

2011.2614.4

Université de Montréal

Étude de l'espace dièdre dans un modèle de prédiction des protéines:
effets d'une distribution ciblée

par

Pierre-Jean L'Heureux

Département de chimie

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 chimie

décembre, 1997

© Pierre-Jean L'Heureux, 1997



QD

3

US4

1998

0.015



Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Étude de l'espace dièdre dans un modèle de prédiction des protéines:
effets d'une distribution ciblée

présenté par:

Pierre-Jean L'Heureux

a été évalué par un jury composé des personnes suivantes:

Benoît Roux, président-rapporteur
John R. Gunn, directeur de recherche
Michel Lafleur, membre du jury

Mémoire accepté le: ...25.04.1998

SOMMAIRE

La prédiction du repliement des protéines est un grand défi pour la science moderne. Du point de vue mécanique, une protéine peut adopter des millions de configurations. Pourtant, dans la nature, nous n'observons que quelques conformations pour une séquence donnée. Nous tentons de prédire cette structure tridimensionnelle.

Le groupe de John Gunn a développé un modèle du repliement des protéines: TRIP. TRIP adopte une vision très hiérarchique d'une recherche conformationnelle, où les structures secondaires sont rigides. En générant des milliers de conformations différentes, il guide l'investigation selon un Monte Carlo-recuit simulé. À chaque niveau de construction dans la hiérarchie, nous avons la possibilité d'ajouter des filtres pour écumer les conformations générées au hasard.

Comme les structures secondaires restent fixes, c'est-à-dire considéré comme connues avant le calcul, les seuls éléments mobiles du modèle sont les acides aminés dans les boucles. De plus, comme les seuls degrés de liberté que possède les acides aminés sont leurs angles dièdres autour du $C\alpha$, il semble essentiel de bien gérer ces angles dièdres.

J'ai, durant ma maîtrise, implanté dans TRIP la possibilité d'utiliser une carte finie de choix d'angles dièdres pour chaque résidu. J'ai donc pu tester les effets de distributions ciblées des angles ϕ et ψ sur la qualité de la prédiction de TRIP. Les angles ϕ et ψ sont les deux angles dièdres autour du $C\alpha$.

La capacité d'utiliser des cartes ciblées permet d'entrer plus d'informations

dans la recherche d'une conformation native. Ceci permet de guider davantage la recherche.

Au niveau des boucles, nous pouvons essayer les cartes ciblées suivantes:

Une seule carte identique pour tous les résidus, avec une variation dans le nombre de choix accessibles, de 676 à 6 points. Entre autre, la carte 676 points possèdent des points non-biologiques, afin de tester le modèle. On conclut qu'une carte de base de 50 points est largement suffisante.

Des cartes différentes selon le type d'acide aminé ne semblent donner aucun résultat concluant si la distribution des points est uniforme, c'est-à-dire lorsqu'un point n'apparaît qu'une seule fois dans la liste. Dans le cas de distribution non-uniforme, nous obtenons des résultats légèrement meilleurs, surtout avec des cartes spécifiques pour la proline.

Une carte différente pour chaque résidu, où l'on ajoute un nuage de points autour du couple d'angles cristallographiques donne les meilleurs résultats. Selon l'importance de la carte de base, TRIP redonne la structure native. Si on varie le nombre de points dans le nuage versus le nombre de points dans la carte de base, nous obtenons une relation plus ou moins linéaire entre le RMS et le pourcentage du poids du nuage. Si on laisse constant le poids du nuage versus la carte de base mais qu'on agrandit le nuage, TRIP conserve une bonne convergence vers la structure native. Si on met des nuages au hasard, TRIP ne choisit pas les nuages.

Au niveau des structures secondaires, nous pouvons essayer les cartes ciblées suivantes: Une carte différente pour chaque résidu, où il n'y a qu'un seul point, soit celui correspondant à la structure native. Le but de cette carte est surtout de faire des tests croisés avec les nuages dans les boucles, afin de voir s'il y a des effets synergiques.

Une carte différentes pour chaque résidu, où il y a deux points, soit

le point cristallographique et le point de la structure secondaire idéale, où TRIP décide énergétiquement lequel est favorable. Ceci montre que la structure cristallographique n'est pas toujours préférée.

Au niveau des retombées de cette étude, nous pouvons envisager d'améliorer substantiellement la qualité de la prédiction de TRIP avec des cartes ciblées. Particulièrement, des nuages de points autour d'un couple d'angles dièdres fourni expérimentalement par RMN semble une avenue très profitable.

Mots clés: protéine, repliement, prédiction, Ramachandran

TABLE DES MATIÈRES

SOMMAIRE	iii
TABLE DES MATIÈRES	vi
LISTE DES TABLEAUX	x
LISTE DES FIGURES	xiii
LISTE DES ABRÉVIATIONS	xvi
CHAPITRE 1: Introduction	1
1.1 Structure générale d'une protéine	5
1.1.1 Les acides aminés	5
1.1.2 Les angles dièdres de la chaîne principale	8
1.1.3 La hiérarchie dans la description d'une structure protéique	9
1.1.3.1 Le modèle des ressorts	13
1.2 Énergie en jeu	13
1.2.1 Le rôle de l'eau dans le repliement d'une protéine	13
1.2.2 Information égale énergie	16
1.2.3 Notion générale d'un potentiel statistique	17
1.2.3.1 Défauts des potentiels statistiques	17

1.3	Le problème de l'explosion combinatoire dans une prédiction de structure protéique tri-dimensionnelle	19
1.3.1	Le principe hiérarchique comme méthode pour jongler avec l'explosion combinatoire	19
1.3.2	Le biaisage des essais, pour freiner l'explosion combinatoire	20
1.4	Le problème des minima multiples dans la fonction d'énergie . . .	21
1.4.1	Le recuit simulé comme méthode pour jongler avec les minima multiples	21
1.4.2	L'algorithme Monte Carlo, pour échantillonner une surface d'énergie	22
1.4.2.1	Application aux protéines de l'algorithme Monte Carlo	23
1.4.3	Frictions entre Monte Carlo et recuit simulé	23
1.5	Quelques notion de RMS, pour calculer la ressemblance entre deux structures	24
CHAPITRE 2: Le modèle		27
2.1	Construction de la protéine	27
2.1.1	Définition de l'espace conformationnel	28
2.1.2	L'acide aminé, la sous-unité dans TRIP	29
2.1.3	La liste de triplets, l'unité de base de la mutation dans TRIP	30
2.1.4	La liste de boucles, un ensemble qui évolue	32

2.1.5	La génération de molécules, l'étape finale avant la recombinaison	33
2.1.6	L'algorithme génétique, pour recombinaison des bonnes solutions éparpillées	33
2.2	La fonction de potentiel utilisé dans ce travail	34
2.2.1	Force hydrophobe	34
2.2.2	Potentiel statistique	34
2.2.3	Volume exclu	35
2.3	Généralité sur les niveaux de filtrage	35
CHAPITRE 3: Problèmes locaux		37
3.1	Choix du niveau de filtrage: espace dièdre des boucles	39
3.1.1	Choix du type de manipulation	40
3.2	MOLFIX	41
3.2.1	Choix du type de RMS	42
3.2.2	Barrières de distance et d'angle	42
3.3	Manipulation de l'espace dièdre des boucles	43
3.3.1	Réduction unilatérale de l'espace conformationnel	44
3.3.2	Carte selon l'espèce chimique, distribution uniforme	50
3.3.3	Carte selon l'espèce chimique, distribution non-uniforme	53
3.3.4	Carte avec les angles MOLFIX dans les boucles	58

3.3.5	Carte avec des nuages autour des angles MOLFIX uniquement dans les boucles	59
3.3.6	Synthèse de l'espace des boucles	62
3.4	Espace dièdre des structures secondaires	62
3.5	Accès aux angles MOLFIX dans tous les espaces dièdres	65
3.5.1	Carte avec les points MOLFIX avec ou sans choix de structures secondaires	66
3.5.2	Carte avec nuage MOLFIX, avec ou sans choix de structures secondaires	69
3.6	Carte avec des nuages au hasard	77
3.7	Synthèse sur l'espace dièdre des boucles combinées aux structures secondaires	79
CHAPITRE 4:	Conclusion	82
RÉFÉRENCES	85
APPENDICE A:	Fichier PDB des protéines cobayes	xviii

LISTE DES TABLEAUX

I	Distances intramoléculaires des acides aminés dans TRIP.	29
II	Angles intermoléculaires des acides aminés dans TRIP.	30
III	Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP.	38
IV	Effet de la réduction uniforme de l'espace dièdre pour un calcul TRIP typique sur la myoglobine.	46
V	Effet de cartes uniformes pour différents acides aminés dans un calcul TRIP typique sur la myoglobine.	52
VI	Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP avec les cartes X et PRO.	54
VII	Effet de cartes non-uniformes pour différents acides aminés dans un calcul TRIP typique sur la myoglobine.	57
VIII	Ventilation de l'énergie pour un calcul TRIP sur la myoglobine avec des cartes d'acides aminés non-uniformes.	60
IX	Effets des couples d'angles MOLFIX uniquement dans les boucles sur un calcul typique de TRIP sur la myoglobine.	60
X	Effets des nuages MOLFIX uniquement dans les boucles sur un calcul typique de TRIP sur la myoglobine.	63

XI	Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP avec des nuages MOLFIX.	63
XII	Effets des structures secondaires MOLFIX sur un calcul typique de TRIP sur la myoglobine.	68
XIII	Effets des couples d'angles MOLFIX avec les structures secondaires MOLFIX sur un calcul typique de TRIP sur la myoglobine.	68
XIV	Ventilation de l'énergie pour un calcul TRIP sur la myoglobine avec et sans choix de structure secondaire, avec les couples d'angles MOLFIX.	70
XV	Effets des nuages MOLFIX sur les boucles, combinés avec des hélices MOLFIX, sur un calcul typique de TRIP sur la myoglobine.	70
XVI	Pourcentage moyen d'utilisation des hélices natives, pendant les calculs TRIP interactifs avec nuages MOLFIX sur la myoglobine.	76
XVII	Effets du poids des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, avec choix de structures secondaires . . .	76
XVIII	Effets de la superficie des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, avec choix de structure secondaire. . .	78
XIX	Effets du poids des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, sans choix de structure secondaire. . . .	78
XX	Effets synergiques des hélices cristallographiques et des nuages MOLFIX dans les boucles sur les prédictions de TRIP pour la 1CTF.	78

XXI Effets de nuages au hasard sur un calcul typique de TRIP sur
la myoglobine. 81

LISTE DES FIGURES

1	Les acides aminés naturels.	6
2	Dimensions de base d'un peptide.	7
3	Les angles ϕ et ψ d'un lien peptidique.	9
4	Carte du modèle rigide de Ramachandran.	10
5	Carte Ramachandran de population hors des structures secondaires ¹ . a) Carte des statistiques sur la proline. b) Carte des statistiques sur la glycine. c) Carte des statistiques sur les résidus précédant une proline. d) Carte des statistiques sur tous les acides aminés naturels, excepté les trois précédents.	11
6	Distribution moyenne du RMS pour des structures compactes générées au hasard et des structures générées par TRIP, ainsi que la déviation moyenne pour les structures générées au hasard.	26
7	Pyramide de la construction d'une protéine dans TRIP.	28
8	Angle ϕ de l'hélice 6 de la structure native de la myoglobine, de la structure générée avec un RMS carbone α et générée par un RMS 6 atomes. L'axe des x correspond au numéro des résidus, à partir du résidu 80.	43

- 9 Les cartes réduites testées dans ce travail. a) 676 points b) 534 points c) 372 points d) 372 points négatifs e) 100 points f) 50 points g) 18 points h) 6 points 45
- 10 Représentation schématique de la myoglobine obtenue par TRIP avec une carte à 6 points, avec le RMS comme seul potentiel. Réalisé avec MOLSCRIPT². 48
- 11 Utilisation typique des couples d'angles dièdres de la carte 534 par un résidu dans une boucle. Ici, le résidu 55 de la myoglobine. Une unité dans l'axe z correspond à une conformation acceptée. 49
- 12 Utilisation typique des couples d'angles dièdres de la carte 534 par un résidu dans une boucle. Coupe horizontale. Une unité dans l'axe z correspond à une conformation acceptée. 50
- 13 Les cartes spécifiques uniformes selon le type d'acide aminé testées dans ce travail. a) Glycine 324 couples. b) Proline 186 couples. c) Résidu avant une proline 201 couples. d) Tous résidus sauf glycine, proline et résidu devant proline 273 couples. 51
- 14 Les cartes spécifiques non-uniformes selon le type d'acide aminé testées dans ce travail a) Glycine 351 couples. b) Proline 322 couples. c) Résidu avant une proline 133 couples. d) Tous résidus sauf glycine, proline et résidu devant proline 415 couples. 55
- 15 Utilisation des couples d'angles dièdres de la carte NPRO non-uniforme pour un calcul TRIP typique sur la myoglobine. Une unité dans l'axe z correspond à une conformation acceptée. 57
- 16 Utilisation des couples d'angles dièdres des cartes avec points MOLFIX. Boucle F de la myoglobine, soit les résidus 54, 55 et 56. Une unité dans l'axe z correspond à une conformation acceptée. 71

- 17 Utilisation des couples d'angles dièdres des cartes avec nuage
MOLFIX. Boucle F de la myoglobine, soit les résidus 54, 55 et
56. Une unité dans l'axe z correspond à une conformation acceptée. 74
- 18 Utilisation des couples d'angles dièdres des cartes avec nuage
sur les bouts des boucles. Boucle F de la myoglobine, soit les
résidus 54, 55 et 56. Une unité dans l'axe z correspond à une
conformation acceptée. 80

LISTE DES ABRÉVIATIONS

CHARMM *Chemistry at HARvard Macromolecular Mechanics*

RMN *Résonance magnétique nucléaire*

RMS *Root Mean Square deviation*

PDB *Brookhaven Protein Data Bank*

TRIP *Programme de prédiction du repliement des protéines. John Gunn*

MOLFIX *Programme qui génère la structure la plus ressemblante possible d'une protéine connue avec un modèle réduit des protéines.*

DC *Dichroïsme circulaire*

Au Père Vachon, ce sage-homme

CHAPITRE 1

Introduction

Il existe un vaste domaine de recherche qui tourne autour du repliement des protéines. Les questions qui y sont débattues nous font réaliser toute l'étendue de notre ignorance. Spécialement au niveau atomique, la description des phénomènes du repliement et de la stabilité des protéines est nébuleuse. Il y a plusieurs approches expérimentales et théoriques qui sont menées présentement dans le monde. Du côté expérimental, on peut diviser le tout ainsi:

1. détermination de la séquence par biologie moléculaire³
2. détermination des structures secondaires par DC ⁴ ou autre spectroscopie (RMN⁵)
3. détermination de structure tridimensionnelle par cristallographie (rayon X/neutron⁶, électronique⁷) ou RMN⁸
4. compréhension et optimisation de la séquence par mutagenèse dirigée⁹
5. analyse du processus du repliement dans un intervalle de 10^{-9} à 10^{-6} sec.
¹⁰ par RMN ou fluorescence

Du côté théorique, nous avons:

1. prédiction de structure secondaire^{11 12 13}
2. prédiction de structure tertiaire
3. étude thermodynamique de la transition de phase ^{14 15 16 17}

Je passerai sous silence le domaine qui gravite autour de la fonction des protéines. Ce champ de connaissances est beaucoup plus développé, mais sort de mon

optique structurale. Il est évident cependant que les deux vont de pair dans la nature.

Comme dans bien d'autres domaines de la recherche scientifique, le besoin pour des modèles intelligibles, qui structurent notre façon de voir un phénomène, est tangible. L'esprit humain aime donner un sens aux phénomènes qui l'entourent. En physique, la méthode scientifique préconise la construction de modèles finis, où nous contrôlons chaque intervenant par des équations mathématiques. La transcription d'un modèle pensé vers une série d'équations mathématiques permet de jouir de la robustesse du système mathématique. La preuve mathématique rend le modèle physique «démonstrable». Le développement de modèles toujours plus précis et plus englobants semble être allé de pair avec le développement des mathématiques. D'autre part, la construction d'un modèle physique qui reproduit les données expérimentales donne un caractère de véracité à ce modèle. Ultimement, un modèle qui rend compte de toutes les données expérimentales d'un phénomène forge notre façon de voir ce phénomène, car il semble être un reflet des véritables forces en présence.

Grâce à l'arrivée des ordinateurs, les mathématiques ont, en quarante ans, plus évolué que durant les deux siècles précédents. Autant au niveau des théories mathématiques que du calcul lui-même, l'ordinateur a permis de grandes avancées. Tellement qu'à certains égards, la frontière entre l'informatique et la mathématique n'existe plus. Parallèlement, plusieurs domaines de recherche, où les données expérimentales semblaient trop disparates pour être unifiées sous un modèle physique, ont vu poindre à l'horizon la possibilité de traduire en langage mathématique, ou à tout le moins informatique, les phénomènes qui les préoccupaient.

La biochimie est un de ces domaines. La biochimie théorique a pour mandat de construire des modèles qui touchent des phénomènes biochimiques. Ici, le repliement des protéines est une avenue de recherche pour la biochimie théorique.

Les protéines ont plusieurs fonctions chez le vivant, nommément: la catalyse par les enzymes, le transport à travers les membranes cellulaires, la communication entre les cellules, réplication, transcription, traduction, rigidification par assemblage, etc ¹⁸ . Si nous avons la possibilité de modéliser les protéines à notre bon vouloir, non pas uniquement de modifier des protéines existantes, mais bien d'en construire de toutes nouvelles, les applications seraient innombrables. Intégrer de nouvelles composantes aux organismes vivants, pour les extraire ou modifier le développement de leur hôte. Une toute nouvelle ère pour les industries chimiques et pharmaceutiques. Où en sommes-nous ? La réponse est rapide: très loin du but.

D'immenses défis jalonnent les recherches interdisciplinaires dans ce domaine. On peut les regrouper en trois sphères:

1. gérer l'explosion combinatoire des possibilités
2. mettre au point une fonction d'énergie universelle et précise
3. intégrer les effets cinétiques et les notions de transition de phase.

Ce mémoire résume mes courtes investigations dans la sphère 1.

Intuitivement, pour prédire par ordinateur la structure d'une protéine, il suffirait d'avoir une fonction de potentiel précise et de faire minimiser la protéine dépliée dans ce potentiel. Malheureusement, la topologie d'un potentiel tout atome pour les protéines est très rêche. La minimisation de la protéine buterait sur d'innombrables minima. Une autre façon intuitive de prédire la structure d'une protéine serait d'essayer toutes les conformations possibles de la protéine, en évaluant l'énergie. Le tout serait de choisir la conformation qui a l'énergie la plus basse. Mais le nombre de conformations possibles est astronomique.

Pour rendre le problème accessible, les scientifiques réduisent le nombre de conformations possibles, en utilisant des modèles simplifiés des protéines. De plus, ils mettent au point des algorithmes intelligents de recherche, souvent à cheval sur la minimisation et une recherche purement au hasard. Dans ce sens, certains

chercheurs ont proposé de découper la protéine en fragments, et de traiter chaque fragment de façon locale (built-up procedure). L'idée étant que l'emboîtement des meilleurs fragments donne la meilleure protéine^{19 20}. Notre approche génère des nouvelles conformations par de petits changements ciblés dans les boucles. Nous préférons cette approche, car:

- elle permet d'utiliser plus facilement des algorithmes qui traitent la protéine de façon globale
- elle exploite plus avantageusement les ressources informatiques à notre disposition.

De façon plus pragmatique, le groupe de John Gunn travaille sur une petite partie du problème: comment prédire la structure tridimensionnelle d'une protéine à partir de sa séquence et de sa structure secondaire (sphère 2).

Au cours de ma recherche, j'ai manipulé l'espace dièdre des boucles et des structures secondaires pour améliorer la qualité de la prédiction. J'ai tout d'abord diminué le nombre de choix possibles, puis ajouté des nuages denses autour des angles dièdres cristallographiques des boucles et finalement utilisé les structures secondaires cristallographiques. J'ai couplé ces méthodes entre elles pour finalement avoir un portrait de l'impact qu'avait la connaissance des angles dièdres dans le calcul de prédiction. Ultiment, nous voyons que des données expérimentales par RMN, qui donnent les angles dièdres à plus ou moins 30°, augmenteront substantiellement la qualité de la prédiction. En termes de réalisation, j'ai tout d'abord écrit un logiciel, PDB2NAT, qui traduit un fichier du Brookhaven Protein Data Bank en fichier lisible par TRIP et MOLFIX. J'ai ensuite développé un logiciel, MOLFIX, qui minimise le RMS d'une protéine construite selon notre modèle réduit par rapport à une structure du Protein Data Bank. J'ai implanté dans TRIP des routines qui lui permettent d'utiliser une carte de Ramachandran par résidu. J'ai écrit un logiciel, CREDIH, qui construit une carte de Ramachandran lisible par TRIP. CREDIH permet d'utiliser les informations de MOLFIX, d'ajouter des nuages de formes diverses, tout cela

en partant d'une carte de base. J'ai finalement ajouté la possibilité pour TRIP de changer de structures secondaires durant la minimisation.

1.1 Structure générale d'une protéine

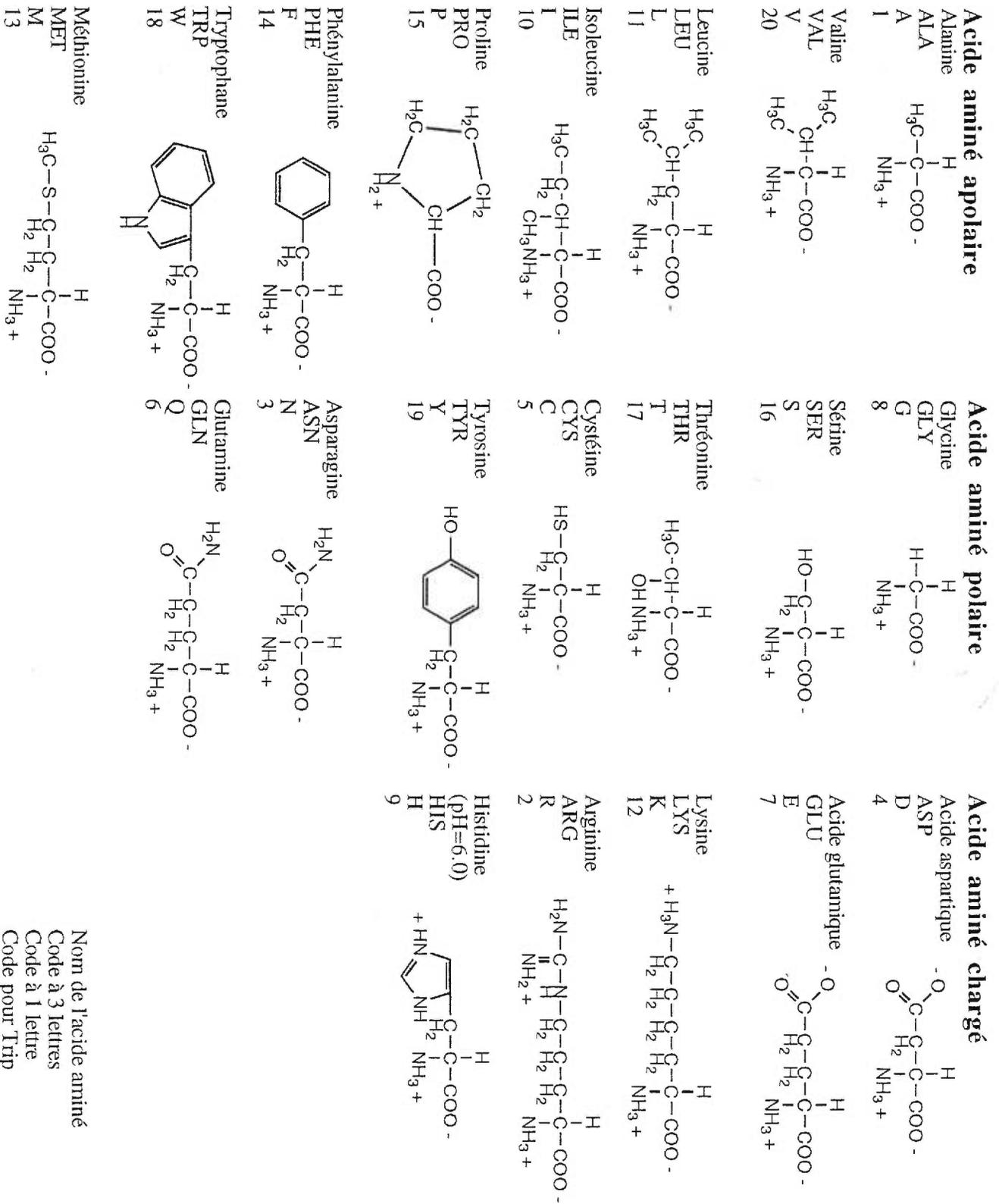
Une protéine est un polymère d'assemblage linéaire d'acides aminés unis par des liaisons peptidiques, qui, sous différentes forces faibles, se recroqueville pour devenir biologiquement actif²¹. Ce qui différencie une protéine d'un polymère synthétique conventionnel est l'unicité de sa conformation repliée. Plus spécifiquement, l'état biologique (natif) est un très petit ensemble de structures tridimensionnelles analogues, parmi une mer de possibilités²². Cette unicité procure des caractéristiques importantes aux protéines, dont leur spécificité d'action. En bref, une séquence, une fonction ³.

1.1.1 Les acides aminés

Un acide aminé est un composé organique qui possède nécessairement deux groupements actifs: un acide et une amine, sur sa chaîne principale. L'acide est à un bout de la chaîne, et l'amine est placée ailleurs sur la chaîne. Le type (α , β , δ , etc.) d'acide aminé est relié à la position de l'amine.

La nature a sélectionné au cours du temps un type d'acide aminé où la fonction amine et la fonction acide sont séparées par un seul carbone. Ce sont les acides α aminés, car le carbone qui attache la fonction amine est le carbone en position α . Le carbone α est chiral, sauf pour la glycine. Dans les acides aminés naturels, la conformation du centre chiral est exclusivement gauche(L).

Il y a 20 acides aminés naturels, qui peuvent, par exemple, être divisés en trois groupes selon leur réactivité chimique: polaire, chargé, hydrophobe. Tous ont le même squelette, excepté la proline. La différence entre les différents acides aminés provient de la chaîne latérale qui pend de ce squelette²³, à partir du carbone α . C'est la chaîne latérale qui donne la spécificité à chaque acide aminé. La réactivité chimique de la chaîne latérale rend uniques les forces physiques qui sont à l'oeuvre entre l'acide aminé et le solvant, ainsi qu'entre l'acide aminé et d'autres acides



Nom de l'acide aminé
Code à 3 lettres
Code à 1 lettre
Code pour Trip

FIGURE 1. Les acides aminés naturels.

aminés .

Le groupement acide et le groupement amine ont des affinités chimiques. L'adjonction d'un acide aminé côté acide sur un autre acide aminé côté amine crée un lien covalent appelé lien peptidique. Nous pouvons, en ajoutant ainsi de suite d'autres acides aminés, créer un polymère linéaire d'acides aminés. C'est dans le choix de la séquence en acide aminé que la nature réussit à modifier la structure d'une protéine. En effet, chaque protéine subit une somme de forces différente de son environnement parce qu'elle a une séquence différente. Deux principes sont à l'oeuvre dans le repliement d'une protéine: la stabilité et la spécificité. Lors de l'évolution, ces deux principes ont guidé le choix des séquences. On peut donc dire que la séquence est la solution dans une recherche qui veut maximiser les deux, pour une fonction donnée.

Étudions plus avant les liaisons peptidiques.

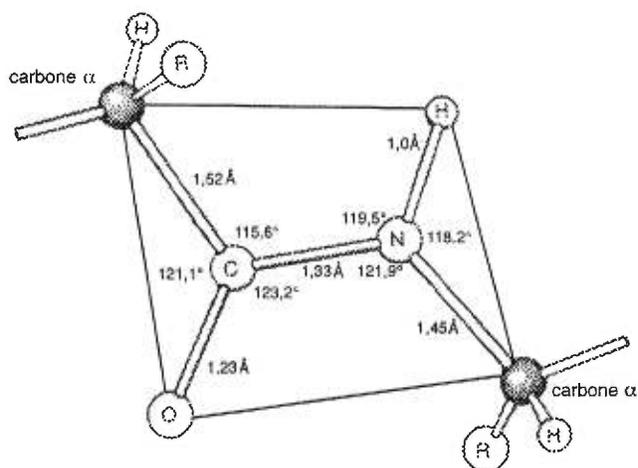


FIGURE 2. Dimensions de base d'un peptide.

Comme la figure 2 nous le montre (tirée de Zubay²³), le lien peptidique peut être considéré comme plan. Le lien peptidique est plan à cause du recouvrement des orbitales π du lien C-N. Parallèlement, une délocalisation des orbitales σ compense la perte de nuage électronique de l'azote. Des forces coulombiennes entre l'azote et le carbone raccourcissent la liaison. Le recouvrement des orbitales

π du lien C-N transforme le double lien du carbonyle. Les atomes d'oxygène, de carbone et d'azote sont en résonance²⁴. Dans un modèle réduit des protéines, il apparaît acceptable de considérer la protéine comme une suite de plan peptidique qui s'attache via le carbone α , puisque dans le PDB l'angle dièdre du lien peptidique est de $-179,5^\circ \pm 3,8$ ^{24 23}.

1.1.2 Les angles dièdres de la chaîne principale

Nous verrons plus loin dans le chapitre 2 que la connaissance des angles dièdres qui s'articulent autour du carbone α est suffisante pour construire le squelette d'une protéine en trois dimensions, si on considère comme plane la liaison peptidique et comme fixe tout autre degré de liberté. Dans une représentation de Newman, on peut visualiser ces angles dièdres. On place le carbone α à l'avant-plan. L'hydrogène pointe vers le haut. L'azote est à l'arrière-plan et caché par le carbone α . L'angle ϕ est celui qu'on calcule avec le carbonyle relié à l'azote. L'angle ϕ est à 180° lorsque le carbone du carbonyle est entre l'hydrogène et le carbone β . Pour l'angle ψ , le carbone α est encore à l'avant-plan, et cache le carbone du carbonyle. On calcule l'angle avec l'azote suivant. Il est à 180° lorsque l'azote suivant est entre l'hydrogène et le carbone β . Lorsque ces deux angles sont à 180° , le peptide est dans sa forme étendue. Une rotation dans le sens des aiguilles d'une montre est positive (fig. 3 tirée du Zubay²³).

Si nous portons ces deux angles sur un graphique, nous construisons un espace à deux dimensions appelé carte de Ramachandran, en reconnaissance du travail de ce pionnier dans le domaine.

Ramachandran a eu l'idée de construire un modèle réduit de peptide où il pouvait faire pivoter de 360° les angles ϕ et ψ ²⁵. Ce faisant, il a démontré qu'à cause de l'encombrement stérique, certaines régions du plan sont interdites (figure 4, tirée du Zubay²³). Plus tard, des statistiques tirées du Protein Data Bank ont confirmé que certaines régions de la carte ne sont pas du tout utilisées. Ces statistiques ont aussi confirmé le caractère particulier de la proline et de la glycine dans les structures protéiques. En effet, leurs cartes d'utilisation de l'espace dièdre sont

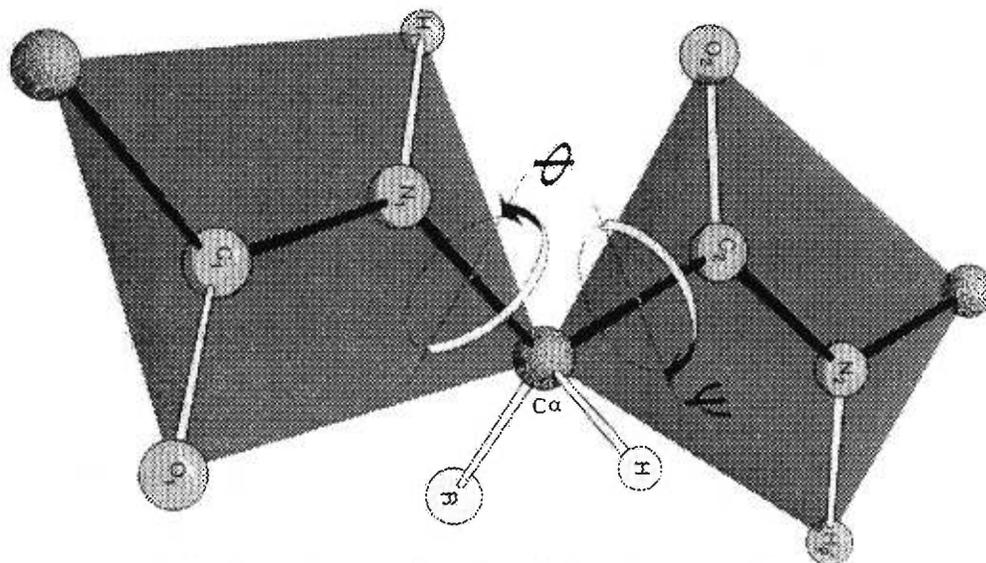


FIGURE 3. Les angles ϕ et ψ d'un lien peptidique.

fondamentalement différentes des dix-huit autres acides aminés²⁶. Dans le cas de la glycine, l'absence de carbone β réduit l'encombrement stérique et permet d'adopter des conformations impossibles pour les autres acides aminés. Dans le cas de la proline, les conformations cis, difficilement accessibles pour les autres acides aminés, deviennent moins défavorables. La figure 5 montrent les statistiques tirées d'un ensemble échantillon du PDB, réalisé par Anthony Keith Felts¹.

En analysant plus avant la carte de Ramachandran, on peut établir des corrélations entre des régions de la carte et des structures peptidiques. Dans la figure 4, ces régions sont fléchées. La relation entre une position sur la carte Ramachandran et une structure en trois dimensions est bidirectionnelle.

1.1.3 La hiérarchie dans la description d'une structure protéique

Les scientifiques ont découpé la description de la conformation d'une protéine en quatre. La structure primaire est la structure covalente, exprimée en termes de séquence de la protéine et de modifications post-transcriptionnelles. La structure secondaire (page 15) regroupe les acides aminés dont le squelette est dans des

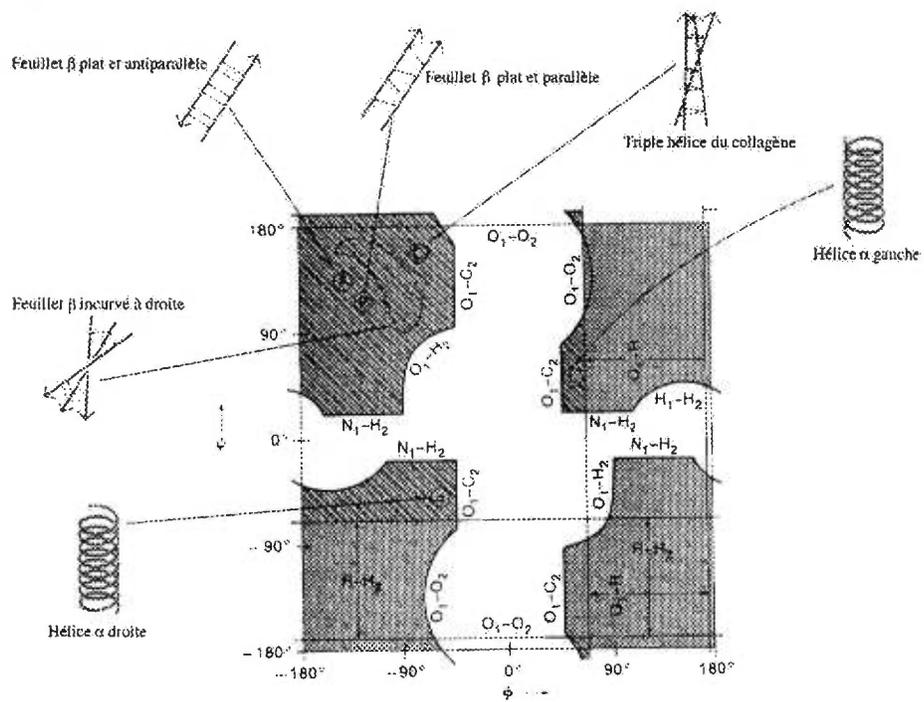


FIGURE 4. Carte du modèle rigide de Ramachandran.

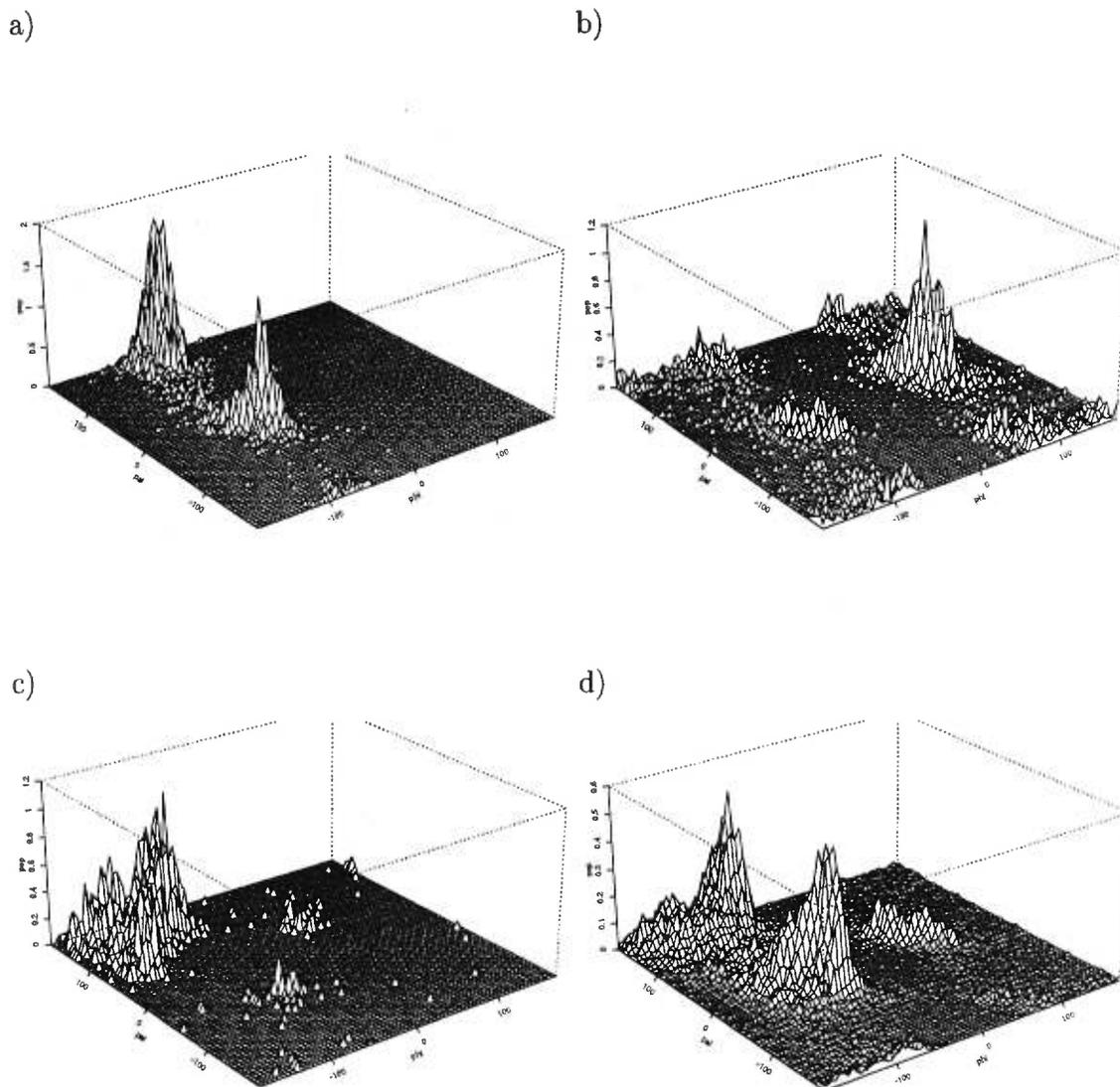


FIGURE 5. Carte Ramachandran de population hors des structures secondaires¹.
a) Carte des statistiques sur la proline. b) Carte des statistiques sur la glycine.
c) Carte des statistiques sur les résidus précédant une proline. d) Carte des
statistiques sur tous les acides aminés naturels, excepté les trois précédents.

unités de structures géométriques répétitives, entre autres l'hélice α et le feuillet β . Dans la figure 4, ceci correspond aux régions du même nom. La structure tertiaire décrit l'ensemble de la structure géométrique. D'un autre point de vue, elle est l'agencement des structures secondaires entre elles. Finalement, la structure quaternaire, réservée aux agrégats de protéines, décrit l'agencement des monomères entre eux.

Plusieurs expériences nous amènent à penser que le secret du repliement est dans la séquence²⁷. C'est elle qui dicte le chemin du repliement, et ultimement décide de la conformation tridimensionnelle finale. Même en donnant une place exagérée à des aides extérieures dans le processus du repliement (chaperonine, etc.) , nous devons admettre la nature spatiale des affinités entre les aides et la protéine, et donc poser la question du lien entre la séquence et la conformation, à tout moment du repliement.

Par exemple, nous avons de forts indicateurs que les structures secondaires et tertiaires se forment de façon coopérative. En effet, plusieurs expériences (expérimentales et théoriques) où on construit un polypeptide avec la séquence d'une hélice provenant d'une protéine aqueuse donnent des peptides avec des conformations dépliées^{28 29}.

En général, dans les protéines globulaires en phase aqueuses, les structures secondaires forment de 40 à 70% de la structure tri-dimensionnelle. Ce pourcentage correspond à la balance énergétique entre former la protéine la plus petite possible (force hydrophobe) et placer les acides aminés polaires le plus à l'extérieur possible (force électrostatique).

Encore une fois, ce travail s'intéresse à prédire la structure tertiaire à partir de la structure primaire et secondaire. Dans sa version originale, TRIP a besoin de la position sur la séquence, de la longueur et du type de structure secondaire. De là, il approxime une géométrie. Cette géométrie reste fixe durant tout le calcul. Ceci nous dispense de traiter explicitement plusieurs forces locales, en plus de réduire dramatiquement le nombre de degrés de liberté.

1.1.3.1 Le modèle des ressorts

Pour faire image, on peut penser à des cylindres en velcro rigides reliés par des ressorts qu'on tortille dans le but d'optimiser les contacts cylindre-cylindre. Dans ce modèle, les seuls éléments flexibles sont les ressorts. Ma recherche consiste, dans un premier temps, à limiter les positions que le ressort peut prendre dans l'espace afin d'augmenter la qualité de la prédiction, et deuxièmement à assouplir les cylindres.

1.2 Énergie en jeu

Bien sûr, la séquence ne se replie pas *in vacuo*. Elle a été sélectionnée et améliorée durant l'évolution pour optimiser son travail dans son milieu: cytoplasme, sang, chime, etc., mais aussi pour se replier dans des milieux précis: appareil de Golgi, membrane cytoplasmique, mitochondrie, chloroplaste, etc. À proprement parler, notre étude du repliement est bien loin de tous ces détails. Il est uniquement intéressant ici de souligner la complexité du processus. Dans ce travail, nous verrons comment gommer toutes ces différences grâce à un potentiel statistique, au prix d'une perte de sensibilité de la prédiction.

1.2.1 Le rôle de l'eau dans le repliement d'une protéine

Comme les protéines qui m'ont intéressé proviennent toutes de la phase aqueuse, nous parlerons un peu de l'eau liquide. Plusieurs caractéristiques physiques donnent à l'eau des propriétés importantes. Premièrement, la molécule possède un dipôle permanent. En se réorientant, les dipôles des molécules d'eau cause un effet d'écran aux forces électrostatiques. Deuxièmement, sa relative polarisabilité est à l'origine des forces d'hydratation. Troisièmement, l'eau peut participer à 4 ponts hydrogènes par molécule (par la suite nommé pont H).

Aujourd'hui encore, malgré la puissance de nos ordinateurs, l'eau reste un système très difficile à modéliser³⁰. On s'accorde pour dire qu'elle est un réseau dynamique basé sur ses ponts H. Ce réseau peut être modifié de plusieurs façons: ajout de sel,

température, etc. Ces phénomènes extérieurs modifient les forces interatomiques qui régissent les propriétés physico-chimiques du solvant. Ce qui est particulier à l'eau, c'est son étonnante rigidité pour un liquide. La moindre perturbation du réseau demande un apport important d'énergie.

L'inclusion d'une molécule, quelle qu'elle soit, dans le réseau de ponts H de l'eau détruira un certain nombre de ces ponts, ou à tout le moins forcera le réseau à s'ajuster. La balance énergétique du phénomène est compliquée à calculer, puisque la rupture de ces ponts cause une augmentation d'enthalpie et à la fois une perte d'entropie, car l'eau perd des possibilités de liaisons en s'organisant autour de la molécule.

La molécule de soluté peut, en compensation, faire des ponts H, abaissant l'enthalpie. Elle peut aussi être chargée, ce qui incitera le dipôle de l'eau à s'orienter conséquemment. Si le rayon de l'ion est petit, il se formera alors une véritable boule d'eau autour de la charge. C'est la force d'hydratation.

Si ni l'un ni l'autre cas n'existent, nous observons un phénomène entropique assez inusité. L'eau liquide tente de satisfaire le maximum de ponts H possible avec elle-même, afin d'abaisser son enthalpie. Ce faisant, elle abaissera son entropie de façon dramatique. Ce gain d'ordre peut même aller jusqu'à la création d'une cage autour de la molécule solvatée (clathrate). Cette perte d'entropie est à l'origine de la force hydrophobe, essence de la structure des protéines globulaires.

Le mélange inévitable de toutes ses forces dans le repliement des protéines rend la description théorique détaillée du phénomène plutôt obscure. En effet, ce fameux réseau de ponts H semble s'étendre sur plusieurs nanomètres de long, rendant complexe la structure même de l'eau autour de la protéine, étant donné la corrélation élevée des H_2O dans le milieu.

Dans une séquence normale de protéine, on retrouvera des acides aminés non-polaires. Ceux-ci sont à l'origine du recroquevillement forcé des protéines. En effet, en subissant la force hydrophobe de l'eau, les chaînes latérales non-polaires (hydrophobes) s'empilent les unes sur les autres afin d'éviter tout contact avec

l'eau. Ce faisant, elles obligent le squelette à se replier.

Le squelette, pour sa part, possède un donneur et un receveur de pont H par résidu. Comme la formation d'un pont H est enthalpiquement favorable, toute façon de maximiser leur nombre sera conservée. Ceci est à l'origine des hélices α , des feuilletts β et d'autres structures ordonnées qu'on voit dans les protéines. Ces structures ordonnées sont formées de ponts H intramoléculaires. On les appelle des structures secondaires.

Des forces de Van der Waals non négligeables se manifestent ensuite, lorsque l'empilement des acides aminés est très serré. Ainsi, le squelette de l'hélice α est stabilisé par ce genre de force.

Il reste maintenant à satisfaire les charges et dipôles des autres chaînes latérales. La plupart s'orienteront vers le solvant, mais des recouvrements internes sont aussi possibles. Le tout sera régi par les lois des forces électrostatiques, enthalpiquement favorables. Il est à noter cependant que l'hydratation des charges sera entropiquement défavorable.

Idéalement, la protéine globulaire en milieu aqueux n'offrira à sa surface que des résidus chargés ou polaires et cachera tous ses résidus hydrophobes. Un véritable milieu sans eau existera à l'intérieur de la protéine³¹.

Ce tour d'horizon rapide des forces en présence nous permet de sentir la délicate balance entre les forces non-covalentes³². Les effets sur la protéine d'un changement physico-chimique du solvant aura des répercussions sensibles sur ce mince équilibre. En effet, si, par un procédé, les ponts H de l'eau deviennent moins forts, les forces hydrophobes, moteur du repliement, seront moins fortes aussi. Autre exemple, si l'eau réussit à entrer dans le coeur de la protéine, elle pourra briser les structures secondaires en s'y intercalant par ponts H. On peut aussi imaginer un milieu moins propice aux interactions électrostatiques, par saturation de l'eau avec des sels. On y verra une précipitation des protéines (Salting out)²³. Bref, on devrait maintenant être convaincu de l'importance du solvant dans l'obtention de la conformation unique et biologiquement active d'une protéine.

Il ne faut cependant pas oublier qu'ultimement, c'est la séquence primaire qui dicte la forme tridimensionnelle de la protéine²⁷. On constate aussi que la balance énergétique du processus n'est pas aussi favorable qu'on pourrait le penser, puisqu'elle est issue de phénomènes thermodynamiquement antagonistes. D'autre part, des expériences ont montré que le repliement s'apparente plus à une transition de phase qu'à un équilibre continu³³.

1.2.2 Information égale énergie

Dans ce travail, nous ferons table rase de toutes ces notions. Premièrement, l'état actuel de la théorie en biophysique ne nous permet pas de modéliser un champ de force qui permettrait de recréer l'environnement d'un bout à l'autre du repliement. En pratique, lorsqu'on utilise les champs de force les plus précis de mécanique moléculaire^{34 35}, on passe une bonne partie de l'expérience à tenter de stabiliser les structures tri-dimensionnelles obtenues de la cristallographie. En effet, les structures sont instables si on ne passe pas par une période de chauffage-équilibrage³⁶. Ces champs de force sont pourtant conçus spécifiquement pour maintenir les structures dans une position biologique. Autre inconvénient majeur, même avec un champ de force correct en dynamique moléculaire, le temps de la simulation serait 6 ordres de grandeur trop petit par rapport au phénomène réel. En effet, les simulations les plus efficaces actuellement font quelques nanosecondes³⁷, tandis que le repliement des protéines prend entre quelques millisecondes et une seconde. Ce problème d'échelle du temps est intrinsèque à la dynamique moléculaire, qui doit calculer itérativement dans un très petit intervalle de temps (10^{-15} sec.) le résultat des équations de mouvement, si elle veut conserver le caractère réversible de sa trajectoire³⁸.

Deuxièmement, le temps de calcul et l'espace disque nécessaires à une évaluation exhaustive de tout l'espace conformationnel avec des champs de force compliqués rendent, d'une façon pratique, le problème incommensurable.

Pour arriver à prédire une structure de protéine malgré cette barrière physique, nous devons prendre le problème sous une tout autre perspective. Nous devons

visualiser l'énergie comme une mesure de la qualité de la prédiction, en transformant toute l'information que nous connaissons d'une protéine en énergie. Le mot énergie devient pour nous synonyme d'indicateur de bon repliement. Toutes les approches *ab initio* ont jusqu'à maintenant échoué. La réalisation à court terme d'un logiciel qui serait capable de prédire la structure tertiaire ne peut être envisagée qu'en intégrant des données expérimentales. Le développement à moyen terme de TRIP suit cette ligne de pensée.

1.2.3 Notion générale d'un potentiel statistique

La plus grosse part de l'énergie que nous utilisons³⁹ pour identifier les conformations est un potentiel statistique tiré des structures cristallographiques⁴⁰⁴¹. L'idée étant de partir de la distribution des acides aminés dans les protéines connues, et d'en tirer un potentiel de forces moyennes. Les bases du raisonnement étant: 1) on ne connaît pas les vraies forces qui régissent le repliement, 2) mais nous avons des structures tri-dimensionnelles de protéines à l'état replié. 3) Nous pouvons donc nous passer de connaître les forces véritables du repliement, 4) en mimant la distribution spatiale des atomes dans les structures connues. Ce potentiel statistique contiendra donc implicitement toutes les forces physiques du repliement.

Nous nous trouvons à synthétiser dans ce potentiel statistique un grand nombre d'informations sur les protéines. Plus précisément, ce potentiel renferme les grandes tendances d'une protéine repliée.

1.2.3.1 Défauts des potentiels statistiques

Plusieurs défauts existent avec cette approche. D'abord ceux liés à la base de données. Pour construire un potentiel statistique, on doit tout d'abord choisir un sous-ensemble de structures représentatives de l'ensemble des protéines connues. Cette échantillonnage de structures cristallographiques de protéines va, à la base, restreindre le champ de nos prédictions. Des biaisages de longueur, de grosseur, de type d'environnement au niveau de l'ensemble échantillon vont

directement transformer la nature du potentiel statistique. Par exemple, le potentiel Bryant&Lawrence⁴² qui a servi pour tout ce travail comprend des protéines de 60 à 200 acides aminés, globulaires, en phase aqueuse.

Jusqu'à récemment, les biochimistes étaient limités à l'étude des protéines solubles. Ces protéines ont l'avantage d'être purifiables facilement et de conserver leur activité biologique après isolement. Nous sommes alors capables de mener nos expériences sur des spécimens purifiés. Il est donc normal que la majorité de nos connaissances sur le repliement des protéines portent sur ces protéines qui naviguent dans l'eau. Les protéines membranaires ne fonctionnent pas expérimentalement sans des bicouches lipidiques. Malheureusement, ces bicouches nous empêchent d'utiliser plusieurs techniques expérimentales. Ce fait a plusieurs implications, dont le principal est le choix de protéines qu'ont fait Bryant&Lawrence. C'est l'origine de la limite dont il était mention plus haut.

Deuxièmement, le type d'analyse statistique utilisé, et surtout le choix de la référence, influent sur la qualité du potentiel. Le problème ici, c'est qu'on ne connaît pas de système de référence vraiment convaincant ⁴³. Troisièmement, le genre d'informations que nous tirons des données cristallographiques, comme des distances moyennes entre les acides aminés, ou des torsions dans l'angle dièdre du lien peptidique, peut affecter la qualité du potentiel. En effet, certains genres d'informations donnent des patrons de données moins significatifs que d'autres⁴⁴.

Quatrièmement, les potentiels statistiques ne sont pas a priori excellents pour des algorithmes de minimisation. Mesurer une ressemblance entre deux distributions est une chose, mais maximiser cette ressemblance en est une autre.

Malgré ces graves limitations, les potentiels statistiques demeurent les potentiels les plus rapidement calculables, une qualité importante dans une approche itérative comme la nôtre.

1.3 Le problème de l'explosion combinatoire dans une prédiction de structure protéique tri-dimensionnelle

Parlons maintenant du problème de l'explosion combinatoire. Pour utiliser la bonne terminologie, le repliement des protéines fait partie de la classe des systèmes NP (non-deterministic polynomial time). En effet, une protéine est un système intrinsèquement hétérogène (les acides aminés très différents), avec plusieurs degrés de liberté fortement corrélés. Dans cette classe de systèmes, les conformations possibles croissent exponentiellement par rapport à N , le nombre de degrés de liberté.

Dans TRIP, nous avons adopté la voie d'une génération aléatoire de structures 3D. L'algorithme augmente donc ses chances de tomber sur la structure native s'il essaie le plus de conformations possibles. Le problème survient lorsque le champ des possibles dépasse largement le champ visité. Prenons par exemple un modèle vraiment réduit des protéines, où les seules parties mobiles sur chaque acide aminé sont les angles dièdres autour du $C\alpha$. Admettons une liste finie de dix choix par acide aminé. Pour une protéine de cent acides aminés, ça nous fait donc 10^{100} possibilités. Comme un calcul typique de prédiction génère dans les 10^{10} possibilités, nous sommes loin du compte par plusieurs ordres de grandeur. C'est ici que les chercheurs peuvent intervenir. Toute réduction du champ des possibles augmentera le ratio des conformations testées versus inconnues, et donc la qualité de la prédiction, dans l'approximation où notre potentiel est bon. Le jeu ici étant à la fois de créer un algorithme «intelligent», qui se guidera dans le champ des possibles, et de tirer avantage de toutes les informations que l'on connaît de la protéine à prédire afin de réduire le champ des possibles.

1.3.1 Le principe hiérarchique comme méthode pour jongler avec l'explosion combinatoire

Une façon de réduire le champ des possibles est d'adopter une vision hiérarchique de la recherche conformationnelle. Dans une telle vision, nous voulons briser l'explosion combinatoire en plusieurs étapes plus facilement manipulables. D'une

façon mathématique,

$$m^N \mapsto ne^{\frac{N}{n}} + M^n \quad (1.1)$$

où m est le nombre de choix pour un élément, N est le nombre d'éléments, n est le nombre de petites étapes et M le nombre de façons de combiner les différentes étapes. Le gain provient de la possibilité de réduire M , afin de le rendre beaucoup plus petit que $e^{\frac{N}{n}}$. Deux facteurs sont importants dans la réduction de M : la génération d'essai et le tamisage.

Concrètement dans notre modèle, nous divisons la protéine en plusieurs fragments qui s'imbriquent les uns dans les autres (p. 28). À chaque niveau, nous procédons à une recherche conformationnelle. Mais les essais d'un niveau supérieur se font à partir des choix effectués dans le niveau inférieur. Pour une itération donnée, nous diminuons ainsi le champ des possibles à couvrir, mais explorons une plus grande proportion de ce champ.

1.3.2 Le biaisage des essais, pour freiner l'explosion combinatoire

Intuitivement, pour juguler l'explosion combinatoire, il suffit de limiter le nombre de choix possibles. Nous pouvons par exemple limiter le nombre de choix possibles d'une variable donnée en lui assignant une liste finie de choix, plutôt que de la traiter avec une fonction continue. Cette liste devrait alors refléter une donnée expérimentale ou un choix du scientifique.

Couplé avec un algorithme hiérarchique, il devient possible de créer des listes finies de choix pour tous les niveaux de recherche. Si en plus le choix des nouveaux essais au niveau le plus bas provient du niveau le plus haut, nous nous retrouvons avec un système auto-cohérent. La recherche conformationnelle n'est donc plus la simple génération au hasard de nouvelles structures, mais une suite de mutations cohérentes par rapport à un potentiel donné.

Finalement, pour décupler la puissance de recherche, la mise en oeuvre d'un algorithme génétique s'avère un choix extrêmement puissant⁴⁵. Lorsque nous avons un problème informatique très complexe à résoudre, mais où la réponse

finale peut être divisée en plusieurs parties, un algorithme génétique est capable de mélanger les parties de solutions ensemble, pour créer des hybrides. Ainsi, de bonnes réponses locales prises à gauche et à droite finissent par devenir la réponse finale souhaitée. Bien sûr, ce ne sont pas tous les genres de problèmes qui peuvent bénéficier d'un algorithme génétique, mais dans le cas des protéines, il est difficile de ne pas l'avoir en tête. Après tout, la génétique est la science qui étudie les gènes et leurs mutations. Plusieurs études tendent à prouver la supériorité de l'approche génétique face à l'approche Monte Carlo seul^{46 47}.

1.4 Le problème des minima multiples dans la fonction d'énergie

Pour compliquer davantage la recherche conformationnelle, le paysage énergétique typique d'une protéine ressemble plus à un terrain de golf qu'à un entonnoir. Dans un tel contexte, avoir l'assurance qu'on se trouve dans le puits de potentiel le plus bas, alors même qu'on n'a pas échantillonné le dixième de la surface accessible, ça devient douteux. On comprend qu'un golfeur aveugle parte avec un gros handicap. On englobe ce problème sous le vocable: le problème des minima multiples. Plusieurs domaines scientifiques font face à ce genre d'obstacle. De nombreuses méthodes ont été développées pour juguler le problème. On peut nommer rapidement:

L'approche Monte Carlo a été développée dans les années 50^{48 49}. Beaucoup des méthodes qui jonglent avec le problème des minima multiples ont donné des fruits dans le problème du repliement des protéines⁵⁰. Pour ce qui est de l'approche Monte Carlo, plusieurs y ont vu l'avantage d'un algorithme simple mais puissant⁵¹
52 53.

1.4.1 Le recuit simulé comme méthode pour jongler avec les minima multiples

Dans ce travail, nous avons choisi le recuit simulé (simulated annealing)⁵⁴ pour minimiser l'énergie. Cet algorithme consiste à placer le système en contact avec un bain de chaleur ayant une température élevée et à graduellement abaisser cette

«température». Les guillemets autour de la température interne veulent souligner le fait qu'elle ne correspond pas à une vraie température physique. C'est une variable interne à l'algorithme, pas au système à l'étude. Dans ce travail, nous abaissons exponentiellement la température du recuit selon:

$$T_k = (T_{init}^{N-k} T_{fin}^k)^{1/N} = T_{init} e^{-Ak} = e^{-A} T_{k-1}, \quad (1.2)$$

où $A = \frac{1}{N} \log(T_{init}/T_{fin})$, N est le nombre total d'itérations, k est l'itération courante et T est la température.

Du point de vue de l'échantillon, une «température» basse diminue «statistiquement» le nombre de choix accessibles. Ainsi, en abaissant la «température», nous transformons le paysage énergétique de telle façon qu'on guide l'ensemble échantillon vers les minima.

1.4.2 L'algorithme Monte Carlo, pour échantillonner une surface d'énergie

Pour implanter le recuit simulé, nous nous servons de l'algorithme Monte Carlo. Grosso modo, l'algorithme Monte Carlo se résume à accepter ou rejeter un essai selon un critère énergétique (Métropolis).

Ainsi, la probabilité qu'une protéine soit dans une conformation i correspond à une distribution de Boltzmann,

$$p_i \propto e^{-\beta E_i}, \text{ où } \beta \text{ est } 1/k_B T \quad (1.3)$$

où i est l'état de la conformation, p_i est la probabilité de cet état et E_i est l'énergie de cet état. À l'équilibre, nous pouvons poser comme condition que la probabilité de passer d'une conformation j à une conformation i , $P_{j \rightarrow i}$, à la relation suivante avec $P_{i \rightarrow j}$,

$$p_j P_{j \rightarrow i} = p_i P_{i \rightarrow j} \quad (1.4)$$

Si on introduit l'équation 1.3 dans la 1.4, nous obtenons:

$$\frac{P_{j \rightarrow i}}{P_{i \rightarrow j}} = e^{-\beta(E_i - E_j)} \quad (1.5)$$

Comme nous désirons un chiffre pour la probabilité de changer d'état, Metropolis a choisi de maximiser le nombre de changements possibles en mettant à 1 la probabilité de passer d'un état plus énergétique à un état moins énergétique. Ceci laisse l'autre probabilité égale à $e^{-\beta(E_i - E_j)}$, si $E_i > E_j$.

Les points distinctifs sont que la génération d'un nouvel essai s'appuie uniquement sur l'essai précédent (créant une chaîne de Markov) et que la décision d'accepter un nouvel essai respecte une distribution statistique de l'énergie⁵⁵ (ici dans l'équation 1.3, la distribution de Boltzman). La base du raisonnement étant qu'un petit ensemble échantillon qui se promène assez longtemps sur la surface d'énergie en suivant cette règle finira par représenter thermodynamiquement l'espace de phase.

1.4.2.1 Application aux protéines de l'algorithme Monte Carlo

L'algorithme Monte Carlo de base est très simple, mais demeure inutilisable dans le domaine des prédictions de protéines. En effet, l'algorithme Monte Carlo a été développé pour échantillonner de façon biaisée un large espace de phase, non pour trouver un minimum global. Comme la conformation d'une protéine ne correspond pas à un ensemble vaste de configurations, mais plutôt à un petit ensemble dans l'espace de phase, il faut modifier le paysage énergétique de celle-ci pour forcer l'ensemble Monte Carlo à trouver ce petit espace de phase. Plusieurs solutions ont été mises au point au cours des années⁵⁰. Dans ce travail, nous nous sommes contentés de l'algorithme de recuit simulé seul.

1.4.3 Frictions entre Monte Carlo et recuit simulé

Il existe cependant un inconvénient de l'algorithme de recuit simulé lorsqu'il est couplé à un Monte Carlo. Lorsqu'une distribution est à l'équilibre, la proportion de la population dans chaque puits d'énergie respecte une distribution de Boltzmann de l'énergie, et chaque saut d'une particule d'une position à une autre respecte l'équilibre. C'est ce qu'on appelle le bilan détaillé («detailed balance»). L'algorithme de recuit simulé détruit l'équilibre du bilan détaillé. Avec un

Monte Carlo seul, nous sommes dans une situation où l'ensemble échantillon est en équilibre dans l'espace de phase. Les moyennes des observables (énergie, spin, etc.) calculées grâce à l'ensemble échantillon correspondent aux moyennes véritables du système, à cette température fictive. Avec le recuit simulé, cette affirmation n'est plus nécessairement vraie. Tout dépend du temps qu'on laisse à l'ensemble échantillon pour s'équilibrer après un abaissement de «température». Un refroidissement rapide brise l'équilibre et précipite l'ensemble échantillon vers les minimums les plus accessibles, sans égard à la distribution d'énergie. D'une certaine façon, ce handicap n'est pas vraiment contraignant, puisque l'utilité du recuit simulé est justement de trouver les minimums d'une surface.

Il reste cependant que l'ensemble échantillon doit alors pouvoir couvrir une large portion de la surface d'énergie pour être sûr de visiter chaque minimum. À cette fin, nous pouvons soit ralentir la baisse de température, soit augmenter le nombre d'échantillons dans l'ensemble. Dans ce travail, mon ensemble échantillon comprenait 64 protéines.

1.5 Quelques notions de RMS, pour calculer la ressemblance entre deux structures

Durant ce travail, mes principaux résultats sont des structures générées par ordinateur. Ces structures sont l'aboutissement d'une recherche conformationnelle, et constituent la «réponse» de notre modèle protéique. Cette «réponse» est modulée par l'énergie de la structure, l'espace conformationnel accessible et les paramètres de départ. Pour juger de la qualité de la «réponse», nous disposons des structures cristallographiques. Grâce à celles-ci, nous pouvons comparer les structures générées par ordinateur et la «vraie structure». Un problème survient lorsque nous voulons jauger la qualité de la prédiction d'une façon quantitative. Nous avons besoin d'une méthode claire et uniforme pour classer par un chiffre le degré de ressemblance d'une structure générée par rapport à la «native» (structure cristallographique).

L'écart-type (Root mean square deviation) est une mesure mathématique de la

ressemblance entre deux ensembles de points. S'il y a des correspondances entre les points des deux ensembles, nous pouvons alors calculer une moyenne qui utilise ces correspondances pour nous donner un indicateur de la ressemblance générale des deux ensembles. Ainsi, pour deux protéines, le fait de calculer la RMS avec les distances provenant des atomes correspondants constitue un indicateur de la similarité entre ces deux protéines. C'est une mesure de la différence structurale. Dans ce travail, il y a deux sortes de RMS. Une RMS qu'on définit par distance et une autre qu'on définit par rotation. La première se calcule comme suit:

$$\Delta d = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (d_{ij} - e_{ij})^2 / n^2} \quad (1.6)$$

$$d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{\frac{1}{2}}$$

où d_{ij} sont les distances interatomiques entre les atomes dans la structure cristalline et e_{ij} sont les distances interatomiques entre les atomes de la structure générée par ordinateur⁵⁶.

La deuxième se décline comme suit:

$$\Delta r = \sqrt{\sum_{i=1}^n (x_i - y_i')^2 / n} \quad (1.7)$$

où les x_i sont les coordonnées de la structure cristallographique et les y_i sont les coordonnées de la structure générée, n étant le nombre d'atomes.

Nous utilisons une version analytique du RMS par rotation⁵⁷. Une correction apportée en 1978 par Kabsch⁵⁸ règle des cas limite où la rotation devenait impropre. Dans ce travail, en l'absence de remarques, la RMS par rotation, comme proposé dans ce dernier article, sera utilisée par défaut.

Cohen et Sternberg ont montré qu'il y a une relation approximativement linéaire entre ces deux mesures.

$$\Delta d = 0.75\Delta r + 0.19 \quad (1.8)$$

Dans le cas de petites molécules, la RMS est un indicateur simple et univoque de la ressemblance entre deux entités. Pour les protéines, qui sont constituées de plusieurs centaines d'atomes, il existe un certain seuil au delà duquel la RMS ne donne plus un portrait juste. Le problème des protéines est qu'elles

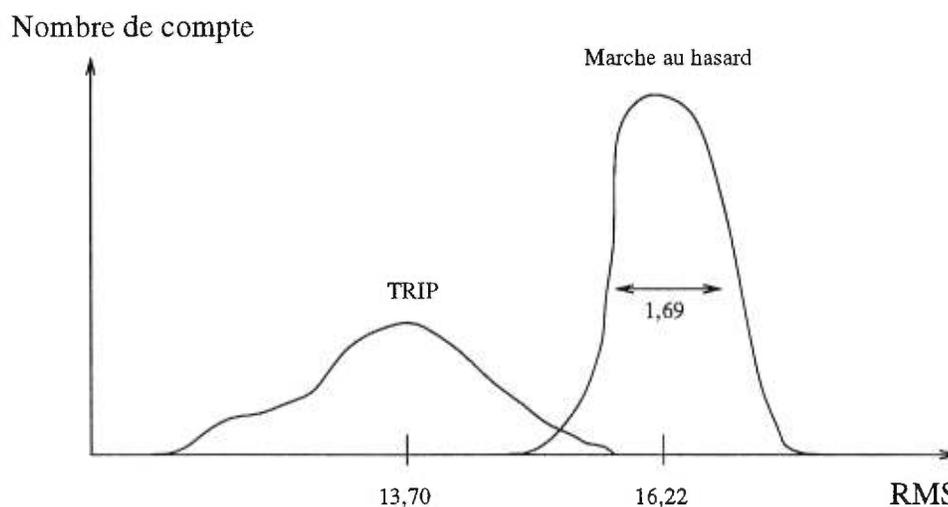


FIGURE 6. Distribution moyenne du RMS pour des structures compactes générées au hasard et des structures générées par TRIP, ainsi que la déviation moyenne pour les structures générées au hasard.

sont globulaires. Ainsi, pour une protéine d'environ 150 acides aminés comme la myoglobine, la distance maximum entre deux atomes tourne autour de 40Å. Ceci implique que des structures compactes, créées au hasard mais sans chevauchement, ont un plateau dans leur valeur de RMS. La relation entre la longueur d'une protéine (N) et la RMS (entre structure native et structures compactes générées par marche au hasard) est⁵⁶:

$$\Delta r = 0.0468 (N) + 9.25 \quad (1.9)$$

Ceci implique que le pire algorithme de prédiction donnerait des structures de 16,22 Å RMS pour la myoglobine. Même en lui donnant des hélices pré-faites, nous n'abaisserions que de 0.8 Å la valeur de la moyenne des RMS.

En portant sur une gaussienne la distribution des valeurs RMS de structures générées(fig. 6), on s'aperçoit que la déviation standard à la moyenne est relativement petite. Ceci pour dire que plus on s'éloigne de la moyenne, moins il y a de structures possibles.

CHAPITRE 2

Le modèle

Dans ce travail, je n'ai pas mis au point l'algorithme principal de prédiction, mais je trouve important d'en tracer les grandes lignes pour justifier mon approche. Je me suis servi du logiciel TRIP⁴⁵. Celui-ci traduit une vision très hiérarchique d'une recherche conformationnelle, en mettant l'accent sur le biaisage des essais Monte Carlo.

Une image très simple de ce que TRIP fait est le modèle des ressorts(p. 13). Dans ce modèle, on considère les structures secondaires comme des cylindres rigides, qui ne change pas durant le calcul. TRIP essaye d'échantillonner les conformations possibles pour les acides aminés dans les boucles afin de placer les cylindres correctement dans l'espace. Si on aide l'échantillonnage des boucles, on aide TRIP à faire des bonnes prédictions.

Il y a deux parties principales dans ce logiciel. La première concerne la construction de la protéine, la deuxième concerne le classement des structures selon un critère énergétique. Dans la version de TRIP qui a servi pour ce travail, on peut dire que la construction se fait dans l'espace dièdre et le classement dans l'espace cartésien (3D).

2.1 Construction de la protéine

Dans TRIP, nous construisons des essais à partir de listes finies de choix. Nous avons des listes d'angles ϕ et ψ pour chaque acide aminé, des listes de triplets d'acides aminés, des listes de segments de triplets correspondant aux boucles et finalement des listes de structures complètes. En biaisant la composition d'une liste de façon cohérente, nous multiplions la puissance de l'approche hiérarchique. Nous allons maintenant expliquer comment nous construisons une protéine à

partir des listes.

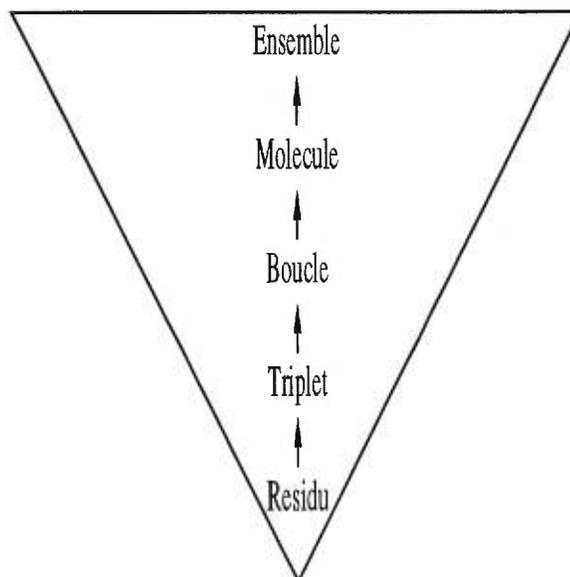


FIGURE 7. Pyramide de la construction d'une protéine dans TRIP.

2.1.1 Définition de l'espace conformationnel

TRIP utilise un modèle réduit des protéines. Toutes les distances (tableau I) et presque tous les angles (tableau II) sont identiques et fixes pour tous les acides aminés et ce, tout le temps de la recherche conformationnelle. Les seuls angles mobiles sont les angles dièdre de la chaîne principale autour du carbone α . Toutes les liaisons peptidiques sont planes. Il est à remarquer que dans la version de TRIP qui a servi pour ce travail, les prolines n'étaient pas cycliques. Leurs chaînes principales sont identiques aux autres. De plus, pour réduire encore davantage la représentation protéique, les chaînes latérales de tous les acides aminés se résument au carbone β . Nous transformons même un des hydrogènes de la glycine en carbone β . Ainsi, la seule différence entre les acides aminés dans TRIP est dans le calcul d'énergie. Ce n'est que dans le calcul d'énergie qu'il y a discrimination entre les différents types d'acide aminé (fig. 1). Une note importante est qu'il n'y a pas de calcul explicite du volume. Celui-ci est intégré dans le potentiel

statistique(p. 35).

TABLEAU I. Distances intramoléculaires des acides aminés dans TRIP.

Atomes	Distances (Å)
$C_\alpha - C'$	1,52
$C' - N$	1,33
$N - C_\alpha$	1,45
$C_\alpha - C_\beta$	1,54
$C' - O$	1,23
$N - H$	1,02

Cette représentation des protéines est à cheval sur une vision minimaliste sur réseau³¹ et une vision tout atomes³⁴. D'une part, conjugué à nos listes finies, nous réduisons l'espace conformationnel accessible à la dimension d'un très gros réseau, sans pour autant être capable de visiter toutes les possibilités. D'autre part, les structures prédites par TRIP sont tri-dimensionnelles et lisibles directement par des logiciels de modélisation moléculaire. Nous sommes à l'interface entre une recherche exhaustive d'un petit espace conformationnel, et une modélisation moléculaire tout atomes.

2.1.2 L'acide aminé, la sous-unité dans TRIP

L'acide aminé n'est pas l'unité de base pour les mutations dans TRIP. C'est pour cela que nous le voyons comme une sous-unité.

Chaque acide aminé est constitué de six atomes: l'azote et son hydrogène, le carbone α , le carbone β , le carbone du carbonyle et l'oxygène. Comme mentionné plus haut, il n'y a que deux degrés de liberté géométrique pour les acides aminés dans TRIP. Ce sont les deux angles dièdres ϕ et ψ . Dans ma version de TRIP, chaque acide aminé a accès à sa propre liste finie de choix d'angles dièdres. Ainsi, nous pouvons définir une carte de Ramachandran identique pour tous les résidus

TABLEAU II. Angles intermoléculaires des acides aminés dans TRIP.

Atomes	Angles
$N - C_{\alpha} - C'$	109,5°
$C_{\alpha} - C' - N$	115,6°
$C' - N - C_{\alpha}$	121,9°
$C_{\alpha} - C' - O$	121,1°
$C_{\alpha} - N - H$	119,0°
$C_{\beta} - C_{\alpha} - C'$	109,5°
$C_{\beta} - C_{\alpha} - C' - N$	120,0°
angle dièdre	

ou une pour chaque type d'acide aminé, ou même utiliser une carte différente pour chaque résidu.

Dans TRIP, pour garder les structures secondaires rigides, on donne aux acides aminés qui composent les structures secondaires une liste de un seul couple d'angle dièdres.

2.1.3 La liste de triplets, l'unité de base de la mutation dans TRIP

Le triplet est constitué de trois acides aminés. Il est l'unité de base pour les mutations. Ceci veut dire que la plus petite différence entre deux structures dans TRIP est un triplet muté.

TRIP doit générer des millions de triplets durant un calcul, pour échantillonner l'espace des conformations. Pour construire un segment peptidique dans l'espace, TRIP utilise un système d'orientation absolue, **U**. Pour chaque carbone α ,

$$\mathbf{U} = \begin{pmatrix} \hat{\mathbf{r}}_{N-C\alpha} \\ \frac{\hat{\mathbf{r}}_{N-H} - \cos \theta \hat{\mathbf{r}}_{N-C\alpha}}{\sin \theta} \\ \frac{\hat{\mathbf{r}}_{N-C\alpha} \times \hat{\mathbf{r}}_{N-H}}{\sin \theta} \end{pmatrix}, \quad (2.1)$$

où $\cos \theta = \hat{\mathbf{r}}_{N-C\alpha} \cdot \hat{\mathbf{r}}_{N-H}$

La construction de la chaîne principale va de carbone α en carbone α en calculant le \mathbf{U} du carbone α suivant grâce à ϕ et ψ . Si \mathbf{T} représente la rotation d'un résidu par rapport à son voisin, on obtient de façon générale

$$\mathbf{U}_{i+1} = \mathbf{U}_i \mathbf{T}_i = \mathbf{T}_1 \mathbf{T}_2 \cdots \mathbf{T}_i \quad (2.2)$$

où $\mathbf{U}_1 = \mathbf{I}$. \mathbf{T} est une fonction calculable à partir des tableaux I et II ainsi que ϕ et ψ ⁵⁹.

Pour créer un essai, nous générons tout d'abord des milliers de triplets, que nous classons ensuite dans des listes de triplets (origine du nom TRIP) selon un critère géométrique. La génération des triplets est aléatoire, mais utilise l'espace dièdre disponible pour chaque acide aminé. Pour le classement des triplets, nous nous servons d'un ensemble de cinq indices géométriques, que nous divisons en intervalles réguliers. Les cinq indices sont:

$$\begin{aligned} q_1 &= \cos^{-1}(\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{r}}) \\ q_2 &= \cos^{-1}(\hat{\mathbf{x}}_2 \cdot \hat{\mathbf{r}}) \\ |q_3| &= \cos^{-1} \left(\frac{\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{x}}_2 - (\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{r}})(\hat{\mathbf{x}}_2 \cdot \hat{\mathbf{r}})}{\sin q_1 \sin q_2} \right), \\ |q_4| &= \cos^{-1} \left(\frac{\hat{\mathbf{y}}_1 \cdot \hat{\mathbf{r}}}{\sin q_1} \right) \\ |q_5| &= \cos^{-1} \left(\frac{\hat{\mathbf{y}}_2 \cdot \hat{\mathbf{r}}}{\sin q_2} \right) \end{aligned} \quad (2.3)$$

où $\{\hat{\mathbf{x}}_1, \hat{\mathbf{y}}_1, \hat{\mathbf{z}}_1\}$ sont les rangées de \mathbf{U}_i , $\{\hat{\mathbf{x}}_2, \hat{\mathbf{y}}_2, \hat{\mathbf{z}}_2\}$ sont les rangées de \mathbf{U}_{i+3} , et $\hat{\mathbf{r}}$ est l'orientation du vecteur $\mathbf{r}_{C\alpha, i+3} - \mathbf{r}_{C\alpha, i}$. Ces coordonnées vont de $0 \leq q_{1,2} \leq \pi$ et $-\pi \leq q_{3,4,5} \leq \pi$ où le signe des dièdres $q_{3,4,5}$ est donné par la relation

$$\frac{q_i}{|q_i|} = \frac{s_i}{|s_i|} \quad (2.4)$$

où

$$\begin{aligned} s_1 &= \hat{\mathbf{r}} \cdot (\hat{\mathbf{x}}_1 \times \hat{\mathbf{x}}_2) \\ s_2 &= \hat{\mathbf{r}} \cdot \hat{\mathbf{z}}_1 \\ s_3 &= \hat{\mathbf{r}} \cdot \hat{\mathbf{z}}_2 \end{aligned} \tag{2.5}$$

sont les triples produits correspondants.

Lorsque vient le temps de faire un essai, TRIP identifie au hasard un triplet à changer. TRIP calcule ensuite les cinq indices pour ce triplet. Le choix d'un nouveau triplet se fera dans la même tranche (v.f. de «bin») que le triplet actuel. Le but de remplacer un triplet par un autre qui lui ressemble étant de faire des petits pas dans l'espace conformationnel.

2.1.4 La liste de boucles, un ensemble qui évolue

Deuxièmement, nous construisons une liste de substituts pour chaque boucle. Une boucle est tout ce qui existe entre deux structures secondaires. Une boucle est constituée d'une suite de triplets qui se chevauchent. Nous classons de nouveau géométriquement les boucles, avec les mêmes cinq indices (éq. 2.3). Il est à noter que nous construisons cette liste à partir des boucles déjà existantes dans l'ensemble. Plus spécifiquement, pour une boucle dans l'échantillon, plusieurs boucles substituts sont créées en changeant un seul triplet à la fois. On s'assure toutefois que les boucles ajoutées à la liste sont à la fois différentes des échantillons de l'ensemble et différentes de ceux dans la liste. Ici, lorsqu'un essai est fait, la liste est unique pour tous les échantillons de l'ensemble. Ainsi, nous pouvons changer plusieurs triplets d'un coup, en choisissant un substitut qui vient d'un autre échantillon. Le filtre spécifique dans le traitement de cette liste est que nous modifions la population dans chaque tranche géométrique selon la population qui existe dans l'ensemble échantillon de structures. Nous nous trouvons du même coup à éliminer d'un trait beaucoup de possibilités farfelues, et à peupler les tranches dont on a vraiment besoin. Il y a donc un biaisage géométrique supplémentaire à cette étape.

2.1.5 La génération de molécules, l'étape finale avant la recombinaison

Troisièmement, nous construisons une molécule en choisissant dans chaque liste de boucles, un essai. Comme les structures secondaires sont rigides, nous obtenons une protéine complète.

Chaque molécule dans l'ensemble évolue d'une façon itérative (Monte Carlo) en essayant différentes boucles de la liste. Deux genres d'essais existent: une seule boucle à la fois ou un segment boucle-structure secondaire-boucle.

Ici, le filtrage est uniquement énergétique. Une fois qu'un choix a été fait, on construit la molécule à partir du début. Ensuite, on évalue l'énergie de cette molécule via une succession de fonctions de potentiel de plus en plus précise. Si la molécule est bloquée dans une de ces fonctions de potentiel, elle est automatiquement rejetée. Ultiment, si la molécule passe à travers tous les tamis énergétiques grossiers (coarse-grained potential), l'acceptation ou le rejet de l'essai est contrôlé par Monte Carlo-recuit simulé. C'est donc à ce niveau que la qualité de la fonction de potentiel se fera le plus sentir.

2.1.6 L'algorithme génétique, pour recombinaison des bonnes solutions éparpillées

Finalement, de façon systématique, un algorithme génétique crée des hybrides à partir des échantillons de l'ensemble. TRIP fait alors évoluer chacun de ces hybrides avec les listes de boucles. L'ensemble des hybrides est beaucoup plus grand (512 individus) que l'ensemble des parents (64 individus). Puis, il sélectionne de façon énergétique un nouvel ensemble de parents. Les nouveaux parents doivent à la fois avoir un lien de parenté avec les anciens parents, être différents des autres nouveaux parents et être bas en énergie.

En résumé:

Monte Carlo des parents → Création des hybrides → Monte Carlo des hybrides
→ Sélection des nouveaux parents.

2.2 La fonction de potentiel utilisé dans ce travail

Il y a plusieurs termes à la fonction de potentiel. Chaque terme possède ses propres paramètres qui font fluctuer le poids de ce terme dans la valeur de l'énergie finale.

2.2.1 Force hydrophobe

Tel qu'expliqué à la section 1.2.1, la force hydrophobe est celle qui compacte la protéine. Dans TRIP, nous avons besoin d'une force qui compacte la protéine aussi. Pour ce faire, nous calculons un terme proportionnel au rayon de giration. Plus la protéine est dépliée, plus l'énergie monte (défavorable). Durant la minimisation, nous faisons décroître linéairement ce terme hydrophobe. En commençant avec une valeur élevée, nous forçons la protéine à se replier. Une fois la minimisation bien entamée, le potentiel statistique est capable de maintenir les protéines repliées sans l'aide de l'énergie hydrophobe.

2.2.2 Potentiel statistique

Pour ce travail, une version tabulée à 4 indices du potentiel de Bryant&Lawrence a été utilisée. Les quatre indices sont les distances entre les carbones α et β de chaque acide aminé. On entend par tabulé, qu'il utilise une immense table qui, pour des intervalles de distances discrétisées, donne une valeur pour un couple d'acides aminés.

Plus subtilement, on obtient la table en testant des milliers de conformations différentes d'un dipeptide qui emploie notre modèle atomique. On place un deuxième dipeptide à une certaine distance et dans une certaine orientation par rapport au premier. En calculant l'énergie Bryant&Lawrence directement, on comptabilise l'énergie pour la table à 4 indices. Comme nous testons des millions de possibilités, nous couvrons presque entièrement les différents cas qui pourraient survenir dans TRIP. Le réel gain de cette méthode est au niveau du temps de calcul. Il est beaucoup plus court de lire dans une table que de calculer une fonction d'énergie.

2.2.3 Volume exclu

Pendant la comptabilisation des énergies pour les différentes conformations, les distances entre les $C\alpha$ et $C\beta$ de chaque dipeptide sont comparées avec une valeur plancher. Si la distance est inférieure à cette valeur plancher, on omet l'énergie pour cette conformation dans la table de potentiel.

Ensuite, lorsque TRIP calcule l'énergie, toutes les distances qui ne possèdent pas de valeur d'énergie dans la table de potentiel se font assigner une valeur par défaut. Cette valeur est positive, ce qui pénalise d'autant la molécule. Le calcul du volume exclu est donc implicite dans le calcul d'énergie.

Il est à remarquer que le fait d'utiliser notre modèle atomique pour créer les dipeptides implique qu'il n'y a pas de chaînes latérales. Notre calcul du volume exclu ne mime pas la présence d'une chaîne latérale, excepté le $C\beta$.

2.3 Généralité sur les niveaux de filtrage

Pour chercher efficacement dans l'espace conformationnel, l'utilisation de toute l'information connue sur la protéine à prédire est très avantageuse. Il faut cependant jauger la pertinence d'une information, pour savoir jusqu'à quel point elle nous est utile. En fonction de cette évaluation, nous devons décider comment l'utiliser dans TRIP. Par son approche hiérarchique, TRIP nous permet avec chaque information de construire des filtres au niveau pertinent de construction. On comprend que des interactions locales entre acides aminés qui se suivent se traitent mieux au niveau des listes de triplets qu'au niveau des molécules.

De plus, à chaque niveau de la construction, deux avenues s'offrent à nous. Nous pouvons soit directement manipuler l'espace accessible, soit créer un potentiel qui exercera une pression sur cet espace. Par exemple, une manipulation directe pourrait être un filtrage systématique de toutes les structures qui n'auraient pas telle ou telle caractéristique. Une manipulation par potentiel interposé donnerait des malus (pénalité d'énergie) ou des bonus d'énergie à des structures qui tendraient ou non vers l'information connue.

La modification d'un espace de construction finit par affecter toute la pyramide de construction. On comprend la voie ascendante. Un changement dans un espace inférieur de construction se répercutera partout dans la chaîne. Mais on a aussi dans TRIP une voie descendante: un changement dans un niveau supérieur de construction a un impact dans le niveau inférieur de construction. Par exemple, la liste de boucles est biaisée vers des structures présentes dans l'ensemble. Donc, indirectement, la fonction de potentiel exerce une pression dans la liste de boucles, vu qu'elle modifie la population de l'ensemble.

On doit donc doser l'élimination de certaines possibilités versus la création de tendance durant la minimisation selon notre confiance dans les différentes sources d'information.

CHAPITRE 3

Problèmes locaux

Le point de départ chronologique de ma recherche a été un résultat du programme PROCHECK⁶⁰.

Le tableau III répertorie 37 mauvais contacts se produisant entre les 146 acides aminés de la 1MBO. De ce nombre, 25 proviennent d'acides aminés voisins d'au plus 4 acides aminés, dont 18 concernent la chaîne principale uniquement. J'en ai déduit qu'il y avait des lacunes au niveau local dans la version 0.2 de TRIP.

Une source possible du problème est l'absence de force explicite au niveau local dans cette version de TRIP. Le potentiel statistique donne un malus lorsque deux atomes sont trop rapprochés, mais il ne l'empêche pas. Nous n'avons pas investi du temps dans une force locale, car nous voulons donner le maximum de flexibilité à la recherche conformationnelle. Cependant, nous sommes conscients que l'absence de force locale à travers toute la minimisation n'est peut-être pas optimale pour empêcher l'encombrement stérique.

Deuxième problème local, une carte de Ramachandran de l'utilisation des angles dièdres dans les boucles me montrait que TRIP n'avait pas la même distribution que les statistiques tirées du PDB (fig.11 et fig.5). Dans un sens, nous étions contents de ce fait, car TRIP a été construit pour optimiser l'arrangement des structures secondaires. Les boucles ne doivent en aucun temps limiter l'arrangement des structures secondaires. Comme le nombre d'angles dièdres nous semblait à prime abord limité, nous n'étions pas surpris d'obtenir des géométries peu probables.

Pourtant, comme la construction de la protéine se passe dans l'espace dièdre, ces problèmes locaux ne devaient pas être en si grand nombre. Comme l'article de Lund *et al.*⁶¹ le mentionne, la minimisation dans l'espace dièdre réduit la

TABLEAU III. Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP.

	Résidu				Résidu			Type de contact	Distance Å
	Num.	Type	Atome		Num.	Type	Atome		
1.	5	GLN	O	→	69	LEU	C	P-P	2,0
2.	6	LEU	C β	→	73	LEU	C β	L-L	1,1
3.	9	HIS	C β	→	70	GLY	C α	L-P	1,5
4.	14	VAL	O	→	15	GLU	C β	P-L	2,5
5.	15	GLU	O	→	16	ALA	C	P-P	2,5
6.	25	ILE	C β	→	111	VAL	C β	L-L	1,6
7.	29	LEU	C β	→	43	PHE	C β	L-L	1,7
8.	29	LEU	O	→	44	LYS	C α	P-P	1,8
9.	32	SER	O	→	34	PRO	N	P-P	2,1
10.	37	LEU	O	→	38	GLU	O	P-P	1,7
11.	37	LEU	C β	→	42	ARG	C β	L-L	1,8
12.	38	GLU	O	→	39	LYS	O	P-P	2,3
13.	41	ASP	O	→	106	GLU	O	P-P	2,5
14.	41	ASP	O	→	110	HIS	C β	P-L	2,4
15.	42	ARG	O	→	43	PHE	C β	P-L	2,5
16.	42	ARG	O	→	110	HIS	C β	P-L	2,3
17.	43	PHE	O	→	44	LYS	C β	P-L	2,5
18.	44	LYS	O	→	45	HIS	O	P-P	1,4
19.	46	LEU	O	→	47	LYS	C β	P-L	2,5
20.	53	LYS	O	→	54	ALA	C	P-P	2,5
21.	54	ALA	O	→	58	LEU	C β	P-L	2,4
22.	55	SER	O	→	56	GLU	C	P-P	2,3
23.	73	LEU	O	→	74	LYS	O	P-P	2,0
24.	73	LEU	O	→	77	GLY	C α	P-P	2,1
25.	78	HIS	O	→	82	GLU	C β	P-L	1,7
26.	87	ALA	O	→	95	LYS	C β	P-L	1,8
27.	90	HIS	C α	→	94	HIS	C	P-P	1,2
28.	90	HIS	C β	→	95	LYS	C α	L-P	1,2
29.	91	ALA	N	→	95	LYS	C β	P-L	1,5
30.	95	LYS	O	→	96	ILE	C	P-P	2,3
31.	96	ILE	O	→	97	PRO	O	P-P	1,4
32.	115	ARG	O	→	116	HIS	C	P-P	2,3
33.	115	ARG	O	→	117	PRO	N	P-P	2,4
34.	116	HIS	O	→	117	PRO	O	P-P	2,2
35.	117	PRO	O	→	118	GLY	C	P-P	2,5
36.	119	ASP	O	→	120	PHE	C	P-P	2,3
37.	145	GLU	C	→	146	LEU	N	P-P	1,3

dimensionnalité, le temps de calcul *et* produit des structures avec des liens et des angles «réalistes».

Pour pallier cette lacune dans le cas des contacts locaux et générer des essais de structures qui ont moins ou même n'ont aucun mauvais contact, une meilleure gestion de l'espace dièdre me paraissait une bonne solution. Deux avenues s'offraient à moi: l'une spatiale, l'autre énergétique. Dans l'avenue spatiale, il était question de manipuler les listes soit en surreprésentant dans la distribution ou en éliminant carrément un choix, selon que l'on veuille avantager ce choix ou lui nuire. Dans l'avenue énergétique, il s'agissait plus subtilement de donner des bonus et des malus à un choix. On peut avoir soit une énergie explicite, qui est calculée à partir de divers coordonnées dans la simulation, i.e. distance interatomique, ou donner un poids énergétique à chaque choix dans une liste. Cette dernière avenue a l'avantage que l'ensemble des essais acceptés au cours de la minimisation respectera la distribution de l'énergie à travers la liste. Cependant, l'avenue énergétique exige une redéfinition de la philosophie de l'évaluation de l'énergie dans l'espace cartésien.

3.1 Choix du niveau de filtrage: espace dièdre des boucles

J'ai décidé, dans un premier temps, d'adopter l'avenue spatiale. En tablant sur l'avenue spatiale, nous pouvions régler les problèmes locaux soit au niveau de l'acide aminé, soit au niveau des boucles.

Pour limiter les conformations de boucles, on aurait pu créer une banque de boucles possibles et limiter les choix à cette banque. Cependant, je voulais conserver le plus de flexibilité possible dans les boucles. Je voulais conserver intact le principe de la minimisation à petits pas. Il me semblait donc nécessaire de conserver la possibilité pour TRIP de créer des boucles «non-naturelles».

J'ai donc choisi de manipuler l'espace dièdre au niveau de l'acide aminé, car ça me donne la possibilité:

- de créer une proportion plus grande de boucles «naturelles» durant la

minimisation

- de ne pas empêcher les boucles «non-naturelles»
- d'améliorer la qualité de la prédiction de TRIP
- d'intégrer des informations expérimentales de RMN dans la prédiction

3.1.1 Choix du type de manipulation

Au niveau de l'acide aminé, il n'y avait pas encore eu d'étude sur l'impact d'une manipulation de l'espace dièdre dans TRIP. Comme nous n'avions aucune information sur le sujet, la construction d'un filtre énergétique semblait un pas trop important. Quant à forcer une distribution d'angles prédéterminée, il faut d'abord savoir quel est le genre de distribution qui existe dans le logiciel pour faire des comparaisons utiles.

Il existe plusieurs types d'informations susceptibles d'améliorer la prédiction de TRIP. La liste qui suit correspond aux différentes manipulations de l'espace diédrique qui se trouvent dans ce travail.

- réduction de la carte de base pour tous les acides aminés dans les boucles
- une carte par type d'acide aminé dans les boucles
- une carte de base avec un choix de plus dans les boucles
- une carte de base avec un nuage de plus dans les boucles
- une carte avec 2 points par résidu dans les structures secondaires

Tous les calculs qui suivent ont été faits avec TRIP version 0.3, et ont duré cinquante itérations. Tous les résultats qui suivent dans ce mémoire correspondent à des moyennes d'au moins trois essais.

3.2 MOLFIX

À plusieurs reprises dans ce travail, nous avons besoin des angles dièdres adoptés par une protéine test. Cependant, il nous faut traduire ces valeurs pour les rendre compatibles avec le modèle réduit des protéines que nous utilisons (section 2.1.1). MOLFIX est le logiciel que j'ai développé pour obtenir la structure la plus basse en RMS possible pour un segment de protéines découpé par l'utilisateur. Ce logiciel fonctionne de façon itérative, par une promenade Monte Carlo-recuit simulé sur les angles dièdres de notre modèle moléculaire, afin que la structure résultante soit le plus proche possible (en RMS) de la vraie structure cristallographique. Ses fichiers d'entrée sont:

native Traduction d'un fichier .pdb en format TRIP

seqnce Séquence de la protéine en chiffre

param.in Paramètres Monte Carlo suivants:

- Le premier et le dernier acide aminé du segment à copier
- Le nombre d'itérations
- Le nombre de départs du générateur aléatoire
- La température initiale et finale du recuit simulé

Ses fichiers de sortie sont:

bestdi_bouX Pour la boucle X, donne la liste des couples d'angles ϕ et ψ optimaux. Lisible par CREDIH.

bestdi_helX Pour la structure secondaire X, donne la liste des couples d'angles ϕ et ψ optimaux. Lisible par CREDIH.

courbe.dat Tableau du RMS à chaque dix itérations.

MOL.pdb Sortie PDB standard du segment à l'étude. Lisible par n'importe quel programme de visualisation moléculaire.

Le but de ce logiciel est simple, mais soulève néanmoins quelques problèmes.

3.2.1 Choix du type de RMS

Le premier problème vient de la fonction RMS. Si on utilise un RMS par rotation qui n'inclut que les carbone α , nous obtenons des structures tordues, avec des angles ϕ et ψ très éloignés de la structure native (fig. 8). Comme le RMS est mon unique potentiel, ses lacunes influent directement sur les résultats.

En utilisant un RMS par rotation qui inclut les six atomes de notre modèle ($C\alpha, C\beta, N, H, C, O$), nous obtenons des valeurs plus raisonnables pour les angles ϕ et ψ (fig. 8).

3.2.2 Barrières de distance et d'angle

Pour forcer davantage le modèle à adopter des angles dièdres «biologiques», j'ai ajouté des garde-fous pour confiner les structures générées dans des conformations proches de la structure cristallographique.

Le premier garde-fou empêche l'encombrement stérique. C'est une simple fonction de volume exclu, où tous les atomes ont un rayon de 2Å. La pénalité est infinie. De façon pratique, nous ne voyons de différence dans les résultats avec cette option, lorsque le deuxième garde-fou est en fonction.

Le deuxième garde-fou limite la différence que peuvent avoir les angles dièdres de la structure générée versus la structure native. Si un angle dièdre de la structure générée dépasse plus ou moins un certain angle, une pénalité à définir est ajoutée à l'énergie. Ce deuxième garde-fou est utile, car il arrive que le calcul numérique de la minimisation vers un RMS bas est instable. En effet, malgré que l'utilisation d'une RMS six atomes minimisent beaucoup les distortions d'angles, il arrivent que les angles dièdres des structures générées s'éloignent trop des angles dièdres cristallographiques. Avec ce garde-fou, nous sommes assurés d'obtenir des structures très proches de la native quant à la position des atomes, tout en respectant les angles dièdres et donc l'allure générale de la chaîne principale.

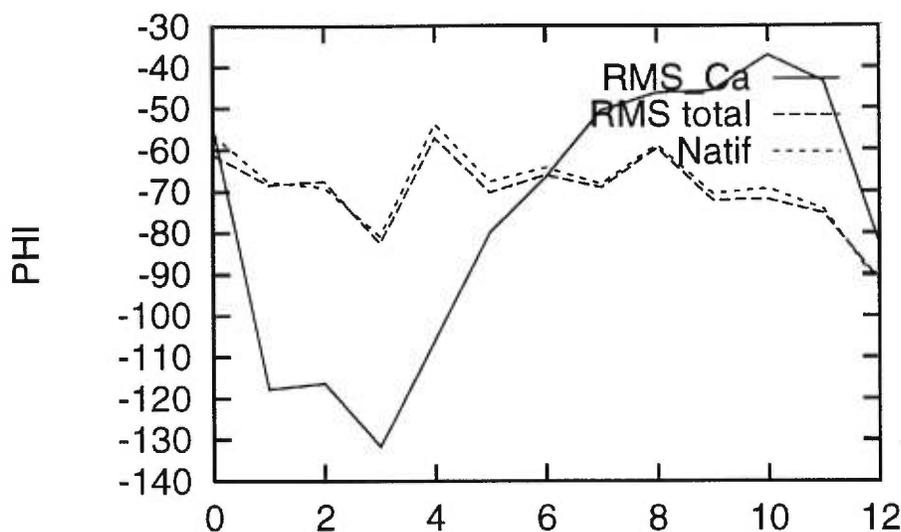


FIGURE 8. Angle ϕ de l'hélice 6 de la structure native de la myoglobine, de la structure générée avec un RMS carbone α et générée par un RMS 6 atomes. L'axe des x correspond au numéro des résidus, à partir du résidu 80.

3.3 Manipulation de l'espace dièdre des boucles

La plupart de mes résultats ont été obtenus en faisant des calculs de prédiction sur la myoglobine de cachalot, coté dans le PDB comme la 1MBO. Des calculs sur la partie C-terminal des unités L7/L12 des ribosomes de *E. coli* (référéncé 1CTF) figure aussi dans ce travail. On peut trouver en annexe les références des fichiers PDB de ces structures.

La 1MBO tel que vu par TRIP possède 146 résidus en tout, dont 37 sont dans les boucles. Les huit structures secondaires qui la composent sont des hélices α . Il y a donc environ 75% de la protéine qui reste fixe durant le calcul. C'est une protéine sans complications, hormis le groupe heme. Comme certaines évidences expérimentales tendent à démontrer que le heme ne transforme pas fondamentalement la structure de la protéine, nous utilisons une myoglobine vidée de son groupe heme. C'est la protéine parfaite pour tester un modèle en développement.

La 1CTF vu par TRIP possède 66 résidus, répartis en six structures secondaires: feuillet, hélice, hélice, feuillet, hélice, feuillet. Il y a 17 résidus dans les boucles, laissant 75% de la protéine fixe durant le calcul. La complication principale de la 1CTF est la présence de feuillets β .

3.3.1 Réduction unilatérale de l'espace conformationnel

Le premier genre de test veut déterminer jusqu'à quel point une carte unique peut satisfaire le besoin de tous les résidus en termes de conformations. Plus spécifiquement, on veut connaître le plus petit nombre de couples d'angles qui donnerait des structures raisonnables. La figure 9 montre les différentes cartes que nous avons essayées.

La carte de 6 points correspond à prendre un point central dans chaque région permise de l'espace de Ramachandran, tel qu'il est proposé par Rooman *et al.*⁶². La carte de 18 points inclut les 6 points précédents, tout en enrichissant les zones statistiquement plus utilisées dans les protéines connues. Ainsi de suite, les cartes 50, 100, 163, 372 et 534 ne sont que des versions plus garnies des cartes moins peuplées. La carte 372 points négatifs est constituée de tous les points de la carte 534 dont l'angle ϕ est négatif. La carte 163 positifs est constituée de tous les points de la carte 534 dont l'angle ϕ est positif, plus le point -65, -40. La carte de 676 points couvre la totalité de l'espace conformationnel, incluant des positions interdites ou absentes des protéines qu'on connaît actuellement.

Comparons les cartes 676, 534 et 372bio, figurant dans le tableau IV. Il n'y a aucune différence significative de RMS entre celles-ci. Pour TRIP, le champ des possibilités avec ces trois cartes est identique. En fait, TRIP est incapable, en moyenne, d'amener l'ensemble échantillon vers la structure native. Pire encore, la moyenne des énergies est largement au-dessus de celle de la native. On verra plus loin qu'il existe beaucoup de conformations pour la myoglobine où l'énergie avoisine celle de la native. Les valeurs élevées d'énergie montrent donc l'incapacité de TRIP à amener l'ensemble échantillon vers des zones plus basses en énergie. TRIP est tout simplement submergé par le nombre astronomique de possibilités.

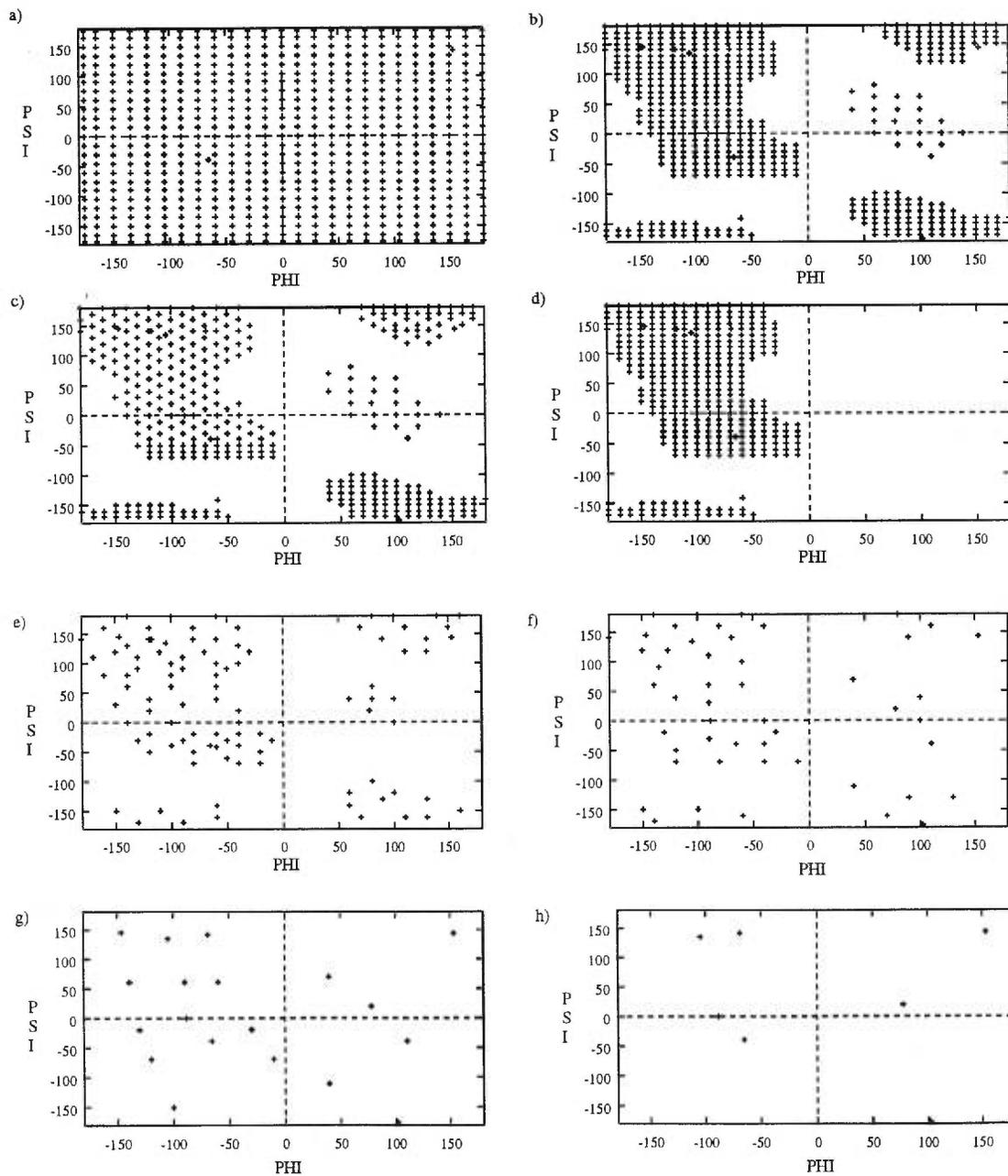


FIGURE 9. Les cartes réduites testées dans ce travail. a) 676 points b) 534 points c) 372 points d) 372 points négatifs e) 100 points f) 50 points g) 18 points h) 6 points

TABLEAU IV. Effet de la réduction uniforme de l'espace dièdre pour un calcul TRIP typique sur la myoglobine.

Nb de couples d'angles dièdres	<RMS>	<Meilleur RMS>	< R^{gir} >	<E>	<Meilleure E>
Native			14,90	-167,18	
676	13,25	10,42	15,24	-152,33	-178,88
534	13,73	8,87	16,47	-147,97	-187,55
372 biol.	13,07	9,69	15,37	-145,76	-178,35
372 nég.	13,37	8,38	15,68	-148,55	-179,92
163 pos.	14,33	12,11	15,92	-146,49	-174,88
100	13,70	10,29	15,29	-157,28	-184,05
50	11,51	9,49	14,97	-167,23	-191,45
18	11,17	9,45	14,34	-167,59	-183,93
6	13,04	12,06	15,69	-161,58	-160,16

D'autre part, si on compare les cartes 100, 50, 18 et 6, nous observons les meilleures énergies dessiner une parabole. Avec 50 points, l'espace dièdre est beaucoup moins grand à couvrir qu'avec 100 points, et TRIP échantillonne une plus grande proportion de l'espace conformationnel. On peut constater ce résultat positif avec les moyennes d'énergie qui avoisinent l'énergie de la native, mais surtout par l'amélioration des meilleures structures en RMS et en énergie. Les meilleures énergies pour les cartes 18 et 6 points remontent, car TRIP manque de choix pour remplir ses listes de boucles. Un indice pour déceler ce problème est le taux d'acceptation Monte Carlo de quarante pour cent à la fin des cinquante itérations d'un calcul à 6 points. Normalement, ce taux est en deçà de un pour cent après vingt itérations. En d'autres termes, TRIP ne converge pas vers un ensemble de l'espace conformationnel lorsqu'il manque de points. Le modèle du ressort (p. 13) exige un nombre suffisant de possibilités pour positionner les structures secondaires. Autre indice, le rayon de giration décrit aussi une parabole. À 6 points, TRIP réussit moins bien à maintenir la protéine compacte. Un bémol cependant, une minimisation du RMS avec une carte 6 points donnent la figure 10 (dessinée avec MOLSCRIPT²), avec un RMS de 3,94, ce qui constitue le minimum absolu pour une carte à 6 points. Le message à tirer de ces quatre cartes est qu'il est avantageux de réduire la dimensionnalité du problème (la carte 50 points obtient vraiment de meilleures énergies et des meilleures RMS), mais pas au point de brimer le modèle du ressort flexible.

Comparons maintenant 372bio, 372 négatifs et 163 positifs. L'idée ici est de frustrer le système en l'empêchant d'utiliser des conformations nécessaires dans ses boucles. Malgré les moyennes de RMS élevées pour la 372bio qui nous empêchent de bien caractériser le phénomène, nous constatons une augmentation du RMS chez la 372nég et la 163pos. Il y a donc vraiment une frustration dans la capacité de générer des structures avec les meilleurs arrangements topologiques, comme nous le souhaitions. Notre modèle a répondu à cette frustration en augmentant les moyennes d'énergie de 372nég et 163pos. Cependant, une

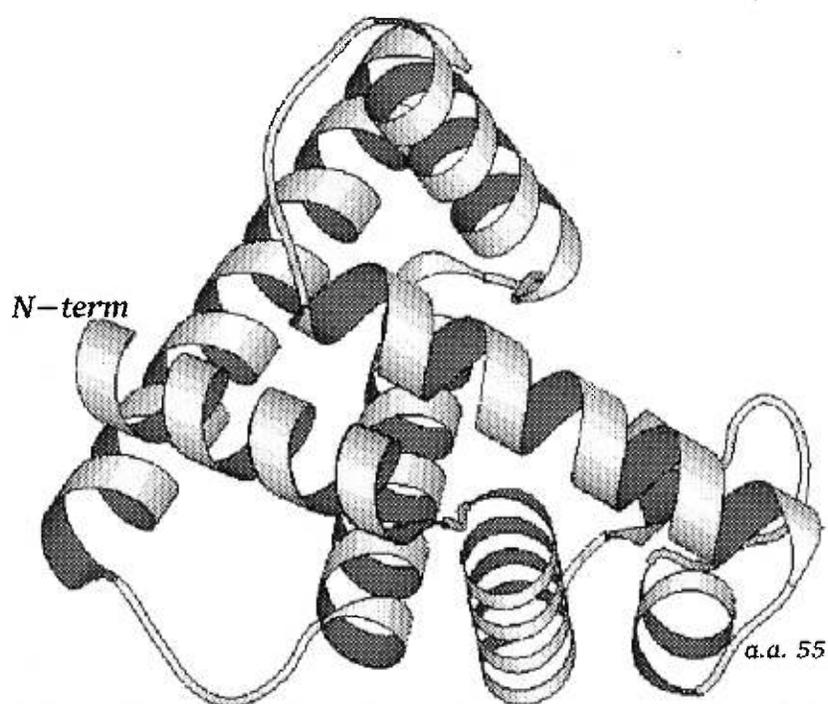


FIGURE 10. Représentation schématique de la myoglobine obtenue par TRIP avec une carte à 6 points, avec le RMS comme seul potentiel. Réalisé avec MOLSCRIPT².

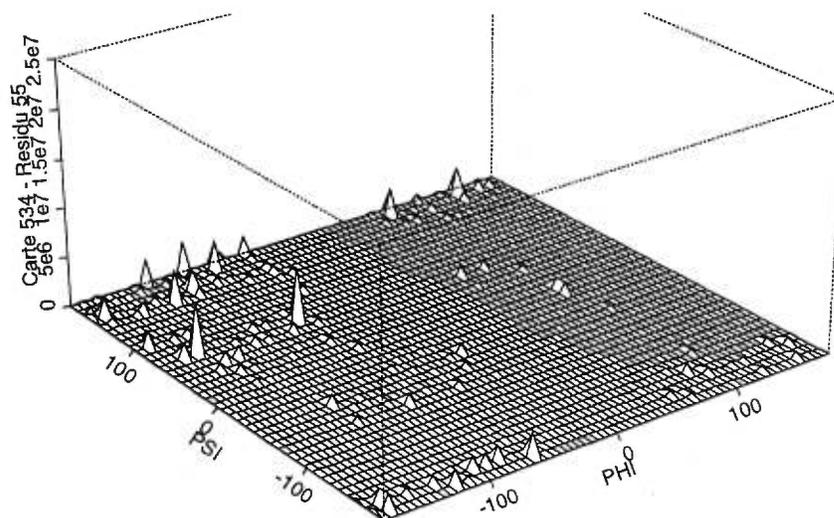


FIGURE 11. Utilisation typique des couples d'angles dièdres de la carte 534 par un résidu dans une boucle. Ici, le résidu 55 de la myoglobine. Une unité dans l'axe z correspond à une conformation acceptée.

augmentation d'une quinzaine d'unités n'est pas vertigineuse. J'interprète ceci comme si TRIP possède assez de flexibilité avec les cartes frustrées pour créer des molécules aussi vraisemblables qu'avec la 372bio. En d'autres termes, notre modèle du ressort(p. 13) est respecté même avec des cartes frustrées.

Pour conclure avec ces cartes de base, il est intéressant de regarder une carte de l'utilisation de l'espace dièdre par TRIP. On peut à cet effet colliger l'utilisation des couples d'angles dièdres à la fin de chaque cycle Monte Carlo. Ça fait trente-deux millions de structures acceptées par le critère Métropolis d'énergie pour un calcul de 50 itérations. La figure 11 montre l'utilisation de la carte 534 par le résidu 55 de la 1MBO.

Nous reverrons plus loin le résidu 55 de la 1MBO pour fin de comparaison. À noter dans la figure 11 le manque d'un patron clair. TRIP n'a pas trouvé d'arrangement plus probable pour ce résidu en particulier. Si cela avait été le cas, nous aurions un large pic pour cette conformation. Plutôt, la conformation finale n'a été que l'aboutissement de la suite d'essais auto-cohérents avec le

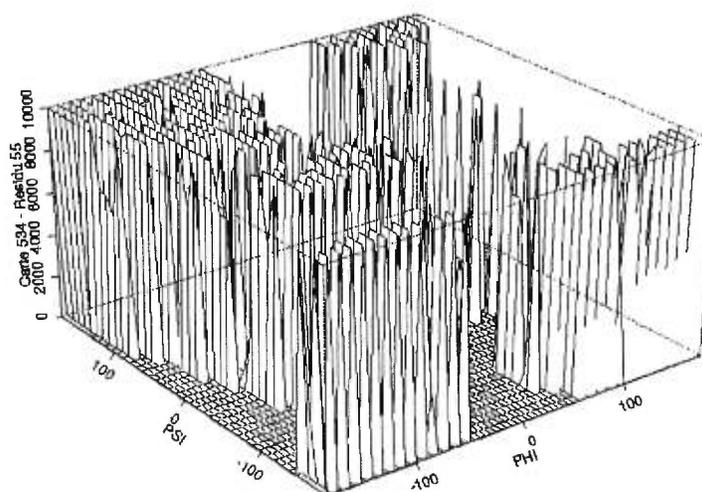


FIGURE 12. Utilisation typique des couples d'angles dièdres de la carte 534 par un résidu dans une boucle. Coupe horizontale. Une unité dans l'axe z correspond à une conformation acceptée.

reste de la protéine. Comme chacune des 64 structures évolue plus ou moins indépendamment, chacune a choisi une conformation différente pour le résidu 55. Si on fait une coupe horizontale du graphique 11 au niveau d'une dizaine de milliers, nous obtenons la figure 12.

Cette figure nous prouve, premièrement, que peu importe les points qu'on place dans une carte, TRIP les essaiera. Deuxièmement, l'utilisation minimum d'un point se situe dans les dizaines de milliers, soit 0.03% de la population totale. Ces résultats nous confortent dans l'idée que la manipulation de l'espace dièdre peut fonctionner et conserve le modèle des ressorts souples. Si TRIP n'avait pas essayé tous les points offerts, il aurait été difficile de poursuivre dans cette voie.

3.3.2 Carte selon l'espèce chimique, distribution uniforme

Ce genre de carte peut nous renseigner sur la réponse de TRIP à une information géométrique spécifique. Ce qu'on entend par distribution uniforme, c'est le fait qu'un point n'apparaît qu'une fois dans la liste. La figure 13 montre les différentes cartes que nous avons utilisées.

Dans le tableau V, GLY correspond à la glycine, PRO à la proline, NPRO à tous

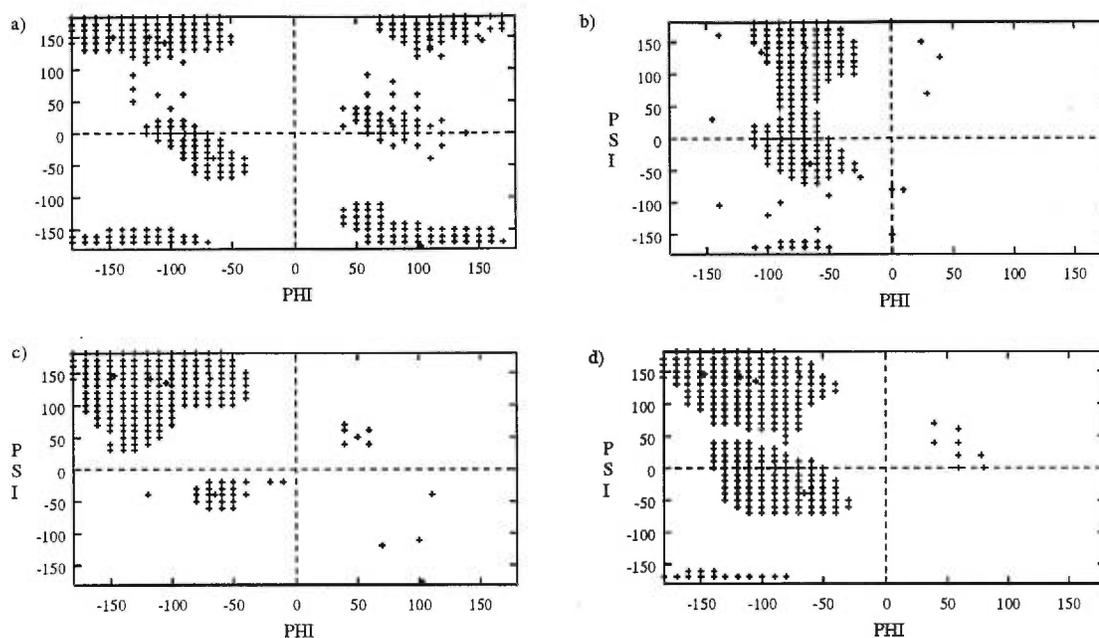


FIGURE 13. Les cartes spécifiques uniformes selon le type d'acide aminé testées dans ce travail. a) Glycine 324 couples. b) Proline 186 couples. c) Résidu avant une proline 201 couples. d) Tous résidus sauf glycine, proline et résidu devant proline 273 couples.

types de résidus précédant la proline et X à tous les résidus exceptés les trois précédents. Cette nomenclature sera respectée dans la suite du mémoire.

Dans le tableau V, lorsqu'un acide aminé n'est pas mentionné, la carte de base 534 lui est assignée.

La première leçon de ce type de carte est que TRIP ne voit presque aucun avantage. TRIP n'a pas plus de facilité à minimiser l'énergie qu'avec la carte de base.

Deuxième leçon et un des rares points positifs est la diminution sensible des RMS pour les cartes PRO et X+PRO. Nous notons par exemple des structures très voisines de la native, autour de 7 Å RMS.

Plus systématiquement, les moyennes de RMS pour les cartes GLY et NPRO sont pires que la carte de base, autant prises seules que combinées. Les énergies un peu plus basses proviennent du nombre de points moins grand. Les cartes X et PRO se démarquent particulièrement. Prises seules, les moyennes de RMS diminuent, mais ensemble, un effet synergique semble améliorer

TABLEAU V. Effet de cartes uniformes pour différents acides aminés dans un calcul TRIP typique sur la myoglobine.

Type de carte	<RMS>	RMS Minimum	<Énergie>	<Énergie Minimum>
Natif du PDB			-167,18	
Carte de base*	13,73	8,87	-147,97	-187,55
GLY (324couples)	14,31	7,31	-154,25	-182,48
NPRO (201couples)	14,25	8,26	-154,55	-178,21
PRO (186couples)	12,43	6,70	-152,95	-185,22
X (273couples)	12,39	7,36	-164,73	-191,99
PRO + NPRO	13,87	10,12	-152,02	-174,94
X + PRO	11,89	7,84	-156,87	-184,70
GLY + PRO + NPRO	14,22	10,89	-167,47	-194,22

* 534 couples

davantage le score. De façon plus précise, on observe avec la carte X des énergies plus proches de la native qu'avec la carte de base 534. La réduction du nombre de points ne peut expliquer seule cette baisse. En comparant avec la carte de base 100 du tableau IV, nous devons admettre qu'il y a une information supplémentaire locale dans la carte X qui aide vraiment la prédiction, malgré l'accroissement du nombre de conformations possibles (de 100 à 273 couples).

En résumé, malgré une réduction de la dimensionnalité, le problème reste NP complet (page 19). Les résultats de TRIP nous montre qu'il est toujours incapable d'arriver à la structure native.

Deuxièmement, lorsqu'une carte contient de l'information utile, TRIP améliore sa capacité de prédiction. Ainsi la carte PRO et X améliore le score en RMS et en énergie, même s'il y a plus de points.

Troisièmement, lorsqu'une carte n'est pas appropriée, qu'elle ne possède pas d'information utile, c'est la stricte dimension (le nombre de points) de la carte qui oriente l'aboutissement du calcul. On peut ainsi imbriquer les tableau IV et V pour la GLY et NPRO. En effet, les cartes GLY et NPRO ne semblent pas contenir d'informations utilisables. Nous voyons alors au niveau du RMS et de l'énergie que les résultats s'intercalent bien dans le tableau IV entre les cartes 372bio et 100 points.

Quatrièmement, le tableau VI montre que le nombre de mauvais contacts baisse. En effet, PROCHECK répertorie en moyenne 20 contacts, dont 5 entre des atomes de la chaîne principale rapprochés d'au moins 4 acides aminés. La stratégie de régler un problème local via les angles dièdres peut donner des fruits. On peut régler des problèmes locaux systématiques par un filtre local systématique. Cette amélioration géométrique se fait peu sentir sur l'énergie, car nous n'avons pas de terme d'interactions locales dans le potentiel.

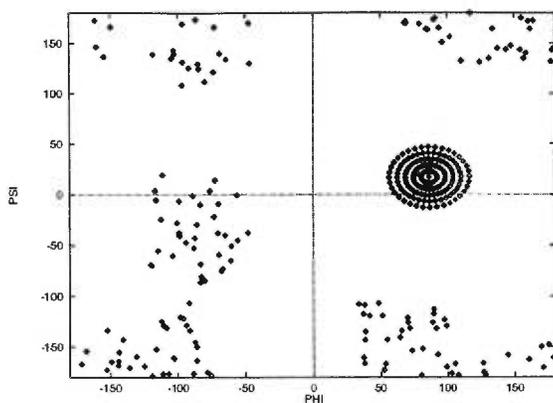
3.3.3 Carte selon l'espèce chimique, distribution non-uniforme

Ce genre de carte peut nous renseigner sur la réponse de TRIP à une information géométrique spécifique.

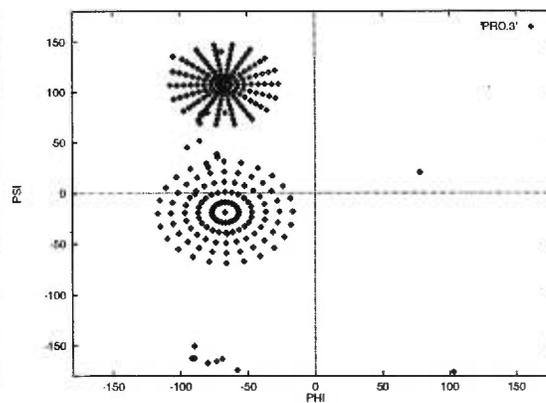
TABLEAU VI. Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP avec les cartes X et PRO.

	Résidu			→	Résidu			Type de contact	Distance Å
	Num.	Type	Atome		Num.	Type	Atome		
1.	14	VAL	O		15	GLU	C β	P-L	2,5
2.	26	LEU	C β		108	ILE	C β	L-L	2,6
3.	31	LYS	O		32	SER	C β	P-L	2,6
4.	35	GLU	O		38	GLU	O	P-P	2,2
5.	39	LYS	O		40	PHE	C	P-P	2,6
6.	67	THR	C β		80	GLU	O	L-P	2,5
7.	67	THR	O		83	LEU	C β	P-L	2,4
8.	67	THR	C β		84	LYS	C β	L-L	2,2
9.	68	ALA	O		110	HIS	C β	P-L	2,2
10.	69	LEU	O		107	ALA	C α	P-P	1,8
11.	72	ILE	C β		106	GLU	O	L-P	0,4
12.	72	ILE	C β		107	ALA	N	L-P	2,3
13.	73	LEU	C α		102	GLU	O	P-P	2,5
14.	73	LEU	C β		103	PHE	O	L-P	1,4
15.	73	LEU	N		106	GLU	C β	P-L	1,0
16.	78	HIS	O		79	HIS	O	P-P	2,5
17.	105	SER	C β		132	LEU	C β	L-L	2,1
18.	113	HIS	O		116	HIS	O	P-P	2,4
19.	120	PHE	O		124	ALA	C β	P-L	2,1
20.	145	GLU	C		146	LEU	N	P-P	1,3

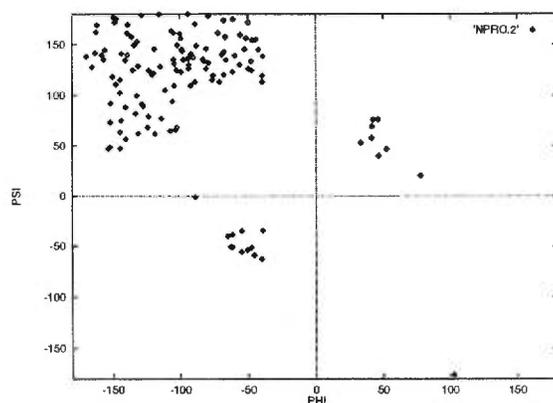
a)



b)



c)



d)

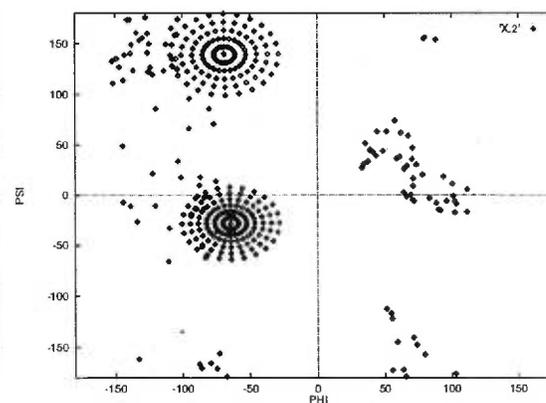


FIGURE 14. Les cartes spécifiques non-uniformes selon le type d'acide aminé testées dans ce travail a) Glycine 351 couples. b) Proline 322 couples. c) Résidu avant une proline 133 couples. d) Tous résidus sauf glycine, proline et résidu devant proline 415 couples.

Les cartes non-uniformes que nous avons utilisées sont très simples. Le concept d'une agglomération de points est surtout présent dans les cercles, où un même point peut être répertorié plusieurs fois dans la liste. Mon but était de grossièrement représenter les statistiques du Dr Felts (fig. 5). En listant un point plus d'une fois dans la liste, on augmente d'autant sa chance d'être utilisé. Comme on donne une carte d'entrée qui a une distribution «naturelle», on s'attend à ce que TRIP respecte cette distribution dans son choix de triplets. Toutes les cartes incluent la carte à 6 points. La figure 14 répertorie les cartes que nous avons utilisées.

Le tableau VII résume les résultats importants. Toutes les cartes non-uniformes sont meilleures que leur équivalent uniforme. Dans les cartes combinées, on note un effet synergique entre les cartes PRO et NPRO. Il ne faut pas oublier qu'il n'y a que 3 prolines dans la 1MBO. On conclut que la distribution de départ de six résidus sur 146 peut améliorer la recherche conformationnelle.

Ce résultat est important, car c'est le premier qui suppose que les boucles ne sont pas de simples ressorts aléatoires. Même s'il n'y a pas beaucoup d'énergie associée aux boucles en tant que telles, l'amélioration de l'espace conformationnel dans les boucles permet à TRIP de faire de meilleures recherches conformationnelles dans l'alignement des structures secondaires. Ce résultat se révèle plus important que ce que l'on recherchait au départ. Il m'a lancé sur la voie de MOLFIX.

Deuxième point important, on s'attend à ce qu'il y ait une amélioration des résultats de TRIP avec les cartes non-uniformes et que les cartes d'utilisation de l'espace dièdre soient des répliques de la carte d'entrée. Or, les résultats ne sont pas aussi univoques que prévu. Les patrons de distribution des résidus qui utilisent la carte X restent plats. Cependant, ceux avec des cartes spécifiques, surtout NPRO et PRO, choisissent un point et l'enrichissent (fig. 15). Nous ne retrouvons jamais le paysage des cartes d'entrée.

Troisièmement, le tableau VII montre que l'avantage des cartes spécifiques ne semble pas directement cumulatif. Lorsqu'on utilise plusieurs cartes différentes, il n'y a pas toujours un effet synergique, comme dans le cas de NPRO+PRO. Ceci rend les conclusions de ce genre de carte très floues. Il y a peut-être un problème local sur la myoglobine à utiliser GLY, PRO et NPRO ensemble. L'utilité des cartes par acide aminé diminuerait de beaucoup si elles figeaient la recherche conformationnelle.

Le tableau VIII divise les valeurs d'énergie moyenne du tableau VII en ses constituants. Nous voyons deux types d'effets. Le premier est la baisse de moitié de l'énergie Hélice \leftrightarrow Boucle pour les cartes modifiées. Même si, en termes absolus, cet abaissement de l'énergie est mineur, l'idée de rendre les triplets plus «naturels»

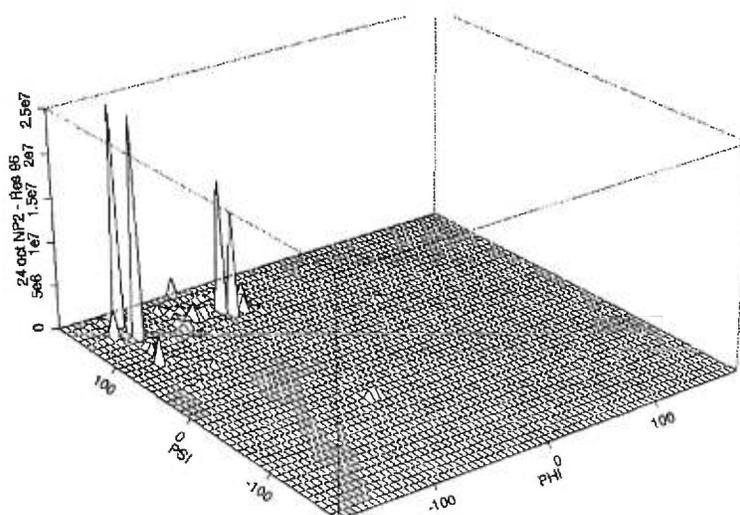


FIGURE 15. Utilisation des couples d'angles dièdres de la carte NPRO non-uniforme pour un calcul TRIP typique sur la myoglobine. Une unité dans l'axe z correspond à une conformation acceptée.

TABLEAU VII. Effet de cartes non-uniformes pour différents acides aminés dans un calcul TRIP typique sur la myoglobine.

Type de carte	<RMS>	RMS Minimum	<Énergie>	<Énergie Minimum>
Natif du PDB			-167,18	
Carte de base*	13,70	8,97	-157,28	-184,05
GLY (351couples)	13,11	6,88	-160,71	-188,22
NPRO (133couples)	13,08	7,53	-161,30	-186,51
PRO (322couples)	12,92	7,31	-161,41	-186,97
X (415couples)	12,12	8,54	-175,97	-204,94
PRO + NPRO	11,74	7,63	-166,66	-188,88
X + PRO	13,51	7,77	-159,20	-187,29
GLY + PRO + NPRO	12,87	8,04	-159,85	-181,32

* 100 couples

porte fruit. Le deuxième est l'abaissement de l'énergie Hélice \leftrightarrow Hélice pour la carte X. De meilleures boucles permettent un meilleur agencement des hélices dans l'espace. À noter que la carte X représente 28 acides aminés sur 37, tandis que PRO+NPRO représente 6 acides aminés. Dans ce dernier cas, la carte modifiée permet probablement de soulager une tension locale dans une des boucles. Dans le cas de la carte X, les effets semblent être petits mais cumulatifs. Toutes les distributions de triplets sentent l'effet d'une carte modifiée. Posons un bémol cependant puisque les calculs X+NPRO+PRO donnent systématiquement de moins bons résultats. Je ne peux alors expliquer comment les effets cumulatifs de la carte X peuvent cesser d'opérer lorsque 6 résidus ont des cartes différentes.

Les réponses de mon modèle très simple pour les cartes d'acides aminés nous amènent bien sûr à carrément utiliser les cartes du Dr Felts. Nous nous posons cependant la question s'il ne vaudrait pas mieux porter en énergie l'axe des z, plutôt que d'augmenter le nombre de points d'entrée. Dans le premier cas de figure, on s'assurerait que la distribution finale ressemble beaucoup plus à la distribution initiale, au prix d'une moins grande flexibilité dans la création de triplets. On peut craindre que cette perte de flexibilité dans les triplets réduisent l'échantillonnage de conformations moléculaires utiles. Il est à remarquer que certains chercheurs n'ont pas noté cette rigidification de la recherche conformationnelle avec des potentiels sur les angles dièdres.^{63 64}

3.3.4 Carte avec les angles MOLFIX dans les boucles

Maintenant, curieux de savoir s'il y avait vraiment une corrélation entre des bonnes boucles et un meilleur agencement des structures secondaires, nous avons développé MOLFIX pour tester TRIP. Le but de MOLFIX(p. 41) est de trouver la structure produite à partir d'un modèle réduit de protéine qui minimise le RMS par rapport à une structure cible. Ceci, nous donne les meilleures angles dièdres de notre modèle réduit. En donnant le meilleur angle à TRIP, nous voulions voir jusqu'à quel point TRIP allait l'utiliser, et jusqu'à quel point il y a une corrélation entre de bonnes boucles et l'agencement des structures secondaires dans l'espace.

Dans un premier temps, un angle MOLFIX a été placé dans des cartes de base de plus en plus clairsemée. Le tableau IX collige les résultats importants. Au niveau des RMS, la moyenne est sensiblement la même pour les cartes 534 à 50. Pour la carte à 18 points, nous obtenons une meilleure moyenne. Si on fait des comparaisons avec le tableau IV, on s'aperçoit que la moyenne des RMS est abaissée d'une unité pour les cartes 534 et 100. Pour les cartes 50 et 18 points avec angle MOLFIX on obtient de nouveau de meilleurs résultats que pour les cartes 50 ou 18 points seules, mais de l'ordre d'une demi-unité. Pour la moyenne des meilleurs RMS (ce qui est différent du meilleur RMS global), nous avons amélioré la performance globale de TRIP. Au niveau de l'énergie, on a encore une amélioration pour toutes les cartes. Il est à noter cependant qu'il n'y a pas de relation de linéarité directe entre l'amélioration du RMS et la diminution de l'énergie. Nous le constatons bien ici.

La leçon de ce type de carte est qu'ajouter un seul point dans une liste n'est pas assez pour radicalement améliorer la prédiction. La nature stochastique de la recherche Monte Carlo, et les barrières que nous avons placées pour éviter une convergence trop rapide de l'algorithme génétique, obligent TRIP à maintenir une population d'ensemble hétérogène au niveau de l'espace dièdre.

Cependant, comme il y a une amélioration, on doit conclure que TRIP utilise les points, mais l'absence de patron d'utilisation nous indique qu'il n'y a pas surutilisation des points MOLFIX (fig.16). TRIP ne fait qu'utiliser les points parce qu'ils sont dans sa liste.

3.3.5 Carte avec des nuages autour des angles MOLFIX uniquement dans les boucles

Dans cette même veine d'essai MOLFIX dans les boucles, la construction des nuages de points autour de l'angle MOLFIX de chaque résidu semblait intéressante. Les nuages sont définis par leur longueur et leur largeur en degrés. Les résultats du tableau X sont déroutants.

Pour la carte 534, ajouter un nuage de points améliore à peine le RMS par rapport

TABLEAU VIII. Ventilation de l'énergie pour un calcul TRIP sur la myoglobine avec des cartes d'acides aminés non-uniformes.

Carte	<Hélice↔Hélice>	<Hélice↔Boucle>	<Boucle↔Boucle>
Carte de base*	-178,32	13,33	5,03
PRO+NPRO	-178,72	7,51	4,55
X	-185,35	4,69	4,69

* 100 couples

TABLEAU IX. Effets des couples d'angles MOLFIX uniquement dans les boucles sur un calcul typique de TRIP sur la myoglobine.

Carte	<RMS>	<RMS Minimum>	<Énergie>	<Meilleure Énergie>
534	12,86	9,05	-155,14	-184,92
100	12,44	9,23	-161,39	-187,90
50	11,24	8,76	-171,25	-190,39
18	10,52	8,25	-175,84	-189,94

à la carte 534 seule, mais empire le RMS par rapport à un point MOLFIX. Le nuage est un handicap, autant un 5×5 qu'un 10×10 . Je ne peux expliquer ce résultat. Au niveau de l'énergie, un point MOLFIX ou un nuage MOLFIX donne la même chose. Mon hypothèse serait qu'ici, TRIP est complètement submergé, et n'est pas capable d'amener l'ensemble vers un consensus. L'information locale supplémentaire qu'on donne à TRIP n'est pas utilisée à sa juste valeur, ne se répercute pas en baisse d'énergie.

Les autres cartes ont beaucoup mieux répondu aux nuages. En fait, plus le pourcentage des nuages augmente, plus le RMS et les énergies diminuent. Dans la carte $100+5 \times 5$, nous observons que le RMS minimum moyen est très bas, or les moyennes d'énergies et de RMS sont sensiblement identiques. Ce qui est déroutant ici, c'est que les moyennes des meilleures énergies restent aussi identiques. Un peu comme si, grâce au nuage, TRIP trouvait un puits de RMS qui ne se matérialise pas en baisse d'énergie. On peut se demander alors pourquoi TRIP y fait plus souvent descendre des structures. Dans la version actuelle de TRIP, nous ne suivons pas durant la minimisation l'évolution de l'énergie pour chaque structure (impossible avec un algorithme génétique). Certains effets de «condensation» sont peut-être à l'oeuvre ici, et expliqueraient ce dernier résultat. Un peu comme si très tôt dans la minimisation, grâce au nuage, l'énergie descendait rapidement dans le puits «natif». En même temps, le reste de l'ensemble s'optimiserait graduellement. À la fin de la minimisation, TRIP n'aurait pas su améliorer fondamentalement les structures qui étaient tombées dans le puits natif, ce qui mènerait à un ensemble de structures possédant la même énergie, mais n'ayant pas les mêmes RMS. Donc, même si l'énergie finale de la structure est identique, le chemin pour y arriver semble subir une discontinuité.

Pour les deux derniers essais, nous observons un plafonnement dans l'amélioration du RMS et des énergies. L'effet d'avoir de bonnes boucles ne peut contre balancer le fait que les structures secondaires sont «idéales». En effet, comme la majeure partie de l'énergie se situe au niveau des structures secondaires, il sera impossible

pour TRIP d'aboutir à la structure native, si celles-ci n'ont pas le bon agencement. Ceci étant dit, il faut souligner que la valeur moyenne des énergies est de 40 unités en dessous de celle de la native. Ceci est un bon résultat en soit, mais signifie aussi que la configuration cristallographique de la 1MBO n'est pas le minimum global de notre fonction d'énergie. Notez qu'à la page 66, une méthode nous permet d'obtenir une structure native dont l'énergie tourne autour du -210.

Finalement, le tableau XI répertorie les mauvais contacts pour un calcul typique avec des nuages MOLFIX dans les boucles. Nous voyons que le nombre total de mauvais contacts diminue dramatiquement. Il ne reste que 6 mauvais contacts entre voisins de 4 acides aminés, dont 4 concernent la chaîne principale uniquement. C'est une amélioration par rapport à la vingtaine de mauvais contacts entre chaîne principale qu'on obtient avec un calcul TRIP de base. Encore une fois, la stratégie de régler un problème local via les angles dièdres peut donner des fruits. Cependant, on ne peut s'attendre à éliminer tous les mauvais contacts.

3.3.6 Synthèse de l'espace des boucles

Nous avons appris que nous pouvons améliorer dramatiquement la qualité de la prédiction en modifiant de façon ciblée l'espace dièdre des boucles. D'une part, la modification selon l'espèce chimique donne de moins bons résultats que prévus, mais n'a pas été exhaustivement couverte. D'autre part, la présence de couples d'angles MOLFIX nous donne l'impression qu'il y a une corrélation plus forte que prévue entre de bonnes boucles et un bon agencement des structures secondaires. Finalement, il y a un plafond à l'amélioration de la prédiction avec ce genre de manipulation, qui est causé par l'utilisation de structures secondaires rigides et «idéales».

3.4 Espace dièdre des structures secondaires

L'idée ici est de savoir jusqu'à quel point le modèle des cylindres est contraignant pour TRIP, dans la perspective d'utiliser MOLFIX pour tous les résidus de la

TABLEAU X. Effets des nuages MOLFIX uniquement dans les boucles sur un calcul typique de TRIP sur la myoglobine.

Carte de base + Nuage	<RMS>	<RMS Minimum>	<Énergie>	<Meilleure Énergie>
534 + 5×5	13,15	8,80	-156,84	-183,52
534 + 10×10	13,01	9,63	-156,15	-185,69
100 + 5×5	11,24	6,44	-161,33	-188,83
50 + 5×5	7,84	3,95	-191,85	-217,54
100 + 10×10	6,01	2,93	-203,73	-221,86
50 + 10×10	5,90	3,63	-202,64	-216,98

TABLEAU XI. Liste des mauvais contacts rapportés par PROCHECK, pour une myoglobine typique générée par TRIP avec des nuages MOLFIX.

	Résidu			→	Résidu			Type de contact	Distance Å
	Num.	Type	Atome		Num.	Type	Atome		
1.	10	VAL	C β		112	LEU	C β	L-L	2,1
2.	13	LYS	C β		115	ARG	C β	L-L	1,9
3.	84	LYS	O		88	GLN	C β	P-L	2,0
4.	89	SER	O		93	LYS	C β	P-L	2,1
5.	90	HIS	O		94	HIS	C	P-P	1,7
6.	113	HIS	O		117	PRO	N	P-P	1,8
7.	113	HIS	C α		120	PHE	C β	P-L	2,5
8.	116	HIS	O		117	PRO	C	P-P	2,6
9.	145	GLU	C		146	LEU	N	P-P	1,3

protéine.

Malgré le fait que TRIP garde rigides les structures secondaires, cette position rigide peut être choisie. Nous pouvons d'une part la choisir avant la simulation, ou d'autre part laisser TRIP choisir durant la minimisation. Lorsque nous laissons TRIP choisir sa structure secondaire, nous appelons le calcul «interactif».

Dans ce travail, nous n'avons que deux choix d'angle: idéal et MOLFIX. Une structure secondaire idéale signifie que les hélices ont un $\phi = -65^\circ$ et un $\psi = -40^\circ$, et que les feuillets anti-parallèles ont un $\phi = -105^\circ$ et un $\psi = 135^\circ$. MOLFIX me donne les angles dièdres de tous les résidus d'une protéine, il suffit d'ajouter à la liste d'angles de TRIP les angles qui appartiennent aux structures secondaires.

Dans le cas d'un choix interactif, TRIP passe 10% de ses essais Monte Carlo à changer les structures secondaires. Lorsque TRIP fait un essai dans les structures secondaires, il échange une structure secondaire idéale pour une structure secondaire MOLFIX ou inversement. La structure secondaire visée est changée en entier.

En modifiant uniquement les structures secondaires, on s'attend à ce que les moyennes de RMS et d'énergie soient abaissées d'une façon beaucoup plus substantielles qu'avec des modifications dans les boucles. En effet, il existe des distorsions dans les structures secondaires. Certaines sont causées par la nature chimique d'un acide aminé (e.g. proline), mais la plupart sont le résultat d'interaction avec le reste de la chaîne. En donnant à TRIP la structure secondaire native, nous donnons la possibilité à TRIP d'agencer plus finement les structures secondaires entre elles. En effet, ceci correspond à la différence entre un casse-tête où toutes les pièces sont de même forme, et un casse-tête où chaque pièce est unique. On comprend qu'une fois repliée, la protéine générée aura de meilleures interactions (baisse d'énergie), et une plus grande ressemblance avec la conformation native (baisse de RMS). On ne peut cependant pas prévoir la présence d'un phénomène négatif dans les premiers moments de la minimisation,

lorsque les structures de l'ensemble évoluent très loin de la conformation native. Le tableau XII montre mes résultats importants. Les hélices natives améliorent le RMS moyen de deux unités et les énergies de 10 unités. Cependant, la possibilité d'échanger les structures secondaires durant la minimisation n'améliore pas substantiellement la qualité de la prédiction. Mon hypothèse est que l'absence d'informations supplémentaires dans les boucles empêche TRIP de converger vers les structures à 6Å que nous avons dans le tableau X. Même si la carte à 100 points possède des points très proches des angles natifs, TRIP ne converge pas. L'idée ici serait que la minimisation semble passer par une phase de «condensation». Il y a un effet synergique entre l'agencement des structures secondaires et la recherche des bonnes boucles.

3.5 Accès aux angles MOLFIX dans tous les espaces dièdres

L'hypothèse qui sous-tend les calculs dans cette section est qu'il y a une corrélation entre l'agencement des structures secondaires et la recherche des bonnes boucles. L'idée est de savoir jusqu'à quel point il faut donner d'informations sur la bonne boucle, via l'espace dièdre, pour que TRIP améliore substantiellement sa capacité de prédiction.

Avant de poursuivre, marquons un point d'arrêt pour répondre à la question: «Quel est l'utilité de ce genre de calcul ? Si on donne en entrée toute la réponse, il n'y a plus besoin de chercher».

L'utilité d'un logiciel comme TRIP sera de faire le pont entre des données expérimentales et une structure tri-dimensionnelle utilisable en dynamique moléculaire. Une méthode strictement théorique est impraticable dans l'état actuel de nos connaissances. Dans cette perspective, il devient intéressant de traduire le maximum de données expérimentales en entrée pour le logiciel. Deux genres de données sont utiles. Les données statistiques sur une banque de données peuvent fournir des potentiels de forces moyennes (section 1.2.3). Des données expérimentales directes peuvent fournir en peu de temps une information

ciblée sur la protéine. Les questions qui restent sont de savoir si l'information expérimentale est assez univoque pour constituer un filtre, et si elle est facilement extractible (le but ultime étant de battre en temps la RMN et la cristallographie). Par rapport à ces questions, les deux sections qui suivent prennent tout leur sens lorsqu'on sait qu'avec la RMN nous pouvons extraire des régions probables d'angles dièdres pour tous les acides aminés d'une protéine^{65 66 67}. Plus exactement, nous obtenons une zone de probabilité autour d'un certain couple d'angles, un peu comme MOLFIX fait avec ses nuages⁶⁸. Il serait très simple de convertir ces données expérimentales en carte MOLFIX.

Dernier point, en ce qui concerne la molécule de référence pour les RMS. Le fait que la conformation native ne soit pas au minimum d'énergie était dérangent. Nous avons cru bon d'obtenir une conformation relaxée de la myoglobine. Une conformation relaxée veut dire minimiser la structure dans notre potentiel. Un calcul TRIP a donc été lancé avec comme seule liste d'angles dièdres un nuage 5×5 autour de l'angle MOLFIX, sans aucune carte de base. Une première conformation, que j'appelle NatCalcBL dans le tableau XIII, a été obtenue. Son énergie est de -214. Cette structure a servi à la fois comme base pour mes angles MOLFIX et comme référence pour le RMS. Nous nous sommes ensuite ravisé pour le RMS, et avons réadopté la conformation cristallographique comme référence. Seul le tableau XIII souffre de cette discontinuité. Plus tard, nous sommes passé à une seconde structure MOLFIX de référence, car sa déviation RMS par rapport à la cristallographique était autour de 1Å au lieu de 2,5, pour une même valeur d'énergie.

3.5.1 Carte avec les points MOLFIX avec ou sans choix de structures secondaires

On commence doucement par l'ajout d'un seul point MOLFIX. Nous obtenons le tableau XIII. Étant donné la référence différente pour le calcul du RMS, il est difficile de commenter adéquatement les différences avec les autres tableaux. On peut constater que les valeurs de RMS et de RMS minimum sont plus grandes

que celles du tableau IX. Au niveau des énergies cependant, on obtient les mêmes résultats qu'avec des hélices idéales et un point MOLFIX. On parle des mêmes énergies, car si on annule l'amélioration des interactions inter-hélices, on tombe sur les mêmes valeurs.

La conclusion logique est que le modèle des hélices cylindriques est un petit peu plus efficace dans les premiers temps de la recherche conformationnelle. En effet, lorsqu'on regarde les valeurs des calculs interactifs, les valeurs de RMS et d'énergies deviennent meilleures que celles du tableau IX. Un peu comme si les hélices cristallographiques demandaient plus d'efforts dans la recherche de la «bonne» boucle. Dans cette perspective, il serait intéressant de suivre l'évolution des hélices durant la minimisation, afin de peut-être voir une transition abrupte dans leur utilisation.

Bien sûr, TRIP n'a visiblement pas atteint les conformations du tableau X. Ma conclusion est donc à prendre avec un bémol.

Un dernier point ici serait de remarquer la progression du rayon de giration. Comme la structure MOLFIX a un rayon de giration plus petit, on s'attend à ce que plus les conformations au hasard utiliseront les angles MOLFIX, plus elles seront petites. C'est ce que nous voyons à la colonne cinq. Sans choix d'hélice, les conformations deviennent de plus en plus petites au fur et à mesure que la proportion des angles MOLFIX augmente. Encore une fois ici, je ne pense pas que TRIP utilise sciemment les points MOLFIX. Il les utilise parce qu'ils sont dans sa liste, mais ne les surutilise pas en tant que «bonnes réponses». Lorsqu'on regarde le rayon de giration des calculs avec choix d'hélices, on s'aperçoit que TRIP préfère de plus grosses conformations. Il est incapable de faire converger d'un coup toutes les boucles, donc il erre dans des minimums locaux.

Le tableau XIV nous renseigne sur la répartition de l'énergie à la fin d'un calcul. On voit bien que les interactions autres que hélice↔hélice sont constantes, peu importe le type de calcul. Le vrai gain d'énergie se situe dans le bon agencement des hélices. Si TRIP avait trouvé les bonnes boucles, on s'attend à ce que les

TABLEAU XII. Effets des structures secondaires MOLFIX sur un calcul typique de TRIP sur la myoglobine.

Carte de base	Interactif	<RMS>	RMS Min.	<Énergie>	<Meilleure Énergie>
534	non	12,36	7,44	-158,19	-185,83
100	non	11,16	6,19	-170,13	-200,09
100	oui	11,10	8,11	-172,49	-195,73

TABLEAU XIII. Effets des couples d'angles MOLFIX avec les structures secondaires MOLFIX sur un calcul typique de TRIP sur la myoglobine.

Carte	Inter.	<RMS>*	<RMS> Min.	< R_g >	<Énergie>	<Meilleure Énergie>
NatCalcBL		2,32		13,80	-214,17	
MOLFIX3		2,50		14,22	-223,08	
100	non	13,12	8,35	14,96	-170,19	-196,70
50	non	14,10	11,23	14,92	-175,41	-196,92
18	non	13,51	12,62	14,50	-171,60	-186,87
100	oui	12,37	9,35	15,39	-176,06	-201,09

* par rapport au fichier MOLFIX d'un calcul TRIP avec nuage MOLFIX (22/02/97)

interactions hélice↔boucle soient réduites de moitié, comme dans le tableau VIII. Encore une fois, le fait de ne pas avoir trouvé les bonnes boucles semble limiter la capacité des calculs sans choix à trouver un minimum local. Il y a une corrélation entre les boucles et les hélices.

La figure 16 montre l'utilisation de l'espace dièdre pour la boucle quatre de la myoglobine. Nous n'observons aucun patron distinctif. TRIP n'a pas fait converger ces résidus vers une région stable. Cependant, si on compare avec la figure 11, nous voyons que des pics un peu plus imposants existent. Sans parler de convergence, nous pourrions qualifier ceci de confluence vers quelques conformations de boucles.

3.5.2 Carte avec nuage MOLFIX, avec ou sans choix de structures secondaires

Sachant qu'il y a des corrélations entre l'obtention des bonnes boucles et l'agencement des structures secondaires, il devient intéressant d'étudier l'impact de la densité et de la grandeur des nuages MOLFIX. En effet, le fait d'avoir des nuages crée des listes de triplets qui sont plus abondants au voisinage du triplet natif. TRIP peut donc plus facilement converger vers les bonnes boucles, et de ce fait mieux agencer les structures secondaires.

Le tableau XV rassemble quelques-uns de mes résultats.

Comparons dans un premier temps les tableaux X et XV pour les cartes 534. Avec un nuage 5×5 , TRIP donne sensiblement les mêmes résultats en RMS et en énergie. TRIP ne conflue pas vers les nuages MOLFIX. Cependant, avec un nuage 10×10 , TRIP abaisse de 10 points son énergie et de 3 points son RMS. Les hélices natives semblent donc affecter la capacité de TRIP à utiliser les nuages MOLFIX. Le rayon de giration moyen nous confirme que TRIP utilise plus sciemment les nuages MOLFIX, c'est-à-dire un peu dans toutes les boucles, car celui-ci est très bas pour une carte 534. Il est à remarquer que la moyenne des RMS minimums est aussi très basse pour une carte 534. Les nuages MOLFIX combinés avec les hélices MOLFIX portent fruit.

TABLEAU XIV. Ventilation de l'énergie pour un calcul TRIP sur la myoglobine avec et sans choix de structure secondaire, avec les couples d'angles MOLFIX.

Carte	<Hélice↔Hélice>	<Hélice↔Boucle>	<Boucle↔Boucle>
100	-178,32	13,33	5,03
100+MF	-189,04	13,63	5,22
100+MF+choix	-195,33	13,09	6,18

TABLEAU XV. Effets des nuages MOLFIX sur les boucles, combinés avec des hélices MOLFIX, sur un calcul typique de TRIP sur la myoglobine.

Carte	<RMS>	<RMS Min.>	< R_g >	<Énergie>	<Meilleure Énergie>
534+5×5	12,45	9,23	15,34	-157,45	-189,59
534+10×10	10,03	6,53	14,71	-169,77	-205,54
100+5×5	7,45	2,53	14,74	-178,40	-216,87
50+10×10(1/2)*	7,27	4,31	14,22	-200,33	-234,75
100+10×10	4,55	1,08	13,79	-226,86	-244,92
0+10×10†	2,67	1,70	14,12	-219,08	-225,90

* Uniquement les hélices B, G et H

† Source de la deuxième référence MOLFIX

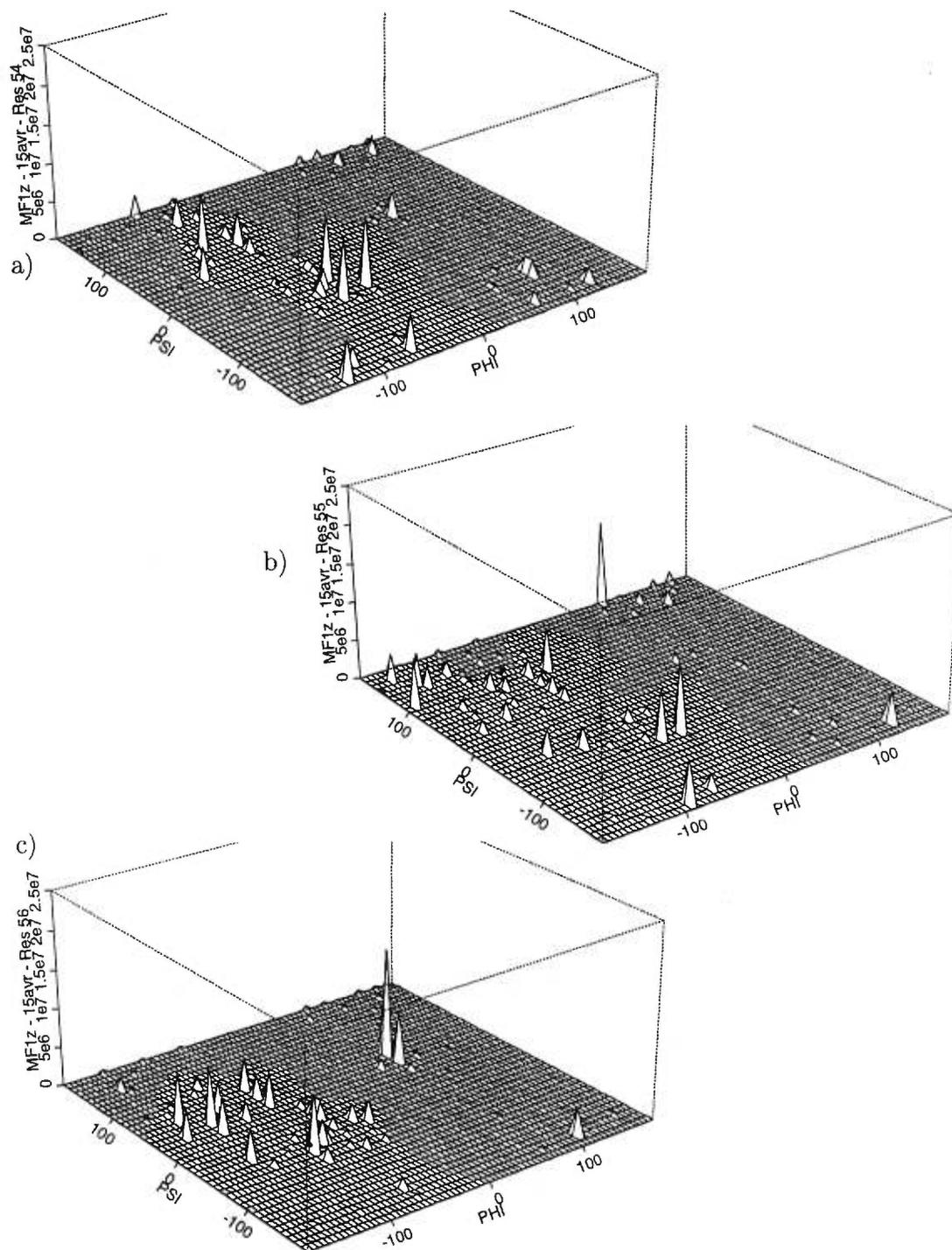


FIGURE 16. Utilisation des couples d'angles dièdres des cartes avec points MOLFIX. Boucle F de la myoglobine, soit les résidus 54, 55 et 56. Une unité dans l'axe z correspond à une conformation acceptée.

Si on compare maintenant les cartes $100+5\times 5$ de tous les tableaux, nous remarquons que les nuages MOLFIX sans hélices MOLFIX sont 10 unités d'énergie plus hautes qu'avec les hélices MOLFIX seules, et que les hélices MOLFIX seules sont 10 unités d'énergie plus hautes qu'avec les nuages MOLFIX combinés. Malgré ce gain de 20 unités d'énergie, nous ne sommes pas encore dans la région des -200. Pourtant, les moyennes de RMS et surtout des RMS minimums sont très basses. Nous pouvons ici vraiment parler de convergence dans l'utilisation des nuages MOLFIX. TRIP surutilise clairement les points dans le nuage MOLFIX. La figure 17 montre un patron d'utilisation typique d'une convergence vers les nuages MOLFIX.

Passons à la carte $50+10\times 10(1/2)$. Ces calculs utilisent un nuage dix par dix dont l'incrément est de deux degrés. Ça donne donc 50 points de plus que les 50 de la carte de base. La vraie particularité de ce calcul est que seules les hélices natives les plus appréciées par TRIP (tableau XVI), c'est-à-dire B,G et H, ont été utilisées. Le résultat est surprenant, puisqu'il donne des ensembles de structures dans la région du -200. Ainsi, avec 3 hélices sur 8 dans la configuration native, TRIP réussit à «condenser» les protéines, comme le prouve le rayon de giration moyen de 14,22. De plus, il est possible qu'en remplaçant les hélices idéales par les hélices natives à la toute fin du calcul, le RMS moyen serait beaucoup plus bas que 7,27, car les bonnes boucles ont été trouvées. À noter qu'un RMS de 6 Å pour la myoglobine implique que la structure générée possède la bonne topologie générale.

Finalement, les deux dernières rangées du tableau XV me permettent d'une part d'affirmer que grâce à MOLFIX, TRIP a trouvé le minimum le plus bas en énergie obtenu jusqu'à ce jour pour la myoglobine avec le potentiel Bryant&Lawrence, et d'autre part que ce minimum inclut des points qui ne proviennent pas du nuage. Les plus basses structures en RMS et en énergie sont obtenues en laissant de la flexibilité dans les boucles.

Nous allons poursuivre avec le tableau XVI, où la capacité de choisir ses hélices est

de nouveau implémenté dans TRIP. Les résultats bruts des énergies se retrouvent dans les tableaux qui suivent. La première découverte est que ce ne sont pas toutes les hélices qui sont préférées. Ce phénomène m'échappe un peu, puisqu'en fait les structures à la sortie sont généralement toutes avec des nuages natifs. Une partie de la réponse vient peut-être du fait que ce pourcentage n'est pas un pourcentage final. Les résultats sont colligés à chaque cycle moléculaire, ce qui implique que si l'ensemble a globalement opté pour une hélice très tôt dans la minimisation, son score sera beaucoup plus grand. C'est dans l'optique de tester cet effet de condensation que le calcul $50+10 \times 10(1/2)$ du tableau XV a été fait, avec les résultats que l'on connaît.

La deuxième découverte est qu'un nuage avec le même nombre de points, mais couvrant une plus grande superficie, a un comportement tout ou rien. En effet, l'utilisation moyenne des hélices avec les nuages 30×30 n'est pas intermédiaires aux deux autres. Soit TRIP utilise l'hélice comme le nuage 10×10 , soit il l'utilise comme un sans nuage. Le calcul des écarts-types nous confirme que la variance sur la moyenne est beaucoup plus grande avec les nuages 30×30 qu'avec les deux autres.

Fort de tous ces résultats sur la façon dont TRIP utilise les informations de MOLFIX, nous avons poussé plus loin l'application des nuages. Dans une optique d'application réelle, nous ne disposerons pas de la protéine cristallographiée. Il faut trouver d'autres source d'information susceptible d'être traduite en nuage. Nous entrevoyons deux sources à l'heure actuelle. D'une part, il y a la modélisation par homologie, d'autre part il y a une technique de RMN⁶⁷ qui utilise le déplacement chimique du ^{13}C . Pour ce faire, il me fallait connaître quels étaient les effets de la densité du nuage et de la dimension du nuage sur le calcul.

D'une part, le résultat expérimental n'est pas toujours de qualité. Une façon simple d'accorder plus ou moins de valeur à ce résultat est d'augmenter ou de diminuer la proportion des points dans le nuage versus les points de la carte de base. Le tableau XVII indique clairement que plus le poids du nuage

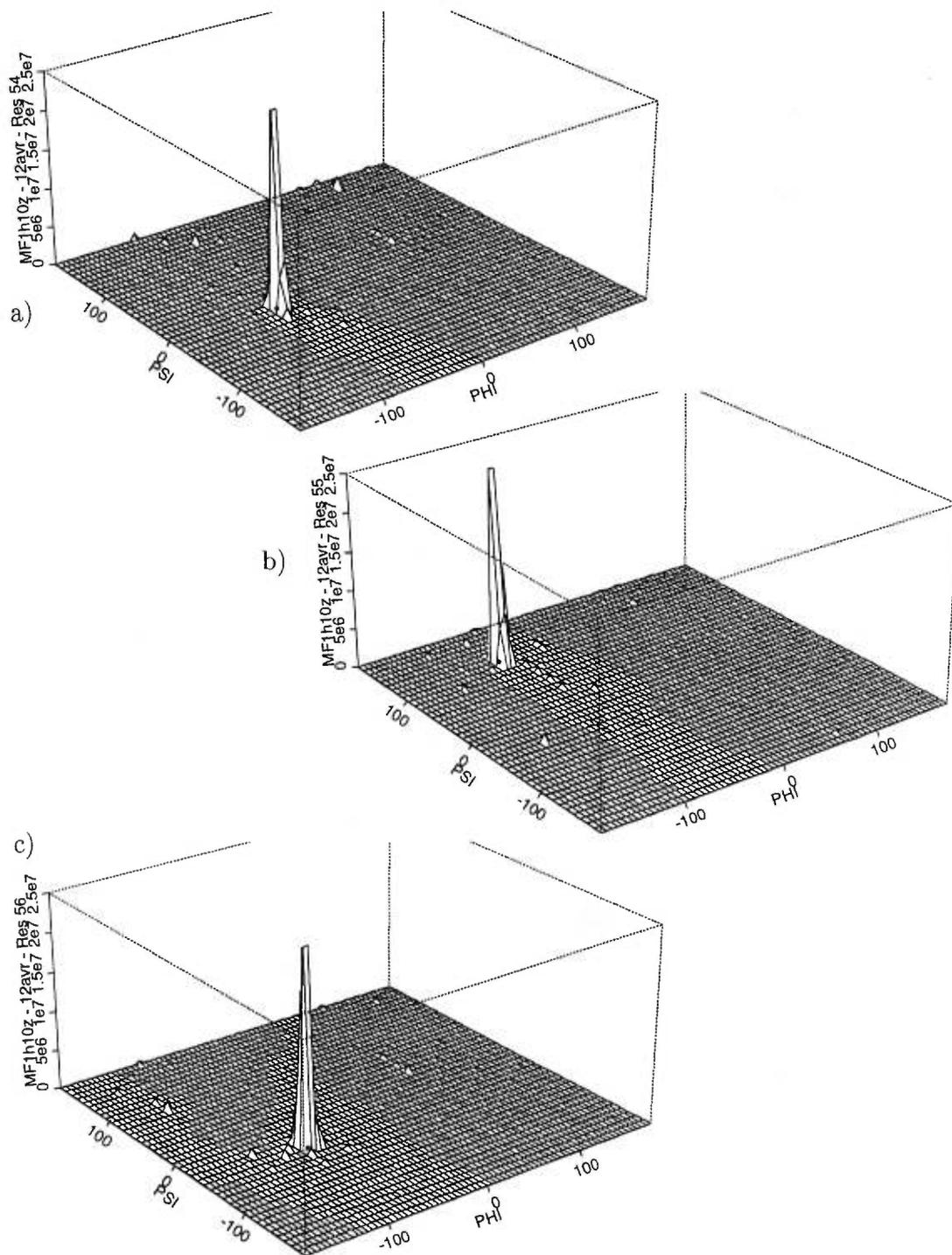


FIGURE 17. Utilisation des couples d'angles dièdres des cartes avec nuage MOLFIX. Boucle F de la myoglobine, soit les résidus 54, 55 et 56. Une unité dans l'axe z correspond à une conformation acceptée.

est important, plus TRIP converge facilement vers la structure native. Et les moyennes d'énergie et les moyennes de RMS empirent avec la décroissance du poids des nuages. Le résultat important ici est qu'un pourcentage de 33% est suffisant pour obtenir un ensemble de structures possédant la bonne topologie. Ce pourcentage devrait être le minimum dans une optique d'application, dans le sens où un pourcentage plus bas indiquerait une trop grande méfiance du résultat. D'autre part, les nuages expérimentaux pourront parfois être assez diffus. Il ne serait pas rare d'avoir un nuage 30×30 , quoique la moyenne se situe plutôt autour de 15×30 . Remarquez que le nuage est asymétrique. Ceci est dû au fait que le nuage est un amalgame des moyennes de distribution expérimentale et de l'incertitude sur la mesure, et l'incertitude sur l'angle ϕ est toujours plus petite]. Le tableau XVIII montre que TRIP minimise très bien l'énergie, peu importe la superficie du nuage. Quant au RMS, la moyenne des structures possède la bonne topologie, avec systématiquement quelques très bonnes structures finales. Il est probablement logique que la moyenne des RMS diminue, car plus on s'éloigne du point MOLFIX, plus la liste de triplets et de boucles divergera de la structure native. Néanmoins, le fait que l'énergie reste sensiblement identique est préoccupant. Il renforce notre scepticisme sur le potentiel d'énergie. En effet, TRIP ne fait pas la différence entre une minimisation qui donne de bons et d'excellents résultats en RMS. Lors d'une prédiction *ab initio*, l'utilisateur ne pourra pas savoir dans quel état de figure il se situe.

La leçon des deux tableaux XVII et XVIII est que l'efficacité d'un nuage tient dans sa concentration autour d'un point, car c'est ainsi que son impact sur les listes de triplets est le plus fort.

Finalement, le tableau XIX vient tempérer nos attentes. La recherche conformationnelle reste un problème NP complet. Lui fournir de l'information via l'espace dièdre n'est pas la panacée universelle. C'est un moyen parmi d'autres pour améliorer l'efficacité de la recherche. Grosso-modo, dans un contexte d'application, cette amélioration se chiffre par une baisse d'une quarantaine

TABLEAU XVI. Pourcentage moyen d'utilisation des hélices natives, pendant les calculs TRIP interactifs avec nuages MOLFIX sur la myoglobine.

Hélice	Point MF	Nuage 10×10	Nuage 30×30 (1/3)
1	57,6	50,3	63,5
2	76,4	80,6	82,5
3	58,3	72,2	65,3
4	40,3	38,0	42,8
5	36,0	56,6	37,5
6	82,3	86,5	94,9
7	88,1	96,9	95,8
8	68,2	93,9	86,2

La carte de base à 100 points

TABLEAU XVII. Effets du poids des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, avec choix de structures secondaires .

Nuage MOLFIX	<RMS>	RMS	<Énergie>
%		minimum	
50	4,14	1,07	-226,358
33	6,48	5,07	-218,01
20	9,76	4,93	-173,33
6,4	11,66	7,38	-171,74

d'unités d'énergie, et une amélioration d'au moins un tiers du RMS.

Jusqu'ici, tous les résultats concernaient des calculs fait sur la myoglobine. Certains calculs similaires ont été réalisés avec un fragment des unités L7/L12 de ribosomes (1CTF). Le tableau XX montre les résultats importants.

Nous voyons très bien au tableau XX que les structures secondaires natives améliorent dramatiquement la recherche de la conformation native. Avec les feuillets et hélices natifs, TRIP trouvent les boucles natives.

3.6 Carte avec des nuages au hasard

Dans cette dernière section, nous abordons le problème qui surviendrait si les nuages fournis à TRIP par RMN n'étaient tout simplement pas bons. On peut tenter de fourvoyer TRIP de plusieurs façons avec des nuages mal placés. Le tableau XXI contient deux types de nuages au hasard qui résumant bien mes découvertes. Le premier type de nuage au hasard est un nuage identique pour tous les résidus, et centré sur le point $\phi = -65$ et $\psi = -40$. La moyenne d'énergie obtenue n'est pas catastrophique. Elle est du même ordre qu'une recherche TRIP sans nuage. Ce qui est important ici est de constater que le rayon de giration reste compact. Si TRIP avait surutilisé les points dans le nuage, les structures générées auraient été beaucoup plus allongées, pour être à la limite une immense hélice α . Donc nous savons que TRIP peut ne pas utiliser un nuage, même s'il constitue 50% de son espace dièdre.

Le deuxième type de nuage au hasard est vraiment au hasard, c'est-à-dire que chaque résidu à un nuage différent choisi au hasard. Si les structures générées avaient surutilisé ces nuages, nous aurions nécessairement des valeurs énormes d'énergie. Or le tableau XXI montre bien que ce n'est pas le cas. TRIP semble en mesure de minimiser l'énergie sans ressentir de frustration de l'espace dièdre. La moyenne des énergies est semblable à celle obtenue en absence de nuage, et le rayon de giration moyen reste celui d'une protéine globulaire. Ceci montre l'avantage d'avoir adopté la voie spatiale plutôt qu'énergétique pour filtrer

TABLEAU XVIII. Effets de la superficie des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, avec choix de structure secondaire.

Nuage	Pas	<RMS>	RMS	<Énergie>
			min.	
10×10	1°	4,14	2,52	-226,36
20×20	2°	6,37	3,05	-225,03
30×30	3°	7,22	3,56	-228,31

Carte de base à 100 points

TABLEAU XIX. Effets du poids des nuages MOLFIX sur un calcul typique de TRIP sur la myoglobine, sans choix de structure secondaire.

Carte	<RMS>	<RMS Min.>	< R_g >	<Énergie>	<Meilleure Énergie>
50+10×10(1/2)	6,56	4,85	14,01	-212,03	-232,14
100+10×10(1/2)	9,76	5,46	14,90	-173,33	-195,89
372+10×10(1/2)	11,66	8,40	15,18	-171,74	-202,91

TABLEAU XX. Effets synergiques des hélices cristallographiques et des nuages MOLFIX dans les boucles sur les prédictions de TRIP pour la 1CTF.

Structure secondaire	<RMS>	<Énergie>
Natif du PDB		-63,09
Idéale [†]	7,95	-64,42
Native	3,84	-65,01
Native [†]	1,46	-64,48

Tous les calculs ont une carte de base à 50 points. [†]Calcul fait avec un nuage 10×10.

l'espace dièdre. Comme il n'y a aucune composante locale à l'énergie, TRIP ne perçoit aucune influence directe des nuages. Les nuages agissent indirectement en peuplant différemment l'espace des boucles.

La grande conclusion des nuages au hasard est que si le nuage est mauvais, TRIP fait comme s'il n'existait pas.

Finalement, la figure 18 conclura cette section. Dans ce calcul, un nuage est placé sur chaque résidu qui commence ou termine une boucle, pour voir l'impact sur les résidus subséquents. En observant les figures 11, 16, 17 et 18, deux conclusions s'imposent. La première est qu'il y a une corrélation. Le résidu 55 a utilisé plus qu'à l'accoutumée la région «feuille β ». La deuxième conclusion est que la corrélation s'éteint rapidement. Le résidu 55 est seul à n'avoir pas de nuage entre les résidus 54 et 56. Pourtant, nous ne voyons pas un pic élevé au-dessus du point natif. L'ajout de nuage semble avoir un effet très local dans l'espace dièdre.

3.7 Synthèse sur l'espace dièdre des boucles combinées aux structures secondaires

En résumé, il y a une corrélation entre la recherche du meilleur agencement des structures secondaires et celui des meilleures boucles. Si une des deux recherches est aidée, l'autre en ressent des effets. Dans ce travail, je m'intéressais à manipuler les espaces dièdres pour améliorer la recherche de bonnes boucles.

La stratégie la plus séduisante est de donner à TRIP des nuages de points autour du couple d'angles de la protéine native. Nous avons vu que ces nuages aident dramatiquement la recherche de bonnes boucles, et conséquemment donnent des structures à toute fin pratique identiques à la native.

Dans une perspective d'application directe de mes recherches, nous avons vu qu'un grand nombre de nuages concentrés provenant de techniques de RMN⁶⁵
⁶⁶ ⁶⁷ ⁶⁸ permettraient à TRIP de systématiquement obtenir la bonne topologie pour l'ensemble de ses structures finales. Mais qu'en revanche, de mauvais nuages ne nuiraient pas à la recherche conformationnelle. Nous sommes donc dans une situation gagnante à tous coups.

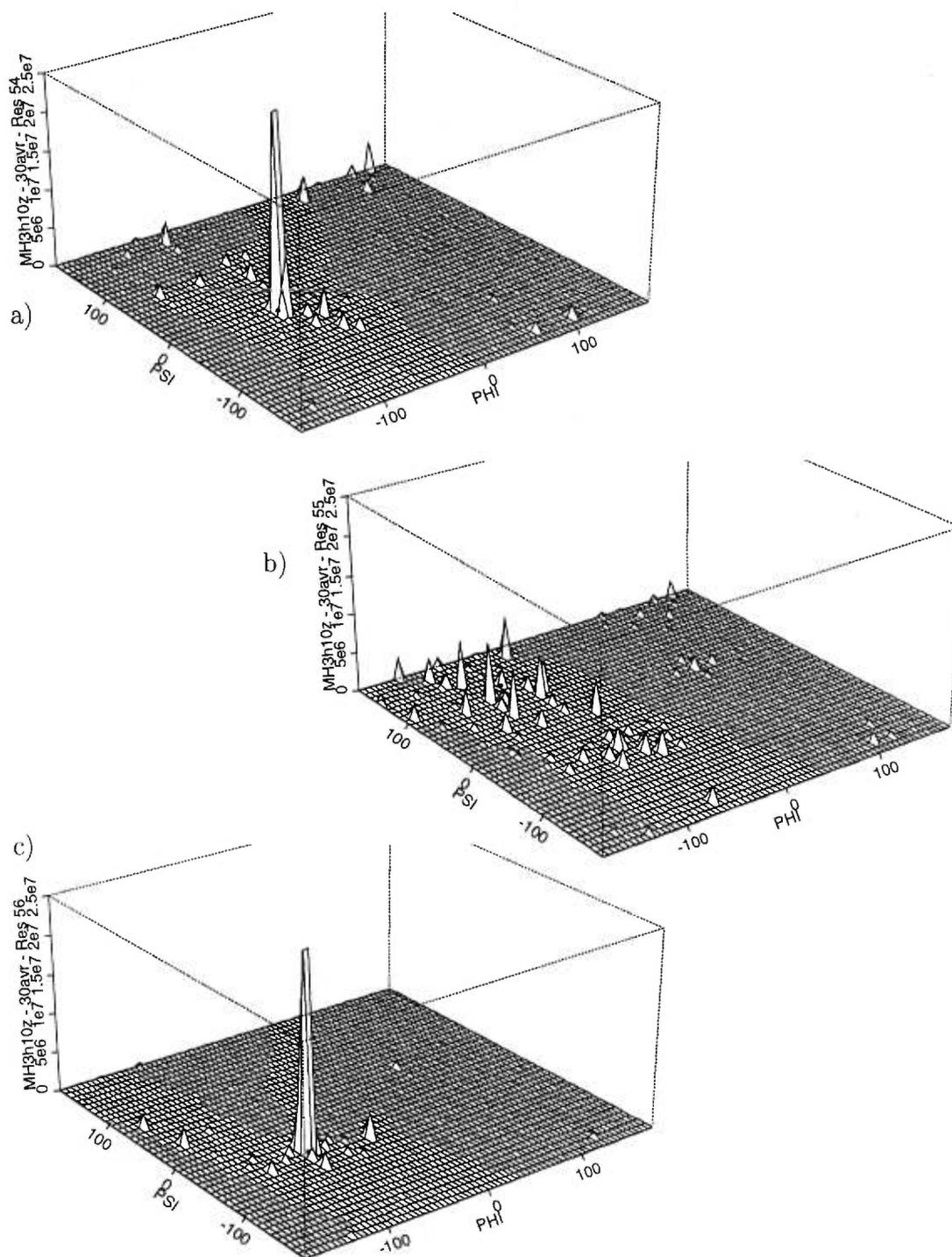


FIGURE 18. Utilisation des couples d'angles dièdres des cartes avec nuage sur les bouts des boucles. Boucle F de la myoglobine, soit les résidus 54, 55 et 56. Une unité dans l'axe z correspond à une conformation acceptée.

TABLEAU XXI. Effets de nuages au hasard sur un calcul typique de TRIP sur la myoglobine.

Type de nuage	$\langle \text{RMS} \rangle$	RMS min.	$\langle R_g \rangle$	$\langle \text{Énergie} \rangle$
MOLFIX	4,55	1,08	13,79	-226,86
Hélice idéale	12,55	8,04	14,89	-153,67
Hasard	12,42	6,86	15,13	-156,36

Tous les nuages font 10×10 sur une carte de base à 100 points.

CHAPITRE 4

Conclusion

Le but de mon travail était de comprendre l'effet d'une distribution ciblée dans l'espace dièdre. Il a été vu tout d'abord que dans sa version de départ, TRIP était dépassé par la dimension du problème. Seule la carte à 50 points commence à alléger l'espace conformationnelle pour TRIP, sans générer trop de boucles non-naturelles.

Les cartes selon l'espèce chimique donnent des résultats mitigés. À part les cartes PRO+NPRO, l'amélioration de la prédiction n'est pas significative.

Lorsqu'on ajoute un point correspondant au couple d'angles natifs, TRIP n'est pas en mesure de converger vers la structure native. On note tout de même une amélioration de la prédiction pour des cartes de bases réduites.

L'utilisation de nuage de points donne les meilleurs résultats. Même avec des hélices idéales, nous obtenons des ensembles de structures de plus en plus proches de la native, au fur et à mesure que nous réduisons la carte de base. Nous notons cependant un plafonnement dans l'amélioration de la prédiction, causé par un mauvais agencement des structures secondaires.

Par suite logique, nous avons testé l'emploi des structures secondaires, en donnant un degré de liberté sur le choix de la structure secondaire. Nous observons que sans apport d'informations dans les boucles, TRIP ne converge pas vers la structure native. Il y a un effet synergique entre le bon agencement des structures secondaires et la recherche de bonnes boucles. Cet effet synergique a une tendance à la condensation. C'est-à-dire que lorsqu'une structure commence à utiliser les bonnes boucles, le reste de la protéine descend rapidement vers le puits de potentiel.

Dans une optique d'application directe, il était intéressant de connaître l'impact

du poids et de la superficie des nuages. Nous avons trouvé qu'il y a une relation quasi linéaire entre le poids relatif des nuages et la capacité de TRIP à trouver la structure native. D'autre part, pour un poids donné, la superficie des nuages ne semble pas affecter la moyenne d'énergie. Cependant, le RMS augmente au fur et à mesure qu'on élargit le nuage. Finalement, l'essai de nuages parfaitement aléatoires nous convainc que TRIP ne subit aucune pression énergétique avec les nuages. Si l'information du nuage est utile, TRIP l'utilisera. Si elle est inutile, TRIP ne l'utilisera pas. Nous sommes dans une situation de gain à tous coups. Nous terminerons cette conclusion par des perspectives d'avenir de ces recherches. Premièrement, il est possible de diminuer des problèmes locaux par des filtres géométriques locaux. Le but initial de mes recherches a donc été en partie satisfait. Mais TRIP ne pourra pas éliminer tous les encombrements stériques par cette approche. En fait, je pense que tant que TRIP ne gèrera pas les chaînes latérales, nous ne pourrons pas améliorer ce résultat.

Deuxièmement, il y a du potentiel encore inexploré dans les cartes par acide aminé. L'amélioration notable avec les cartes non-uniformes donne l'espoir qu'une carte optimal existe pour chaque résidu. Une utilisation plus intelligente du PDB devrait donner des résultats dans ce sens. On peut même imaginer une carte qui se modifierait selon ses voisins immédiats.

Troisièmement, le fait d'avoir découvert une corrélation entre les structures secondaires et les boucles nous amène à penser qu'on pourrait minimiser les structures secondaires en même temps que les boucles. En effet, l'addition de la possibilité d'évolution des structures secondaires permettrait à TRIP de mieux gérer les feuillets β . Les feuillets β , par leur nature très diverse, souffrent beaucoup de la rigidité de TRIP. Une façon souple de régler le problème serait de dynamiquement transformer l'espace dièdre des résidus dans les feuillets durant la minimisation. On pourrait ainsi obtenir des feuillets tordus, ce qui est impossible présentement sans utiliser MOLFIX.

Finalement, l'ultime accomplissement serait de diviser les résidus en trois groupes

au lieu de deux. Il y aurait des résidus correspondant à des centres de nucléation de structures secondaires, des résidus boucles, et des résidus mitoyens entre les deux. Durant la minimisation, un résidu pourrait tour à tour devenir boucle, mitoyen et puis structure secondaire. Ainsi, en manipulant dynamiquement l'espace dièdre, nous pourrions diminuer le nombre de points d'un résidu, et parallèlement modifier la nature de l'espace dièdre selon le besoin. Un mélange dynamique de carte non-uniforme par acide aminé qui évoluerait selon que le résidu soit mitoyen ou boucle, et de nuage selon qu'il soit dans l'un des trois états, décuplerait la puissance de TRIP.

RÉFÉRENCES

- [1] Anthony Keith Felts. Communication privée, (1996).
- [2] Per J. Kraulis. Molscrip: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* **24**, 946–950 (1991).
- [3] Wolfram Saenger. *Principles of nucleic acid structure*. Springer-Verlag, New York, (1984).
- [4] W.C.Jr. Johnson. Secondary structure of proteins through circular dichroism spectroscopy. *Annual Review of Biophysics and Biophysical Chemistry* **17**, 145–166 (1988).
- [5] L. Gomathi et S. Subramanian. Elucidation of secondary structures of peptides using high resolution nmr. *Current Science* **71**(7), 553–567 (1996).
- [6] A. McPherson. *The Preparation and Analysis of Protein Crystals*. Wiley, New York, (1982).
- [7] L.A. Amos, R. Henderson, et P.N. Unwin. Three-dimensional structure determination by electron microscopy of two-dimensional crystals. *Progress in biophysics and molecular biology* **39**, 183–231 (1982).
- [8] K. Wuthrich. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* **243**, 45–50 (1989).
- [9] D. Botstein et D. Shortle. Strategies and applications of in vitro mutagenesis. *Science* **229**, 1193–1201 (1985).
- [10] W.A. Eaton, V. Munoz, P.A. Thompson, C.K. Chan, et J. Hofrichter.

- Submillisecond kinetics of protein folding. *Current Opinion in Structural Biology* **7**(1), 10–14 (1997).
- [11] G. Fasman. *The development of the prediction of protein structure*. Plenum, New York, (1989). p. 193-316.
- [12] S.R. Presnell, B.I. Cohen, et F.E. Cohen. A segment based approach to secondary structure prediction. *Biochemistry* **31**, 983–993 (1992).
- [13] B. Rost et C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Structure, Function and Genetics* **19**, 55–72 (1994).
- [14] S.S. Rao et S.M. Bhattacharjee. Protein folding and spin glass. *Physica A* **224**(1-2), 279–286 (1996).
- [15] A.R. Dinner, A. Sali, M. Karplus, et E. Shakhnovich. Phase diagram of a model protein derived by exhaustive enumeration of the conformations. *Journal of Chemical Physics* **2**, 1444–1451 (1994).
- [16] V.S. Pande, A.Y. Grosberg, et T. Tanaka. On the theory of folding kinetics for short proteins. *Folding & Design* **2**(2), 109–114 (1997).
- [17] G.F. Berriz, A.M. Gutin, et E.I. Shakhnovich. Cooperativity and stability in a langevin model of proteinlike folding. *Journal of Chemical Physics* **106**, 9276–9285 (1997).
- [18] James Darnell, Harvey Lodish, et David Baltimore. *Molecular Cell Biology*. Scientific American Books, Inc., New York, 2^e edition, (1990).
- [19] I. Simon, G. Némethy, et H.A. Scheraga. Conformational energy calculations of the effects of sequence variations on the conformations of two tetrapeptides. *Macromolecules* **11**(4), 797–804 (1978).

- [20] I. Simon, L. Glasser, et H.A. Scheraga. Calculation of protein conformation as an assembly of stable overlapping segments: Application to bovine pancreatic trypsin inhibitor. *Proceedings of the National Academy of Science, USA* **88**, 1661–3665 (1991).
- [21] Jean Pelmont. *Enzymes*. Presse de l'Université de Grenoble, Aubenas d'Ardèche, (1993).
- [22] Thomas E. Creighton. *Proteins: Structures and Molecular Principles*. W.H. Freeman and Co., United States of America, 2^e edition, (1993).
- [23] Geoffrey Zubay. *Biochemistry*. Macmillan Publishing Co., New York, 2^e edition, (1988).
- [24] E.J. Milner-White. The partial charge of the nitrogen atom in peptide bonds. *Protein Science* **6**(11), 2477–2482 (1997).
- [25] G.N. Ramachandran, C. Ramakrishnan, et V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95–99 (1963).
- [26] G.N. Ramachandran et V. Sasisekharan. Conformation of polypeptides and proteins. *Advances in Protein Chemistry* **23**, 284–438 (1968).
- [27] C.B. Anfinsen. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
- [28] L.G. Presta et G.D. Rose. Helix signals in proteins. *Science* **240**, 1632–1640 (1988).
- [29] K.V. Soman, A. Karimi, et D.A. Case. Unfolding of an α -helix in water. *Biopolymers* **31**, 1351–1361 (1991).
- [30] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, et M.L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926–935 (1983).

- [31] K.A. Dill. Dominant forces in protein folding. *Biochemistry* **29**(31), 7133–7155 (1990).
- [32] Jacob Israelachvili. *Intermolecular and Surface Forces*. Academic Press, San Diego, 2^e edition, (1991).
- [33] J.L. Neira, B. Davis, A.G. Ladurner, A.M. Buckle, G.D. Gay, et A.R. Fersht. Towards the complete structural characterization of a protein folding pathway - the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor .2. direct evidence for a nucleation-condensation mechanism. *Folding & Design* **1**(3), 189–208 (1996).
- [34] Brooks, Bruccoleri, et Karplus. Charmm: A program for macromolecular energy, minimization and dynamics calculations. *Journal of Computational Chemistry* **4**(2), 187–217 (1983).
- [35] S. Weiner, P. Weiner, et Kollman Case. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* **106**, 765–784 (1984).
- [36] M.P. Allen et D.J. Tildesley. *Computer simulation of liquids*. Oxford science publications, New York, (1987). p. 168.
- [37] R. Zhou et B.J. Berne. A new molecular dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex systems. *Journal of Chemical Physics* **103**, 9444 (1995).
- [38] D.W. Heermann. *Computer simulation methods in theoretical physics*. Springer-Verlag, Berlin, 2^e edition, (1990).
- [39] J.R. Gunn. Minimizing reduced-model proteins using a generalized hierarchical table-lookup potential function. *Journal of Physical Chemistry* **100**, 3264–3272 (1996).

- [40] A. Monge, E.J.P. Lathrop, J.R. Gunn, P.S. Shenkin, et R.A. Friesner. Computer modeling of protein folding: Conformational and energetic analysis of reduced and detailed protein models. *Journal of Molecular Biology* **247**, 995 (1995).
- [41] G. Casari et M.J. Sippl. Structure-derived hydrophobic potential. *Journal of Molecular Biology* **224**, 725–732 (1992).
- [42] S.H. Bryant et C.E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins: Structure, Function and Genetics* **16**, 92–112 (1993).
- [43] J. Skolnick, L. Jaroszewski, A. Kolinski, et A. Godzik. Derivation and testing of pair potentials for protein folding - when is the quasicheical approximation correct. *Protein Science* **6**(3), 676–688 (1997).
- [44] P.D. Thomas et K.A. Dill. Statistical potentials extracted from protein structures: How accurate are they ? *Journal of Molecular Biology* **257**, 457–469 (1996).
- [45] J.R. Gunn. Sampling protein conformations using segment libraries and a genetic algorithm. *Journal of Chemical Physics* **106**(10), 4270–4281 (1997).
- [46] A. Rabow et H.A. Scheraga. Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator. *Protein Science* **5**, 1800–1815 (1996).
- [47] J.T. Pedersen et J. Moult. Protein folding simulations with genetic algorithms and a detailed molecular description. *Journal of Molecular Biology* **269**(2), 240–259 (1997).
- [48] N. Metropolis. Monte carlo simulated annealing. *Journal of Chemical Physics* **96**, 768–780 (1953).

- [49] William H. Press, editor. *Numerical recipes in FORTRAN : the art of scientific computing*. Cambridge University Press, New York, 2^e edition, (1992).
- [50] H.A. Scheraga. Recent developments in the theory of protein folding - searching for the global energy minimum. *Biophysical Chemistry* **59**, 329–339 (1996).
- [51] Z. Li et H.A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Science, USA* **84**, 6611–6615 (1987).
- [52] A. Kolinski et J. Skolnick. Discretized model of proteins. i.monte carlo study of cooperativity in homopolypeptides. *Journal of Chemical Physics* **97**, 9412–9426 (1992).
- [53] R. Abagyan et M. Totrov. Biaised probability monte carlo conformational searches and electrostatic calculations for peptides and proteins. *Journal of Molecular Biology* **235**, 983–1002 (1994).
- [54] S. Kirkpatrick, C.D.Jr Geddat, et M.P. Vecchi. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
- [55] J.M. Yeomans. *Statistical Mechanics of Phase Transitions*. Oxford University Press, New York, (1992).
- [56] F. Cohen et M. Sternberg. On the prediction of protein structure: The significance of the root-mean-square deviation. *Journal of Molecular Biology* **138**, 321–333 (1980).
- [57] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* **A32**, 922–923 (1976).
- [58] W. Kabsch. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica* **A34**, 827–828 (1978).

- [59] J.R. Gunn et R.A. Friesner. Parallel implementation of a protein structure refinement algorithm. *Journal of Computational Chemistry* **17**(10), 1217–1228 (1996).
- [60] R.A. Laskowski, M.W. MacArthur, D.S. Moss, et J.M. Thornton. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291 (1993).
- [61] O. Lund, J. Hansen, S. Brunak, et J. Bohr. Relationship between protein structure and geometrical constraints. *Protein Science* **5**, 2217–2225 (1996).
- [62] M.J. Rooman, J.-P. Kocher, et S.J. Wodak. Prediction of protein backbone conformation based on seven structure assignments. *Journal of Molecular Biology* **221**, 961–979 (1991).
- [63] J. Kuszewski, A.M. Gronenborn, et G.M. Clore. Improving the quality of nmr and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Science* **5**(6), 1067–1080 (1996).
- [64] B. Cheng, A. Nayeem, et H.A. Scheraga. From secondary structure to three-dimensional structure improved dihedral angle probability distribution function for use with energy searches for native structures of polypeptides and proteins. *Journal of Computational Chemistry* **17**(12), 1453–1480 (1996).
- [65] J. Heller, D.D. Laws, M. Tomaselli, D.S. King, D.E. Wemmer, A. Pines, R.H. Havlin, et E. Oldfield. Determination of dihedral angles in peptides through experimental and theoretical studies of alpha-carbon chemical shielding tensors. *Journal of the American Chemical Society* **119**(33), 7827–7831 (1997).
- [66] D.W. Yang, R. Konrat, et L.E. Kay. A multidimensional nmr experiment for measurement of the protein dihedral angle psi based on cross-correlated

relaxation between (h-alpha-13c-alpha)-h-1 dipolar and c-13' (carbonyl) chemical shift anisotropy mechanisms. *Journal of the American Chemical Society* **119**(49), 11938–11940 (1997).

[67] D. Wishart. Communication privée, avril 1997.

[68] M.J.J. Blommers et W. Jahnke. Direct measurement of dihedral angles with high-resolution nmr spectroscopy. *Angewandte Chemie. Int. Ed. in English* **37**(4), 456–458 (1998).

APPENDICE A

Fichier PDB des protéines cobayes

```

HEADER      OXYGEN STORAGE                      27-AUG-81  1MBO
COMPND      MYOGLOBIN (OXY, { P*H 8.4)
SOURCE      SPERM WHALE (PHYSETER CATODON)
AUTHOR      S.E.V.PHILLIPS
REVDAT      4  30-SEP-83 1MBOC  1      REVDAT
REVDAT      3  19-APR-83 1MBOB  1      REMARK
REVDAT      2  07-MAR-83 1MBOA  1      SEQRES
REVDAT      1  03-FEB-82 1MBO   0
JRNL        AUTH  S.E.V.PHILLIPS
JRNL        TITL  STRUCTURE AND REFINEMENT OF OXYMYOGLOBIN AT 1.6
JRNL        TITL 2 ANGSTROMS RESOLUTION
JRNL        REF   J.MOL.BIOL.                      V. 142  531 1980
JRNL        REFN  ASTM JMOBAK  UK ISSN 0022-2836          070
REMARK      1
REMARK      1 REFERENCE 1
REMARK      1 AUTH  S.E.V.PHILLIPS,B.P.SCHOENBORN
REMARK      1 TITL  NEUTRON DIFFRACTION REVEALS OXYGEN-HISTIDINE
REMARK      1 TITL 2 HYDROGEN BOND IN OXYMYOGLOBIN
REMARK      1 REF   NATURE                          V. 292  81 1981
REMARK      1 REFN  ASTM NATUAS  UK ISSN 0028-0836          006
REMARK      1 REFERENCE 2
REMARK      1 AUTH  S.E.V.PHILLIPS
REMARK      1 TITL  STRUCTURE OF OXYMYOGLOBIN
REMARK      1 REF   NATURE                          V. 273  247 1978
REMARK      1 REFN  ASTM NATUAS  UK ISSN 0028-0836          006
REMARK      1 REFERENCE 3
REMARK      1 AUTH  S.E.V.PHILLIPS
REMARK      1 TITL  THE STRUCTURE OF OXY-*MYOGLOBIN
REMARK      1 REF   ACTA CRYSTALLOGR.,SECT.A        V. 34  55 1978
REMARK      1 REF  2 (SUPPLEMENT)
REMARK      1 REFN  ASTM ACACBN  DK ISSN 0567-7394          108
REMARK      2
REMARK      2 RESOLUTION. 1.6 ANGSTROMS.
REMARK      3
REMARK      3 REFINEMENT. CONSTRAINED RECIPROCAL-SPACE LEAST-SQUARES

```

REMARK 3 (JACK-LEVITT).
REMARK 4
REMARK 4 ERROR IN ATOMIC POSITIONS IS ESTIMATED TO BE APPROXIMATELY
REMARK 4 0.1 ANGSTROMS WHERE B IS LESS THAN 10.0, INCREASING TO 0.3
REMARK 4 ANGSTROMS OR MORE WHERE B IS GREATER THAN 40.0.
REMARK 5
REMARK 5 THE QUANTITY GIVEN IN THE OCCUPANCY FIELD OF THE ATOM AND
REMARK 5 HETATM RECORDS IS THE ELECTRON COUNT.
REMARK 6
REMARK 6 CORRECTION. CORRECT SEQRES RECORD. 07-MAR-83.
REMARK 7
REMARK 7 CORRECTION. INSERT NEW PUBLICATION AS REFERENCE 1 AND
REMARK 7 RENUMBER THE OTHERS. 19-APR-83.
REMARK 8
REMARK 8 CORRECTION. INSERT REVDAT RECORDS. 30-SEP-83.

SEQRES	1	153	VAL	LEU	SER	GLU	GLY	GLU	TRP	GLN	LEU	VAL	LEU	HIS	VAL
SEQRES	2	153	TRP	ALA	LYS	VAL	GLU	ALA	ASP	VAL	ALA	GLY	HIS	GLY	GLN
SEQRES	3	153	ASP	ILE	LEU	ILE	ARG	LEU	PHE	LYS	SER	HIS	PRO	GLU	THR
SEQRES	4	153	LEU	GLU	LYS	PHE	ASP	ARG	PHE	LYS	HIS	LEU	LYS	THR	GLU
SEQRES	5	153	ALA	GLU	MET	LYS	ALA	SER	GLU	ASP	LEU	LYS	LYS	HIS	GLY
SEQRES	6	153	VAL	THR	VAL	LEU	THR	ALA	LEU	GLY	ALA	ILE	LEU	LYS	LYS
SEQRES	7	153	LYS	GLY	HIS	HIS	GLU	ALA	GLU	LEU	LYS	PRO	LEU	ALA	GLN
SEQRES	8	153	SER	HIS	ALA	THR	LYS	HIS	LYS	ILE	PRO	ILE	LYS	TYR	LEU
SEQRES	9	153	GLU	PHE	ILE	SER	GLU	ALA	ILE	ILE	HIS	VAL	LEU	HIS	SER
SEQRES	10	153	ARG	HIS	PRO	GLY	ASP	PHE	GLY	ALA	ASP	ALA	GLN	GLY	ALA
SEQRES	11	153	MET	ASN	LYS	ALA	LEU	GLU	LEU	PHE	ARG	LYS	ASP	ILE	ALA
SEQRES	12	153	ALA	LYS	TYR	LYS	GLU	LEU	GLY	TYR	GLN	GLY			

FTNOTE 1
FTNOTE 1 INSPECTION OF MAPS AT VARIOUS STAGES OF REFINEMENT
FTNOTE 1 INDICATED THAT SEVERAL SIDE-CHAINS AND THE C- AND
FTNOTE 1 N-TERMINAL RESIDUES WERE DISORDERED TO SOME DEGREE. IN
FTNOTE 1 MOST CASES, ONE CONFORMATION WAS VERY MUCH STRONGER THAN
FTNOTE 1 THE OTHER, BUT FOUR SIDE-CHAINS WERE INCLUDED IN THE MODEL
FTNOTE 1 WITH TWO ALTERNATE CONFORMATIONS (VAL 13, LEU 86, LEU 89,
FTNOTE 1 GLN 128).
FTNOTE 2
FTNOTE 2 HOH 305 IS CLOSE TO ATOM O1 OF HEM 546. SEE PAPER CITED
FTNOTE 2 AS JRNL REFERENCE ABOVE.

HET	SO4	545	5	SULFATE
HET	HEM	546	45	PROTOPORPHYRIN IX WITH FE2+ AND O2
FORMUL	2	HOH	*334(H2 O1)	
FORMUL	3	SO4	*O4 S1	
FORMUL	4	HEM	C34 H32 N4 O4 FE1 ++	.
FORMUL	4	HEM	O2	

HEADER RIBOSOMAL PROTEIN 02-SEP-86 1CTF
 COMPND L7(SLASH)*L12 50 S RIBOSOMAL PROTEIN (C-TERMINAL DOMAIN)
 SOURCE (ESCHERICHIA COLI, /MRE 600)
 AUTHOR M.LEIJONMARCK,A.LILJAS
 REVDAT 2 16-JUL-87 1CTFA 1 JRNL
 REVDAT 1 15-JAN-87 1CTF 0
 JRNL AUTH M.LEIJONMARCK,A.LILJAS
 JRNL TITL STRUCTURE OF THE C-TERMINAL DOMAIN OF THE RIBOSOMAL
 JRNL TITL 2 PROTEIN L7(SLASH)*L12 FROM ESCHERICHIA COLI AT 1.7 0
 JRNL TITL 3 ANGSTROMS
 JRNL REF J.MOL.BIOL. V. 195 555 1987
 JRNL REFN ASTM JMOBAK UK ISSN 0022-2836 070
 REMARK 1
 REMARK 1 REFERENCE 1
 REMARK 1 AUTH M.LEIJONMARCK,I.PETTERSSON,A.LILJAS
 REMARK 1 TITL STRUCTURAL STUDIES ON THE PROTEIN L7(SLASH)*L12
 REMARK 1 TITL 2 FROM E. COLI RIBOSOMES
 REMARK 1 EDIT M.BALABAN,J.L.SUSSMAN,W.TRAUB,A.YONATH,R.FELDMANN
 REMARK 1 REF STRUCTURAL ASPECTS OF 761 1981
 REMARK 1 REF 2 RECOGNITION AND ASSEMBLY IN
 REMARK 1 REF 3 BIOLOGICAL MACROMOLECULES
 REMARK 1 PUBL BALABAN INTERNATIONAL SCIENCE SERVICES,REHOVOT
 REMARK 1 REFN ISBN 0-86689-004-1 994
 REMARK 1 REFERENCE 2
 REMARK 1 AUTH M.LEIJONMARCK,S.ERIKSSON,A.LILJAS
 REMARK 1 TITL CRYSTAL STRUCTURE OF A RIBOSOMAL COMPONENT AT
 REMARK 1 TITL 2 2.6 ANGSTROMS RESOLUTION
 REMARK 1 REF NATURE V. 286 824 1980
 REMARK 1 REFN ASTM NATUAS UK ISSN 0028-0836 006
 REMARK 1 REFERENCE 3
 REMARK 1 AUTH A.LILJAS,S.ERIKSSON,D.DONNER,C.G.KURLAND
 REMARK 1 TITL ISOLATION AND CRYSTALLIZATION OF STABLE DOMAINS OF
 REMARK 1 TITL 2 THE PROTEIN L7(SLASH)*L12 FROM ESCHERICHIA COLI 4
 REMARK 1 TITL 3 RIBOSOMES
 REMARK 1 REF /FEBS LETT. V. 88 300 1978 6
 REMARK 1 REFN ASTM FEBLAL NE ISSN 0014-5793 165
 REMARK 2
 REMARK 2 RESOLUTION. 1.7 ANGSTROMS.
 REMARK 3
 REMARK 3 REFINEMENT. BY THE RESTRAINED LEAST SQUARES PROCEDURE OF
 REMARK 3 J. KONNERT AND W. HENDRICKSON (PROGRAM *CORELS*). THE
 REMARK 3 R VALUE IS 0.174 FOR 6617 REFLECTIONS. THE RMS DEVIATION
 REMARK 3 FROM IDEALITY OF THE BOND LENGTHS IS 0.025 ANGSTROMS.
 REMARK 4

REMARK 4 COMPLETE L7/L12 CONTAINS 120 AMINO ACIDS. THIS ENTRY
REMARK 4 PRESENTS COORDINATES FOR THE C TERMINAL FRAGMENT WHICH WAS
REMARK 4 CRYSTALLIZED SEPARATELY. THE C TERMINAL FRAGMENT CONTAINS
REMARK 4 74 RESIDUES CORRESPONDING TO RESIDUES 47 - 120 OF THE
REMARK 4 INTACT MOLECULE. RESIDUES 47 - 52 COULD NOT BE LOCATED IN
REMARK 4 THE ELECTRON DENSITY MAP AND, THEREFORE, THE COORDINATES IN
REMARK 4 THIS ENTRY START WITH RESIDUE 53.
REMARK 5
REMARK 5 WATERS 40 AND 59 ARE IN SPECIAL POSITION (X,X,0) OF THE
REMARK 5 SPACE GROUP. THE OCCUPANCY FOR THESE WATERS SHOULD BE 0.5
REMARK 5 BUT WAS SET TO 0.4 IN THE REFINEMENT.
REMARK 6
REMARK 6 STRUCTURE FACTORS CORRESPONDING TO THESE COORDINATES ARE
REMARK 6 AVAILABLE FROM THE PROTEIN DATA BANK AS A SEPARATE ENTRY.
REMARK 7
REMARK 7 CORRECTION. UPDATE JRNL REFERENCE TO REFLECT PUBLICATION.
REMARK 7 16-JUL-87.

SEQRES 1 74 ALA ALA GLU GLU LYS THR GLU PHE ASP VAL ILE LEU LYS
SEQRES 2 74 ALA ALA GLY ALA ASN LYS VAL ALA VAL ILE LYS ALA VAL
SEQRES 3 74 ARG GLY ALA THR GLY LEU GLY LEU LYS GLU ALA LYS ASP
SEQRES 4 74 LEU VAL GLU SER ALA PRO ALA ALA LEU LYS GLU GLY VAL
SEQRES 5 74 SER LYS ASP ASP ALA GLU ALA LEU LYS LYS ALA LEU GLU
SEQRES 6 74 GLU ALA GLY ALA GLU VAL GLU VAL LYS

FTNOTE 1
FTNOTE 1 THE ELECTRON DENSITY FOR RESIDUE LYS 59 IS POORLY DEFINED
FTNOTE 1 FROM CD TO NZ.
FTNOTE 2
FTNOTE 2 THE ELECTRON DENSITY FOR RESIDUES LYS 81, LYS 108, LYS 120
FTNOTE 2 IS POORLY DEFINED FROM CG TO NZ.
FTNOTE 3
FTNOTE 3 RESIDUE PRO 91 IS A CIS PROLINE.

HET SO4 1 5 SULFATE
FORMUL 2 SO4 O4 S1
FORMUL 3 HOH *62(H2 O1)

HELIX 1 A ASN 64 GLY 77 1 H-BONDS LINK C(I) TO N(I+4)
HELIX 2 B GLY 79 GLU 88 1 H-BONDS LINK C(I) TO N(I+4)
HELIX 3 C SER 99 ALA 113 1 H-BONDS LINK C(I) TO N(I+4)
SHEET 1 I 3 ALA 92 VAL 98 0
SHEET 2 I 3 PHE 54 ALA 60 -1 N PHE 54 0 VAL 98
SHEET 3 I 3 GLU 116 LYS 120 -1 N GLU 116 0 ALA 60

TURN 1 BEB ALA 61 ASN 64 TYPE II(PRIME). GLY IN POS I+1
TURN 2 3HB LEU 86 SER 89 3/10 TURN AT END OF HELIX B
TURN 3 CPT SER 89 ALA 92 TYPE VIB. CIS-PROLINE(91) TURN
TURN 4 PHC LEU 110 ALA 115 PI-BEND AT END OF HELIX C

TURN	5	3HC	GLU	111	GLY	114	3/10	TURN	AT	END	OF	HELIX	C	
CRYST1	54.830		54.830		42.670	90.00	90.00	90.00	P	43	21	2		8
ORIGX1		1.000000		0.000000		0.000000				0.000000				
ORIGX2		0.000000		1.000000		0.000000				0.000000				
ORIGX3		0.000000		0.000000		1.000000				0.000000				
SCALE1		.018238		0.000000		0.000000				0.000000				
SCALE2		0.000000		.018238		0.000000				0.000000				
SCALE3		0.000000		0.000000		.023436				0.000000				