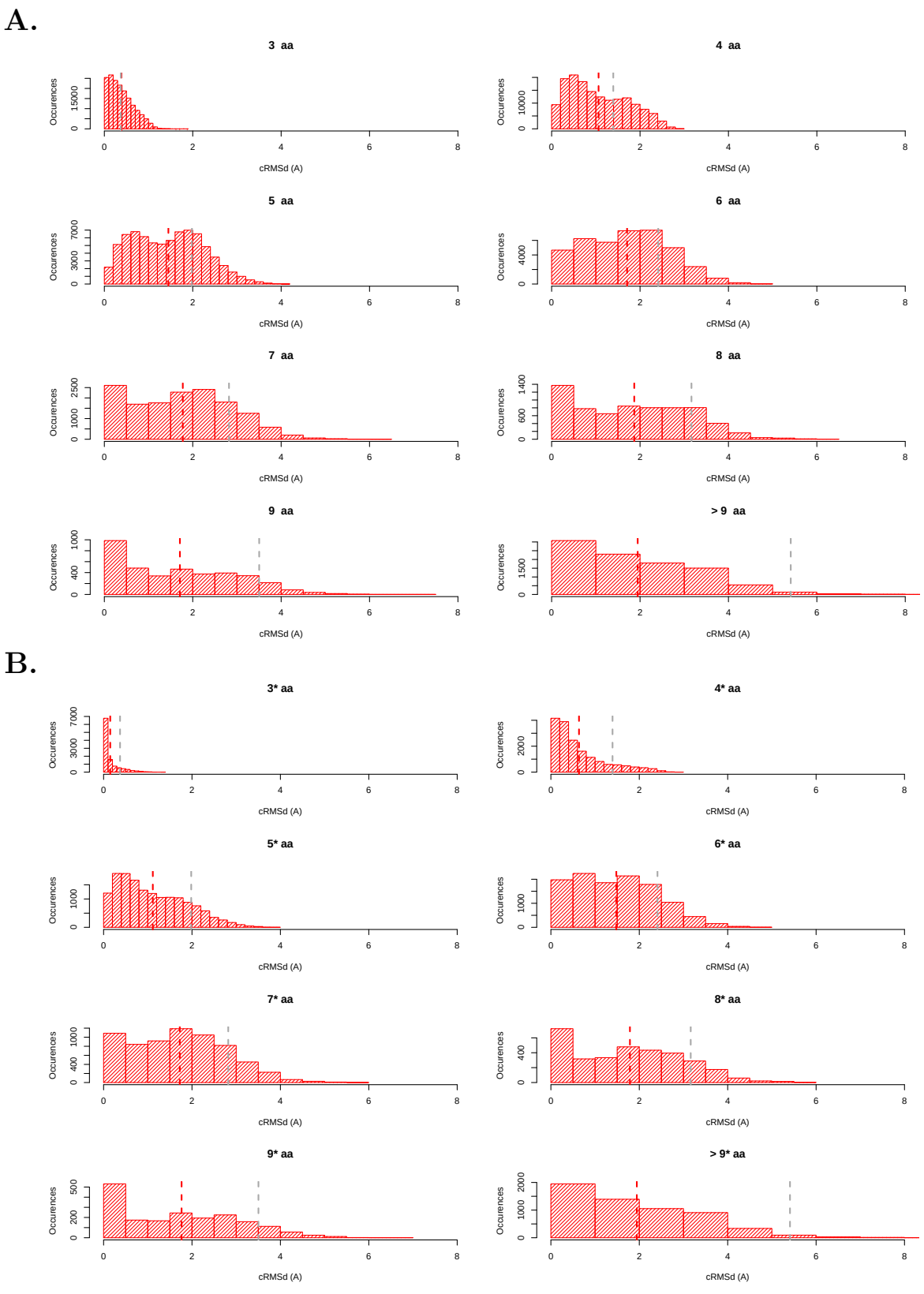


Fig. 6.6: Distribution des cRMSd des fragments SAFrAN. La distribution des cRMSds des fragments SAFrAN par rapport à la structure native de la séquence de recherche sont ici présentés en fonction de leur taille en acides aminés. Les lignes verticales représentent le cRMSd moyen pour la distribution SAFrAN (en rouge), et le hasard (en gris). Dans **A**, tous les fragments candidats sont considérés, et dans **B**, si une position contient plusieurs mots de même taille, seul celui de plus faible cRMSd est considéré.

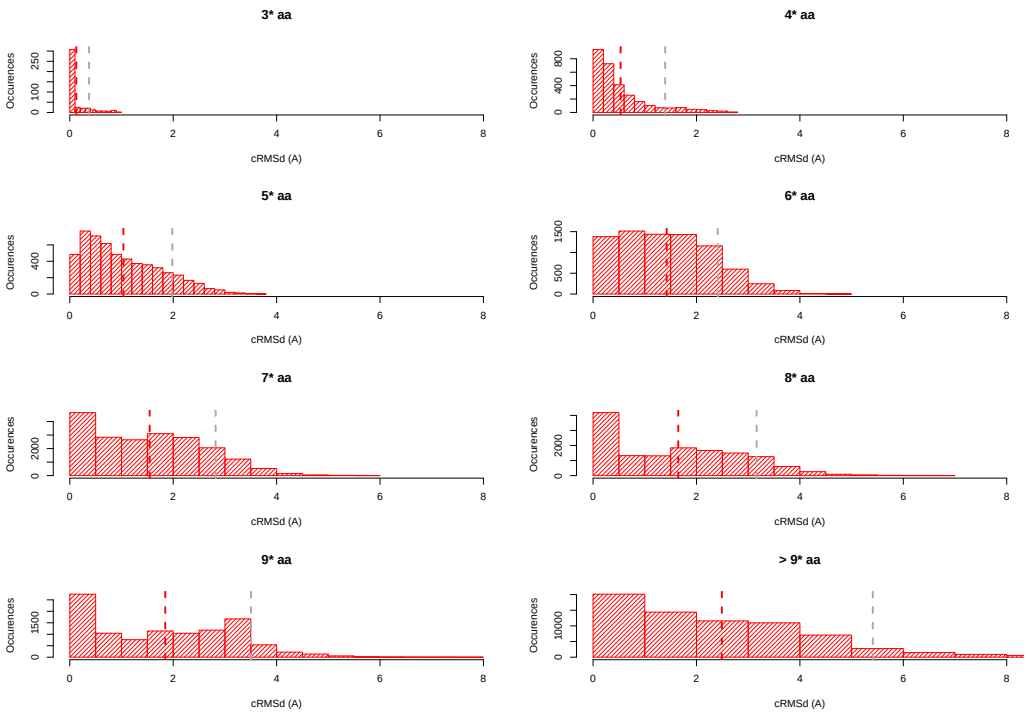


Fig. 6.7: Absence du filtre AA. Les distributions des cRMSds des fragments détectés par SAFrAN SFAA, en comparaison avec la structure native, sont ici présentées en fonction de la taille des fragments. Si une position contient plusieurs mots de même taille, seul celui de plus faible cRMSd est considéré. Les lignes verticales représentent le cRMSd moyen pour la distribution SAFrAN (en rouge), et le hasard (en gris).

lations pour le coeur des hélices α et des feuillets β, mais aussi pour les sorties d'hélices. Pour les mots de taille 7, nous ne pouvons plus distinguer ces deux populations, mais, les coeurs et sortie des hélices α peuplent les zones de faible cRMSd, tandis que les régions de cRMSds intermédiaires sont peuplées par les brins β et les boucles (2 Å de cRMSd en moyenne). Les feuillets β étant des structures plus difficiles à prédire, ces résultats semblent cohérents. Par ailleurs, nous pouvons remarquer que la méthode semble plus performante pour détecter les sorties de structures secondaires (EC, HC) que les entrées (CE, CH).

A titre de comparaison, si l'on analyse ces mêmes distributions sans le filtre sur les acides aminés (figure 6.9), nous pouvons clairement remarquer que les sorties d'hélices sont plus difficiles sans ce dernier qu'avec. Ceci est sûrement dû à la petite taille des fragments détectés avec le FAA.

Prédiction en terme de lettres HMM-SA

Pour SAFrAN, nous pouvons calculer un Neq, de par les fréquences d'apparition des lettres HMM-SA à une position donnée dans l'alignement des fragments candidats avec la

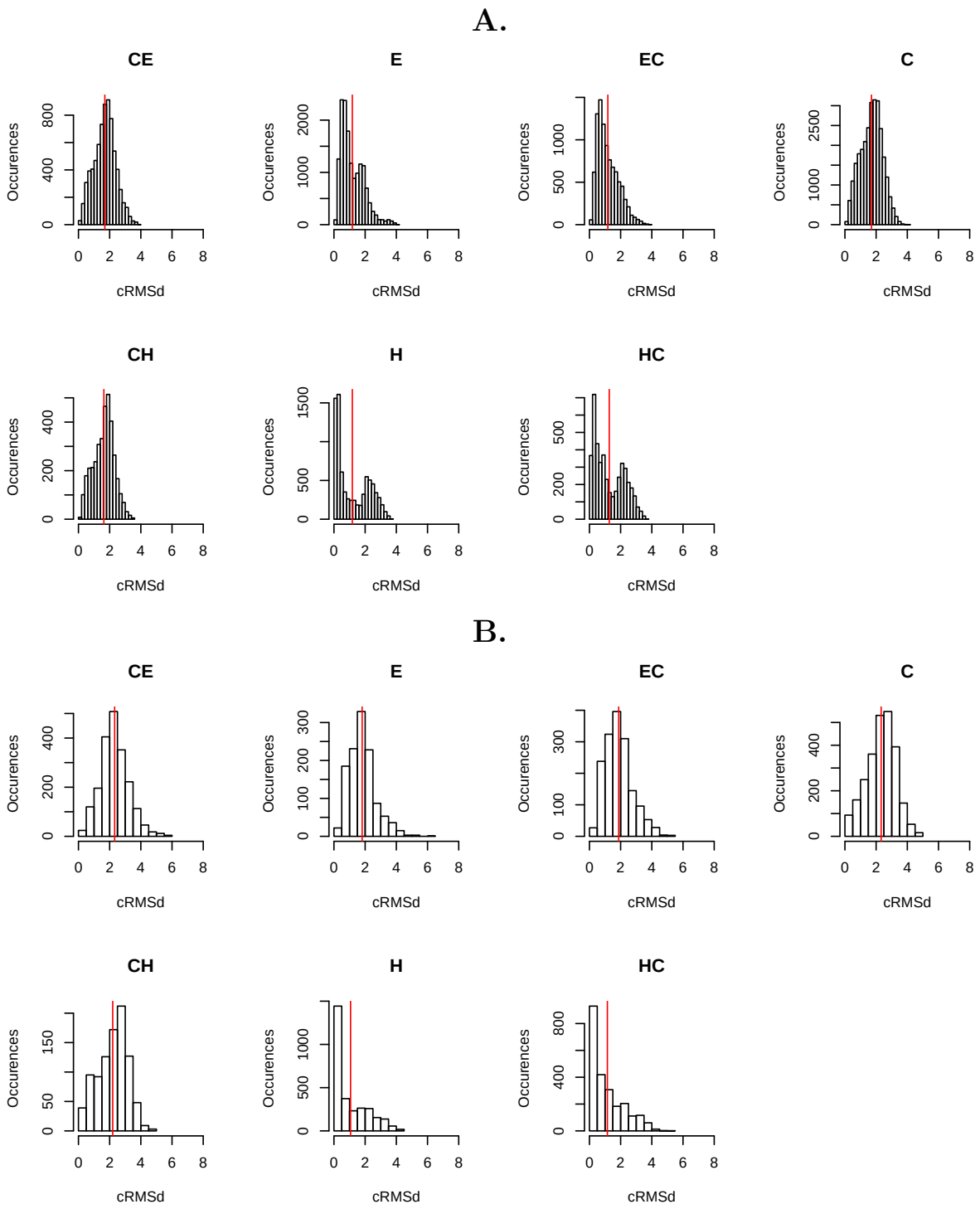


Fig. 6.8: Les fragments de taille 5 et 7 dans les structures secondaires. Pour les fragments de tailles 5 (A) et 7 (B), nous avons ici détaillé la distribution des cRMSds de ces derniers en fonction de leur localisation dans les structures secondaires. Le code choisi est issu du classique code à 3 lettres (E, *Extended*, H, *Helix* et C, *Coils*), ainsi : CE, entrée de feuillet β ; E, coeur du feuillet β ; EC, sortie de feuillet β ; C, boucle ; CH, entrée dans une hélice α ; H, coeur d'une hélice α ; HC, sortie d'une hélice α. Les valeurs moyennes sont ici mises en évidence par une ligne verticale continue rouge.

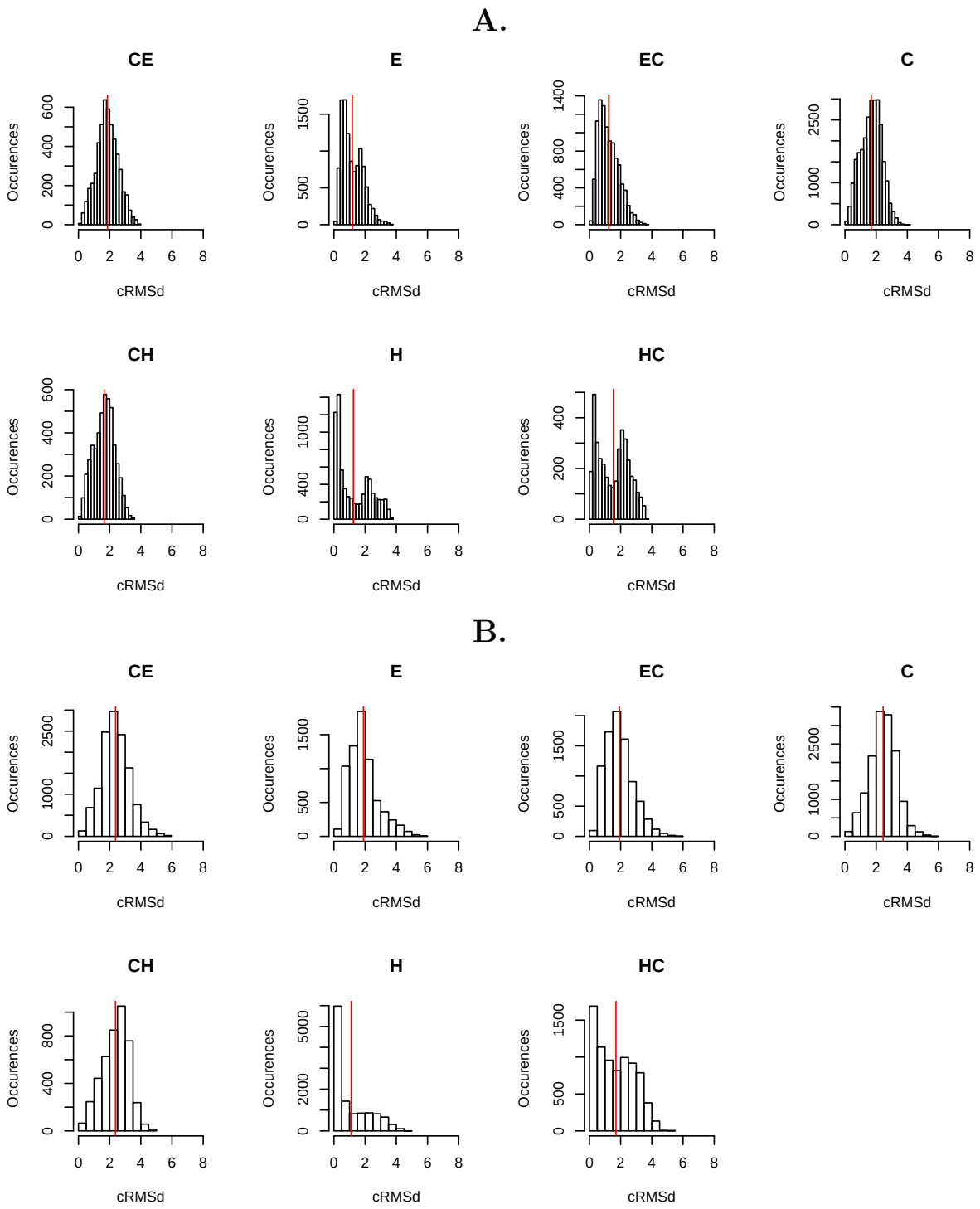


Fig. 6.9: Les fragments de taille 5 et 7 dans les structures secondaires avec SAFrAN SFAA. Pour les fragments de tailles 5 (A) et 7 (B), nous avons ici détaillé la distribution des cRMSds de ces derniers en fonction de leur localisation dans les structures secondaires. Le code choisi est issu du classique code à 3 lettres (E, *Extended*, H, *Helix* et C, *Coils*), ainsi : CE, entrée de feuillet β ; E, coeur du feuillet β ; EC, sortie de feuillet β ; C, boucle ; CH, entrée dans une hélice α ; H, coeur d'une hélice α ; HC, sortie d'une hélice α. Les valeurs moyennes sont ici mises en évidence par une ligne verticale continue rouge.

	Rang 1				**Neq**				
			C				**C**		
		P		**P**		**P**		**P**	$< Neq >$
HMM	0.23	0.25	0.68	0.76	0.67	0.65	0.93	0.94	6.65
SAFrAN$_{FAA}$		0.21		0.73		0.65		0.97	7.28
SAFrAN$_{SFAA}$		0.22		0.70		0.72		0.98	8.41

Tab. 6.1: Performances de prédiction. Dans cette table, sont résumées les performances de prédiction HMM-SA en ne considérant que la lettre la plus probable (Rang 1) ou les Neq meilleures lettres (Neq), sans ou avec la matrice de confusion (C). Dans chacun des cas est envisagée la prédiction conditionnée par PSIPRED (P). Trois cas sont présentés, la prédiction pure à partir du modèle markovien, et les prédiction SAFrAN, avec le filtre sur les acides aminés (SAFrAN$_{FAA}$) et sans ($_{SFAA}$).

séquence requête. Pour le calculer, nous évaluons la fréquence d'apparition E_i de chaque lettre L_i parmi les N lettres présentes :

$$E_i = \frac{L_i}{\sum_{k=1}^{N} L_k} \tag{6.1}$$

Nous pouvons en déduire la valeur de l'entropie de Shannon :

$$H = -\sum_{i=1}^{N} E_i * ln(E_i) \tag{6.2}$$

Le nombre équivalent de lettres est alors :

$$N_{eq} = exp(H) \tag{6.3}$$

Les Neq lettres les plus fréquentes sont toutes statistiquement significatives, ainsi, plus la valeur du Neq est faible, plus la qualité de la prédiction est bonne. Bien que dans le cas présent, ce paramètre soit biaisé par le traitement imposé de la redondance, il nous aide à diminuer la complexité des trajectoires issues de SAFrAN.

SAFrAN étant une méthode de recherche de fragments candidats, il est intéressant d'analyser la qualité de la prédiction par position en terme lettres HMM-SA. Antérieurement à SAFrAN, la performance de la prédiction HMM-SA avait été évaluée sur un ensemble de 80 nouvelles entrées de la PDB n'appartenant pas au jeu d'apprentissage. Les résultats sont présentés dans la table 6.1 (ligne HMM). Si l'on ne considère que la lettre HMM-SA prédite au rang 1, nous avons une valeur de Q_{27} égale 23%. L'utilisation de PSIPRED n'améliore que très peu la pertinence de la prédiction au rang 1 (25%). PSIPRED étant une méthode de prédiction de la structure secondaire des protéines, il a déjà été mentionné, dans un contexte similaire, qu'elle ne semble améliorer principalement que la prédiction de deux états, les hélices α et les feuillets β, ce qui explique le faible

gain observé (Etchebest et al., 2005). Pour Yang and Wang (2003), ce gain est significatif (+7%) à partir du moment où ils ne possèdent que quatre états différents dont deux états spécifiques des structures secondaires.

Pour SAFrAN, la prediction au rang 1 n'a pas de sens : ce que nous cherchons sont des formes de repliement. Nous préférons donc plutôt analyser ces résultats à l'aide de la matrice de confusion 3D entre les états décrite dans l'introduction. Dans ce contexte, le Q_{27} est alors de 76% et 68% avec et sans PSIPRED. De plus, si l'on considère non pas la meilleure prédiction au rang 1, mais les Neq meilleures lettres prédites, alors, sans la matrice de confusion, la prédiction présente une performance similaire à la meilleure prédiction de rang 1 avec la matrice de confusion, soit environ 67%. La valeur moyenne du Neq sur ce jeu test est de 6,65 lettres. PSIPRED diminue les performances de prédictions dans ce cas. Si l'on considère les Neq meilleures prédictions et la matrice de confusion, les performances de prédiction sont alors supérieures à 90%, mais cependant, les trajectoires correspondantes doivent présenter une grande complexité. Globalement, les performances de prédiction semblent comparables à d'autres approches.

de Brevern et al. (2007) ont récemment amélioré leur méthode de prédiction bayesienne (de Brevern et al., 2000; Etchebest et al., 2005), *via* l'extension de mots structuraux (suite de PB). Leur méthode de prédiction semble performante, leur Q_{16} étant de 44 %. Cependant, il est à prendre en considération qu'ils n'ont que 16 états différents dans leur alphabet, et donc 1 chance sur 16 d'avoir une bonne prédiction, soit 6,2%. Dans notre cas, nous avons 27 états, donc un Q_{27} associé au hasard de 3,7%, soit quasiment deux fois moins. Hunter and Subramaniam (2003a) ont défini quant à eux 28 états, et ont une performance de prédiction de 40%. Cependant, comme il l'a déjà été soulevé dans (Etchebest et al., 2005; Sander et al., 2006), leur prédiction est biaisée par les centroïdes les plus fréquents. Ainsi, huit des 28 centroïdes sont prédits correctement avec un taux supérieur à 20%, et seulement quatre au delà de 50%. De plus, onze de leur centroïdes ne sont pas du tout prédits. Sander et al. (2006) aboutissent à une performance de prédiction de 36% pour 27 états représentatifs. Le Q_{27} augmente à 54% lorsqu'ils considèrent 3 classes au lieu de la meilleure, ce qui revient à prendre en considération une version bridée de leur matrice de confusion.

Nous voyons que nous pouvons avoir un Q_{27} de 98% en considérant les Neq valeurs et la matrice de confusion, ce qui représente un gain de 4% par rapport à la prédiction pure, quoi qu'au prix d'un nombre de lettres par position supérieur (8,41 *vs* 6,65).

Une autre façon d'estimer la prédiction induite par SAFrAN est de compter si la lettre exacte ou compatible apparaît à chaque position. Ainsi, sans le filtre sur les acides aminés, sur les 11 lettres proposées en moyenne, dans 75% des cas, SAFrAN a trouvé la lettre HMM-SA exacte pour les positions couvertes, et dans 98% des cas une lettre compatible. Avec le filtre sur les acides aminés, pour 7 lettres proposées par position en moyenne, ces valeurs sont respectivement de 66% et 97%. Ces valeurs sont sensiblement

identiques à celles obtenues en ne considérant que les Neq meilleures lettres. Autrement dit, il est rassurant de voir que les lettres dont la fréquence d'apparition est significative sont les plus pertinentes.

6.2.2 Impact de la méthode de prédiction des structures secondaires

Pour évaluer le gain associé à l'utilisation de PSIPRED dans le conditionnement de la prédiction, nous avons recherché des fragments candidats sur le même jeu de validation, à partir de la prédiction pure (donc sans PSIPRED). Les distributions des cRMSds des fragments de tailles 5 et 7 ainsi obtenues ont été retracées en fonction des structures secondaires (voir la figure 6.10). Nous voyons ici clairement, que, sans PSIPRED, SAFrAN a tendance à peupler les régions de cRMSd comprises entre 2 et 4 Å pour les régions en hélice (H), alors que le cRMSd moyen obtenu avec la prédiction conditionnée par PSIPRED se situe autour de 1 Å. PSIPRED semble donc grandement améliorer la prédiction (et donc la qualité des fragments détectés) pour les régions hélicales. Ceci est moins vrai pour les régions en brin β. Les distributions des cRMSds pour les régions en boucles sont quant à elles identiques avec ou sans le conditionnement de la prédiction par PSIPRED.

6.2.3 Impact du traitement de la redondance

Nous avons procédé à des recherches de fragments candidats par SAFrAN sur le JV, sans traiter la redondance associée, à la fois aux hélices α et aux mots identiques. la figure 6.11 présente les cRMSds des fragments SAFrAN de 7 résidus de long obtenus selon les structures secondaires qu'ils couvrent. Nous voyons bien que les hélices α y sont trop largement représentées. Dans tous les cas, les distributions sont similaires à celles obtenues avec traitement de la redondance, nous avons juste diminué le nombre de fragments par protéine. Cette information de redondance que nous avons choisi d'éliminer pour ne pas détecter toutes les hélices α de la base de donnée contient finalement une information importante qui pourrait nous permettre de monter un indice de confiance par position, ou de tester une approche consensus comme Yang and Wang (2003).

6.2.4 Approximation globale des structures protéiques

Comme l'ont soulevé Sander et al. (2006), il faut prendre en considération que notre objectif final est d'obtenir une structure protéique à partir d'une méthode d'assemblage de fragments. Les trajectoires dérivées de SAFrAN, utilisées directement avec l'algorithme glouton développé par Tuffery and Derreumaux (2005) guidé par un critère cRMSd, permettent de reconstruire les structures natives des protéines du JV avec une précision de 2,2 Å en moyenne. Et ceci pour une complexité de 4,2 lettres compatibles (en terme de transition markovienne) par position. 90% des 322 structures du JV peuvent être recons-

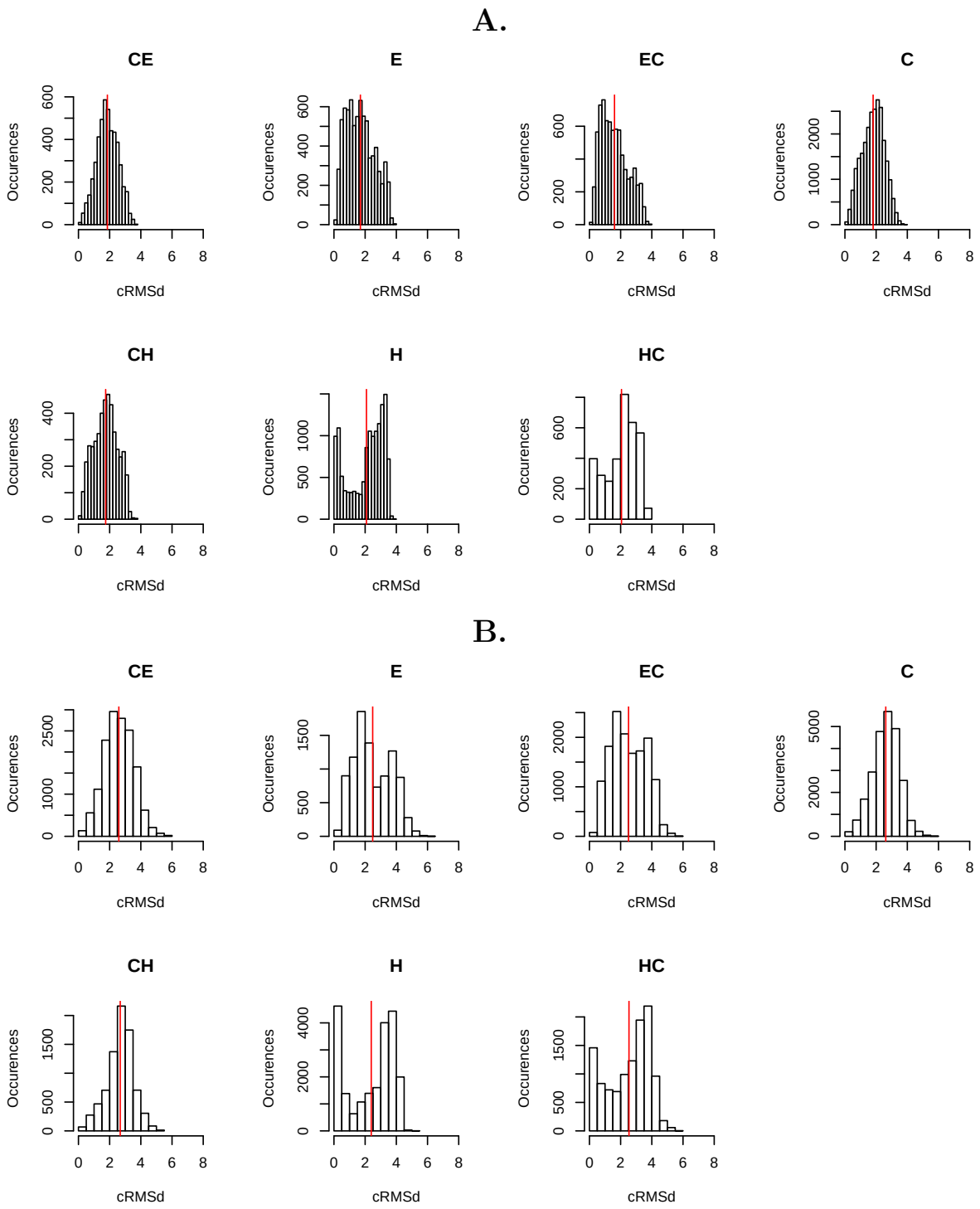


Fig. 6.10: Les fragments de taille 5 et 7 dans les structures secondaires sans utiliser PSIPRED. Pour les fragments de tailles 5 (A) et 7 (B), nous avons ici détaillé la distribution des cRMSds de ces derniers en fonction de leur localisation dans les structures secondaires. Le code choisi est issu du classique code à 3 lettres (E, *Extended*, H, *Helix* et C, *Coils*), ainsi : CE, entrée de feuillet β ; E, coeur du feuillet β ; EC, sortie de feuillet β ; C, boucle ; CH, entrée dans une hélice α ; H, coeur d'une hélice α ; HC, sortie d'une hélice α. Les valeurs moyennes sont ici mises en évidence par une ligne verticale continue rouge.

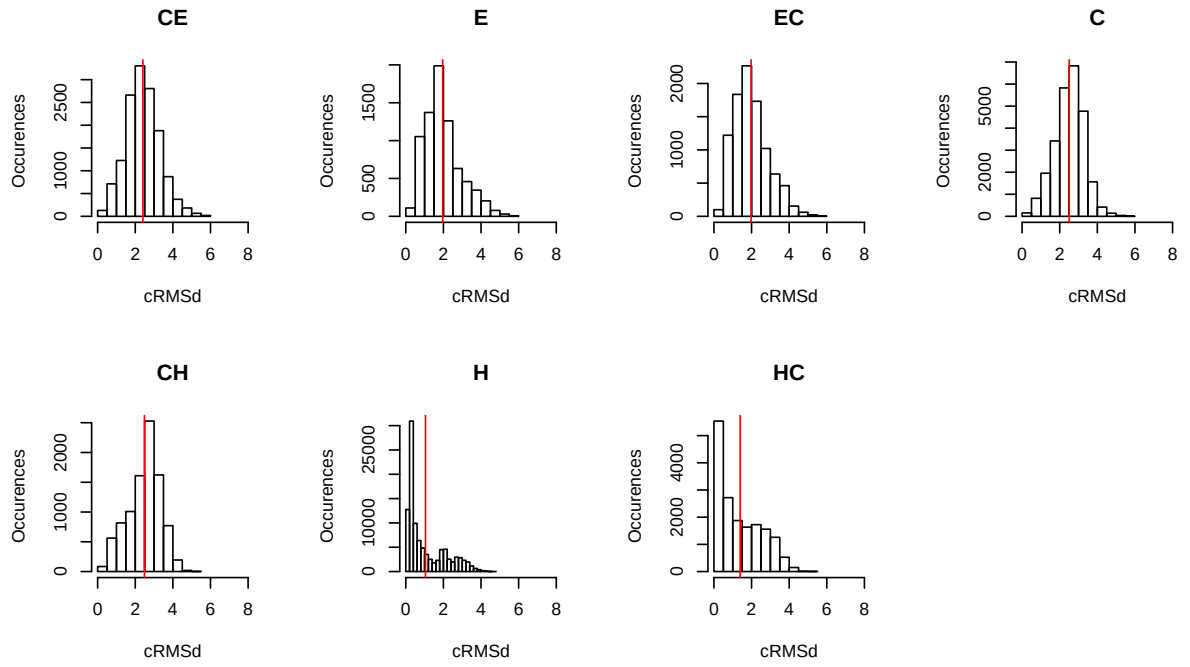


Fig. 6.11: Les fragments de taille 7 dans les structures secondaires, sans traitement de la redondance. Pour les fragments de tailles 7, nous avons ici détaillé la distribution des cRMSds de ces derniers en fonction de leur localisation dans les structures secondaires. Le code choisi est le suivant : CE, entrée de feuillet β ; E, coeur du feuillet β ; EC, sortie de feuillet β ; C, boucle ; CH, entrée dans une hélice α ; H, coeur d'une hélice α ; HC, sortie d'une hélice α. Les valeurs moyennes sont ici mises en évidence par une ligne verticale continue rouge.

truites à moins de 3,3 Å de cRMSd pour une complexité faible (figure 6.12.A). Deux exemples de protéines reconstruites par l'algorithme glouton à partir d'une prédiction SAFrAN sont présentés dans la figure 6.13 C et D.

Il est important de signaler qu'ici, nous utilisons les lettres et non les fragments complets pour guider l'assemblage. Bien que l'algorithme glouton ait été généralisé à la prise en compte de mots et non de lettres, la définition des trajectoires et en particulier les points de jonctions entre les différents mots s'est révelée particulièrement ardue. D'autre part, une procédure d'assemblage rigide de fragments semble peut être trop contrainte pour être compatible avec un schéma de prédiction *ab initio*. Pour cette raison, nous n'avons pas poussé plus avant cette exploration d'assemblage par fragments.

Nous avons aussi testé les trajectoires dérivées des predictions sans le filtre sur les séquences protéiques. Il est apparu que ces trajectoires permettent de prédire la structure de l'ensemble des cibles du JV avec un cRMSd moyen de 1.2 Å, pour une complexité moyenne de 8,3 lettres compatibles par position (figure 6.12.B), *i.e.* pour une complexité

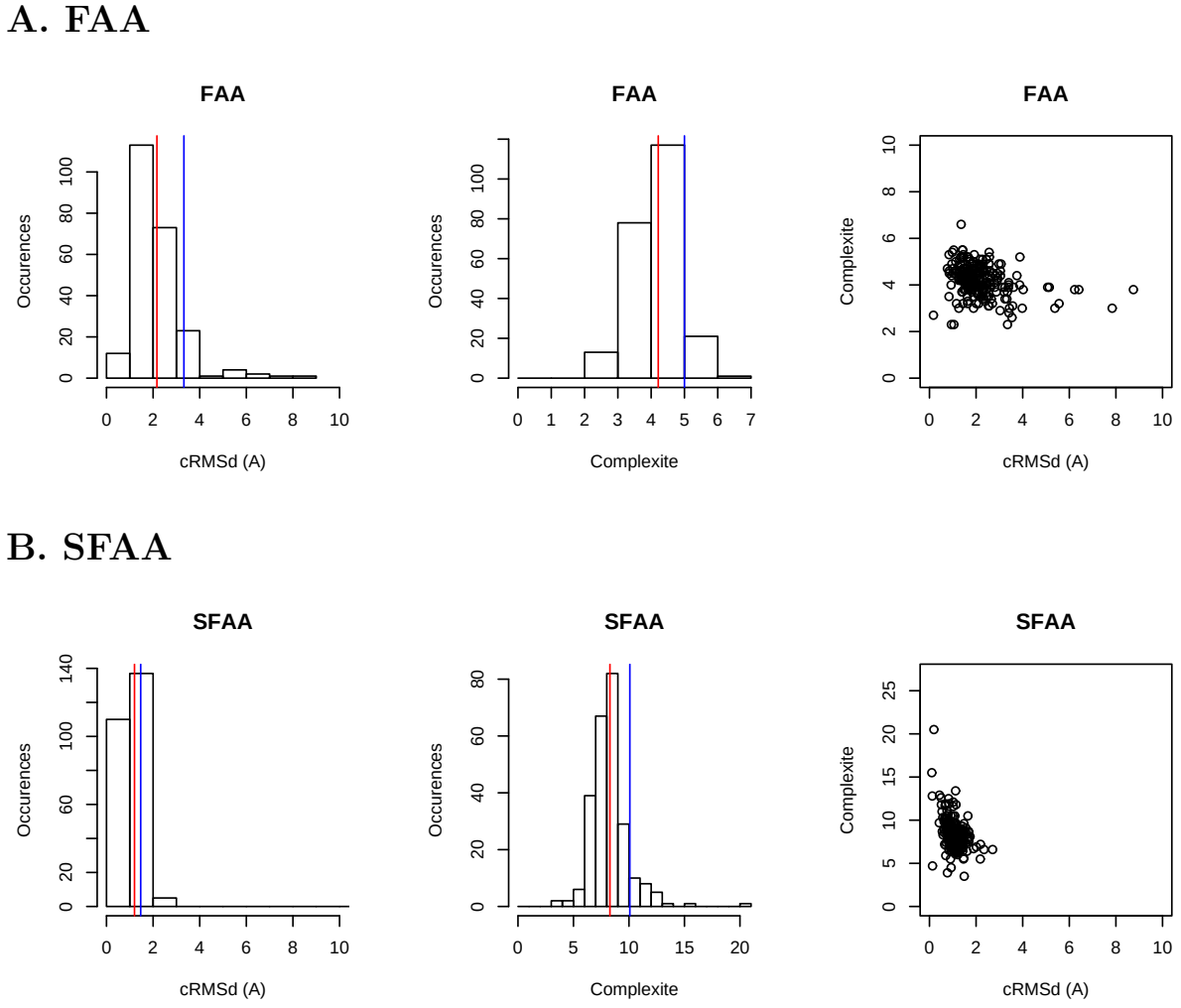


Fig. 6.12: Reconstructions à partir des trajectoires prédites. Nous avons reconstruit l'ensemble des protéines du JV, à partir des trajectoires dérivées des fragments candidats prédits par SAFrAN. Sont présentés les cRMSds des modèles obtenus, ainsi que la complexité des trajectoires données en entrée de l'algorithme glouton (FAA : Filtre acides aminés). Les lignes verticales rouges représentent les valeurs moyennes, tandis que les bleues représentent la valeur du quantile à 90% **A.** Résultats obtenus avec le Filtre sur les Acides Aminés (FAA) et **B.** Sans le Filtre sur les Acides Aminés (SFAA).

deux fois plus grande. Dans le cadre d'une prédiction réelle, nous nous réservons la possibilité d'utiliser ces trajectoires pour leur plus grande souplesse.

La figure 6.12, nous indique qu'il est des cibles de faible complexité que nous n'arrivons pas à reconstuire avec un cRMSd raisonnable. L'analyse de la reconstruction de ces cibles et de la trajectoire prédite, nous a montré, qu'elles sont le résultat d'erreur de prédiction PSIPRED, que SAFrAN n'a pu corriger. C'est le cas de la protéine 2hmcA (314 aa) pour laquelle l'hélice située dans la région 235-256 a été prédite plus longue qu'elle ne l'est en réalité. Cependant le meilleur modèle reconstruit, très proche de la structure native

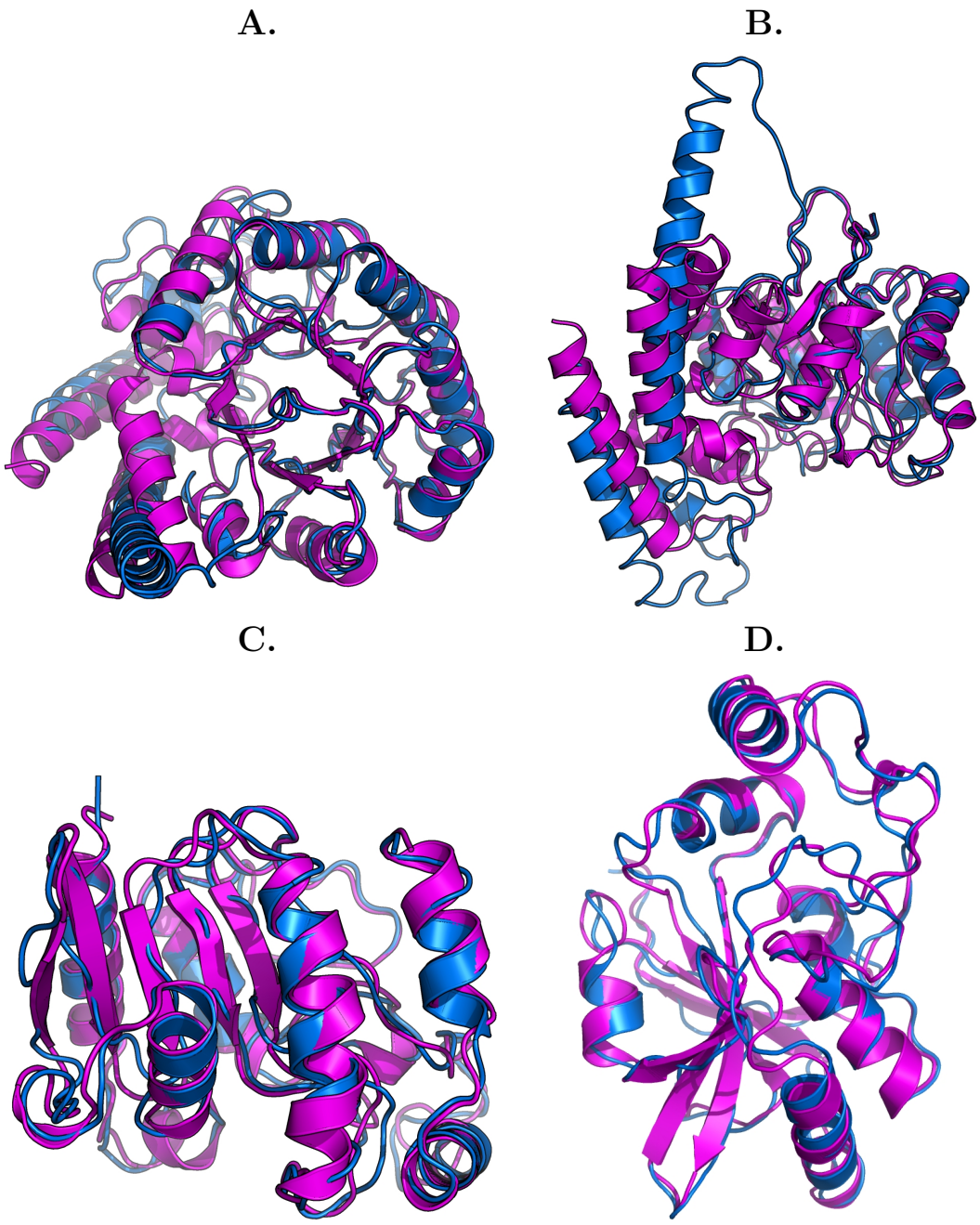


Fig. 6.13: Exemples de reconstructions à partir de prédictions SAFrAN. Le modèle reconstruit (en bleu) pour la protéine 2hmcA (en magenta) à partir de la prédiction SAFrAN présente un cRMSd de 8,7 Å avec cette dernière. Ceci est dû à une erreur de prédiction pour l'hélice 235-256. Vue du dessus (A) et de profil (B). Deux autres exemples de reconstructions (en bleu) sont présentés pour les protéines associées aux codes PDB 2d3yA (C) et 2if6A (D), distantes respectivement de 1,26 et 1,37 Å de cRMSd de leur structure native (magenta).

(voir la figure 6.13 A et B), présente un TM-score de 0,77, ce qui est très bon pour une protéine de cette taille. Notez que le TM-score (Zhang and Skolnick, 2005b) est un score de similitude 3D entre deux protéines compris entre 0 et 1 ; la valeur seuil de 0,5 indique une topologie identique des structures.

La comparaison avec d'autres méthodes semble délicate. Hunter and Subramaniam (2003a) ont obtenu, par leur méthode de prédiction bayesienne, des modèles éloignés de 10 à 50 Å de la structure native. Bien que ces structures puissent être des intermédiaires du repliement protéique, il apparaît que les résultats obtenus à partir des prédictions SAFrAN, couplées à un algorithme performant d'assemblage de fragments, sont encourageants.

6.2.5 Comparaison avec d'autres approches

Une comparaison directe avec les autres méthodes de prédiction locale de la structure des protéines est complexe. Cette difficulté réside dans le fait que chaque méthode ne partage pas le même nombre de fragments, et ces fragments n'ont pas toujours la même taille. Ceci est particulièrement vrai pour SAFrAN qui nous permet d'identifier des fragments de 3 à 15 résidus de long. Cependant, nous pouvons tenter une comparaison en terme de pertinence RMSd ($PR_{1.4}$), critère utilisé par Bystroff and Baker (1998) pour évaluer leur prédiction. Ce critère défini le pourcentage de résidus pour lesquels au moins un fragment s'éloignant de moins de 1.4 Å de la structure native a été prédit. Ce seuil peut être étendu à 1,5 Å ($PR_{1,5}$), voire à 2 Å ($PR_{2,0}$) par soucis de comparaison avec d'autres approches. Pour des fragments de taille 11, Benros et al. (2006) ont une $PR_{1.5}$ de 80.1% s'ils considèrent leurs 5 meilleurs fragments prédits. La méthode de Yang and Wang (2003), qui semble plus performante que Bystroff and Baker (1998), basée sur des fragments de 9 résidus, aboutit à une $PR_{1.4}$ de 62.1% avec un niveau de consensus de 1.

Nous ne pouvons pas comparer les résultats SAFrAN FAA de par la taille des fragments candidats détectés. Ainsi, la $PR_{1.4}$ obtenue avec SAFrAN SFAA, est de 86,6%, et la $PR_{2.0}$ est de 94,8% contre 80,1% pour Benros et al. (2006).

SAFrAN est une méthode de recherche de fragments candidats originale proposant un ensemble de fragments candidats de taille variable compatibles avec une séquence en acides aminés donnée. Sans le filtre sur les acides aminés, 71% des fragments, pour une taille moyenne de 9,3 résidus, couvrent 97,2% de la séquence protéique avec une précision inférieure à 2,5 Å.

6.3 Conclusions de l'étude

SAFrAN permet de prédire un ensemble de fragments candidats décrivant les structures avec une précision inférieure à 2,5 Å. Malgré tout un certain nombre d'améliorations peuvent encore être apportées à la méthode. Nous avons remarqué que la redondance filtrée était une information que nous pourrions dériver en indice de confiance, ce qui manque

pour le moment à la méthode. Par ailleurs, le filtre appliqué aux séquences protéiques des fragments ne semble être pertinent que pour des fragments de taille supérieure à 7 résidus, et a pour conséquence de diminuer la taille moyenne des fragments prédits. Nous allons donc modifier le comportement de ce filtre dans les prochaines versions de SAFrAN.

Tuffery et al. ont déjà entrepris des travaux pour améliorer les performances de la méthode de prédiction. Cette version ne tardera pas à être intégrée à SAFrAN pour en améliorer la performance.

Les perspectives principales sont doubles : (i) nous savons que cet ensemble de fragments candidats prédits peut décrire avec précision des conformations locales. Cependant nous n'en avons pas encore pu en exploiter le potentiel. Le point sur lequel nous nous penchons à l'heure actuelle est de pouvoir utiliser directement cet ensemble de fragments dans une méthode de reconstruction pour restreindre significativement l'espace conformationnel, (ii) enfin, SAFrAN pourrait être utilisé pour aider à la résolution de structures expérimentales obtenues à faible résolution, ou en cas de données expérimentales incomplètes.

Troisième partie

OPEP : un potentiel énergétique pour guider le repliement protéique.

Dans une optique de prédiction *ab initio*, nous avions besoin d'un potentiel générique pour évaluer la pertinence des modèles générés. Le potentiel OPEP semblait un choix approprié par sa performance, et, de par le fait que ce soit un modèle gros grain, donc moins couteux en terme de ressources computationelles que des modèles tous atomes. Dans cette partie, nous allons exposer les différentes étapes de l'implémentation du champ de force gros grain OPEP dans l'algorithme glouton pour générer des modèles protéiques.

La mécanique d'assemblage dans l'algorithme glouton est basée uniquement sur les coordonnées des carbones alpha, ainsi, le premier problème à résoudre a été d'étendre la représentation des résidus au modèle gros grain. En parallèle de la méthode de génération de modèles protéiques, ce point a donné lieu au développement de SABBAC, une méthode de reconstruction de structures protéiques complètes à partir de la trace (ensemble des carbones α) d'une protéine.

Afin d'améliorer le pouvoir de reconnaissance d'OPEP, nous avons ensuite entrepris de l'optimiser, afin qu'il puisse mieux distinguer le bassin énergétique de la structure native de celui de conformations non natives.

Et enfin, la discrétisation de l'espace conformationnel, inhérante à l'utilisation d'un alphabet structural combiné à une procédure d'assemblage rigide, nous a conduit à modifier la formulation et quelques paramètres de la dernière version d'OPEP implémentée dans l'algorithme glouton.

Adaptation de l'algorithme de reconstruction au modèle gros grain

Dans le potentiel OPEP, les atomes de la chaîne principale sont représentés explicitement, tandis que les chaînes latérales sont représentées par une sphère dont le rayon et la position dépendent de la nature du résidu considéré. A partir d'un ensemble de coordonnées que sont les carbones alpha, il nous faut donc générer une structure protéique complète. C'est un problème classique qui a été beaucoup étudié par le passé que nous détaillerons dans le chapitre suivant.

La génération des coordonnées se déroule en deux étapes : (i) reconstruction des liaisons peptidiques à partir de la trace des carbones alpha, et (ii) positionnement des chaînes latérales (voir la figure 6.14).

La chaîne principale

Les coordonnées des atomes de la liaison peptidique sont directement issues des 155 prototypes de l'alphabet structural HMM-SA. Ainsi, la matrice de transformation (calculée sur les carbones α) utilisée pour ajouter un nouveau prototype au modèle en croissance,

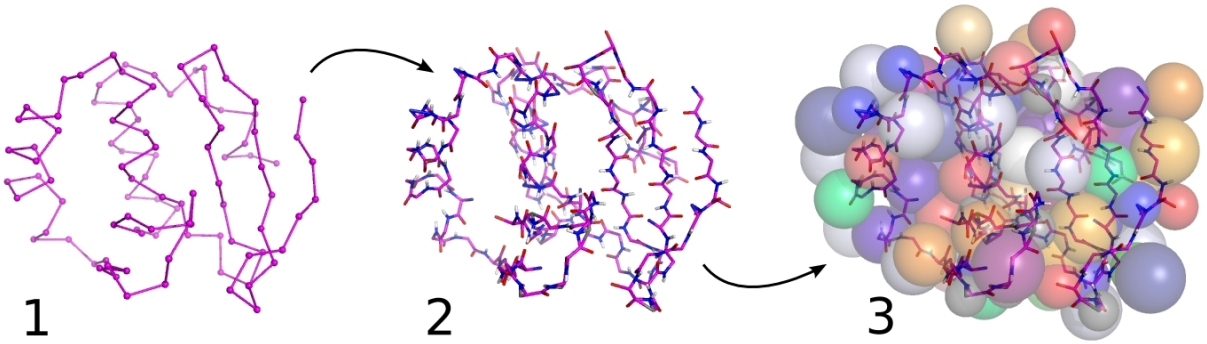


Fig. 6.14: Les étapes nécessaires à l'implémentation de sOPEP dans l'algorithme glouton. L'algorithme glouton étant une méthode de reconstruction basée sur les carbones alpha, il a été nécessaire d'adapter la représentation des modèles dans ce dernier pour y implémenter sOPEP. Dans un premier temps, les liaisons peptidiques sont reconstruites, puis, les sphères représentant les chaînes latérales sont positionnées.

est appliquée aux atomes lourds (N_i, C'_i et O_i) de ce même prototype, mais aussi au groupement carbonyle (C'_{i-1}, O_{i-1}) du résidu précédent dans le cas d'une reconstruction de l'extrémité N vers C terminale. Cette procédure permet de garantir la fermeture de la liaison peptidique, cependant, elle a tendance à éloigner la longueur de liaison CA_{i_1}–C'_{i-1} de sa valeur canonique. C'est cependant la solution qui est apparue comme la plus satisfaisante dans le cadre d'un assemblage rigide pour former un réseau de liaisons hydrogène satisfaisant.

Les coordonnées de l'atome d'hydrogène HN sont générées par l'intermédiaire d'un repère local similaire à celui précédemment décrit pour le positionnement des CL (section 5.1). Néanmoins, ce repère est défini ici par les atomes C'_{i-1}, N_i et CA_i, si l'on considère le positionnement de l'atome HN_i. De ce fait, l'extrémité N terminale du modèle est dépourvue d'hydrogène.

Les chaînes latérales

Les sphères représentant les chaînes latérales sont ensuite positionnées avec la même méthode et les mêmes paramètres que ceux décrits lors de la présentation d'OPEP (voir l'introduction).

Cette adaptation de la méthode de reconstruction est un point crucial de la méthode. En effet, les centroïdes étant positionnés conditionnellement aux coordonnées des atomes lourds de la chaîne principale de chaque résidu, il est essentiel que ces dernières soient pertinentes pour permettre (i) de former un réseau de liaisons hydrogène compatible avec la trace du modèle (formation des structures secondaires), et (ii) des interactions cohérentes entre les chaînes latérales.

Nous allons maintenant vous présenter SABBAC, méthode dérivée de ces travaux d'adaptation de l'algorithme glouton au champ de force gros grain OPEP.

Chapitre 7

SABBAC : on line Structural Alphabet-based BackBone reconstruction from Alpha-Carbon trace

Il existe un certain nombre de champs d'applications dans lesquels la structure des protéines, représentée par des modèles gros grain, aurait besoin d'être étendue à un modèle tous atomes. C'est le cas par exemple des méthodes *ab initio* ou *de novo* de génération de modèles protéiques, ou d'analyse en modes normaux. Ce besoin peut aussi se faire sentir dans le cadre du raffinement de données expérimentales obtenues à faible résolution.
Partant d'un modèle gros grain d'une structure protéique, qui nous permet d'avoir une information sur la position approximative des résidus dans l'espace, la génération d'une structure protéique complète est décomposée en deux étapes : (i) générer les coordonnées des atomes de la chaîne principale, et (ii) positionner les chaînes latérales. La première étape peut être réalisée en partant d'une trace de carbones α, et c'est d'ailleurs le cas le plus étudié (Purisima and Scheraga, 1984; Jones and Thirup, 1986; Reid and Thornton, 1989; Claessens et al., 1989; Correa, 1990; Holm and Sander, 1991; Levitt, 1992; Rey and Skolnick, 1992; Payne, 1993; Liwo et al., 1993; van Gelder et al., 1994; Mathiowetz and Goddard, 1995; Gan et al., 1996; Milik et al., 1997; van Hooft and Holtje, 2000; Feldman and Hogue, 2000; Kazmierkiewicz et al., 2002; Iwata et al., 2002; Adcock, 2004). Parmi les différentes approches proposées, certaines explorent l'espace conformationnel de l'unité peptidique pour produire une chaîne principale complète. C'est le cas par exemple de la méthode développée par Payne (Payne, 1993) qui détermine la rotation optimale de l'unité peptidique en utilisant un potentiel de force moyenne dépendant des résidus adjacents. La plupart des méthodes, comme Holm and Sander (1991) et Adcock (2004), repose sur l'assemblage de fragments extraits de librairies issues de structures connues. Ces fragments sont ensuite assemblés pour produire une chaîne peptidique complète en ajustant au mieux la trace en carbones α. Ces méthodes utilisent un critère énergétique

ou géométrique pour guider la recherche. Les méthodes basées sur des fragments sont en général confrontées à la limitation suivante : pour augmenter la pertinence de la méthode, il est nécessaire de maintenir une collection large et à jour de fragments. Le serveur MaxSprout (Holm and Sander, 1991) a été un des premiers serveurs proposés pour s'attaquer à ce genre de problématique ; il est d'ailleurs régulièrement mis à jour à partir des structures protéiques disponibles.

Dans cette partie, nous allons introduire SABBAC (*on line Structural Alphabet-based BackBone reconstruction from Alpha-Carbon trace*), une nouvelle approche basée sur une méthode de sélection de fragments et leur assemblage pour reconstruire une structure protéique complète à partir de la seule définition de la position des carbones alpha. SABBAC utilise l'encodage de la trace des carbones *alpha* dans l'espace de l'alphabet structural HMM-SA (Camproux et al., 2004), pour sélectionner à chaque position de la structure, un sous-ensemble de fragments candidats parmi les 155 décrivant toutes les lettres de l'alphabet structural (*i.e.* permettant de décrire toutes les conformations de toutes les structures protéiques). Ces fragments sont ensuite assemblés en utilisant l'algorithme glouton décrit en introduction, guidé par un critère énergétique de sOPEP v1.0.

7.1 Méthodes

7.1.1 Librairie de fragments dépendante de la structure

Etant donné que nous avons l'ensemble des coordonnées des carbones alpha de la structure à reconstruire, il nous est possible de définir à chaque position les lettres de l'alphabet structural HMM-SA la décrivant au mieux par l'algorithme de Viterbi. En ne sélectionnant qu'une seule lettre décrivant la conformation de chaque fragment de quatre carbones alpha consécutifs, le nombre moyen de fragments envisagé est de 5 à 6 par position.

7.1.2 Reconstruction de la liaison peptidique

Pour générer rapidement les coordonnées des atomes N, HN, C' et O de la liaison peptidique, nous avons suivi la procédure décrite par Milik et al. (1997). Les coordonnées sont précalculées dans un repère local défini par trois carbones alpha consécutifs. Comme les prototypes associés aux différentes lettres de l'alphabet structural HMM-SA ont une longueur de quatre résidus, nous pouvons ainsi définir deux repères différents pour ces derniers. Pour chaque prototype, nous pouvons choisir la première, la deuxième ou la troisième liaison peptidique. Il s'est avéré que la liaison peptidique centrale, dans le repère défini par les trois derniers carbones alpha, était la plus stable et la plus pertinente (données non présentées). Pour les positions N et C terminales de la chaîne principale,

nous utilisons respectivement la première et la dernière liaison peptidique dans ce même repère local.

7.1.3 A la recherche d'une combinaison idéale de fragments

Une fois l'ensemble des fragments candidats déterminé à chaque position, nous utilisons l'algorithme glouton (Tuffery et al., 2005; Tuffery and Derreumaux, 2005) pour rechercher leur combinaison idéale. Brièvement, à chaque position nous reconstruisons toutes les combinaisons de liaisons peptidiques, ces solutions sont alors triées selon un critère énergétique (décrit ci-dessous), et nous ne conservons que les 10 solutions de plus basse énergie pour l'itération suivante. Cette taille de pile retenue a été apprise de reconstructions de structures connues. A la différence de l'algorithme glouton original, la structure croît ici de l'extrémité N vers C terminale, sans itérations.

7.1.4 La fonction d'énergie

Pour guider la recherche, nous utilisons un critère énergétique dérivé d'OPEP (Maupetit et al., 2007). Historiquement, c'est la première version de sOPEP que nous avons implémenté pour guider l'algorithme glouton. A cette époque, l'optimisation du champ de force n'avait pas encore abouti, donc aucun poids n'était appliqué sur chacun des termes énergétiques. Cependant, les positions des centroïdes et les valeurs de rayons utilisées sont les valeurs optimisées identiques aux versions 3.0 d'OPEP (voir la table 8.1 pour les valeurs).

La formulation de cette version de la fonction d'énergie est la suivante :

$$E = E'_{non-liante} + E_{liaisons-H} + E'_{locale} \tag{7.1}$$

$$\begin{aligned} E'_{non-liante} \;=\; & \sum_{1,4} E_{VdW} + \sum_{CP',CP'} E_{VdW} + \sum_{CP',C\alpha} E_{VdW} \\ & + \sum_{CP,CL} E_{VdW} + \sum_{CL,CL} E_{VdW} \end{aligned} \tag{7.2}$$

$$E'_{locale} = E_{PhiP} + E_{VC\alpha} + E_{Trans} \tag{7.3}$$

Pour les besoins de SABBAC, nous avons modifié la formulation originale de sOPEP v1.0 : (i) Le terme énergétique associé aux liaisons hydrogène ($E_{liaisons-H}$) est conservé dans sa formulation originale, (ii) la formulation de l'énergie non liante ($E'_{non-liante}$) est ici modifiée pour ignorer les interactions entre carbones alpha, non discriminantes puisque les coordonnées de la trace sont identiques pour toutes les solutions, et (iii) les interactions

locales sont spécifiques à SABBAC. Cette dernière est composée de trois termes que sont E_{PhiP} pour la contribution des angles dièdres ϕ dans leur valeurs positives,

$$E_{PhiP} = \sum_{1=1}^{N} E^i_{PhiP} \tag{7.4}$$

$$E^i_{PhiP} = \begin{cases} 0, \text{ si } \phi < 0 \\ \text{si } \phi > 0 \begin{cases} +0,5 \text{ kCal/mol} \\ -0,3 \text{ kCal/mol, si le résidu } i \text{ est une glycine.} \end{cases} \end{cases} \tag{7.5}$$

$E_{V_{C\alpha}}$, permettant de garantir que la valence des carbones α ne dévie pas trop de sa valeur canonique ($V^{ref}_{C\alpha} = 110^\circ$), la constante de force k ayant été fixée à $1,25 * 10^{-3}$ kCal/(mol.degre),

$$E_{V_{C\alpha}} = \sum_{1=1}^{N} E^i_{V_{C\alpha}} \tag{7.6}$$

$$E^i_{V_{C\alpha}} = \begin{cases} 0, \text{ si } (V_{C\alpha} - V^{ref}_{C\alpha}) < 20 \\ k * (V_{C\alpha} - V^{ref}_{C\alpha})^2 \text{ , sinon.} \end{cases} \tag{7.7}$$

et enfin, E_{Trans} la pseudo énergie de transition entre les prototypes de l'alphabet structural HMM-SA,

$$E_{Trans} = \sum_{1=1}^{N-1} -log(\ p_{\ i,\ i+1}\ ) \tag{7.8}$$

avec $p_{\ i,\ i+1}$ la probabilité que le prototype sélectionné à la position $i+1$ suive le prototype sélectionné à la position i. Les prototypes sont ici les 155 sous-conformations des lettres de l'alphabet structural HMM-SA. L'ensemble de ces probabilités a été estimé sur un jeu de protéines non redondantes issues de PDB, partageant moins de 30% d'identité de séquence et encodées dans l'espace de l'alphabet structural HMM-SA.

7.2 Résultats - Discussion

7.2.1 Performance de la méthode

Nous avons évalué les performances de SABBAC sur un nombre important de structures. Quelques résultats sont présentés dans la table 7.1. Dans la partie supérieure de la table sont présentés les résultats obtenus pour un jeu de structures discuté par Adcock (2004) permettant une comparaison avec des approches antérieures. La partie inférieure de la table, quant à elle, contient les résultats obtenus pour de nouvelles entrées de la PDB, n'ayant donc pas servi à l'apprentissage de notre jeu de fragments.

PDB	L	SABBAC	MaxSprout	bb
Jeu utilisé par Adcock				
111m	154	0,21	0,42	0,91
1ctf	68	0,33	0,73	0,85
1igd	61	0,37	0,44	0,68
1omd	107	0,40	0,41	0,77
1sema	58	0,48	0,34	1,00
1tima	247	0,58	0,60	0,97
1ubq	76	0,34	0,38	0,96
2cts	437	0,38	0,45	0,86
2lym	129	0,38	0,44	0,98
2mhr	118	0,44	0,54	0,88
2pcy	99	0,44	0,54	0,91
2wrp	104	0,29	0,42	0,87
4pti	58	0,37	0,44	0,81
5nll	138	0,38	0,46	0,85
μ	132	0,38	0,47	0,89
σ	101	0,09	0,10	0,08
Jeu de nouvelles entrées PDB				
1pxza	346	0,48	0,54	0,96
1rkia	101	0,56	0,44	0,88
1s7la	177	0,34	0,36	0,86
1t70a	255	0,44	0,50	0,95
1txoa	235	0,39	0,38	0,96
1v0ed	666	0,51	0,45	0,89
1v7ba	175	0,25	0,41	0,87
1vb5b	255	0,30	0,42	0,84
1vkca	149	0,34	0,33	0,82
1vr4a	103	0,46	0,59	1,00
1vr9a	121	0,45	0,45	0,79
1wmha	83	0,27	0,28	0,82
1wpbg	168	0,42	0,35	0,86
1wmia	88	0,34	0,42	0,81
1x6ja	88	0,40	0,36	0,76
1xb9a	108	0,48	0,51	0,81
1xe0b	107	0,58	0,62	0,90
μ	190	0,41	0,44	0,88
σ	144	0,10	0,09	0,07
Ensemble				
μ	**164**	**0,40**	**0,45**	**0,87**
σ	**128**	**0,09**	**0,10**	**0,07**

Tab. 7.1: Les performances de SABBAC - Comparaison avec d'autres méthodes. La qualité de la chaîne principale reconstruite est ici estimée en calculant le RMSd (en Å) de la chaîne principale reconstruite *versus* native (tous les atomes lourds sont ici considérés). Deux jeux de données sont présentés, le premier est un jeu discuté par Adcock dans Adcock (2004), et le second est composé de récentes entrées de la *Protein Data Bank* (Berman et al., 2000a). Sont aussi présentés les résultats obtenus par les méthodes MaxSprout (Holm and Sander, 1991) et bb (Adcock, 2004).

Pour chaque protéine, à titre de comparaison, sont indiqués les résultats obtenus par MaxSprout (Holm and Sander, 1991) et *bb* (Adcock, 2004) ; pour MaxSprout, nous avons utilisé le serveur en ligne disponible à l'adresse suivante : www.ebi.ac.uk/maxsprout, et pour *bb*, nous avons téléchargé le programme éponyme disponible sur la toile à l'adresse mccammon.ucsd.edu/˜adcock/bb.html. Pour ce dernier, sont présentés les résultats bruts, *ie* sans minimisation ultérieure.
SABBAC apparaît globalement légèrement meilleur que les deux approches MaxSprout et *bb* aussi bien sur le jeu utilisé par Adcock que sur les nouvelles entrées PDB, et ce pour des protéines de tailles allant de 60 à 600 résidus.

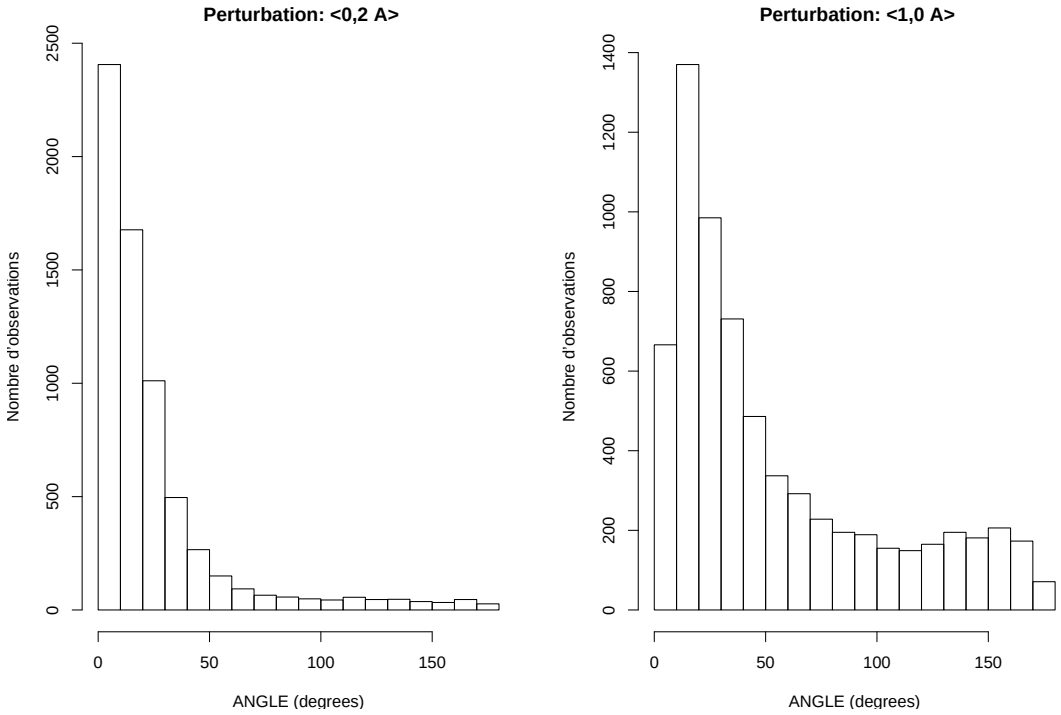


Fig. 7.1: Deviation du plan de la liaison peptidique. Sont présentées ici les valeurs de déviation du plan de la liaison peptidique reconstruite par rapport à la structure native dans le cas de traces en carbones *alpha* perturbées en moyenne de 0,2 Å (à gauche) à 1.0 Å (à droite).

Nous avons aussi testé les performances de SABBAC pour des traces en carbones *alpha* perturbées. Pour ce faire, nous avons perturbé les traces en carbones *alpha* des structures de la table 7.1 en moyenne de 0,2 Å et de 1 Å. La figure 7.1 présente les distributions des angles des plans peptidiques des structures reconstruites à partir de ces traces perturbées avec le plan peptidique de la structure native. Même pour des traces perturbées en moyenne de 1 Å, SABBAC reste performant. Nous insistons sur le fait qu'une orientation correcte du plan peptidique permet un positionnement correct des chaînes latérales.

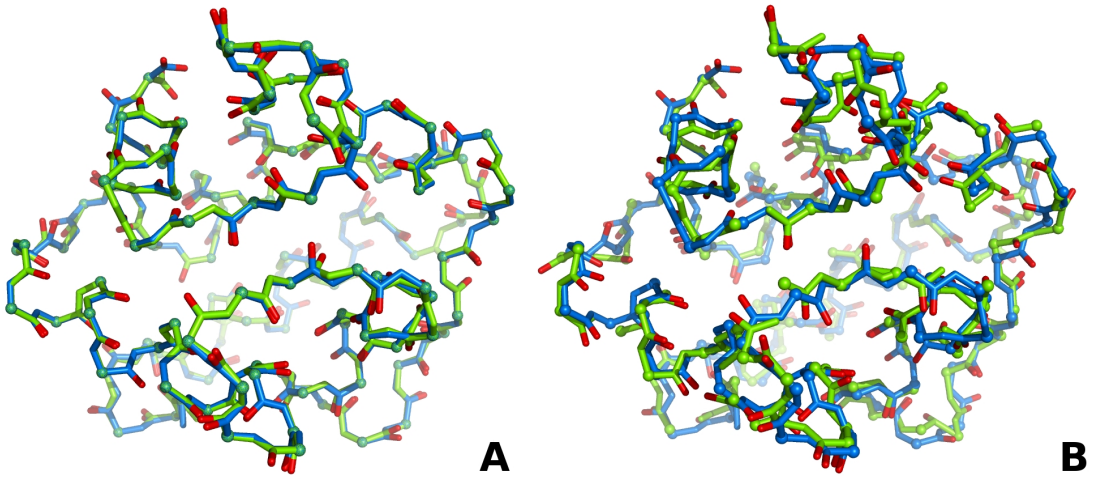


Fig. 7.2: Un exemple de reconstruction : l'oncomoduline (1omd). A. Les structures native et reconstruite par SABBAC superposées. **B.** Structure native dont la trace a été perturbée de 0,8 Å en moyenne et structure reconstruite par SABBAC, superposées. La structure native est en bleu et la structure reconstruite en vert. Les oxygènes sont en rouge dans les deux structures.

La figure 7.2, montre un exemple de reconstruction SABBAC dans le cas de l'oncomoduline (code PDB 1omd), pour la trace native (figure 7.2A) et pour une trace perturbée en moyenne de 0,8 Å (figure 7.2B). L'orientation des atomes d'oxygène (en rouge sur la figure) de la chaîne principale nous permet de voir la pertinence des reconstructions SABBAC pour les traces natives et perturbées.

Enfin, nous avons aussi considéré la reconstruction de séries de modèles issus de l'expérience CASP6 (Moult et al., 2005). Parmi l'ensemble des cibles, nous avons retenu toutes celles correspondant à des cibles complètes (les domaines sont exclus) ne comportant pas de fragments manquants. Pour chaque cible, nous avons considéré le meilleur modèle et le modèle de rang 5, selon la classification par le score GDT_TS. Les modèles incomplets ayant été ignorés, cela résulte en jeu de 31 cibles (60 modèles) se répartissant dans les catégories CASP comme suit : 14 cibles de modélisation par homologie, 13 cibles de reconnaissance de repliement, et 4 cibles de nouveau repliement. Pour chaque modèle, nous avons reconstruit sa chaîne principale à l'aide de SABBAC et de MaxSprout. Pour MaxSprout, nous n'avons obtenu de résultats que pour 57 modèles. La figure 7.3 trace, pour chaque modèle reconstruit par les deux méthodes, la fraction des plans peptidiques inférieure à respectivement 10 et 40 degrés par rapport à la structure native. Ces valeurs sont tracées en fonction de la pertinence du modèle par rapport à la structure native, évaluée à l'aide du TM-score (Zhang and Skolnick, 2005b). Pour rappel, une valeur de TM-score de 1 induit une correspondance parfaite entre le modèle et la structure native, et plus la valeur de ce score est basse, plus le modèle s'éloigne de la structure native.

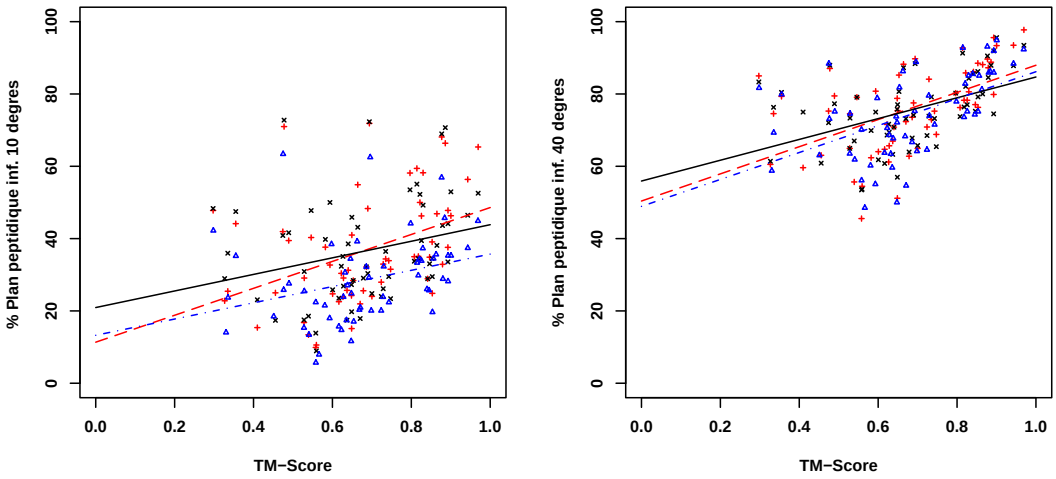


Fig. 7.3: Performance des reconstruction MaxSprout et SABBAC sur des modèles issus de l'expérience CASP6. Pour les 31 cibles considérées, nous avons sélectionné le meilleur modèle et le modèle de rang 5. Les modèles pour lesquels la fraction du plan peptidique dévie de moins de 10 degrés par rapport à la structure native de référence sont présentés à gauche, et de moins de 40 degrés à droite. Ces proportions sont tracées en fonction du TM-score du modèle considéré. Les + correspondent aux modèles CASP6, les triangles bleus aux reconstructions MaxSprout, et les **x** aux reconstructions SABBAC. Les lignes correspondent aux regressions linéaires obtenues pour chaque méthode.

Comme nous pouvons le constater, SABBAC semble plus performant que MaxSprout pour les déviations inférieures à 10 degrés. L'écart entre MaxSprout et SABBAC diminue pour des déviations inférieures à 40 degrés. Si l'on compare SABBAC aux méthodes ayant généré les modèles CASP, il semble que SABBAC propose de meilleurs résultats pour des TM-scores faibles, ce qui suggère une certaine robustesse de la méthode face à des perturbations importantes de la chaîne principale. A *contrario*, pour des modèles très proches de la structure native (TM-scores proches de 1,0), les modèles CASP semblent plus pertinents. Cependant, il faut aussi prendre en compte le fait que le processus de génération des modèles diffère grandement, SABBAC ne procédant à aucun raffinement de la structure.

7.2.2 Le service en ligne

Le serveur SABBAC intègre toutes les étapes nécessaires pour générer une structure protéique complète à partir d'une trace en carbones alpha au format PDB. Dans la version actuelle, tous les atomes autres que les carbones *alpha* sont ignorés. Les groupements hétéro "classiques" telle que la sélénométhionine, etc, sont maintenant pris en compte et remplacés par leurs équivalents canoniques. Le service ayant été connecté à une version améliorée de SCit (Gautier et al., 2004), l'utilisateur peut choisir de recontruire ou non les

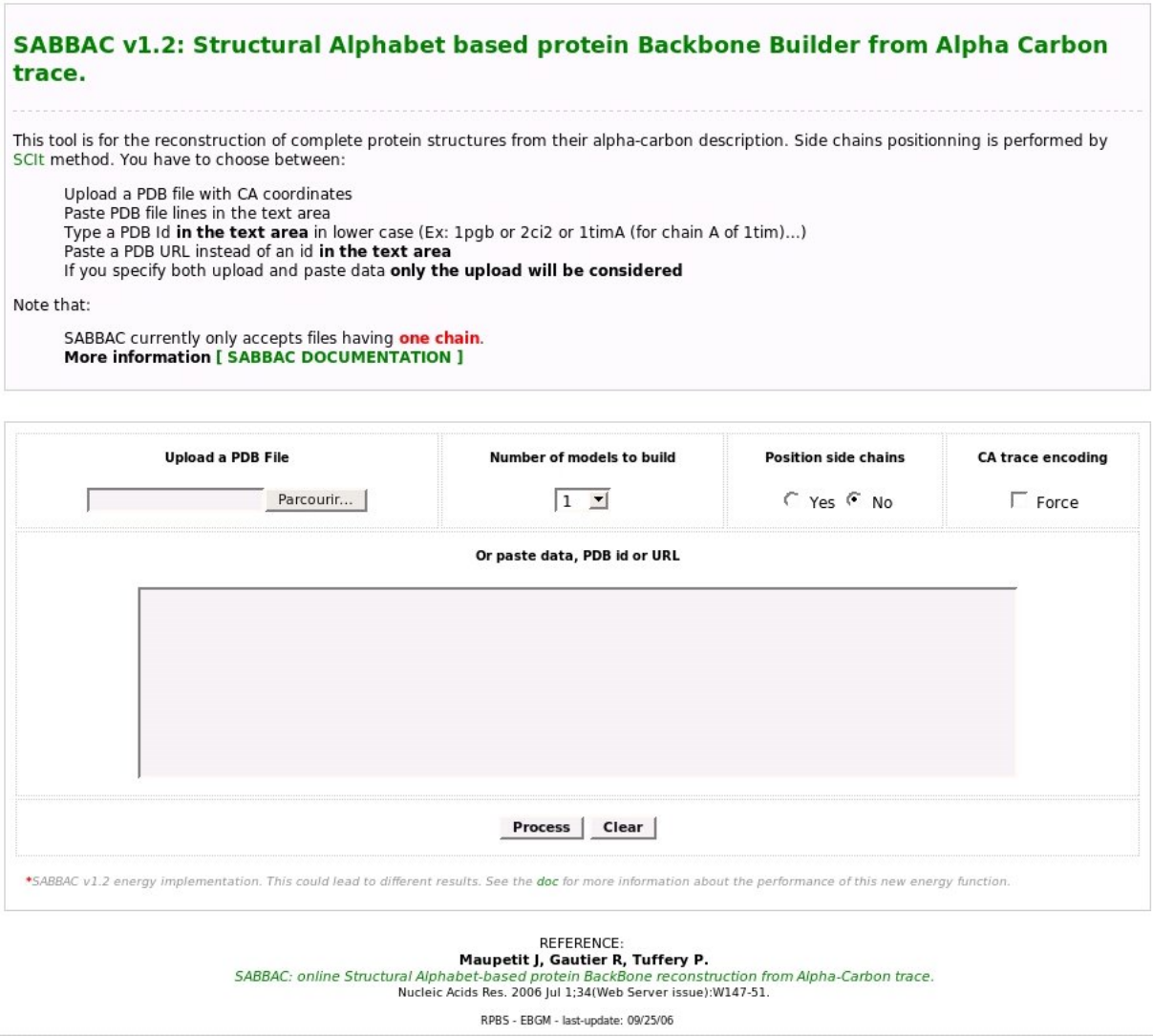
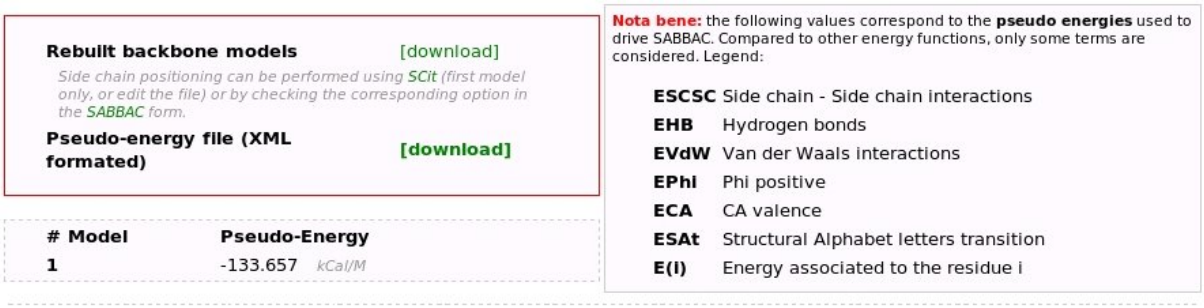


Fig. 7.4: Le formulaire SABBAC. Voici une capture d'écran du service en ligne de la méthode SABBAC disponible à l'adresse : bioserv.rpbs.jussieu.fr/cgi-bin/SABBAC. Dans cette interface, l'utilisateur peut envoyer ses données *via* un fichier ou par copier-coller. Il peut choisir le nombre de modèles qu'il veut que SABBAC génère pour lui, et s'il veut que les chaînes latérales soient reconstruites. Enfin, pour les traces incorrectes dont les carbones *alpha* consécutifs sont éloignés, l'utilisateur peut cocher l'option "CA trace encoding : Force".

chaînes latérales. Cette version rapide de SCit sélectionne la conformation la plus probable d'une chaîne latérale étant donné la conformation de la chaîne principale, en ignorant les conformations de chaînes latérales entrant en collision entre elles ou avec la chaîne principale. Par ailleurs, comme la pile de l'algorithme glouton est de 10, l'utilisateur peut choisir d'obtenir le meilleur (rang 1) ou les N meilleurs modèles, N étant inférieur à 10. Le résultat de SABBAC (voir figure 7.5) est un fichier PDB contenant les coordonnées atomiques de la chaîne principale (et des chaînes latérales). Pour chaque reconstruction sont fournis les détails énergétiques par modèle et par résidu sous la forme d'un fichier XML.



Rebuilt backbone models [download]

Side chain positioning can be performed using SCit (first model only, or edit the file) or by checking the corresponding option in the SABBAC form.

Pseudo-energy file (XML formated) **[download]**

# Model	Pseudo-Energy
1	-133.657 *kCal/M*

Nota bene: the following values correspond to the **pseudo energies** used to drive SABBAC. Compared to other energy functions, only some terms are considered. Legend:

ESCSC	Side chain - Side chain interactions
EHB	Hydrogen bonds
EVdW	Van der Waals interactions
EPhi	Phi positive
ECA	CA valence
ESAt	Structural Alphabet letters transition
E(i)	Energy associated to the residue i

#Model	#Res	ESCSC	EHB	EVdW	EPhi	ECA	ESAt	E(i)	E(i) < 0	E(i) > 0
0	0	0.033	-3.705	-2.281	0.000	0.000	0.000	-5.953		
0	1	0.055	-0.001	-2.511	0.000	0.000	0.027	-2.429		
0	2	-1.641	-4.600	-1.924	0.000	0.000	0.112	-8.054		
0	3	-2.238	-1.782	-3.113	0.000	0.000	0.194	-6.939		
0	4	-0.496	-0.011	-2.169	0.000	0.000	0.027	-2.649		
0	5	0.486	-1.368	-1.723	0.000	0.000	0.000	-2.606		
0	6	0.361	-1.138	-1.953	0.000	0.000	0.000	-2.730		
0	7	0.158	-1.345	-2.257	0.000	0.000	0.000	-3.444		
0	8	-0.503	-1.449	-2.072	0.000	0.000	0.000	-4.024		
0	9	0.210	-2.562	-2.117	0.000	0.000	0.000	-4.470		
0	10	0.144	-1.617	-2.388	0.000	0.000	0.000	-3.861		
0	11	0.449	-1.909	-2.769	0.000	0.000	0.000	-4.229		
0	12	-1.532	-2.849	-3.895	0.000	0.000	0.000	-8.276		
0	13	0.142	-1.377	-2.465	0.000	0.000	0.000	-3.700		
0	14	-1.318	-0.480	-1.965	0.000	0.000	0.179	-3.585		
0	15	-0.328	-0.564	-2.176	0.000	0.000	0.068	-3.001		
0	16	-0.863	-1.408	-2.576	0.000	0.000	0.000	-4.847		

Fig. 7.5: Un exemple de résultat de SABBAC. Les détails énergétiques peuvent être téléchargés sous la forme d'un fichier XML ou analysés directement sur la page de résultats. La partie droite indique la contribution relative du résidu i dans l'énergie du modèle.

Du point de vue du temps d'exécution, en utilisant l'algorithme glouton, le temps de calcul augmente linéairement avec la taille de la protéine à reconstruire. Ceci est maintenant d'autant plus vrai que le calcul de l'énergie a été optimisé pour être lui aussi linéaire (dans la version 1.2 actuellement en service). Pour les protéines de petite taille (moins de 60 résidus), le calcul prend quelques secondes, et peut prendre quelques minutes pour des protéines plus grosses, ces valeurs dépendant aussi de la charge du serveur.

7.3 Conclusions de l'étude

Dans sa version courante, SABBAC apparaît comme étant une bonne alternative au serveur de référence MaxSprout qui a été pionnier en la matière. En moyenne, SABBAC fournit une reconstruction de la chaîne principale plus pertinente, et ses performances pour des traces perturbées restent stables. C'est un point fort de SABBAC qui apparaît capable de fournir une reconstruction pertinente pour des traces dégénérées. Dans

certains cas, MaxSprout peut fournir des structures avec des parties manquantes si les conformations sont trop éloignées de celles observées dans sa banque de fragments. De plus, les performances de SABBAC reposent sur un ensemble très restreint de fragments dépendant de l'encodage de la protéine cible dans l'espace de l'alphabet structural HMM-SA. Enfin, SABBAC reconstruit des structures complètes dans un temps raisonnable.
Il est intéressant de noter que Gront et al. (2007) ont récemment développé une nouvelle méthode de reconstruction de structures protéiques complètes à partir de la seule trace en carbones α de cette dernière. Cette méthode autorise la déviation des carbones α et permet de reconstruire des structures éloignées en moyenne de 0,42 Å de cRSMd de la structure native.

Chapitre 8

Optimisation d'OPEP

Lors de ma thèse, mon but a été de tester OPEP sur un nombre significatif de protéines, et de déterminer sa capacité à pouvoir distinguer des conformations natives ou proches de la structure native (PSN) de structures non natives. Dans ce chapitre, nous allons présenter le modèle gros grain, ainsi que sa formulation, par rapport à la version originale. Puis, dans la suite de ce chapitre, nous décrirons la procédure d'optimisation que nous avons suivi, l'analyse du pouvoir de reconnaissance d'OPEP optimisé et nous discuterons enfin des propriétés physiques de ses paramètres.

8.1 Matériels et méthodes

8.1.1 Les paramètres des centroïdes

Le mode de placement des centroïdes a cependant évolué dans les versions 3 du champ de force. Dans le modèle initial, chaque centroïde était défini par (i) sa longueur de liaison, (ii) l'angle de cette liaison virtuelle et (iii) l'angle dièdre impropre par rapport aux carbones *alpha* adjacents. Dans la version courante d'OPEP, chaque chaîne latérale est maintenant placée dans un repère local défini par les atomes lourds de la chaîne principale de chaque résidu : N, $C\alpha$ et C' (voir figure 5.3).

Pour chaque type d'acide aminé, la position moyenne de son centroïde a été calculée sur une banque de rotamères développée au sein de l'EBGM (Gautier et al., 2004). Il est à noter que le calcul de la moyenne est pondéré par la fréquence de chacun des rotamères. Pour déterminer le rayon de chaque sphère représentant les chaînes latérales, nous avons procédé en deux étapes. Les valeurs initiales des rayons sont estimées en calculant la distance moyenne séparant le centroïde du résidu considéré, et le plus lointain de ses atomes lourds en excluant le carbone β. Ces valeurs sont optimisées selon un critère des moindres carrés pour respecter au mieux la valeur moyenne de la distribution des distances intercentroïdes de résidus en contact. Cette valeur moyenne a été estimée sur une banque

de 2248 structures PDB non redondantes à 30% pour les 210 types d'interactions possibles, deux résidus non séquentiels étant considérés comme en contact si la distance qui les sépare est inférieure à la somme des rayons de Van Der Waals de leurs atomes les plus proches, augmentée de 1 Å.
Les cas particuliers de l'alanine et de la glycine sont traités à part : aucun centroïde n'est considéré pour la glycine, et dans le cas de l'alanine, les paramètres standards du groupement méthyl sont utilisés. La proline ayant une géométrie trop spécifique, tous ses atomes sont représentés explicitement dans OPEP. Pour sOPEP, nous avons tout de même considéré un centroïde pour ce résidu, ainsi que pour la glycine. Dans ce dernier cas, le rayon associé est celui de l'atome d'hydrogène (1,964 Å).
Les paramètres de positionnement des centroïdes et les rayons obtenus pour représenter les chaînes latérales sont présentés dans la table 8.1.

res	r_0 (Å)	$r_{C\alpha-Sc}$ (Å)	$\widehat{N.C\alpha.Sc}$	$\widehat{Sc.C\alpha.C'}$
ALA	2,287	1,52	116,60	111,10
CYS	2,426	1,95	108,50	117,65
ASP	2,707	2,14	110,74	119,46
GLU	2,968	2,77	110,95	120,87
PHE	3,33	2,62	110,72	124,15
GLY	-	-	-	-
HIS	3,078	2,60	109,93	124,39
ILE	2,939	2,27	109,16	118,96
LYS	3,143	3,11	112,92	121,57
LEU	2,928	2,40	112,72	124,94
MET	2,999	2,70	113,54	121,70
ASN	2,797	2,16	109,66	121,47
PRO	2,669	1,81	68,77	133,11
GLN	3,01	2,76	111,24	122,32
ARG	3,411	3,59	112,07	119,79
SER	2,334	1,77	106,78	107,98
THR	2,675	1,90	107,21	114,11
VAL	2,880	1,44	126,10	99,66
TRP	3,67	2,86	117,40	119,81
TYR	3,382	2,79	112,54	123,56

Tab. 8.1: OPEP : paramètres du positionnement des chaînes latérales. Pour chaque type de résidu, sont présentés, le rayon de la sphère (r_0) et les paramètres de positionnement du centroïde que sont, la distance $C\alpha$ centroïde ($r_{C\alpha-Sc}$), l'angle $\widehat{N.C\alpha.Sc}$ et l'angle $\widehat{Sc.C\alpha.C'}$.

Les valeurs de rayons obtenues semblent en accord avec les données de la littérature (Levitt, 1976). Cependant, nous avons constaté qu'une telle procédure ne prévient pas de toutes les collisions pouvant intervenir entre les chaînes latérales. Ainsi, 3,5 % des paires de résidus en contact voient leurs sphères latérales s'interpénétrer. Nous discuterons plus

précisément de ce problème, lors de l'optimisation du champ de forces.

8.1.2 La méthode d'optimisation

Dans les méthodes de prédiction de la structure des protéines, un champ de force idéal doit discriminer les structures natives des structures non natives pour toutes les séquences en acides aminés. Ceci peut être réalisé en optimisant le *ratio* T_f/T_g, où T_f et T_g sont les températures de transition de repliement (*folding transition temperature*) et vitreuse (*glass transition temperature*), ou en maximisant le Z-score entre les structures natives et non natives (Fujitsuka et al., 2004). Cependant, nous avons choisi de ne pas utiliser de tels critères dans la procédure d'optimisation. Parce que le bassin énergétique natif consiste en un ensemble de conformations proches de la structure native en équilibre, il semble approprié d'apprendre aussi de ces conformations proches de la structure native (PSN) (Zhang et al., 2003). Dans le cadre de l'optimisation d'OPEP, nous avons suivi le travail de Qiu et Elber (Qiu and Elber, 2005) qui requiert que, pour toutes les protéines cibles, la structure native (N) soit la structure de plus basse énergie, et que les structures PSN soient à la fois de plus haute énergie que la structure native, et de plus basse énergie que les autres leurres ou *decoys* (L). L'ensemble de ces conditions peut être résumé par les inégalités suivantes :

$$E(S_n, L_i) - E(S_n, N) > 0, \quad \text{pour tout i et n} \tag{8.1}$$

$$E(S_n, PSN_j) - E(S_n, N) > 0, \quad \text{pour tout j et n} \tag{8.2}$$

$$E(S_n, L_i) - E(S_n, PSN_j) > 0, \quad \text{pour tout i, j et n} \tag{8.3}$$

L'énergie de la conformation C est $E(S, C)$, S_n étant la séquence de la protéine n, N est la conformation native, L_i la $i^{\text{ème}}$ conformation leurre et PSN_j la $j^{\text{ème}}$ conformation PSN. Ainsi, pour chaque protéine, la fonction de score (FS) vaut -1 pour chaque inégalité satisfaite, et 0 sinon.

Pour résoudre les équations 8.1, 8.2 et 8.3 pour toutes les proteines du jeu d'apprentissage (JA), les 261 poids du potentiel OPEP sont optimisés par l'intermédiaire d'un algorithme génétique avec deux contraintes. Premièrement, une interaction de Van der Waals donnée, ne peut passer d'un potentiel de type 6-12 à un potentiel de type -6 durant le processus d'optimisation, et vice versa. Deuxièmement, la variation des paramètres $CL-CL$ ne peut évoluer qu'au sein d'intervalles fixés selon l'interaction considérée. Nous détaillerons ce point dans la suite de ce chapitre.
L'algorithme génétique implémenté utilise des opérateurs classiques que sont la mutation et le simple *crossing-over*, avec un taux de mutations de 20% diminuant de 0,25% toutes

les 10 itérations, et un taux de *crossing-over* de 14%. D'autres opérateurs plus complexes, tels que le double *crossing-over* ont été testés sans amélioration notable des résultats. Nous utilisons une taille de population de 80 chromosomes et l'élitisme est fixé à 13% pour la sélection. Tous ces paramètres sont issus d'une analyse de sensibilité. La convergence d'une simulation d'algorithme génétique est considérée comme atteinte si le nombre total d'inégalités résolues ne change pas durant 100 générations. Pour chaque cycle d'optimisation, 100 simulations sont réalisées avec différentes graines aléatoires.
Pour garantir que l'algorithme génétique ne soit pas bloqué dans un *minima* local, nous avons réalisé 50 simulations de MC de type recuit simulé de 50.000 pas chacun, en prenant comme point de départ les meilleurs paramètres déterminés par l'algorithme génétique, en les faisant varier un par un à chaque pas. Aucune amélioration n'a été observée, nous indiquant ainsi que l'algorithme génétique semble robuste et approprié pour notre approche.

8.1.3 Le jeu d'apprentissage

Le jeu d'apprentissage est constitué des 11 entrées PDB suivantes : 1abz, 1dv0, 1e0m, 1orc, 1pgb, 1qhk, 1shg, 1ss1, 1vii, 2ci2, 2cro, et du peptide *betanova* (Kortemme et al., 1998). La dernière entrée 1pgbF couvre l'épingle à cheveux β de 1pgb (résidus 41 à 56). Ces protéines ont été sélectionnées selon trois critères : (i) elles sont stables en solution, sans besoin d'un facteur supplémentaire tel qu'un pont disulfure ou un acide aminé non standard, (ii) elles couvrent un spectre suffisant de classes de protéines (α, β et α/β), (iii) elles ont des chaînes de taille raisonnable (entre 16 et 65 résidus) facilitant ainsi leur échantillonnage conformationnel, et (iv) elles ne présentent *a priori* pas de structure quaternaire stabilisant la chaîne.
Comme il a déjà été démontré (Mirny and Shakhnovich, 1996; Qiu and Elber, 2005; Lu and Skolnick, 2001), la qualité des leurres utilisés est essentielle dans un processus d'optimisation. Ces leurres doivent couvrir le plus largement possible l'espace conformationnel, avoir une énergie basse pour pouvoir rentrer en compétition avec la structure native et avoir été générés par différents protocoles pour minimiser l'influence de l'algorithme utilisé. Dans la présente étude, nous avons combiné quatre protocoles pour générer des conformations PSN et non natives pour chaque cible du jeu d'apprentissage.

Dynamique moléculaire : partant de la structure expérimentale, des simulations de dynamique moléculaire utilisant le champ de force tous atomes GROMOS ont été effectuées avec le package GROMACS (Berendsen et al., 1995; Spoel et al., 2005). Toutes les simulations ont été réalisées en solvant explicite à 600K pendant 1 ns. Toutes les structures générées ont été regroupées en classe avec une distance seuil de 2,5 Å. En moyenne, 100 structures par cible sont considérées, déviant de 1 à 12 Å de la structure expérimentale.

Jeu	Exp.	M	L	**Nombre de leurres**				**cRMSd** (Å)		**TM-score**		α	β
				DM	ENF	G	TOT	*min*	*max*	*min*	*max*	%	%
Jeu d'apprentissage complet													
1abz	N	1	38	49	281	278	608	2,6	11,9	0,14	0,58	46,9	5,82
betanova	-	-	20	347	299	283	930	0,7	12,2	0,01	0,59	10,36	23,87
1dv0	N	1	45	118	209	275	602	1,9	13,6	0,16	0,76	35,81	4,91
1e0m	N	1	37	59	213	183	451	1,5	12,8	0,14	0,82	17,14	17,33
1orc	X	1	64	79	89	262	430	0,9	13,6	0,22	0,92	26,69	20,00
1pgb	X	1	56	67	179	289	535	0,9	36,9	0,12	0,92	26,57	20,41
1pgbF	X	1	16	28	276	300	604	0,2	11,1	0,01	0,99	10,31	15,05
1qhk	N	1	47	59	189	277	525	3,3	31,1	0,13	0,78	24,22	10,03
1shg	X	1	57	41	601	286	928	1,3	41,7	0,12	0,85	16,99	21,98
1ss1	N	1	60	90	128	212	430	2,2	13,8	0,19	0,85	39,77	4,49
1vii	N	1	36	45	250	270	565	1,6	10,9	0,13	0,68	37,41	4,33
2ci2	X	6	65	30	166	298	494	2,2	34,9	0,11	0,87	21,52	16,33
2cro-*fisa*†	X	3	65	25	-	-	525	2,2	12,3	0,23	0,99	61,75	0,02
Jeu de validation complet													
1bba-*lmds*†	N	1	36	-	-	-	500	0,9	9,2	0,41	0,82	47,33	0,00
1ctf-*4state*†	X	2	68	-	-	-	616	0,5	12,5	0,24	0,88	33,56	4,09
1ctf-*lattice*†	X	2	68	-	-	-	977	5,1	10,5	0,25	0,51	34,65	0,26
1ctf-*lmds*†	X	2	68	-	-	-	489	3,3	15,1	0,28	0,68	51,28	21,16
1ctf-*semfold*†	X	2	68	-	-	-	996	4,4	9,2	0,28	0,59	50,70	0,63
1f4i	N	1	45	11519•	-	-	11519	0,1	36,9	0,12	0,99	4,74	7,37
1fsd	N	1	28	84	319	290	693	1,4	10,7	0,13	0,80	34,49	4,68
1khm-*semfold*†	N	1	73	-	-	-	998	3,9	7,9	0,32	0,58	44,31	2,22
1r69	X	1	63	72	99	281	452	0,1	27,7	0,23	0,99	46,81	3,51
1r69-*4state*†	X	1	63	-	-	-	670	0,8	11,4	0,25	0,94	36,63	0,36
1r69-rosetta†	X	1	61	-	-	-	998	0,2	14,9	0,27	0,99	63,13	0,00
1s04-*casp6*†	N	1	110	27	-	-	213	0,6	56,5	0,15	0,98	28,19	15,66
1te7-*casp6*†	N	1	103	18	-	-	222	1,8	38,3	0,18	0,88	22,69	21,79
1ubi-*lmds*v2†	X	1	76	61	-	-	361	2,0	15,4	0,22	0,81	15,54	30,44
2cro-*4state*†	X	3	65	25	-	-	697	0,1	9,3	0,25	0,99	36,33	0,19
2cro-*lmds*†	X	3	65	25	-	-	525	0,1	12,9	0,25	0,99	63,85	0,06

Tab. 8.2: Description des jeux complets d'apprentissage et de validation. Exp. désigne la méthode expérimentale ayant permis de résoudre la structure native de la cible, N pour RMN et X pour cristallographie aux rayons X. M est l'état oligomérique de chaque protéine comme il est défini par l'E-MSD (*the European Bioinformatics Institute Macromolecular Structure Database* (Boutselakis et al., 2003)), L est la longueur de la protéine (en acides aminés), TOT est le nombre total de leurres générés pour chaque protéine, par dynamique moléculaire (DM), enfilage (ENF) et algorithme glouton (G) ou issus de jeux publiques (†). Les colonnes suivantes représentent les valeurs maximales et minimales de cRMSd et TM-score observées pour chaque jeu de leurres en comparaison avec la structure native. Sont aussi présentés les pourcentages moyens de résidus situés dans une hélice α ou un brin β identifiés par le programme *Stride* (Frishman and Argos, 1995). Le symbole • indique les leurres générés par des simulations ART-OPEP.

Enfilage : pour chaque protéine du jeu d'apprentissage, en moyenne 200 leurres ont été générés par une méthode simple d'enfilage sans *gaps*. Les structures matrices utilisées incluent les entrées PDB 1pgb, 1shg, 2ci2, 1csp, 5fd1, 2acy, 1ctf, 1ubq et 3chy. Cette liste est gardée petite car les méthodes d'enfilage ne sont pas très performantes en reconnaissance de repliement, mais, cependant, les minima locaux générés sont intéressants, car ils ont des caractéristiques proches d'une structure native.

Algorithme glouton : nous avons également utilisé l'algorithme glouton décrit en introduction (Tuffery et al., 2005; Tuffery and Derreumaux, 2005), couplé à la fonction

d'énergie de type Go ou au critère RMSd pour générer des leurres à la fois de très haute et très faible énergie. Ces conformations dévient de 0,05 à 20 Å de la structure native.

Decoys 'R' us : enfin, nous avons aussi utilisé la base publique de leurres Decoys 'R' us (Samudrala and Levitt, 2000). Pour le jeu d'apprentissage, nous n'avons utilisé que les leurres de la cible 2cro du jeu *fisa* (noté par la suite 2cro-*fisa*). Ces leurres ont été générés par l'assemblage de fragments protéiques selon une procédure de recuit simulé (Simons et al., 1997).

Un autre point essentiel de notre procédure d'optimisation est d'assigner le caractère PSN ou non native à une conformation. Il est très courant dans les études de repliement protéique d'utiliser le cRMSd ou la fraction de contacts natifs. Or des analyses des surfaces d'énergie libre par des méthodes de graphes et de réseaux ont clairement montré les limitations de ces paramètres (Caflisch, 2006). Dans ce travail, nous avons utilisé le programme TM-align, un algorithme rapide et performant pour aligner les structures protéiques (Zhang and Skolnick, 2004b, 2005b). Le TM-score associé à un seuil de 0,5 nous permet de distinguer les structures natives ($> 0,5$) des structures non natives ($< 0,5$). Ainsi, tous les leurres ont été alignés avec leur structure native correpondante, et nous avons selectionné cinq structures PSN avec des TM-scores variant entre $0,5$ et $1,0$. Pour 1shg, la distribution étroite des TM-scores nous a conduit à ne sélectionner que 4 conformations PSN.
Dans la suite de ce chapitre, le jeu d'apprentissage non redondant correspond à la structure native, cinq (ou quatre) structures PSN et les leurres non natifs (TM-score $< 0,5$), alors que le jeu d'apprentissage complet inclut tous les leurres indépendamment de leur TM-score, autrement dit, il contient en plus toutes les structures PSN.
La table 8.2 analyse les caractéristiques structurales du jeu d'apprentissage complet en termes de cRMSd et de TM-score en comparaison avec l'état natif, et le pourcentage de résidus dans des structures secondaires de type α ou β. Il est à noter que les valeurs extrêmes de ces paramètres ne changent pas entre les jeux non redondant et complet, et que toutes les structures ont été minimisées en utilisant OPEP jusqu'à ce que la norme du gradiant soit inférieure à $0,01$ kCal/mol.Å. Globalement, la densité d'états obtenus montre que l'espace conformationnel échantillonné n'est pas limité à un bassin d'attraction spécifique.

8.1.4 Le jeu de validation

Pour tester le pouvoir discriminant d'OPEP, nous avons considéré plusieurs jeux de leurres parmi lesquels *4-state reduced* (Park and Levitt, 1996; Samudrala and Levitt, 2000), *lmds* et *lmdsv2* (Samudrala and Levitt, 2000), *semfold* (Samudrala and Levitt, 2002), *fisa* (Simons et al., 1997), *lattice-ssfit* (Xia et al., 2000), Rosetta (Tsai et al., 2003)

et *casp6* (Moult et al., 2005).
Toutes les cibles de ces jeux de leurres n'ont pas été utilisées, car (i) connaissant nos ressources de calculs, nous ne pouvions raisonnablement pas minimiser 120 cibles pour un total de 210.000 conformations leurres, avec une norme de gradiant de 0,01 kCal/mol.Å, (ii) une cible fait partie de notre jeu d'apprentissage (2cro-*fisa*), (iii) certaines protéines sont multidomaines (*e.g.* 1fc2), (iv) se fixent à un ligand tel que le calcium (*e.g.* 4icb, 1e68) ou le fer (4rxn), (v) sont stabilisées par des ponts disulfure (c'est le cas de deux des sept protéines du jeu *4-state reduced*) ce qui nécessiterait de paramètriser l'état oxydé des cystéines, ou ont des extrémités terminales tronquées (Rosetta).
Ainsi, nous avons considéré les cibles suivantes : *lmds* (1bba, 1ctf, 2cro), *lmdsv2* (1ubi), *4-state reduced* (1ctf, 1r69, 2cro), *lattice-ssfit* (1ctf), *semfold* (1ctf, 1khm), Rosetta (1r69) et *casp6* (1s04, 1te7). Pour 1cft *semfold*, 1ctf *lattice-ssfit* et 1khm *semfold*, nous n'avons pas retenu l'ensemble des conformations des jeux qui contenaient respectivement 11.399, 2000 et 21.079 leurres, mais un sous-ensemble d'environ 1.000 leurres couvrant un large spectre de TM-scores et de cRMSd. Remarquez que 1ctf est la seule cible selectionnée qui est stabilisée par un ligand SO_4^{2-}, et 1r69 possède 63 acides aminés dans la structure native et dans le jeu *4-state reduced*, mais 61 acides aminés dans le jeu Rosetta (les deux acides aminés C-terminaux sont manquants).
De plus, des leurres ont été générés pour les entrées PDB 1fsd et 1r69 avec les trois méthodes précédemment décrites. Nous avons aussi regénéré des leurres DM pour les cibles 1s04, 1te7, 1ubi, et 2cro. Et, finalement, des leurres ont été obtenus pour 1f4i par l'intermédiaire de simulations ART-OPEP (Malek and Mousseau, 2000; Wei et al., 2002). Les simulations ART-OPEP avaient pour point de départ une conformation complètement étendue et ont duré 20.000 pas avec les paramètres d'OPEP avant optimisation.
Cet ensemble de 10 protéines distinctes, correspondant à 16 cibles, et contenant 20.926 leurres, correspond au jeu complet de validation (JV). Comme pour le jeu d'apprentissage, nous avons généré un jeu de validation non redondant. Les propriétés structurales du jeu de validation sont résumées table 8.2. Nous pouvons remarquer que, comme pour le jeu d'apprentissage, un large éventail de valeurs pour le cRMSd et TM-score est couvert.

La table 8.3 montre le nombre total de contacts CL-CL et la fraction de chaque type d'acides aminés dans les structures secondaires des leurres des JA et JV. Bien que nous utilisions un nombre total de 22 protéines (28.553 leurres), quelques paires d'interactions CL-CL ne sont que très peu peuplées dans le JA et le JV. Sur un total de 210 paires, 10 d'entre elles ont un nombre de contacts inférieur à 50. Ce résultat était attendu pour la paire Cys-Cys, puisque nous n'avons pas considéré de protéines stabilisées par des ponts disulfure. La faible population observée dans le JA pour les paires Cys-Ser, His-Phe, His-Gly, His-Met, His-Gln, His-Ser et His-Val n'est pas retouvée dans le JV. Et inversement, la faible occurence de la paire His-His n'est observée que pour le JV. Ceci montre clairement

JV/JA	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	3416 89836	1532	2920	5692	5428	3516	204	5936	7222	13478	3912	2994	702	4290	3102	1550	5612	6436	1176	3540
CYS	28246	**0** 138	668	766	162	78	**0**	218	748	1378	246	892	466	376	98	**44**	416	910	776	622
ASP	32084	2426	796 28642	1698	1340	1750	108	448	5402	1384	586	2166	1490	1118	1590	1842	3794	1748	1042	2190
GLU	122876	18134	74108	1154 73224	2830	2496	94	4124	12798	5174	2286	3440	1100	4664	6178	2686	7670	2550	1060	2374
PHE	136940	29442	34466	57020	1822 23116	2254	**8**	3076	3284	8656	2260	2150	1900	2694	2850	1704	3284	3660	952	3184
GLY	47274	706	8560	21682	34738	1104 9628	**50**	1166	5054	3534	870	3752	538	1434	3192	964	4148	4442	2000	4204
HIS	2194	38	1950	5686	4642	604	**0** 824	196	166	162	**44**	604	62	**10**	68	**48**	252	**0**	160	226
ILE	115988	14024	12728	67924	87720	12882	6612	1794 22652	3686	8968	1336	816	466	3420	2384	1868	1954	3862	716	2884
LYS	79474	4422	68266	251094	18750	24792	1156	39906	2940 21496	8064	1696	4238	1238	5612	4116	3054	8220	4548	2492	4380
LEU	358120	20358	29590	81190	272606	87196	2764	210078	65324	8602 259172	4628	4184	3242	6990	6348	4290	5252	7624	3374	2424
MET	15366	936	1592	12602	10744	5092	838	15184	4182	30226	178 2356	228	340	1162	862	920	2466	1126	486	716
ASN	65442	4454	32632	91958	55940	9326	330	7296	47790	43442	2630	2006 25732	432	3132	1302	1896	3774	818	1932	2864
PRO	20876	336	7348	13810	6630	4872	10524	13748	6064	25612	4234	3544	284 4850	936	2024	936	688	1826	1156	766
GLN	59220	1064	9976	77786	21774	11798	1134	45194	52296	60014	8804	42852	7766	2738 24468	1642	3690	4272	2790	1118	2624
ARG	25796	926	41636	103220	19822	13552	7222	31148	37720	51934	7424	8686	17682	20140	1932 15896	814	2944	2068	1268	1584
SER	47080	3196	19964	66516	27754	12170	3872	38052	31264	66456	1972	14708	15114	43492	19428	244 14776	2224	1612	876	1366
THR	23800	804	8564	23334	13764	18580	5934	21380	19214	35364	5382	7538	10064	24094	22072	17138	6400 11406	4586	3272	5812
VAL	108760	2840	24072	44636	50290	42562	2556	91130	57408	151560	17466	6038	15804	10052	38516	36876	17462	1078 46808	2498	3920
TRP	5856	1618	10122	2160	2210	3472	548	3572	2524	16834	3644	4542	1620	1922	5098	4482	2532	12674	74 0	2242
TYR	46228	2396	4878	16264	32274	9634	2532	24698	8400	39872	7044	7820	9130	11470	10596	5018	7028	22162	2092	1930 2716
JV α	18,453	0,025	5,643	11,502	0,890	2,904	0,332	5,761	15,230	13,956	0,452	1,234	1,038	5,429	5,304	3,213	1,926	5,440	0,600	0,661
JA α	13,899	0,738	1,174	10,846	4,214	1,802	0,147	6,843	10,433	14,342	2,987	3,222	0,865	8,590	5,674	4,733	5,182	2,231	1,046	1,030
JV β	1,071	0,006	0,246	0,636	0,281	0,335	0,000	0,503	0,441	0,662	0,005	0,058	0,034	0,102	0,080	0,122	0,249	0,989	0,007	0,021
JA β	0,328	0,007	0,274	0,467	0,677	0,514	0,000	0,275	0,668	0,333	0,059	0,355	0,014	0,115	0,181	0,400	1,174	0,418	0,164	0,837

Tab. 8.3: Les contacts entre chaînes latérales et le pourcentage de résidus dans des structures secondaires de type α et β pour les jeux complets d'apprentissage et de validation. Les valeurs de la partie supérieure droite correspondent au comptage des contacts CL-CL au sein du JA, et les valeurs de la partie inférieure gauche au sein du JV. JV α,β et JA α,β correspondent aux pourcentages de résidus dans l'état α et β dans le JV et le JA. Les paires de contacts < 50 sont mises en gras.

que notre choix de protéines constituant les JA et JV n'est pas dénié d'un quelconque biais. Par contre, il est intéressant de noter que le contact Cys-His n'est que rarement vu à la fois dans le JA et le JV, et dans notre PDB non redondante de 2.248 structures. Ceci suggérant une faible occurence de cette paire dans les structures protéiques. Afin de prendre en compte le biais précédemment décrit dans notre distribution des contacts CL-CL, l'optimisation de la matrice de contact proposée par Betancourt et Thirumalai est réalisée sous contrainte, et les poids $w_{CL,CL}$ sont maintenus entre 0,7 et 1,3.

8.2 Résultats - Discussion

8.2.1 Le pouvoir discriminant d'OPEP optimisé

TARGET	L_{TOT}	PSN	FS_{max}	$FS_{\text{début}}$	FS_{fin}	$Gain$	NS
1abz	445	5	-2675	-2380	-2613	233	62
betanova	618	5	-3713	-3658	-3649	-9	64
1dv0	378	5	-2273	-2153	-2256	103	17
1e0m	310	5	-1865	-1056	-1317	261	548
1orc	175	5	-1043	-886	-943	57	100
1pgb	304	5	-1829	-1823	-1824	1	5
1pgbF	526	5	-3161	-3096	-3124	28	37
1qhk	416	5	-2501	-2442	-2445	3	56
1shg	631	4	-3159	-3132	-3143	11	16
1ss1	162	5	-977	-536	-857	321	120
1vii	430	5	-2585	-2288	-2542	254	43
2ci2	189	5	-1139	-918	-936	18	203
2cro-*fisa*	422	5	-2537	-2504	-2452	-52	85
Total	5006	64	-29457	-26872	-28101	1229	1356

Tab. 8.4: Résultats de l'optimisation sur le jeu d'apprentissage non redondant. L_{TOT} est le nombre total de leurres. PSN est le nombre de structures PSN utilisées durant le processus d'optimisation. FS_{max} est la valeur maximale que peut atteindre la fonction de score. $FS_{\text{début}}$ et FS_{fin} sont les valeurs de la fonction de score au début, et à la fin de l'optimisation. $Gain$ est le gain apporté par l'optimisation, et NS le nombre d'inégalités encore non satisfaites à la fin de l'optimisation.

La table 8.4 présente les résultats de l'optimisation de la version 3.1 d'OPEP sur le JA non redondant. Une fonction de score totale de -29457 indique que toutes les inégalités des équations 8.1, 8.2 et 8.3 sont satisfaites. Ainsi toutes les structures natives sont les structures de plus basse énergie, et les structures PSN ont une énergie inférieure aux leurres non natifs. Durant le processus d'optimisation, la FS diminue de -26872 à -28101. Bien que l'amélioration ne soit pas optimale, 48% des inégalités non satisfaites le deviennent à la fin du processus d'optimisation. La version optimisée d'OPEP permet de mieux distinguer les structures natives ou PSN pour 11 protéines (la FS diminue de 1

pour 1pgb et de 321 pour 1ss1) et semble performante pour *betanova* et 2cro-*fisa* (la FS augmente de 9 et 52 respectivement). Les paramètres optimisés pour les interactions des chaînes latérales, les liaisons hydrogène, et les propensités α/β sont donnés en annexe A.1.

CIBLE	$E(N)$	$E(L_{min})$	$E(PSN_{min})$	N_{cl}	PSN_{cl}	Rc
Jeu d'apprentissage complet						
1abz	-87.3	-93.9	-91.4	30	11	Non
betanova	-29.0	-34.0	-29.0	26	29	Non
1dv0	-107.2	-106.0	-106.0	1	2	Oui
1e0m	-50.1	-60.1	-54.0	34	14	Non
1orc	-135.5	-143.9	-143.9	5	1	Oui
1pgb	-141.8	-156.1	-156.1	21	1	Oui
1pgbF	-30.4	-30.7	-30.7	17	1	Oui
1qhk	-100.1	-93.1	-88.6	1	4	Oui
1shg	-136.3	-139.4	-139.4	5	1	Oui
1ss1	-121.8	-115.5	-115.5	1	2	Oui
1vii	-71.2	-71.6	-71.6	2	1	Oui
2ci2	-173.3	-166.9	-166.9	1	2	Oui
2cro-*fisa*	-141.9	-160.8	-160.1	115	2	Non
Jeu de validation complet						
1bba-*lmds*	-48.8	-57.7	-57.7	378	1	Oui
1ctf-*4state*	-180.3	-179.2	-179.2	1	2	Oui
1ctf-lattice	-180.3	-155.5	-145.5	1	11	Oui
1ctf-*lmds*	-180.3	-183.1	-177.8	2	3	Non
1ctf-*semfold*	-180.3	-181.8	-181.8	3	1	Oui
1f4i	-78.4	-90.3	-90.3	6	1	Oui
1fsd	-55.3	-58.4	-58.4	12	1	Oui
1khm-*semfold*	-161.8	-141.6	-141.1	1	4	Oui
1r69	-140.5	-140.6	-140.6	2	1	Oui
1r69-*4state*	-140.5	-136.2	-136.2	1	2	Oui
1r69-rosetta	-139.6	-140.0	-140.0	3	1	Oui
1s04-*casp6*	-286.7	-309.1	-302.7	32	2	Non
1te7-*casp6*	-254.2	-280.5	-280.5	24	1	Oui
1ubi-*lmdsv2*	-165.8	-179.2	-179.2	69	1	Oui
2cro-*4state*	-141.9	-153.8	-153.8	28	1	Oui
2cro-*lmds*	-141.9	-153.8	-153.8	37	1	Oui

Tab. 8.5: Performance d'OPEP 3.1 sur les jeux d'apprentissage et de validation. $E(N)$ est l'énergie de la structure native, $E(L_{min})$ est la plus basse énergie obtenue pour la cible (la structure native exclue) et $E(PSN_{min})$ est l'énergie la plus basse des structures PSN. N_{cl} est le classement de la structure expérimentale, PSN_{cl} est le classement de la structure PSN de plus basse énergie. Rc est l'état de reconnaissance, "Oui" indiquant que la structure de plus basse énergie est une structure native ou PSN. Les valeurs soulignées mettent en évidence les cibles pour lesquelles l'état de reconnaissance a changé lors du processus d'optimisation.

Considérons maintenant les JA et JV (voir table 8.5). Les tracés des énergies OPEP

optimisé contre le TM-score pour toutes les protéines sont dans l'annexe A.2, ainsi que l'énergie OPEP optimisé contre le cRMSd pour *Betanova*.
Au total, 3.179.377 inégalités doivent être satisfaites sur l'ensemble des JA et JV complets. Avant optimisation, 2.439.934 inégalités sont satisfaites (et 739.443 non satisfaites), et, après optimisation, 2.748.877 inégalités sont satisfaites (430.500 non satisfaites). Ainsi, 42% des inégalités non satisfaites ont été résolues à la fin du processus d'optimisation. Dans la table 8.5, est indiqué le classement de la structure expérimentale, et l'état de reconnaissance, *i.e.* la capacité d'OPEP à reconnaître une structure native ou PSN comme étant celle de plus basse énergie parmi l'ensemble des leurres d'une cible donnée. La structure expérimentale est classée première pour quatre protéines du JA (1dv0, 1qhk, 1ss1, et 2ci2), et quatre protéines du JV (1ctf-*4state*, 1ctf-*lattice*, 1khm-*semfold* et 1r69-*4state*). Par ailleurs, une structure PSN est classée première pour cinq autres cibles du JA, et 10 cibles du JV. Donc, globalement, la version 3.1 d'OPEP est capable de distinguer des structures natives ou PSN de structures non natives pour 23 des 29 cibles de nos jeux d'apprentissage et de validation, *i.e.* 69% (9/13) des cibles du JA, et 87% (14/16) de celles du JV.

8.2.2 Les cibles problématiques

Pour expliquer pourquoi OPEP ne parvient pas à reconnaître l'état natif de six cibles, nous avons regardé au niveau atomique quelles pouvaient en être les raisons. Parmi ces six cibles, quatre appartiennent au JA (1abz, *betanova*, 1e0m et 2cro-*fisa*), et deux au JV (1ctf-*lmds* et 1s04-*casp6*). Pour ces cibles, la figure 8.1 montre les structures expérimentales et de plus basse énergie superposées.

Pour les cibles d'apprentissage 1abz et 1e0m, le leurre d'énergie la plus basse est une conformation PSN. En effet, cette structure a un TM-score de 0,49 (voir l'annexe A.2). De plus elle ne dévie que de 2,3 Å de cRMSd de la structure RMN (figure 8.1A). De la même manière, le leurre de plus basse énergie pour la protéine de 34 résidus 1e0m, partage un TM-score de 0,43 avec la structure RMN, ce qui correspond, dans ce cas, à un cRMSd de 6,8 Å. Malgré cette haute valeur de cRMSd, la topologie de ce leurre est native (ceci nous est fortement suggéré par la valeur du TM-score), et si l'on exclue les extrémités de la protéine, qui sont connues pour être de basse "résolution" en RMN (Macias et al., 2000), le cRMSd de la région 6-31 tombe à seulement 2,2 Å (figure 8.1C). Le gain énergétique du leurre vient surtout de l'extension du troisième brin β de deux résidus.

Pour l'ensemble de leurres de la protéine de 65 résidus 2cro-*fisa*, l'écart énergétique entre le leurre de plus basse énergie et les structures PSN n'est que marginal, de l'ordre de k_BT. Le meilleur leurre (avec TM-score de 0,45) dévie de 6 Å (cRMSd) de la struc-

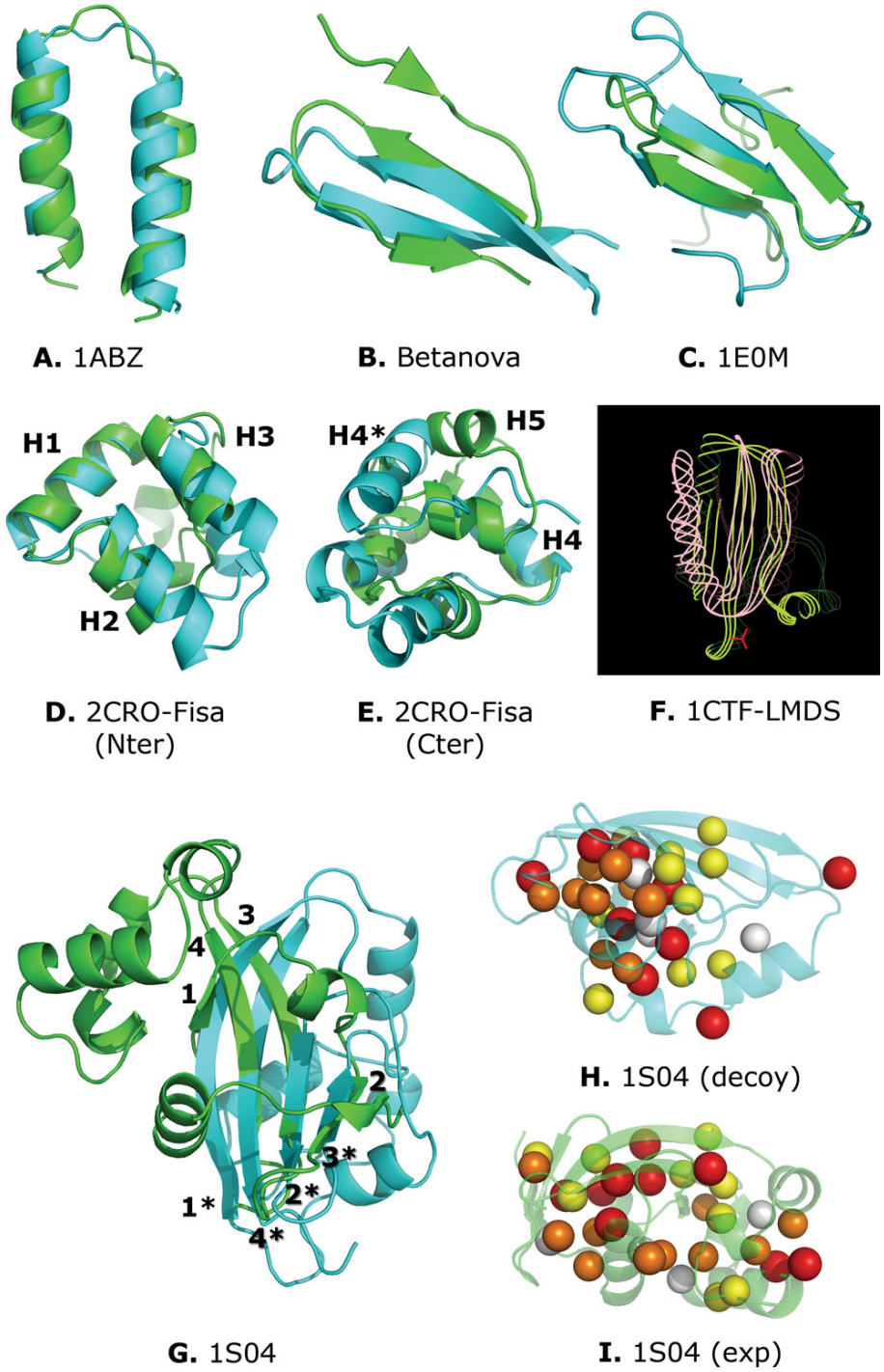


Fig. 8.1: Les cibles problématiques. Sont superposées ici les structures expérimentales et les leurres de plus basse énergie. La structure native est en vert pour toutes les cibles. Pour 2cro-*fisa* (vues D et E), les deux vues mettent en évidence les différences structurales au niveau des régions N- et C-terminales. H# correspond à la $\#^{\text{ème}}$ hélice de la structure expérimentale, et H#* à la $\#^{\text{ème}}$ hélice du leurre. Pour 1ctf-*lmds* (vue F), la particule rouge représente la position de l'ion SO_4^{2-}. Pour la cible 1s04, trois vues sont proposées : la vue G montre l'échange des brins β, et les vues H et I illustrent la différence d'empaquetage des chaînes latérales non-polaires, avec les isoleucines en rouge, les valines en jaune, les phénylalanines en blanc, et les leucines en orange. Ces images ont été générées par PyMol (DeLano, 2002) pour les vues A à E, et G à I, et XmMol (Tuffery, 1995) pour la vue F. Figure extraite de (Maupetit et al., 2007).

ture cristalline (figures 8.1D,E). La meilleure conformation PSN (TM-score de 0,58) dévie de 4,2 et 2,7 Å de cRMSd de la structure expérimentale pour les régions 1-65 et 3-61 respectivement. Alors que les structures native et PSN sont décrites par un faisceau de cinq hélices, le leurre n'en a que quatre. La cinquième hélice est dépliée, et les quatre premières hélices s'éloignent de leur position dans le cristal de 4.3 Å (figures 8.1D,E). Il est cependant à noter que la structure expérimentale de la cible 2cro présente des interactions moléculaires au sein de l'unité cellulaire de la structure cristalline (assemblage homo-trimérique, selon l'EMSD (Boutselakis et al., 2003)).

Alors que le peptide *betanova* est censé adopter une conformation en feuillet β à trois brins en solution (Kortemme et al., 1998), la structure d'énergie minimale est un motif β en épingle à cheveux (figure 8.1B). Ce leurre est issu de l'enfilage de la séquence de vingt résidus du peptide *betanova* dans ce motif de 1pgb. Le gain énergétique associé à cette conformation est de 5 kCal/mol par rapport à la structure RMN ; les contributions des liaisons hydrogène à deux et quatre corps compensent la perte des interactions CL-CL. Cette préférence pour la conformation en épingle à cheveux est surprenante, mais il existe un certain nombre d'études théoriques et expérimentales indiquant que la conformation feuillet β à trois brins n'est pas majoritaire en solution. Contrairement aux premières études expérimentales, la population de feuillets β a été réestimée à environ 10% dans l'eau à 283 K (de la Paz et al., 2001). Par ailleurs, des simulations tous atomes en solvant explicite ont montré que l'état idéal en feuillet β à trois brins est en équilibre dynamique avec d'autres conformations désordonnées dans une échelle de temps de 100 ns (Soto and Colombo, 2004). D'autre part, il est intéressant de noter que, pour un autre peptide, en l'occurence le fragment C-terminal de 1pgb, des analyses en graphes de transition des surfaces d'énergie libre de repliement suggèrent que l'épingle à cheveux β est un état plus peuplé que le feuillet β à trois brins (Caflisch, 2006). L'ensemble de ces observations permet d'expliquer les deux conformations que nous avons observées en compétition.

Le leurre *lmds* de plus basse énergie observé pour la protéine de 68 résidus 1ctf possède un TM-score de 0,42 avec la structure native et présente cependant une topologie non native. Ce leurre est stabilisé de 3 et 5 kCal/mol, respectivement, par rapport à la structure PSN de plus basse énergie et la structure native. En terme de cRMSd, ce leurre s'éloigne de la structure native de 8,7 Å. Comme il a été discuté plus haut, la structure cristalline possède un ligand SO_4^{2-} formant une liaison hydrogène avec l'azote de la chaîne latérale de la lysine 65, et l'azote de la liaison amide de la glycine 62 (numérotation PDB). Il a été suggéré que cet environnement souffré aurait une implication fonctionnelle (Leijonmarck and Liljas, 1987). L'observation fine de cette même région du leurre nous indique qu'il n'y a pas de place pour insérer un ligand SO_4^{2-} à proximité de la lysine 65 et de la glycine 62. De ce fait, la présence du ligand pourrait empêcher de conduire à cette conformation.

Enfin, le leurre 1s04-*casp6* de plus basse énergie diffère de 23 kCal/mol par rapport à la structure native, et 6 kCal/mol par rapport à la structure PSN de plus basse énergie. Comme nous pouvons l'observer sur la figure 8.1G, le leurre, de topologie non native, est caractérisé par une interversion des brins β 2 et 3 et par une déstabilisation des hélices α (24% des résidus dans une conformation α *versus* 39% dans la structure native). De plus le feuillet β est moins déformé et les brins étendus (29% de conformation β *versus* 25% dans la structure native). Comme il est montré sur les figures 8.1H-I, ce changement de topologie est associé, à la fois à une stabilisation par les liaisons hydrogène, mais aussi à un coeur hydrophobe compact.

Pour résumer, nous pouvons considérer que OPEP reconnaît correctement 1abz (selon le critère cRMSd), le calcul en solution pour 1ctf et 2cro ne semble pas approprié, et OPEP échoue clairement pour 1s04-*casp6*. Ceci suggère que la balance entre les liaisons hydrogène et les interactions CL-CL pourrait être réévaluée, mais avant de se pencher sur ce problème, quatre questions doivent être soulevées.

8.2.3 Robustesse des paramètres

La première question à laquelle nous pouvons répondre est : quelle est l'influence du nombre de cibles du jeu d'apprentissage sur les paramètres d'OPEP (Mirny and Shakhnovich, 1996) ?
Pour y répondre, nous avons répété le processus d'apprentissage avec cette fois ci l'ensemble des JA et JV, en partant des paramètres optimaux issus de l'optimisation initiale réalisée sur le JA seul. L'optimisation, après 100 simulations d'algorithme génétique, à conduit à un gain négligable de $+0,05\%$ d'inégalités satisfaites. Appliqué aux JA et JV complets, le gain total atteint $+2,2\%$. Par ailleurs, le coefficient de corrélation avec l'ensemble des poids initiaux est de 0,98, et le pouvoir de reconnaissance est identique. De plus, le rang de la structure native pour les cibles problématiques est 28 *vs.* 30 pour 1abz, 33 *vs.* 34 pour 1e0m, 31 *vs.* 32 pour 1s04-*casp6* et 116 *vs.* 115 pour 2cro-*fisa*. L'ensemble de ces résultats nous indique que les paramètres optimisés d'OPEP semblent robustes et insensibles au nombre de protéines utilisées pour l'apprentissage.

8.2.4 Comparaison avec le champ de force DOPE

Il semble essentiel d'évaluer comment se place OPEP par rapport à d'autres champs de force de la littérature en terme performance de classement de conformations natives. Ceci est loin d'être une tâche aisée en considérant le fait que toutes les études n'utilisent pas forcément les mêmes jeux de cibles, les mêmes critères, ou les mêmes leurres. Cependant, Shen et Sali ont récemment testé le potentiel statistique tous atomes DOPE

CIBLE	**OPEP**		**DOPE**				
	Z	R	Z	R	N_{cl}	PSN_{cl}	Rc
Jeu d'Apprentissage Complet							
1abz	-1,87	-0,635	-0,56	-0,677	186	8	Non
betanova	-1,34	-0,555	1,12	-0,048	789	87	Non
1dv0	-3,06	-0,604	-1,85	-0,762	23	1	Oui
1e0m	-1,69	-0,389	-1,39	-0,358	35	5	Non
1orc	-1,67	-0,753	-1,88	-0,830	2	1	Oui
1pgb	-1,88	-0,858	-1,23	-0,780	53	1	Oui
1pgbF	-2,32	-0,755	-1,41	-0,409	42	31	Non
1qhk	-2,93	-0,509	-1,30	-0,584	13	1	Oui
1shg	-2,07	-0,776	-1,79	-0,590	12	1	Oui
1ss1	-2,69	-0,601	-1,44	-0,751	29	1	Oui
1vii	-2,05	-0,574	-0,23	-0,583	247	6	Non
2ci2	-1,90	-0,698	-1,18	-0,786	11	1	Oui
2cro-*fisa*	-0,75	-0,493	-0,57	-0,451	152	4	Non
μ	-2,02	-0,631	-1,06	-0,585	-	-	-
σ	0,63	0,132	0,83	0,223	-	-	-
Jeu de Validation Complet							
1bba-*lmds*	0,57	-0,212	0,22	-0,190	305	1	Oui
1ctf-*4state*	-2,36	-0,775	-2,32	-0,800	1	2	Oui
1ctf-lattice	-4,79	-0,251	-4,75	-0,281	1	4	Oui
1ctf-*lmds*	-2,54	-0,160	-3,05	-0,093	1	40	Oui
1ctf-*semfold*	-2,53	-0,274	-2,12	-0,326	6	1	Oui
1f4i	-2,14	-0,513	-2,13	-0,691	3	1	Oui
1fsd	-2,07	-0,718	0,62	-0,698	498	1	Oui
1khm-*semfold*	-4,63	-0,356	-2,24	-0,361	13	3	Non
1r69	-1,77	-0,787	-1,61	-0,795	4	1	Oui
1r69-*4state*	-2,58	-0,737	-2,20	-0,785	4	1	Oui
1r69-rosetta	-2,42	-0,381	-1,35	-0,417	89	1	Oui
1s04-*casp6*	-1,24	-0,586	-1,54	-0,561	5	6	Non
1te7-*casp6*	-1,28	-0,674	-1,20	-0,580	27	2	Non
1ubi-*lmdsv2*	-1,01	-0,801	-1,64	-0,890	60	1	Oui
2cro-*4state*	-1,86	-0,698	-1,58	-0,775	52	1	Oui
2cro-*lmds*	-1,30	-0,514	-0,45	-0,289	164	4	Non
μ	-2,12	-0,527	-1,71	-0,533	-	-	-
σ	1,29	0,226	1,25	0,254	-	-	-
Jeux d'Apprentissage et de Validation Complets							
μ	-2,07	-0,574	-1,42	-0,557	-	-	-
σ	1,03	0,194	1,11	0,237	-	-	-

Tab. 8.6: Comparaison des champs de force OPEP et DOPE. Pour chaque cible des JA et JV complets, sont reportés les Z-scores (Z) et le coefficient de correlation (R) entre les énergies et les TM-scores en utilisant OPEP 3.1 ou DOPE (Eramian et al., 2006). La moyenne (μ) et l'écart type (σ) sont donnés pour le JA, le JV et l'ensemble des JA et JV. N_{cl} est le rang de la structure native, PSN_{cl} de la structure PSN de plus basse énergie. Rc rapporte le statut de reconnaissance de DOPE, "Oui" indiquant que la structure de plus basse énergie est la structure native ou une structure PSN.

(*Discrete Optimized Protein Energy model*) (Eramian et al., 2006; Shen and Sali, 2006), et ont comparé sa performance de reconnaissance à cinq autres fonctions de score, que sont Rosetta (Simons et al., 1997, 1999; Misura et al., 2006), DFIRE (Zhou and Zhou, 2002) and ModPipe (Melo et al., 2002). DOPE identifie correctement 28 structures natives sur 32 cibles, contre 27 pour DFIRE, 14 pour Rosetta, et 19, 7 et 18 pour les trois versions de ModPipe.

Dans ce travail, nous avons calculé les énergies de l'ensemble des nos leurres avec le champ de force "tous-atomes" DOPE. Pour ce faire, il nous a fallu reconstruire les chaînes latérales de nos leurres avec le programme SCWRL3 (Canutescu et al., 2003), étant donné que dans notre modèle gros grain, nous n'avons qu'un centroïde pour les représenter. Les énergies DOPE ont été calculées grâce à l'outil "decoy" de la *Protein Library* librement distribué sur le réseau à l'adresse suivante : http ://protlib.uchicago.edu (Eramian et al., 2006). Parce que cette reconstruction des chaînes latérales peut entraîner un biais dans l'analyse en comparaison avec les résultats publiés de DOPE, nous avons aussi reporté les résultats de DOPE sur les leurres publiques non minimisés (voir la figure A.6 en annexe A.3). Nous insistons sur le fait que notre but n'est pas ici de classer OPEP par rapport à DOPE, car notre analyse n'est pas basée sur l'ensemble des leurres ayant servis à l'évaluation de DOPE. Et, de plus, DOPE est connu pour être moins efficace sur les petites protéines (de moins de 50 résidus) et les structures résolues par RMN (Shen and Sali, 2006).

La table 8.6 compare les performances d'OPEP et de DOPE sur chacune de nos cibles, selon quatre critères distincts : le Z-score, calculé comme étant la distance, en écart-type, entre l'énergie de la structure native E_N et l'énergie moyenne des leurres ($Z_{\text{score}} = (E_N - < E_D >)/\sigma_{E_D}$), le coefficient de corrélation entre les énergies et les cRMSds ou les TM-scores, et le rang de la structure native ou PSN de plus basse énergie. La figure A.5 en annexe A.3 trace les énergies *vs.* TM-scores et les énergies *vs.* cRMSds pour toutes les cibles, en utilisant DOPE et OPEP.

Pour les leurres minimisés, OPEP identifie correctement 23 conformations natives ou PSN sur 29 cibles, alors que DOPE en identifie correctement 19. OPEP et DOPE présentent tous les deux des valeurs de Z-score négatives, sauf pour la cible 1bba-*lmds* du côté OPEP (mais la plupart des leurres sont en réalité PSN, voir la figure A.5), et pour *beta-nova*, 1bba-*lmds* et 1fsd du côté de DOPE. Globalement, le Z-score moyen est légèrement inférieur du côté d'OPEP (-2,07 *vs* -1,42 en utilisant DOPE), les écarts-types restant similaires (1,03 *vs* 1,11 en utilisant DOPE). Il est intriguant de noter que les valeurs moyennes de Z-score sur les JA et JS sont très similaires. Les coefficients de corrélation entre les énergies et les TM-scores et les énergies et les cRMSds sont quasi identiques. En utilisant

à la fois les JA et JV, les coefficients de corrélation moyens sont -0,574 *vs.* -0,557 pour l'énergie contre le TM-score et 0,543 *vs.* 0,557 pour l'énergie contre le cRMSd en utilisant OPEP et DOPE respectivement.
En partant des leurres publiques disponibles (*Decoys 'R' Us*, Rosetta et CASP6), *ie.* sans aucune minimisation, ni repositionnement des chaînes latérales, sur quatorze cibles, DOPE identifie correctement dix d'entre elles, mais échoue pour 1bba-ldms, 1khm-*semfold*, 1s04-*casp6* et 1te7-*casp6*, alors qu'il échoue pour cinq cibles sur 14 en utilisant les cibles minimisées. Le Z-score moyen est de -2,95, avec une valeur extrême de 15 pour 1bba-ldms, et -10,1 pour 1ctf-lattice. Les valeurs correspondantes sur les leurres minimisés sont de -1,77, 0,62 et -4,75. Sur ces mêmes cibles non minimisées, OPEP échoue pour trois d'entre elles (2cro-*fisa*, 1ctf-*lmds* et 1s04-*casp6*), et donc identifie correctement onze cibles sur quatorze, avec une valeur moyenne de Z-score de -2,05. Les tracés des énergies *vs.* TM-scores et cRMSds sont présentés en annexe A.3, figure A.6.

8.2.5 Impact des propensités

Si nous poussons plus loin l'analyse de la performance d'OPEP 3.1, nous pouvons nous poser la question de savoir : quel impact aurait sur la reconnaissance une valeur nulle pour les potentiels de propensité en hélice α et en feuillet β ?
Cette version d'OPEP pour ne numéro 3.2. Dans ce but, nous avons reitéré le processus d'optimisation sur le JA non redondant, et analysé les résultats sur les JA et JV complets. Nous avons constaté que le nombre d'inégalités satisfaites après optimisation ne varie que très peu par rapport à la version 3.1 : 2.583.679 sont ici maintenant satisfaites contre 2.748.877 dans la version 3.1. Cependant, la table 8.7 montre que l'identification de la structure d'énergie minimale reste identique pour toutes les cibles, et le rang de la structure native ne varie substantiellement que pour 1f4i (elle est classée 125 pour la version 3.2 contre 6 pour la version 3.1). Dans son ensemble, nous ne pouvons distinguer cette version 3.2 de la version 3.1 sur la base de l'écart énergétique entre la structure native ou PSN et le reste des conformations non natives.

8.2.6 OPEP est-il pertinent pour des études cinétiques et thermodynamiques ?

Bien que la capacité d'un champ de force à distinguer des structures natives de structures mal repliées soit essentielle dans le repliement protéique, cela ne garantie pas que ce champ de force procure une description précise de la dynamique de notre protéine autour et hors du bassin natif.
Autour du minima, les fluctuations thermiques des atomes sont bien décrites. Ceci a été démontré par des simulations de dynamique moléculaire utilisant la version 3 d'OPEP : les trois modèles protéiques simulés ont montré des valeurs de RMSF (*Root Mean Square*

CIBLE	$E(N)$	$E(L_{min})$	$E(PSN_{min})$	N_{cl}	PSN_{cl}	Rc
Jeu d'Apprentissage Complet						
1abz	-75,8	-81,7	-78,8	31	11	Non
betaNonva	-29,1	-36,3	-29,0	33	36	Non
1dv0	-103,3	-103,2	-103,2	1	2	Oui
1e0m	-49,2	-59,3	-52,5	38	16	Non
1orc	-137,1	-146,1	-146,1	5	1	Oui
1pgb	-142,4	-156,6	-156,6	19	1	Oui
1pgbF	-29,9	-30,2	-30,2	17	1	Oui
1qhk	-100,0	-97,1	-88,5	1	6	Oui
1shg	-140,8	-144,0	-144,0	5	1	Oui
1ss1	-119,1	-111,5	-111,5	1	2	Oui
1vii	-71,1	-72,3	-72,3	5	1	Oui
2ci2	-175,0	-168,5	-168,5	1	2	Oui
2cro-*fisa*	-136,6	-155,4	-154,7	122	3	Non
Jeu de Validation Complet						
1bba-*lmds*	-46,5	-55,4	-55,4	384	1	Oui
1ctf-*4state*	-174,3	-175,8	-175,8	2	1	Oui
1ctf-lattice	-174,3	-150,6	-138,5	1	12	Oui
1ctf-*lmds*	-174,3	-178,6	-173,7	2	3	Non
1ctf-*semfold*	-174,3	-177,2	-177,2	3	1	Oui
1f4i	-75,0	-86,9	-86,9	125	1	Oui
1fsd	-53,7	-56,6	-56,6	11	1	Oui
1khm-*semfold*	-162,6	-143,5	-143,0	1	3	Oui
1r69	-136,4	-136,4	-136,4	2	1	Oui
1r69-*4state*	-136,4	-1330	-133,0	1	2	Oui
1r69-rosetta	-135,9	-137,0	-136,3	3	1	Oui
1s04-*casp6*	-284,3	-309,5	-301,9	33	2	Non
1te7-*casp6*	-256,2	-281,6	-281,6	25	1	Oui
1ubi-*lmdsv2*	-169,0	-183,4	-183,4	70	1	Oui
2cro-*4state*	-136,6	-148,9	-148,9	30	1	Oui
2cro-*lmds*	-136,6	-148,9	-148,9	38	1	Oui

Tab. 8.7: Performance d'OPEP 3.2 sur les jeux d'apprentissage et de validation. Pour la légende voir la table 8.5.

Fluctuation) comparables à celles trouvées avec des champs de forces tous atomes de mécanique moléculaire (Derreumaux and Mousseau, 2006). Hors du minima, OPEP doit générer des barrières énergétiques et des états de transition consistants avec les calculs de mécanique quantique et de DM. L'étude détaillée du repliement d'un motif de type épingle à cheveux de seize résidus de long, en utilisant des simulations ART-OPEP (version 2), a permis d'identifier deux chemins de repliement proches de ceux observés lors d'études théoriques antérieures (Wei et al., 2004). De plus, des simulations de dynamique moléculaire, avec la version 3 d'OPEP, ont permis d'identifier des mécanismes d'agrégation de fibres amyloïdes (quatre chaînes du fragment Aβ_{16-22}) compatibles avec les résultats

expérimentaux (Derreumaux and Mousseau, 2006).
Mais de la même manière, ces résultats ne garantissent pas que OPEP soit pertinent en terme de thermodynamique. Cependant, les récents résultats de quatre études utilisant cette version 3 d'OPEP sont très encourageants : (i) les simulations de MC du fragment Aβ_{21-30} en solution sont très proches des structures expérimentales RMN (Chen et al., 2006), (ii) des simulations de dynamique moléculaire à 300 K ont mis en évidence que le faisceau de trois hélices et le domaine B1 de la protéine G sont stables sur une échelle de temps de 50 ns (Derreumaux and Mousseau, 2006), (iii) les simulations REMD-OPEP (*Replica Exchange Molecular Dynamics*) du motif en épingle à cheveux de 1pgb (1pgbF) ont permis de calculer une température de fusion de 295 K contre 297 K déterminée expérimentalement (résultats non publiés), et (iv) en utilisant huit répliques avec T variant de 287 K à 500 K, les surfaces d'énergie libre déterminées par REMD-OPEP, pour les dimères de Aβ_{16-22} à 310 K, sont très proches de celles générées par des simulations REMD tous atomes en solvant explicite (Wei et al., 2007).

8.3 Conclusions de l'étude

Nous avons ré-évalué les paramètres de la fonction d'énergie OPEP sur 29 cibles avec un total de 28553 leurres. Ces leurres ont été produits par différents protocoles ou extraits de jeux de leurres publics pour limiter l'impact de la méthode les ayant générés. Les paramètres d'OPEP sont optimisés par un algorithme génétique et de MC couplés à une fonction de score nécessitant que la structure native et les structures PSN soient de plus basse énergie que les leurres.
Vu dans son ensemble, OPEP identifie correctement 24 conformations natives ou PSN sur 29 cibles. Ce résultat semble significatif, puisque la fonction d'énergie tous atomes DOPE, qui est connue pour être plus performante que cinq champs de forces reconnus, présente des capacités de reconnaissance similaires. OPEP ne reconnaît pas le bassin énergétique de cinq cibles que sont 1e0m, 2cro-*fisa*, 1ctf-*lmds*, *betanova* et 1s04-*casp6*. Cependant, la structure de 1ctf a été résolue avec ion sulfate à proximité, et la structure de 2cro présente des effets d'empaquetage au sein du cristal. Ceci suggère que les ions et les contacts entre sous-unités doivent être pris en compte dans notre potentiel, pour pouvoir identifier ces structures natives (McConkey et al., 2003). Par contre, les cibles 1e0m, *betanova* et 1s04 ne présentent pas de telles caractéristiques.
Distinguer de telles structures nécessiterait une fonction d'énergie plus complexe. Deux voies pourraient être explorées pour ce faire : inclure un terme de plus haut degré pour les paires de dièdres ϕ-ψ (Sims and Kim, 2006), et/ou introduire un terme d'interaction à plusieurs corps entre les chaînes latérales (Shimizu and Chan, 2002).
Une autre explication de ces échecs pourrait résider dans le fait que les structures expérimentales ne sont pas associées à l'énergie minimale effective, mais au minimum global de

l'énergie effective et de l'entropie conformationnelle (Caflisch, 2006).

Chapitre 9

Implémentation d'OPEP dans l'algorithme glouton

Une conséquence du couplage d'OPEP avec l'algorithme glouton est lié au processus d'assemblage rigide, qui impose la géométrie de valence de l'assemblage. Pour cette raison, l'ensemble des termes énergétiques du potentiel gros grain OPEP ne sont pas pris en compte. Cette version simplifiée d'OPEP, implémentée dans l'algorithme glouton, sera notée sOPEP (*simplified Optimized Potentiel for Efficient protein structure Prediction*) dans le reste de ce manuscrit. Un certain nombre de problèmes se posent de par la discrétisation de la recherche. OPEP a été conçu dans un espace continu, et l'algorithme glouton évolue dans un espace discret.

Une fois le potentiel mis en place, nous avons entrepris de modifier les paramètres du potentiel d'interaction entre les chaînes latérales, trop stringents pour la méthode de génération des modèles. L'ensemble des paramètres a ensuite été optimisé, et est maintenant utilisé en production. Cette formulation sera notée sOPEP v2.0 par opposition à la version initiale de sOPEP (v1.0) implémentée dans l'algorithme glouton, et la version optimisée sOPEP v2.1 par analogie à la version 3.1 optimisée d'OPEP.

9.1 La formulation de sOPEP

La formulation initiale de sOPEP (v1.0) que nous avons implémentée dans l'algorithme glouton était directement issue de la version 3.0 du champ de force OPEP, *i.e.* dans sa version non optimisée. Les termes énergétiques associés à la géométrie du modèle ne semblant, de prime abord, pas pertinent dans la méthode de reconstruction, ils n'ont pas été considérés. Seule la pénalité énergétique associée aux angles dièdres $\Phi > 0$. Par ailleurs, dans le cadre du développement d'une méthode de modélisation par homologie, nous avons ajouté un potentiel énergétique E_{RMSd}, nous permettant d'associer une pseudo-énergie pénalisant toute déviation du modèle par rapport à une structure matrice.

Ce potentiel est inactivé en prédiction *ab initio*.

La formulation de sOPEP est donc la suivante :

$$E = E'_{locale} + E_{non-liante} + E_{liaisons-H} + E_{RMSd} \tag{9.1}$$

avec

$$E'_{locale} = \sum_{\phi} k_{\phi\psi}(\phi - \phi_o)^2 \tag{9.2}$$

Les formulations de $E_{non-liante}$ et $E_{liaisons-H}$ étant strictement identiques à celles d'OPEP v3.0, présentées dans les équations 5.5, et 5.7 (E_{LH1}),5.10 (E_{LH2}) respectivement.

Le terme E_{RMSd} a été mis en place dans le cadre du développement d'une méthode de modélisation par homologie utilisant l'algorithme glouton qui sera présentée dans la suite de ce manuscrit. Sa formulation est la suivante :

$$E_{RMSd} = k_R * \sum_{i=0}^{N} \begin{cases} (d_i - d_0)^2, & \text{si } d_i > d_0 \\ 0, & \text{sinon} \end{cases} \tag{9.3}$$

avec la constance de force $k_R = 0,8$ kCal.mol^{-1}.Å^{-1}, d_i la distance entre le $i^{\text{ème}}$ carbone *alpha* de la matrice (superposée au modèle en croissance) et le carbone *alpha* du modèle correspondant, imposé par l'alignement des séquences du modèle et de la matrice, et $d_0 = 1$ Å, la distance critique à partir de laquelle la déviation par rapport à la matrice devient pénalisante.

9.2 Re-paramétrisation du potentiel entre les chaînes latérales

Comme nous l'avons déjà évoqué à plusieurs reprises, le potentiel de force moyenne s'établissant entre les chaînes latérales semble trop stringent pour l'algorithme glouton, conduisant bien souvent à des modèles protéiques avec un faible degré de compacité. Nous étions alors obligés de tronquer chaque interaction CL-CL à une valeur maximale de 2 kCal/mol pour partiellement résoudre le problème. Nous avons donc récemment entrepris de le re-paramétrer afin qu'il soit plus en corrélation avec les distributions des distances d'interaction CL-CL observées.

9.2.1 Le cas général

Afin d'optimiser les rayons des sphères représentant les chaînes latérales, nous avons calculé, sur une PDB non redondante à 30%, la distribution des distances intercentroïdes lorsque deux résidus sont considérés comme étant en interaction. Ces données sont les même que celles nous ayant servi à optimiser les rayons d'OPEP v3.0 (section 5.1). L'objectif à atteindre a été de trouver une formulation qui nous permette d'avoir un potentiel plus permissif, s'ajustant au mieux sur ces distributions.

D'une manière générale, le potentiel de force moyenne s'exerçant entre les chaînes latérales (équation 5.6) peut s'exprimer comme suit :

$$E_{CL,CL}(r_{ij}) = \begin{cases} -\epsilon_{ij} * C(r_{ij})^6, \text{ pour } \epsilon_{ij} < 0, \\ (C(r_{ij})^{12} - 2 * C(r_{ij})^6) \text{ sinon.} \end{cases} \tag{9.4}$$

avec $C(r_{ij}) = \frac{r^0_{ij}}{r_{ij}}$. Sans changer la forme du potentiel, ni le degré, nous pouvons introduire un nouveau paramètre p_{ij}, tel que :

$$C(r_{ij}) = \frac{r^0_{ij} - p_{ij}}{r_{ij} - p_{ij}}, \quad \text{pour } \epsilon_{ij} > 0. \tag{9.5}$$

Ce nouveau paramètre nous permet d'imposer la valeur seuil gR^0_{ij} à partir de laquelle le potentiel devient pénalisant ($E > 0$). Ainsi, si l'on résoud l'équation :

$$E(gR^0_{ij}) = 0 \tag{9.6}$$

nous pouvons en déduire la valeur de p_{ij} pour $\epsilon_{ij} > 0$:

$$p_{ij} = \frac{r^0_{ij} - \sqrt[6]{2} * gR^0_{ij}}{1 - \sqrt[6]{2}} \tag{9.7}$$

Dans le cas où $\epsilon_{ij} < 0$, nous avons appliqué un facteur d'échelle permettant de décaler le potentiel vers des valeurs plus faibles en fonction de gR^0_{ij}, ainsi, $C(r_{ij})$ s'exprime par

$$C(r_{ij}) = \frac{2 * gR^0_{ij} - r^0_{ij}}{r_{ij}}, \quad \text{pour } \epsilon_{ij} < 0. \tag{9.8}$$

Les valeurs gR^0_{ij} sont fixées, pour chaque interaction, à la valeur du quantile correspondant à une probabilité de 0,1 pour $\epsilon_{ij} > 0$ et 0,2 pour $\epsilon_{ij} < 0$. L'ensemble des paramètres gR^0_{ij} est donné en annexe dans la table A.7, ainsi que les 210 tracés de cette nouvelle formulation, dans la figure A.8. Deux exemples sont présentés dans la figure 9.1 : nous pouvons remarquer que, aussi bien dans le cas d'interactions défavorables (Arginine-Arginine où $\epsilon_{ij} < 0$) que favorables (Tryptophane-Tryptophane où $\epsilon_{ij} > 0$), la formulation d'origine (OPEP v3.0) est très stringente et pénalise près de la moitié des interactions observables dans un ensemble de structures expérimentales.

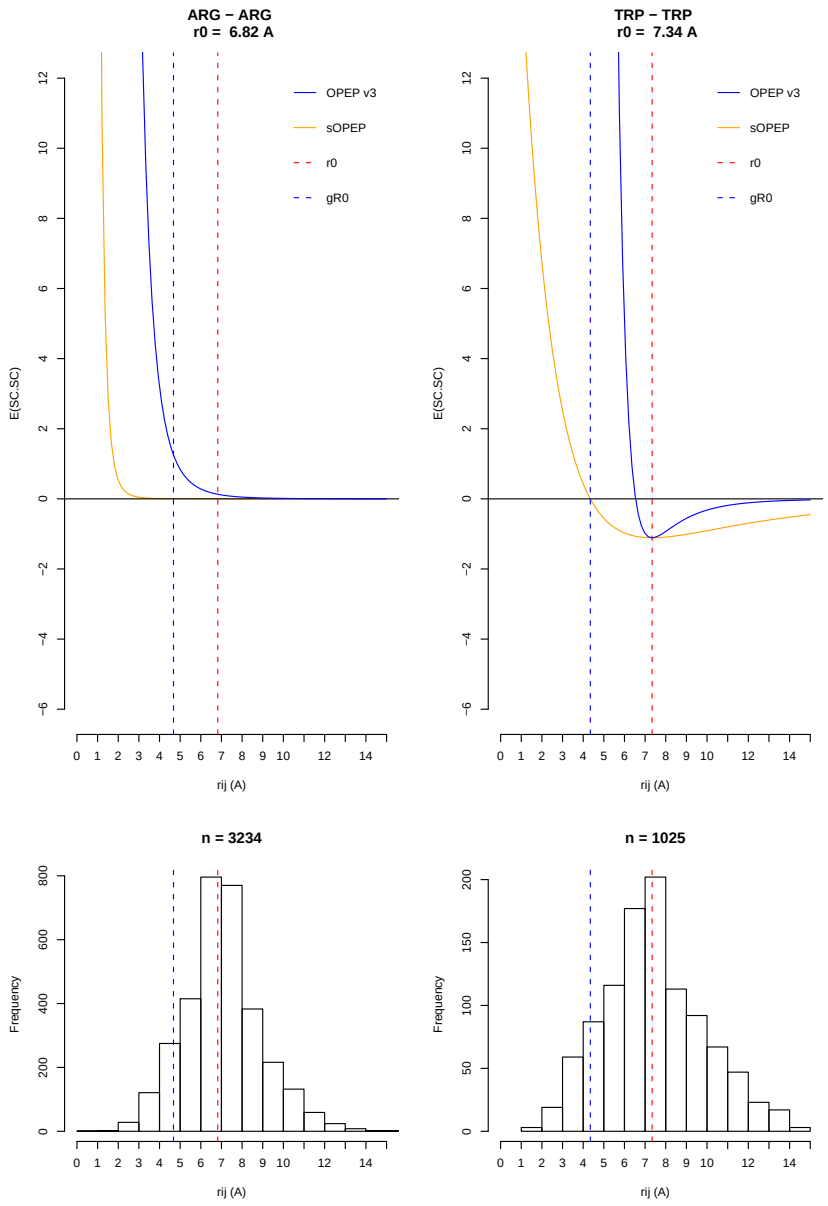


Fig. 9.1: La nouvelle formulation du potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).

9.2.2 Optimisation des nouveaux paramètres

Matériels et méthodes

Une fois la formulation et les nouveaux paramètres de sOPEP v2.0 fixés, nous les avons réoptimisés en suivant le même protocole d'optimisation que celui utilisé pour OPEP v3.0 (chapitre 8). Cette version optimisée est notée 2.1. Partant du vecteur optimal obtenu pour OPEP v3.1, nous avons réalisé à nouveau des cycles d'optimisation avec les contraintes suivantes : (i) seuls les 210 poids associés aux interactions CL-CL sont optimisés avec la

contrainte de rester dans l'intervale [0,7 ; 1,3], (ii) les poids associés aux termes de géométrie ne sont pas optimisés et sont initialisés à 0, et (iii) le JA est identique au JA v3.0 (section 8.1.3), à l'exception de *betanova*, jugée trop versatile pour être utilisée comme structure de référence dans le cadre d'une optimisation. Dans ce cas précis, les cibles du JA sont donc 1abz, 1dv0, 1e0m, 1orc, 1pgb, 1pgbf, 1qhk, 1shg, 1ss1, 1vii, 2ci2 et 2cro-*fisa*.

Résultats

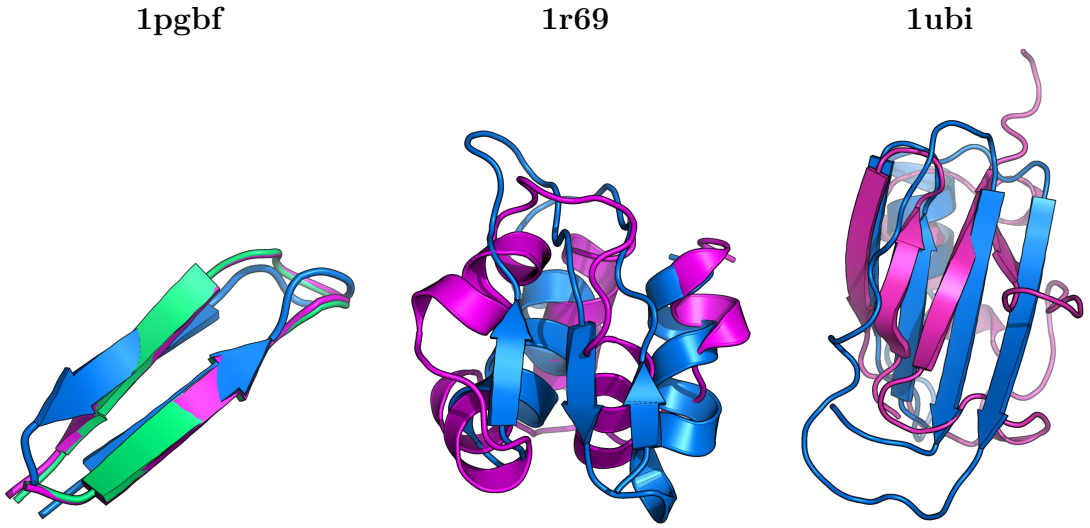


Fig. 9.2: Optimisation sOPEP, les cibles problématiques. La structure native (en magenta) est superposée au modèle de plus basse énergie (en bleu). Pour 1pgbf, le modèle PSN de plus basse énergie (en vert) est aussi superposé à la structure native.

Les tracés de l'énergie *versus* le TM-score pour cette optimisation sont présentés en annexe, dans la section A.6. Globalement, cette implémentation de sOPEP v2.1 semble discriminante pour 22 cibles sur les 29 totales, *versus* 23 pour OPEP v3.1. Dans l'ensemble, le pouvoir discriminant de sOPEP est quasi identique à celui de la version 3.1. Pour les cibles, 1abz et 2cro-*fisa* qui n'étaient pas réellement problématiques, la structure de plus basse énergie est maintenant dans une conformation PSN. La cible du JA 1pgbf présente une répartition Energie/TM-score similaire à celle obtenue pour OPEPv3.1 (voir l'annexe A.6), cinq leurres présentant des énergies faibles pour des TM-scores autour de 0,3. Cependant, il existe un ensemble de conformations PSN ayant des énergies très proches (moins de 0,3 kCal/mol), et comme nous pouvons le voir sur la figure 9.2, la topologie du leurre est bonne, le feuillet β étant positionné à l'identique dans la structure native. Encore une fois, il semble que le TM-score ne soit pas pertinent pour des structures de petite taille.

Pour la cible 1r69, deux conformations non natives présentent des énergies proches de celle de la structure native. Ces deux leurres d'enfilage ont une topologie de type

CIBLE	$E(N)$	$E(L_{min})$	$E(PSN_{min})$	N_{cl}	PSN_{cl}	Rc
Jeu d'apprentissage complet						
1abz	-80.219	-91.181	-91.181	40	1	Oui
1dv0	-95.409	-98.162	-98.162	4	1	Oui
1e0m	-51.217	-70.616	-63.795	76	14	Non
1orc	-135.379	-145.924	-145.924	4	1	Oui
1pgb	-146.410	-156.836	-156.836	16	1	Oui
1pgbf	-28.729	-29.123	-28.821	24	6	Non
1qhk	-113.684	-101.289	-101.289	1	1	Oui
1shg	-142.738	-140.733	-140.733	1	1	Oui
1ss1	-122.505	-123.510	-123.510	2	1	Oui
1vii	-53.101	-68.875	-68.875	201	1	Oui
2ci2	-126.608	-149.012	-149.012	144	1	Oui
2cro-*fisa*	-143.160	-153.970	-153.970	89	1	Oui
Jeu de validation complet						
betanova	-28.102	-36.347	-29.770	39	26	Non
1bba-*lmds*	-47.360	-55.987	-55.987	480	1	Oui
1ctf-*4state*	-160.834	-160.388	-160.388	1	1	Oui
1ctf-lattice	-160.834	-130.221	-123.022	1	27	Oui
1ctf-*lmds*	-160.834	-170.021	-168.701	35	2	Non
1ctf-*semfold*	-160.834	-154.179	-154.179	1	1	Oui
1f4i	-91.694	-104.325	-104.325	12	1	Oui
1fsd	-40.334	-54.825	-54.825	205	1	Oui
1khm-*semfold*	-164.513	-145.704	-145.704	1	1	Oui
1r69	-124.212	-126.991	-124.386	3	2	Non
1r69-*4state*	-124.212	-120.701	-120.701	1	1	Oui
1r69-baker	-122.570	-125.152	-125.152	4	1	Oui
1s04	-326.089	-337.288	-329.939	8	3	Non
1te7	-273.909	-294.394	-294.394	27	1	Oui
1ubi	-159.392	-181.350	-169.136	122	13	Non
2cro-*4state*	-143.160	-155.135	-155.135	28	1	Oui
2cro-*lmds*	-143.160	-153.970	-153.970	31	1	Oui

Tab. 9.1: Performance de sOPEP sur les JA et JV complets. $E(N)$ est l'énergie de la structure native, $E(L_{min})$ est la plus basse énergie obtenue pour la cible (la structure native exclue) et $E(PSN_{min})$ est l'énergie la plus basse des structures PSN. N_{cl} est le classement de la structure expérimentale, PSN_{cl} est le classement de la structure PSN de plus basse énergie. Rc est l'état de reconnaissance, "Oui" indiquant que la structure de plus basse énergie est une structure native ou PSN. Les valeurs soulignées mettent en évidence les cibles pour lesquelles l'état de reconnaissance a changé entre la version 3.1 d'OPEP et sOPEP v2.1.

$\beta\alpha\beta\beta\alpha$ au lieu des 5 hélices α de la structure native. Leur position se justifie par la forte contribution énergétique des liaisons hydrogène associées au feuillet β.

Le leurre de plus basse énergie pour la cible 1ubi présente une topologie proche de la structure native, mais le brin 4 (le plus à droite) est parallèle alors qu'il devrait être anti-parallèle (voir la figure 9.2). Comme pour la cible 1r69, une forte coopérativité des

liaisons hydrogène β est responsable d'un tel classement.

Ces résultats nous suggèrent que l'optimisation à une légère tendance à favoriser les brins β.

Conclusions

La version optimisée de sOPEP v2.1 semble avoir un pouvoir discriminant similaire à la OPEP v3.1 sur un ensemble de modèles protéiques dans un espace continue. La pertinence d'une telle formulation dans un espace discret semble encore à déterminer.

9.2.3 Les ponts disulfure

L'implémentation de sOPEP v1.0 ne prenait pas en compte les ponts disulfure. Comme nous avons reparamétré les interactions CL-CL, nous en avons profité pour implémenter des paramètres spécifiques aux interactions entre cystéines.

Matériels et Méthodes

Les paramètres. Dans la paramétrisation classique, les cystéines oxydées ou réduites sont confondues en une seule distribution des interactions CYS-CYS. Pour la paramétrisation spécifique aux ponts disulfure, nous avons séparé les cystéines oxydées et réduites en deux distributions distinctes. Ainsi, nous avons re-paramétrisé le potentiel pour les cystéines réduites avec le même protocole déjà présenté (section 9.2.1) sur la distribution des distances d'interactions de cystéines réduites.
Pour le potentiel associé aux ponts disulfure, nous avons testé deux formulations spécifiques. Dans les deux cas, nous avons fixé la valeur minimale du potentiel (E_{SS}^{min}) à -5 kCal/mol (Chen, 1999).

La première formulation (notée sOPEP v2.1.2) est identique à celle présentée dans la section précédente (notée sOPEP v2.1.1), si ce n'est que, pour maîtriser la valeur minimale du potentiel, nous avons ajouté un facteur d'échelle S à l'équation 9.4, ainsi, cette expression devient, pour $\epsilon_{ij} > 0$:

$$E_{SS}(r_{ij}) = S * \epsilon_{ij} \left(C(r_{ij})^{12} - 2 * C(r_{ij})^{6} \right) \tag{9.9}$$

Si l'on résoud l'equation $E_{SS}(r_{ij}^{0}) = E_{SS}^{min}$, nous pouvons en déduire la valeur de S, $S = -E_{SS}^{min}/\epsilon_{ij}$. L'équation précédante devient donc :

$$E_{SS}(r_{ij}) = -E_{SS}^{min} \left(C(r_{ij})^{12} - 2 * C(r_{ij})^{6} \right) \tag{9.10}$$

Avec cette première formulation, nous ne pénalisons pas les cystéines qui devraient être impliquées dans un pont disulfure et qui restent réduites, nous avons donc envisagé

une deuxième formulation du second ordre pour le potentiel associé aux ponts disulfure (notée sOPEP v2.1.3). Sachant que cette deuxième formulation doit aussi respecter les contraintes de la distribution des distances intercentroïdes pour les cystéines oxydées, nous avons déduit l'expression suivante :

$$E_{SS}(r_{ij}) = S * (r_{ij} - r_{ij}^0)^2 + E_{SS}^{min} \tag{9.11}$$

si l'on résoud l'équation $E_{SS}(gR_{ij}^0) = 0$, nous pouvons en déduire la valeur de S :

$$S = -E_{SS}^{min}/(gR_{ij}^0 - r_{ij}^0)^2 \tag{9.12}$$

L'équation 9.11 devient alors :

$$E_{SS}(r_{ij}) = E_{SS}^{min} \left(1 - \left(\frac{r_{ij} - r_{ij}^0}{gR_{ij}^0 - r_{ij}^0}\right)^2\right) \tag{9.13}$$

Les trois formulations du potentiel d'interactions CYS-CYS, en plus de la re-paramétrisation spécifique aux cystéines réduites sont tracées dans la figure 9.3.

Le jeux de validation. Afin de valider la nouvelle formulation de l'énergie associée aux ponts disulfure, nous avons constitué un jeu de validation composé de 16 structures expérimentales présentées dans la table 9.2. Partant d'une liste de structures PDB partageant toutes moins de 30% d'identité de séquence, nous avons sélectionné toutes les structures RX de moins de 50 résidus ayant au moins un pont disulfure. Les structures possédant plus d'un fragment sont ignorées. Les structures résultantes ont entre 2 et 5 ponts disulfure, et se répartissent comme suit en terme de classe : 8 β, 1 α, 6 α/β et 1 structure sans structure secondaire canonique.

L'opérateur zip. Afin d'apporter de la diversité dans la procédure de génération des modèles, et de dépasser la contrainte de faire croître linéairement les structures d'une extrémité à l'autre, nous avons implémenté un nouvel opérateur de reconstruction. Ce nouvel opérateur nous permet de commencer la reconstruction d'une position donnée dans la trajectoire, et de faire croître la structure alternativement par son extrémité N terminale et C terminale respectivement (voir la figure 9.4). Deux nouveaux paramètres sont introduits avec cet opérateur : d la position de départ, et A l'amplitude de reconstruction à chaque itération, *i.e.* le nombre de fragments à ajouter avant de changer le sens de croissance. Ainsi, par exemple, pour une protéine de 60 résidus de long (et donc une trajectoire de 57 positions), si $d = 34$, $A = 2$ et que nous commençons la reconstruction de l'extrémité N vers C terminale, à la fin de la troisième itération, nous avons reconstruit les résidus 34 à 39, puis à la cinquième itération, les résidus 32 à 39, etc. Une fois une extrémité atteinte, la reconstruction se poursuit alors normalement uniquement dans l'extrémité opposée.

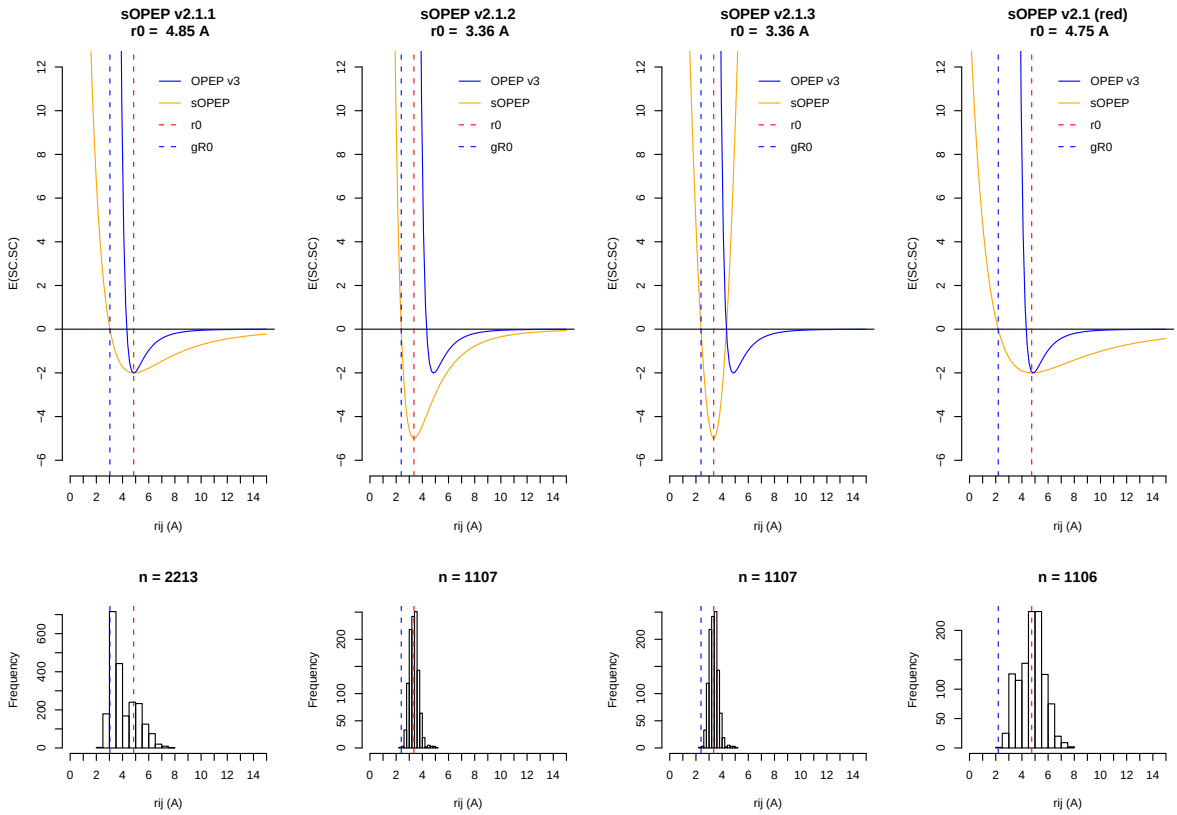


Fig. 9.3: sOPEP : Le potentiel associé aux ponts disulfure. Sont présentés sur ces graphiques les tracés des potentiels associés aux interactions CYS-CYS (en orange), en fonction des distributions des distances intercentroïdes en interaction dans la partie inférieure. De gauche à droite, (i) le potentiel classique sOPEP v2.1.1 ne distinguant pas les cystéines réduites d'oxydées, (ii) le potentiel classique modifié sOPEP v2.1.2 pour les cytéines oxydées, (iii) le potentiel sOPEP v2.1.3 pour les cystéines oxydées toujours, et (iv) la re-paramétrisation du potentiel pour les cystéines réduites. L'ancienne formulation du potentiel issue d'OPEP v3 est dans tous les cas tracée en bleu dans les graphiques.

L'avantage d'un tel opérateur est qu'il permet de mettre en place des contacts au sein de la structure lors des premières étapes du repliement, et ainsi le guider vers la solution native. Cependant, les reconstructions sont alors très dépendantes du choix de la position de départ. Nous avons remarqué, lors de nos tests de reconstruction sur un ensemble de peptides, que seules certaines positions départ nous permettaient de former un motif hélice-boucle-hélice, les autres positions ne conduisant qu'à une longue hélice.
Dans une optique *ab initio*, la position de départ est tirée au hasard, ainsi, la contrepartie d'un tel opérateur, réside dans le fait qu'il faille lancer un grand nombre de simulations indépendantes.
Nous avons testé cet opérateur dans le cadre du développement d'une approche hiérarchique, et il semblerait qu'il conduise à de meilleurs résultats qu'une reconstruction linéaire des structures. C'est la raison pour laquelle, nous l'avons utilisé ici pour tester le potentiel

#	PDB	Res.	L	Protéine	# SS	% α	% β
1	1ajj	1,70	37	*LDL receptor ligand-binding module*	3	-	0,108
2	1bhp	1,70	45	*Beta-purothionin*	4	0,422	0,133
3	1clvI	2,00	32	*Alpha-amylase inhibitor*	3	-	0,188
4	1edmB	1,50	39	*Epidermal growth factor-like domain from human factor IX*	3	-	0,256
5	1ejgA	0,54	48	*Crambin*	3	0,435	0,087
6	1fd3A	1,35	41	*Human beta-defensin 2*	3	0,171	0,268
7	1fleI	1,90	47	*Elafin*	4	-	0,213
8	1gl1I	2,10	34	*Protease inhibitor LCMI II*	3	-	0,324
9	1h59B	2,10	45	*Insulin-like growth factor binding protein*	2	0,156	0,244
10	1ijuD	1,40	36	*Human beta-defensin-1*	3	0,194	0,333
11	1oc0B	2,28	37	*Somatomedin b domain of vitronectin*	4	-	-
12	1p9gA	0,84	41	*Antifungal protein*	5	-	0,195
13	1q9bA	1,50	43	*Hevein*	4	-	0,140
14	1xu1R	1,90	38	*Tumor necrosis factor receptor superfamily member 13*	3	-	0,211
15	1xu2R	2,35	36	*Tumor necrosis factor receptor superfamily member 17*	3	0,139	0,222
16	2erl	1,00	40	*Pheromone Er-1*	3	0,700	-

Tab. 9.2: Ponts disulfure : le jeu de validation. Pour chacune de ces protéines appartenant au jeu de validation, sont présentés : son identifiant PDB (PDB), la résolution de la structure RX expérimentale en Å (Res.), la longueur de la protéine en acides aminés (L), son nom (Protéine), le nombre de ponts disulfure qui la constitue (# SS), et la proportion de résidus en hélice α et feuillet β.

associé aux ponts disulfure. Nous reviendrons sur la performance de cet opérateur lors du chapitre 10.

Simulations. Pour chacune des cibles du JV et pour chaque version du potentiel associé aux ponts disulfure, nous avons effectué 30 simulations de repliement avec l'algorithme glouton guidé par sOPEP v2. Pour chaque simulation, nous utilisons l'opérateur zip en tirant aléatoirement la position de départ, la première croissance s'effectuant de l'extrémité C vers N terminale. Un seul cycle est réalisé avec une pile de 3000 (aléatoire à partir de 1000), suivi d'une simulation de MC de 300.000 pas permettant de muter ponctuellement les prototypes. Les trajectoires (description en terme de lettres HMM-SA) données en entrées sont des trajectoires floues obtenues par l'encodage *forward-backward* avec une probabilité supérieure à 10^{-6}.

Résultats

Les trois formulations du potentiel ont été testées sur l'ensemble des 15 peptides du jeu de validation (voir la table 9.2). Si l'on raisonne en terme de cRMSd moyen sur

				sOPEP v2.1.1				sOPEP v2.1.2				sOPEP v2.1.3			
	Native			R		SS		R		SS		R		SS	
PDB	**# SS**	$\mathbf{C}_p$	$\mathbf{C}_{p3}$	cRMSd	# SS	cRMSd	# SS	cRMSd	# SS	cRMSd	# SS	cRMSd	# SS	cRMSd	# SS
1ajj	3	15,3	6,2	2,375	1	2,375	1	**2,034**	**3**	**2,034**	**3**	**2,459**	**3**	**2,459**	**3**
1bhp	4	9,8	6,2	2,946	1	2,946	1	2,232	3	4,276	4	**2,195**	**4**	**2,195**	**4**
1clvI	3	13,6	5,2	4,865	0	4,923	1	2,624	2	4,668	3	2,801	2	3,012	3
1edmB	3	15,9	5,9	6,212	0	7,189	1	**5,491**	**3**	**5,491**	**3**	**3,457**	**3**	**3,457**	**3**
1fd3A	3	17,5	7,2	2,694	0	2,694	0	2,671	1	6,241	3	4,595	0	4,722	3
1fleI	4	17,8	6,6	5,797	1	6,742	2	5,725	3	5,725	3	4,458	2	4,907	4
1gl1I	3	12,6	5,7	5,731	0	7,282	2	**3,854**	**3**	**3,854**	**3**	**5,014**	**3**	**5,014**	**3**
1h59B	2	12,4	6,0	3,556	0	6,890	1	3,974	1	4,266	2	**2,239**	**2**	**2,239**	**2**
1ijuD	3	14,5	5,9	3,894	0	4,821	1	3,503	1	6,067	2	4,315	0	6,394	3
1oc0B	4	14,9	6,3	1,375	0	1,645	2	**1,272**	**4**	**1,272**	**4**	**1,237**	**4**	**1,237**	**4**
1p9gA	5	15,9	5,8	4,056	0	5,073	1	2,691	2	5,221	4	3,233	4	3,804	5
1q9bA	4	13,9	5,5	6,302	0	6,534	2	**3,044**	**4**	**3,044**	**4**	**3,045**	**4**	**3,045**	**4**
1xu1R	3	14,0	6,9	3,175	0	6,872	1	**1,623**	**3**	**1,623**	**3**	**1,386**	**3**	**1,386**	**3**
1xu2R	3	13,4	6,9	2,672	0	2,926	1	1,488	2	2,470	3	**1,765**	**3**	**1,765**	**3**
2erl	3	12,6	7,0	1,604	0	1,631	1	1,565	2	1,579	3	**1,476**	**3**	**1,476**	**3**
μ	-	-	-	*3,817*	*5,6*	*4,703*	*36,3*	*2,919*	*73,8*	*3,855*	*94,8*	*2,912*	*79,8*	*3,141*	*100,0*
δ	-	-	-	*1,635*	*11,6*	*2,147*	*15,7*	*1,377*	*25,7*	*1,739*	*11,1*	*1,246*	*35,7*	*1,554*	*0,0*

Tab. 9.3: Détection des ponts disulfure sur un ensemble de 16 peptides. Pour chacune des protéines du JV sont présentés : son code PDB (PDB), le nombre de ponts disulfure présents dans la structure (# SS), la complexité de la trajectoire HMM-SA associée en prenant en compte tous les prototypes (C_p) ou seulement trois prototypes par lettre (C_{p3}). Nous avons simulé le repliement de ces protéines avec trois versions différentes du champ de force sOPEP pour le potentiel associé aux ponts disulfure. Pour chacune de ces séries de simulations, sont présentés, le cRMSd et le nombre de ponts disulfure dans le modèle, si l'on prend comme meilleur modèle le plus faible cRMSd (R) ou celui qui possède le plus de ponts disulfure natifs (SS). Les cibles pour lesquelles le modèle de plus faible cRMSd possède tous les ponts disulfure natifs sont mises en gras.

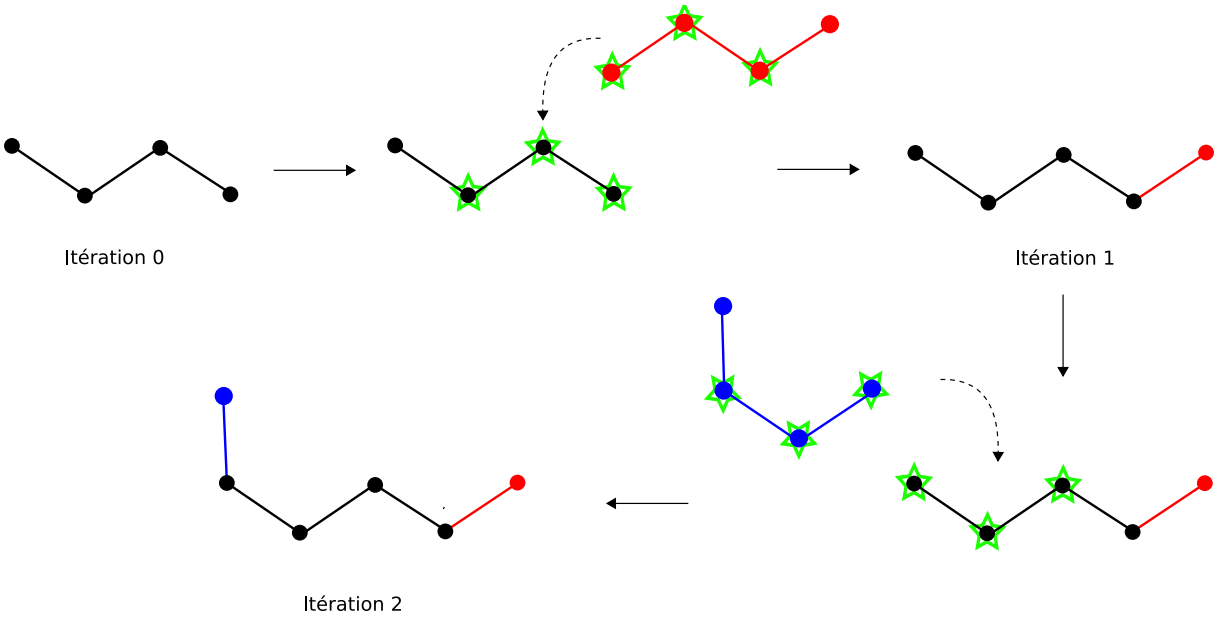
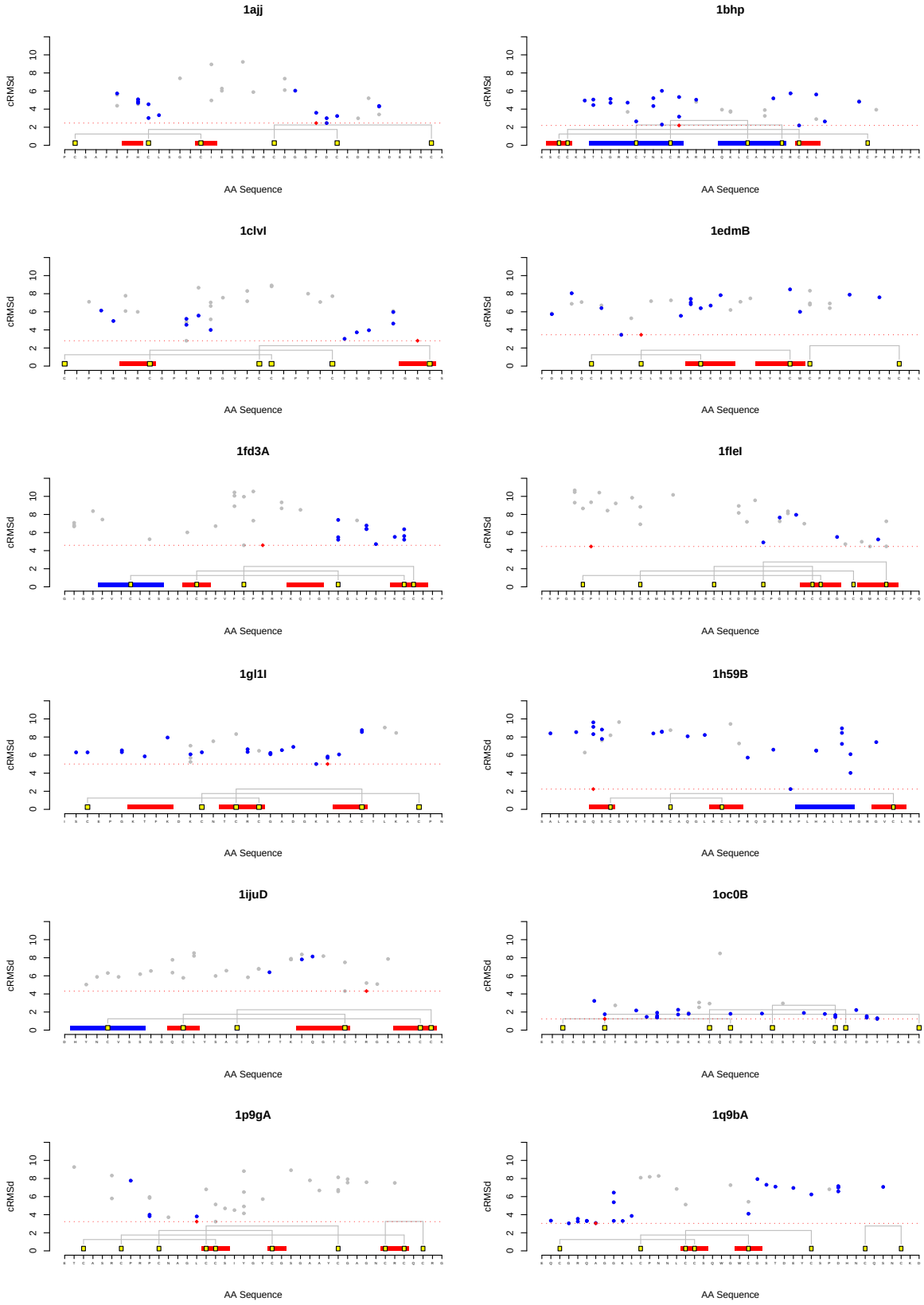


Fig. 9.4: L'opérateur de reconstruction "zip". L'opérateur zip nous permet à partir d'une position donnée dans la trajectoire de reconstruire alternativement les extrémités N et C terminale, par la même méthode de superposition que celle présentée dans la section 4.1.1 (les carbones α utilisés sont entourés d'une étoile verte). Dans le cas qui est présenté, nous avons une amplitude (A) de reconstruction de un.

les meilleurs modèles cRMSd, il semble que les nouvelles formulations sOPEP v2.1.2 et sOPEP v2.1.3 permettent d'accroître la qualité des modèles d'environ un angström en moyenne (3,8 Å dans la formulation sOPEP v2.1.1 contre 2,9 Å pour les deux autres). Bien que les structures de plus faible cRMSd ne forment pas l'ensemble des ponts disulfure. Cette adaptation du potentiel semble guider l'algorithme glouton vers un bassin énergétique favorable. Les deux nouvelles formulations sOPEP v2.1.2 et sOPEP v2.1.3 génére des modèles de qualité équivalente en terme de cRMSd (2,9 Å), mais la formulation sOPEP v2.1.3 permet de mieux forcer la détection des ponts disulfure, puisque pour 10 cibles sur 15, les modèles de plus faible cRMSd possèdent le profil natif de pont disulfure (*vs* 6 pour la formulation sOPEP v2.1.2). L'écart entre les meilleurs modèles en terme de cRMSd et en terme de nombre de ponts disulfure natifs devient alors très faible : 0,2 Å sur les moyennes, contre 0,9 Å pour la formulation sOPEP v2.1.2. Les nouvelles formulations nous permettent de générer des modèles avec l'intégralité des ponts disulfure natifs alors que la formulation classique sOPEP v2.1.1 n'en trouve que 36%, et avec des cRMSd plus élévés (4,7 Å en moyenne).

Pour la formulation sOPEP v2.1.3, sont présentés des résultats graphiques des simulations dans la figure 9.5. Pour l'ensemble des cibles, il semblerait que l'échantillonnage des positions de départ par l'opérateur zip couvre l'ensemble de la séquence. En ce qui concerne la position de départ, aucune conclusion ne peut être tirée quant à sa localisation


1ajj
1bhp
1clvl
1edmB
1fd3A
1flel
1gl1l
1h59B
1ijuD
1oc0B
1p9gA
1q9bA
cRMSd
AA Sequence

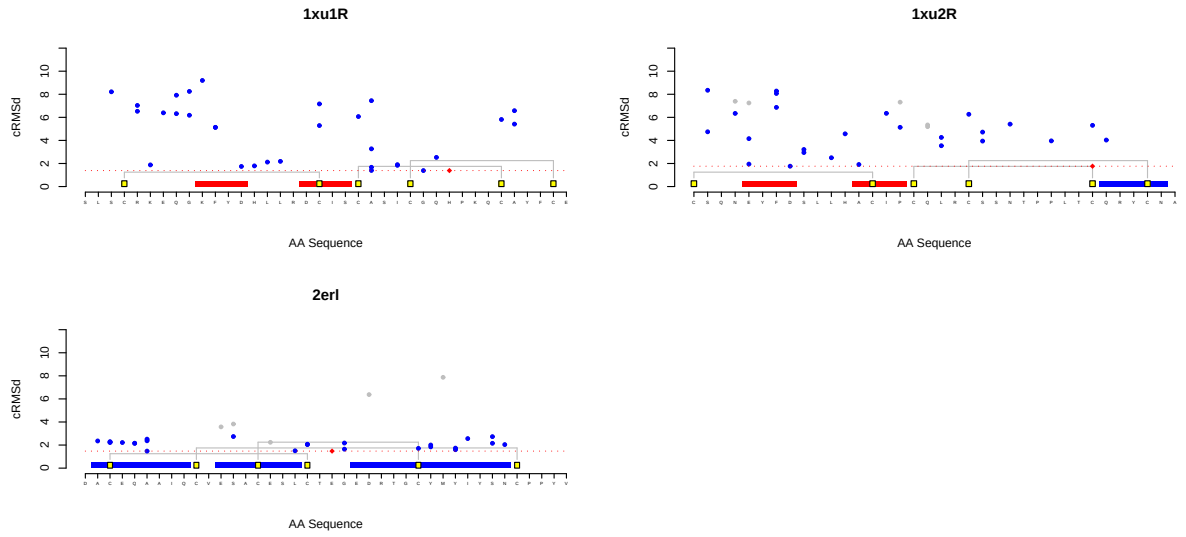


Fig. 9.5: sOPEP v2.1.3 analyse des simulations. Pour chaque cible, chaque point correspond au résultat d'une simulation, tracé en fonction du point de départ de l'opérateur zip le long de la séquence de la cible, et du cRMSd du modèle généré. Les zones de couleur tracées le long de la séquence correspondent aux structures secondaires : les brins β en rouge, et les hélices α en bleu. Les cystéines sont représentées par des carrés jaunes, les pontages gris indiquant les ponts disulfure. Les modèles possédant l'ensemble des ponts disulfure natifs sont tracés en bleu (gris étant la couleur par défaut). Le modèle de plus faible cRMSd est représenté quant à lui par un diamant rouge, la ligne pointillé horizontale représentant cette valeur extrême.

idéale, des modèles pertinents pouvant être générés en partant du coeur des structures secondaires ou en dehors, à proximité des cystéines oxydées ou à distance.
Il est intéressant de remarquer que, contrairement à ce que l'on aurait pu penser, la qualité des modèles n'est pas en relation directe avec le nombre de ponts disulfure contraignant la reconstruction : pour la cible 1p9gA, les cinq ponts disulfure natifs trouvés ont généré un modèle éloigné de 3,8 Å de la structure native. Plus généralement, dans de nombreux cas, nous pouvons replier des structures ayant le bon profil de ponts disulfure et une topologie éloignée de la structure native (cRMSd $>$ 6 Å). Dans le cas de 1ijuD, nous n'avons trouvé que trois fois la bonne combinaison de ponts disulfure sur l'ensemble des trente simulations, et l'écart entre le meilleur modèle cRMSd et le meilleur modèle ayant le bon profil de ponts disulfure, est important (2 Å).

Conclusions de l'étude

Nous avons ajouté un nouveau potentiel associé à la connectivité des cystéines dans le champ de force sOPEP. La formulation sOPEP v2.1.2, la plus physique, permet de retrouver 95 % des ponts disulfure natifs d'un jeu de 16 peptides, pour un cRMSd moyen des modèles de 3,8 Å, alors que la formulation classique (sOPEP v2.1.1) n'en trouve que

36 % pour un cRMSd moyen de 4,7 Å. La formulation sOPEP v2.1.3 déformant trop le paysage énergétique, ne sera pas retenue en production.

Nous savons que l'algorithme glouton, guidé par la version actuelle du potentiel sOPEP, a des difficultés pour générer des conformations natives pour des protéines de taille importante, ainsi, de raisonner sur des protéines de petite taille nous a permis d'en limiter l'impact, et de pouvoir apprécier le gain associé à une formulation adaptée aux ponts disulfure. Dans le cadre d'une prédiction de structure, nous pourrions intégrer l'information de connectivité des cystéines dans la méthode globale HMM-SA, comme GDAP (*Genomic Disulfide Analysis Program*) (O'Connor and Yeates, 2004; Mallick et al., 2002) ou une méthode récemment développée par Lu et al. (2007).

9.3 Conclusions

Dans le cadre du développement d'une méthode de prédiction *ab initio*, nous avons intégré une version simplifiée du champ de force OPEP (sOPEP) dans notre méthode de reconstruction de modèles protéiques. Un certain nombre de tests de génération de modèles ont mis en évidence que la formulation du potentiel entre les chaînes latérales était problématique dans une représentation discrète de la structure des protéines. Ainsi, nous avons re-paramétré ce potentiel pour qu'il s'adapte mieux au distributions de distances intercentroïdes observées de chaînes latérales en interaction. L'ensemble de ces paramètres ont ensuite été optimisés similairement au modèle continu.

La re-paramétrisation de ce potentiel a d'ailleurs été l'occasion de développer un potentiel spécifique aux ponts disulfure permettant de trouver 95 % des ponts disulfure natifs d'un ensemble de 16 structures expérimentales.

La seule information des ponts disulfure revient finalement à une carte de contacts partielle, et, nous savons que l'algorithme glouton est capable de générer des topologies natives, même pour des cibles de grande taille, guidé par un critère de type Go (Tuffery and Derreumaux, 2005). Ainsi, il pourrait être intéressant d'intégrer une contribution énergétique basée sur des contraintes de distances entre certains résidus sélectionnés (comme c'est le cas dans un certain nombre d'approches de modélisation par homologie, tel que le programme MODELLER (Sali and Blundell, 1993)) pour guider l'algorithme glouton vers des solutions toujours plus pertinentes.

Dans la suite de ce manuscrit, nous allons vous présenter les premiers tests de simulation de repliement que nous avons effectué avec sOPEP v2.1 couplé à l'algorithme glouton pour le repliement de super-structures secondaires, de peptides et de structures protéiques complètes.

Quatrième partie

Applications de la méthode HMM-SA

Chapitre 10

Essais de modélisation *de novo* de la structure des protéines

La première fois où nous avons commencé à avoir l'ensemble des briques nous permettant de passer de la séquence à la structure 3D, a correspondu à la période de CASP7 (été 2006). Après avoir présenté les résultats que nous avons obtenu lors de cette compétition, nous allons brièvement exposer quelques tests récents concernant la mise en place d'une procédure hiérarchique pour le repliement de protéines.

10.1 Premiers essais : CASP7

Comme nous l'avons déjà évoqué plusieurs fois lors de ce manuscrit, CASP est une expérience internationale permettant de comparer entre elles les différentes méthodes de prédiction de la structure des protéines. Elle a lieu tous les deux ans depuis 1994 (Moult et al., 1997, 1999, 2001, 2003, 2005, 2007).

Depuis l'édition originale, le but ultime de l'expérience est de prédire la structure tertiaire des protéines, à partir de la séquence en acides aminés. Différentes catégories ont ensuite fait leur apparition, comme :

- la prédiction de contacts
- la prédiction des régions désordonnées
- la prédiction de domaines
- la prédiction de fonction
- l'évaluation de la qualité des modèles (depuis CASP7)
- et le raffinement de modèles (depuis CASP7 aussi).

La prédiction de la structure tertiaire peut être découpée en trois catégories selon le pourcentage d'identité de séquence que l'on peut trouver entre la séquence en acides aminés à modéliser et les structures disponibles dans les base de données. A plus fort taux d'identité, les cibles appartiennet à la catégorie *High Accuracy Template Based Modeling*

(HA-TBM), puis, à moins fort taux d'identité de séquence, les organisateurs parlent de *Template Based Modeling* (TBM), et enfin, la dernière catégorie réputée comme étant la plus complexe : *Free Modeling* (FM). Pour chaque cible, chaque participant inscrit peut proposer jusqu'à cinq modèles classé par ordre de pertinence. Cette évaluation est propre au groupe qui soumet ses modèles, le modèle 1 ayant leur préférence.

Dans cette partie, nous allons présenter les résultats que nous avons obtenu par catégorie, pour les cibles pour lesquelles nous avons concouru.

10.1.1 Méthodes

La méthode HMM-SA.

Lors de notre participation à la septième édition de CASP, nous avons pu tester pour la première fois la méthode HMM-SA dans son ensemble, pour des cas réels de prédiction. D'une manière générale, la méthode se découpe en deux étapes que sont (i) la prédiction d'un ensemble de fragments candidats par SAFrAN, puis (ii) l'assemblage de cet ensemble de fragments par l'algorithme glouton guidé par le potentiel gros grain sOPEP.

De SAFrAN vers l'algorithme glouton. Au moment de CASP7, l'algorithme glouton n'étant pas encore prêt à gérer des fragments de taille variable, nous avons décidé de n'utiliser que des lettres et non des mots de l'alphabet structural HMM-SA. Ainsi, chacun des mots (ou fragments) prédits par SAFrAN est systématiquement découpé en lettres. La trajectoire résultante est alors filtrée en terme de transition markovienne pour garantir qu'il n'existe pas d'extrémités borgnes et pour réduire la complexité de la recherche.

Un algorithme glouton moins gourmand (!) Du point de vue des ressources computationnelles, l'algorithme glouton est très gourmand en espace mémoire. Par exemple, une simulation de repliement pour une protéine de 300 résidus peut prendre plus de deux giga-octets de RAM. Ainsi, le facteur limitant dans notre cas, n'est pas le nombre de CPU (*Central Process unit*) à notre disposition, mais le nombre de barettes mémoire. Ainsi, pour pouvoir traiter un grand nombre de simulations, il nous fallait réduire la complexité des trajectoires données en entrée. Nous avons donc décidé de réduire le nombre maximum de prototypes décrivant chaque lettre de l'alphabet structural HMM-SA à trois. Ceci diminue la complexité d'un facteur 2, le nombre de prototypes total étant de 74 au lieu des 155 initiaux. Nous avons eu l'agréable surprise de remarquer que, dans le cadre d'une prédiction, ce gain de temps ne semble pas trop affecter le niveau de description de l'alphabet structural.

sOPEP v1.0. La formulation de sOPEP, pour les interactions de chaînes latérales, était encore à cette époque identique aux versions 3 d'OPEP. Ceci conduisant trop souvent à des collisions entre ces dernières, nous avions introduit une troncature arbitraire de chaque interaction CL-CL à 2 kCal/mol. De la même manière, les interactions de type Van der Waals, entre atomes de la chaîne principale, étaient troncquées à 2 kCal/mol. Par ailleurs, il est à noter, que cette version n'était pas encore optimisée.

Les prémices d'une approche hiérarchique. Pour les cibles de la catégorie FM, nous avons pu constater que de procéder à de courtes simulations sur des régions déterminées, nous permettait d'obtenir des motifs cohérents de super-structures secondaires. Les solutions obtenues pour ces régions peuvent ensuite être imposées dans les simulations de la structure complète. Cette méthodologie, semblant donner de meilleurs résultats qu'une reconstruction complète brute, revient finalement à pousser à l'extrême le principe des préfiltres mis en place dans l'algorithme glouton (Tuffery and Derreumaux, 2005). Cependant, dans ce cas ci, nous imposons un prototype unique pour chaque position de chaque région simulée.
Cette méthode requiert encore une expertise humaine importante pour déterminer (i) quelle région doit on simuler indépendamment ? et (ii) quelles solutions doit on conserver pour générer un modèle complet ?
Lors de CASP7, la réponse à ces questions est restée du domaine de l'intuition. Cependant, la détermination des régions à pré-simuler a été établie par rapport aux prédictions des structures secondaires des cibles par PSIPRED (Jones, 1999b), et des modèles prédits par Robetta (Chivian et al., 2003; Kim et al., 2004; Chivian et al., 2005; Chivian and Baker, 2006).

Post traitement des modèles à soumettre Les modèles générés par l'algorithme glouton étant dépourvus de chaînes latérales, chacun d'entre eux est ensuite reconstruit par SABBAC (Maupetit et al., 2006) (voir la section 7). Ensuite, afin de corriger les quelques erreurs de géométrie pouvant survenir par la méthode, une minimisation courte (1000 pas de minimisation avec la méthode du gradient conjugué) est réalisée par l'intermédiaire du logiciel GROMACS (Spoel et al., 2005; Kutzner et al., 2007).

La modélisation par homologie

Il est un certain nombre de cibles pour lesquelles nous avons utilisé une structure matrice pour générer des modèles. En effet, notre participation à CASP a été l'occasion de mettre en place GreedyHomol, une euristique pour nous appuyer sur des structures connues lorsque l'identité de séquence cible-matrice le permet.

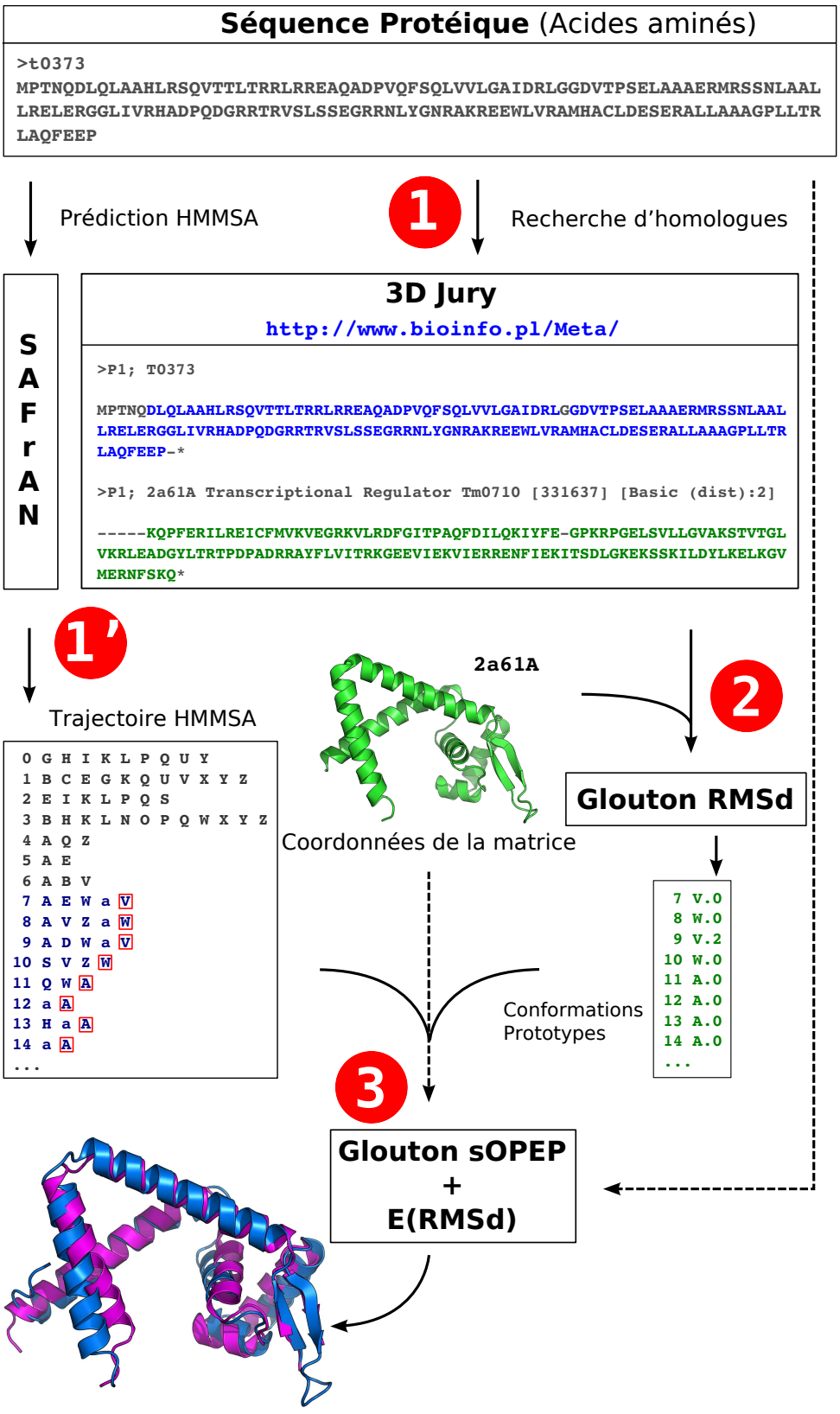


Fig. 10.1: GreedyHomol : La méthode modélisation par homologie. Dans ce schéma, sont présentées les différentes étapes de la méthode de modélisation par homologie que nous avons développé pendant CASP7. L'algorithme glouton s'aide ici des coordonnées et de l'encodage d'une structure matrice pour générer un modèle protéique.

Comme nous l'avons déjà précisé dans l'introduction, les premières étapes d'une méthode de modélisation par homologie sont : (i) identifier des homologues dont la structure est connue, (ii) raffiner l'alignement de séquence entre la cible et la matrice et (iii) générer un modèle protéique à partir de cet alignement et des coordonnées de la matrice, modèle qui, on l'espère, tend à s'éloigner de la matrice pour se rapprocher de la structure de la cible. GreedyHomol suit globalement ces différentes étapes (voir la figure 10.1) :

1. Partant de la séquence en acides aminés, nous recherchons des homologues structuraux par une méthode *ad hoc*, telle que 3D-Jury (Ginalski et al., 2003a), PipeAlign (Plewniak et al., 2003) ou encore une méthode basée sur l'alphabet structural HMM-SA (travaux en cours).
2. De l'étape précédente, nous obtenons un alignement entre les séquences protéiques de la cible et d'une structure matrice sélectionnée. Il nous faut donc préparer les données de la matrice pour pouvoir les introduire dans l'algorithme glouton. Nous devons donc :
   - Calculer la trajectoire optimale de la structure matrice (algorithme Viterbi).
   - Reconstruire la structure matrice par l'algorithme glouton, en mode RMSd, pour en déduire les prototypes de chaque lettre structurale pouvant au mieux la décrire.
   - Extraire les coordonnées de la matrice pour les régions alignées avec la cible et les prototypes qui en découlent (déterminés à l'étape précédente).
3. Des simulations de repliement avec l'algorithme glouton, guidé par le critère énergétique sOPEP, permettent de générer un modèle protéique complet. La trajectoire donnée en entrée est déduite des fragments candidats prédits par SAFrAN, réduite, dans les régions alignées, aux seuls prototypes correspondants de la matrice. Par ailleurs, en plus d'imposer la conformation des régions alignées, nous utilisons les coordonnées correspondantes de la matrice pour le calcul de la contribution énergétique E_{RMSd}, qui nous permet de contraindre l'algorithme à ne pas s'éloigner de la structure matrice.

Comme toute méthode de modèlisation par homologie, l'étape la plus sensible de GreedyHomol est la recherche d'une structure matrice et d'un alignement pertinent entre cette dernière et la cible à modéliser. Cependant, l'avantage de GreedyHomol est qu'il est indépendant de la méthode utilisée pour réaliser cette étape, la seule information dont il a besoin étant un alignement de séquences.

10.1.2 Résultats

Performances de SAFrAN

Afin d'évaluer la pertinence de la prédiction SAFrAN, une fois CASP7 terminée et les structures des cibles rendues publiques, nous avons reconstruit ces dernières par l'algo-

Cat.	Cible	PDB	L	Classe	N	Matrice	Id	G	Gap_{max}	C_p	C_{p3}	PFR	PFR_3	M	TM	cRMSd
HA-TBM	t0288	2gzv	93	a/b	3	1n7eA	23,2	2	3	28,0	15,2	1,06	1,21	2	0,75	4,00
	t0297	2hsj	211	a	2	1es9A	20,9	9	8	17,4	11,0	1,83	1,99	1	0,69	7,80
	t0302	2h33	132	a	2	1zv4X	42,4	4	7	11,1	7,7	2,16	1,96	1	0,74	3,15
	t0305	2h4v	297	a/b	1	1fprA	40,7	16	9	20,9	11,5	2,04	3,07	1	0,70	21,89
	t0308	2h57	165	a/b	2	1mr3F	40,6	4	2	25,3	14,5	1,39	1,81	1	0,79	4,20
	t0315	2gzx	257	a/b	1	1j6oA	38,5	3	3	18,6	11,4	1,26	1,67	1	0,70	5,32
	t0317	2hcm	163	a	2	1m3gA	27,0	2	11	23,9	13,6	1,08	1,65	1	0,60	7,91
	t0322	2hbo	157	a/b	1	1vh5A	11,5	6	15	27,1	15,5	1,29	1,74	1	0,57	7,26
	t0373	2hr3	147	a	5	2a61A	17,7	3	5	10,9	7,3	2,50	2,53	3	0,71	3,67
TBM	t0283	2hh6	112	a	4	1kgnA	6,2	3	3	15,5	10,7	1,22	1,45	3	0,31	17,21
	t0369	2hkv	148	a	1	2f22A	14,2	4	6	22,2	13,4	1,47	1,92	1	0,62	8,52
FM	t0300	2h3r	102	a	5	-	-	-	-	17,8	11,0	0,63	0,94	5	0,35	15,61
	t0353	2hfq	85	a/b	2	-	-	-	-	27,5	15,6	0,59	0,90	1	0,41	11,10
	t0358	2hjj	87	a/b	4	-	-	-	-	27,0	14,5	0,68	0,92	3	0,27	10,52
	t0383	2hng	127	a/b	5	-	-	-	-	22,9	13,0	1,62	1,74	2	0,42	10,42

Tab. 10.1: Résultats obtenus pour les cibles CASP7 concourues. Pour chaque cible, pour laquelle nous avons concouru à CASP7 (Cible), sont présentés l'identifiant PDB de la structure expérimentale (PDB), sa longueur (L) en acides aminés, sa classe (Classe), sa catégorie CASP (Cat.), le nombre de modèles soumis (N), la matrice utilisée pour les cibles de modélisation par homologie (Matrice), ainsi que le pourcentage d'identité de séquence entre la cible et la matrice (Id), le nombre de *gaps* dans l'aligmement (G), et la taille en acides aminés du plus long *gap*, la complexité de la trajectoire prédite en considérant tous les prototypes (C_p) ou seulement trois prototypes par lettre (C_{p3}), le plus faible cRMSd que l'on peut obtenir à partir de cette trajectoire en utilisant tous les prototypes (PFR), ou seulement trois prototypes par lettre (PFR_3), et enfin, le classement du meilleur modèle parmi les modèles soumis (M), son TM-score (TM) et son cRMSd.

rithme glouton, guidé par le critère cRMSd, en partant des trajectoires prédites. Cette procédure nous permet d'estimer les meilleurs modèles que nous pouvions obtenir à partir de la prédiction SAFrAN. Nous pouvons constater que pour les quinze cibles modélisées, la trajectoire prédite permet de conduire à des solutions PSN, le cRMSd moyen étant de 1,4 Å, avec des valeurs s'échelonnant entre 0,59 et 2,5 Å, lorsque l'on considère les 155 prototypes HMM-SA (voir la table 10.1). Ceci pour une complexité moyenne raisonnable de 21,1 pour des protéines dont la longueur moyenne est de 152 acides aminés (Tuffery et al., 2005). Si l'on ne considère maintenant que trois prototypes maximum par lettre HMM-SA, la complexité moyenne est alors de 12,4. Avec cette limitation du nombre de prototypes, le cRMSd moyen des reconstructions des structures expérimentales à partir des trajectoires prédites selon un critère RMSd est, cette fois, de 1.7 Å avec des valeurs comprises entre 0,9 et 3,07 Å. Alors que le gain en terme de complexité (et donc de temps de calcul), est de l'ordre d'un facteur 2, la diminution de la performance de description n'est que de l'ordre de 20%. Depuis CASP7, nous avons donc gardé cette limitation du nombre de prototypes utilisés dans l'algorithme glouton.

Nous pouvons donc conclure que globalement, les prédictions SAFrAN semblent pertinentes, puisqu'elles peuvent conduire l'algorithme glouton à un ensemble de structures PSN, lorsque celui-ci est guidé par un critère approprié.

Cibles *High Accuracy Modelling* (HA-TBM)

La catégorie HA-TBM a été l'occasion de tester GreedyHomol pour des pourcentages d'identité de séquence allant de 11,5 à 42,4%. Des représentations des modèles que nous avons obtenus, superposés aux structures expérimentales, sont présentées dans la figure 10.2.
Globalement, si l'on se réfère aux TM-scores (table 10.1) obtenus pour chacun de ces modèles, comparés à la structure expérimentale, tous ont une topologie native, le TM-score moyen étant de 0,69. Si l'on change de référentiel, et que l'on analyse les cRMSds des modèles prédits en comparaison toujours avec la structure native, les valeurs semblent plus dispersées : le cRMSd moyen est de 7,24 Å si l'on considère les 9 modèles, pour des valeurs comprises entre 3,15 et 21,89 Å, et passe à 5,41 Å si l'on ignore la valeur extrême de 21,89 Å, obtenue pour la cible t0305. Au final, 55% des modèles générés ont un cRMSd inférieur à 5,3 Å par rapport à leur structure expérimentale.

Il est intéressant de constater qu'il n'existe pas de relation directe entre la qualité des modèles obtenus et le pourcentage d'identité de séquence puisque, par exemple, pour la cible t0373, qui partage 17,7 % d'identité de séquence avec la matrice 2a61A, le modèle produit partage un cRMSd de 3,67 Å avec la structure expérimentale de la cible, tandis que, pour la cible t0305, qui possède un des plus forts taux d'identité de séquence (40,7

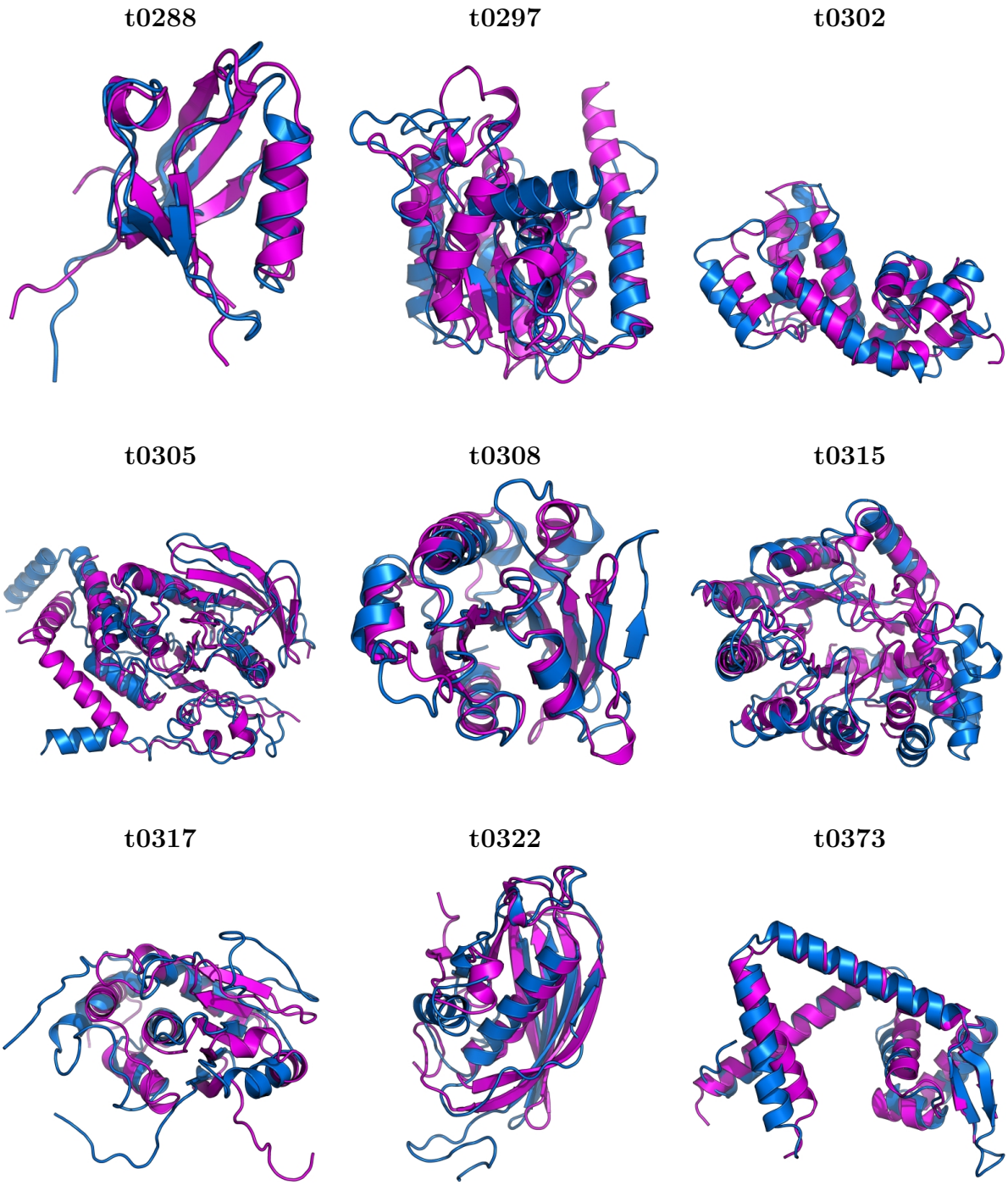


Fig. 10.2: Meilleurs modèles obtenus pour les cibles HA-TBM. Les meilleurs modèles obtenus pour la catégorie HA-TBM sont ici présentés en bleu, superposés à la structure expérimentale en magenta.

%), le cRMSd du modèle GreedyHomol est de 21,89 Å. Il semblerait plutôt que la qualité des modèles produits soit en rapport avec la complexité de l'alignement entre la cible et la matrice (voir les alignements en annexe, dans la section B.1). Ainsi, pour la cible t0305, le nombre important de gaps (16) dans l'alignement a fait totalement diverger le modèle par rapport à la structure native. Cependant le terme énergétique E_{RMSd} a forcé le modèle à se superposer sur les coordonnées de la matrice données en entrée, ce qui explique le fort TM-score obtenu (0,70) ; score qui est basé principalement sur des contraintes de distances géométriques. Dans une moindre mesure, le problème est similaire pour la cible t0297, alors que pour les cibles t0317 et t0322, le coeur de repliement est natif, les gaps dans l'alignement étant situés aux extrémités N et C terminales. Les modèles obtenus pour ces cibles problématiques, nous indiquent clairement, que la formulation du terme énergétique E_{RMSd} est problématique. Cependant, il s'avère qu'il soit nécessaire, car, lorsqu'il est inactivé, les topologies générées étaient bien souvent disperses.

Avec des alignements peu complexes, les cibles t0288, t0302, t0308, t0315 et t0373 sont très proches de la structure native, pour des identités de séquence allant de 17,7 % à 42,4 %. De prime abord, ces résultats peuvent sembler satisfaisants, cependant, il ne sont pas assez performants si on les compare aux principales méthodes de modélisation comparative (voir les classements des modèles parmi l'ensemble des modèles CASP soumis pour ces cibles, dans la table 10.2). D'un autre côté, l'approche a été implémentée et testée en même temps, et nécessite encore quelques efforts pour devenir compétitive. On peut se demander si une méthode gros grain telle que la notre est appropriée pour ce genre de cas de modélisation.

Cibles *Template Based Modelling* (TBM)

Les cibles TBM ayant aussi été réalisées avec GreedyHomol, elles souffrent des mêmes défauts que les cibles HA-TBM, en plus d'un taux d'identité de séquence moyen faible par rapport à ces dernières.

La cible 283 est un faisceau de cinq hélices alpha, les hélices 1 et 2 ont correctement été repliées dans le modèle, mais une inversion de topologie s'opère entre les hélices 3 et 4 du modèle conduisant à une topologie non native. Le choix de la matrice est à remettre en cause, sa topologie étant trop éloignée de la structure expérimentale à prédire. Notre méthode de modélisation par homologie nous permet difficilement de sortir de la structure matrice.

La cible 369 est aussi un faisceau de cinq hélices alpha, les hélices 5, 4 et 2 ont été correctement formées dans le modèle, alors que l'hélice 3 y est inexistante, et l'hélice 1 a été rompue. L'hélice 3, qui n'existe pas non plus dans la matrice, n'a pa été trouvée par

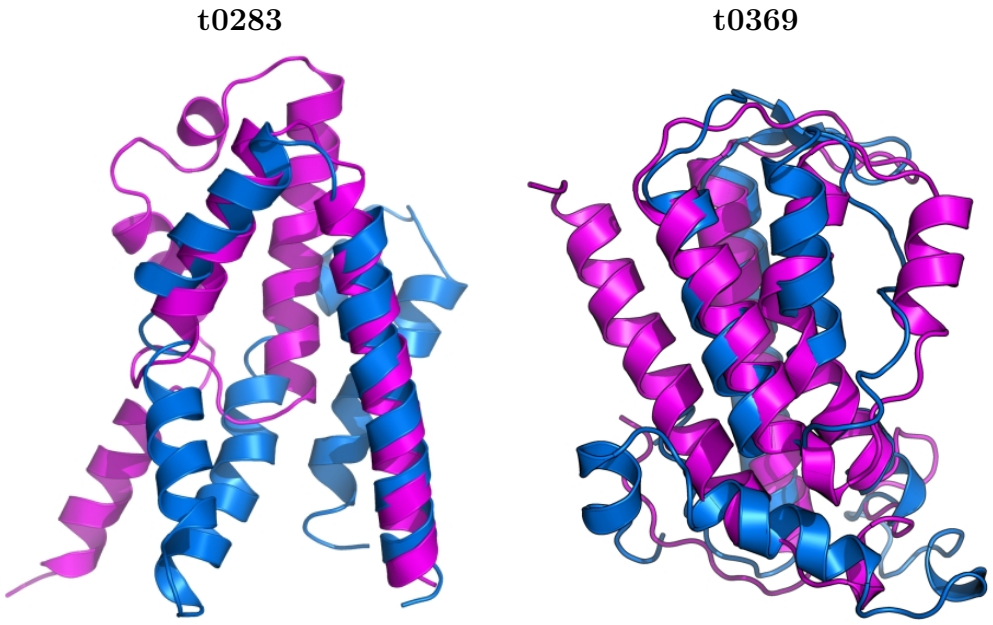


Fig. 10.3: Meilleurs modèles obtenus pour les cibles TBM. Les meilleurs modèles obtenus pour la catégories TBM sont ici présentés en bleu, superposés à structure expérimentale en magenta.

l'algorithme glouton, et l'hélice 1 pourtant présente a été rompue probablement à cause du long gap présent dans cette zone (voir l'alignement en annexe B.1).

La méthode de génération des modèles étant à faible résolution, il semblerait qu'elle puisse être plus compétitive dans la catégorie TBM que HA-TBM. Nous n'avons cependant pu la tester que sur deux cibles pour lesquelles les résultats n'ont pas été encore convaincants.

Cibles *Free Modelling* (FM)

Les cibles FM ont été les premiers réels tests de prédiction *ab initio* de la méthode HMM-SA dans son ensemble. Ce sont les modèles qui nous ont le plus surpris en terme de performance. L'approche hiérarchique a été mise en place lors de la résolution de ces modèles, pour lesquels nous avons réalisé des simulations sur des régions données, dont nous avons "assemblé" les résultats pour former des modèles complets.
Du point de vue du cRMSd, chacun de ces modèles est loin de la structure native, tous partageant un cRMSd supérieur à 10 Å avec cette dernière. Cependant, une analyse plus fine des topologies nous montre que les modèles sont plus pertinents qu'il n'y parait.

Pour la cible t0300, l'hélice *alpha* 1 est à l'opposé de sa position dans la structure native, et l'hélice 2, trop longue n'a pas été rompue pour se replier sur elle-même, mais

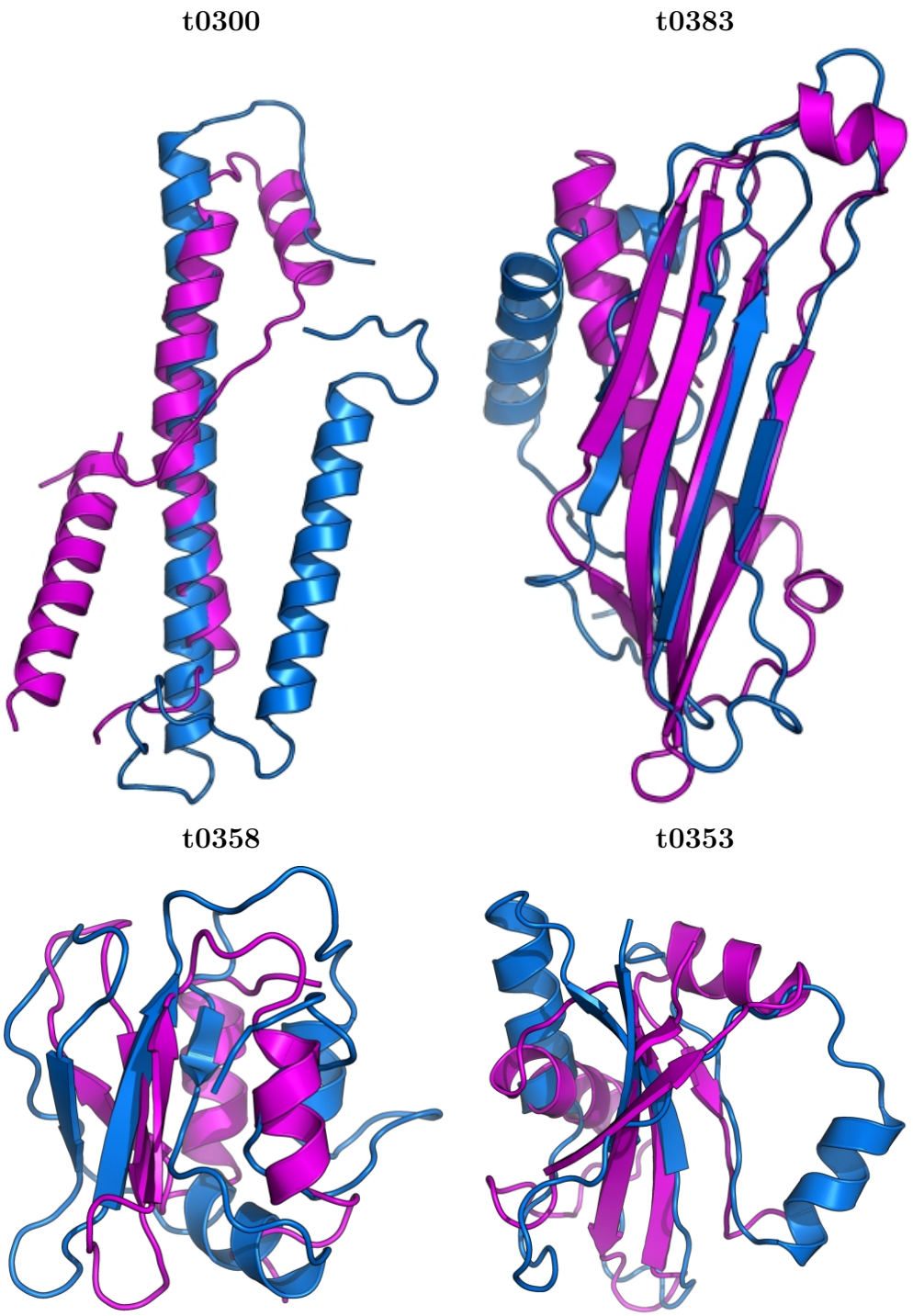


Fig. 10.4: Meilleurs modèles obtenus pour les cibles FM. Les meilleurs modèles obtenus pour la catégorie FM sont ici présentés en bleu, superposés à la structure expérimentale en magenta.

cependant, la topologie globale est satisfaisante (voir la figure 10.4).

Comme peut en attester le TM-score obtenu pour la cible t0353, la topologie du modèle obtenu est très proche de la structure native : l'hélice N terminale et le feuillet à

Cat.	Cible	N	E.	B.	R*	T.	Z.	Z*	J.	L.
HA-TBM	t0288	144	127	3	71	6	18	13	23	12
	t0297	129	119	14	50	2	6	18	40	1
	t0302	140	130	7	94	48	5	4	121	3
	t0305	129	126	25	49	21	4	1	39	14
	t0308	133	119	12	68	41	14	38	111	58
	t0315	128	123	22	50	16	19	20	38	1
	t0317	136	129	11	28	22	7	8	6	38
	t0322	137	125	51	25	13	7	1	27	84
	t0373	137	98	2	45	10	20	26	88	16
TBM	t0283	135	68	1	19	4	7	9	2	20
	t0369	136	93	3	29	5	1	6	35	45
FM	t0300	134	67	2	18	42	10	4	5	75
	t0353	145	65	19	8	22	1	2	17	46
	t0358	144	99	70	20	40	6	2	21	8
	t0383	145	76	43	45	9	13	18	17	28

Tab. 10.2: Classement CASP7 par cibles. Pour chaque cible CASP7 pour laquelle nous avons soumis des modèles, sont présentés, le nombre de groupes qui ont soumis des modèles (N), et le classement du meilleur modèle soumis par un ensemble sélectionné de groupes : EBGM (E.), Baker (B.), *Robetta server* (R*), TASSER (T.), Zhang (Z.), *Zhang server* (Z*), Jones-UCL (J.) et Lee (L.), selon le critère GDT_TS (Venclovas et al., 2003).

quatre brins sont quasi complètement formés (le réseau de liaisons hydrogène est incomplet), seule l'orientation de l'hélice C terminale a été inversée.

Le modèle généré pour la cible t0358 a une topologie plus complexe que la structure expérimentale avec un feuillet béta à quatre brins au lieu de trois, et les hélices sont mal placées, car nous n'avons pas trouvé les interactions hydrophobes s'établissant entre les résidus des brins et des hélices.

En ce qui concerne la cible t0383, l'hélice C-terminale n'a pas été correctement positionnée, mais le feuillet à quatre brins d'une topologie complexe 1-4-3-2 a été trouvée par l'algorithme glouton. A titre de comparaison, aucun des modèles générés par Robetta n'a obtenu les quatre brins β correctement positionnés.

10.1.3 Conclusions de l'étude

Si l'on se réfère au score GDT_TS (Venclovas et al., 2003), utilisé comme score de référence pour évaluer les modèles soumis à CASP (Vincent et al., 2005), nous pouvons classer les groupes ayant participé pour chaque cible. La table 10.2 présente le classement que nous avons obtenu pour chaque cible, ainsi que le classement des méthodes recon-

nues comme étant les plus performantes, telles que Rosetta (Rohl et al., 2004; Simons et al., 1997, 1999), du groupe de David Baker, et son serveur associé Robetta, TASSER (Zhang et al., 2005; Zhou and Skolnick, 2007; Lee and Skolnick, 2007; Borreguero and Skolnick, 2007) du groupe de Jeffrey Skolnick, I-TASSER (Wu et al., 2007) du groupe de Yang Zhang, et le serveur associé (http ://zhang.bioinformatics.ku.edu/I-TASSER/), GenTHREADER (Jones, 1999a; McGuffin and Jones, 2003) du groupe de David T. Jones, et enfin, le groupe de Joojung Lee (Lee et al., 2005; Kim et al., 2005).

En ce qui concerne les cibles HA-TBM, GreedyHomol est loin d'être compétitif ; un certain nombre d'améliorations sont encore à apporter à la méthode pour y introduire plus de fléxibilité. Néanmoins, nous n'avons pas encore testé la nouvelle implémentation du champ de force sOPEP 2.1 conjointement à cette dernière. Une fois ce problème de la topologie de la protéine résolu, nous envisageons d'utiliser des librairies de conformations de boucles disponibles, telle WLoop (Kwasigroch et al., 1996; Wojcik et al., 1999), encodées dans l'espace de l'alphabet structural HMM-SA, pour explorer l'espace conformationnel.

Les modèles proposés pour les cibles FM sont globalement situés dans la première moitié de l'ensemble des modèles, et sont compétitifs lorsqu'on les compare aux méthodes les plus performantes. La méthodologie employée à CASP semble donc être prometteuse. Ainsi, les leçons que nous avons tirées de CASP sont multiples : (i) les prédictions SA-FrAN semblent suffisamment performantes pour obtenir une solution native-proche, (ii) nous pouvons diminuer la complexité de la recherche en réduisant le nombre de prototypes utilisés pour chacune des lettres HMM-SA, (iii) employer une approche hiérarchique pour replier des protéines semble conduire l'algorithme glouton vers des solutions plus pertinentes, et enfin (iv) l'implémentation du critère énergétique sOPEP dans l'algorithme glouton nécessitait encore quelques améliorations.

10.2 Vers une approche hiérarchique

Cette approche hiérarchique suggérée lors de l'expérience CASP7 s'inscrit dans la lignée d'un certain nombre d'autres approches que nous allons présenter.

Au delà du concept d'alphabet structural, si l'on se place à un niveau supérieur de l'architecture protéique, apparaît la notion de *foldon*, *i.e.* des unités protéiques à repliement autonome. Ces foldons peuvent adopter leur structure tertiaire native en l'absence du reste de la chaîne protéique (Haspel et al., 2003a). Associés à un certain nombre de résidus clés du repliement protéique, ces foldons sont des motifs récurrents au sein des structures protéiques. De ce fait, suivant un des modèles du repliement protéique (Fersht and Daggett, 2002), nous considérons qu'une protéine adopte sa conformation native selon

un processus hiérarchique : dans un premier temps, des noyaux de repliement se forment (les *foldons*), puis s'assemblent pour former une structure tertiaire complète.

Basées sur ce principe, à différents niveaux d'organisation de la structure des protéines, sont apparues les notions de *Tight End Fragments*, *Folding Units*, *Building Blocks* et *Protein Chunks*. Nous allons présenter brièvement ces travaux pionniers, initiateurs d'une approche hiérarchique dans le cadre de la prédiction de la structure des protéines.

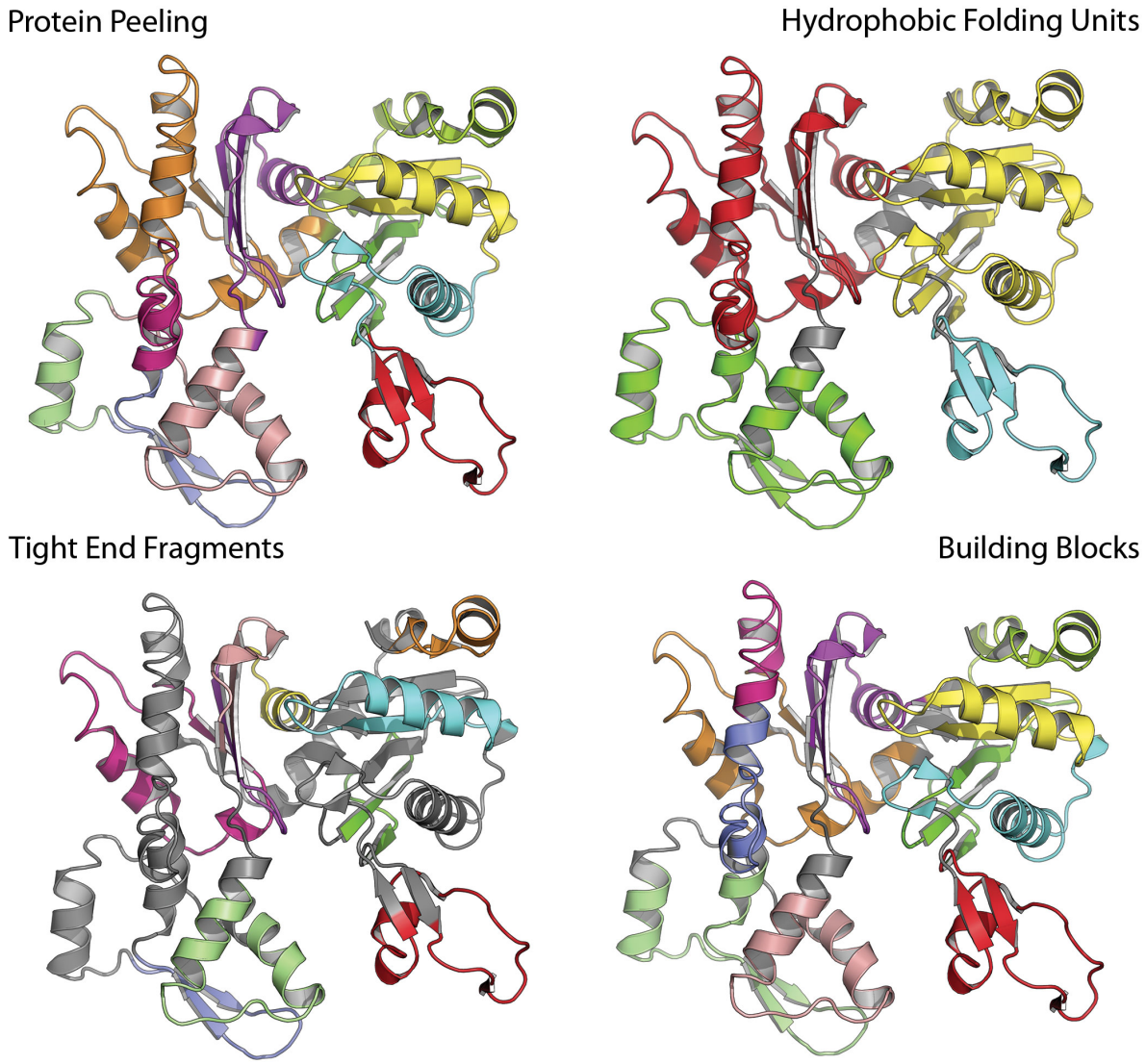


Fig. 10.5: L'architecture protéique à un niveau mésoscopique. L'architecture des protéines peut être vue à un niveau mésoscopique, comme un assemblage d'unités à repliement autonome. Différentes visions de la même protéine (code PDB : `1atnA`), associées à différentes méthodes, sont ici présentées. Chaque couleur représente une unité.

Trifonov et al ont proposé que la structure d'une protéine globulaire peut être assimi-

lée à un ensemble de *closed-loops units* (Berezovsky et al., 2000; Berezovsky and Trifonov, 2001b; Berezovskyi and Trifono, 2001) appellées TEF (*Tight End Fragments*). Cette vision des boucles est très différente de la définition communément admise de regions connectant des structures secondaires ; dans ce cas précis, une *closed-loop* est définie comme étant un segment de la chaîne polypeptidique qui émerge et replonge dans le coeur de la structure protéique avec ses extrémités proches. L'analyse de la répartition des *closed loops* au sein des séquences en acides aminés de différentes familles protéiques a montré que leur longueur typique est de 30 acides aminés, et qu'elles sont un constituant majeur des structures protéiques (entre 40 et 85 %). Les TEF ne sont pas systématiquement des super-structures secondaires, certaines d'entre elles se terminant au milieu d'un feuillet β ou d'une hélice α (voir la figure 10.5). Les extrémités des TEF sont des régions clés de la structure protéique : (i) elles forment des sites de multiples contacts (Berezovsky and Trifonov, 2001b) entre les extrémités N et C terminales de différentes boucles, et (ii) elles sont préferentiellement constituées de topohydrophobes (Lamarine et al., 2001; Berezovsky et al., 2001), *i.e.* des acides aminés très hydrophobes situés à des positions clés chez tous les membres d'une famille partageant le même repliement (Poupon and Mornon, 1998, 1999). Les résidus des extrémités interagissant peuvent être considérés comme des verrous de Van der Waals pour la structure (Berezovsky and Trifonov, 2001a), maintenant le coeur de la protéine en maintenant proches les TEF. Il est possible de prédire la localisation des TEF en prédisant la localisation des positions topohydrophobes (Chomilier et al., 2004; Papandreou et al., 2004).

Beaucoup d'efforts ont été entrepris ces trentes dernières années pour diviser les protéines en domaines ou en *folding units* (FU) (Go, 1981; Lesk and Rose, 1981; Wodak and Janin, 1981; Sowdhamini and Blundell, 1995). Tsai and Nussinov (1997a,b) ont défini les FU par leur structure compacte, un coeur hydrophobe, leur stablité thermodynamique et leur repliement indépendant. Les FU diffèrent des domaines protéiques de par le fait qu'elles peuvent avoir beaucoup de contacts entre elles au sein de la structure native (principalement des interactions hydrophiles). Ainsi, un domaine structural peut être constitué de plusieurs *Hydrophibic Folding Units* ou HFU (voir la figure 10.5). L'idée sous-jacente est bien qu'il exite une hiérarchie dans le processus de repliement. Ainsi, si l'on inverse le processus de repliement, nous pouvons assimiler les chaînes protéiques à un assemblage de domaines, eux mêmes constitués de FU, que l'on peut séparer en *Building Blocks* (BB), que nous décrirons dans le paragraphe suivant. Contrairement aux études plus anciennes qui ne se focalisaient que sur le critère de compacité des segments protéiques pour identifier des FU, Tsai et Nussinov ont introduit un score plus complet basé sur les propriétés physiques des protéines pour évaluer la pertinence des fragments candidats. Leur fonction de score est composée de quatre termes que sont la compacité (dont la définition est inspirée de Zehfus and Rose (1986)), l'isolement, l'hydrophobicité et la segmentation.

Plus récemment, Gelly et al. (2006a,b) ont proposé une approche originale : le *Protein Peeling*. Cette méthode, similaire à une procédure de classification hiérarchique, découpe itérativement la séquence protéique en unités protéiques. La partition qui maximise l'indépendance des sous-unités en terme de contacts est conservée pour l'itération suivante. Cette indépendance est évaluée par l'indice de partition, qui lui même est dérivé du coefficient de corrélation de Matthews (Matthews, 1976). Cette méthode donne des résultats similaires aux autres approches existantes, et serait confortée par résultats expérimentaux ayant permis d'évaluer les limites des noyaux de repliement.

Dans tous les cas, le but de ces méthodes est d'identifier les sites possibles de nucléation durant le processus de repliement. Ainsi, certaines FU seraient de bonnes candidates pour décrire l'état de globule mou (*molten globule*), ou peuvent être associées à une fonction biologique.

Descendant encore d'un niveau dans l'organisation de la structure protéique, toujours en respectant la conception de repliement hiérarchique, Tsai et al. (1999, 2000) ont défini la notion de *Building Blocks* (BB). Les BB sont des fragments protéiques continus fréquents au sein des structures. Ils sont composés d'une ou de plusieurs structures secondaires, pouvant être des super-structures secondaires. Ils héritent des propriétés des HFU et de leur fonction de score, excepté pour ce qui est de la segmentation. Les structures protéiques sont découpées itérativement jusqu'à ce qu'aucun fragment de taille inférieure ne puisse être trouvé, ces fragments sont les BB. A chaque itération, tous les fragments d'une certaine taille sont évalués avec la fonction de score précédemment décrite, et les fragments présentant le meilleur score sont conservés pour la prochaine itération. A la fin de la procédure un arbre peut être construit, arbre qui transcrit un probable chemin de repliement de la protéine, partant des noyaux de repliement (BB) pour aller jusqu'à la structure protéique complète, en passant par les HFU. La validité du modèle de Tsai et al. a été mis à l'épreuve en comparant les BB identifiés pour un certain nombre de protéines et les fragments identifiés par des expériences de protéolyse ménagée (Tsai et al., 2002). La bonne adéquation des deux méthodes suggère qu'elles pourraient toutes les deux être utilisées pour étudier l'initiation du repliement protéique. Par ailleurs, la classification des BB a montré que des protéines de la même famille conservent généralement les mêmes BB (Haspel et al., 2003a). Cependant, des protéines de différentes familles peuvent aussi avoir des BB en commun, mais c'est la combinaison des BB qui engendre la spécificité. Il a par ailleurs été montré que certains BB sont essentiels pour la structure protéique et agissent comme des fragments chaperons intramoléculaires conservés durant le temps de vie de la protéine. Ces BB particuliers n'ont pas de longueur, de structures secondaires ou de composition en acides aminés spécifiques. Nous savons qu'ils se situent préférentiellement à l'extrémité C-terminale, qu'ils présentent une proportion importante d'acides aminés

hydrophobes, et une surface d'interaction non négligeable avec les autres BB.
La relation avec les positions topohydrophobes serait à analyser pour mieux caractériser leur relation avec les TEF. De la même manière que les TEF, l'analyse des BB peut être un bon point de départ pour prédire la structure des protéines. Dans ce but, Haspel et al. (2003b) ont développé une méthode de prédiction et d'assemblage des BB partant de la séquence en acides aminés de la protéine et d'un ensemble de classes de BB.

Enfin, une autre approche suivant ce concept hiérarchique a été développée par Zhou and Skolnick (2007). Cette méthode dérivée de TASSER (Zhang et al., 2005; Zhou and Skolnick, 2007; Lee and Skolnick, 2007; Borreguero and Skolnick, 2007) se base sur la notion de *Protein Chunks* (PCs). Les PCs peuvent être assimilés à des super-structures secondaires : ils sont constitués de trois structures secondaires régulières consécutives (hélice α ou feuillet β) incluant les boucles mettant en connexion ces structures secondaires. Dans cette méthode hybride de TASSER, les PCs sont ajoutés à l'ensemble des fragments séléctionnés par la méthode SP^3 (Zhou and Zhou, 2005) pour servir de matrice lors de la génération des modèles. Les PCs sont reconstruits par une méthode d'insertion de fragments similaire à notre algorithme glouton ou Rosetta (Simons et al., 1997). Cette procédure hiérarchique semble avoir donné de meilleurs résultats lors de CASP7 que la méthode TASSER seule (Zhou et al., 2007), et permet de rendre accessible des cibles de grande taille.

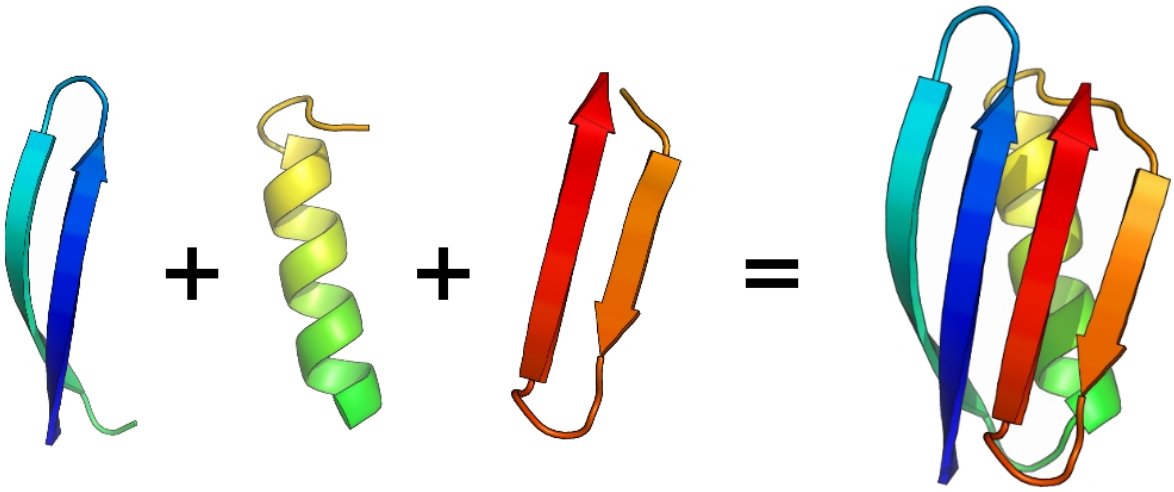


Fig. 10.6: Principe de l'approche hiérarchique. Dans le cadre d'une approche hiérarchique, les structures sont découpées en sous-ensembles reconstruits séparément puis assemblés pour former une protéine complète.

L'ensemble de ces méthodes met en avant le potentiel d'une méthode hierarchique. Tout en respectant ce modèle, il est possible de simuler le repliement des super-structures

secondaires, puis de les assembler pour former une structure protéique complète (voir la figure 10.6). C'est une hypothèse que nous avons testé pour la première fois lors de CASP7, et qui nous a permis d'obtenir des modèles cohérants pour des cibles de tailles comprises entre 85 et 127 résidus. Par ailleurs, comme nous avons pu le voir dans l'ensemble de ces méthodes, il reste à définir quel niveau de découpage serait le plus approprié ?

Dans la section suivante, nous allons présenter les premiers tests de prédiction que nous avons réalisé pour évaluer le potentiel d'une approche hiérarchique. Deux types de tests succints ont été envisagés : dans un premier temps, nous avons rejoué une cible CASP7, puis, nous avons testé l'approche sur un ensemble de nouvelles entrées de la PDB n'ayant pas été apprises par SAFrAN.

10.2.1 Matériels et méthodes

CASP7 *reloaded*

Afin d'évaluer l'approche hiérarchique dans un cas réel, nous avons voulu rejouer CASP7 pour une cible : la cible t0358. Par ailleurs, cela nous a aussi permis d'évaluer les progrès réalisés depuis.

```
  |               |       |               |         |                    |
C 85146566653311000043128864799999908983898541668980025678887508453 3
P CCCCHHHHHHHHHHEEEEEECCCCEEEEEEECCCCCEEEEEECCCCCHHHHHHHHHHHCCCCC
A GPFTRRQAQAVTTTYSNITLEDDQGSHFRLVVRDTEGRMVWRAWNFEPDAGEGLNRYIRTSGIRTD
S     HHHHHHHHHHTTEEEE  TTTTTEEEEETTTT EEEEETTTTTTHHHHHHHHHTTTT
```

Fig. 10.7: Découpage de la cible t0358. La cible t0358 a été découpée selon les prédiction PSIPRED de CASP7, en trois régions dont les extrémités sont marquées par trois couleurs distinctes : rouge (1-23), bleu (16-47) et vert (38-66). Les trois premières lignes représentent la prédiction PSIPRED contenant le degré de confiance dans la prediction (C), la prédiction de structure secondaire dans son code à 3 lettres (P) et la séquence en acides aminés (A). La dernière ligne (S) correspond à l'assignation des structures secondaires par le logiciel STRIDE (Frishman and Argos, 1995).

La cible t0358 a été découpée en trois parties chevauchantes composées de 2 ou 3 structures secondaires prédites par PSIPRED lors de CASP7 (voir la figure 10.7) : régions 1-23 ($\alpha - \beta$), 16-47 ($\beta - \beta - \beta$) et 38-66 ($\beta - \alpha$). Nous n'avons considéré que le segment de la cible résolu expérimentalement, des résidus 10 à 75.

Pour chacune des régions identifiées, nous avons procédé 30 simulations indépendantes. Les fragments ainsi repliés servent ensuite de fragments d'entrée pour de nouvelles simu-

lations de repliement. Afin d'accroître le nombre de chemins possibles, les fragments sont itérativement tronqués dans les zones de chevauchement puis ajoutés à l'ensemble de fragments. Les lettres prédites par SAFrAN dans les zones de chevauchements sont aussi ajoutées pour rendre les transitions plus souples.

En parallèle, nous avons aussi effectué des simulations de la structure complète pour évaluer la perte ou le gain associé à l'approche hiérarchique.

Jeu de nouvelles entrées PDB

Nous avons aussi initié des tests de l'approche hiérarchique sur un ensemble de 11 nouvelles entrées PDB appartenant au JV de SAFrAN. Les protéines selectionnées ont une taille comprise entre 63 et 121 résidus. Cinq protéines sont principalement β (2cwrA, 2extA, 2g3rA, 2iteA et 2j73A), trois sont principalement α (2ictA, 2o35A et 2o38A) et trois sont de type α/β (2obpA, 2i4aA et 2iojA).

Chacune de ces protéines a été découpée en fragments comprenant chacun trois structures secondaires (définies à l'aide du logiciel STRIDE) avec une structure secondaire chevauchante commune à deux fragment adjacents. La taille moyenne des fragments résultants est de 33 résidus.

Pour chaque fragment, nous avons procédé à 30 simulations de repliement gloutonsOPEP v2.1 avec l'opérateur zip.

10.2.2 Résultats

La cible t0358

Approche globale. Les simulations de repliement de la cible t0358 dans son ensemble nous ont permis d'obtenir un modèle éloigné de 5,6 Å de la structure native pour un TM-score de 0,46. A titre de comparaison, si l'on considère tous les modèles CASP soumis pour cette cible, classés en terme de cRMSd, notre modèle serait classé 37$^{\text{ème}}$ sur 592 modèles, soit entre iTASSER (9) et le groupe de Joojung Lee (17) d'un côté et chunck-TASSER (64), Robetta (116) et Rosetta (158) de l'autre. Ce résultat est sans comparaison avec les modèles que nous avions soumis lors de CASP7, le meilleur modèle étant éloigné de 10 Å de la structure native.

Approche hiérarchique. Pour chaque région de la cible t0358, nous avons obtenu des fragments pertinents, éloignés d'enviton 4 Å de la structure native (respectivement 4,2 - 4,2 et 3,4 Å). Le feuillet β central a été le plus difficile à replier : sur 30 simulations, nous n'avons obtenu une topologie correcte que 2 fois. Les simulations de repliement de la cible dans sa globalité, à partir de ces fragments, n'ont pas conduit à un modèle de topologie native pour cette cible. La raison de cet echec doit se situer au niveau des

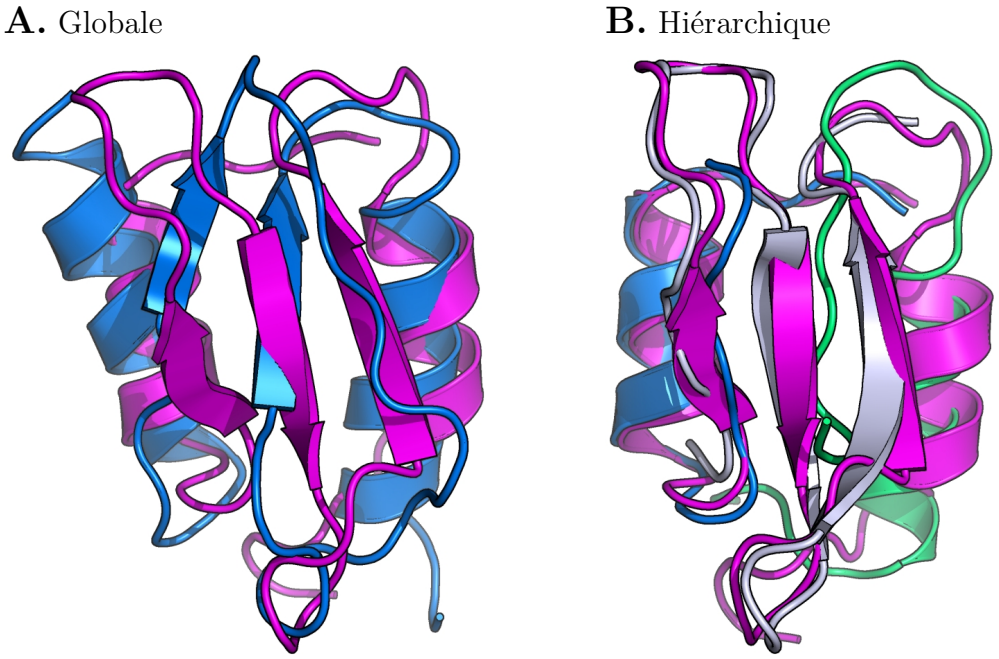


Fig. 10.8: CASP7 reloaded, la cible t0358. La structure expérimentale de la cible t0358 (Code PDB : 2hjj) est ici représentée en magenta. Dans **A**, le meilleur modèle obtenu par l'approche globale, *i.e.* non hiérarchique, est en bleu. Dans le cas de l'approche hiérarchique (**B**), les meilleurs fragments repliés sont ici superposés à la structure native : 1-23 ($\alpha - \beta$) en bleu, 16-47 ($\beta - \beta - \beta$) en blanc et 38-66 ($\beta - \alpha$) en vert.

zones de transition entre les fragments. Une procédure d'assemblage moins rigide est ici à envisager.

Les nouvelles entrées PDB

Les résultats obtenus pour les fragments issus de protéines récentes de la PDB sont résumés dans la table 10.3. Le cRMSd moyen des fragments générés est de 5.4 Å pour une taille moyenne des fragments de 33 résidus. Il est des fragments pour lesquels, l'algorithme glouton couplé à sOPEP v2.1 a réussi à trouver le coeur de repliement natif. C'est le cas de la cible 2obpA (voir la figure 10.9) pour laquel le faisceau de trois hélices est bien formé. Cependant, dans les cas d'erreur de prédiction, il semblerait que ce soit la procédure de découpage qui soit à remettre en question. Prenons l'exemple du premier fragment de la cible 2extA (voir la figure 10.9). Dans la topologie de cette protéine, les brins passent alternativement d'un feuillet à l'autre, cependant comme nous avons replié les trois premiers brins indépendament du reste de la protéine, des derniers ont eu tendance à former un feuillet à trois brins isolé. Il apparait donc que le nombre de structures secondaires consécutives à replier indépendamment soit un paramètre à optimiser.

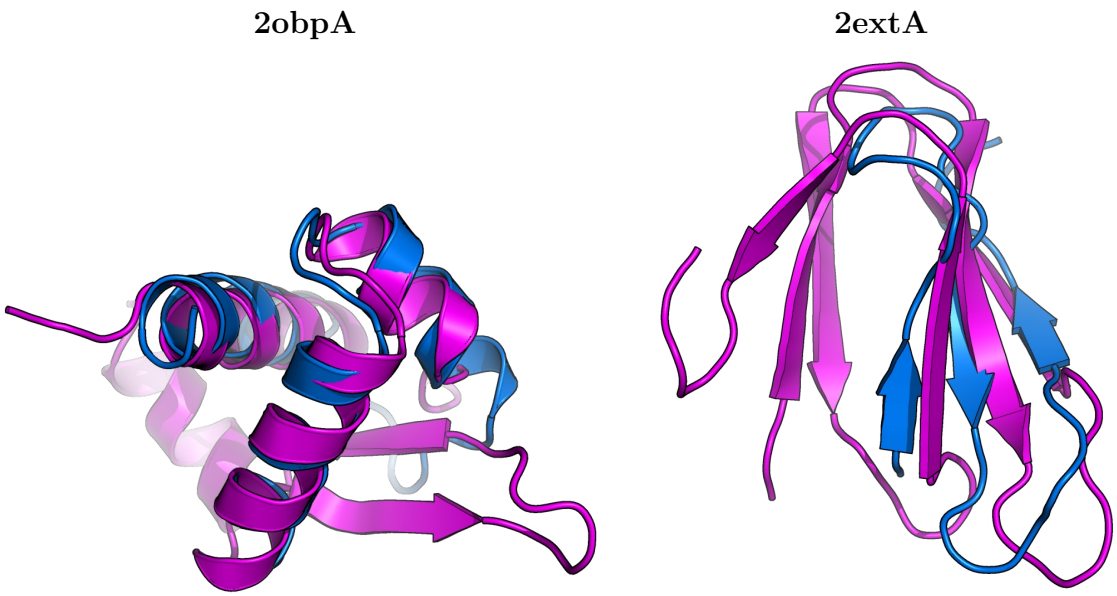


Fig. 10.9: Deux exemples de reconstruction hiérarchique. Le fragment prédit (en bleu) est superposé à la structure native (en magenta).

10.2.3 Conclusions préliminaires et perspectives

Les quelques tests récents que nous avons pu réaliser, pour un ensemble de structures n'appartenant pas aux banques d'apprentissage, ont mis en évidence la nécessité d'avoir une stratégie générique efficace de découpage et de population des zones de transition entre les fragments pour pallier la rigidité de la procédure d'assemblage. Une fois les fragments générés, nous tombons dans la même problématique que SAFrAN, *i.e.* de pouvoir générer des structures protéiques complètes cohérentes à partir d'un ensemble de fragments de taille variable, dont on sait qu'ils sont pertinents.

Une procédure de raffinement dans un espace continu serait ici à envisager. Cependant, la cible CASP7 (t0358) pour laquelle nous avons généré des modèles complets, nous a permis de mettre en évidence les progrès réalisés en un an, en dehors du contexte hiérarchique. Cette approche hiérarchique, encore en cours de développement, semble indiquer que l'algorithme glouton, couplé à la version courante de sOPEP v2.1, à partir des prédiction SAFrAN, est pertinent pour des coeurs de repliement correctement identifiés.

PDB	P_d	P_f	cRMSd	PDB	P_d	P_f	cRMSd
2cwrA	0	36	5.74	2iteA	0	33	6.62
	31	53	4.84		24	56	4.16
	48	77	5.78		49	70	3.97
	69	96	5.98		64	97	7.96
M			**5.58**		88	120	7.65
2extA	0	31	5.01		M		**6.07**
	25	51	4.75	2j73A	0	37	7.07
	45	62	4.53		34	52	3.14
	M		**4.76**		45	80	7.14
2g3rA	0	34	6.58		70	102	4.83
	29	48	2.86		M		**5.54**
	47	79	6.00	2o35A	0	55	8.25
	67	97	7.69		41	78	2.85
	92	118	5.72		M		**5.55**
	M		**5.77**	2o38A	0	45	8.96
2i4aA	0	27	5.31		37	88	9.62
	21	58	3.99		M		**9.29**
	52	81	5.08	2obpA	0	50	3.58
	76	106	3.45		37	67	3.71
	M		**4.46**		63	80	3.93
2ictA	0	42	7.75		M		**3.74**
	34	93	6.72				
	M		**7.23**				
2iojA	0	16	5.25				
	12	29	5.71				
	26	52	3.94				
	47	74	2.99				
	68	97	3.33				
	94	119	3.47				
	M		**4.11**				

Tab. 10.3: Tests de l'approche hiérarchique sur des protéines. Pour chaque fragment sont donnés les positions de départ (P_d) et de fin (P_f) de la région correspondante dans la protéine (PDB), ainsi que le meilleur cRMSd du modèle généré (cRMSd). M représente le cRMSd moyen sur l'ensemble des fragments d'une protéine.

Chapitre 11

Du repliement de peptides

Il est apparu lors de ces dix dernières années que les peptides ont un potentiel fort en terme d'applications biologiques. Une application thérapeutique prometteuse est apparue avec le développement de peptides pénétrant dans la cellule, qui peuvent apporter des protéines ou des acides nucléiques dans cette dernière (Zorko and Langel, 2005; Tréhin and Merkle, 2004; Dietz and Bahr, 2004; Deshayes et al., 2005). Cependant, il semblerait que la théorie d'Anfinsen (Anfinsen, 1973) soit mise à mal dans le cadre des peptides. En effet, leur taille courte induit qu'ils n'ont pas le même degré de stabilité que les protéines (Thomas et al., 2006). Un certain nombre de méthodes de prédiction de leur structure existent cependant. Pepstr (Kaur et al., 2007) utilise une méthode spécifique de prédiction des structures secondaires, optimisée pour la sélection des coudes β, et un raffinement par le champ de force AMBER (Case et al., 2005). PepLook (Thomas et al., 2006) utilise une recherche *de novo* d'un minimum énergétique par une procédure itérative de type Boltzmann-Stochastique pour sélectionner aléatoirement des couples de dièdres (ϕ, ψ) dérivés des *Protein Blocks* (de Brevern et al., 2007, 2000).

Nous avons voulu évaluer la pertinence de l'approche HMM-SA couplée à la nouvelle implémentation de sOPEP v2.1, dans le cadre de la prédiction de la structure d'un ensemble de 32 peptides ayant des classes et des topologies variées.

11.1 Matériels et méthodes

11.1.1 Le jeu de validation de peptides

Ce jeu de validation de 32 peptides (voir la table 11.1) a été constitué à partir du site de la PDB (requête avancée, disponible à l'adresse suivante : http ://www.rcsb.org/pdb/search/advSearch.do), selon les critères suivants : (i) leur taille est comprise entre 15 et 50 résidus, (ii) ils ne sont associés à aucun ligand, (iii) ne sont constitués que d'une seule chaîne protéique, et (iv) ne doivent pas partager plus de 30% d'identité de séquence. Le filtre de redondance en terme de séquence ne semblant pas fonctionnel, nous avons raffiné

#	PDB	Exp.	R	L	Peptide	# SS	% α	% β
Peptides principalement alpha (16)								
1	1abz	RMN	1,659	38	*De novo designed peptide*	0	0.816	-
2	1aie	RX	1,500	31	*p53 tetramerization domain*	0	0,645	-
3	1akg	RX	1,100	16	*Alpha-conotoxin PNIB*	2	0,438	-
4	1bde	RMN	2,752	33	*C-terminal half of HIV-1 VPR*	0	0,848	-
5	1bgk	RMN	0,787	37	*Sea anemone toxin (BGK)*	3	0,243	-
6	1brv	RMN	1,389	19	*Immunodominant region of Protein G*	2	0,316	-
7	1erc	RMN	0,660	40	*Pheromone ER-1*	3	0,650	-
8	1erd	RMN	1,092	40	*Pheromone ER-2*	3	0,475	-
9	1ery	RMN	0,644	39	*Pheromone ER-11*	3	0,667	-
10	1f4i	RMN	3,240	45	*UBA2 DNA binding protein*	0	0.578	-
11	1fsd	RMN	1,956	28	*De novo designed peptide*	0	0.357	-
12	1gw3	RMN	*n.d.*	46	*Helix-hinge-helix structural motif in human apolipoprotein A-I*	0	0,435	-
13	1ifp	RX	3,100	44	*Inovirus strain PF3 major coat protein assembly*	0	0,955	-
14	1l2y	RMN	0,779	20	*Trp-Cage miniprotein construct TC5b*	0	0.450	-
15	1vii	RMN	*n.d.*	36	*Subdomain from chicken Villin headpiece*	0	0.583	-
16	1vpu	RMN	2,343	45	*HIV-1 VPU cytoplasmic domain*	0	0,489	-
Peptides principalement beta (11)								
1	1agg	RMN	5,679	48	*Omega-AGA-IVB*	4	-	0,125
2	1axh	RMN	1,115	37	*Atracotoxin-HVI*	3	-	0,324
3	1az6	RMN	3,447	36	*Cellulose-binding domains of cellobiohydrolase I*	2	-	0,250
4	1bnb	RMN	3,548	38	*Beta-defensin 12*	3	-	0,395
5	1eit	RMN	1,431	36	*Mu-agatoxin*	4	-	0,417
6	1fsb	RMN	1,463	40	*EGF domain of P-selectin*	3	-	0,400
7	1j4m	RMN	*n.d.*	14	*De novo designed peptide*	0	-	0.429
8	1lfc	RMN	0,973	25	*Lactoferricin*	1	-	0,320
9	1pdc	RMN	*n.d.*	45	*Collagen-binding type II domaine*	2	-	0,133
10	1qdp	RMN	1,033	42	*Robustoxin*	4	-	0,143
11	2b38	RMN	1,981	31	*Kalata B8*	3	-	0.194
Peptides alpha/beta (5)								
1	1bbg	RMN	*n.d.*	40	*Ragweed pollen allergen*	4	0,150	0,250
2	1bhi	RMN	3,248	38	*Transactivation domain of CREBP1/ATF-2*	0	0,316	0,105
3	1bk8	RMN	0,642	50	*Aesculus hippocastanum antimicrobial protein 1*	4	0,200	0,360
4	1chl	RMN	1,342	36	*Chlorotoxin*	4	0,250	0,167
5	1gps	RMN	1,426	47	*Gamma 1-P thionins*	4	0,234	0,362

Tab. 11.1: Approche hiérarchique : le jeu de validation. Pour chaque peptide sont donnés, son identifiant PDB (PDB), la méthode de résolution expérientale (Exp.), sa résolution (structures RX) ou pseudo-résolution (structures RMN) (R) en Å, sa taille en acides aminés (L), son nom (Peptide), le nombre de ponts disulfure qui le constitue (# SS), et le pourcentage de résidus impliqués dans une structure de type hélice α (% α) ou feuillet β (% β). *n.d.* donnée non disponible (structure RMN avec un seul modèle proposé).

la liste résultante à l'aide du logiciel PISCES (Wang and Dunbrack, 2003). Nous avons ensuite sélectionné ce sous-ensemble de 32 peptides en ne conservant que ceux possédant des structures secondaires, et en ignorant les topologies redondantes. Pour les structures RMN comprenant plusieurs modèles, nous avons calculé la structure moyenne du peptide en calculant les coordonnées moyennes de chaque atome, sur l'ensemble des peptides superposés. Nous avons ensuite évalué la pseudo-résolution de ces structures comme étant le plus grand écart de cRMSd entre la structure moyenne et les modèles.

11.1.2 Description locale des structures

PDB	**3D 10^{-6}**		**SAFrAN**		**PDB**	**3D 10^{-6}**		**SAFrAN**	
	R	**C_{p3}**	**R**	**C_{p3}**		**R**	**C_{p3}**	**R**	**C_{p3}**
1abz	0,46	7,30	0,34	35,00	1agg	0,61	7,60	1,04	19,60
1aie	0,30	5,70	0,36	24,20	1axh	0,47	5,60	0,68	23,90
1akg	0,25	5,30	0,61	15,40	1az6	0,62	6,50	0,96	19,00
1bde	0,26	5,80	0,28	30,20	1bnb	0,69	6,70	1,62	16,50
1bgk	0,60	7,20	0,81	20,30	1eit	0,65	7,10	1,15	15,20
1brv	0,41	6,80	1,19	13,00	1fsb	0,61	6,40	1,17	16,40
1erc	0,67	7,10	1,95	12,20	1j4m	0,54	4,70	0,45	14,60
1erd	0,46	7,00	1,42	18,80	1lfc	0,48	5,80	1,03	20,10
1ery	0,62	6,50	1,16	22,50	1pdc	0,83	6,10	1,00	14,60
1f4i	0,38	7,30	0,51	28,50	1qdp	0,74	6,60	1,48	15,40
1fsd	0,49	6,90	1,47	12,20	2b38	0,44	5,00	1,32	16,20
1gw3	0,42	7,90	0,85	29,50	1bbg	0,56	6,70	1,08	17,10
1ifp	0,27	7,70	0,82	29,10	1bhi	0,50	7,40	0,96	18,10
1l2y	0,36	7,30	1,66	13,40	1bk8	0,52	7,30	1,14	17,20
1vii	0,42	6,50	0,80	18,90	1chl	0,81	6,30	0,97	12,50
1vpu	0,53	8,90	0,59	36,30	1gps	0,52	7,60	0,95	14,90
					μ	*0,52*	*6,71*	*0,99*	*19,71*
					σ	*0,15*	*0,90*	*0,40*	*6,62*

Tab. 11.2: Propriétés des trajectoires floues et prédites. Pour chacune des trajectoires, floue (3D 10^{-6}) ou prédite (SAFrAN) sont données sa complexité (C_{p3}) en ne considérant que trois prototypes maximum par lettre HMM-SA, et le meilleur cRMSd (R) en Å, que l'on peut atteindre avec cette trajectoire.

Nous avons testé deux séries de trajectoires pour les simulations de repliement avec l'algorithme glouton. La première série est une description floue des structures : toutes les lettres ayant une probabilité supérieure à 10^{-6} par l'encodage des structures avec l'algorithme de *forward-backward* sont considérées. Ces trajectoires, peu complexes, nous permettent de tester le champ force couplé à l'algorithme glouton, dans un cas de prédiction idéale. La deuxième série de trajectoires est issue de la prédiction SAFrAN à partir de la séquence en acides aminés. Pour rappel : afin de constituer ces trajectoires, les mots HMM-SA (fragments protéiques), identifiés par SAFrAN, sont discrétisés en lettres

à chaque position de la séquence, puis, les trajectoires sont filtrées en terme de transition markovienne, pour éviter les extrémités borgnes. La complexité des trajectoires, en utilisant trois prototypes maximum par lettre HMM-SA, ainsi que le cRMSd de la meilleure reconstruction que l'on puisse obtenir à partir de cette dernière ont été résumés dans la table 11.2. Nous savons que les trajectoires floues contiennent la solution optimale (0.5 Å en moyenne) pour une complexité faible (toujours inférieure à 9). Les trajectoires prédites sont en moyenne trois fois plus complexes, et conduisent à une solution qui s'éloigne de moins de 2 Å de la structure native (1 Å en moyenne).

11.1.3 Les simulations de repliement

Les simulations de repliement ont été réalisées avec le même protocole que celui utilisé pour replier des fragments dans l'approche hiérarchique. Pour chaque peptide et chaque condition, 30 simulations indépendantes ont été réalisées. Pour chaque simulation, un cycle de l'algorithme glouton est suivi de 300.000 pas de MC pour raffiner le modèle. Dans le cas des simulations zip, le point de départ le long de la séquence est aléatoire, et, dans le cas contraire, nous reconstruisons toujours à partir de l'extrémité C-terminale. L'information de connectivité des ponts disulfure est donnée ou non en entrée de l'algorithme glouton selon les simulations.

11.2 Résultats

11.2.1 A partir d'une trajectoire floue

Globalement, l'approche linéaire, sans introduire l'information de connectivité des ponts disulfure (table 11.3, colonne sOPEP v2.1.1), nous permet de reconstruire l'ensemble des peptides avec un cRMSd moyen, par rapport à la structure native, de 4,3 Å. Treize des peptides reconstruits sont éloignés de moins de 3 Å de la structure expérimentale. L'utilisation de l'opérateur Zip, dans les simulations de repliement des peptides, est associée à un gain moyen de 1 Å de cRMSd sur les modèles prédits. Pour 28 des 32 peptides, l'opérateur zip a conduit à une solution plus pertinente que l'approche linéaire. Ainsi, 18 cibles prédites ont un cRMSd inférieur à 3 Å par rapport à la structure expérimentale.

Que l'on soit dans l'approche linéaire ou zip, lorsque les ponts disulfure sont connus (formulation sOPEP v2.1.2), ils ne semblent pas améliorer la qualité des modèles générés (en terme de cRMSd). De plus, dans les deux cas, pour les modèles ayant des ponts disulfure, seul 50% d'entre eux ont un cRMSd plus faible avec la formulation sOPEP v2.1.2 par rapport à la formulation sOPEP v2.1.1.

		Linéaire								**Zip**							
		sOPEP v2.1.1				**sOPEP v2.1.2**				**sOPEP v2.1.1**				**sOPEP v2.1.2**			
PDB	S_n	R	S	R'	S'	R	S	R'	S'	R	S	R'	S'	R	S	R'	S'
Peptides principalement Alpha																	
1abz	0	2,291				2,263				2,198				2,073			
1aie	0	2,944				3,063				2,831				2,916			
1akg	2	0,852	0	1,049	2	0,676	2	0,676	2	0,870	0			0,562	2		
1bde	0	1,809				1,809				1,689				1,729			
1bgk	3	2,975	0			1,634	3			2,195	1			1,445	3		
1brv	2	2,147	0			1,455	2			1,710	0			1,401	2		
1erc	3	5,485	1			4,931	1	7,551	2	3,851	1	3,851	1	2,174	3		
1erd	3	2,580	0			5,962	2			2,698	1	3,186	2	2,044	3		
1ery	0	2,097				1,804				1,660				1,694			
1f4i	0	8,598				8,359				5,133				5,056			
1fsd	0	2,432				5,148				2,202				2,234			
1gw3	0	4,770				5,176				3,524				3,425			
1ifp	0	3,104				3,103				2,585				2,585			
1l2y	0	2,380				2,364				2,177				2,115			
1vii	0	8,002				8,063				2,147				2,079			
1vpu	0	6,149				6,762				3,555				3,610			
Peptides principalement beta																	
1agg	4	7,025	0	8,669	2	6,716	3	8,580	4	5,794	1	5,794	1	7,124	1	7,745	4
1axh	3	2,550	3			2,581	3			1,999	0	3,081	2	2,497	3		
1az6	2	4,627	0			6,331	2			4,131	0	7,321	1	6,233	2		
1bnb	3	5,578	0			8,595	0	8,985	2	5,570	0	6,930	1	5,347	3		
1eit	4	7,598	0			5,935	4			4,920	0	5,798	1	5,221	2	6,036	4
1fsb	3	6,351	0	6,393	1	5,981	3	5,981	3	4,627	0	6,628	1	2,511	3		
1j4m	0	0,933				0,861				0,854				0,861			
1lfc	1	5,075	0			7,372	1			2,088	0			5,029	1		
1pdc	2	7,349	0			6,582	2			7,365	0			6,687	2		
1qdp	4	6,803	0	8,776	1	6,943	4			5,869	0	6,922	1	5,872	4		
2b38	3	1,941	0	2,132	2	2,042	3			1,727	2			1,298	2	1,315	3
Peptides alpha/beta																	
1bbg	4	4,054	0	9,160	2	7,813	4			3,030	0	4,558	1	3,922	3	4,085	4
1bhi	0	4,645				5,595				3,353				4,819			
1bk8	4	6,330	1			9,220	3			6,886	0	8,401	1	5,669	1	7,821	4
1chl	4	3,122	1			2,430	4			2,484	0	2,827	1	2,330	3	2,445	4
1gps	4	5,018	2			4,745	1	8,290	3	4,313	0	5,656	1	4,425	4		
μ		*4,300*		*4,586*		*4,760*		*5,023*		*3,314*		*3,776*		*3,343*		*3,465*	
σ		*2,204*		*2,508*		*2,562*		*2,756*		*1,710*		*2,149*		*1,846*		*2,037*	

Tab. 11.3: Reconstruction de peptides à partir d'une trajectoire floue. Pour l'ensemble des peptides du jeu de validation sont présentés les cRMSds (R) les plus faibles obtenus pour les modèles générés, en comparaison avec la structure native ou moyenne, pour les structures résolues par RMN comprenant plusieurs modèles, et le nombre de ponts disulfure natifs du modèle (S). Si le meilleur modèle maximisant le nombre de ponts disulfure natifs est différent du meilleur modèle cRMSd, son cRMSd (R') et le nombre de ponts disulfure natifs qu'il comporte (S') sont présentés dans la colonne de droite. Si la connectivité des ponts disulfure est connue, sOPEP v2.1.2 est utilisée. Dans le cas contraire, nous utilisons par défaut sOPEP v2.1.1.

La reconstruction de cet ensemble de modèles semble plus pertinente pour les structures α, pour lesquelles nous avons 12 modèles sur 16 dont le cRMSd est inférieur à 3 Å, que pour les structures β ou α/β pour lesquelles nous avons respectivement 4 modèles sur

11 et 2 modèles sur 5 dont le cRMSd est inférieur à 3 Å.

Les cibles problématiques. La cible principalement *alpha* prédite avec le plus fort cRMSd est 1f4i, s'éloignant de 5.1 Å de la structure native. Cependant, comme nous pouvons le voir dans la figure 11.1, seules la position et la longueur de l'hélice 3 semblent fausses, la topologie de la structure étant globalement maintenue.

Pour les structures β, 1agg, 1az6 et 1bnb, les modèles résolus ont une grande variabilité (voir les pseudo-résolutions de la table 11.1), et nous n'avons pu les replier correctement. La cible 1agg n'est que très peu structurée, sa topologie est hautement contrainte par les ponts disulfure qui la compose. La meilleure reconstruction en terme de cRMSd, qui ne contient pas les ponts disulfure natifs, a une topologie plus éloignée de la structure native que la meilleure reconstruction possédant tous les ponts disulfure. L'alignement des structures par TM-align nous indique que 33 des 48 résidus s'alignent dans l'espace (sans décalage de la séquence), et sont éloignés de 3,5 Å en moyenne.

Pour la meilleure reconstruction de 1az6, la topologie du brin β C-terminal est respectée, mais le réseau de liaisons hydrogène n'a pu se former correctement, la conformation de la chaîne principale étant fausse dans cette zone. Dans la meilleure reconstruction cRMSd, la même erreur est présente, car, ce feuillet non canonique, n'est pas représenté par les lettres spécifiques des brins β dans l'espace HMM-SA. La présence des deux ponts disulfure natifs dans le meilleur modèle sOPEP v2.1.2 empêche totalement d'explorer de telles topologies pour l'extrémité C-terminale. Nous pouvons faire les mêmes remarques pour la cible 1bnb : la topologie du feuillet semble respectée, mais le réseau de liaisons hydrogène des brins β, trop contraint dans la structure expérimentale n'est pas correctement formé.

Similairement à 1agg, pour la cible 1eit, dans le meilleur modèle sOPEP v2.1.1, la topologie de l'extrémité C-terminale est inversée, alors que dans le meilleur modèle sOPEP v2.1.2, bien que la connectivité des ponts disulfure soit respectée, c'est l'extrémité N-terminale dont la topologie est inversée (voir la figure 11.1).

Dans le cas de la cible 1fsb, la connectivité des ponts disulfure permet d'améliorer grandement la pertinence du modèle (2,511 Å *versus* 4,627 Å avec la formulation sOPEP v2.1.1). Dans tous les cas, nous arrivons à former le brin β, mais dans le cas sOPEP v2.1.1 les cystéines n'arrivent pas à se voir pour former les ponts disulfure natifs. Cependant, dans tous les cas, la topologie du peptide est relativement proche, les TM-scores des deux modèles étant de 0,47 par rapport à la structure expérimentale.

Pour le peptide α/β 1bk8, même dans la meilleure reconstruction cRMSd, bien qu'ayant une conformation étendue à ce niveau, nous n'avons pas pu mettre en place le réseau natif de liaisons hydrogène nécessaire à la formation du feuillet β. En effet, ceci est dû à la présence d'un pli (*kink*) pour les brins 2 et 3 dans la structure RMN (Fant et al., 1999).

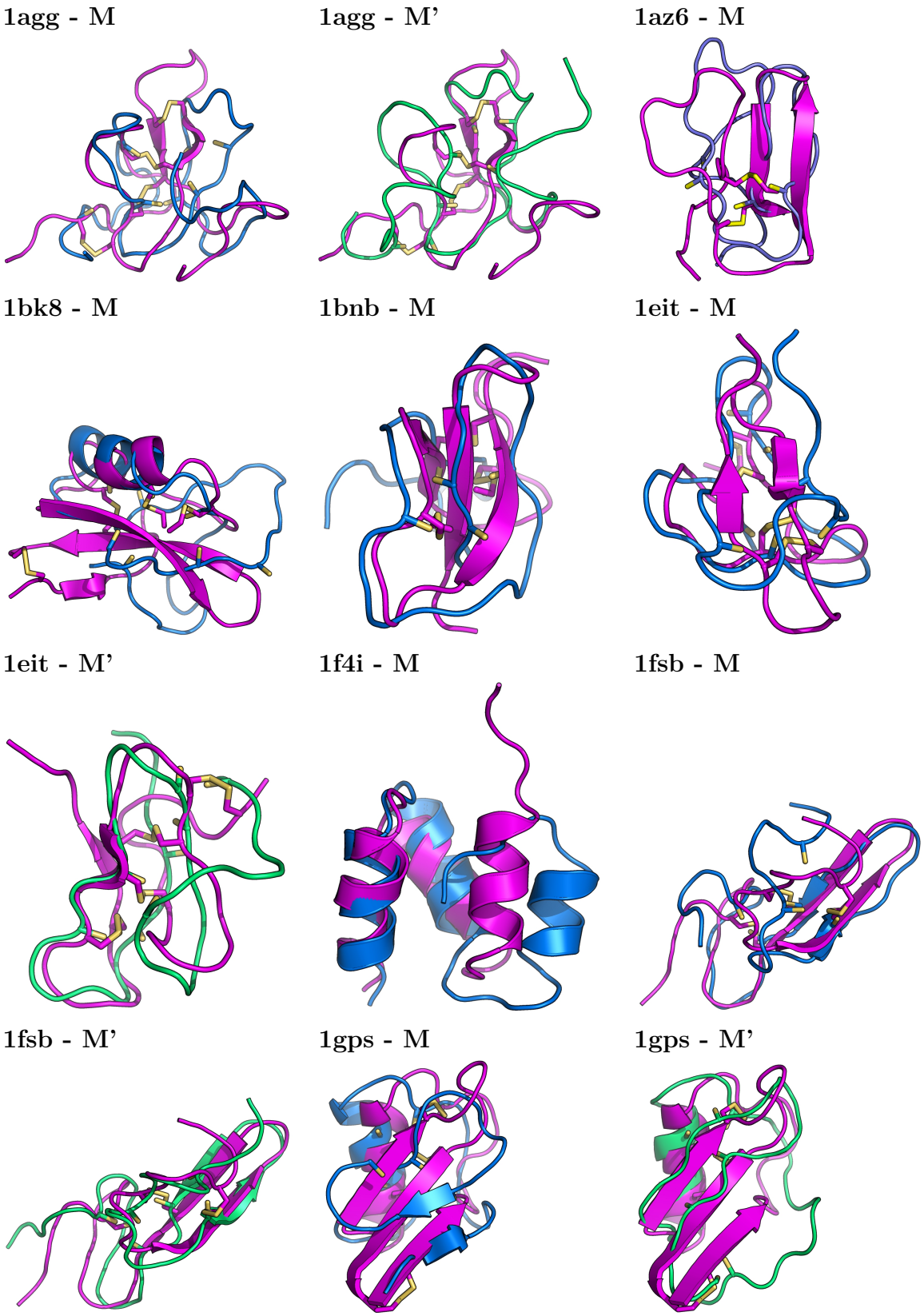


Fig. 11.1: Trajectoires floues : les peptides problématiques. Pour chaque peptide, la structure native, en magenta, est superposée au meilleur modèle en terme de cRMSd (M) en bleu, ou au meilleur modèle maximisant le nombre de ponts disulfure natifs (M') en vert.

Ainsi, dans les modèles générés, l'hélice est correctement constituée, mais nous n'avons pu obtenir une topologie native (voir la figure 11.1). Il se peut que cette limitation soit due au nombre restreint de prototypes que nous utilisons. Ce point reste à préciser dans des analyses ultérieures.

La cible 1gps reconstruite avec la version sOPEP v2.1.1 du champ de force présente une topologie native. Cependant, la conformation de l'extrémité C-terminale empêche le brin β 2 de former les liaisons hydrogène natives avec le brin 3. Il est à noter que, dans le meilleur modèle sOPEP v2.1.2, la formation du pont disulfure entre les cystéines 3 et 47 déstabilise totalement la formation du premier brin.

Globalement, si l'on ne considère que les cibles n'ayant pas de ponts disulfure, le cRMSd moyen des meilleurs modèles générés par l'algorithme glouton, couplé à la version actuelle de sOPEP, tombe à 2,6 Å. Sur des peptides, la présence de ponts disulfure semble contraindre énormément les structures. L'algorithme glouton, couplé à sOPEP v2.1, a ainsi du mal à reproduire un tel paysage énergétique.

11.2.2 A partir d'une trajectoire prédite

		sOPEP v2.1.1				**sOPEP v2.1.2**			
PDB	S_n	R	S	R'	S'	R	S	R'	S'
Peptides principalement Alpha									
1abz	0	2,950				2,833			
1aie	0	3,615				5,241			
1akg	2	1,668	0			1,749	2		
1bde	0	2,058	0			2,015			
1bgk	3	5,547	0	5,639	1	5,200	2		
1brv	2	3,039	0	3,823	1	2,733	1	2,911	2
1l2y	0	4,364				3,994			
1vii	0	4,383				4,962			
1vpu	0	5,930				6,069			
Peptides principalement beta									
1j4m	0	1,321				1,415			
Peptides alpha/beta									
1chl	4	3,632	0	5,129	1	3,539	2		

Tab. 11.4: Reconstruction de peptides à partir d'une trajectoire prédite. Nous n'avons utilisé dans ce cas-ci que l'opérateur Zip. Voir la table 11.4 pour la légende.

Le cRMSd moyen obtenu pour l'ensemble de ces modèles à partir des prédiction SA-FrAN est de 6,3 Å. L'application de la méthode HMM-SA dans son ensemble pour les peptides de notre jeu de validation n'a pas donné les résultats escomptés. Nous avons choisi de ne présenter que les meilleurs résultats obtenus pour 11 des 32 peptides du JV

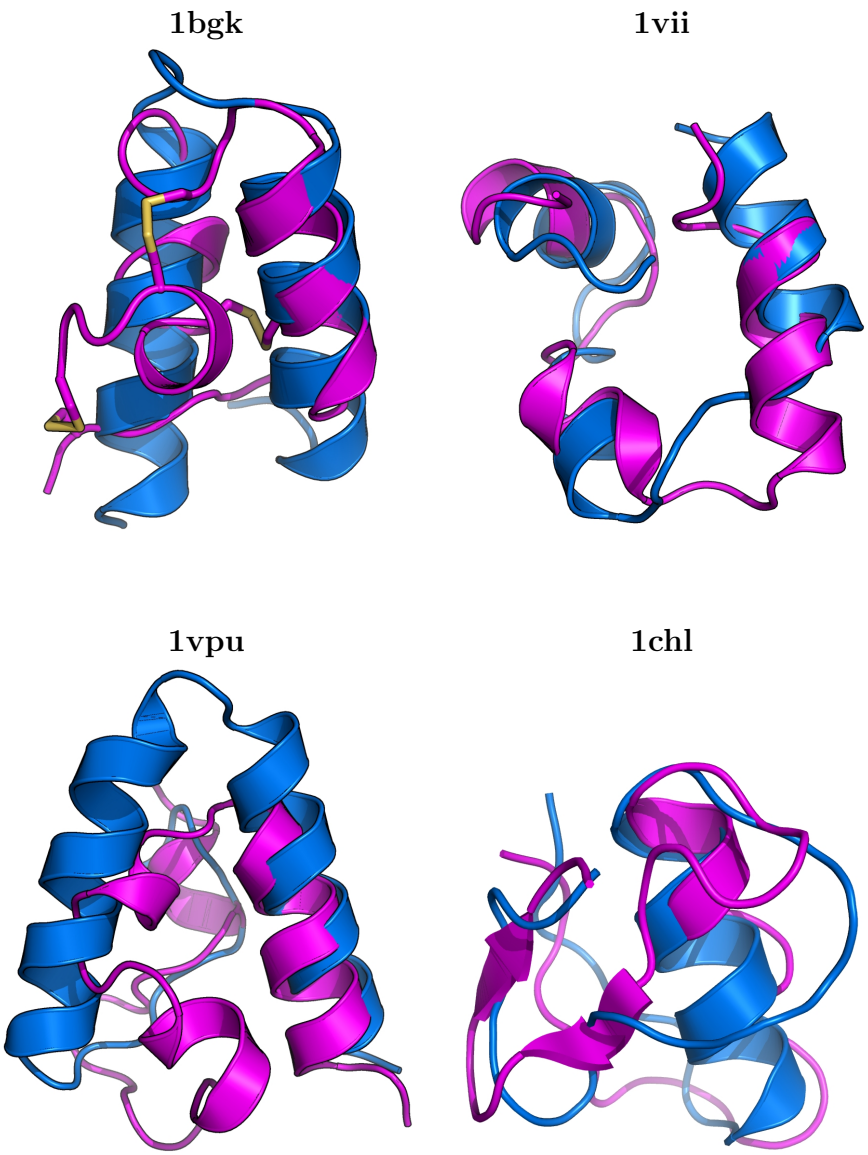


Fig. 11.2: Trajectoires prédites : quelques exemples. Pour chaque peptide, la structure native, en magenta, est superposée au meilleur modèle en terme de cRMSd (M) en bleu.

dans la table 11.4. Nous savons, des reconstructions réalisées par l'algorithme glouton en critère cRMSd, que les trajectoires prédites par SAFrAN, bien que complexes, nous permettent d'approximer la structure native à 1 Å en moyenne. Autrement dit, nous avons la solution dans la prédiction SAFrAN, mais nous n'avons pas pu la générer. Deux points de la méthode peuvent alors être remis en question : l'exploration de l'espace conformationnel par l'algorithme glouton ou le champ de force OPEP.

Pour des trajectoires complexes, il semblerait que l'algorithme glouton, guidé par sOPEP, ait tendance à générer des hélices qui, une fois formées, ne peuvent être rompues au profit de la formation de ponts disulfure, comme c'est le cas pour 1bgk (voir la figure 11.2). Pour le peptide 1vii, s'éloignant de la structure native de 4,4 Å, nous avons réussi à

obtenir une topologie proche de la structure native (figure 11.2). Le TM-score du meilleur modèle est de 0,42. Dans le cas de 1vpu, la première hélice amphipathique semble bien formée et orientée, et la seconde hélice semble mieux défini que dans la structure RMN, et ne serait pas incohérente (Willbold et al., 1997). Pour le peptide 1chl, la topologie du peptide a été retrouvée, et l'hélice correctement formée, mais les brins ne sont pas formés (figure 11.2).

Le sous-ensemble des 11 peptides analysés ici semble suggérer que nous avons moins de difficulté à gérer les motifs structuraux de type hélice α que feuillet β, sauf quand ils sont canoniques (cas du peptide 1j4m replié à 1,3 Å de la structure native).

11.2.3 Analyse de la pertinence de sOPEP dans un espace discret

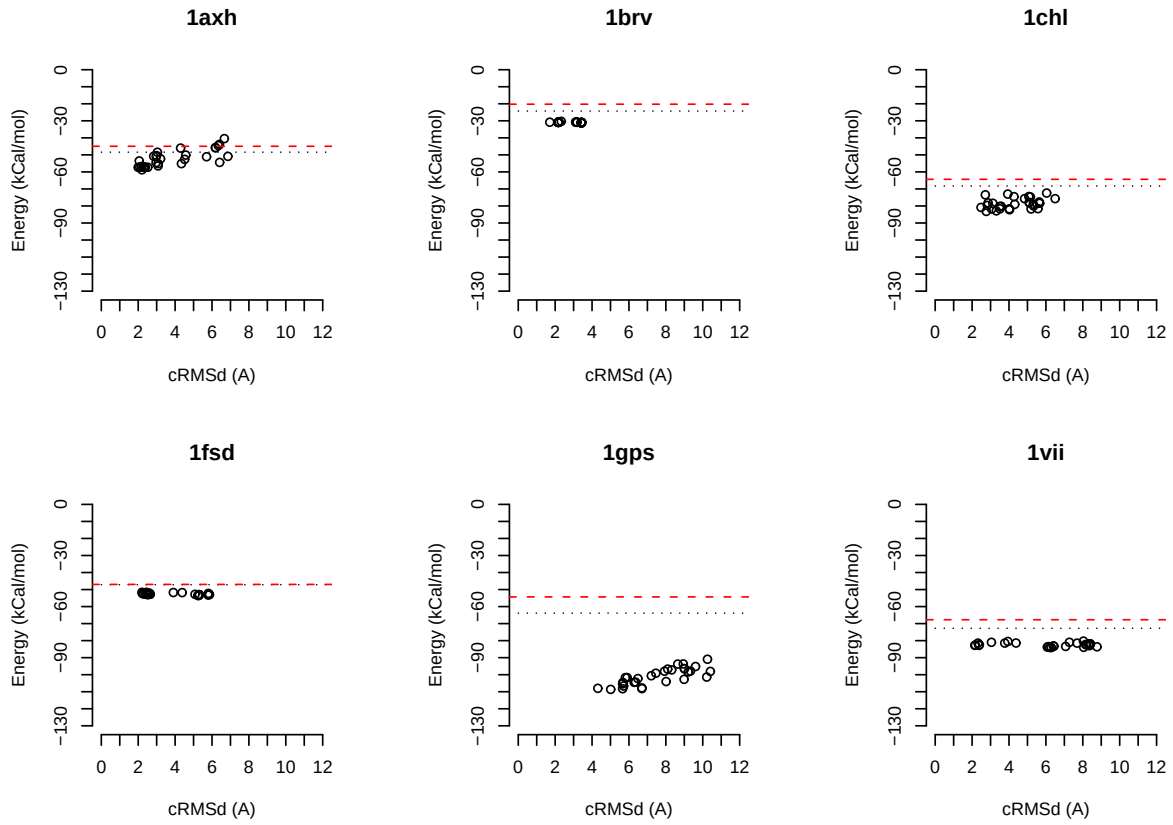


Fig. 11.3: Le pouvoir discriminant de sOPEP v2.1 sur les peptides du JV. Pour quelques cibles du JV sont présentées les valeurs d'énergies obtenues pour les modèles générés, en fonction du cRMSd que partagent ces derniers avec la structure native. La ligne horizontale rouge (noire) en pointillé représente l'énergie de la structure native lorsqu'au maximum 3 (l'ensemble des) prototypes sont considérés par lettre HMM-SA.

Nous avons montré que sOPEP v2.1 dans un espace continu a un pouvoir de reconnaissance de la structure native similaire à OPEP v3.1. Cependant, nous n'avions jusqu'alors pas étudié en détail ce pouvoir de reconnaissance dans un espace discret. Ainsi, nous avons

analysé la relation entre les énergies obtenues et les cRMSds des modèles générés pour la formulation sOPEP v2.1. Quelques exemples pour les trajectoires floues sont fournis dans la figure 11.3. L'energie de la structure native dans l'espace discrétisé a été évaluée en reconstruisant ces dernières avec l'algorithme glouton et sOPEP v2.1 pour une trajectoire exacte. Cependant, une telle procédure ne garantit pas que la conformation générée sera proche de la structure native. Nous avons donc considéré le potentiel E_{RMSd} pour limiter les déviations par rapport à la structure native. La valeur énergétique associée à ce potentiel a ensuite été otée de l'énergie totale pour en déduire l'énergie de référence de la cible.

Comme nous pouvons le constater, et c'est valable pour l'ensemble des 32 peptides du JV, la structure native n'est jamais la structure de plus basse énergie. De plus, la corrélation entre cRMSds et énergies est mauvaise : dans le cas de 1vii par exemple le même bassin énergétique correspond à un intervalle de cRMSds allant de 2 à 9 Å ; cependant pour le peptide 1gps, la tendance semble aller dans le bon sens.

Bien qu'OPEP puisse générer des structures de topologie native, il semblerait que sa capacité à distinguer des conformations natives de conformations non natives soit, dans ce cas précis, à remettre en cause. Nous pouvons nous demander si cette observation n'est pas la conséquence d'une discrétisation trop importante de l'espace conformationnel étant donné que nous n'utilisons ici que 74 prototypes différents au lieu des 155 initiaux. Ainsi, de la même manière que celle précédemment énoncée, nous avons recalculé les énergies de référence pour chaque cible (ligne horizontale noire de la figure 11.3). Bien qu'il soit apparu, pour certaines cibles, que cette ouverture de l'espace conformationnel puisse diminuer significativement l'énergie du modèle de référence, les modèles prédits ont toujours une énergie inférieure à la structure native. Pour nous conforter dans l'idée que nous n'avions pas eu tort de ne pas considérer les potentiels de géométrie (E_{locale}) d'OPEP v3.1 dans sOPEP v2.0, nous les avons considérés pendant les simulations de repliement, aucune amélioration n'a été observée sur la qualité des modèles générés. Il apparait ainsi que l'optimisation de sOPEP dans un espace continu ne semble pas adaptée à son utilisation dans espace discret.

Bien qu'il faille encore s'assurer que ces observations ne soient pas inhérentes au jeu de peptides testé, deux voies peuvent être explorées. La première réside dans l'optimisation de sOPEP sur un ensemble de leurres discrétisés dans l'espace HMM-SA par la même procédure d'optimisation qu'OPEP v3.0. La deuxième solution envisageable serait une procédure de raffinement dans un espace continu avec sOPEP v2.1.

11.3 Conclusions de l'étude

La recontruction d'un ensemble non redondant de peptides, appartenant à des classes protéiques diverses, a pu être réalisée, par l'intermédiaire de notre version améliorée de l'algorithme glouton, couplée à une version simplifiée d'OPEP dont la formulation du potentiel entre chaînes latérales a été optimisée. A partir d'une description floue des structures, nous avons pu reconstruire ces peptides avec un cRMSd moyen de 3,3 Å. Pour 18 cibles sur 32, nous avons généré un modèle à moins de 3 Å de la conformations native. Plus de 90 % des peptides du jeu de validation étant des structures RMN, ces valeurs de cRMSd n'ont une portée que limitée, nous envisageons donc dans un avenir proche de vérifier si les modèles générés pour ces peptides vérifient les contraintes NOE expérimentales inclues dans la base de données BMRB (Seavey et al., 1991).

La formulation actuelle du potentiel associé aux ponts disulfure ne permet de former que 30 % des ponts disulfure natifs quand nous ne les connaissons pas *a priori*. Il reste encore des efforts à faire en ce sens. Deux voies ont été explorées : (i) lorsque deux cystéines se rapprochent, nous avons fait évoluer la formulation de la version 2.1.1 vers la version 2.1.2 , et (ii) pour éviter de former des noyaux de cystéines, un terme répulsif a été ajouté comme suggéré dans Derreumaux (1999). Ces deux ajouts à la fonction d'énergie n'ont en rien amélioré la pertinence des modèles générés à partir des prédictions SAFrAN. Par ailleurs, le jeu de validation comprend 13 structures sur 32 sans ponts disulfure, dont seulement une dans les catégories β et α/β. Il serait donc intéressant d'augmenter le nombre de structures sans ponts disulfure dans ces catégories pour débiaiser l'analyse.

L'analyse du pouvoir discriminant de sOPEP sur ces peptides nous a permis de mettre en évidence la nécessité d'optimiser à nouveau sOPEP sur un ensemble de leurres évoluant dans un espace discret, et de pouvoir raffiner les modèles générés dans un espace continu.

Les résultats obtenus en prédiction pourraient suggérer que les peptides n'obéissent pas aux mêmes règles de repliement que les protéines, et nécessiteraient donc un paramétrage spécifique du champ de force. Thomas et al. (2006) ont révélé que Robetta (Chivian et al., 2005) présenterait les même faibles performances en comparaison avec des méthodes de prédiction optimisées pour les peptides.

Cinquième partie

Conclusions et perspectives

Lors de mon travail de thèse, nous avons consacré beaucoup de temps à la mise en place du potentiel gros grain OPEP dans notre algorithme de reconstruction de modèles protéiques, et à son optimisation dans un modèle continu. La version optimisée du champ de force OPEP semble avoir un pouvoir discriminant équivalent au potentiel tous atomes DOPE qui est plus performant que cinq champs de force reconnus.

L'implémentation d'OPEP dans l'algorithme glouton, a été l'occasion de développer SABBAC, une méthode originale de reconstruction de modèles tous atomes à partir de la seule trace des protéines. SABBAC présente des performances équivalentes au logiciel pionnier de référence MaxSprout et est plus performant que la méthode *bb* développée par SA Adcock.

Une fois la version simplifiée du potentiel OPEP intégrée dans la méthode de prédiction HMM-SA, nous avons pu tester la méthode de prédiction dans sa globalité lors de l'expérience CASP7. Notre participation à cette compétition internationale, nous a permis de mettre en évidence les points forts et les points faibles de la méthode. Le développement d'une méthode de modélisation par homologie en parallèle de la compétition n'a pas donné de résultats très pertinents en comparaison des méthodes existantes reconnues. Par contre, les cibles traitées à très faible taux d'identité de séquence ont donné des modèles intéressants. A l'issue de CASP7, nous avons donc entrepris de résoudre les problèmes rencontrés lors de la compétition et de formaliser les méthodes employées.

Ainsi, lors de ma dernière année de thèse, nous avons entrepris de re-paramétrer le potentiel d'interaction entre les chaînes latérales pour que ses paramètres correspondent au mieux aux distributions observées. Nous avons aussi ajouté un modèle pour les ponts disulfure. Par ailleurs, nous avons conjointement mis en place un nouvel opérateur dans l'algorithme de reconstruction permettant de dépasser les contraintes d'une reconstruction linéaire des structures protéiques.

En parallèle, j'ai contribué à l'amélioration et la validation de SAFrAN une méthode performante de recherche fragments candidats fournissant une prédiction pertinente d'un ensemble de fragments protéiques compatibles avec une séquence en acides aminés.

Enfin, mes travaux de ces dernières semaines ont porté sur le développement d'une approche hiérarchique pour reconstruire des modèles protéiques complets, en suivant les travaux de Skolnick et al. ayant fournit de très bons résultats lors de la dernière expérience CASP. Bien que ces travaux ne soient pas encore complètement aboutis, nous avons dores et déjà pu mettre en évidence les progrès réalisés depuis l'expérience CASP, en rejouant une cible avec la version courante de la méthode HMM-SA.

Les derniers résultats de prédiction obtenus sont prometteurs, bien qu'il faille encore investir des efforts pour améliorer la pertinence du champ de force sOPEP dans l'algorithme glouton. Une approche hiérarchique est en cours de développement. Le point sur lequel je me penche encore porte sur le développement d'un nouvel opérateur de reconstruction à partir de mots de longueur variable. Ce qui nous permettrait d'utiliser directement les mots prédits par SAFrAN ou générés par l'approche hiérarchique pour former des modèles protéiques complets (voir la figure 11.4).

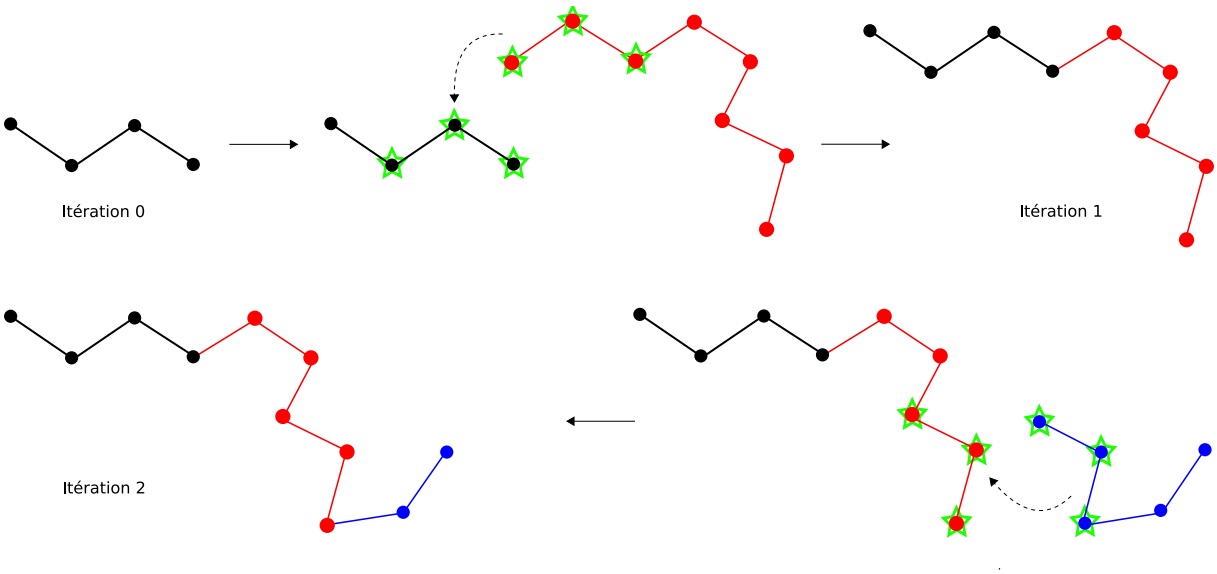


Fig. 11.4: Reconstruction à partir de mots HMM-SA. L'algorithme glouton est maintenant capable de reconstruire des modèles protéiques à partir de mots HMM-SA (fragments de plus de quatre résidus de long). La procédure de superposition est toujours la même : les trois derniers carbones *alpha* du modèle reconstruit sont superposés aux trois premiers carbones *alpha* du fragment à ajouter, si l'on considère une croissance de l'extrémité N vers C terminale.

A l'issue de ce travail de thèse, plusieurs perspectives peuvent être proposées. La première voie qu'il serait intéressant d'explorer serait la mise en place d'une étape de raffinement des modèles, par le biais du champ de force sOPEP ou d'un champ de force tous atomes dans un espace continu. La deuxième voie serait celle de la modélisation comparative : il serait intéressant d'y intégrer l'ensemble des outils HMM-SA ayant été développés au laboratoire, parmi lesquels SAFrAN, glouton-sOPEP et une méthode de prédiction de boucles.

Sixième partie

Bibliographie

Bibliographie

R Abagyan and M Totrov. Biased probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol*, 235(3) :983–1002, January 1994.

S A Adcock. Peptide backbone reconstruction using dead-end elimination and a knowledge-based forcefield. *J Comput Chem*, 25(1) :16–27, January 2004.

S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res*, 25(17) :3389–402, September 1997.

C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96) : 223–30, July 1973.

C B Anfinsen and H A Scheraga. Experimental and theoretical aspects of protein folding. *Adv Protein Chem*, 29 :205–300, 1975.

P Argos and J Palau. Amino acid distribution in protein secondary structures. *Int J Pept Protein Res*, 19(4) :380–93, April 1982.

I Bahar and R L Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol*, 266(1) :195–214, February 1997.

P Baldi, S Brunak, P Frasconi, G Soda, and G Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11) :937–46, November 1999.

R L Baldwin. Intermediates in protein folding reactions and the mechanism of protein folding. *Annu Rev Biochem*, 44 :453–75, 1975.

R L Baldwin and G D Rose. Is protein folding hierarchic ? i. local structure and peptide folding. *Trends Biochem Sci*, 24(1) :26–33, January 1999.

P A Bates, L A Kelley, R M MacCallum, and M J Sternberg. Enhancement of protein modeling by human intervention in applying the automatic programs 3d-jigsaw and 3d-pssm. *Proteins*, Suppl 5 :39–46, 2001.

L E Baum, T Petrie, G Soules, and N Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41 : 164–171, 1970.

C Benros, A G de Brevern, C Etchebest, and S Hazout. Assessing a novel approach for predicting local 3d protein structures from sequence. *Proteins*, 62(4) :865–80, March 2006.

H J C Berendsen, D Van Der Spoel, and R Van Drunen. Gromacs : A message-passing parallel molecular dynamics implementation. *Comp Phys Comm*, 91 :43–56, 1995.

I N Berezovsky and E N Trifonov. Van der waals locks : loop-n-lock structure of globular proteins. *J Mol Biol*, 307(5) :1419–26, April 2001a.

I N Berezovsky and E N Trifonov. Loop fold nature of globular proteins. *Protein Eng*, 14 (6) :403–7, June 2001b.

I N Berezovsky, A Y Grosberg, and E N Trifonov. Closed loops of nearly standard size : common basic element of protein structure. *FEBS Lett*, 466(2-3) :283–6, January 2000.

I N Berezovsky, V M Kirzhner, A Kirzhner, and E N Trifonov. Protein folding : looping from hydrophobic nuclei. *Proteins*, 45(4) :346–50, December 2001.

I N Berezovskyi and E Trifono. Protein structure and folding : a new start. *J Biomol Struct Dyn*, 19(3) :397–403, December 2001.

H M Berman, T N Bhat, P E Bourne, Z Feng, G Gilliland, H Weissig, and J Westbrook. The protein data bank and the challenge of structural genomics. *Nat Struct Biol*, 7 Suppl :957–9, November 2000a.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res*, 28(1) :235–42, January 2000b.

M R Betancourt and D Thirumalai. Pair potentials for protein folding : choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*, 8(2) :361–9, February 1999.

T L Blundell, B L Sibanda, M J Sternberg, and J M Thornton. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111) :347–52, March 1987.

P J Bond and M S P Sansom. Insertion and assembly of membrane proteins via simulation. *J Am Chem Soc*, 128(8) :2697–704, March 2006.

P J Bond, J Holyoake, A Ivetac, S Khalid, and M S P Sansom. Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J Struct Biol*, 157(3) :593–605, March 2007.

R Bonneau, C E M Strauss, C A Rohl, D Chivian, P Bradley, L Malmstrom, T Robertson, and D Baker. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol*, 322(1) :65–78, September 2002.

J M Borreguero and J Skolnick. Benchmarking of tasser in the ab initio limit. *Proteins*, 68(1) :48–56, July 2007.

N S Boutonnet, A V Kajava, and M J Rooman. Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins*, 30(2) :193–212, February 1998.

H Boutselakis, D Dimitropoulos, J Fillon, A Golovin, K Henrick, A Hussain, J Ionides, M John, P A Keller, E Krissinel, P McNeil, A Naim, R Newman, T Oldfield, J Pineda, A Rachedi, J Copeland, A Sitnov, S Sobhany, A Suarez-Uruena, J Swaminathan, M Tagari, J Tate, S Tromm, S Velankar, and W Vranken. E-msd : the european bioinformatics institute macromolecular structure database. *Nucleic Acids Res*, 31(1) : 458–62, January 2003.

J U Bowie, R Luthy, and D Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016) :164–70, July 1991.

P Bradley, L Malmstrom, B Qian, J Schonbrun, D Chivian, D E Kim, J Meiler, K M S Misura, and D Baker. Free modeling with rosetta in casp6. *Proteins*, 61 Suppl 7 : 128–34, 2005.

B R Brooks, R E Bruccoleri, B D Olafson, D J States, S Swaminathan, and M Karplus. Charmm : A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*, 4(4) :187–217, 1983.

S Brown, N J Fawzi, and T Head-Gordon. Coarse-grained sequences for protein folding and design. *Proc Natl Acad Sci U S A*, 100(19) :10712–7, September 2003.

N V Buchete, J E Straub, and D Thirumalai. Development of novel statistical potentials for protein fold recognition. *Curr Opin Struct Biol*, 14(2) :225–32, April 2004a.

N V Buchete, J E Straub, and D Thirumalai. Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. *J Mol Graph Model*, 22(5) :441–50, May 2004b.

C Bystroff and D Baker. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281(3) :565–77, August 1998.

C Bystroff, V Thorsson, and D Baker. Hmmstr : a hidden markov model for local sequence-structure correlations in proteins. *J Mol Biol*, 301(1) :173–90, August 2000.

A Caflisch. Network and graph analyses of folding free energy surfaces. *Curr Opin Struct Biol*, 16(1) :71–8, February 2006.

A C Camproux and P Tuffery. Hidden markov model-derived structural alphabet for proteins : the learning of protein local shapes captures sequence specificity. *Biochim Biophys Acta*, 1724(3) :394–403, August 2005.

A C Camproux, P Tuffery, J P Chevrolat, J F Boisvieux, and S Hazout. Hidden markov model approach for identifying the modular framework of the protein backbone. *Protein Eng*, 12(12) :1063–73, December 1999.

A C Camproux, R Gautier, and P Tuffery. A hidden markov model derived structural alphabet for proteins. *J Mol Biol*, 339(3) :591–605, June 2004.

A A Canutescu, A A Shelenkov, and R L Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9) :2001–14, September 2003.

D A Case, T E Cheatham, T Darden, H Gohlke, R Luo, K M Merz, A Onufriev, C Simmerling, B Wang, and R J Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16) :1668–88, December 2005.

A W Chan, E G Hutchinson, D Harris, and J M Thornton. Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci*, 2(10) :1574–90, October 1993.

W Chen, N Mousseau, and P Derreumaux. The conformations of the amyloid-beta (21-30) fragment can be described by three families in solution. *J Chem Phys*, 125(8) :084911, August 2006.

Y Z Chen. Modified self-consistent harmonic approach to thermal fluctuational disruption of disulfide bonds in proteins. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 60(5 Pt B) :5938–42, November 1999.

G Chikenji, Y Fujitsuka, and S Takada. A reversible fragment assembly method for de novo protein structure prediction. *J Chem Phys*, 119 :6895–6903, 2003.

F Chiti, M Stefani, N Taddei, G Ramponi, and C M Dobson. Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950) : 805–8, August 2003.

D Chivian and D Baker. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res*, 34(17) : e112, 2006.

D Chivian, D E Kim, L Malmstrom, P Bradley, T Robertson, P Murphy, C E M Strauss, R Bonneau, C A Rohl, and D Baker. Automated prediction of casp-5 structures using the robetta server. *Proteins*, 53 Suppl 6 :524–33, 2003.

D Chivian, D E Kim, L Malmstrom, J Schonbrun, C A Rohl, and D Baker. Prediction of casp6 structures using automated robetta protocols. *Proteins*, 61 Suppl 7 :157–66, 2005.

J Chomilier, M Lamarine, J P Mornon, J H Torres, E Eliopoulos, and N Papandreou. Analysis of fragments induced by simulated lattice protein folding. *C R Biol*, 327(5) : 431–43, May 2004.

K C Chou. Prediction and classification of alpha-turn types. *Biopolymers*, 42(7) :837–53, December 1997.

M Christen, P H Hunenberger, D Bakowies, R Baron, R Burgi, D P Geerke, T N Heinz, M A Kastenholz, V Krautler, C Oostenbrink, C Peter, D Trzesniak, and W F van Gunsteren. The gromos software for biomolecular simulation : Gromos05. *J Comput Chem*, 26(16) :1719–51, December 2005.

M Claessens, E Van Cutsem, I Lasters, and S Wodak. Modelling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng*, 2(5) :335–45, January 1989.

C Clementi, P A Jennings, and J N Onuchic. Prediction of folding mechanism for circular-permuted proteins. *J Mol Biol*, 311(4) :879–90, August 2001.

M Coincon, A Heitz, L Chiche, and P Derreumaux. The beta-alpha-beta-alpha-beta elementary supersecondary structure of the rossmann fold from porcine lactate dehydrogenase exhibits characteristics of a molten globule. *Proteins*, 60(4) :740–5, September 2005.

A Colubri. Prediction of protein structure by simulating coarse-grained folding pathways : a preliminary report. *J Biomol Struct Dyn*, 21(5) :625–38, April 2004.

P E Correa. The building of protein structures from alpha-carbon coordinates. *Proteins*, 7(4) :366–77, 1990.

G E Crooks and S E Brenner. Protein secondary structure : entropy, correlations and prediction. *Bioinformatics*, 20(10) :1603–11, July 2004.

J A Cuff, M E Clamp, A S Siddiqui, M Finlay, and G J Barton. Jpred : a consensus secondary structure prediction server. *Bioinformatics*, 14(10) :892–3, 1998.

C Czaplewski, A Liwo, J Pillardy, S Oldziej, and H A Scheraga. Improved conformational space annealing method to treat beta-structure with the unres force-field and to enhance scalability of parallel implementation. *Polymer*, 45 :677–686, 2004a.

C Czaplewski, S Oldziej, A Liwo, and H A Scheraga. Prediction of the structures of proteins with the unres force field, including dynamic formation and breaking of disulfide bonds. *Protein Eng Des Sel*, 17(1) :29–36, January 2004b.

J A R Dalton and R M Jackson. An evaluation of automated homology modelling methods at low target template sequence similarity. *Bioinformatics*, 23(15) :1901–8, August 2007.

A G de Brevern and S Hazout. 'hybrid protein model' for optimally defining 3d protein structure fragments. *Bioinformatics*, 19(3) :345–53, February 2003.

A G de Brevern, C Etchebest, and S Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3) :271–87, November 2000.

A G de Brevern, C Etchebest, C Benros, and S Hazout. "pinning strategy" : a novel approach for predicting the backbone structure in terms of protein blocks from sequence. *J Biosci*, 32(1) :51–70, January 2007.

M L de la Paz, E Lacroix, M Ramirez-Alvarado, and L Serrano. Computer-aided design of beta-sheet peptides. *J Mol Biol*, 312(1) :229–46, September 2001.

W L DeLano. The pymol molecular graphics system, 2002. URL `http://www.pymol.org`.

P Derreumaux. Insight into protein topology from monte carlo simulations. *J Chem Phys*, 117(7) :3499–3503, 2002.

P Derreumaux. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *J Chem Phys*, 111 :2301–2310, 1999.

P Derreumaux. Generating ensemble averages for small proteins from extended conformations by monte carlo simulations. *Phys Rev Lett*, 85(1) :206–9, July 2000.

P Derreumaux and N Mousseau. Coarse-grained protein molecular dynamics simulations. *J Chem Phys*, 126 :025101–6, 2006.

S Deshayes, M C Morris, G Divita, and F Heitz. Cell-penetrating peptides : tools for intracellular delivery of therapeutics. *Cell Mol Life Sci*, 62(16) :1839–49, August 2005.

G P H Dietz and M Bahr. Delivery of bioactive molecules into the cell : the trojan horse approach. *Mol Cell Neurosci*, 27(2) :85–131, October 2004.

K A Dill, K M Fiebig, and H S Chan. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci U S A*, 90(5) :1942–6, March 1993.

K A Dill, S B Ozkan, T R Weikl, J D Chodera, and V A Voelz. The protein folding problem : when will it be solved ? *Curr Opin Struct Biol*, 17(3) :342–6, June 2007.

L E Donate, S D Rufino, L H Canard, and T L Blundell. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures : a database for modeling and prediction. *Protein Sci*, 5(12) :2600–16, December 1996.

D Douguet and G Labesse. Easier threading through web-based comparisons and cross-validations. *Bioinformatics*, 17(8) :752–3, August 2001.

H J Dyson and P E Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6(3) :197–208, March 2005.

R C Edgar and K Sjolander. Satchmo : sequence alignment and tree construction using hidden markov models. *Bioinformatics*, 19(11) :1404–11, July 2003.

R J Ellis. Moledular chaperones ten years. introduction. *Semin Cell Dev Biol*, 11(1) :1–5, February 2000.

R J Ellis. Molecular chaperones : assisting assembly in addition to folding. *Trends Biochem Sci*, 31(7) :395–401, July 2006.

D Eramian, M Shen, D Devos, F Melo, A Sali, and M A Marti-Renom. A composite score for predicting errors in protein structure models. *Protein Sci*, 15(7) :1653–66, July 2006.

J Espadaler, N Fernandez-Fuentes, A Hermoso, E Querol, F X Aviles, M J E Sternberg, and B Oliva. Archdb : automated protein loop classification as a tool for structural genomics. *Nucleic Acids Res*, 32(Database issue) :D185–8, January 2004.

C Etchebest, C Benros, S Hazout, and A G de Brevern. A structural alphabet for local protein structures : improved prediction methods. *Proteins*, 59(4) :810–27, June 2005.

F Fant, W F Vranken, and F A Borremans. The three-dimensional solution structure of aesculus hippocastanum antimicrobial protein 1 determined by 1h nuclear magnetic resonance. *Proteins*, 37(3) :388–403, November 1999.

G Favrin, A Irback, and S Mohanty. Oligomerization of amyloid abeta16-22 peptides using hydrogen bonds and hydrophobicity forces. *Biophys J*, 87(6) :3657–64, December 2004.

M Feig, W Im, and C L Brooks-III. Implicit solvation based on generalized born theory in different dielectric environments. *J Chem Phys*, 120(2) :903–11, January 2004.

H J Feldman and C W Hogue. A fast method to sample real protein conformational space. *Proteins*, 39(2) :112–31, May 2000.

N Fernandez-Fuentes, E Querol, F X Aviles, M J E Sternberg, and B Oliva. Prediction of the conformation and geometry of loops in globular proteins : testing archdb, a structural classification of loops. *Proteins*, 60(4) :746–57, September 2005.

N Fernandez-Fuentes, B Oliva, and A Fiser. A supersecondary structure library and search algorithm for modeling loops in protein structures. *Nucleic Acids Res*, 34(7) : 2085–97, 2006a.

N Fernandez-Fuentes, J Zhai, and A Fiser. Archpred : a template based loop structure prediction server. *Nucleic Acids Res*, 34(Web Server issue) :W173–6, July 2006b.

A R Fersht and V Daggett. Protein folding and unfolding at atomic resolution. *Cell*, 108 (4) :573–82, February 2002.

J S Fetrow. Omega loops : nonregular secondary structures significant in protein function and stability. *FASEB J*, 9(9) :708–17, June 1995.

J S Fetrow, M J Palumbo, and G Berg. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins*, 27(2) :249–71, February 1997.

D Fischer. Hybrid fold recognition : combining sequence derived properties with evolutionary information. *Pac Symp Biocomput*, pages 119–30, 2000.

E Fischer. Einfluss der configuration auf die wirkung der enzyme. *Berichte der Deutschen Chemischen Gesellshaft*, 27 :2985–2993, 1894.

A Fiser and A Sali. *Comparative protein structure modeling.*, pages 167–206. Marcel Dekker, Inc., 2003.

C A Floudas. Computational methods in protein structure prediction. *Biotechnol Bioeng*, 97(2) :207–13, June 2007.

F Forcellino and P Derreumaux. Computer simulations aimed at structure prediction of supersecondary motifs in proteins. *Proteins*, 45(2) :159–66, November 2001.

I Friedberg, T Harder, R Kolodny, E Sitbon, Z Li, and A Godzik. Using an alignment of fragment strings for comparing protein structures. *Bioinformatics*, 23(2) :e219–24, January 2007.

D Frishman and P Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4) :566–79, December 1995.

Y Fujitsuka, S Takada, Z A Luthey-Schulten, and P G Wolynes. Optimizing physical energy functions for protein folding. *Proteins*, 54(1) :88–103, January 2004.

K Gan, P Alexander, J M Coxon, J A McKinnon, and G A Worth. The reconstruction of a protein backbone from ca coordinates. *Biopolymers*, 41 :381–389, 1996.

J Garnier, D J Osguthorpe, and B Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, 120(1) :97–120, March 1978.

J Garnier, J F Gibrat, and B Robson. Gor method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266 :540–53, 1996.

R Gautier, A C Camproux, and P Tuffery. Scit : web tools for protein side chain conformation analysis. *Nucleic Acids Res*, 32(Web Server issue) :W508–11, July 2004.

J C Gelly, A G de Brevern, and S Hazout. 'protein peeling' : an approach for splitting a 3d protein structure into compact fragments. *Bioinformatics*, 22(2) :129–33, January 2006a.

J C Gelly, C Etchebest, S Hazout, and A G de Brevern. Protein peeling 2 : a web server to convert protein structures into series of protein units. *Nucleic Acids Res*, 34(Web Server issue) :W75–8, July 2006b.

J F Gibrat, J Garnier, and B Robson. Further developments of protein secondary structure prediction using information theory. new parameters and consideration of residue pairs. *J Mol Biol*, 198(3) :425–43, December 1987.

K Ginalski, A Elofsson, D Fischer, and L Rychlewski. 3d-jury : a simple approach to improve protein structure predictions. *Bioinformatics*, 19(8) :1015–8, May 2003a.

K Ginalski, J Pas, L S Wyrwicz, M von Grotthuss, J M Bujnicki, and L Rychlewski. Orfeus : Detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res*, 31(13) :3804–7, July 2003b.

F Glover. Tabu search : Part i. *ORSA J Comput*, 1 :190–206, 1989.

M Go. Correlation of dna exonic regions with protein structural units in haemoglobin. *Nature*, 291(5810) :90–2, May 1981.

J Greer. Comparative modeling methods : application to the family of the mammalian serine proteases. *Proteins*, 7(4) :317–34, 1990.

D Gront and A Kolinski. Hcpm program for hierarchical clustering of protein models. *Bioinformatics*, 21(14) :3179–80, July 2005.

D Gront, S Kmiecik, and A Kolinski. Backbone building from quadrilaterals : a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J Comput Chem*, 28(9) :1593–7, July 2007.

J R Gunn. Sampling protein conformations using segment libraries and a genetic algorithm. *J Chem Phys*, 106(10) :4270–4281, 1997.

H Guo and D R Salahub. Cooperative hydrogen bonding and enzyme catalysis. *Angew Chem Int Ed Engl*, 37 :2985–2990, 1998.

F Guyon, A C Camproux, J Hochez, and P Tuffery. Sa-search : a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res*, 32(Web Server issue) :W545–8, July 2004.

U H E Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett*, 281 :140–150, 1997.

U H E Hansmann and Y Okamoto. Prediction of peptide conformation by multicanonical algorithm : New approach to the multiple-minima problem. *J Comput Chem*, 14(11) : 1333 – 1338, 1993.

B S Hartley. Homologies in serine proteinases. *Philos Trans R Soc Lond B Biol Sci*, 257 (813) :77–87, February 1970.

N Haspel, C J Tsai, H Wolfson, and R Nussinov. Hierarchical protein folding pathways : a computational study of protein fragments. *Proteins*, 51(2) :203–15, May 2003a.

N Haspel, C J Tsai, H Wolfson, and R Nussinov. Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci*, 12(6) :1177–87, June 2003b.

S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22) :10915–9, November 1992.

D A Hinds and M Levitt. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci U S A*, 89(7) :2536–40, April 1992.

L Holm and C Sander. Database algorithm for generating protein backbone and side-chain co-ordinates from a c alpha trace application to model building and detection of co-ordinate errors. *J Mol Biol*, 218(1) :183–94, March 1991.

L Holm and C Sander. Mapping the protein universe. *Science*, 273(5275) :595–603, August 1996.

C G Hunter and S Subramaniam. Protein local structure prediction from sequence. *Proteins*, 50(4) :572–9, March 2003a.

C G Hunter and S Subramaniam. Protein fragment clustering and canonical local shapes. *Proteins*, 50(4) :580–8, March 2003b.

E G Hutchinson and J M Thornton. A revised set of potentials for beta-turn formation in proteins. *Protein Sci*, 3(12) :2207–16, December 1994.

Y Iwata, A Kasuya, and S Miyamoto. An efficient method for reconstructing protein backbones from alpha-carbon coordinates. *J Mol Graph Model*, 21(2) :119–28, October 2002.

T Jiang, Q Cui, G Shi, and S Ma. Protein folding simulations of the hydrophobic–hydrophilic model by combining tabu search with genetic algorithms. *J Chem Phys*, 119(8) :4592–4596, 2003.

D T Jones. Genthreader : an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, 287(4) :797–815, April 1999a.

D T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, 292(2) :195–202, September 1999b.

D T Jones. Successful ab initio prediction of the tertiary structure of nk-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins*, Suppl 1 :185–91, 1997.

David T Jones and Liam J McGuffin. Assembling novel protein folds from super-secondary structural fragments. *Proteins*, 53 Suppl 6 :480–5, 2003.

T A Jones and S Thirup. Using known substructures in protein model building and crystallography. *EMBO J*, 5(4) :819–22, April 1986.

W L Jorgensen and J Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J Am Chem Soc*, 110 :1657–1666, 1988.

W L Jorgensen, D S Maxwell, and J Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*, 118 :11225–11236, 1996.

K Karplus, C Barrett, M Cline, M Diekhans, L Grate, and R Hughey. Predicting protein structure using only sequence information. *Proteins*, Suppl 3 :121–5, 1999.

K Karplus, R Karchin, J Draper, J Casper, Y Mandel-Gutfreund, M Diekhans, and R Hughey. Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53 Suppl 6 :491–6, 2003.

M Karplus and D L Weaver. Protein folding dynamics : the diffusion-collision model and experimental data. *Protein Sci*, 3(4) :650–68, April 1994.

H Kaur, A Garg, and G P S Raghava. Pepstr : A de novo method for tertiary structure prediction of small bioactive peptides. *Protein Pept Lett*, In Press, 2007.

R Kazmierkiewicz, A Liwo, and H A Scheraga. Energy-based reconstruction of a protein backbone from its alpha-carbon trace by a monte-carlo method. *J Comput Chem*, 23 (7) :715–23, May 2002.

L A Kelley, R M MacCallum, and M J Sternberg. Enhanced genome annotation using structural profiles in the program 3d-pssm. *J Mol Biol*, 299(2) :499–520, June 2000.

D Kihara and J Skolnick. The pdb is a covering set of small protein structures. *J Mol Biol*, 334(4) :793–802, December 2003.

D Kihara, H Lu, A Kolinski, and J Skolnick. Touchstone : an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci U S A*, 98(18) :10125–30, August 2001.

D E Kim, D Chivian, and D Baker. Protein structure prediction and analysis using the robetta server. *Nucleic Acids Res*, 32(Web Server issue) :W526–31, July 2004.

H Kim and H Park. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng*, 16(8) :553–60, August 2003.

S Y Kim, S B Lee, and J Lee. Structure optimization by conformational space annealing in an off-lattice protein model. *Phys Rev E Stat Nonlin Soft Matter Phys*, 72(1 Pt 1) : 011916, July 2005.

L N Kinch, J O Wrabl, S S Krishna, I Majumdar, R I Sadreyev, Y Qi, J Pei, H Cheng, and N V Grishin. Casp5 assessment of fold recognition target predictions. *Proteins*, 53 Suppl 6 :395–409, 2003.

S Kirkpatrick, C D Gelatt, and M P Vecchi. Optimization by simulated annealing. *Science*, 220(4598) :671–680, May 1983.

J L Klepeis and C A Floudas. Analysis and prediction of loop segments in protein structures. *Comp Chem Eng*, 29 :423–436, 2005.

J L Klepeis and C A Floudas. Free energy calculations for peptides via deterministic global optimization. *J Chem Phys*, 110 :7491–7512, 1999.

J L Klepeis and C A Floudas. Ab initio prediction of helical segments in polypeptides. *J Comput Chem*, 23(2) :245–66, January 2002.

J L Klepeis and C A Floudas. Prediction of beta-sheet topology and disulfide bridges in polypeptides. *J Comput Chem*, 24(2) :191–208, January 2003a.

J L Klepeis and C A Floudas. Astro-fold : a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys J*, 85(4) :2119–46, October 2003b.

J L Klepeis, M J Pieja, and C A Floudas. A new class of hybrid global optimization algorithms for peptide structure prediction : integrated hybrids. *Comp Phys Comm*, 151 :121–140, 2003a.

J L Klepeis, M J Pieja, and C A Floudas. Hybrid global optimization algorithms for protein structure prediction : alternating hybrids. *Biophys J*, 84(2 Pt 1) :869–82, February 2003b.

P Koehl and M Delarue. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat Struct Biol*, 2(2) :163–70, February 1995.

P Koehl and M Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239(2) :249–75, June 1994.

P Koehl and M Delarue. Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol*, 6(2) :222–6, April 1996.

A Kolinski. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol*, 51(2) :349–71, 2004.

A Kolinski and J M Bujnicki. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins*, 61 Suppl 7 :84–90, 2005.

A Kolinski and J Skolnick. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. Chapman & Hall, August 1996.

A Kolinski, M R Betancourt, D Kihara, P Rotkiewicz, and J Skolnick. Generalized comparative modeling (genecomp) : a combination of sequence comparison, threading, and

lattice modeling for protein structure prediction and refinement. *Proteins*, 44(2) :133–49, August 2001.

P A Kollman, R Dixon, W Cornell, T Fox, C Chipot, and A Pohorille. *The Development/Application of a 'Minimalist' Organic/Biochemical Molecular Mechanic Force Field using a Combination of ab Initio Calculations and Experimental Data*, volume 3, pages 83–96. Escom, The Netherlands, 1997.

R Kolodny and M Levitt. Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*, 68(3) :278–85, March 2003.

R Kolodny, P Koehl, L Guibas, and M Levitt. Small libraries of protein fragments model native protein structures accurately. *J Mol Biol*, 323(2) :297–307, October 2002.

J Kopp and T Schwede. Automated protein structure homology modeling : a progress report. *Pharmacogenomics*, 5(4) :405–16, June 2004.

T Kortemme, M Ramirez-Alvarado, and L Serrano. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science*, 281(5374) :253–6, July 1998.

J Kosinski, I A Cymerman, M Feder, M A Kurowski, J M Sasin, and J M Bujnicki. A "frankenstein's monster" approach to comparative modeling : merging the finest fragments of fold-recognition models and iterative model refinement aided by 3d structure evaluation. *Proteins*, 53 Suppl 6 :369–79, 2003.

A Krogh, M Brown, I S Mian, K Sjolander, and D Haussler. Hidden markov models in computational biology. applications to protein modeling. *J Mol Biol*, 235(5) :1501–31, February 1994.

N Kurt, T Haliloglu, and C A Schiffer. Structure-based prediction of potential binding and nonbinding peptides to hiv-1 protease. *Biophys J*, 85(2) :853–63, August 2003.

C Kutzner, D van der Spoel, M Fechner, E Lindahl, U W Schmitt, B L de Groot, and H Grubmuller. Speeding up parallel gromacs on high-latency networks. *J Comput Chem*, 28(12) :2075–84, September 2007.

J M Kwasigroch, J Chomilier, and J P Mornon. A global taxonomy of loops in globular proteins. *J Mol Biol*, 259(4) :855–72, June 1996.

M Lamarine, J P Mornon, N Berezovsky, and J Chomilier. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions : towards an efficient punctuation of protein folding ? *Cell Mol Life Sci*, 58(3) :492–8, March 2001.

J Lee and H A Scheraga. Conformational space annealing by parallel computations : Extensive conformational search of met-enkephalin and of the 20-residue membrane-bound portion of melittin. *Intl J of Quantum Chem*, 75 :255–265, 1999.

J Lee, H A Scheraga, and S Rackovsky. New optimization method for conformational energy calculations on polypeptides : Conformational space annealing. *J Comput Chem*, 18 :1222–1232, 1997.

J Lee, H A Scheraga, and S Rackovsky. Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, 46(2) :103–16, August 1998.

J Lee, A Liwo, D R Ripoll, J Pillardy, and H A Scheraga. Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, Suppl 3 :204–8, 1999.

J Lee, J Pillardy, C Czaplewski, Y A Arnautova, D R Ripoll, A Liwo, K D Gibson, R J Wawak, and H A Scheraga. Efficient parallel algorithms in global optimization of potential energy functions. *Comp Phys Comm*, 128 :399–411, 2000.

J Lee, D R Ripoll, C Czaplewski, J Pillardy, W J Wedemeyer, and H A Scheraga. Optimization of parameters in macromolecular potential energy functions by conformational space annealing. *J Phys Chem B*, 105 :2323–2347, 2001.

J Lee, S Y Kim, K Joo, I Kim, and J Lee. Prediction of protein tertiary structure using profesy, a novel method based on fragment assembly and conformational space annealing. *Proteins*, 56(4) :704–14, September 2004.

J Lee, S Y Kim, and J Lee. Protein structure prediction based on fragment assembly and parameter optimization. *Biophys Chem*, 115(2-3) :209–14, April 2005.

S Y Lee and J Skolnick. Development and benchmarking of tasser(iter) for the iterative improvement of protein structure predictions. *Proteins*, 68(1) :39–47, July 2007.

M Leijonmarck and A Liljas. Structure of the c-terminal domain of the ribosomal protein l7/l12 from escherichia coli at 1.7 a. *J Mol Biol*, 195(3) :555–79, June 1987.

A M Lesk and G D Rose. Folding units in globular proteins. *Proc Natl Acad Sci U S A*, 78(7) :4304–8, July 1981.

J F Leszczynski and G D Rose. Loops in globular proteins : a novel category of secondary structure. *Science*, 234(4778) :849–55, November 1986.

C Levinthal. Are there pathways for protein folding ? *Journal de Chimie Physique et de Physico-Chimie Biologique*, 65 :44–45, 1968.

M Levitt. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol*, 226(2) :507–33, July 1992.

M Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1) :59–107, June 1976.

M Levitt and A Warshel. Computer simulation of protein folding. *Nature*, 253(5494) : 694–8, February 1975.

P Lio, N Goldman, J L Thorne, and D T Jones3. Passml : combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14(8) :726–33, 1998.

A Liwo, M R Pincus, R J Wawak, S Rackovsky, and H A Scheraga. Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. *Protein Sci*, 2(10) :1697–714, October 1993.

A Liwo, S Oldziej, M R Pincus, NewAuthor3, S Rackovsky, and H A Scheraga. A united-residue force field for off-lattice protein-structure simulations. i. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem*, 18 :849–873, 1997a.

A Liwo, S Oldziej, M R Pincus, NewAuthor3, S Rackovsky, and H A Scheraga. A united-residue force field for off-lattice protein-structure simulations. ii. parameterization of short-range interactions and determination of weights of energy terms by z-score optimization. *J Comput Chem*, 18 :874–887, 1997b.

A Liwo, C Czaplewski, J Pillardy, and H A Scheraga. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J Chem Phys*, 115 :2323–2347, 2001.

A Liwo, P Arlukowicz, C Czaplewski, S Oldziej, J Pillardy, and H A Scheraga. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape : application to the unres force field. *Proc Natl Acad Sci U S A*, 99(4) :1937–42, February 2002.

C H Lu, Y C Chen, C S Yu, and J K Hwang. Predicting disulfide connectivity patterns. *Proteins*, 67(2) :262–70, May 2007.

H Lu and J Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3) :223–32, August 2001.

R Luthy, J U Bowie, and D Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364) :83–5, March 1992.

M J Macias, V Gervais, C Civera, and H Oschkinat. Structural analysis of ww domains and design of a ww prototype. *Nat Struct Biol*, 7(5) :375–9, May 2000.

A D Mackerell. Empirical force fields for biological macromolecules : overview and issues. *J Comput Chem*, 25(13) :1584–604, October 2004.

A D Mackerell, B Brooks, C L Brooks-III, L Nilsson, B Roux, Y Wong, and M Karplus. *CHARMM : The Energy Function and Its Parameterization with an Overview of the Program*, volume 1, pages 271–277. John Wiley & Sons : Chichester, 1998.

R Malek and N Mousseau. Dynamics of lennard-jones clusters : a characterization of the activation-relaxation technique. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 62(6 Pt A) :7723–7728, 2000.

P Mallick, D R Boutz, D Eisenberg, and T O Yeates. Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *Proc Natl Acad Sci U S A*, 99(15) :9679–84, July 2002.

J Martin, J F Gibrat, and F Rodolphe. Analysis of an optimal hidden markov model for secondary structure prediction. *BMC Struct Biol*, 6 :25, 2006.

J Martin, A G de Brevern, and A C Camproux. In silico local structure approach : a case study on outer membrane proteins. *Proteins*, In Press, 2007.

J Martin, L Regad, H Lecornet, and A C Camproux. Structural deformation upon protein-protein interaction : a structural alphabet approach. *BMC Struct Biol*, Submited.

A M Mathiowetz and W A Goddard. Building proteins from c alpha coordinates using the dihedral probability grid monte carlo method. *Protein Sci*, 4(6) :1217–32, June 1995.

B W Matthews. X-ray crystallographic studies of proteins. *Annu Rev Phys Chem*, 27 : 493–523, 1976.

J Maupetit, R Gautier, and P Tuffery. Sabbac : online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. *Nucleic Acids Res*, 34(Web Server issue) :W147–51, July 2006.

J Maupetit, P Tuffery, and P Derreumaux. A coarse-grained protein force field for folding and structure prediction. *Proteins*, 69(2) :394–408, November 2007.

B J McConkey, V Sobolev, and M Edelman. Discrimination of native protein structures using atom-atom contact scoring. *Proc Natl Acad Sci U S A*, 100(6) :3215–20, March 2003.

L J McGuffin and D T Jones. Improvement of the genthreader method for genomic fold recognition. *Bioinformatics*, 19(7) :874–81, May 2003.

F Melo, R Sanchez, and A Sali. Statistical potentials for fold assessment. *Protein Sci*, 11 (2) :430–48, February 2002.

E Michalsky, A Goede, and R Preissner. Loops in proteins (lip)–a comprehensive loop database for homology modelling. *Protein Eng*, 16(12) :979–85, December 2003.

C Micheletti, F Seno, and A Maritan. Recurrent oligomers in proteins : an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40(4) :662–74, September 2000.

A D Michie, C A Orengo, and J M Thornton. Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol*, 262(2) :168–85, September 1996.

M Milik, A Kolinski, and J Skolnick. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *J Comput Chem*, 18 :80–85, 1997.

E J Milner-White. Recurring loop motif in proteins that occurs in right-handed and left-handed forms. its relationship with alpha-helices and beta-bulge loops. *J Mol Biol*, 199 (3) :503–11, February 1988.

L A Mirny and E I Shakhnovich. How to derive a protein folding potential ? a new approach to an old problem. *J Mol Biol*, 264(5) :1164–79, December 1996.

A E Mirsky and L Pauling. On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci U S A*, 22(7) :439–47, July 1936.

K M S Misura, D Chivian, C A Rohl, D E Kim, and D Baker. Physically realistic homology models built with rosetta can be more accurate than their templates. *Proc Natl Acad Sci U S A*, 103(14) :5361–6, April 2006.

S Miyazawa and R L Jernigan. Estimation of effective inter-residue contact energies from protein crystal structures : quasi-chemical approximation. *Macromolecules*, 18 :534–552, 1985.

M Monnigmann and C A Floudas. Protein loop structure prediction with flexible stem geometries. *Proteins*, 61(4) :748–62, December 2005.

J Moult. Predicting protein three-dimensional structure. *Curr Opin Biotechnol*, 10(6) : 583–8, December 1999.

J Moult, T Hubbard, S H Bryant, K Fidelis, and J T Pedersen. Critical assessment of methods of protein structure prediction (casp) : round ii. *Proteins*, Suppl 1 :2–6, 1997.

J Moult, T Hubbard, K Fidelis, and J T Pedersen. Critical assessment of methods of protein structure prediction (casp) : round iii. *Proteins*, Suppl 3 :2–6, 1999.

J Moult, K Fidelis, A Zemla, and T Hubbard. Critical assessment of methods of protein structure prediction (casp) : round iv. *Proteins*, Suppl 5 :2–7, 2001.

J Moult, K Fidelis, A Zemla, and T Hubbard. Critical assessment of methods of protein structure prediction (casp)-round v. *Proteins*, 53 Suppl 6 :334–9, 2003.

J Moult, K Fidelis, B Rost, T Hubbard, and A Tramontano. Critical assessment of methods of protein structure prediction (casp)–round 6. *Proteins*, 61 Suppl 7 :3–7, 2005.

J Moult, K Fidelis, A Kryshtafovych, B Rost, T Hubbard, and A Tramontano. Critical assessment of methods of protein structure prediction-round vii. *Proteins*, October 2007.

N Mousseau and P Derreumaux. Exploring the early steps of amyloid peptide aggregation by computers. *Acc Chem Res*, 38(11) :885–91, November 2005.

A G Murzin, S E Brenner, T Hubbard, and C Chothia. Scop : a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247 (4) :536–40, April 1995.

M Nanias, M Chinchio, S Oldziej, C Czaplewski, and H A Scheraga. Protein structure prediction with the unres force-field using replica-exchange monte carlo-with-minimization ; comparison with mcm, csa, and cfmc. *J Comput Chem*, 26(14) :1472–86, November 2005.

S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3) :443–53, March 1970.

G Nuel, J Martin, L Regad, and A C Camproux. Pattern statistics in sets of biological sequences. *Algorithms for Molecular Biology*, Submitted, Submited.

B D O'Connor and T O Yeates. Gdap : a web tool for genome-wide protein disulfide bond prediction. *Nucleic Acids Res*, 32(Web Server issue) :W360–4, July 2004.

T Ohlson, B Wallner, and A Elofsson. Profile-profile methods provide improved fold-recognition : a study of different profile-profile alignment methods. *Proteins*, 57(1) : 188–97, October 2004.

S Oldziej, C Czaplewski, A Liwo, M Chinchio, M Nanias, J A Vila, M Khalili, Y A Arnautova, A Jagielska, M Makowski, H D Schafroth, R Kazmierkiewicz, D R Ripoll, J Pillardy, J A Saunders, Y K Kang, K D Gibson, and H A Scheraga. Physics-based

protein-structure prediction using a hierarchical protocol based on the unres force field : assessment in two blind tests. *Proc Natl Acad Sci U S A*, 102(21) :7547–52, May 2005.

B Oliva, P A Bates, E Querol, F X Aviles, and M J Sternberg. An automated classification of the structure of protein loops. *J Mol Biol*, 266(4) :814–30, March 1997.

C A Orengo and W R Taylor. A local alignment method for protein structure motifs. *J Mol Biol*, 233(3) :488–97, October 1993.

C A Orengo, A D Michie, S Jones, D T Jones, M B Swindells, and J M Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8) :1093–108, August 1997.

S B Ozkan, G A Wu, J D Chodera, and K A Dill. Protein folding by zipping and assembly. *Proc Natl Acad Sci U S A*, 104(29) :11987–92, July 2007.

N Papandreou, I N Berezovsky, A Lopes, E Eliopoulos, and J Chomilier. Universal positions in globular proteins. *Eur J Biochem*, 271(23-24) :4762–8, December 2004.

B Park and M Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2) :367–92, May 1996.

B H Park and M Levitt. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, 249(2) :493–507, June 1995.

L Pauling and R B Corey. Configuration of polypeptide chains. *Nature*, 168(4274) :550–1, September 1951.

V Pavone, G Gaeta, A Lombardi, F Nastri, O Maglio, C Isernia, and M Saviano. Discovering protein secondary structures : classification and description of isolated alpha-turns. *Biopolymers*, 38(6) :705–21, June 1996.

P W Payne. Reconstruction of protein conformations from estimated positions of the c alpha coordinates. *Protein Sci*, 2(3) :315–24, March 1993.

F Pearl, A Todd, I Sillitoe, M Dibley, O Redfern, T Lewis, C Bennett, R Marsden, A Grant, D Lee, A Akpor, M Maibaum, A Harrison, T Dallman, G Reeves, I Diboun, S Addou, S Lise, C Johnston, A Sillero, J Thornton, and C Orengo. The cath domain structure database and related resources gene3d and dhs provide comprehensive domain family information for genome analysis. *Nucleic Acids Res*, 33(Database issue) :D247–51, January 2005.

F M G Pearl, C F Bennett, J E Bray, A P Harrison, N Martin, A Shepherd, I Sillitoe, J Thornton, and C A Orengo. The cath database : an extended protein family resource for structural and functional genomics. *Nucleic Acids Res*, 31(1) :452–5, January 2003.

J Pei, R Sadreyev, and N V Grishin. Pcma : fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, 19(3) :427–8, February 2003.

D Petrey, Z Xiang, C L Tang, L Xie, M Gimpelev, T Mitros, C S Soto, S G Fischman, A Kernytsky, A Schlessinger, I Y Y Koh, E Alexov, and B Honig. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, 53 Suppl 6 :430–5, 2003.

U Pieper, N Eswar, F P Davis, H Braberg, M S Madhusudhan, A Rossi, M Marti-Renom, R Karchin, B M Webb, D Eramian, M Y Shen, L Kelly, F Melo, and A Sali. Modbase : a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res*, 34(Database issue) :D291–5, January 2006.

J Pillardy, C Czaplewski, A Liwo, W J Wedemeyer, J Lee, D R Ripoll, P Arlukowicz, S Oldziej, Y A Arnautova, and H A Scheraga. Development of physics-based energy functions that predict medium-resolution structures for proteins of the alpha, beta and alpha/beta structural classes. *J Phys Chem B*, 105 :7299–7311, 2001.

F Plewniak, L Bianchetti, Y Brelivet, A Carles, F Chalmel, O Lecompte, T Mochel, L Moulinier, A Muller, J Muller, V Prigent, R Ripp, J C Thierry, J D Thompson, N Wicker, and O Poch. Pipealign : A new toolkit for protein family analysis. *Nucleic Acids Res*, 31(13) :3829–32, July 2003.

N Pokala and T M Handel. Energy functions for protein design : adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J Mol Biol*, 347(1) :203–27, March 2005.

G Pollastri, D Przybylski, B Rost, and P Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, 47(2) :228–35, May 2002.

A Poupon and J P Mornon. Predicting the protein folding nucleus from sequences [correction of a sequence]. *FEBS Lett*, 452(3) :283–9, June 1999.

A Poupon and J P Mornon. Populations of hydrophobic amino acids within protein globular domains : identification of conserved “topohydrophobic” positions. *Proteins*, 33(3) :329–42, November 1998.

S J Prestrelski, A L Williams, and M N Liebman. Generation of a substructure library for the description and classification of protein secondary structure. i. overview of the methods and results. *Proteins*, 14(4) :430–9, December 1992.

D Przybylski and B Rost. Improving fold recognition without folds. *J Mol Biol*, 341(1) : 255–69, July 2004.

E O Purisima and H A Scheraga. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers*, 23(7) :1207–24, July 1984.

J Qiu and R Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins*, 61(1) :44–55, October 2005.

L R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc of the IEEE*, 77 :57–285, 1989.

K R Rajashankar and S Ramakumar. Pi-turns in proteins and peptides : Classification, conformation, occurrence, hydration and sequence. *Protein Sci*, 5(5) :932–46, May 1996.

L Regad, F Guyon, J Maupetit, P Tuffery, and A C Camproux. A hidden markov model applied to the protein 3d structure analysis. *Computational Statistics Data and Analysis*, In press.

L S Reid and J M Thornton. Rebuilding flavodoxin from c alpha coordinates : a test study. *Proteins*, 5(2) :170–82, 1989.

D Reith, M Putz, and F Muller-Plathe. Deriving effective mesoscale potentials from atomistic simulations. *J Comput Chem*, 24(13) :1624–36, October 2003.

A Rey and J Skolnick. Efficient algorithm for the reconstruction of a protein backbone from the a-carbon coordinates. *J Comput Chem*, 13 :443–456, 1992.

J S Richardson. The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34 :167–339, 1981.

J S Richardson and D C Richardson. Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240 :1648–52, 1988.

J S Richardson, E D Getzoff, and D C Richardson. The beta bulge : a common small unit of nonrepetitive protein structure. *Proc Natl Acad Sci U S A*, 75(6) :2574–8, June 1978.

C S Ring, D G Kneller, R Langridge, and F E Cohen. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol*, 224(3) :685–99, April 1992.

C A Rohl, C E M Strauss, K M S Misura, and D Baker. Protein structure prediction using rosetta. *Methods Enzymol*, 383 :66–93, 2004.

M J Rooman, J Rodriguez, and S J Wodak. Automatic definition of recurrent local structure motifs in proteins. *J Mol Biol*, 213(2) :327–36, May 1990.

G D Rose, L Gierasch, and J A Smith. *Turns in Peptides and Proteins.*, volume 37 of *Advances in Protein Chemistry*. Academic Press, New York, 1985.

B Rost and C Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19(1) :55–72, May 1994.

B Rupp. *High Throughput Protein Crystallography.*, chapter 5, pages 61–104. CRC Press, July 2005. ISBN 0824753356.

L Rychlewski, L Jaroszewski, W Li, and A Godzik. Comparison of sequence profiles. strategies for structural predictions using sequence information. *Protein Sci*, 9(2) : 232–41, February 2000.

R Sadreyev and N Grishin. Compass : a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol*, 326(1) :317–36, February 2003.

A Sali and T L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3) :779–815, December 1993.

R Samudrala and M Levitt. Decoys 'r' us : a database of incorrect conformations to improve protein structure prediction. *Protein Sci*, 9(7) :1399–401, July 2000.

R Samudrala and M Levitt. A comprehensive analysis of 40 blind protein structure predictions. *BMC Struct Biol*, 2 :3, August 2002.

R Sanchez and A Sali. Advances in comparative protein-structure modelling. *Curr Opin Struct Biol*, 7(2) :206–14, April 1997.

O Sander, I Sommer, and T Lengauer. Local protein structure prediction using discriminative models. *BMC Bioinformatics*, 7 :14, 2006.

S Santini, G H Wei, N Mousseau, and P Derreumaux. Exploring the folding pathways of proteins through energy landscape sampling : Application to alzheimer's beta-amyloid peptide. *Internet Electron J Mol Des*, 2 :564–577, 2003.

S Santini, N Mousseau, and P Derreumaux. In silico assembly of alzheimer's abeta16-22 peptide into beta-sheets. *J Am Chem Soc*, 126(37) :11509–16, September 2004a.

S Santini, G H Wei, N Mousseau, and P Derreumaux. Pathway complexity of alzheimer's beta-amyloid abeta16-22 peptide assembly. *Structure*, 12(7) :1245–55, July 2004b.

J Schuchhardt, G Schneider, J Reichelt, D Schomburg, and P Wrede. Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng*, 9(10) :833–42, October 1996.

G E Schulz. Structural rules for globular proteins. *Angew Chem Int Ed Engl*, 16 :23–32, 1977.

G Schwartz. Estimating the dimension of a model. *Annals of Statistics*, 6 :461–464, 1978.

T Schwede, J Kopp, N Guex, and M C Peitsch. Swiss-model : An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13) :3381–5, July 2003.

B R Seavey, E A Farr, W M Westler, and J L Markley. A relational database for sequence-specific protein nmr data. *J Biomol NMR*, 1(3) :217–36, September 1991.

M Shen and A Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 15(11) :2507–24, November 2006.

J Shi, T L Blundell, and K Mizuguchi. Fugue : sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, 310(1) :243–57, June 2001.

A Y Shih, A Arkhipov, P L Freddolino, and K Schulten. Coarse grained protein-lipid model with application to lipoprotein particles. *J Phys Chem B*, 110(8) :3674–84, March 2006.

S Shimizu and H S Chan. Anti-cooperativity and cooperativity in hydrophobic interactions : Three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. *Proteins*, 48(1) :15–30, July 2002.

B L Sibanda and J M Thornton. Conformation of beta hairpins in protein structures : classification and diversity in homologous structures. *Methods Enzymol*, 202 :59–82, 1991.

N Siew, A Elofsson, L Rychlewski, and D Fischer. Maxsub : an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, 16(9) :776–85, September 2000.

N Siew, Y Azaria, and D Fischer. The orfanage : an orfan database. *Nucleic Acids Res*, 32(Database issue) :D281–3, January 2004.

K T Simons, C Kooperberg, E Huang, and D Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268(1) :209–25, April 1997.

K T Simons, R Bonneau, I Ruczinski, and D Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins*, Suppl 3 :171–6, 1999.

G E Sims and S H Kim. A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring. *Proc Natl Acad Sci U S A*, 103(12) :4428–32, March 2006.

M J Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4) :859–83, June 1990.

J Skolnick. In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol*, 16(2) :166–71, April 2006.

J Skolnick and A Kolinski. Dynamic monte carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol*, 221(2) :499–531, September 1991.

J Skolnick, L Jaroszewski, A Kolinski, and A Godzik. Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct ? *Protein Sci*, 6(3) :676–88, March 1997.

J Skolnick, D Kihara, and Y Zhang. Development and large scale benchmark testing of the prospector 3 threading algorithm. *Proteins*, 56(3) :502–18, August 2004.

T F Smith and M S Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1) :195–7, March 1981.

T F Smith, M S Waterman, and W M Fitch. Comparative biosequence metrics. *J Mol Evol*, 18(1) :38–46, 1981.

P Soto and G Colombo. Characterization of the conformational space of a triple-stranded beta-sheet forming peptide with molecular dynamics simulations. *Proteins*, 57(4) : 734–46, December 2004.

R Sowdhamini and T L Blundell. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci*, 4(3) : 506–20, March 1995.

D Van Der Spoel, E Lindahl, B Hess, G Groenhof, A E Mark, and H J C Berendsen. Gromacs : fast, flexible, and free. *J Comput Chem*, 26(16) :1701–18, December 2005.

R Srinivasan and G D Rose. Ab initio prediction of protein structure using linus. *Proteins*, 47(4) :489–95, June 2002.

R Srinivasan and G D Rose. Linus : a hierarchic procedure to predict the fold of a protein. *Proteins*, 22(2) :81–99, June 1995.

R Srinivasan, P J Fleming, and G D Rose. Ab initio protein folding using linus. *Methods Enzymol*, 383 :48–66, 2004.

J K Sun and A J Doig. Addition of side-chain interactions to 3(10)-helix/coil and alpha-helix/3(10)-helix/coil theory. *Protein Sci*, 7(11) :2374–83, November 1998.

C L Tang, L Xie, I Y Y Koh, S Posy, E Alexov, and B Honig. On the role of structural information in remote homology detection and sequence alignment : new methods using hybrid sequence profiles. *J Mol Biol*, 334(5) :1043–62, December 2003.

A Thomas, S Deshayes, M Decaffmeyer, M H Van Eyck, B Charloteaux, and R Brasseur. Prediction of peptide structure : how far are we ? *Proteins*, 65(4) :889–97, December 2006.

J L Thorne, N Goldman, and D T Jones. Combining protein evolution and secondary structure. *Mol Biol Evol*, 13(5) :666–73, May 1996.

V Tozzini. Coarse-grained models for proteins. *Curr Opin Struct Biol*, 15(2) :144–50, April 2005.

A Tramontano and V Morea. Assessment of homology-based predictions in casp5. *Proteins*, 53 Suppl 6 :352–68, 2003.

R Tréhin and H P Merkle. Chances and pitfalls of cell penetrating peptides for cellular drug delivery. *Eur J Pharm Biopharm*, 58(2) :209–23, September 2004.

C J Tsai and R Nussinov. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci*, 6(1) :24–42, January 1997a.

C J Tsai and R Nussinov. Hydrophobic folding units at protein-protein interfaces : implications to protein folding and to protein-protein association. *Protein Sci*, 6(7) :1426–37, July 1997b.

C J Tsai, S Kumar, B Ma, and R Nussinov. Folding funnels, binding funnels, and protein function. *Protein Sci*, 8(6) :1181–90, June 1999.

C J Tsai, J V Maizel, and R Nussinov. Anatomy of protein structures : visualizing how a one-dimensional protein chain folds into a three-dimensional shape. *Proc Natl Acad Sci U S A*, 97(22) :12038–43, October 2000.

C J Tsai, P P de Laureto, A Fontana, and R Nussinov. Comparison of protein fragments identified by limited proteolysis and by computational cutting of proteins. *Protein Sci*, 11(7) :1753–70, July 2002.

J Tsai, R Bonneau, A V Morozov, B Kuhlman, C A Rohl, and D Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 53(1) :76–87, October 2003.

V Tsui and D A Case. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers*, 56(4) :275–91, 2000.

P Tuffery. Xmmol : an x11 and motif program for macromolecular visualization and modeling. *J Mol Graph*, 13(1) :67–72, 62, February 1995.

P Tuffery and P Derreumaux. Dependency between consecutive local conformations helps assemble protein structures from secondary structures using go potential and greedy algorithm. *Proteins*, 61(4) :732–40, December 2005.

P Tuffery, F Guyon, and P Derreumaux. Improved greedy algorithm for protein structure reconstruction. *J Comput Chem*, 26(5) :506–13, April 2005.

M Tyagi, P Sharma, C S Swamy, F Cadet, N Srinivasan, A G de Brevern, and B Offmann. Protein block expert (pbe) : a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res*, 34(Web Server issue) :W119–23, July 2006.

Y Udea, H Taketomi, and N Go. Studies of protein folding, unfolding, and fluctuations by computer simulation. ii. 3-dimensional lattice model for lysozyme. *Biopolymers*, 17 : 1531–1548, 1978.

R Unger and J L Sussman. The importance of short structural motifs in protein structure analysis. *J Comput Aided Mol Des*, 7(4) :457–72, August 1993.

R Unger, D Harel, S Wherland, and J L Sussman. A 3d building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4) :355–73, 1989.

B Urbanc, L Cruz, F Ding, D Sammond, S Khare, S V Buldyrev, H E Stanley, and N V Dokholyan. Molecular dynamics simulation of amyloid beta dimer formation. *Biophys J*, 87(4) :2310–21, October 2004.

C W van Gelder, F J Leusen, J A Leunissen, and J H Noordik. A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. *Proteins*, 18(2) :174–85, February 1994.

P A van Hooft and H D Holtje. Construction of a full three-dimensional model of the transpeptidase domain of streptococcus pneumoniae pbp2x starting from its calpha-atom coordinates. *J Comput Aided Mol Des*, 14(8) :719–30, November 2000.

C Venclovas, A Zemla, K Fidelis, and J Moult. Assessment of progress over the casp experiments. *Proteins*, 53 Suppl 6 :585–95, 2003.

M Vendruscolo, E Kussell, and E Domany. Recovery of protein structure from contact maps. *Fold Des*, 2(5) :295–306, 1997.

J J Vincent, C H Tai, B K Sathyanarayana, and B Lee. Assessment of casp6 predictions for new and nearly new fold targets. *Proteins*, 61 Suppl 7 :67–83, 2005.

D Vitkup, E Melamud, J Moult, and C Sander. Completeness in structural genomics. *Nat Struct Biol*, 8(6) :559–66, June 2001.

G Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *Journal fur die Reine und Angewandte Mathematik*, 133 :97–178, 1907.

B Wallner and A Elofsson. All are not equal : a benchmark of different homology modeling programs. *Protein Sci*, 14(5) :1315–27, May 2005.

A Wallqvist and M Ullner. A simplified amino acid potential for use in structure predictions of proteins. *Proteins*, 18(3) :267–80, March 1994.

G Wang and R L Dunbrack. Pisces : a protein sequence culling server. *Bioinformatics*, 19(12) :1589–91, August 2003.

G Wang, Y Jin, and R L Dunbrack. Assessment of fold recognition predictions in casp6. *Proteins*, 61 Suppl 7 :46–66, 2005.

J J Ward, L J McGuffin, B F Buxton, and D T Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13) :1650–5, September 2003.

G H Wei, N Mousseau, and P Derreumaux. Exploring the energy landscape of proteins : a characterization of the activation relaxation technique. *J Chem Phys*, 117 :11379–11387, 2002.

G H Wei, P Derreumaux, and N Mousseau. Sampling the complex energy landscape of a simple beta-hairpin. *J Chem Phys*, 119 :6403–6406, 2003.

G H Wei, N Mousseau, and P Derreumaux. Complex folding pathways in a simple beta-hairpin. *Proteins*, 56(3) :464–74, August 2004.

G H Wei, N Mousseau, and P Derreumaux. Computational simulation of the early steps of protein agregation. *Prion J*, 2007.

D B Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, 70(3) :697–701, March 1973.

D Willbold, S Hoffmann, and P Rosch. Secondary structure and tertiary fold of the human immunodeficiency virus protein u (vpu) cytoplasmic domain in solution. *Eur J Biochem*, 245(3) :581–8, May 1997.

C M Wilmot and J M Thornton. Analysis and prediction of the different types of beta-turn in proteins. *J Mol Biol*, 203(1) :221–32, September 1988.

R T Wintjens, M J Rooman, and S J Wodak. Automatic classification and analysis of alpha alpha-turn motifs in proteins. *J Mol Biol*, 255(1) :235–53, January 1996.

S J Wodak and J Janin. Location of structural domains in protein. *Biochemistry*, 20(23) : 6544–52, November 1981.

J Wojcik, J P Mornon, and J Chomilier. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J Mol Biol*, 289(5) :1469–90, June 1999.

S Wu, J Skolnick, and Y Zhang. Ab initio modeling of small proteins by iterative tasser simulations. *BMC Biol*, 5 :17, 2007.

Y Xia, E S Huang, M Levitt, and R Samudrala. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol*, 300(1) :171–85, June 2000.

J Xu, M Li, D Kim, and Y Xu. Raptor : optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1(1) :95–117, April 2003.

Y Xu and D Xu. Protein threading using prospect : design and evaluation. *Proteins*, 40 (3) :343–54, August 2000.

A S Yang and L Y Wang. Local structure prediction with local structure-based sequence profiles. *Bioinformatics*, 19(10) :1267–74, July 2003.

J M Yang and C H Tung. Protein structure database search and evolutionary classification. *Nucleic Acids Res*, 34(13) :3646–59, 2006.

L Yang, C H Tan, M J Hsieh, J Wang, Y Duan, P Cieplak, J Caldwell, P A Kollman, and R Luo. New-generation amber united-atom force field. *J Phys Chem B*, 110(26) : 13166–76, July 2006.

G Yona and M Levitt. Within the twilight zone : a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5) :1257–75, February 2002.

K Yue and K A Dill. Constraint-based assembly of tertiary protein structures from secondary structure elements. *Protein Sci*, 9(10) :1935–46, October 2000.

M Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*, 12(6) :1271–82, June 2003.

B Zagrovic, E J Sorin, and V Pande. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J Mol Biol*, 313(1) :151–69, October 2001.

M H Zehfus and G D Rose. Compact units in proteins. *Biochemistry*, 25(19) :5759–65, September 1986.

C Zhang, S Liu, H Zhou, and Y Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Sci*, 13(2) :400–11, February 2004.

X Zhang, J S Fetrow, W A Rennie, D L Waltz, and G Berg. Automatic derivation of substructures yields novel structural building blocks in globular proteins. *Proc Int Conf Intell Syst Mol Biol*, 1 :438–46, 1993.

Y Zhang and J Skolnick. Spicker : a clustering approach to identify near-native protein folds. *J Comput Chem*, 25(6) :865–71, April 2004a.

Y Zhang and J Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4) :702–10, December 2004b.

Y Zhang and J Skolnick. The protein structure prediction problem could be solved using the current pdb library. *Proc Natl Acad Sci U S A*, 102(4) :1029–34, January 2005a.

Y Zhang and J Skolnick. Tm-align : a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Res*, 33(7) :2302–9, 2005b.

Y Zhang, A Kolinski, and J Skolnick. Touchstone ii : a new approach to ab initio protein structure prediction. *Biophys J*, 85(2) :1145–64, August 2003.

Y Zhang, A K Arakaki, and J Skolnick. Tasser : an automated method for the prediction of protein tertiary structures in casp6. *Proteins*, 61 Suppl 7 :91–8, 2005.

H Zhou and J Skolnick. Ab initio protein structure prediction using chunk-tasser. *Biophys J*, 93(5) :1510–8, September 2007.

H Zhou and Y Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*, 11(11) :2714–26, November 2002.

H Zhou and Y Zhou. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins*, 58(2) : 321–8, February 2005.

H Zhou, S B Pandit, S Y Lee, J Borreguero, H Chen, L Wroblewska, and J Skolnick. Analysis of tasser-based casp7 protein structure prediction results. *Proteins*, August 2007.

B H Zimm and J K Bragg. Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys*, 31 :526–531, 1959.

M Zorko and U Langel. Cell-penetrating peptides : mechanism and kinetics of cargo delivery. *Adv Drug Deliv Rev*, 57(4) :529–45, February 2005.

M Zuker and R L Somorjai. The alignment of protein structures in three dimensions. *Bull Math Biol*, 51(1) :55–78, 1989.

Septième partie

ANNEXES

Annexe A

Le champ de force OPEP

A.1 Les paramètres d'OPEP 3.1.

$w_{hb1-4} * \varepsilon_{hb1-4}$	$w_{hb1>4} * \varepsilon_{hb1>4}$	$w_{\alpha}^{coop} * E_{\alpha}^{coop}$	$w_{\beta}^{coop} * E_{\beta}^{coop}$
1.126	0.822	-0.625	-2.416

Tab. A.1: OPEP 3.1 Paramètres des liaisons hydrogène.
Voir la section 5.2 et les équations 5.7, 5.11 et 5.12.

-	$w_\alpha^R * E_\alpha^R$	$w_\beta^R * E_\beta^R$
CYS	0.18	0.17
LEU	-0.28	0.17
VAL	0.11	-0.35
ILE	0.11	-0.23
MET	-0.11	0.33
PHE	0.12	-0.32
TYR	0.14	-0.35
LYS	-0.18	0.34
ARG	-0.09	0.33
PRO	0.46	0.33
GLY	0.44	0.25
ALA	-0.07	0.27
GLN	0.06	0.28
HIS	0.25	0.29
ASN	0.13	0.38
ASP	0.15	0.28
GLU	-0.22	0.33
SER	0.08	0.26
THR	0.23	-0.39
TRP	0.09	0.32

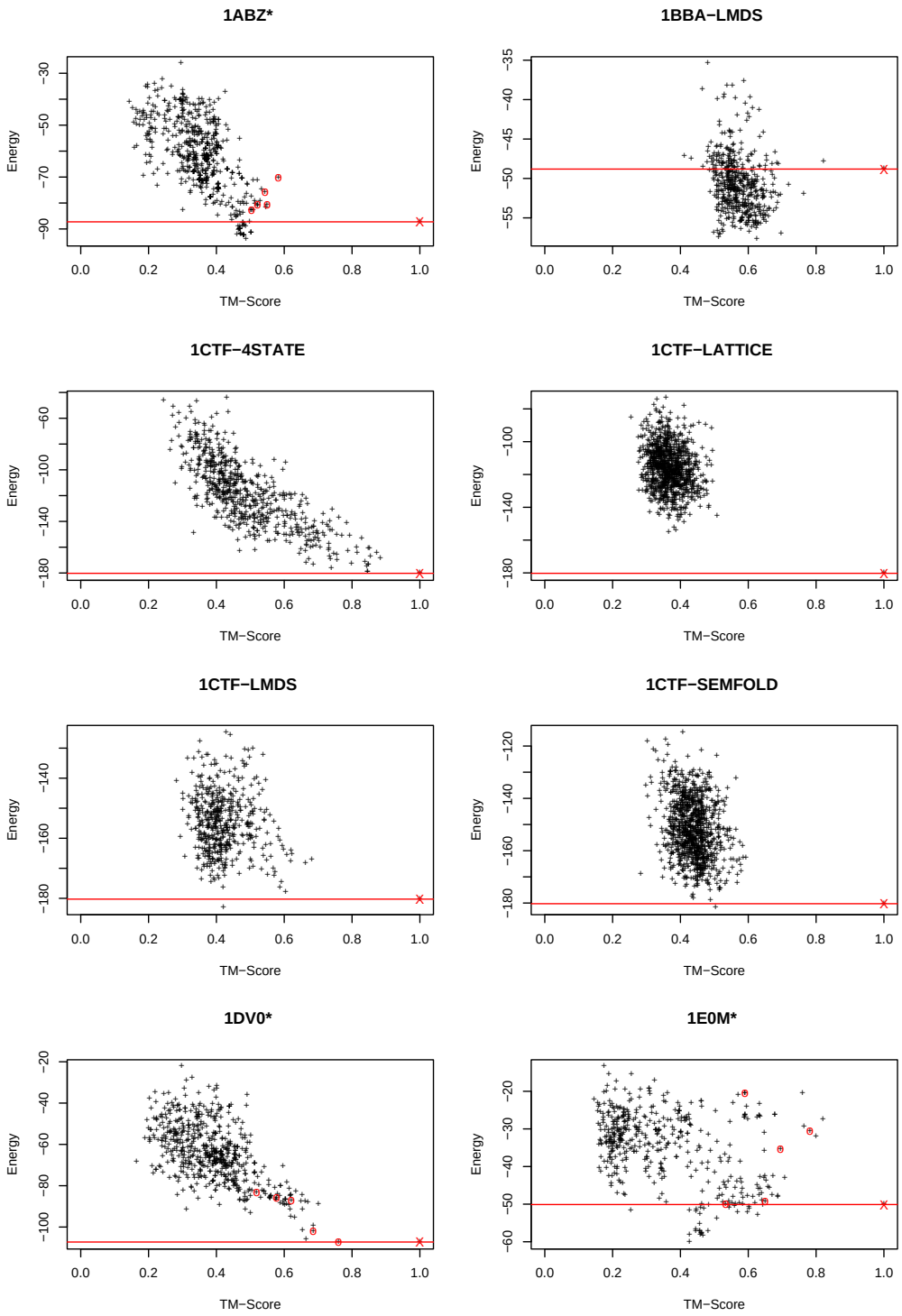
Tab. A.2: OPEP 3.1 propensités des hélices α et feuillets β.
Voir la section 5.2 et les équations 5.11 et 5.12.

-	Cys	Phe	Leu	Trp	Val	Ile	Met	His	Tyr	Ala	Gly	Pro	Asn	Thr	Ser	Arg	Gln	Asp	Lys	Glu
Cys	1.90	0.59	0.74	1.13	0.68	0.70	0.70	0.26	0.22	0.39	0.13	0.24	-0.39	-0.01	-0.14	-0.39	-0.06	-0.55	-0.49	-0.68
Phe		1.16	1.22	0.95	1.05	1.02	0.83	0.27	0.71	0.45	-0.16	0.29	-0.39	-0.01	-0.15	-0.12	0.06	-0.64	-0.15	-0.52
Leu			1.31	0.94	1.08	1.23	0.83	-0.15	0.59	0.42	-0.22	0.11	-0.48	-0.01	-0.37	-0.08	-0.11	-0.87	-0.23	-0.45
Trp				0.96	0.84	0.83	1.27	0.67	0.81	0.55	0.31	1.10	0.13	-0.02	-0.10	0.61	0.15	-0.09	0.33	0.20
Val					1.07	1.04	0.64	-0.26	0.42	0.55	-0.06	0.12	-0.59	-0.01	-0.33	-0.27	-0.24	-0.98	-0.23	-0.54
Ile						0.95	0.76	-0.25	0.43	0.52	-0.29	-0.08	-0.81	-0.01	-0.50	-0.25	-0.21	-0.69	-0.27	-0.59
Met							0.63	0.23	0.48	0.33	-0.12	0.22	-0.44	-0.02	-0.47	-0.22	0.01	-0.85	-0.32	-0.30
His								0.46	0.28	-0.33	-0.34	0.08	-0.12	-0.01	-0.20	-0.06	-0.31	0.33	-0.39	0.14
Tyr									0.43	0.20	0.05	0.51	-0.02	-0.01	-0.09	0.46	0.25	0.09	0.50	0.22
Ala										0.30	0.05	-0.10	-0.35	-0.01	-0.23	-0.35	-0.32	-0.48	-0.25	-0.59
Gly											0.27	0.01	-0.14	-0.02	-0.15	-0.19	-0.28	-0.26	-0.17	-0.70
Pro												0.10	-0.17	-0.01	-0.22	0.03	0.06	-0.37	-0.15	-0.39
Asn													0.05	-0.01	-0.20	-0.03	0.07	0.16	0.16	0.01
Thr														-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
Ser															-0.18	-0.17	-0.37	-0.01	-0.14	-0.14
Arg																-0.18	0.18	0.93	-0.75	1.17
Gln																	-0.19	-0.17	0.24	-0.14
Asp																		-0.41	0.63	-0.59
Lys																			-0.61	0.83
Glu																				-0.66

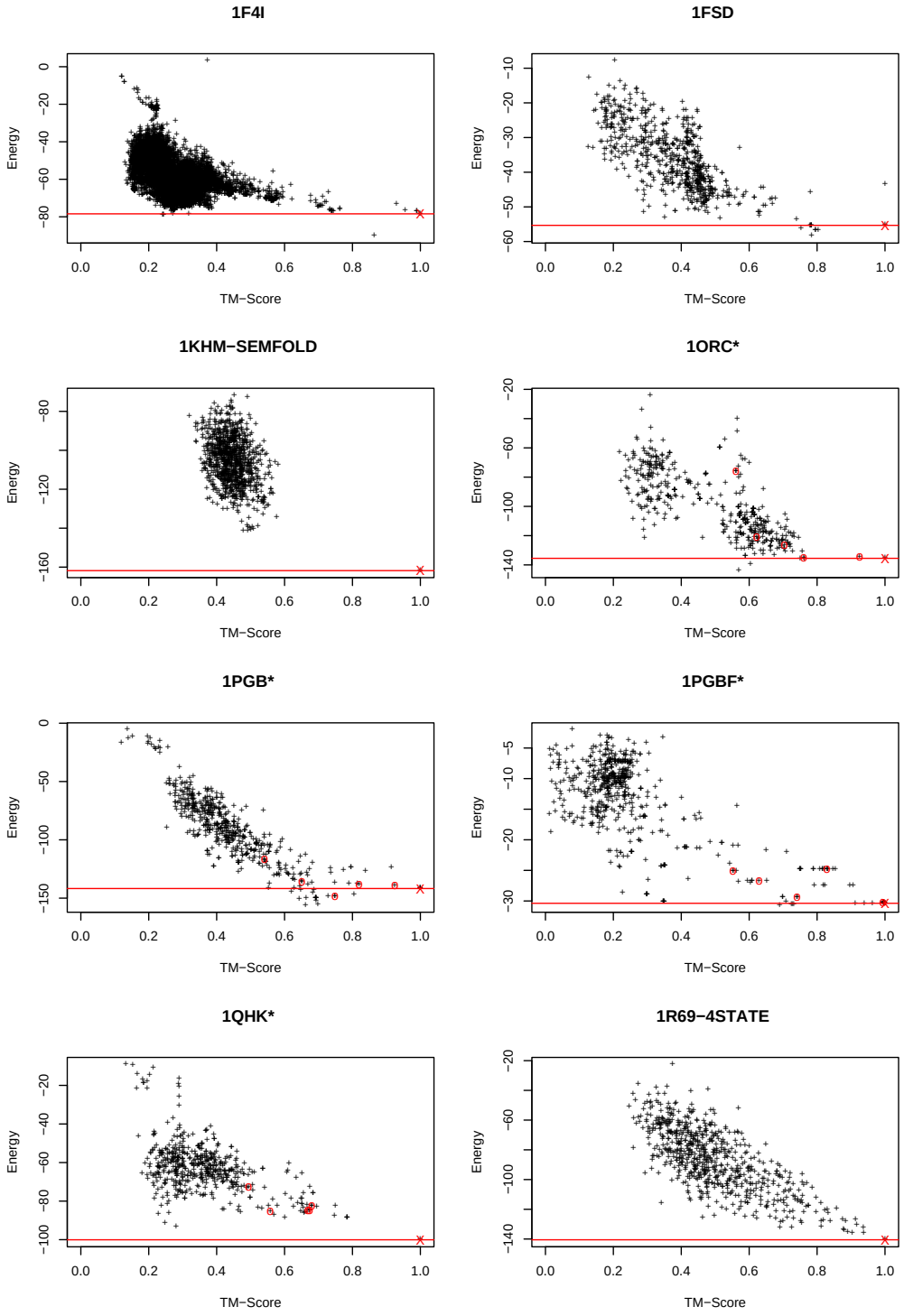
Tab. A.3: OPEP 3.1 : les paramètres de interactions CL-CL

Les valeurs suivantes correspondent au produit des poids $w_{CL,CL}$ obtenus lors du processus d'optimisation, et les valeurs initiales de la matrice de contact de Betancourt et Thirumalai (Betancourt and Thirumalai, 1999). Voir la section 5.2.

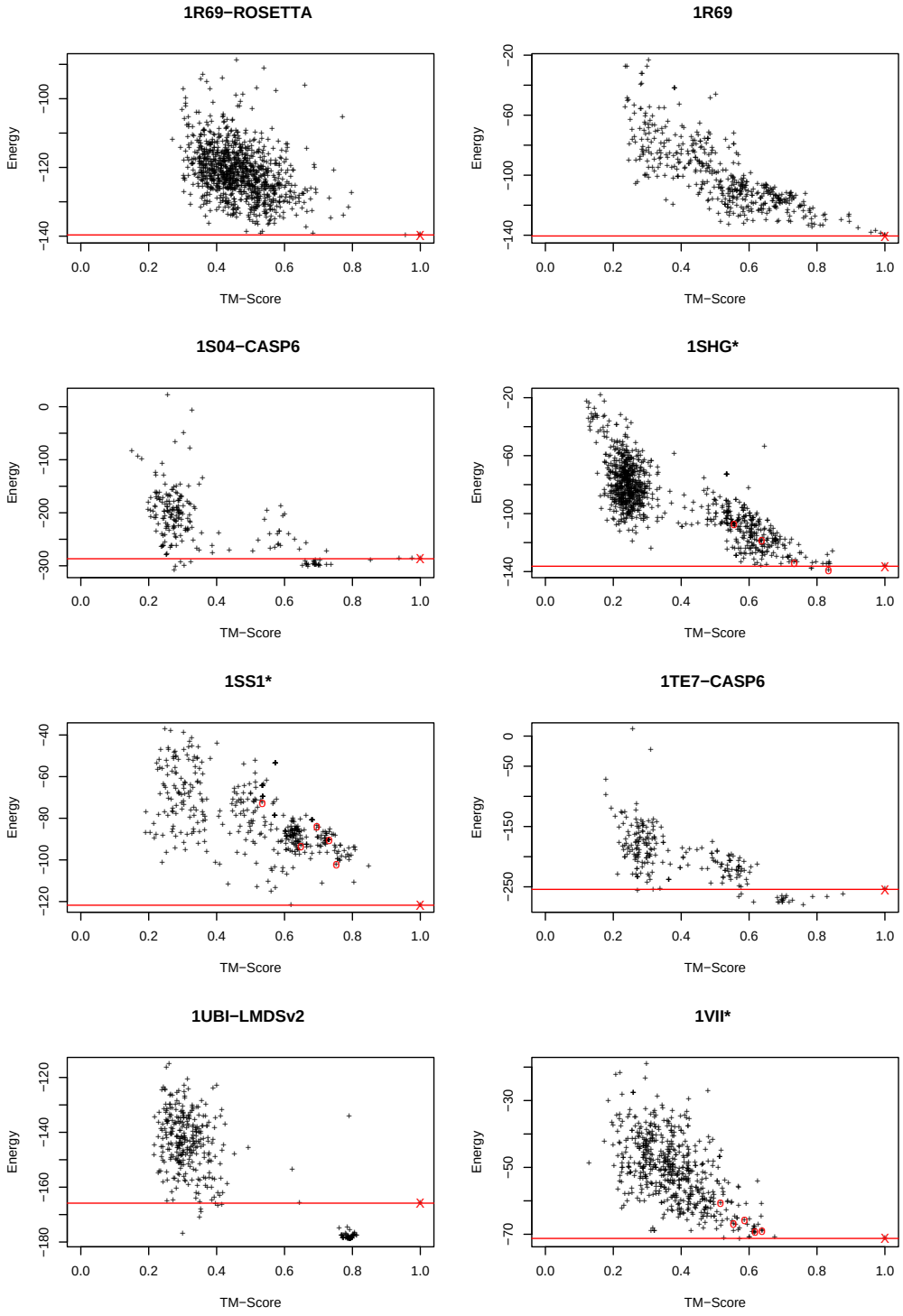
A.2 Le pouvoir discriminant d'OPEP 3.1



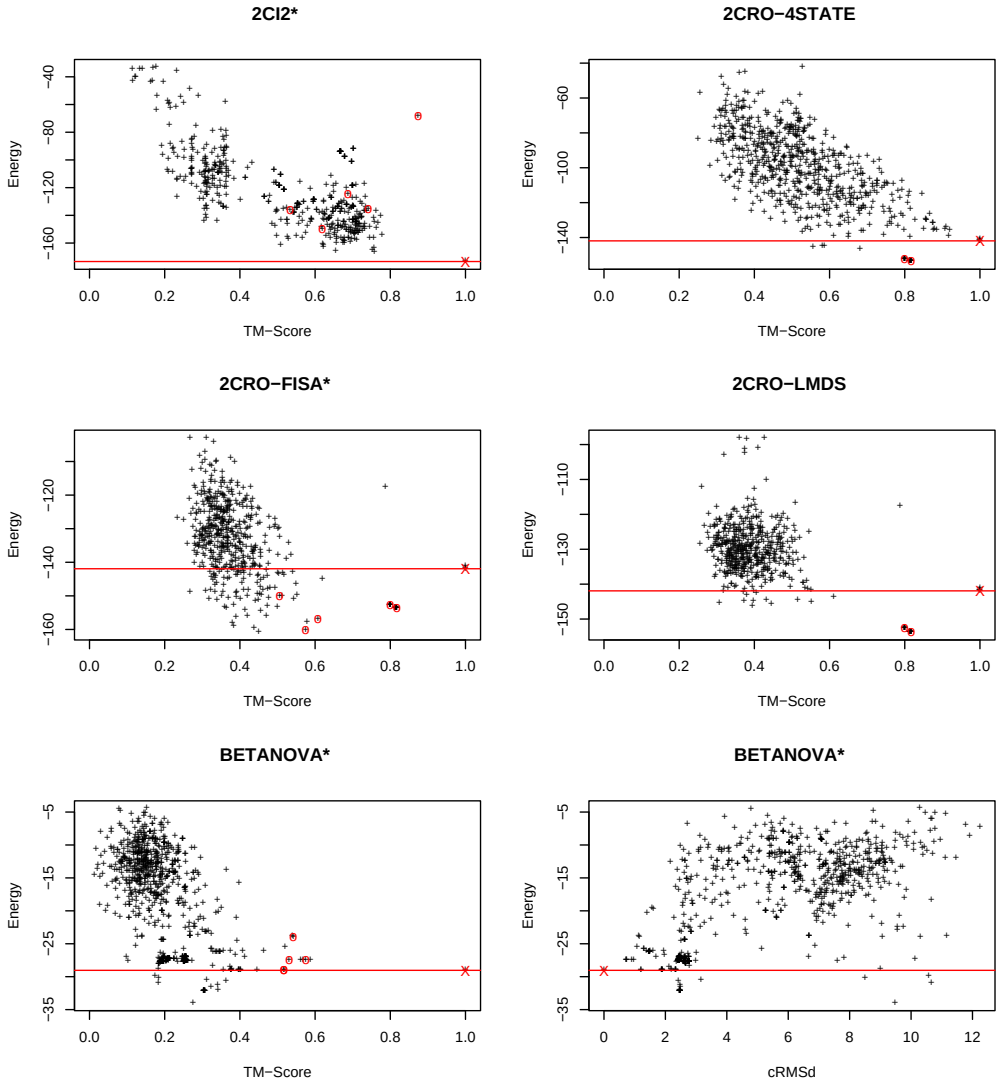
Tab. A.4: OPEP 3.1 : énergie *versus* TM-score. Voici les graphiques de l'energie *versus* le TM-score obtenus pour la version 3.1 d'OPEP. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd.



Tab. A.4: OPEP 3.1 : énergie *versus* TM-score. Voici les graphiques de l'energie *versus* le TM-score obtenus pour la version 3.1 d'OPEP. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd.

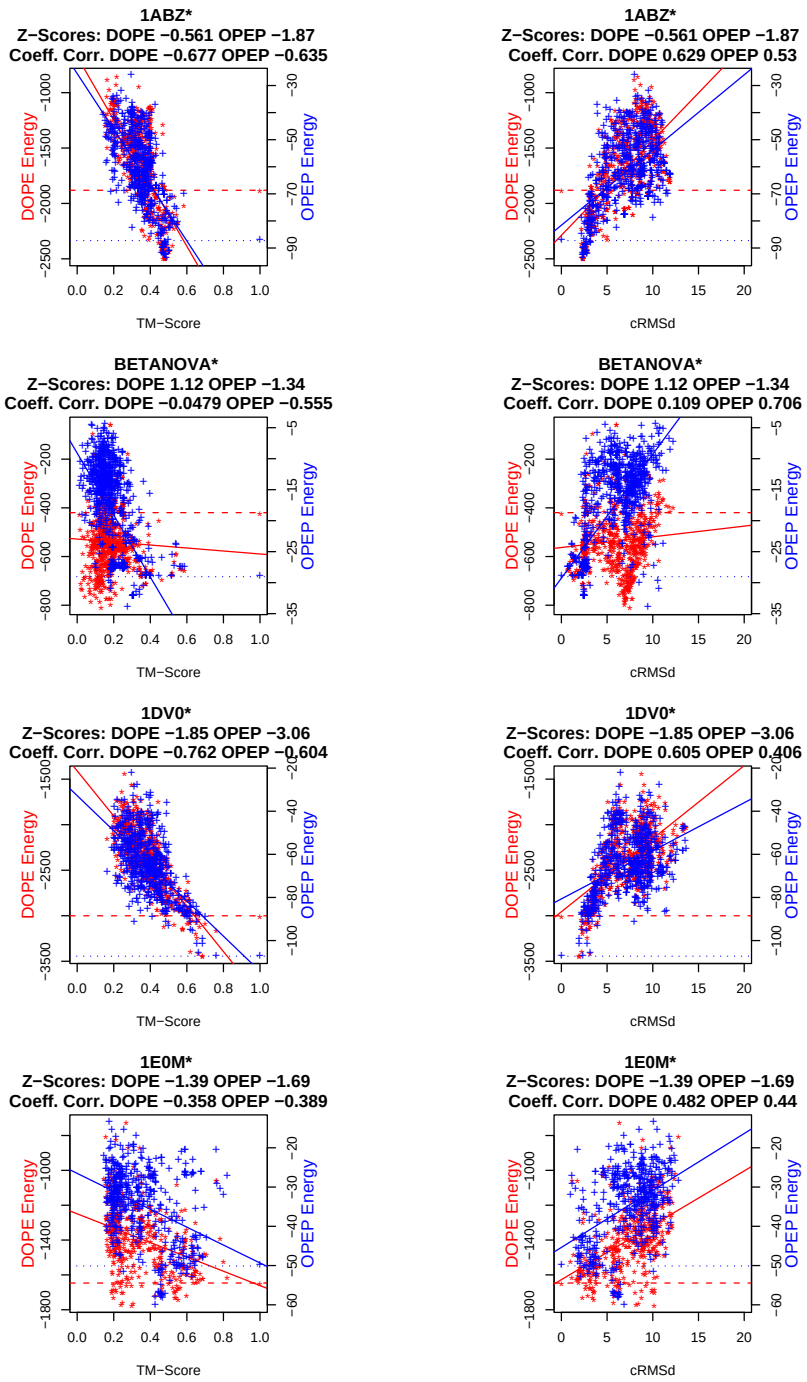


Tab. A.4: OPEP 3.1 : énergie *versus* TM-score. Voici les graphiques de l'energie *versus* le TM-score obtenus pour la version 3.1 d'OPEP. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd.

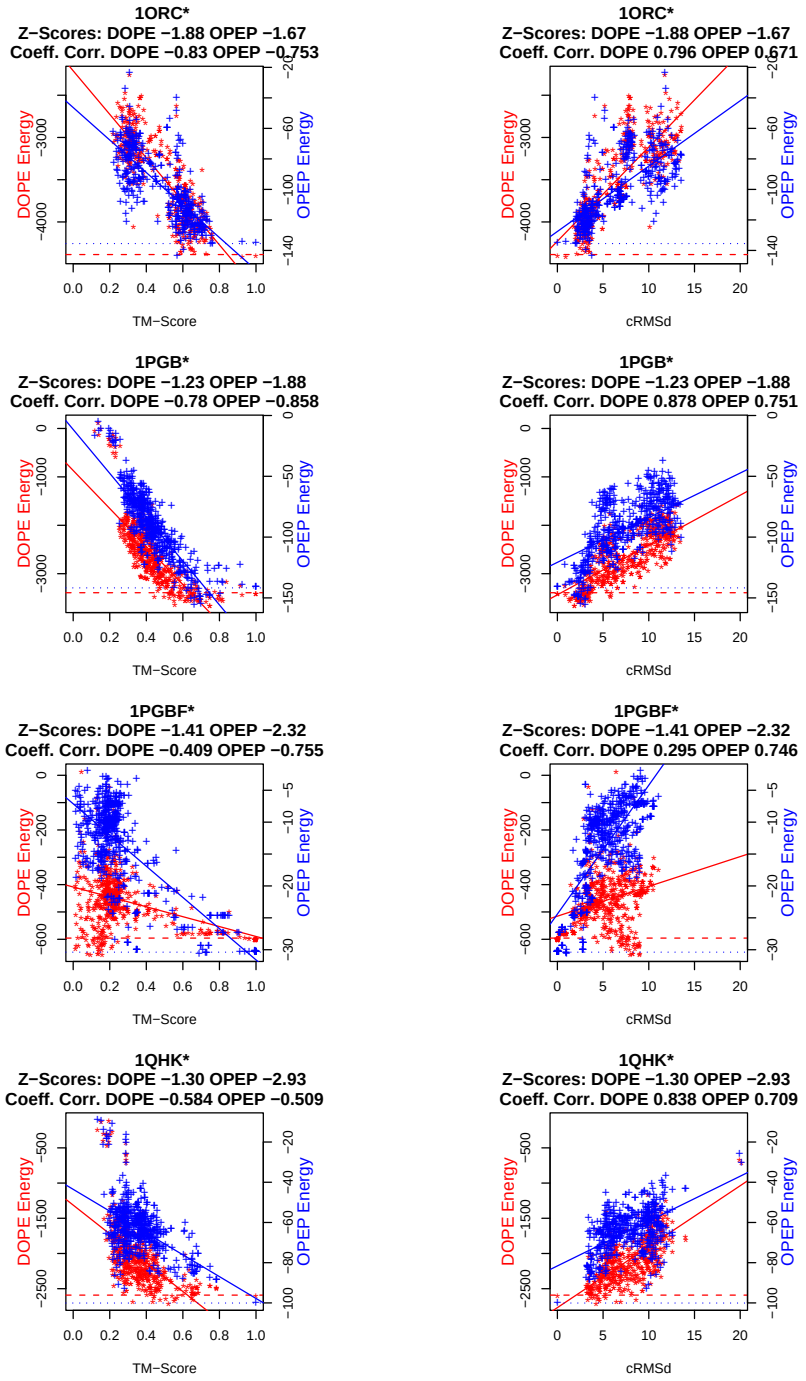


Tab. A.4: OPEP 3.1 : énergie *versus* TM-score. Voici les graphiques de l'energie *versus* le TM-score obtenus pour la version 3.1 d'OPEP. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd.

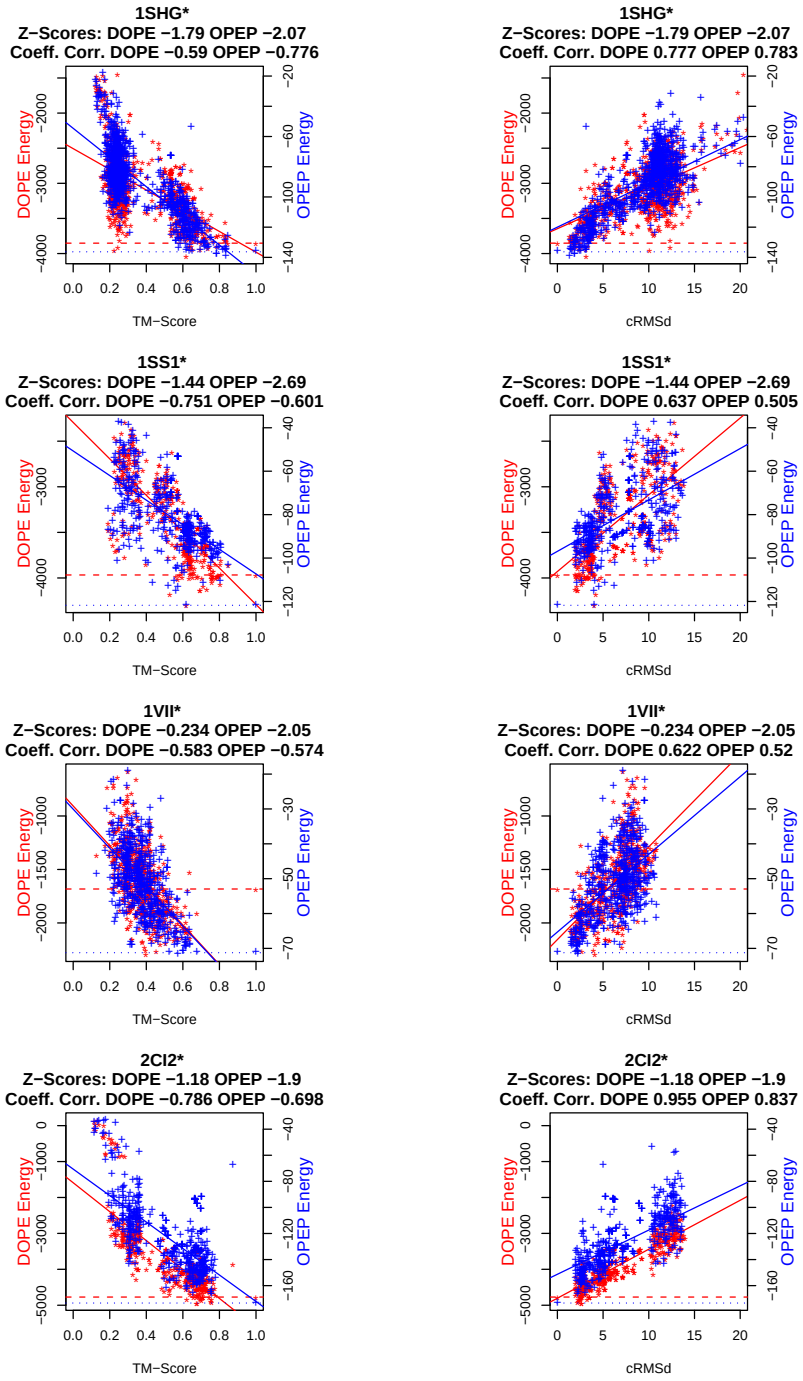
A.3 Comparaison DOPE *vs.* OPEP v3.1 sur les leurres minimisés.



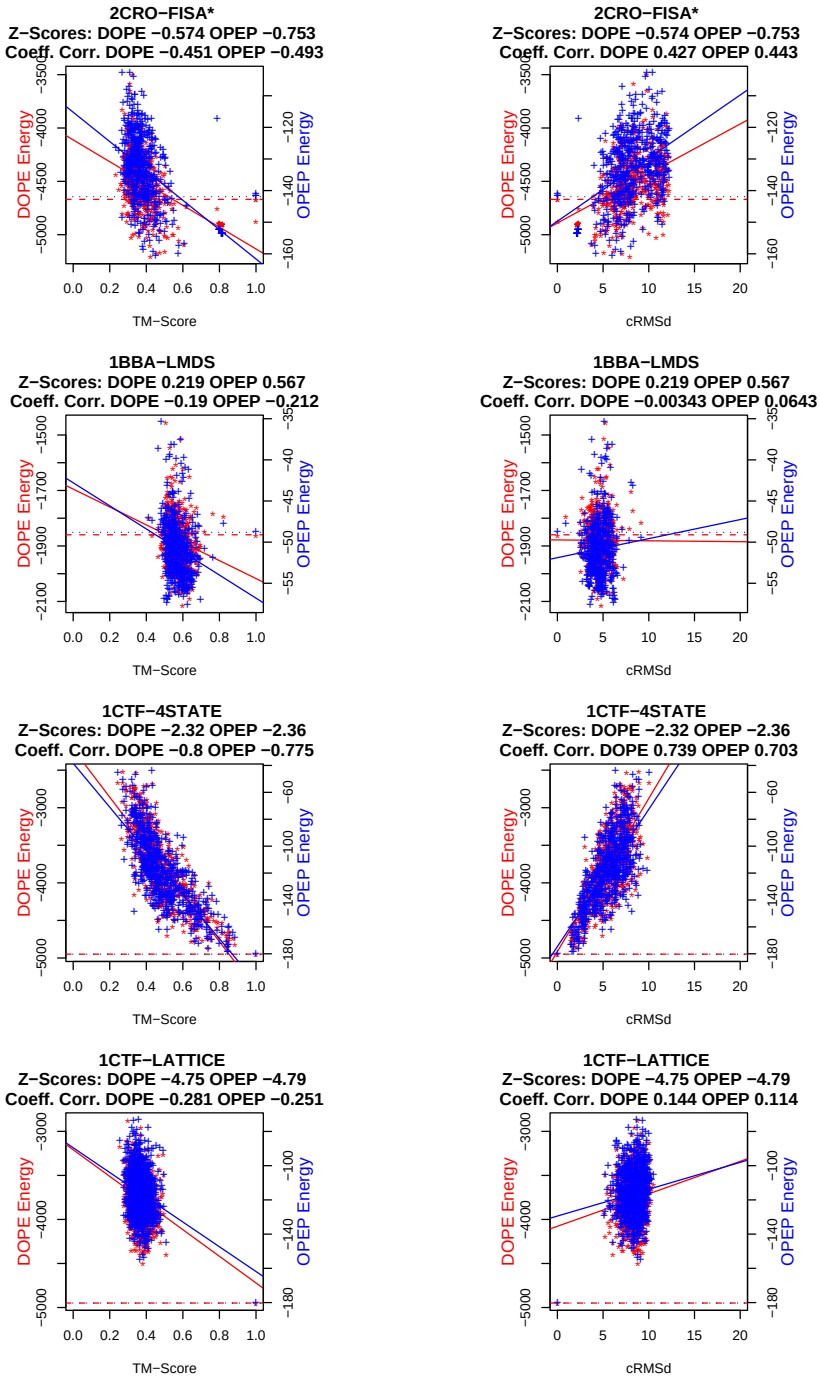
Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.



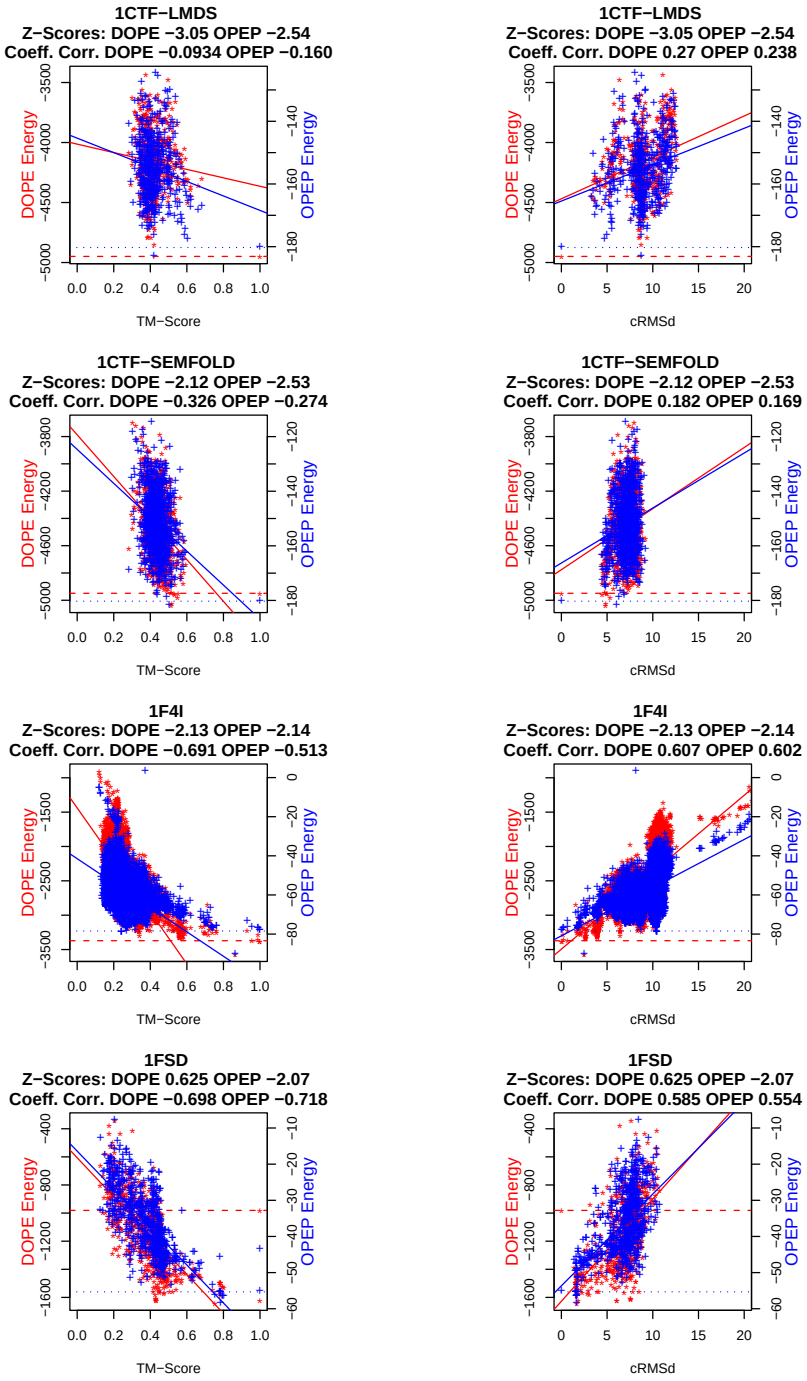
Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.



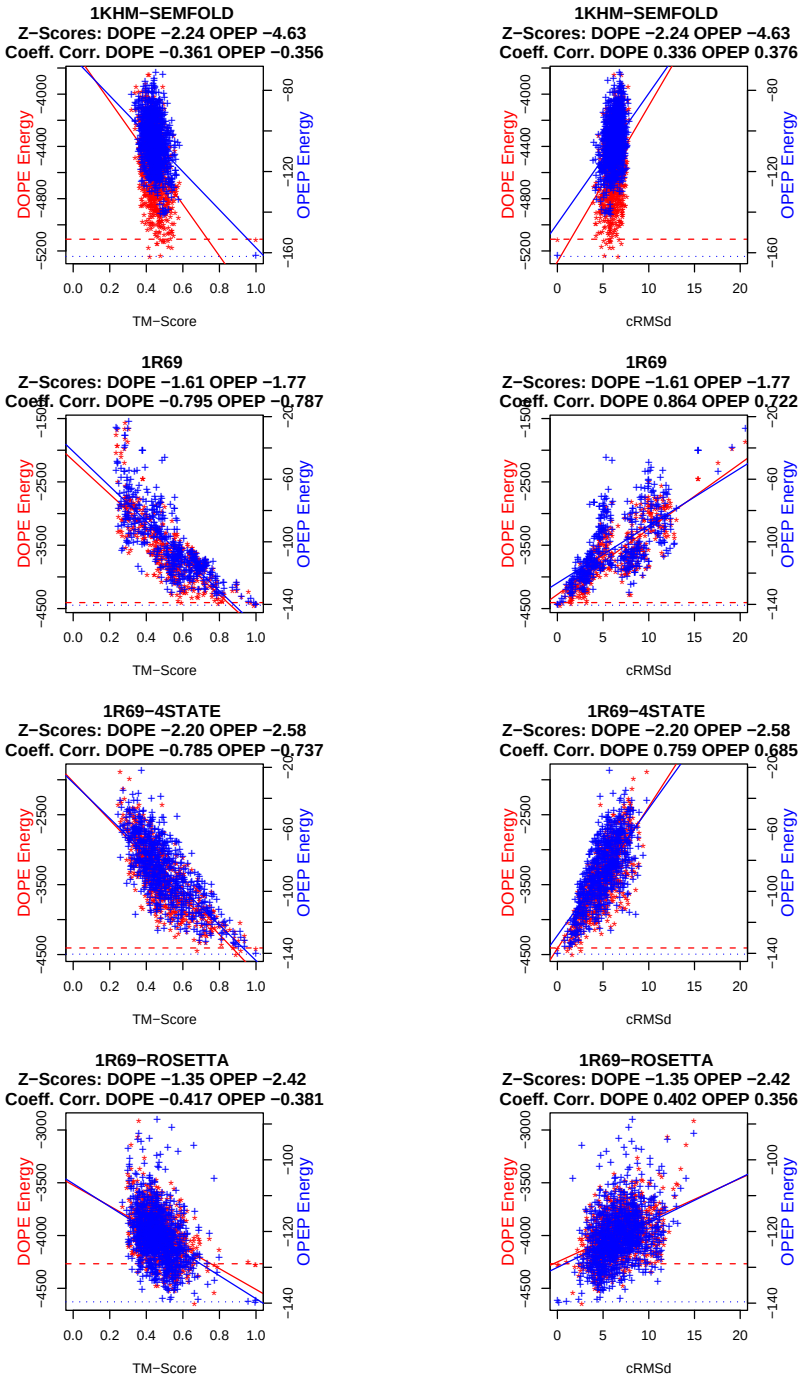
Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.



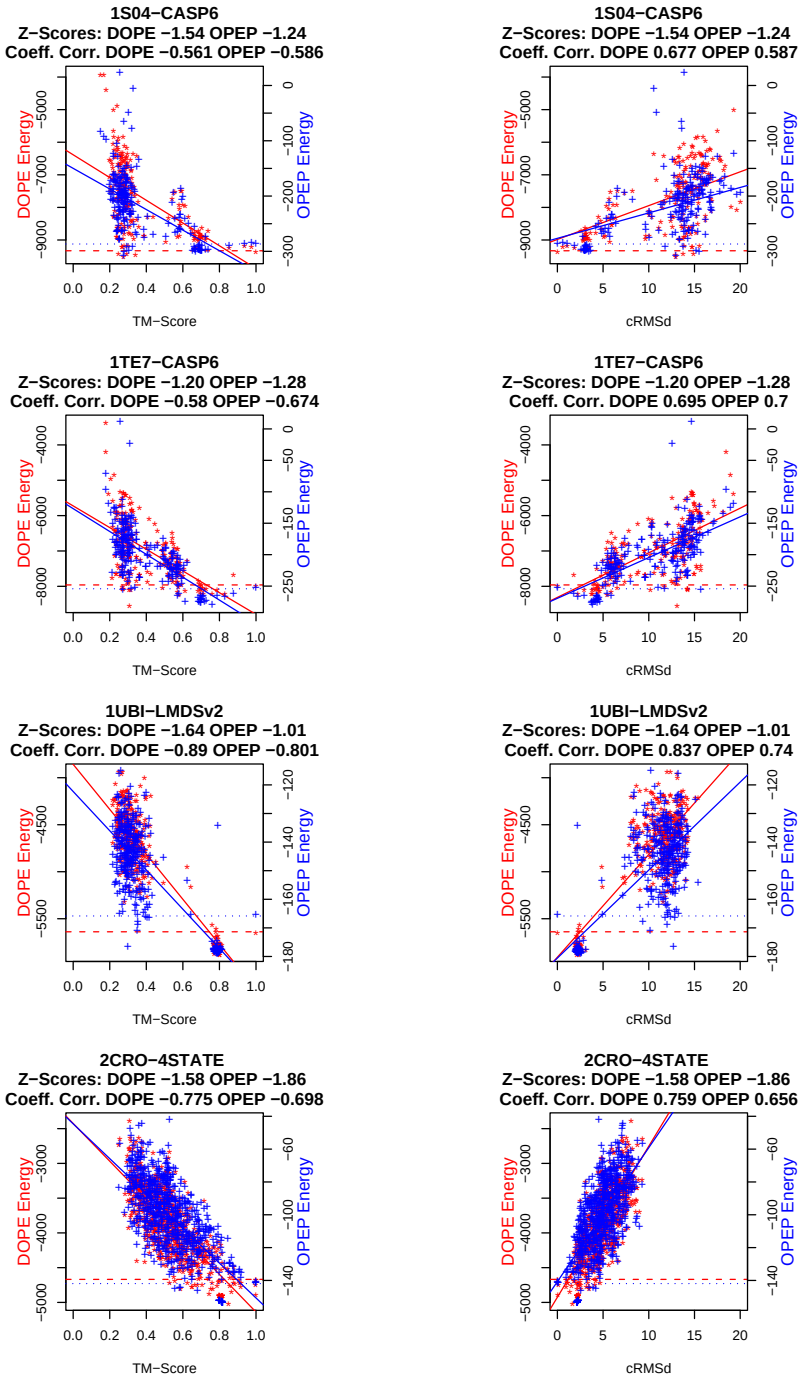
Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.



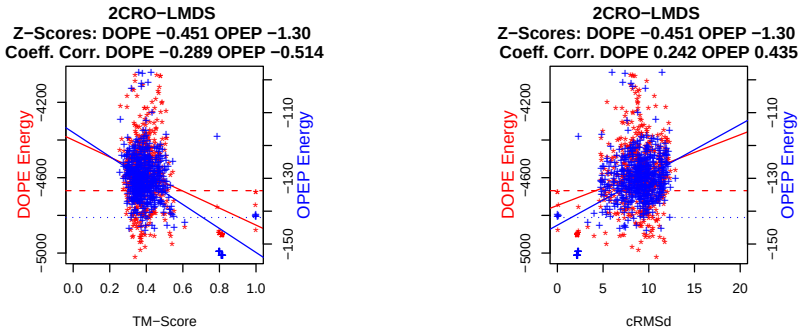
Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.



Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.

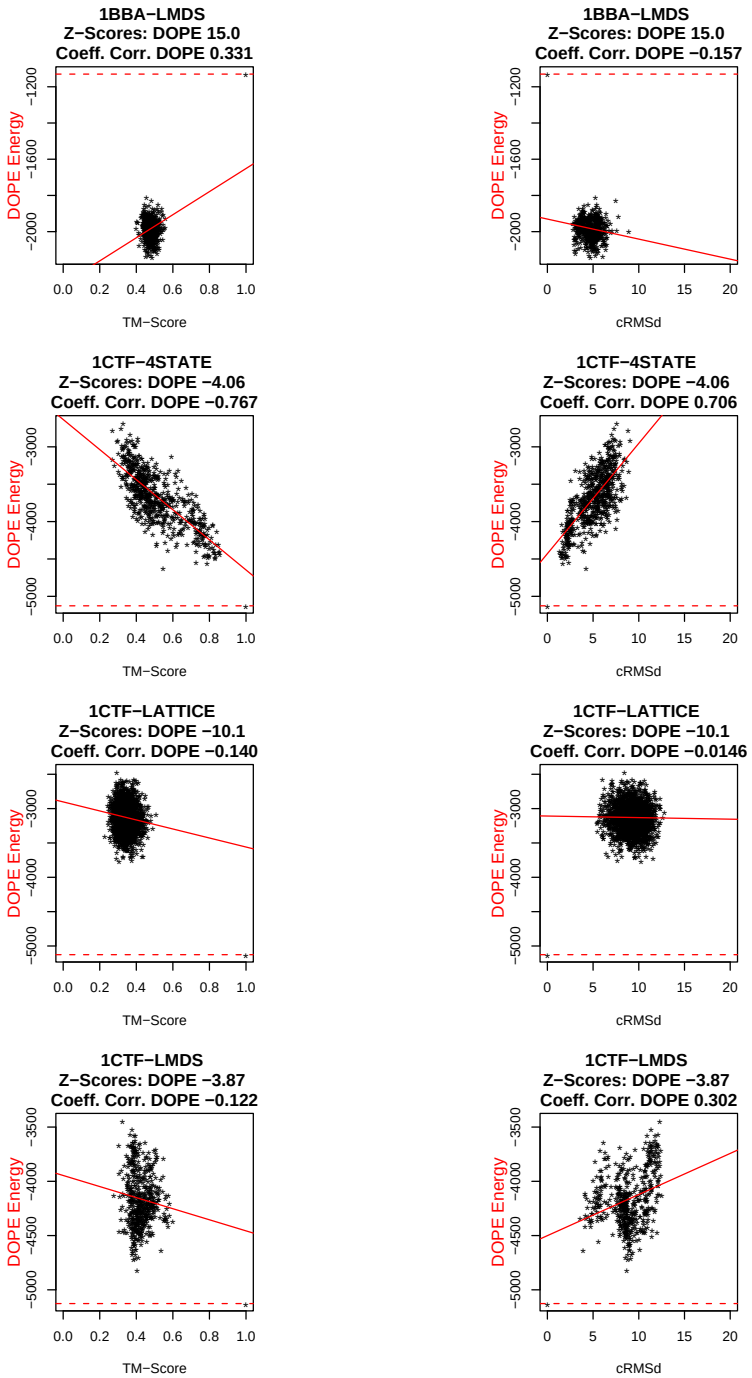


Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.

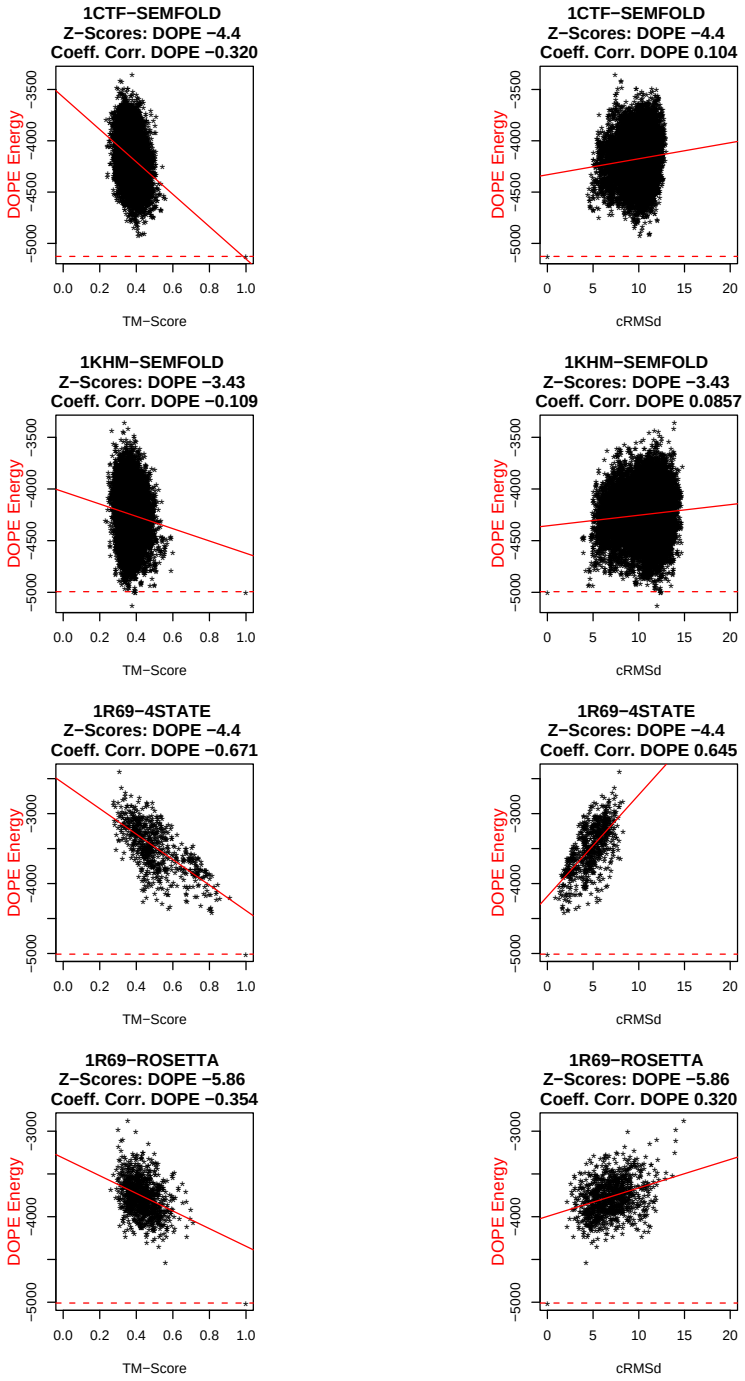


Tab. A.5: OPEP 3.1 *vs.* DOPE : les leurres minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant OPEP 3.1 (croix bleues) et DOPE (étoiles rouges). Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native. Les noms des proteines appartenant au jeu d'apprentissage se terminent par une étoile.

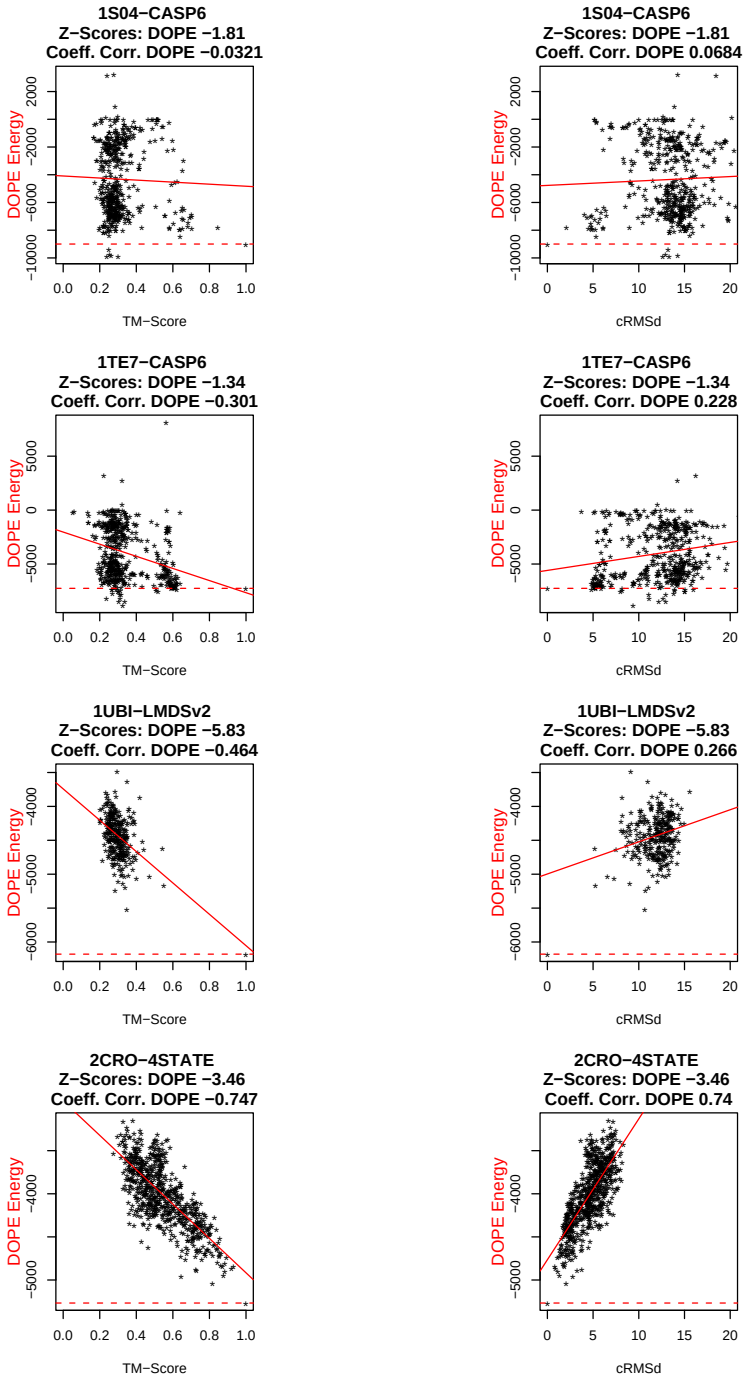
A.4 Le pouvoir discriminant de DOPE sur les leurres non minimisés pour les jeux de leurres publiquement disponibles.



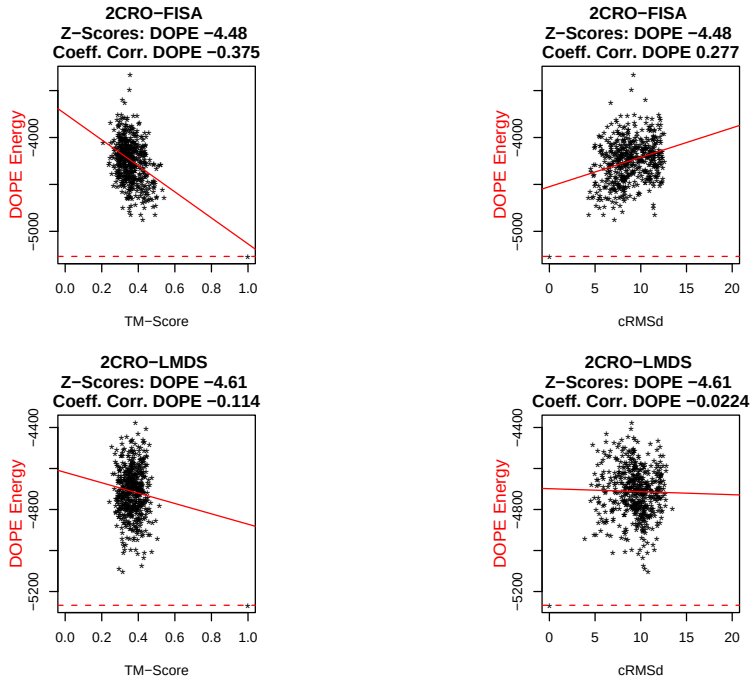
Tab. A.6: Le Pouvoir discriminant de DOPE sur les leurres publics non minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant DOPE. Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native.



Tab. A.6: Le Pouvoir discriminant de DOPE sur les leurres publics non minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant DOPE. Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native.



Tab. A.6: Le Pouvoir discriminant de DOPE sur les leurres publics non minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant DOPE. Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native.



Tab. A.6: Le Pouvoir discriminant de DOPE sur les leurres publics non minimisés. Sur ces graphiques sont tracées les énergies contre les TM-scores et les énergies contre les cRMSds en utilisant DOPE. Les lignes continues correspondent à la régression linéaire, et les lignes discontinues à l'énergie de la structure native.

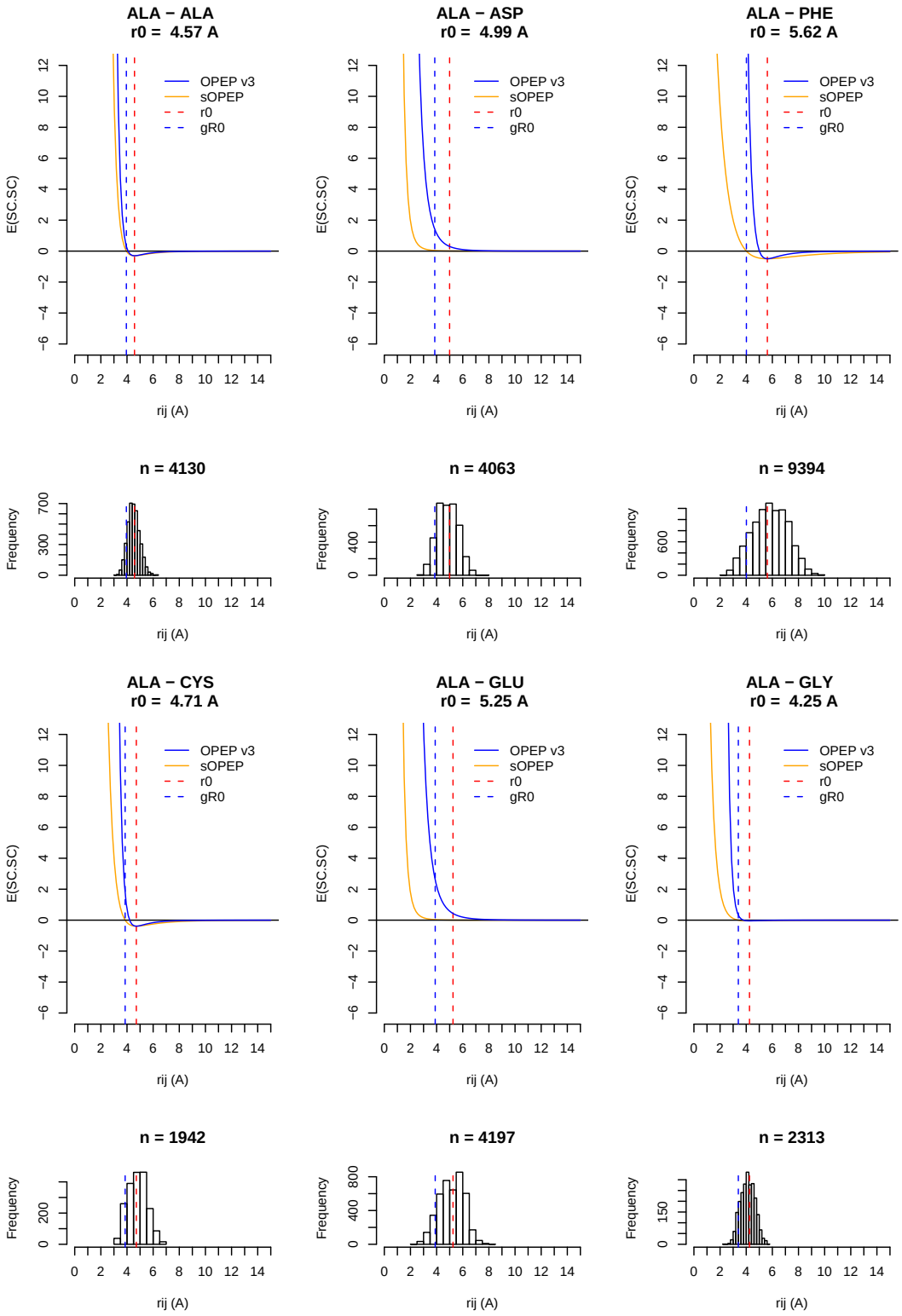
A.5 sOPEP v2.0 : la nouvelle formulation du potentiel CL-CL

A.5.1 Valeurs du paramètre gR^0_{ij}

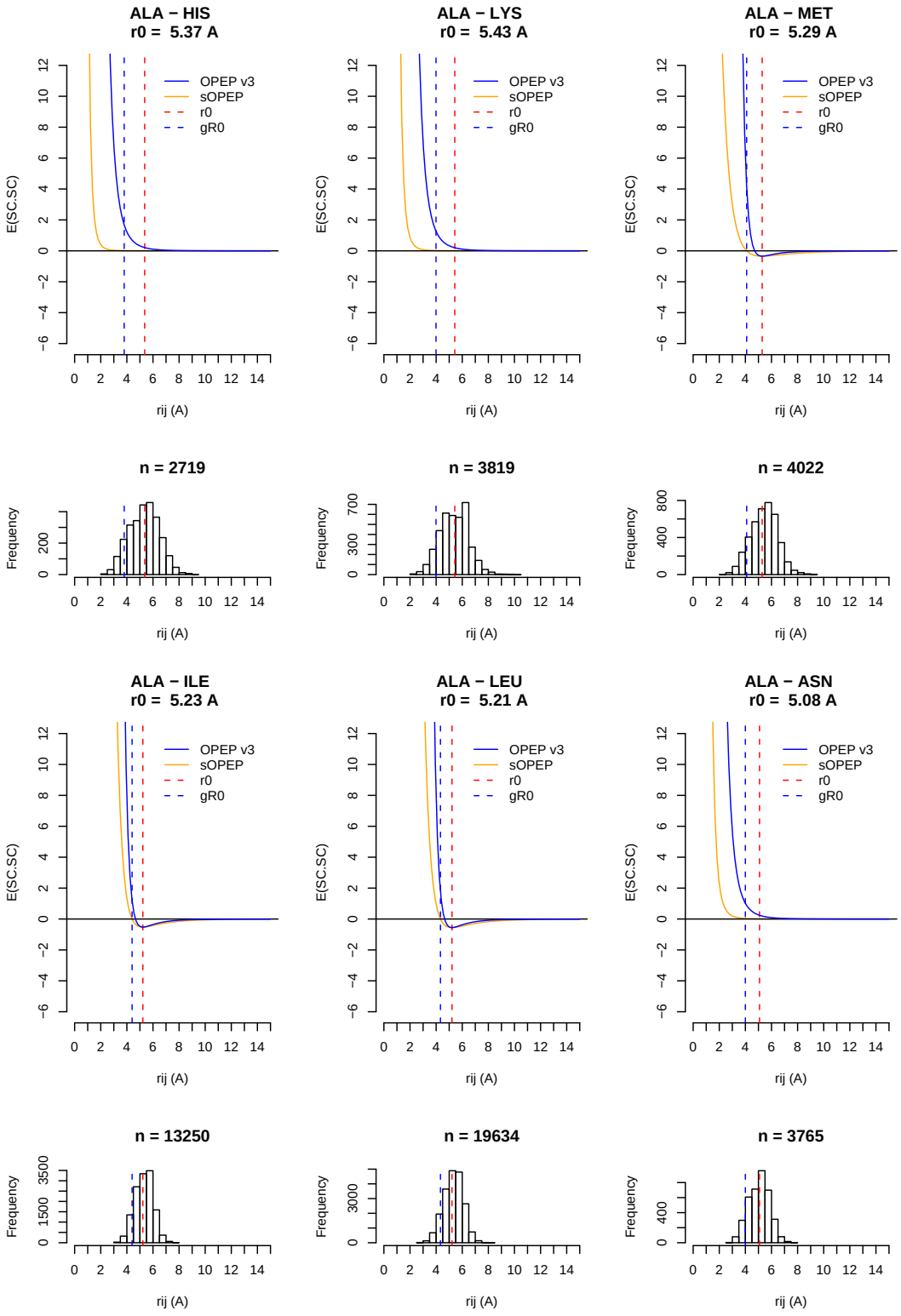
-	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	3,95	3,86	4,14	4,30	4,01	3,40	4,28	4,40	4,44	4,33	4,10	4,33	4,45	4,39	4,60	4,02	4,40	4,45	3,97	3,79
CYS	-	3,04	4,33	4,25	3,97	3,53	4,20	4,40	4,60	4,30	3,91	4,52	4,35	4,35	4,58	4,18	4,44	4,43	4,11	3,73
ASP	-	-	4,62	4,82	4,75	3,84	4,34	4,76	4,27	4,55	4,45	4,26	4,72	4,96	4,41	4,35	4,52	4,77	5,23	4,46
GLU	-	-	-	5,32	4,86	3,96	4,39	4,85	4,37	4,81	4,63	4,29	4,82	5,25	4,55	4,46	4,67	4,83	4,49	4,48
PHE	-	-	-	-	4,26	4,25	4,05	4,39	4,80	4,38	4,27	4,88	4,05	4,11	4,97	4,59	4,97	4,61	4,31	4,10
GLY	-	-	-	-	-	2,87	4,03	4,31	4,13	4,33	4,14	4,01	3,99	4,08	4,27	3,66	3,97	4,20	3,70	3,46
HIS	-	-	-	-	-	-	4,43	4,64	4,79	4,66	4,05	4,97	4,05	4,98	4,93	4,53	4,71	4,74	3,96	4,08
ILE	-	-	-	-	-	-	-	4,80	4,91	4,82	4,55	4,88	5,08	4,92	5,03	4,64	5,04	4,92	4,25	4,15
LYS	-	-	-	-	-	-	-	-	5,46	4,84	4,80	4,40	4,94	4,33	5,58	4,69	4,90	4,97	3,84	4,09
LEU	-	-	-	-	-	-	-	-	-	4,82	4,55	4,72	4,66	4,84	4,99	4,62	5,03	4,92	4,23	4,12
MET	-	-	-	-	-	-	-	-	-	-	4,25	4,67	4,39	4,09	4,98	4,49	4,87	4,66	4,14	4,01
ASN	-	-	-	-	-	-	-	-	-	-	-	4,36	4,85	4,35	5,08	4,43	4,65	4,83	4,32	5,08
PRO	-	-	-	-	-	-	-	-	-	-	-	-	4,62	4,39	4,50	4,48	4,73	4,42	3,93	4,08
GLN	-	-	-	-	-	-	-	-	-	-	-	-	-	5,15	4,36	4,55	4,70	4,89	4,39	4,17
ARG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5,60	4,73	4,93	5,09	3,89	4,16
SER	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4,07	4,31	4,62	4,87	4,71
THR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4,57	5,00	5,26	4,95
VAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4,93	4,51	4,34
TRP	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4,34	4,07
TYR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3,88

Tab. A.7: Le paramètre gR^0_{ij} pour les 210 interactions CL-CL . Pour chacune des 210 interactions entre les chaînes latérales, sont présentées les valeurs du paramètre gR^0_{ij} pour la version 2 de sOPEP.

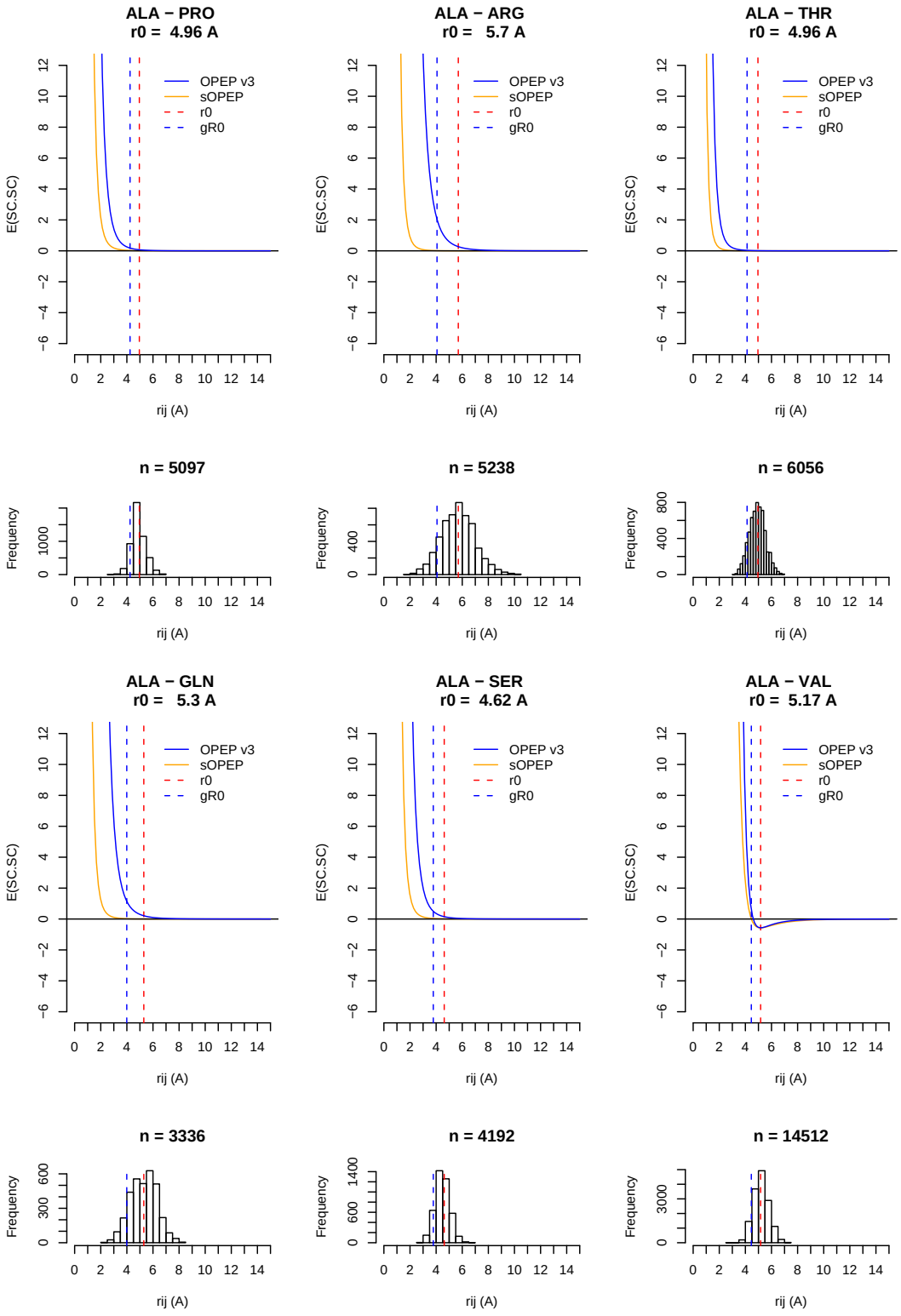
A.5.2 Tracés de la nouvelle formulation du potentiel



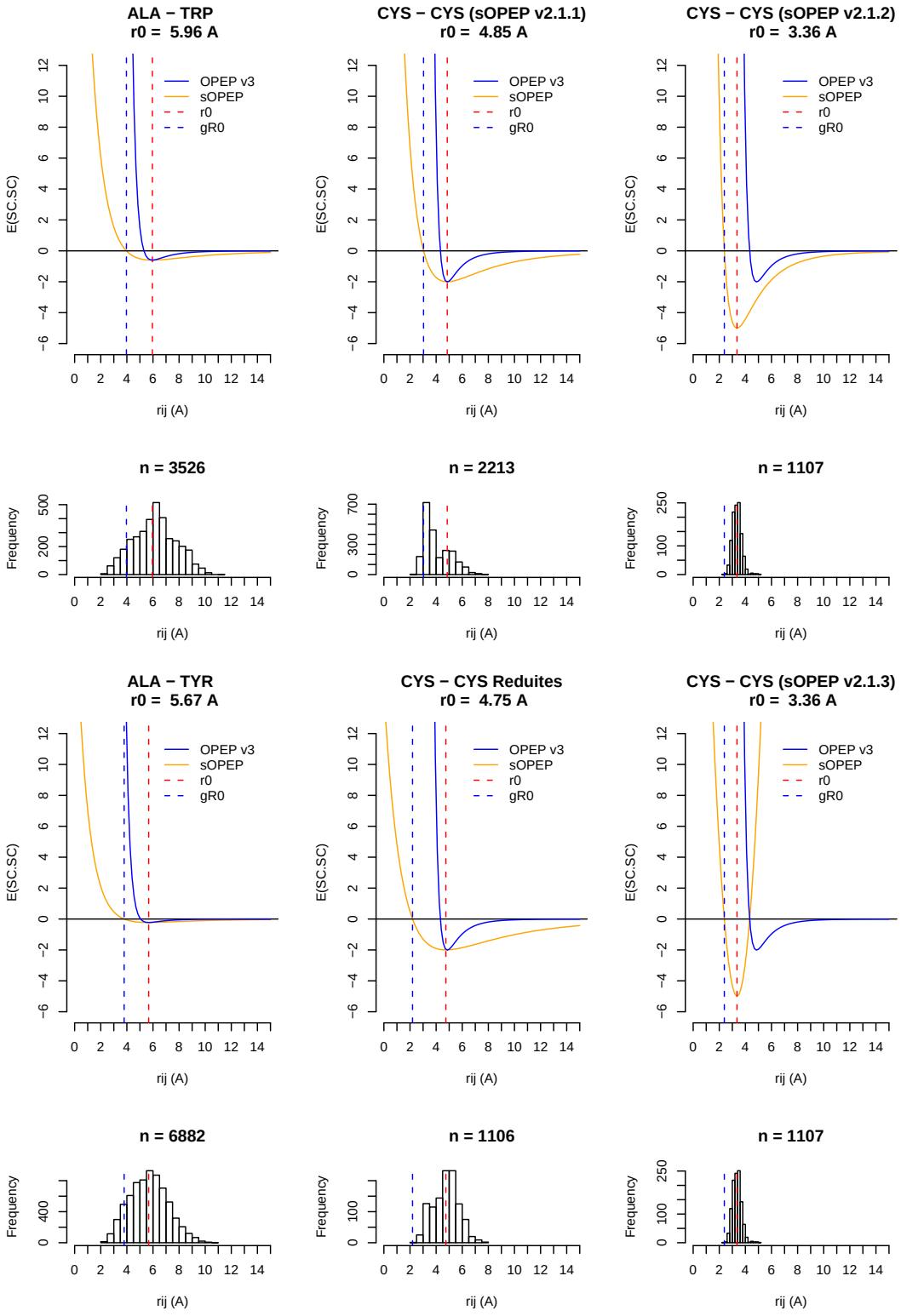
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



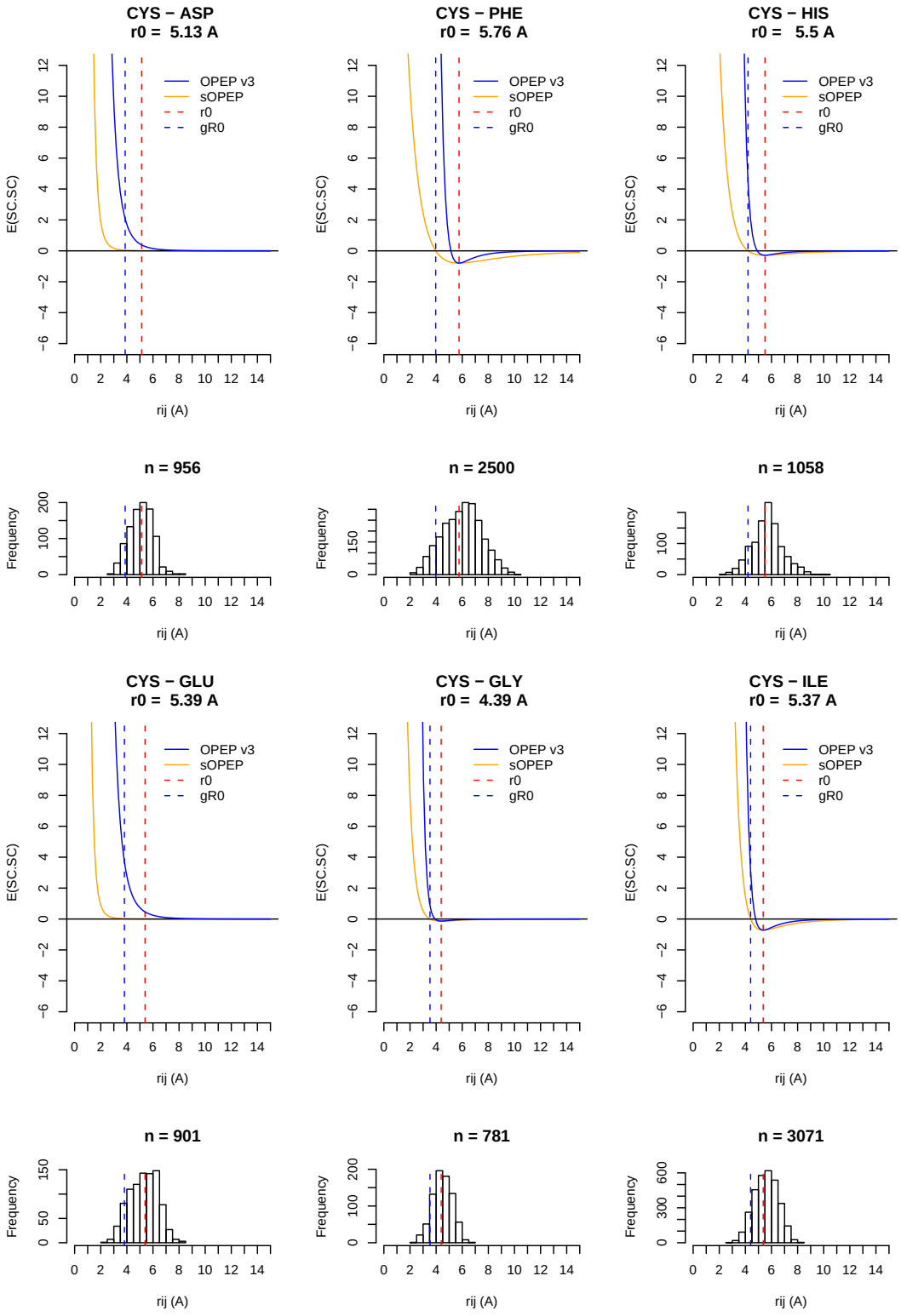
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



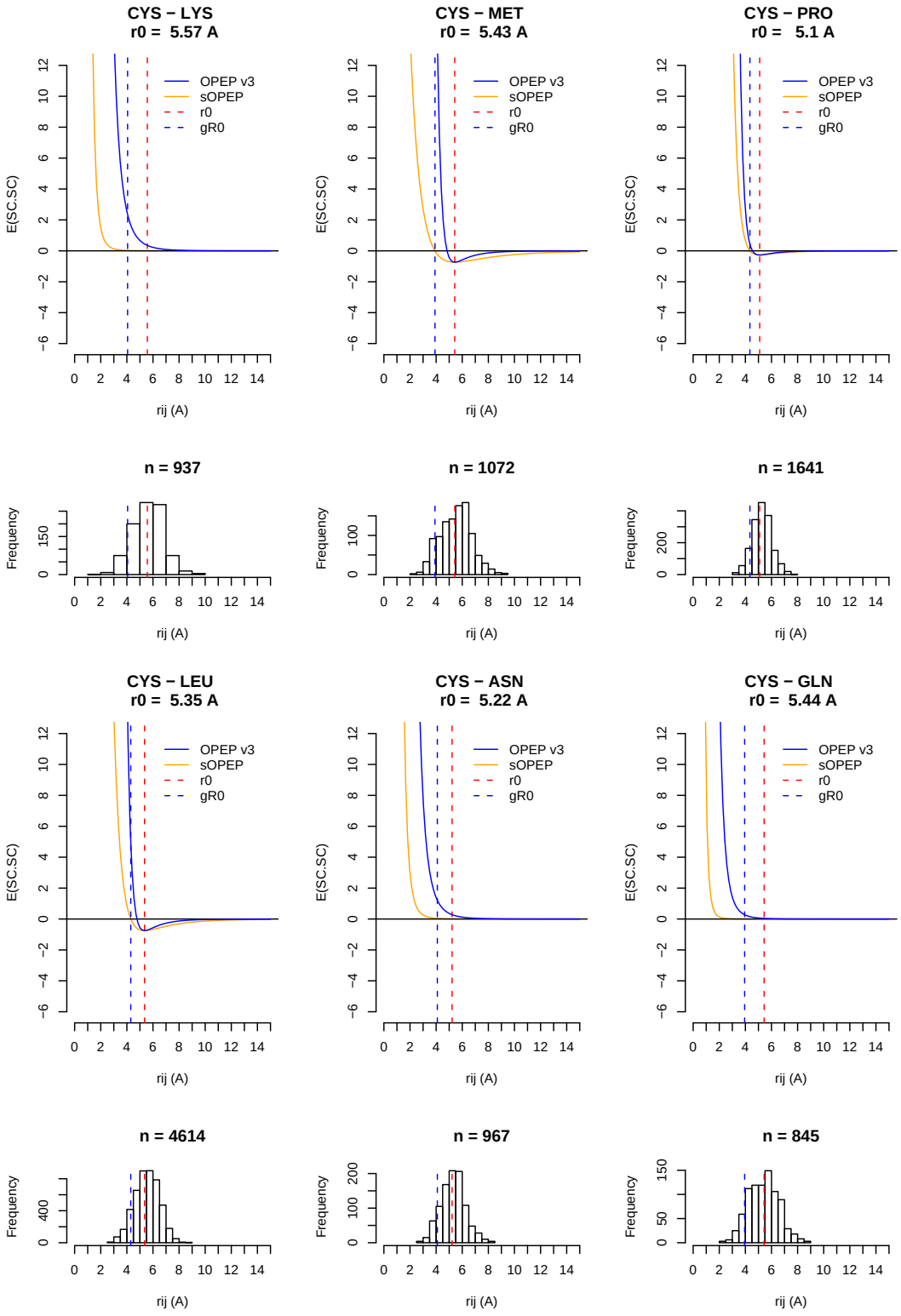
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



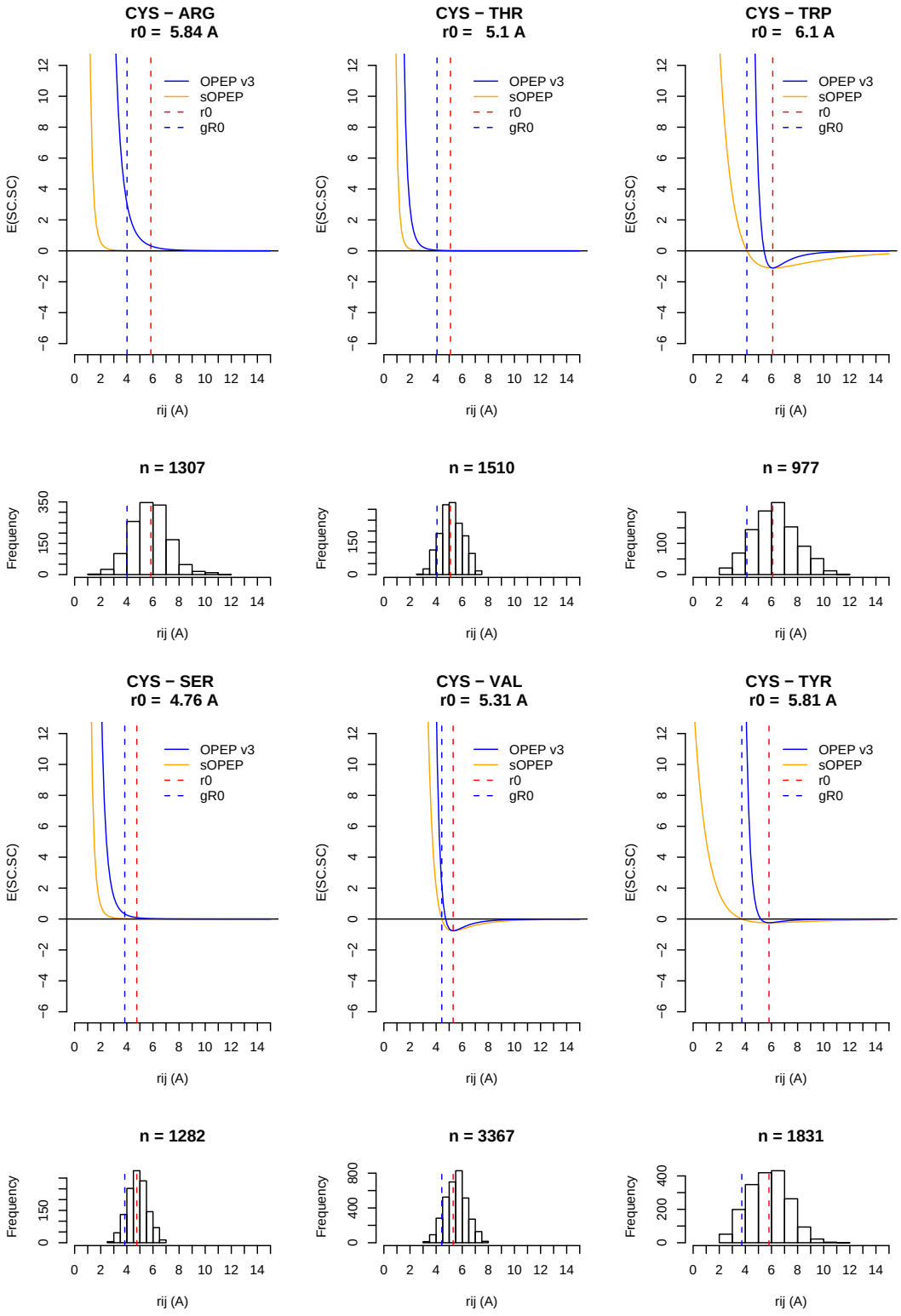
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



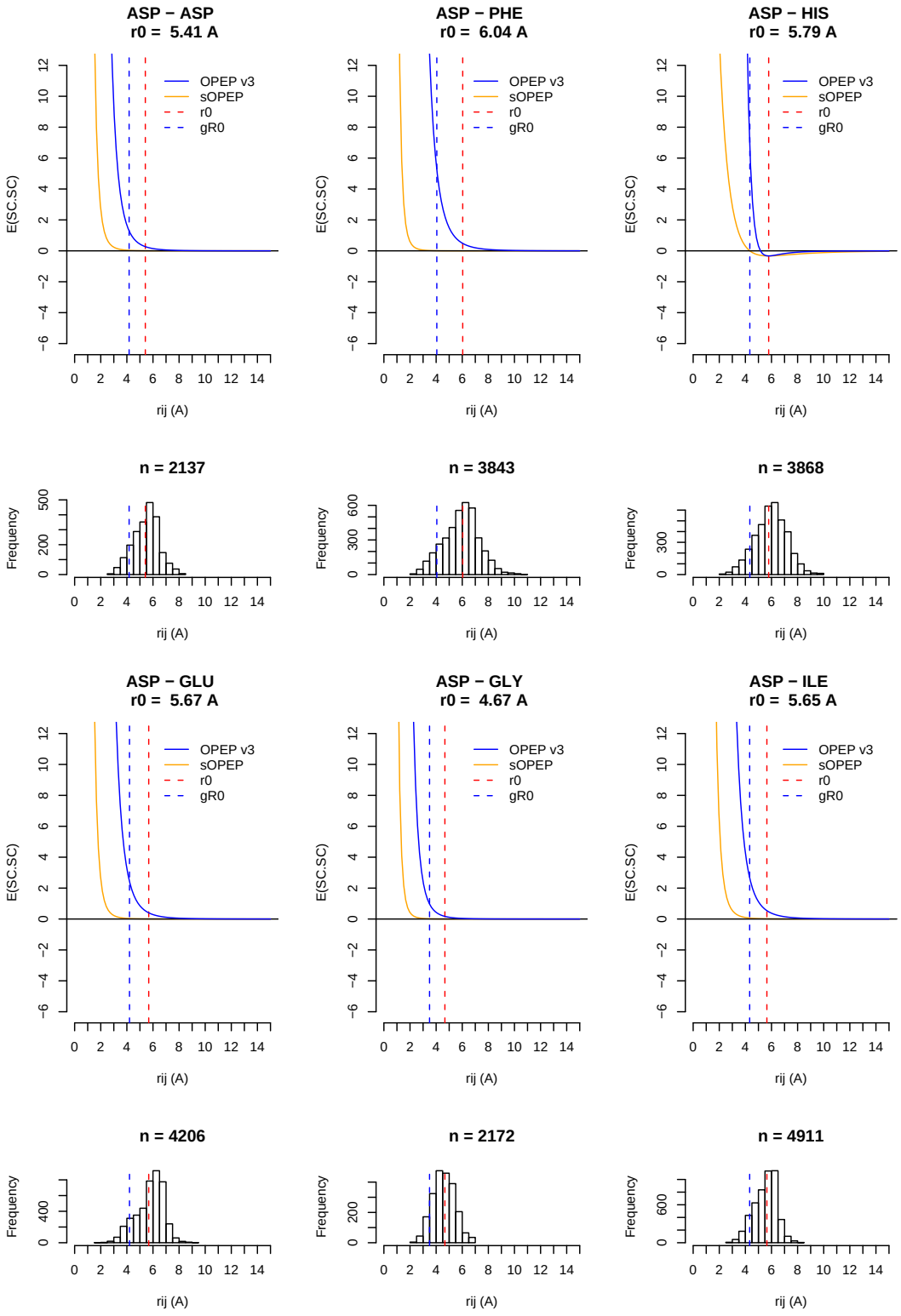
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



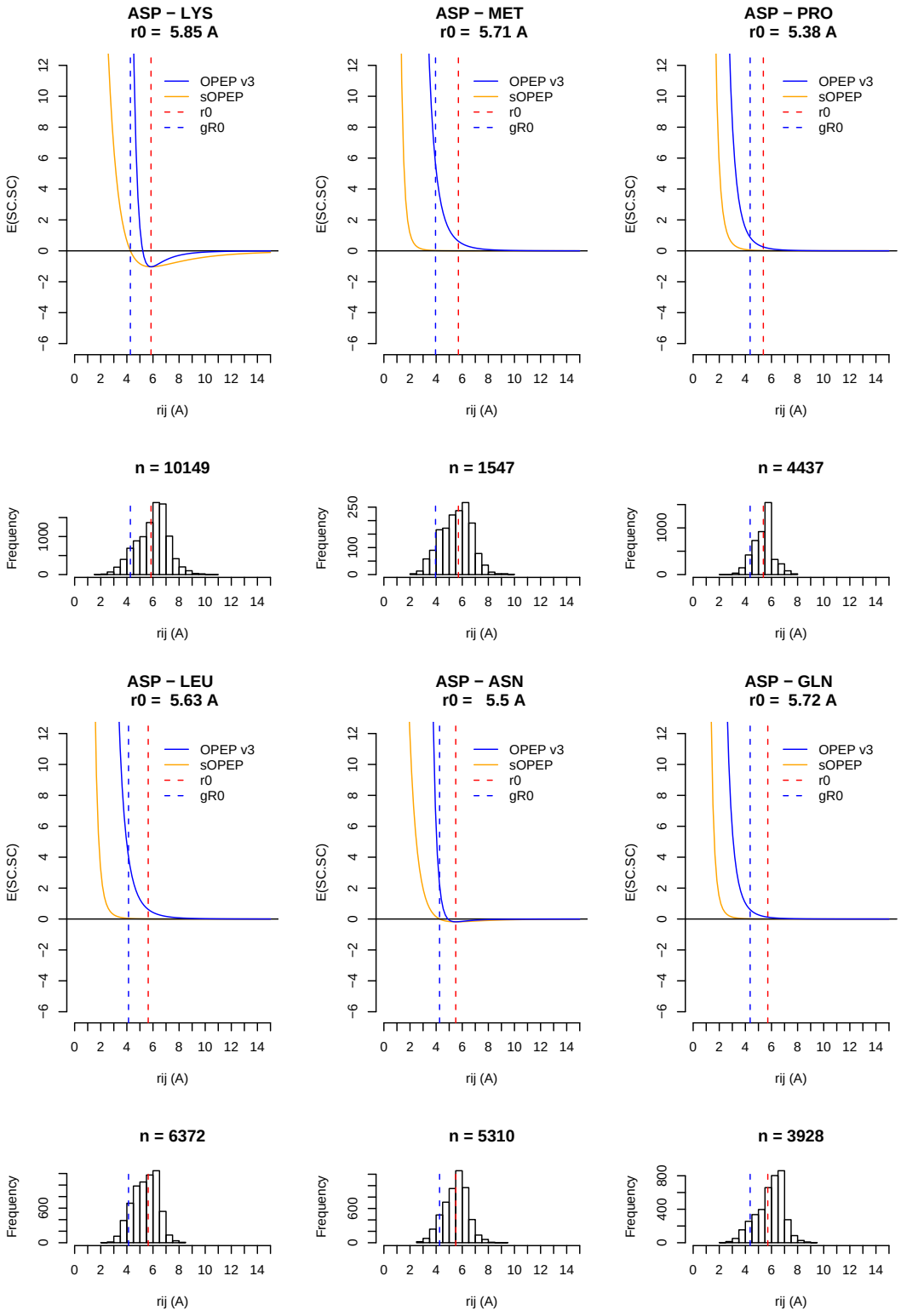
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



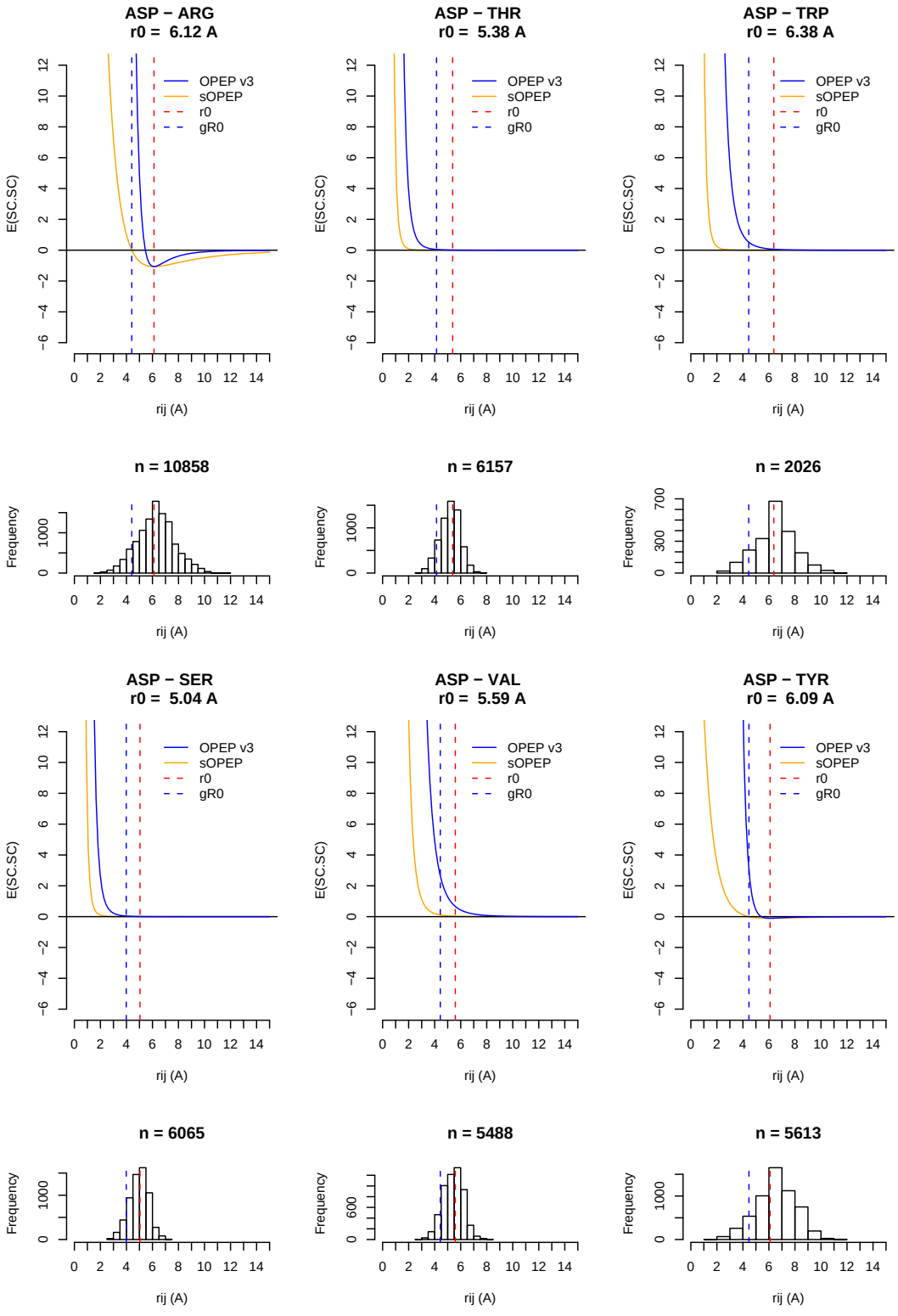
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



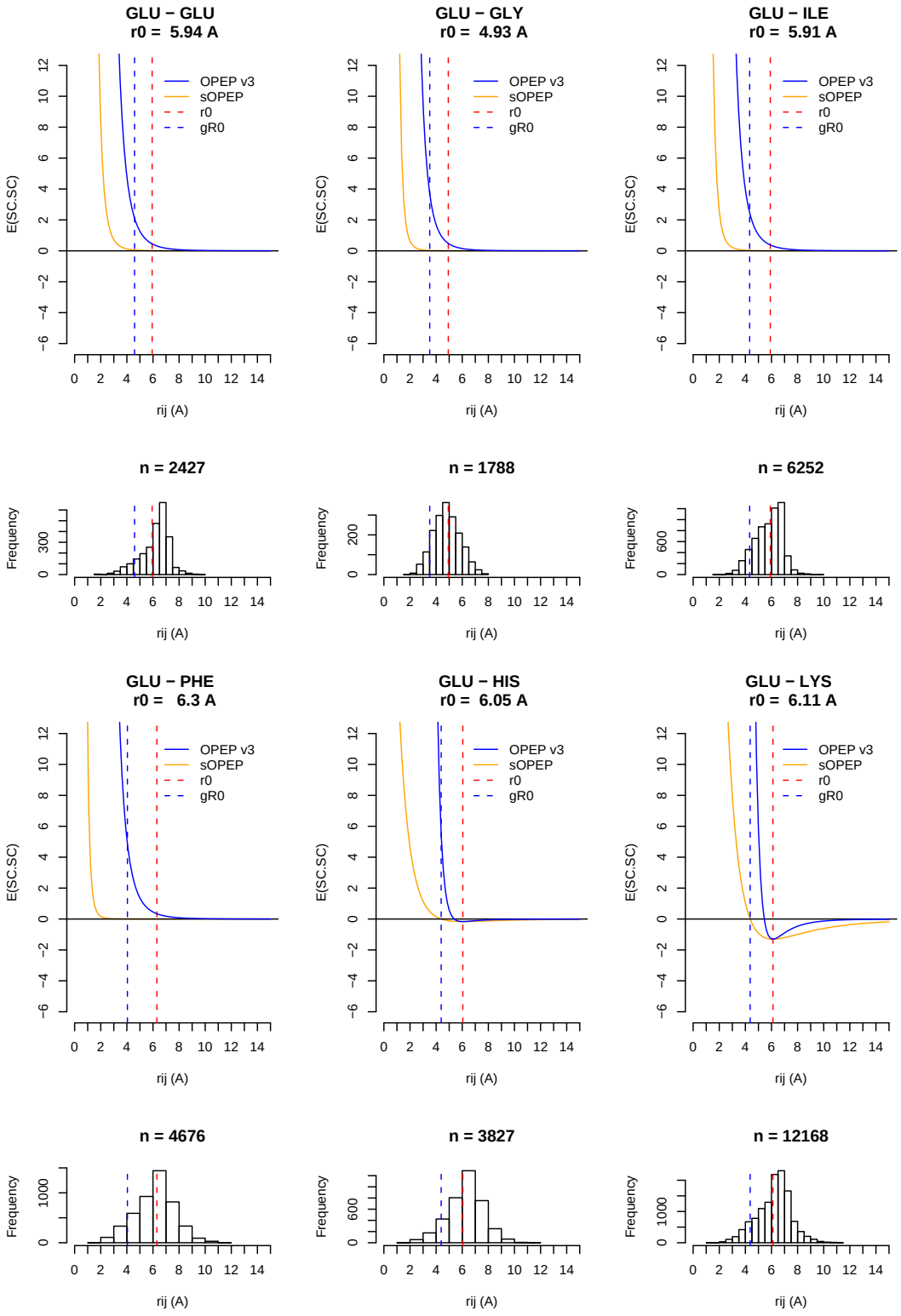
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



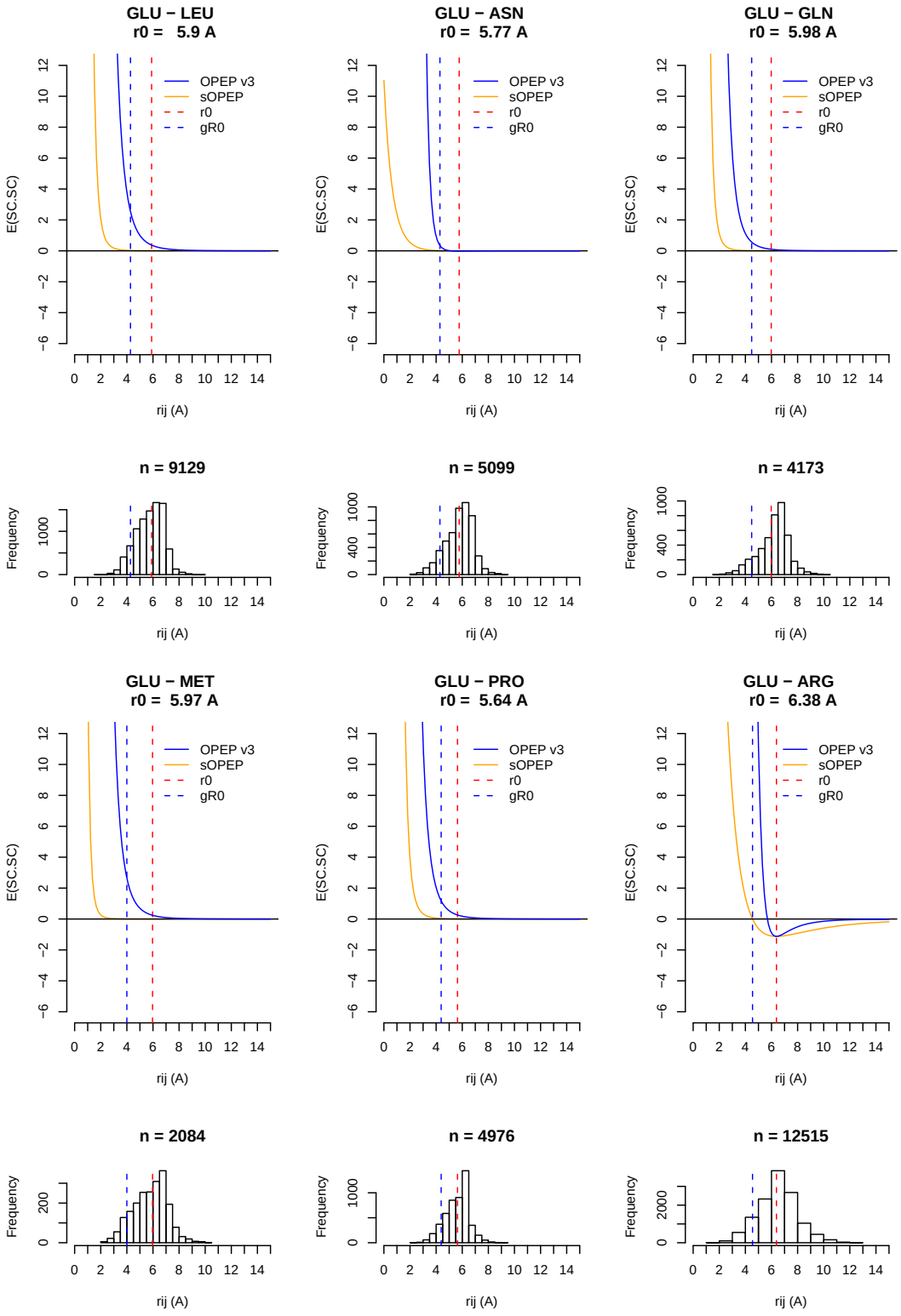
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



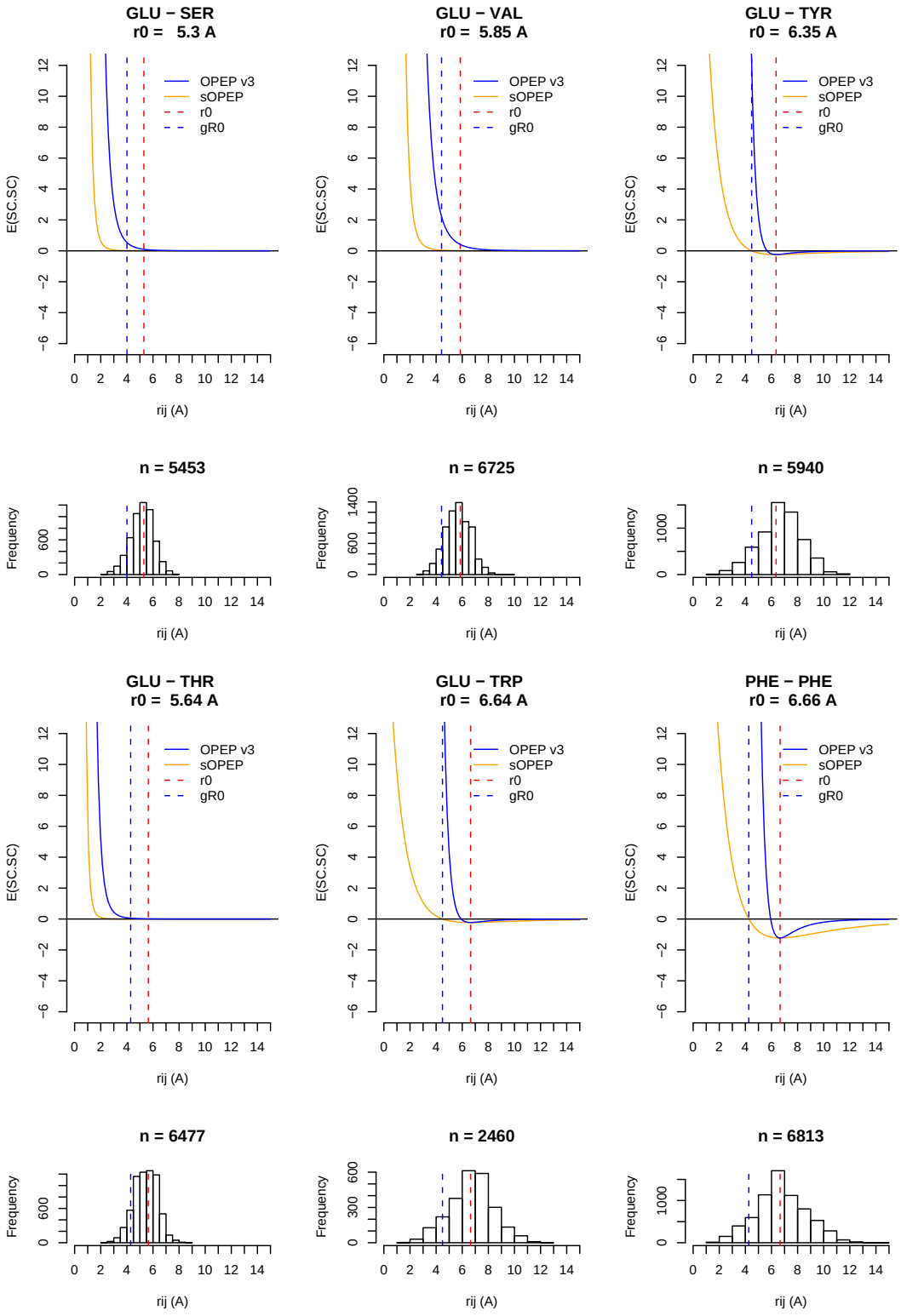
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



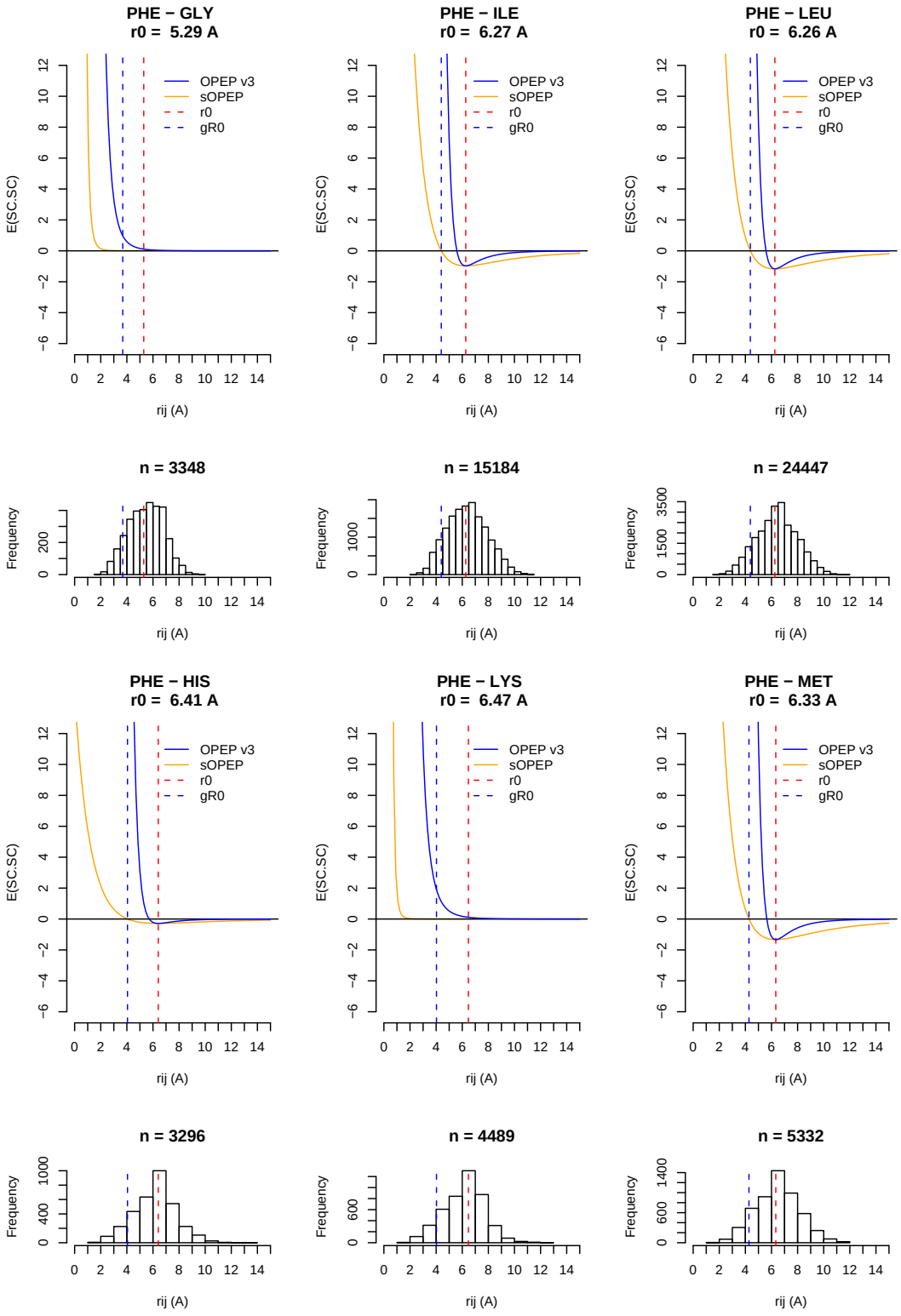
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



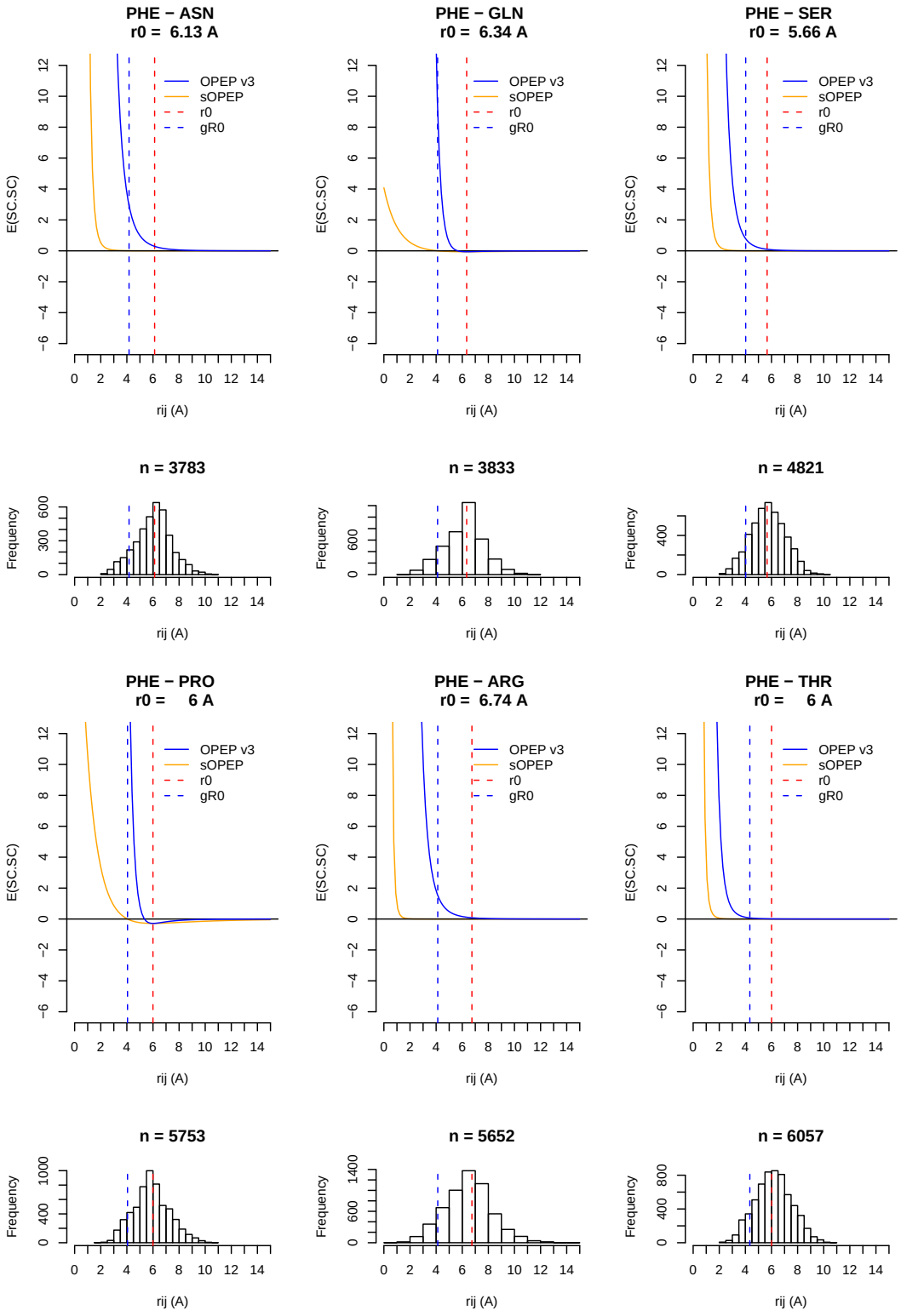
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



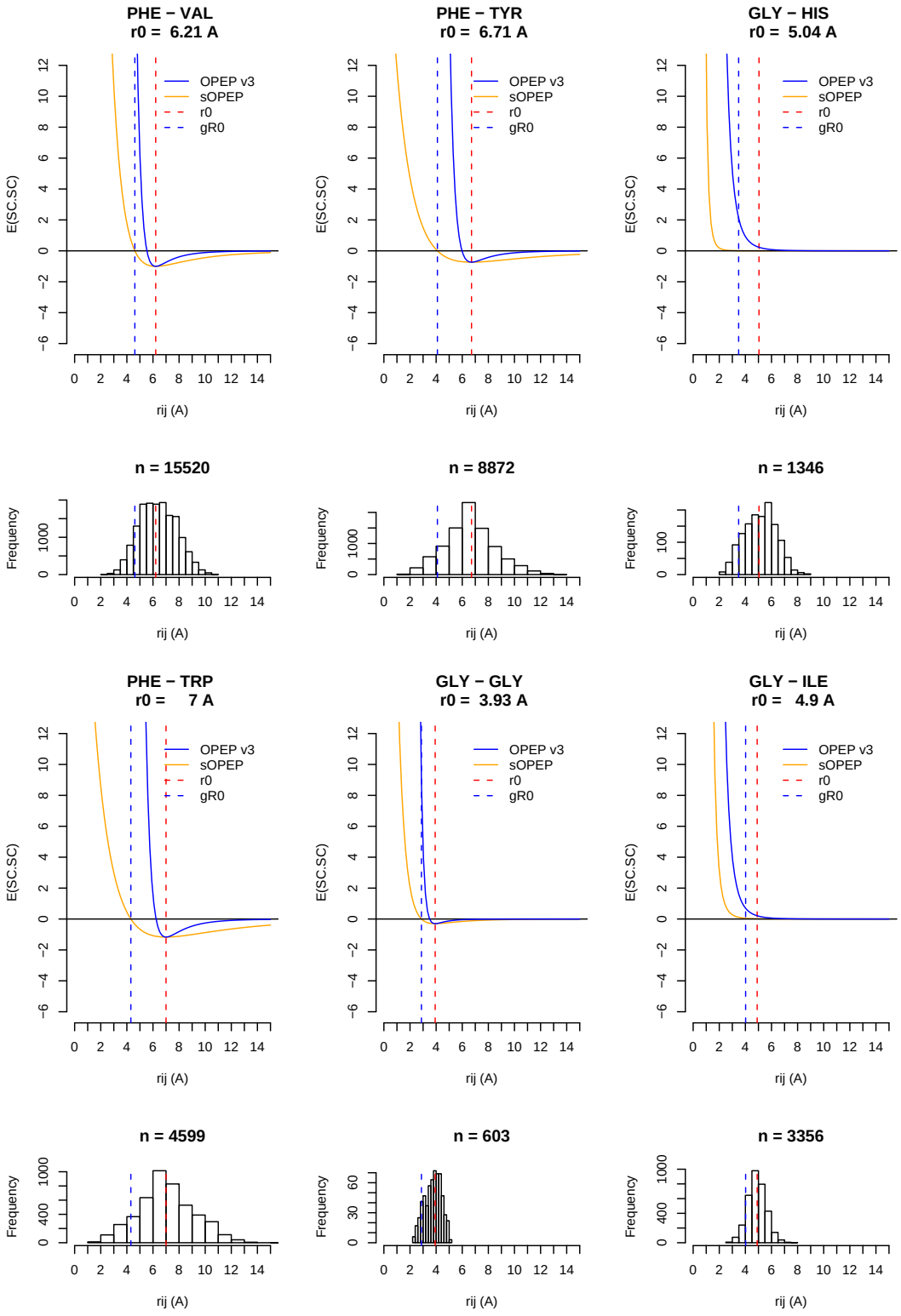
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



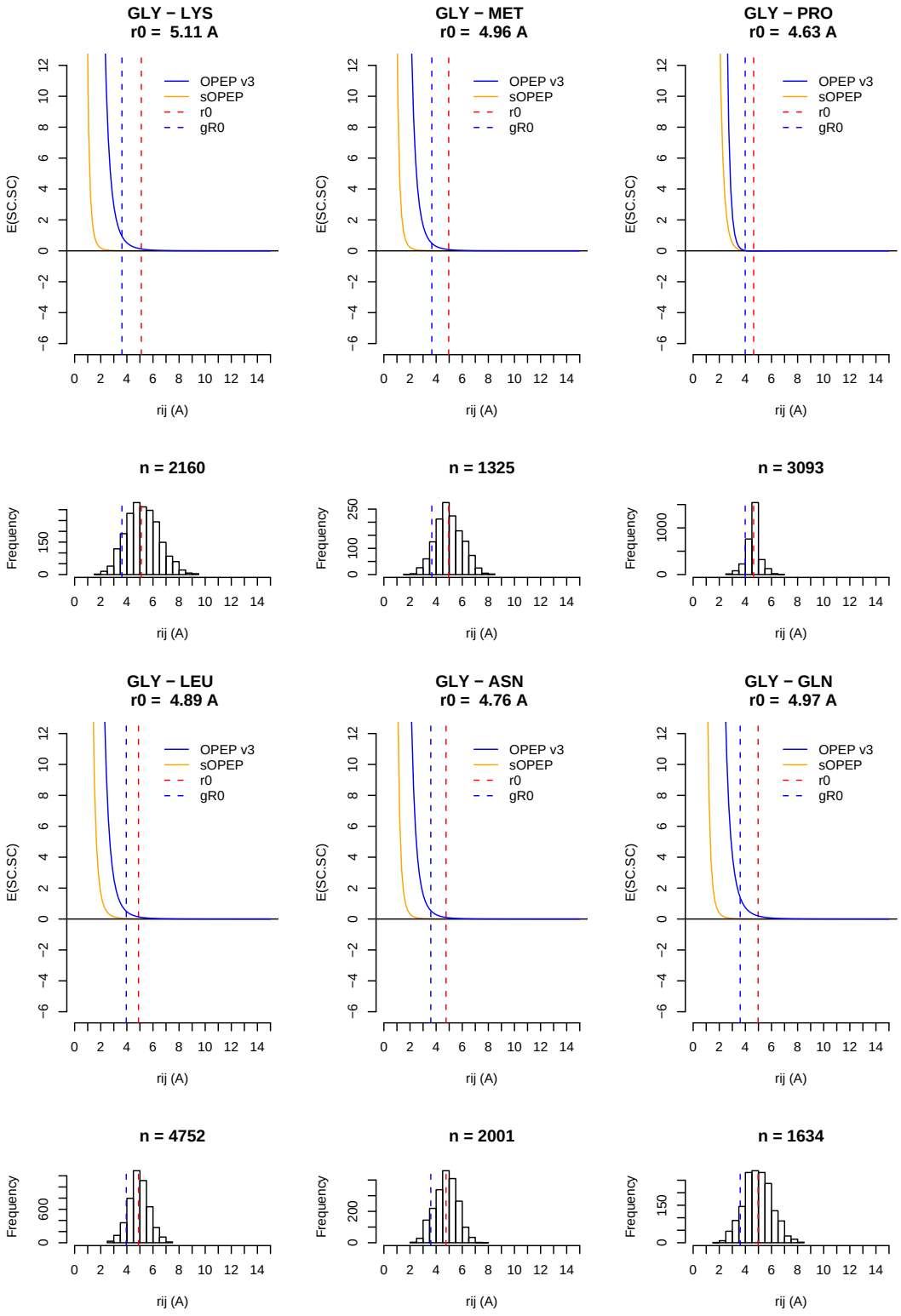
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



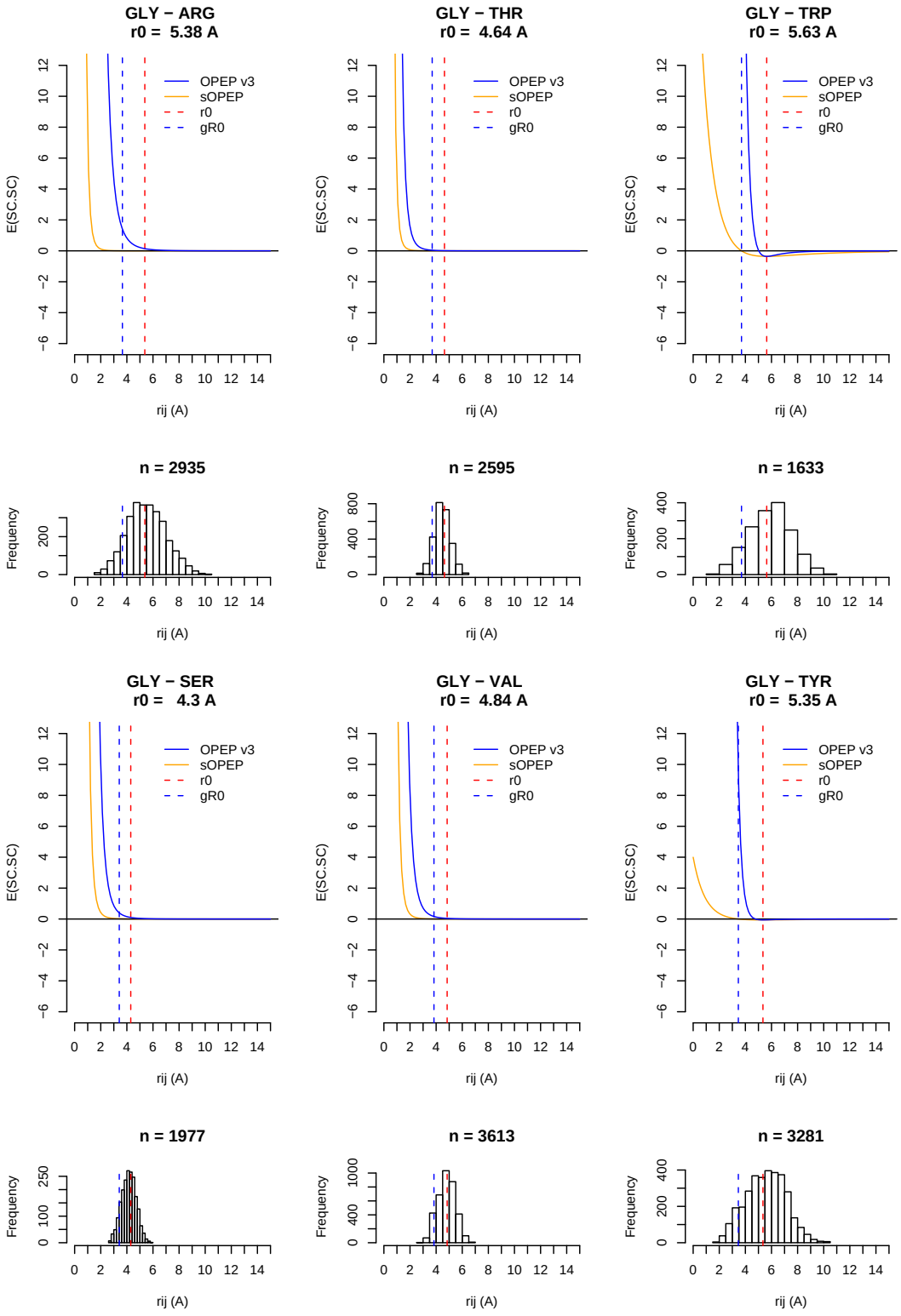
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



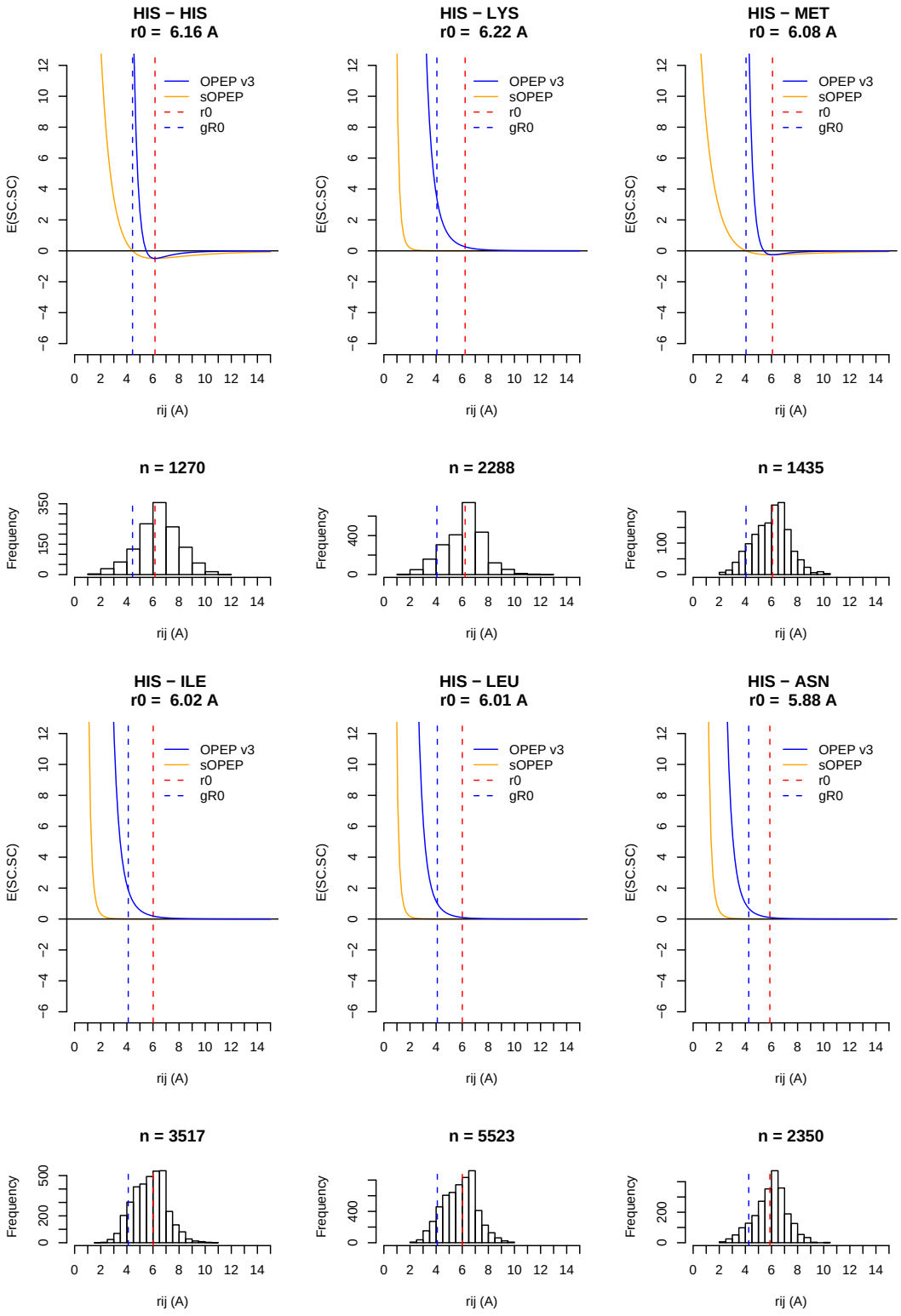
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



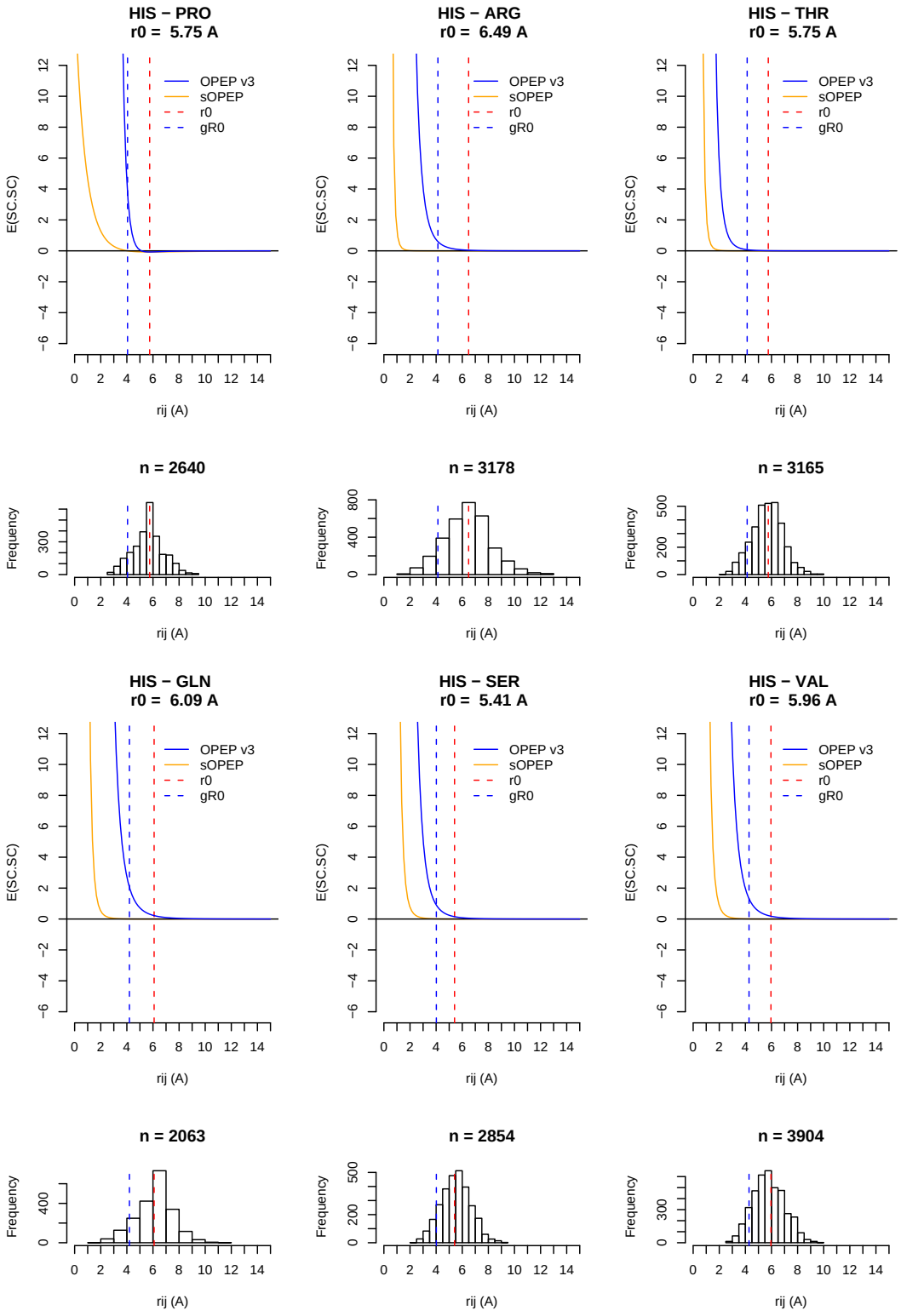
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



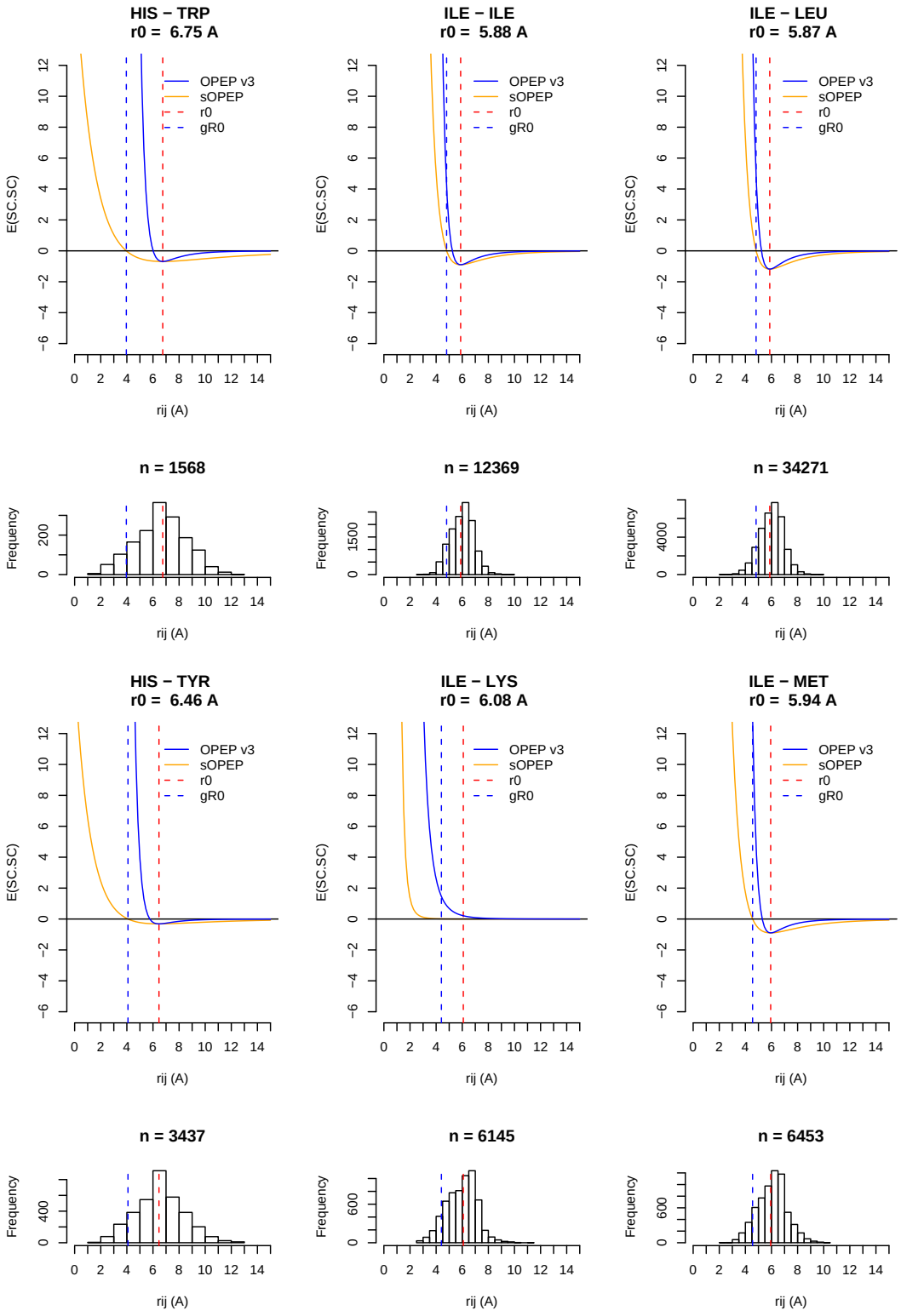
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



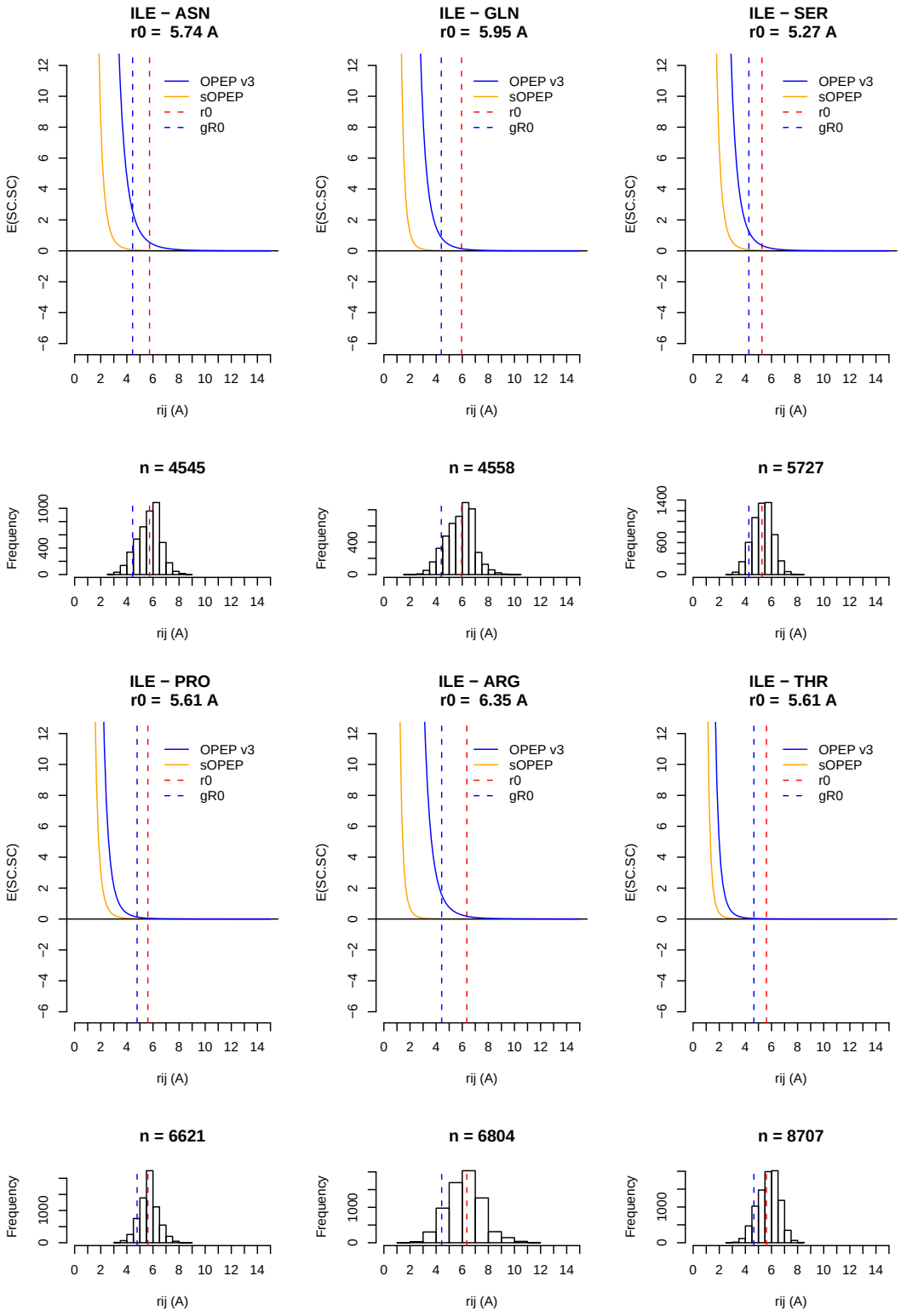
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



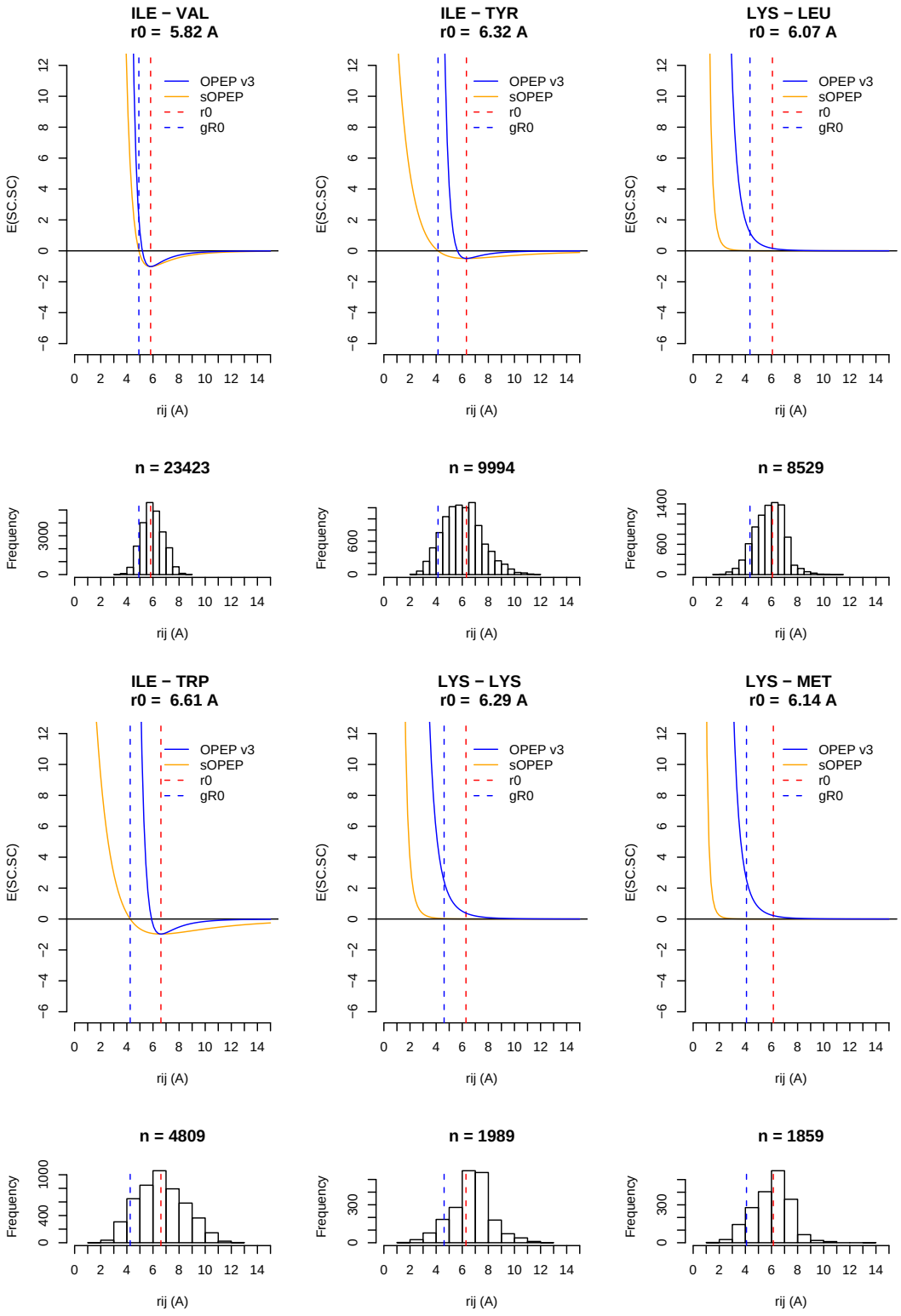
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



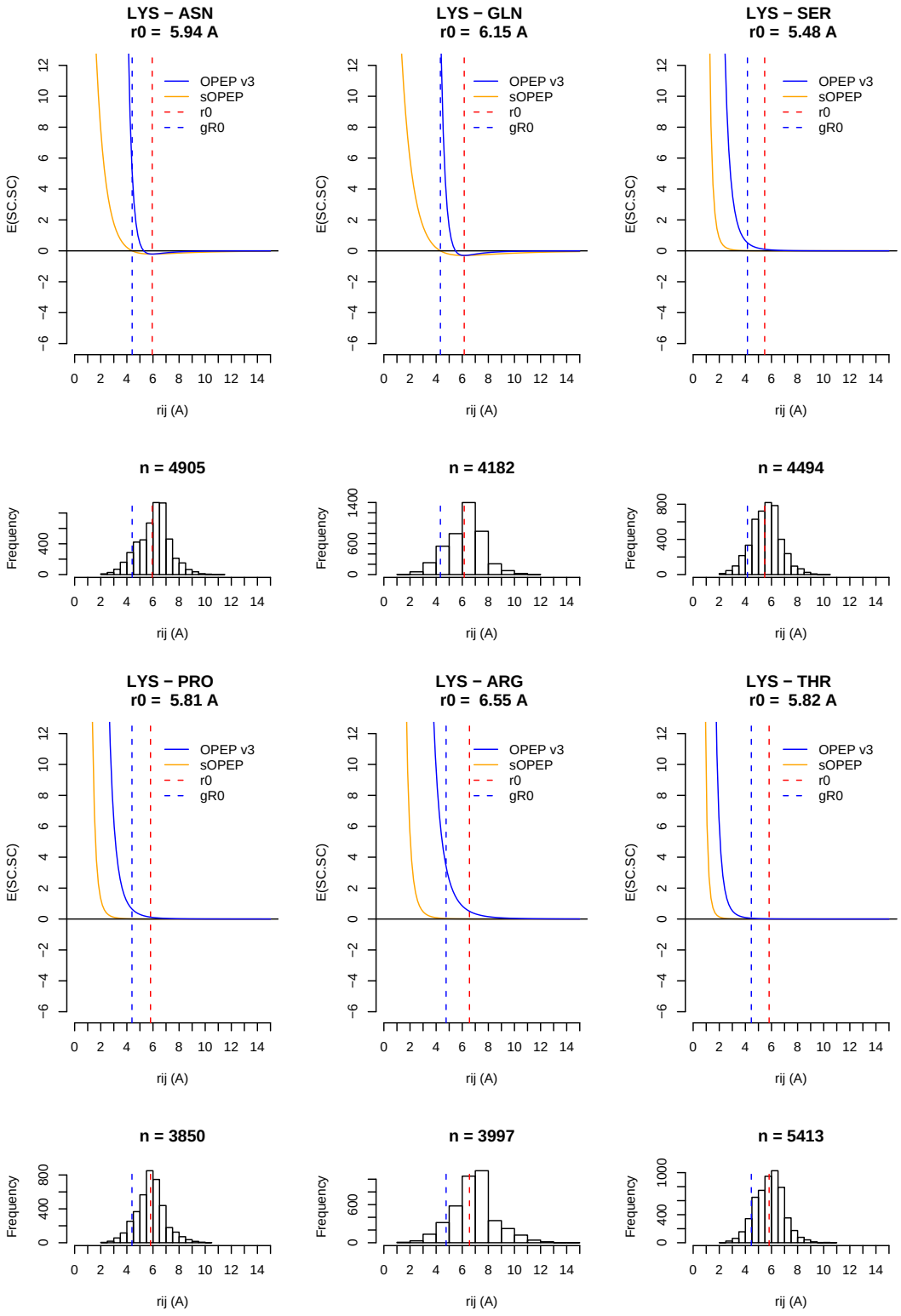
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



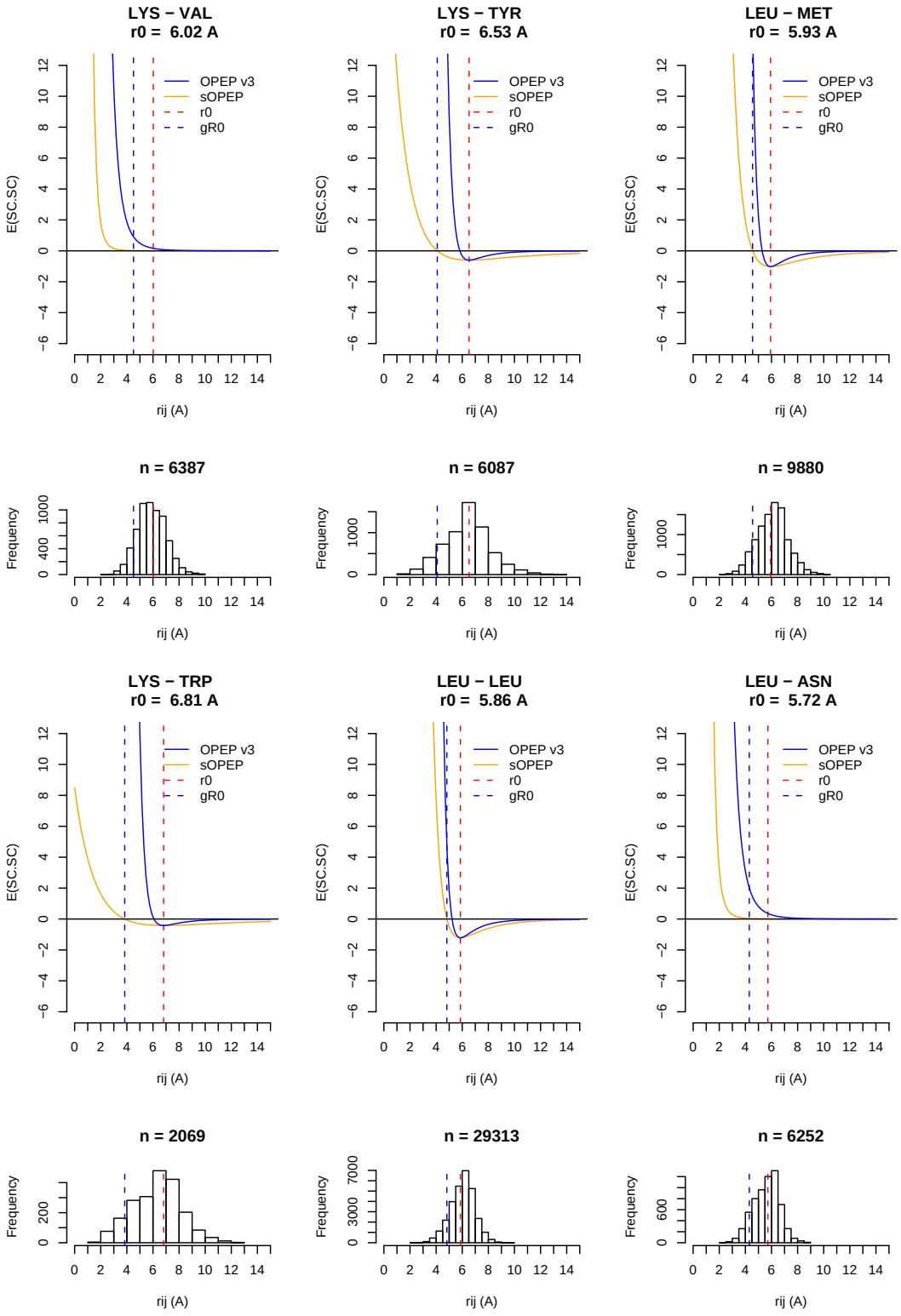
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



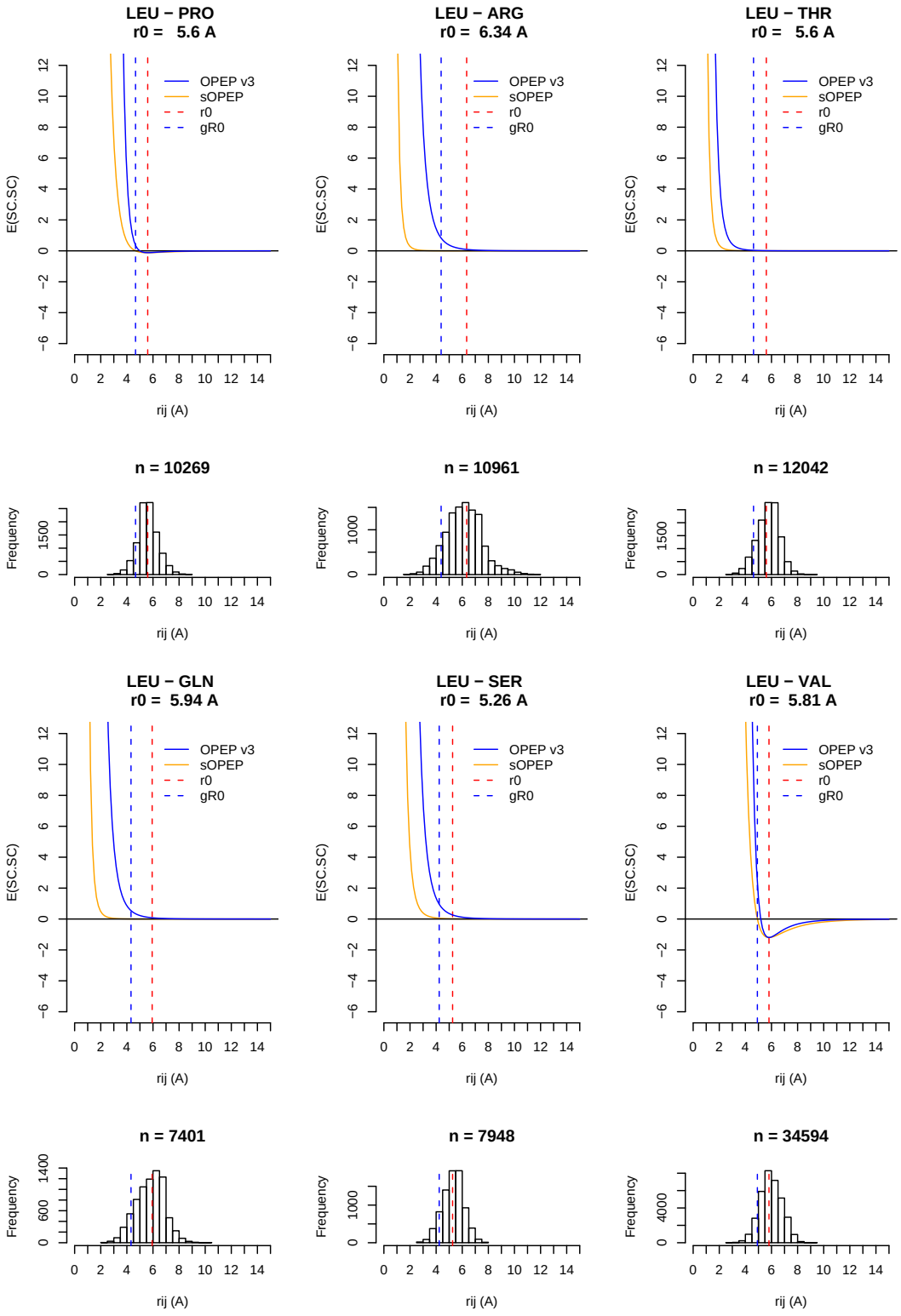
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



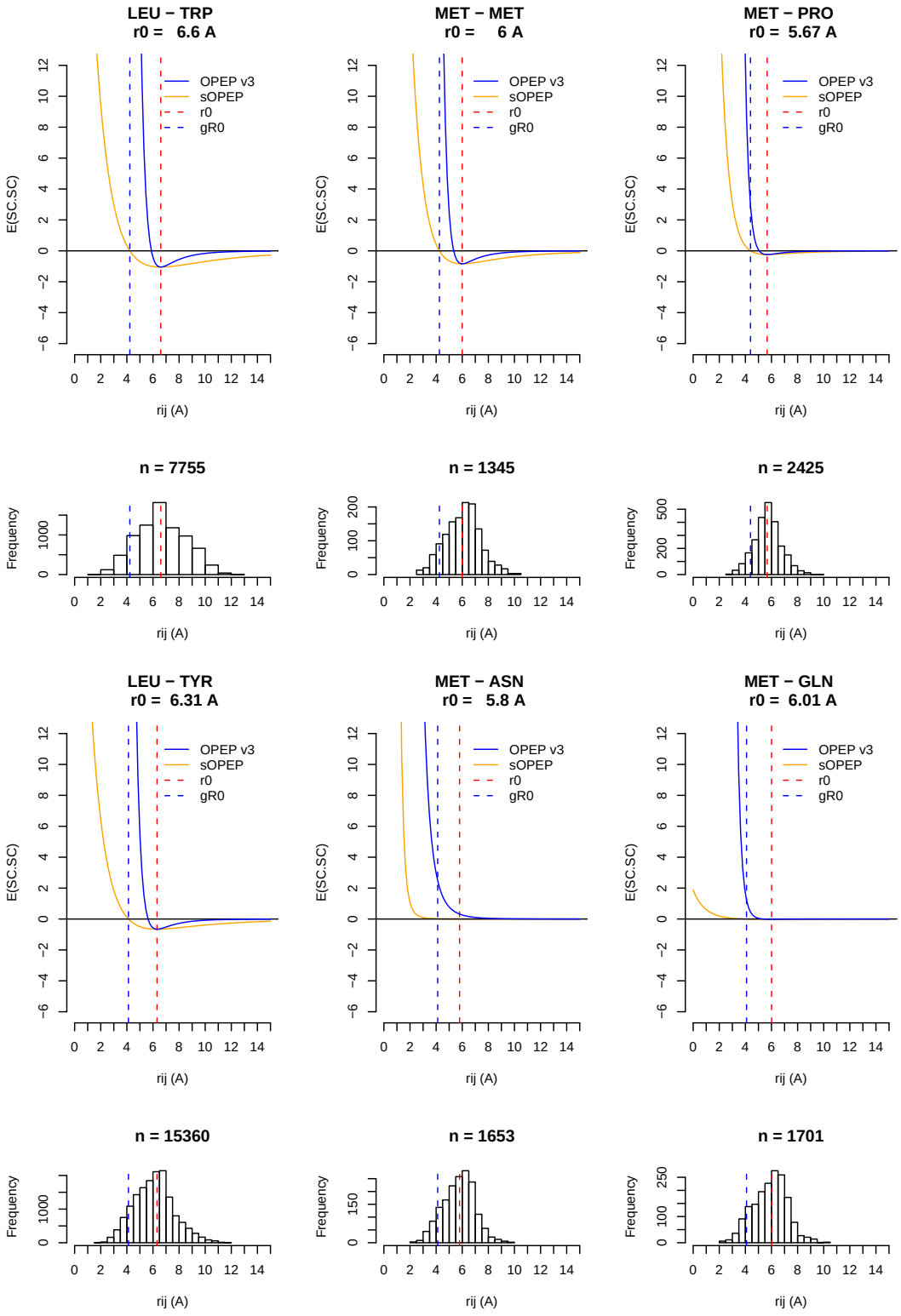
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



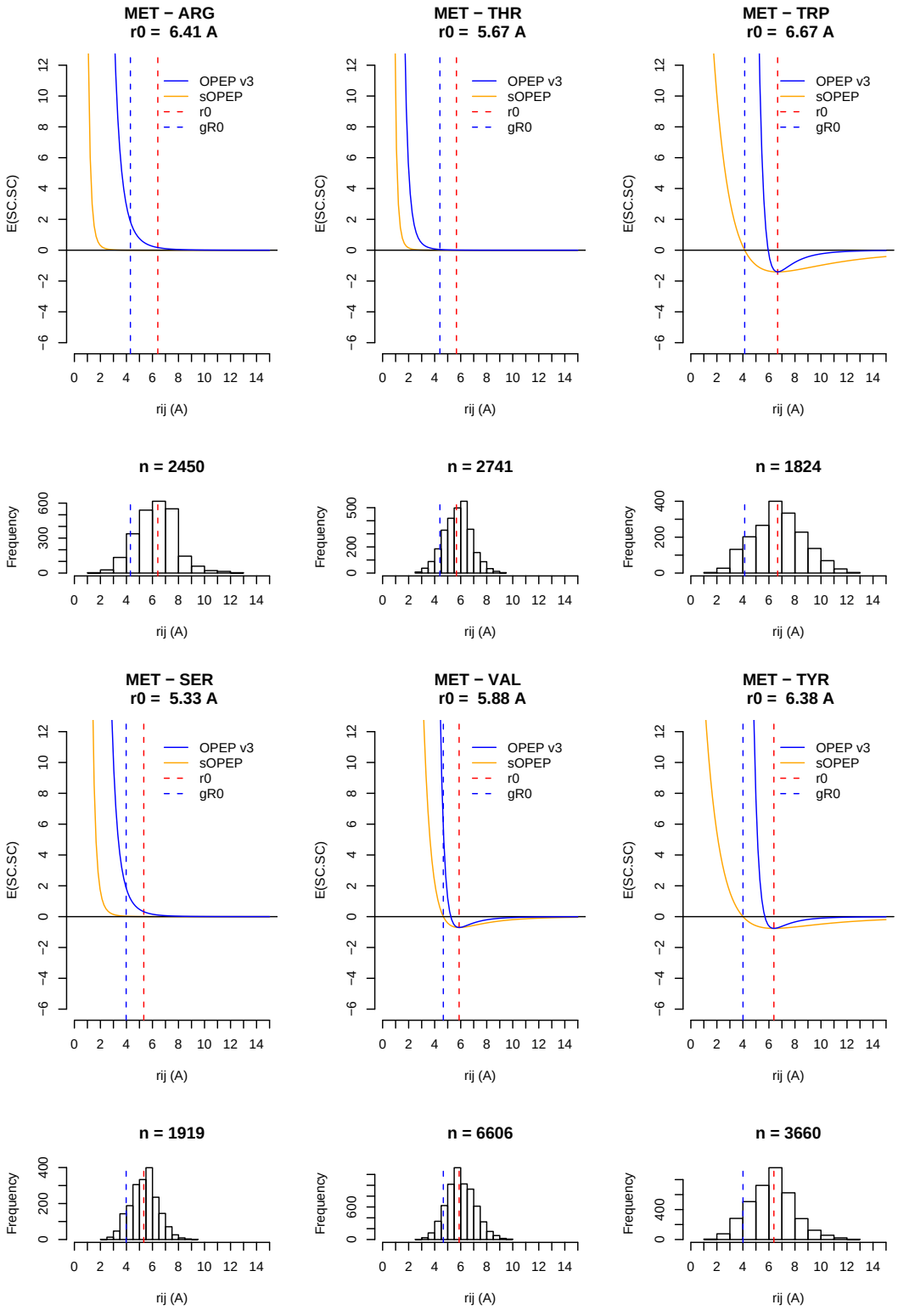
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



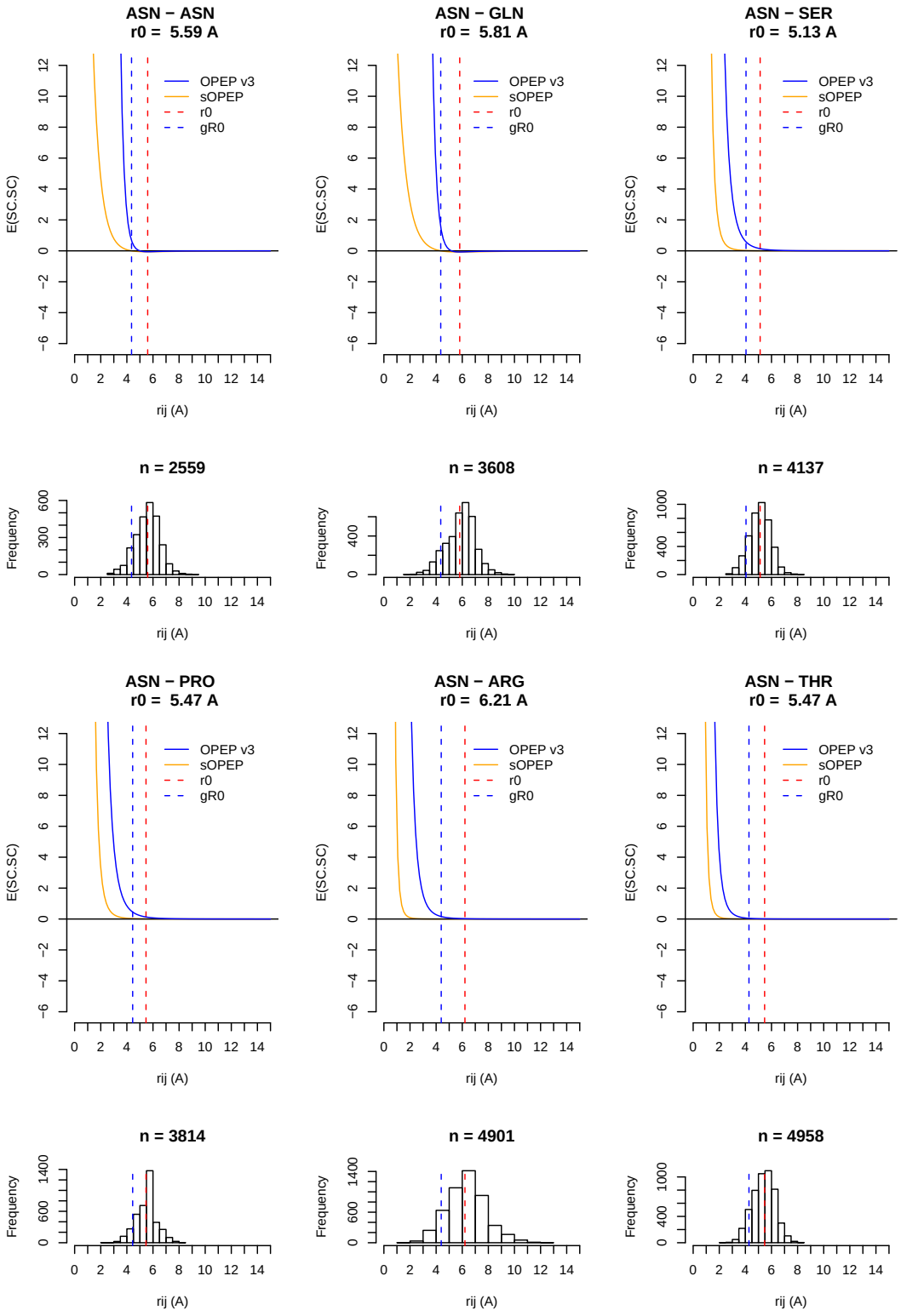
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



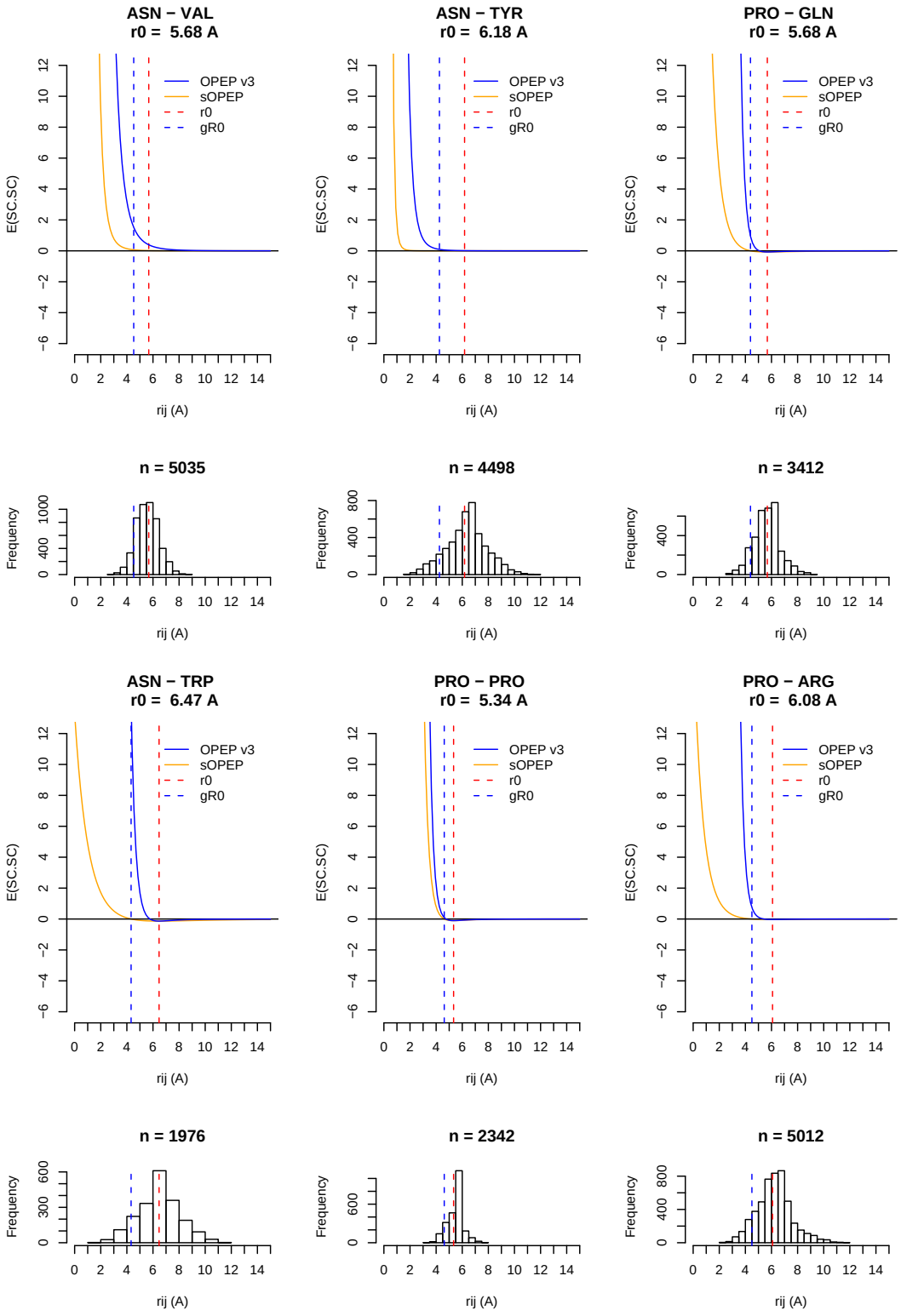
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



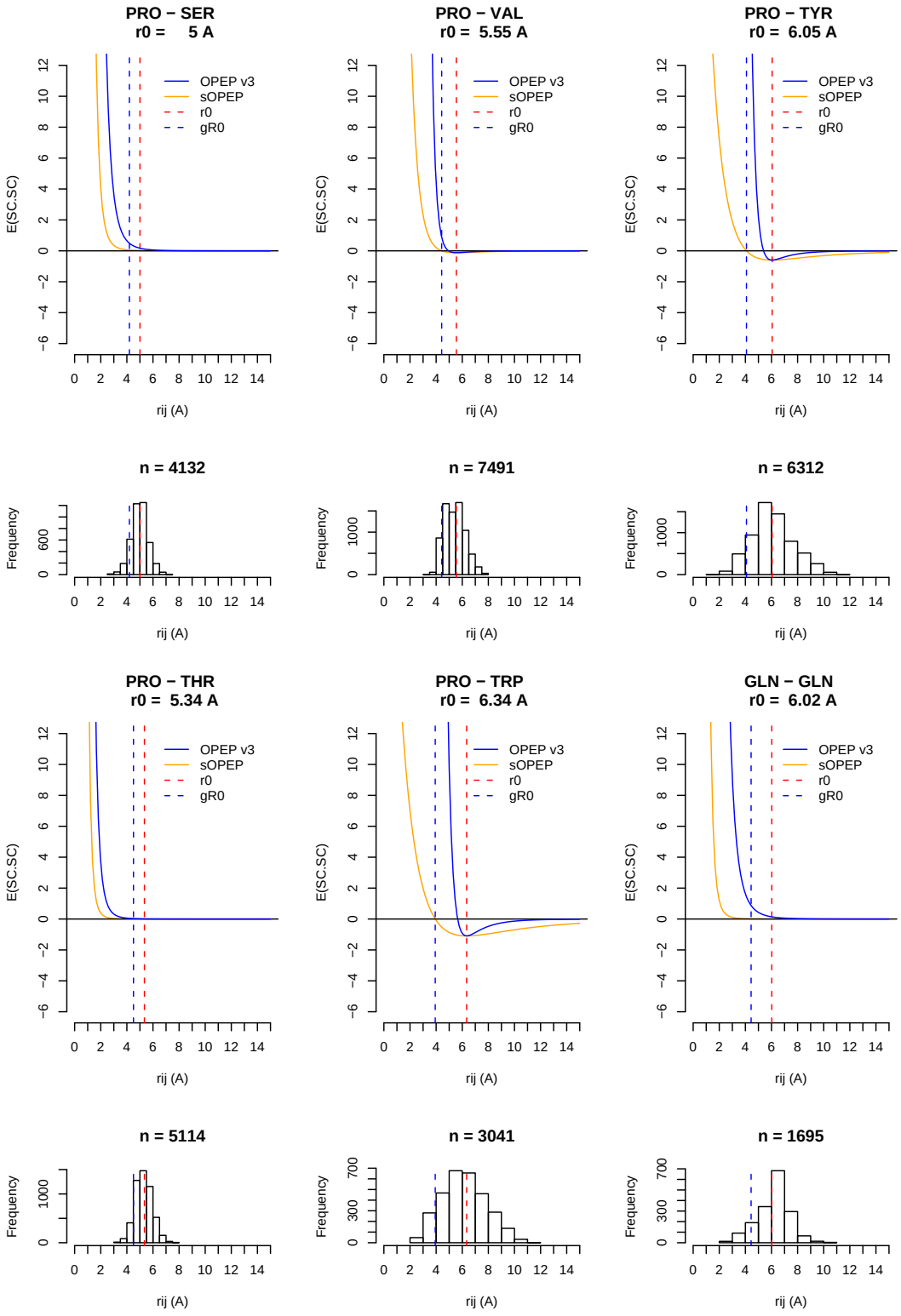
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



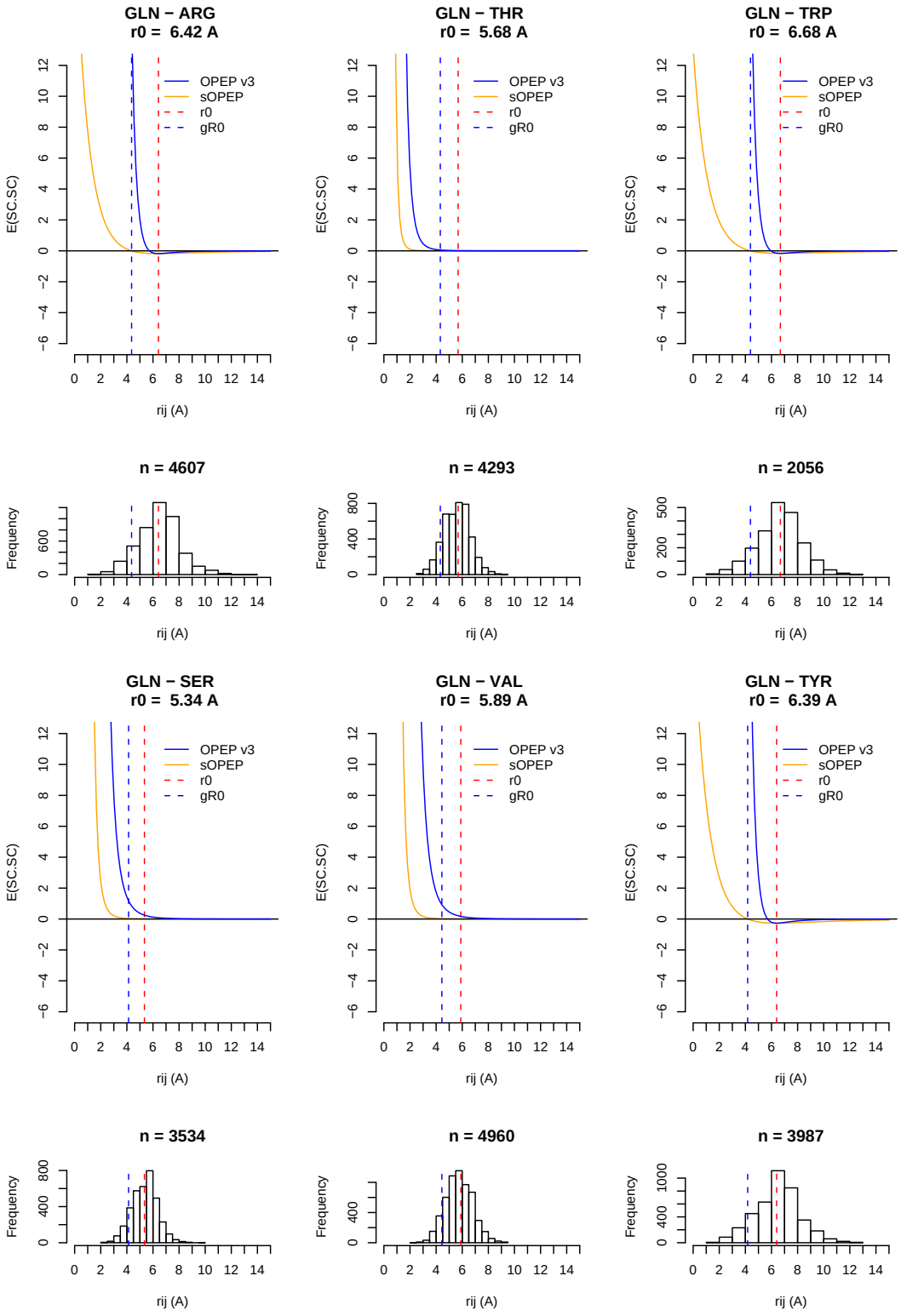
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



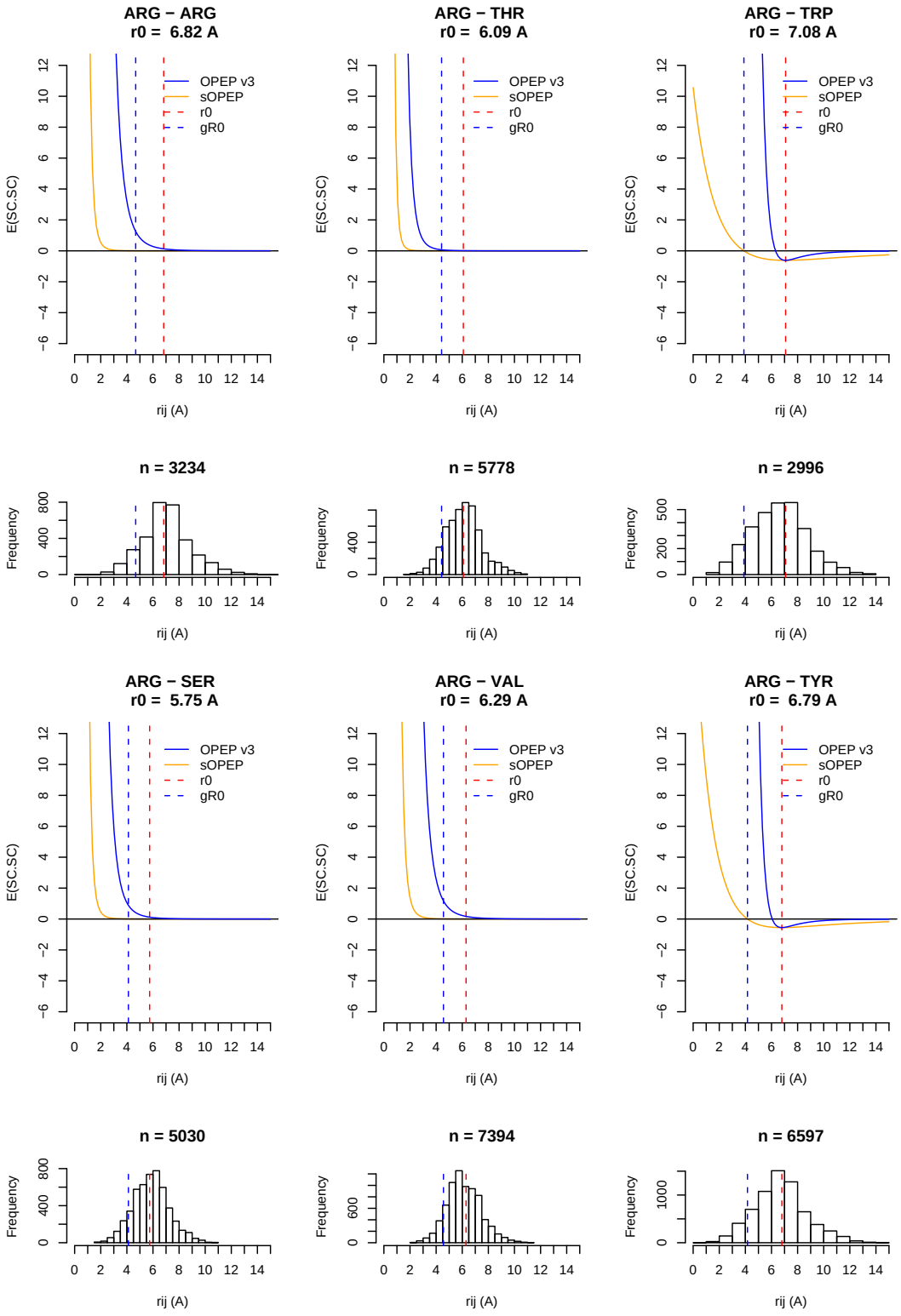
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



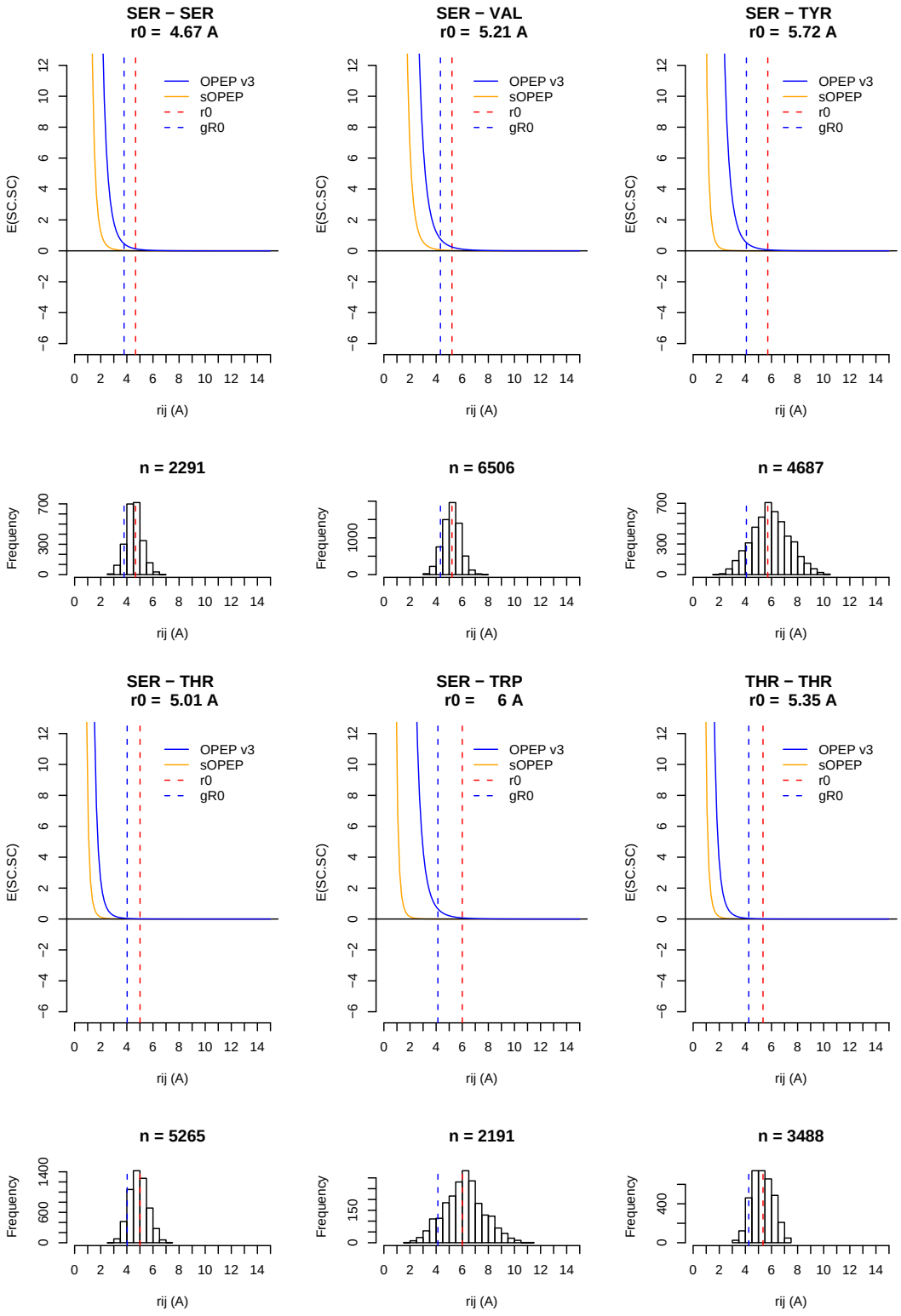
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



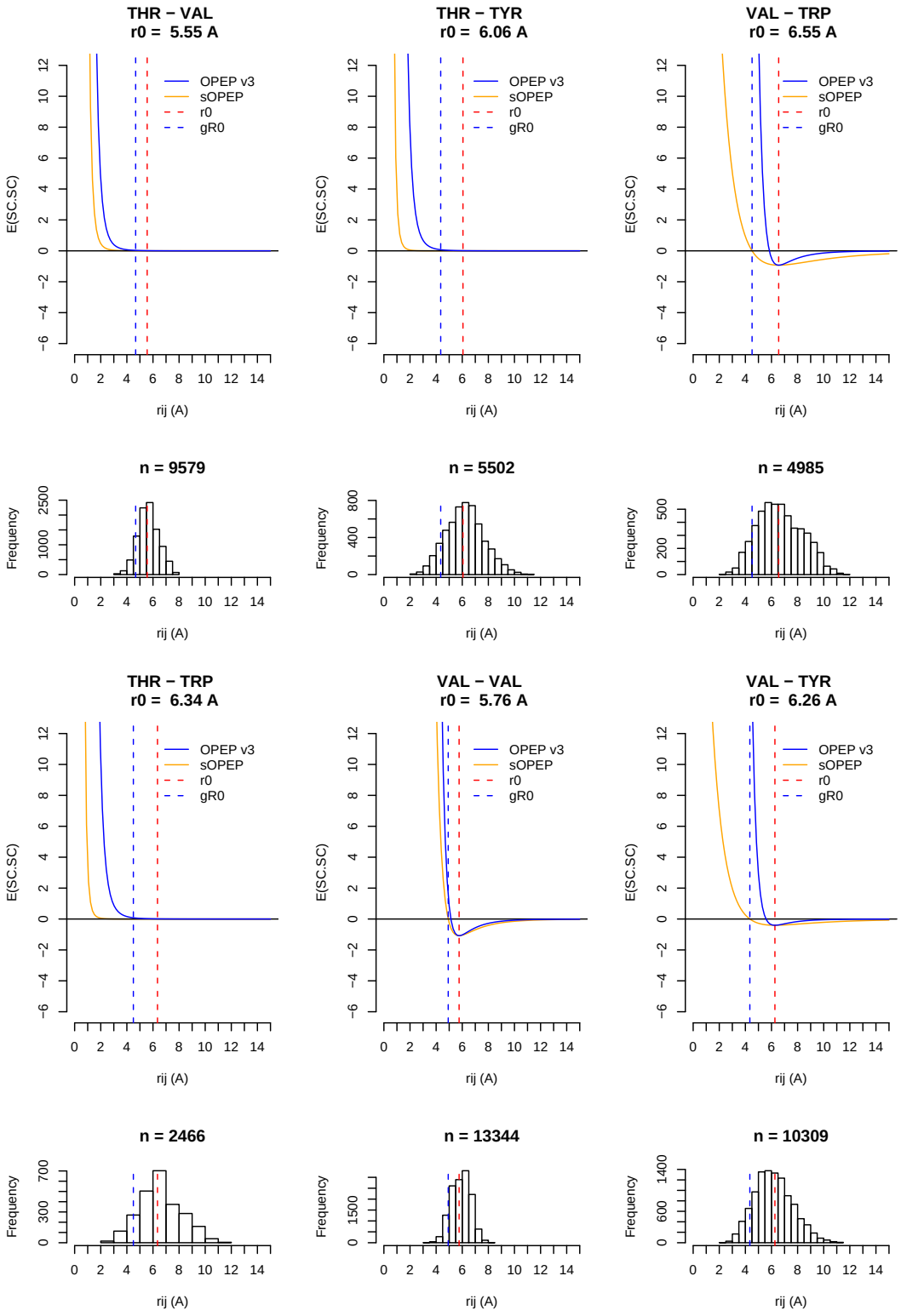
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



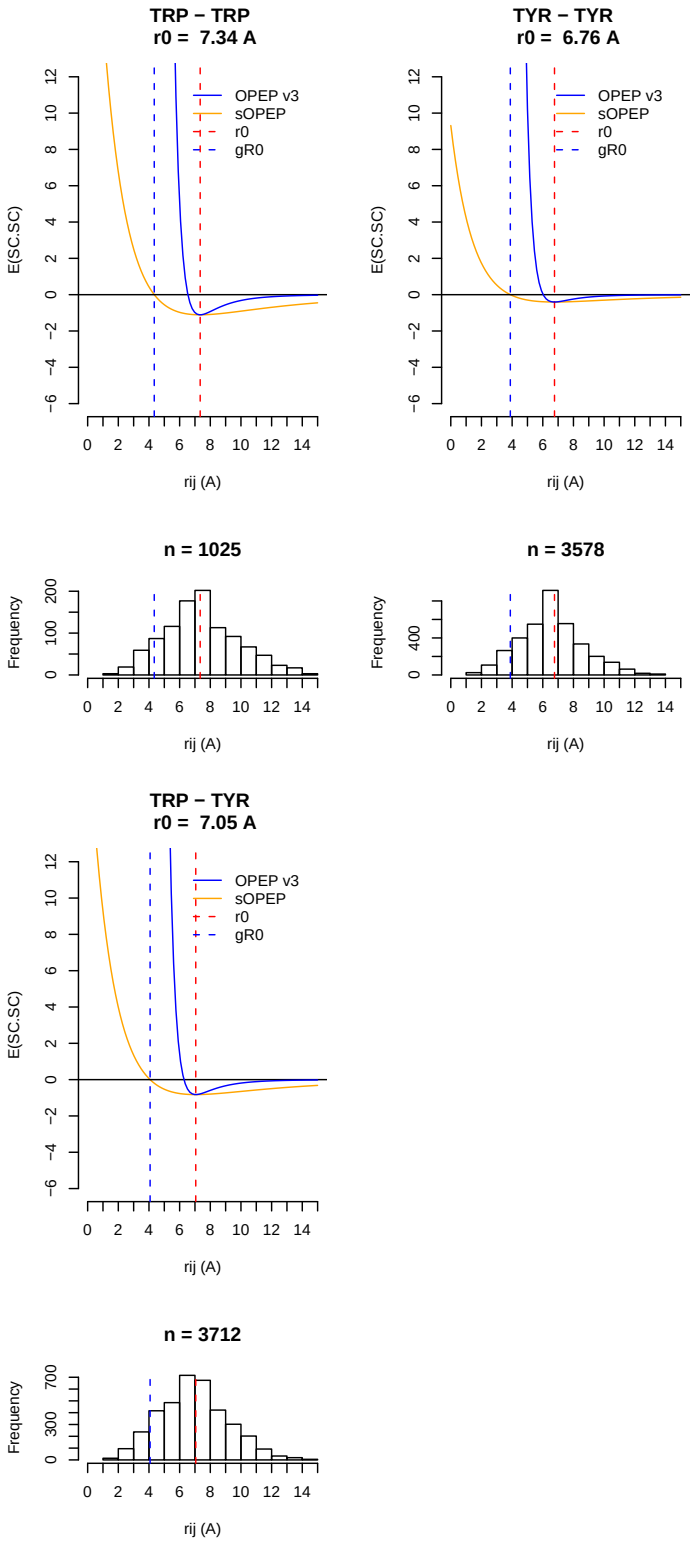
Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).



Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).

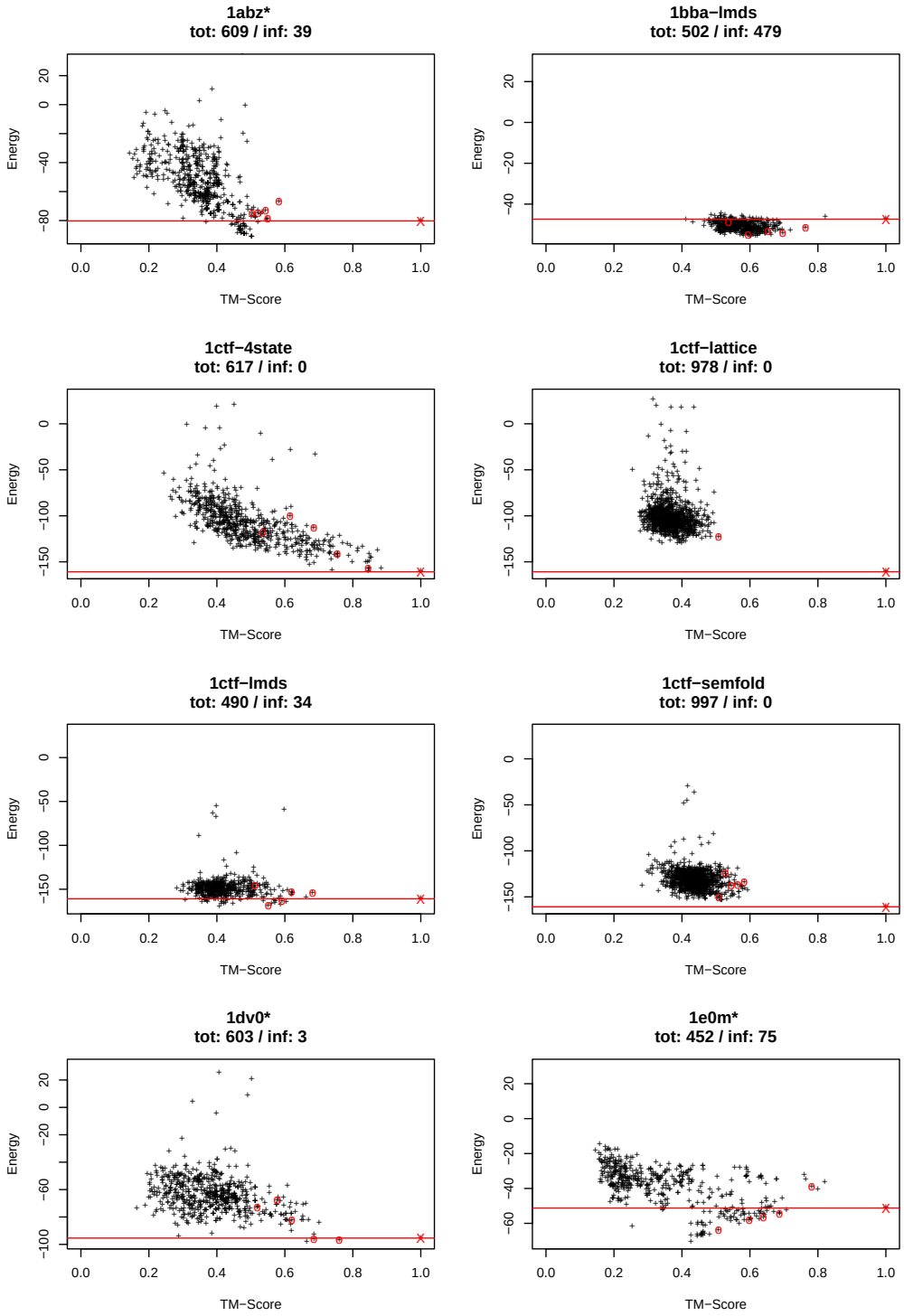


Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).

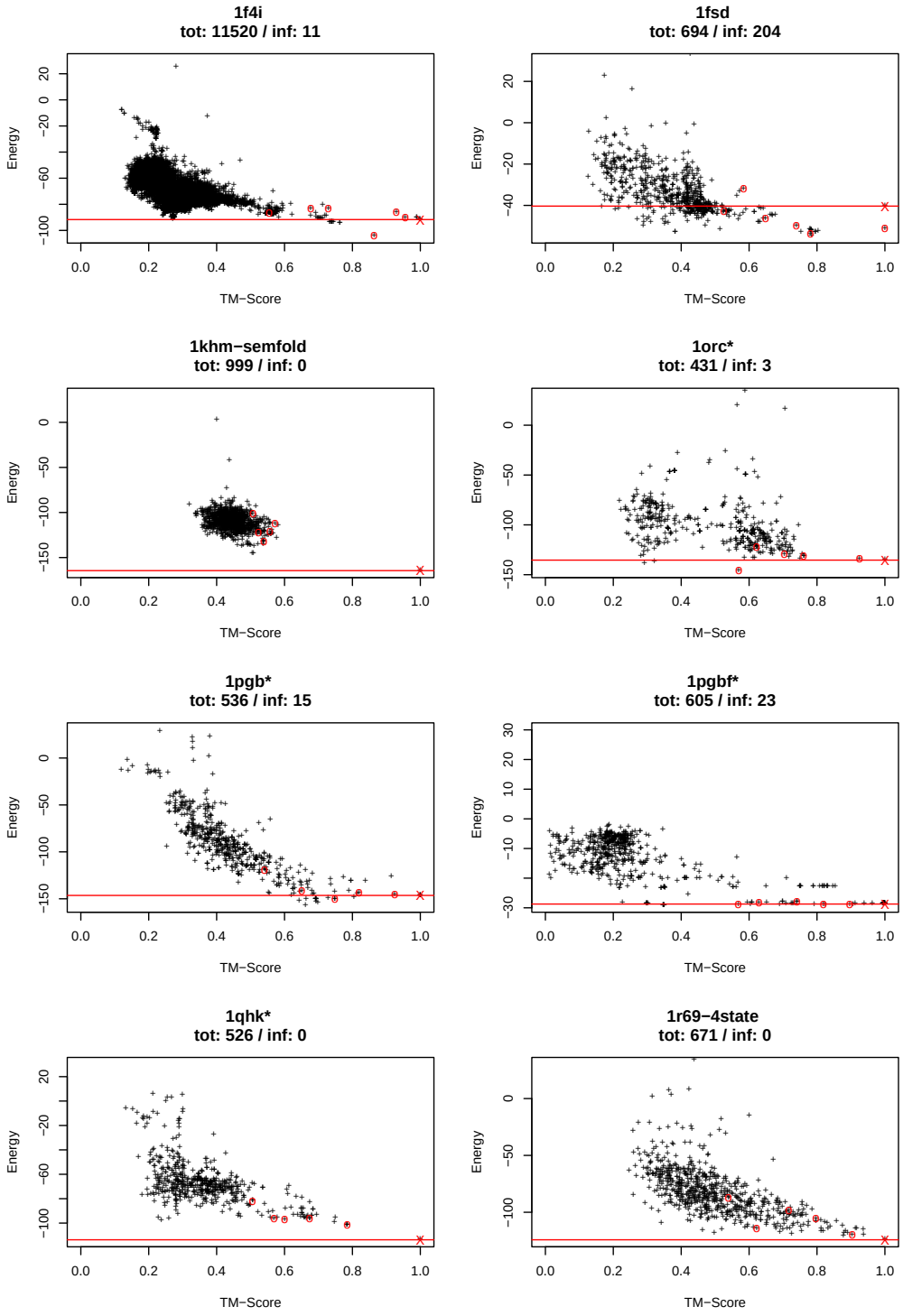


Tab. A.8: sOPEP v2.0 : le nouveau potentiel CL-CL. Pour chaque type d'interaction de type CL-CL sont présentés, le potentiel associé dans sa nouvelle formulation (en orange), et avec la formulation OPEP v3 (en bleu). Les traits discontinus verticaux correspondent aux valeurs $gR0$ (bleu) et r_0 respectivement (rouge).

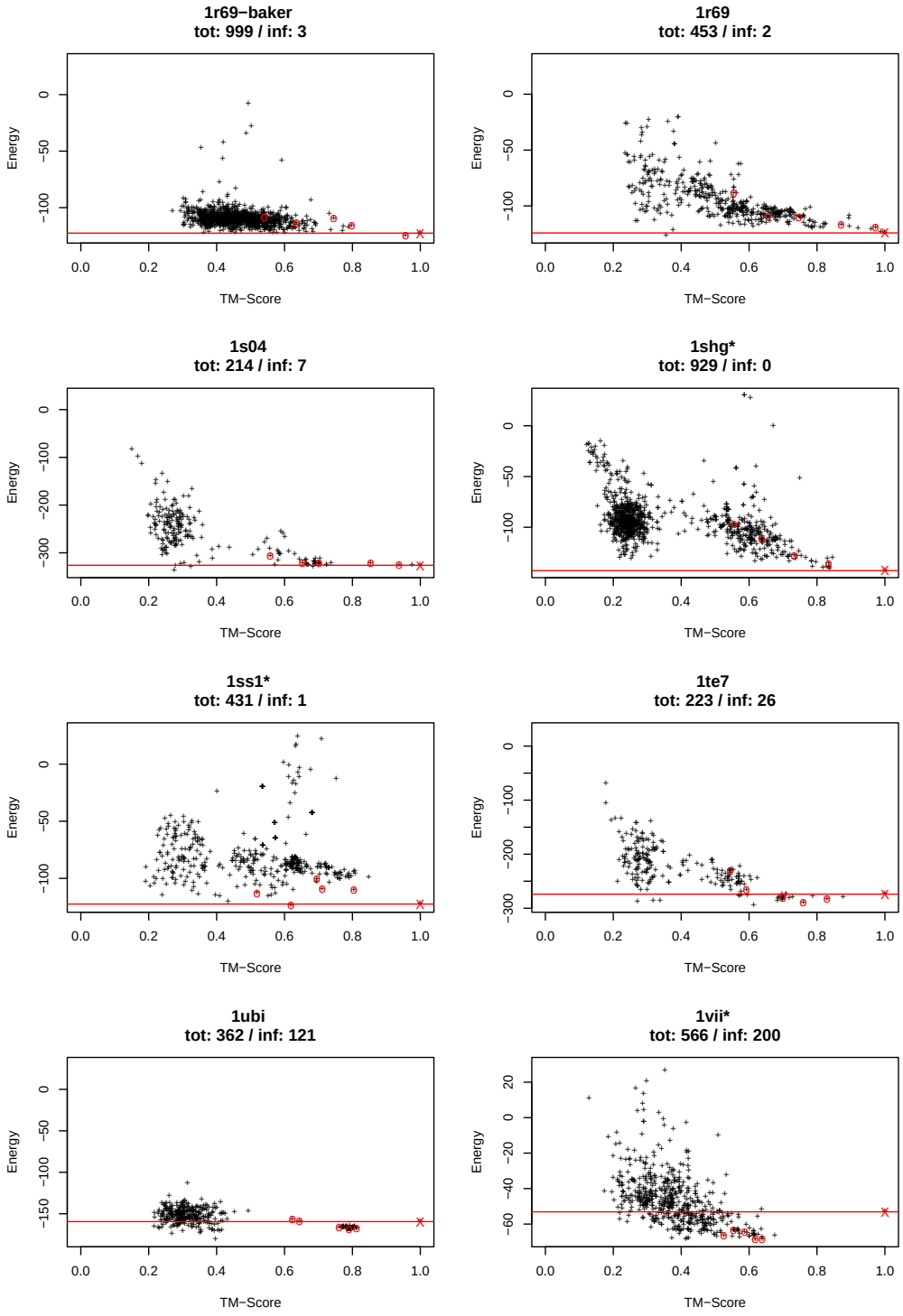
A.6 Le pouvoir discriminant de sOPEP v2.1



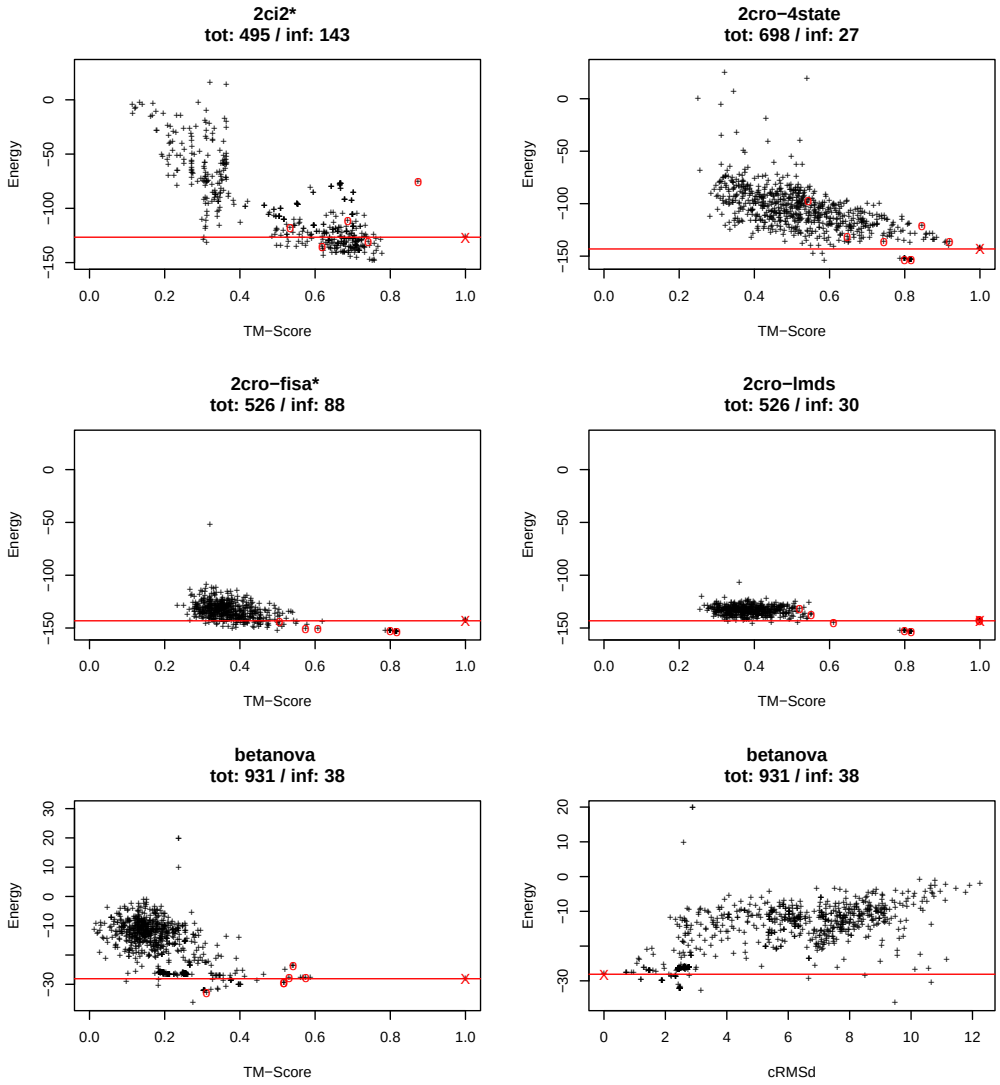
Tab. A.9: sOPEP v2.1 : énergie *versus* TM-score. Voici les graphiques obtenus de l'energie *versus* le TM-score avec sOPEP v2.1. Les cibles utilisées pendant le processus d'apprentissage sont notées d'une étoile. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd (voir dernière page de cette table).



Tab. A.9: sOPEP v2.1 : énergie *versus* TM-score. Voici les graphiques obtenus de l'energie *versus* le TM-score avec sOPEP v2.1. Les cibles utilisées pendant le processus d'apprentissage sont notées d'une étoile. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd (voir dernière page de cette table).



Tab. A.9: sOPEP v2.1 : énergie *versus* TM-score. Voici les graphiques obtenus de l'energie *versus* le TM-score avec sOPEP v2.1. Les cibles utilisées pendant le processus d'apprentissage sont notées d'une étoile. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd (voir dernière page de cette table).



Tab. A.9: sOPEP v2.1 : énergie *versus* TM-score. Voici les graphiques obtenus de l'energie *versus* le TM-score avec sOPEP v2.1. Les cibles utilisées pendant le processus d'apprentissage sont notées d'une étoile. Les structures PSN ayant servi à l'apprentissage sont entourées d'un cercle rouge. La ligne rouge horizontale indique l'énergie de la structure native marquée d'un x rouge. Pour la cible *betanova*, nous avons aussi tracé l'énergie *versus* le cRMSd (voir dernière page de cette table).

Annexe B

CASP7

B.1 Les alignements des cibles de modélisation par homologie

Pour chacun de ces alignements, la séquence de la cible est en première position, et la séquence de la matrice en deuxième. Les identités dans l'alignement sont en rouge.

t0283

```
>t0283
-MSFIEKMIGSLNDKREWKAMEARAKALPKEYHHAYKAIQKYMWTSGGPTDWQDTKRIFG
GILDLFEEGAAEGKKVTDLTGEDVAAFCDELMKDTKTWMDKYRTKLNDSIGRD
>1kgnA
EAVYTNIAFMESVHAKSYSNIFMTLASTPQINEAFRWSEENENLQRKAKIIMSYYN---G
DDPLKKKVASTLLESFLFYSGFYLPMYLSSRAKLTNTADIIRLIIRDESVHG-
```

t0288

```
>t0288
GAIIYTVELKRYGGP-LGITISGTEEPFDPIIISSLTKGGLAERTGAIHIGDRILAINSS
SLKGKPLSEAIHLLQMAGETVTLKIKKQTDAQPASS
>1n7eA
SMVPGKVTLQKDAQNLIGISIGGGAQYCPCLYIVQVFDNTPAALDGTVAAGDEITGVNGR
SIKGKTKVEVAKMIQEVKGEVTIHYNKLQYYKV---
```

t0297

```
>t0297
---MAVQLLENWLLKEQEKIQTKYRHLNHISVVEPNILFIGDSIVEYYP----LQELFGT
SKTIVNRGIRGYQTGLLLENLDA-HLYGGAVDKIFLLIGTNDIGKDVPVNEALNNLEAII
QSVARDYPLTEIKLLSILPVNEREEYQQAVYIRSNEKIQNWNQAYQ-ELASAYMQVEFVP
VFDCLTDQAGQLKKEYTTDG-LHLSIAGYQALSKSLKDYLY
>1es9A
ENPASKPTPVQDVQGDGKWMSLHHRFVADSKDKEPEVVFIGDSLVQLMHQCEIWRELFSP
-LHALNFGIGGDSTQHVLWRLENGELEHIRPKIVVVWVGTNN--HGHTAEQVTGGIKAIV
QLVNERQPQARVVVLGLLPRGQHPNPLREKNRRVNELVRAALAGHPR--------AHFLD
ADPGFVHSDGTIS-HHDMYDYLHLSRLGYTPVCRALHSLLL
```

t0302

```
>t0302
SM-------VSPEEAVKWGESFDKLLSHRDGLEAFTRFLKTEFSEENIEFWIACEDFKKS
KGPQQIHLKAKAIYEKFIQTDAPKEVNLDFHTKEVITNSITQPTLHSFDAAQSRVYQLME
QDSYTRFLKSDIYLDLMEG--
>1zv4X
--LYFQSMNPTAEEVLSWSQNFDKMMKAPAGRNLFREFLRTEYSEENLLFWLACEDLKKE
QNKKVIEEKARMIYEDYISIL-PKEVSLDSRVREVINRNLLDPNPHMYEDAQLQIYTLMH
RDSFPRFLNSQIYKSFVESTA
```

t0305

```
>t0305
Y--FQSM-KQFVKHIGELYSNNQHGFSEDFEEVQRCTADMNITAEHSNHPENKHKNRYIN
ILAYDHSRVKLRPLPGKDSK--HSDYINANYVD----GYNK-AKAYIATQGPLKSTFEDF
WRMIWEQNTGIIVMITNLVEKGRRKCDQYWPTENSEE-YGNIIVTLKSTKIHAC--YTVR
RFSIRNTKVKKGQKGNPKGRQNERVVIQYHYTQWPDMGVPEYALPVLTFVRRSSAAR--M
PETGPVLVHCSAGVGRTGTYIVIDSMLQQIKDKST---VNVLGFLKHIRTQRNYLVQTEE
QYIFIHDALLEAI------LG
>1fprA
WEEFESLQKQEVK--------NLH---------QRL---------EGQRPENKGKNRYKN
ILPFDHSRVILQ---GRDSNIPGSDYINANYIKNQLLGPDENAKTYIASQGCLEATVNDF
WQMAWQENSRVIVMTTREVEKGRNKCVPYWPEVGMQRAYGPYSVT-------NCGEHDTT
EYKLRTLQVSPLDNGDLI-----REIWHYQYLSWPDHGVPSEPGGVLSFLDQINQRQESL
PHAGPIIVHSSAGIGRTGTIIVIDMLMENISTKGLDCDIDIQKTIQMVRAQRSGMVQTEA
QYKFIYVAIAQFIETTKKKLE
```

t0308

```
>t0308
EVHVLCLGLDNSGKTTIINKLKPSNAQSQNILPTIGFSIEKFKSSSLSFTVFDMSGQGRY
RNLWEHYYKEGQAIIFVIDSSDRLRMVVAKEELDTLLNHPDIKHRRIPILFFANKMDLRD
AVTSVKVSQLLCLENIK-DKPWHICASDAIKGEGLQEGVDWLQDQI
>1mr3F
EMRILMVGLDGAGKTTVLYKLKLG--EVITTIPTIGFNVETVQYKNISFTVWDVGGQDRI
RSLWRHYYRNTEGVIFVIDSNDRSRIGEAREVMQRMLNEDELR--NAVWLVFANKQDLPE
AMSAAEITEKLGLHS-IRNRPWFIQSTCATSGEGLYEGLEWLSNNL
```

t0315

```
>t0315
MLIDTHVHLNDEQYDDDLSEVITRAREAGVDRMFVVGFNKSTIERAMKLIDEYDFLYGII
GWHPVDAIDFTEEHLEWIESLAQHPKVIGIGEMGLDYHWDKSPADVQKEVFRKQIALAKR
LKLPIIIHNREATQDCIDILLEEHAEEVGGIMHSFSGSPEIADIVTNKLNFYISLGGPVT
FKNAKQPKEVAKHVSMERLLVETDAPYLSPHPYRGKRNEPARVTLVAEQIAELKGLSYEE
VCEQTTKNAEKLFNLNS
>1j6oA
-MVDTHAHLHFHQFDDDRNAVISSFEENNIEFVVNVGVNLEDSKKSLDLSKTSDRIFCSV
GVHPHDAKEVPEDFIEHLEKFAKDEKVVAIGETGLDFFRNISPAEVQKRVFVEQIELAGK
LNLPLVVHIRDAYSEAYEILRTESLPEKRGVIHAFSSDYEWAKKFID-LGFLLGIGGPVT
YPKNEALREVVKRVGLEYIVLETDCPFLPPQPFRGKRNEPKYLKYVVETISQVLGVPEAK
VDEATTENARRIFL---
```

t0317

```
>t0317
MGTSEAAPPPFARVAPALFIGNARAAGATELLVRAGITLCVNVSRQQPGPRAPGVAELRV
PVFDDPAEDLLTHLEPTCAAMEAAVRDGGSCLVYCKNGRSRSAAVCTAYLMRHRGHSLDR
AFQMVKSARPVAEPNLGFWAQLQKYEQTLQAQAILPREPIDPE
>1m3gA
-------QGGPVEILPYLFLGSCSHSSDLQGLQACGITAVLNVSASCPNHFEGLFRYKSI
PVEDNQMVEISAWFQEAIGFIDWVKNSGGRVLVHSQAGISRSATICLAYLMQSRRVRLDE
AFDFVKQRRGVISPNFSFMGQLLQFETQVLCH-----------
```

t0322

```
>t0322
MSDDLTDAQTAAIPEGFSQLNWSRGFGRQIGPLFEHREGPGQARLAFRVEEHHTNGLGNC
HGGMLMSFADMAWGRIIS--LQKSYSWVTVRLMCDFLSGAKLGDWVEGEGELISEEDMLF
TVRGRIWA-GERTLITGTGVFKALSARKPRPGELAYKEEA
>1vh5A
-SLIWKRKITLEALNAMGE-----GNMVGFLDIRFEHIGDDTLEATMPVDSRTKQPFGLL
HGGASVVLAESIGSVAGYLCTEGEQKVVGLEINANHVRSAREG-RVRGVCKPLHLGSRHQ
VWQIEIFDEKGRLCCSSRLTTAILE---------------
```

t0369

```
>t0369
MTDWQQALDRHVGVGVRTTRDLIRLIQPEDWDKRPISGKRSVYEVAVHLAVLLEADLRIA
TGATADEMAQFYAVPVLPEQLVDRLDQSWQYYQDRLMADFSTETTYWGVTDSTTGWLLEA
AVHLYHHRSQLLDYLNLLGYDIKLDLFE
>2f22A
-MDTNGVLY-----AANMTNALAKEIPESKWDIQLIPELGTLRKLFIHIV---RVRDVYR
DGLKTGSIKFPGRLASDEHRLLDELERSMEELVFEFKQTTFNSIKMGENYLSIMELLGTV
IQHEGIHQGQYYVALKQSGINL------
```

t0373

```
>t0373
MPTNQDLQLAAHLRSQVTTLTRRLRREAQADPVQFSQLVVLGAIDRLGGDVTPSELAAAE
RMRSSNLAALLRELERGGLIVRHADPQDGRRTRVSLSSEGRRNLYGNRAKREEWLVRAMH
ACLDESERALLAAAGPLLTRLAQFEEP-
>2a61A
-----KQPFERILREICFMVKVEGRKVLRDFGITPAQFDILQKIYFE-GPKRPGELSVLL
GVAKSTVTGLVKRLEADGYLTRTPDPADRRAYFLVITRKGEEVIEKVIERRENFIEKITS
DLGKEKSSKILDYLKELKGVMERNFSKQ
```

Résumé : Dans l'ère post-génomique, de nombreuses protéines identifiées par leur séquence demeurent de structure inconnue, non résolue expérimentalement, et non accessibles aux méthodes de modélisation comparative. L'objet de mon travail de thèse a été d'explorer une approche de prédiction *de novo* de la structure des protéines. Cette méthode est fondée sur le concept d'alphabet structural, c'est à dire une description de la structure locale des protéines utilisant un nombre réduit de conformations prototypes. A partir de la seule séquence en acides aminés de la structure à prédire, nous avons mis en place une méthode de prédiction de fragments candidats de taille variable, couvrant plus de 98,6% de la structure de la protéine pour une taille moyenne 6,7 résidus. Les fragments prédits nous permettent d'approximer les structures protéiques avec une précision moyenne de 2,2 Å. L'assemblage de ces fragments est réalisé par un algorithme glouton. Le champ de force OPEP a été optimisé puis implémenté dans l'algorithme glouton pour évaluer la pertinence des modèles générés. L'évaluation, en aveugle, de la méthode a été réalisée, pour la première fois, lors de notre participation à CASP7, ce qui nous a permis d'identifier les faiblesses de la méthode. A l'heure actuelle, l'amélioration du champ de force et de la procédure d'assemblage des fragments, nous permet, dans certains cas, de donner des résultats au niveau ou meilleurs que les serveurs réputés du domaine.

Abstract : In a post-genomic context, plenty of proteins identified by their sequence have no experimentaly resolved structure, and fall out the range of application of comparative modelling methods. The goal of my PHD thesis has been to explore a new *de novo* protein structure prediction approach. This approach is based on the concept of structural alphabet, *i.e.* of a local description of protein architecture by using a small number of prototype conformations. Starting from the amino acids sequence of the protein to model, we have developped a candidate fragments prediction method covering more than 98.6% of the protein structure with an average length of 6.7 residues. This set of predicted fragments can approximate the protein structures with a precision of less than 2.2 Å. A greedy algorithm have been developped in the laboratory to assemble fragments. The OPEP force field has been optimized and then implemented in the greedy assembling procedure to evaluate the relevance of the generated models. Our participation to the CASP7 experiment came out some weaknesses of the method. For now, the improvement of the OPEP force field and the fragment assembly procedure leeds us to generate, in some cases, models as relevant or better than other famous protein structure prediction servers.