Université de Montréal

Utilisation de banques de données structurelles dans le raffinement des boucles lors de la prédiction de structures tertiaires de protéines

par
Eric Martineau
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

avril 2001

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Utilisation de banques de données structurelles dans le raffinement des boucles lors de la prédiction de structures tertiaires de protéines

présenté par :
Eric Martineau

a été évalué par un jury formé des personnes suivantes :

Thomas H. Ellis,	président rapporteur
John R. Gunn,	directeur de recherche
Michel Lafleur,	membre du jury

Mémoire accepté le : _____

SOMMAIRE

Une protéine se définit comme une macromolécule biologique d'une très grande complexité structurelle. Chaque séquence d'acides aminés confère à la protéine une structure unique ainsi que son rôle dans un organisme vivant. Certains segments d'acides aminés dans une protéine empruntent des structures bien définies afin de minimiser localement l'énergie à l'interne de la molécule en maximisant le nombre d'interactions favorables et simultanément en minimisant le nombre d'interactions défavorables. Ainsi, on identifie à l'interne de ces macromolécules, des structures secondaires, dont les principales sont l'hélice α (qui ressemble à un ressort) et le feuillet β (composé de brins d'acides aminés positionnés entre eux de façon parallèle ou antiparallèle). Les acides aminés qui ne font pas partie des structures secondaires se retrouvent dans les boucles. Ces dernières ne suivent aucun patron structurel prédéterminé. Les conformations s'agenceront elles aussi de manière à minimiser l'énergie de la molécule. Enfin, la combinaison de ces différents types de structures forment la structure tertiaire de la protéine, qui se veut en fait sa structure tridimensionelle.

Chaque séquence d'acides aminés tendera donc à se replier en une structure tridimensionelle compacte de grande stabilité. Or, le mécanisme selon lequel ce processus d'effectue comporte encore certains points nébuleux. Entre autres mots, comment une molécule ayant tant de possibilités au niveau de sa conformation, se retrouve à son minimum d'énergie dans un laps de temps si court? Au niveau des sciences théoriques, la question semble encore plus significative que l'énergie de chaque conformation possible devrait être calculée. L'âge de l'univers ne serait pas suffisant pour un tel calcul.

Pour contourner ces difficultés, les scientifiques développèrent un modèle plus simpliste d'une protéine, éliminant ainsi plusieurs conformations improbables énergétiquement,

voire impossibles. Malgré la quantité de conformations intrinsèquement rejetées par ce modèle simpliste d'une protéine, le nombre de structures possibles demeure d'envergure. La recherche conformationelle doit donc couvrir un grand nombre de structures possibles tout en se déroulant sur un période de temps raisonnable. Cette recherche conformationelle se déroule sur des structures tridimensionelles construites au hasard à partir de la séquence en acides aminés et de l'emplacement des structures secondaires.

Afin d'améliorer le modèle, un nouvel élément peut être apporté au niveau de la construction des structures générées au hasard. Ces structures générées au hasard ne représentent pas bien la réalité. Les boucles ne suivent pas des patrons structurels bien définis tels que l'hélice α ou le feuillet β, mais certains éléments issus de la nature pourraient potentiellement guider la construction. De cette façon, l'inclusion d'éléments de géométrie provenant de structures naturelles de protéines telles que celles de la *Protein Data Bank* pourrait donner un sens au processus de construction.

L'inclusion de ce nouveau type d'informations nécessite la confection d'une banque de données de structures. De cette dernière, des banques de données de sous-structures devront être considérées afin de bien cibler divers éléments géométriques dans les boucles. Les deux sous-structures considérées seront donc le triplet (séquences de trois acides aminés) et les boucles (de longueurs variables). La comparaison et l'analyse des différentes distributions de triplets et de boucles issues de la nature par rapport à celles ne comportant aucun élément naturel serviront à prédire si l'apport de ce type d'informations peut véritablement guider le processus de construction et éventuellement améliorer le modèle afin de trouver la structure native.

Évidemment ces distributions peuvent être manipulées et ce, par l'intermédiaire de critères de sélection à différents niveaux. Pour concilier l'information de la banque de données de structures naturelles avec la séquence d'acides aminés de la protéine à l'étude, il faut utiliser des critères d'homologie dans le but d'accepter ou de rejeter des

géométries de structures provenant de la distribution naturelle. Au niveau des triplets, deux critères d'acceptation seront mis à contribution. Le premier, un critère d'acceptation basé sur la similarité au niveau des séquences en acides aminés. Le second, pour sa part, constitue un critère d'acceptation au niveau de la structure avoisinante de chacun des triplets. Finalement, pour les boucles, le critère d'acceptation repose sur l'environnement physicochimique tridimensionel de ces boucles.

Au terme de cette étude, on pourra constater les différences majeures qui existent entre les distributions biologiques et non-biologiques. Ces distributions peuvent également se modifier durant les simulations afin d'optimiser la qualité de l'information requise pour améliorer les résultats d'une simulation. Dépendemment du niveau de sélection de ces critères, il semble y avoir une nette amélioration au niveau des structures de protéines calculées dans certains cas. L'optimisation doit être accomplie à chaque nouvelle protéine, mais les résultats dans l'ensemble sont probants.

mots clés: protéine, repliement, distributions biologiques, homologie.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DE FIGURES

LISTE DES ABRÉVIATIONS

BLAST	*programme qui effectue de l'alignement local de séquences, Altschul, Gish, Miller, Myers et Lipman.*
CHARMM	*Chemistry at HARvard Macromolecular Mechanics*
NACCESS	*programme de calcul d'aire enfouie de résidus, Hubbard et Thornton.*
MOLFIX	*programme qui génère la structure la plus ressemblante possible d'une protéine selon un modèle réduit, L'Heureux.*
RMN	*Résonance Magnétique Nucléaire*
RMS	*Root Mean Square deviation*
PAM	*matrices de probabilité mutationnelle selon Dayhoff (Point Accepted Mutation)*
PDB	*Brookhaven Protein Data Bank*
TRIP	*Programme de prédiction du repliement des protéines, Gunn.*

À tous ceux et celles qui ont favorisé mon entropie personnelle

CHAPITRE 1

Introduction

L'homme fut toujours curieux de connaître son environnement proche et lointain, de l'infiniment petit à l'infiniment grand, de l'événement le plus simple à l'événement le plus complexe. Les phénomènes qui se produisent dans l'univers représentent des objets de contemplation et de fascination qui nous poussent à chercher toujours plus profondément afin de les comprendre. Cette quête du savoir, amorcée il y a plusieurs siècles, se poursuit aujourd'hui compte tenu du fait que nombre de questions subsistent toujours quant aux diverses facettes notre univers.

On peut aborder ce questionnement sous différents angles. Une façon de le cerner consiste en l'énumération de toutes les recherches qui tournent autour des quatre grandes forces dans l'univers, la force gravitationnelle (qui régit le mouvement des planètes et des astres par exemple), la force électromagnétique (qui agit au niveau de la structure des molécules), et les forces nucléaires forte et faible (responsables de la structure des divers noyaux atomiques). Ces quatre forces couvrent effectivement les différentes organisations dans l'univers, de l'échelle du microscopique au macroscopique.

Le groupe de recherche dont je fais partie s'intéresse aux structures tridimensionelles des protéines globalement régies par les forces électromagnétiques. Une protéine se définit comme un polymère biologique de taille variable dont l'unité de base se nomme acide aminé. Ce polymère, grâce à des enchevêtrements de structures primaires (les boucles) et de structures secondaires (comme les hélices et les feuillets), tend à se replier en une forme globale appelée structure tertiaire. Cette forme lui permet de minimiser son énergie globale et donc d'être plus stable dans son environnement. Il existe en fait plusieurs types d'interactions possibles dans les forces électromagnétiques (par exemple: les interactions de Van der Waals, les liaisons covalentes et la force hydrophobe). Je survolerai plus tard les forces impliquées dans le repliement des protéines. Plusieurs domaines de recherches

gravitent autour des protéines que ce soit au niveau de la détermination de la séquence des acides aminés, au niveau de la détermination des structures secondaires et tertiaires ou encore au niveau de l'étude de leurs rôles au sein d'un organisme vivant et je passe sous silence moult autres sujets de recherche. Peu importe le champ des recherches des divers scientifiques, tous s'entendent pour dire que les protéines jouent des rôles cruciaux dans le fonctionnement d'un organisme vivant.

1.1. L'origine des protéines: l'ADN [1,2]

Non seulement l'origine des protéines provient de l'ADN (acide désoxyribonucléique) mais l'origine de la vie est aussi tributaire de l'ADN. Il existe une relation fondamentale entre l'ADN, l'ARN (acide ribonucléique) et les protéines. Dans chaque noyau cellulaire, on dénote la présence de chromosomes composés de milliers de segments de molécules ADN. Une hélice d'ADN caractérisée par son double brin se compose, en moyenne, de 250 millions de paires de bases (adénosine, cytosine, guanine et thymine).

Le processus de réplication de l'ADN s'effectue avec une efficacité stupéfiante; une hélice d'ADN se dédouble en quelques minutes seulement. Les erreurs se font rares, soit environ une toutes les 10^{10} paires de bases. La réplication de l'ADN constitue un processus enzymatique responsable de la formation de la double hélice (en sens inverse) édifiée de façon à favoriser l'appariement complémentaire entre les bases.

L'ARN se caractérise par une hélice à brin simple. Les trois types d'ARN permettent d'établir le pont entre l'information génétique originant de l'ADN et la formation des protéines. Dans un premier temps, l'ARN messager relaie l'information génétique du noyau cellulaire au ribosome par un processus nommé transcription. L'ARN messager comporte toute l'information nécessaire à la synthèse d'une protéine sous forme codée. L'ARN de transfert pour sa part, se compose de codons (séquences de trios de bases (adénosine, cytosine, guanine et uracile)) . Il existe 4^3 codons possibles, dont 61 représentent un code pour un acide aminé spécifique et 3 représentent des signaux d'arrêt.

Ainsi, grâce à ces processus, la biosynthèse de diverses biomolécules se produit au sein des organismes vivants.

1.2. Vue d'ensemble d'une protéine

Les protéines, aussi considérées comme des polymères biologiques, sont composées de diverses combinaisons de longueur variable de 20 acides aminés. L'acide aminé, lui, se définit comme le monomère qui forme la protéine. L'alignement selon lequel les acides aminés sont ordonnés constitue l'origine de la structure de la protéine et par le corollaire sa fonction propre au sein d'un organisme, bien qu'au début, ce fait ne semblait pas évident. Leurs rôles sont multiples et diversifiés; les protéines peuvent contrôler certaines fonctions génétiques, être des catalyseurs importants dans plusieurs réactions chimiques, assurer le transport membranaire de certaines substances, reconnaître des sites actifs et se lier de façon non-covalente à d'autres biomolécules, et j'en passe.[3] Dans ces macromolécules plutôt complexes dues à leurs dimensions, on y retrouve certaines sous-structures communes qui ont pour but de favoriser thermodynamiquement l'ensemble de la structure protéique. On traite donc de la structure des protéines selon quatre niveaux d'organisation (voir figure 1).

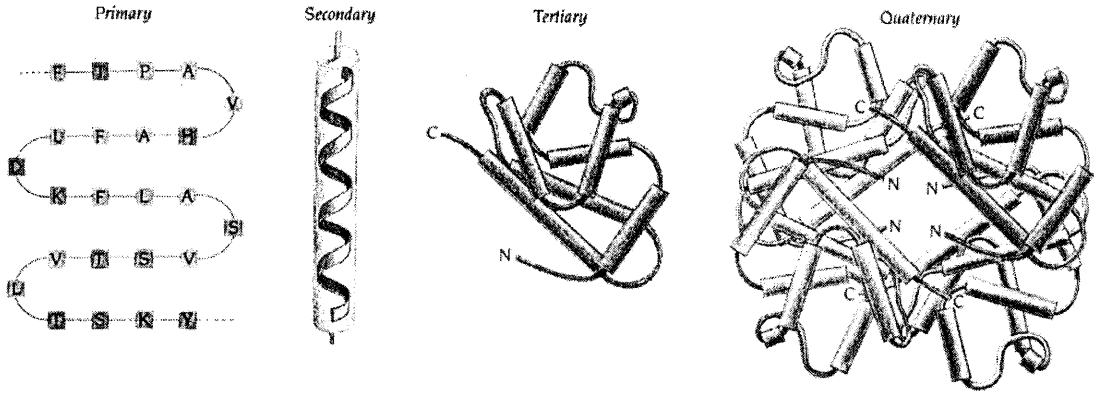


FIGURE 1: Les 4 niveaux d'organisation de la protéine.

1.2.1. La structure primaire d'une protéine

La structure primaire se définie comme l'arrangement linéaire des acides aminés, on la dénomme en fait: séquence. Comme il a été mentionné plus haut, les 20 acides aminés possèdent des propriétés physiques et chimiques propres à chacun. Au fait qu'est-ce qu'un acide aminé? L'appellation acide aminé tire son origine du fait qu'il s'agisse d'une molécule organique composée d'une part d'un groupement carboxyl (-COOH) et d'autre part d'un groupement amino ($-NH_2$). Ces deux groupements sont liés de façon covalente à un carbone chiral (sauf pour la glycine) flanqué d'un hydrogène et d'un autre groupement appelé chaîne latérale. L'ensemble de ces groupements, en excluant la chaîne latérale, se nomme chaîne principale. Il s'agit en fait d'un tronc commun pour chacun des acides aminés. Ce carbone chiral possède une configuration gauche (L) dans le cas des acides aminés naturels. C'est dans cette chaîne latérale que se trouve principalement les propriétés de polarité, d'acidité, de réactivité chimique et d'hydrophobicité.

1.2.2. Le lien peptidique

Dans un polypeptide ou une protéine, les acides aminés sont liés les uns aux autres par un lien covalent situé entre le carbone carboxyl d'un acide aminé et l'azote de l'acide aminé suivant. On dénomme communément ce lien covalent, le lien peptidique (figure 2, tirée du Zubay[4]).

Le lien peptidique possède un certain caractère de double liaison (dû aux orbitales π du lien C-N), ce qui restreint la rotation autour du lien N-CO. C'est donc ce lien rigide qui fait en sorte que plusieurs atomes se retrouvent dans le même plan. Évidemment, la présence d'un grand nombre de liens peptidiques réduit considérablement le nombre de degrés de liberté de la chaîne principale. On compte deux degrés de liberté d'importance sur la chaîne principale: les angles ϕ et ψ. Il s'agit d'angles situés de part et d'autre du

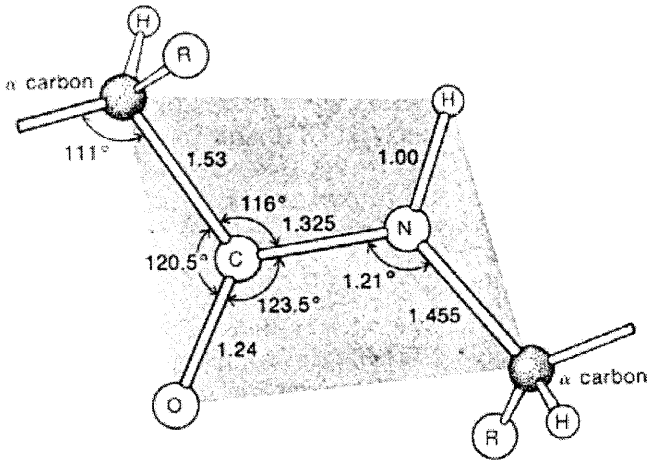


FIGURE 2: Le lien peptidique et ses dimensions

carbone α. Par convention l'angle ϕ se définit comme l'angle de rotation autour du lien azote-carbone α. L'angle ψ, quant à lui se définit comme l'angle de rotation autour du lien carbone α-carbonyl. (voir figure 3 tirée du Zubay[4] à la page suivante). Une rotation dans le sens anti-trigonométrique représente un angle positif. Ces angles ϕ et ψ s'ajustent de façon à minimiser l'énergie de la chaîne principale en amenuisant entre autre l'encombrement stérique et en augmentant le nombre d'interactions stabilisantes comme les ponts H.

Évidemment, comme les acides aminés comportent des différences notables les uns les autres, certaines combinaisons d'angles dièdres ne sont pas permises. Les combinaisons permises varient d'un résidu à l'autre. Par exemple, la glycine ne possède pas de carbone

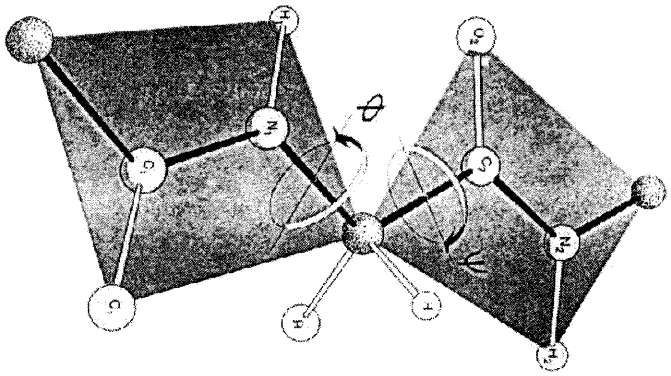


FIGURE 3: Les angles ϕ et π.

β mais seulement un hydrogène agissant comme une petite chaîne latérale qui n'interfère que peu avec la chaîne principale selon diverses combinaisons d'angles dièdres. Cependant, le tryptophane, qui possède une chaîne latérale énorme et cyclique ne peut qu'adopter certaines combinaisons à défaut de quoi, des contacts entre les atomes de la chaîne latérale et des atomes de la chaîne principale auraient lieu.

C'est Ramachandran[5], le premier qui a eu l'idée d'étudier cet espace conformationel des angles ϕ et ψ. Il a tout simplement effectué une rotation de 360° degrés des angles ϕ et ψ pour chaque acide aminé. Ces résultats sont représentés sous forme de cartes[6] des angles ϕ en fonction des angles ψ. Par ces cartes, il a démontré que certaines régions demeurent inaccessibles à cause de l'encombrement stérique. On peut voir à la figure 4, une carte du modèle rigide de Ramachandran. On peut aussi observer les zones permises et interdites ainsi que les espaces où l'on retrouve des angles correspondant à des structures secondaires.

1.2.3. Un survol des propriétés physiques et chimiques des divers acides aminés[7,8]

Il s'agit maintenant de survoler chacune des structures des 20 acides aminés, question de bien comprendre plus tard certains choix effectués dans le cadre de cet ouvrage[8]. Évidemment, il ne sera pas question d'entrer dans les détails les plus subtils, mais bien de s'attarder sur les structures et les tendances générales de réactivité et d'acidité. On peut voir les différents acides aminés à la figure 5 tirée du mémoire de maîtrise de Pierre-Jean L'Heureux[9]. Le tableau I tiré du livre de Creighton[10] présente les différents volumes et masses de ces acides aminés.

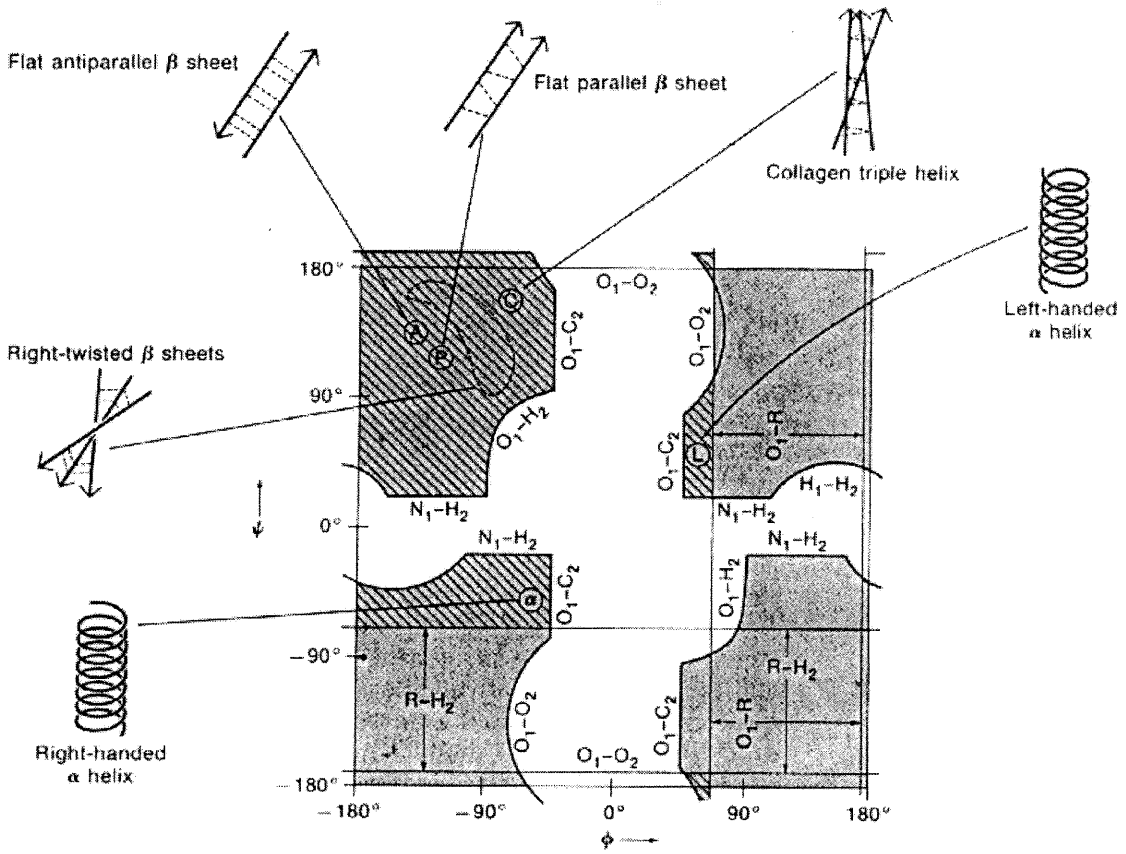


FIGURE 4: Carte de Ramachandran et ses zones spécifiques.

Si on regarde d'abord la glycine, on constate qu'il s'agit sans contredit du plus petit acide aminé de tous. Il s'agit également du seul acide aminé qui ne possède pas de carbone chiral car sa chaîne latérale est formée d'un hydrogène. Cette propriété lui confère d'ailleurs la meilleure flexibilité de tous les acides aminés. Il fait parti des résidus hydrophiles et aura tendance à se situer à l'extérieur des protéines.

Ensuite viennent les résidus aliphatiques soit: l'alanine, la leucine, la valine et l'isoleucine. Ces quatre acides aminés ont certaines propriétés communes. Ils ne possèdent pas de caractères réactifs sur leurs chaînes latérales dû à leur caractère non-polaire. Par contre ils forment des interactions thermodynamiquement favorables entre

eux. Fait notable, l'isoleucine possède un second carbone asymétrique, ce dernier étant situé sur la chaîne latérale.

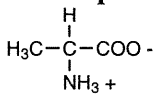
TABLEAU I: Masses atomiques et volumes de Van der Waal des acides aminés naturels

Acide aminé	Masse de l'acide aminé (daltons)	Volume de Van der Waal ($Å^3$)
Alanine	71.09	67
Arginine	156.19	148
Asparagine	114.11	96
Acide aspartique	115.09	91
Cystéine	103.15	86
Glutamine	128.14	114
Acide glutamique	129.12	109
Glycine	57.05	48
Histidine	137.14	118
Isoleucine	113.16	124
Leucine	113.16	124
Lysine	128.17	135
Méthionine	131.19	124
Phénylalanine	147.18	135
Proline	97.12	90
Sérine	87.08	73
Thréonine	101.11	93
Triptophane	186.21	163
Tyrosine	163.18	105
Valine	119.40	105

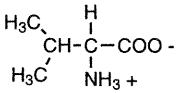
FIGURE 5: Les acides aminés naturels

Acide aminé apolaire

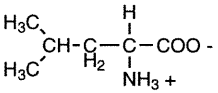
Alanine
ALA
A
1



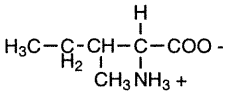
Valine
VAL
V
20



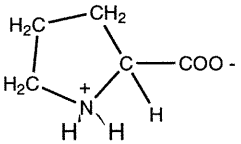
Leucine
LEU
L
11



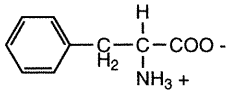
Isoleucine
ILE
I
10



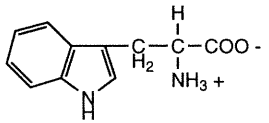
Proline
PRO
P
15



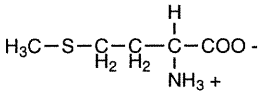
Phénylalanine
PHE
F
14



Tryptophane
TRP
W
18

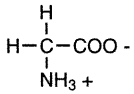


Méthionine
MET
M
13

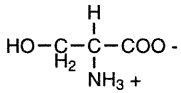


Acide aminé polaire

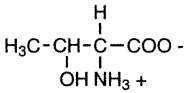
Glycine
GLY
G
8



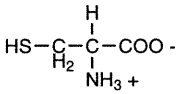
Sérine
SER
S
16



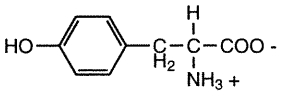
Thréonine
THR
T
17



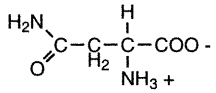
Cystéine
CYS
C
5



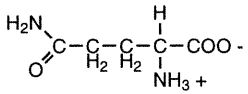
Tyrosine
TYR
Y
19



Asparagine
ASN
N
3

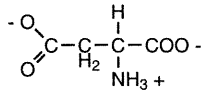


Glutamine
GLN
Q
6

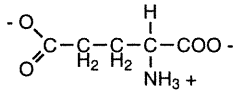


Acide aminé chargé

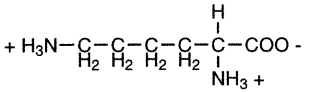
Acide aspartique
ASP
D
4



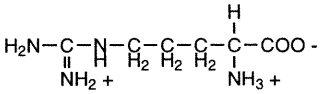
Acide glutamique
GLU
E
7



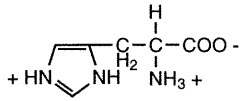
Lysine
LYS
K
12



Arginine
ARG
R
2



Histidine
(pH=6.0)
HIS
H
9



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

On dénote aussi l'existence d'un résidu cyclique: la proline. Sa chaîne latérale aliphatique est liée de façon covalente à l'azote. La chaîne principale de la proline ne possède pas d'hydrogène amide pour l'utiliser comme donneur de ponts H. Sa nature cyclique impose des contraintes de rotations autour du carbone α, ce qui entraîne d'importants effets stériques ayant un impact majeur sur la conformation de la chaîne principale. Normalement, les carbones α s'éloignent l'un de l'autre (causant ainsi une configuration trans), mais dans le cas de la proline, cette préférence est moins prononcée puisque l'azote est lié à deux autres carbones au lieu d'un seul.

Viennent ensuite les résidus hydroxy, c'est-à-dire la sérine et la thréonine. Ils possèdent de petites chaînes latérales aliphatiques à l'exception du groupement OH polaire. Normalement, ces résidus ne réagissent pas plus que l'éthanol lorsqu'ils participent à certaines réactions chimiques. Dernier détail pour ce type d'acides aminés, la thréonine contient un carbone asymétrique sur sa chaîne latérale en plus de son carbone α.

On rencontre aussi des acides aminés dit acides, soit l'acide aspartique et l'acide glutamique. Ces résidus ne diffèrent que par un $-CH_2-$ sur leur chaîne latérale, mais leurs propriétés chimiques diffèrent considérablement. L'impact de cette différence se situe au niveau des interactions avec la chaîne principale. Au pH physiologique, ces résidus adoptent un caractère ionisé et polaire.

Il existe aussi une classe d'acides aminés possédant la forme amide de l'acide aspartique et de l'acide glutamique, soit respectivement l'asparagine et la glutamine. Bien qu'il s'agisse d'une forme amide, ces acides aminés existent naturellement. Ces acides aminés ne réagissent que très peu, ne s'ionisent pas malgré leur caractère très polaire et se comportent comme de bons donneurs et de bons accepteurs de ponts H.

Une autre classe regroupe les acides aminés basiques, soit la lysine et l'arginine. La chaîne hydrophobe de la lysine comportant un azote à l'extrémité s'ionise au pH physiologique. Il y a toujours un segment sur cet acide aminé qui n'est pas ionisé. L'arginine, elle, possède un groupement guanido ionisé à tous les pH physiologiques

possibles. On dénote de la résonance sur la chaîne latérale, ce qui lui confère plus de stabilité au niveau thermodynamique. Finalement, ces acides aminés sont peu réactifs.

L'histidine est un acide aminé structurellement bien différent des autres, ce qui le laisse dans une classe à part. La chaîne imidazole possède plusieurs propriétés spéciales qui font de cet acide aminé un bon catalyseur nucléophile. L'amide tertiaire situé sur sa chaîne latérale possède un caractère très réactif. C'est un résidu large et basique qui, lorsqu'il se trouve dans sa forme non-ionisée, possède un groupement –NH électrophile et donneur de ponts H ainsi qu'un amine tertiaire (–N=) nucléophile et accepteur de ponts H.

L'avant-dernière classe d'acides aminés regroupe les résidus aromatiques, soit la phénylalanine, la tyrosine et le tryptophane. Ces trois résidus sont responsables de presque toute l'absorption de l'ultra-violet et des propriétés fluorescentes. La phénylalanine possède un cycle semblable au benzène, ce qui implique une chaîne latérale non-polaire et non-réactive dans l'environnement des protéines. La tyrosine contient un groupement hydroxy qui rend cet acide aminé très réactif. Enfin, le tryptophane est le plus fluorescent et le plus large des acides aminés aromatiques.

La dernière classe regroupe celle des acides aminés contenant un soufre, c'est-à-dire la méthionine et la cystéine. La méthionine possède une longue chaîne latérale non-polaire et relativement non-réactive. Le soufre comporte un caractère nucléophile, quoi qu'il puisse être protoné dans les protéines. Le groupement thiol de la cystéine possède un caractère très réactif (une oxydation se produit lorsque deux groupements thiol forment une liaison covalente à deux soufres) et il s'ionise à des pH légèrement alcalins. Cet acide ressemble beaucoup à la sérine, à la seule différence près que le groupe $S\text{-}CH_3$ remplace le groupe OH sur sa chaîne latérale.

1.2.4. La structure secondaire des protéines

Certaines séquences d'acides aminés donnent lieu à des conformations caractérisques de la chaîne peptidique. Il s'agit du deuxième niveau d'organisation des protéines (la structure secondaire) constitué de sous-structures régulières dont les deux principales, l'hélice α et le feuillet β. L'hélice α[11] fut proposée par L. Pauling et R. Corey en 1951 suite à des expériences de diffraction des rayons-X. La structure ressemble à un ressort possédant 3.7 acides aminés par tour. Cette structure est stabilisée par des ponts H entre l'hydrogène lié à l'azote et le groupement carbonyl quatre unités plus loin sur la chaîne. Les chaînes latérales sont alors dirigées vers l'extérieur du dit ressort.

Le feuillet β[12] est en fait une conformation de la chaîne principale où les brins du feuillets sont placés parallèlement les uns aux autres et ce, en sens inverse ou non. La forme en escalier permet de limiter au maximum l'encombrement stérique. On définit le feuillet parallèle comme un feuillet dans lequel les chaînes liées par des liaisons hydrogènes sont dirigées dans le même sens. Le feuillet anti-parallèle quant à lui, se compose de chaînes liées par des liaisons hydrogènes, sauf que ces chaînes sont en sens inverse l'une par rapport à l'autre. La différence entre les deux types de feuillets se situe dans la formation des ponts H entre les brins.

1.2.5. La structure tertiaire et quaternaire des protéines

On répertorie deux types de structures tertiaires. Il y a des protéines fibreuses, telles le collagène[13], qui possèdent une forme plutôt linéaire (plus rares) et il existe des protéines globulaires[14]. Dans ces dernières de forme grossièrement sphérique, on retrouve des structures secondaires reliées entre elles par des segments d'acides aminés. On remarque aussi certains endroits de la chaîne où on dénote la présence de ponts bi-sulfure. Chez les protéines solubles, les acides aminés chargés et polaires se situent généralement à l'extérieur tandis que les acides aminés hydrophobes et non-chargés se retrouvent plutôt

au coeur de la forme globulaire de la protéine. Les structures secondaires comportent généralement entre 40 et 70% des acides aminés de la protéine. Dans le même ordre d'idées, ces structures relativement compactes et basses en énergie contribuent beaucoup à la stabilisation énergétique des protéines. Sans l'apport de ces structures secondaires, les protéines ne pourraient se replier en des formes aussi compactes. Enfin, la structure quaternaire[15] est en fait un amas de plusieurs chaînes peptides. Il s'agit en fait d'un dérivé de l'arrangement des sous-unités entre elles dans l'espace.

1.3. Les forces impliquées dans les protéines[16]

Maintenant que la hiérarchie structurale est bien définie, on peut se demander quelles forces électromagnétiques sont impliquées pour qu'une séquence d'acides aminés se replie en un enchevêtrement globulaire? D'abord, il est important de bien différencier les interactions à courte portée, les interactions à longue portée, les interactions locales et les interactions non-locales. C'est la distance qui distingue les interactions à courte portée (comme les attractions Lennard-Jones et les répulsions) et les interactions à longue portée (comme les interactions ion-dipôle). L'énergie de ces interactions dépend de la distance (r) selon la relation r^{-p}. Si p est inférieur ou égal à 3, il s'agit d'interactions à longue portée. Dans le cas échéant, si p est supérieur à 3, il s'agit d'interactions à courte portée. Cette valeur de p est déterminée naturellement par l'intégrale de l'énergie totale qui converge pour les interactions à courte portée et diverge pour les interactions à longue portée. La position des résidus permet de distinguer les interactions locales et les interactions non-locales. Deux résidus voisins ou proches sur une même chaîne interagiront localement entre eux. Par contre, s'ils sont situés sur des chaînes différentes (ou distancés pas plusieurs acides aminées entre eux), il s'agira d'une interaction non-locale (voir la figure 6). De ce fait, il existe des interactions à courte et longue portées pour des interactions locales et non-locales.

En ce qui concerne la nature des forces, il semble évident après plusieurs décennies de recherche que l'hydrophobicité représente la force dominante et directrice dans le repliement des protéines. La stabilisation énergétique se produisant lorsque les groupes

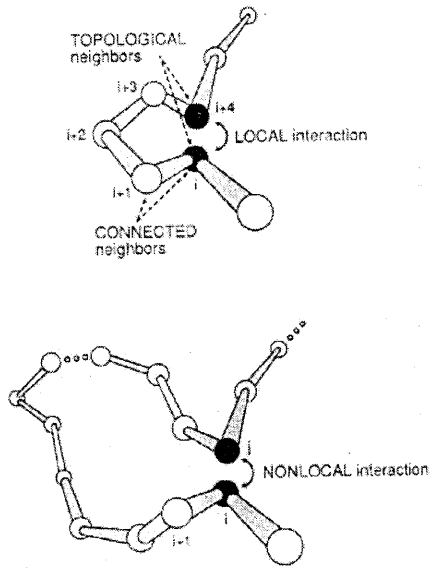


FIGURE 6: Les interactions locales et non-locales.

hydrophobes se rassemblent les uns les autres et réduisent leur exposition au solvant est à l'origine de des interactions hydrophobiques. Ce type d'interactions représente un exemple de processus d'ordre qui est stabilisé par une tendance vers un plus grand désordre du solvant.[17] En définitive, les acides aminés hydrophobes se regroupent en un centre apolaire laissant ainsi les acides aminés hydrophiles, situés à l'extérieur, interagir avec le solvant polaire.

Les autres forces impliquées sont plus faibles mais contribuent quand même à la géométrie de la structure. En milieux acide et basique, les charges électrostatiques semblent déstabiliser la structure de la protéine. Les forces électrostatiques (F) sont régies par l'équation:

$$F = \frac{Q \cdot Q'}{4\pi\varepsilon_0 r^2} \qquad (1.1)$$

où: Q et Q' représentent les charges pontuelles

ε_0 est la constante de permittivité du vide (8.854×10^{-12} $N^{-1}m^{-2}C^2$)

r est la distance entre les deux charges ponctuelles

Pour des pH aux environs de 7.0, il semble que l'appariement ionique stabilise les structures protéiques. Au niveau des acides aminés polaires, les interactions de Van der Waals et les ponts H sont effectifs. Dans l'ensemble, il s'agit d'une faible contribution du point de vue de la stabilité thermodynamique, toutefois, ces interactions sont importantes au niveau de la spécificité, par exemple au niveau de la formation de structures secondaires.

Évidemment, les forces directrices sont confrontées à des forces qui contribuent à la structure protéique et à sa stabilité. Cette force provient principalement de la perte d'entropie conformationelle non-locale due à des contraintes stériques dans l'état replié. Tandis que l'hydrophobicité est responsable de la structure interne de la protéine, les contraintes stériques sont responsables de l'architecture interne. La seule raison expliquant que la structure interne native est explicitement dépendante de la séquence d'acides aminés peut être largement attribuable à l'hydrophobicité. Il n'existe que très peu de façons de configurer une chaîne pour maximiser le nombre de contacts non-polaires.

Il y a deux types d'entropie impliquées dans un système protéique. La première se nomme l'entropie locale, elle provient de l'énergie responsable des conformations des résidus interreliés entre eux.

$$S_i = \ln z_i + < \varepsilon (\phi_i, \psi_i, \chi_i, \chi_{i+1}) > / kT \qquad (1.2)$$

où: z est la fonction de partition ($z_i = \int \exp(-\varepsilon/kT) \cdot d\tau$)

ε est l'énergie du peptide (qui est fonction des angles ϕ,ψ et χ)

k est la constante de Boltzmann ($1.3806503 \times 10^{-23}$ J K^{-1})

T est la température en kelvin

Cette entropie regroupe les entropies vibrationelles et rotationelles internes des petites molécules. Bien que ces contributions soient petites, elles restent indépendantes des propriétés globales comme le rayon de gyration. En d'autres mots, l'entropie totale égale la sommes des entropies individuelles.

Le deuxième type est l'entropie non-locale qui dépend du nombre de configurations des chaînes ρ comme une fonction de densité de segments de chaînes. Voici la formule qui la décrit:

$$S = -k \int \rho \ln \rho \, d\rho \qquad (1.3)$$

La variable ρ exprime le nombre de monomères divisé par le volume occupé par la chaîne. Étant donné que le repliement débute d'un état dénaturé se repliant vers un état natif beaucoup plus compact, ceci implique une perte importante d'entropie non-locale durant le processus.

1.4 La détermination expérimentale de ces différentes structures

Certaines méthodes de la biologie moléculaire[18] permettent la détermination de la structure primaire des protéines[19] . La spectroscopie RMN [20] et le dichroïsme circulaire [21] permettent de déterminer les structures secondaires. Les structures tertiaires peuvent aussi être élucidées expérimentalement, soit par RMN [22], soit par la microscopie électronique [23] ou par la cristallographie par rayons-X [24].

Cependant, bien que ces méthodes soient efficaces dans certains cas, il n'en reste pas moins qu'elles sont coûteuses en main-d'oeuvre et en instrumentation. De plus, du côté expérimental, on rencontre d'importants obstacles. Par exemple, la détermination par cristallographie une protéine membranaire se veut fastidieuse parce qu'il est difficile de cristalliser la membrane. De son côté, la spectroscopie RMN nous donne des spectres illisibles dès que la protéine devient trop grande. Compte tenu du nombre astronomique de protéines possibles, ces méthodes deviennent rapidement insuffisantes afin de suivre le rythme du séquençage. C'est ici que les sciences théoriques représentent un espoir à moyen terme.

1.5 L'utilité des sciences théoriques dans le repliement des protéines

Dans la deuxième moitié du vingtième siècle, l'avènement de l'ordinateur a profondément influencé le cours des sciences. Et l'influence ne cesse de s'accroître à mesure que l'on crée de nouveaux processeurs toujours plus rapides et dotés d'une capacité de gérer une quantité de mémoires vive et morte impressionante. À l'aide des principes de la physique construit à partir de solides bases mathématiques, les scientifiques ont élaboré diverses méthodes de calcul pour venir entre autres appuyer les sciences expérimentales. L'ordinateur vient au service des scientifiques en exécutant plusieurs millions d'opérations par seconde accomplissant ainsi des calculs autrefois impossibles à résoudre parce qu'ils auraient été trop coûteux en temps.

Le domaine de recherche des protéines est un de ceux qui peut être appuyé par les sciences théoriques. Le groupe de recherche du professeur John R. Gunn s'investit à prédire des structures tridimensionnelles de protéines. Plusieurs groupes de recherche ont commencé à se servir des outils théoriques pour tenter d'élucider les mystères entourant le domaine des protéines. Là aussi, on fait face à d'importantes limites dont il sera question un peu plus loin. Quel est l'état de la situation du côté théorique? On ne peut pas affirmer que l'on est capable de trouver les bonnes structures, mais bien que les structures élucidées sont moins mauvaises qu'auparavant.[25]

CHAPITRE 2

Les méthodes théoriques et le fonctionnement de TRIP

La quête de la structure native tridimensionnelle d'une protéine est parsemée d'embûches. Les moult degrés de liberté inclus dans ces macromolécules posent effectivement un problème important. La surface de potentiel multidimensionelle nécessite un temps de calcul considérable afin de l'échantillonner. En fait, dans le modèle numérique créé par les théoriciens[26,27,28], il n'existe qu'une structure native et elle se trouve au minimum global d'une fonction de potentiel universelle.

Trois difficultés majeures ressortent toutefois de ce modèle. D'abord, compte tenu des 2N degrés de liberté (où N représente le nombre de résidus) d'une protéine, le nombre de conformations possibles est faramineux. Ensuite, la fonction de potentiel consiste en un grand nombre d'interactions similaires. En d'autres mots, le manque de diversité dans les interactions nous prive de structures comportant des différences physiques propres à la native. On peut avoir l'impression que chacune des protéines échantillonées se ressemblent et que la native est noyée dans cet ensemble. Finalement, cette fameuse fonction de potentiel n'est en fait qu'une approximation empirique de la fonction de potentiel idéale tirée des calculs de la mécanique quantique ou de la mécanique statistique. Une allégorie proposée par le professeur Gunn résume bien ces trois difficultés. Le premier volet du problème se compare à chercher une aiguille dans une botte de foin. Le deuxième volet du problème nous enfonce davantage dans cette direction, ainsi dans ce cas, on cherche une aiguille dans une botte d'aiguilles. Le dernier volet, en fait le plus sordide, courrone le tout en nous rappelant, qu'on ne connaît même pas la bonne définition de l'aiguille recherchée!

2.1. Méthodes quantiques et méthodes empiriques[29]

La physique moderne constitue la base de diverses méthodes afin de s'attaquer aux multiples problèmes que nous causent les molécules. En effet, pour une molécule donnée, étudier ses propriétés électroniques, ses états de transition lors de réactions chimiques, sa structure, ses divers mouvements moléculaires ne sont en fait qu'une fraction des champs de recherche impliqués dans ce domaine. Dans la cas des protéines, on peut s'intéresser aussi à la thermodynamique et à la cinétique du repliement à partir d'une structure dégénérée vers une structure compacte native.

À une extrémité, on peut utiliser les méthodes dites quantiques. Originellement issues de l'équation de Schrödinger pour les atomes et plus tard généralisée pour les molécules, la mécanique quantique traite de la distribution électronique selon les orbitales. Avec ce type de méthode, la paramétrisation des équations est généralisée et l'équation de Schrödinger peut être résolue. Évidemment, elle comporte aussi de sérieux désavantages. Elle ne s'applique qu'aux petits systèmes moléculaires compte tenu du temps de calcul considérable exigé. Il faut constamment avoir en tête que ce type de méthode ne tient pas compte des noyaux mais de la distribution électronique autour de ceux-ci. Les noyaux sont considérés fixes par l'approximation de Born-Oppenheimer qui stipule que les électrons bougent beaucoup plus rapidement que les noyaux. Ceci implique évidemment un plus grand nombre de variables à traiter.

À l'autre extrémité se situe les méthodes empiriques basés sur la mécanique moléculaire. Le principal avantage de la mécanique moléculaire réside en sa capacité à gérer de gros systèmes en respectant les lois de la physique classique. Les avantages de ce type de méthode deviennent les inconvénients de l'autre. La paramétrisation des équations pose un problème majeur. Plusieurs forces agissent au niveau d'une protéine, le poids de chacune de ces forces n'est pas necéssairement évident à ajuster. Les forces électromagnétiques sont responsables des liaisons entre les atomes. Ces forces sont qualifiées d'harmoniques et doivent être paramétrisées pour chaque mouvements

moléculaire (élongation de liaisons, torsions, rotations et vibrations entre autres). Également à prendre en considération, la formation des ponts H. Il s'agit d'un autre type d'énergie de liaison, mais dans ce cas, le lien formé est beaucoup plus fragile que dans le cas d'un lien covalent. De plus, il faut ajouter l'existence d'interactions de Van der Waals, qui, pour leur part, se définissent comme des forces non-liées. Des atomes partiellement chargées interagissent avec d'autres parties de la molécule portant des charges partielles, sans toutefois former des liaisons covalentes. Non seulement, la nature des interactions diffèrent, mais leurs effets doivent être respectés.

La paramétrisation de ces équations s'obtient souvent des calculs de la mécanique quantique. Même en mécanique moléculaire, certains termes sont en relation étroite avec la mécanique quantique. De ce fait, la détermination des géométries de molécules s'effectuent beaucoup plus rapidement avec les méthodes empiriques comparativement aux méthodes quantiques. Dans certaines situations, on peut même y inclure les effets de solvatation.

Certains groupes de recherches ont développé des méthodes dites semi-empiriques. Dans ce cas, on effectue les calculs selon les recettes de la mécanique quantique en ne considérant que les électrons de valence. Ce type de méthodes se veut un compromis entre les deux extrêmes que représentent respectivement la mécanique quantique et la mécanique classique. Issue des précédentes méthodes, la série de fonctions de potentiel représentant les diverses forces impliquées est sommée et consiste en ce que l'on nomme un champ de forces. CHARMM[30], un champ de forces développé par Karplus est abondamment utilisé pour modéliser des molécules biologiques et des macromolécules.

2.2. Les méthodes stochastiques

Dans certains cas, comme dans notre groupe de recherche, le but premier reste de déterminer le plus précisément possible la structure native sans connaître les détails de son repliement de l'état dénaturé vers son état natif. Ainsi, il est donc possible d'utiliser

des méthodes de recherche stochastique. Il s'agit en fait de naviguer au hasard sur une surface de potentiel et de rejeter ou d'accepter une structure selon des critères énergétiques. L'utilisation de potentiels statistiques peut s'avérer fortement utile dans notre quête.

2.2.1. Les potentiels statistiques

La table de potentiel utilisée dans le cadre de nos recherches est de type statistique. Un potentiel statistique n'utilise pas d'équations théoriques afin de calculer, voire prédire, des effets énergétiques comme c'est le cas d'un champ de forces. Un potentiel statistique, comme son nom l'indique, est basé sur des observations. Par exemple, une liste de structures non-redondante est choisie selon certains critères, et une table de valeurs d'énergie basée sur cet ensemble de structures est compilée de façon statistique. Ainsi, pour utiliser un potentiel statistique lors d'une simulation, aucune fonction analytique n'est requise afin de calculer l'effet d'une interaction. Le concept de potentiel de force moyenne est d'importance capitale ici. Le potentiel de force moyenne se définit comme la différence de potentiel thermodynamique due aux interactions à une distance donnée. Dans le cas du potentiel empirique de Sippl[31], basé sur des structures de protéines résolues par rayons-X, la différence de potentiel thermodynamique s'exprime sous la forme suivante:

$$\Delta E^{ab}(r) = -kT \cdot \ln[f^{ab}(r)/f(r)] \qquad (2.1)$$

où: a et b correspondent à un des 20 acides aminés

$f^{ab}(r)$ est la fréquence relative de la paire a-b à une distance r

$f(r)$ est la fréquence normée de n'importe quelle paire d'acides aminés à une distance r.

Dans un potentiel de force moyenne, on utilise le résultat observé par l'ensemble des interactions impliquées dans un système sans prendre en considération chaque type d'interaction et sa contribution respective. En fait, on peut déterminer le potentiel de force moyenne si on connaît exactement chacune des forces impliquées et leur contribution respective. On l'exprime comme une combinaison linéaire d'un ensemble de fonctions de base normalisées ψ_i (r):

$$\Delta E^{ab}(r) = \Sigma_{i=1}^{q} c^{ab}{}_i \psi_i(r) \qquad (2.2)$$

Cette fonction d'énergie doit être maximisée afin de connaître éventuellement les valeurs des coefficients c^{ab}. Il s'agit d'une séparation minimale qui exclut les paires de résidus directement liés par la chaîne principale. L'équation 2.3 représente une valeur de potentiel pour une paire d'acides aminés qui se décrit comme la somme de deux valeurs d'hydrophobicité individuelle. Ainsi, donc l'équation 2.7, devient avec le terme ajouté:

$$E = \sum_{a\text{-}b\geq 20} (h_a + h_b + h_0) \mid r_a - r_b \mid \qquad (2.3)$$

Le potentiel de Sippl se veut une observation physique des éléments attractifs et répulsifs pour des paires d'acides aminés hydrophobes et hydrophiles. Les valeurs de ce type de potentiel offrent l'avantage d'une représentation fidèle de l'équilibre entre la protéine et le solvant. Par contre, ici encore, l'avantage pour une méthode devient l'inconvénient pour l'autre; ce potentiel ne peut servir pour la dynamique moléculaire. Cependant, ce défaut ne nous affecte pas puisque nous n'effectuons aucune dynamique moléculaire mais bien de la modélisation suivant un processus stochastique. Dans Trip, on utilise ce potentiel basé sur des distributions de paires d'acides aminés. On retrouve ce potentiel sous la forme de l'équation 2.8.

2.2.2. Le niveau de précision des potentiels statistiques[32]

Même dans le domaine d'utilisation où l'on se situe, il faut se méfier de certains aspects des potentiels statistiques. La principale faiblesse des potentiels statistiques réside dans le

postulat stipulant que la fréquence d'une paire d'acides aminés X est indépendante des autres paires d'acides aminés. Dans le volume relativement petit qu'occupe une protéine, l'espace nécessaire aux autres acides aminés impose une importante contrainte sur la position possible de chaque paire. Il en découle que comparativement à un vrai potentiel physique, les potentiels statistiques dépendent de la longueur et de la nature de la chaîne. Autre problème notable, les potentiels statistiques sont construits à partir de banques de données de structures déterminées expérimentalement. Étant donné que la plupart des structures élucidées jusqu'à présent sont des protéines globulaires en phase aqueuse, le potentiel rencontre déjà ses limites. Ce potentiel statistique s'avère donc inefficace pour élucider des protéines membranaires par exemple.

Ce type de potentiel ne peut donc pas représenter quantitativement les énergies réellement impliquées entre les paires d'acides aminés. C'est pourquoi lorsque l'on veut utiliser un potentiel statistique pour modéliser des protéines, il faut ajouter de l'information afin d'améliorer la qualité des prédictions de structures, à défaut de quoi, les résultats sont rarement satisfaisants. Néanmoins, les potentiels statistiques demeurent une référence essentielle pour les méthodes stochastiques dû à leur efficacité afin d'effectuer des calculs rapidement.

2.3. L'explosion combinatoire[33]

Comme il a été mentionné au début de ce chapitre, le nombre astronomique de degrés de liberté des protéines nous cause un problème d'ampleur. De plus, plusieurs degrés de liberté sont fortement corrélés entre eux (en d'autres mots, la variation d'un des degrés de liberté peut amener plusieurs autres degrés de liberté à changer); un acide aminé supplémentaire à l'extrémité d'une chaîne et le nombre de conformations possibles vient de croître exponentiellement.

On peut exprimer le phénomème de l'explosion combinatoire sous forme mathématique pour voir par la suite comment peut-on le gérer. Il s'agit d'effectuer un filtrage

hiérarchique. Soit une protéine de n boucles dont chacune possède N/n résidus. On prend les $m^{N/n}$ conformations de la boucle et on applique un filtrage pour en retenir M. On considère que seulement les M meilleures conformations risquent de participer dans le minimum global. On a donc M^n conformations pour la chaîne de n boucles, au lieu de $m^{N/n}$ comme au début. Si $M^n << m^{N/n}$, on gagne considérablement en temps de calcul. Par contre si $M^n = m^{N/n}$, on se retrouve à la case départ et il n'y a aucune optimisation. Le but ici consiste donc à minimiser M^n, c'est-à-dire trouver le nombre minimum de configurations à retenir. Voyons maintenant comment Trip applique cette idée pour minimiser M. C'est ici en partie ce qui explique pourquoi Trip travaille au niveau des triplets et de boucles de façon hiérarchique.

2.4. Trip et son fonctionnement[34]

Voilà une question préliminaire que l'on peut être tenté de se demander: Pourquoi le nom Trip? Ce nom est simplement issu de la désignation triplet (séquence de trois acides aminés). Mais de ce nom, émane un sens beaucoup plus profond, il soutend en fait l'importance de la philosophie hiérarchique avec laquelle il fut conçu. Dans Trip, on construit et raffine des protéines selon plusieurs niveaux. On génère d'abord des structures à partir de certaines prémices pour ensuite les tamiser par l'intermédiaire de filtres énergétiques. Comme données de départ, on fournit à Trip la séquence des acides aminés et les endroits des structures secondaires. On débute ensuite la construction des protéines en commençant à traiter les résidus, unité de base dans la hiérarchie protéique.

2.4.1. Les listes finies de choix de géométries de résidus

Comme on a vu au premier chapitre, un acide aminé contient plusieurs degrés de liberté. Dans la modélisation de protéine, il s'agit d'un nombre beaucoup trop grand à traiter. C'est pourquoi on utilise un modèle réduit de protéines, c'est-à-dire qu'on ne considère que deux degrés de libertés effectifs par acide aminé, soit les angles ϕ et ψ. Les autres

degrés de liberté sont négligés et les distances des différents liens chimiques demeurent fixes durant la simulation. Dans le cadre de mes recherches, le modèle ne tenait pas compte des chaînes latérales non plus. La notion de chaîne latérale a été réduite au carbone β et ce, même pour un des hydrogènes de la glycine. À noter que la proline ne porte pas de chaîne latérale cyclique dans Trip. Une version ultérieure de Trip tient maintenant compte des chaînes latérales[35], mais elle n'a pas été utilisée dans le cadre de cet ouvrage. Ainsi chaque acide aminé dans les boucles possède une liste finie de paires d'angles qui lui est possible d'adopter. Ces listes sont érigées sous forme de cartes de Ramachandran. La situation diffère pour les acides aminés inclus dans les structures secondaires. En fait, Trip considère officiellement deux types de structures secondaires, l'hélice α et le feuillet β (parallèle et anti-parallèle). Pour ces acides aminés, une seule paire d'angles par structure secondaire est utilisée. Le but est de garder les structures secondaires rigides durant le calcul, question de réduire l'espace de phases. On tire l'information géométrique sur les structures secondaires à partir des données expérimentales[36] comparativement à d'autres groupes de recherche qui ont décidé de minimiser les structures secondaires et tertiaires en même temps[37,38]. Déjà au niveau des résidus, on vient de réduire considérablement l'espace conformationelle accessible.

2.4.2. Le niveau le plus simpliste de la hiérarchie: les triplets

Trip se sert des triplets pour générer des structures. La manipulation des triplets constitue la première étape de la construction de la structure protéique. Des triplets sont générés en choisissant trois états conformationels d'acides aminés. Chaque triplet est immédiatement accepté ou rejeté dépendemment si les coordonnées des extrémités du dit triplet correspondent à une région permise dans l'espace conformationel des triplets. Le triplet lui-même s'appuie sur l'espace dièdre de chaque acide aminé[39]. Pour obtenir une distribution de la population de triplets, on doit effectuer un classement selon des critères géométriques bien définis. On peut voir à la figure 7, la forme des coordonnées géométriques utilisées pour le classement. Le vecteur R désigne les coordonnées entre le premier carbone α et le dernier carbone α. Comme on peut le voir à la figure 7, les deux

premières coordonnées correspondent aux angles polaires entre chaque résidu et le vecteur de séparation. La troisième coordonnée se rattache à l'angle dièdre des deux axes polaires selon le vecteur de séparation. Finalement, les deux derniers représentent des rotations de chacun des résidus autour de leur propre axe. Chacun de ces 5 éléments sera discrétisé et les 5 indices constituent des "boîtes de classement". Chacune de ces boîtes contient des géométries de triplets qui lui sont propres.

$$q_1 = \left(\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{R}}\right)$$
$$q_2 = \left(\hat{\mathbf{x}}_2 \cdot \hat{\mathbf{R}}\right)$$
$$|q_3| = \cos^{-1}\left(\frac{\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{x}}_2 - \left(\hat{\mathbf{x}}_1 \cdot \hat{\mathbf{R}}\right)\left(\hat{\mathbf{x}}_2 \cdot \hat{\mathbf{R}}\right)}{\sqrt{(1-q_1^2)(1-q_2^2)}}\right)$$
$$|q_4| = \cos^{-1}\left(\frac{\hat{\mathbf{y}}_1 \cdot \hat{\mathbf{R}}}{\sqrt{1-q_1^2}}\right)$$
$$|q_5| = \cos^{-1}\left(\frac{\hat{\mathbf{y}}_2 \cdot \hat{\mathbf{R}}}{\sqrt{1-q_2^2}}\right)$$

$$\hat{\mathbf{x}} = \hat{\mathbf{r}}_{N-C_\alpha}$$
$$\hat{\mathbf{y}} = \frac{\hat{\mathbf{r}}_{N-H} - \cos\theta\,\hat{\mathbf{r}}_{N-C_\alpha}}{\sin\theta}$$
$$\hat{\mathbf{z}} = \frac{\hat{\mathbf{r}}_{N-C_\alpha} \times \hat{\mathbf{r}}_{N-H}}{\sin\theta}$$

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

FIGURE 7: Forme mathématique des coordonnées q1 à q5.

Ainsi, la population des triplets dans chacune des boîtes selon chacune des 5 coordonnées représente la distribution des géométries de triplets acceptés qui servira dans les étapes hiérarchiques subséquentes. À l'échelle de l'espace de phases, remplacer un triplet par un autre qui lui ressemble permet de faire un petit pas dans l'espace conformationnel.

2.4.3. Le pas hiérarchique suivant: les boucles

La méthode pour travailler avec les boucles comporte certaines similitudes avec l'étape hiérarchique précédente. Les buts atteints sont cependant quelque peu différents. Les

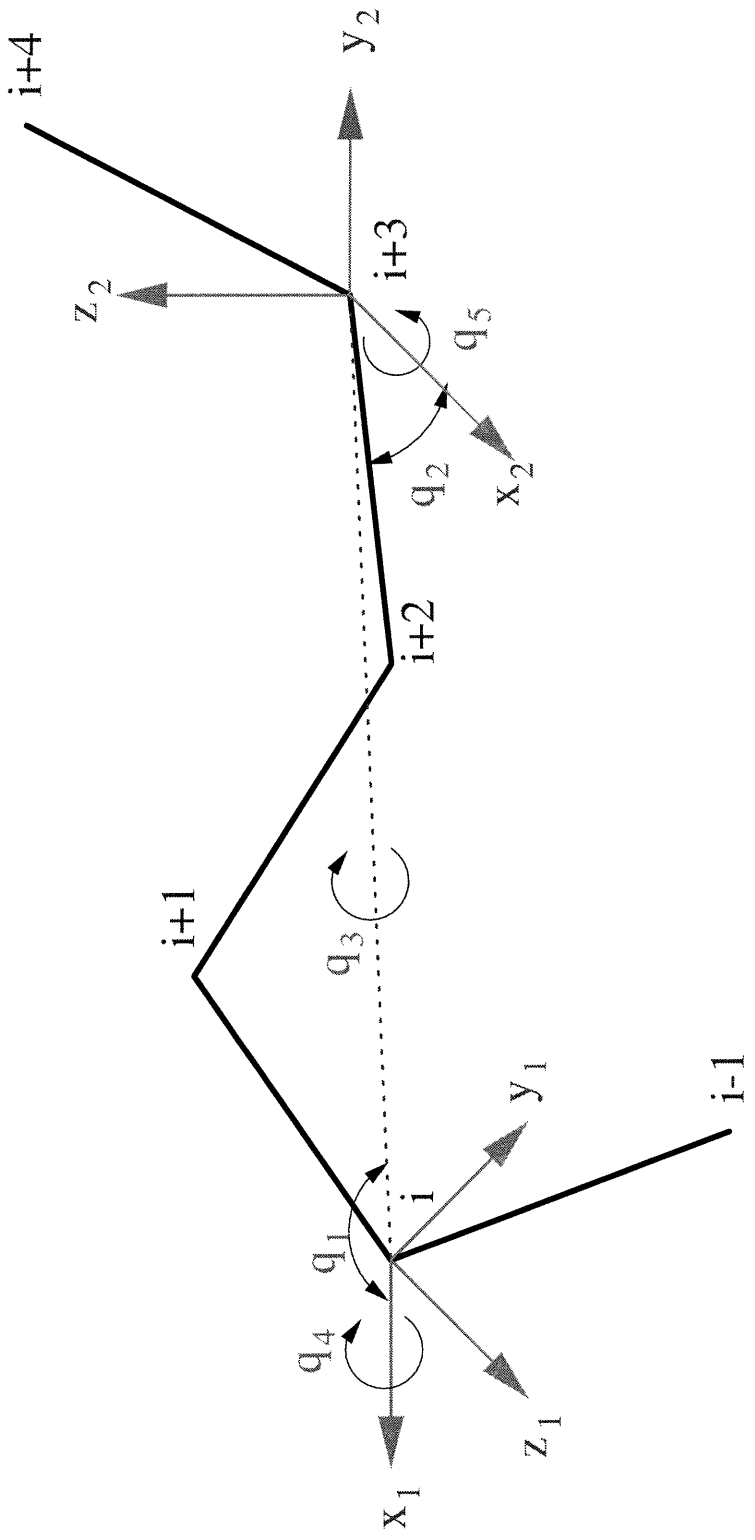


FIGURE 8: La géométrie d'une boucle avec ses coordonnées q1 à q5.

boucles sont définies dans Trip comme un enchevêtrement de triplets se retrouvant entre les structures secondaires. Ici encore, on effectue des modifications sur des boucles en mutant certains triplets, c'est-à-dire échanger une géométrie d'un triplet ciblé pour une autre géométrie de triplet. Trip effectue ensuite une comparaison de structures de boucles modifiées par rapport à la géométrie de la boucle de départ. Les boucles formées doivent être différentes de la géométrie de départ sans nier toute ressemblance de son origine. Le classement de ces boucles s'effectue selon le même système que les triplets. On retrouve ainsi une liste de boucles ayant évoluées en terme de géométrie tout au long du calcul. La méthode de filtrage employée ici permet d'ajouter du poids sur les boîtes importantes (celles qui comportent des géométries possibles). C'est en fait le deuxième biaisage géométrique durant le processus.

2.4.4. La dernière étape hiérarchique: la génération de molécules

La construction de la protéine globale s'effectue à cette étape. Il s'agit de choisir une boucle parmi celles retenues dans la liste et de faire un essai. Ces boucles sont insérées entre les structures secondaires préalablement établies fixes. L'acceptation ou le rejet de cette nouvelle molécule est établie sur un critère purement énergétique issu de la fonction de potentiel statistique. La molécule générée évoluera dans un ensemble suivant les pas de l'algorithme Monte Carlo. Une structure ne répondant pas au critère énergétique de sélection sera systématiquement rejeté. À l'inverse, une structure qui a passé tous les critères énergétique (coarsed-grained potential) verra son avenir contrôlé par l'intermédiaire d'un algorithme Monte Carlo Recuit-Simulé.

2.5. L'algorithme Monte Carlo et le recuit simulé

L'algorithme monte Carlo constitue un outil puissant afin d'échantillonner une surface de potentiel. La difficulté de naviguer dans un espace de phases réside dans la recherche du minimum global. Puisqu'il existe plusieurs variations, c'est ainsi que l'on considère la

partie de la surface de potentiel. On postule ici que ce même manège répété sur une période de temps suffisamment longue finira par représenter statistiquement et thermodynamiquement l'espace de phases en question.[42]

Afin de rechercher le minimum global, un recuit simulé a été combiné à l'algorithme. Le recuit simulé est traduit de l'anglais "simulated annealing" qui indique que l'on soumettra notre système à l'effet d'une température virtuelle. L'abaissement graduel de cette température bloquera l'accès à un certain nombre de choix préalablement accessibles. On peut imager la situation suivante comme une réaction où l'on doit fournir de la chaleur afin de franchir la barrière de potentiel entre les réactifs et les produits. L'abaissement de la température à un moment donné de la réaction tranchera d'un côté des molécules de produits et de l'autre les produits de départ n'ayant pas réagi. Dans Trip, la fonction d'abaissement de température est exponentielle et s'exprime de la façon suivante:

$$T_k = (T_i^{N-k} \cdot T_f^{k})^{1/N} = T_i e^{-Ak} = e^{-A} T_{k-1} \qquad (2.9)$$

où: $A = (1/N) \cdot \log(T_i / T_f)$

N est le nombre d'itérations

k est l'itération courante

T est la dite température virtuelle

Il faut prendre soin de rappeler ici que la température est dite virtuelle, elle n'a aucun lien avec la température physique du système à l'étude.

Le recuit simulé comporte quand même certains inconvénients lorsque juxté à l'algorithme Monte Carlo. Le Monte Carlo seul respecte ce que l'on appelle en mécanique statistique le bilan détaillé, c'est-à-dire que chaque "mouvement" Monte Carlo sur l'espace de phases respecte l'équilibre énergétique. Le cas n'est plus necéssairement vrai avec la combinaison du recuit simulé. Le nombre d'essais trop petit ne permettant pas d'atteindre l'équilibre constitue le véritable problème. Si le recuit simulé s'effectue sans être en équilibre, il devient impossible de trouver le minimum

global. De plus, certains termes énergétiques dans Trip font en sorte que l'énergie calculée n'a pas de signification physique dans quelconque système thermodynamique. Malgré ce manque de signification physique, la fonction d'énergie reste toujours la fonction à minimiser.

2.6. Gérer un ensemble de molécules grâce à l'algorithme génétique

Les structures protéiques qui existent à ce stade ont passé à travers les divers filtres énergétiques. Maintenant, il reste à améliorer le mieux possible ces structures potentiellement près de la structure native. La stratégie ici est d'utiliser un algorithme génétique. On effectue une modification sur les 64 structures sélectionnées (les dit parents) par l'intermédiaire de l'algorithme Monte Carlo. C'est en fait une création de structures hybrides. L'ensemble des hybrides (les dit enfants) est beaucoup plus grand que celui des parents, soit 512 structures. On refait ici le même manège, on procède à des modifications puis on retient un sous-ensemble canonique de 64 parents pour générer à nouveau une génération fraîche de structures protéiques. Ainsi les nouveaux parents possèdent des ressemblances avec leurs prédécesseurs et l'on peut former des enfants qui seront d'énergie plus basse.

2.7. Évaluation de structures protéiques

La meilleure façon que l'on connaisse à présent afin d'évaluer des structures consiste en une comparaison qui oppose l'énergie de la structure et son RMSD (root-mean-square deviation) ou en français, sa déviation moyenne quadratique. Il s'agit en fait de comparer la ressemblance entre deux ensembles de points. Il existe deux types de RMS, le premier est une comparaison des distances. Voici l'expression mathématiques qui la décrit:

$$\Delta d = (\sum_{i=1}^{N} \sum_{j=1}^{N} (d_{ij} - e_{ij})^2/n^2)^{½} \qquad (2.10)$$

où: $d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{½}$

En mots, d_{ij} correspond aux distances interatomiques dans la structures élucidée expérimentalement et e_{ij} correspond aux distances interatomiques de la structure générée par l'ordinateur.[43]

La deuxième méthode, dite de rotation va comme suit:

$$\Delta r = (\sum_{i=1}^{n} (x_i - y_i)^2 / n)^{1/2} \qquad (2.11)$$

Cette méthode ici nécessite d'abord une superposition optimale des deux structures que l'on veut comparer puisque x représente les coordonnées de la structure connue et y représente les coordonnées de la structure à l'étude, n étant le nombre d'atomes.

La mesure de RMS dans le cadre de mes recherches provient d'une solution analytique de la méthode des rotations.[45] Une correction apportée par Kabsch gère le cas où des rotations impropres seraient en cause.[46] Qu'est-ce que nous indique le RMS? Au-delà de 10 Å, cette mesure ne peut absolument rien nous indiquer sur la qualité de la présente structure. Ces structures ne sont aucunement comparables à la native. Dans l'intervalle allant de 5Å à 9Å, on se situe dans une zone grise. Dans cette zone, on ne peut rien conclure. La protéine générée est peut-être très près de la native à l'exception d'une structure secondaire inversée. En deçà de 5Å, on est en présence d'une structure probablement très près de la native. Normalement, sur un graphique de l'énergie en fonction du RMS, on devrait observer une certaine corrélation entre ces deux valeurs pour plusieurs protéines.

2.8. Tirer profit de l'information biologique

L'architecture de Trip, bien qu'efficace à la base, nécessite des informations supplémentaires afin d'effectuer des simulations numériques dans le but d'élucider une structure. Présentement, Trip crée ses propres banques de triplets au hasard. À partir de

certains choix dans cette banque, une liste de boucles sera érigée. Il pourrait être intéressant de voir ce que Trip peut faire avec un échantillonnage différent. C'est à ce moment que l'on peut avoir recours à la meilleure information possible, celle de mère Nature. Cette prémice constitue la fondation de mon projet de recherche. Ainsi le chapitre trois traitera de l'apport d'une banque de géométries naturelles de triplets au fonctionnement de Trip et le chapitre quatre sera élaborée autour d'une banque de géométries de boucles naturelles. Le principe de fonctionnement de base restera inchangé à l'exception du poids qui sera porté sur la distribution biologique de la population des triplets et des boucles.

CHAPITRE 3

L'apport d'une distribution biologique de géométries de triplets

3.1. Un aperçu de la méthodologie

Afin d'apporter de l'information biologique au niveau hiérarchique des triplets dans le logiciel Trip, il faut suivre un certain protocole dans le but d'élaborer une banque de données. De cette banque de données, il suffira d'extraire l'information géométrique nécessaire afin d'alimenter Trip. Des modifications devront être apportées à Trip pour qu'il puisse inclure et jongler avec ces nouvelles données. Finalement, des simulations seront lancées pour vérifier si la nouvelle distribution biologique oriente l'algorithme dans sa recherche d'une bonne structure tridimensionelle.

3.2. Utiliser une liste de triplets adéquate

En premier lieu, on veut travailler sur un ensemble de protéines non-redondant au niveau structurel. Le but consiste à obtenir la plus grande diversité possible de structures afin de couvrir un large éventail de géométries de triplets. Un ensemble de structures redondantes permettrait d'attribuer un poids à certaines géométries de triplets. Mais étant donné que nous n'avons pas une approche probabilistique, il n'est point question de considérer un poids variable. Il n'est pas nécessaire d'assembler un lot de structures parfaitement résolues par une méthode quelconque, puique l'on veut les coordonnées internes de ces triplets et non les coordonnées cartésiennes de chacun des atomes des résidus. On passe ainsi de 54 coordonnées cartésiennes par triplet (pour la chaîne principale) à 5 coordonnées internes, soit une réduction importante du nombre de degrés de liberté à traiter dans Trip. À la limite, si certains atomes essentiels au calcul des coordonnées internes sont manquants, il demeure possible d'utiliser des routines pour compléter ces trous.

Pour trouver un ensemble de structures protéiques présentant différentes identités séquentielles selon un certain critère de résolution, la page web du groupe de Roland Dunbrack fut consultée. C'est à partir de cette liste de protéines déjà établie que la banque de données sera construite. La méthode employée par le groupe de Dunbrack pour construire ces listes de protéines se décline comme suit. D'abord, l'alignement local des séquences s'effectue par l'intermédiaire du programme BLAST[47,48]. Les alignements globaux présentent un défaut majeur comparativement aux alignements locaux. Dans un alignement global, l'identité séquentielle se calcule sous forme d'un pourcentage selon une séquence cible. Comme les protéines présentent souvent différents domaines, il est possible qu'un des domaines soit très similaire à la cible mais que les autres diffèrent considérablement. Ainsi, on perd l'essentiel de l'information de similitude entre les structures, d'où l'utilisation de l'algorithme de BLAST. L'algorithme, qui ressemble à celui de Hobohm & Sander[49], débute en triant toutes les séquences selon des critères de résolution (de la meilleure à la pire) et de longueur (de la plus longue à la plus courte). Ensuite, les séquences qui ne répondent pas aux critères mentionnés sont éliminées de la liste. La première séquence de la liste est définie comme étant la cible. Par la suite, l'algorithme retire toutes les séquences dépassant un seuil limite d'identité. Le processus est répété pour toutes les séquences subséquentes dans la liste.

3.3. Statistiques d'ensemble sur les chaînes retenues et les triplets

À ce moment, il faut être conscient de deux faits lorsque l'on construit une banque de données. D'abord, la qualité de l'information structurelle d'une archive ne peut être meilleure que l'information de la source. La recherche de redondance peut cependant être un indicateur de qualité du matériel contenu dans cette banque. Ensuite, on doit respecter l'information de la source sans la modifier. Seule l'annotation de divers aspects selon un langage préalablement établi et la mise à jour régulière de la banque de données peut aider à assimiler et à compléter l'information déjà présente.[50]

Maintenant qu'on possède une liste de protéines adéquate, on peut procéder à l'élaboration d'une banque de données de protéines. Pour les besoins du présent travail, une liste contenant 2376 chaînes ayant un seuil d'acceptation d'identité séquentielle de 90% ou moins, un facteur de résolution de 3.0 Å ou moins et un facteur R selon BLAST (soit une pénalité attribuée par résidu pour un trou dans la séquence) de 1.0. Le seuil d'acceptation séquentielle dans ce cas-ci signifie que lorsque l'on compare les structures primaires entre elles, au maximum, 90% des acides aminés se situant à une position spécifique dans une séquence se retrouveront au même endroit dans une autre séquence de l'ensemble choisi. Le seuil d'acceptation d'identité séquentielle semble élevé à première vue, mais c'est un mal necéssaire, si on veut un vaste échantillonnage. Le facteur de résolution pour sa part signifie qu'il s'agit de la limite inférieure de l'écart observable entre deux atomes. En d'autres mots, on considère qu'au-delà de 3.0 Å, on ne connaît pas suffisament d'informations sur la situation exacte des atomes pour considérer la protéine. Grâce à cette liste, une banque de données de protéines sera élaborée. De cette banque de données de protéines, un banque de triplets issu de boucles sera élaborée. Cette banque de triplets contiendra de l'information essentielle pour Trip afin de modifier le poids de la distribution des géométries de triplets. On considérera chaque triplet selon sa séquence en acides aminés, sa structure (selon que chaque acide aminé et ses deux voisins immédiats de part et d'autre font partie d'une boucle, d'un feuillet ou d'une hélice) et finalement de ses coordonnées internes. On considère des voisins de part et d'autre du triplet afin de vérifier ultérieurement à quel point la structure des voisins influence la géométrie d'un triplet.

Au total, quelques 184 661 triplets furent extirpés des boucles des protéines retenues. La figure 9 illustre l'échantillonnage de diverses combinaisons d'acides aminés. L'abscisse est décrite sous forme d'entier pour chaque combinaison de triplet par la formule:

$$\Sigma_{m=1,8000}\, m = \Sigma_{i=1,20}\, \Sigma_{j=1,20}\, \Sigma_{k=1,20}\, (aa)_i (aa)_j (aa)_k \qquad (3.1)$$

où les indices de 1 à 20 correspondent au code Trip pour les 20 acides aminés (voir figure 5 au chapitre premier).

À première vue, le graphique semble montrer une distribution abondante dans la grande majorité des cas. Le graphique que l'on peut observer à la figure 9 semble démontrer des patrons de distribution, mais ces patrons n'ont aucune signification puisque les codes de triplets sont énumérés selon un ordre purement mathématique. Les triplets les plus abondants, contiennent généralement les résidus glycine, leucine, proline ou une combinaison de ces résidus.

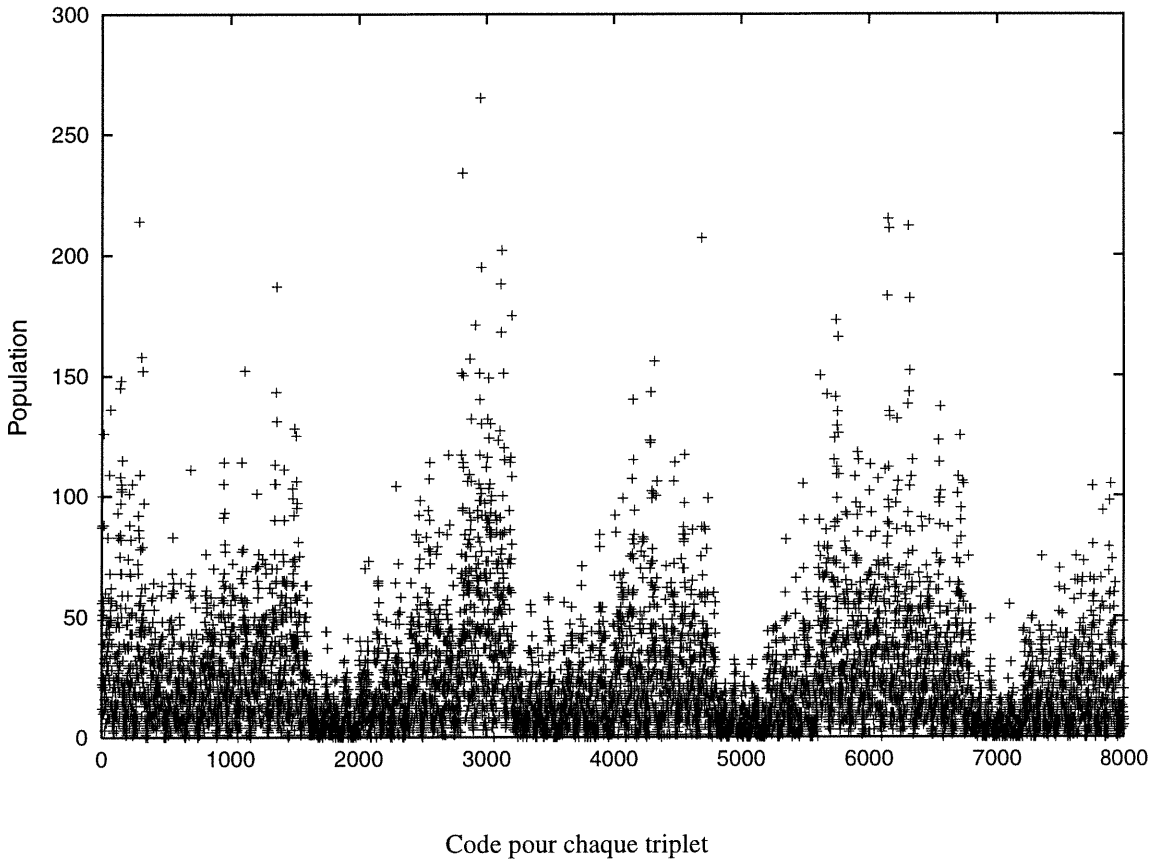


FIGURE 9: Distribution des divers triplets selon leur séquence.

Malgré la densité de population des diverses séquences de triplets, certains demeurent rares, voir inexistants. En fait, 546 combinaisons de triplets sur 8000 n'existent pas dans la banque de données. De ces combinaisons manquantes, 22.3% sont des triplets

contenant un tryptophane, 17.2% sont des triplets comportant une cystéine et 16.1% des triplets sont composés du résidu méthionine. L'absence du tryptophane se justifie par sa taille. Les deux autres chaînons manquants sont en fait les résidus contenant un soufre dans leur structure.

Un deuxième critère contribuera à déterminer cette distribution, il s'agit d'un critère de résidus avoisinants. Ce critère est établi selon la nature des deux voisins de chaque côté des triplets. On regroupe la nature structurelle de ces voisins en trois ensembles, soit ceux utilisés dans Trip. Un ensemble regroupant les hélices α, un ensemble regroupant les feuillets β parallèles et anti-parallèles et un ensemble appelé boucle, incluant tout le reste. On traite ici avec des séquences fixes: triplets (séquence de trois acides aminés), deux acides aminés précédent le triplet et deux acides aminés suivant le triplet. On définit ainsi des classes de voisins pour chacune des combinaisons possibles. On dénote chaque combinaison plausible de voisinage de triplets. Par exemple, un triplet débutant une boucle précédé d'une hélice α se verra noté hélice - hélice - triplet - boucle - boucle. Il en va de même avec d'autres type de voisins. Il existe toutefois des combinaisons impossibles qu'il faut retrancher. Par exemple, boucle - boucle - triplet - feuillet - hélice. Il est physiquement impossible d'avoir un feuillet d'un seul acide aminé. Le tableau II montre les différentes combinaisons de voisins possibles portant l'indication 1 pour les résidus d'une boucle, l'indication 2 pour les résidus d'un feuillet et l'indication 3 pour les résidus d'une hélice. Il y a ainsi 25 codes d'environnement représentant chacun les acides aminées ceinturant le triplet.

La figure 10 à la page 40 illustre bien la présence de chacun des types possibles. Cette figure illustre en toute logique que l'environnement le plus fréquent se situe lorsque le triplet est ceinturé de résidus inclus dans des boucles. Parmi les autres combinaisons de voisins présentes dans des proportions de 5 à 10% de l'ensemble, on retrouve, des combinaisons de voisins formées d'un ou deux résidus se retrouvant dans des feuillets. Les plus rares sont les environnements mixtes, hélices et feuillets ou encore les environnements formés à 75% ou 100% d'hélices.

TABLEAU II: Les 25 environnements de triplets possibles

Résidus précédents un triplet	Résidus suivants un triplets	Code d'environnement
1-1	1-1	1
2-1	1-1	2
3-1	1-1	3
2-2	1-1	4
3-3	1-1	5
1-1	1-2	6
2-1	1-2	7
3-1	1-2	8
2-2	1-2	9
3-3	1-2	10
1-1	1-3	11
2-1	1-3	12
3-1	1-3	13
2-2	1-3	14
3-3	1-3	15
1-1	2-2	16
2-1	2-2	17
3-1	2-2	18
2-2	2-2	19
3-3	2-2	20
1-1	3-3	21
2-1	3-3	22
3-1	3-3	23
2-2	3-3	24
3-3	3-3	25

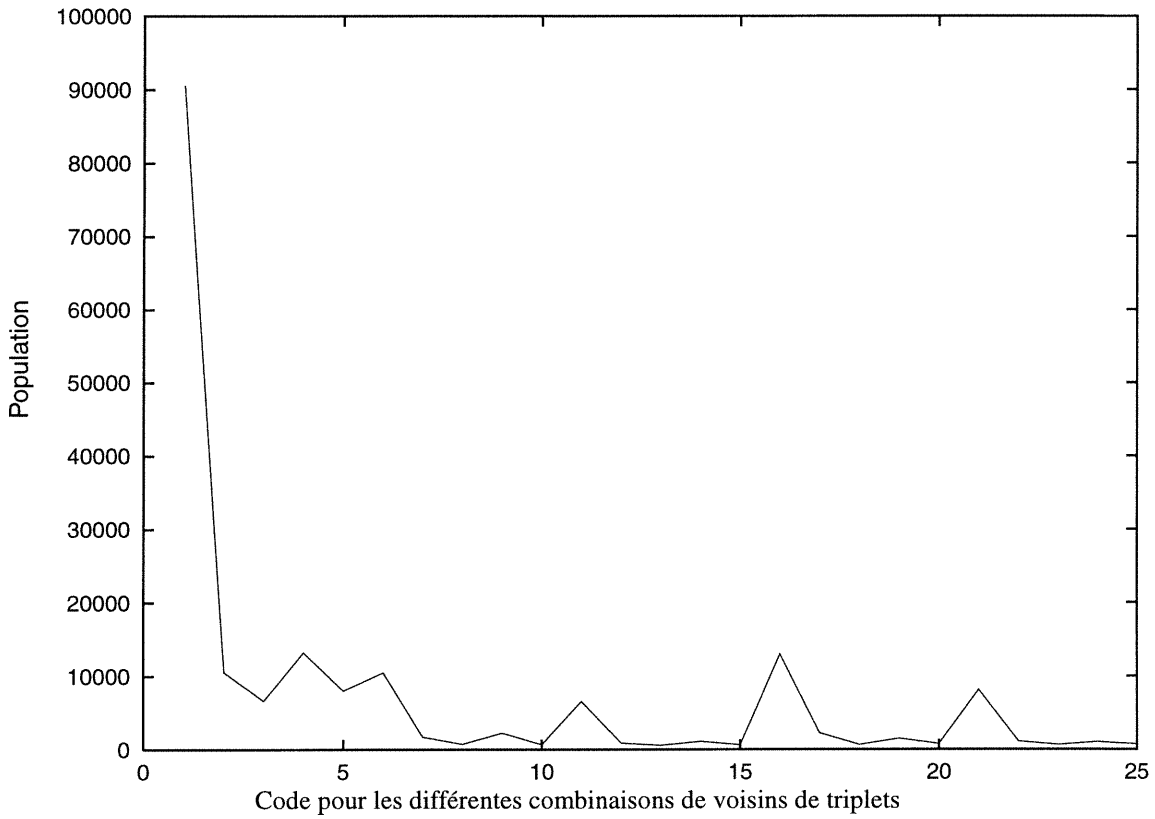


FIGURE 10: Distribution des différentes combinaisons de voisins de triplets

D'une part, on possède une liste de triplets de toutes les combinaisons de voisins plausibles, mais dont certaines séquences sont inexistantes. Ceci n'empêche en rien la poursuite du travail puisque l'on fonctionne avec des seuils d'acceptation basés sur des comparaisons de séquence et de combinaison de voisins et non sur la présence absolument nécessaire de chacun des types de triplets.

La dernière statistique à prendre en considération consiste à savoir si la nature géométrique des triplets réussit à combler l'espace conformationnel de classement de Trip. Qu'est-ce que l'espace conformationnel de classement? Cet espace se définit par un nombre de boîtes de classement (le paramètre b dans Trip). Ce paramètre désigne l'espace conformationnel par degré de liberté donnant au total b^5 boîtes de classement. La grandeur de chaque division est de 360°/b (pour les angles dièdres), et 2/b pour les cosinus d'angles polaires. Pour l'instant, ce paramètre b fut optimisé à 4 suite à une

utilisation du logiciel sans distribution d'informations biologiques. On négocie donc présentement avec un espace conformationnel de 1024 divisions. La figure 11 qui suit nous montre qu'effectivement, les triplets de la distribution naturelle, réussissent à couvrir l'ensemble des divisions de l'espace conformationnel.

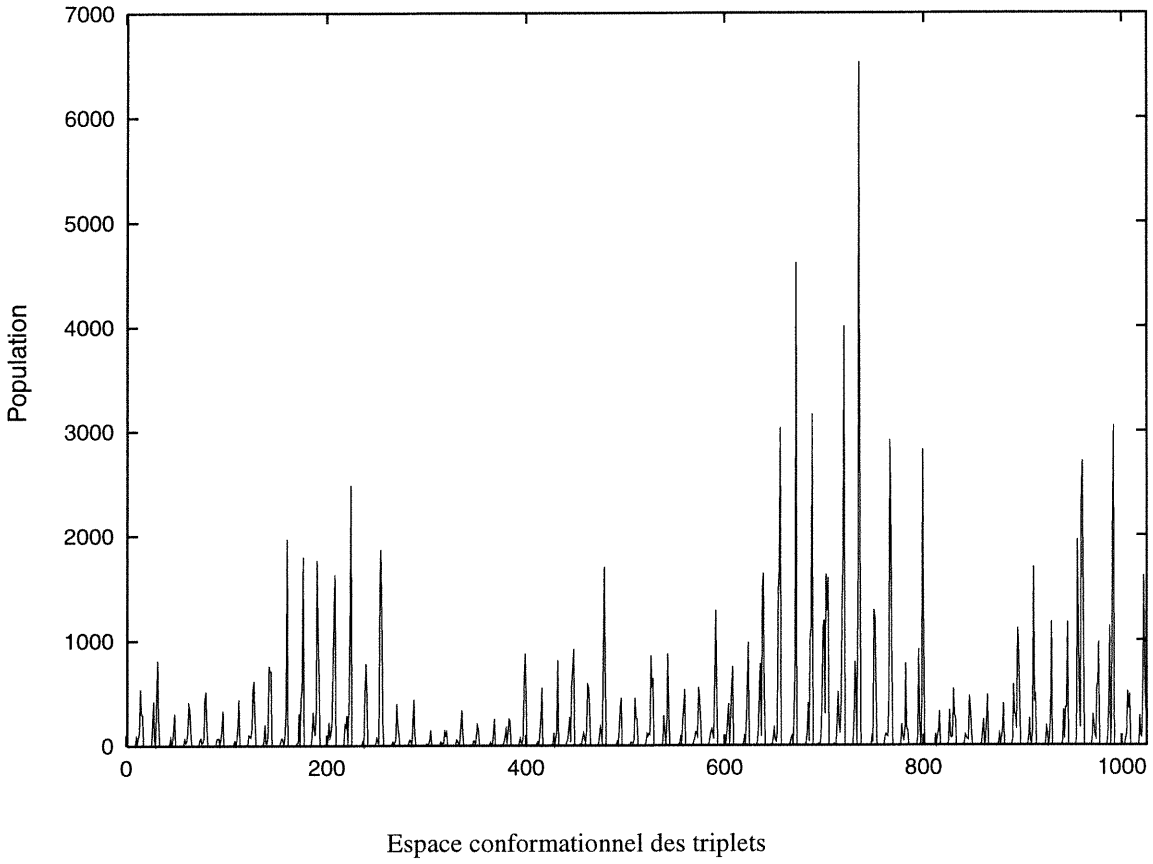


FIGURE 11: Distribution des géométries naturelles de triplets dans l'espace conformationel de Trip.

Certaines boîtes sont plus remplies que d'autres, c'est le fruit de la distribution biologique. L'emphase de cette distribution permettra à Trip d'orienter son processus de construction de triplets. Il s'agit maintenant d'ajuster cette distribution en fonction des seuils d'acceptation pour la séquence du triplet et le voisinage du triplet.

3.4. Homologie de séquence et de structure

Pour ajuster des seuils d'acceptation, il faut utiliser un indicateur de similarité intermédiaire entre l'information de la banque de données d'informations biologiques et l'information contenue dans Trip. L'outil auquel on a recourt dans le cadre de ces recherches se nomme homologie. On cherche ainsi un indicateur afin de savoir si deux structures sont suffisamment homologues entre elles pour éventuellement les accepter ou les rejeter. Pour l'homologie de séquence, les matrices de Dayhoff s'avèrent adéquates.

Les matrices de Dayhoff[51], aussi appelées PAM (Point Accepted Mutations) constituent l'élément clé d'un modèle représentatif des mutations d'acides aminés au cours de l'évolution. Ces mutations observées dans la nature ne sont valides qu'à deux conditions: d'une part, il existe une dépendance en rapport avec la quantité de mutations dans une même portion d'un gène produisant un acide aminé dans une protéine. D'autre part, l'espèce subissant cette mutation doit "accepter" cette mutation comme une nouvelle forme prédominante. Dans le modèle de Dayhoff, ces mutations sont exprimées sous forme matricielle selon des diverses échelles de temps. Ces matrices permettent la comparaison des 20 acides aminés les uns par rapport aux autres, donc au total, 400 scores indiquant une probabilité de mutation entre deux acides aminés. Plus le score est élevé, plus la mutation entre ces deux acides aminés est probable. Les éléments non-diagonaux de la matrice répondent à l'équation:

$$M_{ij} = (\lambda\, m_{ij}\, A_{ij}) \cdot (\Sigma_i\, A_{ij})^{-1} \qquad (3.2)$$

où: A_{ij} représente un élément de la matrice PAM

λ est une constante de proportionalité

m_j représente la mutabilité du j ième acide aminé

Les éléments diagonaux, eux, prennent les valeurs fixées par: $M_{jj} = 1 - \lambda m_j$.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	9730	0	31	24	5	34	37	42	5	3	5	18	19	5	54	99	45	0	0	32
R	0	9981	5	0	0	13	0	0	17	0	0	23	18	2	0	1	0	0	0	0
N	14	7	9701	36	0	20	7	10	24	4	2	19	1	0	10	51	17	0	0	4
D	13	0	45	9757	0	27	96	8	6	0	2	8	1	0	1	26	2	0	0	4
C	1	0	0	0	9928	0	0	1	0	2	0	0	11	0	0	12	3	0	0	6
Q	12	14	15	16	0	9736	24	4	14	4	2	9	11	0	11	13	10	0	0	5
E	21	0	9	95	0	40	9726	13	4	4	4	13	1	0	17	15	12	0	0	7
G	40	0	22	13	3	11	22	9870	1	0	2	5	0	0	17	42	8	0	0	7
H	2	19	20	4	0	15	3	0	9865	4	3	6	0	3	0	10	5	11	4	1
I	1	0	3	0	3	4	3	0	4	9703	22	4	22	14	2	3	14	0	0	70
L	4	0	3	3	0	4	7	2	6	52	9899	6	99	19	0	5	7	0	0	24
K	17	65	37	13	0	23	21	5	14	9	6	9845	11	0	6	22	14	0	4	13
M	2	7	0	0	5	4	0	0	0	7	14	2	9672	5	0	5	2	0	0	12
F	2	3	0	0	0	0	0	0	4	18	10	0	18	9879	0	5	2	30	74	2
P	23	0	9	1	0	13	13	7	0	3	0	0	3	0	9850	11	5	0	0	4
S	59	2	67	28	27	22	16	26	17	4	3	14	23	6	15	9598	69	0	0	7
T	30	0	25	3	8	20	14	6	8	24	5	10	11	3	8	76	9759	0	0	20
W	0	0	0	0	0	0	0	0	4	0	0	0	0	8	0	0	0	9941	7	0
Y	0	0	0	0	0	0	0	0	4	0	0	2	0	51	0	0	0	17	9909	0
V	27	0	7	5	18	12	10	6	3	156	22	12	82	3	8	9	25	0	0	9783

FIGURE 12a: Matrice PAM pour une distance évolutive de 2, ce qui signifie qu'il y a deux mutations acceptées par 100 acides aminés. Les éléments de la matrice ont été multipliés par 10000 pour des raisons de simplicité.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	14	6	11	11	7	11	11	13	7	8	6	9	8	4	13	12	12	1	2	9
R	2	26	3	2	1	4	2	1	6	1	1	7	3	1	1	2	2	1	1	2
N	5	4	7	6	2	5	5	5	6	3	2	5	3	1	5	5	5	1	1	3
D	6	3	7	12	2	7	10	6	5	3	3	5	3	1	5	6	5	1	1	4
C	2	1	2	1	41	1	1	1	1	2	1	1	3	1	1	3	2	0	0	3
Q	4	4	4	4	2	7	4	3	4	2	2	4	3	1	4	4	4	1	1	3
E	6	3	6	10	2	7	10	6	4	4	3	5	4	1	6	6	5	1	1	4
G	12	4	10	10	5	9	10	26	5	5	4	7	5	2	11	11	9	1	1	6
H	3	7	5	3	1	4	3	2	20	2	2	4	2	2	2	3	3	5	3	2
I	3	2	2	2	3	3	2	2	2	9	6	3	6	4	2	3	4	1	2	7
L	5	4	5	4	4	5	5	3	5	15	34	5	18	10	3	5	6	3	5	12
K	8	19	10	9	3	10	9	6	9	6	5	21	7	3	7	8	8	1	3	7
M	1	1	1	1	1	1	1	1	1	2	3	1	3	1	1	1	1	0	1	2
F	2	2	1	1	1	1	1	1	3	5	5	1	5	31	1	2	2	18	29	3
P	6	2	4	4	2	5	5	5	2	3	2	3	2	1	18	5	4	0	0	3
S	7	4	7	7	6	6	7	7	5	4	3	6	5	2	7	7	7	1	1	5
T	8	4	7	6	6	7	7	6	5	7	5	6	6	3	6	8	11	1	1	7
W	0	0	0	0	0	0	0	0	2	0	0	0	0	5	0	0	0	48	5	0
Y	1	1	1	0	0	1	0	0	2	2	2	1	2	20	0	1	1	14	41	1
V	8	4	6	6	9	7	6	6	5	16	11	6	12	5	6	7	9	1	2	17

FIGURE 12b. Matrice de Dayhoff PAM256. Pour simplifier l'apparence, les valeurs affichées sont multipliées par 100.

En regardant les figures 12a et 12b représentant respectivement les matrices de Dayhoff PAM 2 et PAM 256, on constate rapidement des différences au niveau des éléments diagonaux. Les chiffres acompagnant les PAM correspondent au nombre de mutations par 100 acides aminés. Sur une distance évolutionnaire de 2 PAM, les probabilités de mutations d'un acide aminé par lui-même sont plutôt élevées comparativement aux probabilités exprimées dans la PAM256. La logique derrière ces matrices stipule que la probabilité de mutation sera qualitativement proportionnelle au degré de ressemblance aux niveaux chimique et physique. C'est ainsi que, par exemple, le tryptophane ne possèdent que de très faibles chances d'être muté par rapport à la quasi-totalité des acides aminés.

Dans le cadre de mes recherches, le choix d'une matrice PAM sera basé sur certaines ressemblances entre familles d'acides aminés et non selon l'échelle de temps impliquée. La matrice PAM256 semble bien répondre à ces critères de ressemblances qualitatifs. Les critères sont établis sur la définition des familles au premier chapitre. Ainsi, par exemple, les acides aminés possédant un soufre seront avantagés pour se muter entre eux. Au niveau de la diagonale, il semble plus raisonnable de considérer ceux de la PAM256. La probabilité de mutation d'un acide aminé par lui-même reste la plus élevée sans toutefois masquer par son importance tous les autres scores exprimés par les éléments non-diagonaux.

Ces proportionalités mutationelles représentent un bon point de départ afin de choisir des géométries de triplets selon un critère séquentiel. Un score sera calculé pour chacun des triplets. Ce score s'exprime en fait comme la somme sur tous les résidus du triplets comparés un par un avec la séquence cible. Un seuil d'acceptation sera fixé dès le départ d'une simulation; c'est ce dernier, par l'intermédiaire de la matrice de Dayhoff, qui décidera de la distribution biologique et géométrique des triplets.

Afin d'évaluer un seuil d'acceptabilité au niveau du voisinage d'un triplet, on doit employer un autre type de matrice que celle de Dayhoff. En fait, il faudra créer une nouvelle matrice de comparaison de voisinage de triplet.

Pour mesurer une ressemblance entre deux types de voisinage de triplet, il semble intéressant d'utiliser le RMS (calculé avec tous les atomes). Avec cette mesure, on peut comparer cinq types de voisinage de part et d'autre du triplet (voir le tableau II pour les différents types de voisinages). Évidemment on rejette les combinaisons où un seul résidu est classé feuillet ou hélice (ce cas impliquerait une erreur dans les données brutes). La comparaison de toutes les possibilités formera une matrice 25 par 25. Contrairement aux matrices de Dayhoff qui comportent un facteur de population naturelle, cette matrice de mesures de RMS est symétrique. La taille de cette matrice justifie en partie la longueur de fenêtre pour le voisinage de part et d'autre du triplet. Si on considère trois acides aminés de chaque côté, on se retrouve (avant élimination) avec $(3^2)^2$, c'est-à-dire 81 choix. Ceci implique une matrice 81 par 81 qui exige beaucoup trop de sous-classes à traiter inutilement. À l'inverse, traiter un seul acide aminé de chaque côté ne suffit pas. Un seul voisin ne renseigne pas suffisamment sur l'entourage du triplet. Un fenêtre comparative de sept acides aminés, soit le triplets plus deux voisins au début et à la fin de ce dernier, sera l'idéale pour le critère de structures avoisinantes de triplets. Il s'agit maintenant de prendre triplet par triplet dans la banque de données et d'effectuer une mesure de RMS pour ensuite la compiler dans une catégorie particulière.

Comme la banque contient 184661 triplets avec leur environnement, un peu plus de 30 milliards de paires de structures seront évaluées, ce qui représente un bon échantillonage statistique. Lorsque les comparaisons seront accomplies, on procède à une vérification de la normalité de la distribution en prenant le couple d'environnement le plus disparate, le couple le plus similaire et un couple de ressemblance moyenne. Il suffit de d'effectuer une étude graphique de la distribution des structures en fonction d'intervalles finies de RMS. Si cette étude révèle une distribution normale, il serait possible de noter la matrice sous forme de cote Z pour des raisons de simplicité d'interprétation. La cote Z se décline comme suit:

$$Z = (\langle x \rangle - \sigma) / n \qquad (3.3)$$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0.073	1.028	0.086	1.368	0.166	0.783	-0.174	0.020	-0.259	-0.462	0.990	-0.459	-0.230	-0.454	0.440	0.411	-0.142	-0.485	0.134	-0.365	0.428	-0.159	0.148	-0.165	0.571
2	1.028	2.082	1.259	2.529	1.391	1.793	0.622	1.263	0.586	0.685	2.379	0.575	1.182	0.805	1.944	1.394	0.572	0.703	1.298	0.594	1.647	0.846	1.506	0.916	2.105
3	0.086	1.259	0.123	1.671	0.200	0.962	-0.264	-0.090	-0.374	-0.528	1.315	-0.478	-0.351	-0.453	0.433	0.472	-0.278	-0.662	-0.028	-0.397	0.506	-0.327	0.017	-0.329	0.590
4	1.367	2.529	1.671	2.839	1.872	2.157	1.011	1.733	0.918	1.274	2.877	0.892	1.850	1.021	2.533	1.780	1.152	1.358	1.696	1.061	2.183	1.337	2.104	1.323	2.579
5	0.165	1.391	0.200	1.872	0.281	1.081	-0.267	0.000	-0.340	-0.677	1.399	-0.670	-0.383	-0.518	0.375	0.564	-0.284	-0.568	0.103	-0.392	0.504	-0.303	-0.151	-0.266	0.624
6	0.782	1.793	0.961	2.157	1.081	1.502	0.538	1.023	0.471	0.541	2.017	0.421	0.962	0.514	1.598	1.128	0.544	0.506	1.042	0.359	1.447	0.768	1.382	0.774	1.766
7	-0.174	0.622	-0.264	1.011	-0.267	0.538	-1.251	-0.714	-1.221	-1.224	0.454	-0.871	-1.063	-0.764	-0.330	0.170	-1.074	-1.189	-0.721	-0.629	-0.207	-0.892	-0.657	-0.737	-0.185
8	0.019	1.263	-0.090	1.733	-0.001	1.023	-0.714	-0.534	-0.841	-0.801	1.302	-0.626	-1.056	-0.610	-0.114	0.462	-0.605	-1.129	-0.473	-0.562	0.150	-0.968	-0.701	-0.772	-0.147
9	-0.259	0.586	-0.374	0.919	-0.340	0.471	-1.221	-0.841	-1.411	-1.172	0.354	-0.942	-1.113	-0.910	-0.351	0.121	-0.972	-1.182	-0.803	-0.691	-0.304	-1.005	-0.757	-0.773	-0.383
10	-0.462	0.685	-0.528	1.274	-0.677	0.541	-1.224	-0.801	-1.172	-1.696	0.583	-1.221	-1.697	-1.347	-1.161	-0.031	-1.315	-1.742	-0.830	-1.053	-0.406	-1.243	-1.050	-1.199	-0.806
11	0.989	2.379	1.315	2.877	1.399	2.017	0.454	1.302	0.354	0.583	2.838	0.458	1.107	0.533	1.910	1.472	0.445	0.514	1.140	0.629	1.819	0.785	1.529	0.652	2.268
12	0.459	0.575	-0.478	0.892	-0.670	0.421	-0.871	-0.626	-0.942	-1.221	0.458	-0.932	-1.191	-1.527	-0.825	-0.071	-0.657	-1.450	-0.669	-0.858	-0.214	-0.920	-0.613	-1.249	-0.589
13	-0.230	1.182	-0.351	1.850	-0.383	0.962	-1.063	-1.056	-1.113	-1.697	1.107	-1.191	-2.008	-1.108	-0.914	0.300	-1.159	-1.888	-0.850	-0.807	-0.295	-1.474	-1.625	-1.125	-0.744
14	-0.454	0.805	-0.453	1.021	-0.518	0.514	-0.764	-0.610	-0.910	-1.347	0.533	-1.527	-1.109	-2.444	-1.031	-0.049	-0.446	-1.130	-0.636	-0.733	-0.171	-0.683	-0.974	-1.465	-0.791
15	0.440	1.944	0.433	2.533	0.375	1.598	-0.330	-0.114	-0.351	-1.161	1.910	-0.825	-0.914	-1.031	-0.411	0.859	-0.333	-0.831	-0.253	-0.220	0.550	-0.434	-0.958	-0.675	0.007
16	0.411	1.394	0.473	1.780	0.564	1.128	0.169	0.462	0.121	-0.031	1.472	-0.072	0.300	-0.049	0.859	0.712	0.149	-0.016	0.505	-0.150	0.884	0.273	0.653	0.246	1.027
17	-0.142	0.572	-0.278	1.152	-0.284	0.544	-1.074	-0.605	-0.972	-1.315	0.445	-0.657	-1.159	-0.446	-0.333	0.149	-1.389	-1.298	-0.649	-0.821	-0.278	-0.918	-0.603	-0.549	-0.194
18	-0.485	0.703	-0.662	1.358	-0.568	0.506	-1.189	-1.129	-1.182	-1.742	0.514	-1.450	-1.888	-1.130	-0.831	-0.016	-1.298	-1.724	-0.774	-1.144	-0.588	-1.407	-1.525	-0.980	-0.604
19	0.134	1.298	-0.028	1.696	0.103	1.042	-0.721	-0.473	-0.803	-0.830	1.140	-0.669	-0.850	-0.636	-0.253	0.505	-0.649	-0.774	-0.536	-0.554	0.183	-0.671	-0.619	-0.534	-0.282
20	-0.365	0.594	-0.397	1.061	-0.392	0.359	-0.629	-0.562	-0.691	-1.050	0.629	-0.858	-0.807	-0.733	-0.220	-0.150	-0.821	-1.144	-0.554	-1.114	-0.052	-0.762	-0.451	-0.665	-0.035
21	0.427	1.647	0.506	2.183	0.504	1.447	-0.207	0.150	-0.303	-0.406	1.819	-0.214	-0.295	-0.170	0.550	0.884	-0.278	-0.588	0.183	-0.052	0.707	-0.403	0.130	-0.195	0.597
22	-0.159	0.846	-0.327	1.337	-0.303	0.768	-0.892	-0.967	-1.005	-1.243	0.785	-0.920	-1.474	-0.683	-0.434	0.273	-0.918	-1.407	-0.671	-0.762	-0.403	-1.767	-1.238	-1.070	-0.901
23	0.148	1.506	0.017	2.104	-0.151	1.382	0.657	-0.701	-0.757	-1.050	1.529	-0.612	-1.625	-0.974	-0.958	0.653	-0.603	-1.525	-0.619	-0.451	0.130	-1.238	-0.983	-0.881	-1.157
24	-0.164	0.916	-0.329	1.323	-0.266	0.774	-0.737	-0.772	-0.773	-1.199	0.652	-1.249	-1.125	-1.465	-0.675	0.246	-0.549	-0.980	-0.534	-0.665	-0.195	-1.070	-0.881	-0.736	-0.986
25	0.571	2.105	0.590	2.579	0.624	1.766	-0.185	-0.147	-0.383	-0.806	2.268	-0.589	-0.744	-0.791	0.007	1.027	-0.194	-0.604	-0.282	-0.035	0.596	-0.901	-1.157	-0.986	-0.328

FIGURE 13: Matrice de similarité basée sur le RMS pour les structures avoisinantes.

où: <x> est la moyenne

σ est l'écart-type soit: $\sigma = ((x - \langle x \rangle)^2 / n^2)^{1/2}$

n est le nombre de compte

Après investigation, la distribution ne se voulait pas normale, la cote Z ne peut donc pas être utilisée pour noter la matrice. On aurait pu utiliser une autre méthode pour représenter cette matrice, mais il semble que les données brutes de RMS sont amplement suffisantes pour effectuer les simulations. Remarque importante à souligner, contrairement aux matrices de Dayhoff, un score élevé lors d'une comparaison représente une grande dissemblance entre les structures. On peut voir cette matrice de similarité structurelle à la figure 13. Quelques cas limites doivent être mentionnés (pour les codes de voisinages, se référer au tableau II). D'abord les couples de voisins les plus différents, vers un RMS de 1.77 Å, on retrouve les comparaisons 17-12 et 17-14 (soit des triplets ceinturés à 75% de feuillets avec des triplets entourés de fragments de feuillets et d'hélices). Dans la même lignée, vers un RMS de 1.71 Å, les comparaisons 18-8, 17-25, 19-5 et 19-8 (qui sont eux aussi des mélanges d'hélices et de feuillets). À l'autre extrême, les couples les plus semblables sont des couples mixtes dont les comparaisons sont faites entre eux. Ainsi, on retrouve les couples 14-14, 23-13 et 12-13 à des RMS entre 1.24 Å et 1.29 Å. Avec une matrice couvrant chacune des comparaisons de voisinage possible sur une plage de RMS allant de 1.24 Å à 1.78 Å, nous possédons un outil efficace afin d'orienter la construction des triplets selon un critère structurel.

3.5. Méthode employée pour déterminer les seuils optimaux

D'abord, le critère séquentiel sera déterminé et optimisé pour trois protéines différentes. Premièrement, la myoglobine (1mbo)[52], protéine dont Trip traite 146 résidus et possédant 8 hélices α comme structures secondaires. Deuxièmement, la CheY (3chy)[53], dont la structure traitable par Trip comporte 121 résidus répartis entre autres dans 10 structures secondaires (5 hélices α et 5 feuillets β). Finalement la glutarédoxine (1aba)[54], protéine possédant une structure modélisable de 86 résidus répartis partiellement parmi 7

structures secondaires (3 hélices α et 4 feuillets β). Avec ce choix de protéines, on couvre trois grandeurs structurelles et diverses combinaisons de feuillets et d'hélices. À noter, l'absence d'une protéine possédant exclusivement des feuillets β en tant que structure secondaire. La raison est simple, Trip possède un module afin de bien gérer la présence d'hélices α mais pas les feuillets β.

Le critère séquentiel sera déterminé en lançant diverses simulations avec des seuils d'acceptation différents. Les seuils d'acceptation serviront à effectuer une sélection parmi les géométries de triplets naturels à considérer dans la nouvelle distribution biologique de Trip. Une fois que ce seuil sera déterminé et optimisé pour chacune des trois protéines, le critère structurel sera à son tour optimisé. Préalablement aux lancements de simulations, il est essentiel de regarder attentivement les différentes distributions géométriques offertes par Trip et par la banque de triplets naturels. On souhaite découvrir que la distribution biologique des triplets soit en fait un sous-ensemble de la distribution des géométries offertes par Trip. Pour ce faire, on doit comparer les graphiques de chacune des coordonnées internes représentant le nombre de d'occurences de la dite coordonnée en fonction de l'espace conformationel des coordonnées internes. Des graphiques ont été produits pour la distribution biologique et la distribution générée par Trip.

Avant de commencer l'analyse des graphiques présentés aux figures 14 à 23, un détail se doit d'être mentionné à propos de ce qu'on appelle l'espace conformationel des coordonnées internes. Il s'agit du nombre de division d'un certain intervalle d'angles. Pour les coordonnées internes q1 et q2, cet espace va de 0 à π (on couvre 180°). Pour les coordonnées internes q3, q4 et q5, cet espace va de $-\pi$ à π (on couvre 360°). Pour l'échelle des degrés des coordonnées internes q1 et q2, on constate sur le graphique qu'ils vont de π à 0, mais que ces valeurs d'angles sont toujours positives. Les définitions de q1 et de q2 sont des produits de deux vecteurs. Ces deux vecteurs sont en fait le cosinus d'un angle. Ainsi, pour les besoins de ces graphiques, on a utilisé la fonction arccosinus de ce produit vectoriel afin de trouver une valeur d'angle. Comme le domaine d'une fonction arccosinus est compri entre 0 et π inclusivement, les valeurs d'angles ne peuvent jamais être négatives.

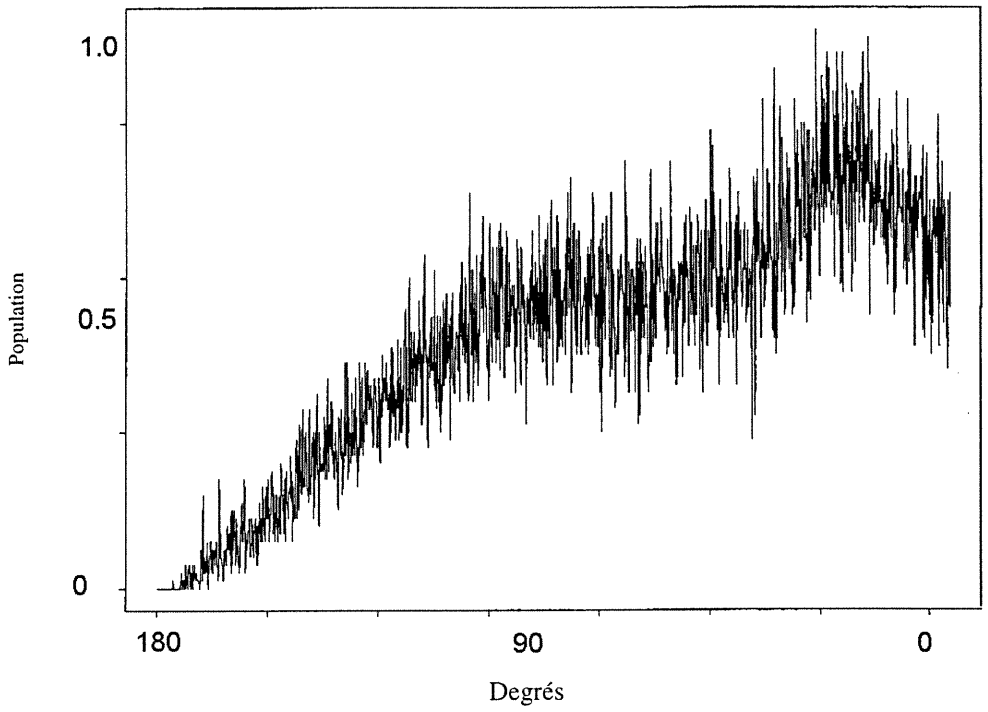


FIGURE 14. Distribution de la coordonnée interne q1 générée par Trip.

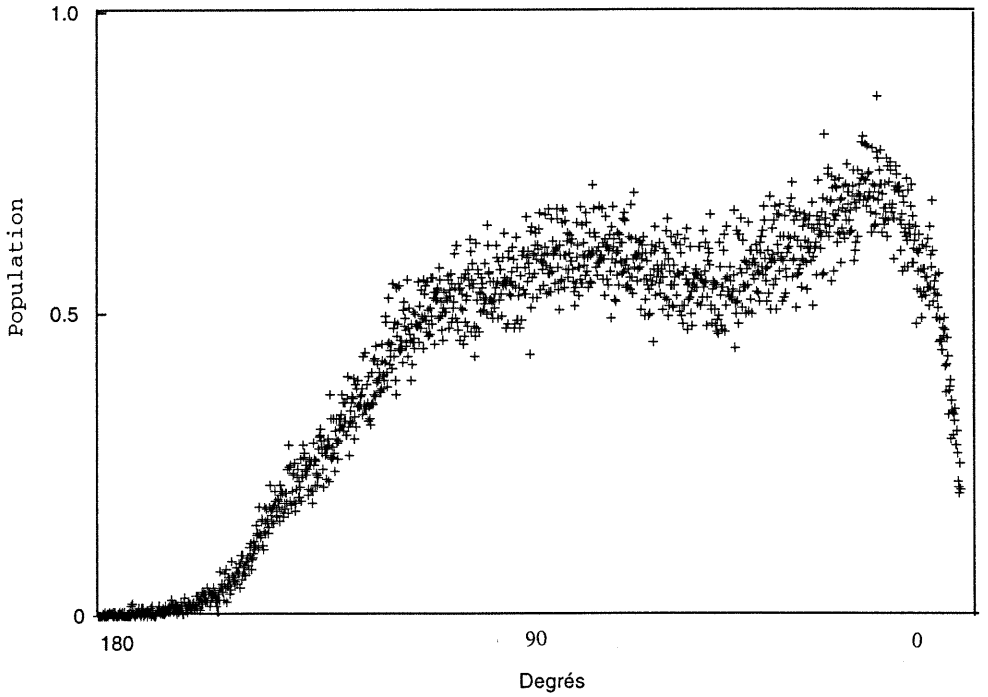


FIGURE 15. Distribution de la coordonnée interne q1 naturelle.

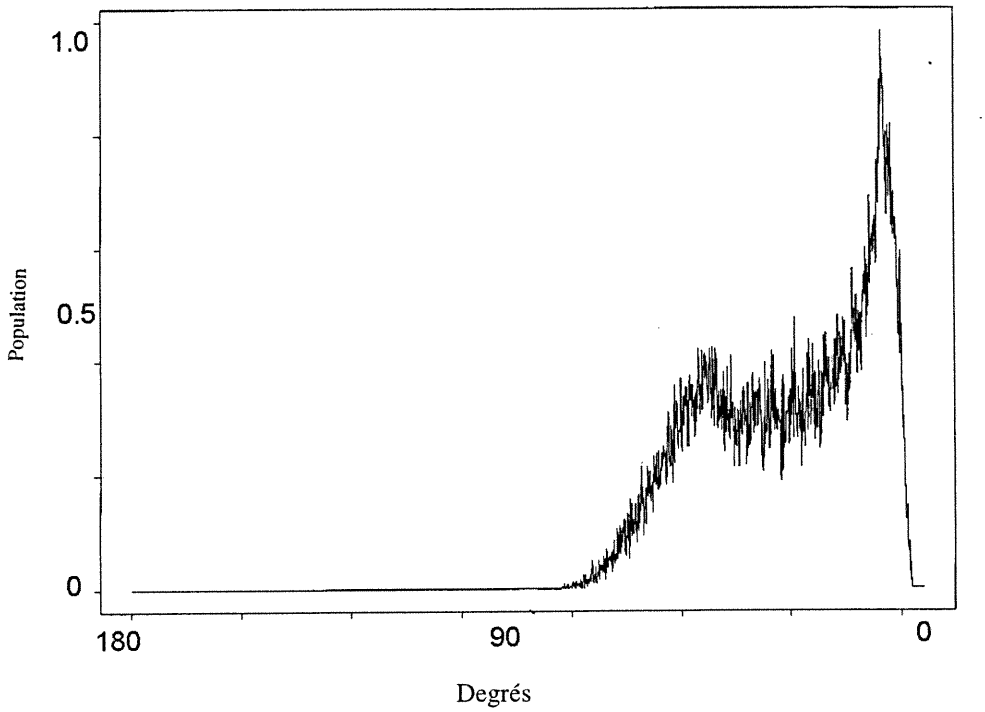


FIGURE 16. Distribution de la coordonnée interne q2 générée par Trip.

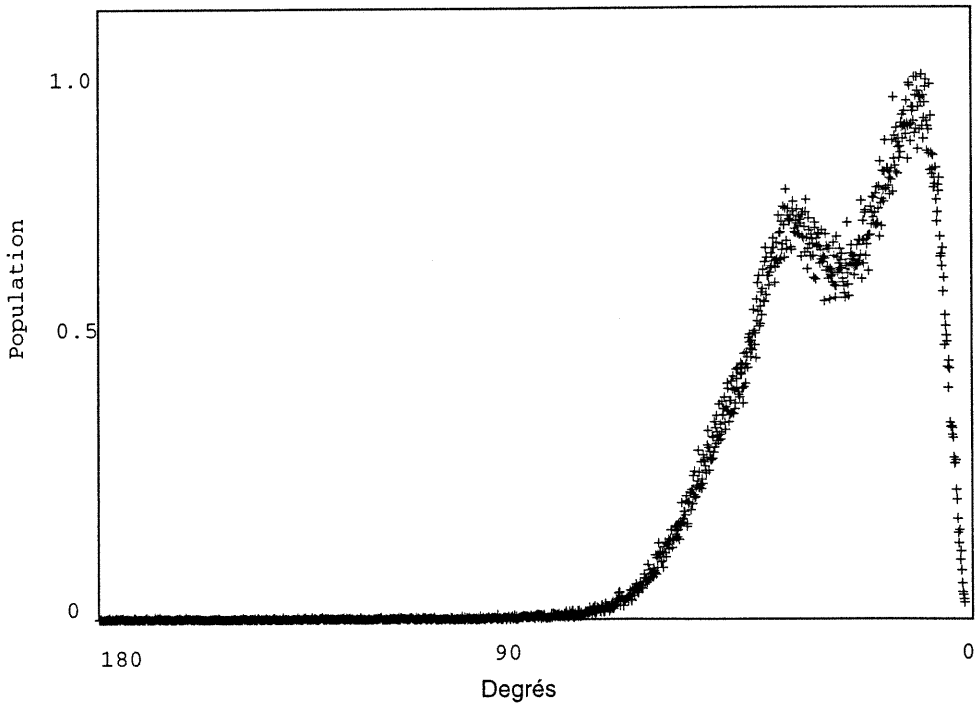


FIGURE 17. Distribution de la coordonnée interne q2 naturelle.

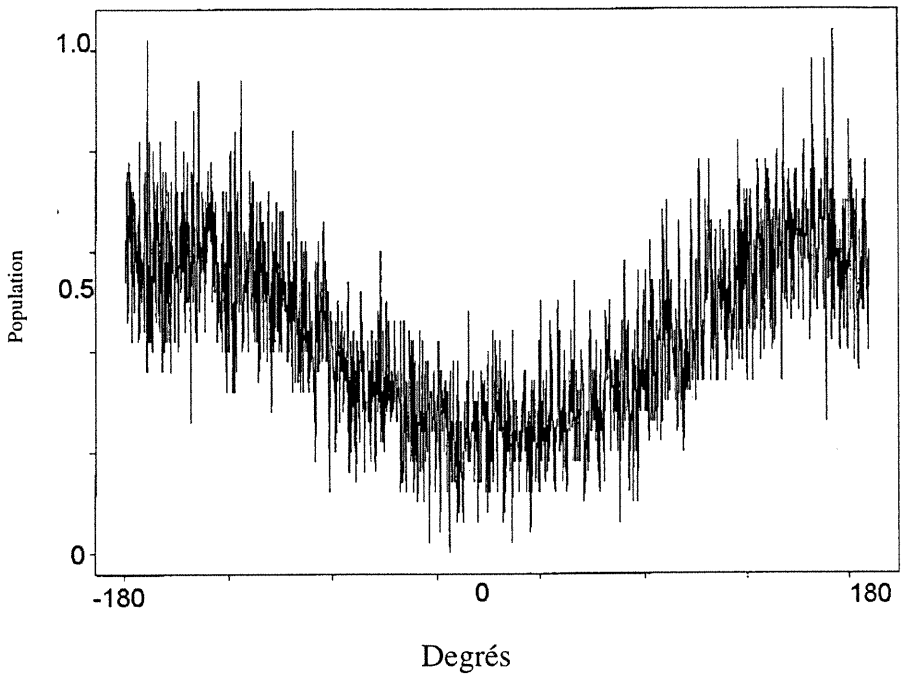


FIGURE 18. Distribution de la coordonnée interne q3 générée par Trip.

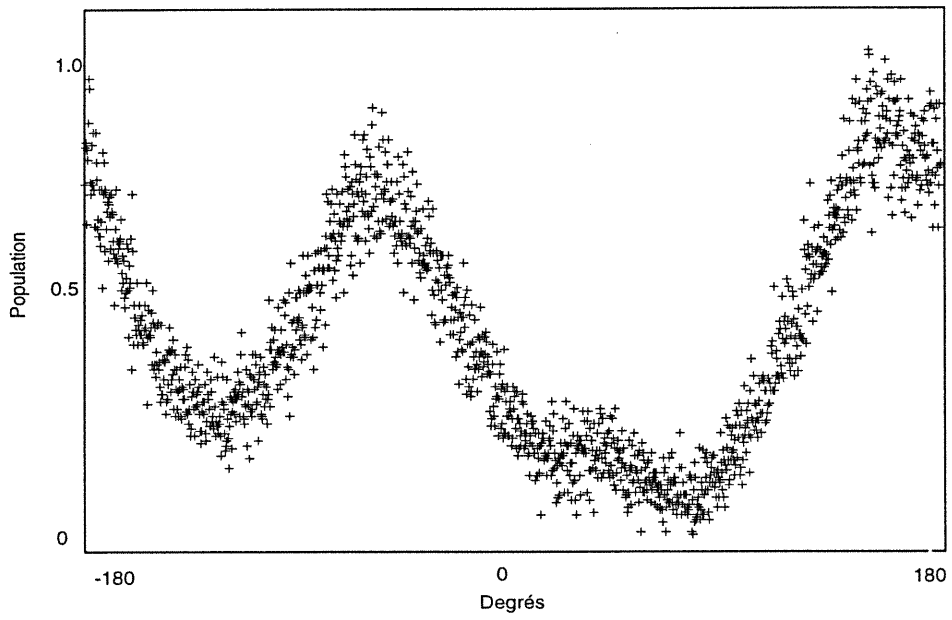


FIGURE 19. Distribution de la coordonnée interne q3 naturelle.

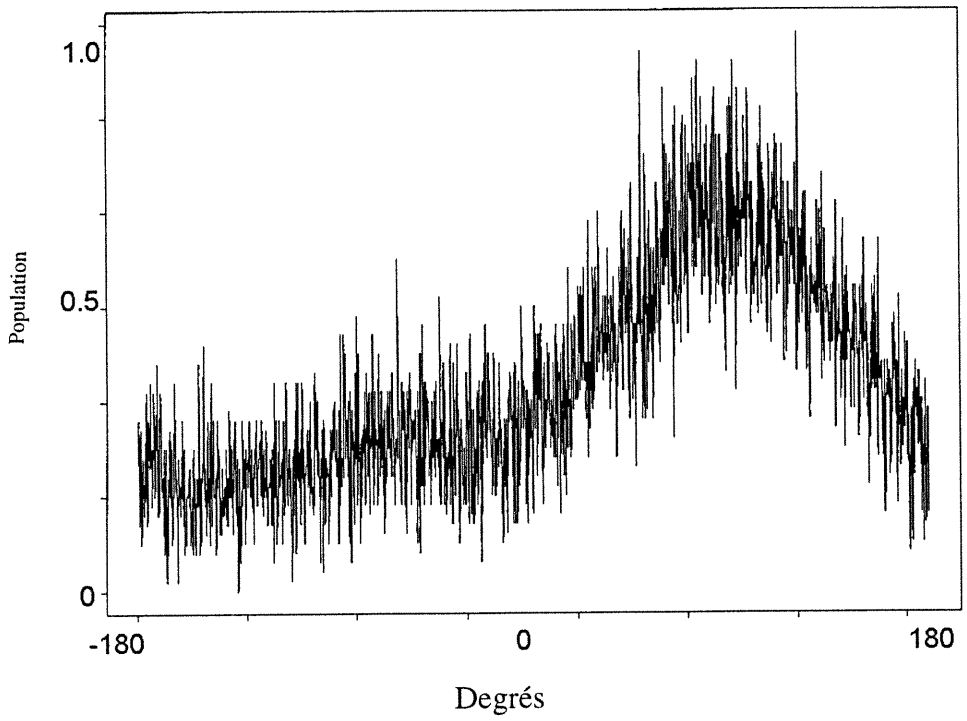


FIGURE 20. Distribution de la coordonnée interne q4 générée par Trip.

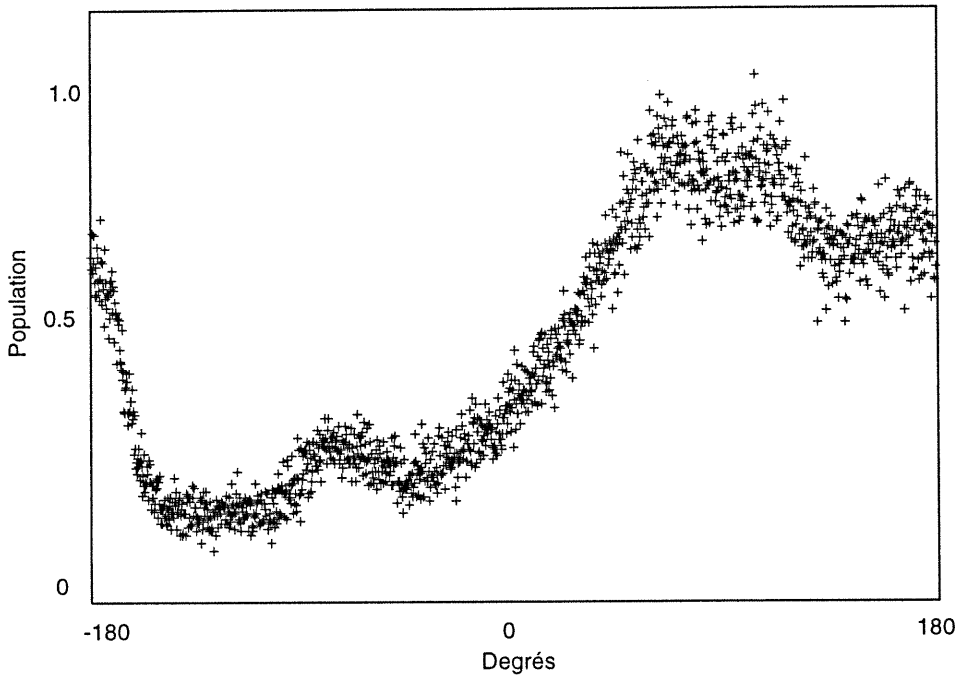


FIGURE 21. Distribution de la coordonnée interne q4 naturelle.

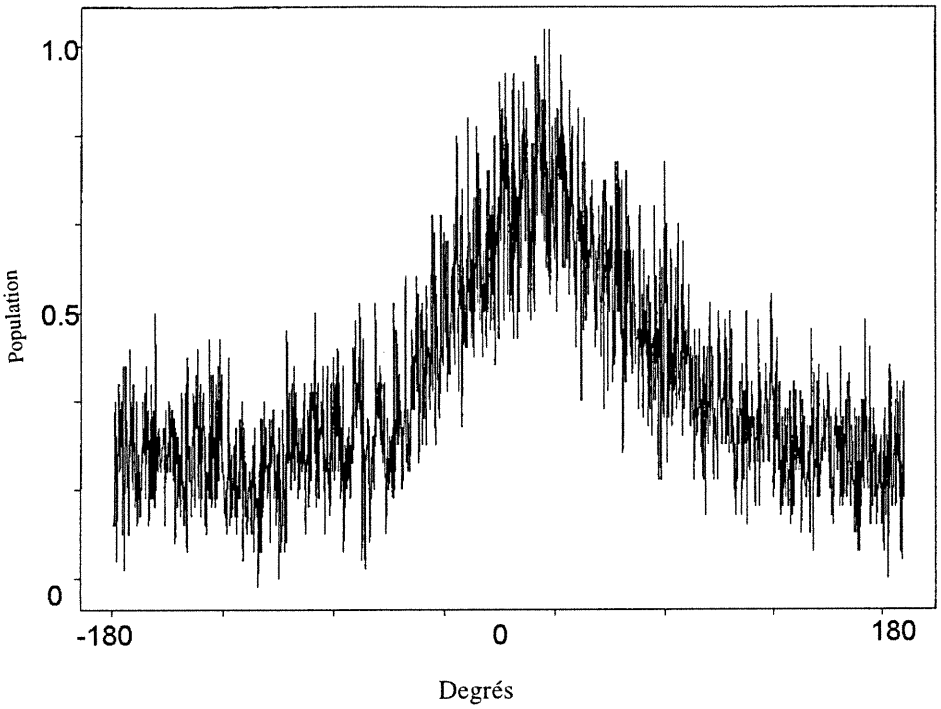


FIGURE 22. Distribution de la coordonnée interne q5 générée par Trip.

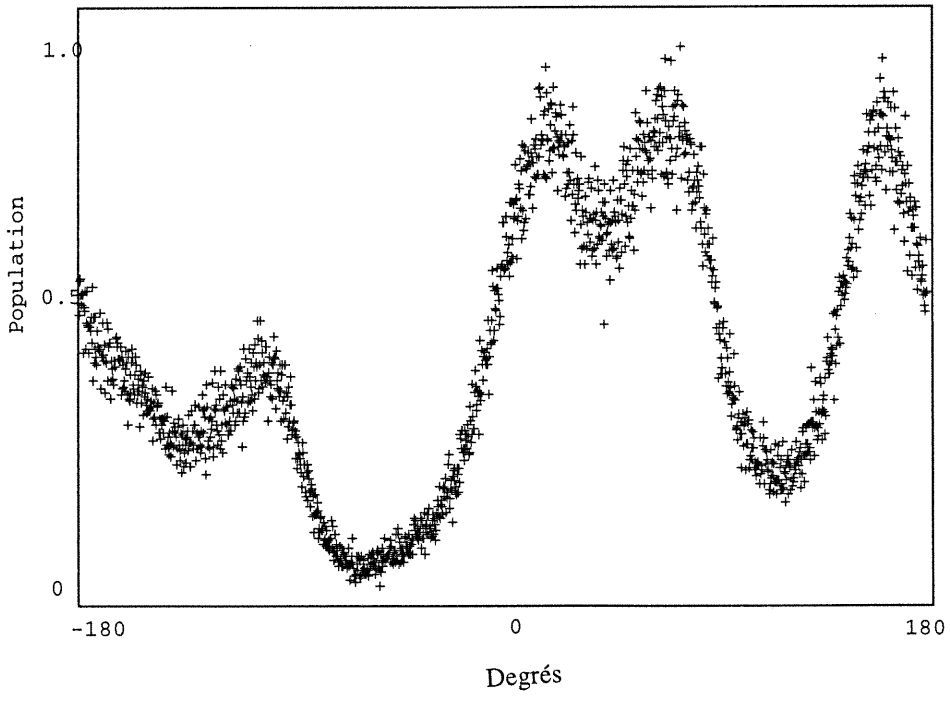


FIGURE 23. Distribution de la coordonnée interne q5 naturelle.

D'abord pour les coordonnées q1, les deux graphiques (figures 14 et 15) ont sensiblement la même allure aux mêmes endroits. Les deux seules différences notables se situent de -180° à -155° et de 5° à 0°. Dans ces deux régions, on retrouve moins cette configuration de q1 dans la distribution biologique comparativement à la distribution générée par Trip.

Pour les coordonnées q2, on constate d'abord une forme un peu inusitée (voir les figures 16 et 17). Cette forme découle de la définition de la coordonnée q2 (voir les figures 7 et 8 au chapitre 2) qui ne permet pas aux coordonnées de se situer au delà d'environ 90°. Les deux distributions de q2 se ressemblent, à l'exception du fait que la montée en occurences s'effectue plus loin dans la répartition des angles. Fait notable lorsque l'on compare les distributions de q2 entre elles, la région allant de 60° à 0° contient plus de conformations dans le cas de la distribution biologique. De ce fait, on constate que Trip ne donnait pas suffisamment d'importance en poids à cette région. La principale différence entre la distribution de la coordonnée q1 et celle de la coodonnée q2 réside en l'appartenance de l'azote précédent le dernier carbone α du triplet. L'azote précédent le premier carbone α n'appartient pas à ce dernier selon la définition mathématique de q1, ainsi ce carbone peut se positioner sur à peu près n'importe quel choix d'angles. L'azote du dernier carbone α fait partie intégrante du triplet selon la définition mathématique donnée. Ainsi, une zone lui est totalement interdite puisque ça soutenderait qu'il reviendrait sur ses pas pour joindre son carbone α correspondant.

Viennent ensuite les distributions de la coordonnée q3 (figures 18 et 19). C'est à ce moment que l'on commence à entrevoir l'importance de l'apport d'informations biologiques. Dans le distribution générée par Trip, on observe une forme "U" évasée alors que la distribution naturelle arbore une double forme de "U". Dans la région d'angles allant de -155° à -125° ainsi que de 5° à 105°, Trip semble générer des coordonnées q3 qui n'ont pas d'origine naturelle. On abaisse ici le nombre d'informations à traiter plus tard à l'interne du logiciel. D'autant plus que dans la distribution naturelle, on observe autour de -67.5°, une protubérance indiquant l'importance de cette configuration.

Les distributions des coordonnées q4 (figures 20 et 21) quant à elles se ressemblent à quelques détails près. L'allure reste sensiblement la même à l'exception du début et de la fin de la répartition des angles. Dans la distribution biologique, les régions allant de -180° à -155° et de 120° à 180° sont peuplées plus significativement. Signe encore ici que la distribution des géométries non-naturelles demeure insuffisante, la distribution biologique apporte donc un bonus à ces régions.

Finalement, la distribution des coordonnées q5 (voir les figures 22 et 23) est celle qui réserve le plus de surprises. Du côté non-biologique, on voit simplement une pointe parmi le nuage de points situés entre 85° et 135°. En ce qui concerne la distribution biologique, on observe beaucoup de variation du début à la fin. Certains endroits pointent des manques de la part de Trip (par exemple autour de -90°, de 30° et de 160° et d'autres endroits montrent que Trip n'apporte que des informations superflues et redondantes inutilement (comme vers -45° et 135°).

À la vue de ces graphiques, il devient notable que l'apport d'informations biologiques améliorera probablement les structures de triplets générées par Trip. C'est le poids de la distribution qui changera. En d'autres mots, cette distribution orientera Trip différemment dans son processus de construction de triplets. Ces triplets devraient plus tard améliorer le processus de construction des boucles.

3.6. Optimisation des paramètres d'inclusion

Compte tenu de la pléïade de variables avec lesquelles Trip doit jongler, certaines peuvent être optimisées. C'est le cas du paramètre b dans l'espace conformationel par degré de liberté donnant au total b^5 boîtes de classement. La grandeur de chaque division est de 360°/b (pour les angles dièdres), et 2/b pour les cosinus d'angles polaires. C'est aussi le cas du paramètre qui traduit le nombre de triplets maximal à traiter dans chaque liste de triplets durant une simulation.[55]

Pour déterminer le nombre optimal de boîtes de classement à traiter, on procède simplement à l'élaboration de graphiques illustrant le comportement du poids de la distribution biologique versus le nombre de boîtes de classement et de graphiques illustrant le comportement du poids de la distribution statistique de Trip en fonction du nombre de boîtes de classement. En traçant ces graphiques selon les différentes grandeurs du paramètre b, on peut étudier la quantité de pics se pointant au delà du bruit statistique.

À un extrême, un nombre de boîtes trop petit ne permet pas suffisamment à Trip de bénéficier de l'apport d'une distribution biologique (voir les figures 24 et 25). À cet extrême, certe une petite différence se fera ressentir, ne serait-ce que de comparer l'amplitude des pics d'une représentation par rapport à l'autre. On constate à nouveau que la distribution biologique est en fait un sous-ensemble de la distribution statistique de Trip. De surcroît, on remarque une redondance inutile dans cette dernière. Toutefois, les boîtes étant trop grandes, il est difficile de rendre justice de façon minutieuse au poids de la distribution biologique.

À l'autre extrême, si on essaie de gérer un trop grand nombre de boîtes, ces dernières seront très peu peuplées puisque la même distribution se répartie sur plus de divisions (voir les figures 26 et 27). Il en résulte un grand bruit statistique dans lequel sont noyés les foyers de populations importants de la distribution statistique de Trip. Dans le cas de la distribution biologique, on se retrouve avec un bruit statistique plus faible, mais à l'inverse de la distribution non-naturelle, la quantité de pics se démarquant du bruit de fond augmente avec le nombre de classeurs. C'est ainsi que l'on constate l'importance d'optimiser le paramètre b afin de maximiser l'influence du poids de la distribution biologique.

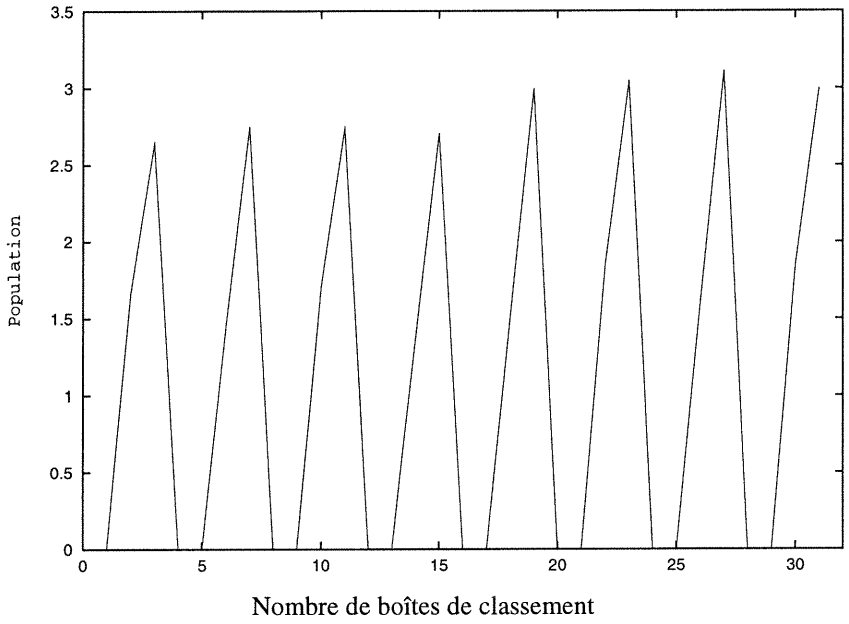


FIGURE 24: Distribution des coordonnées internes générées par Trip selon un système de 32 boîtes de classement.

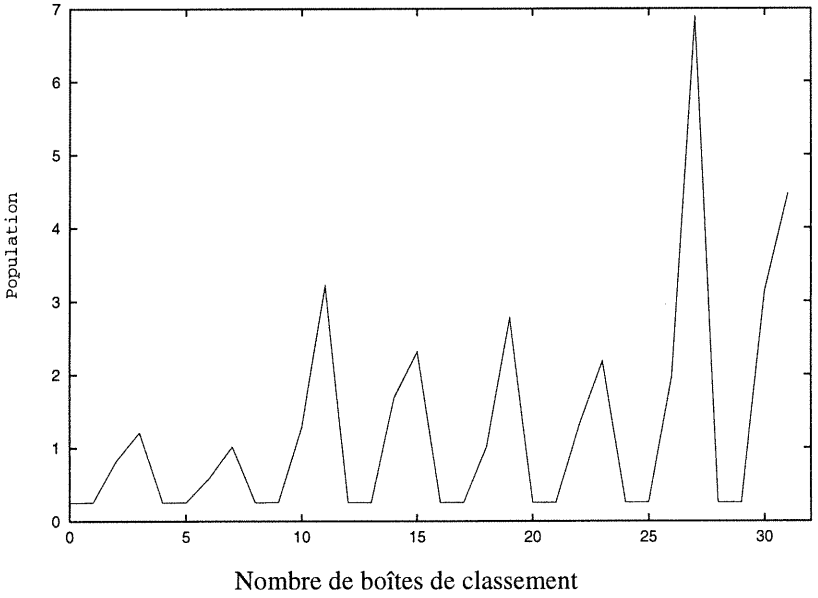


FIGURE 25: Distribution des coordonnées internes naturelles selon un système de 32 boîtes de classement.

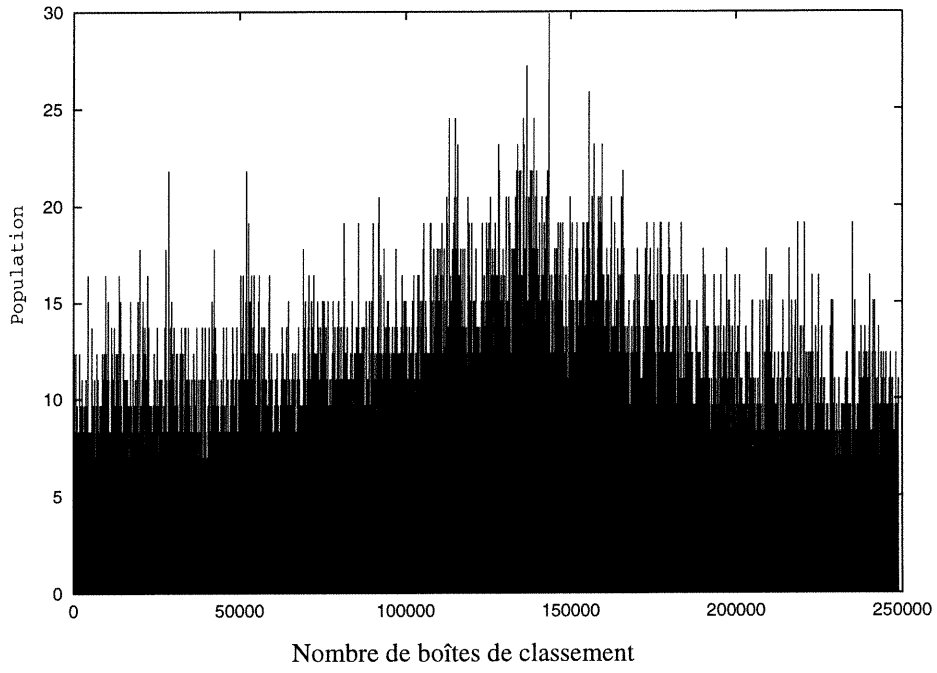


FIGURE 26: Distribution des coordonnées internes générées par Trip selon un système de 248832 boîtes de classement.

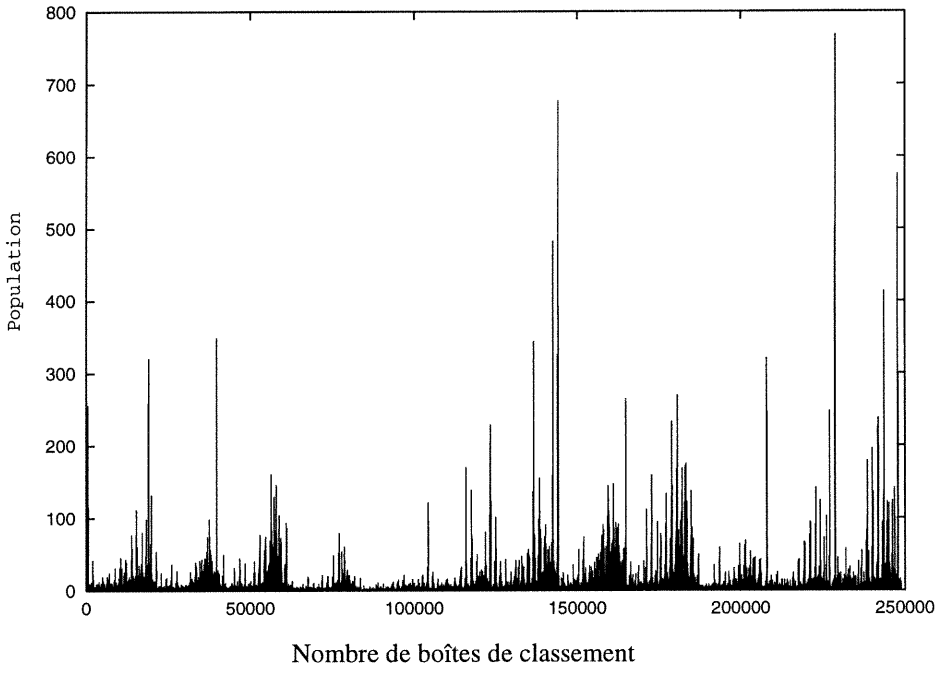


FIGURE 27: Distribution des coordonnées internes naturelles selon un système de 248832 boîtes de classement.

On peut donc dès maintenant, mettre en évidence ces constatations sous forme graphique. D'abord on pointe l'attention vers le rapport des boîtes vides de la distribution statistique de Trip par rapport à la distribution biologique, et ce en fonction du paramètre b (voir figure 28). Il en découle une ascension suivi d'un plateau débutant vers une valeur de b égalant 11. Lorsque le paramètre b égale 7, le rapport du nombre de boîtes de classement vides de la distribution biologique par rapport au nombre de boîtes vides de la distribution non-naturelle est de un pour un. À cette valeur, une distribution n'est pas désavantagée par rapport à l'autre. Présentement, Trip fonctionne avec une valeur de b égale à 4. En conservant cette valeur la distribution biologique se trouve désavantagée par rapport à la distribution statistique puisque l'information contenue dans le peu de boîtes de classement n'est pas suffisamment bien répartie. En d'autres mots, on comprime inutilement l'information biologique. Si au contraire, on fait appel à un b plus grand que 7, on se retrouve avec des boîtes vides inutilement, on ne possède pas la spécificité nécessaire afin de les peupler.

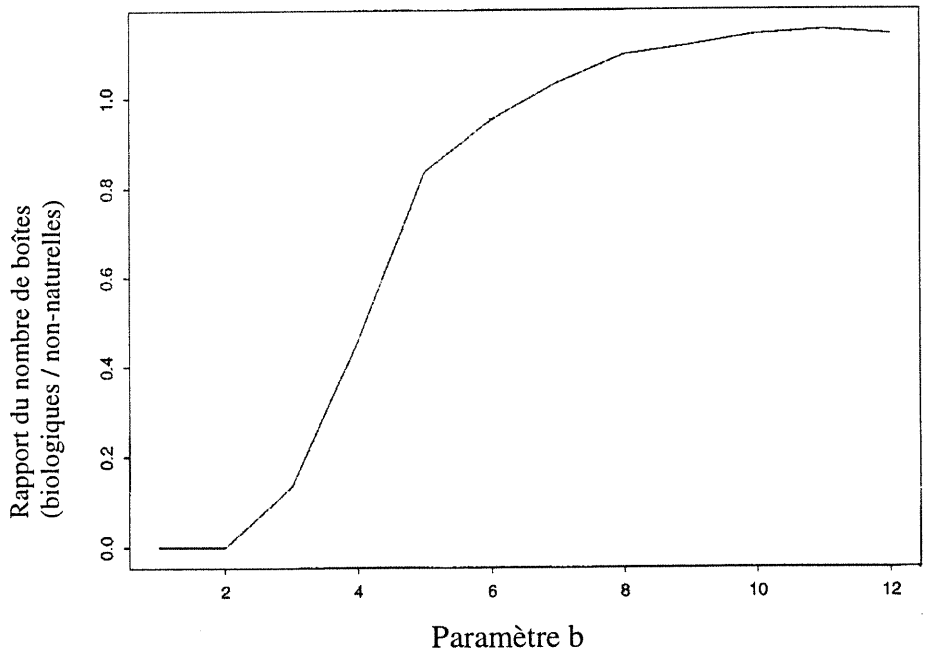


FIGURE 28. Fraction des boîtes de classement de triplets peuplées en fonction du nombre de divisions pour le paramètre b.

Deuxième fait intéressant à porter en graphique, le nombre de pics dans la distribution biologique qui sont significativement plus élevés que la moyenne des pics non-nuls dans la distribution statistique de Trip par rapport au nombre de boîtes significatives en fonction du paramètre b (voir la figure 29 ci-dessous).

Une courbe généralement descendante depuis le départ jusqu'à la fin du graphique se perçoit. En mots, la signification de ce comportement démontre que plus le nombre de boîtes est élevé, moins les boîtes de classement significativement peuplées de la distribution biologique se font ressentir. Au delà d'une valeur de b égale à 7, on dilue beaucoup trop la population des conformations. La conjugaison de ces deux graphiques confirme qu'une valeur de b égale à 7 semble optimale afin de maximiser la contribution d'une distribution biologique.

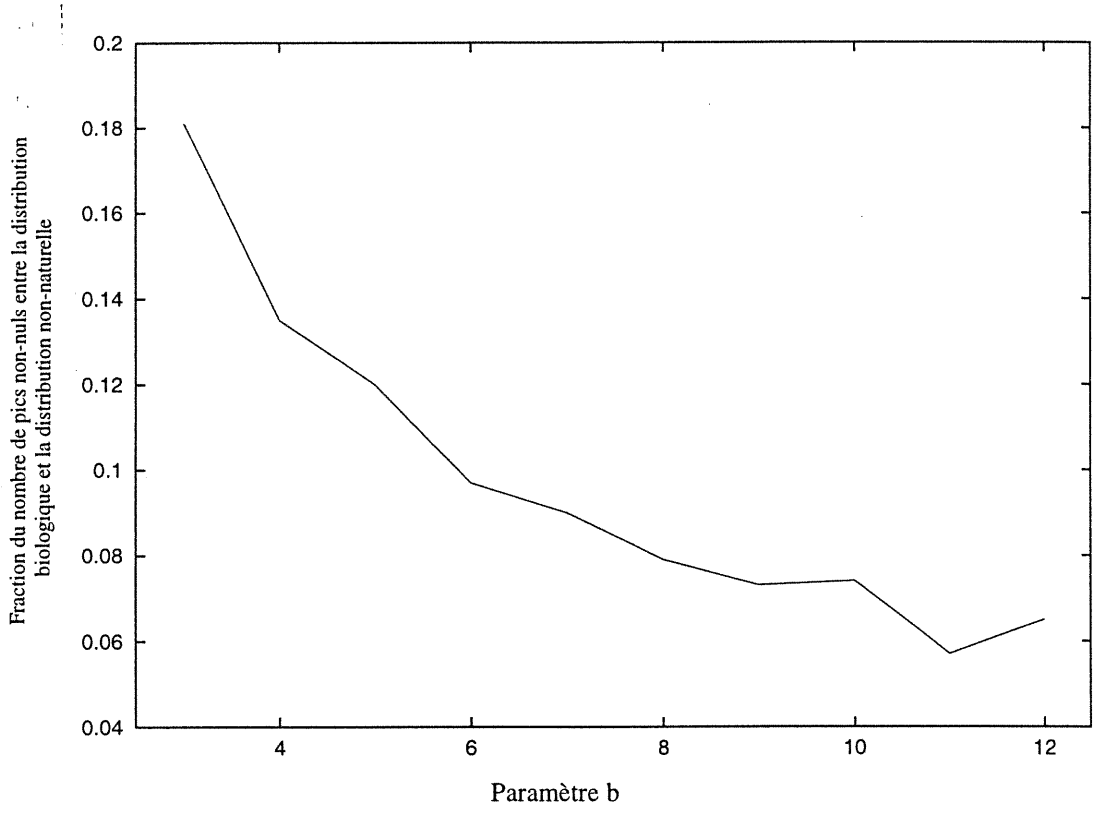


FIGURE 29. Quantité des boîtes de classement peuplées significativement dans la distribution biologique versus les boîtes de rangement de population moyenne de la distribution statistique non-naturelle versus le paramètre b.

Pour l'optimisation du nombre de triplets par listes à distribuer dans les boîtes, il faut comprendre que si on a plus de boîtes, il est normale d'utiliser plus d'éléments pour les remplir à leur juste valeur. Des boîtes mal remplies ne représentent pas nécessairement bien le poids de la nouvelle distribution. Les test à effectuer consiste à vérifier si le nombre de triplets est suffisant compte tenu du nombre de boîtes de classement à remplir. Le graphique suivant illustre le résultat de ce test (voir la figure 30).

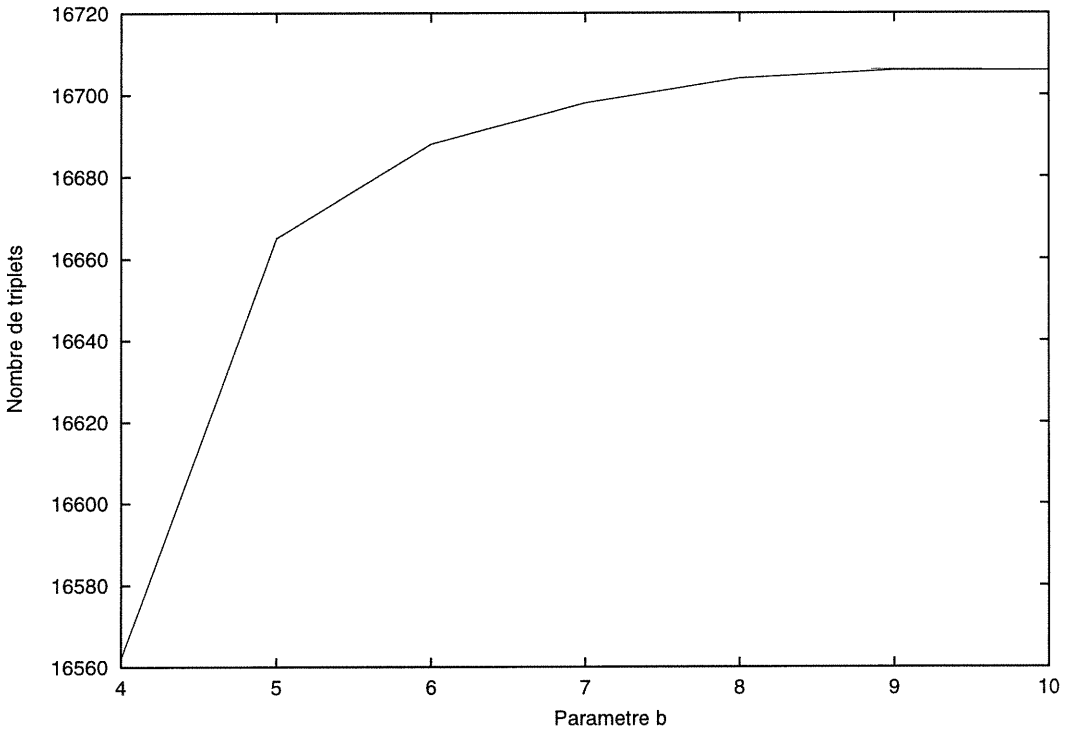


FIGURE 30. Population de triplets nécessaire afin de combler équitablement la quantité de boîtes (b^5)de classement disponibles.

Un plafond est rapidement atteint pour une valeur de 16700 triplets. Évidemment, on peut utiliser un plus grand nombre de triplets mais cela ne peut qu'être nuisible sur les performances de calculs puisque plus d'éléments doivent être gérés par le logiciel sans toutefois apporter de l'information supplémentaire utile. À partir de ce point, on peut commencer à lancer des simulations.

3.7. Procédure pour l'accumulation de résultats

Afin d'optimiser l'utilisation de la distribution naturelle de triplets, on a recourt à l'utilisation de seuils d'acceptation. Les seuils d'acceptation se définissent dans le cas présent comme une barrière. Grâce au scores calculés pour chaque triplet à l'aide de la matrice de Dayhoff pour l'homologie de séquence ou à l'aide de la matrice de RMS pour les structures avoisinantes, on peut décider d'inclure la géométrie de triplet correspondante dans la nouvelle distribution biologique ou non. Pour les scores selon les matrices de Dayhoff, il s'agit de mesures de similitude, donc on accepte le triplet si son score est au delà de la barrière fixée. Pour les scores selon la matrice de RMS, il s'agit d'une mesure de dissemblance, donc on accepte le triplet si son score est en-deçà de la barrière fixée. Ces seuils d'acceptation sont fixés au préalable de chaque simulation.

Il s'agit d'abord d'optimiser le premier seuil d'acceptation (celui pour l'homologie de séquence) pour chacune des trois protéines test. Ensuite, à partir de ces résultats optimisés, on procédera à l'optimisation du deuxième seuil, soit celui qui dépend de l'homologie structures avoisinantes. Les seuils d'acceptation optimaux seront probablement différents d'une protéine à l'autre. Chacune des simulations pour les différents seuils d'acceptation sera lancée 3 fois afin d'apporter de la reproductibilité aux résultats obtenus.

Avant de procéder aux simulations avec la nouvelle distribution biologique, des calculs nominaux ont été lancés pour chacune des protéines test. C'est à partir de ces résultats que l'on pourra constater s'il y a amélioration au niveau énergétique et au niveau du RMS. Bien que les valeurs énergétiques des conformations soient calculées par l'intermédiaire du potentiel de Sippl, le nombre d'approximations utilisées reste considérable. Ainsi les valeurs d'énergie possèdent des unités arbitraires. La comparaison de ces valeurs énergétiques avec les valeurs physiques d'énergie n'auraient aucun sens. Ainsi donc, la variation entre les valeurs d'énergie des calculs nominaux et les valeurs d'énergie des

calculs avec la nouvelle distribution seront à prendre en considération. On peut voir ces valeurs d'énergie et de RMS au tableau III.

TABLEAU III: Énergies et RMS obtenus à partir de calculs nominaux pour les trois protéines test.

Protéine	Meilleure énergie	Meilleure énergie moyenne	Meilleur RMS (Å)	Meilleur RMS moyen (Å)
3chy	-118.82	-73.57	9.78	13.52
1mbo	-25.66	8.45	8.59	12.37
1aba	-92.99	-69.06	8.33	11.35

Voyons maintenant les résultats obtenus pour les 3 protéines avec différents seuils d'homologie de séquence. Le tableau IV témoigne des meilleures énergies, les meilleurs RMS, les meilleures énergies moyennes et les meilleurs RMS moyens pour chacune des protéines en fonction de seuil d'acceptation d'homologie de séquence.

TABLEAU IV. Énergies et RMS optimaux obtenus pour diverses simulations sur trois protéines test.

Protéine	Meilleure énergie	Meilleure énergie moyenne	Meilleur RMS (Å)	Meilleur RMS moyen (Å)
3chy	-140.54	-101.66	7.47	10.71
1mbo	-34.98	0.84	6.41	11.78
1aba	-114.76	-80.90	6.30	8.64

Les résultats obtenus ne sont pas tous excellents. Une spécification à apporter sur les résultats inscrits dans le tableau IV: toutes les énergies sont notées par rapport à une structure native ayant 0 comme énergie. Il en va de même avec les mesures de RMS. Au niveau énergétique, la 3chy donne d'excellents résultats avec des structures générées par Trip dont l'énergie moyenne tourne autour de -101.66. De plus, il faut noter la présence

d'une structure ayant -140.54 unités d'énergie. Des résultats semblables sont obtenus pour la 1aba, elle qui comporte à la fois des feuillets et des hélices comme sa consoeur. La bête noire au niveau énergétique est la 1mbo. Malgré quelques structures dont l'énergie est négative, la plupart des structures générées par Trip n'arrivent pas à l'énergie de la structure nominale malgré l'absence de feuillets β. Ces résultats sont probablement la conséquence de la présence de trois boucles de trois résidus. La distribution biologique des triplets ne donne peut-être pas le poids voulu sur une conformation qui aide Trip à l'interne. Les coordonnées internes acceptées témoignent de cette hypothèse. Ainsi les graphiques suivants correspondent au coordonnées internes des triplets qui ont été acceptées par un certain seuil d'homologie. Dans le cas suivant, il s'agit des coordonnées extirpées à partir de la 3chy selon un seuil d'acceptation de 0.50. On peut observer un comportement similaire pour les autres protéines.

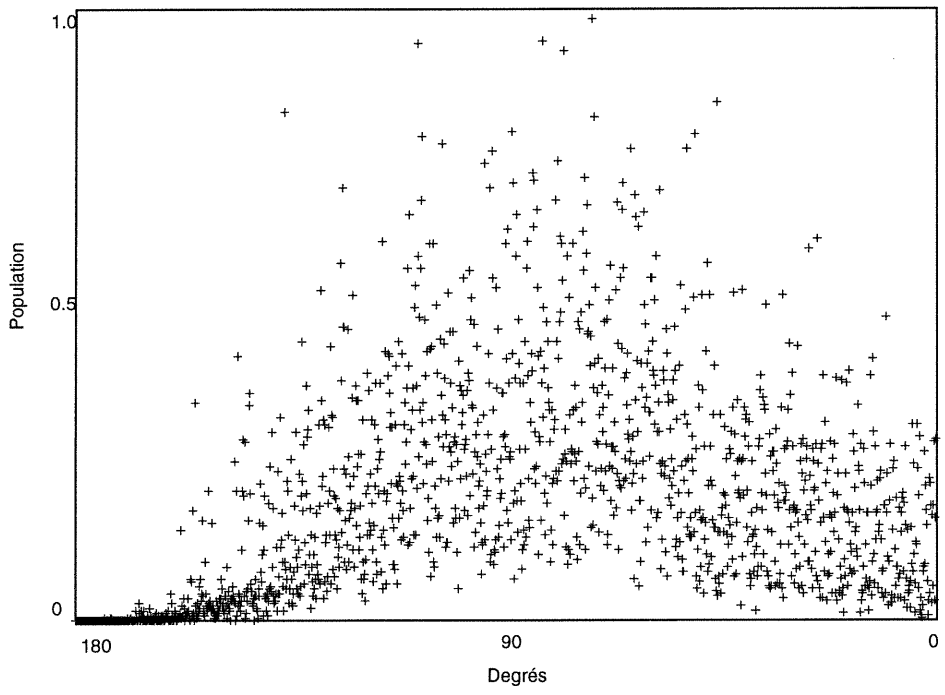


FIGURE 31. Distribution de la coordonnée interne q1 des triplets acceptés.

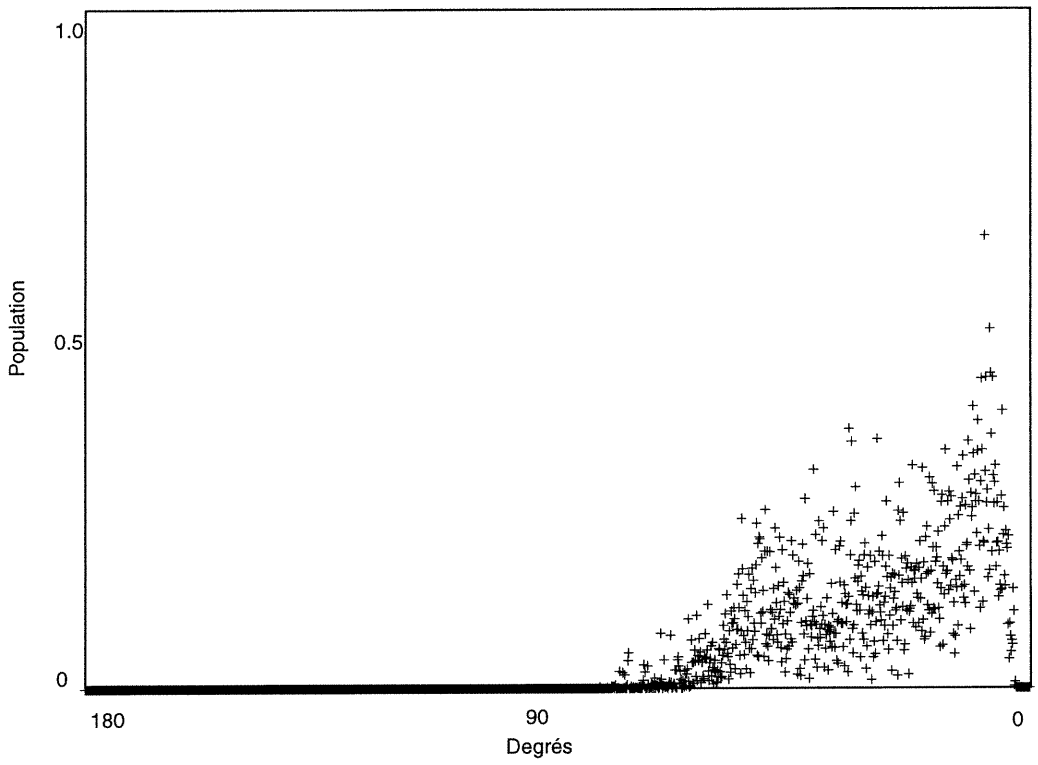


FIGURE 32. Distribution de la coordonnée interne q2 des triplets acceptés.

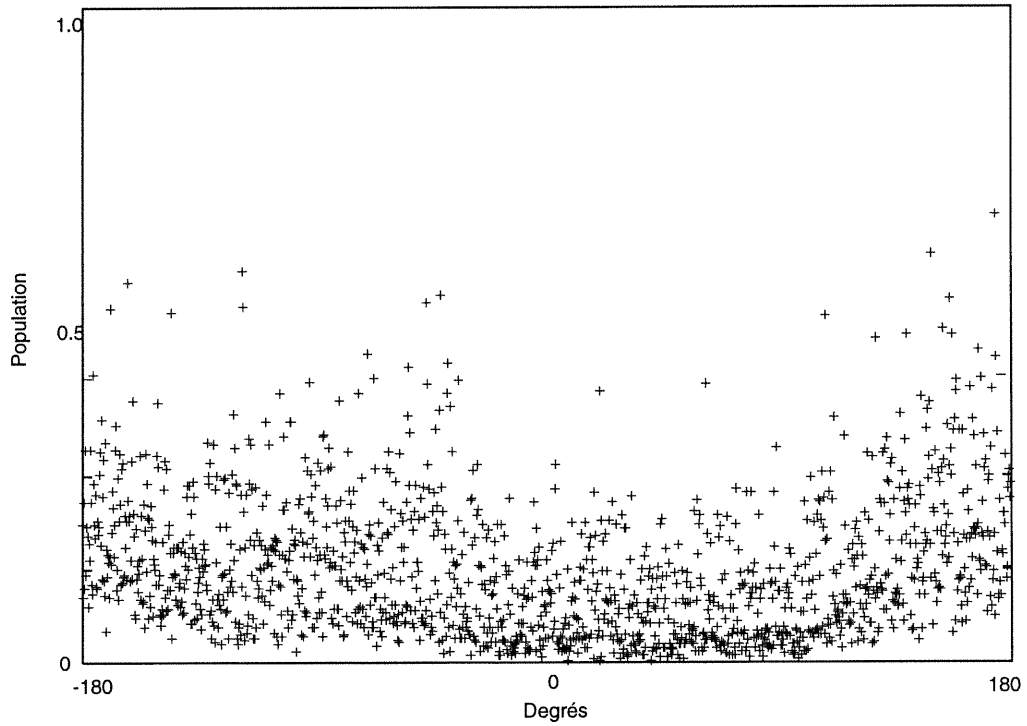


FIGURE 33. Distribution de la coordonnée interne q3 des triplets acceptés.

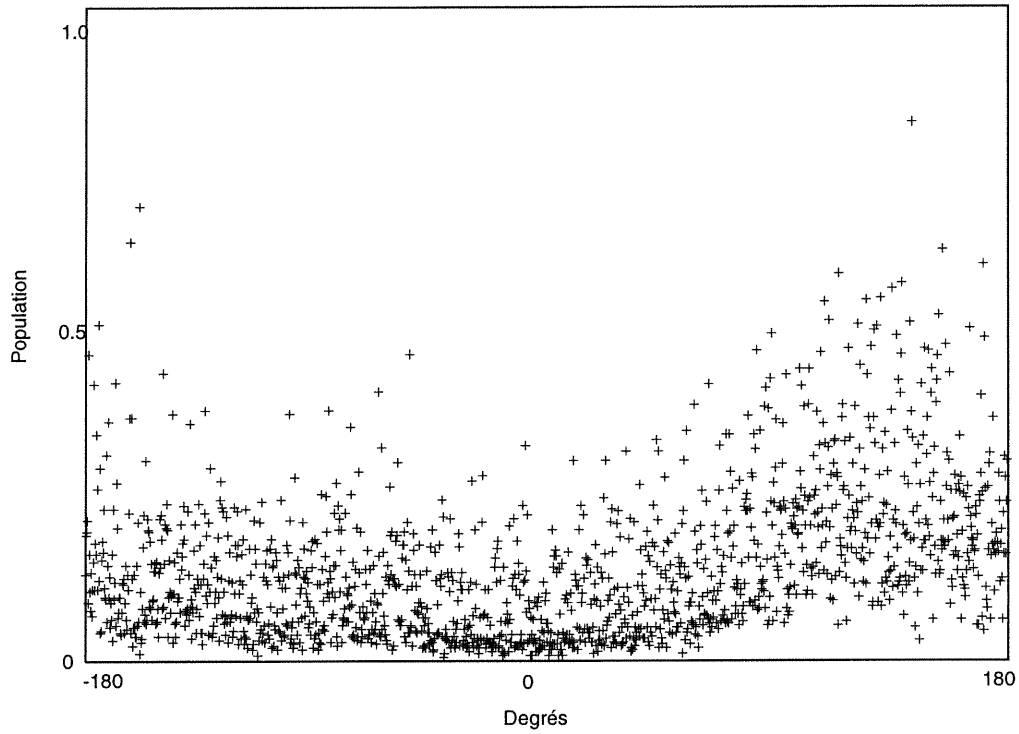


FIGURE 34. Distribution de la coordonnée interne q4 des triplets acceptés.

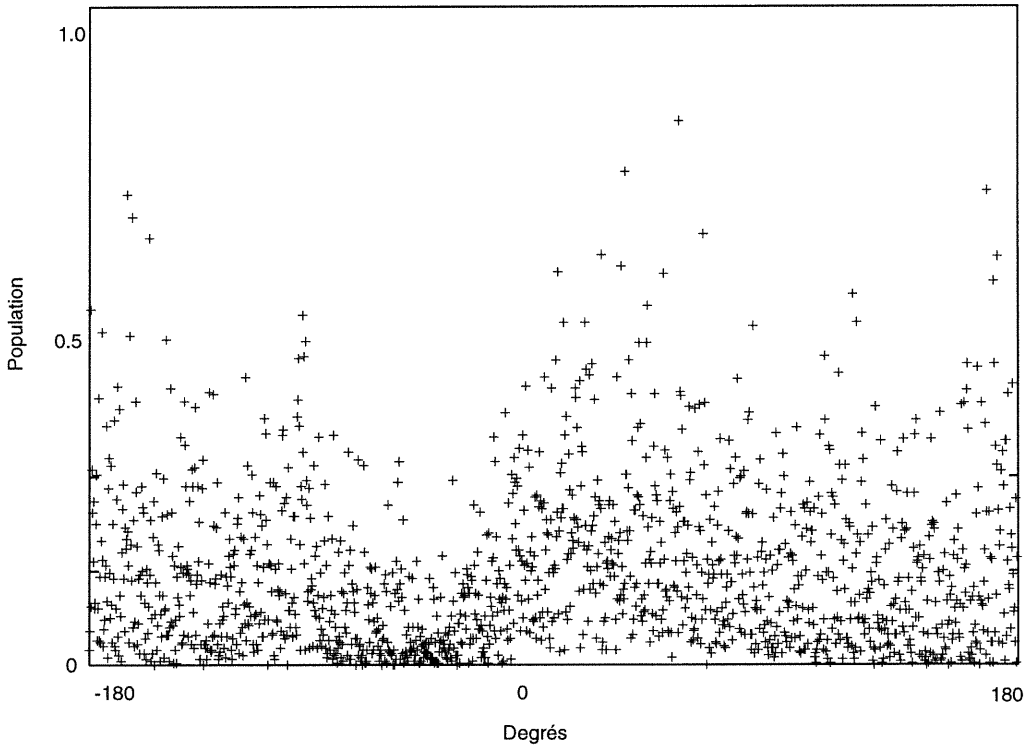


FIGURE 35. Distribution de la coordonnée interne q5 des triplets acceptés.

Des figures précédentes, on remarque que grossièrement, la forme des distributions de chacune des coordonnées internes q1 à q5 se ressemblent par rapport aux distributions initiales des coordonnées internes q1 à q5. Toutefois, il y a une sur-utilisation d'une boîte de classement pour chacune de ces coordonnées. En effet, il existe un angle différent pour chacune des coordonnées dont la population est de 10 à 15 fois plus élevée que la deuxième boîte de classement la plus remplie. Ce détail est négligé sur les graphiques puisqu'il ne permettait pas de bien voir les distributions des triplets acceptés. Tout porte à croire que cette exception s'applique à un résidu et non à une moyenne sur la séquence, puisque s'il s'agissait de plusieurs résidus, on obtiendrait probablement une structure insolite. Trip semble faire une utilisation normale de la population des autres boîtes puisque la tendance générale montre que les distributions de ces boîtes ressemblent globalement aux distributions de départ.

En ce qui concerne la distribution des énergies et des RMS moyens en fonction des seuils d'homologie de séquence, les graphiques suivants (figures 36 à 41) illustrent l'évolution de ces paramètres. On remarque qu'il n'y a pas de consensus pour les seuils d'acceptation d'homologie de séquence d'une protéine à l'autre. D'autant plus que dans certains cas, ce seuil d'acceptation n'est pas le même entre l'énergie et le RMS d'une même protéine. Afin de choisir un seuil d'acceptation optimal, on doit se baser sur les résultats de RMS d'abord et avant tout. Les résultats de RMS sont calculés selon une formule stricte et de vraies distances entre chacun des atomes. L'énergie, elle, n'a aucune signification physique selon les lois de l'électromagnétisme et de la mécanique. Normalement, une bonne simulation devrait abaisser à la fois le RMS et l'énergie, cependant prendre seulement l'énergie comme facteur d'optimisation du seuil d'homologie de séquence serait une grave erreur. Ainsi pour la 3chy, on choisit un seuil de 0.50 (qui correspond aussi avec le minimum d'énergie). Il y a aussi un puits vers un seuil de -0.50, mais au niveau énergétique, il semble que seul le critère de 0.50 réponde à un minimum. Pour la 1mbo, le graphique du RMS ne démontre à peu près aucune amélioration peu importe le seuil. Par contre, si on s'aide avec le graphique de l'énergie, on pourrait avoir tendance à penser qu'un seuil d'acceptation de 0.75 serait fort acceptable compte tenu des résultats obtenus

avec cette protéine. Bien qu'en RMS, le seuil optimal semble être déterminé à -0.25, l'énergie correspondant à ce seuil est des plus élevée, on doit donc rejeter cette valeur. Finalement pour la 1aba, on choisit un seuil d'acceptation de 0.50 bien qu'au niveau de l'énergie moyenne, il ne s'agisse pas de la valeur la plus basse. Quoique dans l'ensemble, cette valeur d'énergie est sous la moyenne des valeurs sur ce graphique. Ainsi, le critère d'homologie de séquence est optimisé. Il faut maintenant optimiser le critère d'homologie pour les structures avoisinantes. Le tableau V montre les résultats optimaux obtenus. Premier constat, ce tableau ne diffère que très peu avec le tableau IV. Les seules valeurs qui changent ont un meilleur RMS minimal pour la 3chy et un meilleur RMS moyen pour la 1mbo. À part ces deux exceptions, les résultats sont les mêmes. La raison est simple toutefois; d'un côté, on améliore la qualité de la distribution selon le critère structurel mais de l'autre côté, le seuil enlève de plus en plus de choix de géométries de triplets. La qualité s'en trouve améliorée certes mais le faible choix ne permet pas à Trip de construire des boucles qui le favoriseront dans sa quête de la structure native.

TABLEAU V. Énergies et RMS optimaux obtenus pour diverses simulations sur trois protéines test avec les deux critères d'homologie optimisés

Protéine	Meilleure énergie	Meilleure énergie moyenne	Meilleur RMS (Å)	Meilleur RMS moyen (Å)
3chy	-140.54	-101.66	6.27	10.71
1mbo	-34.98	0.84	6.41	10.80
1aba	-114.76	-80.90	6.30	8.64

Les graphiques (figures 42 à 47) illustrent l'évolution des énergies moyennes et des RMS moyens en fonction du seuil d'acceptation d'homologie de structure. Un bémol à apporter à ces graphiques: l'échelle de l'abscisse n'est pas conséquente avec les valeurs notées. La justification provient du fait qu'entre 1.50 et 1.60, il y a beaucoup de ségrégation dans la sélection des structures selon ce critère d'homologie de structures avoisinantes. Ce fait contraste avec les valeurs du critère de sélection pour l'homologie de séquence dont la ségrégation s'étend à peu près également entre -1 et 1. Afin de mieux voir les fluctuations

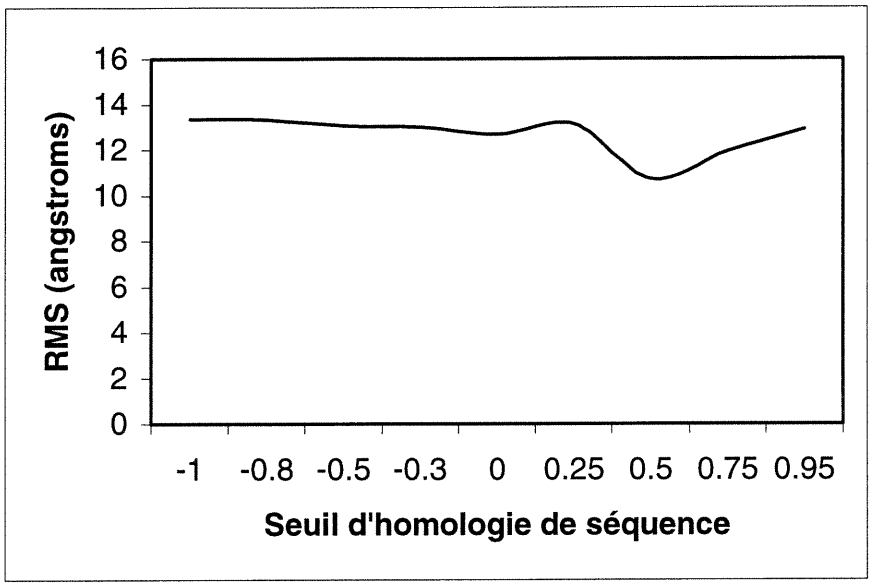


FIGURE 36. Évolution du RMS moyen en fonction du seuil d'homologie de séquence pour la 3chy.

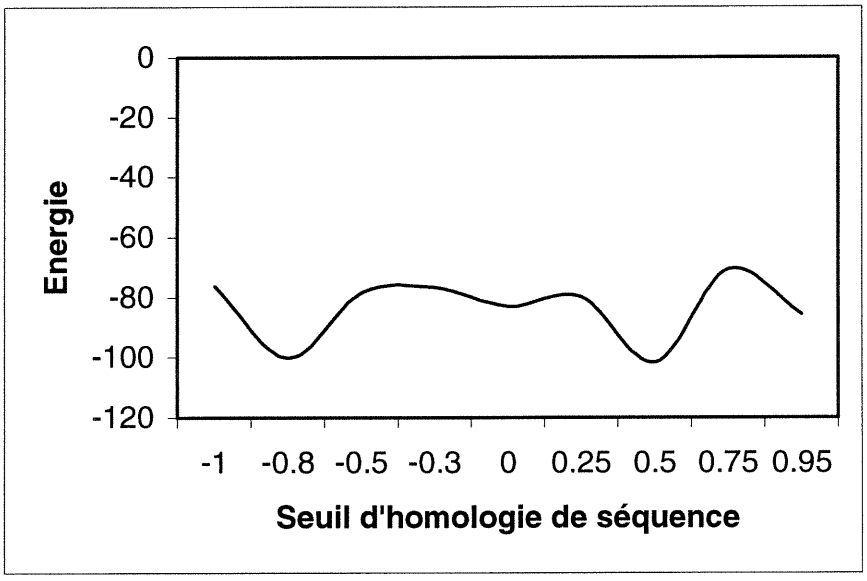


FIGURE 37. Évolution de l'énergie moyenne en fonction du seuil d'homologie de séquence pour la 3chy.

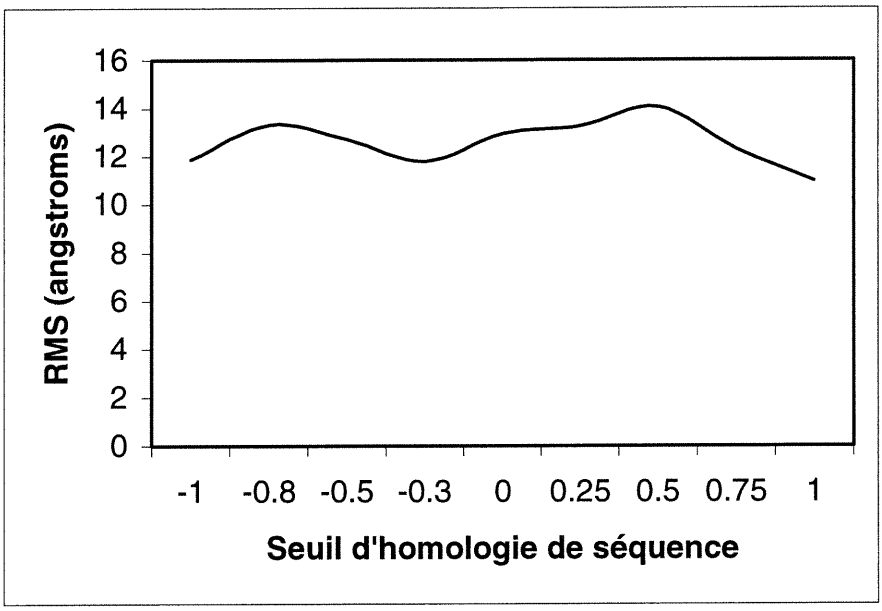


FIGURE 38. Évolution du RMS moyen en fonction du seuil d'homologie de séquence pour la 1mbo.

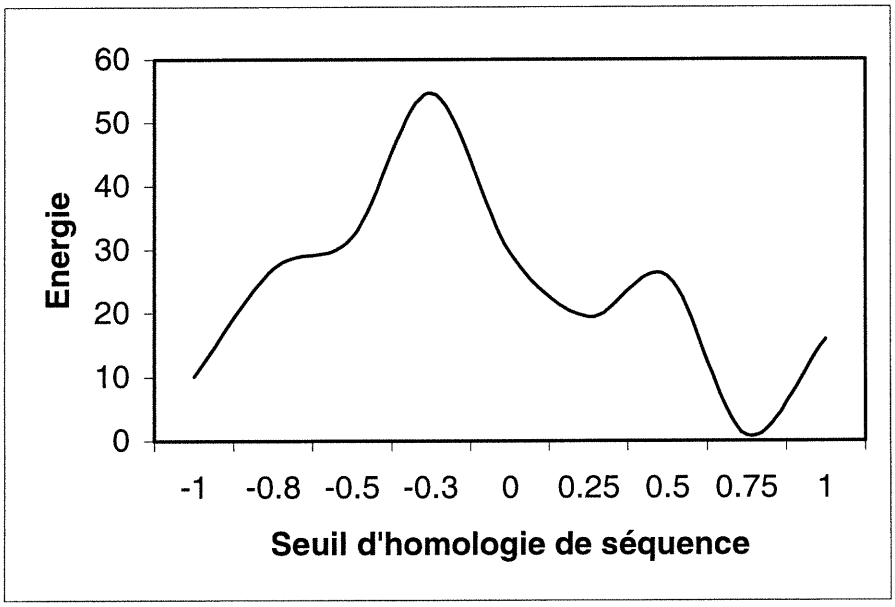


FIGURE 39. Évolution de l'énergie moyenne en fonction du seuil d'homologie de séquence pour la 1mbo.

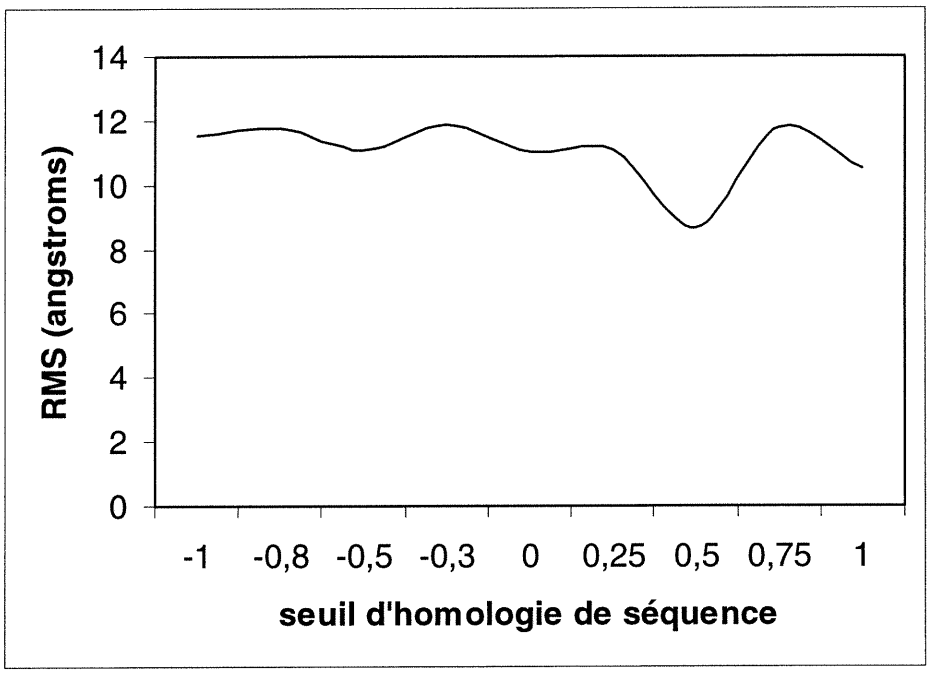


FIGURE 40. Évolution du RMS moyen en fonction du seuil d'homologie de séquence pour la 1aba.

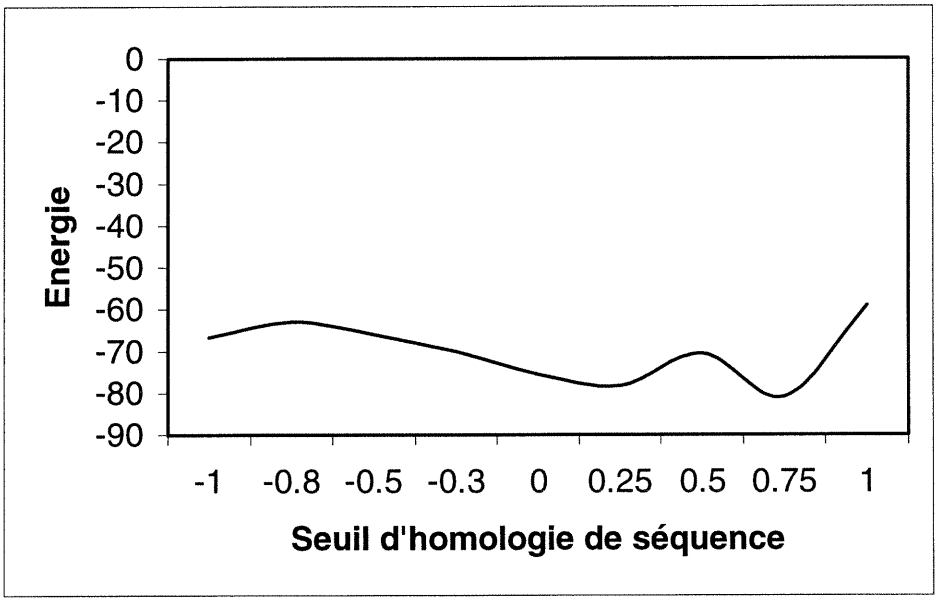


FIGURE 41. Évolution de l'énergie moyenne en fonction du seuil d'homologie de séquence pour la 1aba.

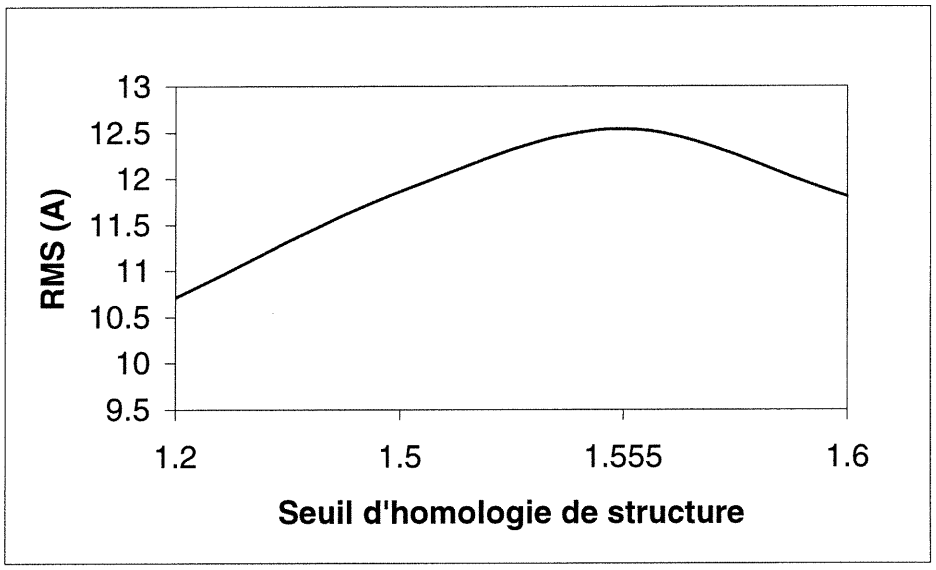


FIGURE 42. Évolution du RMS moyen en fonction du seuil d'homologie de structure optimisé pour la 3chy.

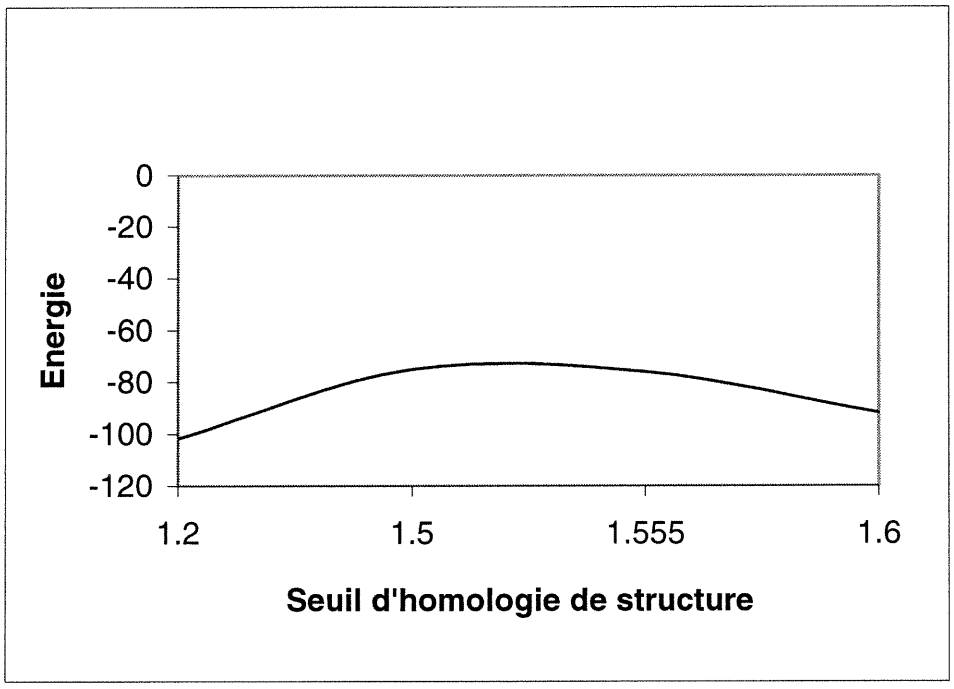


FIGURE 43. Évolution de l'énergie moyenne en fonction du seuil d'homologie de structure optimisé pour la 3chy.

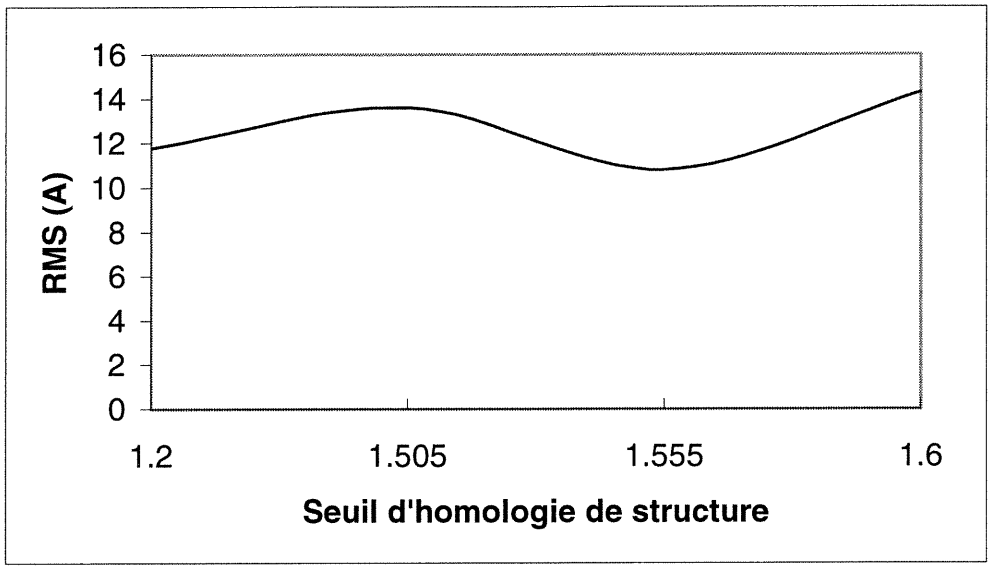


FIGURE 44. Évolution du RMS moyen en fonction du seuil d'homologie de structure optimisé pour la 1mbo.

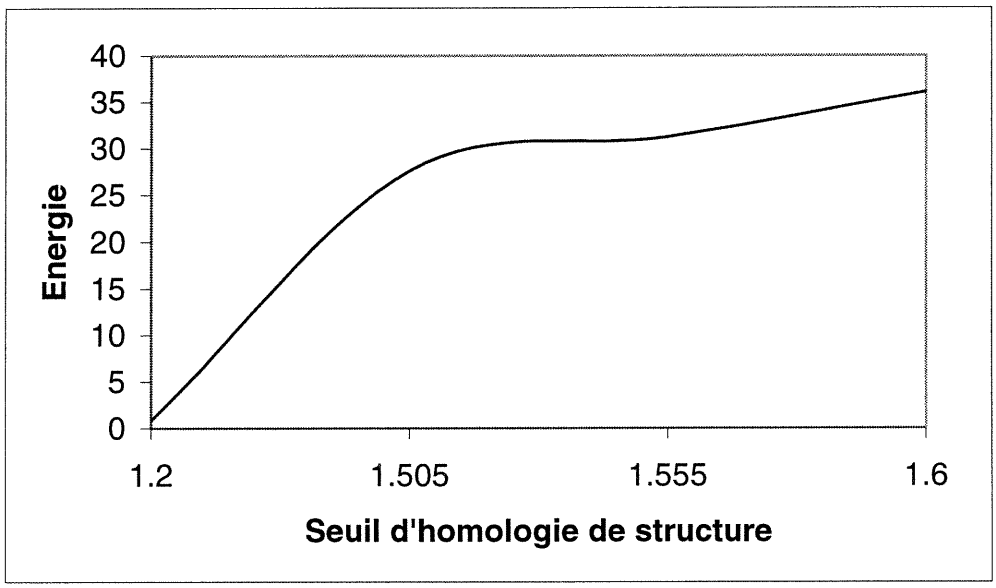


FIGURE 45. Évolution de l'énergie moyenne en fonction du seuil d'homologie de structure optimisé pour la 1mbo.

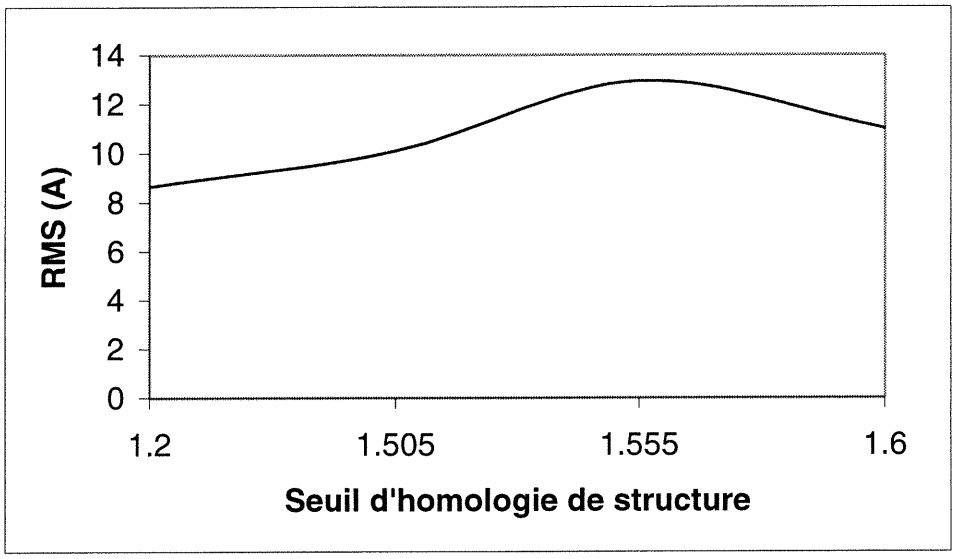


FIGURE 46. Évolution du RMS moyen en fonction du seuil d'homologie de structure optimisé pour la 1aba.

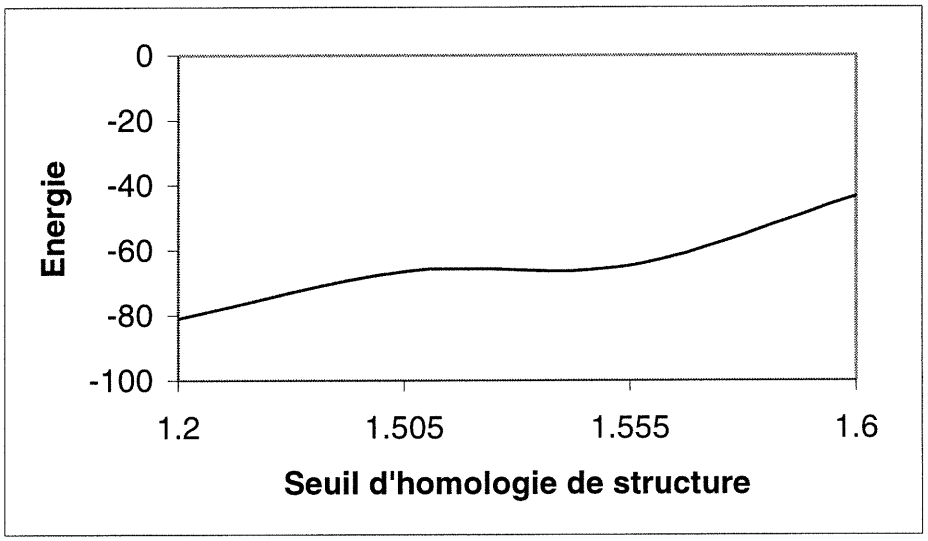


FIGURE 47. Évolution de l'énergie moyenne en fonction du seuil d'homologie de structure optimisé pour la 1aba.

d'énergie et de RMS, les valeurs ont été placé à égale distance les unes des autres.

On remarque dans le cas de la 3chy (figures 42 et 43), qu'il y a nettement un minimum en RMS et en énergie lorsque le seuil d'acceptation est très permissif. Pour la 1mbo (figures 44 et 45), le RMS présente deux puits peu profonds, mais le graphique de l'énergie nous montrent clairement que le seuil le plus permissif encore une fois représente le choix le plus avantageux. Finalement, la 1aba (figures 46 et 47), comme la 3chy, montre un net avantage encore une fois à utiliser un seuil très permissif au niveau de la structure.

3.8. Apport de la distribution biologique

Deux faits sont à mentionner avant de passer définitivement aux conclusions de ce chapitre. D'abord, comme les seuils d'acceptation optimaux en séquence et en structure étaient parfois trop stringents, on a dû aider le calcul. Dans certains cas, ces critères étaient tellement restrictifs, qu'aucun modèle de triplet ne pouvait être accepté dans la construction de boucles. Pour remédier à ce problème, un seuil d'acceptation ajustable durant du calcul fut implémenté pour l'homologie de structures. Ainsi, lorsque la routine passe toute la liste de géométries de triplets et qu'aucun n'a été sélectionné, le seuil d'acceptation est automatiquement ajusté de façon à accepter un certain nombre de géométries de triplets dans la distribution. Résultat: certains triplets se doivent d'accepter des géométries qui sont moins bonne qualité que le seuil l'exige.

Pour terminer cette section, il ne reste qu'à jeter un coup d'oeil quant à l'amélioration due à l'apport d'une distribution de géométries de triplets biologiques comparativement à l'ancienne distribution statistique de Trip. Les tableaux III, IV et V témoignent de l'avantage d'utiliser une distribution biologique au lieu d'une distribution non-naturelle comme précédemment.

Les résultats démontrés sont concluants. La distribution biologique donne nettement de meilleurs RMS en abaissant les valeurs des calculs nominaux de 2.81Å pour la 3chy et les valeurs de RMS minimales de 3.51Å pour la 3chy et de 2.03 Å pour la 1aba. Les

valeurs énergétiques aussi suivent avec une amélioration sur les scores énergétiques moyens de 28.09 unités d'énergie pour la 3chy et de 11.84 unités d'énergie pour la 1aba. Seule la 1mbo semble poser des problèmes due à la présence de sa boucle de 3 résidus. Toutefois, les résultats obtenus avec la distribution biologique sont améliorés par rapport aux simulations nominales. Conclusion: chaque protéine est un cas particulier et doit être testée mais il semble qu'il y ait une nette amélioration grâce à l'apport d'une distribution de géométries de triplets biologiques.

CHAPITRE 4

Distribution naturelle de boucles

Maintenant que Trip possède une distribution naturelle de géométries de triplets, il sera question de lui imposer une liste de géométries de boucles naturelles. Cette nouvelle liste de boucles fournira à Trip des modèles de référence. Durant le processus de construction de boucles, une fonction se chargera de vérifier selon certains critères, la ressemblance entre une boucle générée et un choix de boucles prédéterminé dans la liste prévue à cet effet. Trip pourra ensuite s'inspirer de ces géométries de boucles afin d'apporter des mutations sur les boucles qu'il a lui-même générées. La liste de boucles naturelles tire son origine de la même banque de données de protéines qui a servit précédemment à élaborer la liste de triplets naturels.

4.1. La banque de données de boucles

Après une vérification de la même banque de données de protéines qui a servi précédemment pour les triplets, certaines structures doivent être rejetées. D'abord celles qui possèdent trop de trous dans leur boucles et celles dont le fichier de la Protein Data Bank est truffé d'erreurs. De surcroît, les boucles contienant des résidus prolines de configuration cis seront aussi éliminés car le traitement exige des outils de configuration plus complexes et trop coûteux en temps pour le peu de structures à traiter. Après épuration, on compte 27956 boucles de longueurs variées dans la liste. La figure 48 montre la distribution de ces boucles de géométries naturelles. Les boucles de un et deux résidus sont éliminées puisque la plus petite unité hiérarchique dans Trip est un triplet. On voit à partir de cette distribution que les boucles de trois et quatre résidus sont en très grand nombre et que par la suite, la courbe de distribution diminue assez rapidement à mesure que la longueur des boucles se fait plus importante. L'éventail possible est

évidemment plus restreint que celui qu'offre la distribution des triplets, mais comme il s'agit d'un guide d'orientation pour Trip et non du nombre total et fini de choix de boucles, cette distribution demeure acceptable et traitable.

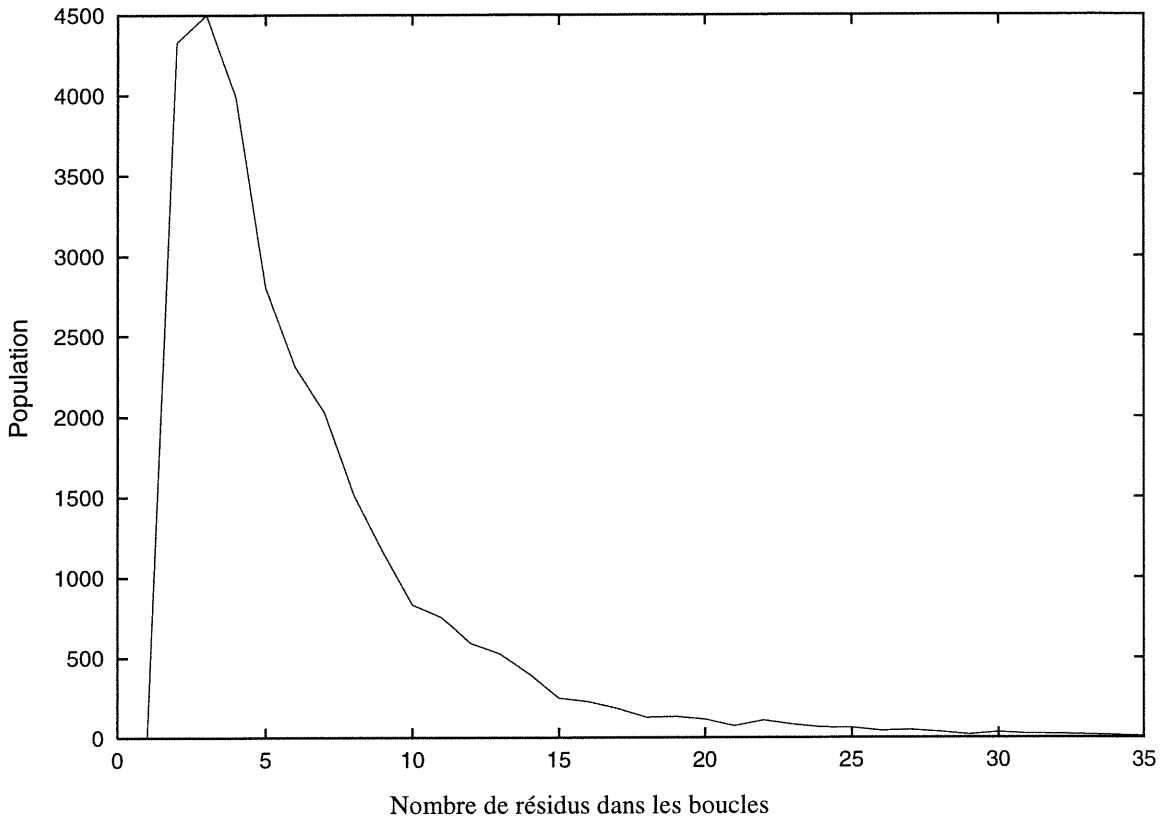


FIGURE 48. Distribution des boucles exprimée selon le nombre d'occurences en fonction de la longueur individuelle des boucles.

Cette liste de boucles servira à établir une distribution biologique de géométries de boucles pour Trip. Afin de savoir si une structure de boucle est acceptée ou non dans la nouvelle distribution, on aura recourt encore cette fois à l'homologie. Cette fois l'homologie sera de type environnement tridimensionel versus séquence en acides aminés. Comme précédemment avec les triplets, un seuil d'acceptation devra être établi de raffiner la distribution des boucles. La méthode d'homologie dont il est question pour cette section a été établie par Eisenberg. En voici donc une description.

4.2 Les profils d'environnement d'Eisenberg[56,57]

À l'origine, les profils d'environnement d'Eisenberg contribuent à la résolution du problème du repliement inverse. Cette méthode constiste à déterminer la séquence d'acides aminés qui se replie selon une structure tridimensionelle déjà connue. Ce procédé permet de détecter des similarités structurelles entre des familles de protéines qui ne démontrent aucune ressemblance séquentielle détectable. Dans le cas présent, on applique les travaux d'Eisenberg sur la liste de boucles issues de l'ensemble de structures non-redondantes de protéines. Une matrice de comparaison sera construite à partir de l'information extraite de ces boucles. Cette matrice contiendra les scores relatant chacun des acides aminés aux paramètres d'environnement d'Eisenberg. Il s'agit du troisième type d'homologie qui servira au coeur du processus hiérarchique de Trip. Ces profils d'environnements se définissent en trois segments, l'aire enfouie d'un résidu, la fraction polaire de sa chaîne latérale et la classe d'environnement à laquelle ce résidu appartient.

4.2.1. L'aire enfouie d'un résidu

Premièrement, on tient compte de l'aire enfouie des résidus dans une protéine, c'est-à-dire, qui est inaccessible au solvant. En d'autres mots, on parle ici de l'aire de la chaîne latérale enfouie par les autres atomes de la protéine. Il faudra déterminer l'aire accessible au solvant et la retrancher de l'aire totale. L'aire accessible au solvant se calcule selon la méthode de Lee et Richards[58] dans laquelle il s'agit de placer des sphères imaginaires de solvant autour de chaque atome de la protéine en question. Ces sphères de solvant, dotées du rayon de Van der Waals d'une molécule d'eau, sont insérées autour des résidus, là où c'est possible. Évidemment, s'il est imposible d'insérer une molécule d'eau, on qualifiera cet aire de atome comme enfouie. Une "sonde" vérifie à tous les 0.75Å si l'espace ponctuel dans lequel elle se trouve touche une des sphères de solvant ou non. Si ce point se situe à l'intérieur d'une sphère de n'importe quel atome de solvant, cette région permettait à la molécule de solvant de s'y insérer. Dans ce cas, on dénote cette aire

accessible à l'eau. Dans le cas inverse, il s'agit d'une aire enfouie. L'aire accessible au solvant (AAS) est donnée par la formule suivante[56]:

$$AAS = (N_{acc} / N_{total}) \cdot Aire_{total} \qquad (4.1)$$

où: N_{acc} représente le nombre de points accessibles au solvant;
N_{total} représente le nombre total de points échantillonnés;
$Aire_{total}$ représente l'aire totale de sphères de solvant pour cet atome;

Le programme NACCESS[59] de Hubbard et Thornton fut utilisé pour déterminer ce critère de surface enfouie.

4.2.2. La fraction polaire des chaînes latérales

La deuxième étape consiste à calculer quelle fraction de l'aire des chaînes latérales est couverte par des atomes polaires. L'auteur définit par atome polaire, tout atome pouvant donner ou recevoir un pont-H. Comme précédemment, une "sonde" passe à tous les 0.75Å et vérifie s'il y a présence d'atomes polaires en cet espace ponctuel. Donc, pour chaque atome de la chaîne latérale, cette "sonde" vérifiera s'il existe un atome polaire avoisinant la dite chaîne latérale. Ainsi, il est possible de calculer quelle est l'aire couverte par des atomes polaires. Ensuite, cette aire couverte par des atomes polaires est divisée par une valeur d'aire totale calculée pour la dite chaîne latérale. Cette valeur d'aire totale pour la chaîne latérale d'un acide aminé X est tirée d'un calcul d'aire de chaîne latérale sur un tripeptide Gly-X-Gly[60]. Ainsi, la fraction polaire des chaînes latérales F, est définie selon la formule suivante[56].

$$F = N_p / N_{total} \qquad (4.2)$$

où: N_p représente le nombre de points échantillonnés recouvert par des atomes polaires ou exposés au solvant;

N_{total} est le nombre total de points échantillonnés;

4.2.3. Les classes d'environnements

À partir de ces deux derniers critères, il est possible de classer chacun des acides aminés d'une protéine dans des classes d'environnements bien définies. Au total, 18 classes environnements sont délimitées par ces critères, soit 6 pour les boucles, 6 pour les feuillets et 6 pour les hélices. Comme cette partie du projet de recherche porte uniquement sur les boucles, les classes d'environnements pour les structures secondaires seront éliminées. Les 6 classes possibles sont représentées au tableau VI avec leur délimitations respectives pour chacun des deux critères.

TABLEAU VI: Les 6 classes d'environnements et leur délimitations selon Eisenberg[56].

Aire enfouie du résidu en $Å^2$ (A)	Fraction polaire de la chaîne latérale (F)	Classe d'environnement selon Eisenberg
A >114	F < 0.45	B1
A >114	0.45 ≤ F < 0.58	B2
A >114	F ≥ 0.58	B3
40 < A ≤ 114	F < 0.67	P1
40 < A ≤ 114	F ≥ 0.67	P2
A ≤ 40	F ≥ 0.67	E

4.3. L'association entre Trip et l'information d'Eisenberg

À l'aide de ces 6 classes d'environnement et de la séquence en acides aminés des boucles que Trip construit, il sera question d'effectuer du "threading". Ce terme (portant la mauvaise traduction française d' "enfilement") consiste à comparer deux structures, résidu par résidu, selon un certain critère d'homologie. Dans le présent cas, il est question

d'une homologie de type séquence-profil d'environnement. Une matrice compilée à partir des critères d'Eisenberg doit donc être construite à partir de la liste de boucles. Cela consiste à compiler pour chaque acide aminé dans les boucles de la liste, son pourcentage de présence dans chacune des classes d'environnement. Voici une représentation graphique d'une matrice formée à partir des boucles de 300 des 2364 protéines de la banque de données (voir la figure 49). La légende associant les différents tons avec les 6 classes est affichée au bas de la figure 49.

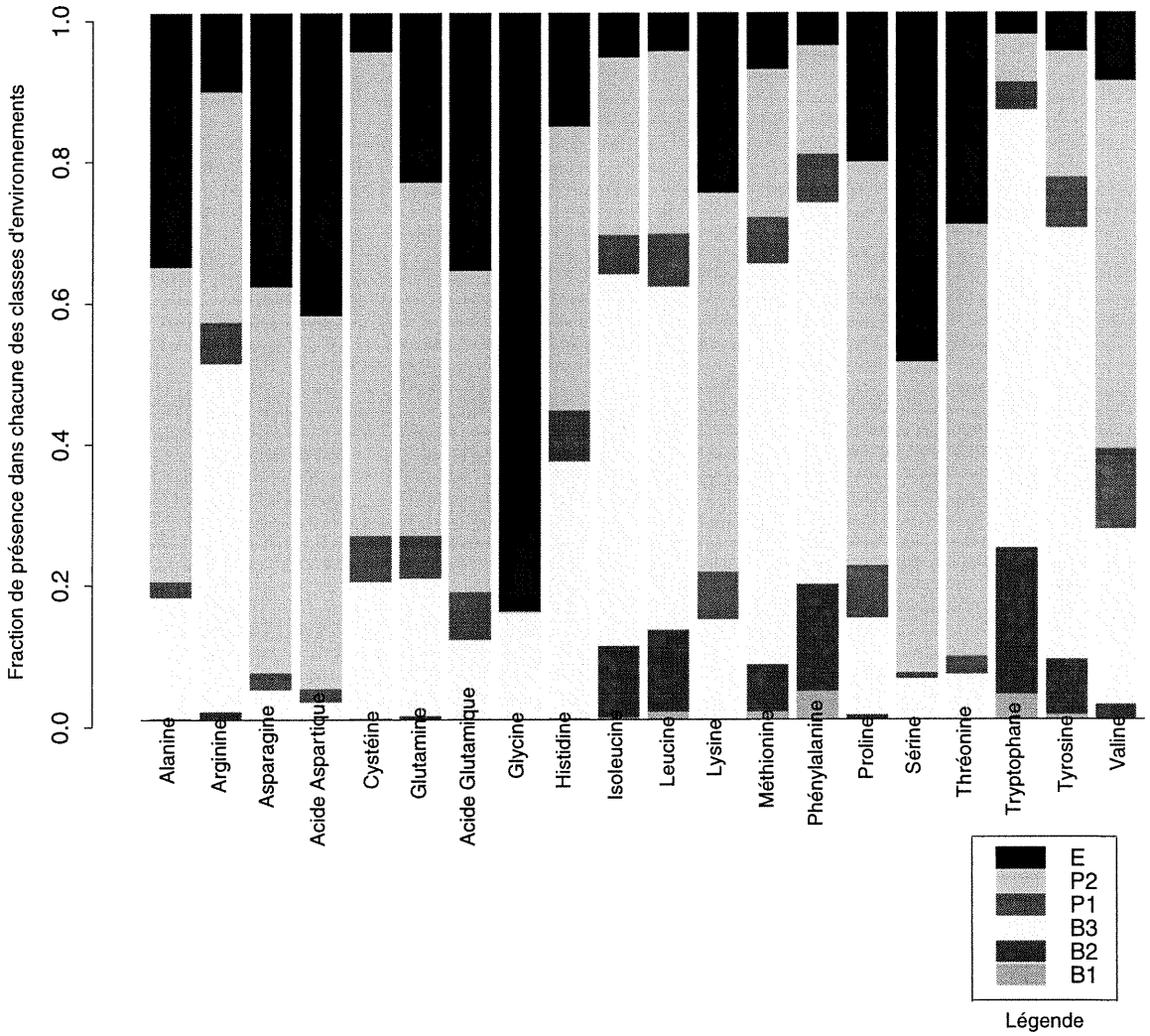


FIGURE 49. Histogramme des acides aminés et de leur pourcentage de présence dans chacune des 6 classes d'environnements d'Eisenberg.

En regardant certains faits saillants de cet histogramme (figure 49), on remarque que ces classes représentent bien la réalité. Au début du chapitre premier, la nature des 20 acides aminés naturels fut discutée. On peut ici confimer ce qui a été mentionné précédemment. Par exemple, la glycine se retrouve principalement dans la classe E, dans la classe B3 ainsi qu'une infime partie dans la classe P2. Ces 3 classes sont caractérisées par une fraction polaire très grande. Chose véridique puisque la glycine possède la plus petite chaîne latérale de tous. Ces classes nous indiquent aussi que la glycine se retrouve principalement en surface et dans le coeur de la protéine. Cet acide aminé met donc sa grande flexibilité en jeu lorsqu'il est question de ramifier les extrémités vers le milieu structurel. Une remarque d'ordre générale se doit d'être mentionnée dû au fait qu'on ne regarde que les acides aminés dans les boucles. Ceci dit, très peu d'acides aminés ont la propriété de se retrouver au coeur de la protéine. De ceux-là on répertorie la phénylalanine, le tryptophane, l'isoleucine, la tyrosine et la leucine principalement. À l'inverse, tous les acides aminés ont au minimum une probabilité non-nulle de se retrouver en surface qu'ils soient polaires ou non. Dernier détail d'ordre général, la classe la plus présente semble être la P2. C'est dans cette classe que l'on retrouve un enfouissement moyen et une fraction d'aire chaîne latérale accessible au solvant maximum. C'est dans cette région délimitée par la classe P2 que l'on retrouve la majorité des enchevêtrements structurels.

4.4. Détails d'implémentation

Pour des raisons de simplicité, avant même de se lancer dans un processus d'optimisation plus approfondi, il est de mise de trouver une méthode rapide et efficace afin d'implémenter les informations contenues dans les boucles vers Trip. Ce dernier gère les boucles selon une liste finie de choix d'angles dièdres. Pour l'instant, une carte de Ramanchandran possédant 100 couples d'angles dièdres est suffisante pour effectuer des simulations. L'apport de couples d'angles supplémentaires n'améliore aucunement les résultats. Trip ne possède pas suffisamment de finesse pour utiliser une carte plus vaste.

Avec l'apport d'une liste naturelle de boucles, il devient nécessaire d'inclure une carte plus permissive. Le problème se résume comme suit. Dans la nature, tous les couples d'angles inscri sur les cartes de Ramanchandran initialement sont permises selon différentes probabilités. Trip fonctionne en approximant la géométrie des boucles selon la carte dièdre qui lui est imposée. Or, comme il s'agit d'importer dans Trip de l'information biologique, une carte de 100 angles dièdres est largement insuffisante afin d'obtenir des structures similaires entre le produit brut provenenant de la PDB et l'information qui doit entrer dans Trip. On doit utiliser une carte comportant au minimum 1093 couples d'angles. Ainsi, en effectuant une mesure de RMS entre les structures biologiques brutes et les structures transformées en format TRIP, on voit clairement que la carte de 1093 couples d'angles permet d'obtenir des structures beaucoup plus similaires.

Les figures 50 et 51 illustrent respectivement la carte d'angles dièdres à 100 points et la carte d'angles dièdres à 1093 points. La carte à 1093 couples possède des points qui n'existent pas dans la nature. Mais il ne faut pas oublier qu'il s'agit ici d'approximations, la carte utilisée dans Trip ne possède pas une infinité de points. Si une structure a besoin d'une paire d'angles inexistante sur la carte, une approximation au couple le plus près sera de mise. Le logiciel MOLFIX[46] conçu par Pierre-Jean L'Heureux effectue ce transfert entre l'information brute et ce que Trip peut assimiler, selon une carte d'angles dièdres choisie. Le but du logiciel MOLFIX consiste à trouver la meilleure approximation pour une structure brute en utilisant que les liens et les angles du modèle réduit des protéines au lieu de prendre les angles tels quels.

4.5. Efficacité des profils d'environnement

Afin de vérifier l'efficacité de ce type d'homologie, quelques tests statistiques s'imposent. D'abord, on peut simuler le problème du processus de repliement inverse. On choisit une séquence cible de 4 acides aminés au hasard dans la banque de boucles naturelles et on

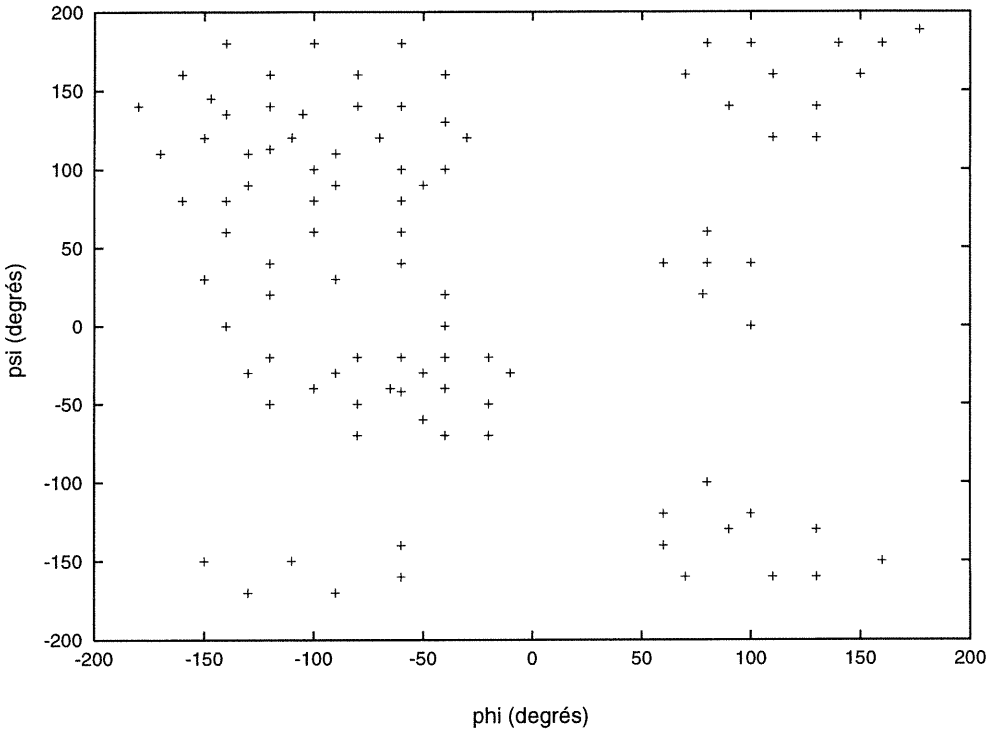


FIGURE 50. Carte de Ramanchandran à 100 paires d'angles.

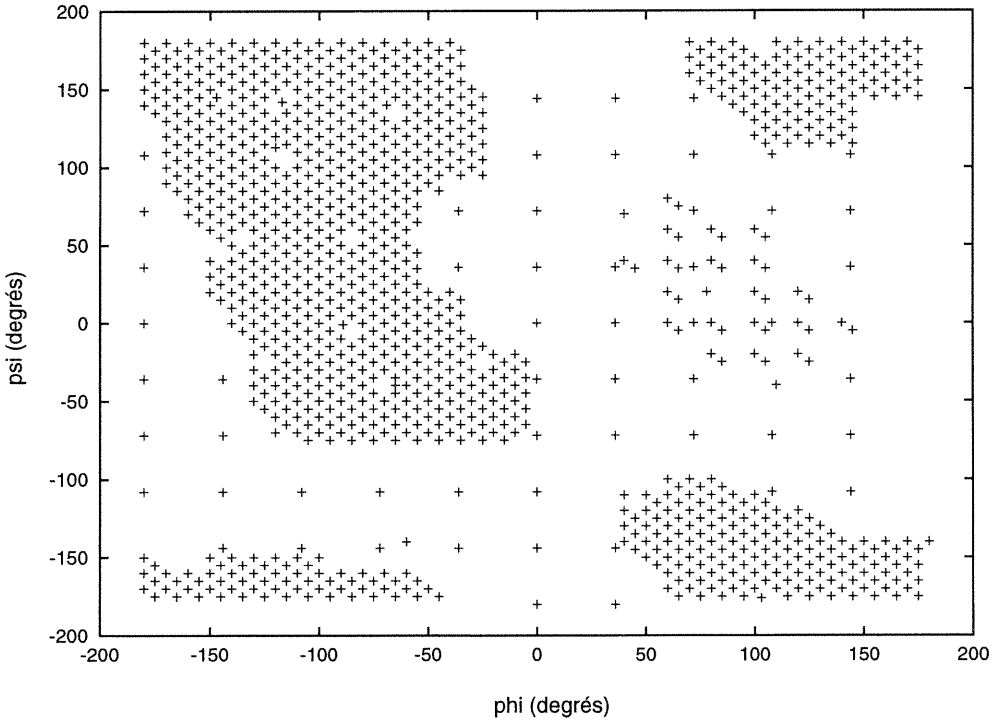


FIGURE 51. Carte de Ramanchandran à 1093 paires d'angles.

regarde parmi la liste de boucles naturelles celles de même longueur. À chaque nouvelle boucle de la liste, on calcule d'abord le RMS entre les deux boucles en cause, puis, on calcule un score entre ces deux mêmes boucles. Le score, E, se calcule selon la formule suivante:

$$E = -\sum_i \ln (p_i + c) \qquad (4.3)$$

où: p_i est la probabilité de retrouver l'acide aminé i dans une classe d'environnement;
c est une constante fixée à 0.01;

Ce score s'interprète comme une fonction d'énergie dans laquelle, plus sa valeur est basse, plus les deux structures comparées sont compatibles. Puisque certains acides aminés ont une probabilité nulle de se retrouver dans une classe d'environnement donnée, une constante c est ajoutée au score. Sa valeur fut fixée en regardant les probabilités les plus faibles dans la matrices d'homologie selon Eisenberg. De cette façon, on évite d'obtenir des logarithmes de zéro, donc des valeurs impossibles. Ce processus sera répété pour 3 autres boucles cibles de 4 résidus tirées de la banque de données. Une version de ce test sera aussi effectuée avec 6 boucles de 10 acides aminés. On utilise 6 patrons de boucles de 10 résidus au lieu de 4 pour obtenir une distribution plus représentative puisqu'il y a

TABLEAU VII: Patrons de boucles à 4 et à 10 résidus (exprimés en code de 1 lettre) utilisés pour les graphiques de comparaison.

Patrons de boucles de 4 résidus	Patrons de boucles de 10 résidus
GKDF	GDPDIGWYFK
PSVL	GHSNPEEFYW
LLNN	APVCGDTTGS
DGIR	LDLMQVPSHT
----	YAGAAVDELG
----	DRIGELKSGD

moins de boucles de 10 résidus que de boucles de 4 résidus. Voici au tableau VII, les différentes séquences qui ont été choisies au hasard pour effectuer ces tests.

On peut voir ces graphiques de comparaison du RMS versus les scores selon les profils d'environnements d'Eisenberg (voir figure 49), aux figures 53 et 54. À la figure 53, pour l'étude des boucles de 4 résidus, on peut voir trois régions distinctes. La première région, la plus dense, se retrouve approximativement entre 0 et 3 Å de RMS. La limite supérieure en score est environ de 15 unités. C'est dans cette région que l'on retrouve les structures les plus compatibles avec ce type d'homologie. Dans la deuxième région, bordée de 3 à 5 Å en RMS et de 9 à 25 unités en score, on retrouve des structures moins intéressantes qu'on pourra rejeter lors de la sélection dans Trip. Finalement, le nuage supérieur de points contient des structures nullement intéressantes pour l'étude. Dans le présent cas, il est clair que le seuil d'acceptation s'arrête à environ 8 en unités de score et à 3 Å en RMS. En deçà de cette valeur, on peut considérer ces boucles comme des modèles à l'interne de Trip. La figure 54 montre pour sa part, un graphique d'une allure complètement différente. Comparativement, au cas des boucles de 4 résidus, on ne voit aucune focalisation. On dénote grossièrement deux régions différentes dont l'une possède de bons RMS et l'autre de mauvais RMS. Les scores de ces structures s'étalent du minimum au maximum pour le nuage de bons RMS. À noter, les 6 points qui ont un RMS de 0Å sont les comparaisons des patrons par rapport à eux-mêmes. Pour le nuage de mauvais RMS, l'étalement est sensiblement plus circonscrit, mais dans ce cas, les structures ne sont d'aucun intérêt.

De cette situation, surgit un problème: à partir de quel score peut-on discriminer les structures qui serviront de modèles? Cette discrimination était évidente pour les boucles à 4 résidus, puisqu'il y avait une région clairement délimitée, où on pouvait trouver de bonnes structures. Il semble évident de trouver un seuil d'acceptation pour les boucles plus petites, mais l'état de la situation ne semble pas le même pour les boucles de plus grande envergure. Cette situation, bien que triste, reste néanmoins normale. Trouver une structure compatible pour un patron de petite taille semble beaucoup plus facile, compte

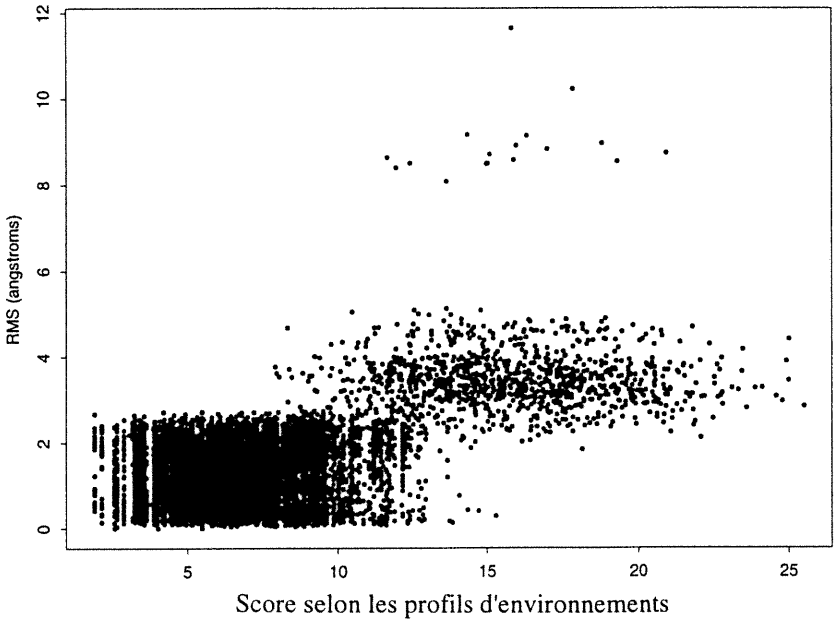


FIGURE 52. Étude du RMS en comparaison avec le score selon les profils d'environnement de 4 patrons de boucles de 4 résidus différents.

FIGURE 53. Étude du RMS en comparaison avec le score selon les profils d'environnement de 6 patrons de boucles de 10 résidus différents.

tenu du nombre de boucles disponibles. De plus, il est plus évident de trouver un match quasi-parfait lorsqu'on a affaire avec de petites boucles, compte tenu du nombre restreint de combinaisons possibles d'acides aminés.

Pour vérifier cette situation, on doit effectuer un autre test statistique afin de déterminer des seuils d'acceptation potentiels. On procédera à une étude de la population de structures par strates de RMS. Cette méthode sera répétée pour des patrons de boucles de 4 résidus et des boucles de patrons de boucles de 10 résidus. Du même coup, on peut vérifier si la distribution de boucles biologiques peut orienter Trip en lui fournissant des modèles de qualité. Pour vérifier ce fait, on demandera à Trip de générer 500 boucles au hasard à partir d'une boucle native, comme il le fait au cours d'un calcul standard. On disposera de ces résultats de la même façon que précédemment, selon des graphiques de population en fonction du RMS. Pour faciliter les comparaisons, les populations seront exprimées en fraction (donc en population relative), afin de conserver une même échelle de grandeur. D'abord, si on regarde la figure 55, on constate un maximum de structures autour de 2 Å de RMS. Ceci signifie que Trip peut générer au hasard des boucles de 4 résidus dont la majorité ressemble au patron suggéré avec un écart de 2 Å seulement. En comparant ce résultat avec celui de la figure 56 (où sont représentées les populations relatives naturelles), on voit que la distribution biologique de boucles peut fournir quelques modèles de bonne qualité à Trip puisque le maximum se trouve en-deçà de 1 Å de RMS. Il s'agit cependant d'une région très restreinte autour de 1 Å de RMS, ce qui sousentend que peu de modèles seront suggérés à Trip. En soit, ce n'est pas mauvais, mais il faut garder en tête que plus le nombre de modèles suggérés est grand, plus Trip possède de choix afin d'apporter des mutations à ses boucles. De plus, ces graphiques confirment les seuils de discrimination trouvés par le premier test statistique représenté à la figure 53.

Pour les boucles de 10 résidus, la situation est plus intéressante. La figure 58 montre un maximum de population de structures autour de 4 Å. Le RMS est évidemment plus élevé puisqu'il s'agit de comparaisons de structures plus longues entre elles. Le plus intéressant survient lorsque l'on compare ce graphique avec la figure 57, où est représentée la

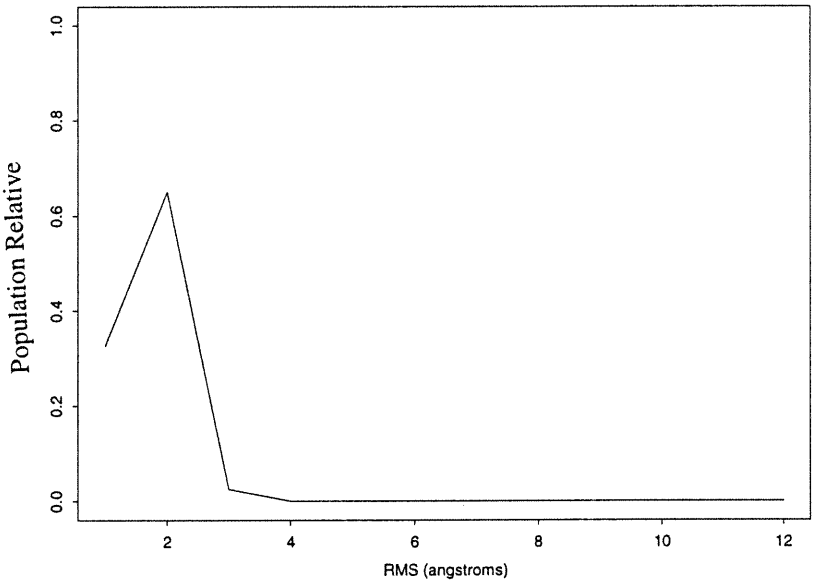


FIGURE 54. Étude de la population relative au hasard de structures de 4 résidus selon l'étalement de leur RMS.

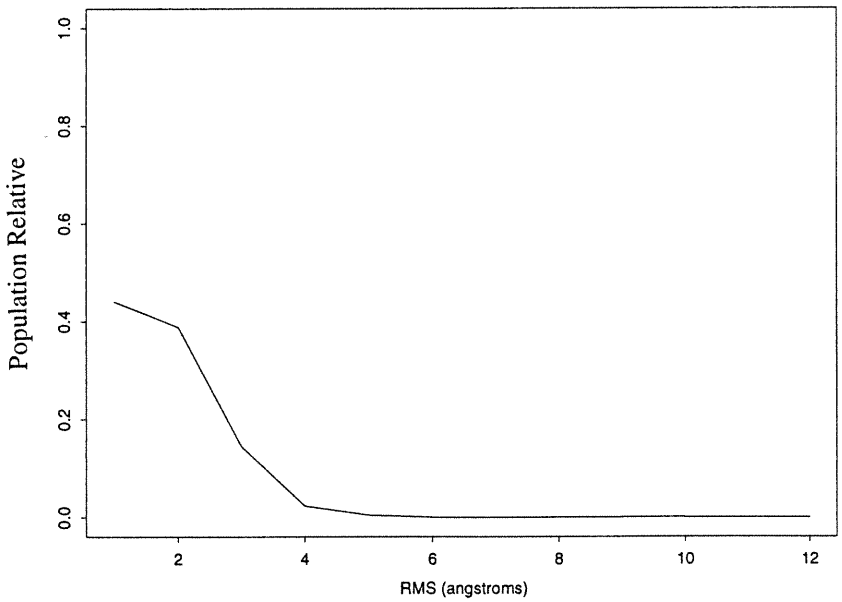


FIGURE 55. Étude de la population relative naturelle de structures de 4 résidus selon l'étalement de leur RMS.

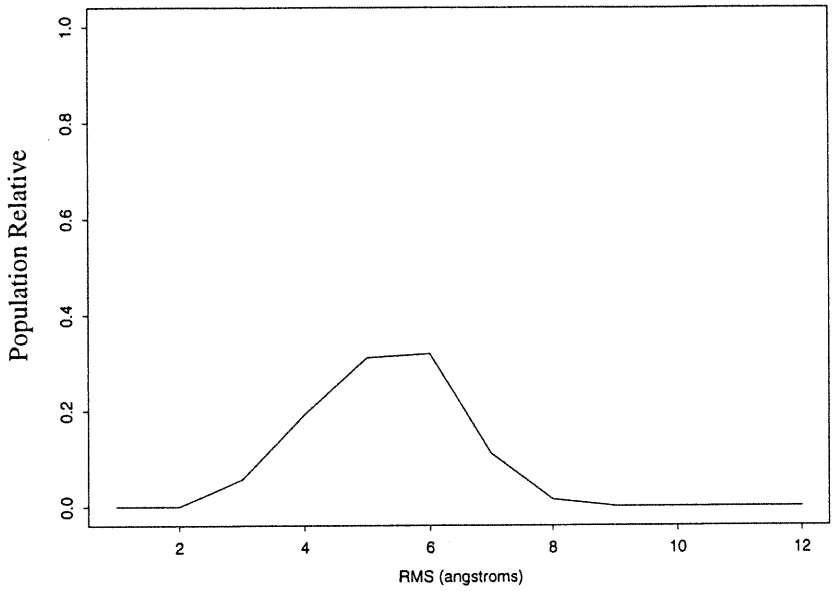


FIGURE 56. Étude de la population relative au hasard de structures de 10 résidus selon l'étalement de leur RMS.

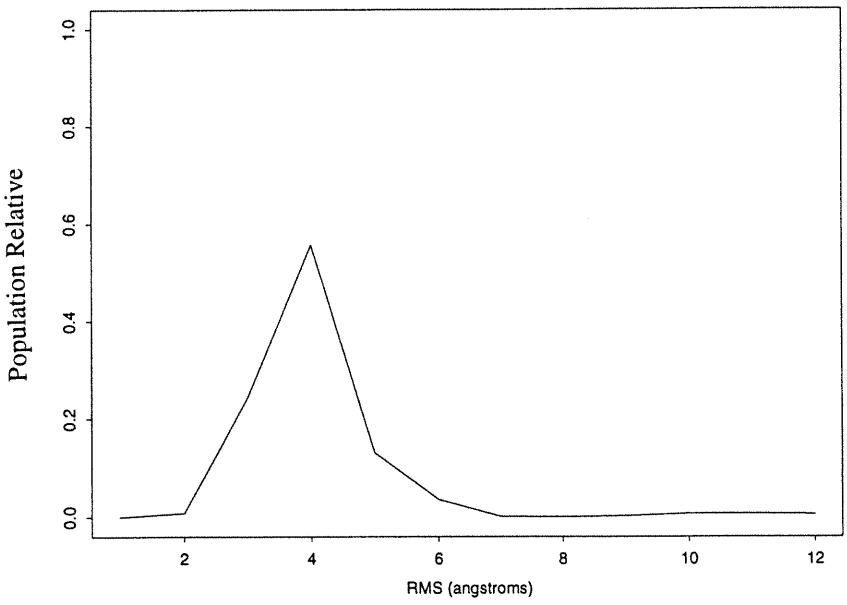


FIGURE 57. Étude de la population relative naturelle de structures de 10 résidus selon l'étalement de leur RMS.

population relative issue de Trip. Le maximum se trouve clairement décalé dans un espace de 5 à 6 Å. Dans ce cas, la distribution biologique contribue pleinement à Trip en lui apportant des modèles de boucles. De plus, comparativement au premier test statistique proposé à la figure 54, on peut pleinement décider d'un seuil de discrimination pour suggérer des modèles. Ainsi, les structures à moins de 5 Å peuvent entièrement considérées comme modèles potentiels pour Trip.

4.6. Conclusions préliminaires sur les profils d'environnement

À prime abord, les profils d'environnements peuvent sembler adéquats pour l'information que l'on veut introduire dans Trip. Il est possible de soumettre à Trip des modèles qui orienteront ses travaux de construction de boucles. Ce n'est pas toujours évident si on ne tient compte que du RMS en fonction des scores selon les profils d'environnement, mais un regard plus approfondi sur les populations relatives de structures que Trip peut générer par rapport aux populations relatives de structures que la distribution naturelle fournies, peut nous guider vers un seuil de discrimination pour les boucles.

Toutefois, ces avantages possèdent leur équivalent négatif. D'une part, la distribution de boucles naturelles de 4 résidus possède un certain nombre de modèles à suggérer à Trip compte tenu de la quantité présente dans cette distribution. Mais l'intervalle restreint de RMS le prive de plusieurs d'entre elles. D'autre part, la distribution de boucles naturelles de 10 résidus possède peu de structures à fournir à Trip, mais l'intervalle d'acceptation en RMS est plus large. Donc, dans les deux cas, on ne possède qu'un petit nombre de structures à suggérer comme modèles. Bien qu'on ne s'attende pas à en avoir une pléiade, un nombre plus important serait apprécié. D'autant plus que lors de simulations éventuelles, on aura à tester différents seuils d'acceptation afin d'ajuster la quantité et la qualité des modèles, un peu comme il a été fait avec les triplets. On risque ainsi de se retrouver avec trop peu de structures adéquates. Il en découle que la distribution biologique de boucles n'aura à peu près pas d'effet sur la construction. Aussi, il faudra

effectuer des calculs "pré-Trip" afin d'ajuster la liste de boucles à chaque nouvelle protéine. Mais, dans l'ensemble, les profils d'Eisenberg bruts semblent être une bonne source d'informations à fournir à Trip.

Chapitre 5

Conlusion

Le but du travail se définissait en deux étapes. Premièrement, étudier si l'apport d'une distribution naturelle de géométries de triplets peut orienter le processus de construction de boucles dans les protéines. Deuxièmement, vérifier si une distribution naturelle de géométries de boucles dont la sélection interne s'effectue selon des classes d'environnement, peut guider Trip dans sa quête de structure native.

La version 0.38 de Trip débute son processus hiérarchique de construction de structures en choisissant des triplets parmi une liste qu'il a lui-même créée depuis le poids d'une distribution de géométries non-naturelles. Une distribution naturelle fut compilée à partir d'une liste de structures protéiques non-redondantes. Les géométries de triplets inclus dans les boucles furent aussi extirpées de cette liste. Le poids de cette distribution s'ajuste selon deux critères de sélection pour les triplets. Le premier est un critère d'homologie de séquence selon une matrice de Dayhoff. Le deuxième est un critère homologie de structures avoisinantes aux triplets. Ces derniers se doivent d'être optimisés pour chacune des protéines. Deux paramètres ont dû être ajustés à l'interne de Trip afin que l'algorithme subisse pleinement le poids de la distribution biologique. Ainsi, le nombre de boîtes de classement et la quantité maximale de triplets à gérer par boîtes furent modifiés.

Les résultats obtenus montrent qu'en optimisant chacun des seuils d'acceptation pour les triplets, une fluctuation de la mesure de RMS et de l'énergie sont à noter pour trois protéines différentes. De plus, en comparant ces résultats avec ceux obtenus par des calculs Trip sans la distribution biologique, une nette amélioration de fait sentir au niveau du RMS et de l'énergie de la structure comparativement à sa structure native.

Dans le deuxième volet, une liste de géométries de boucles naturelles issues du même ensemble que celui de l'origine des triplets était en cause. Cette fois, les boucles servent de modèle à Trip au deuxième pas hiérarchique de son processus de construction. La sélection des boucles s'effectue selon les profils d'environnements d'Eisenberg. Ce type de threading permet la comparaison entre la séquence des classes d'environnement formées à partir de l'aire enfouie de résidus et de la surface polaire des chaînes latérales des résidus. Il semble que ce type de profils soient potentiellement efficaces afin d'apporter de l'information à Trip. Les structures de la distribution naturelle de boucles semblent partiellement de meilleure qualité par rapport à celle que Trip peut générer au hasard. Toutefois, il serait intéressant d'avoir un plus grand échantillonnage de boucles puisque lorsque ce n'est pas la quantité qui bloque l'apport d'informations, c'est la marge de manœuvre du RMS qui limite le nombre de boucles à inclure.

De ces travaux ainsi que des résultats qui en découlent, certains projets de recherche pourraient naître dans un avenir rapproché. Au niveau des boucles, un sujet d'étude consiste à vérifier si une distribution élargie de boucles non-redondantes permettrait l'implémentation de cette information naturelle au sein de Trip. D'abord, développer une méthode pour effectuer des calculs pré-simulations spécifiques à chaque boucle rencontrée dans les protéines à l'étude. Ensuite ajuster Trip pour qu'il intègre optimalement ces informations naturelles dans son processus de construction de boucles. Finalement, on se doit de lancer plusieurs simulations sur différentes protéines afin de vérifier l'efficacité de cet apport d'informations. Normalement, on espère que cela pourra guider Trip vers sa quête de structure tertiaire native.

REMERCIEMENTS

À John R. Gunn, pour m'avoir laissé la chance de participer à ses recherches. Pour sa patience et son appui malgré la situation difficile qui afflige l'avenir du groupe. Merci pour tout!

À ceux qui ont contribué financièrement et matériellement à l'avancement du groupe (Université de Montréal, CERCA, PENCE, CRSNG, FCAR, Chemical Computing Group) et au personnel de soutien de l'Université de Montréal et du CERCA.

À Pierre-Jean L'Heureux, qui fut mon mentor depuis mon entrée au sein du groupe. Ton esprit scientifique, ta générosité, ton âme de philosophe et ton sens des responsabilités font de toi un sage.

À Benoît Cromp, qui grâce à son esprit cartésien, réussit à trouver l'aiguille dans la botte de foin. Ton rire communicatif réchauffe l'atmosphère du groupe.

À mes parents, Claudette et Jacques, que j'adore et qui m'ont tout donné pour réussir dans la vie ainsi qu'à ma famille proche qui a cru en moi, je vous suis éternellement reconnaissant.

À mes grands ami(e)s, (Alexandre Caron, Isabelle Valade, Steve Sarmento et Dominik Herbart) avec qui je partage la pluie et le beau temps.

À Virginie Landreville, l'ange dont les ailes me couvrent en cette fin de maîtrise.

À tout ceux et celles que je ne mentionne point mais dont la présence m'a été bénéfique.

Bibliographie

1. J-R. Beaudry, Génétique générale. Décarie Editeur Inc, 1985, 501 p.
2. K. Arms et P.S. Camp, Biologie Générale, Éditions Études Vivantes, 1993, chapitre 2, pp.167-324.
3. J. Darnell, H. Lodish and D. Baltimore, Molecular Cell Biology, 2ème edition, W.H. Freeman & Co., 1990, chapitre 2, pp.44-45.
4. G. Zubay, Biochemistry, 2ème édition, MacMillan Publishing Co., 1988.
5. G.N. Ramachandran, C. Ramachandran et V. Sasisekharan, Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology,1963,**7**, pp.95-99.
6. G.N. Ramachandran et V. Sasisekharan, Conformation of polypeptides and proteins. Advances in Protein Chemistry, 1968, 23, pp.284-438.
7. D. Voet et J.G. Voet, Biochimie, De Boeck Université (2e édition), 1998, p.149.
8. T. E. Creighton, Proteins: Structures and Molecular Properties, 2ème edition, W.H. Freeman & Co., 1993, pp.6-17.
9. P-J. L'Heureux, Étude de l'espace dièdre dans un modèle de prédiction des protéines : effet d'une distribution ciblée, Université de Montréal, 1997, p. 6.
10. T. E. Creighton, Proteins: Structures and Molecular Properties, 2ème edition, W.H. Freeman & Co., 1993, p 4.
11. J. F. Leszczynski et G. D. Rose, Loops in globular proteins: a novel category of secondary structure, Science, 1986, 234, pp.849-855.
12. F. R. Salemme, Structural properties of protein β-sheets, Progress in Biophysics and Molecular Biology, 1983, 42, pp.95-133.
13. E. Y. Jones et A. Miller, Analysis of structural design features in collagen, Journal of Molecular Biology, 1991, 218, pp.209-219.
14. C. Chotia, Principles that determine the structures of proteins, Annual Review in Biochemistry, 1984, 53, pp.537-572.
15. T. E. Creighton, Protein Function. A practical approach., Academic Press (3e édition), 1975, pp.226-411.

16. K. A. Dill, Dominant Forces in Protein Folding, Biochemistry, 1990, 29, pp.7133-7155.
17. P. Atkins, Physical Chemistry, 5ème édition, W.H. Freeman & Co., 1993, p. 973.
18. A. S. Bhown, Protein/Peptide sequence analysis: current methodologies, CRC Press, 1988.
19. W. Saenger, Principles of Nucleic Acid Structure, Springer-Verlag, 1984.
20. L. Gomathi et S. Subramanian, Elucidation of Secondary Structures of Peptides Using High Resolution RMN, Current Science, 1996, 7, pp.553-567.
21. WC Johnson, Secondary Structure of Proteins Through Circular Dichroism Spectroscopy, Annual Review of Biophysics and Biophysical Chemistry, 1988, 17, pp.145-166.
22. 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, 1982, 39,183-231.
23. K. Wuthrich. Protein Structure Determination in Solution By Nuclear Magnetic Resonance Spectroscopy, Science, 1989, 243, pp. 45-50.
24. A. McPherson, The Preparation and Analysis of Protein Crystals, Wiley, 1982.
25. P. Koehl et M. Levitt, Theory and simulation: Can theory challenge experiment?, Current Opinion in Structural Biology, 1999, pp.155-156.
26. J. N. Onuchic, Z. Luthey-Schulten et P. G. Wolynes, Theory of protein folding: the energy landscape perspective, Annual Review in Physical Chemistry, 1997, 48, pp.545-600.
27. K. A. Dill et H. S. Chan, From Levinthal to pathways to funnels, Natural Structural Biology, 1997, 4, pp.10-19.
28. E. L. Shakhnovich, Protein design: a perspective from simple tractable models, Folding Design, 1998, 3, pp.R45-R58.
29. H. Dugas, Principes de base en Modélisation Moléculaire, 5e édition, Librairie de l'Université de Montréal, 2000, chapitre 3, pp. 47-148.
30. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, CHARMM: A program for macromolecular energy, minimization

and dynamics calculations, Journal of Computational Chemistry, 1983, 4, pp.187-217.

31. G. Casari et M. J. Sippl, Structure-derived Hydrophobic Potential: Hydrophobic Potential Derived From X-ray Structures of Globular Proteins is able to identify Native Folds, Journal of Molecular Biology, 1992, 224, pp.725-732.
32. P. D. Thomas et K. A.Dill, Statistical Potentials extracted from protein structures: How accurate are they? Journal of Molecular Biology, 1996, 257, pp.457-469.
33. J.T. Ngo et J. Marks, Computational complexity of a problem in Molecular structure prediction, Protein Engeneering, 1992, 5, pp.313-321.
34. J. R. Gunn, Sampling protein conformations using segment librairies and a genetic algorithm, Journal of Chemical Physics, 1997, 106, pp.4270-4281.
35. A. S. Lemak et J. R. Gunn, Rotamer specific potentials of Mean Force for Residue pair interactions, Journal of Physical Chemistry, 2000, 13, pp.1097-1107.
36. F. E. Cohen, T. J. Richmond et F. M. Richards, Protein folding: evalutation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example, Journal of Molecular Biology, 1980, 137, pp.9-22.
37. A. Kolinski et J. Skolnick, Monte Carlo simulations of protein folding I. Lattice model and interaction scheme, Proteins, 1994, 18, pp.338-253.
38. A. Kolinski et J. Skolnick, Monte Carlo simulations of protein folding II. Application to protein A, ROP, and crambin, Proteins, 1994, 18, pp.353-366.
39. P-J. L'Heureux, Étude de l'espace dièdre dans un modèle de prédiction des protéines: effets d'une distribution ciblée, Université de Montréal, 1997, pp.19-21.
40. N. Metropolis, Monte Carlo simulated annealing, Journal of Chemical Physics, 1953, 96, pp.768-780.
41. Z. Li et H. A. Scheraga, Monte Carlo Minimization approach to the multiple-minima problem in protein folding, Proceedings of the National Academy of Sciences USA, 1987, 84, 6611-6615.
42. B. Yeomans, Statistical Mechanics of phase transitions, Oxford Press, 1992, chapitre 7, pp.95-104.
43. S. Kirpatrick, C. D. Gelatt et M. P. Vecchi, Optimization by simulated annealing, Science, 1983, 220, pp.671-680.

44. F. E. Cohen et M. J. Sternberg, On the prediction of protein structure: the significance of the root-mean square deviation, Journal of Molecular Biology, 1980, 138, pp.321-333.

45. W. Kabsch, A solution for the best rotation to relate two sets of vectors. Acta Cristallographia, 1976, A32, pp.922-923.

46. W. Kabsch, A solution of the solution for the best rotation to relate two sets of vectors. Acta Cristallographia, 1978, A34, pp.827-828.

47. S. Karlin et S.F. Altschul, Applications and Statistics for multiple high-scoring segments in molecular sequences. Proceedings of the National Academy of Sciences of USA, 1993, 90, pp.5873-5877.

48. S. Karlin et S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences of USA, 1990, 87, pp.2264-2268.

49. U. Hobohm, M. Scharf, R. Schneider et C. Sander, Selection of representative protein data sets, Protein Sciences, 1993, 1, pp.409-417.

50. D. Frishman, K. Heumann, A. Lesk et H-W. Mewes, Comprehensive, comprehensible, distributed and intelligent databases: current status. Bioinformatics, 1998, 14, pp.551-561.

51. M. O. Dayhoff, Atlas of Protein Sequence and Structure, volume 5, 1978, Biochemical Research Foundation, Washington DC.

52. S. E. Phillips, Structure and refinement of oxymyoglobin at 1.6Å, Journal of Molecular Biology, 1980, 142, pp.531.

53. K. Volz et P. Matsumura, Crystal structure of Escherichia Coli CheY refined at 1.7Å resolution, Journal of Biological Chemistry, 1991, 266, pp.15511.

54. H. Eklund, M. Ingelman, B. O. Soderberg, T. Uhlin, P. Nordlun, M. Nikkola, U. Sonnerstam, T. Joelson et K. Petratos. Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. Journal of Molecular Biology, 1992, 228, pp.596.

55. J. R. Gunn. Trip User's Guide version 0.3, Université de Montréal, 1996, p.13.

56. J. U. Bowie, R. Lüthy et D. Eisenberg, A Method to Identify Protein Sequences That Fold into a Known Three Dimensional Structure, Science, 1991, 253, pp.164-170.

57. R. Lüthy, J. U. Bowie et D. Eisenberg, Assessment of protein models with Three-dimensional profiles, Nature, 1992, 356, pp.83-85.

58. B. Lee et F. M. Richards, The Interpretation of Protein Structures: Estimation of Static Accessibility, Journal of Molecular Biology, 1971, 55, pp.379-400.

59. S. J. Hubbard et J. M. Thornton, (NACCESS), Computer Program, Department of Biochemistry and Molecular Biology, University College London.

60. D. Eisenberg, M. Wesson et M. Yamashita, Interpretation of Protein Folding and Binding with Atomic Solvation Parameters, Chemica Scripta, 1989, 29A, pp.217-221.

61. P-J. L'Heureux, Étude de l'espace dièdre dans un modèle de prédiction des protéines : effet d'une distribution ciblée, Université de Montréal, 1997, p.

62. S. T. Prigge, A. S. Kolhekar, B. A. Eipper et L.M. Amzel, Amidation of bioactive peptides: the structure of peptidylglycine α-hydroxylating monooxygenese Science, 1997, 278, pp.1300.

63. J. R. Gunn. Trip User's Guide version 0.3, Université de Montréal, 1996, p.23.