Université de Montréal

Méthodes de simulations moléculaires accélérées: application et développement

par
Jean-François St-Pierre

Département de Biochimie (Bio-informatique)
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en Bio-informatique

Décembre, 2006

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Méthodes de simulations moléculaires accélérées: application et développement

présenté par:

Jean-François St-Pierre

a été évalué par un jury composé des personnes suivantes:

François Major
président-rapporteur

Normand Mousseau
directeur de recherche

Radu Ion Iftimie
membre du jury

Mémoire accepté le 6 mars 2007

RÉSUMÉ

L'étude *in silico* du repliement des protéines, tout comme l'étude de leur flexibilité, est encore aujourd'hui limitée par les capacités matérielles des ordinateurs modernes et par l'efficacité des méthodes de simulation employées. Alors que certaines techniques de distribution des simulations de dynamique moléculaire ont été développées afin de permettre une exécution parallélisée de l'étude de systèmes complexes, d'autres techniques se concentrent sur la simplification des processus d'exploration des conformations. La technique d'activation-relaxation (ART nouveau) fait partie de ces méthodes simplifiées qui permettent l'exploration activée de la surface énergétique du repliement des protéines. Couplée au potentiel à représentation réduite OPEP, la technique a été utilisée pour l'étude du repliement de petites protéines et de l'agrégation de peptides. Dans ce mémoire, nous vous présentons tout d'abord l'étude du repliement du domaine B de la protéine A de *Staphylococcus aureus* utilisant la méthode ART-OPEP. Cette protéine de 60 acides aminés et dont la structure compte un triplet d'hélices-α est un modèle populaire auprès des théoriciens et des expérimentateurs intéressés par les mécanismes de repliement. Nos travaux démontrent que la surface énergétique est plus complexe que prévu, étant peuplée d'entonnoirs menant à des structures aux topologies similaires à d'autres protéines. Nous présentons par la suite l'implémentation dans ART nouveau de deux algorithmes permettant de rigidifier les degrés de liberté peu flexibles que sont les longueurs de liens et les angles de valence. Nous démontrons que ces algorithmes ne permettent pas de gagner en rapidité sur la méthode ART nouveau actuelle.

Mots clés: protéine staphylococcale A, ART nouveau, OPEP, FRODA, simulation géométrique, corps rigides

ABSTRACT

The study of the folding mechanism of proteins is currently limited by the current computer technologies and by the *in silico* simulation methods. While certain techniques of workload distribution have been devised to study more complex systems, other techniques rely on the simplification of the different process involved in folding. The activation-relaxation technique (ART Nouveau) is one of those simplified activated methods that enable the navigation of the energetic surface. In conjunction with the reduced representation energy potential OPEP, ART has been used to study the folding of small proteins and the aggregation process of peptides. In this work, we present our results of the folding trajectories of the B domain of *Staphylococcus aureus'* protein A using ART-OPEP. This 60 amino acids protein forms a bundle of three α-helices and is well known and studied by both theorists and experimentalists interested in its folding process. Our results present a different picture of the energy landscape of this protein, one that is punctuated by funnels leading to topologies found in other proteins. We then continue by presenting the implementation of two rigid bodies algorithms in the ART nouveau method. These algorithms are used to eliminate the less interesting degrees of freedom that are bond lengths and bond angles. Our analysis of these methods are that they do not provide a noticeable gain in performance over the original ART nouveau.

Mots clés staphylococcal protein A, ART nouveau, OPEP, FRODA, geometrical simulations, rigid bodies

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES APPENDICES

LISTE DES SIGLES

ART	Technique d'Activation-Relaxation
BdpA	Domain B de la protéine staphylococcale A
DM	Dynamique Moléculaire
GuHCl	Hydrochloride de Guanidine
IgG	Immunoglobulin G
MC	Monte-Carlo
MM	Mécanique Moléculaire
MRSA	Methicilin Resistant Staphyloccocus aureus
Pont-H	Pont Hydrogène
QM	Mécanique Quantique
RMN	Résonance Magnétique Nucléaire
DMER	Dynamique Moléculaire avec Échange de Répliques
RMSD	Root Mean Square Distance
SpA	Protéine staphylococcale A

À ma famille.

REMERCIEMENTS

J'aimerais remercier mon directeur de recherche, le professeur Normand Mousseau, pour sa patience, son support, ainsi que pour les opportunités qu'il m'a gracieusement présentées. J'adresse un remerciement particulier à Philippe Derreumaux pour son aide généreuse et ses points de vue judicieux.

Je tiens aussi à remercier les membres de notre laboratoire ainsi que mes collègues de biochimie pour leur aide et leurs réponses à mes nombreuses questions, soit Lillianne Dupuis, Guanhong Wei, Geneviève Boucher, Kevin Smith, Myrian Grondin et Mathieu Coinçon.

Je voudrais souligner le support moral et la compréhension grandement appréciée des membres de ma famille.

Finalement, je remercie la contribution financière du programme stratégique de bourses de formation des IRSC en bio-informatique, les bourses d'excellence biT.

CHAPITRE 1

INTRODUCTION

La vie, telle que nous la connaissons, est articulée par les protéines. Les dernières évaluations du génome humain prédisent un nombre de gènes variant entre 20000 et 25000 [Ste04]. De ce nombre, la portion étant exprimée reste inconnue, mais on présume qu'une vaste majorité de ces gènes peut être transcrite en ARNm, puis traduite en protéines. Les protéines sont ubiquistes à la vie et jouent tous les rôles : elles servent de canaux traversant la membrane de la cellule ou d'intermédiaires aux messages hormonaux, de squelette structural intracellulaire, de moteurs dans les cellules musculaires, elles composent les anticorps servant à la reconnaissance des corps intrus et elles sont sécrétées dans nos cheveux. De plus, elles jouent le rôle de catalyseur impliqué dans divers cycles du métabolisme.

Bien qu'il est aujourd'hui possible de déterminer la structure native des protéines par l'étude du spectre à rayons X de la protéine cristallisée ou encore par résonance magnétique nucléaire, les outils expérimentaux d'étude du processus de repliement des protéines sont peu nombreux et ne fournissent que des données partielles. Ceci découle du fait que les protéines se replient sur des temps très courts de l'ordre des microsecondes pour les plus rapides jusqu'aux secondes pour les plus lentes. Il est donc difficile de caractériser les intermédiaires de courte durée de vie peuplant la trajectoire de repliement. Or ces intermédiaires peuvent être importants du point de vue médical. Par exemple, on croit que les oligopeptides des protéines Aβ40 et Aβ40 qui sont des intermédiaires à la formation de fibres amyloïdes de la maladie d'Alzheimer sont plus toxiques que les fibres elles-mêmes. Aussi, si le processus de repliement d'une protéine passe par un état intermédiaire stable, cet intermédiaire peut être une cible intéressante d'inhibition de la fonction de la protéine.

Le développement d'outils informatiques de simulation physique des protéines tel que la dynamique moléculaire nous a permis d'augmenter nos connaissances sur la cinétique du repliement de très courtes protéines. Cependant, malgré la crois-

sance en puissance du matériel informatique au cours des 25 dernières années, il est encore aujourd'hui impossible de simuler une courte protéine au repliement rapide dans toute la complexité du processus par des techniques traditionnelles. Pour y arriver, on utilise des méthodes de parallélisation du calcul permettant de diviser les tâches entre plusieurs processeurs, ou encore des simplifications du processus tel que les représentations réduites du nombre d'atomes de la protéine ou l'utilisation d'un solvant implicite éliminant du fait même plusieurs centaines, voir milliers de molécules d'eau. Il existe aussi les méthodes d'échantillonnage accéléré dites de Monte-Carlo qui ne tiennent pas compte de la thermodynamique du système. La technique d'activation-relaxation ART développée par le professeur Mousseau et ses collaborateurs [BM96] est une de ces méthodes. Dans sa dernière implémentation nommée ART nouveau, la méthode a été utilisée pour étudier le repliement de courtes protéines et l'agrégation de peptides. La méthode, couplée à un potentiel énergétique générique à représentation réduite OPEP [Der99], parcourt la surface énergétique des protéines à la recherche de points de selle de premier ordre reliant deux minima locaux.

Dans un premier volet de ce mémoire, nous présentons nos résultats les plus récents sur l'utilisation de ART nouveau pour étudier le repliement d'une protéine bien étudiée par les groupes de recherche intéressés par le repliement, soit le domaine B de la protéine A. Nous débutons au chapitre 2 par une introduction aux aspects biochimiques et numériques du repliement des protéines, puis au chapitre 3 une revue de la littérature sur la protéine A, suivi par l'article présentant nos résultats et d'une discussion au chapitre 5 sur l'apport scientifique de cet article.

Par la suite, nous nous intéressons à la possibilité d'optimiser la technique ART nouveau pour la rendre plus efficace. Pour y arriver, nous tentons de fixer les degrés de liberté les moins intéressants lors des changements de conformation d'une protéine, soit la longueur des liens et des angles de valence. Nous examinons au chapitre 6 deux algorithmes qui rigidifient ces coordonnées internes des protéines tout en laissant libre les angles dièdres, ces angles de torsions qui définissent la configuration d'une protéine.

CHAPITRE 2

CONCEPTS BIOCHIMIQUES ET OUTILS NUMÉRIQUES

2.1 Aspects biochimiques

2.1.1 Structure des protéines

La fonction jouée par une protéine est intimement reliée à sa structure tridimensionnelle et à sa composition atomique. Les protéines sont des polymères linéaires d'acides aminés, aussi appelés résidus, qui adoptent une conformation spatiale spécifique dite repliée. Il existe 20 acides aminés conventionnels qui partagent un même squelette carboné (Figure 2.1) et qui se distinguent par la composition de leur chaîne latérale (Tableau 2.1.1). Chaque acide aminé contribue six atomes au squelette en plus de ceux de la chaîne latérale. De ces six atomes, trois sont liés de façon linéaire. Il s'agit de l'azote du groupement aminé, du carbone-α liant un hydrogène aliphatique et la chaîne latérale, et le carbone du groupement carbonyle. Le lien unissant le groupement carbonyle d'un résidu au groupement aminé du suivant se nomme le lien peptidique. Ce lien est en fait harmonique et partage un doublet d'électrons avec l'oxygène du carbonyle. Il en résulte que les atomes des groupements aminés et carbonyles sont situés sur un même plan, nommé le plan peptidique (Figure 2.2).

La séquence d'acides aminés formant une protéine se nomme la structure primaire. Les angles dièdres ϕ et ψ du squelette sont les deux seuls degrés de liberté influençant la conformation spatiale du polymère, le ou les angles χ de la chaîne latérale ne jouant qu'un rôle important au niveau de l'encombrement stérique. À cause de l'encombrement stérique des atomes de la chaîne carbonée et des premiers atomes des chaînes latérales, certaines valeurs des angles ϕ et ψ sont préconisées. Ces jeux de valeurs donnent naissance à des éléments de structure spatiale spécifiques tels que les hélices α et les feuillets β (Figure 2.3). Le diagramme de Ramachandran donne une représentation des régions où les combinaisons d'angles

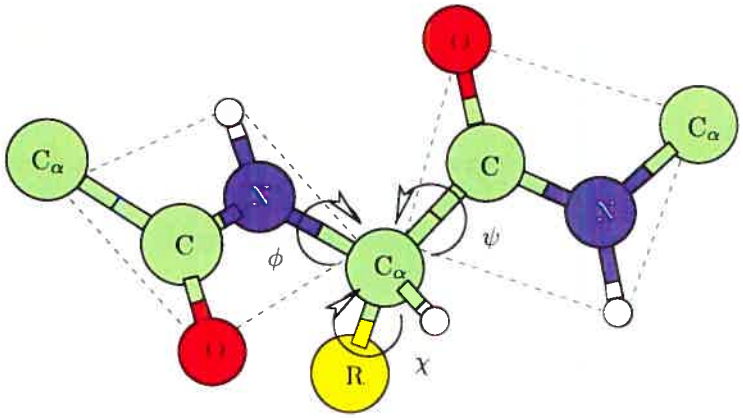

FIG. 2.1 – Squelette carboné des acides aminés. Les atomes non identifiés sont des hydrogènes. Les symboles ϕ et ψ correspondent aux angles dièdres libres de la chaîne carbonée alors que le symbole χ correspond au premier angle dièdre de la chaîne latérale. Images extraites du livre de Wales [Wal03]

dièdres sont favorisées (Figure 2.4 (a) et (b)). L'espace entre ces régions peut être peuplé de quelques paires d'angles, mais ceux-ci sont généralement énergétiquement défavorisés.

Le diagramme de Ramachandran ne s'applique pas à tous les résidus. En effet, deux résidus ont une structure les différenciant notablement des autres. Tout d'abord, la glycine a la particularité que sa chaîne latérale n'est composée que d'un seul atome d'hydrogène. Au contraire des autres acides, la glycine n'est pas un acide aminé à carbone-α chiral. De ce fait, elle ne manifeste pas le même encombrement stérique que les autres acides aminés et peut adopter une vaste gamme de combinaison d'angle ϕ et ψ (Figure 2.4 (c)). L'autre acide aminé particulier est la proline qui possède un cycle à 5 atomes comprenant le carbone-α et l'azote du groupement aminé (qui de ce fait n'a pas d'hydrogène). L'angle dièdre ϕ de ces résidus est ainsi restreint et ne peut adopter que des valeurs variant entre -80 et -100 degrés (Figure 2.4 (d)). Il crée aussi un encombrement stérique prononcé sur le résidu le précédant, limitant les valeurs d'angles dièdres que ce dernier peut occuper (Figure 2.4 (e)).

TAB. 2.1 – Nom et composition des chaînes latérales des acides aminés

Acide Aminé	Abrév.	Chaîne Latérale	Polarité	Charge	Aliphatique / Aromatique
Alanine	Ala/A	$-CH_3$	non-polaire		Aliphatique
Arginine	Arg/R	$-(CH_2)_3NHC(NH)NH_2$	polaire	basique	
Asparagine	Asn/N	$-CH_2CONH_2$	polaire	neutre	
Aspartate	Asp/D	$-CH_2COOH$	polaire	acide	
Cystéine	Cys/C	$-CH_2SH$	non-polaire		
Glutamine	Gln/Q	$-CH_2CH_2CONH_2$	polaire	neutre	
Glutamate	Glu/E	$-CH_2CH_2COOH$	polaire	acide	
Glycine	Gly/G	$-H$	non-polaire		Aliphatique
Histidine	His/H	$-CH2-C3H3N2$	polaire	basique	
Isoleucine	Ile/I	$-CH(CH_3)CH_2CH_3$	non-polaire		Aliphatique
Leucine	Leu/L	$-CH_2CH(CH_3)_2$	non-polaire		Aliphatique
Lysine	Lys/K	$-(CH_2)_4NH_2$	polaire	basique	
Méthionine	Met/M	$-CH_2CH_2SCH_3$	non-polaire		Aliphatique
Phénylalanine	Phe/F	$-CH_2C_6H_5$	non-polaire		Aromatique
Proline	Pro/P	$-CH_2CH_2CH_2-$	non-polaire		
Serine	Ser/S	$-CH_2OH$	polaire		Aliphatique
Thréonine	Thr/T	$-CH(OH)CH_3$	polaire		Aliphatique
Tryptophane	Trp/W	$-CH_2C_8H_6N$	non-polaire		Aromatique
Tyrosine	Tyr/Y	$-CH_2-C_6H_4OH$	non-polaire		Aromatique
Valine	Val/V	$-CH(CH_3)_2$	non-polaire		Aliphatique

Il existe deux autres niveaux de structure utilisés pour décrire les protéines. Tout d'abord, la structure tertiaire est définie par l'assemblage des éléments discrets de structures secondaires dans une forme compacte et stable. Ce niveau de structure implique souvent des interactions hydrophobes entre des résidus séquentiellement distants stabilisant un coeur hydrophobe, mais aussi des ponts hydrogène entre les chaînes latérales de résidus polaires. Le niveau de structure quaternaire, quant à lui, est défini par l'agrégation de protéines, homogènes ou hétérogènes, dans un complexe protéinique actif. Certaines protéines n'ont pas de structure quaternaire puisqu'elles ne s'associent pas aux autres protéines.

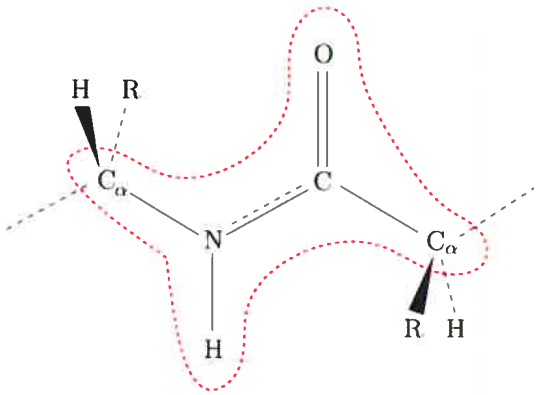

FIG. 2.2 – Lien et plan peptidique reliant deux acides aminés consécutifs.

2.1.2 Théories du repliement

Les travaux d'Anfinsen dans les années 50 et 60 sur la relation entre la séquence et la structure de la ribonucléase pancréatique bovine permirent d'établir que seule la séquence joue un rôle déterminant sur la structure finale et lui valurent le prix Nobel de chimie de 1972 [Anf73]. Or, Levinthal dans son fameux paradoxe nous dit que puisque l'espace conformationnel des protéines est énorme, il est impossible qu'une protéine puisse se replier par simple recherche aléatoire [Lev69]. Il existe donc un processus de repliement encodé dans la séquence des protéines et qui est suffisant pour les replier dans des temps biologiques.

La clé de ce processus réside en partie dans la structure native. En se basant sur les principes de la thermodynamique, on peut conclure que pour qu'une structure repliée existe avec une forte prédominance sur les autres structures, il faut que cette première ait une énergie libre plus basse que les autres. Ainsi, une fois repliée, la structure de plus basse énergie sera choisie la majorité du temps. C'est la structure native.

En se basant sur cette définition de la structure native, on peut se représenter la surface d'énergie libre comme un entonnoir au large cou représentant l'entropie conformationnelle de l'état dénaturé, et un resserrement accompagné par la dimi-

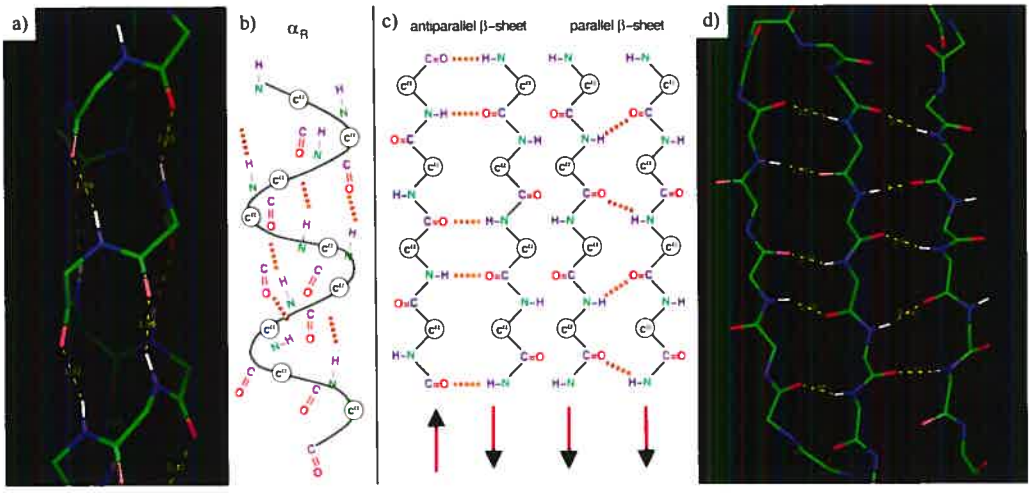

FIG. 2.3 – Représentations atomiques des éléments de structure secondaire : les hélices α (a) et (b), les feuillets β (c) et (d). Les feuillets gauches ont des brins antiparallèles et les feuillets droits ont des brins parallèles. Les images (a) et (d) sont générées à l'aide de l'outil Pymol [DeL02], les images (b) et (c) proviennent du livre de Schlick [Sch02]

nution de l'enthalpie configurationnelle (Figure 2.5). La surface de repliement n'est pas lisse. Les minima locaux y sont nombreux et on peut définir le processus de repliement par le parcours de ces minima locaux. En soi, la théorie de l'entonnoir ne peut expliquer les processus relativement rapides du repliement des protéines. Sur une séquence aléatoire, on ne dénote généralement pas d'états natifs. Ceci est expliqué par le phénomène de frustration : plus d'une structure peut être stabilisée par un ensemble d'éléments de structure secondaire et tertiaire différents et l'énergie de chacune de ces structures n'est pas assez basse pour en stabiliser une favorablement. Ainsi, la protéine passe son temps à changer d'une conformation stable à l'autre. Les protéines naturelles par contre sont un phénomène émergeant de millions d'années d'évolution. Puisque par des méthodes évolutives la nature réussit à générer des protéines dont la forme active correspond à la structure native, des chercheurs ont proposé que l'évolution de ces protéines aurait eu un effet de sélection favorisant celles qui possèdent une frustration minimale [BOSW95].

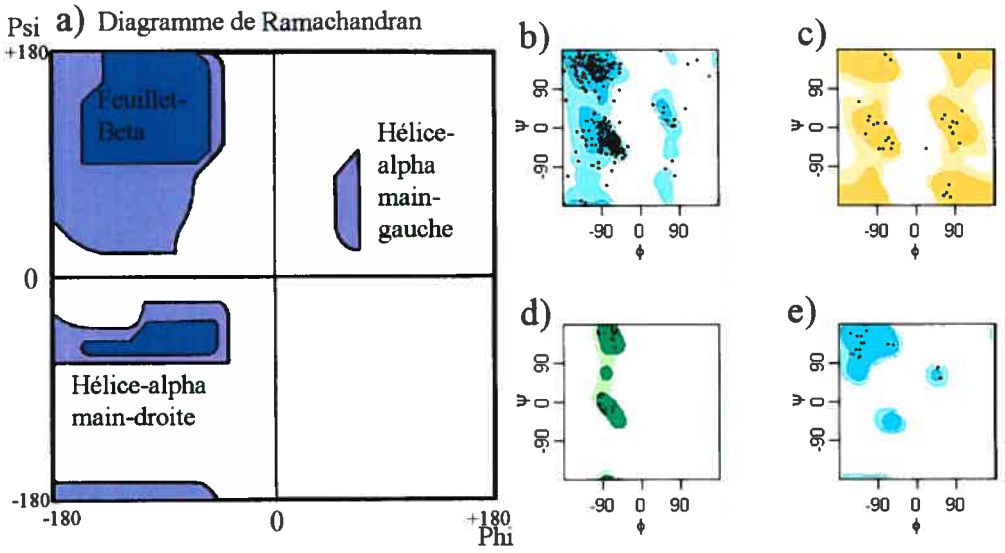

FIG. 2.4 – Diagramme de Ramachandran des valeurs des angles dièdres permises et des éléments structures secondaires associés (a). Aussi un exemple de diagramme général (b) et pour trois cas exception : Les glycines(c), les prolines (d) et les pré-prolines(e)

Ainsi, on suppose que les acides aminés formant des contacts stabilisant l'état natif ne peuvent former d'autres contacts lors du repliement. L'évolution se chargerait donc de générer des protéines stables, mais aussi qui se replient rapidement.

Bien que nous ayons à présent une bonne représentation de la surface d'énergie libre, nous ne savons toujours pas quels sont les processus consécutifs ayant lieu au cours du repliement. Voit-on un rassemblement rapide des résidus hydrophobes dans un coeur stable tout en exposant les résidus hydrophiles au solvant ? Est-ce que cette étape précède ou succède à la formation de structure secondaire ? Plusieurs modèles ont été proposés.

Sous le modèle du squelette en armature (Framework) [Pti87, KB90], on prédit que la structure secondaire, plus particulièrement les hélices-α, se forment rapidement et qu'il s'en suit une étape lente d'agencement des éléments de structure secondaire pour former la structure tertiaire.

Un autre modèle similaire au squelette en armature est celui de la diffusion-

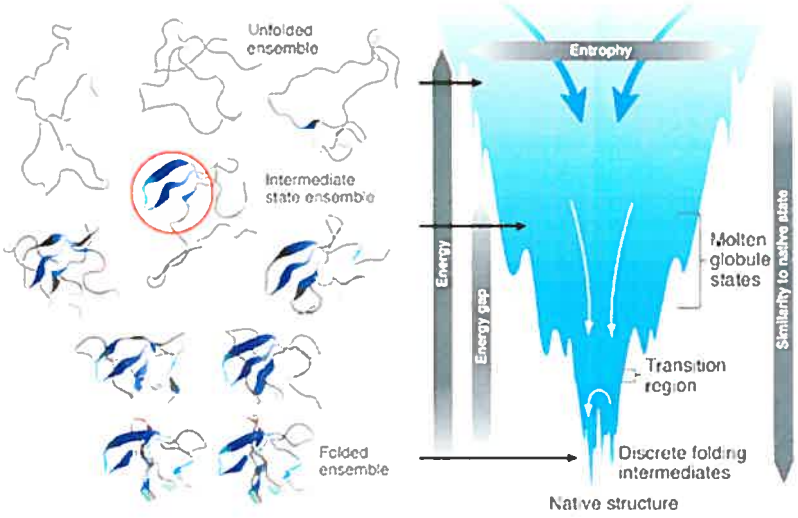

FIG. 2.5 – Exemple de surface d'énergie libre (droite) associée au repliement de src-SH3 (gauche). L'image provient de l'article de Brooks *et al.* [BOW01]

collision [KW94] qui suppose que plusieurs microdomaines peuvent se former et se dissiper aléatoirement jusqu'à ce qu'ils entrent en collision et se stabilisent mutuellement. Les microdomaines peuvent être définis comme étant des éléments de structure secondaire ou des coeurs hydrophobes isolés. Les structures secondaire et tertiaire se formeraient donc parallèlement. Certains auteurs considèrent ces deux modèles comme équivalents malgré leur différence [Fer97].

Dans le modèle de nucléation-condensation [Wet73, Wet90], on suppose que seulement certains éléments de structure secondaire forment un microdomaine qui sert de noyau sur lequel le reste de la protéine peut se stabiliser pour former concomitamment des éléments de structure secondaire et tertiaire.

Le dernier modèle est celui de l'effondrement hydrophobe [DBY+95]. Dans ce modèle, les acides aminés hydrophobes cherchent à se stabiliser mutuellement en formant rapidement un coeur hydrophobe. Ce coeur serait dominé par des interactions de longues portées natives et l'étape limitante serait la formation de la structure secondaire.

De ces quatre modèles, aucun n'est privilégié dans l'explication du repliement de toutes les protéines. Les modèles peuvent s'entrecroiser : il est possible qu'une protéine ait un repliement que l'on considérerait comme étant de type diffusion-collision dans une section où la séquence de résidus affiche une préférence notable à la formation de structure secondaire. Une autre partie de la même protéine riche en acides aminés hydrophobes pourrait quant à elle se replier sous le modèle de l'effondrement. On note cependant une nette distinction dans le profil énergétique des différents modèles de repliement. Selon Gianni*et al.* [GGK+03], le repliement suivant le modèle de diffusion-collision passe par un intermédiaire métastable d'énergie libre plus faible que l'ensemble dénaturé alors que le processus de nucléation-condensation voit une trajectoire comptant un seul état de transition issu de la formation parallèle de la structure secondaire et tertiaire (Figure 2.6).

2.1.3 Outils expérimentaux d'étude du repliement

2.1.3.1 Analyse des valeurs-Φ

L'une des méthodes favorites pour étudier le repliement des protéines est la méthode des valeurs-Φ qui permet de connaître les interactions entre les acides aminés à l'état de transition [MKSF89]. Celle-ci étudie l'effet de mutations clés sur la constante cinétique de repliement et sur la stabilité énergétique de la structure native. Les valeurs-ϕ sont donc données par :

$$\phi = \frac{RT \ln (k_{wt}/k_{mut})}{\Delta\Delta G_{D-N}} \tag{2.1}$$

où R est la constante des gaz parfaits et T est la température, k_{wt}/k_{mut} est le ratio entre les constantes de repliement de la protéine naturelle (Wild Type) et de la protéine mutée, et $\Delta\Delta G_{D-N}$ est la différence d'énergie libre entre la structure dépliée et repliée. Ces valeurs sont expérimentalement déterminées par la méthode de titrage d'un dénaturant avec examen par spectre de dichroïsme circulaire (CD) et par la méthode des sauts de températures à différentes concentrations de dénaturant en mesurant le spectre de fluorescence. Un marqueur de fluorescence

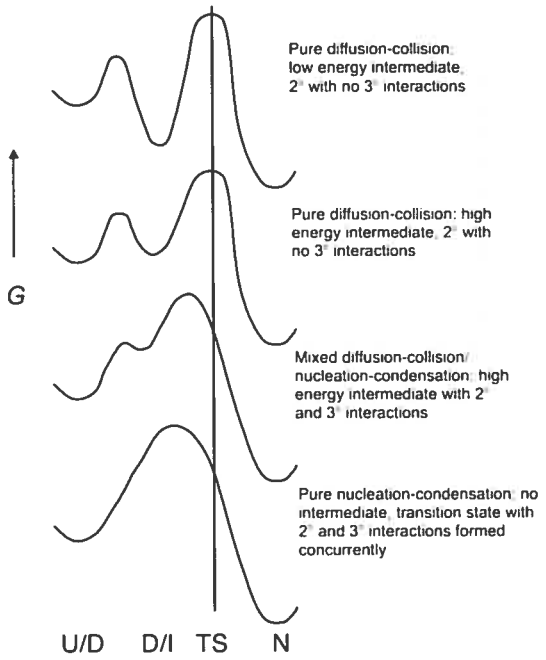

FIG. 2.6 – Diagramme simplifié d'énergie libre pour les modèles de repliement de nucléation-condensation pure à diffusion-collision pure. Le passage du premier vers le dernier est obtenu en augmentant la probabilité de formation de structure secondaire. Deux modèles de diffusion-collision sont donc possible, dépendamment si l'énergie de l'intermédiaire est plus haut ou plus bas que l'énergie de l'état dénaturé. Image extraite de l'article de Gianni*et al.* [GGK+03]

doit donc préférablement être présent, ou il faudra alors le rajouter.

Si un acide aminé muté n'est pas impliqué dans une interaction stabilisatrice à l'étape de transition, il s'en suit que $k_{wt} = k_{mut}$ et que $\phi = 0.0$. À l'inverse, un acide aminé important au site actif induira une valeur-ϕ se rapprochant de 1.0. Puisque l'état de transition est par définition le point dans la trajectoire de repliement où la protéine a une probabilité égale de se replier ou de se déplier, on suppose que toutes les interactions impliquées à l'état de transitions existent aussi dans la forme repliée. Ainsi, en choisissant de muter des acides aminés stabilisant la structure secondaire ou tertiaire de la structure native, on peut identifier quels éléments de structure sont présents à l'état de transition.

Pour étudier la structure secondaire, on choisit de muter des acides aminés

exposés au solvant. Le processus est double : on mute chacun de ces acides aminés pour une valine, puis ensuite pour une glycine. Puisque les mutations Val→Gly ont un faible impact sur les interactions à longue portée, en mesurant la valeur-ϕ entre ces deux mutants, on obtient de l'information sur les interactions de courte portée de la structure secondaire des hélices. Dans les cas de la structure tertiaire, on choisira de muter des résidus en contact dans le coeur hydrophobes par d'autres résidus hydrophobes de plus petite taille, de préférence des valines ou des glycines.

La méthode des valeurs-ϕ possède des forces et des faiblesses. Pendant de nombreuses années, elle s'est imposée comme étant la méthode privilégiée d'étude des états de transition. Cependant, elle est basée sur deux hypothèses qui ne sont pas encore prouvées. La première étant que les mutations n'altèrent pas la stabilité des états dénaturés. La justification apportée est que cet ensemble d'états affiche une variété de configurations qui ne favorisent aucune interaction de façon notable. La seconde hypothèse suppose que seules les interactions natives stabilisent l'état de transition. Or, on peut imaginer des cas où la valeur-ϕ associée à un acide aminé provient d'une interaction autre que celles retrouvées dans la forme native. Par exemple, l'état de transition pourrait être formé d'une majorité d'éléments de structure secondaire qui doivent pivoter et se réaligner pour atteindre la configuration native.

2.2 Aspects numériques

2.2.1 Modèles numériques

Pour étudier le repliement des protéines, il faut d'abord avoir une représentation spatiale de celles-ci ainsi qu'un modèle exprimant l'énergie associée à une configuration donnée.

Il n'y a pas de règle déterminant le degré de réalisme que doit maintenir une représentation. Certains modèles sont intentionnellement vagues (les modèles à gros grains, les représentations réduites). Il est en effet possible d'émettre des hypothèses sur le repliement d'une protéine en simulant chaque élément de sa structure secon-

daire comme étant une bille sur une ficelle pouvant interagir avec d'autres billes. Les conclusions apportées par ces études peu coûteuses en temps de simulation peuvent servir de base pour, par la suite, étudier des éléments spécifiques du repliement.

La majorité des modèles de représentations sont plus réalistes. Les modèles tout-atomes et tout-atomes lourds (excluant les hydrogènes) sont les plus coûteux à utiliser, mais présentent un niveau de réalisme qui n'est surpassé que par les simulations de mécanique quantique/mécanique moléculaire (QM/MM) impraticables sur des systèmes aussi gros que les protéines et leur solvant. Les modèles à représentation réduite offrent un compromis entre la rapidité et le réalisme en regroupant plusieurs atomes sous une seule bille.

Un niveau de simplification supplémentaire peut être atteint en utilisant une représentation implicite du solvant. Ce type de représentation exclut les atomes des molécules d'eau du solvant et les remplace par un champ de force qui peut être couplé à un calcul de la surface accessible de la protéine au solvant. Les méthodes implicites ont cependant le défaut de ne pas tenir compte des mouvements qui nécessitent l'expulsion de molécules d'eau. En effet, certains auteurs notent une relation inverse entre la viscosité du solvant et le taux de repliement, indiquant que le déplacement du solvant est une étape limitante [PB98, JGHS99]. Par contre, l'option de l'utilisation d'un solvant explicite impose ses propres limites sur les temps de simulation pouvant être atteints. Lors de simulations avec solvant explicite, on enveloppe la protéine dans une boîte de molécules d'eau à bordures périodiques suffisamment volumineuse pour pouvoir contenir la protéine à l'état dénaturé et pour qu'il n'y ait pas d'interaction entre un atome et sa copie dans la boîte voisine. Ce processus implique l'utilisation de plusieurs milliers de molécules de solvant pour une protéine de petite taille, augmentant de façon significative le temps de calcul.

Au modèle de la protéine est jumelé un potentiel énergétique, aussi appelé champ de force moyen puisque son gradient est un vecteur de dimension $3N$ exprimant les forces moyennes en x, y z pour les N atomes du système. À chaque structure ou conformation de la protéine est associée une énergie configurationnelle. Ce terme énergétique est déterminé *ab initio* ou par des valeurs expérimentales pour

les différentes interactions entre les atomes de la protéine et du solvant. Prenons l'exemple de l'équation du potentiel AMBER [PC03] :

$$V(r^N) = \sum_{liens} \frac{1}{2} k_b (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{1}{2} V_n [1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ 4\epsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \tag{2.2}$$

Dans l'ordre affiché, on trouve les sommations : 1) d'un terme harmonique sur la longueur des liens interatomiques, 2) d'un terme harmonique sur les angles de valence, 3) d'un terme en $1 + \cos \Delta\theta$ pour les angles dièdres, 4) d'une double sommation d'un terme 12-6 de Lennard Jones pour les interactions de Van-der-Waals entre tous les atomes distants et d'un terme électrostatique pour les atomes chargés. Ces termes sont similaires à ceux retrouvés dans d'autres potentiels tout-atomes comme GROMOS [Sco99] ou CHARMM [BK83], mais aussi dans les potentiels à représentation réduite tels que celui utilisé dans notre laboratoire, le potentiel OPEP [Der99]. Ils se distinguent souvent par le choix des paramètres.

Le potentiel OPEP est un potentiel générique à représentation réduite. Sa représentation atomique est composée de tous les atomes de la chaîne carbonée à l'exception des hydrogènes aliphatiques que l'on retrouve sur le carbone-α et sur les carbones des prolines. Les configurations du cycle de ces mêmes prolines sont approximées par un choix d'angles de valence imposant une conformation plane et un angle ϕ de -90 degrés. Toutes les autres chaînes latérales sont représentées par une bille de volume de Van-der-Waals approprié et de centre de masse correspondant à la position moyenne rencontrée dans des structures natives extraites de la base de données de protéine PDB [BKW+77]. Dans ce potentiel, l'interaction du solvant est simulée implicitement dans les constantes du potentiel. L'équation de celui-ci est similaire à celle d'AMBER ci-haut tout en contenant des termes

supplémentaires pour les interactions atome-bille et bille-bille. Étant donné que l'encombrement stérique provenant des chaînes latérales sous leur représentation réduite est diminué, un terme harmonique est aussi présent pour contraindre les angles dièdres libres ϕ et ψ des résidus dans des positions en accord avec les valeurs du diagramme de Ramachandran.

2.2.2 Méthodes numériques

2.2.2.1 Dynamique moléculaire

Une fois que le choix d'un potentiel énergétique et d'un modèle de représentation de la protéine est fait, le choix d'une méthode de simulation s'impose. Plusieurs méthodes sont offertes, mais la plupart des études utilisent une variante de la dynamique moléculaire (DM). Cette méthode repose sur l'intégration numérique des équations du mouvement de Newton :

$$a(t_i) = -\nabla E(X(t_i))/M(X) = F(X(t_i))/M(X) \tag{2.3}$$

$$V(t_i + \Delta t) = V(t_i) + a(t_i)\Delta t \tag{2.4}$$

$$X(t_i + \Delta t) = X(t_i) + V(t_i)\Delta t + a(t_i)\Delta t^2 \tag{2.5}$$

où $X(t_i)$ est le vecteur de position atomique cartésien de l'ensemble du système au temps t_i, $V(t_i)$ est le vecteur vitesse, $a(t_i)$ est le vecteur d'accélération, $-\nabla E(X(t_i))$ est le gradient de l'énergie, équivalent au vecteur de force $F(X(t_i))$, M(X) est la matrice diagonale des masses des atomes de X et Δt est le temps d'intégration.

Plusieurs algorithmes d'intégration numérique sont disponibles, mais la famille des intégrateurs de Verlet présente une conservation de la cinétique la rendant populaire. En partant d'une position atomique initiale $X(t_i)$, d'une position atomique la précédant dans le temps par Δt $(X(t_i - \Delta t))$, et d'une accélération qui nous vient du champ de force $a(t_i) = \frac{F(X(t_i))}{M(x)}$ on peut calculer la position de notre système au

temps $t = t_0 + \Delta t$:

$$X(t_i + \Delta t) = 2X(t_i) - X(t_i - \Delta t) + a\Delta t^2 \tag{2.6}$$

L'algorithme a ceci de particulier qu'il met à jour les vitesses des atomes du temps $t_i - \Delta t$ au temps t_i, donc toujours avec un cycle de retard sur la position actuelle :

$$V(t_i) = \frac{X(t_i + \Delta t) - X(t_i - \Delta t)}{2\Delta t} \tag{2.7}$$

Une variante populaire de l'algorithme de Verlet est le "Leapfrog", ou saut de grenouille, dans lequel la direction du pas menant à la position $X(t_i + \Delta t)$ est déterminée par la vitesse au temps $t_i + \frac{\Delta t}{2}$:

$$V(t_i + \frac{\Delta t}{2}) = V(t_i - \frac{\Delta t}{2}) + a(t_i)\Delta t \tag{2.8}$$

$$X(t_i + \Delta t) = X(t_i) + V(t_i + \frac{\Delta t}{2})\Delta t \tag{2.9}$$

$$V(t_i + \Delta t) = V(t_i + \frac{\Delta t}{2}) + a(t_i + \Delta t)\frac{\Delta t}{2} \tag{2.10}$$

La méthode de dynamique moléculaire est donc exacte si l'on choisit un temps d'intégration suffisamment petit pour caractériser tous les modes vibratoires existants dans les protéines. Le problème réside dans le fait que le repliement est un processus multi-échelle : d'un côté, les vibrations les plus rapides ont des fréquences de l'ordre de 10^{-14}. De l'autre, le processus de repliement de petites molécules s'effectue sur un temps de l'ordre de 10^{-5}. Si l'ont veut caractériser le processus de repliement dans son ensemble, on doit utiliser un temps d'intégration de $1fs = 10^{-15}s$ ($2fs$ si on impose des contraintes sur les hydrogènes). Ceci implique que l'on doive échantillonner en moyenne 10^{10} pas d'intégration. Le nombre peut sembler petit étant donné la puissance des ordinateurs modernes, mais c'est sans compter sur le fait que chaque pas d'intégration fait appel au calcul du potentiel énergétique

qui a un ordre d'exécution de $O(N^2)$ où N est le nombre d'atomes du système. On comprend alors l'intérêt porté aux nouvelles méthodes accélérées.

2.2.2.2 ART nouveau

Depuis déjà dix ans, la méthode d'activation-relaxation, développée par notre groupe en collaboration avec Barkema, est utilisée pour l'étude des verres et les semi-conducteurs amorphes [MB98, MDBM01] ainsi que pour l'agrégation de courts peptides et le repliement des protéines [WMD02, SMD04, MD05]. La technique est en fait une méthode générique de parcours des surfaces énergétiques. Son processus activé consiste en quatre étapes qui trouvent un chemin continu et physique reliant deux minima énergétiques et passant par un point de selle de premier ordre. Ces quatre étapes se regroupent en deux phases, l'une que l'on nomme Activation qui a pour but de trouver les points de selle à partir d'un minimum local, et l'autre, la Relaxation qui à partir du point de selle trouvé converge vers un minimum local différent du précédent. Examinons en détail les quatre étapes.

a) Sortie du bassin harmonique : .

Le système étudié, qu'il soit une protéine ou un réseau cristallin, débute le cycle itératif dans un minimum local d'énergie. Ce minimum prend la forme d'un bassin harmonique dans toutes les directions. C'est à dire, que la dérivée première de l'énergie (le gradient de force) est égale à zéro, indiquant un minimum, alors que la dérivée seconde de l'énergie est positive dans toutes les directions. Pour arriver à détecter un point de selle, la structure doit être déplacée de façon à augmenter son niveau d'énergie suffisamment pour que l'analyse de la surface énergétique puisse trouver des directions de courbure négative. Une direction aléatoire est donc choisie et un nombre d'atomes variant entre 20% et 100% de la séquence est déplacé. Du vecteur de déplacement, on retire la composante qui est parallèle au vecteur de forces reliées aux éléments rigides de la molécule (les termes énergétiques des longueurs de liens et des angles de valence et angles impropres). Ainsi, on espère que le déplacement suivra une direction dans laquelle seuls les angles dièdres libres seront modifiés. Un nombre de pas minimal doit être déterminé pour chaque système. On

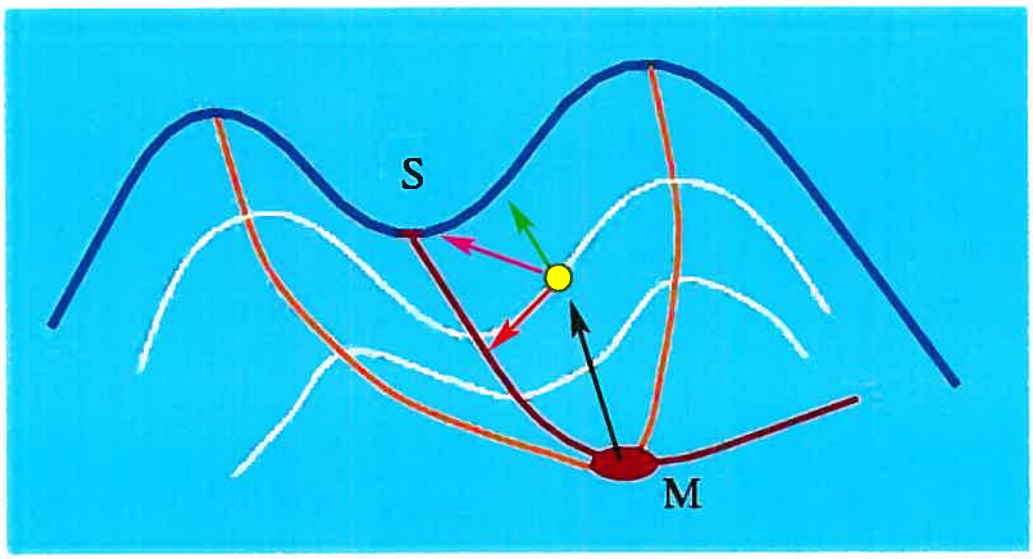

FIG. 2.7 – Exemple de trajet emprunté par ART nouveau pour atteindre un point de selle à partir d'un minimum. Suivant une direction aléatoire, la configuration est poussée hors du bassin harmonique (flèche noire) jusqu'au point où un vecteur propre de valeur propre négative est détecté (jaune). La configuration est alors poussée dans le sens de ce vecteur propre (flèche verte) tout en minimisant sont énergie dans l'hyperplan perpendiculaire (flèche rouge) resultant en un mouvement menant au point de selle (flèche rose).

ne lancera l'analyse de la courbure de la surface énergétique qu'après que ces pas auront été parcourus, la raison étant qu'une analyse trop hâtive peut détecter des points de selle qui ne connectent pas deux minima différents. Si cela se produit, le cycle doit alors être recommencé.

Après que la structure ait été poussée suffisamment hors du bassin harmonique, on lancera la méthode de Lanczos servant à détecter la courbure négative de la surface [Lan88]. Cette méthode itérative d'ordre O(N) peut extraire les vecteurs propres de plus faible valeur propre d'une matrice Hessienne sans avoir à la calculer en entier. Les vecteurs propres ainsi identifiés pointent dans les directions où la courbure de la surface énergétique est la plus faible. Notre algorithme continue à faire des pas l'éloignant du bassin harmonique jusqu'à ce que la méthode de Lanczos ait trouvé un vecteur propre dont la valeur propre est sous un seuil négatif choisi

en fonction du système étudié. La découverte de ce vecteur propre lance la seconde étape du cycle.

b) Convergence vers un point de selle :

À l'aide du vecteur propre de valeur propre négative, la structure est poussée dans la direction du point de selle. Cette étape faisant partie de la phase d'activation a contre-intuitivement l'effet de diminuer l'énergie de la protéine. En effet, l'étape précédente du cycle doit souvent déformer la molécule et augmenter son énergie plus haut que le point de selle avant de détecter la direction menant à ce dernier. La direction donnée par le vecteur propre n'est pas non plus le chemin de plus basse énergie. Par analogie, on peut décrire ce vecteur comme étant tangentiel à une intersection hypothétique entre deux montagnes en forme de cône. Or, pour passer par le point de selle, il faut se déplacer dans la direction du vecteur ainsi qu'être dans le sillon d'intersection. Pour y arriver, on procédera par une minimisation des forces qui sont perpendiculaires au vecteur propre après chaque pas pris dans la direction du vecteur. Cette minimisation de l'énergie dans l'hyperplan perpendiculaire a pour effet de rapprocher la structure du sillon menant au point de selle. On mesure aussi la norme du vecteur de force à tous les pas afin de déterminer la position du point de selle. Tout comme les minima des bassins harmoniques, les points de selle de premier ordre ont pour caractéristique d'avoir une somme des forces égale à zéro, la raison étant que le point de selle est un minimum énergétique dans toutes les directions sauf dans une direction où il est alors un maximum.

c) Saut du point de selle et convergence vers le nouveau minimum local :

Une fois le point de selle identifié, la prochaine étape consiste à pousser la configuration dans la direction l'éloignant du minimum précédent. Pour démontrer que la phase d'activation est un processus réversible, il suffit de pousser la structure dans l'autre direction. Lorsque la structure a été poussée (dans une direction ou dans l'autre), elle se retrouve spontanément exposée à un vecteur de force non-nulle. À l'aide de d'algorithme de minimisation comme une dynamique moléculaire amortie ou du gradient conjugué, on peut trouver une nouvelle configuration d'énergie

minimale.

d) Acceptation ou refus de la nouvelle structure :

Afin de déterminer si la configuration sera acceptée ou refusée, on utilise un critère de Métropolis définit par :

$$p_{accepter} = min(1, \exp \frac{\Delta E}{k_B T}) \tag{2.11}$$

où $p_{accepter}$ est la probabilité d'accepter la nouvelle structure, ΔE est la différence entre l'énergie de l'état final et celui de l'état initial, k_B est la constante de Boltzmann et T est la température simulée. On voit qu'une conformation dont l'énergie est plus basse que la précédente sera automatiquement acceptée.

Le seul critère sous le contrôle du chercheur à cette étape est le choix d'une température convenant à la simulation désirée. Cette température n'a pas d'impact sur la surface énergétique et ne sert qu'à accepter ou refuser des conformations échantillonnées. Une valeur élevée aura pour effet d'accepter des structures affichant de grands écarts positifs d'énergies, ce qui causera le dépliement de la molécule. Aussi, de petites valeurs auront pour effet de piéger la structure dans un minimum local entouré de minima de plus haute énergie. De plus, puisque différentes molécules démontrent des écarts d'énergie inter-minima différents, le choix de cette température doit être fait en fonction de la molécule étudiée.

Les défauts de la méthode ART nouveau sont aussi ses qualités : de par la navigation de la surface énergétique en passant par des événements de transition directe reliant des minima énergétiques, la méthode est extrêmement efficace. Cependant, en négligeant l'entropie du système, la méthode ne possède pas l'information sur l'énergie libre des configurations nécessaire pour déterminer la probabilité de chaque configuration à l'équilibre thermodynamique. Pour l'instant, il n'est pas encore prouvé que ART nouveau souscrive au bilan détaillé, c'est-à-dire que l'équation suivante n'est pas satisfaite :

$$\forall_{ij} P_{i \rightarrow j} \times Q_i = P_{j \rightarrow i} \times Q_j \tag{2.12}$$

où $P_{a \rightarrow b}$ est la probabilité de transiter de l'état a à l'état b et Q_b est la probabilité que la protéine se trouve à l'état b à l'équilibre thermodynamique. Bien que les événements de transition obtenues par ART nouveau soient physiques et possibles, il est impossible de déterminer leur probabilité réelle. En théorie, la méthode ART nouveau peut choisir un point de selle étroit et d'énergie élevée puisque son choix n'est basé que sur la courbure de la surface énergétique menant à ce point. D'autres points de selle de plus faible énergie et plus larges pourraient être préconisés par une méthode dirigée par l'entropie. Toutefois, le succès de la méthode, couplé au potentiel OPEP, à élucider les trois mécanismes de repliement de l'épingle β [WDM03, WMD04a] ainsi qu'à identifier la structure détectée par RMN de l'agrégation de poly-peptides KFFE [WMD04b, WMD04c] démontre que la méthode n'est pas limitée par cette contrainte. Ces publications démontrent que la méthode et son potentiel sont suffisamment génériques pour replier des protéines aux structures secondaires principalement en feuillet-β. Au chapitre 4, il vous sera présenté l'étude du repliement d'une protéine de structure secondaire tout-α à l'aide de la même combinaison ART nouveau - OPEP.

CHAPITRE 3

REVUE DE LA PROTÉINE A

Le choix d'un modèle d'étude du repliement des protéines est influencé par plusieurs facteurs :

- La taille de la protéine, son nombre d'acides aminés et d'atomes.
- L'aisance avec laquelle on peut appliquer les protocoles expérimentaux à son étude.
- L'indépendance du processus de repliement face aux protéines chaperonnes ou autres molécules de l'environnement de repliement.
- L'intérêt porté par la communauté scientifique et la possibilité de comparer ses résultats.

Le domaine B de la protéine A de *Staphylococcus aureus* (BdpA) répond à tous ces critères. L'intérêt scientifique porté à la protéine staphylococcale A (SpA) remonte aux années soixante alors que l'on découvre l'affinité de cette dernière à l'extrémité FC des molécules d'immunoglobuline [DKWQ69], plus particulièrement à l'immunoglobuline G (IgG) [LLM70]. En plus d'expliquer un des mécanismes de la bactérie qui la rend moins vulnérable à la phagocytose par les leucocytes par inhibition de l'opsonisation, cette découverte a ouvert la porte à l'utilisation de SpA dans les techniques de microscopie électronique à l'immunoferritine [BDWB77], les techniques de séparations cellulaires [GSS75], comme réactif dans les essais immunoenzymatiques sur support solide (ELISA) [God78], et même dans les procédures de diagnostic du virus de la rubéole [MRBW76]. Une recherche par mot-clé de "Staphylococcal protein A" à l'aide de l'outil PubMed de Medline nous donne plus de 3400 résultats pour la seule période 1977-2006, démontrant la quantité d'information disponible aux chercheurs désirant étudier cette protéine.

Du point de vue structural, la protéine membranaire A est composée de cinq domaines homologues, chacun pouvant lier IgG des mammifères [MAN+86], nommés E, D, A, B, et C et apparaissant dans cet ordre dans la séquence. Les domaines E

et B, pris individuellement, sont les deux fragments de la protéine qui sont le plus souvent utilisés dans les études de repliement.

Le domaine B, qui compte soixante acides aminés, possède une structure composée de trois hélices alpha interconnectées par deux tours et dont l'enroulement des hélices l'une par rapport à l'autre est main-gauche (rotation horaire suivant l'axe d'enroulement, Figure 3.1). Il fait donc partie de la classe des protéines tout-alpha. Ce fragment se replie *in vitro* sans l'aide de protéine chaperonne. Bien que la structure cristallographique du domaine B liée au fragment FC est disponible depuis 1981 [Dei81], il fallut attendre 1992 avant d'avoir accès à la structure seule en solution obtenue par résonance magnétique nucléaire (RMN) [GTS+92].

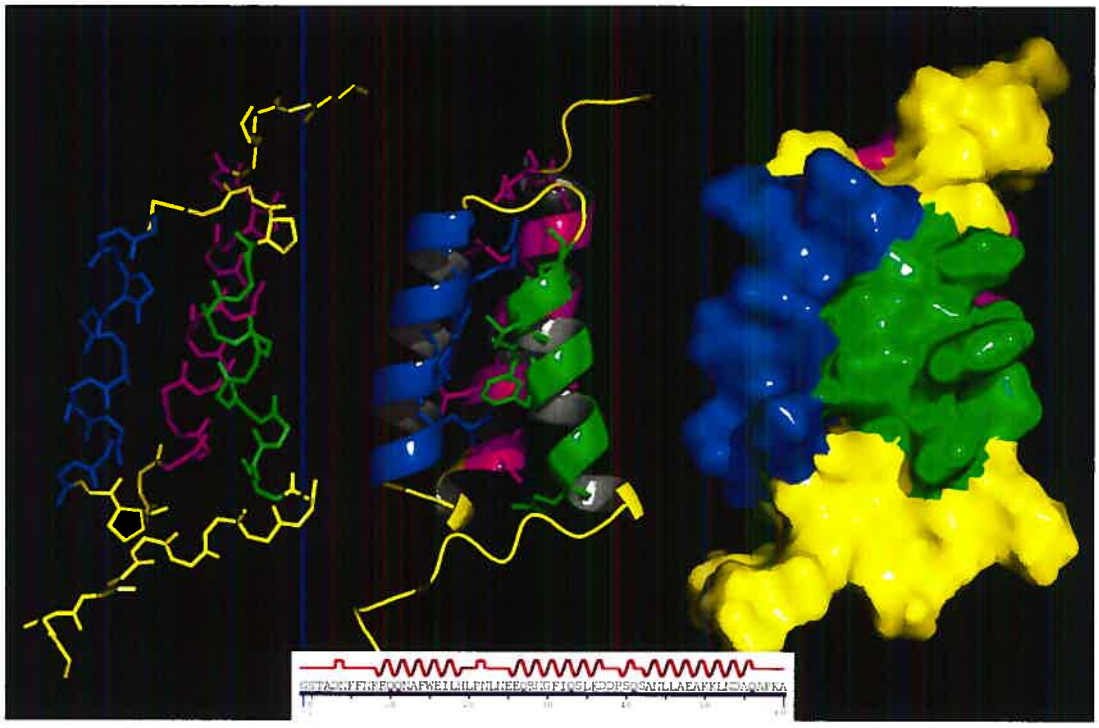

FIG. 3.1 – Structure tridimensionnelle du domaine B de la protéine A du mutant Tyr15Trp (PDB : 1SS1). À gauche, la chaîne carbonée en absence d'hydrogènes et de chaînes latérale. Au centre, représentation cartoon de l'enroulement des hélices et chaînes latérale des résidus du coeur hydrophobe. À droite, volume CPK tout-atomes.

3.1 Études expérimentales

On dénombre cinq publications importantes étudiant par des méthodes expérimentales le repliement du domaine B, dont une portant sur un double-mutant du domaine B nommé le domaine Z [DDM+04]. Par le biais de méthodes utilisant le marquage par échange des hydrogènes (H) pour des deutériums (D), Bai *et al.* évaluèrent tout d'abord quels étaient les acides aminés qui protégeaient le mieux leurs hydrogènes liés aux azotes du squelette carboné à l'équilibre thermodynamique [BKDW97]. Les protons impliqués dans un pont hydrogène (pont-H) étant moins susceptibles d'être échangés pour des deutérons, il est possible de déterminer quels acides aminés protégés dans la forme native ne sont plus protégés lors d'un passage à l'état dénaturé. Du fait que l'hélice H2 et H3 démontraient une plus grande protection de leurs protons, ils conclurent que ces deux hélices n'étaient pas dépliées dans aucun état intermédiaire en équilibre avec la forme native. De plus, leur examen du spectre de dichroïsme circulaire des fragments du domaine B composé d'une seule hélice ou d'une paire d'hélices démontre une prédisposition des fragments correspondant à l'hélice H3 seule et à la paire H2-H3 d'adopter une structure secondaire en hélice. Ils en déduisirent que s'il existait un intermédiaire dans le processus de repliement du domaine B, celui-ci devait être formé par les hélices H2 et H3. Cependant, leurs essais pour détecter un intermédiaire stable à l'aide de marquage par impulsion et échange H/D ne permirent d'identifier l'intermédiaire puisque la protéine atteignait sa forme native à l'intérieur des 6 ms du temps mort de l'appareil. Ils introduisirent alors la mutation de Ile16 en Trp16 à l'intérieur de la séquence de l'hélice H1 afin de déstabiliser l'état natif et favoriser le présumé intermédiaire, mais n'arrivèrent pas à identifier un état intermédiaire peuplé. Ils en conclurent que la protéine se replie suivant une trajectoire à deux états sans intermédiaires de basse énergie.

L'équipe de Myers poursuivit la recherche de la constante de repliement du domaine B en examinant par RMN le signal du résidu histidine HIS19 en fonction de la concentration de dénaturant [MO01]. Les résultats de l'expérience menée à

des concentrations d'hydrochloride de guanidine (GuHCl) variant entre 1.6 et 4.3 M prennent la forme d'une relation en chevron dont les deux droites se croisant correspondent aux taux de repliement k_f et de dépliement k_u. Par interpolation linéaire, ils obtinrent les valeurs $k_f = -120000s^{-1}$ et $k_u = -68s^{-1}$ en absence de dénaturant, ce qui permet d'établir la demi-vie du repliement à environ $6\mu s$ à l'aide de l'équation de réaction de premier ordre $t_{1/2} = \frac{\ln 2}{k_f}$ (puisque $k_f >> k_u$). La valeur de k_f est aussi supportée par leur expérience de Carr-Purcell-Meiboom-Gill [MG58] établissant le taux cinétique en examinant le temps de relaxation de la magnétisation transversale T_2 de la bande de HIS18.

Là où les expériences de dénaturation par GuHCl nous donnent une image thermodynamique du repliement, la dénaturation par saut de température possède une résolution temporelle suffisante pour avoir une image de la cinétique du dépliement. Dans leur expérience, Vu *et al.* induisirent à l'aide d'un laser des sauts de température à tous les 100 ms en examinant la bande d'absorbance infrarouge de l'amide I' à tous les 23 ns. Cette bande correspond aux étirements du lien du groupe carbonyle C=O sur le squelette carboné et elle varie en fonction de la présence ou absence d'éléments de structure secondaire. De plus, cette bande augmente en fréquence lorsque les résidus en hélice font partie d'un coeur hydrophobe, ce qui permet d'évaluer l'évolution de la structure tertiaire. Par l'examen de ces deux bandes, ils observèrent un dépliement suivant une relation bi-exponentielle caractéristique des repliements à trois états. L'état intermédiaire ainsi identifié serait composé de fragments d'hélice solvatés en absence de structure tertiaire. Sa formation rapide en 90 ns serait suivie par l'étape limitante d'assemblage des hélices pour former le coeur hydrophobe et les interactions à longue portée avec une durée de vie de $9\mu s$.

De tous les travaux expérimentaux, celui ayant suscité le plus d'intérêt est celui de l'analyse des valeurs-ϕ fait par Sato *et al.* [SRDF04]. La méthode développée par le groupe [MKSF89] étudie l'effet de mutations clés sur la constante cinétique de repliement et sur la stabilité énergétique de la structure native (voir 2.1.3.1 pour une description de la méthode). Puisque la séquence de BdpA ne possède pas de marqueur fluorescence naturel, ils introduisirent la mutation de Tyr15 pour

Trp15(Fig. 3.1). La structure native du mutant a été élucidée par RMN et est disponible dans la PDB sous le code 1SS1. Dans une étude subséquente, ils prouvèrent à l'aide de deux autres mutants que leurs résultats étaient reproductibles et non biaisé par le choix de la première mutation [SRF06]. En mesurant les valeurs-ϕ du passage Val→Gly de 16 acides aminés situés sur les hélices et exposés au solvant, ainsi qu'en procédant à 35 mutations sur 13 acides aminés hydrophobes, ils purent identifier les éléments de structure secondaire et tertiaire présents à l'état de transition. Le portrait dressé de la structure secondaire est une hélice H2 complètement formée alors que H1 est dénaturée et que H3 a quelques acides aminés en hélice α dans sa portion N-terminale. Malgré le fait que l'hélice H1 est peu structurée, ils détectèrent de nombreuses interactions de longue portée entre les résidus Phe14, Ile17 et Leu18 au centre et dans la portion C-terminale de H1 avec les résidus. Alors que le tour T1 n'est pas bien défini, la structure du tour T2 et les contacts de longue portée l'avoisinant sont établis. À partir de cette image de l'étape de transition, ils conclurent que le mécanisme de repliement suivait le modèle de nucléation-condensation à deux états.

Comme on peut le constater, les études expérimentales présentent deux portraits différents du repliement de BdpA. L'étude des fragments d'une ou deux hélices de Bai *et al.* démontre que l'hélice H3 est la plus stable et que la paire d'hélices H2-H3 possède le plus de chance d'être repliée, H1 étant la moins stable des trois. De leur côté, Sato *et al.* donnent une représentation de l'état intermédiaire où H2 est déjà formée avec seulement la partie de N-terminale de H3 est structurée. Leur méthode permettant aussi d'extraire de l'information sur la structure tertiaire, ils aperçoivent des interactions de longue portée entre H1 et H2 et un tour T2 bien formé.

3.2 Études théoriques

Du point de vue théorique, d'énormes efforts ont été portés à élucider le processus de repliement du domaine B de la protéine A. Les méthodes utilisées étant

presque aussi variées que les résultats obtenus, il est important de déterminer une classification de ces derniers. Dans son article de revue, Wolynes proposa de regrouper les résultats de simulations selon deux familles [Wol04]. Tout d'abord, il regroupe ensemble les résultats identifiant à l'état de transition une structure dominée par une hélice H2 bien formée et des interactions spécifiques de longue portée avec H1 tel qu'observé par Sato *et al.* La deuxième famille serait composée des simulations retrouvant une séquence de repliement où les hélices H3, H2 et H1 se formeraient dans cet ordre, tel qu'observé par Bai *et al.*

La première famille ne dénombre que quatre études théoriques. Tout d'abord, on note les simulations de dynamique moléculaire de Boczko *et al.* [BB95]. Leur étude porta sur un fragment plus court du domaine B de 46 acides aminé, le fragment 10–55, dans lequel les acides aminés des queues N-terminale et C-terminale ont été enlevées[1]. Leur approche démontre une certaine ingéniosité étant donné les moyens informatiques limités de l'époque. Puisqu'une simulation tout-atomes de DM n'aurait pu simuler le repliement de BdpA de la forme dénaturée à la forme native, 78 simulations furent exécutées à partir d'états différents choisis en fonction de leur ressemblance à l'état natif. Les auteurs définissent le rayon de giration (R_{gir}) comme étant la coordonnée de réaction du repliement. Leurs 78 structures initiales sont réparties dans l'intervalle [R_{gir} = 9.3 Å, R_{gir} = 14.0 Å] correspondant à l'état natif et l'état dénaturé respectivement. Les simulations sont alors exécutées jusqu'à ce qu'il y ait une intersection entre chacune d'entre elles permettant de tracer un chemin de repliement reliant les deux extrêmes. En se basant sur cette simulation concaténée, ils tracèrent le portrait du repliement en fonction de R_{gir}. Le processus identifié commence à des valeurs de R_{gir} entre 12 Å et 12.5 Å par l'assemblage des hélices H1 et H2 partiellement repliées pour former les interactions de longue portée similaires à celles identifiées par les valeurs-ϕ. Entre 11 Å et 12 Å, les hélices H1 et H2 continuent à se former en stabilisant leur interaction. L'hélice H3 n'apparaît

[1]Dans un soucis d'économie de temps de calcul et de simplifications du processus de repliement, pratiquement toutes les simulations portées sur le domaine B utilisent le fragment 10–55. Tous les cas mentionnés dans ce chapitre utilisant ce fragment, nous ne mentionneront que les exceptions.

qu'à des valeurs de R_{gir} plus petites que 11 Å.

Le groupe de Brooks reproduisit l'expérience en 1997, cette fois-ci avec un solvant explicite, en utilisant une méthode différente de jonction des données des simulations, la méthode d'analyse par histogramme pondéré (WHAM) [Boc94]. Encore une fois, ils notèrent à l'état de transition une structure contenant plusieurs ponts hydrogène natifs dans les régions de H1 et H2 caractéristiques d'hélices partiellement formées. Ils notèrent aussi la présence fréquente du contact Phe13–Leu34 entre H1 et H2, mais aussi Leu34–Leu44 et Leu34–Leu44 entre H2 et H3. L'hélice H3 par contre n'est pas bien formée. Finalement, ils constatèrent que le bassin de l'état natif était large, permettant l'existence de structures ayant de 8 à 25 contacts natifs et de 10 à 23 ponts-H natifs.

Deux autres études trouvèrent des résultats similaires en utilisant les méthodes de DM. La première par García et Onuchic utilisa la méthode de d'échantillonnage par échange de répliques (DMER) qui consiste à exécuter en parallèle plusieurs simulations de DM à différentes températures et à échanger les configurations en se basant sur un critère de probabilité [GO03]. Cependant, leur méthodologie est douteuse puisque dans les 82 états initiaux utilisés, 44 ont un RMSD avec l'état natif de 4.0 Å ou moins, ce qui correspond à la distance entre l'état natif et l'état de transition. Il serait donc surprenant que les simulations ainsi exécutées échantillonnent adéquatement l'espace de repliement.

La dernière étude de cette famille utilise elle aussi une méthode parallélisée de DM. Cheng *et al.* utilisèrent une méthode guidée dans laquelle on exécute plusieurs simulations à partir d'un seul état initial jusqu'à ce critère de synchronisation soit satisfait [CYWL05]. À ce moment, toutes les simulations sont relancées à partir d'un des états sélectionnés dans les simulations précédentes. Le critère de sélection utilisé favorise les contacts natifs et défavorise les contacts non-natifs. Puisque le critère de sélection est utilisé peu fréquemment comparativement au temps de simulation, ils émettent l'hypothèse que le processus de repliement trouvé n'est que faiblement biaisé. Bien que les résultats du groupe soient en accord avec ceux obtenus par l'analyse des valeurs-ϕ, il est important de noter le parallèle entre leur

critère de sélection et à l'hypothèse employée dans l'analyse des valeurs-ϕ selon laquelle seuls les contacts natifs peuvent exister à l'état de transition. Si le but des chercheurs avait été de trouver un état de transition composé d'interactions natives comme celui proposé par le groupe de Sato, ils n'auraient pu choisir meilleur critère de sélection.

La seconde famille de simulations compte plus d'une dizaine résultats en accord avec les probabilités individuelles des hélices et des paires d'hélices à se replier définies par Bai *et al.* [BKDW97] et nous ne citerons que quelques unes. Les méthodes sont variées et certaines méritent plus que les autres de s'y attarder. Tout comme dans les simulations présentées ci-haut, les études portent en majorité sur le fragment 10–55. Dans les expériences de dynamique moléculaire, on note tout d'abord le travail de Jang *et al.* qui ont utilisé un modèle tout-atomes avec solvant implicite [JKSP03]. Leur choix d'utiliser une température de repliement de 400 K sous-entends que leur potentiel énergétique surestime le point de fusion de la protéine qui a été évalué empiriquement à 340 K. Leur expérience devrait normalement échantillonner des états dénaturés. Or, ils observent des trajectoires de repliement rapide pour BdpA (14 ns) ainsi que pour la protéine HP-36. Leurs simulations de BdpA débutent par un effondrement rapide de la structure étendue suivi par la formation des trois hélices en parallèle. Ils dénotent cependant que les hélices H3 et H2 se forment et s'assemblent légèrement avant l'hélice H1, conformément aux probabilités mentionnées ci-haut.

La plus vaste étude théorique portée sur la protéine A est l'oeuvre du groupe de Pande et de son réseau distribué *Folding@Home*. Dans une publication récente, Jayachandran *et al.* exécutèrent 4900 simulations parallèles de dynamique moléculaire tout-atomes en utilisant le potentiel énergétique GROMACS96 et un solvant explicite à une température de 300 K. Tout comme García et Onuchic, leur choix d'états initiaux laisse à désirer : deux structures initiales sont respectivement à 3.3 Å et 3.8 Å de distance RMSD de la structure native, les 47 autres structures ayant une distance RMSD moyenne de 5.8 ± 1.4 Å de l'état natif. Malgré cette proximité alarmante de l'état natif, à peine 0.1% de leurs simulations atteignent une distance

RMSD de 2.0 Å avec l'état natif. Cependant, leur grand nombre de simulation leur permit d'agglomérer leurs résultats pour en extraire l'information concernant les probabilités de formations et de déformations individuelles des hélices. Leur première constatation est que la formation des hélices est faiblement corrélée entre elles. Ils observent que chaque hélice se forme et se déforme fréquemment dans l'ensemble d'états dénaturés et ils y voient un signe d'un processus de repliement selon le modèle de diffusion-collision. Aussi, en traçant un diagramme d'état discret correspondant à la présence ou l'absence des différentes combinaisons d'hélices, ils observèrent que l'hélice H2 avait la plus grande probabilité de se former en premier (60%). Cependant, lorsqu'un état contenant la paire d'hélices H2 et H3 est présent, dans la majorité des cas (70%) l'hélice H3 était la première formée. Ceci semble indiquer que l'hélice H2 n'est que partiellement stable en l'absence de H3. Aussi, l'état natif est dans 75% des cas créés par l'ajout de H1 à la paire d'hélices H2-H3, conformément à l'expérience de Bai *et al.* Ils utilisèrent par la suite une méthode statistique de corrélation des interactions de longue portée qui précèdent la formation d'éléments de structure secondaire. Il est intéressant de noter que dans le cas de la protéine A, près de 50% des contacts inter-résidus pouvant prédire la formation de chaque hélice ne sont pas des contacts existants dans la forme native. Cette information pourrait être interprété comme étant une preuve de frustration dans la protéine puisque si des contacts non-natifs sont utilisés comme échafaud pour l'assemblage d'éléments natifs, ces mêmes contacts doivent par la suite être brisés pour permettre un réassemblage natif des contacts.

Plusieurs autres études ont été faites en utilisant des représentations réduites. Notamment, le travail de Berriz *et al.* dans lequel les acides aminés sont définis en entier par une seule bille connectée à ses voisins par une tige rigide [BS01], les études du groupe de Scherage utilisant le potentiel UNRES dans lequel les acides aminés sont représentés par trois billes(carbone-α, plan peptidique et chaîne latérale) [LKS05, KLS06], ainsi que le travail de Favrin *et al.* [FIW02] utilisant un potentiel dont la représentation spatiale est identique, à un atome près sur les prolines, à celle d'OPEP [Der99]. Berriz *et al.* par dynamique de Langevin et à l'aide

d'un potentiel biaisé trouvèrent que les hélices H2 et H3 se formaient individuellement avant de s'assembler à l'état de transition. Khalili *et al.* en utilisant lui aussi la dynamique de Langevin avec son potentiel non-biaisé arrive à la même conclusion [KLS06]. Favrin *et al.* et leurs simulations Monte-Carlo arrivent à reproduire la stabilité relative des fragments d'hélice individuels telle que vue par Bai *et al.*

Ces méthodes présentent des faiblesses qui les rendent moins intéressantes. Tous d'abord, il n'est pas encore établi à quel point les méthodes basées sur un potentiel biaisé donnent des résultats conformes avec le repliement réel des protéines. Dans le cas de l'utilisation d'un potentiel à représentation réduite biaisé tel que celui de Berriz *et al.*ou encore celui de Zhou *et al.* [ZL02, LZ02], le niveau d'approximation est double et la validité des résultats est mise en doute. Quant à lui, le potentiel UNRES fût utilisé pour replier plusieurs petites protéines aux structures secondaires variées, mais leur résultats préliminaires semblent indiquer une préférence du potentiel pour les structures en hélices-α, identifiant difficilement les structures natives contenant des éléments de structure en feuillet-β [LKS05]. Dans leur résultats plus récents, la structure native de la protéine A qu'ils trouvent par simulation n'est pas entièrement repliée [KLS06]. Le potentiel réduit de Favrin *et al.* ne définit que cinq type de chaînes latérales et leurs simulations identifient elles-aussi un minimum global énergétique différent de la forme native [FIW02].

D'autres travaux numériques traitant du domaine B de la protéine A ne seront pas examinés dans ce travail. Il s'agit de travaux utilisant des méthodes non-conventionnels [SOB99, GES02, IKW02, IS06] ou peu documentés [KSS02]. Leurs références sont toutefois incluses par souci d'exhaustivité.

CHAPTER 4

ARTICLE: THE COMPLEX FOLDING PATHWAYS OF PROTEIN A SUGGEST A MULTIPLE-FUNNELLED ENERGY LANDSCAPE

Jean-François St-Pierre, Normand Mousseau
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec), Canada
Philippe Derreumaux
Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique et Université Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France

Folding proteins into their native states requires the formation of both secondary and tertiary structures. Many questions remain, however, as to whether these form into a precise order, and various pictures have been proposed that place the emphasis on the first or the second level of structure in describing folding. One of the favorite test models for studying this question is the B domain of protein A, which has been characterized by numerous experiments and simulations. Using the activation-relaxation technique (ART nouveau) coupled with a generic energy model (OPEP), we generate more than 50 folding trajectories for this 60-residue protein. While the folding pathways to the native state are fully consistent with the funnel-like description of the free energy landscape, we find a wide range of mechanisms in which secondary and tertiary structures form in various orders. Our non-biased simulations also reveal the presence of a significant number of non-native β and α conformations both on and off-pathway, including the visit, for a non-negligible fraction of trajectories, of fully-ordered structures resembling the native state of non-homologous proteins.

Keywords: protein folding — Monte-Carlo simulations — folding pathways — staphylococcal protein A

4.1 INTRODUCTION

Folding proteins is a complex process that requires the formation of both secondary and tertiary structures. The order by which these levels of structure assemble during folding is still unclear, however. For small proteins, it has been suggested that folding takes place cooperatively, following a funnel-like energy surface leading to the native fold, the global lowest energy conformation [LMO92]. Under the minimal frustration principle, the protein folds to its global minimum efficiently without getting trapped in local minimal energy conformations rising from discordant energy signals [BOSW95]. Recent experimental and theoretical works indicate that this model could offer an oversimplified picture of folding and that folding pathways could be much more diverse (see, for example, Refs. [Sko05, LKS05, SFM06, RFBM06]).

Using the Activation-Relaxation Technique (ART nouveau) [MM00] in conjunction with a generic Optimized Potential for Efficient peptide structure Prediction (OPEP) [Der99], we investigate the folding pathways of the full-length B domain of the staphylococcal protein A, a fast-folding 60-residue sequence that has been used extensively in the quest to elucidate protein folding. This left-handed bundle [GTS$^+$92] of three helices (H1-H3) has been the subject of multiple experimental and theoretical studies exploring various ingredients of folding and offering sometimes conflicting views on the process. For example, while quench-flow and NMR experiments show a two-state process with a population of low-energy intermediates below 0.6 % on the folding pathway [BKDW97, MO01], laser-induced jump analysis detect short-lived intermediates characterized by nascent helices within 90 ns and preceding the μs-scale formation of the tertiary structure [VMOD04]. However, if initial unfolding study shows that the first helix to unfold was H1, followed by H2 and H3 [BPS$^+$94], temperature jump [DDM$^+$04] and ϕ-value [SRDF04, SRF06] analysis conclude that H2 is nearly fully formed at the transition state, with H1 and H3 being only partially ordered.

On the theoretical side, while the full-length protein has received little attention,

with only on-lattice [KS94], unfolding [AD00] and biased folding [GES02] simulations, considerable work has focused on the core fragment 10-55 using a range of methods [LKS05, BB95, GBB97, LLS99, SOB99, ZK99, BS01, FIW02, KSS02, LZZ02, ZL02, GO03, JKSP03, VRS03, CYWL05, KLS06, JVGP06, Der00]. These include Monte-Carlo [KS94, VRS03, Der00], molecular dynamics using all-atom representations in explicit solvent [AD00, BB95, GO03, JVGP06] or implicit solvent [LZZ02, ZL02, JKSP03, CYWL05], reduced protein representation MD [ZK99, FIW02] and Langevin dynamics [LKS05, SOB99, BS01, KLS06], as well as some novel methods like conformational space annealing [LLS99], stochastic difference equation algorithm [GES02], application of the diffusion-collision model [MO01, IKW02], or the statistical mechanical model [IS06]. Following Wolynes [Wol04], we can identify two families of simulation results: a first group of studies [BB95, GBB97, GO03, CYWL05] identifying helix H2 and the formation of specific long-range interaction between the first and second helix (H1-H2) at the transition state, and a second group of studies finding that H3 is the first to form followed by H2 and H1. While the former studies are consistent with the picture provided by the ϕ-value analysis, the majority of the studies [AD00, GES02, ZK99, BS01, FIW02, KSS02, LZZ02, JKSP03, JVGP06, IKW02, KLR$^+$05] are more consistent with the observed helical stabilities of the individual helices which identifies H3 as the most stable helix [BKDW97].

Here, we report on 52 folding ART nouveau-OPEP trajectories for this protein. Coupled with Monte-Carlo [Der97, Der98], the OPEP force field has been used to predict the structures of 20 peptides [Der99, FD01] and the 56-residue domain B1 of protein G [Der02] with models within 3.0Å Cα root-mean square deviations (RMSD) from experiments.

Coupled with ART-nouveau, we predict folding mechanisms for a 16-amino acid β-hairpin consistent with other methods [WMD04a] and aggregation mechanisms for multi-chain amyloid systems [SMD04, MDG05, MMD06], consistent with IR spectroscopy [PD05].

4.2 Methods and details of simulation

4.3 ART-OPEP simulations

ART nouveau [MM00, WMD04a] builds folding trajectories by exploring the energy landscape in a four-step iterative process. Starting from a local minimal-energy conformation, the structure is deformed and brought to an adjacent first-order saddle point by following the direction of lowest negative curvature of the energy landscape. The conformation is then relaxed into a new local minimum. This new conformation is accepted based on the Metropolis criterion, $p_{\text{accepted}} = \min(1, \exp -\Delta E/k_B T)$, where T is the Metropolis temperature and k_B the Boltzmann's constant. Since thermal fluctuations are not included in these local minima, this temperature does not reflect a physical temperature. It is therefore adjusted to ensure a proper sampling of the conformation space. Here, we use a Metropolis temperature of 900 K for folding, (referred to as T_h) resulting in a 50% acceptation rate. In some cases, we pursue the refinement of the lowest-energy structures found during folding at a temperature of 600 K (referred to as T_l). While ART nouveau does not simulate a well-defined thermodynamical ensemble, the method generates continuous pathways going from minimum to minimum through adjacent saddle points and all trajectories represent physically-possible folding pathways [WMD04a]. Comparisons with other simulations carried out at 300 K for a 16 amino acid β-hairpin [IC06] and amyloid peptides [FIM04, MN06] show indeed that the ART-generated trajectories capture the overall folding and aggregation pathways, even though, due to the lack of detailed balance, the relative rate of each of these pathways cannot be established.

The OPEP energy potential [Der99] is an off-lattice coarse-grained implicit solvent force field for protein related simulations in which all the side-chain atoms of a non-proline amino acid are represented by a single bead with appropriate van der Waals radius and hydrophobic/hydrophilic character. All backbone atoms are represented with the exception of the aliphatic hydrogens. The underlying energy function includes terms for four types of interactions: (1) an harmonic term

for each bond angle, bond length, improper angle of the peptide bond and side-chain to maintain proper geometry. (2) a 12-6 potential for pairs of hydrophobic or oppositely charged side-chains coupled to a repulsive 6-potential for other side-chain interactions, (3) a backbone two-body and four-body (cooperative) hydrogen-bonding term, (4) an excluded-volume term between main chain atoms and between main chain atoms and side-chain beads. Effects of the solvent are included in the above parameters.

4.3.1 Details of the simulations

The sequence used is the full Y15W mutant (accession pdb: 1SS1 [SRDF04]) that was engineered for Φ-value analysis. This 60 amino-acid protein of 964 atoms is represented by 368 beads using the OPEP coarse-grained model.

A total of 52 simulations were launched from three initial conformations. Twelve simulations were initiated from the fully extended state (structure EX0) and ran at T_h for about three months on a modern IBM PowerPC processor. Each simulation generated about 50000 events, with around 23000 accepted events each. All of these simulations show a collapse of the extended structures, within the first 100 accepted events (Supplemental Fig. II.1 (a)). Two additional groups of 20 simulations each were started from two random coiled conformation, RH0 and LH0, with opposite initial "handedness" of the coil formation to reduce the impact of the initial sate (Supplemental Fig. II.1 (b) and (c)). These 40 simulations are run at the same Metropolis temperature T_h for about 30 000 events, leading to about 13 000 accepted events for each run. The lowest-energy structure of each of these 40 runs was further refined for an additional 8000 events at T_l. In what follows, all event numbers refer to accepted events only.

All secondary structure analysis was done using the dssp program [KS83].

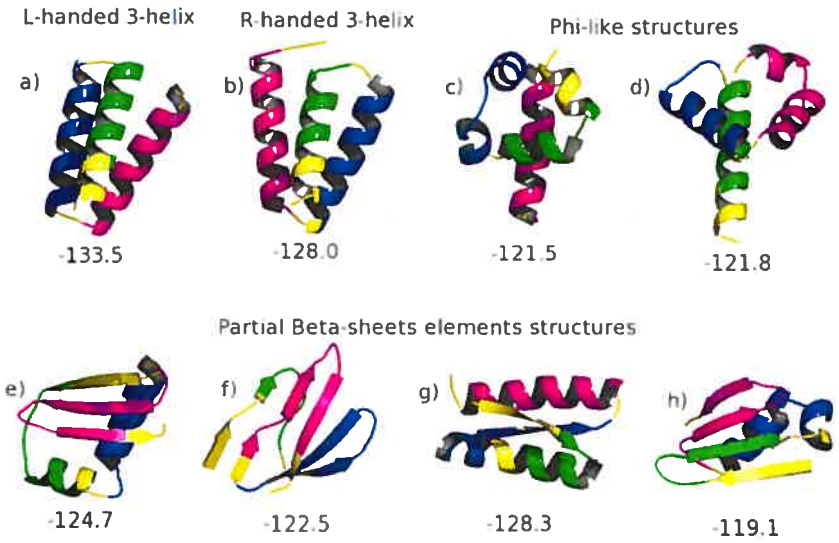

Figure 4.1: The structure of lowest energy found through independent simulations at 900K followed by a energy refining simulations at 600K. (a) The left-handed native-like bundle found in 4 simulations; (b) the right-handed mirror image found 7 times; (c) ϕ-like structures with complete H3 helix — 3 simulations; and (d) ϕ-like structures with complete H1 helix — 4 simulations. (e) to (h) show representative structures of families of conformations with a significant β-sheet components: (e) found in 8 simulations; (f) in 3 simulations; (g) in 5 simulations; and (h) in 2 simulations. Not represented here are the 16 simulations who did not find a configuration of energy lower then -116.0 kcal/mol.

4.4 Results

4.4.1 Structures of lowest energy

We start from the NMR structure 1SS1, characterized by a disordered N-terminal region (residues 1-9), three helices H1, H2 and H3 at residues GLN10-HIS19, GLU25-ASP37 and SER42-GLN56, respectively and two turns spanning LEU20-ASN24 and ASP38-GLN41, respectively. Energy relaxation locates a local minimum (Supplemental Fig. II.2), with an energy of -116.3 kcal/mol at a Cα RMS distance of 1.2 Å.

In contrast, the lowest-energy structure (LE1) predicted by our folding simu-

lations has an energy -133.5 kcal/mol. LE1 displays the NMR topology, but the generated left-handed triple helix bundle is characterized by longer and shifted helices [H1 : PHE6–ILE17 vs. GLN10–HIS19 by NMR, H2: PRO21–LYS36 vs. GLU25-ASP37 by NMR and H3: PRO39–GLU56 vs. SER42-GLU56 by NMR]. The turns are shorten to LEU18–LEU20 for T1 and ASP37–ASP38 for T2 (Fig. 4.1 (a), Supplemental Fig. II.3). Overall, LE1 deviates from the NMR structure by 6.8, 5.2 and 2.9 Åusing the residues 1-60, 10-55 and 25-60, respectively. This deviation, which involves the first 24 residues, is therefore mostly due to the shift in T1. This excess of predicted secondary structures does not result from a minor role of tertiary interactions. The formation of helices contributes 35% of the total energy of the 60-residue protein, and previous Monte Carlo simulations of the 10-55 fragment using the same OPEP parameters generated low-energy structures deviating by 2.9 Åfrom the NMR structure [Der00]. It is possible that the increase in helical content results from the presentation of the conformations in minima and the underestimation of the entropic effects. It is not known if a similar relaxation occurs with other off-lattice potentials as, to our knowledge, all simulations used the 10-55 truncated variant.

In what follows, LE1 is referred to as the native structure and used instead of the NMR structure for the evaluation of the native secondary structure and native contacts in the graphs. The color coding used in the figures reflects this definition: green, blue and magenta are associated with the helices H1, H2 and H3, respectively, and yellow is used for the other elements of the structure.

Figure 4.1 shows representative conformations of the lowest energy structures found in 36 of the 52 simulations that reach an energy below -116 kcal/mol. While there are some considerable fluctuations, especially among the less structured conformations, we can regroup all these structures into four classes. The first two classes (panels a and b) correspond to the native three-helix bundle and its right-handed image. The left-handed native-like bundle is found in 4 simulations (run 18, 25, 32, 35) out of 52. These native-like structures range in energy from -120.8 to -133.5 kcal/mol. Its right-handed mirror image occurs also in 4 simulations within

an energy range of -122.5 to -128.0 kcal/mol (3 other less-ordered simulations reach a state with the same topology as the right-handed mirror image and an energy below -116 kcal/mol). Although the secondary structures are almost identical between the left- and right-handed structures, there is little overlap between their contact maps (Supplemental Fig. II.4). Non-native right-handed conformations were also reported by other studies [LLS99, FIW02, VRS03].

The third class of conformations displays an all α content (panels c and d) and are called ϕ-shaped structures for their resemblance with the Greek character. They have either a fully formed H1 or H3 helix at the center, with the remaining broken helices forming a loop over it. The fully formed H3 ϕ-shaped topology (panel c) occurs in 3 simulations, with an energy ranging from -119.8 to -121.5 kcal/mol while the full H1 version is found 4 simulations and has a similar energy range (-118.3 to -121.8 kcal/mol). Both conformations occur with varying helix breaking points and coiling handedness. Interestingly, this structure corresponds to the lowest-energy structures found by Vila *et al.* using ECEPP/3 and various solvation models [VRS03].

Finally, 50% of the trajectories locate conformations with all β or mixed α/β content. The lowest-energy conformations reached by eight simulations (panel e) have a three stranded β-sheet formed in the C-terminal region interacting with partially formed helices H1 and H2 and an energy ranging from -116.7 to -126.3 kcal/mol. Three simulations (panel f) locate an all β topology (-117.8 to -122.5 kcal/mol); five simulations (panel g) locate a β-sheet between the N-terminal and H2 region while H1 and H3 helices are partially formed (-118.5 to -128.3 kcal/mol). The last two runs (panel h) show conformations with a partial H2 helix above a four-stranded β-sheet and energies of -118.4 and -119.1 kcal/mol.

4.4.2 Pathways leading to the native structure

We first look at the 4 trajectories, R18, R32, R35 and R25, that reach a left-handed three-helix bundle matching the experimental topology. Figure 4.2 presents the folding R18 trajectory, initiated from structure LH0. After a rapid collapse,

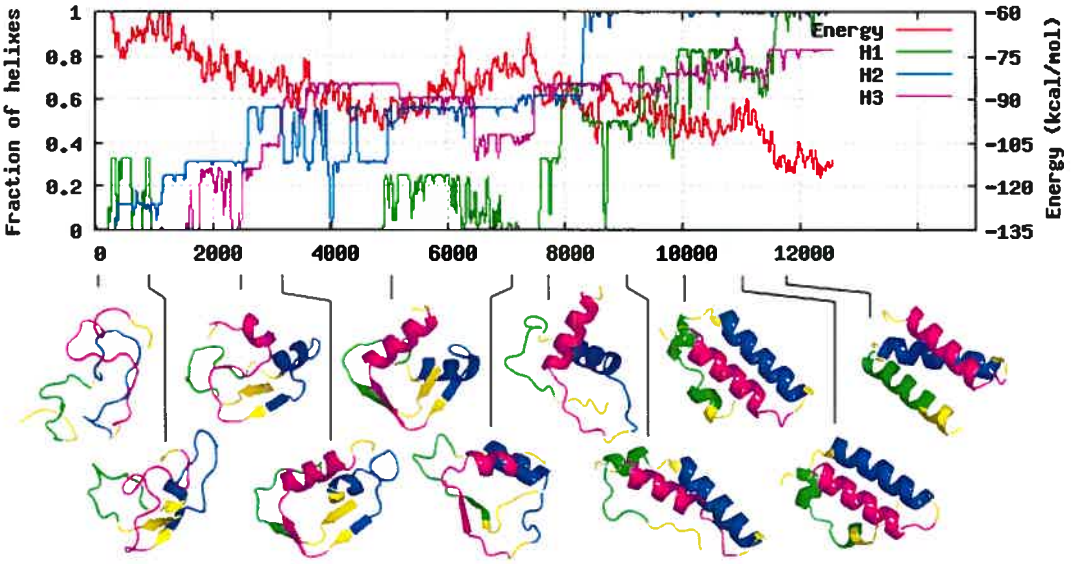

Figure 4.2: Evolution of trajectory # 18 . Top graphs: formation of the helical regions H1(green), H2(blue), and H3(magenta) into helices and evolution of the energy level (red). All lines are smoothed by a Gaussian of width $\sigma = 0.1$

a two-stranded β-sheet spanning SER0–ASP3 and LEU35–ASP38 of H2 forms at accepted event 1000 (see the snapshots of Fig. 4.2). The H3 region, free to move, rapidly enrolls into a partial helical structure between events 2500 and 4000. However, the full formation of H3 is prevented by another pair of parallel β-strands spanning residues LYS5–ASN7 of the N-terminal region and SER40–ALA43 of the T2 and H3 regions that are broken and reformed (events 3300 to 7000). Simultaneously, H3 changes its orientation from parallel to antiparallel with respect to the H2 region (events 5000 to 8500), going over an energy barrier. This rotation is accompanied by the formation of the H2 helix followed by a partial organization of H1 in a downhill energy process. The structure of lowest energy (-121.4 kcal/mol) is found at event 11800. A refinement run at T_l allows the system to relax down to -133.5 kcal/mol. Until event 8000, the structure is dominated by non-native interactions, with no-more than 25 % of native contacts (not shown); this number

grows rapidly after the H3 reorientation, which allows the formation of H2 and H1.

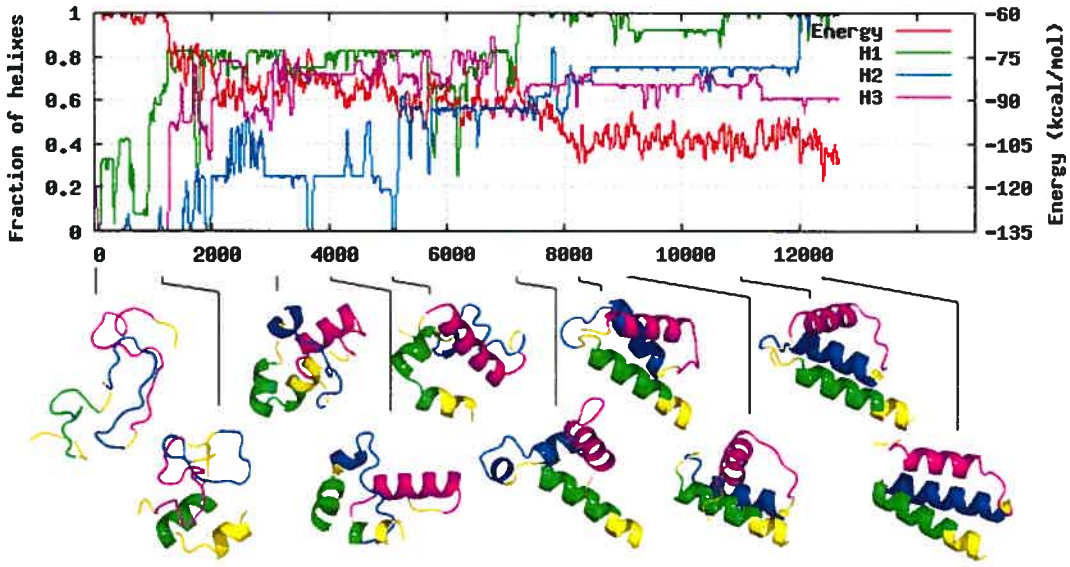

Figure 4.3: Evolution of trajectory #32 . Top graphs: formation of the helical regions H1(green), H2(blue), and H3(magenta) into helices and evolution of the energy level (red). All lines are smoothed by a Gaussian of width $\sigma = 0.1$

Run R32 also starts from the structure LH0 and shares a number of similarities with R18 (Fig. 4.3). In the first 4000 events, H1 and the loop regions around it are stabilized into interacting helical fragments, leaving the H3 region free to grow at full length. As H3 rotates with respect to the H2 region, H1 grows to form evolving structures with 60 to 80% of residues in α-helical conformation. At this point, H1 and H3 do not share any contacts and are well exposed to the solvent as are T1 and the N-terminal regions that also adopt α-helical structures. At around event 5000, ALA49 and LEU52 of H3 come into contact with ILE17 and LEU18 of H1 while the H2 regions is freed from the core and exposed to the solvent. Contacts rapidly move to ALA13 and ILE17 of H1 and ALA49 and LEU52 of H3 while the H2 helix is formed. Reptation movements occur between the H1 and N-terminal regions and the H2 region while the H3 rotates around its anchor point on H1 (events 7000 to

12000). The structure of lowest energy found at event 12425 is at -120.9 kcal/mol (-125.5 kcal/mol after refinement at T_l).

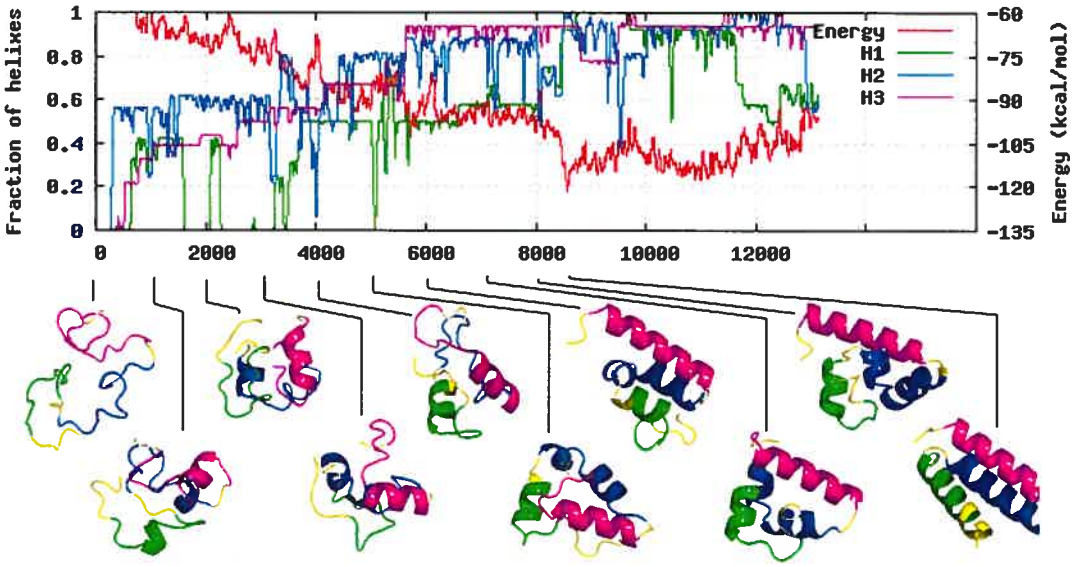

Figure 4.4: Evolution of trajectory #35 . Top graphs: formation of the helical regions H1(green), H2(blue), and H3(magenta) into helices and evolution of the energy level (red). All lines are smoothed by a Gaussian of width $\sigma = 0.1$

R35 (Fig. 4.4) is initiated from the right-handed random coil RH0 structure. No specific mechanism to reduce the initial preference is identifiable although by the 1000th event, the general coiling of the structure has shifted to a more left-handed prone conformation. H2 and H3 have up to 50% of their residues in α state within the 4000 first events. From the 6000th through the 8000th event, H3 is almost fully formed while H1 and H2 oscillate between 60 and 80% of their helical content. The final folding step is initiated around event 8300 when the two helical fragments of H1 coalesce to form a single helix. Subsequently, H2 forms a full helix, and then helix H3 completes its formation. The process only takes 250 steps during which H3 rotates on its axis thus breaking its contacts with the two residues of H1. The

lowest-energy structure is found at event 8568 at -127.1 kcal/mol (-128.8 kcal/mol after refinement at T_l).

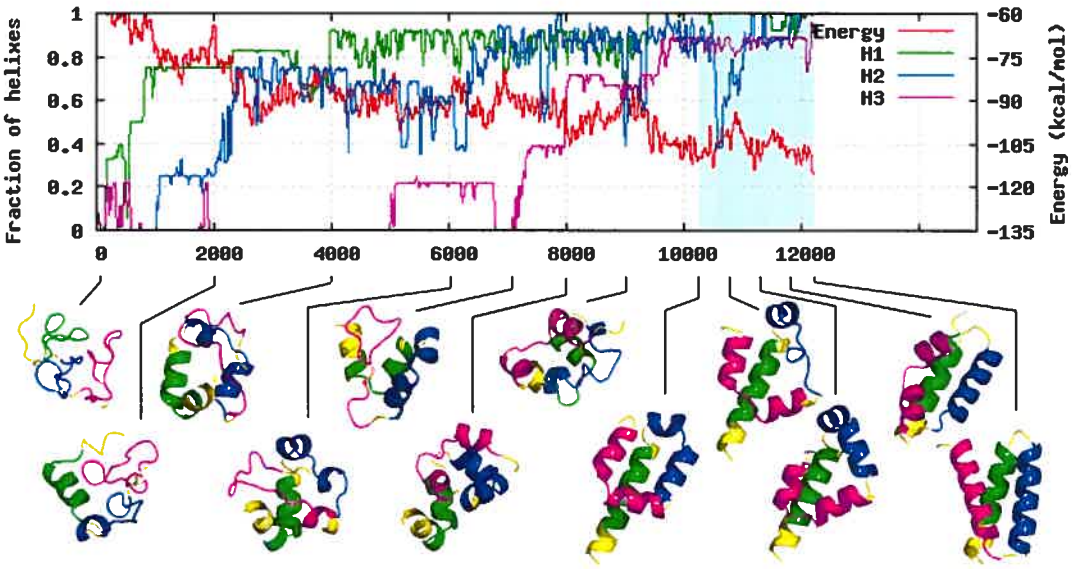

Figure 4.5: Evolution of trajectory #25 . Top graphs: formation of the helical regions H1(green), H2(blue), and H3(magenta) into helices and evolution of the energy level (red). The blue box of (d) indicates that a lower temperature criterion was used for that part of the simulation. All lines are smoothed by a Gaussian of width $\sigma = 0.1$

Fig. 4.5 shows the folding pathway for R25, which starts from LH0. At the end of the higher T run, namely after 10141 accepted events, the structure is caught in a ϕ-shaped conformation of energy -114.7 kcal/mol. From this point, however, the simulation at T_l manages to finish the relaxation and brings the protein into its native topology, with an energy of -120.8 kcal/mol. R25 is the first trajectory to fold to the native state by a pathway in which the H1 helix is formed first and maintained at over 60% through all the simulation. Also unusual is the fact that H2 partially forms before H3. In the following simulation at T_l, the ϕ-shaped structure proceeds into the native state by first forming H2, followed by H3. This overcoming

of the ϕ-shaped conformation is interesting since seven other simulations end with a similar structure, suggesting that this intermediate structure might be important.

4.4.3 Aggregated results

To identify the underlying process regulating these diverse folding pathways, we examine the evolution of the formation of helices based on a discrete definition [JVGP06]. A segment is considered to be α-helical if 80% of its residues remain helical for at least four consecutive ART events, which is sufficient to remove most of the fluctuations. The presence or absence of the three helices thus define a populated state (this definition does allow some of the most structured ϕ-shaped conformations to be counted in the three folded helix state.)

Examining the agglomerated transitions for the 22 simulations that find low energy conformations with high α-helical content (Table 4.1 (Top)), we see that helix H3 is the most likely to be formed first from an unordered state (38%) while H1 and H2 have lower probability (29% and 23%). These three one-helix states have a high probability of returning to an unordered state (H1:55%, H2:58% and H3:42%), and they show a preference for the formation of two-helix states with H2 or H3 has second folded helices (HUU→HHU: 21%, HUU→HUH: 15%), (UHU→HHU: 9%, UHU→UHH: 30%), (UUH→UHH: 35%, UUH→HUH: 22%).

However, a notable fraction of the one-helix HUU structures transits toward the three-helix state without intermediates (7%), indicating a case of parallel formation for the H2 and H3 helices. Looking at the fractions of states that transit to the three-helix state, we see that a majority has the H3 helix formed in a two-helix state (UHH:40%, HUH:29%) (Supplemental Table I.1).

If we restrict the analysis to the four simulations that converge onto the native state, we find a slightly different picture (Table 4.1 (Bottom)). The one-helix state UUH that has the highest probability of forming first in the agglomerated data from 22 trajectories (38%) has the lowest probability to form first in the 4 trajectories that find the native conformation (7%) (Table4.1(Bottom)). However, we see that an even greater majority of the fraction of states that transit to the three-helix

Table 4.1: Transition probability between secondary structures averaged over the 22 simulation with a majority of α content (top) and the 4 simulations finding the native structure (bottom). Line and colons represent the starting and ending states respectively. States are defined by a triple character string indicating if each of the 3 helices H1, H2 and H3 are formed (H) or unformed (U) respectively. The POP line and column indicate the population of the associated state. Lines are normalized by the sum of outgoing transitions from this state.

CONF:	TO:	UUU	UUH	UHU	UHH	HUU	HUH	HHU	HHH
FROM:	POP:	116483	48095	14451	19805	19571	16980	4376	18409
UUU	116483		**0.38**	**0.23**	**0.09**	**0.29**			
UUH	48095	**0.42**			**0.35**		**0.22**		
UHU	14451	**0.58**			**0.30**			**0.09**	0.02
UHH	19805	**0.12**	**0.40**	**0.16**					**0.32**
HUU	19571	**0.55**					**0.15**	**0.21**	**0.07**
HUH	16980		**0.41**		0.01	**0.17**			**0.40**
HHU	4376			**0.11**	0.03	**0.39**			**0.47**
HHH	18409		0.01	0.01	**0.38**	**0.07**	**0.29**	**0.22**	
FROM:	POP:	19532	1612	3156	2872	5655	5791	475	3842
UUU	19532		**0.07**	**0.40**	0.03	**0.50**			
UUH	1612	**0.26**			**0.10**		**0.64**		
UHU	3156	**0.76**			**0.20**	0.01		0.01	0.01
UHH	2872	0.03	**0.07**	**0.22**				0.01	**0.66**
HUU	5655	**0.67**		0.01			**0.18**	**0.10**	0.04
HUH	5791		**0.39**			**0.28**			**0.33**
HHU	475			**0.08**	**0.12**	**0.35**			**0.46**
HHH	3842	0.01			**0.54**	0.04	**0.23**	**0.18**	

state contain the H3 helix (UHH: 55%, HUH: 24%, HHU: 15%) suggesting that H3 is statistically important for folding (Supplemental Table I.2).

Another measure of tertiary structure formation is the correlation that can arise in the formation of hydrophobic long-range contacts. An analysis of set of native contacts as a function of the total number of residues in α-helical conformation reveals that only the hydrophobic pair LEU18–LEU23 is present in all simulations, with a presence probability exceeding 70% for structures containing 28 to 40 residues in α-helical conformation. Since this pair involves two helical residues at the extremities of the T1 turn, these residues are sequentially nearer then any other native pair. This result suggests that there there is a wide distribution in the

order in which tertiary structure forms during folding but that the stabilization of turns could be important for accelerating this process [BS01].

4.5 Discussion and Conclusions

4.5.1 Richness of folding pathways

The 52 folding simulations of the B-domain of protein A show a rich diversity of pathways and structures. Looking at the 22 simulations that reach an all α-helix conformation, the probability of a given helix to be formed in the state preceding the triple helix state is 47%, 62%, and 69% for helix H1, H2, and H3 respectively. In the case of the 4 simulations folding into the native state, these percentages are 39%, 70%, and 79%. Statistically, therefore, these results are in agreement with the majority of reported simulations that also observe that H3 is the most stable helix on the pathway near the native structure. However, the individual folding pathways observed here are much more diverse and do not support the existence of a folding trajectory with a unique sequence of events, consistent with the folding results on protein A obtained using a structure-derived potential [KSS02]; which pathway dominates should depend on the details of the experimental folding condition.

Even though folding pathways are multiple, the evolution of the average C_α RMSD as a function of the number of α-helical residues is clearly linear with the number of events (Supplemental Fig. II.5). This indicates that the helix formation occurs as the tertiary structures fall into place. Folding is therefore highly cooperative while respecting the funnel picture [OW04]: the protein collapses first and then rearranges itself step by step, going through a number of intermediate states dominated by non-native interactions.

4.5.2 Structures of intermediates on and off folding pathways

The richness of the folding process is also seen in the on-pathway intermediates sampled by the simulations, with transient formation of non-native α and β secondary structures such as the β-sheets seen in the first 5000 events of R18

(Fig. 4.2). While these elements are not obligatory for folding — of the 22 simulations finding an all-α lowest energy structure 11 pathways contained some β-sheets elements in the first 5000 events — they provide strong evidence of residual frustration in the folding landscape. This frustration has already been observed by others [Sko05, LKS05, GO03].

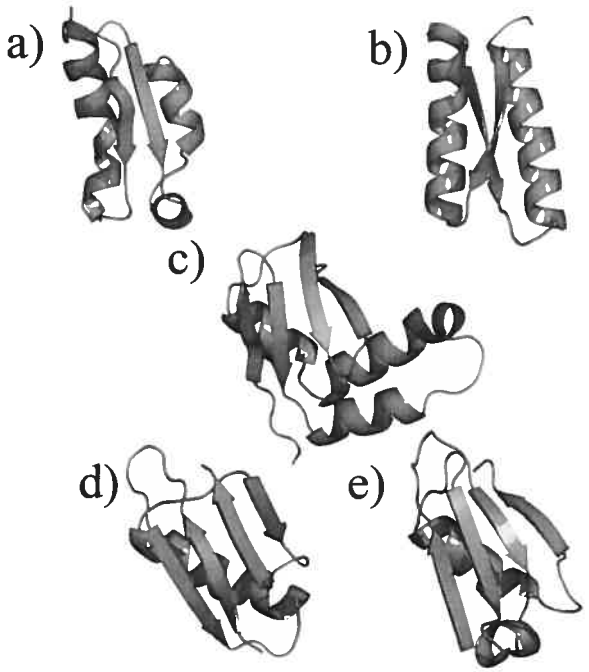

Figure 4.6: Top, protein G (a), protein 1UDX (c) and structure of lowest energy (-119.1 kcal/mol) from simulation #46 (b); Bottom, two Rossman folds, one from Staphylococcal peptidyl-cysteine decarboxylase EpiD (d), the other found through simulation (-128.3 kcal/mol) (e). Both pairs of homologous structures are aligned without gaps in sequence.

During folding, simulations also reach a number of ordered metastable structures in addition to the native and mirror three-helix bundles. Interestingly, these structures are not as random as expected, but can be detected in the PDB for non-homologous sequences using structural alignment tools [YG04]. For instance, structure 1(g), of lowest-energy among the α/β predicted topologies (-128.3 kcal/mol),

displays a high similarity (P-value = 0.0486) with the fragment 1-52 (a partial Rossman motif) of the 174 residues Staphylococcal peptidyl-cysteine decarboxylase EpiD (PDB no: 1G5Q) (Fig. 4.6 (a) and (b)). Similarly, our predicted topology (h) of Fig. 4.1 shares high structural similarity (P-value = 0.009) with the 1UDX structure shown Fig. 4.6(c) as well as with the 56-residue α/β Streptococcal protein G (PDB no: 1GB1, P-value = 0.0262), although we see a key difference in the packing of the two β-hairpins with 1GB1 (Fig. 4.6(e)). This specific result is interesting in view of recent directed-evolution analysis. These experiments showed that it is possible to pass from the three-helix bundle to protein G topology with an homology of 59 % between the mutated sequences [AROB05]. The authors showed that the energy gap between the two alternative folds is > 6 kcal/mol while we find 14 kcal/mol in our simulations.

This experimental result suggests that the topologies found with ART-OPEP can be believed both in their structure and overall energy ranking. Recent comparisons on OPEP using ART nouveau and parallel replica show that the energy difference can be used as a good guide for the respective weight at equilibrium [WMD07].

If we apply this criterion here, and supposing that the entropy for both motifs are identical, the right-handed conformation (family b) and the partial Rossman motif of family (g), standing at 5.2 kcal/mol above the native state should therefore occur about 10000 times less often then the latter in equilibrium at 300 K; the probability of occurrence for the other patterns shown in Fig. 4.1 is even lower.

Even though these occupation probabilities are only indicative, they clearly suggest that the transient topologies observed here should be very short-lived and cannot be observed experimentally, unless specific stabilizers or antibodies for recognition are used. Nevertheless, their appearance during unbiased folding simulations tells us something about the structure of the energy landscape. It is generally assumed under the minimal frustration principle that these off-pathway minimal energy conformations would not be visited. As a result, the energy landscape is often described by a single funnel [LMO92, BOSW95].

Our simulations show, however, that a number of low-energy structures with

α/β or all-β topologies exist and should be sampled in the test tube, albeit with a very small probability. These results are in agreement with the topomer-sampling model of Debe *et al.* [DCG99] which was reformulated somewhat in the backbone-based model of Rose *et al.* [RFBM06]. In these models, the family of structures that can be visited is totally determined by the backbone, reducing considerably the complexity of folding. A protein can therefore explore multiple non-native topologies and enter the native funnel, determined by the details of the side-chain interactions, at various levels of organisation, in a way which is well captured by the network and graph picture derived from the analysis of a rich heterogeneous denatured state ensemble for of the β-hairpin [Caf06].

In summary,while it is possible to identify a relatively well-defined pathway leading to folding by averaging over the trajectories, analysis of the individual runs underlines the presence of a rich diversity of intermediate structures such as the ϕ-shaped conformation [VRS03, KLS06].

Our simulations also indicate two levels of frustration in the energy landscape. The first level is residual in character and is illustrated by the presence of intermediates with transient non-native secondary structures en route to the native state [Sko05, LKS05, GO03]. The second level of frustration is of much larger amplitude which suggests a multiple-funnelled energy landscape with the dominant funnel belonging to the native state. This description supports the picture of a large but limited set of backbone-determined structures that are sampled by the protein before reaching the intermediate state that leads it to proper folding [RFBM06, DCG99].

JFSP and NM acknowledge partial support from Natural Sciences and Engineering Research Council of Canada, the Canada Research Chair Fund and the Fonds québécois de recherche sur la nature et les technologies. NM is also grateful to CRNS for a *poste rouge*. PD acknowledges funding from CNRS and Université of Paris 7. All calculations were performed with the support of the Réseau québécois de calcul de haute performance.

CHAPITRE 5

APPORT SCIENTIFIQUE DE L'ARTICLE

La première conclusion que l'on puisse tirer de l'article précédent est que la méthode ART nouveau, couplée au potentiel énergétique OPEP, est effectivement une méthode sans biais vers la structure native pouvant être utilisée dans l'étude du repliement des protéines. Tout d'abord, mentionnons le relativement haut taux de succès de la méthode alors que 4 simulations sur 52 réussissent à trouver une conformation hautement similaire à la forme native malgré l'utilisation de la séquence complète de 60 acides aminés, et que 3 autres simulations atteignent une conformation identifiée comme étant un intermédiaire près de la forme native (les conformations en forme de ϕ). Le fait que la forme native trouvée par la méthode soit aussi la conformation de plus basse énergie valide le choix des paramètres du potentiel générique OPEP. Ces mêmes paramètres ont aussi été utilisés dans les autres simulations de notre groupe identifiant la forme native de l'épingle β [WDM03, WMD04a] ainsi que des agrégats de peptides KFFE dont la topologie de plus basse énergie est en accord avec les données expérimentales [WMD04b, WMD04c]. Nous en concluons que le biais provenant de l'utilisation du potentiel OPEP sur nos simulations est négligeable.

Les méthodes exactes affichent des taux de réussite notamment plus faibles et cela malgré le fait qu'ils utilisent le fragment de 10-55 de BdpA plus facile à replier. Par exemple, Jayachandran *et al.* dans leurs travaux de DM avec un potentiel tout-atomes et un solvant explicite n'ont eu un taux de succès que de 0.1% alors que leurs états initiaux étaient à proximité de l'état replié [JVGP06]. De l'autre côté, certains groupes utilisant des représentations réduites ont noté des taux de succès très élevés. Citons l'exemple récent de Khalili *et al.* qui, à l'aide du potentiel réduit UNRES, ont pu générer 380 trajectoires de repliement complètes sur 400 exécutions (taux de succès de 95%) et cela dans un très court temps-processeur (de l'ordre de 30 heures-processeur par simulation) [KLS06]. Par contre, le bassin natif qu'ils

observent a une forme en ϕ et une distance RMSD sur les $C\alpha$ de 3.75 Å de la forme native, indiquant que leur représentation ou leur calibration du potentiel ne sont pas adéquates. Dans ce sens, nos simulations présentent un compromis intéressant entre les méthodes précises et lourdes en temps-processeur et les méthodes rapides basées sur des potentiels qui font possiblement trop d'approximations.

En absence de l'aspect thermodynamique, notre méthode a pourtant réussi à générer des résultats en accord avec plusieurs publications expérimentales et théoriques précédentes. Par exemple, les probabilités individuelles des trois hélices à se former en l'absence des deux autres sont en accord avec les probabilités expérimentalement identifiées qui suggèrent que H3 est l'hélice la plus stable seule et que la paire H2–H3 est stabilisée [BKDW97]. Aussi, l'examen des données agglomérées des quatre simulations trouvant la forme native sont en accord avec les résultats Jayachandran *et al.* indiquant que l'hélice H2 fluctue beaucoup (probabilité de 40% de se former seule 76% de se déformer). Une différence importante de notre étude à l'inspection de l'hélice H1 dans les 4 cas de trajectoires menant à l'état natif démontre que cette hélice fluctue plus que H2. Les populations des états contenant H1 seule ou en combinaison avec l'hélice H3 surpassent les autres populations d'une hélice ou d'une paire d'hélices. Toutefois, ces états comprenant H1 affichent aussi de hautes probabilités de se déformer. Une explication possible de ce phénomène peut être donnée si l'on considère la définition des hélices utilisées : nous avons en effet redéfini dans notre étude les fragments des hélices pour que la nomenclature reflète notre état natif plus structuré que l'état expérimental. Cette définition a pour effet de rallonger l'hélice H1 et de la décaler dans la zone N-terminale : d'une définition Arg10–His19, on passe à la définition Phe6–Ile17. Les effets de cette translation dans la zone N-terminale sur l'information recueillie à propos de la fluctuation de l'hélice H1 sont indéterminés. Bien que ART sous-estime l'entropie du système, il reste que les régions intrinsèquement flexibles et exposées au solvant ont une plus grande probabilité d'être déformées dans des événements ART, ce qui peut expliquer les fluctuations dans cette région.

Un autre facteur est le nombre limité de trajectoires disponibles. Avec seulement

quatre trajectoires trouvant l'état natif, les statistiques compilées sur l'évolution des hélices n'ont pas une probabilité élevée de donner une image entièrement fiable du processus de repliement moyen. Puisque les trajectoires sont multiples et que les états définit discrètement en fonction de la structure secondaire des régions H1, H2 et H3 démontre une grande variété de conformations, nous devons concéder que les statistiques obtenues n'ont qu'une valeur qualitative. Avec ce modèle discret du repliement, et en ne comptant que les transitions qui rajoute une ou des hélices à la structure, on dénombre 10 trajectoires menant de l'état déplié UUU à l'état natif HHH. Le nombre de simulation disponible est donc insuffisant pour échantillonner adéquatement ces 10 trajectoires et obtenir des statistiques de formation aux marges d'erreur raisonnables.

Aussi, il est possible que le biais de la méthode ART nous donne une fausse image des trajectoires de repliement préférées par la protéine A. Un aspect visible du biais de la méthode est démontré par la diminution rapide du rayon de gyration dans les 100 premiers événements lancées à partir de la forme étendu. Puisque ART sous-estime l'entropie du système, les molécules ont tendance à s'écraser et à maintenir un rayon de gyration faible tout au long de des simulations. L'effet est visible jusque dans notre définition de la structure native qui ne comporte pas d'extrémités flexibles tel qu'observée dans la structure expérimentale. Bien qu'étant présent, il nous est difficile de déterminer l'impact que peut produire ce biais sur les trajectoires de repliement. Un autre facteur pouvant fausser nos observation des trajectoires de repliement provient du biais par lequel ART choisi les points de selle visités, bien qu'il est difficile d'émettre des hypothèse sur l'effet de ce biais dans nos simulations.

L'aspect le plus intéressant de notre étude provient des simulations qui n'ont pas trouvé l'état natif ou un état lui ressemblant. Une première catégorie de structures identifiées est l'image miroir de la forme native dans laquelle l'assemblage des hélices a un arrangement main-droite relativement à l'axe d'enroulement. Cette forme fut identifiée par d'autres groupes, notamment par Favrin *et al.* dont le potentiel a une représentation atomique similaire à OPEP [FIW02]. Dans leur cas, la forme

main-droite possède une énergie plus faible que la forme native alors que nous observons le contraire. Le fait que leur forme miroir ait une énergie plus basse peut être expliqué par la non-spécificité des termes énergétiques des billes de la chaîne latérale utilisés. Là où le potentiel OPEP définit une bille de chaîne latérale aux propriétés optimisées pour chaque type d'acide aminé, Favrin *et al.* n'utilise que cinq définitions de billes pour représenter les dix-huit types acides aminés réduits. Cependant, le fait que les deux modèles ont une expression spatiale similaire peut en partie expliquer les raisons pour lesquelles la structure miroir est échantillonnée par nos deux groupes.

La catégorie de simulation la plus surprenante est celle qui échantillonne des topologies correspondant à l'état natif d'autres protéines. À notre connaissance, aucune autre étude portée sur la protéine A n'a su échantillonner des structures α/β ou encore tout-β. Sous le principe de la frustration minimale, ce genre d'état ne devrait pas exister de façon stable à l'équilibre thermodynamique. Or, la différence en énergie entre la structure α/β de plus basse énergie (le motif de Rossman) et la structure native est de 5.2 kcal/mol, ce qui signifie qu'à l'équilibre la structure passe $\sim 10^{-4}$ du temps dans l'état de plus haute énergie. De plus, puisque le critère de température utilisé lors de l'échantillonnage de ces structures est le même que celui utilisé pour replier vers la forme native, le chemin de repliement qui les a trouvées est composé de transitions choisies sous le même biais de la méthode. Une population de $\sim 10^{-4}$ de ces états n'aurait pas un impact significatif sur la fonction d'une protéine. Cependant, nous croyons que ces états peuvent être des cibles potentielles d'inhibition : leur énergie plus élevée provenant en partie de contacts non-optimaux, il est possible qu'un ligand puisse être développé pour stabiliser la conformation non-native suffisamment pour inhiber sa fonction. Coïncidamment, la protéine A est un facteur de virulence de *S. aureus* qui joue un rôle nécessaire à la prolifération de la bactérie. Dans le contexte actuel d'augmentation des infections nosocomiales et en tenant compte du fait qu'il existe une souche de *S. aureus* résistante à la méthicilline (MRSA), nous croyons que notre hypothèse mérite d'être approfondie.

Outre l'aspect pharmacologique, ces états de basse énergie sont aussi intéressants du point de vue de la caractérisation des surfaces d'énergie. Des études récentes ont proposé une vision différente des théories du repliement qui sont en accord avec nos résultats. Selon Rose *et al.*, il y a un paradoxe dans la théorie actuelle de l'entonnoir où une protéine se replie en formant de façon continue de nouvelles interactions la stabilisant [RFBM06]. Étant donné le temps nécessaire pour un seul mouvement de rotation d'un angle dièdre, ils croient qu'une protéine ne peut se replier en évitant les pièges cinétiques. De plus, selon eux, la perte de l'entropie conformationnelle de l'ensemble dénaturé lors du processus de repliement ne peut être expliquée par un processus continu de diminution de l'énergie. En se basant sur le fait qu'il n'existe vraisemblablement que quelques milliers de topologies de structure secondaire et tertiaire possibles, ils énoncent une théorie du repliement basée sur les interactions du squelette carboné. Plus précisément, ils pointent le fait qu'un seul pont hydrogène non-satisfait et non-stabilisé par les interactions avec le solvant cause une augmentation de l'énergie d'environ 5 kcal/mol alors qu'un pont H non-satisfait, mais stabilisé par le solvant aqueux contribue une augmentation de 1 à 2 kcal/mol. Avec une différence d'énergie libre d'à peine 10 kcal/mol entre l'ensemble d'états dénaturés et l'état natif, un petit nombre de pont-H non-satisfait seraient suffisent pour que la structure ne puisse stabiliser l'état natif. Toujours selon leur théorie, l'ensemble dénaturé serait caractérisé par un échantillonnage discret des topologies possibles, en accord avec les résultats de repliement obtenu par échantillonnage de topomère [DCG99]. Nos résultats supportent ce nouveau modèle : à l'exception de 3 simulations trouvant des conformations tout-β, les 3 familles de minimum local affichant des éléments de structure secondaire en feuillet β ont des topologies existant dans d'autres protéines. Puisque certaines de ces topologies sont identifiées par des outils tels que FATCAT [YG04] comme étant des fragments de protéines plus volumineuses, nous ne pouvons conclure que ces structures sont des domaines complets. Une raison pouvant expliquer que la méthode ART nouveau soit la seule à identifier ces structures non-natives pourrait être le fait qu'elle sous-estime l'entropie du système, favorisant ainsi des conformations possédant plus d'éléments de

structure secondaire. Or, le modèle du repliement dirigé par le squelette exige que la stabilisation énergétique des ponts hydrogène soit plus grande que les valeurs actuellement acceptées par la communauté scientifique. Ainsi, la sous-estimation de l'entropie a pour effet secondaire de stabiliser les ponts hydrogène sans pour autant modifier leur valeur énergétique, ce qui satisfait l'exigence du modèle.

En conclusion, la méthode ART nouveau, couplée au potentiel réduit OPEP semble être un outil approprié pour identifier les configurations non-natives, mais bien définies, qui sont potentiellement échantillonnables sur la surface énergétique. Par contre, une note doit être apportée au choix d'un critère de température. Nous avons dit plus haut que les états non-natifs visités étaient probables puisqu'ils avaient été atteints à l'aide d'un critère de température permettant de replier la protéine. Le critère de température de Métropolis choisi dans ce travail n'a pas été validé afin de déterminer s'il était optimal pour la protéine A. Nous savons seulement qu'un critère de 600 K ne peut sortir des pièges profonds de certains minima locaux alors que l'on peut trouver le minimum global à 900 K. Il serait donc intéressant dans une prochaine étude de déterminer si un choix de température légèrement plus élevé ou plus bas que 900 K influence la proportion de structures non-natives de basse énergie.

CHAPITRE 6

MÉTHODES ACCÉLÉRÉES

L'étude du repliement des protéines et de l'agrégation de peptides, ou même la simulation à l'équilibre thermodynamique des sites actifs des enzymes ont, dans les quinze dernières années, fait des avancées remarquables. La découverte d'astuces permettant l'accélération des simulations d'échantillonage et de dynamique moléculaire tel que la méthode d'échange des répliques [SO99] ou encore la méthode de la dynamique des ensembles [SP01] permettent aujourd'hui d'étudier des trajectoires de repliement plus longues. Dans cette dernière, plusieurs simulations sont lancées en parallèle à partir de conditions initiales quasi identiques et sont exécutées jusqu'à l'apparition d'un événement considéré rare tel que la sortie d'un bassin de configurations vers un autre bassin. Les simulations sont alors arrêtées et on conserve la trajectoire intéressante. Dans le cas des protéines où l'on croit que le nombre d'événements de transition inter-bassin suit une loi de Poisson, le temps nécessaire à une transition suit une loi exponentielle. À l'aide de cette propriété, on peut approximer le temps moyen de transition d'un événement donné par la somme des temps des simulations parallèles au moment où elles ont été arrêtées.

Cette méthode de parallélisation n'est pas limitée par les temps de communications entre les ordinateurs, ce qui la rend distribuable sur des réseaux où les protocoles de communication ne sont pas contrôlés, tels que le réseau d'ordinateurs personnels de grande échelle qu'est l'Internet. De plus, contrairement aux méthodes traditionnelles de parallélisation dans lesquelles la tâche du calcul du gradient de force d'un point est distribuée à plusieurs processeurs d'un même superordinateur, la méthode de dynamique des ensembles affiche une accélération du temps de calcul nécessaire qui est linéairement proportionnel au nombre de processeurs utilisés jusqu'à une valeur critique correspondant au temps minimal requis pour transiter d'un bassin à un autre par le chemin le plus court possible. Ainsi, si le temps moyen de repliement d'une protéine est de 10 μs et que le temps minimal pour qu'une

seule protéine parcoure la trajectoire de repliement la plus courte est de 10 ns, l'utilisation de plus de 1000 simulations parallèles atteindra le seuil de dégradation de l'efficacité où l'accélération devient sous-linéaire. Or, en tenant compte qu'un ordinateur moderne peut simuler 2 ns par jour de la trajectoire de repliement d'une protéine, la méthode de dynamique des ensembles permet en théorie de trouver une trajectoire de repliement en 5 jours de calcul sur 1000 processeurs là où normalement 1 an et demi seraient nécessaires pour un seul.

La méthode de dynamique des ensembles est un pas en avant pour la dynamique moléculaire et elle est à la base du réseau hautement distribué *Folding@home* du groupe de recherche de Pande [LSSP02]. Cependant, bien que la méthode permet aujourd'hui d'étudier des systèmes plus complexes sur des temps compatibles avec les exigences de publication de résultats, l'obtention de ces résultats repose sur la disponibilité de temps-processeur de millier d'ordinateurs personnels. Puisqu'il faut convaincre les propriétaires de ces ordinateurs de partager leur temps de calcul, la ressource est limitée par l'efficacité du plan de marketing appliqué. À ce problème se rajoute un coût environnemental souvent négligé : bien que le groupe vende l'image d'une contribution scientifique au coût négligeable des cycles-processeurs qui seraient normalement "perdus" par les propriétaires d'ordinateurs personnels, il est bon de se rappeler qu'un processeur utilisé à sa pleine capacité consomme jusqu'à 30 watts de plus qu'un processeur en attente. Les "cycles perdus" des 188760 processeurs actifs[1] consomment en moyenne entre 1.8 et 2.8 mégawatts d'électricité[2], l'équivalent d'une ville américaine de 5100 à 7500 habitants[3]. Ce nombre peut sous-estimer le coût environnemental réel puisqu'il ne tient pas compte des gens qui contribuent plus que leurs "cycles inutilisés" en laissant

[1]Valeurs obtenue en date du 20 décembre 2006 sur le site web du groupe : http ://folding.stanford.edu/

[2]Estimation basée sur une différence de consommation de 20 à 30 watt entre un processeur en attente et un processeur pleinement utilisé et sur un temps d'activité de 12 heures par jour par processeur.

[3]Estimation basée sur la consommation moyenne d'un ménage américain en l'an 2004 de 958 watts par habitation par jour et sur une occupation moyenne de l'an 2000 de 2.59 habitants par habitation

leur ordinateur allumé lorsqu'ils ne l'utilisent pas. En bref, le développement de méthodes accélérés permettant des gains de temps de simulation et ne reposant pas sur une parallélisation intensive du processus de calcul aurait pour avantage de diminuer l'impact environnemental en plus de permettre l'étude de systèmes de plus grande envergure. À ceci se rajoute le problème du biais introduit par la méthode : puisque l'exploration d'un bassin de configuration est arrêté après la première transition, seul les événements aux transitions rapides sont échantillonnés. Bien qu'il puisse y avoir un lien entre la rapidité d'une transition et sa probabilité, il est nécessaire d'échantillonner les transitions plus lentes qui permettent de sortir des pièges entropiques où une configuration peut visiter exclusivement un sous-groupe de bassins de configuration.

La méthode ART nouveau, introduite en 2.2.2.2, est une méthode activée qui permet d'échantillonner une surface énergétique de façon efficace. On entend par activation que la méthode recherche des points de selle pour faciliter le passage d'un bassin de configuration à un autre élimine entièrement l'étape lente de la dynamique moléculaire qui consiste à échantillonner les configurations d'un bassin confiné. Étant donné que la méthode parcourt le chemin le plus court menant d'un minimum local à un point de selle, les fluctuations thermiques aléatoires n'y sont pas simulées, sous-estimant ainsi l'entropie et ne satisfaisant pas au bilan détaillé. Ces lacunes ne diminuent pas le fait qu'elle ait pu être utilisée avec succès dans l'étude de systèmes aussi complexes que la protéine A présentée dans l'article du chapitre 4, et cela avec un temps de calcul d'à peine deux mois-processeur par simulation. Cependant, l'étude de systèmes de plus grande envergure est encore aujourd'hui difficile étant donné l'élément limitant de la méthode, les appels au potentiel énergétique, affiche un temps d'exécution de $O(N^2)$ où N est le nombre d'atomes. Puisque l'optimisation du potentiel est limitée par le degré de réalisme souhaité, une autre avenue d'optimisation est de diminuer le nombre d'appels à la routine du potentiel. Pour y arriver, deux options non-exclusives sont possibles. Dans les mouvements dirigés, on peut :

- Prendre des pas de plus grande taille entre les appels au potentiel.

- Prendre des pas nécessitant moins de corrections perpendiculairement à la direction du mouvement.

La méthode ART nouveau est présentement implémentée dans un contexte où le travail est effectué en coordonnées cartésiennes. La méthode n'est pourtant pas limitée à ce choix de coordonnées et en combinaison avec un potentiel énergétique en coordonnées internes, la méthode ARTIST fonctionne dans ce jeu de coordonnées [YLM+06]. En théorie, les méthodes de simulations activées et les méthodes dirigées peuvent bénéficier grandement de l'utilisation de coordonnées internes : ce choix de coordonnées permet de fixer plus facilement certains degrés de liberté tels que la longueur des liens et la valeur des angles de valence qui ne varie que très peu lors du repliement des protéines. Cependant, un problème avec les méthodes en coordonnées internes est qu'un mouvement local tel que la rotation autour d'un axe dièdre a des répercussions non-locales pouvant causer des sauts énergétiques difficiles à amortir.

Dans ce qui suit, nous vous présentons une description des degrés de liberté des protéines, suivi par l'examen de trois méthodes d'imposition de contraintes permettant d'isoler les degrés de liberté nécessaires au repliement. Deux de ces algorithmes ont été implémentées et testés dans la méthode ART nouveau dans le but de déterminer si l'imposition de contraintes sur les mouvements atomiques pouvait améliorer l'efficacité de ART.

6.1 Mouvement dans les protéines

À leur plus simple définition, les protéines sont des polymères linéaires munis de courtes ramifications. Comme il a été discuté plus tôt, les mouvements observés lors d'un changement de conformations d'une protéine sont principalement des rotations autour des axes dièdres correspondants aux angles ϕ et ψ de la chaîne carbonée et aux angles χ des chaînes latérales (voir figure 2.1). Les longueurs de lien et les angles de valence observés dans les structures cristallines de protéines ne varient que par 1 à 2% de leur valeur d'équilibre. À l'exception des résidus

précédant les prolines, les atomes du plan peptidique adoptent presque toujours une conformation trans et un angle dièdre ω de -180 degrés.

Les mouvements en coordonnées cartésiennes prennent le plus souvent la forme de vecteurs de translations linéaire. Or, les degrés de liberté de la chaîne carbonée s'expriment par des rotations autour de liens axiaux. Il s'en suit que les mouvements de translation déforment la géométrie rigide autant que les angles dièdres. Prenons l'exemple de l'éthane (Figure 6.1), la plus petite molécule non plane possédant un angle dièdre. Si la position des atomes du groupement CH_3 de l'arrière-plan est fixe, d'application d'un vecteur de force sur un hydrogène du CH_3 d'avant-plan aura pour effet de produire une modification de l'angle de valence et de la longueur du lien de celui-ci avec le carbone (Figure 6.1 (b)). Cependant, si on projette le mouvement cartésien sur les coordonnées internes de l'éthane, on peut rigidifier la longueur des liens et les angles de valence et produire un mouvement net correspondant à une rotation autour de l'axe carbone–carbone (Figure 6.1 (c)). On obtient le même mouvement en coordonnées cartésiennes, mais seulement après avoir rééquilibré la valeur des angles et liens par des étapes de minimisations. Puisque chaque pas de minimisation nécessite un calcul du gradient d'énergie d'ordre $O(N^2)$, un mouvement en coordonnées internes peut potentiellement être plus efficace puisqu'il est plus aisé de contrôler le stress appliqué à la structure rigide.

Une structure quelconque de N atomes possède $3N$ coordonnées cartésiennes, correspondant aux valeurs en x, y et z de chaque atome. Le jeu de coordonnées internes nécessaire pour représenter la même structure compte $3N-6$ coordonnées indépendantes correspondant aux degrés de libertés internes. À cela se rajoutent trois coordonnées de positionnement et trois coordonnées d'orientation spatiale de la structure.

Il existe plus de coordonnées internes à une structure que ce qu'il est nécessaire pour la représenter ; seulement pour les coordonnées internes faisant la relation entre deux corps, on compte $N \times (N-1)/2$ possibilités. Il est donc primordial de choisir un jeu de coordonnées permettant une représentation complète de l'espace

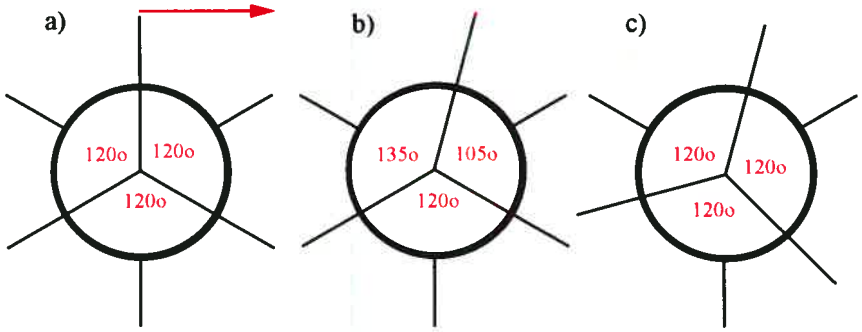

FIG. 6.1 – Molécule d'éthane (a) à sa position d'équilibre avant l'application d'une force, (b) après l'application d'un mouvement en coordonnées cartésiennes, (c) après le même mouvement en coordonnées internes rigidifiées. Les valeurs des angles sont calculées à partir d'une projection des liens sur le plan occupé par le carbone d'avant-plan et ne représente pas les valeurs à l'équilibre du tétraèdre tridimensionnel du groupe CH_3

cartésien de dimension $3N$ tout en étant significatif et faiblement corrélé. Un choix populaire de coordonnées internes est la représentation par la matrice Z et sa variante la matrice Z délocalisée. Un exemple est donné pour l'éthane dans le tableau 6.1. Dans la matrice Z délocalisée, la première ligne donne la position cartésienne du premier atome, la seconde ligne donne la distance du second atome avec le premier atome ainsi que deux angles d'orientations de la molécule, la troisième ligne donne la distance entre le troisième et deuxième atomes, l'angle entre les trois premiers atomes, et un angle dièdre d'orientation de la molécule. Ainsi, les trois premières lignes jouent le double rôle de positionner et d'orienter la molécule dans l'espace ainsi que de donner les coordonnées internes des 3 premiers atomes. Toutes les lignes qui suivent ont le même format : une longueur de lien avec un atome directement connecté, un angle de valence et un angle dièdre. Cette représentation est suffisante pour décrire la structure de n'importe quelle protéine.

Le passage de coordonnées cartésiennes à coordonnées internes se fait de façon linéaire par simple application de formules de géométrie. Pour connaître le mouvement interne associé à un déplacement cartésien, il est nécessaire de projeter le

TAB. 6.1 – Représentation cartésienne et interne (matrice Z délocalisée) d'une molécule d'éthane ($H_3C - CH_3$)

Coord. Cartésiennes

H_1	x_{H_1}	y_{H_1}	Z_{H_1}
H_2	x_{H_2}	y_{H_2}	Z_{H_2}
H_3	x_{H_3}	y_{H_3}	Z_{H_3}
C_1	x_{C_1}	y_{C_1}	Z_{C_1}
C_2	x_{C_2}	y_{C_2}	Z_{C_2}
H_4	x_{H_4}	y_{H_4}	Z_{H_4}
H_5	x_{H_5}	y_{H_5}	Z_{H_5}
H_6	x_{H_6}	y_{H_6}	Z_{H_6}

Matrice Z

H_1	x_{H_1}	y_{H_1}	Z_{H_1}
C_1	$Lien_{H_1-C_1}$	$Angle_{C_1-H_1-X}$	$Angle_{C_1-H1-Y}$
C_2	$Lien_{C_2-C_1}$	$Angle_{C_2-C_1-H_1}$	$Diedre_{C_2-C_1-H_1-X}$
H_4	$Lien_{H_4-C_2}$	$Angle_{H_4-C_2-C_1}$	$Diedre_{H_4-C_2-C_1-H_1}$
H_5	$Lien_{H_5-C_2}$	$Angle_{H_5-C_2-C_1}$	$Diedre_{H_5-C_2-C_1-H_1}$
H_6	$Lien_{H_6-C_2}$	$Angle_{H_6-C_2-C_1}$	$Diedre_{H_6-C_2-C_1-H_1}$
H_2	$Lien_{H_2-C_1}$	$Angle_{H_2-C_1-C_2}$	$Diedre_{H_2-C_1-C_2-H_4}$
H_3	$Lien_{H_3-C_1}$	$Angle_{H_3-C_1-C_2}$	$Diedre_{H_3-C_1-C_2-H_4}$

vecteur de déplacement sur les vecteurs de mouvements maximaux de chaque coordonnée interne. Par exemple, le cas le plus simple est celui du lien covalent entre deux atomes A_1 et A_2. Dans ce cas, un vecteur de déplacement aura un impact maximal sur la longueur du lien $A_1 - A_2$ si les déplacements de chacun de ces deux atomes sont parallèles au vecteur du lien tout en étant de direction opposée, signifiant un étirement ou une compression du lien. On a donc les vecteurs d'amplitude maximale :

$$S_{A_1-A_2} = \overrightarrow{A_1A_2} \tag{6.1}$$

et

$$S_{A_2-A_1} = -S_{A_1-A_2} = \overrightarrow{A_2A_1} \tag{6.2}$$

Le passage du déplacement cartésien à interne est donné par les produits scalaires :

$$\Delta_{A_1-A_2} = S_{A_2-A_1} \cdot \Delta_{A_1} + S_{A_1-A_2} \cdot \Delta_{A_2} \tag{6.3}$$

Une liste des équations de vecteurs d'amplitude maximale pour les angles de

valence, les angles dièdres et les angles impropres est donnée par Wilson dans son livre [WDC80]. Ces équations ne sont malheureusement pas applicables aux cas où l'on désirerait fixer certains degrés de liberté internes pour n'en permettre que d'autre. Bien que l'on puisse immobiliser les longueurs de liens ou les angles de valence à l'aide de ces équations, le système ainsi simulé n'est plus physique : dans une simulation normale, un lien qui est comprimé lors d'un déplacement génèrera des vecteurs de force dans les deux directions qui seront rééquilibrés lors du prochain déplacement. Il s'en suit qu'après rééquilibration, les deux atomes du lien se seront déplacés dans la direction du premier mouvement ayant comprimé le lien. Ce processus doit être émulé dans les systèmes où des contraintes sont imposées : les forces qui normalement modifieraient un degré de liberté interne rigidifié doivent être transmises le long de la chaîne carbonée et être transformées en mouvement de degrés de liberté souples.

Il est cependant possible de travailler dans l'espace des angles dièdres avec des méthodes d'échantillonnage de type Monte-Carlo. Il est alors possible de faire varier la valeur des angles dièdres pour générer les conformations spontanées (ie. : qui ne sont pas reliées entre-elle par un chemin continu établi). Cependant, ces méthodes doivent faire face à un autre problème relié au mouvement en coordonnées internes : contrairement aux déplacements en coordonnées cartésiennes, les mouvements effectués en coordonnées internes ont des effets non-locaux. Ainsi, sur une chaîne linéaire dont la conformation spatiale est étendue, l'allongement de n'importe quel lien aura pour effet d'éloignes les atomes des deux extrémités. Similairement, la modification d'un angle dièdre par 3 degrés pourra résulter en un déplacement relatif de 0.5 Å quelques 10 Å plus loin sur la chaîne. Pour éviter que des collisions stériques se produisent lors d'un déplacement dans l'espace des angles dièdres, soit il faut utiliser une taille de pas plus petite, soit on doit avoir recours à des méthodes plus complexes telles que celles ici présentées.

6.2 Méthode de Bystroff

Cette première méthode à avoir été implémentée dans ART nouveau et publiée en 2001 propose une solution exacte au transfert des déplacements d'un système de référence cartésiens à l'espace des angles dièdres internes [Bys01]. Pour y arriver, toutes les paires d'atomes doivent être examinées afin de déterminer le changement en distance interne entre chaque paire :

$$dD_{ij} = (d\vec{x}_j - d\vec{x}_i) \cdot u_{ij} \tag{6.4}$$

où $d\vec{x}_i$ et $d\vec{x}_j$ sont les vecteurs de déplacement appliqués sur les atomes i et j, et u_{ij} est le vecteur unitaire pointant de i vers j. Selon Bystroff, un mouvement inter-atomique non-nul s'exprime au niveau de la protéine par une altération des angles dièdres K_{ij} séparant séquentiellement la paire d'atomes. Le déplacement inter-atomique peut alors être exprimé par la somme des dérivées de la distance en fonction des angles de rotation libres séparant la paire d'atomes :

$$dD_{ij} = \sum_{k \epsilon K_{ij}} \frac{dD_{ij}}{\theta_k} \theta_k \tag{6.5}$$

où θ_k est l'angle dièdre k séparant la paire d'atomes i et j. Le terme de dérivée de la somme précédente est détaillé sous la forme :

$$\frac{dD_{ij}}{\theta_k} = u_{ij} \cdot (u_k \otimes \vec{r}_{kj}) \tag{6.6}$$

où u_k est le vecteur unitaire de l'axe de rotation du dièdre k et $\vec{r}_{kj}$ est un vecteur reliant un point sur l'axe de rotation à l'atome j. Le produit vectoriel $u_k \otimes \vec{r}_{kj}$ correspond au vecteur d'amplitude maximale du mouvement puisqu'il est tangentiel à l'axe de rotation. La projection par produit scalaire du vecteur unitaire de mouvement u_{ij} sur le vecteur d'amplitude maximale donne une valeur proportionnelle à la capacité de l'angle d'accommoder le mouvement inter-atomique par une rotation.

Puisque dans un système à N atomes il y a $N(N-1)/2$ paires d'atomes et K angles dièdres, la résolution des déplacements angulaires est déterminée par un système de $N(N-1)/2$ équations linéaires et de K inconnus où $N > K$. La méthode proposée par Bystroff pour résoudre le problème est celle des moindres carrés. Il bâtit donc une matrice du système et la porte au carré :

$$m_{k(ij)} = u_{ij} \cdot (u_k \otimes \vec{r}_{kj}) \tag{6.7}$$

$$M = m \times m^{\dagger} \tag{6.8}$$

La matrice carrée M de taille K est ensuite inversée et multipliée par le vecteur des déplacements projetés sur les vecteurs d'amplitude maximale :

$$\vec{A}_k = \sum_i \sum_j (u_k \otimes \vec{r}_{kj})((d\vec{x}_j - d\vec{x}_i) \cdot u_{ij}) \tag{6.9}$$

$$M^{-1}\vec{A} = d\vec{\theta} \tag{6.10}$$

6.2.1 Analyse

La méthode de Bystroff n'est pas la plus efficace des méthodes que nous présentons. Rappelons que chaque mouvement est accompagné d'un calcul du gradient de force d'un ordre d'exécution $O(N^2)$, peu importe le choix de coordonnées. Dans ce cas-ci, on voit que la matrice m est de taille $K \times (N(N-1)/2)$ où K est le nombre d'angles dièdres et N est le nombre d'atomes. L'ordre d'espace mémoire est donc de $O(N^3)$. Cependant, on peut démontrer qu'environ $N(N-1)/2$ cases sont occupées, donnant un ordre d'espace de $O(N^2)$. Ceci découle du fait que pour toute paire d'atomes, seuls les angles dièdres séquentiellement entre ces deux atomes ont une valeur non-nulle dans la matrice m. Il va de soit que l'ordre d'exécution de la matrice m est aussi de $O(N^2)$. En se basant sur cette propriété, la multiplication de la matrice m par sa transposée $m^{\dagger}$ pour obtenir M peut être exécutée en temps

$O(N^3)$ par un jeu d'optimisation éliminant les multiplications de termes de valeur nulle. La double sommation de la construction du vecteur $\vec{A}$ nécessite un temps d'exécution de $O(N^2)$, et l'étape la plus longue, l'inversion de la matrice M est elle aussi d'ordre $O(N^3)$ en utilisant la décomposition de Cholesky [PFP+86]. Puisque toutes ces étapes sont indépendantes et séquentielles, l'ordre global d'exécution est l'ordre le plus long de $O(N^3)$.

Notez que notre analyse est en désaccord avec l'évaluation de l'auteur qui situe l'ordre d'exécution à $O(N^2)$. Lors de nos essais, les appels à l'algorithme de Bystroff étaient significativement plus lents que nos appels au potentiel énergétique d'un temps d'exécution de $O(N^2)$. De plus, l'auteur dans une conversation par courriel nous a indiqué que les déplacements inter-atomiques pouvaient aussi bien être calculés à partir du gradient de force qu'à partir des termes spécifiques d'attraction ou de répulsion des paires d'atomes. Rappelons que le gradient de force est un vecteur de taille $3N$ exprimant les forces totales en x, y, et z qui sont imposées à chaque atome. Ainsi, en utilisant les vecteurs de forces du gradient pour déterminer le déplacement inter-atomique, une partie de la composition de ces vecteurs provient d'interactions autres que la paire d'atomes impliquée. Lors d'essais sur un modèle simplifié à trois atomes (A,B,C) interagissant et séparés par deux angles dièdres, l'utilisation du gradient de force démontra que la composante d'attraction de l'atome B vers l'atome A est aussi perçue par l'algorithme comme une force de répulsion de l'atome B envers l'atome C, bien que ces deux atomes n'aient pas de terme énergétique.

Le problème n'existe pas lorsqu'on utilise des termes énergétiques à deux corps pour définir les déplacements inter-atomiques. Cependant, ART nouveau est une méthode utilisant des vecteurs unifiés de taille $3N$ pour les forces ainsi que pour les directions données par l'analyse de la courbure énergétique. Il fut donc conclu qu'il était impossible d'utiliser l'algorithme de Bystroff avec notre méthode pour les mouvements unifiés qui ne peuvent être exprimés sous forme de termes inter-atomiques. L'algorithme peut cependant être utilisé lors de l'étape de minimisation en modifiant le potentiel à fin d'obtenir une matrice de termes de force au lieu du

gradient de force, mais sont temps d'exécution d'ordre $O(N^3)$ le rend prohibitif et n'offre pas d'avantages substantiels puisque les appels au potentiel énergétique que l'algorithme remplace ont un ordre de $O(N^2)$.

6.3 Méthode SHAKE

Malgré le fait que nous ne l'ayons pas implémenté dans ART nouveau, la méthode SHAKE [RCB77] est ici mentionnée de par sa popularité d'utilisation dans les simulations de dynamique moléculaire. Développée à la fin des années 70 dans le but d'accélérer l'étude des molécules n-alkane alors considérées complexe, SHAKE est une adaptation de l'algorithme d'intégration de Verlet qui corrige les mouvements qui ne respectent pas un jeu de contraintes. Les contraintes, qui peuvent être imposées sur n'importe quelle coordonnée interne, sont habituellement imposées seulement sur la longueur des liens, laissant les angles libres. De par le formalisme des multiplicateurs de Lagrange, on redéfinit l'algorithme de Verlet (Eq. 2.8) comme étant :

$$V(t_i + \frac{\Delta t}{2}) = V(t_i - \frac{\Delta t}{2}) + a(t_i)\Delta t + \frac{\Delta t}{M(X)} G(t - \frac{\Delta t}{2}) \tag{6.11}$$

où G est le vecteur de forces correctives pour maintenir les contraintes. Les termes de force pour chaque atome sont donnés par :

$$G_i(t) = \sum_{k=1}^{L} \lambda_k \nabla_i \sigma_k \tag{6.12}$$

où L est le nombre de contraintes impliquant l'atome i, λ_k est le multiplicateur de Lagrange et σ_k est la contrainte imposée. Une contrainte de distance entre deux atomes i et j aurait la forme :

$$\sigma_{ij} = d_{ij}^2 - \overline{d_{ij}}^2 = 0 \tag{6.13}$$

où d_{ij} est la distance entre les atomes i et j et $\overline{d_{ij}}$ est la distance à l'équilibre pour ces deux atomes. Il en résulte un système d'équations linéaires où les λ_k sont les

inconnus.

Dans un système à K contraintes, une résolution analytique serait de l'ordre de $O(K^3)$ provenant du coût d'inversion d'une matrice de taille K. La méthode préconisée pour résoudre le système d'équations est par une méthode itérative. On construit quatre listes de contraintes de façon à ce qu'aucune contrainte d'une liste ne partage un atome avec une autre contrainte de la même liste. Ce nombre de liste provient du fait que les carbones hybridés sp3 peuvent établir quatre liens avec des atomes voisins. Par itération sur les quatre listes, on choisira des valeurs de λ_k qui rétablissent les contraintes jusqu'à ce qu'il y ait convergence et que les écarts aux contraintes tombent sous un seuil prédéterminé.

La méthode SHAKE a inspiré plusieurs autres méthodes : RATTLE, une version qui contraint à la fois les positions atomiques et les vecteurs de vitesse [And83], SETTLE, une version analytique exacte de RATTLE qui peut être utilisée sur les petites molécules telles que les molécules d'eau [MK92], et aussi QSHAKE, qui utilise le formalisme des quaternions pour modéliser des fragments rigides et les contraindre.

6.4 Méthode FRODA

La seconde méthode à avoir été implémentée par notre groupe afin de tenter d'accélérer les simulations d'ART nouveau est une méthode de simulation géométrique nommée FRODA [WMHT05]. Cette méthode a été développée dans le but d'étudier les degrés de flexibilité de larges protéines et est jumelée à la méthode de percolation des régions de flexibilité FIRST [JRKT01]. Cette dernière n'a toujours pas été implémentée dans des simulations de dynamique moléculaire. La méthode affiche quelques ressemblances à l'imposition de contrainte de SHAKE en ceci qu'elle est une méthode itérative de correction des positions atomiques. Cependant, là où SHAKE est utilisé sur des contraintes de quelques atomes tels que des liens et angles de valences, FRODA est appliqué sur des fragments rigidifiés de trois atomes ou plus. Un exemple de corps rigide à l'intérieur d'une protéine serait

le groupe de 5 atomes ayant en son centre un carbone-α tétraédrique et ses quatre voisins.

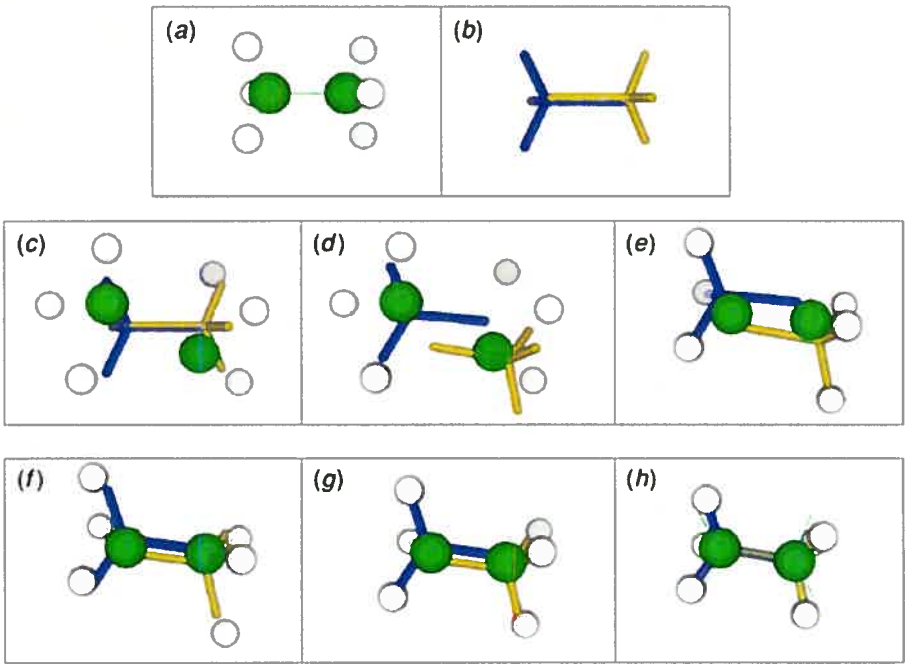

FIG. 6.2 – Itérations des étapes d'imposition de contraintes de FRODA sur une molécule d'éthane. À partir d'une configuration atomique à l'équilibre (a), on définit un ensemble de corps fantômes décrivant les parties rigide(b). Suite à un mouvement des atomes (c), les corps fantômes sont repositionnés de façon à diminuer la distance RMSD (d), avant que la position des atomes soit réétablie sur les corps fantômes (e). Ces deux étapes sont itérées (f et g) jusqu'à convergence (h). Image extraite de l'article de Wells [WMHT05]

Un exemple d'application de FRODA sur une molécule d'éthane est donné dans la figure 6.2. L'étape d'initialisation de la méthode consiste en la définition des corps rigides (Fig. 6.2 (b)) et à leur apposition sur les atomes. Le choix des atomes formant un corps fantôme rigide est à la discrétion de l'utilisateur et peut comprendre tous les atomes d'un domaine s'il est désirable de retirer tous les degrés de liberté de ce dernier. Habituellement, on choisira un groupe d'atomes liés entre eux par un atome central et qui affiche un comportement rigide. Ce groupe d'atome, s'il est retiré de la molécule, peut donc inclure des angles dièdres si ces angles sont

rigides tel que l'angle ω du plan peptidique. Fait important : les corps rigides ont des atomes partagés avec leurs voisins. Dans le cas de l'exemple de l'éthane, on note que les deux atomes de carbones en vert font parti des deux corps rigides. Lorsqu'une paire d'atomes est ainsi partagée, le lien unissant cette paire forme l'axe d'un angle dièdre libre. Si plus de deux atomes sont partagés, la paire de corps fantômes devient rigide ou semi-rigide dépendamment du degré d'imposition des contraintes.

Une fois définis, les atomes d'un corps fantôme se voient attribuer une valeur correspondant à la position des atomes réels. Ces positions atomiques sont initialement déterminées comme minimisant l'énergie des liens covalents et des angles de valence à l'intérieur du fantôme. Ils peuvent cependant être redéfinis au besoin lorsqu'il est préférable pour les atomes d'adopter une position qui dévie de l'équilibre, par exemple dans un minimum local où la structure est plus compacte et où les contraintes rigides sont moins respectées.

Lorsqu'une simulation est exécutée et que des atomes sont déplacés (Fig. 6.2 (d)), le cycle d'imposition des contraintes est initié par le positionnement du corps fantôme au point et à l'orientation qui minimise la distance RMSD entre les atomes du fantôme et les atomes réels (Fig. 6.2 (d)). On génère alors un vecteur de déviation de la position réelle vis-à-vis de la position fantôme. Les atomes partagés entre plusieurs corps fantômes se voient attribués une déviation qui est la somme des déviations à chaque fantôme auquel ils contribuent. L'étape suivante est donc d'appliquer un mouvement aux atomes réels dans la direction du vecteur de déviation qui aura pour effet de minimiser l'écart entre les atomes réels et leur représentation fantôme. L'itération des étapes (d) et (e) a pour effet de graduellement diminuer la déviation entre les atomes réels et leurs fantômes jusqu'à ce qu'il en résulte une nouvelle conformation aux contraintes respectées (Fig. 6.2 (h)) qui ne diffère de la configuration initiale que par une valeur différente de l'angle dièdre.

6.4.1 Implémentation

Pour une première version de FRODA-ART, nous sommes intéressés à simuler tous les mouvements de torsion autour des axes dièdres permis par le potentiel. La définition des corps fantômes est aussi dépendante de la représentation réduite du potentiel OPEP (voir 2.2.2). À l'exception des prolines, les corps fantômes de tous les acides aminés sous OPEP ont la même représentation : un corps rigide liant la bille de la chaîne latérale au carbone-α central et aux atomes d'azote et de carbone des plans peptidiques voisins (Fig. 6.3 (a) et (e)). Le carbone-α de ces corps est aussi partagé avec deux autre corps qui forment les plan peptidiques voisins (Fig. 6.3 (d)). Le cas de la proline à ceci de particulier qu'il incorpore l'angle ϕ et le contraint à une valeur de -90 degrés (Fig. 6.3 (c)), ce qui a pour effet de libérer l'angle ω puisque celui-ci n'est pas inclu dans le corps rigide du plan peptidique pré-proline (Fig. 6.3 (b)).

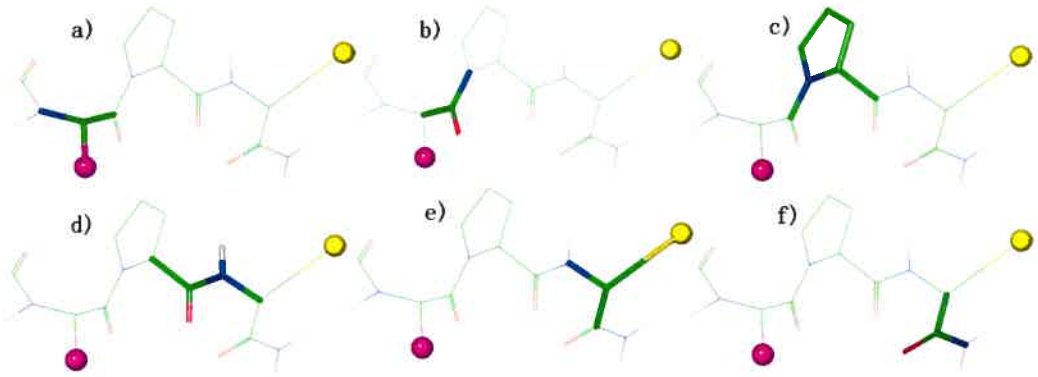

FIG. 6.3 – Différents corps rigides sur un peptide de trois acides aminés dans la représentation réduite d'OPEP.

Il existe trois types de mouvements dans la méthode ART nouveau qui nécessitent d'être considérés lors de l'application de contraintes par apposition des corps fantômes (décrites en détaille en 2.2.2.2. Tout d'abord, dans la phase d'activation, le cas le plus simple d'un mouvement dirigé est celui de par lequel la configuration est poussée à l'extérieur d'un bassin harmonique (a)). La direction alors employée

est un vecteur aléatoire duquel on a retiré la composante parallèle au vecteur de forces rigides découlant de la modification de longueurs de liens et d'angles de valence. Ici, la méthode FRODA n'a pas à être précise puisque le but visé n'est que de déformer la configuration initiale suffisamment pour que la méthode de Lanczos puisse détecteur un vecteur propre de valeur propre négative pointant vers un point de selle.

En second lieu, l'étape suivant la sortie du bassin harmonique consiste à suivre le vecteur propre de plus faible valeur propre jusqu'à ce que la configuration du point de selle soit atteinte (section 2.2.2.2 b)). Normalement, la direction du vecteur propre ne varie que faiblement lors du parcours du chemin menant au point de selle. Par contre, les vecteurs propres doivent être calculés dans un espace non-contraint afin d'assurer la stabilité de l'algorithme de Lanczos. Puisque le vecteur propre pointe dans une direction où la dérivée seconde de l'énergie est négative, la composante dans la direction du vecteur de forces rigides, dont les termes harmoniques ont une dérivée seconde positive, devrait être minime.

Le dernier type de mouvement à examiner est celui de l'étape de minimisation (section 2.2.2.2 c)). Ici l'algorithme couramment utilisé est une dynamique moléculaire amortie. L'algorithme consiste à prendre des pas avec une vitesse croissante dans la direction du vecteur de force. Le vecteur de vitesse qui est suivit, construit à partir des contributions successives des vecteurs de force, est amorti à chaque pas par un terme de friction virtuel et est remis à zéro lorsque l'orientation de ce vecteur est opposée à celle du vecteur de force (lorsque leur produit scalaire est négatif). De par la topologie rugueuse de la surface énergétique, les mouvements de cette étape sont rarement rectilinéaires du début à la fin : plus on approche de la valeur minimale du bassin harmonique, plus le vecteur de force est souvent remis à zéro. La minimisation apporte une difficulté supplémentaire à l'apposition de corps rigides : l'état d'énergie minimale d'un bassin correspond aussi à l'état le plus compact. Les liens et angles de valence de certains atomes centraux ont souvent une valeur déviant de la valeur à l'équilibre.

6.4.2 Analyse

Suite à l'implémentation de l'algorithme FRODA dans ART nouveau, quelques examens de la stabilité à l'équilibre ont été effectués. La structure choisie pour ces examens est l'épingle β utilisée lors des premières publications de l'application de ART nouveau sur des protéines [WDM03, WMD04a]. Cette protéine de 16 acides aminés et de 99 atomes sous sa représentation réduite OPEP possède 32 angles dièdres libres. Puisqu'elle ne contient pas de prolines, sa représentation en corps fantômes FRODA compte 16 corps fantômes de 4 atomes dont le centre est un carbone-α, 15 corps de 6 atomes pour les plans peptidiques, puis 2 corps pour les extrémités N et C.

Lors du premier test, on assigne une position d'équilibre aux corps fantômes sur une chaîne étendue de l'épingle. La configuration étendue est préalablement relaxée et son énergie minimisée est de -4.252 kcal/mol. La contribution rigide à cette énergie provenant des angles de valence, des longueurs de liens et des angles de torsion est de 0.837 kcal/mol. Pour les tests subséquents, on considère cette énergie comme étant la valeur à l'équilibre des corps rigides. Les autre contributions énergétiques sont pour l'instant ignorées puisque nous sommes intéressés tout d'abord par l'aspect géométrique du comportement de l'algorithme FRODA.

Pour déterminer l'erreur numérique d'application des corps rigides, on examine la distance RMSD qu'engendre la correction du vecteur de position par apposition des corps fantômes. Puisque les corps fantômes et leurs atomes réels sont parfaitement superposés, la correction n'implique qu'une déviation de la distance RMSD de 2.5×10^{-14} Å qui ne varie pas en fonction du nombre d'itérations. Un examen similaire a été fait pour l'algorithme de conversion de coordonnées cartésiennes à coordonnées internes sous la représentation de la matrice Z, suivi par la conversion inverse. Les multiples opérations géométriques créent une déviation de 3.7×10^{-13} Å de la distance RMSD par application ou itération. Ces valeurs sont dans les marges d'erreur reliées à l'utilisation de variables à virgule flottante de dimension double.

On pousse par la suite l'étude de cette déviation due à l'erreur numérique en examinant la correction par apposition des corps fantômes lorsque les atomes de la chaîne sont dans une position d'équilibre, mais que leur orientation spatiale diffère de celle des corps fantômes. Pour y arriver, on induit une modification des valeurs des angles dièdres libres en passant par la matrice Z tout en conservant les autres coordonnées internes dans leur position d'équilibre. Cinq conformations sont générées à partir de la forme étendue dans lesquelles tous les angles dièdres libres sont modifiés par une valeur moyenne de 18, 36, 54, 72 et 90 degrés. Les courbes de déviation de distance RMSD suite à l'apposition des corps fantômes sur ces cinq structures sont présentées en figure 6.4. On remarque que dans tous les cas, les premières itérations de la méthode FRODA n'ont qu'un faible impact mesuré par une faible déviation de la distance RMSD d'au plus 1.8×10^{-12} Å signifiant que les corps fantômes sont bien réorientés avant d'être apposés sur la position de leurs atomes réels. La déviation notée lors de l'augmentation du nombre d'itérations se situe entre 6.7×10^{-17} Å et 3.8×10^{-16} Å, ce qui peut aussi être attribué à la précision des registres du processeur utilisé.

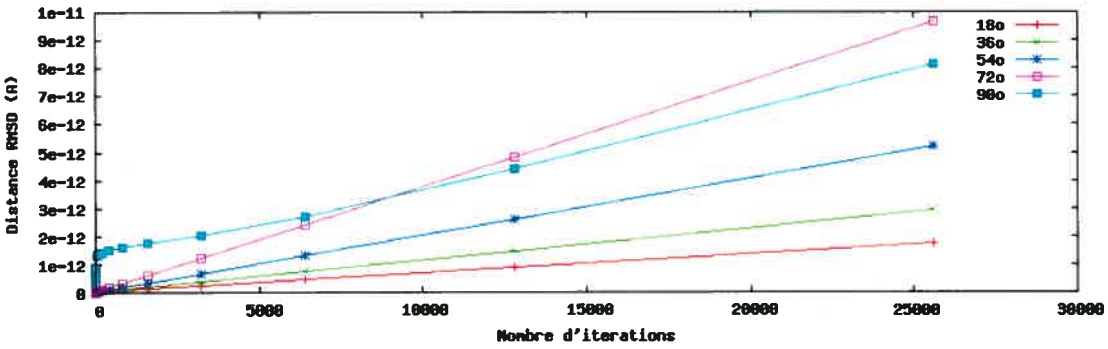

FIG. 6.4 – Effet du nombre d'itérations de la méthode FRODA sur cinq structures dont les atomes sont déjà à l'équilibre avec leur représentation fantôme

Les examens précédents permettent d'établir la précision de la méthode FRODA lorsque des modifications à la structure dans l'espace des angles dièdres sont apportées. Les tests qui suivent servent à établir le comportement de la méthode

lorsque les degrés de liberté considérés rigides sont altérés. Le plus simple des tests envisageables consiste à parcourir linéairement le trajet reliant deux conformations à l'équilibre. Pour ce faire, nous définissons la structure initiale comme étant la forme étendue examinée précédemment, et la structure finale par une conformation générée en modifiant les angles dièdres libres par une valeur moyenne de 5 degrés. Les structures initiale et finale sont distancées par un RMSD de 1.22 Å, mais elles sont toutes les deux à l'équilibre avec leur position fantôme, affichant des énergies rigides de 0.84 et 0.86 kcal/mol respectivement. Étant donné que les mouvements internes aux protéines sont principalement des rotations d'angle dièdre, le vecteur de déplacement entre la position finale et initiale $\vec{V_{f-i}}$ n'est qu'une estimation de la moyenne des déplacements infinitésimaux linéaires reliant les deux conformations. Un profil énergétique des degrés de liberté rigide des conformations identifiables par combinaison linéaire de ces deux structures démontre que le vecteur les séparant est un trajet peut probable puisqu'il culmine par un point de selle de 360.7 kcal/mol (Fig. 6.5).

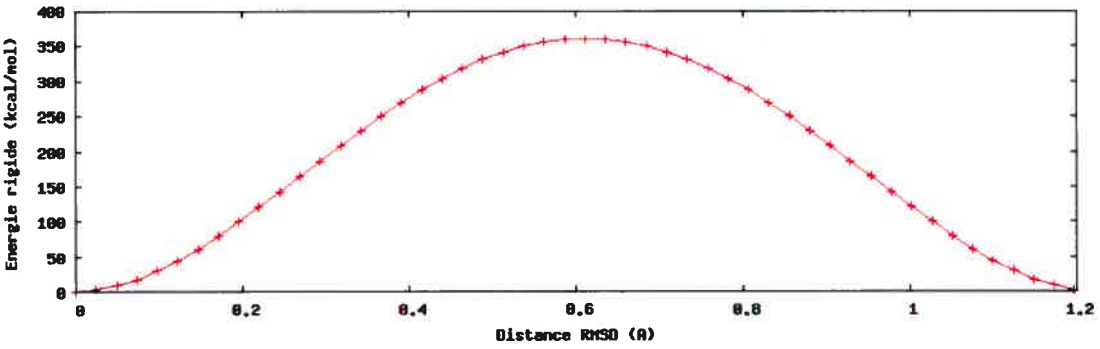

FIG. 6.5 – Énergie rigide observée lors d'un déplacement linéaire entre deux conformations aux corps fantômes à l'équilibre mais distancées par 1.22 Å

La norme du vecteur $\vec{V_{f-i}}$ est de 12.1 Å et la distance moyenne séparant chaque atome des deux conformations est de 0.93 Å. Dans le test suivant, on parcourt cette distance en cinq bonds dans la direction de $\vec{V_{f-i}}$ et de longueur $0.2 \times ||\vec{V_{f-i}}||$ corrigée par la méthode FRODA. Afin de conserver la contribution rigide à l'énergie sous un

seuil de 0.9 kcal/mol (8% au-dessus de la valeur d'équilibre), la méthode FRODA doit être itérée 30 fois. Après les cinq bonds corrigés, la structure trouvée est à 0.07 Å de distance RMSD de la structure finale, ce qui est envisageable puisque $\vec{V_{f-i}}$ pointe dans la direction générale du mouvement souhaité et qu'aucune autre contribution viennent gêner le déplacement.

Afin de mieux caractériser le nombre d'itération nécessaire pour des mouvements qui ne sont pas aussi optimisés, l'expérience est répétée avec un vecteur de direction $\vec{U}$ définit par :

$$\vec{U} = \vec{V_{f-i}} + k \times \vec{V_p} \tag{6.14}$$

où $\vec{V_p}$ est un vecteur unitaire perpendiculaire à $\vec{V_{f-i}}$, c'est-à-dire que $\vec{V_p} \cdot \vec{V_{f-i}} = 0$, et k est un entier entre 5 et 25. L'utilisation d'un vecteur de direction comprenant une contribution perpendiculaire à $\vec{V_{f-i}}$ induit une déformation volontaire des degrés de liberté rigides. L'expérience est répétée à 10 reprises en variant le choix de $\vec{V_p}$ pour générer les courbes en figure 6.6. À l'examen de ce dernier, on remarque que le pas dans la direction $\vec{U}$ sans composante perpendiculaire a $\vec{V_{f-i}}$ ($k = 0$) nécessite 4.45 fois moins d'itération pour que sa composante d'énergie rigide passe sous la barre de 1 kcal/mol comparativement au cas où $k = 25$, mais aussi 4.3 fois moins d'itération pour passer sous la barre de 0.9 kcal/mol. On en conclut que le processus n'est pas Markovien, c'est a dire que le passage d'un état S_i d'énergie E_i à l'état S_{i+1} dépend de l'énergie E_{i-1} de l'état précédent. De plus, puisque les termes des composantes rigides de l'énergie sont harmoniques, on peut conclure que le nombre d'itérations nécessaires à la méthode FRODA pour passer sous un seuil énergétique prédéterminé est proportionnel au carré de la déviation des coordonnées internes rigidifiées avant l'application itérative de la méthode.

Suite à ces observations préliminaires, nous pouvons mieux établir le compromis entre le temps d'exécution et la précision de la méthode FRODA, tous deux dépendant du nombre d'itération d'apposition des corps fantômes. Afin que la méthode FRODA-ART puisse être comparée avec l'algorithme ART nouveau ori-

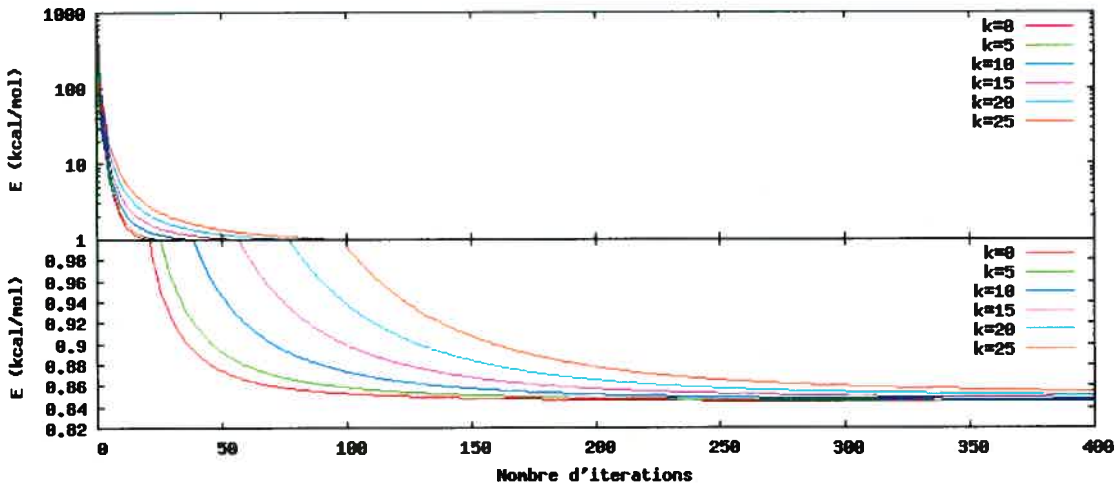

FIG. 6.6 – Énergie rigide moyenne observée lors du premier pas de taille et direction $0.2 \times \vec{U}$ en fonction de la contribution perpendiculaire $k \times \vec{V}_p$ et du nombre d'itérations de la méthode FRODA.

ginal, on choisira un jeu de paramètres communs pour effectuer des essais avec les deux algorithmes. On limite le nombre d'itérations de la méthode FRODA à 20 au maximum et à 5 au minimum. Ce nombre d'itérations est variable : afin d'économiser du temps de calcul, on interrompra la méthode FRODA si l'énergie des composantes rigides de la structure passe sous le seuil de 10% plus élevé que la valeur de l'énergie rigide à l'équilibre lorsque les corps fantômes ont été assignés. Aussi, pour assurer une plus grande stabilité, les positions des corps fantômes sont réassignées à la fin de chaque cycle de minimisation pour que leur valeur reflète la déviation du minimum local visité. Examinons maintenant en détail les différences entre les méthodes FRODA-ART et ART nouveau étape par étape.

a) Sortie du bassin harmonique :

Pour la première étape du cycle de ART, la méthode FRODA-ART affiche un net avantage sur la méthode ART nouveau. On se rappelle que la première étape consiste à déformer la structure dans une direction aléatoire pour quitter le bassin harmonique du minimum local. Pour l'épingle-β, cette parti rapide du cycle ne consomme que 5% du temps de calcul d'un cycle complet. Cette étape est aussi

l'instigatrice de la majorité de la distance parcourue lors du passage d'un minimum local au suivant. Bien que l'étape nécessite trois fois plus de temps de calcul sous FRODA-ART que sous ART nouveau, l'application de FRODA permet de parcourir une plus grande distance tout en gardant le contrôle sur le niveau énergétique de la protéine. En moyenne, on note que les structures ainsi générées ont une distance RMSD du minimum quitté 16% supérieure (RMSD moyen de 1.15 Å au lieu de 0.99 Å tout en observant une augmentation moyenne de l'énergie globale 4.25 fois moindre (163 kcal/mol pour FRODA-ART comparativement à 695 kcal/mol pour ART nouveau). Cette propriété peut être exploitée afin de générer des mouvements d'amplitude toujours grande. Contextuellement, on pourrait décider de générer des événements de transition de plus grande amplitude lorsque la configuration est prise dans un bassin de conformations séparé des autres conformations par des états intermédiaires de plus haute énergie. Les mouvements couvrant une plus longue distance RMSD auront alors une meilleure probabilité de sortir du piège. Cependant, la conséquence de ces mouvements de grande amplitude est qu'on ne peut plus être certain si on a sauté par-dessus des minima locaux sans les avoir échantillonnés. Malgré son coût de calcul supérieur, il reste avantageux d'utiliser FRODA-ART pour sortir des bassins harmoniques simplement pour profiter de la valeur énergétique moins élevée de la structure générée qui peut diminuer le temps de calcul de la partie suivante du cycle de ART.

b) Convergence vers un point de selle :

Une fois que la configuration a été poussée hors du bassin harmonique d'un minimum local et qu'on a identifié un vecteur propre de valeur propre négative pointant dans la direction du point de selle, le cycle suivant de la méthode ART consiste à se diriger vers le point de selle. Ce trajet se parcourt par l'itération de deux étapes. Tout d'abord, un pas est pris dans la direction du point de selle en suivant le vecteur propre identifié. Ensuite, une série de pas sont pris afin de minimiser l'énergie les forces perpendiculaires au vecteur propre. Le trajet ainsi parcouru se rend au point de selle avec une énergie de l'hyperplan perpendiculaire au vecteur propre qui est minimisée. Le point de selle peut être identifié en

examinant la norme des vecteurs de forces parallèles et perpendiculaires : lorsque ces normes passent sous un seuil de 0.1 et $0.5kN/mol$ respectivement, l'étape de convergence vers le point de selle est arrêtée et on passe à l'étape suivante. Outre le calcul du vecteur propre par la méthode de Lanczos [Lan88], l'étape prenant le plus de temps consiste à la minimisation des forces perpendiculaires. À l'implémentation de FRODA-ART, nous avions espoir de pouvoir éliminer l'étape de minimisation perpendiculaire. Cependant, cette étape du cycle de ART est difficilement utilisée conjointement avec FRODA. Le premier problème provient de la méthode de Lanczos. Cette dernière identifie les vecteurs propres de plus faible valeur propre de la matrice Hessienne par approximation. Les données qui lui sont nécessaires sont des vecteurs de forces provenant de conformations obtenues par une série de mouvements perpendiculaires les uns aux autres. Pour implémenter FRODA dans cette partie de l'algorithme de Lanczos, il faudrait que ces vecteurs de mouvements soient perpendiculaires entre-eux dans le sous-espace des coordonnées internes. Cela ne suffit pas de corriger à l'aide de FRODA les positions atomiques obtenues après un de ces mouvements perpendiculaires dans l'espace cartésien. L'algorithme de Lanczos doit donc être entièrement calculé en coordonnées cartésiennes et les vecteurs propres ainsi générés seront eux aussi dans cet espace. En soit, ceci n'est qu'un inconvénient mineur puisque les vecteurs propres de valeurs propres négatives ne peuvent contenir énormément de contributions déformant des degrés de liberté rigides tels que les longueurs de liens et angles de valence : les transitions d'états étant dominées par des variations des angles dièdres libres, les autres degrés de liberté ne peuvent être altérés que faiblement.

Ceci nous mène au deuxième problème de la méthode FRODA dans l'étape de convergence vers le point de selle. La méthode FRODA est très efficace pour minimiser l'énergie des degrés de liberté rigides des protéines, mais il en va autrement lorsque l'énergie qui doit être minimisée provient d'interactions à longue portée tel que les termes de Van der Waals et électrostatiques. On se rappelle que le point de selle est identifié à l'aide de la norme des vecteurs de force perpendiculaires et parallèles au vecteur de Lanczos. Or, sous la méthode FRODA-ART, il est quasiment

impossible d'atteindre les seuils utilisés sous ART nouveau pour deux raisons. Tout d'abord, la rigidité de la structure crée une certaine frustration lorsqu'on tente de minimiser finement les forces près du point de selle. Ensuite, la position des atomes dans les corps rigides a été fixé au dernier minimum local dans une configuration localement à l'équilibre. Suite au mouvement généré par ART, le point de selle auquel l'algorithme tente de converger est en moyenne distant de ~ 1 Å du minimum local où les corps fantômes ont été rééquilibrés. On choisit donc des valeurs seuils plus permissives pour permettre d'identifier le point de selle, soit une norme 5 fois plus grande pour la norme du vecteur parallèle et 6 fois plus grande pour la norme du vecteur perpendiculaire. Malgré ces valeurs permissives qui ne garantissent pas qu'on ait identifié un vrai point selle, l'algorithme FRODA-ART prend un temps comparable à la méthode ART nouveau pour converger aux points de selle, affichant faible gain de rapidité de l'ordre de 10% dans les cas où l'on trouve un point de selle. Cependant, on dénombre 20% des essais qui ne trouvent pas le point de selle dans le nombre d'itérations requis, ce qui a pour effet d'augmenter le coût moyen en temps de calcul à 3.4 fois plus élevé que le coût moyen de la méthode ART nouveau. Une optimisation dynamique du nombre d'itérations permises pour les cycles qui ne convergent pas vers le point de selle pourrait permettre de diminuer ce coût, mais pas suffisamment pour que la méthode devienne moins coûteuse que ART nouveau.

Notons cependant que le but visé de diminuer le nombre d'appels au potentiel énergétique a été atteint par la méthode FRODA-ART. Pour les événements qui réussissent à converger dans le nombre d'itérations allouées, on note une diminution de ce nombre d'appels par un facteur de 13 fois. Les appels à la méthode FRODA par contre rétablissent la consommation en temps de calcul à un niveau similaire à la méthode ART nouveau. Puisque nous utilisons un nombre d'itérations fixe pour la méthode FRODA, son ordre de temps de calcul est de $O(N)$, ce qui signifie que des gains en efficacité sont envisageable lors de simulations sur des protéines de plus grande taille.

c) Saut du point de selle et convergence vers le nouveau minimum local :

L'étape de minimisation est aussi affectée par la difficulté de la méthode FRODA à converger vers une structure où les forces s'annulent. Tout d'abord, la méthode arrive difficilement a converger, même en lui accordant un temps illimité, et se stabilise habituellement à une énergie d'environ 5 kcal/mol plus élevée que le minimum local à proximité. Tout comme le vecteur propre de l'étape précédente, la méthode présente une incapacité de se déplacer parallèlement au vecteur de force lorsque la configuration devient plus compacte. Il s'en suit que l'algorithme de dynamique moléculaire amortie ne peut construire et maintenir un vecteur de vitesse suffisamment longtemps avant qu'apparaisse un vecteur de force dont la direction est contraire au vecteur de vitesse. Cette frustration n'est pourtant pas attribuée aux positions des atomes enregistrés dans les corps fantôme : lors d'un test où on connaît déjà la position des atomes dans le minimum local à proximité, une structure est minimisé en utilisant FRODA-ART et la position des corps fantômes du minimum local précédent. La même structure est ensuite reminimisée en utilisant cette fois-ci la position des corps fantômes du minimum local déterminée par minimisation en coordonnées cartésiennes. Bien que le biais dans le second cas est favorable à la découverte du minimum local, la structure minimisée par FRODA-ART avec les corps fantômes idéaux converge elle aussi vers un plateau énergétique d'à peu près le même niveau.

La méthode de minimisation par dynamique moléculaire amortie de FRODA-ART peut par contre être utilisée en collaboration avec la méthode cartésienne dans un rapport de temps de calcul de 10%-90% : on débute le processus de minimisation par quelques pas avec FRODA qui ont pour effet de minimiser rapidement les forces des composantes rigides tout en engendrant des rotations autour des angles dièdres ayant le moins de contraintes stériques de leur environnement. Puis, on poursuit la minimisation en coordonnées cartésiennes pour optimiser les contacts et les mouvements fins qui mènent au minimum local. Malheureusement, cette méthode n'affiche qu'un gain d'au plus 10% sur le temps de calcul en coordonnées

cartésiennes.

6.5 Discussion

Des trois méthodes présentées dans ce chapitre, la méthode d'apposition des corps fantôme FRODA [WMHT05] nous semblait la mieux adaptée pour être implémentée dans ART nouveau. La méthode de Bystroff [Bys01], bien qu'étant une solution exacte au problème de la distribution des forces inter-atomiques sur une chaîne, exige un ordre d'exécution de $O(N^3)$ qui la rend trop lente dans le contexte d'ART-OPEP où N est le nombre d'atomes. De son côté, SHAKE [RCB77] a un aspect similaire à FRODA dans le sens que les contraintes sont imposées par un processus itératif d'optimisation. Cependant, la définition du corps fantôme d'un seul plan peptidique de 6 atomes à l'aide de contraintes SHAKE nécessiterait la définition de 5 contraintes de longueur de liens, 4 contraintes d'angles de valence et 3 contraintes d'angles dièdres. L'algorithme de FRODA qui rigidifie des blocs entiers d'atomes présente une méthode optimale du traitement des changements de conformation dans l'espace des angles dièdres libres. De plus, son utilisation implique un temps de calcul d'ordre $O(N)$ si on fixe le nombre d'itérations maximales. Dans le cas où la méthode est itérée jusqu'à ce que la composante d'énergie rigide tombe sous un seuil prédéterminé, l'ordre d'exécution déterminé expérimentalement est de $O(N \times D^2)$ où D est est la longueur du mouvement influençant une coordonnée rigide.

Notre examen de la méthode FRODA sur une petite protéine de 16 acides aminés démontre que cette dernière ne peut être utilisée efficacement dans toutes les étapes de la méthode ART nouveau. Son utilisation apporte un avantage net seulement dans l'étape de sortie du bassin harmonique alors qu'elle permet des mouvements de plus grande amplitude tout en gardant le contrôle sur la valeur de l'énergie de la structure. Dans le contexte des autres étapes, la nécessité de converger vers un point spécifique de la surface énergétique dépasse les capacités de l'algorithme.

Bien que les conclusions de notre analyse ne soient pas favorables à l'utilisation de la méthode FRODA dans le contexte de la méthode accélérée ART nouveau, l'algorithme pourra possiblement trouver une utilisation dans d'autres méthodes non-activées. Par exemple, FRODA pourrait remplacer jusqu'à un certain degré l'algorithme SHAKE dans les simulations de dynamique moléculaire. Sur de gros systèmes où l'on désire diminuer les degrés de liberté des atomes qui ne sont pas à proximité d'un site actif ou d'une région flexible d'intérêt, FRODA pourrait définir des corps fantômes pour des domaines entiers. Nous proposons de pousser l'expérimentation de la méthode dans les projets futurs de notre laboratoire traitant de l'étude de la flexibilité des protéines dans leur configuration native.

CONCLUSION

Dans ce travail, nous avons examiné l'application de la méthode accélérée de navigation de la surface énergétique ART nouveau au repliement d'une protéine de 60 acides aminés. Bien que nous ayons trouvé plusieurs trajectoires menant à la structure native, nous n'avons pu déterminer un ordre de formation des hélices qui soit fortement favorisé sur les autres. Nos données reflètent cependant les probabilités expérimentales de formation des hélices isolées qui assignent une plus grande stabilité à l'hélice H3 et à la paire d'hélices H2-H3.

De plus, nos résultats démontrent que cette protéine peut adopter des configurations stables, mais de topologies différentes de celle de la structure native. Ces données sont en accord avec une théorie récente du repliement qui assigne une plus grande importance à la stabilisation de la chaîne carbonée des protéines [RFBM06]. Dans cette théorie, l'état dénaturé est caractérisé par un échantillonnage d'un nombre discret de topologies non-natives en équilibre avec l'état natif dont la topologie et les contacts tertiaires sont favorisés énergétiquement. Nos données suggèrent que ces topologies peuvent exister et qu'il semble y avoir un degré de frustration dans la protéine A qui n'est pas permis par les théories traditionnelles de l'entonnoir. Il serait donc intéressant de valider la stabilité de ces conformations à l'aide de potentiel tout-atomes et de méthodes pouvant évaluer leur énergie libre dans un système à solvant explicite tel que la dynamique moléculaire.

Nous croyons aussi que cette propriété d'échantillonner des topologies différentes devrait être examinée plus en profondeur, car elle constitue une nouvelle voie d'inhibition : en stabilisant ces protéines dans une conformation non-native, il serait possible d'inhiber la fonction de celles-ci. Cette avenue serait avantageuse pour l'inhibition de protéines dont les inhibiteurs traditionnels présentent un niveau de toxicité *in vivo* les rendant impropres à la consommation. Outre l'aspect pharmacologique, l'inhibition par stabilisation des topologies non-natives nous semble être la meilleur façon de prouver expérimentalement leur existence. Dans cette optique, l'étude *in silico* de ces topologies à fin de déterminer si elles possèdent des

sites qu'un ligand pourrait lier serait alors la première étape menant à l'utilisation d'outil de docking de ligand.

Bien que la méthode ART nouveau utilisée soit une méthode accélérée qui permet d'obtenir des résultats relativement rapidement, nous étions aussi intéressés à optimiser la méthode pour la rendre encore plus efficace. Pour ce faire, nous avons implémenté deux algorithmes de rigidification des degrés de liberté moins intéressants que sont les angles de valence et les longueurs de liens. En théorie, ces méthodes auraient pu nous permettre de générer des mouvements efficaces en ce qui concerne les degrés de liberté réels des protéines, leurs angles dièdres. La première méthode implémentée, la méthode analytique de Bystroff [Bys01] s'est révélée inutilisable avec la méthode ART puisque pour être exacte cette dernière nécessite l'utilisation de termes de force interatomiques spécifiques au lieu des vecteurs de déplacement de ART. De plus, son temps de calcul étant de l'ordre $O(N^3)$ où N est le nombre d'atomes, son utilisation pour diminuer les appels au potentiel énergétique d'ordre $O(N^2)$ rendent la méthode ART moins efficace. Nous avons alors étudié un second algorithme de rigidification qui impose des corrections aux déviations des degrés de liberté rigides par une méthode itérative.

La méthode FRODA [WMHT05] des corps fantômes est nettement plus rapide que celle de Bystroff, affichant un ordre de $O(N)$ lorsqu'elle est utilisée avec un nombre limité d'itérations. Cependant, lorsqu'on utilise une valeur seuil de l'énergie interne aux corps fantôme pour déterminer le nombre d'itérations nécessaire, la méthode est d'ordre $O(N \times D^2)$ où D est la distance de la déviation à corriger. Son utilisation efficace ne permet donc pas d'augmenter outre mesure la taille des pas déformant les coordonnées rigides. Lors d'examens de son utilisation dans ART nouveau, nous avons déterminé que la méthode FRODA ne pouvait apporter des gains en efficacité sur la méthode originale. Nous observons que les mouvements dirigés de convergence vers un point de selle suivi par ceux de convergence vers un minimum local sont plus efficaces dans l'espace des coordonnées cartésiennes que dans celui des angles dièdres. La correction des positions des atomes des corps fantômes est incompatible avec les mouvements fins nécessaires à la minimisation de l'énergie

qui font converger la position des atomes vers des conformations toujours plus compactes. Cependant, les mouvements de sortie des bassins harmoniques, c'est-à-dire les mouvements d'expansion d'une structure compacte, sont mieux contrôlés par la méthode FRODA qui garde à l'équilibre les coordonnées internes rigidifiées. Ceci nous porte à croire que FRODA puisse être un substitut intéressant à l'algorithme d'apposition de contraintes SHAKE [RCB77] utilisé dans les simulations de dynamique moléculaire. Aussi, FRODA nous semble être une méthode appropriée pour rigidifier des éléments de structure complets dans des simulations de flexibilité de site actif où une petite région de la protéine est examinée. Bien qu'elle ne permette pas de gains d'efficacité auprès de la méthode ART, des implémentations alternatives où la définition des corps rigides serait modifiée restent envisageables. La méthode se rajoute donc à nos outils de simulations jusqu'à ce qu'elle trouve une utilisation appropriée.

BIBLIOGRAPHIE

[AD00] D. O. V. Alonso and V. Daggett. Staphylococcal protein a : Unfolding pathways, unfolded states, and differences between the b and e domains. *Proc Natl Acad Sci USA*, 97(1) :133–138, 2000.

[And83] H. C. Andersen. Rattle : A "velocity" version of the shake algorithm for molecular dynamics calculations. *J Comput Phys*, 52(1) :24–34, 1983.

[Anf73] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181 :223–230, 1973.

[AROB05] P. A. Alexander, D. A. Rozak, J. Orban, and P. N. Bryan. Directed evolution of highly homologous proteins with different folds by phage display : Implications for the protein folding code. *Biochemistry-Us*, 44(43) :14045–14054, 2005.

[BB95] E. M. Boczko and III Brooks, C. L. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science*, 269(5222) :393–396, 1995.

[BDWB77] T. Bachi, G. Dorval, H. Wigzell, and H. Binz. Staphylococcal protein a in immunoferritin techniques. *Scand J Immunol*, 6(3) :241–6, 1977.

[BK83] B. Brooks and M. Karplus. Harmonic dynamics of proteins - normal-modes and fluctuations in bovine pancreatic trypsin-inhibitor. *Proc Natl Acad Sci-Biol*, 80(21) :6571–6575, 1983.

[BKDW97] Y. W. Bai, A. Karimi, H. J. Dyson, and P. E. Wright. Absence of a stable intermediate on the folding pathway of protein a. *Protein Sci*, 6(7) :1449–1457, 1997.

[BKW+77] F. C. Bernstein, T. F. Koetzle, G. J. Williams, Jr. Meyer, E. F., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank : a computer-based archival file for macromolecular structures. *J Mol Biol*, 112(3) :535–42, 1977.

[BM96] G. T. Barkema and N. Mousseau. Event-based relaxation of continuous disordered systems. *Phys Rev Lett*, 77(21) :4358–4361, 1996.

[Boc94] Brooks C. L. III Boczko, E. M. Constant-temperature free energy surfaces for physical and chemical processes. *J Phys Chem*, 97 :4509–4513, 1994.

[BOSW95] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding : a synthesis. *Proteins*, 21(3) :167–195, 1995.

[BOW01] C. L. Brooks, J. N. Onuchic, and D. J. Wales. Statistical thermodynamics : Taking a walk on a landscape. *Science*, 293(5530) :612–613, 2001.

[BPS+94] S. P. Bottomley, A. G. Popplewell, M. Scawen, T. Wan, B. J. Sutton, and M. G. Gore. The stability and unfolding of an igg binding protein based upon the b domain of protein a from staphylococcus aureus probed by tryptophan substitution and fluorescence spectroscopy. *Protein Eng*, 7(12) :1463–1470, 1994.

[BS01] G. F. Berriz and E. I. Shakhnovich. Characterization of the folding kinetics of a three-helix bundle protein via a minimalist langevin model. *J Mol Biol*, 310(3) :673–685, 2001.

[Bys01] C. Bystroff. An alternative derivation of the equations of motion in torsion space for a branched linear chain. *Protein Eng*, 14(11) :825–828, 2001.

[Caf06] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Curr Opin Struct Biol*, 16(1) :71–78, 2006.

[CYWL05] S. M. Cheng, Y. D. Yang, W. R. Wang, and H. Y. Liu. Transition state ensemble for the folding of b domain of protein a : A comparison of distributed molecular dynamics simulations with experiments. *J Phys Chem B*, 109(49) :23645–23654, 2005.

[DBY+95] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principles of protein folding–a perspective from simple exact models. *Protein Sci*, 4(4) :561–602, 1995.

[DCG99] D. A. Debe, M. J. Carlson, and III Goddard, W. A. The topomer-sampling model of protein folding. *Proc Natl Acad Sci USA*, 96(6) :2596–2601, 1999.

[DDM+04] G. Dimitriadis, A. Drysdale, J. K. Myers, P. Arora, S. E. Radford, T. G. Oas, and D. A. Smith. Microsecond folding dynamics of the f13w g29a mutant of the b domain of staphylococcal protein a by laser-induced temperature jump. *Proc Natl Acad Sci USA*, 101(11) :3809–3814, 2004.

[Dei81] J. Deisenhofer. Crystallographic refinement and atomic models of a human fc fragment and its complex with fragment b of protein a from staphylococcus aureus at 2.9- and 2.8-a resolution. *Biochemistry*, 20(9) :2361–70, 1981.

[DeL02] W. L. DeLano. The pymol molecular graphics system, 2002.

[Der97] P. Derreumaux. Folding a 20 amino acid alpha beta peptide with the diffusion process-controlled monte carlo method. *J Chem Phys*, 107(6) :1941–1947, 1997.

[Der98] P. Derreumaux. Finding the low-energy forms of avian pancreatic polypeptide with the diffusion-process-controlled monte carlo method. *J Chem Phys*, 109(4) :1567–1574, 1998.

[Der99] P. Derreumaux. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *J Chem Phys*, 111(5) :2301–2310, 1999.

[Der00] P. Derreumaux. Generating ensemble averages for small proteins from extended conformations by monte carlo simulations. *Phys Rev Lett*, 85(1) :206–209, 2000.

[Der02] P. Derreumaux. Insight into protein topology from monte carlo simulations. *J Chem Phys*, 117(7) :3499–3503, 2002.

[DKWQ69] J. H. Dossett, G. Kronvall, Jr. Williams, R. C., and P. G. Quie. Antiphagocytic effects of staphylococcal protein a. *J Immunol*, 103(6) :1405–10, 1969.

[FD01] F. Forcellino and P. Derreumaux. Computer simulations aimed at structure prediction of supersecondary motifs in proteins. *Proteins*, 45(2) :159–166, 2001.

[Fer97] A. R. Fersht. Nucleation mechanisms in protein folding. *Curr Opin Struct Biol*, 7(1) :3–9, 1997.

[FIM04] Giorgio Favrin, Anders Irback, and Sandipan Mohanty. Oligomerization of amyloid abeta16-22 peptides using hydrogen bonds and hydrophobicity forces. *Biophys. J.*, 87(6) :3657–3664, 2004.

[FIW02] G. Favrin, A. Irback, and S. Wallin. Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins*, 47(2) :99–105, 2002.

[GBB97] Z. Y. Guo, C. L. Brooks, and E. M. Boczko. Exploring the folding free energy surface of a three-helix bundle protein. *Proc Natl Acad Sci USA*, 94(19) :10161–10166, 1997.

[GES02] A. Ghosh, R. Elber, and H. A. Scheraga. An atomically detailed study of the folding pathways of protein a with the stochastic difference equation. *Proc Natl Acad Sci USA*, 99(16) :10394–10398, 2002.

[GGK+03] S. Gianni, N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. White, M. L. DeMarco, V. Daggett, and A. R. Fersht. Unifying features in protein-folding mechanisms. *Proc Natl Acad Sci USA*, 100(23) :13286–13291, 2003.

[GO03] A. E. Garcia and J. N. Onuchic. Folding a protein in a computer : An atomic description of the folding/unfolding of protein a. *Proc Natl Acad Sci USA*, 100(24) :13898–13903, 2003.

[God78] J. W. Goding. Use of staphylococcal protein a as an immunological reagent. *J Immunol Methods*, 20 :241–53, 1978.

[GSS75] V. Ghetie, G. Stalenheim, and J. Sjoquist. Cell separation by staphylococcal protein a-coated erythrocytes. *Scand J Immunol*, 4(5-6) :471–7, 1975.

[GTS+92] H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, and I. Shimada. Three-dimensional solution structure of the b domain of staphylococcal protein a : comparisons of the solution and crystal structures. *Biochemistry*, 31(40) :9665–9672, 1992.

[IC06] Hideo Imamura and Jeff Z. Y. Chen. Dependence of folding dynamics and structural stability on the location of a hydrophobic pair in ?-hairpins. *Proteins*, 63(3) :555–570, 2006.

[IKW02] S. A. Islam, M. Karplus, and D. L. Weaver. Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J Mol Biol*, 318(1) :199–215, 2002.

[IS06] K. Itoh and M. Sasai. Flexibly varying folding mechanism of a nearly symmetrical protein : B domain of protein a. *Proc Natl Acad Sci USA*, 103(19) :7298–7303, 2006.

[JGHS99] M. Jacob, M. Geeves, G. Holtermann, and F. X. Schmid. Diffusional barrier crossing in a two-state protein folding reaction. *Nat Struct Biol*, 6(10) :923–6, 1999.

[JKSP03] S. M. Jang, E. Kim, S. Shin, and Y. Pak. Ab initio folding of helix bundle proteins using molecular dynamics simulations. *J Am Chem Soc*, 125(48) :14841–14846, 2003.

[JRKT01] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44(2) :150–165, 2001.

[JVGP06] G. Jayachandran, V. Vishal, A. E. Garci'a, and V. S. Pande. Local structure formation in simulations of two small proteins. *J Struct Biol*, In Press, Corrected Proof, 2006.

[KB90] P. S. Kim and R. L. Baldwin. Intermediates in the folding reactions of small proteins. *Annu Rev Biochem*, 59 :631–60, 1990.

[KLR+05] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga. Molecular dynamics with the united-residue model of polypeptide chains. i. lagrange equations of motion and tests of numerical stability in the microcanonical mode. *J Phys Chem B*, 109(28) :13785–13797, 2005.

[KLS06] M. Khalili, A. Liwo, and H. A. Scheraga. Kinetic studies of folding of the b-domain of staphylococcal protein a with molecular dynamics and a united-residue (unres) model of polypeptide chains. *J Mol Biol*, 355(3) :536–547, 2006.

[KS83] W. Kabsch and C. Sander. Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12) :2577–2637, 1983.

[KS94] A. Kolinski and J. Skolnick. Monte carlo simulations of protein folding. i. lattice model and interaction scheme. *Proteins*, 18(4) :338–352, 1994.

[KSS02] E. Kussell, J. Shimada, and E. I. Shakhnovich. A structure-based method for derivation of all-atom potentials for protein folding. *Proc Natl Acad Sci USA*, 99(8) :5343–5348, 2002.

[KW94] M. Karplus and D. L. Weaver. Protein folding dynamics : the diffusion-collision model and experimental data. *Protein Sci*, 3(4) :650–68, 1994.

[Lan88] Lanczos. In *Applied Analysis*. dover, New York, 1988.

[Lev69] C. Levinthal. How to fold graciously. In *Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House, Monticello*, page 22. University of Illinois Press, Urbana, Illinois, 1969.

[LKS05] A. Liwo, M. Khalili, and H. A. Scheraga. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci USA*, 102(7) :2362–2367, 2005.

[LLM70] I. Lind, I. Live, and B. Mansa. Variation in staphylococcal protein a reactivity with gamma g-globulins of different species. *Acta Pathol Microbiol Scand [B] Microbiol Immunol*, 78(6) :673–82, 1970.

[LLS99] J. Lee, A. Liwo, and H. A. Scheraga. Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field : Application to the 10-55 fragment of staphylococcal protein a and to apo calbindin d9k. *Proc Natl Acad Sci USA*, 96(5) :2025–2030, 1999.

[LMO92] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels : A kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci USA*, 89(18) :8721–8725, 1992.

[LSSP02] S. M. Larson, C. D. Snow, M. Shirts, and V. S. Pande. Folding@ home and genome@ home : Using distributed computing to tackle previously intractable problems in computational biology. In Richard Grant, editor, *Computational Genomics*. Horizon Press, 2002.

[LZ02] A. Linhananta and Y. Q. Zhou. The role of sidechain packing and native contact interactions in folding : Discontinuous molecular dynamics folding simulations of an all-atom g(o)over-bar model of fragment b of staphylococcal protein a. *J Chem Phys*, 117(19) :8983–8995, 2002.

[LZZ02] A. Linhananta, H. Y. Zhou, and Y. Q. Zhou. The dual role of a loop with low loop contact distance in folding and domain swapping. *Protein Sci*, 11(7) :1695–1701, 2002.

[MAN+86] T. Moks, L. Abrahmsen, B. Nilsson, U. Hellman, J. Sjoquist, and M. Uhlen. Staphylococcal protein a consists of five igg-binding domains. *Eur J Biochem*, 156(3) :637–43, 1986.

[MB98] N. Mousseau and G. T. Barkema. Traveling through potential energy landscapes of disordered materials : The activation-relaxation technique. *Phys Rev E*, 57(2) :2419–2424, 1998.

[MD05] N. Mousseau and P. Derreumaux. Exploring the early steps of amyloid peptide aggregation by computers. *Acc Chem Res*, 38(11) :885–91, 2005.

[MDBM01] N. Mousseau, P. Derreumaux, G. T. Barkema, and R. Malek. Sampling activated mechanisms in proteins with the activation-relaxation technique. *J Mol Graph Model*, 19(1) :78–86, 2001.

[MDG05] N. Mousseau, P. Derreumaux, and G. Gilbert. Navigation and analysis of the energy landscape of small proteins using the activation-relaxation technique. *Phys Biol*, 2(4) :S101–7, 2005.

[MG58] S. Meiboom and D. Gill. Modified spin-echo method for measuring nuclear relaxation times. 29(8) :688–691, 1958.

[MK92] S. Miyamoto and P. A. Kollman. Settle : An analytical version of the shake and rattle algorithm for rigid water models. *J Comput Chem*, 13(8) :952–962, 1992.

[MKSF89] A. Matouschek, Jr. Kellis, J. T., L. Serrano, and A. R. Fersht. Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, 340(6229) :122–6, 1989.

[MM00] 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.

[MMD06] A. Melquiond, N. Mousseau, and P. Derreumaux. Structures of soluble amyloid oligomers from computer simulations. *Proteins*, 65(1) :180–91, 2006.

[MN06] Buyong Ma and Ruth Nussinov. Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. *Curr Opin Chem Biol*, 10(5) :445–452, 2006.

[MO01] J. K. Myers and T. G. Oas. Preorganized secondary structure as an important determinant of fast protein folding. *Nat Struct Biol*, 8(6) :552–558, 2001.

[MRBW76] H. Mallinson, C. Roberts, and G. B. Bruce White. Staphylococcal protein a ; its preparation and an application to rubella serology. *J Clin Pathol*, 29(11) :999–1002, 1976.

[OW04] J. N. Onuchic and P. G. Wolynes. Theory of protein folding. *Curr Opin Struc Biol*, 14(1) :70–75, 2004.

[PB98] K. W. Plaxco and D. Baker. Limited internal friction in the rate-limiting step of a two-state protein folding reaction. *Proc Natl Acad Sci USA*, 95(23) :13591–6, 1998.

[PC03] J.W. Ponder and D.A Case. Force fields for protein simulatioin. *Adv. Prot. Chem.*, 66 :27–85, 2003.

[PD05] S. A. Petty and S. M. Decatur. Intersheet rearrangement of polypeptides during nucleation of beta-sheet aggregates. *Proc Natl Acad Sci USA*, 102(40) :14272–14277, 2005.

[PFP+86] W. H. Press, B. P. Flannery, Satwh Press, B. P. Flannery, S. A. Teukolsky, and Wtvwt Vetterling. Numerical recipes, 1986.

[Pti87] O. B. Ptitsyn. Protein folding : Hypotheses and experiments. *J of Prot Chem*, V6(4) :273–293, 1987.

[RCB77] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes. *J Comput Phys*, 23(3) :327–341, 1977.

[RFBM06] George D. Rose, Patrick J. Fleming, Jayanth R. Banavar, and Amos Maritan. A backbone-based theory of protein folding. *Proc Natl Acad Sci USA*, 103(45) :16623–16633, 2006.

[Sch02] Tamar Schlick. *Molecular Modeling and Simulation, An Interdisciplinary Guide.* Springer, New-York, 2002.

[Sco99] Hunenberger P.H. Tironi I.G. Mark A.E. Billeter S.R. Fennen J. Torda A. E. Hubert T. Kruger P. Van Gunsteren W. F. Scott, W.R.P. The gromos biomolecular simulation program package. *J Phys Chem*, 103 :3596–3607, 1999.

[SFM06] M. Sadqi, D. Fushman, and V. Munoz. Atom-by-atom analysis of global downhill protein folding. *Nature*, 442(7100) :317–321, 2006.

[Sko05] J. Skolnick. Putting the pathway back into protein folding. *Proc Natl Acad Sci USA*, 102(7) :2265–2266, 2005.

[SMD04] 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) :11509–11516, 2004.

[SO99] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*, 314(1-2) :141–151, 1999.

[SOB99] J-E. Shea, J. N. Onuchic, and III Brooks, C. L. Exploring the origins of topological frustration : Design of a minimally frustrated model of fragment b of protein a. *Proc Natl Acad Sci USA*, 96(22) :12512–12517, 1999.

[SP01] M. S. Shirts and V. S. Pande. Mathematical foundations of ensemble dynamics. *Phys Rev Lett*, 86 :4983–4987, 2001.

[SRDF04] S. Sato, T. L. Religa, V. Daggett, and A. R. Fersht. Testing protein-folding simulations by experiment : B domain of protein a. *Proc Natl Acad Sci USA*, 101(18) :6952–6956, 2004.

[SRF06] S. Sato, T. L. Religa, and A. R. Fersht. Phi-analysis of the folding of the b domain of protein a using multiple optical probes. *J Mol Biol*, 360(4) :850–864, 2006.

[Ste04] L. D. Stein. Human genome : End of the beginning. *Nature*, 431(7011) :915–916, 2004.

[VMOD04] D. M. Vu, J. K. Myers, T. G. Oas, and R. B. Dyer. Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein. *Biochemistry-Us*, 43(12) :3582–3589, 2004.

[VRS03] J. A. Vila, D. R. Ripoll, and H. A. Scheraga. Atomically detailed folding simulation of the b domain of staphylococcal protein a from random structures. *Proc Natl Acad Sci USA*, 100(25) :14812–14816, 2003.

[Wal03] D. J. Wales. *Energy Landscapes*. Cambridge University Press, Cambridge, 2003.

[WDC80] E. B. Wilson, J. C. Decius, and P. C. Cross. *Molecular Vibrations : the Theory of Infrared and Raman Vibrational Spectra*. Dover Publications, New York, 1980.

[WDM03] G. H. Wei, P. Derreumaux, and N. Mousseau. Sampling the complex energy landscape of a simple beta-hairpin, 2003.

[Wet73] D. B. Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA*, 70(3) :697–701, 1973.

[Wet90] D. B. Wetlaufer. Nucleation in protein folding–confusion of structure and process. *Trends Biochem Sci*, 15(11) :414–415, 1990.

[WMD02] 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(24) :11379–11387, 2002.

[WMD04a] G. Wei, N. Mousseau, and P. Derreumaux. Complex folding pathways in a simple beta-hairpin. *Proteins*, 56(3) :464–474, 2004.

[WMD04b] G. Wei, N. Mousseau, and P. Derreumaux. Exploring the early steps of aggregation of amyloid-forming peptide kffe. *J Phys-Condens Matt*, 16(44) :S5047, 2004.

[WMD04c] G. Wei, N. Mousseau, and P. Derreumaux. Sampling the self-assembly pathways of kffe hexamers. *Biophys J*, 87(6) :3648–3656, 2004.

[WMD07] G.H. Wei, N. Mousseau, and P. Derreumaux. *in preparation*, 2007.

[WMHT05] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe. Constrained geometric simulation of diffusive motion in proteins. *Phys Biol*, 2(4) :S127–S136, 2005.

[Wol04] P. G. Wolynes. Latest folding game results : Protein a barely frustrates. *Proc Natl Acad Sci USA*, 101(18) :6837–6838, 2004.

[YG04] Y. Ye and A. Godzik. Fatcat : a web server for flexible structure comparison and structure similarity searching. *Nucleic acids research*, 32(Web Server issue) :W582–585, 2004.

[YLM+06] M. R. Yun, R. Lavery, N. Mousseau, K. Zakrzewska, and P. Derreumaux. Artist : an activated method in internal coordinate space for sampling protein energy landscapes. *Proteins*, 63(4) :967–75, 2006.

[ZK99] Y. Q. Zhou and M. Karplus. Interpreting the folding kinetics of helical proteins. *Nature*, 401(6751) :400–403, 1999.

[ZL02] Y. Q. Zhou and A. Linhananta. Thermodynamics of an all-atom off-lattice model of the fragment b of staphylococcal protein a : Implication for the origin of the cooperativity of protein folding. *J Phys Chem B*, 106(6) :1481–1485, 2002.

Annexe I

Matériel supplémentaire en référence dans l'article

I.1 Tableaux supplémentaires

TAB. I.1 – Transition probability between secondary structures averaged over the 22 simulation with a majority of α content. Line and columns represent the starting and ending states respectively. The POP line and column indicate the population of the associated state. Columns are normalized by the sum of incoming transitions to this state.

CONF :	TO :	UUU	UUH	UHU	UHH	HUU	HUH	HHU	HHH
FROM :	POP :	116483	48095	14451	19805	19571	16980	4376	18409
UUU	116483		**0.44**	**0.59**	**0.12**	**0.56**			
UUH	48095	**0.38**			**0.41**		**0.42**		0.01
UHU	14451	**0.23**			**0.16**			**0.12**	0.01
UHH	19805	**0.09**	**0.34**	**0.30**			0.01	0.02	**0.40**
HUU	19571	**0.29**					**0.18**	**0.40**	**0.06**
HUH	16980		**0.21**			**0.15**			**0.29**
HHU	4376			**0.08**	0.01	**0.21**			**0.22**
HHH	18409			0.02	**0.30**	**0.08**	**0.38**	**0.45**	

TAB. I.2 – Transition probability between secondary structures averaged over the 4 simulation that find the native state. Line and columns represent the starting and ending states respectively. The POP line and column indicate the population of the associated state. Columns are normalized by the sum of incoming transitions to this state.

CONF :	TO :	UUU	UUH	UHU	UHH	HUU	HUH	HHU	HHH
FROM :	POP :	19532	1612	3156	2872	5655	5791	475	3842
UUU	19532		**0.26**	**0.76**	0.06	**0.70**			
UUH	1612	**0.07**			0.06		**0.41**		
UHU	3156	**0.41**			**0.22**	0.01		0.04	0.01
UHH	2872	0.01	**0.13**	**0.20**				0.04	**0.55**
HUU	5655	**0.49**		0.01			**0.30**	**0.38**	0.05
HUH	5791		**0.62**			**0.17**			**0.24**
HHU	475			0.03	0.04	**0.09**			**0.15**
HHH	3842				**0.62**	0.03	**0.30**	**0.54**	

Annexe II

Matériel supplémentaire en référence dans l'article

II.1 Figures supplémentaires

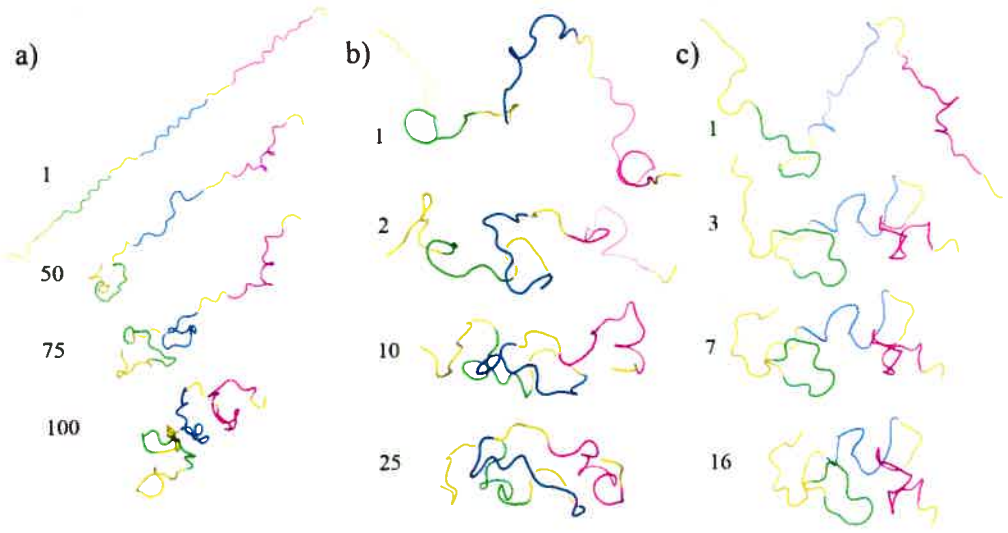

FIG. II.1 – Rapid collapse of the extended structures prior to the apparition of any secondary structure element. a) The completely extended structure used in the first 12 simulations shows generally a right-hand coiled random coil in the first 100 accepted events. Forty simulations were also initiated from a left-handed random coil b) and a right-handed coil c)

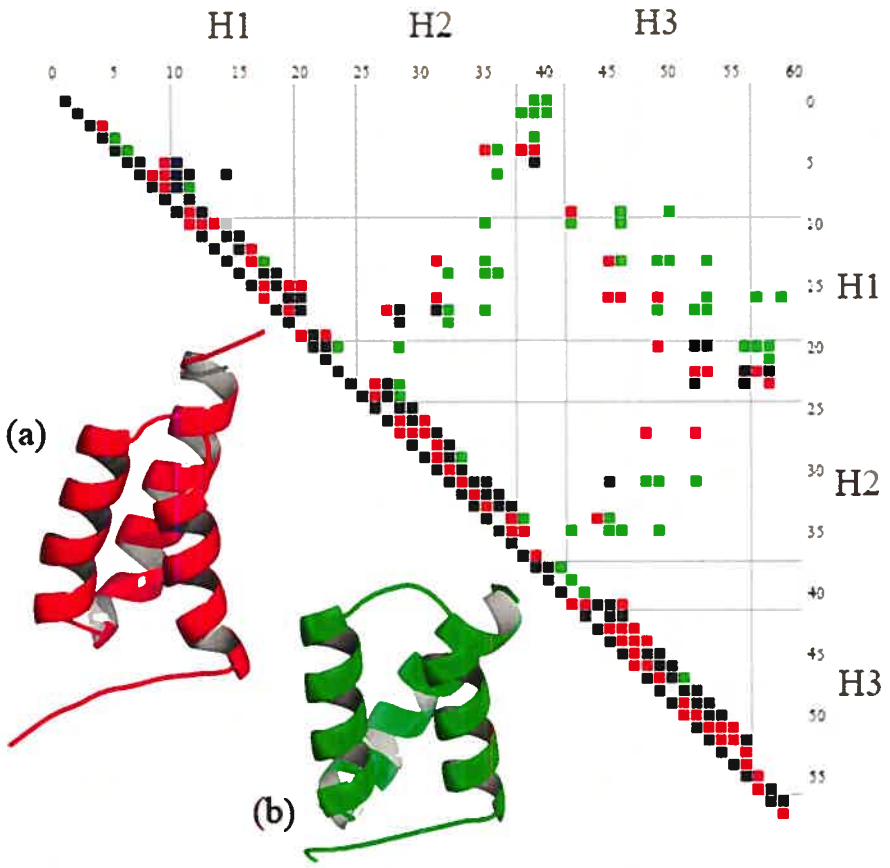

FIG. II.2 – Structure of protein A. (a) Experimental structure described with the OPEP potential; (b) minimized structure, after 10 ART nouveau steps using a Metropolis criterion of 300K, at -116.3 kcal/mol.

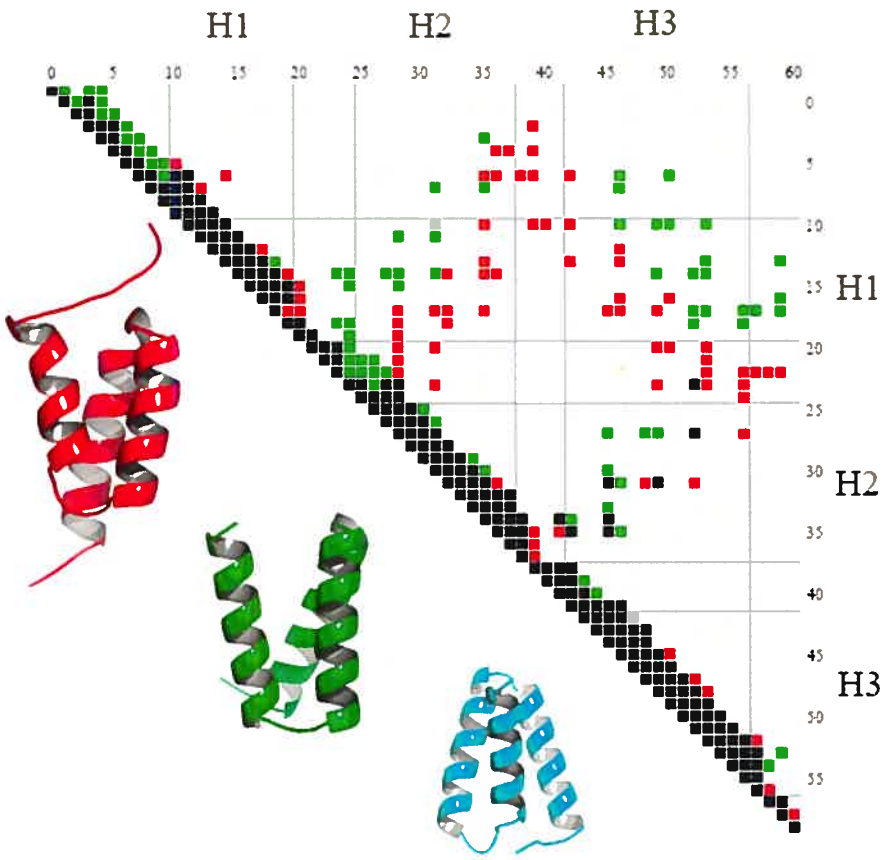

FIG. II.3 – Contact maps of the native structure obtained from NMR (red) and the left-handed structure of lowest energy found through a refinement simulation with a metropolis criterion of 600K (green). Also shown in cyan is the structure of lowest energy presenting a bundle of three helix with opposite coiling to the native conformation. Shared contacts are drawn black.

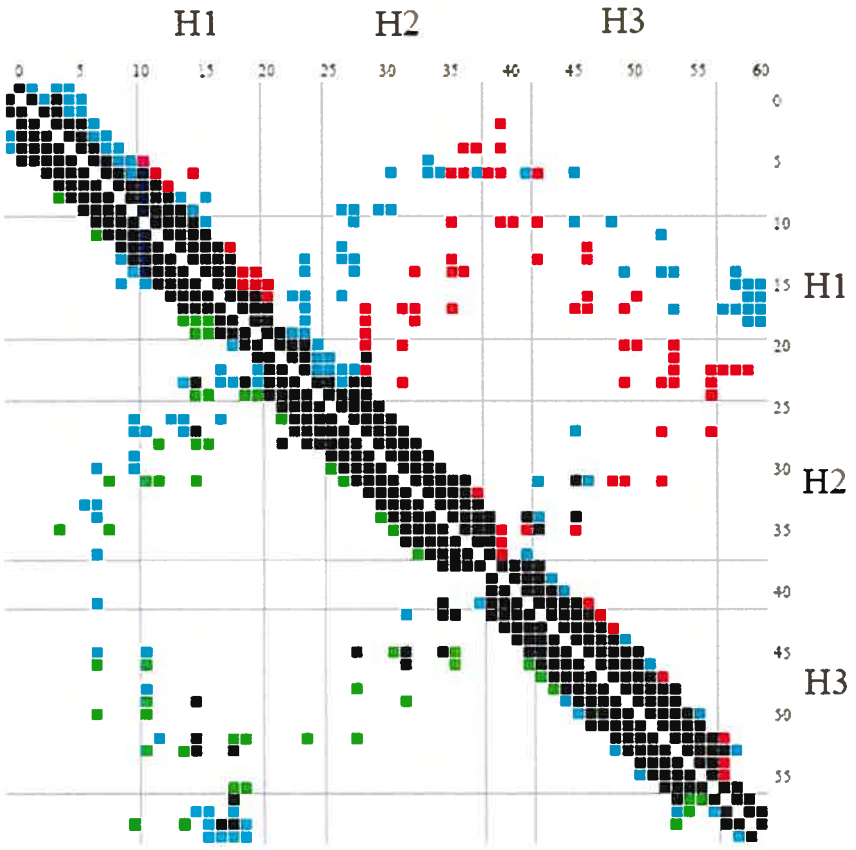

FIG. II.4 – Contact maps of the native conformation with the right-handed conformation of lowest energy (upper-right) and of the right-handed and left-handed lowest energy (lower-left). The native contacts are shown in red, those of the right-handed minimum in cyan and of the left-handed minimum in green. Shared contacts are drawn black.

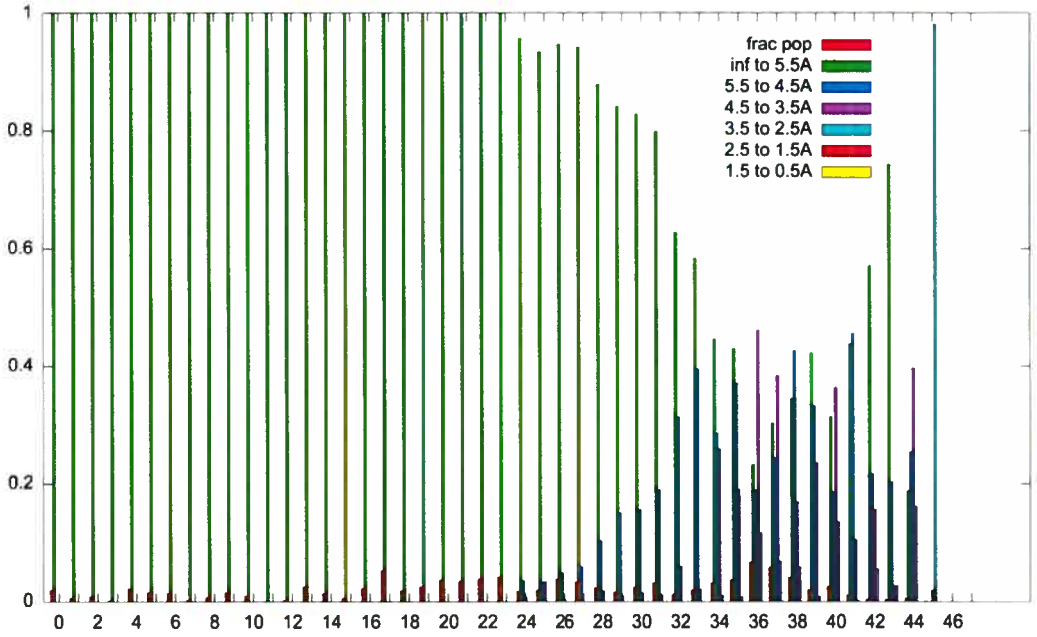

FIG. II.5 – Distribution of the RMS distance to the global lowest energy structure at various number of helical residues formed. In red is the corresponding fraction of the population

Annexe III

Accord des coauteurs de l'article

Nom de l'étudiant : Jean-François St-Pierre

Titre du programme : Maîtrise en bio-informatique, 2-468-1-0

Titre de l'article : The complex folding pathways of protein A suggest a multiple-funneled energy landscape

Liste des coauteurs : Jean-François St-Pierre, Normand Mousseau, Philippe Derreumaux

Status : Soumis pour publication au journal : Proceedings of the National Academy of Sciences of the United States of America

Déclaration des coauteurs :

À titre de coauteur de l'article identifié ci-dessus, je suis d'accord pour que Jean-François St-Pierre inclue cet article dans son mémoire de maîtrise qui a pour titre : *Méthodes de simulations moléculaires accélérées : application et développement*

Coauteur : Normand Mousseau

Signature : [redacted]

Date : 26 mars 2007

Coauteur : Philippe Derreumaux

Signature :

Date :

Gmail

JF SP [redacted]

Autorisation de l'utilisation d'un article dans le memoire de Jean-Francois St-Pierre

Philippe Derreumaux [redacted] JF St-Pierre [redacted]

Chere Mme Meunier, Chere Gertraud

Je vous souhaite une tres bonne année 2007.

Coauteur de l'article "The complex folding pathways of protein A suggest a multiple-funneled energy landscape" sousmis pour fin de publication au journal "Proceedings of the National Academy of Sciences of the United States of America", je suis d'accord pour que Jean-Francois St-Pierre inclut cet article dans son mémoire de maitrise en Bio-informatique qui a pour titre "Methodes de simulations moléculaires accélerées : application et développement"

Tres cordialement,
Philippe

Prof. Philippe Derreumaux
UFR de Biochimie, Universite Paris VII - Denis Diderot
Directeur du Laboratoire de Biochimie Theorique,
UPR 9080 CNRS, Institut de Biologie Physico-Chimique,
13 rue Pierre et Marie Curie, 75005 Paris
Tel 33 1 58 41 51 72
FAX 33 1 58 41 50 26