MASTER EN SCIENCES BIOLOGIQUES DES ORGANISMES ET ÉCOLOGIE

Elaboration d'une stratégie d'alignement fiable d'une séquence de protéine de structure connue et d'une séquence de faible homologie, en vue de sa modélisation

Léonard, Nadia

Award date:
2000

FACULTES UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR

Faculté des Sciences

Elaboration d'une stratégie d'alignement fiable d'une séquence de protéine de structure connue et d'une séquence de faible homologie, en vue de sa modélisation.

Mémoire présenté pour l'obtention du grade de
licencié en Sciences biologiques

Nadia LEONARD
Juin 2000

Facultés Universitaires Notre-Dame de la Paix
FACULTE DES SCIENCES
Secrétariat du Département de Biologie
Rue de Bruxelles 61 - 5000 NAMUR
Téléphone: + 32(0)81.72.44.18 - Téléfax: + 32(0)81.72.44.20
E-mail: joelle.jonet@fundp.ac.be - http://www.fundp.ac.be/fundp.html

Elaboration d'une stratégie d'alignement fiable d'une séquence de protéine de structure connue et d'une séquence de faible homologie, en vue de sa modélisation

LEONARD Nadia

Résumé

Le but de ce travail était d'élaborer une stratégie fiable d'alignement pairé pour la modélisation de protéines de faible homologie avec leur patron (*template*). Dans ce but, nous avons consacré nos efforts à optimiser l'alignement séquence cible-template fourni au programme de modélisation. Pour ce faire, nous avons modélisé 9 protéines en utilisant trois types d'alignement : i) un alignement construit sur base de huit alignements pairés provenant de 4 programmes d'alignement, ii) un alignement construit sur base de 12 alignements pairés et de l'alignement pairé fourni par PSI-BLAST, iii) uniquement l'alignement pairé fourni par PSI-BLAST.
Les modèles obtenus ont été comparés aux structures cristallographiques. Ce travail montre que notre méthodologie visant à optimiser l'alignement initial améliore le processus modélisation.

Mémoire de licence en Sciences biologiques
Juin 2000
Promoteur: E. Depiereux

Remerciements

Au terme de cette longue nuit , le soleil se lève, le labo tente de rester éveillé, et sa seule certitude est que l'incroyable est arrivé, ... Plus que quelques heures et les dés seront jetés... Plus que quelques jours et nous partirons vers d'autres horizons, tristes de quitter cet univers qui, en quelques mois, nous était devenu familier...ou, qui sait, avides d'y demeurer pour des années...

Je voudrais tout d'abord remercier le Professeur Eric Depiereux pour m'avoir accueillie dans son laboratoire.

Ensuite, je désirerais témoigner toute ma reconnaissance au professeur Xavier De Bolle pour avoir toujours été là quand il le fallait. Merci, Xa, pour tes encouragements et pour ta bonne humeur.

Mais non, Cri, je ne t'ai pas oublié ! Comment ne pas te remercier pour avoir pu supporter mes sautes d'humeur et mes 36000 questions ? Merci de m'avoir fait découvrir l'étrange univers de la modélisation ... Et quel esprit logique ! Comme tu dis, on l'a ou on ne l'a pas...

Merci, Katalin, le moulin à paroles de BMS, pour m'avoir toujours remonté le moral quand il le fallait même quand je m'inquiétais pour des bêtises...

Nathalie et Jean-Marc, un merci tout particulier pour l'aide précieuse que vous m'avez apportée ces derniers jours.

Merci, Etienne, pour tes bêtises, mais bon sang, arrête un peu de frimer avec tes 3 ordinateurs ! ! !

Je voudrais remercier Cindy et Olivier pour leur bonne humeur et leurs mails qui m'ont bien fait rigoler.

Sandrine, comment ne pas te remercier, pour nos longues conversations qui n'en finissaient pas et pour avoir toujours eu les mots qu'il fallait ? On devrait te nommer Sage du labo !

Valérie, comment fais-tu pour être aussi excitée ? On devrait t'appeler la *queen* du labo. En tout cas, j'espère qu'on se reverra autre part que sur la planète Mercury.

Merci également à tous les membres de l'URBM, pour l'ambiance conviviale qu'ils font régner dans le labo. Merci aux autres mémorants de l'URBM : Amélie, Chantal, Marie, Jacques, Sandrine, Valérie, Lionel, Philippe et Benoît.

Enfin, je tiens à remercier mes parents et mes amis, qui m'ont encourgée durant ces années mémorables .

TABLE DES MATIERES

ABREVIATIONS

BLOSUM	BLOcks Substitution Matrix
CASP	Critical Assessment of techniques for protein Structure Prediction
CMH	Complexe Majeur d'Histocompatibilité
DSSP	Dictionary of Secondary Structure Protein
HLH	Helix Loop Helix
HTH	Helix Turn Helix
HSP	Heat Shock Proteins
MB	Mega Bytes
MSI	Molecular Simulation Inc.
MHz	Mega Hertz
MIPS	Multy initial processors
NCBI	National Center for Biotechnology Information
ORF	Open reading Frame
pSCR	predicted Structurally Conserved Regions
PAM	Point Accepted Mutation
PIR	Protein Indentification Ressource
PDB	Protein Data Bank
RAM	Random Accessible Memory
RMSD	Root Means Square
RMSD	Root Means Square Distance
RMN	Résonnance Magnétique Nucléaire
SCR	Structurally Conserved Regions
SS	Structure Secondaire
TCP/IP	Transmission Control Protocol/Internet Protocol
VR	Variables Regions
WWW	World Wide Web

AVANT-PROPOS

A l'aube du troisième millénaire, la biologie moléculaire vit une véritable révolution. En effet, le séquençage à grande échelle de toute une série de génomes d'organismes tant procaryotes qu'eucaryotes génère une quantité d'information qui croît de jour en jour. C'est pour faire face à cet afflux d'informations que s'est développée une nouvelle discipline alliant biologie et informatique : la bioinformatique.

En génomique et protéomique, la bioinformatique a, dans un premier temps, été appliquée à l'analyse de données fournies par les séquences nucléotidiques et protéiques. D'autre part, la détermination de la structure tridimensionnelle par diffraction des rayon X ou par RMN a permis l'élaboration de nouvelles techniques de prédiction de la structure des protéines. Celles-ci se basent, notamment, sur la similarité de structure entre deux séquences homologues ayant un pourcentage d'identité suffisant (modélisation par homologie).

Il est néanmoins important de garder à l'esprit qu'il existe un réel fossé quantitatif entre les informations structurales (environ 12000 structures protéiques décrites à ce jour, issue 92 de PDB) et le nombre de séquences disponibles dans les banques de données (de l'ordre du demi million). Ce déficit de structures ne fait que s'aggraver depuis l'explosion des différents programmes de séquençage de génomes. Ceci conduit les biologistes à explorer de nouvelles stratégies pour convertir au plus vite les données des séquences protéiques en données structurales, afin d'en tirer des informations assez fiables que pour émettre ou encore étayer des hypothèses plausibles quant à la fonction de ces protéines.

Notons que la bioinformatique recouvre de nombreux autres domaines qui ne seront pas traités ici. Citons néanmoins :

- la difficulté que pose la gestion des banques de données, qui requiert l'expertise de l'utilisateur. D'une part, la redondance que l'on rencontre dans des banques telles que GenBank (banque nucléotidique) résulte de l'archivage de données provenant de différentes sources. D'autre part, la fiabilité des données structurales est assez relative : les

structures déterminées expérimentalement sont plus fiables que des données obtenues par modélisation).

- la mise au point de logiciels spécifiques pour l'analyse automatique des séquences de génomes entiers (détection d'ORF, de sites de régulation,...).

Enfin, signalons que les banques de données ainsi que de nombreux programmes bioinformatiques sont disponibles via Internet, ce qui demeure un atout non négligeable de cette science.

CHAPITRE I: LES DIFFERENTS NIVEAUX DE STRUCTURE DES PROTEINES.

Avant d'aborder les méthodes de prédiction des structures protéiques, il paraît utile, dans un premier temps, de se remémorer différents niveaux de structure des protéines.

1. LES STRUCTURES SECONDAIRES.

1.1. LES STRUCTURES SECONDAIRES : DEFINITION.

Lorsqu'un polypeptide s'étend, on constate que les interactions entre résidus ne sont pas aléatoires mais qu'elles tendent à favoriser la création de structures plus ou moins régulières et communes à bon nombre de protéines. On les appelle structures secondaires par opposition aux structures primaires, qui désignent uniquement la séquence en acides aminés d'une protéine .

Avant de passer en revue les différents types de structures secondaires, nous allons nous intéresser au squelette de la protéine qui déterminera sa structuration..

1.2. LE SQUELETTE PROTEIQUE.

L'enchaînement des chaînes principales des acides aminés définit le squelette (*backbone*) de la protéine. La structuration de ces aminoacides n'est pas sans contraintes et ce sont précisément celles-ci que nous allons détailler dans les points qui suivent.

1.2.1. La liaison peptidique.

Comme mentionné plus tôt, les chaînes principales sont reliées entre elles par des liens covalents communément appelés liaisons peptidiques, formés par condensation des fonctions amine et carboxyle respectivement de l'un et de l'autre résidu engagés dans la liaison. En conséquence, seules les extrémités C et N terminales des protéines gardent une de ces fonctions libre.

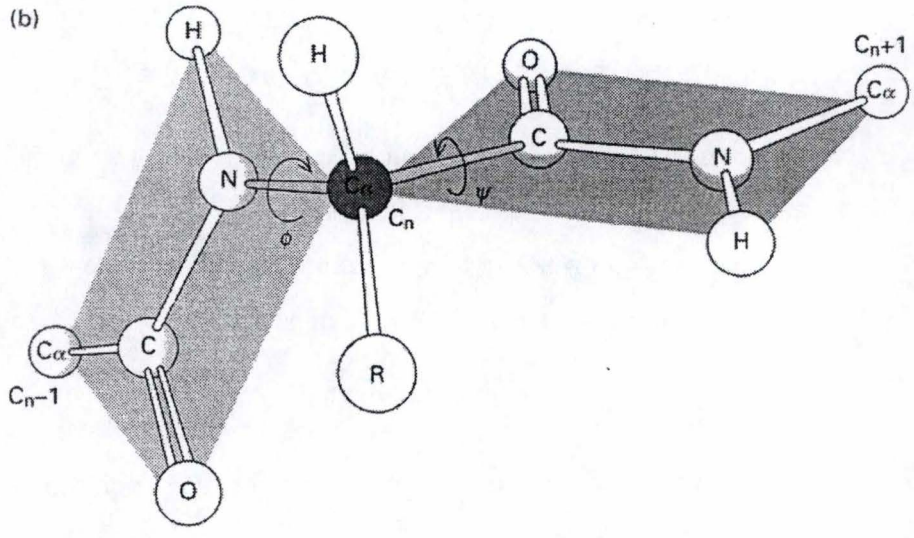


Figure 1. Localisation des différents angles de torsion le long de la chaîne peptidique.

La liaison peptidique présente un caractère double-partiel (voir figure 1) qui peut se résumer comme étant l'apparition d'un équilibre entre les fonctions amide et imine, équilibre dû à la délocalisation de la charge de l'azote sur l'oxygène.

Par conséquent, les atomes qui se trouvent de part et d'autre de la liaison peptidique se situent en général dans le même plan. Ceci constitue déjà une contrainte pour le repliement ultérieur de la protéine.

1.2.2. Les angles de torsion

Lors de l'assemblage d'un polypeptide, le principe d'émergence se vérifie : les propriétés que possèdent ce polypeptide sont radicalement plus complexes que celles d'un de ses composants pris isolément. Ces nouvelles propriétés sont le résultat, non seulement, des interactions entre chaînes latérales (interactions coulombiennes, de Van Der Waals, ponts H, ...) mais aussi des angles de torsion qui en sont, somme toute, la conséquence.

En effet, ces angles, définis entre certains groupes d'atomes qui exercent une rotation autour de liaisons covalentes, confèrent aux résidus une position qui sied mieux à l'équilibre entre les différentes forces en présence de manière à doter la protéine d'une conformation stable.

Décrivons brièvement les trois types d'angles de torsion qui caractérisent le squelette d'une protéine : (voir figure 1)

- **l'angle ω** se situe au niveau de la liaison peptidique et vaut presque toujours 180° , ce qui entraîne que les C_α des deux résidus impliqués dans la liaison de même que leurs prolongements adoptent le plus souvent une configuration dite en *trans*, où les chaînes latérales interagissent le moins possible. On observe parfois des angles ω de 0° où les résidus sont en *cis* mais, en général, le rapport *trans/cis* est de plus de mille si on analyse un grand nombre de structures de protéines. Notons, toutefois, que la proline fait exception : on n'y observe que quatre fois plus de configurations *trans* que de configurations *cis*.

- **l'angle ψ** correspond à la liaison $C\text{-}C_\alpha$. Il ne peut prendre que certaines valeurs, qui remplissent les conditions les plus favorables d'un point de vue stérique.

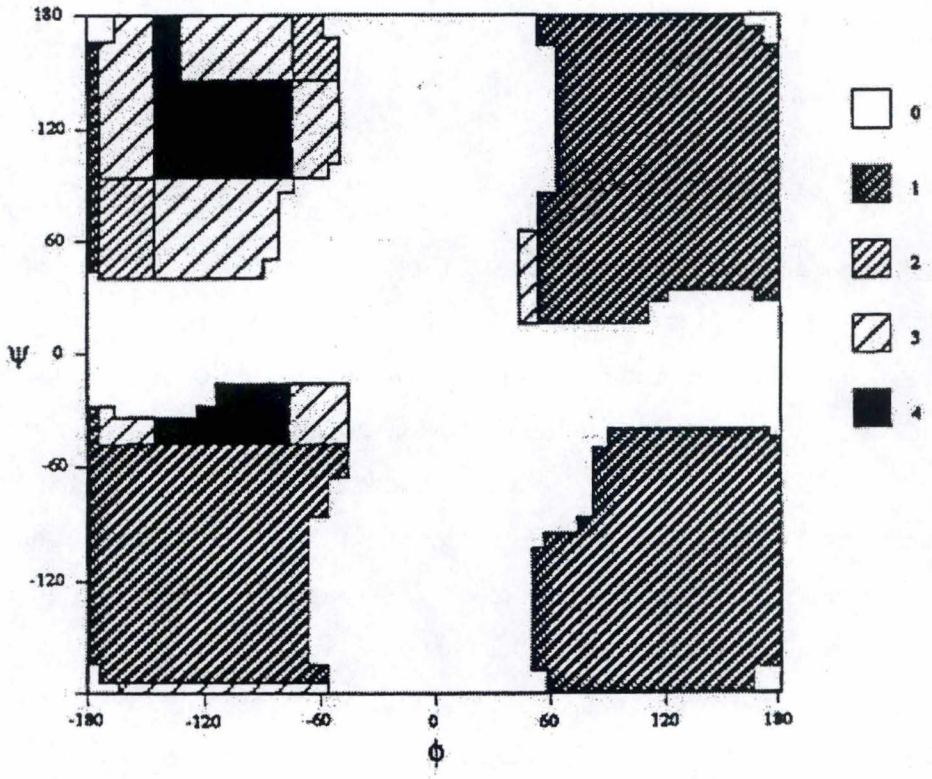


Figure 2. Graphe de Ramachandran indiquant la répartition théorique des angles de torsion ϕ et ψ pour les acides aminé. Gly occupe les régions de 1 à 4 – Ala de 2 à 4 – Val et Ile ne se retrouvent que dans la région 4. Les autres résidus à grande chaînes se situent tous dans les régions 3 et 4.ts angles de torsion le long de la chaîne peptidique.

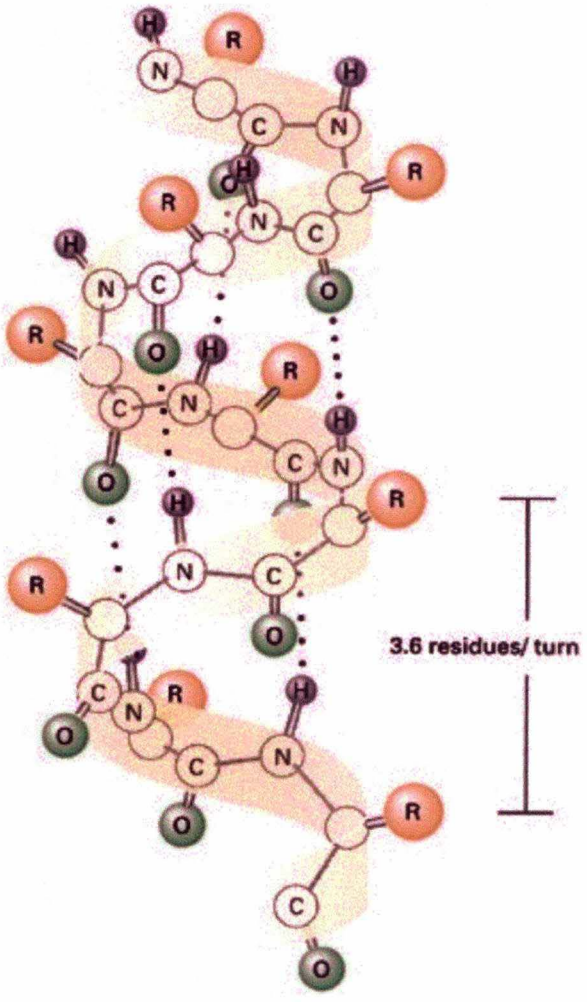


Figure 3. Représentation shématique de l'hélice α ($3,6_{13}$) classique.

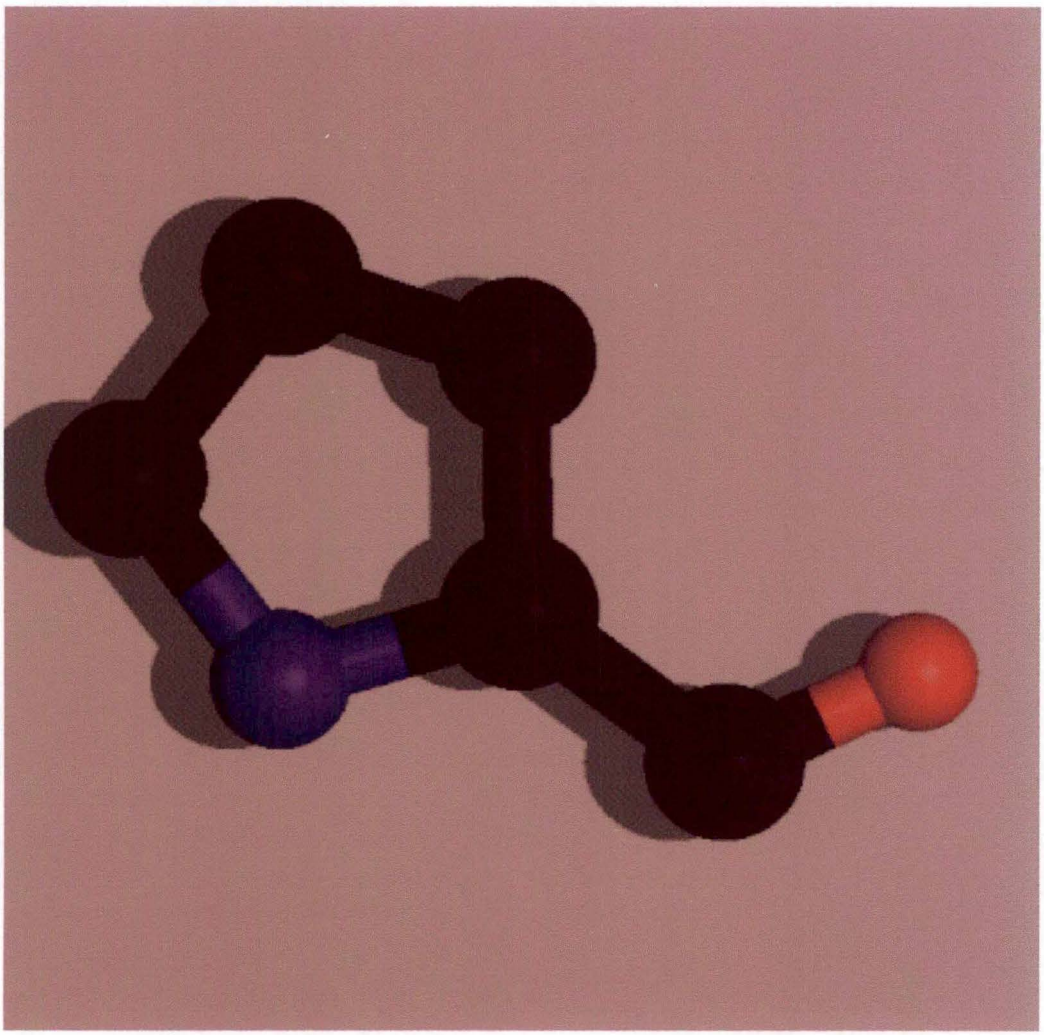


Figure 4. Representation shematique d une proline

- **l'angle φ** s'établit entre le C_α et le N. Pour les mêmes raisons que ψ, il ne prend pas toutes les valeurs possibles.

On a l'habitude de regrouper et de visualiser les valeurs des angles φ et ψ dans un diagramme appelé **diagramme de Ramachandran** (voir figure 2). Remarquons :

- que les valeurs prises par ces angles dépendent du type de structure secondaire auquel on a affaire ;

- que la glycine peut prendre beaucoup plus de valeurs d'angles φ et ψ de par la petitesse de sa chaîne latérale.

Outre les angles de torsion qui caractérisent le squelette, il ne faut pas oublier que des angles de torsion se forment également dans les chaînes latérales et sont désignés par le symbole $\chi_{(j)}$ où l'indice (j) indique la position de la liaison par rapport à la chaîne principale. Ces angles prennent aussi un nombre limité de valeurs.

1.3. CLASSIFICATION ET DESCRIPTION DES STRUCTURES SECONDAIRES.

On classe généralement les structures secondaires en conformations régulières et irrégulières.

1.3.1. Conformations régulières.

On distingue essentiellement trois grands types de conformations régulières : les hélices α,les plans β et les autres conformations régulières.

1.3.1.1. L'hélice α

Caractéristiques générales (voir figure 3)

Ce type de structure est de loin le plus connu. Comme son nom l'indique , l'hélice α est formée d'un squelette enroulé en hélice.

En réalité, il existe des hélices gauches et droites mais de par la configuration L des acides aminés, seule l'hélice α est représentée dans les protéines, pour des raisons de stabilité.

Dans une hélice α, les résidus s'enroulent de manière telle que tous les C=O situés de part et d'autre des liaisons peptidiques soient tournés d'un côté alors que tous les N-H se trouvant à hauteur de ces mêmes liaisons sont orientés vers le côté opposé pour obtenir une disposition colinéaire des atomes de N, H et O. Ceci a pour conséquence la formation de liaisons par pont hydrogène entre l'oxygène du C=O et l'azote du N-H situé trois résidus plus loin. Ces ponts H sont parallèles entre eux et contribuent fortement à stabiliser la conformation de l'hélice.

Chaque tour de spire contient en moyenne 3.6 résidus, ce qui représente 0.54 nm sur l'axe de l'hélice. On appelle cette distance pas de l'hélice.

Composition en acides aminés

La composition en acides aminés (de 10 à 15 par hélice) est biaisée. On constate, en effet, que certains résidus s'y retrouvent plus fréquemment que d'autres. Par exemple, l'alanine et la leucine y seront bien représentés alors que la glycine (étant donné le nombre élevé de valeurs d'angles φ et ψ qu'elle peut prendre), ou encore les résidus chargés (qui ont tendance à se repousser), ne s'y trouvent que très rarement car ils déstabiliseraient l'hélice. La proline, quant à elle, se localisera rarement à l'intérieur d'une hélice α car elle impose des angles qui déformeraient cette dernière et sa structure la rend incapable de contribuer à la formation de ponts hydrogènes (voir figure 4). Par contre, on la détectera très souvent en début d'hélice probablement car elle favorise son repliement initial.

Repliement de l'hélice α

La formation d'une hélice α à partir d'une conformation de départ appelée *random coil* se déroule en deux étapes : la nucléation et le *zippering*.

La nucléation est l'étape limitante en terme de temps et d'énergie. Il s'agit de la formation du premier pont H et donc, du premier tour de spire. Cette étape est favorisée par la présence de proline.

Le *zippering* (fermeture éclair) est beaucoup plus rapide et termine le repliement de l'hélice α, à l'image d'une tirette qui se ferme facilement après en avoir emboîté les deux bouts.

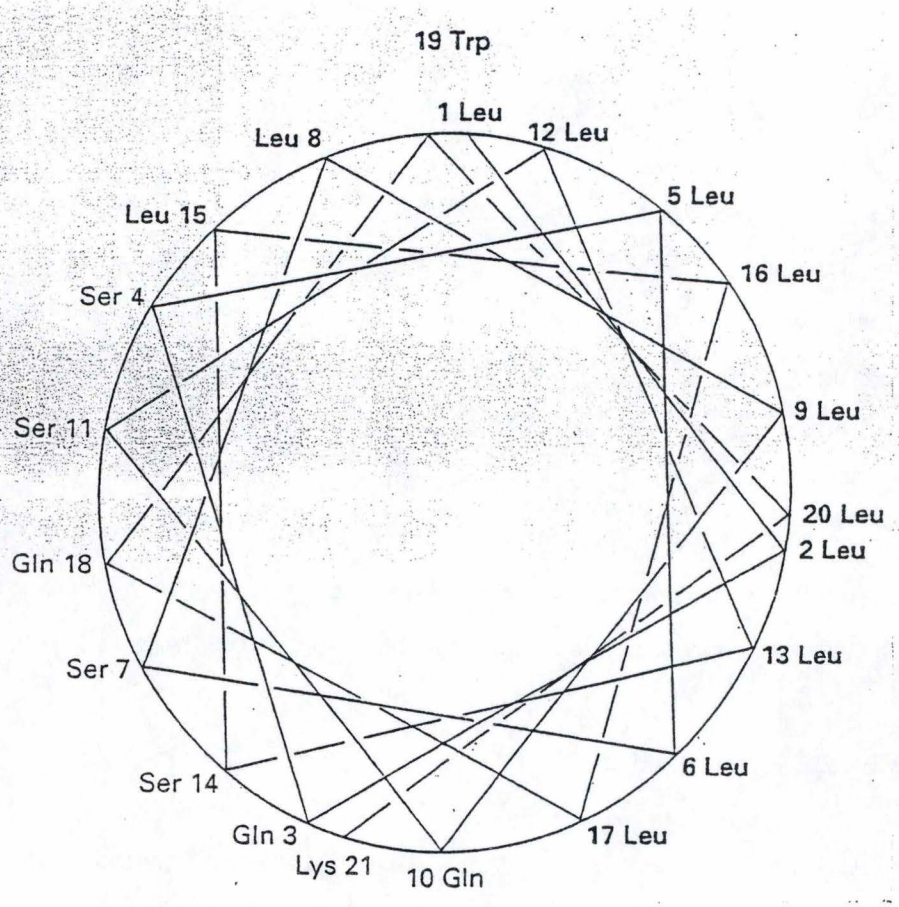


Figure 5. Représentation d'une roue hélicoïdale d'une hélice α , où les positions des chaînes latérales sont montrées en projection sous l'axe de l'hélice

Disposition des résidus dans une hélice α

On observe, dans la plupart des hélices α (en l'occurrence les hélices amphiphiles), que tous les résidus hydrophobes se tournent vers l'intérieur de telle sorte qu'ils y forment une coque hydrophobe alors que les résidus hydrophiles se concentrent à l'extérieur de l'hélice. Notons cependant que les résidus seront disposés en fonction de leur environnement et du rôle que joue la protéine. Par exemple, dans les hélices transmembranaires, la plupart des résidus en contact avec la membrane seront hydrophobes.

On peut visualiser ce phénomène sur une représentation en roue hélicoïdale (*helical wheel*, voir figure 5).

Polarité de l' hélice α

L'orientation de toute la série de C=O et N-H susmentionnée est telle que l'hélice α apparaît comme étant un dipôle où prennent naissance des charges δ^+ aux deux extrémités de l'hélice.

Cette polarité s'avère être d'une importance capitale notamment pour la fixation d'un groupement phosphate dans de nombreuses enzymes au niveau de l'extrémité δ^+. D'autres protéines comme les protéines liant le sulfate (*sulfate binding proteins*) stabilisent un groupement sulfate au moyen de trois hélices orientées de manière adéquate. Une protéine périplasmique impliquée dans le transport actif du sulfate chez *Salmonella typhimurium* en est exemple (Pflugrath and Quiocho 1988) .

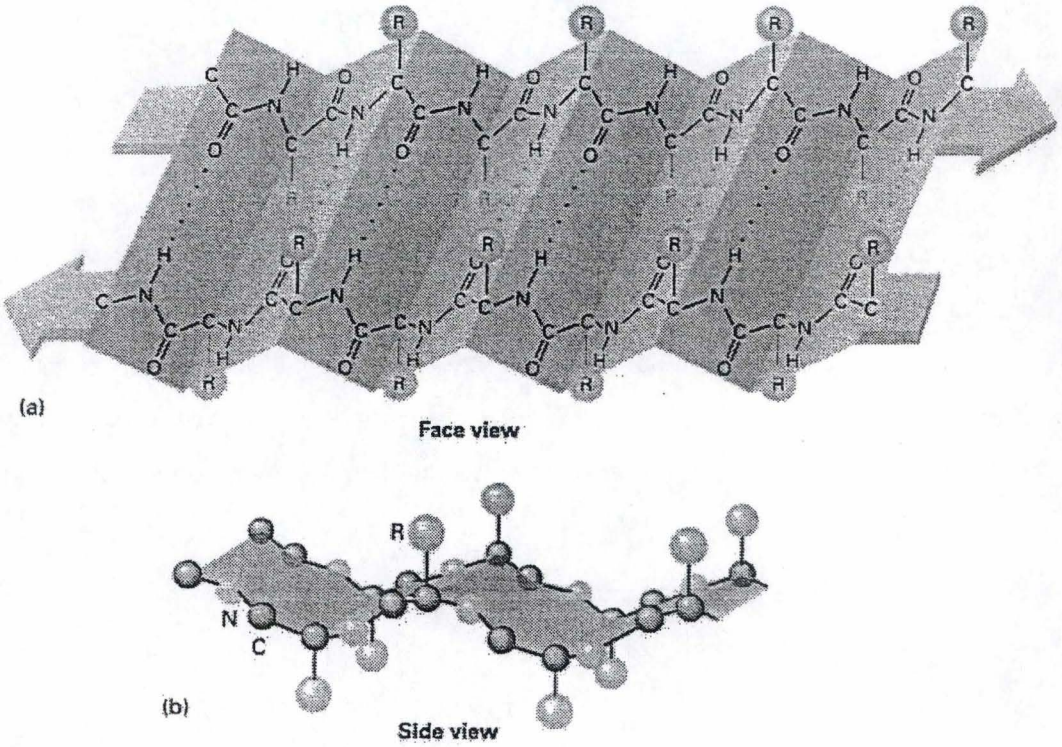
Angles φ et ψ

Les angles φ et ψ prennent respectivement les valeurs de –60 et –50° dans l'hélice α

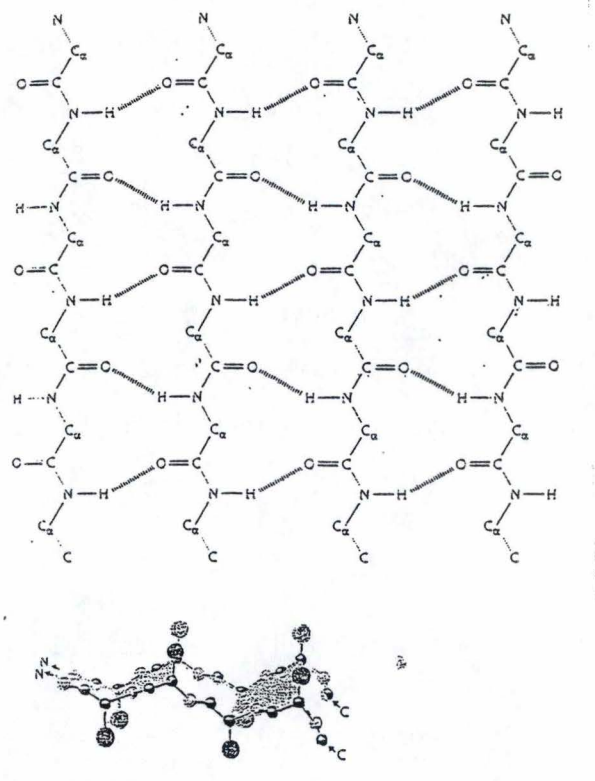
1.3.1.2. Les plans β

Caractéristiques générales

Autre type de structure secondaire que l'on rencontre dans les protéines, le plan ou feuillet β est le résultat de l'assemblage de brins β, chaînes polypeptidiques contenant en



(A)



(B)

Figure 6. Représentation schématique d'un feuillet β antiparallèle (A) et d'un feuillet β parallèle (B).

moyenne six résidus. D'aspect légèrement plissé , il recouvre une plus grande surface que l' hélice α pour un même nombre de résidus (voir figure 6).

Il existe deux types de plan β : ils sont parallèles ou antiparallèles selon que les brins qui les constituent soient orientés dans le même sens ou dans des sens alternativement opposés. Des plans mixés, mélanges de ces deux types sont parfois rencontrés.

Les plans β antiparallèles sont stabilisés par des ponts H parallèles entre eux et ceci conduit à une meilleure stabilité conformationnelle par rapport aux plans β parallèles.

Composition en acides aminés

Comme pour l'hélice α, les fréquences des acides aminés ne se révèlent pas toutes identiques dans les plans β.

Les résidus chargés et de grande taille y sont rares mais la proline est bien représentée malgré son effet disloquant sur les feuillets.

Angles φ et ψ

Les angles φ et ψ prennent respectivement les valeurs moyennes de –120 et +120°.

Disposition des résidus.

Comme pour l'hélice α, la disposition des chaînes latérales des résidus sera fonction de leur environnement hydrophile ou hydrophobe.

1.3.2. Les autres conformations régulières.

1.3.2.1. L'hélice 3_{10}

Cette hélice droite est beaucoup plus étroite que l'hélice α (trois résidus par tour) et possède de valeurs d' angles φ et ψ de –60 et –30°. Lorsqu'elle est présente, on retrouve fréquemment ce type de structure aux extrémités C terminales des hélices α mais ces structures ne sont jamais très longues.

1.3.2.2. Les coudes

Les coudes ou *turns* sont de courtes structures secondaires en U stabilisées par un pont H entre les résidus n et n+1 dont le rôle est de connecter les structures secondaires entre elles pour permettre un repliement optimal de la protéine.

Formés de quatre résidus, les β *turns* forment un pont H entre le premier et le troisième résidu et contiennent très souvent de la proline ou de la glycine dont les angles de torsion permettent, pour la première, de relier deux brins β antiparallèles et, pour la seconde, d'induire un changement de direction de la chaîne polypeptidique.

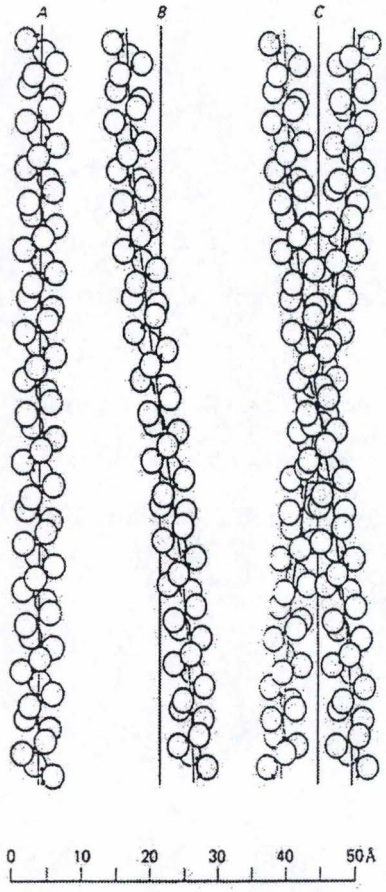
Semblables aux β *turns*, les γ *turns* ne sont formés que par trois résidus.

1.3.3. Conformations irrégulières.

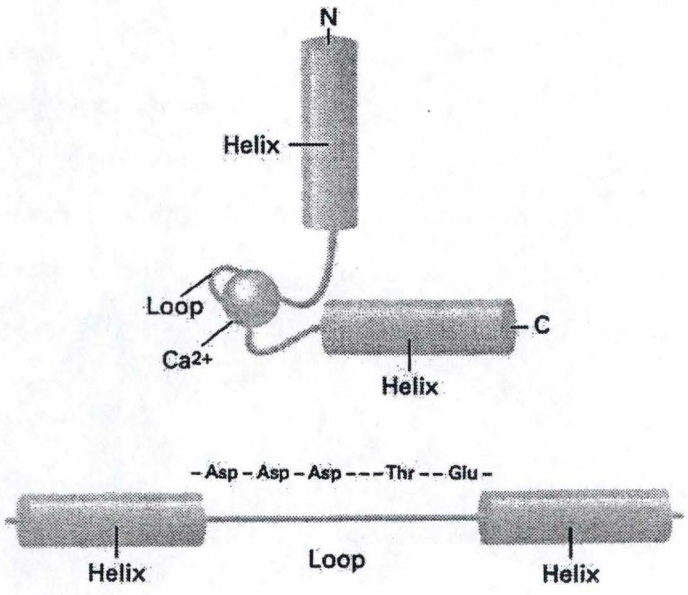
Les conformations irrégulières sont des fragments polypeptidiques de longueur variable en forme de boucles ou *loop*s. Ces boucles ont pour but de relier les différents éléments de structure et surtout de les maintenir au centre de la protéine.

Contrairement aux plans β et aux hélices α, les boucles contiennent, de par leur situation en surface de la protéine, un nombre assez important de résidus hydrophiles et chargés, ce qui leur permet d'interagir avec les solvants.

Enfin, il est important de signaler dès maintenant que les régions qui délimitent ces boucles sont, en général , peu ou pas conservées d'un point de vue évolutif. Toutefois, dans certains cas, les boucles auraient un rôle fonctionnel (phosphorylation, ...) et seraient, dès lors, invariables (Lodish 1997). Par exemple, les structures des régions hypervariables des immunoglobulines sont conservées chez tous les vertébrés (Barre, Greenberg *et al.* 1994).



a) Représentation schématique d'un coiled coil α–helix



b) Représentation schématique d'une hélice α - loop - helice α

2. SUPERSTRUCTURES SECONDAIRES

On applique à un arrangement particulier de structures secondaires fréquemment observées dans les protéines le terme de superstructure secondaire ou motifs.

Dans cette section, nous donnerons un bref aperçu des superstructures secondaires les plus courantes. Celles-ci sont illustrées dans la figure 7.

2.1. LES COILED COIL/ α HELIX

Dans ces superstructures secondaires, deux hélices α s'enroulent l'une sur l'autre de façon à obtenir une super hélice gauche.

2.2. HELICE /BOUCLE/HELICE (HLH) ET HELICE/TURN/HELICE 5HTH)

Ces conformations, où deux hélices sont reliées par une boucle ou par un *turn*, sont souvent observées dans des protéines fixant le calcium ou liant l'ADN (facteurs de transcription,régulateurs,...).Les motifs HTH semblent les plus fréquents dans les protéines liant l'ADN bien que des motifs HLH jouent le même rôle.

2.3. CLEF GRECQUE

Ce motif résulte de la juxtaposition de brins β antiparallèles liés par des boucles. On les retrouve dans les plans β antiparallèles.

2.4. MOTIFS βαβ ET ββ

Dans le motif βαβ , deux brins β sont reliés par une chaîne irrégulière ou une hélice α.

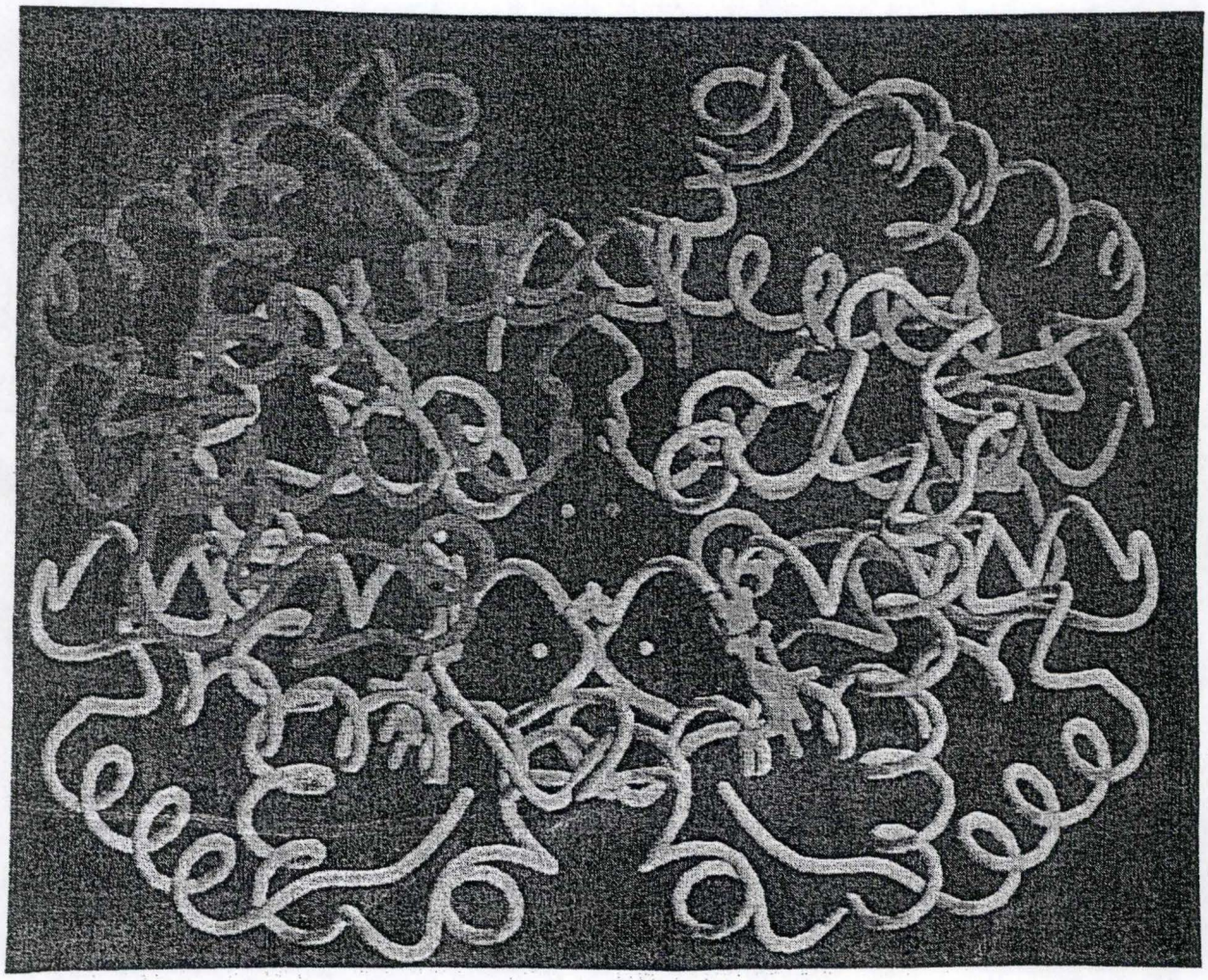
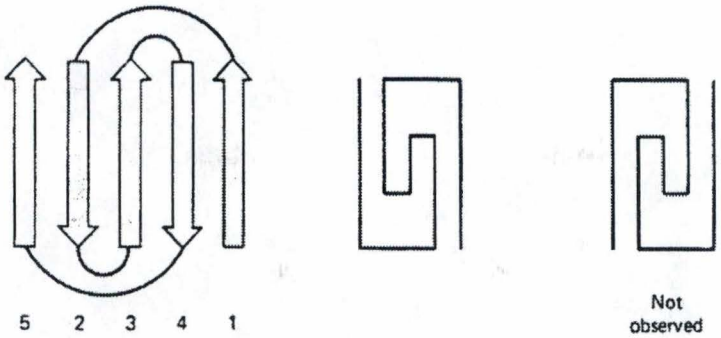
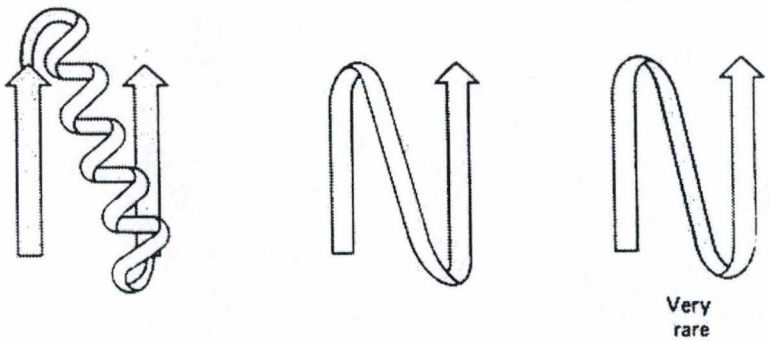


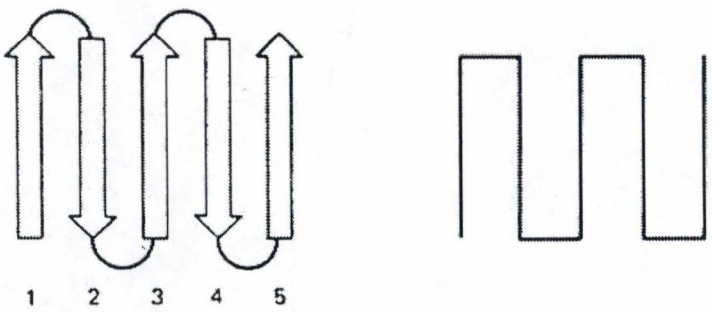
Figure 8. Modèle squelettique de l'hémoglobine tétramérique.



c) Représentation schématique du motif clef grecque



d) Représentation schématique du motif βαβ



e) Représentation schématique du motif ββ

Figure 7. a), b), c), d), e) . Représentations shématiques de diverses superstructures courantes.

Le motif ββ,quant à lui, consiste en un arrangement de deux brins β antiparallèles connectés par une boucle.

3. STRUCTURE TERTIAIRE

Dans une protéine, les structures et superstructures secondaires s'organisent en une structure tertiaire composée de un ou plusieurs domaine(s). Chaque domaine est responsable d'une fonction particulière dans cette protéine.

Le repliement en structure tertiaire a pour effet de rapprocher spatialement des résidus fortement éloignés au niveau de la séquence, ce qui rend possible leur positionnement très précis, qualité requise pour assurer une activité optimale de la protéine.

Par ailleurs, les contacts entre domaines créent de nouvelles interactions entre résidus, ce qui a pour conséquence, notamment pour les enzymes, leur fixation plus aisée à leur(s) substrat(s), coenzyme(s), ... Nous pouvons illustrer ceci en prenant comme exemple les déshydrogénases. Dans ces protéines, le rapprochement de deux zones hydrophiles appartenant à deux domaines identiques favorise la création d'un milieu plus favorable à la fixation du NAD^+, qui est essentiellement une molécule hydrophile (X. De Bolle, communication personnelle).

4. STRUCTURE QUATERNAIRE

De nombreuses protéines ne deviennent fonctionnelles que lorsque plusieurs chaînes polypeptidiques s'assemblent en une conformation que l'on définit comme étant une structure quaternaire.

On parle alors de protéines oligo- ou multimériques, dont les sous-unités (monomères) sont soit identiques soit différentes. Dans ce second cas, l'association de sous-unités de structures différentes permettra à la protéine de cumuler des fonctions qui pourront être complémentaires. L'hémoglobine (voir figure 8) en est un exemple.

CHAPITRE II : PREDICTION DES STRUCTURES TRIDIMENSIONNELLES DE PROTEINES

1. INTRODUCTION

La connaissance de la structure tridimensionnelle d'une protéine est un outil qui se révèle d'une importance capitale pour mieux comprendre sa fonction, ses interactions avec d'autres protéines dans l'organisme que l'on étudie, ainsi que les effets phénotypiques de ses mutations (Tramontano 1998). Elle permet entre autres d'identifier les résidus impliqués dans la catalyse, la liaison ou la stabilité structurale, d'examiner les interactions protéines-protéines et de corréler les mutations génotypiques et le phénotype (Sauder, Arthur *et al.* 2000).

Or, comme nous l'avions déjà précisé dans l'avant-propos, le nombre de structures protéiques connues est faible par rapport à la quantité de séquences disponibles dans les banques de données, quantité pour laquelle on observe une croissance exponentielle (Attwood and Parry-Smith 1999).

Les techniques expérimentales de détermination des structures, à savoir la diffraction des rayons X et la RMN, ne peuvent être à elles seules utilisées à grande échelle, et ce, parce que:

- d'une part, la diffraction des rayons X, bien qu'étant la technique la plus précise, exige un cristal de bonne qualité, c'est-à-dire grand, individuel, bien ordonné et contenant des protéines intactes. Ceci exclut presque toujours des protéines instables comme par exemple certaines protéases et les protéines insolubles telles les protéines membranaires;
- d'autre part, la RMN ne se prête qu'à des protéines de poids moléculaire inférieur à environ 30 kDa.

C'est pour faire face à cet état de fait que différentes méthodes de prédiction de structures protéiques ont pris leur essor. Ces méthodes ont, en effet, l'avantage d'être

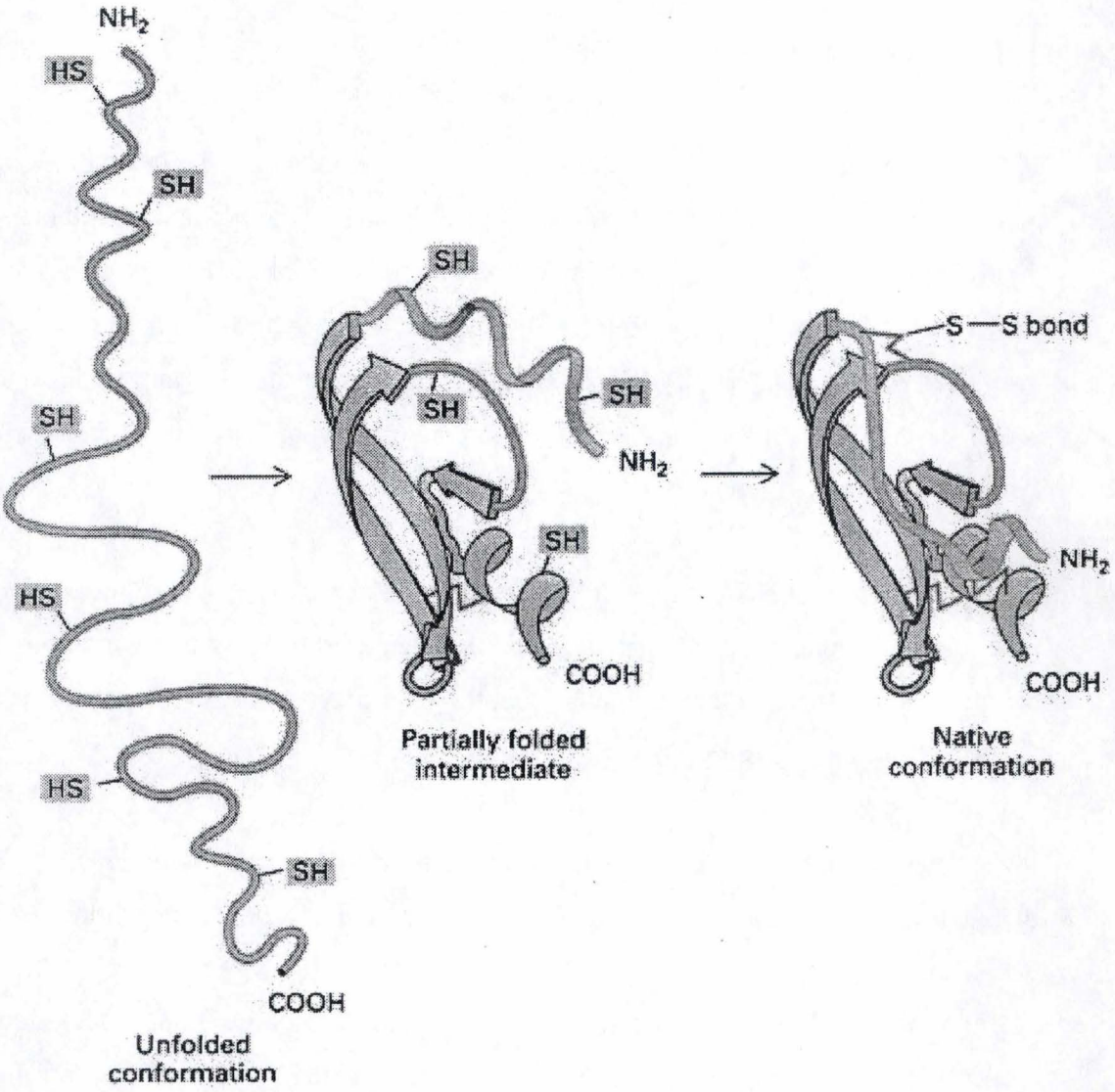


Figure 10. Représentation schématique du repliment d'une protéine

beaucoup plus rapides que la détermination de structures par diffraction des rayons X ou par RMN.

Dans ce chapitre, les caractéristiques essentielles des méthodes de prédiction de structures protéiques seront brièvement exposées. L'accent sera mis sur la modélisation par homologie, qui fera l'objet de ce mémoire.

2. PROBLEME DU REPLIEMENT DES PROTEINES

C'est en 1961 que Anfinsen (Anfinsen 1973) a montré que les ribonucléases étaient capables *in vitro* de se replier après dénaturation. Cette expérience suggérait que toute l'information nécessaire pour qu'une protéine adopte sa conformation native était encodée par sa structure primaire (voir figure 10) (Attwood and Parry-Smith 1999) .

Ce principe est à l'origine des méthodes dites *ab initio* (Sternberg, Bates 1999) qui tentent de prédire la structure de protéines uniquement sur base de l'information contenue dans leur séquence.

Cependant, les règles qui régissent le repliement ou *folding* des protéines sont loin d'être complètement élucidées (Attwood and Parry-Smith 1999). En effet, des expériences similaires à celles de Anfinsen ne se sont vérifiées que pour une faible proportion de protéines. De plus, le repliement de la majorité des protéines qui intéressent les biologistes moléculaires et cellulaires fait intervenir, outre la séquence, d'autres facteurs.

Par exemple, les protéines multimériques et multidomaines ne peuvent pas se replier correctement après dénaturation : des régions hydrophobes ont souvent tendance à s'agréger de manière à atteindre un minimum d'énergie, ce qui provoque la formation de structures ne correspondant plus à la structure native (Serrano, communication personnelle).

Parmi ces facteurs affectant le repliement des protéines, les protéines chaperones (qui font partie des protéines impliquées dans la protection contre les chocs thermiques :HSP, *Heat Shock Proteins*) jouent un rôle important. A l'heure actuelle, les proportions de protéines qui se replient avec ou sans l'aide de chaperones ne sont pas connues (Ellis , 1991). Cependant, les chaperones se retrouvent aussi bien chez les eucaryotes que chez les procaryotes, ce qui montre leur importance. Elles interviennent dans la formation des ponts disulfures nécessaires, par exemple, pour le repliement des molécules de CMH de classe I (Farmery, Allen *et al.* 2000). Elles empêchent aussi l'adoption de structures non fonctionnelles par les protéines notamment lors de leur adressage vers

les compartiments intracellulaires (par exemple HSP70 dans les mitochondries). Elles favorisent également le repliement de nombreuses protéines au cours de leur biosynthèse: en se fixant sur la protéine en formation et empêchent la formation accidentelle d'agrégats au cours du processus de repliement (par exemple chez *E. Coli*, le système de chaperones DnaK-DnaJ-GrpE (Diamant, Peres Ben-Zvi *et al.* 2000).

Les cofacteurs (tels l'hème dans les cytochromes) ou les ions (comme le zinc dans les motifs en doigt de zinc liant l'ADN présents dans les facteurs de transcription (Branden and Tooze 1991) influencent également la conformation et/ou la multimérisation des protéines.

D'autres facteurs extérieurs à la séquence sont aussi à prendre en compte : les interactions avec d'autres protéines, les modifications chimiques, le pH, la salinité, l'accessibilité des résidus au solvant, la température...

L'existence de facteurs extérieurs à la séquence intervenant dans le repliement des protéines montrent que le postulat de Anfinsen n'est pas correct pour toute une série de protéines. Dès lors, il est souvent impossible de prédire la structure d'une protéine à partir de sa séquence.

C'est pourquoi de telles méthodes, bien qu'étant en constante évolution (Tramontano 1998) (A. Fiser *et al.*, 1999 unpublished), n'ont pas encore prouvé leur efficacité pour la prédiction de structures protéiques. Nous n'en ferons dès lors pas usage dans le cadre de ce travail.

D'autres méthodes, ont été développées pour contourner ce problème :

- la prédiction de structures secondaires dans les protéines ;
- la modélisation par homologie et par reconnaissance de *fold*.

3. PREDICTION DE STRUCTURES SECONDAIRES

Le but des méthodes de prédiction de structures secondaires est de prédire la localisation de ces structures au sein des protéines. Dans cette optique, la recherche s'est concentrée sur trois types d'approches (Attwood and Parry-Smith 1999):

- les **méthodes statistiques** basées sur le principe que des types différents d'acides aminés ont différentes probabilités de se trouver dans diverses structures secondaires.

  La plus connue, pour des raisons historiques, est celle de Chou et Fasman (Chou and Fasman 1974) qui se base sur une détermination statistique des occurrences préférentielles des 20 acides aminés dans trois états structuraux (hélice

α, plan β et coils). Ces valeurs ont été déterminées à partir des occurrences des résidus dans les structures scondaires de 15 protéines non homologues de structures connues. Mais cette méthode est très peu fiable (<environ 50 % (Kabsch and Sander 1983) car la taille des échantillons utilisés à cet effet était inadéquate. D'autres méthodes, telles la méthode GOR III (Garnier, Osguthorpe *et al.* 1978) (Gibrat, Garnier *et al.* 1987) se basent sur le même principe mais tiennent compte, en outre, des interactions entre les résidus et leur environnement local. Cependant, ces types de méthodes, qui ne reposent que sur l'information d'une seule séquence, ne prédisent précisément les structures secondaires que dans à peu près 60% des cas. Certaines méthodes intégrant l'information fournie par des alignements de plusieurs séquences homologues ont permis d'augmenter la précision des prédictions.

- les **méthodes basées sur des critères physico-chimiques** (hydrophobicité , charge,...) DSC en est un exemple (King 1997).
- les méthodes qui utilisent les caractéristiques de structures connues de protéines homologues pour assigner une structure secondaire à la protéine d'intérêt.

Une de ces méthodes utilise les réseaux neuronaux constitués de différentes unités intégrant leurs propres données et les transformant en réponses qui sont transmises aux autres unités connectées en parallèle. Ces neurones sont entraînés sur un jeu de protéines tests (de structure connue). Le réseau cherche quelles y sont les relations entre les acides aminés et les structures qu'ils peuvent former dans un contexte particulier. Ensuite, à partir de la séquence d'intérêt, il calcule les probabilités de formation d'une structure particulière par un résidu donné. Citons par exemple PHD (Rost, Sander *et al.* 1994) .

PSI PRED est un autre logiciel qui utilise les réseaux neuronaux pour prédire les structures secondaires à partir de l'information fournie par l'alignement multiple de séquences homologues à la séquence d'intérêt proposé par PSI-BLAST (cfr. modélisation par homologie). A l'heure actuelle, il s'agit de la méthode qui donne les meilleurs résultats de prédiction de structure secondaire (environ 77% d'exactitude sur les protéines testées).

4. MODELISATION PAR HOMOLOGIE

Le but de la modélisation par homologie est de construire un modèle tridimensionnel d'une protéine de structure inconnue que l'on appellera protéine cible ou

séquence cible (en anglais ***target*** ou ***query***) basé sur la similarité de séquence avec une ou plusieurs protéine(s) de référence ou *template*(s), dont la structure est connue (A.Fiser *et al.*,1999 , unpublished).

La fonction d'une protéine ne peut pas toujours être inférée lorsqu'on effectue une recherche par similarité de séquences. En effet, bien que l'on observe une étroite relation entre la similarité entre séquences et la conservation de leur structure (Chothia C. 1986), la conservation de la fonction n'est pas toujours vraie. Cependant, de nombreuses caractéristiques fonctionnelles peuvent être prédites par une telle recherche. En effet, on peut distinguer plusieurs niveaux de fonctions : la fonction cellulaire ou physiologique d'une protéine, sa fonction biochimique (par exemple l'activité catalytique d'une enzyme), et l'ensemble des caractéristiques moléculaires (fixation à un substrat, liaison à l'ADN,...) Sous cet éclairage, l'identification de zones conservées entre séquences homologues permet de suggérer certaines caractéristiques fonctionnelles.

Par exemple, si un domaine HTH est conservé entre la séquence d'intérêt et sa famille de protéines homologues, cette protéine se liera très probablement à l'ADN. Par contre, des investigations expérimentales seront nécessaires pour déterminer s'il s'agit d'un inhibiteur ou d'un activateur de la transcription.

On peut alors se demander ce qu'un modèle de la structures tridimensionnelle d'une protéine peut apporter de plus qu'une comparaison de séquence.

Dans certains cas, certaines caractéristiques fonctionnelles ne peuvent être détectées directement par similarité de séquences. En effet, il se peut que certains résidus qui ne semblaient pas conservés au niveau de la séquence le soient au niveau structural.

Par exemple, il est possible qu'un site actif soit très conservé au niveau structural mais certains résidus qui y sont très proches spatialement soient très éloignés dans la séquence d'intérêt, du fait de l'insertion d'un loop alors que dans son template, le loop n'était pas présent.

Par ailleurs, la connaissance de la structure protéiques permet d'infirmer certaines hypothèses soulevées lors d'une recherche par similarité. Elle peut révéler que certains résidus supposés être impliqués dans une fonction ne le sont pas.

Par exemple, ces résidus peuvent êtres exposés au solvant au lieu d'être enfouis. Il se peut aussi qu'ils soient camouflés par une structure que l'on ne pouvait pas détecter par similarité.

De plus, si la fonction de la protéine a déjà été déterminée, sa structure permet de planifier rapidement des expériences de mutagenèse dirigée en ciblant les résidus

impliqués dans le site actif. Ces mutations permettent de montrer si un ou plusieurs résidus sont impliqués dans la fonction. Elles permettent également d'étudier la stabilité d'une protéine en fonction des acides aminés qui ont été substitués. Ceci ouvre à des applications telles que l'ingénierie d'enzymes thermostables, par exemple.

Enfin, si le modèle est très précis, son étude détaillée peut aider à la conception de molécules (médicaments, ...) se fixant spécifiquement à la protéine cible. Il permet également d'étudier les interactions protéines-ligand, protéines-protéines, protéines ADN, ...

La modélisation par homologie s'avère être la méthode de modélisation la plus prometteuse car elle peut s'appliquer avec une précision raisonnable à un nombre dix fois plus élevé de séquences que ne le permettent les méthodes expérimentales (Sanchez and Sali 1997). Il est important de souligner qu'il s'agit, actuellement, de la méthode de modélisation la plus précise (CASP1,2,3).

La condition *sine qua non* pour l'utilisation d'une telle méthode est le partage d'un pourcentage d'identités d'acides aminés significatif entre la séquence d'intérêt et au moins une séquence de structure connue (Tramontano 1998). Si c'est le cas, ces protéines ont de fortes chances d'être homologues, c'est-à-dire qu'elles dérivent d'un ancêtre commun (Attwood and Parry-Smith 1999), et, bien qu'elles aient divergé au cours de l'évolution, elles adopteront très probablement la même conformation générale (Doolittle 1981). Ceci est dû au fait que la structure de protéines appartenant à une même famille est plus conservée que leur séquence (Lesk and Chothia 1980). Par conséquent, si l'on peut détecter une similarité entre la séquence cible et une séquence de structure connue, la première peut être modélisée à partir de la seconde.

On définit comme **orthologues** les protéines homologues qui sont supposées jouer le même rôle dans différents organismes. Elles ont évolué verticalement d'un organisme à l'autre à partir d'un ancêtre commun. Remarquons qu'une protéine présente dans un organisme peut être l'orthologue d'une famille de protéines de même fonction appartenant à un autre organisme (Henikoff, Greene *et al.* 1997).

Les protéines **paralogues,** quant à elles, sont issues de duplications géniques au sein d'un même organisme. Elles n'ont pas nécessairement la même fonction car suite à des mutations, ont pu apparaître des divergences fonctionnelles. Il est intéressant de noter que l'acquisition d'une nouvelle spécificité ou d'une fonction modifiée après duplication génique peut être détectable par

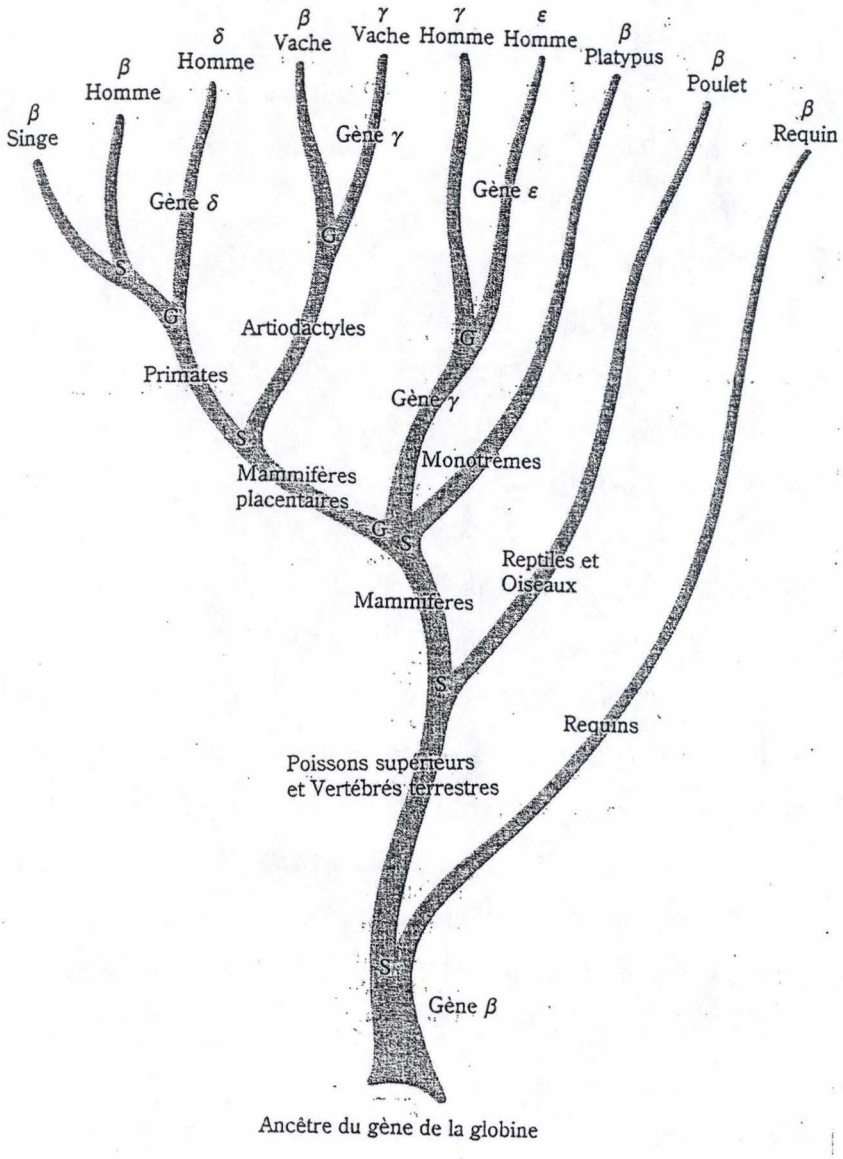


Figure 9. Arbre phylétique de l'évolution des gènes de la globine β. Quelques gènes nouveaux apparaissent par duplication (embranchement G) ; d'autres naissent par mutation après divergences de deux espèces (bifurcation S). (illustration de Patricia Johnson).

comparaison de séquences. Par exemple, les α-globines sont plus reliées entre elles qu'elles ne le sont aux β-globines (voir figure 9) (Henikoff, Greene *et al.* 1997).

Les **protéines chimères** résultent de la duplication et de l'évolution des modules correspondant aux domaines protéiques suivant divers mécanismes de réarrangements géniques, c'est-à-dire des transferts horizontaux (Henikoff, Greene *et al.* 1997).

4.1. RECHERCHE DE SEQUENCES HOMOLOGUES A LA SEQUENCE CIBLE

4.1.1. Banques de données

La toute première étape consiste en une recherche de séquences et de structures dans différentes banques de données afin d'identifier le ou les *template* (s) potentiel(s) et de détecter un maximum de séquences homologues, de façon à obtenir un alignement le plus fiable possible (cfr. point suivant).

A cet effet, de nombreuses banques de données sont disponibles sur le réseau :

- des banques nucléotidiques (telles GENBANK). Ces banques ne seront pas utilisées dans notre cas ;
- des banques de séquences protéiques: par exemple SWISS-PROT et PIR ou des banques non redondantes comme la banque nr du NCBI.
- des banques de structures tridimensionnelles (dont la plus connue est PDB : Protein Data Bank, Brookhaven National Laboratory, Cambridge ; USA) La *Pending List* de PDB est une liste régulièrement renouvelée de protéines dont la structure est connue mais n'est révélée au public qu'après un certain laps de temps.

La prolifération des banques de données sur le réseau pose certains problèmes : pour choisir la banque la plus adéquate, on peut se demander laquelle de celle-ci est la plus précise, celle qui est le plus mise à jour ou encore quel format elle utilise…

Parmi les banques de données protéiques, NRL-3D (la banque de séquences de PDB) a l'avantage d'être directement reliée à l'information structurale mais, de ce fait, limite très fort la recherche par similarité de séquences. L'utilisation d'autres banques de données est donc requise :PIR, par exemple, contient beaucoup d'informations mais ses

annotations sont relativement pauvres. SWISSPROT, quant à elle, fournit d'excellentes annotations mais englobe moins de séquences que PIR.

Les banques composites compilant diverses sources primaires et ayant la propriété de contenir des séquences non redondantes sont très utiles pour la recherche en banque de données. En effet, les séquences identiques ou très fortement similaires (à quelques résidus près) en ont été exclues. Ces banques rendent la recherche plus efficace et plus rapide puisqu'elle restreint le nombre de banques à consulter.

4.1.2. Programmes de recherche en banques de données

Les programmes qui effectuent ces recherches sont également disponibles sur le web, comme par exemple, BLAST (Altschul, 1990), PSI BLAST (Altschul, Madden *et al.* 1997) et FASTA (Pearson 1990). Leur principe général consiste en une comparaison pairée entre la séquence d'intérêt et chaque séquence de la base de données pour y détecter des zones de similarité. Les séquences retenues sont affichées par ordre de similarité. De plus, les programmes cités ci-dessus fournissent un alignement pairé entre la séquence cible et les séquences ainsi trouvées. Le fonctionnement des programmes que nous avons utilisés est exposé ci-dessous.

BLAST est un programme rapide de recherche de séquences similaires à la séquence d'intérêt dans une banque de données.

Son principe est de rechercher, pour chacune des séquences de la banque, des segments (appelés mots ou *words*) similaires à des segments de même taille définis préalablement dans la séquence d'intérêt. A chaque comparaison pairée est attribué un score issu d'une matrice de substitution (voir matériel et méthodes) qui augmente avec la similarité. Ensuite, BLAST procède à l'extension des mots fixés de façon à obtenir un alignement pairé entre la séquence cible et chacune des autres séquences.

L'ancienne version de BLAST présentait la caractéristique de fournir des alignements sans gaps.

Les gaps ou indels sont des espaces qui permettent de décaler les séquences à aligner pour apparier certaines régions n'apparaissant pas directement comme similaires. En effet, au cours de l'évolution, il est probable que deux séquences homologues aient

perdu (délétion) ou acquis (insertion) un ou plusieurs résidus. Le fait de ne pas introduire de gaps pourrait, dès lors, camoufler une réelle similarité.

Cet inconvénient a été levé avec l'introduction de Gapped BLAST.

Les résultats d'une recherche avec BLAST contiennent l'intitulé des séquences qu'il a trouvées par ordre de similarité et une Expected value (E value) qui correspond à l'estimation de la fréquence d'occurrence des séquences de la banque de données qui auraient par hasard le même score que la séquence d'intérêt. Plus elle est faible, plus les séquences sont similaires. En dessous de 0,001, on peut dire que les séquences ont de très fortes chances d'être homologues.

Différentes versions de BLAST existent selon que les séquences soient nucléotidiques (BLASTn), nucléotidiques traduites (BLASTx) ou protéiques (BLASTp). On peut également choisir la banque de données sur laquelle seront réalisées les recherches.

PSI BLAST est une extension de BLAST qui suit une approche hybride incorporant à la fois les éléments des méthodes d'alignements pairé et multiple de séquences.

Dans une première étape, le programme effectue une recherche en base de données (i.e. un simple gapped BLAST). PSI BLAST crée automatiquement un profil spécifique pour chaque acide aminé de la séquence d'intérêt. Après cette première étape, le processus est répété, mais cette fois-ci, on réalise la recherche en base de données en utilisant le profil spécifique. On recommence le processus (par itérations) jusqu'à convergence du nombre de séquences retrouvées.

Un profil est une table de scores spécifique de la position qui y inclut l'information contenue dans un alignement. Il définit quels résidus sont permis pour des positions données :quelles sont les positions conservées ou dégénérées ou quelles positions/régions peuvent tolérer des insertions ou des délétions.

Psi BLAST a l'avantage de trouver plus de séquences homologues à la séquence cible que BLAST (Altschul, Madden *et al.* 1997). En revanche, son désavantage est que lorsqu'une séquence non homologue est prise en compte pour la création du profil, celui-ci est faussé. Cela conduira à l'incorporation d'autres séquences non homologues lors des itérations suivantes.

4.2. ALIGNEMENT DE SEQUENCES.

Cette étape permet d'aligner les zones similaires entre la séquence cible et le *template* dans le but de délimiter les régions prédites comme étant structuralement conservées (pSCR: predicted Structurally conserved regions) (Vinals, De Bolle *et al.* 1995); (de Fays, Tibor *et al.* 1999).

En général, l'alignement fourni par une recherche en base de données n'est pas optimal (Venclovas, Ginalski *et al.* 1999) et n'inclut souvent que des régions de haute similarité entre la séquence d'intérêt et les séquences homologues. Il est donc nécessaire de réaligner le *template* sélectionné à la séquence cible.

Si le pourcentage d'identités entre les deux séquences dépasse approximativement 50% , un alignement pairé, qui ne compare que ces deux séquences, sera suffisant car il ne sera pas trop entaché d'erreurs. Par contre, si ce pourcentage est inférieur à 50 %, et surtout en dessous de 30 %, un alignement multiple sera nécessaire car les erreurs d'alignement pairé ne font que s'accentuer lorsque les séquences alignées sont de moins en moins similaires (Johnson and Overington 1993) .

Un alignement multiple s'opère entre la séquence cible, le ou les *template*(s) et l'ensemble des séquences homologues qu'on lui fournit, de façon à avoir un aperçu de la famille de protéines à laquelle appartient la séquence cible et de rendre l'alignement le plus fiable possible. L'alignement multiple se base sur le fait qu'une similarité de séquences est plus hautement significative si elle est partagée par plusieurs séquences (Depiereux and Feytmans 1992). De fait, les alignements multiples peuvent réduire significativement le nombre d'alignements alternatifs qui pourraient se produire (Venclovas, Ginalski *et al.* 1999) .

Notons que la plupart des erreurs dans la modélisation par homologie dériveront d'un alignement erroné (CASP1,2,3). En effet, on comprend facilement que la qualité de l'alignement influence dramatiquement la fiabilité du modèle (Vinals, De Bolle *et al.* 1995). De fait, la plupart des erreurs sont dues à la position incorrecte d'insertions et de délétions qui ont pour effet de décaler l'alignement. Dès lors, il est possible, surtout lorsque le pourcentage d'identités entre les séquences cible et *template* est faible, d'extraire un consensus de divers programmes d'alignement multiple. Ce consensus sera en général

plus fiable que l'utilisation d'une seule de ces méthodes. (Thompson, Plewniak *et al.* 1999). Il faut également veiller à utiliser le plus d'informations expérimentales sur la famille de la séquence cible et du *template*. La prédiction de structures secondaires et la reconnaissance de fold (voir point suivant) sont des informations précieuses qui peuvent être prises en compte pour l'optimisation de l'alignement final (Kabsch and Sander 1983). Divers programmes d'alignement pairé et multiple existent. Ceux que nous utiliserons seront décrits dans la partie « matériel et méthodes ». Voici un aperçu très général de ceux-ci.

Il existe deux types d'alignements multiples : les alignements locaux et les alignements globaux. En général, les programmes d'alignement global donnent de meilleurs résultats excepté en présence de larges extensions N et C terminales et d'insertions internes (Thompson, Plewniak *et al.* 1999).

Les programmes d'alignement local essaient de localiser des similarités locales entre séquences. Par conséquent, ils restreignent l'alignement aux segments qui sont les plus similaires. Match-Box et Dialign en sont des exemples. L'intérêt des ces méthodes est qu'ils calculent un score de confiance pour chaque position alignée.

En revanche, les programmes d'alignement global alignent les séquences sur toute leur longueur. Parmi ceux-ci, les programmes d'alignement progressif s'appuient sur le fait que des séquences similaires sont liées évolutivement. Ces programmes alignent d'abord les séquences qui leur sont fournies deux à deux en assignant à chaque comparaison un score de similarité basé sur une matrice de scores. Ensuite, les séquences les plus similaires sont regroupées sur base d'un arbre phylogénétique. Ensuite, les autres séquences sont rajoutées progressivement par ordre de similarité de telle sorte qu'un alignement final soit obtenu entre toutes les séquences. Citons par exemple ClustalW (Thompson, Higgins *et al.* 1994), Map (Huang 1994), MULTALIN (Corpet 1988) et PIMA (Smith and Smith 1992). Dans cette catégorie de programmes, ClustalW est un des meilleurs programmes actuels (Thompson, Plewniak *et al.* 1999) ; (Briffeuil, Baudoux *et al.* 1998).

Remarquons que d'autres méthodes d'alignement global existent. Les méthodes qui utilisent des stratégies itératives comme les algorithmes génétiques sont très performantes sauf lorsque l'on introduit une séquence trop divergente par rapport à un groupe de séquences fortement similaires.

Le résultat de cette étape donne donc un alignement pairé entre la séquence cible et la séquence de référence où seront déterminées les pSCR.

4.3. CONSTRUCTION DU MODELE TRIDIMENSIONNEL DE LA PROTEINE CIBLE

4.3.1. Détermination des (p) SCR sur base des alignements de séquences obtenus

4.3.2. Assignation des coordonnées du *template* à la séquence cible pour les régions conservées

Celle-ci se fait par superposition des régions conservées de la séquence cible à la structure des régions équivalentes dans le *template*. Le résultat de cette étape fournit un ensemble de coordonnées spatiales, généralement des Cα dans les zones conservées, et ce, pour la séquence cible. A ce stade, on obtient un modèle partiel du squelette de la protéine, car les régions variables correspondant le plus souvent aux *loops* ne sont pas encore modélisées.

4.3.3. Prédiction des loops:

Certaines techniques de modélisation des *loops* utilisent des banques de données de structures de loops extraits de structures cristallographiques (Bates and Sternberg 1999). Parmi ces structures, le fragment qui remplit le mieux les trous entre les régions conservées est sélectionné. Cette sélection se fait par superposition des régions variables de la séquence cible à chacun des fragments de même longueur répertoriés dans la librairie. La structure la plus énergétiquement favorable est alors choisie. D'autres méthodes recherchent la meilleure conformation que pourrait adopter le fragment de séquence en utilisant la dynamique moléculaire et la minimisation d'énergie (voir plus loin).

4.3.4. Positionnement des chaînes latérales

Le positionnement des chaînes latérales se fait en leur attribuant les conformations prises par les chaînes latérales équivalentes dans le *template* et, si ce n'est pas possible, se

base sur les conformations préférentielles qu'elles peuvent prendre (Bates and Sternberg 1999) ; (Dunbrack 1999). Elle dépend fortement de l'alignement et de la qualité de la structure du *template* (Venclovas, Ginalski *et al.* 1999).

Pour des raisons stériques, les angles de torsion des chaînes latérales prennent des valeurs limitées. De ce fait, ces dernières adoptent des conformations préférentielles appelées rotamères. Des graphes similaires au plot de Ramachandran reprennent des distributions d'angles retrouvées pour chacune des chaînes latérales dans les meilleurs structures cristallographiques. Les rotamères correspondent aux zones de haute densité d'angles. Pour modéliser les chaînes latérales, les rotamères sont assignés à chacune de celles-ci et celui qui apparaît le plus énergétiquement favorable est alors adopté. Cependant des états non rotamériques (i.e. non énergétiquement favorables) sont systématiquement observés dans les structures cristallographiques des protéines, généralement à cause des interactions tertiaires avec les autres chaînes latérales (Schrauber, Eisenhaber *et al.* 1993). Dès lors, le positionnement des chaînes latérales lors de la prédiction de structures tridimensionnelles n'est pas optimal.

4.3.5. Optimisation du modèle par minimisation d'énergie et dynamique moléculaire.

L'hypothèse thermodynamique du repliement protéique suppose qu'une protéine adopte sa conformation native à son minimum d'énergie libre de Gibbs. Cette énergie est difficilement mesurable mais on peut la relier à l'énergie potentielle de la protéine. Le calcul de l'énergie potentielle de telles molécules est fastidieux et irréalisable en pratique si on veut utiliser les méthodes très précises (mécanique quantique). Néanmoins, on peut approximer cette énergie grâce à des fonctions empiriques beaucoup plus simples du nom de champ de forces (*force field).* Celui-ci décrit l'ensemble des interactions subies par chaque atome d'une protéine (les interactions de van der Waals et électrostatiques). Lorsque la dérivée de cette fonction par rapport aux coordonnées atomiques est égale à 0, la fonction est à un minimum.

Le but de la minimisation d'énergie est de ramener la fonction de chaque atome à un minimum. Dans le cadre de la modélisation par homologie, la minimisation d'énergie tente de réajuster la conformation du modèle initial pour atteindre un minimum que l'on espère le plus proche de l'énergie de la structure native. Différents ajustements de la

structure (les itérations) ont lieu. La minimisation d'énergie présente l'inconvénient de ne pas pouvoir franchir les barrières énergétiques. Or, l'ensemble des conformations que peut prendre une protéine comprend toute une série de minima locaux séparés par des barrières énergétiques. Une simple minimisation d'énergie a, dès lors, pour effet de ramener la conformation initiale du modèle à un minimum d'énergie qui ne correspond pas obligatoirement au minimum auquel on s'attend. En conséquence, la conformation générée ne s'écarte pas beaucoup de celle du modèle initial.

La dynamique moléculaire, pour sa part, permet de franchir les barrières énergétiques en apportant virtuellement de l'énergie cinétique à la protéine. Le déplacement des atomes que cette énergie occasionne permet de générer toute une série de conformations différentes. Ensuite, une minimisation d'énergie permet de retrouver, en principe, la conformation native correspondant au minimum global. L'exploration de l'espace conformationnel se fait aussi par itérations.(avec alternance de dynamique moléculaire et de minimisation d'énergie). Cependant, il faudrait environ 30 milliards d'années pour simuler le repliement complet d'une protéine. Il est donc impossible d'explorer tout son ensemble conformationnel. En pratique, seules certaines conformations sont explorées et un minimum global est rarement atteint.

Remarquons que la minimisation d'énergie et la dynamique moléculaire n'améliorent pas toujours le modèle. En effet, bien qu'étant plus énergétiquement favorable, la conformation du modèle « optimisé » peut s'écarter de la structure native. Par conséquent, elles peuvent être sources d'erreurs (CASP1,2,3).

4.3.6. Evaluation du modèle sur base de critères énergétiques et géométriques

WHAT-CHECK (Vriend and Sander 1993) et PROCHECK (Laskowski, Rullmannn *et al.* 1996) sont des programmes de vérification de la vraisemblance de toute une série de caractéristiques géométriques dans une structure. Parmi celles-ci, citons les angles de torsion et de valence, la chiralité, la longueur des liaisons, ... Ces programmes comparent chacune des caractéristiques aux distributions statistiques observées dans de nombreuses protéines de structure connue. Cependant, mêmes si ces contraintes sont respectées, le modèle n'est pas pour autant toujours correct. Il est, en effet, possible de construire des modèles qui sont validés par de tels programmes mais qui n'ont aucune

signification biologique. De plus, des conformations inhabituelles mais pas erronées sont parfois prises dans la structure native et s'avérent capitales pour la fonction d'une protéine.

Des programmes comme Verify 3D (Luthy, Bowie *et al.* 1992) utilisent un potentiel de forces moyennes pour vérifier si le modèle adopte une structure énergétiquement favorable.

La construction des champs de force extrapole les données thermodynamiques dérivées de systèmes simples aux systèmes macromoléculaires que sont les protéines. Celle-ci se base sur l'hypothèse selon laquelle le comportement de systèmes complexes résulte de la combinaison du même type d'interactions que dans des systèmes simples. A l'inverse, la théorie « déductive » présume que les forces en présence sont beaucoup plus compliquées. Suivant cette approche, les structures des protéines sont prises comme seule source d'information pour extraire les forces et potentiels qui stabilisent les protéines. On appelle ces potentiels les « potentiels de forces moyennes». Leur calcul se base sur les considérations suivantes: à l'équilibre, un système moléculaire se situe à un minimum d'énergie. Cependant, pour une molécule donnée, plusieurs conformations correspondant à des états énergétiques différents sont possibles. La distribution statistique de ces conformations est gouvernée par la loi de Boltzmann, qui relie l'énergie libre aux probabilités d'occurrence des différents états énergétiques observés pour ces conformations.

$$E(r) = -kT \ln\left[f(r)\right]$$

où r est la distance
E(r) est l'énergie à la distance r
k est la constante de Boltzmann
T est le température

Dès lors, la loi inverse de Boltzmann (reprenant la même équation mais inversée) permet de retrouver l'énergie libre d'une molécule (son potentiel de force moyenne) à partir des probabilités d'occurrence d'un état particulier parmi les différents états énergétiques possibles. Pour l'ensemble des états possibles, on obtient alors une courbe d'énergie. Ceci peut s'appliquer aux protéines : on peut calculer le potentiel de force moyenne d'un *fold* connu à partir des probabilités d'occurrence des chacun des 20 acides amines vis à vis de tous les autres acides aminés. Ce potentiel correspond à l'énergie

minimum de la protéine (puisqu'elle est à l'état natif). Pour vérifier la vraisemblance d'un modèle, il suffit d'additionner les énergies de chaque résidu calculées à partir des probabilités d'occurrences d'interactions entre un résidu de la structure protéique et les autres. Le score ainsi calculé représente l'énergie libre de la structure protéique. Si cette énergie est positive, la protéine n'est pas dans une conformation stable. Cela signifie que le modèle n'est pas correct.

4.4. COMPARAISON DU MODELE DE LA PROTEINE CIBLE A SA STRUCTURE REELLE

Il est intéressant de comparer les modèles de protéines de structure connue à leur structure cristallographique. De cette façon, il est possible d'évaluer la précision du modèle, et de déterminer quelles sont les sources d'erreurs les plus fréquentes lors de la modélisation. Pour ce faire, le critère le plus communément usité est le RMSD (root mean square of distance). Celui-ci est calculé après superposition des deux structures par des logiciels tels Insight II. Le RMSD représente la moyenne des carrés des distances entre chaque paire d'atomes superposés et est exprimé en Å.

Les CASP (*Critical Assessment of techniques for Structure Prediction*) sont des sessions bisannuelles qui existent depuis 1994. Leur but est d'évaluer les méthodes existantes de prédiction de structures de protéines. Avant la session, des séquences de la *Pending List* de PDB sont soumises à la communauté scientifique. Les participants, qui ne connaissent pas leur structure réelle sont alors tenus de les modéliser. Ensuite, les modèles de chaque protéine test sont comparés à sa structure réelle en utilisant divers critères dont principalement le RMSD. Nous nous attacherons à comparer les performances de notre méthode aux meilleurs expert en modélisation de ce concours international.

4.4.1. Qualité et utilité d'un modèle prédit par homologie

La qualité d'un tel modèle dépend du pourcentage d'identités partagé entre la séquence cible et son (ses) *template(s)*.

- S'il est supérieur à 50 %, la plupart des caractéristiques structurales du modèle seront bien prédites.
- De 50 à 25 %, la précision du modèle tend à diminuer.
- S'il est inférieur à 25 %, la construction du modèle ne sera pas techniquement plus difficile mais sa précision sera moins bonne (Tramontano 1998). Dans ce dernier cas, pour obtenir un modèle raisonnable, il sera dès lors nécessaire de combiner plusieurs méthodes tout en tenant compte de toutes les informations disponibles : famille de la protéine cible, contexte génétique, ...

En d'autres termes, plus les résidus sont différents d'une protéine à une autre, plus la prédiction de structure sera difficile.

On peut classer les qualités des modèles en trois catégories (Peitsch 1996):

- les modèles basés sur des alignements incorrects entre la séquence cible et le(s) *template*(s). De tels modèles sont souvent utiles lorsque les erreurs ne se localisent pas dans les régions bien conservées telles les sites actifs des enzymes.
- Les modèles construits à partir d'alignements corrects mais pour lesquels la séquence d'intérêt et le(s) *template*(s) partagent une similarité faible ou moyenne (<70 %). Ces modèles s'avèrent très utiles pour la planification d'expériences de mutagenèse dirigée mais ils ne permettent pas d'étudier en détail leur liaison à des ligands.
- Les modèles de protéines partageant un fort taux d'identités (>70 %) avec leur(s) *tempalte*(s). Ceux-ci sont indiqués, par exemple, pour comparer des protéines variant d'une espèce à l'autre. Ces comparaisons peuvent faciliter la recherche d'inhibiteurs spécifiques d'une structure présente uniquement dans une espèce donnée.

Sachant cela, il est important de remarquer que même des modèles de faible résolution sont souvent utiles pour se poser des questions biologiques. En effet, la plupart des caractéristiques fonctionnelles peuvent souvent être suggérées à partir de caractéristiques structurales d'un modèle (A. Fiser *et al.*, 1999 unpublished). Donc, même si l'on sait que le modèle est loin de la réalité, il est possible d'évaluer la fiabilité de certaines de ses parties et d'en tirer des informations intéressantes à propos de la protéine cible (Tramontano 1998).

5. MODELISATION PAR RECONNAISSANCE DE FOLD

Il n'est pas toujours possible de trouver une séquence homologue de structure connue lors d'une recherche en base de données. La probabilité de détecter une structure connue pour une séquence prise au hasard dans un génome se situe entre 20 et 30 % (Huynen and van Nimwegen 1998) ; (Jones, Tress *et al.* 1999).

Lorsqu'on est incapable de détecter une homologie entre la séquence d'intérêt et une séquence de structure connue, on peut se tourner vers les méthodes dites de reconnaissance de fold (qui seront utilement complétées par la prédiction de structure secondaire). Ces méthodes sont basées sur le fait que deux protéines peuvent adopter des folds très similaires sans pour autant avoir une similarité de séquence ou de fonction. Ceci suggère que le nombre de folds serait limité (certainement par les contraintes physico-chimiques). Il est estimé à 1000 (Chothia and Murzin 1993). Dès lors, pour prédire la structure d'une protéine, on peut se poser la question de savoir si la séquence considérée peut adopter un des folds connus.

On distingue deux types de méthodes de reconnaissance de fold :

- le **threading** se base sur des considérations énergétiques (le potentiel de force moyenne, voir partie « Matériel »).

Chaque famille de folds connus est caractérisée par une énergie potentielle (le potentiel de force moyenne) résultant de la somme des interactions entre résidus.

La séquence d'intérêt est alors « enfilée » sur chacune des structures de la librairie de folds. Le programme recherche alors l'alignement structural qui sera le plus énergétiquement favorable.

De telles méthodes sont effectuées par des programmes tels ProFIT (Sippl 1993) ou THREADER (Jones, Taylor *et al.* 1992).

- le **pseudo-threading**. .quant à lui, décrit le fold en terme d'environnement de chaque résidu dans la structure : la structure secondaire locale, l'accessibilité au solvant et le degré de polarité des atomes. Cet environnement est plus conservé que l'identité du résidu en elle-même et cette méthode peut donc détecter plus d'interactions séquence-structure que les méthodes basées uniquement sur la comparaison de séquences.

Notons que les méthodes de reconnaissances de fold ne sont pas très précises : environ 50 % des folds sont assignés correctement. Toutefois, ce pourcentage n'est valable que si l'on ne tient compte que du premier fold donné. Il augmente si l'on choisit un fold parmi les dix premiers folds proposés...

Pour limiter les erreurs, il faut donc veiller à

- combiner les résultats de différents programmes ;
- vérifier si le fold choisi est plausible en collectant le plus d'informations structurales sur sa famille et en s'assurant qu'il concorde avec les prédictions de structure secondaire.

CHAPITRE III: MATERIEL ET METHODES

Les recherches effectuées durant ce travail ont toutes été réalisées sur des stations de travail Silicon Graphics Octane duo MIPS R10000 à 225Mhz et INDIGO 2 MIPS R4400 à 150 Mhz fonctionnant sur les systèmes d'exploitation IRIX 6.5 et IRIX 5.3 ayant respectivement 512 et 64 MB de RAM

1. RECHERCHE EN BANQUES DE DONNEES

1.1. BANQUES DE DONNEES.

1.1.1. PDB (Protein Data Bank Brookhaven National Laboratories, Cambridge, USA)

Toutes les protéines modélisées au cours de ce mémoire sont issues de PDB. Il s'agit d'une banque de structures tridimensionnelles de protéines, d'acides nucléiques, de carbohydrates et d'une variété de complexes déterminés expérimentalement par diffraction des rayons X et par RMN. Le site Web de PDB (http://www.pdb.bnl.gov) permet non seulement de retrouver la structure tridimensionnelle d'une protéine de PDB à partir d'un code de quatre caractères qui lui a été assigné mais aussi d'obtenir des informations dont:

- une brève description de la manière dont a été déterminée sa structure (technique utilisée, date de soumission, nom des auteurs, résolution, ...)
- des commentaires sur sa structure proprement dite : fonction, nombre de chaînes et de domaines, cofacteurs, structures primaire et secondaire, site actif,...
- une liste de caractéristiques pour chaque résidu (numéro, coordonnées 3D, ...)

1.1.2. Banque de données non redondante

Les recherches effectuées tout au long de ce travail ont été réalisées dans une banque non redondante : la banque "nr" formée au NCBI. Elle reprend toutes les séquences CDS (séquences codantes) traduites de Genbank, ainsi que les séquences non redondantes

des banque de données PDB, SWISS PROT, PIR et PRF rendues publiques tous les 30 jours.

1.2. PROGRAMMES DE RECHERCHE EN BANQUES DE DONNEES

1.2.1. BLAST (Basic Local Alignment Search Tool)

BLAST (Altschul, Gish *et al.* 1990) est un programme rapide de recherche de séquences similaires à la séquence d'intérêt dans une banque de données. En plus des séquences qu'il trouve, il donne un alignement local pairé entre chacune des séquences homologues et la séquence d'intérêt.

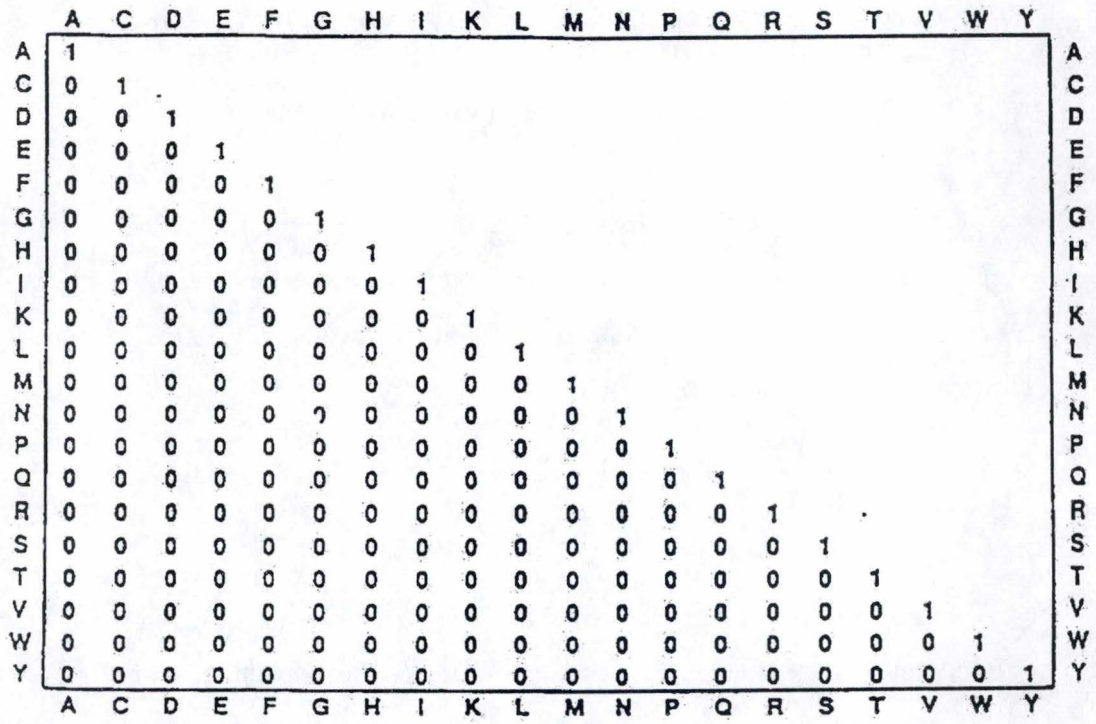
1.2.2. PSI BLAST

PSI-BLAST (Altschul, Madden *et al.* 1997) est également un programme de recherche en base de données. Il construit un profil spécifique pour chaque position de la séquence d'intérêt. En effet, dans une première étape, le programme effectue une recherche en base de données (i.e. un simple gapped BLAST). Puis PSI BLAST crée automatiquement un profil pour la séquence d'intérêt avec ses homologues les plus proches. Après cette première étape, le processus est répété, mais cette fois-ci, on réalise la recherche en base de données en utilisant le profil. On recommence le processus (par itérations) jusqu'à convergence du nombre de séquences retrouvées.

Un profil est une table de scores pour chaque position d'un alignement. Il définit quels résidus sont permis pour des positions données : quelles sont les positions conservées ou dégénérées ou quelles positions/régions peuvent tolérer des insertions ou des délétions. Ce profil va permettre de rechercher un plus grand nombre d'homologues au cours de différentes itérations (Altschul, Madden *et al.* 1997). Il fournit également des alignements pairés semblables à ceux proposés par BLAST mais ceux-ci incluent en général plus de résidus.

1.2.3. PURGE

Purge(Lawrence, Altschul *et al.* 1993) est un programme qui utilise une stratégie itérative pour extraire des séquences homologues non redondantes d'un ensemble de séquences.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1																			
C	0	1																		
D	0	0	1																	
E	0	0	0	1																
F	0	0	0	0	1															
G	0	0	0	0	0	1														
H	0	0	0	0	0	0	1													
I	0	0	0	0	0	0	0	1												
K	0	0	0	0	0	0	0	0	1											
L	0	0	0	0	0	0	0	0	0	1										
M	0	0	0	0	0	0	0	0	0	0	1									
N	0	0	0	0	0	0	0	0	0	0	0	1								
P	0	0	0	0	0	0	0	0	0	0	0	0	1							
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	1						
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1					
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 11. Matrice Identité

Cet algorithme effectue une recherche transitive pour déterminer quelles sont les séquences reliées entre elles. Si une recherche pairée trouve que deux séquences A et B sont similaires, et qu'une seconde recherche trouve que B est reliée à une troisième séquence, C, alors, A et C peuvent être reliées. A travers une série de recherches par BLAST, toutes les interactions de ce type sont déduites de façon itérative jusqu'à ce qu'aucune nouvelle séquence ne soit trouvée.

Dans ce mémoire, Purge a été utilisé pour extraire un ensemble de séquences équidistantes en terme de similarité.

2. ALIGNEMENT DE SEQUENCES

2.1. LES MATRICES DE SCORES

Différents programmes d'alignement de séquences utilisent des tableaux appelés **matrices de scores**. Celles-ci sont représentées par des tableaux 20 x 20 permettant de comparer les acides aminés deux à deux en fonction de leur similarité et de leur attribuer un score.

Il existe toute une série de matrices de scores construites, par exemple, sur base d'échelles de caractéristiques physico-chimiques des divers acides aminés, de critères énergétiques ou encore en fonction du taux de mutation d'un résidu par un autre.

Nous pouvons distinguer deux représentations de matrices de scores : les matrices de **similarité** et les matrices de **distance**. Les premières attribuent un score élevé aux paires d'acides aminés similaires alors que les paires les moins similaires seront indiquées par un score faible. Par contre, les secondes sont des matrices pour lesquelles plus les acides aminés sont différents, plus leurs scores sont élevés. Il est néanmoins possible de transformer les matrices de similarité en matrices de distance en soustrayant les valeurs de scores maximales et en changeant de signe.

La matrice identité octroie un score de 1 ou de 0 selon qu'ils soient respectivement identiques ou non. Elle permet de calculer le pourcentage d'identité qui est souvent utilisé pour calculer une similarité globale entre deux séquences (voir figure 11).

Parmi les matrices les plus connues, on notera la famille des PAMs (Dayhoff M. O. 1972) qui se basent sur les fréquences de mutations acceptées lors de l'évolution. Par exemple, la matrice PAM120 est adéquate pour comparer des séquences ayant subi 1,2 fois (120%) plus de mutation acceptées que leur nombre d'acides aminés. On notera également la famille des BLOSUMs (Henikoff and Henikoff 1992) qui

sont construites à partir des fréquences de remplacement des acides aminés dans des alignements de séquences de référence.

2.2. PROGRAMMES D'ALIGNEMENT PAIRE

2.2.1. Align

Align (Pearson 1990) est un programme d'alignement pairé qui utilise la méthode de Needleman et Wunsch (programmation dynamique) (Needleman and Wunsch 1970) pour réaliser l'alignement. Nous n'entrerons pas dans les détails de fonctionnement de l'algorithme.

Paramètres par défaut:

- matrice de score : BLOSUM 50
- Gap opening penalty : -12
- Gap extending penalty : -2

2.3. PROGRAMMES D'ALIGNEMENT MULTIPLE.

2.3.1. Match-Box

Match-Box (Depiereux, Baudoux *et al.* 1997) est un logiciel d'alignement multiple local, c'est-à-dire qu'il aligne des séquences par fragments similaires afin d'y déterminer les régions prédites comme étant structuralement conservées. Il comporte deux parties:

- EXPLORE permet de détecter si des séquences sont similaires et, de ce fait, s'il est adéquat de les aligner ultérieurement. Une analyse permet de visualiser des groupes de séquences similaires.

- ALIGN est conçu pour réaliser l'alignement des séquences. Il est composé de trois algorithmes, à savoir le *scanning*, le *matching* et le *screening*. Le scanning effectue une

analyse des séquences de manière à déterminer les meilleurs paramètres qui doivent être utilisés dans la suite du programme. Le *matching* effectue toutes les comparaisons pairées de segments de séquences pour constituer des boîtes de segments similaires. Dans le *screening*, les boîtes *a priori* les meilleures sont alors sélectionnées pour constituer l'alignement final.

Match-Box fourni un coefficient de confiance compris entre 1 et 9 (par ordre décroissant de certitude). Cet indice est proportionnel aux taux de faux positifs observés lors de l'alignement de familles de séquences dont les alignements de référence sont connus. La confiance est un élément important pour déterminer si les zones ont été correctement alignées. Ceci représente un avantage qu'il est pertinent de relever.

2.3.2. Clustal W

Clustal W (Thompson, Higgins *et al.* 1994) est un programme d'alignement multiple global de séquences très utilisé. Il est issu de Clustal V (Higgins, Bleasby *et al.* 1992).

A partir des séquences trouvées lors d'une recherche en base de données, il calcule simultanément un ensemble d'alignements pairés en comparant chaque séquence à toutes les autres séquences. Suite à ces comparaisons, une matrice de distance qui reflète la similarité est calculée. Cette matrice permet la construction d'un **arbre phylogénétique** (*i.e.* une représentation graphique des relations évolutives présumées entre groupes d'organismes). Celui-ci peut éventuellement être pondéré de telle sorte que soient favorisées les séquences très fortement similaires. Il sert de base pour la construction de l'alignement. Ce dernier débute par l'alignement pairé des deux séquences les plus similaires, puis un nouvel alignement est opéré avec la séquence la plus similaire et ainsi de suite.

Si un nombre conséquent de séquences doit être ajouté, l'addition de *gaps* est inévitable, pour insérer les séquences divergentes dans l'alignement. La différence principale entre Clustal W et Clustal V est, pour la version la plus récente, une réduction locale des *gaps* dans les régions hydrophiles : les indels sont favorisées dans les structures secondaires formant les boucles plutôt que dans les structures régulières susceptibles d'être conservées. Ceci a été réalisé pour éviter d'aligner des régions qui, en réalité, ne sont pas similaires, l'apport de *gaps* augmentant le nombre d'identités entre deux séquences. En

outre, une pénalité (*gap penalty*) est calculée pour chaque *gap* en fonction de sa localisation probable dans l'une ou l'autre structure.

2.3.3. Multalin

Multalin (Corpet 1988) est un programme d'alignement multiple de séquences qui utilise une stratégie progressive (comme ClustalW). Cette stratégie consiste à aligner toutes les paires de séquences pour en retirer un score de similarité. Ce score est utilisé pour construire un arbre phylogénétique qui servira à guider la construction de l'alignement multiple. L'alignement multiple est alors construit en commençant par aligner les séquences puis groupes de séquences les plus proches évolutivement. Les groupes se forment avec des séquences qui ont déjà été alignées.

2.3.4. Dialign

Dialign (Morgenstern, Frech *et al.* 1998) utilise une stratégie de recherche de segments similaires. Elle consiste à trouver des points d'ancrages qui seraient des parties de l'alignement global optimal en faisant toute une série de comparaisons pairées. Au plus le nombre de points d'ancrage trouvés est grand, au plus le temps calcul sera réduit. Dialign effectue toutes les comparaisons pairées de segments puis assemble les paires de segments en un alignement final. Cette méthode essaie de trouver des similarités dans les séquences et restreint l'alignement aux segments de séquences qui sont plus similaires entre eux qu'on ne pourrait l'espérer par hasard. Tout comme Match-Box, cette méthode fourni un indice de certitude à chaque position alignée.

2.3.5. PIMA (Pattern-Induced Multi-sequence Alignment)

PIMA (Smith and Smith 1992) est un autre programme d'alignement progressif qui utilise des patterns (séquences consensus) pour représenter les groupes de séquences déjà alignées: chaque fois qu'il réalise un alignement pairé entre les deux séquences les plus similaires, il définit cet alignement par une séquence consensus et utilise ce consensus pour réaliser l'alignement pairé suivant. Il produit deux alignements car il utilise deux manières différentes pour générer l'arbre phylogénétique lui servant de guide.

3. PROGRAMMES DE MODELISATION.

3.1. MODELLER

MODELLER (Sali and Blundell 1993) est un programme qui modélise les structures 3D de protéines en respectant les contraintes spatiales telles les distances entre atomes, les angles de liaisons , de torsion, ...

Ce programme est utilisé le plus souvent pour la modélisation de protéines par homologie. Son utilisateur lui fournit un alignement préalable entre la séquence d'intérêt et son *template*. Le modèle est alors construit automatiquement. Néanmoins, il peut aussi construire un alignement puis un modèle si on lui soumet les deux séquences à aligner.

Dans notre cas, nous avons utilisé MODELLER avec les paramètres par défaut pour modéliser chaque cas-test à partir de chacun des alignements qui lui étaient donnés.

Le résultat du modèle est un ensemble de fichiers relatifs à la modélisation. Parmi ceux-ci, un fichier contenant les coordonnées du modèle est disponible sous format PDB. Son utilisation permet de visualiser la protéine modélisée au moyen de programmes comme Insight II (Insight© and Homology© programs of molecular simulation, San Diego).

4. PROGRAMMES D'EVALUATION DES MODELES

4.1. PROCHECK ET WHATCHECK

Procheck (Laskowski, Rullmannn *et al.* 1996) et Whatcheck (Vriend and Sander 1993) sont des programmes qui évaluent les qualités stéréochimiques et structurales d'une structure protéique donnée en évaluant de sa fiabilité locale, résidu par résidu.

Ce logiciel produit un ensemble de diagrammes (par exemple un diagramme de Ramachandran) ainsi qu'une liste détaillée de chaque caractéristique évaluée pour chaque résidu.

Leur principe est de comparer les valeurs de diverses caractéristiques structurales (angles de valence, longueurs de liaisons, angles Phi et Psi, stéréochimie, ...) aux distributions de ces grandeurs observées dans des protéines déterminées avec une bonne résolution (inférieure à 2.5 Å). On peut classer les valeurs obtenues pour la structure d'intérêt en différentes classes suivant leur écart par rapport à la moyenne: à 1, 2, 3, 4 et

plus de 4 écarts types. Les valeurs distantes de plus de 4 écarts types par rapport à la moyenne sont considérées comme exceptionnelles et, donc, ont très peu de chance d'être observées. Elles seront alors considérées comme très mauvaises.

4.2. VERIFY 3D

Verify 3D (Luthy, Bowie *et al.* 1992) utilise un potentiel de forces moyennes pour vérifier si le modèle adopte une structure énergétiquement favorable.

Cette approche consiste à ne prendre que les structures des protéines comme seule source d'information pour extraire les forces et potentiels qui stabilisent les protéines. On appelle ces potentiels les « potentiels de forces moyennes». Leur calcul se base sur les considérations suivantes: à l'équilibre, un système moléculaire se situe à un minimum d'énergie. Cependant, pour une molécule donnée, plusieurs conformations correspondant à des états énergétiques différents sont possibles. La distribution statistique de ces conformations est gouvernée par la loi de Boltzmann, qui relie l'énergie libre aux probabilités d'occurrence des différents états énergétiques observés pour ces conformations.

$$E(r) = -kT \ln[f(r)]$$

où r est la distance
E(r) est l'énergie à la distance r
k est la constante de Boltzmann
T est le température

Dès lors, la loi inverse de Boltzmann (reprenant la même équation mais inversée) permet de retrouver l'énergie libre d'une molécule (son potentiel de force moyenne) à partir des probabilités d'occurrence d'un état particulier parmi les différents états énergétiques possibles. Pour l'ensemble des états possibles, on obtient alors une courbe d'énergie. Ceci peut s'appliquer aux protéines : on peut calculer le potentiel de force moyenne d'un *fold* connu à partir des probabilités d'occurrence des chacun des 20 acides amines vis-à-vis de tous les autres acides aminés. Ce potentiel correspond à l'énergie minimum de la protéine (puisqu'elle est à l'état natif). Pour vérifier la vraisemblance d'un modèle, il suffit d'additionner les énergies de chaque résidu calculées à partir des

probabilités d'occurrence d'interactions entre un résidu de la structure protéique et les autres. Le score ainsi calculé représente l'énergie libre de la structure protéique. Si cette énergie est positive, la protéine n'est pas dans une conformation stable. Cela signifie que le modèle n'est pas correct.

5. PROGRAMMES DE VALIDATION DES MODELES.

5.1. INSIGHT II

InsightII (Insight© programs of molecular simulation, San Diego)

est un logiciel qui permet entre autres de visualiser des molécules en trois dimensions à partir des coordonnées cristallographiques (issues du fichier PDB pour les protéines) ou de coordonnées de modèles réalisés préalablement.

Dans ce mémoire, nous avons également utilisé ce logiciel pour comparer nos modèles à la structure réelle de chaque cas-test. Pour ce faire, nous avons superposé la trace des deux structures comparées en utilisant la commande « superimpose » du menu « Transform ». Le RMSD était alors calculé automatiquement par le programme.

Pour calculer les RMSD locaux, le module Homology de MSI (Homology© programs of molecular simulation, San Diego)

nous a permis d'effectuer les étapes suivantes :

- l'alignement pairé automatique des séquences considérées
- l'alignement manuel des structures
- la constitution de boîtes contenant 9 paires de résidus.
- le calcul du RMSD pour chaque boite de 9 acides aminés de l'alignement pairé.

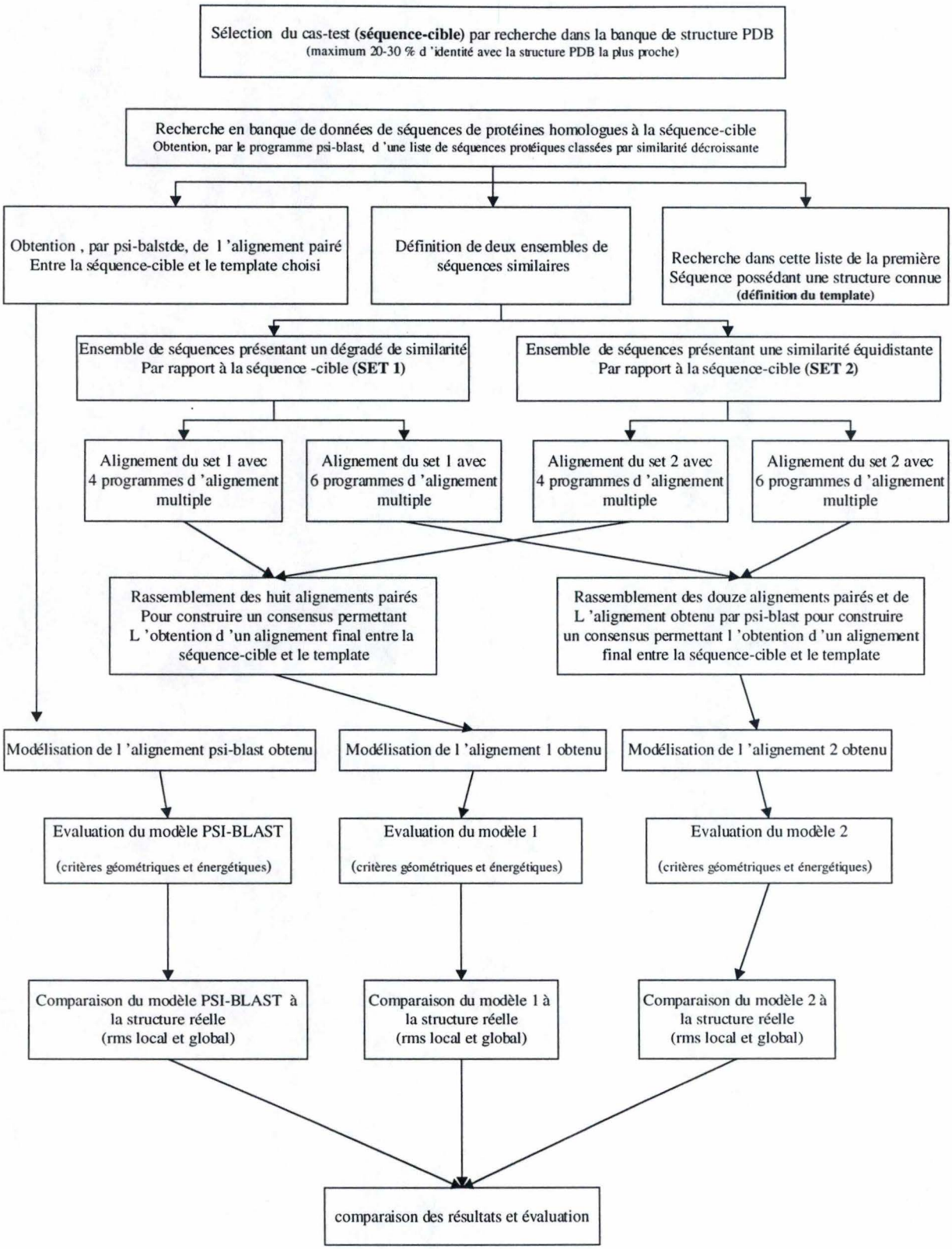
BUT DU MEMOIRE

Ce travail consiste en l'élaboration d'une stratégie fiable d'alignement pairé pour la modélisation de protéines de faible homologie avec leur *template* (*i.e.* dans la *twilight zone*). Cette stratégie sera ensuite utilisée dans le cadre de la mise au point d'un système expert de modélisation.

La *Twilight Zone (Doolittle 1986)* est une zone de similarité de séquences (comprise entre 20 et 30 % d'identités) dans laquelle les alignements peuvent sembler plausibles visuellement mais ne sont pas toujours statistiquement significatifs. Autrement dit, il y a des chances que ces alignements aient été produits par hasard. Il est donc impératif d'utiliser les méthodes de détection d'homologues les plus sensibles (comme PSI-BLAST) afin d'éviter les faux positifs. De plus, il faut s'assurer que l'alignement réalisé entre les séquences homologues est optimal, ce qui limite le risque de modélisation erronée. C'est pourquoi notre stratégie essaie d'améliorer les alignements pairés *target-template*. En dessous de 20 % d'identités, les programmes de comparaison de séquences ne parviennent plus à détecter des similarités significatives. C'est la *Midnight Zone* (Rost 1997).

A l'heure actuelle, il n'existe aucun programme automatique qui puisse prédire de façon fiable la structure tridimensionnelle de protéines partageant un pourcentage d'identités inférieur à 30 % avec une structure *template*. C'est la raison pour laquelle l'optimisation de la modélisation des protéines dans la *twilight zone* et son automatisation se révèlent d'un grand intérêt

Stratégie de modélisation



CHAPITRE IV: METHODOLOGIE

1. SELECTION DES CAS-TEST

Les protéines modélisées au cours de ce mémoire (*i.e.* les cas-test) sont toutes issues de la PDB. Leur structure étant connue, il était, dès lors, possible de comparer les modèles générés à leur structure réelle .

Nous nous sommes focalisés sur la modélisation par homologie de protéines partageant un faible pourcentage d'identités avec leur *template*. Les cas-test ont été sélectionnés de la manière suivante:

- Toutes les séquences de PDB ont été alignées deux à deux par le programme ALIGN (avec les paramètres par défaut). Les protéines choisies partageaient un pourcentage d'identités compris entre environ 20 et 30 % avec la structure de PDB la plus proche.
- Chacun des cas-test devait contenir au moins 120 résidus de manière à un domaine fonctionnel complet. Dans la mesure ou nous ne nous sommes intéressés qu'à la modélisation à partir d'un seul *template*, des protéines trop longues, contenant plusieurs domaines n'ont pas été reprises. En effet, même si la prédiction structurale des domaines était souvent correcte, l'orientation des domaines les uns par rapport aux autres se révélerait certainement moins conservée.

Nous aurions pu utiliser plusieurs *templates*. En effet, l'usage de plusieurs *templates* augmente souvent la précision du modèle. Cependant, comme nous voulions modéliser les cas-test le plus rapidement possible, nous avons choisi de ne sélectionner qu'un seul *template*. Evidemment, l'alignement éventuel de la séquence d'intérêt à plusieurs *templates* ne sera pas à négliger et devra être testé lors de l'élaboration du système expert automatisé.

Les cas-test ayant été définis, la méthodologie élaborée pour la modélisation de chacun d'eux est décrite ci-dessous.

2. MODELISATION PAR HOMOLOGIE

Pour rappel, nous avons utilisé la modélisation par homologie pour modéliser nos protéines car c'est celle qui est la plus précise (CASP, ref). Les principales étapes sont :

- la recherche en base de données de séquences homologues à la séquence cible
- la sélection du *template*
- l'alignement multiple de la séquence cible, du *template* et de leurs séquences homologues
- l'extraction d'un alignement pairé entre la séquence cible et le *template* à partir des alignements multiples
- la modélisation de la séquence cible sur base de son alignement à la structure du *template*
- la vérification du modèle par des programmes qui évaluent la plausibilité d'une structure sur base de critères énergétiques ou géométriques.
- la comparaison du modèle à la structure réelle en calculant les RMSD globaux et locaux entre les deux structures superposées.

2.1. RECHERCHE EN BASE DE DONNEES ET SELECTION DU TEMPLATE

La recherche de séquences similaires à chaque séquence cible a été réalisée en utilisant la banque de **séquences** non redondantes (nr) du NCBI : dans chaque cas, la séquence d'intérêt a été soumise au programme PSI-BLAST. Ce programme permet de retrouver plus de séquences homologues que BLAST et FASTA (Altschul, Madden *et al.* 1997). Celui-ci a alors fonctionné jusqu'à convergence du nombre de séquences retrouvées. Les paramètres utilisés ont été : matrice BLOSUM 62 , coût des gaps par résidus = 1, lambda ratio =1

Lors de cette recherche, seules les séquences de Evalue < 0,001 ont été retenues afin de s'assurer avec assez de certitude de leur similarité par rapport à la séquence cible.

Ensuite, le *template* a été sélectionné comme étant la première séquence de **structure** connue classée par PSI-BLAST. Remarquons qu'il ne correspondait pas obligatoirement à la séquence de structure connue la plus similaire à la séquence cible. Toutefois, ceci ne constitue pas un problème en soi car bien que le *template* ne soit pas toujours le plus similaire à la séquence cible, les structures sont beaucoup plus conservées que les séquences. Et le but escompté est bien de prédire la structure de la séquence d'intérêt.

Notons cependant que si aucune séquence de structure connue (hormis la séquence d'intérêt) n'était trouvée lors de cette recherche, les protéines d'intérêt n'ont pas été modélisées. Il faut, par ailleurs, remarquer qu'elles auraient pu être modélisées par des méthodes telles la reconnaissance de *fold* mais là n'était pas le but de notre travail.

Les cas-test sélectionnés partageaient entre 20 et 30 % d'identités avec la séquence de PDB la plus similaire. Or, aux alentours de ces pourcentages, le nombre d'identités entre deux protéines peut être dû au hasard et ne reflète pas nécessairement une réelle similarité (Rost and O'Donoghue 1997). Et donc, lors de la recherche en base de données, il est possible que PSI-BLAST ne retrouve pas de séquence de structure connue similaire.

A partir du groupe de séquences similaires retrouvées par PSI BLAST, deux ensembles de séquences ont été constitués pour faire fonctionner les programmes d'alignement dans deux conditions différentes :

- un ensemble de séquences similaires présentant un dégradé de similarité (l'ensemble 1).
- un ensemble de séquences équidistantes du point de vue de la similarité (l'ensemble 2).

En effet, Thompson (Thompson, Plewniak *et al.* 1999) a montré que les programmes d'alignement multiple ont des performances différentes dans ces deux conditions. Et donc, si dans ces deux conditions, des programmes alignent la séquence cible et le *template* de façon identique, un consensus des deux alignements augmente la confiance de la prédiction (Briffeuil, Baudoux *et al.* 1998). C'est pour cette raison que la plupart de nos modélisations se sont basées sur un consensus de divers alignements.

L'ensemble 1 regroupait les séquences sélectionnées au cours de l'étape précédente à l'exception des séquences trop courtes (dont les résidus recouvraient moins de 80 % des résidus de la séquence d'intérêt).

Comme nous essayions de modéliser le plus possible de résidus de la séquence cible, nous avons écarté les séquences trop courtes sachant d'office qu'elles n'allaient pas être alignées sur toute la longueur de la séquence cible.

L'ensemble 2 a été produit à partir de l'ensemble 1 duquel ont été soustraites les séquences trop similaires entre elles. Ce deuxième ensemble a été créé au moyen du programme PURGE (Lawrence, Altschul *et al.* 1993) de manière à obtenir un ensemble d'au moins 5 séquences.

En outre, chacun des deux ensembles générés devait contenir au maximum 50 séquences (pour que les alignements ultérieurs ne prennent pas trop de temps). S'ils en contenaient plus, les 50 premières séquences ont été gardées.

Enfin, nous nous sommes assurés que chaque ensemble contenait le *template* et la séquence d'intérêt. Dans le cas contraire, nous avons veillé à les rajouter dans le deux ensembles de séquences.

2.2. *ALIGNEMENT DE SEQUENCES*

Dans le cadre de ce travail, nous avons utilisé :

- d'une part, l'alignement pairé généré par PSI BLAST entre la séquence cible et le *template* au cours de la recherche en base de données.

Psi BLAST alignant plus de résidus que BLAST, et pour une précision identique (Dunbrack 1999) , nous avons choisi l'alignement qu'il proposait à convergence du nombre de séquences.

- d'autre part, les alignements multiples effectués sur les deux ensembles précisés ci-dessus par divers programmes :Match-Box, ClustalW, Multalin, Dialign et PIMA. Un alignement pairé *target /template* a été extrait de chacun des alignements multiple.

Les alignements pairés *target-template* ont été utilisés pour la construction des consensus. Ces derniers feront l'objet du point suivant.

2.2.1. Elaboration du consensus

Comme nous l'avons signalé, la construction d'un consensus à partir de plusieurs alignements semble augmenter la confiance des résultats (*i.e.* la confiance des régions prédites comme étant structuralement conservées). C'est pourquoi la stratégie de modélisation développée durant ce travail a privilégié l'usage de consensus. Pour évaluer notre méthode, nous avons voulu comparer le modèle construit à partir de l'alignement fourni par PSI BLAST à ceux qui ont été générés à partir de consensus (Dunbrack 1999).

Pour chaque cas-test, deux consensus ont été construits :

- le premier (consensus 1) était issu de **8 alignements pairés** entre les séquences cible et de référence. Ceux-ci provenaient des alignements multiples réalisés par 4 programmes (Match-Box, Clustal W, Dialign et Multalin)sur les ensembles de séquences 1 et 2.
- Le deuxième (consensus 2) résultait, quant à lui, de **13 alignements pairés** *target-template* : les 8 alignements décrits ci-dessus, 4 alignements pairés extraits d'alignements multiples exécutés par PIMA (deux alignements par ensemble de séquences)ainsi que l'alignement de PSI-BLAST.

2.2.2. Réalisation et description d'un consensus

Pour élaborer chaque consensus, les alignements pairés considérés ont d'abord été édités dans un éditeur de séquences et sauvés sous format FASTA. Cet éditeur a été ouvert en utilisant Seaview sous UNIX.

Chaque alignement a alors été réajusté à la séquence d'intérêt du premier alignement de façon à obtenir, d'un côté, la séquence d'intérêt, et de l'autre, l'ensemble des séquences de référence (il s'agit du même *template*) qui lui ont été alignées par les différents programmes. La séquence du *template* a ensuite été chargée. C'est sur cette séquence qu'a été construit le consensus.

2.2.3. Conditions pour le choix de la position des acides aminés

Les conditions à remplir pour le choix de la position de chaque acide aminé lors de l'élaboration d'un consensus étaient les suivantes :

Premièrement, les zones d'alignement les plus sûres devaient être définies et gelées en fonction du nombre maximal d'acides aminés identiques situés à la même position dans le *template* **et** prédits à la même position dans les alignements.

Par exemple, ci-dessous, le résidu A colorié en jaune est aligné 5 fois à la même position et se situe à la même place dans la séquence du *template* (entre les résidus I et T)

```
query          QAVEFTPADP AENEIQVENK AIGINFIDTY IRSGLYP.PP SLPSGLGTEA
gi|1942871     SIEEIEVAPP KAHEVRIKII ATAVCHTDAY TLSGADP.EG CFPVILGHEG
gi|1942871     IEEIEVA.PP KAHEVRIKII ATAVCHTDAY TLSGADP.EG CFPVILGHEG
gi|1942871     SIEEIEVAPP KAHEVRIKII ATAVCHTDAY TLSGADP.EG CFPVILGHEG
gi|1942871     .VAWEAGKPL SIEEIEVAPP KAHEVRIKII ATAVCHT.DA YTLSGADPEG
```

```
gi|1942871      SIEEIEVAPP KAHEVRIKII ATAVCHTDAY TLSGADPEGC FPVILGHEGA
gi|1942871      .......... .......... .......... .......... ..........
gi|1942871      .IEEIEVAPP KAHEV.IKII ATAVCHTDAY TLSGAD..EG CFPVILGHEG
gi|1942871      PLSIEEIEVA PPKAHEVRIK IIATAVCHTD AYTL....EG CFPVILGHEG
consensus       SIEEIEVAPP KAHEVRIKII ATAVCHTDAY TLSGADP.EG CFPVILGHEG
```

S'il y avait incertitude (*i.e.* si ce nombre maximal était identique pour des résidus situés à la même position dans la séquence mais alignés à des positions différentes), il fallait sélectionner les zones alignées les plus longues.

Dans l'exemple présenté ci-dessous, le résidu R colorié en jaune est aligné deux fois à la position choisie dans le consensus. Ce résidu est aussi aligné deux fois à la position suivante. La première position a été choisie car la zone (en gris foncé) que les deux programmes alignaient identiquement autour du résidu considéré (colorié en jaune) était plus longue que celle qui entourait l'autre position (en gris clair).

```
query           CQWAK.ALG. .A.KLIGTV. GTAQKAQS.. ..A..LK... AG.AWQVINY
gi|1942871      IMGCK.VAGA .S.RIIGVD. INKDKFAR.. ..A..KE... FG.ATECINP
gi|1942871      IMGCK.VAG. .A.SRIIGV. DINKDKFAR. .....AK... EFG.ATECIN
gi|1942871      IMGCK.VAG. AS.RIIGVD. INKDKFAR.. ..A..KE... FG.ATECINP
gi|1942871      GLGGV.GLA. .V.IMGCKV. AGASRIIGVD INK..DKFA. ..RAKE...F
gi|1942871      MGCKV.AGA. .SRIIGVDI. NKDKFARA.. ..K..EF... GA.TECINPQ
gi|1942871      .......... .......... .......... .......... ..........
gi|1942871      GCKVA.GAS. .R.IIG..VD INKDKFA... ...KEFG... AT.ECINPQD
gi|1942871      IMGCK.VAG. .A.SRIIGV. DINKD.AR.. ..A..KE... FG.ATECI..
consensus       IMGCK.VAG. .ASRIIGVD. INKDKFAR.. ..A..KE... FG.ATECINP
```

Si ces zones étaient de longueur identique, il fallait alors choisir le maximum d'acides aminés identiques situés à la même position dans les alignements mais pas nécessairement par rapport aux acides aminés contigus dans la séquence du *template.*

Enfin, si certains résidus n'étaient pas encore alignés, il fallait, autant que possible, les aligner en comptant le nombre d'acides aminés identiques à l'acide aminé de la séquence d'intérêt pour une position donnée.

2.2.4. Attribution des scores

A chaque position alignée dans le consensus, un score directement proportionnel au nombre de résidus situés à la même position dans le *template* a été attribué.

- Pour le consensus 1, les scores étaient compris entre 0 (aucune position alignée ou une position alignée par un seul programme) et 8 (résidus situés à des positions identiques dans le *template* et dans tous les alignements pairés)
- Pour le consensus 2, les scores allaient de 0 à 13.

2.3. MODELISATION

Les différents alignements pairés obtenus ci-avant ont été réalisés en vue d'être utilisés pour la modélisation de chaque protéine cible.

Pour chaque cas-test, trois modèles ont été construits :

- le premier, à partir du consensus 1,
- le deuxième à partir du consensus 2,
- le troisième à partir de l'alignement pairé produit par PSI BLAST.

2.3.1. Construction de chaque modèle à partir de l'alignement target-*template*

Chaque alignement entre la séquence cible et le *template* ayant été réalisé, la modélisation de la séquence d'intérêt pouvait être accomplie sur base de l'un de ceux-ci.

Pour ce faire, chaque alignement a été transféré dans un fichier spécialement conçu pour le programme de modélisation MODELLER.

Ensuite, la séquence cible a été modélisée automatiquement par le programme MODELLER, avec les paramètres par défaut.

2.4. EVALUATION DES MODELES

Chaque modèle a été évalué par deux programmes d'évaluation : Procheck, Whatcheck. Pour rappel, Procheck et Whatcheck évaluent les caractéristiques géométriques du modèle en les comparant à la distribution statistique de ces mêmes caractéristiques dans les meilleures structures connues. Cependant, le fait que les caractéristiques des modèles satisfassent aux critères d'évaluation définis par ces

programmes n'établit en aucun cas que le modèle est correct. Si ces critères ne sont pas remplis (*i.e.* si certaines caractéristiques géométriques se situent en dehors de la distribution statistique observée dans les structures connues), il faudrait s'assurer que ces invraisemblances se retrouvent dans les régions qui présentent les mêmes caractéristiques structurelle inhabituelles du *template*. Si ce n'est pas le cas, c'est que le modèle comporte des erreurs ou que de nouvelles invraisemblances sont apparues parallèlement à l'apparition de caractéristiques fonctionnelles différentes de celles du *template*. Cette dernière possibilité ne peut être vérifiée que par la comparaison à la structure réelle.

2.5. VALIDATION DES MODELES

Pour valider les modèles, le RMSD global entre chaque cas test et sa structure réelle a été calculé. Le RMSD permet de mesurer si deux structures sont proches. Dans notre cas, nous voulions vérifier si chaque modèle se rapprochait de la structure réelle de chaque protéine considérée. Pour ce faire, nous avons utilisé le logiciel Insight II. Celui-ci nous a permis

- de visualiser les modèles de chaque cas-test (à partir des coordonnées stockées dans un fichier créé par MODELLER) et leur structure réelle (sur base du fichier PDB).
- de superposer les traces (correspondant aux Cα) de chaque modèle à celles de la structure cristallographique de la protéine correspondante.

A partir de cette superposition, le RMSD global a été calculé automatiquement.

Un RMSD local, c'est-à-dire pour chaque paire de résidus superposés, a également été calculé. En effet, lorsque certaines zones ne sont pas prédites précisément (ces zones correspondent souvent aux *loops* généralement peu conservés, ou aux extrémités), leur modélisation est assez aléatoire. Ceci a pour conséquence de faire croître très rapidement le RMSD global. Dès lors, ce dernier ne nous donne qu'une idée très approximative de la validité du modèle. Le RMSD local, quant à lui, donne une idée beaucoup plus précise de la précision de la prédiction. Puisqu'il est calculé pour chaque résidu, il permet de déterminer si certaines régions ont été prédites correctement.

Nous avons calculé un RMSD local pour chaque paire de segments en balladant une fenêtre de 9 résidus sur l'alignement modèle-structure. Ensuite, nous avons défini cinq

zones de RMSD (0.0-0.5 Å,0.5-1.0 Å,1.0-1.5 Å,1.5-2.0 Å,2.0-10 Å) puis nous avons calculé les pourcentages de présence de chaque zone dans l'alignement modèle-structure. Nous nous retrouvons alors avec une distribution de RMSD correspondant aux différentes fenêtres. Au plus le pourcentage de fenêtres dans les zones de faible RMSD est grande, au plus le modèle est proche de la structure cristallographique, localement.

CHAPITRE V: RESULTATS ET DISCUSSION

1. SELECTION DES CAS TESTS

Voici, ci-dessous une brève description des cas-test qui ont été modélisés. Cette description reprend, dans l'ordre, le nom PDB de chaque séquence d'intérêt, sa fonction et l'organisme duquel elle est issue :

- 1AMY : α-1,4 glucan-4-glucanhydrolase (α-amylase) de*Hordeum vukgare*
- 1LXA : ADP N-acétyl glucosamine acétyltransférase de *Pseudomonas aeruginosa*
- 1BMTA : Méthionine synthase (domaines se liant à la vitamine B12) : chaîne A de *Escherischia coli*
- 3PTE : D-alanyl – Dalanine carboxypeptidase (transpeptidase) *Streptomyces* sp R161
- 1D2F : aminotransférase probable, enzyme qui dégrade l'inducteur du système du maltose chez *Escherischia coli.*
- 1QORA : quinone oxydoréductase complexée au NADPH de *Escherischia coli.*
- 1DUPA : déoxyuridine 5'-triphosphate nucléotide synthase (DUTPase) de *Escherischia coli.*
- 1OXA : cytochrome P450 de *Saccharopylospora erythraea*
- 1NEC : NAD(P)H nitroréductase insensible à l'O2 de *Entrerobacter cloacae*

L'ordre dans lequel nous présentons les résultats permet de classer les modèles obtenus des moins fiables aux plus fiables. Cet ordre repris dans la discussion permettra une meilleure compréhension de l'ensemble des résultats et des conclusions que l'on peut en tirer.

2. MODELISATION

2.1. *1AMY (α-1,4 GLUCAN-4-GLUCANHYDROLASE (α-AMYLASE) DE* HORDEUM VUKGARE)

La recherche en banque de données effectuée au moyen de PSI-BLAST nous a permis de sélectionner comme *template* une α-amylase de *Tenebrio molitor* : 1JAE. Le pourcentage d'identités (calculé par ALIGN) entre ces deux séquences est de 18,4 %.

- le **modèle 1** est issu de l'alignement consensus 1 (consensus de 8 alignements pairés avec 4 programmes d'alignement)
- le **modèle 2** a été produit à partir du consensus 2 (consensus de 13 alignements pairés avec 6 programmes d'alignement + l'alignement de PSI-BLAST)
- le **modèle PSI** a été élaboré à partir de l'alignement pairé de PSI-BLAST.

Nous garderons ces dénominations pour les modèles de toutes les protéines-test.

2.1.1. Comparaison des modèles à la structure réelle

Les RMSD calculés pour les trois modèles construits pour cette première protéine sont présentés ci-dessous. Il est important de noter que plus le RMSD est faible, plus la structure du modèle se rapproche de la structure réelle.

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI - BLAST
11,90	11,77	12,28

Nous remarquons que pour cette protéine, le modèle 2 semble montrer une très légère amélioration par rapport aux deux autres modèles. Cependant, comme les RMSD sont très élevés, les trois modèles sont de toute façon incorrects. On considère généralement qu'un modèle est bon si son RMSD par rapport à la structure réelle est inférieur à 3,8 Å (Jones and Kletwegt, 1999); d'autre part, nous considérerons les modèles ayant un RMSD de plus de 7 Å comme très mauvais car dans la zone de RMSD allant de 3.5 à 7 Å, de grandes erreurs très localisées peuvent influencer fortement le RMSD global

(voir le cas de **3pte**). Dans ce dernier cas, les distributions des RMSD locaux seront plus informatifs.

Les pourcentages de résidus pour lesquels le RMSD local se situe dans le classes définies ci-dessous (voir tableau) donne une idée de l'exactitude du modèle.

RMSD local (Å)	% résidus Modèle 1	% résidus Modèle 2	% résidus Modèle PSI-BLAST
0,000-0,499	10,67	17,66	13,88
0,500-0,999	11,17	15,38	12,85
1,000-1,499	4,96	4,56	4,37
1,500-1,999	5,96	5,70	7,46
>2	67,26	56,70	61,44

Les résultats montrent que dans les trois modèles, le pourcentage de résidus pour les quels le RMSD **local** est supérieur à 2 Å représente plus de la moitié des résidus modélisés. Or , au dessus de cette distance, on considère que les modèles sont trop éloignés de la structure réelle (Briffeuil, 1998). Dès lors, bien qu'une légère amélioration soit observée pour les modèles 1 et 2 par rapport au modèle PSI-BLAST, la qualité du modèle reste très médiocre.

2.1.2. Vérification de la vraisemblance des modèles

La vraisemblance des angles φ et ψ des trois modèles a été vérifiée en utilisant le programme Procheck.

Ne sont montrés ici (et pour les protéines qui suivent) que les pourcentages des valeurs d'angles φ et ψ situés dans des zones de haute probabilité (i.e. en dessous d'un écart-type par rapport à la moyenne des angles φ et ψ présents dans la plupart des protéines) et dans les zones non permises (i .e. en dehors de 4 écarts-types). Les écarts-types sont dénommés par la lettre « s » ; Le modèle construit à partir de l'alignement de PSI-BLAST est noté « modèle PSI ».

Si les modèles contiennent 90 % des valeurs d'angles φ et ψ comprises entre la moyenne et un écart-type, on peut dire qu'ils satisfont aux critères définis par le programme (Vriend and Sander 1993; Laskowski, Rullmannn *et al.* 1996). Il faut

néanmoins garder à l'esprit qu'un modèle n'ayant aucune signification biologique peut correspondre à de tels critères.

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	69,0	71,2	75,2
Valeurs non permises >4s	5,4	2,4	1,9

Les résultats montrent que les valeurs des angles de torsion du modèle issu de l'alignement de PSI-BLAST se rapprochent plus des valeurs attendues que celles des autres modèles. Ceci est dû au fait que PSI-BLAST n'aligne que les résidus qu'il considère comme similaires et, donc, restreint la modélisation aux zones les plus sûres, *a priori*. Pour la raison qui vient d'être évoquée, les valeurs non permises de ces angles sont moindres que pour les deux autres modèles. Toutefois, remarquons que le modèle 2 comporte deux fois moins de valeurs non permises que les modèle 1.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	84,8
Valeurs non permises (>4 s)	0,0

Les pourcentages des angles de la structure réelle ne satisfont pas aux critères édictés par le programme d'évaluation, ce qui prouve bien que l'évaluation statistique de ces valeurs n'est pas absolue.

2.2. *1LXA : ADP N-ACETYL GLUCOSAMINE ACETYLTRANSFERASE D'* ESCHERISCHIA COLI

Au cours de la recherche en banque de données, le *template* défini pour la protéine cible était 1XAT : une hexapeptide xenobiotic aminotransférase. Le pourcentage d'identités calculé par ALIGN pour ces deux séquences est de 18,2%.

2.2.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI
16,05	17,09	13,47

Les RMSD présentés ci-dessous sont très médiocres pour les trois modèles. Bien que le RMSD relatif au modèle PSI-BLAST soit beaucoup plus faible que celui qui a été calculé pour les autres modèles, il reste beaucoup top élevé que pour rendre compte d'une éventuelle fiabilité du modèle. Nous avons, en effet, montré que pour une protéine (1AMY) dont les modèles présentaient un RMSD global d'environ 11 Å, les pourcentages de résidus pour lesquels les RMSD locaux dépassaient 2 représentaient plus de la moitié du total des résidus de la protéine cible. Dès lors, les RMSD locaux n'ont pas été calculés pour la présente protéine. Les deux protéines, 1AMY et 1LXA sont comparables puisqu'elles partagent environ 18 % d'identités avec leur *template*.

2.2.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	70,9	63,0	73,9
Valeurs non permises	0,8	0,0	0,7

Les valeurs d'angles ϕ et φ se trouvant dans les zones de haute probabilité sont très éloignées du seuil des 90 %, ce qui nous laisse penser qu'elles comportent des invraisemblances.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	85,1
Valeurs non permises	0,5

Comme pour la protéine 1AMY, la structure réelle présente aussi un pourcentage de valeurs hautement permises inférieur au seuil du programme (90%).

2.3. 1BMTA :METHIONIE SYNTHASE (DOMAINES SE LIANTT A LA VITAMINE B12) :CHAINE A

Le *template* sélectionné pour cette protéine lors de la recherche en base de données était une méthylmalonyl CoA mutase de Propionibacterium freudenreichii shermanii : 5REQ. Les deux protéines partagent un pourcentage d'identités de 9,9 %.

2.3.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle PSI -BLAST
21,19	19,60

Comme les alignements opérés pour le modèles 1 et le modèle PSI-BLAST étaient très médiocres (chaque programme alignait les séquences trop différemment) et que les RMSD globaux calculés pour ces modèles étaient très élevés, nous n'avons pas jugé utile de construire un modèle à partir du consensus 2. En effet, bien qu'une amélioration ait été probable dans ce cas, il était impossible qu'elle augmente de façon significative la fiabilité du modèle. Un tel résultat doit être imputé, en grande partie, au pourcentage d'identités entre la séquence d'intérêt et le *template* : il se situe dans la *Midnight Zone*. Dans cette zone, la question de la faisabilité d'un alignement se pose. Dès lors, la modélisation par homologie est peut-être inappropriée en deçà d'un certain pourcentage d'identités.

Puisque le RMSD global des modèles élaborés était fort élevé, il était inutile de calculer les RMSD locaux leur correspondant.

2.3.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	79,2	79,7
Valeurs non permises	2,3	2,6

L'analyse des pourcentages de valeurs d'angles φ et ψ montre que les valeurs hautement favorables sont très éloignées du seuil d'évaluation proposé par le programme. De plus, 2 % des valeurs des angles considérés sont très défavorables.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	93,6
Valeurs non permises	0,0

Lorsque l'on compare les valeurs du paragraphe précédent aux valeurs obtenues pour la structure réelle, une nette différence est observée, ce qui nous laisse penser que les modèles obtenus pour la protéine considérée sont très médiocres. Cependant, cette affirmation n'aurait pas pu être prouvée si nous ne connaissions pas la structure réelle de ladite protéine. Ce qui souligne encore une fois la difficulté de validation d'un modèle théorique. Néanmoins, les alignements étant déjà très mauvais, on pouvait s'attendre à des modèles très peu fiables.

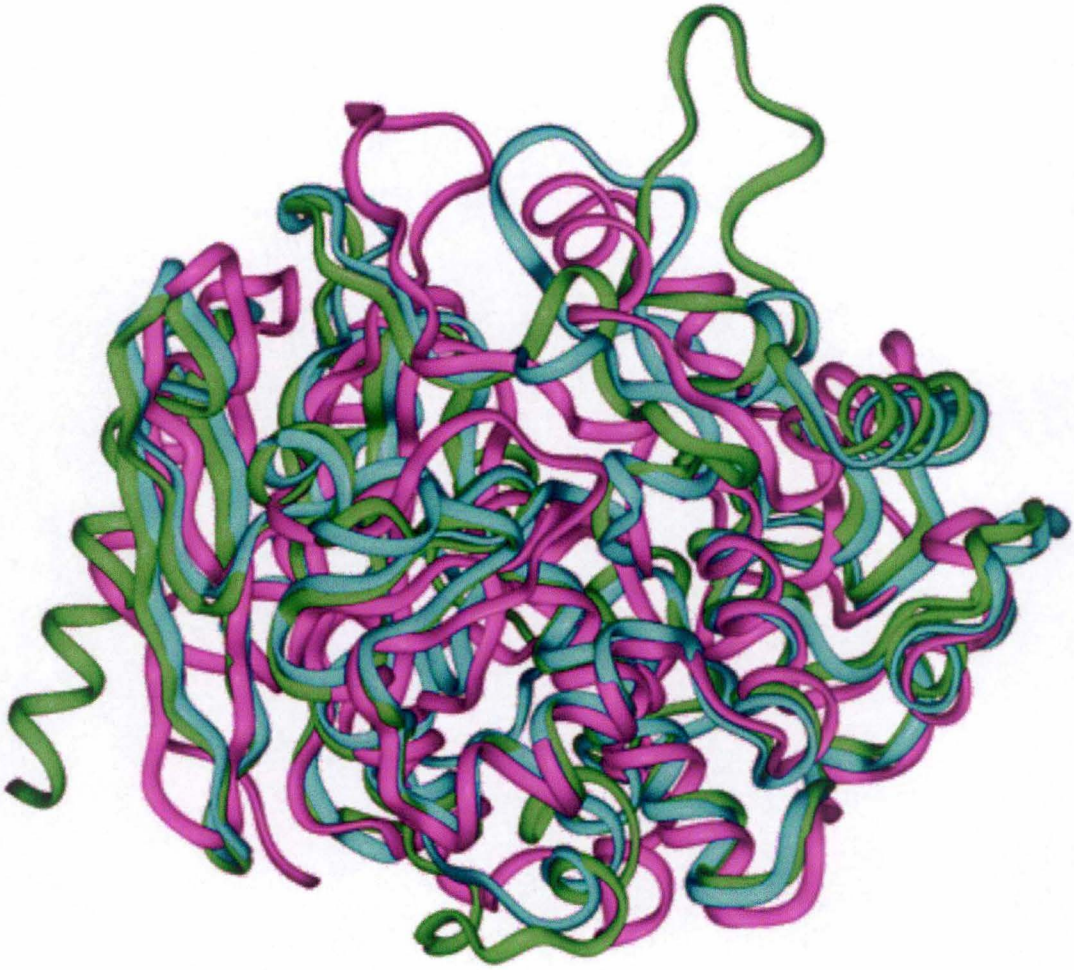


Figure 12. Superposition de structures 3D de 3pte (modèle 2 en vert, modèle PSI en bleu et structure cristallographique en magenta). On peut voir dans la partir supérieure que modèle 2 se décalle fortement de la structure cristallographique.

2.4. *3PTE : D-ALANYL – DALANINE CARBOXYPEPTIDASE (TRANSPEPTIDASE)*

STREPTOMYCES *SP R161*

Cette transpeptidase partage 20 % d'identités avec 1GCE, une β-lactamase d'*Enterobacter cloacae*.

2.4.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI -BLAST
6,79	6,61	4,86

Le RMSD global du modèle PSI-BLAST est significativement plus faible que les RMSD des deux autres modèles. Rappelons cependant que PSI-BLAST n'aligne que les résidus qu'il considère comme similaires.

RMSD local (Å)	Modèle 1	Modèle 2	Modèle PSI-BLAST
0,000-0,499	16,72	21,67	17,72
0,500-0,999	23,89	22,91	21,43
1,000-1,499	10,79	8,98	10,58
1,500-1,999	7,40	7,43	6,35
>2	41,19	39,01	43,92

En revanche, lorsque l'on examine les pourcentages de résidus en fonction des RMSD locaux, on se rend compte que contrairement à ce que suggéraient les RMSD globaux, le modèle 2 est plus proche de la structure réelle que le modèle PSI-BLAST. En effet, on y retrouve un plus grand nombre de résidus pour lesquels le RMSD est inférieur à 1Å que dans le modèle PSI-BLAST. Le fait que le RMSD global soit plus petit pour le modèle PSI-BLAST s'explique de la manière suivante : dans une des régions mal modélisées, le modèle PSI-BLAST présente une structure qui se rapproche de l'hélice α de la structure réelle. Par contre, pour le modèle 2, cette région s'éloigne très fort de la structure réelle, ce qui a pour effet l'augmentation rapide du RMSD global. Cependant, d'autres parties y sont mieux modélisées que pour le modèle PSI-BLAST. Le modèle 1 est comparable au modèle PSI-BLAST.(Figure 12)

L'utilité de la mesure des RMSD locaux en vue de mieux comparer des modèles se révèle de toute sa splendeur.

2.4.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	73,4	76,8	76,7
2,1	3,3	2,4	2,1

Les valeurs d'angles de torsion dans les zones fortement permises sont bien en deça de 90%, et ce ,pour les trois modèles.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	92,8
Valeurs non permises	2,4

Lorsque l'on effectue la comparaison des modèles à la structure réelle, sur base des valeurs d'angles de torsion, on voit qu'il existe encore une nette différence. Encore une fois, une minimisation d'énergie avec un seuil de convergence peu contraignant pourrait améliorer les scores obtenus pour les critères géométriques.

2.5. *1D2F : AMINOTRANSFERASE PROBABLE, ENZYME QUI DEGRADE L'INDUCTEUR DU SYSTEME DU MALTOSE CHEZ* ESCHERISCHIA COLI.

Le *template* utilisé pour la modélisation de cette protéine est une aspartate aminotransférase de *Thermus aquaticus thermophilus* : 1BKG. Le pourcentage d'identités partagé par ces deux protéines est de 20,7 selon ALIGN.

2.5.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI -BLAST
3,39	2,91	2,79

L'analyse des RMSD globaux calculés pour cette protéine montre une très petite amélioration lorsque l'on compare le modèle PSI-BLAST aux deux autres modèles. L'amélioration est plus nette pour le modèle PSI-BLAST lorsqu'il est comparé au modèle 1.

RMSD local (Å)	% résidus Modèle 1	% résidus Modèle 2	% résidus Modèle PSI
0,000-0,499	26,59	25,97	29,09
0,500-0,999	25,76	28,45	32,41
1,000-1,499	8,31	16,30	13,85
1,500-1,999	7,48	12,98	12,19
>2	31,86	16,30	12,47

Les pourcentages de résidus pour lesquels le RMSD est inférieur à 1 sont plus élevés pour le modèle PSI-BLAST que pour les modèles 1 et 2. Cependant il faut remarquer que :

- globalement, on observe une nette amélioration du modèle 2 par rapport au modèle 1 puisqu'on y retrouve environ deux fois moins de résidus de RMSD supérieur à deux.
- PSI-BLAST n'alignant que les régions qu'il estime similaires, il est normal que le modèle qui en découle contienne moins d'invraisemblances puisque les extrémités non alignées par ce programme n'ont pas été modélisées.

2.5.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles	Pourcentage d'angles	Pourcentage d'angles

	phi et psi modèle 1 (%)	phi et psi modèle 2 (%)	phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	76,4	75,5	77,6
Valeurs non permises	1,5	1,7	1,5

On n'observe pas de différence significative dans les trois modèles en ce qui concerne les valeurs d'angles φ et ψ hautement favorables ni pour les valeurs non permises.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	88,3
Valeurs non permises	0,3

Pour la structure réelle de cette protéine, le pourcentage des valeurs des angles considérés sont plus faibles que ce à quoi l'on s'attend et quelques valeurs non permises sont observées. Ces observations signifient, encore une fois, que les seuils statistiques définis par Procheck ne sont pas toujours valables.

2.6. *1QORA : QUINONE OXYDOREDUCTASE COMPLEXEE AU NADPH DE* ESCHERISCHIA COLI

Cette enzyme partage 23,5 % d'identités avec 1 TEH, une alcool déshydrogénase de classe III, plus spécifiquement une formaldéhyde déshydrogénase dépendante de glutathione humaine.

2.6.1. Comparaison des modèles à la structure réelle

RMSD modèle1	RMSD modèle2	RMSD modèle PSI-BLAST
3,769	3,25	4,28

Comparés au modèle PSI-BLAST, les modèles 1 et 2 se rapprochent plus de la structure réelle, globalement. Une légère amélioration est observée pour le modèle 2 par rapport au modèle 1.

RMSD local (Å)	Modèle 1	Modèle 2	Modèle PSI-BLAST
0,000-0,499	14,19	9,43	11,11
0,500-0,999	30,32	39,94	42,09
1,000-1,499	17,42	20,44	18,52
1,500-1,999	10,00	9,12	5,05
>2	28,06	21,07	23,23

Cependant, au vu des pourcentages de résidus calculés en fonction des RMSD locaux, on ne peut plus différencier le modèle 2 du modèle PSI-BLAST. Cependant, PSI-BLAST n'aligne que les résidus situés dans la zone qui lui semble similaire, ce qui diminue le risque de mauvaise modélisation, et donc les RMSD.

2.6.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	72,8	73,5	74,1
Valeurs non permises	1,1	1,4	73,5

Les valeurs caractérisant les angles de torsion considérés sont rencontrées en plus grand nombre dans les zones hautement probables pour le modèle PSI que pour les autres modèles.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	91,8
Valeurs non permises	0,0

Les valeurs d'angles de torsion correspondent aux critères d'évaluation par le programme utilisé.

2.7. *1DUPA :DEOXYURIDINE 5'-TRIPHOSPHATE NUCLEOTIDE SYNTHASE (DUTPASE) DE* ESCHERISCHIA COLI.

Cette protéine partage 29 % d'identités avec son *template*, 1DUT, une polyprotéine protéase (rétropepsine). Ce *template* est codé par le gène *pol* du virus d'immunodéficience féline. Ses fonctions précises sont : réverse trancriptase, DUTPase, ribonucléase H.

2.7.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI
3,00	2,99	4,47

Les RMSD globaux calculés pour le modèle 1 et pour le modèle 2 sont quasi identiques. Ils sont significativement plus faibles que celui qui a été calculé pour le modèle PSI-BLAST. Dans ce cas, tout particulièrement, la mesure des RMSD locaux prend tout son sens : elle permet de différencier deux modèles qui, en apparence, étaient d'une précision identique.

RMSD local (Å)	Modèle 1	Modèle 2	Modèle PSI-BLAST
0,000-0,499	35,34	32,26	27,50
0,500-0,999	16,38	17,74	16,67
1,000-1,499	9,48	15,32	11,67
1,500-1,999	25,86	14,52	14,17
>2	12,93	20,16	30,00

Le tableau ci-dessus révèle que le modèle 1 se rapproche un peu plus de la structure réelle que le modèle 2 car il contient moins de résidus pour lesquels le RMSD dépasse 2 Å. Ces deux modèles sont bien meilleurs que le modèle PSI-BLAST. De plus, le modèle

construit à partir de l'alignement fait par PSI-BLAST contient beaucoup de résidus dont le RMSD local est supérieur à 2 Å.

2.7.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	79,2	84,2	79,6
Valeurs non permises	0,0	4,0	3,1

Les valeurs caractérisant les angles de torsion considérés sont rencontrées en plus grand nombre dans les zones hautement probables pour le modèle 2 que pour les autres modèles.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	93,8
Valeurs non permises	1,8

La structure réelle de la protéine étudiée compte des angles de torsion dont les valeurs situées dans le zones hautement favorables dépassent le seuil de 90 % défini par PROCHECK, ce qui montre que les modèles ne sont pas optimaux. Une minimisation d'énergie pourrait éventuellement, dans ce cas précis, être utile pour corriger les valeurs d'angles ϕ et φ erronées.

2.8. 1OXA : CYTOCHROME P450 DE SACCHAROPYLOSPORA ERYTHRAEA (CHAINE A)

La recherche en banque de données nous a permis de déterminer un *template* pour notre protéine : 1CMN, cytochrome P450. Le pourcentage d'identités partagé par les deux séquences est de 30,3 %.

2.8.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle 2	RMSD modèle PSI -BLAST
3,90	3,84	3,92

Le RMSD global est légèrement plus faible pour le modèle 2 par rapport aux RMSD calculés pour le modèle 1 et le modèle PSI-BLAST.

RMSD local (Å)	Modèle 1	Modèle 2	Modèle PSI-BLAST
0,000-0,499	40,31	40,85	37,44
0,500-0,999	25,26	26,07	26,88
1,000-1,499	11,99	12,78	12,06
1,500-1,999	6,12	4,26	7,29
>2	16,33	16,04	16,33

L'analyse des RMSD locaux permet de montrer que le modèle 2 se rapproche plus de la structure réelle, comparé au modèle 1 et au modèle PSI-BLAST respectivement.

2.8.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle 2 (%)	Pourcentage d'angles phi et psi modèle PSI (%)
Valeurs hautement permises (1s)	85,5	80,8	82,3
Valeurs non permises	1,8	1,7	0,3

Les valeurs des angles de torsion pour les trois modèles sont inférieurs au seuil de 90% du programme. Remarquons que, si le modèle 2 est le meilleurs dans l'analyse des RMSD locaux, il présente le moins de valeurs hautement permises d'angles ϕ/ψ. De nouveau, une minimisation d'énergie avec un seuil de convergence pas trop faible pourrait

corriger les valeurs d'angles les moins probables sans pour autant changer fondamentalement les valeurs de RMSD.

	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	90,5
Valeurs non permises	0,0

On remarque que la structure réelle présente des valeurs hautement probables supérieures au seuil de 90%, ce qui démontre que les modèles pourraient peut-être encore améliorés.

2.9. 1NEC :NITROREDUCTASE DE ENTEROBACTER CLOACAE

Cette protéine partage 33 % d'identités avec 1VFRA , une NAD(P)H FMN oxydoréductase impliquée dans la bioluminescence de *Vibrio fixsheri.* Dans le cas de ce modèle, les RMSD globaux sont très bons. Dès lors, nous n'avons pas réaliser le modèle 2 car d'après les résultats précédents, celui-ci a toujours un RMSD inférieur à celui qu'obtient le plus mauvais des deux autres modèles. Comme dans ce cas, PSI-BLAST a le RMSD le plus mauvais, le modèle 2 aurait probablement été meilleur que le modèle construit à partir de PSI-BLAST.

2.9.1. Comparaison des modèles à la structure réelle

RMSD modèle 1	RMSD modèle PSI -BLAST
2,144	2,685

Le RMSD global calculé pour le modèle 1 est significativement plus faible que celui du modèle PSI-BLAST. Pour cette protéine, la création d'un consensus augmente la fiabilité du modèle. Notons que le pourcentage d'identités qu'elle partage avec son *template* se situe au delà de la Twilight Zone, ce qui rend les alignements plus faciles. Remarquons, pour ce cas-test, que notre méthode améliore de façon significative l'alignement pairé cible-*template* et donc le modèle qui en découle.

RMSD local (Å)	Modèle 1	Modèle PSI-BLAST
0,000-0,499	46,63	45,85
0,500-0,999	23,08	18,05
1,000-1,499	10,10	11,71
1,500-1,999	5,77	6,83
>2	14,42	17,56

De plus, lorsque les pourcentages de résidus présentant un RMSD donné par rapport à la structure réelle sont comparés, nous confirmons les résultats obtenus sur base du RMSD global. En effet, le modèle PSI BLAST contient un nombre plus important de résidus dont le RMSD dépasse 2Å.

2.9.2. Vérification de la vraisemblance des modèles

	Pourcentage d'angles phi et psi modèle 1 (%)	Pourcentage d'angles phi et psi modèle PSI (%)	Structure réelle de la protéine cible
Valeurs hautement permises (1s)	85,2	82,8	93,3
Valeurs non permises	1,5	1,0	0,0

Les valeurs d'angles φ etψ montrent que les modèles diffèrent de la structure réelle. On voit cependant que le pourcentage de ces angles dans les zones hautement probables se rapproche de celui de la structure réelle. Néanmoins, ce modèle ne peut pas être considéré comme « bon » par le programme d'évaluation. Une minimisation d 'énergie pourrait, dès lors, être envisagée, de façon à corriger les valeurs erronées de ces angles.

3. DISCUSSION

L'analyse des trois modèles obtenus pour chaque protéine (voir tableaux 1.a et 1.b) et leur comparaison à la structure réelle nous permettent de dégager les observations suivantes :

1.a	nom de la protéine à modéliser	template choisi	% d'identités entre la séquence-cible et le template	MODELE 1 RMSD 1 Global	MODELE 2 RMSD 2 Global	MODELE 3 RMSD PSI Global	meilleur modèle sur base du RMSD global
moins de 20 %	1AMY	1JAE	18,40%	11,9	11,77	12,28	modèle 2
	1LXA	1XAT	18,20%	16,05	17,09	13,47	modèle PSI
	1BMTA	5REQ	9,90%	21,19	pas calculé	19,6	modèle PSI
plus de 20 %	3PTE	1GCE	20%	6,79	6,61	4,86	modèle PSI
	1D2F	1BKG	20,70%	3,39	2,91	2,79	modèle PSI
	1QORA	1TEH	23,50%	3,769	3,25	4,28	modèle 2
	1DUPA	1DUT	29%	3	2,99	4,47	modèle 2
	1OXA	1CMN	30,30%	3,9	3,84	3,92	modèle 2
	1NEC	1VFRA	33%	2,144	pas calculé	2,685	MODELE 1

1.b				MODELE 1		MODELE 2		MODELE PSI		
	nom de la protéine à modéliser	template choisi	% d'identités entre la séquence-cible et le template	% des résidus dont le le RMSD local est inférieur à 1A°	% des résidus dont le le RMSD local est supérieur à 2A°	% des résidus dont le le RMSD local est inférieur à 1A°	% des résidus dont le le RMSD local est supérieur à 2A°	% des résidus dont le le RMSD local est inférieur à 1A°	% des résidus dont le le RMSD local est supérieur à 2A°	meilleur modèle selon les RMSD locaux
moins de 20 %	1AMY	1JAE	18,40%	21,84	67,26	33,04	56,7	26,73	61,44	MODELE 2
	1LXA	1XAT	18,20%	pas calculé	pas calculé	pas calculé	pas calculé	pas calculé	pas calculé	
	1BMTA	5REQ	9,90%	pas calculé	pas calculé	pas calculé	pas calculé	pas calculé	pas calculé	
plus de 20 %	3PTE	1GCE	20%	40,61	41,19	44,58	39,01	39,15	43,92	MODELE 2
	1D2F	1BKG	20,70%	52,35	31,86	54,42	16,3	61,5	12,47	MODELE PSI
	1QORA	1TEH	23,50%	44,51	28,06	49,37	21,07	53,2	23,23	MODELE PSI = MOD 2
	1DUPA	1DUT	29%	51,72	12,93	50	20,16	44,17	30	MODELE 1
	1OXA	1CMN	30,30%	65,57	16,33	66,92	16,04	64,32	16,33	MODELE 2
	1NEC	1VFRA	33%	69,71	14,42	pas calculé	pas calculé	63,9	17,56	MODELE 1

Tableau : 1.a)Comparaison des modèles à la structure réelle en calculant le RMS global(RMSD).

1.b)Comparaison des modèles à la structure réelle en calculant le RMS local(RMSD).

- de 20 % = moins de 20 % d'identité avec le template

+ de 20 % = plus de 20 % d'identité avec le template

3.1. *EXISTENCE DE LA* MIDNIGHT ZONE

Les trois premières protéines modélisées présentent un pourcentage d'identités avec leur *template* compris entre 9,9 % et 18,4 %. L'analyse des RMSD locaux et globaux des modèles obtenus met en évidence qu'aucun de ces modèles n'est fiable quelle que soit la méthode utilisée. Ces résultats reposent la question de la faisabilité de modélisation d'une protéine lorsque celle-ci partage moins de 20 % d'identités avec son *template* (cfr définition de la *midnight zone*, voir **but du mémoire**). Peut-être est-il illusoire de vouloir modéliser des protéines par homologie en-deça d'un certain seuil d'identités. Actuellement, aucune approche n'a permis d'abaisser ce seuil.

3.2. RMSD LOCAUX ET GLOBAUX

3.2.1. Tous les modèles

La comparaison des RMSD globaux montre que l'alignement fourni par notre méthode (modèles 1 et 2) est majoritairement le meilleure (tableau 1.a : 5 modèles sur 9). Cependant, il nous semble nécessaire de séparer les modèles issus de protéines partageant moins de 20 % d'identités (midnight zone) et les modèles issus de protéines partageant plus de 20% d'identités vu que la limite d'utilisation de notre méthode se situe dans la twilight zone (20-30% d'identités). Dans la suite de la discussion, nous analyserons donc plus en détails les résultats obtenus pour les protéines 3PTE, 1D2F, 1QORA, 1DUPA, 1OXA et 1NEC. Cette analyse nous permettra de mieux comparer notre méthode à l'alignement fourni par PSI-BLAST, méthode actuellement la meilleure d'après le CASP 3 (Dunbrack 1999).

3.2.2. Modèles corrects (%id>20%)

L'analyse des RMSD globaux des six dernières protéines de notre liste (tableau 1.a) montre que 66% des modèles les plus précis sont issus de notre méthode.

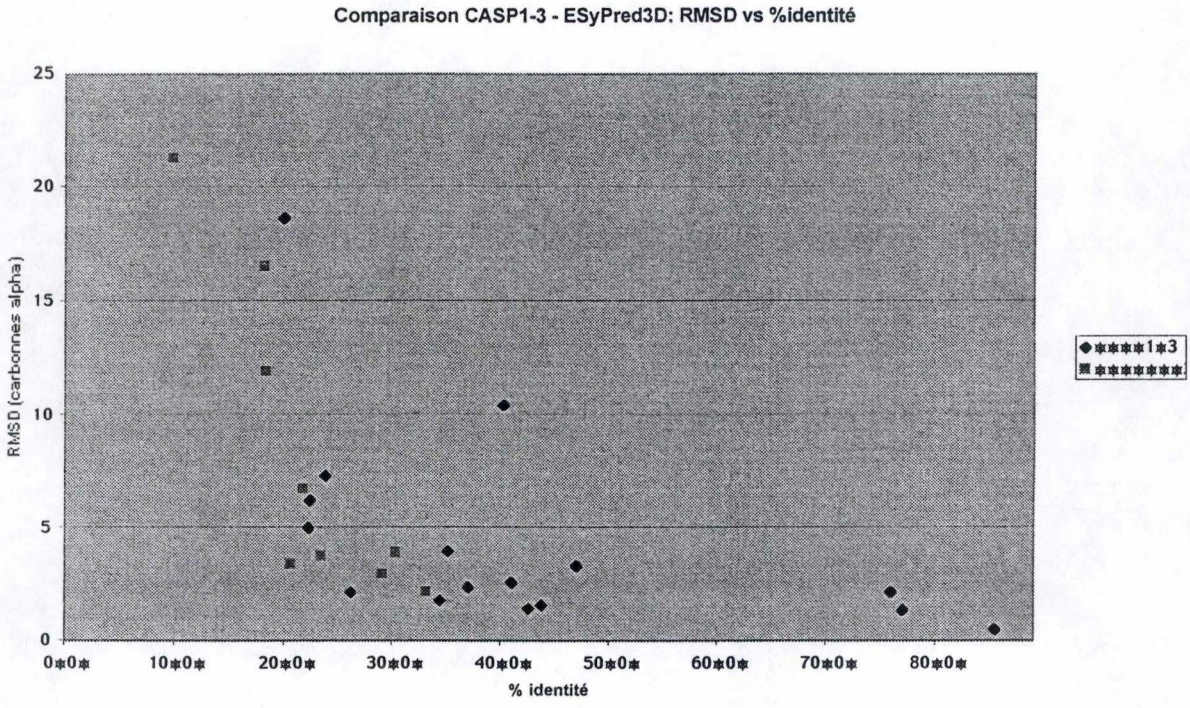


Figure 13. Comparaison des résultats obtenus aux cours des CASP 1-3 (carrés) et de ceux obtenus avec notre méthode implémentée dans leprogramme EsyPred (losanges). Le graphe représente le pourcentage d'identité en fonction du RMS global.

Si l'on analyse les résultats obtenus pour les RMSD locaux, les observations effectuées sur l'analyse des RMSD globaux sont confirmées. En effet, cinq des six meilleurs modèles ont été réalisés en créant un alignement pairé séquence-*template* basé sur un consensus de plusieurs programmes. Une comparaison des résultats des modèles 1 et 2 montre que l'utilisation de consensus de six programmes d'alignement multiple et, de l'alignement pairé target-*template* de PSI-BLAST donne de meilleurs résultats que lorsque les modèles sont construits à partir du consensus de huit alignements pairés n'incluant pas PSI-BLAST

3.3. COMPARAISON AUX CASPS

Enfin, les résultats que nous obtenons sont comparables aux *meilleurs* modèles évalués au CASP. Les modèles prédits dans le cadre du CASP constituent des modèles de référence.

Sur le graphe (figure 13), nous observons que le RMSD global calculé entre chacun des modèles diminue de façon inversement proportionnelle en fonction du pourcentage d'identités partagé par chaque cible et son *template*.

Les RMSD calculés pour les modèles décrits ci-dessus se situent aux alentours des RMSD des meilleurs modèles soumis au CASP. Il est donc probable que notre approche basée sur la création de consensus d'un grand nombre de programmes optimise l'alignement *target-template* et tout le processus de modélisation qui s'ensuit. Remarquons toutefois que :

- nous n'avons testé que quelques protéines. Nos observations ne peuvent, par conséquent, être inférées à toutes les protéines.
- les consensus ont été construits manuellement. Des erreurs d'alignements sont donc possibles. Dès lors, il est probable que l'automatisation de ce processus optimise l'alignement final.
- les RMSD relatifs à nos modèles ont été calculés pour toute la structure, y compris les parties non modélisées, alors que les RMSD correspondant aux modèles du CASP ne tenaient compte que de la partie modélisée.

CHAPITRE VI: CONCLUSIONS ET PERSPECTIVES

Au terme de ce travail, nous avons donc modélisé neuf protéines en utilisant une nouvelle approche. En effet, nous avons consacré nos efforts à optimiser l'alignement séquence-cible/*template* fourni au programme de modélisation.

Pour ce faire, nous avons utilisé :

- un alignement construit sur base de huit alignements pairés provenant de quatre programmes d'alignement (modèle 1)
- un alignement construit sur base de douze alignements pairés provenant de six programmes d'alignement auxquels nous avons ajouté l'alignement fourni par le programme PSI-BLAST (modèle 2)
- l'alignement fourni automatiquement par le programme PSI-BLAST

L'analyse des trois modèles obtenus pour chaque protéine (tableau 1) et leur comparaison à la structure réelle nous permettent de dégager les observations suivantes :

- la question de la faisabilité de modélisation d'une protéine lorsque celle-ci partage moins de 20 % d'identités avec son *template* est posée.
- Nos approches (modèles 1 et 2) fournissent des modèles plus précis que PSI-BLAST tant pour les RMSD globaux ou locaux. Par ailleurs, l'utilisation de consensus de six programmes d'alignement multiple et, de l'alignement pairé target-*template* de PSI-BLAST donne de meilleurs résultats que lorsque les modèles sont construits à partir du consensus de huit alignements pairés n'incluant pas PSI-BLAST
- Enfin, les résultats que nous obtenons sont comparables aux *meilleurs* modèles évalués au CASP, considérés comme des modèles de référence.

Dès lors, bien que nous ne puissions affirmer que notre méthode constitue la meilleure approche, ces conclusions peuvent nous mener à des investigations plus poussées.

Perspectives :

Ces investigations impliquent tout d'abord la modélisation d'un nombre significatif de protéines–test en utilisant notre méthodologie de manière automatique. Elles permettront de montrer si cette méthodologie donne des résultats plus intéressants que les méthodes actuelles.

De plus, de nombreux tests visant l'amélioration de la méthode sont à prévoir.

1. Il sera nécessaire de modifier et de tester les paramètres utilisés par les programmes de recherche en banque de données, d'alignements et de modélisation. La détermination des paramètres optimaux permettra d'améliorer les différentes étapes de la modélisation par homologie.
2. Ensuite, il faudra inclure la possibilité d'utiliser plusieurs *templates* dans le cas de protéines multidomaines. Pour construire le consensus, il sera nécessaire :
   - d'évaluer ce que peut apporter l'utilisation d'un plus grand nombre de programmes d'alignement de séquences et surtout des programmes les plus performants
   - de pondérer les alignements en donnant plus de poids aux méthodes les plus performantes. Il faudra comparer les modèles construits à partir de consensus d'alignements pondérés différemment et non pondérés afin de déterminer si la pondération est nécessaire. Si c'est le cas, il sera utile de préciser quelle sera la pondération optimale.
3. d'inclure la prédiction de structures secondaires et la reconnaissance de *fold* de façon à mieux sélectionner le (s) *template* (s) et à optimiser l'alignement
4. Bien qu'une minimisation d'énergie n'influence pas la précision du modèle, il serait quand même intéressant d'en réaliser en prenant des critères de convergence de plus en plus stricts de façon à observer s'il y a amélioration des critères géométriques sans changer fondamentalement les RMSD locaux et globaux et , si cela s'observe, à quel critère de convergence ce phénomène s'observe.
5. Si notre approche se confirme comme améliorant la précision du modèle par amélioration de l'alignement séquence/*template*, il sera envisageable d'automatiser entièrement la procédure afin d'éviter les erreurs humaines, notamment lors de l'établissement du consensus.
6. Finalement, le test par excellence pour valider une nouvelle méthode est de participer au CASP/CAFASP et de comparer les résultats obtenus par notre méthode à ceux obtenus par les autres méthodes.

BIBLIOGRAPHIE

Altschul, S. F., W. Gish, *et al.* (1990). "Basic Local Alignment Search Tool." Journal of Molecular Biology **215**: 403-410.

Altschul, S. F., T. L. Madden, *et al.* (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleid Acids Research **25**(17): 3389-3402.

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

Attwood and Parry-Smith (1999). "Bioinformatics." .

Barre, S., A. S. Greenberg, *et al.* (1994). "Structural conservation of hypervariable regions in immunoglobulins evolution." Nat Struct Biol **1**(12): 915-20.

Bates, P. A. and M. J. Sternberg (1999). "Model building by comparison at CASP3: Using expert knowledge and computer automation." Proteins **37**(S3): 47-54.

Branden and Tooze (1991). "Introduction à la structure des protéines." .

Briffeuil, P., G. Baudoux, *et al.* (1998). "Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance reliability of predictions." Bioinformatics **14**(4): 357-66.

Chothia, C. and A. G. Murzin (1993). "New folds for all-beta proteins." Structure **1**(4): 217-22.

Chothia C., L. A. M. (1986). "The relation between the divergence of sequence and structure in proteins." EMBO **5**((4)): 823-826.

Chou, P. Y. and G. D. Fasman (1974). "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins." Biochemistry **13**(2): 211-22.

Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." Nucl. As. Res.(16): 10881-10890.

Dayhoff M. O., E. R. V., Park C.M. (1972). "Amodel of evolutionary change in proteins." Atlas of protein sequence and structure. **5**: 89-99.

de Fays, K., A. Tibor, *et al.* (1999). "Structure and function prediction of the Brucella abortus P39 protein by comparative modeling with marginal sequence similarities." Protein Eng **12**(3): 217-23.

Depiereux, E., G. Baudoux, *et al.* (1997). "Match-Box server: a multiple sequence alignment tool placing emphasis on reliability." Comput. Appl. Biosci. **13**(3): 249-256.

Depiereux, E. and E. Feytmans (1992). "MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences." Comput Appl Biosci **8**(5): 501-9.

Diamant, S., A. Peres Ben-Zvi, *et al.* (2000). "Size-Dependent Disaggregation of Stable Protein Aggregates by the DnaK Chaperone Machinery." J Biol Chem.

Doolittle (1986). " Of URFs and ORFs: a primer on how to analyse derived amino acid sequences." University Science Books **19**: 15-18.

Doolittle, R. F. (1981). "Similar amino acid sequences:." chance or common ancestry. **214**: 1449-1459.

Dunbrack, R. L., Jr. (1999). "Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL." Proteins **Suppl**(3): 81-7.

Farmery, M. R., S. Allen, *et al.* (2000). "The role of ERp57 in disulfide bond formation during the assembly of major histocompatibility complex class I in a synchronized semipermeabilized cell translation system [In Process Citation]." J Biol Chem **275**(20): 14933-8.

Garnier, J., D. J. Osguthorpe, *et al.* (1978). "Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins." Journal of Molecular Biology **120**(1): 97-120.

Gibrat, J. F., J. Garnier, *et al.* (1987). "Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs." Journal of Molecular Biology **198**(3): 425-443.

Henikoff, S., E. A. Greene, *et al.* (1997). "Gene families: the taxonomy of protein paralogs and chimeras." Science **278**(5338): 609-14.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc. Natl. Acad. Sci. USA **89**: 10915-10919.

Higgins, D. G., A. J. Bleasby, *et al.* (1992). "CLUSTAL V: improved software for multiple sequence alignment." Comput Appl Biosci **8**(2): 189-91.

Huang, X. (1994). "On global sequence alignment." CABIOS **10**(3): 227-235.

Huynen, M. A. and E. van Nimwegen (1998). "The frequency distribution of gene family sizes in complete genomes." Mol Biol Evol **15**(5): 583-9.

Johnson, M. S. and J. P. Overington (1993). "A structural basis for sequence comparisons. An evaluation of scoring methodologies." J Mol Biol **233**(4): 716-38.

Jones, D. T., W. R. Taylor, *et al.* (1992). "A new approach to protein fold recognition." Nature **358**(6381): 86-89.

Jones, D. T., M. Tress, *et al.* (1999). "Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure." Proteins **37**(S3): 104-111.

Kabsch, W. and C. Sander (1983). "How good are predictions of protein secondary structure?" FEBS Lett **155**(2): 179-82.

King, R. D. (1997). "DSC: public domain protein secondary structure predication." Comput. Appl. Biosci. **13**(4): 473-474.

Laskowski, R. A., J. A. Rullmannn, *et al.* (1996). "AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR." J Biomol NMR **8**(4): 477-86.

Lawrence, E., S. F. Altschul, *et al.* (1993). "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment." Science **262**: 208-214.

Lesk, A. M. and C. Chothia (1980). "How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins." J. Mol. Biol. **136**(3): 225-270.

Lodish, B., Berk,Zipursky,Matsudaira,Darnell (1997). "Structure et fonction des protéines." Biochimie moléculaire de la cellule.: 51-100.

Luthy, R., J. U. Bowie, *et al.* (1992). "Assessment of protein models with three-dimensional profiles." Nature **356**(6364): 83-5.

Morgenstern, B., K. Frech, *et al.* (1998). "DIALIGN: Finding local similarities by multiple sequence alignment." Bioinformatics **14**(3): 290-294.

Needleman, S. B. and C. D. Wunsch (1970). "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." Journal of Molecular Biology **48**: 443-453.

Pearson, W. R. (1990). "Rapid and Sensitive Sequence Comparison with FASTP and FASTA." Methods in Enzymology **183**: 63-98.

Peitsch, M. C. (1996). "ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling." Biochem Soc Trans **24**(1): 274-9.

Pflugrath, J. W. and F. A. Quiocho (1988). "The 2 A resolution structure of the sulfate-binding protein involved in active transport in Salmonella typhimurium." J Mol Biol **200**(1): 163-80.

Rost, B. (1997). "Protein structures sustain evolutionary drift." Fold Des **2**(3): S19-24.

Rost, B. and S. O'Donoghue (1997). "Sisyphus and prediction of protein structure." Comput Appl Biosci **13**(4): 345-56.

Rost, B., C. Sander, *et al.* (1994). "PHD--an automatic mail server for protein secondary structure prediction." CABIOS **10**(1): 53-60.

Sali, A. and T. L. Blundell (1993). "Comparative protein modelling by satisfaction of spatial restraints." Journal of Molecular Biology **234**(3): 779-815.

Sanchez, R. and A. Sali (1997). "Advances in comparative protein-structure modelling." Current Opinion in Structural Biology **7**: 206-214.

Sauder, J. M., J. W. Arthur, *et al.* (2000). "Large-scale comparison of protein sequence alignment algorithms with structure alignments." Proteins **40**(1): 6-22.

Schrauber, H., F. Eisenhaber, *et al.* (1993). "Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins." J Mol Biol **230**(2): 592-612.

Sippl, M. J. (1993). "Recognition of errors in three-dimensional structures of proteins." Proteins **17**(4): 355-362.

Smith, R. F. and T. F. Smith (1992). "Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling." Protein Eng. **5**(1): 35-41.

Thompson, J. D., D. G. Higgins, *et al.* (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Thompson, J. D., F. Plewniak, *et al.* (1999). "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs." Bioinformatics **15**(1): 87-8.

Thompson, J. D., F. Plewniak, *et al.* (1999). "A comprehensive comparison of multiple sequence alignment programs." Nucleic Acids Res **27**(13): 2682-90.

Tramontano, A. (1998). "Homology Modeling with Low Sequence Identity." METHODS: A Companion to Methods in Enzymology. **14**: 293-300.

Venclovas, C., K. Ginalski, *et al.* (1999). "Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins." Proteins **37**(S3): 73-80.

Vinals, C., X. De Bolle, *et al.* (1995). "Knowledge-Based Modeling of the D-Lactate Dehydrogenase Three-Dimensional Structure." Proteins: Structure, Function, and Genetics. **21**: 307-318.

Vriend and Sander (1993). "Quality-control of protein models-directional atomic contact analysis." Journal of applied crystallography **993**(26): 47-60.