DOCTEUR EN SCIENCES

Développement d'une méthode automatique fiable de modélisation de la structure tridimensionnelle des protéines par homologie et application au protéome de Brucella Melitensis

Lambert, Christophe

Award date:
2003

Awarding institution:
Universite de Namur

FACULTES UNIVERSITAIRES
NOTRE-DAME DE LA PAIX **NAMUR**

FACULTE DES SCIENCES
DEPARTEMENT DE BIOLOGIE

Développement d'une méthode automatique fiable de modélisation de la structure tridimensionnelle des protéines par homologie et application au protéome de *Brucella melitensis*

Dissertation présentée par
Christophe Lambert
en vue de l'obtention du grade
de Docteur en Sciences

Composition du jury:
Xavier De Bolle (FUNDP)
Eric Depiereux (Promoteur, FUNDP)
Martine Raes (FUNDP)
Jean-Louis Ruelle (GlaxoSmithKline Biologicals, Rixensart)
Jacques van Helden (SCMB, ULB, Bruxelles)

2003

Rempart de la Vierge, 13
B - 5000 Namur (Belgique)

Imprimé en Belgique
ISBN: 2-87037-425-9
Dépôt légal: D / 2003 / 1881 / 37

Facultés Universitaires Notre-Dame de la Paix

Faculté des Sciences

Rue de Bruxelles, 61 B-5000 Namur, Belgique

Développement d'une méthode automatique fiable de modélisation de la structure tridimensionnelle des protéines par homologie et application au protéome de *Brucella melitensis*

Par Christophe Lambert

La connaissance de la structure tridimensionnelle (3D) des protéines est une information capitale. Néanmoins, le nombre de protéines dont la structure 3D a été déterminée expérimentalement est cent fois plus faible que le nombre de protéines connues aujourd'hui. Cet écart ne pourra pas être comblé, car les techniques expérimentales de détermination de structure (diffraction de rayons X et résonance magnétique nucléaire) sont coûteuses et lentes (un an de travail en moyenne pour une seule protéine).

Un moyen d'obtenir plus rapidement la structure 3D de protéines est de la prédire par des moyens bioinformatiques. La technique de prédiction la plus précise actuellement est la modélisation par homologie. Celle-ci est basée sur la similitude de structure entre deux protéines de séquences similaires. L'étape critique de cette méthode est l'étape d'alignement entre la séquence à modéliser et une séquence similaire de structure connue.

Notre travail a consisté tout d'abord en la conception d'une nouvelle méthode d'alignement pairé très fiable. Cette méthode a ensuite été incluse dans un système automatique de modélisation par homologie: la bonne qualité des structures prédites par le système trouve en partie son origine dans le programme d'alignement utilisé.

Enfin, nous avons appliqué notre système de modélisation automatique à la modélisation de toutes les protéines déduites du génome d'une bactérie pathogène étudiée dans notre unité de recherche: *Brucella melitensis*. Cela nous a conduit à créer une banque de données structurales et fonctionnelles consacrée au génome de cette bactérie. Cette banque de données est devenue un outil de travail indispensable pour plusieurs équipes de recherche européennes qui étudient *Brucella melitensis*.

Facultés Universitaires Notre-Dame de la Paix

Faculté des Sciences

Rue de Bruxelles, 61 B-5000 Namur, Belgium

Developing a reliable automatic method to predict the three-dimensional structure of proteins by homology modeling and application to the *Brucella melitensis* proteome

by Christophe Lambert

The three-dimensional (3D) structure of proteins is an important information. Nevertheless, the number of proteins for which the 3D structure has been determined experimentally is hundred times smaller than the number of currently known proteins. This gap will never been filled because the experimental techniques to determine protein structure (X-ray diffraction and nuclear magnetic resonance) are expensive and slow (one year of work on average for only one protein).

One way to obtain the 3D structure of proteins faster is to predict it using bioinformatics tools. The most accurate technique is currently homology modelling. It is based on the structural similarity between two proteins having similar sequences. The critical step of this technique is the alignment step between the target sequence and a similar sequence of known structure, used as template.

Our work has consisted primarily to design a new reliable pairwise alignment method. This method has been included in an automatic homology modelling system: the high quality of structures predicted by the system comes partly from the alignment program used.

Finally, we have applied our automatic modelling system to the modelling of all proteins deduced from the genome of a pathogenic bacteria studied in our research unit: *Brucella melitensis*. This has lead us to create a structural and functional databank focused on *Brucella*'s genome. This databank has become an essential tool for European research teams that study *Brucella melitensis*.

Table des matières

Index des figures

Index des tableaux

Remerciements

Comme il est de coutume à la fin d’un long et difficile travail, je vais remercier ceux sans qui il n’aurait pas été possible. Il est toujours délicat d’écrire cette partie car personne ne doit être oublié et ceux qui sont remerciés se trouveront peut-être injustement placés en telle ou telle position par rapport à un autre. De plus, l’objet du remerciement peut parfois être ironique, parfois être sincère et souvent, politiquement correct. Après ces quelques petites réflexions, voici mes remerciements.

Tout d’abord, je voudrais remercier le professeur Eric Depiereux de m’avoir accepté dans son laboratoire et de m’avoir soutenu tout au long de mon doctorat. Ces années de thèse m’ont permis de développer mes capacités d’expression, aussi bien écrite qu’orale, des capacités de gestion d’équipe et une certaine autonomie dans mon travail. Je le remercie également de m’avoir aidé dans un moment particulièrement difficile de l’année 1998.

Je remercie les professeurs Jean Vandenhaute, Xavier De Bolle et Jean-Jacques Letesson pour leurs conseils, leur curiosité scientifique et leur intérêt pour mon travail. Si je devais retenir deux mots, ce serait: "Oser rêver". En effet, si le scientifique doit connaître beaucoup de notions, seule sa créativité pourra faire avancer ses recherches.

Je ne trouverai probablement pas les mots assez forts pour remercier Guy Baudoux. Merci pour m'avoir indiqué le chemin à suivre, merci pour toutes tes connaissances bibliographiques, merci pour tes nombreux conseils, merci pour tout ce temps passé à discuter et à corriger mes manuscrits. Sans toi, ma thèse ne serait probablement pas aussi aboutie et je n'aurais certainement pas pu la mener à son terme.

Le mémoire de Nadia fut le point de départ de ma thèse. Elle a montré que mon approche sur ESyPred3D était meilleure que d'autres méthodes existantes. Sans son acharnement à trouver les erreurs, mes programmes seraient moins performants et, surtout, buggés. Tu as fait énormément de travail, que ce soit pendant ton mémoire ou pour les CASPs. Tu as toujours corrigé mes manuscrits et ton sens littéraire a beaucoup aidé la lisibilité de mes documents. Je n'oublierai jamais ces moments difficiles pendant lesquels tu étais déprimée. Je n'ai pas toujours su trouver les mots justes, mais sache que sans toi, sans tes corrections et sans tes idées audacieuses, ce travail ne serait pas ce qu'il est. Merci pour ton amitié, merci pour tout et merci pour toi.

Les mémoires de Nicolas et Jean-Marc sont à la base des améliorations de Match-Box. Je remercie particulièrement Nicolas et

Isabelle pour ce parcours que nous avons suivi et que nous suivrons ensemble. Merci pour toutes ces soirées passées à discuter, à boire et à manger. Merci pour votre amitié et à bientôt pour une course de poussettes.

Johan, Katalin, Bernard et CRo, je vous remercie pour toutes ces discussions autour d'un bon verre. Merci pour vos nombreux conseils et merci pour cette bonne ambiance au laboratoire. Un merci tout particulier à Katalin et Bernard, vous comptez beaucoup pour moi.

Je remercie Etienne d'avoir supporté ce "petit homme" parfois dérangeant et surtout dérangé. Ce "petit homme" n'aurait peut-être pas su prendre sa place en URBM sans tes conseils et sans ces nombreuses discussions dans notre bureau. Merci pour ces folies passagères qui amélioraient l'ambiance du laboratoire.

Merci à Benjamin et Aïko pour ces discussions sur la guerre en Irak, sur l'informatique et sur les différentes perspectives de la bioinformatique. Merci plus particulièrement à Benjamin pour cette fabuleuse expérience qu'est la création de notre entreprise.

Je remercie tous les membres de l'URBM pour m'avoir soutenu et accompagné jusqu'au terme de cette thèse. Vous m'avez parfois demandé beaucoup de travail et c'était difficile sur le moment même de se sentir exploité. Mais au bout du compte, ces travaux ont créé une grande expérience que je pourrai mettre à profit plus tard.

Je m'en voudrais d'oublier tous ceux sans qui je n'aurais pas pu suivre des congrès ou participer à des écoles d'hiver ou d'été en Europe ou aux Etats-Unis. Merci donc au F.R.I.A., au F.N.R.S., à l'O.T.A.N., au fonds Adrien Bauchau, à la *Fundación* Juan March et à Lambda+.

Ce travail n'aurait pas été possible sans mes parents qui m'ont toujours soutenu pendant mes études, malgré ces années difficiles en candidature, en licence et en 1998. Je ne vous remercierai jamais assez pour tout ce que vous avez fait pour moi, pour toute cette culture que vous avez mise à ma disposition, pour cette joie de vivre, et pour cet amour dans votre couple. Merci à Jérôme et Géraldine de m'avoir supporté parfois à leurs dépens. Merci pour votre amitié et merci pour toute cette richesse que vous avez apportée à ma personnalité.

Et bien non, Monique, mon amour, je ne t'ai pas oubliée, je gardais la meilleure pour la fin. Je te remercie pour tout cet amour que tu me portes, merci pour tes nombreuses relectures et tes nombreux conseils. Je n'aurais peut-être jamais commencé cette thèse sans toi. Merci pour toutes ces années passées et futures pendant lesquelles nous vivrons une aventure formidable car, "osons le dire, vivre à deux, c'est merveilleux" (Tybo and Goupil, 1996). Néanmoins, il ne faudrait pas oublier non plus que "vivre à trois c'est extra" (Tybo and Goupil, 1996).

Enfin, je remercie tous ceux que j'ai oubliés de citer ci-dessus, je les remercie collectivement pour nos discussions fructueuses et pour l'expérience qu’ils m'ont apportée. Parmi eux, je retiendrai Burkhard Rost, Carla Vinals, Cindy Castado, Didier Belhomme, Ernest Feytmans, Jacques Van Helden, Jean-Louis Ruelle, Jean-Luc Pellequier, Joëlle Jonet, José Remacle, le laboratoire de Chimie Moléculaire Structurale, le laboratoire de Chimie Théorique Appliquée, Marc Dieu, Marcus Marti-Renom, Martine Raes, Michel Dieu, Shoshana Wodak et Volker Eyrich.

Abréviations

3D	Trois dimensions ou tridimensionnel
Å	Angström
a.a.	**a**cide **a**miné
ADN	**A**cide **D**ésoxyribo**N**ucléique
ARN	**A**cide **R**ibo**N**ucléique
ASG	**A**ligné **S**ans ***G**ap*
BLOSUM	***BLO**cks **SU**bstitution **M**atrix*
bp	***b**ase **p**air*
BMMT	**B**oîte **M**odifiée par **M**atch-**T**al
BOMB	**B**oîte **O**riginale de **M**atch-**B**ox
BSMT	**B**oîte **S**électionnée par **M**atch-**T**al
CAFASP	***C**ritical **A**ssessment of **F**ully **A**utomated **S**tructure **P**rediction*
CAPRI	***C**ritical **A**ssessment of **PR**ediction of **I**nteractions*
CASP	***C**ritical **A**ssessment of techniques for protein **S**tructure **P**rediction*
CDS	***CoD**ing **S**equence*
CSS	**C**onservation de la **S**tructure **S**econdaire
COFFEE	***C**onsistency based **O**bjective **F**unction **F**or alignm**E**nt **E**valuation*
DP	***D**educed **P**rotein* ou protéine déduite
DRX	**D**iffraction des **R**ayons **X**
EMBL	***E**uropean **M**olecular **B**iology **L**aboratory*
ESyPAli	***E**xpert **Sy**stem for **P**airwise **Ali**gnment*
ESyPAliNN	***E**xpert **Sy**stem for **P**airwise **Ali**gnment using **N**eural **N**etworks*
ESyPred3D	***E**xpert **Sy**stem for **Pred**iction of **3D** protein structures*

E-value	***E**xpected **value***
FNRS	**F**onds **N**ational de la **R**echerche **S**cientifique
FORTRAN	*IBM mathematical **FOR**mula **TRAN**slation system*
FRIA	**F**onds de formation à la **R**echerche dans l'**I**ndustrie et dans l'**A**griculture
FUNDP	**F**acultés **U**niversitaires **N**otre-**D**ame de la **P**aix
GB	***G**iga**B**yte*, 1 073 741 824 octets
GCC	***G**NU **C**ompiler **C**ollection*
GCG	***G**enetics **C**omputer **G**roup*
GDUS	***G**oal **D**irected **U**nidirectional **S**earch*
GHz	**G**iga**H**ert**z**, 1 000 000 000 Hertz
GNU	***G**NU is **N**ot **U**nix*
GPCR	***G P**rotein-**C**oupled **R**eceptor*
H	**H**ydrogène
HD	***H**ard **D**isk*
HMM	***H**idden **M**arkov **M**odel*
HSP	***H**eat **S**hock **P**rotein*
HTH	***H**elix-**T**urn-**H**elix* ou hélice-coude-hélice
HTML	***H**yper **T**ext **M**arkup **L**angage*
IBM	***I**nternational **B**usiness **M**achine Inc.*
ISS	***I**ntermediate **S**equence **S**earch*
KEGG	***K**yoto **E**ncyclopedia of **G**enes and **G**enomes*
LAAO	**L-A**mino**A**cid **O**xydase
MAO	**M**ono**A**mine **O**xydase
matching_EL	***matching**_SF* tenant compte de l'**E**nvironnement **L**ocal
matching_MB	***matching*** originel de **M**atch-**B**ox
matching_PSSM	***matching**_SF* utilisant les **PSSM** générés après recherche dans la base de données *nr* du NCBI par le programme PSI-BLAST
matching_SF	***matching*** originel de Match-Box, mais **S**ans **F**iltre

	statistique
MB	***Mega*Byte**, 1 048 576 octets
MDM	***M*utation *D*ata *M*atrix**
MHz	**M**ega**H**ert**z**, 1 000 000 Hertz
MIPS	***M*illion *I*nstruction *P*er *S*econd**
µm	micromètre ou micron
mm	**mi**lli**mè**tre
MULTICS	***MULT*iplexed *I*nformation and *C*omputing *S*ervice**
MW	***M*olecular *W*eight** ou masse moléculaire
NA	**N**on **A**pplicable
NCBI	***N*ational *C*enter for *B*iology *I*nformation**
nr	***n*on *r*edundant**
OTAN	**O**rganisation du **T**raité de l'**A**tlantique **N**ord
PAM	***P*oint *A*ccepted *M*utation**
PAO	**P**oly**A**mine **O**xydase
pCDS	***p*redicted *CoDi*ng *S*equence**
PDB	***P*rotein *D*ata *B*ank**
pH	**p**otentiel **H**ydrogène
pI	**p**oint **I**soélectrique
pSCR	***p*redicted *S*tructurally *C*onserved *R*egion**
PSSM	***P*osition *S*pecific *S*coring *M*atrix**
RAM	***R*andom *A*cces *M*emory**
RBS	***R*ibosome *B*inding *S*ite**
RMN	**R**ésonance **M**agnétique **N**ucléaire
RMSD	***R*oot *M*ean *S*quare *D*eviation**
SCE	**S**omme des **C**arrés des **E**carts
screening_MB	***screening*** originel de **M**atch-**B**ox
screening_NS	***screening*** utilisant une **N**ouvelle **S**tratégie (maximisation de la longueur de l'alignement)
screening_SS	***screening*** utilisant les **S**tructures **S**econdaires

SGBD	**S**ystème de **G**estion de **B**ase de **D**onnées
SGI	***S**ilicon **G**raphics **I**nc.*
SI	**S**équence d'**I**ntérêt
SS	**S**tructure **S**econdaire
SSC	**S**équence de **S**tructure **C**onnue
trEMBL	***Tr**anslated* **EMBL**
UNIX	***U**NIX is **N**ot mult**ICS***
URBM	**U**nité de **R**echerche en **B**iologie **M**oléculaire

Avant-propos

En ce début de troisième millénaire, la biologie moléculaire vit une véritable révolution. En effet, le séquençage de toute une série de génomes d'organismes tant procaryotes qu'eucaryotes génère une quantité de séquences nucléiques et protéiques qui double approximativement tous les quatorze mois (Baxevanis, 2003). C'est pour faire face à cet immense flux d'informations que s'est développée une nouvelle discipline alliant biologie et informatique: la bioinformatique.

Des méthodes bioinformatiques ont donc été développées pour traiter les données rencontrées tant dans l'analyse des séquences nucléotidiques (provenant des génomes) que dans l'analyse des séquences protéiques. Parmi ces méthodes, citons l'assemblage de courtes séquences nucléotidiques (*reads*) pour reconstituer un génome séquencé, la prédiction des séquences codantes (CDS: *CoDing Sequence*) dans les génomes, la prédiction de la structure secondaire des séquences d'ARN, la gestion des banques de données de séquences, la recherche par similarité dans ces banques de données, l'alignement multiple de séquences protéiques, la prédiction de la structure secondaire des protéines, la prédiction de l'accessibilité au solvant des acides aminés et la prédiction de la structure tridimensionnelle (3D) des protéines.

La structure 3D des protéines est une source d'informations essentielle pour mieux comprendre leurs fonctions, leurs interactions avec d'autres substances (ligands, protéines, ADN, ...) et les effets phénotypiques des mutations (Tramontano, 1998). Quand la structure d'une protéine est bien comprise, il devient possible d'expliquer la modification de ses propriétés physico-chimiques dans différentes conditions expérimentales. C'est ainsi que la compréhension des relations entre la structure et la fonction d'une protéine est une des clés des nouvelles avancées thérapeutiques (Glen and Allen, 2003)

Néanmoins, il existe un profond fossé quantitatif entre le nombre de séquences protéiques connues et les informations structurales disponibles. A ce jour, pour environ un million et demi de séquences dans la banque de données non redondante (*nr*) du *National Center for Biology Information* (NCBI), environ 22.000 structures protéiques sont recensées dans la *Protein Data Bank* (PDB) (Westbrook *et al*., 2003). Cette relative rareté de données structurales ne fera que de s'accentuer avec l'augmentation du nombre de projets de séquençage de génomes.

Dès lors, pour émettre ou étayer des hypothèses plausibles concernant la fonction d'innombrables protéines dans des domaines aussi variés que la pharmacologie, la biotechnologie, l'industrie agro-alimentaire

et même l'étude des écosystèmes, les biologistes ont exploré des stratégies pour convertir au plus vite les séquences protéiques en informations structurales fiables. Une de ces techniques, la modélisation par homologie, se base sur la similarité de structure entre deux séquences protéiques homologues (dérivant d'un ancêtre commun) partageant un taux d'acides aminés identiques (pourcentage d'identité) élevé.

Dans cette thèse, nous allons montrer comment nous avons tenté de perfectionner une méthode d'alignement de séquences en vue d'améliorer la prédiction automatique de la structure 3D des protéines. Nous commencerons par rappeler quelques notions concernant les différents niveaux de structure des protéines et nous décrirons largement les différentes techniques d'alignement de séquences. Nous évoquerons les méthodes de prédiction de la structure secondaire des protéines et détaillerons les différentes méthodes de prédiction de leur structure tertiaire. Après ces rappels, nous présenterons nos objectifs, puis entamerons la description de notre travail. Celui-ci a d'abord porté sur le développement d'une méthode très fiable d'alignement pairé. Ce développement a ensuite permis de réaliser un système automatique de modélisation par homologie. Ce dernier a finalement été utilisé dans la construction d'une base de données structurales et fonctionnelles consacrée au génome de la bactérie *Brucella melitensis*. Nous terminerons cette thèse par une conclusion, incluant des perspectives pour de nouvelles recherches.

I. Introduction

Cette première partie va principalement rappeler les notions essentielles qui seront utiles à la bonne compréhension des techniques dont le développement fait l'objet de cette thèse. Dans la première section, la notion de structure protéique sera revue en rappelant les différents niveaux de structure des protéines. Mon travail ayant consisté principalement à améliorer l'alignement de séquences protéiques, je passerai ensuite en revue dans une deuxième section les différentes méthodes existantes et pointerai leurs avantages et leurs faiblesses. Dans une troisième section, les réseaux neuronaux seront présentés de manière non exhaustive afin de faciliter la compréhension de l'utilisation qui en est faite dans la suite de ce travail. Cette thèse étant également centrée sur la prédiction de la structure tertiaire des protéines, les principales techniques de prédiction seront expliquées, dans une quatrième section, en mettant l'accent sur la modélisation par homologie, son amélioration étant le but de cette thèse. Enfin, puisque notre technique de modélisation automatique a été appliquée à la modélisation des protéines du génome de *Brucella melitensis*, une cinquième et dernière section sera consacrée à la présentation de cette bactérie pathogène.

I.1. Structure des protéines

La structure des protéines est une donnée essentielle pour la compréhension de la fonction des protéines. On décompose généralement cette structure en différents niveaux: les structures primaire, secondaire, tertiaire et quaternaire. Une section sera consacrée à chacun de ces niveaux de structure.

I.1.1. STRUCTURE PRIMAIRE DES PROTÉINES

I.1.1.1. Définition

Les protéines sont des hétéropolymères linéaires d'acides aminés appelés également polypeptides. La séquence des acides aminés (a.a. ou aussi résidus) définit la structure primaire. Dans les organismes vivants, les acides aminés intervenant dans la composition des protéines sont au nombre de 20. Ils ont tous en commun une chaîne principale (Figure 1) composée d'un carbone central (carbone α ou C_α) portant une fonction carboxyle, une fonction amine (NH_2), un atome d'hydrogène (H) et une chaîne latérale (R) qui détermine les propriétés physico-chimiques spécifiques de chaque acide aminé (voir Figure 2).

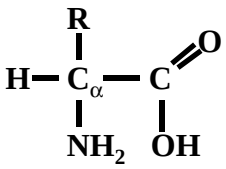


Figure 1: Chaîne principale commune à tous les acides aminés naturels.

L'atome central est un carbone asymétrique uniquement présent sous la forme lévogyre (L), sauf dans la glycine pour laquelle la chaîne latérale est uniquement constituée d'un atome d'hydrogène. Contrairement aux autres acides aminés, la proline forme un cycle par une liaison entre l'extrémité de la chaîne latérale et l'atome d'azote de la fonction amine.

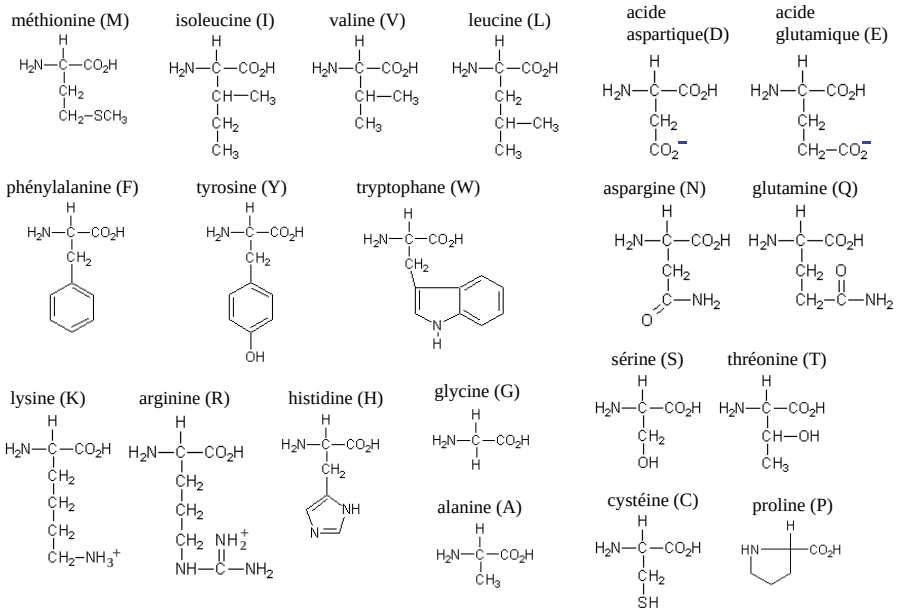


Figure 2: Représentation des 20 acides aminés naturels.

I.1.1.2. La liaison peptidique

Les acides aminés des protéines sont reliés entre eux 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.

L'enchaînement des chaînes principales des acides aminés définit le squelette (aussi appelé *backbone*) de la protéine. La conformation spatiale des acides aminés dépend de certaines contraintes que nous allons détailler dans les points qui suivent.

La liaison peptidique présente un caractère double partiel (voir Figure 3) dû à la délocalisation du doublet de l'azote sur l'oxygène du carboxyle. 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. Le squelette de la protéine et les chaînes latérales peuvent également prendre différentes orientations dans l'espace par rotation autour des liaisons chimiques simples.

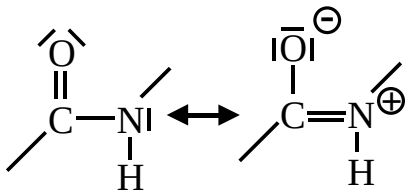


Figure 3: Structures de résonance montrant le caractère double partiel de la liaison peptidique.

I.1.1.3. Les angles de torsion

L'assemblage d'un polypeptide fait apparaître des propriétés plus complexes que celles de ses composants pris séparément à cause des interactions entre chaînes latérales (interactions de Coulomb, de van der Waals, ponts d'hydrogène, …). Ces interactions ne sont possibles que par la variation des angles de torsion des chaînes principales.

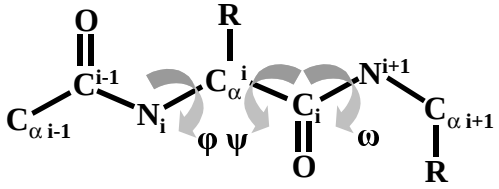


Figure 4: Portion de squelette protéique. Les angles ω, φ et ψ y sont représentés.

Décrivons brièvement les trois types d'angles de torsion qui caractérisent le squelette d'une protéine (voir Figure 4):

- **L'angle ω** ($C\alpha_i$ - C_i - N_{i+1} - $C\alpha_{i+1}$) se situe au niveau de la liaison peptidique et vaut presque toujours 180°. Ceci entraîne que les C_α des deux résidus impliqués dans la liaison et leurs prolongements adoptent le plus souvent une configuration *trans*, dans laquelle l'interaction des chaînes latérales est minimale. On observe rarement des résidus en configuration *cis* (angle ω de 0°). En effet, le rapport entre le nombre de configurations *trans* et *cis* observées dans les structures connues est supérieur à mille. La proline fait exception avec un rapport *trans*/*cis* approximatif de quatre. Cette exception est due à la liaison de la chaîne latérale de la proline sur l'atome d'azote, qui rend la position *trans* moins favorable que dans les autres acides aminés.
- **L'angle φ** (C_{i-1} - N_i - $C\alpha_i$ - C_i) ne peut prendre que certaines valeurs (Figure 5), qui remplissent les conditions les plus favorables d'un point de vue stérique.

- **L'angle ψ** (N_i - $C\alpha_i$ - C_i - N_{i+1}) ne prend pas toutes les valeurs possibles (Figure 5) pour les mêmes raisons que l'angle φ.

Le **diagramme de Ramachandran** (Ramachandran *et al.*, 1963) définit les plages de valeurs autorisées des angles φ et ψ des protéines (voir Figure 5).

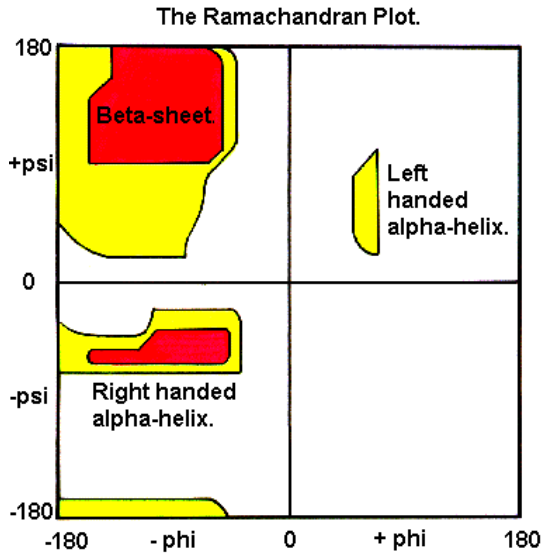


Figure 5: Diagramme de Ramachandran. Les valeurs d'angles φ et ψ colorées en rouge représentent 66% des valeurs adoptées par une catégorie, celles en jaune, représentent 95% des valeurs.

Des angles de torsion sont également définis dans les chaînes latérales. Ils sont désignés par $\chi_{(j)}$ où l'indice *j* indique la position de la liaison par rapport au carbone α. Dans les structures des protéines, ces angles sont restreints à un nombre limité de valeurs. On peut retrouver les principales valeurs des angles $\chi_{(j)}$ dans les banques de rotamères (Dunbrack, 2002).

I.1.2. STRUCTURE SECONDAIRE DES PROTÉINES

I.1.2.1. Définition

Les structures secondaires sont des régions du squelette protéique déterminées par des séquences spécifiques d'angles, φ et ψ, et par des réseaux de ponts hydrogène caractéristiques. On les classe généralement en conformations régulières et irrégulières. Et on distingue essentiellement trois

grands types de conformations régulières: les hélices α, les plans β et les autres conformations régulières.

I.1.2.2. Hélice α

Les hélices α (Figure 7) sont des structures hélicoïdales où l'oxygène du carbonyle (C=O) d'un résidu en position *n* dans la séquence est lié par pont hydrogène (H) au groupe amine (N-H) du résidu *n*+4. 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 et représente 0.54 nm sur l'axe de l'hélice. On appelle cette distance le pas de l'hélice. Dans l'hélice α, les angles φ et ψ prennent respectivement les valeurs moyennes de –57 et –47° et l'angle ω reste à environ 180°.

Un exemple de protéine contenant des hélices α est la thioredoxine d'*Escherichia coli* (Figure 6).

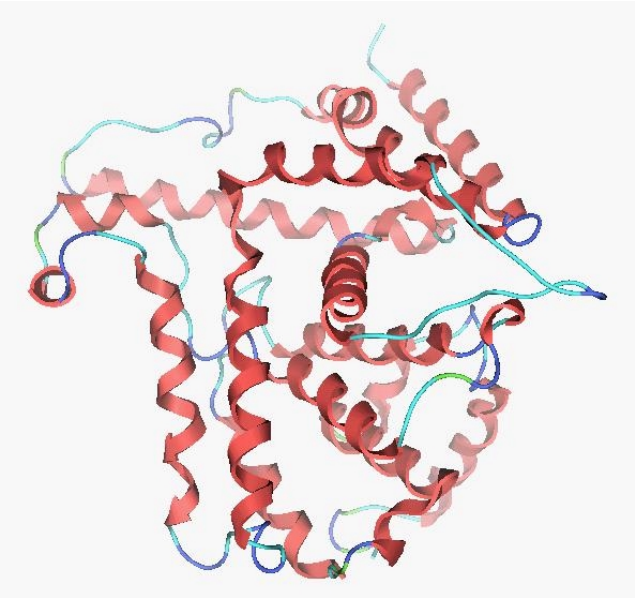


Figure 6: Structure de la thioredoxine d'*Escherichia coli* (PDB ID: 1f0j). Les hélices α sont colorées en rouge.

I.1.2.3. Plan β

Le plan ou feuillet β (Figure 7) est le résultat de l'assemblage de brins β qui interagissent entre eux via des ponts H entre les groupes C=O d'un brin et NH de l'autre brin. Dans les brins β, les angles φ et ψ prennent respectivement les valeurs moyennes de –140 et +135° et l'angle ω reste à

environ 180°. Les brins β ont une longueur de 6 résidus, en moyenne. D'aspect légèrement plissé, le feuillet β recouvre une plus grande surface que l'hélice α pour un même nombre de résidus.

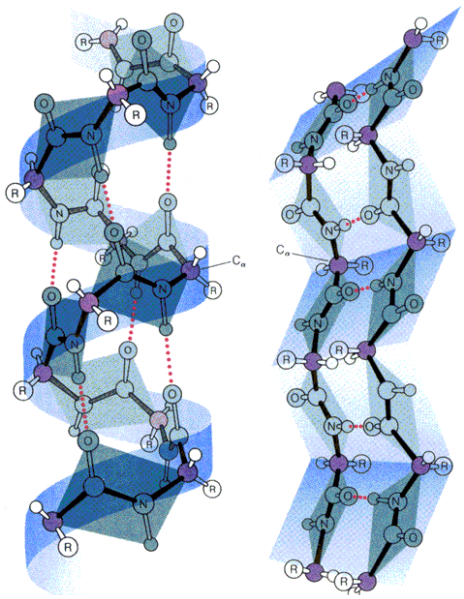


Figure 7: Représentation d'une hélice α et d'un feuillet β.

Suivant l'agencement des ponts H entre les brins, les brins β peuvent se combiner en sens parallèles ou antiparallèles selon que les brins qui les constituent sont orientés dans le même sens ou alternativement dans des sens opposés. Les plans β antiparallèles sont stabilisés par des ponts H parallèles entre eux, conduisant à une meilleure stabilité conformationnelle par rapport aux plans β parallèles. Des plans mixtes, mélanges de ces deux types, sont parfois rencontrés.

Un exemple de protéine contenant des brins β est le domaine catalytique de la phosphodiesterase 4B2B humaine (Figure 8).

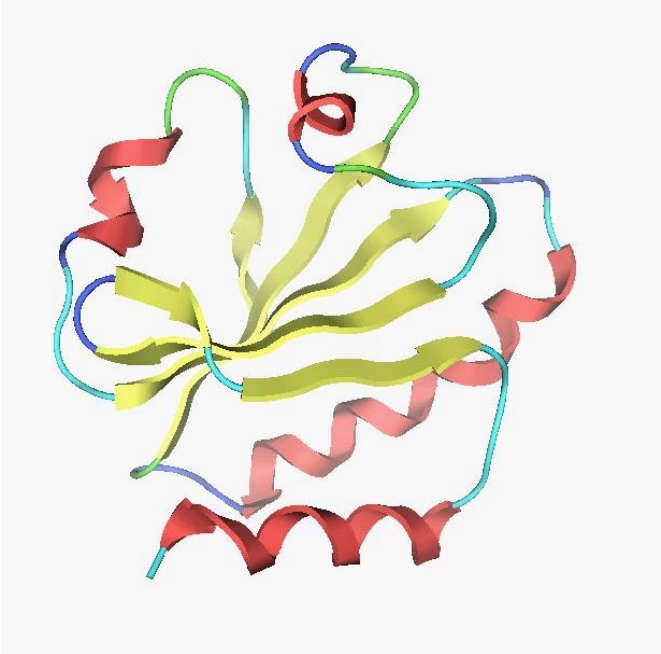


Figure 8: Structure du domaine catalytique de la phosphodiesterase 4B2B humaine (PDB ID: 2trx). Les brins β sont colorés en jaune et les hélices α sont colorées en rouge.

I.1.2.4. Autres conformations régulières

Les hélices α et les brins β sont les structures les plus fréquemment rencontrées dans les protéines. Néanmoins, d'autres structures régulières moins répandues existent: l'hélice 3_{10}, les coudes ou *turns*, les β *turns* et les γ *turns*.

L'hélice 3_{10} possède trois résidus par tour, est beaucoup plus étroite que l'hélice α et possède des valeurs d'angles φ et ψ de –60 et –30°. Lorsqu'elle est présente, on la trouve aux extrémités C terminales des hélices α mais ces structures ne sont jamais plus longues que deux tours.

Les coudes ou *turns* sont de courtes structures secondaires en U stabilisées par un pont H entre les résidus *n* et *n*+2 et dont le rôle est de connecter les hélices α et les brins β en changeant la direction de la chaîne polypeptidique. Ils permettent ainsi la formation de protéines (globulaires) compactes (Rose *et al*., 1985).

Composés de quatre résidus, les β *turns* forment un pont H entre le premier et le troisième résidu et contiennent très souvent une proline ou une glycine dont les angles de torsion permettent de relier deux brins β

antiparallèles et d'induire un changement de direction de la chaîne polypeptidique.

Semblables aux β *turns*, les γ *turns* ne sont constitués que de trois résidus.

I.1.2.5. Conformations irrégulières

Les conformations irrégulières sont des fragments polypeptidiques de longueur variable en forme de boucles (*loop*s). Ces boucles ont pour but de relier les différents éléments de structure secondaire de la protéine.

Contrairement aux plans β et aux hélices α, les boucles situées principalement en surface de la protéine contiennent un nombre assez important de résidus hydrophiles et chargés qui leur permettent d'interagir avec le solvant.

Ces segments de la séquence sont en général peu conservés d'un point de vue évolutif, sauf dans certains cas où ils ont un rôle fonctionnel (phosphorylation, …). Dans ce cas, les mutations y sont beaucoup plus rares (Lodish *et al*., 1997). Par exemple, les structures de certaines boucles des immunoglobulines sont conservées chez tous les vertébrés (Barre *et al*., 1994).

I.1.3. STRUCTURE TERTIAIRE DES PROTÉINES

Les éléments de structure secondaire forment des arrangements géométriques spécifiques appelés « motifs structuraux », « domaines » ou « super structures secondaires » (Sibanda and Thornton, 1985; Efimov, 1991; Efimov, 1991). L'arrangement tridimensionnel de ces domaines est appelé structure tertiaire de la protéine (*fold*).

Le repliement en structure tertiaire (*folding*) a pour effet de rapprocher spatialement des résidus fortement éloignés au niveau de la séquence. Ce rapprochement rend possible le positionnement très précis des chaînes latérales, qualité requise pour assurer une activité optimale de la protéine (fixation plus aisée d'un (de) substrat(s) ou d'un (de) coenzyme(s) par les enzymes). Par exemple, dans les déshydrogénases, le rapprochement de deux zones hydrophiles appartenant à deux domaines identiques favorise la création d'un milieu plus favorable à la fixation du NAD^+, molécule essentiellement hydrophile (De Bolle *et al*., 1995).

I.1.4. STRUCTURE QUATERNAIRE DES PROTÉINES

Les structures quaternaires sont des assemblages de plusieurs chaînes polypeptidiques. Ces assemblages peuvent contenir deux unités (dimère) ou plus (multimère), jusqu'à plusieurs centaines comme dans les capsides de virus. La structure quaternaire, très souvent nécessaire à la fonctionnalité de la protéine, est souvent organisée de manière symétrique pour permettre la formation de larges complexes avec un nombre réduit de monomères différents. Les monomères interagissent entre eux via des liaisons chimiques faibles (interactions hydrophobes et/ou contacts polaires, ponts H et intervention d'ions) pour stabiliser la structure quaternaire. Dans certains cas, le monomère doit adapter sa conformation spatiale pour permettre ces contacts.

De plus, des fonctionnalités multiples sont accrues en formant des complexes multi-enzymatiques qui catalysent un ensemble de réactions. Enfin, certains assemblages complexes jouent non seulement un rôle fonctionnel, mais aussi un rôle structural au niveau cellulaire (par exemple, la tubuline).

I.2. Méthodes d'alignement de séquences protéiques

Publication présentée (voir Annexe 1):

C. Lambert, J.-M. Van Campenhout, X. De Bolle and E. Depiereux
Review of common sequence alignment methods: clues to enhance reliability
Current Genomics **4**(2): 131-146 (2003)

I.2.1. INTRODUCTION

La bioinformatique offre à la biologie moléculaire un ensemble de possibilités d'analyses qui transforment fondamentalement de nombreuses stratégies de recherche. Régulièrement, des logiciels et sites *web* proposent de nouvelles approches pour des analyses en protéomique et en génomique. Parmi ces techniques, l'alignement de séquences occupe une position centrale. Il est utilisé pour assembler les fragments de séquences provenant des projets de séquençage, prédire la position des séquences codantes dans les génomes (Salzberg *et al*., 1998; Delcher *et al*., 1999), prédire la fonction des protéines par des recherches de similarité dans des banques de données (Pearson and Lipman, 1988; Altschul *et al*., 1990), comparer des génomes apparentés et tenter de reconstituer la phylogénie de plusieurs protéines ou séquences nucléotidiques (Felsenstein, 1989). Un alignement de séquences offre en lui-même une source importante d'information sur la variabilité des acides aminés à des positions clés de la séquence. Les acides aminés les plus conservés correspondent à des résidus essentiels impliqués dans des sites catalytiques, ou dans la stabilité de la structure, ou encore dans les interactions de la protéine avec d'autres molécules.

La qualité d'un alignement de séquences influence considérablement la fiabilité de toutes les méthodes qui se basent sur la corrélation entre la conservation de la structure des protéines et celle de leur séquence. Ainsi, obtenir un alignement de qualité est un préalable essentiel à la modélisation par homologie (Sanchez and Sali, 1998), à la reconnaissance de *fold* (Bates and Sternberg, 1999), à la prédiction *de novo* (Simons *et al*., 1999; Bonneau *et al*., 2001), à la prédiction de la structure secondaire (Rost and Sander, 1993), celle de l'accessibilité au solvant (Rost and Sander, 1995) et celle des hélices transmembranaires (Rost *et al*., 1996).

Dans cette section, nous présenterons d'abord une grande partie des algorithmes d'alignement de séquences décrits dans la littérature. Ensuite, nous nous focaliserons sur l'évaluation de leurs performances.

I.2.2. DÉFINITION

Un alignement de séquences est une table à deux dimensions dans laquelle les lignes représentent les séquences protéiques ou nucléotidiques et les colonnes représentent les positions des résidus (Figure 9). Les séquences sont placées sur cette grille de telle manière que:

(a) la position absolue des résidus, c'est-à-dire leur position dans l'ordre de la séquence non alignée, soit préservée

(b) les résidus similaires dans toutes les séquences soient placés dans la même colonne en introduisant, si nécessaire, des cellules vides appelées *gap* (trou).

```
              10        20        30        40        50        60        70
               +         +         +         +         +         +         +
1ALC  kqftkcelsqnlyd--idgygrialpelictmfhtsgydtqai--vendesteyglfqisnalwckssqs
1LZ1  kvfercelartlkrlgmdgyrgislanwmclakwesgyntratnynagdrstdygifqinsrywcndgkt
2LZ2  kvygrcelaaamkrlgldnyrgyslgnwvcaakfesnfnthatn-rntdgstdygilqinsrwwcndgrt
2LZT  kvfgrcelaaamkrhgldnyrgyslgnwvcaakfesnfntqatn-rntdgstdygilqinsrwwcndgrt


              80        90       100       110       120       130       140
               +         +         +         +         +         +         +
1ALC  pqsrnicditcdkflddditddimcakkil-dikgidywiahkalctekleqwlcek---
1LZ1  pgavnachlscsallqdniadavacakrvvrdpqgirawvawrnrcqnrdvrqyvqgcgv
2LZ2  pgsknlcnipcsallssditasvncakkiasggngmnawvawrnrckgtdvhawirgcrl
2LZT  pgsrnlcnipcsallssditasvncakkivsdgngmnawvawrnrckgtdvqawirgcrl
```

Figure 9: Exemple d'alignement de séquences de lysozymes de *Homo sapiens* (1LZ1), de *Meleagris gallopavo* (2LZ2) et de *Gallus gallus* (2LZT), et d'une alpha-lactalbumine (1ALC) de *Papio hamadryas cynocephalus*. Les séquences ont des longueurs différentes et il est nécessaire de placer des *gaps* pour aligner les séquences. Les positions des résidus dans l'alignement sont représentées sous la forme d'une règle, avec une numérotation de 10 en 10. Les résidus similaires sont positionnés dans la même colonne.

La position d'un résidu dans l'alignement de séquences est appelée sa position relative et les résidus situés dans la même colonne auront la même position relative. Par contre, à moins que les séquences soient toutes identiques, les positions absolues des résidus d'une même colonne seront presque toujours différentes. Par conséquent, la position absolue est une propriété de la séquence tandis que la position relative est une propriété de l'alignement. Un programme d'alignement de séquences peut dès lors être considéré comme une fonction capable de transformer des positions absolues en positions relatives. Dans ce point de vue, les alignements sont de simples modèles mathématiques dont la forme peut être modifiée grâce aux variations de différents paramètres.

De très nombreuses méthodes ont été développées pour aligner des séquences. Les méthodes classiques utilisent des principes mathématiques simples: comparaison de petits segments ou programmation dynamique. Les

nouvelles approches utilisent des algorithmes développés au départ pour la physique, les mathématiques et l'intelligence artificielle: traitement du signal, modèles cachés de Markov (*Hidden Markov Models* ou HMM), algorithmes génétiques, optimisation *monte carlo*, statistiques bayesiennes et réseaux neuronaux.

D'autre part, on peut regrouper les algorithmes d'alignement suivant deux stratégies différentes. La première recherche à établir une similarité entre des séquences prises dans leur intégralité (alignement global) et la seconde s'intéresse seulement à des régions similaires (alignement local). Ces stratégies correspondent à des contextes biologiques différents. L'alignement global ne se justifie que si l'on considère des séquences supposées largement similaires sur toute leur longueur. La recherche de similarité locale se base sur le fait que les sites fonctionnels (sites catalytiques des enzymes, par exemple) sont localisés dans des régions relativement courtes et bien conservées malgré des insertions, mutations ou délétions dans les autres régions de la séquence. Néanmoins, il faut noter que de courts segments de séquences similaires n'ont pas nécessairement la même structure 3D (Simons *et al*., 1997).

I.2.3. ALIGNEMENT DE DEUX SÉQUENCES

I.2.3.1. Matrices de scores

Toutes les techniques de comparaison de séquences sont basées sur le concept d'alignement, qui définit la relation entre les résidus des séquences comparées. Des résidus alignés sont considérés comme ayant la même fonction ou la même origine phylogénétique.

Dans beaucoup d'applications de comparaisons de séquences, la meilleure solution est sélectionnée parmi toutes les solutions possibles, en se basant sur un score calculé suivant un certain modèle de similarité. Cette optimisation implique généralement l'assignation d'un score à chaque alignement ou sous-alignement.

Généralement, le score associé à un alignement est calculé en prenant la somme des valeurs (poids) associées à chaque paire d'acides aminés alignés. Ces poids sont extraits de tables appelées "matrices de scores". Dans ces matrices, on trouve des poids pour toutes les paires d'acides aminés possibles. Ces poids expriment soit une "dissemblance" ou "distance" entre acides aminés, soit, plus fréquemment, une similarité. La matrice de scores de similarité la plus simple est une matrice unitaire, assignant +1 aux identités et 0 aux différences. Une matrice basée sur le code génétique refléterait le nombre maximum de nucléotides que les codons partagent entre eux pour deux acides aminés. Les poids seraient alors +3

pour les identités, puis +2, +1 et 0 pour les codons n'ayant aucun nucléotide en commun.

D'autres matrices de scores plus complexes ont été construites. Dans ces dernières, les poids des résidus identiques et non identiques sont établis d'après leurs taux de substitution observés dans des familles de protéines dont la structure ou la fonction est conservée ou d'après la similarité de leurs propriétés physico-chimiques. L'utilisation de ces matrices permet d'augmenter la sensibilité du processus d'alignement, spécialement dans les cas où la similarité des séquences à aligner est très faible.

Nous n'allons pas présenter exhaustivement toutes les matrices de scores existantes mais nous allons nous focaliser sur deux des plus populaires: les matrices de mutations de Dayhoff (*mutation data* ou MD) (Dayhoff *et al*., 1978) et la série des matrices BLOSUM (Henikoff and Henikoff, 1992).

I.2.3.1.1. Matrices de Dayhoff (Mutation Data Matrix, MDM*)*

Le modèle de mutation des acides aminés de Dayhoff (Dayhoff *et al*., 1978) est Markovien; c'est-à-dire que la probabilité de mutation à n'importe quel endroit de la protéine est indépendante de son histoire précédente. Selon ce modèle, les matrices de scores sont dérivées d'une matrice de probabilités de transition dans laquelle chaque élément donne la probabilité qu'un acide aminé A soit remplacé par un acide aminé B en une unité de changement évolutif. Dans cette matrice, les éléments de la diagonale donnent les probabilités que les acides aminés restent inchangés. La somme des éléments de la diagonale donne la probabilité qu'il n'y ait aucun changement pendant l'intervalle évolutif représenté. Dans le cas de la matrice de Dayhoff, la matrice de probabilité a été normalisée pour que cette probabilité corresponde à 99%. L'unité d'évolution représentée par la matrice de probabilité de transition correspond donc à une mutation acceptée pour 100 sites de mutation ou 1 unité PAM (*Point Accepted Mutation*). Des matrices de probabilités de transition correspondant à des intervalles plus larges de distance évolutive peuvent être obtenues en multipliant de façon répétée la matrice de probabilité de transition originelle par elle-même. La matrice de scores MDM PAM 250 donne ainsi des poids de similarité correspondant à environ 20% d'acides aminés identiques restant entre deux séquences.

Chaque élément de la matrice de scores est calculé selon la formule suivante:

$$S_{A,B} = \log\left(\frac{P(A/B)}{P(A)\cdot P(B)}\right)$$

où $S_{A,B}$ est la valeur de l'élément de la matrice de scores correspondant aux acides aminés A et B

P(A/B) est la probabilité de mutation de l'acide aminé A par l'acide aminé B

P(A) est la probabilité de trouver l'acide aminé A

P(B) est la probabilité de trouver l'acide aminé B

En effet, pour des comparaisons de séquences, il est utile d'employer une matrice dans laquelle les éléments reflètent le rapport entre la probabilité d'échange d'un acide aminé par un autre et la probabilité que cette substitution survienne par hasard. Quand une protéine est comparée à une autre, position par position, ces rapports sont multipliés pour calculer une probabilité pour l'alignement complet. Cependant, du point de vue computationnel, il est plus intéressant d'additionner les logarithmes (représentés par des nombres entiers, via un facteur d'échelle) de ces nombres. Ainsi, la matrice PAM 250 contient les logarithmes des rapports de probabilité correspondant à 250 unités PAM (*log odds matrix*). Dans cette matrice, les valeurs supérieures à zéro indiquent une forte chance de mutation, une valeur de zéro est neutre (aléatoire) et les valeurs inférieures à zéro indiquent une faible chance de mutation.

Le but de la plupart des alignements est de découvrir des similarités faibles; par exemple, identifier des relations dans la *Twilight Zone* (Doolitle, 1986; Rost, 1999): zone de pourcentage d'identité située entre 20% et 30% où deux protéines peuvent encore avoir une structure similaire et où les performances des programmes d'alignement de séquences chutent brutalement. La matrice MD pour 250 PAMs est donc devenue la matrice par défaut dans beaucoup de programmes d'analyse de séquences puisqu'elle reflète un niveau de 20% d'identité. Cependant, il est en principe préférable d'utiliser une matrice qui correspond à la distance évolutive réelle entre les séquences que l'on veut comparer. Mais cette méthode est limitée puisqu'elle requiert une connaissance *a priori* de la distance évolutive et, donc, de connaître à l'avance ce qu'on recherche. Une bonne pratique est d'établir une stratégie dans laquelle différentes matrices PAM sont utilisées.

I.2.3.1.2. Les matrices BLOSUM

Les matrices proposées par Dayhoff sont limitées parce que les taux de substitution sont dérivés d'alignements peu nombreux et où le pourcentage d'identité est de l'ordre de 85%. Pourtant, le but des alignements de séquences est de détecter des relations plus distantes entre les séquences.

De manière à représenter plus explicitement ces relations distantes, Henikoff et Henikoff (Henikoff and Henikoff, 1992) ont dérivé un ensemble de matrices de scores à partir des groupes de segments alignés (appelés blocs en français et *blocks* en anglais) rassemblés dans la banque de données BLOCKS (Henikoff and Henikoff, 1996; Henikoff *et al*., 1999). Les matrices ont été calculées à partir des groupes de segments BLOCKS dans lesquels le pourcentage d'identité entre les paires de segments est supérieur à une valeur seuil. Par exemple, les segments regroupés avec un pourcentage d'identité supérieur ou égal à 80% sont utilisés pour générer la matrice BLOSUM 80 (*BLOcks SUbstitution Matrix*), ceux ayant un pourcentage d'identité supérieur ou égal à 62% pour la matrice BLOSUM 62, et ainsi de suite.

On peut observer des différences d'alignement suivant que l'on utilise les matrices de scores BLOSUM ou Dayhoff. Ces différences peuvent être cruciales dans la *Twilight Zone* où la détection de similarités faibles est l'objectif central. Une comparaison des performances d'un grand nombre de matrices de scores pour l'alignement pairé a été publiée par Vogt *et al.* (Vogt *et al*., 1995).

I.2.3.2. Algorithme de Needleman-Wunsch

En 1970, Needleman et Wunsch (Needleman and Wunsch, 1970) ont proposé un algorithme qui détermine un alignement optimal parmi tous les alignements possibles qui peuvent être générés pour deux séquences. Cet algorithme utilisait une technique appelée programmation dynamique (Bellman, 1957), qui est une méthode d'optimisation dans laquelle on construit une solution à un problème en résolvant des sous-problèmes plus petits mais similaires. Il fonctionnait de la manière décrite ci-dessous.

Soit *A* et *B*, deux séquences de longueurs *m* et *n*, et désignons par *A*(*i*) et *B*(*i*), le i^{ème} résidu dans ces deux séquences. Supposons que nous assignons pour chaque paire de résidus possible des deux séquences, un poids, *wt*, qui reflète la similarité entre les deux résidus:

$$wt(i, j) = weight[A(i), B(j)]$$

Il est évident que, dans l'alignement optimal, la somme des poids assignés aux paires de résidus, est maximum. Les poids sont extraits d'une matrice de scores telle que présentée ci-dessus. Pour trouver la somme

maximum des poids, l'algorithme de Needleman – Wunsch construit une matrice à deux dimensions, *L*, dans laquelle chaque cellule (*i*,*j*) correspond à la paire de résidus *A*(*i*), *B*(*j*) pour i=1, …, m et j=1, …, n. Le but est alors de trouver un chemin dans L qui maximise la somme des poids (*wt*). Needleman et Wunsch appellent ce chemin le chemin d'appariement maximum. Deux cellules consécutives de ce chemin sont (*i*,*j*) et (*x*,*j*+1) pour $i < x \leq m$ ou (*i*,*j*) et (*i*+1,*y*) pour $j < y \leq n$.

La somme des poids est effectuée à partir de la première cellule (1,1), colonne par colonne et rangée par rangée de la manière suivante:

$$L(i, j) = \max\{L(i-1, j-1) + wt(i, j), L(x, j-1) - g, L(i-1, y) - g\}$$
$$(i < x \leq n, j < y \leq n)$$

La valeur de *L*(*i*,*j*) est ainsi la somme de *wt*(*i*,*j*) et de la valeur maximum des cellules déjà calculées qui précèdent (*i*,*j*). De manière à éviter un nombre excessif de *gaps*, une pénalité *g* est soustraite de la valeur maximum de contribution sauf si elle vient de la cellule (*i*-1,*j*-1). Calculée de cette manière, la valeur de *L*(*m*,*n*) contient alors la valeur maximum de la somme des poids de l'alignement complet. Les résidus alignés de manière optimale peuvent être retrouvés en effectuant en arrière la recherche du chemin qui a mené au score de la cellule *L*(*m*,*n*). Le résultat obtenu est un alignement global, et optimal (au sens de l'algorithme utilisé) car son score est le plus élevé possible.

I.2.3.3. Algorithme de Smith-Waterman

L'algorithme de Needleman et Wunsch produit des résultats corrects pour des séquences qui partagent une similarité sur toute leur longueur. Cependant, si on considère deux séquences qui partagent une homologie lointaine, seules de courtes régions seront encore similaires. Ainsi, aucun alignement global satisfaisant ne pourra être trouvé. En 1981, Smith et Waterman (Smith and Waterman, 1981) ont décrit une méthode, connue communément comme l'algorithme de Smith-Waterman, pour trouver ces régions communes. Comme la technique de Needleman et Wunsch (Needleman and Wunsch, 1970), c'est une approche basée sur une matrice de poids *L* et une recherche en arrière est utilisée pour reconstruire l'alignement local optimal avec *gaps*.

Néanmoins, dans le cas de l'algorithme de Smith-Waterman, *L*(*i*,*j*) est obtenu par la formule suivante:

$$L(i, j) = \max\{L(i-1, j-1) + wt(i, j), L(x, j-1) - g, L(i-1, y) - g, 0\}$$
$$(i < x \leq n, j < y \leq n)$$

La différence essentielle entre les deux algorithmes que nous venons de décrire est que, dans le cas de l'algorithme de Smith-Waterman, la matrice contient une cellule de valeur maximum qui peut ne pas correspondre à la partie C-terminale des séquences. La cellule correspondant au score maximum représente le point final d'un alignement de segments tel qu'aucune autre paire de segments n'existe avec un score de similarité supérieur. Pour obtenir cet alignement de segments, il suffit de remonter le chemin suivi pour aboutir à ce score et de s'arrêter lorsque la valeur de la cellule rencontrée est de 0. Ainsi, l'algorithme de Smith-Waterman est une méthode d'alignement local plutôt que global.

La méthode de Smith-Waterman a été à la base de plusieurs autres algorithmes et est utilisée comme référence lorsqu'on compare différentes techniques d'alignement pairé.

I.2.3.4. Autres programmes d'alignement pairé de séquences

La recherche de similarités de séquences dans des banques de données est effectuée à l'aide d'algorithmes rapides qui détectent des similarités locales entre paires de séquences. BLAST (Altschul *et al.*, 1990), BLAST2 (Altschul *et al.*, 1997) et FastA (Pearson and Lipman, 1988; Pearson, 1990) sont les plus connus. Ils utilisent des règles heuristiques (règles empiriques qui aident à trouver la bonne solution).

Le programme PSI-BLAST (Altschul *et al.*, 1997) fonctionne par itérations. Dans la première itération, une simple recherche avec BLAST2 (Altschul *et al.*, 1997) est effectuée. Elle permet de retrouver un ensemble de séquences similaires à la séquence d'intérêt. Ensuite, chaque séquence retrouvée est alignée à la séquence d'intérêt dans un alignement multiple, et un profil en est créé. Ce profil contient la fréquence relative de chaque acide aminé pour chaque position de la séquence d'intérêt et représente la variabilité des acides aminés dans la famille de la séquence d'intérêt. A partir de ce profil et de la matrice de scores utilisée pour effectuer la recherche, il est possible de calculer une matrice de scores spécifique de chaque position de la séquence d'intérêt (*Position Specific Scoring Matrix* ou PSSM) (Altschul *et al.*, 1997). On a donc, pour chaque acide aminé de la séquence d'intérêt, une table de 20 scores qui représentent au mieux les probabilités de mutation de ce résidu vers un autre résidu.

Dans la seconde itération et dans les suivantes, une version modifiée de BLAST2 utilise la PSSM pour rechercher de nouvelles séquences similaires. Celles-ci sont intégrées à l'ensemble des séquences découvertes lors des itérations précédentes et, sur cette base, une nouvelle PSSM est calculée et utilisée pour rechercher de nouvelles séquences. Le processus

s'arrête automatiquement lorsque le programme ne découvre plus de nouvelles séquences.

I.2.4. ALIGNEMENT MULTIPLE DE SÉQUENCES

I.2.4.1. Complexité computationnelle

Les techniques d'alignement pairé nécessitent généralement un temps de calcul et un espace mémoire proportionnel au produit des longueurs des séquences qui vont être comparées: $O(m_1 m_2)$ où O représente l'ordre de grandeur du temps calcul pris par l'algorithme, et m_1 et m_2 sont les longueurs des séquences. En généralisant ces techniques pour aligner trois séquences, on obtient un temps calcul dépendant de $O(m_1 m_2 m_3)$ où m_3 est la longueur de la troisième séquence.

Quand on considère n séquences, la complexité du temps calcul devient $O(m_1 m_2 \ldots m_n)$, où m_n est la longueur de la dernière séquence de l'ensemble de comparaisons, ou de manière plus concise $O(m^n)$, où n est le nombre de séquences et m est la longueur moyenne des séquences. Donc, dans le cas de méthodes d'alignement simultané à n séquences, le temps nécessaire au calcul d'un alignement multiple augmente exponentiellement avec le nombre de séquences à aligner.

Plusieurs approches combinent la programmation dynamique à des heuristiques. De telles techniques utilisent soit l'alignement de toutes les paires de séquences, soit l'alignement de chaque séquence à une séquence spécifique, ou bien l'alignement des séquences dans un ordre arbitraire ou encore l'alignement des séquences suivant l'ordre d'un arbre phylogénétique qui sert de guide (méthodes progressives). Les résultats de ces méthodes tendent à ne pas être optimaux et requièrent au moins n-1 alignements pairés.

Rappelons qu'un alignement sera dit optimal si son score correspond à l'optimum global de la fonction de score utilisée par la méthode d'alignement. Un alignement optimal n'est cependant pas nécessairement celui qui reflète le mieux la réalité biologique.

I.2.4.2. Méthodes simultanées

I.2.4.2.1. Algorithmes de programmation dynamique à N dimensions

Il est assez facile de calculer un alignement multiple optimal en généralisant les méthodes d'alignement pairé à n séquences (Sankoff, 1975; Waterman *et al.*, 1976). Puisque le temps calcul et l'espace mémoire

nécessaire sont proportionnels au produit des longueurs des séquences, ces algorithmes ne pourraient être appliqués qu'à l'alignement de 3 à 6 séquences. Des algorithmes ont été proposés pour effectuer un alignement de trois séquences par programmation dynamique (Fredman, 1984; Murata *et al*., 1985; Gotoh, 1986). Ceux-ci diffèrent les uns des autres par le traitement et la pénalisation des *gaps* (Altschul, 1989). Etant données les limitations liées au temps calcul et à la quantité de mémoire utilisés par ces méthodes, des techniques d'accélération ont été imaginées.

I.2.4.2.2. Accélération des méthodes N-dimensionnelles

Puisque les méthodes d'alignement pairé fonctionnent plus rapidement que les méthodes N-dimensionnelles, une accélération peut être introduite en utilisant l'information provenant des N(N-1)/2 combinaisons d'alignements pairés pour construire l'alignement multiple final.

I.2.4.2.2.1. Algorithmes réduisant l'espace de recherche

Le principe de ces algorithmes est de réduire l'espace de recherche pour accélérer le calcul de l'alignement multiple final. Carillo et Lipman (Carillo and Lipman, 1988) ont proposé de réduire l'espace de recherche en analysant un alignement multiple produit par des comparaisons pairées et en recherchant l'alignement optimal sans que les positions de cet alignement ne soient modifiées de plus de D résidus. Spouge (Spouge, 1989; Spouge, 1991) a proposé d'utiliser pour cette tâche, l'algorithme A* (Hart *et al*., 1968) (ou GDUS, *Goal Directed Unidirectional Search*: cet algorithme recherche le chemin optimal en sélectionnant à chaque nœud d'un graphe, la branche la plus proche du but à atteindre) pour résoudre le problème de recherche de l'alignement multiple optimal. Le programme MSA (Lipman *et al*., 1989; Gupta *et al*., 1995) utilise aussi bien les restrictions de Carillo-Lipman que l'algorithme A*.

Un autre programme, DCA (Stoye *et al*., 1997), utilise une approche en trois étapes. La première consiste à diviser les séquences à aligner en deux, formant ainsi deux ensembles de séquences plus petites à aligner. Cette première étape est répétée autant de fois que nécessaire. Dans la deuxième étape, ces groupes de séquences plus petites sont alignés en utilisant le programme MSA. Enfin, dans la troisième étape, les différents alignements multiples produits par MSA sont rassemblés pour former un alignement multiple final des séquences de départ.

I.2.4.2.2.2. Méthodes d'identification de segments similaires

La principale stratégie de ces méthodes est de trouver des courts segments alignés (points d'ancrage) qui font partie de l'alignement global optimal en utilisant des informations tirées d'alignements pairés locaux. Plus

le nombre de points d'ancrage trouvés est élevé, plus le temps de calcul nécessaire pour calculer l'alignement multiple est faible. Les programmes utilisant cette stratégie (Waterman *et al*., 1976; Johnson and Doolittle, 1986; Sobel and Martinez, 1986; Santibanez and Rohde, 1987; Vingron and Argos, 1989; Schuler *et al*., 1991; Vingron and Argos, 1991; Vingron and Pevzner, 1995; Depiereux *et al*., 1997; Morgenstern *et al*., 1998) n'utilisent pas nécessairement un algorithme de programmation dynamique pour découvrir les points d'ancrage. Deux de ces programmes sont bien connus: Match-Box (Depiereux and Feytmans, 1992; Depiereux *et al*., 1997) et DIALIGN (Morgenstern *et al*., 1998; Morgenstern, 1999) assemblent les segments conservés dans toutes les séquences (Match-Box) ou les paires de segments (DIALIGN) en un alignement local final (qui n'est donc pas optimal). Ces méthodes restreignent l'alignement multiple à des segments de séquences qui sont significativement similaires, c'est-à-dire pour lesquels la similarité n'est pas due au hasard. Le grand intérêt de ces programmes est qu'ils fournissent un indice de confiance pour chaque position alignée.

I.2.4.3. Méthodes progressives

Beaucoup de méthodes d'alignement multiple (Hogeweg and Hesper, 1984; Waterman and Perlwitz, 1984; Barton and Sternberg, 1987; Feng and Doolittle, 1987; Corpet, 1988; Taylor, 1988; Hein, 1989; Higgins and Sharp, 1989; Smith *et al*., 1990; Gotoh, 1993; Huang, 1994) exploitent le fait que des séquences similaires sont fort probablement liées du point de vue phylogénétique. Dans une première étape, ces programmes génèrent tous les alignements pairés nécessaires pour produire les informations requises pour la création d'un arbre phylogénétique qui servira de guide. Ensuite, les séquences sont alignées par paires en suivant l'ordre des branches de l'arbre guide. On parle d'alignement progressif car les séquences les plus similaires sont alignées les premières pour former des groupes, puis les séquences plus distantes sont ajoutées progressivement par après. Les groupes sont ensuite alignés entre eux pour générer l'alignement multiple final. Toutes les méthodes diffèrent par la construction de l'arbre guide, le système de score, la façon de représenter et de calculer les alignements intermédiaires et de calculer l'alignement multiple final. Dans certains programmes (Thompson *et al*., 1994; Thompson *et al*., 1994; Taylor, 1995), il est possible d'inclure de l'information additionnelle comme des structures secondaires (prédites ou observées) ou la propension de formation de *gaps*. Dans cette catégorie, le programme probablement le plus connu est ClustalW (Thompson *et al*., 1994). D'autres programmes bien connus d'alignement progressif sont MAP (Huang, 1994), MULTAL (Taylor, 1988), MULTALIGN (Barton and Sternberg, 1987), MULTALIN (Corpet, 1988) et PIMA (Smith *et al*., 1990; Smith and Smith, 1992). Dans toutes ces méthodes progressives, l'alignement final n'est pas optimal.

I.2.4.4. Méthodes d'affinement itératif

La principale source d'erreurs dans les méthodes progressives est l'impossibilité de corriger un alignement intermédiaire calculé dans les premières phases de la procédure d'alignement. Les méthodes d'affinement itératif tentent de minimiser ce risque d'erreur en modifiant les alignements intermédiaires ou le modèle servant à les construire jusqu'à convergence du processus (c'est-à-dire que l'alignement ou le modèle n'est plus modifié d'une itération à l'autre).

I.2.4.4.1. Hidden Markov Model

Les *Hidden* Markov *Models* (HMM) ou modèles cachés de Markov ont été très utilisés dans les systèmes de reconnaissance vocale (Rabiner, 1989). Un HMM consiste en une répétition de trois états: émission (E_i), insertion (I_i) et délétion (d_i). Ces états sont interconnectés, avec pour chaque état, un symbole de sortie observable (Krogh *et al.*, 1994) (Figure 10). Quand le modèle est dans

- un état E_i, il émet de manière aléatoire un symbole choisi dans une table de symboles associée à E_i,
- un état I_i, il émet de manière aléatoire un symbole choisi dans une table de symboles associée à I_i,
- un état d_i, il n'émet aucun symbole.

A cette liste, il faut également ajouter un état initial (*début*), qui reflète la mise en marche du modèle, et un état final (*fin*), qui indique la fin d'utilisation du modèle. A chaque état est associée une table de probabilités de transition et, pour les états E_i et I_i, d'une table de probabilité d'émission. Les probabilités d'émission sont les probabilités d'émettre un symbole extrait d'une table caractéristique de cet état. Les probabilités de transition sont les probabilités de passer d'un état à un état suivant.

Par exemple, la séquence d'états, I_0 E_1 E_2 I_2 d_3 d_4 E_5 I_5 I_5, permet de générer la séquence d'acides aminés suivante: « VENTEST ». Or ce que nous pouvons observer est uniquement la séquence « VENTEST » et la séquence des états qui a permis de générer cette séquence nous est inconnue. Dans un alignement de séquences, « VENTEST » sera probablement représentée par « VENT--EST », à cause des deux délétions d_3 et d_4. La séquence des états est cachée et seule la séquence des symboles peut être observée (Rabiner, 1989).

Le HMM peut trouver simultanément un alignement et un modèle de probabilité des substitutions, insertions et délétions qui sont autocohérents. L'alignement le plus probable entre une séquence et un HMM est calculé par l'algorithme de Viterbi (Viterbi, 1967). Le HMM est calculé (ou entraîné) de

manière itérative en effectuant un alignement pairé entre chaque séquence et le HMM qui est progressivement calculé. Les méthodes utilisant les HMMs (Baldi *et al.*, 1994; Krogh *et al.*, 1994) sont néanmoins limitées par les facteurs suivants (Hughey and Krogh, 1996):

- Elles nécessitent un grand nombre de séquences (N>50) pour assurer un entraînement optimal.
- Lors de l'entraînement, la procédure ne garantit pas l'obtention du HMM optimal. Elle peut donc converger vers un minimum local. Par conséquent, l'alignement calculé peut ne pas être optimal.
- Les HMMs ne représentent pas correctement les régions riches en *gaps*.

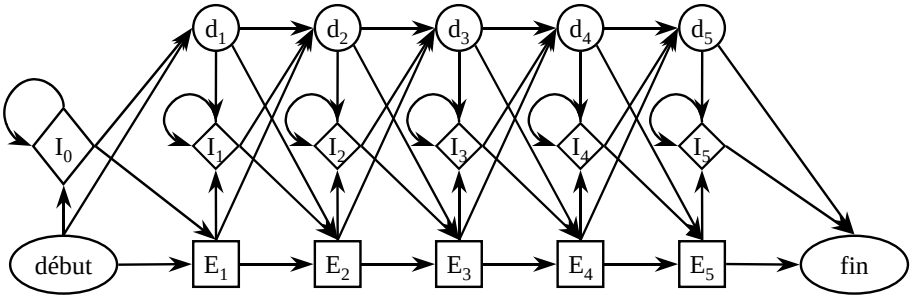


Figure 10: Exemple de modèle caché de Markov appliqué à l'alignement de séquences. Chaque nœud *i* a un état d'émission (E_i), un état d'insertion (I_i) et un état de délétion (d_i). Chaque état possède une certaine probabilité d'émettre un symbole, et des probabilités de transition (représentées par flèches). La séquence des états est cachée puisque seule la séquence des symboles émis est observable.

I.2.4.4.2. Méthodes progressives par affinement itératif

Beaucoup de méthodes d'affinement itératif (Sankoff *et al.*, 1976; Subbiah and Harrison, 1989; Berger and Munson, 1991; Hirosawa *et al.*, 1995) essayent d'optimiser un score calculé pour l'alignement. Le système de score suppose alors que plus le score est important, meilleur sera l'alignement. La méthode PRRP proposée par Gotoh (Gotoh, 1996) n'utilise pas ce principe. Néanmoins, cette méthode est une des plus performantes du point de vue de la qualité des résultats (Thompson *et al.*, 1999). Le principe de cette méthode peut se résumer comme suit:

- Le programme calcule tous les alignements pairés et génère un arbre en utilisant ces alignements. Un alignement multiple est ensuite construit en suivant ce premier arbre.
- Le programme itère jusqu'à convergence en suivant le processus décrit ci-après. Un arbre est déduit des alignements pairés extraits de l'alignement multiple. Ce nouvel arbre sert ensuite de guide pour la

construction d'un nouvel alignement multiple. Le processus est arrêté quand il n'y a pas de différence dans l'alignement multiple d'une itération à l'autre.

I.2.4.5. Méthodes stochastiques

Dans plusieurs domaines scientifiques, des méthodes stochastiques (qui utilisent une variable aléatoire) servent à résoudre des problèmes d'optimisation. Le calcul d'un alignement multiple optimal étant un problème d'optimisation, on y a appliqué deux types de méthodes: le *simulated annealing* (Kirkpatrick *et al*., 1983) et les algorithmes génétiques.

Le *simulated annealing* (Kirkpatrick *et al*., 1983) ou « recuit simulé » tire profit de l'analogie entre le procédé de réchauffement physique des solides pour améliorer leur cristallinité et l'optimisation combinatoire des systèmes complexes. Le simulated annealing est un processus itératif au cours duquel on répète deux opérations:

1. la modification aléatoire du système étudié
2. une acceptation ou le rejet de cette modification.

Les modifications apportées au système au cours de l'opération (1) sont caractérisées par une différence d'énergie (ou pseudo-énergie ou score) ΔE. Si ce ΔE est favorable au système, la modification qui l'a produite est acceptée. Par contre, si ΔE est défavorable, la modification sera acceptée avec une probabilité $p = \exp(-\Delta E/T_n)$ où T_n est un paramètre de contrôle diminuant à chaque itération n. Dans le cas des alignements multiples, chaque modification de l'état est réalisée en déplaçant des *gaps* ou des blocs de *gaps* (Lukashin *et al*., 1992; Ishikawa *et al*., 1993; Kim *et al*., 1994).

Les méthodes GIBBS (Lawrence *et al*., 1993) et PROBE (Neuwald *et al*., 1997) utilisent une stratégie basée sur un échantillonnage itératif: à chaque itération du programme, on compare le score de l'alignement retenu à la distribution des scores obtenue en générant un grand nombre d'alignements avec insertions aléatoires des *gaps*. On peut comparer cette méthode à celles de maximisation des espérances (*expectation maximization*) utilisées dans certains programmes d'alignement multiple local (Lawrence and Reilly, 1990; Cardon and Stormo, 1992). L'approche de tous ces auteurs est basée sur le principe de manque d'information (Goodman, 1974): la probabilité des positions non observées peut être inférée au travers de l'application du théorème de Bayes sur les alignements de séquences observés.

Les algorithmes génétiques simulent le processus naturel d'évolution des gènes: mutations, *crossovers* et sélection. Chaque génération d'individus (ici, d'alignements) est composée d'une population de taille fixe. Les

individus sont évalués par leur adaptation (ici, un score caractérisant l'alignement) et ceux qui sont les plus adaptés ont une meilleure chance de produire une descendance. Le premier a avoir réalisé cette approche pour le calcul de l'alignement multiple était Tajima (Tajima, 1993).

En 1996, Notredame *et al.* (Notredame and Higgins, 1996) ont décrit une approche d'alignement multiple utilisant un algorithme génétique appelé SAGA. Cette méthode utilise une population d'alignements de séquences qui change de manière quasi évolutive et qui améliore graduellement son adaptation. Celle-ci est mesurée par une fonction objectif qui évalue la qualité de l'alignement multiple. L'attrait de cette approche est la capacité d'optimiser n'importe quelle fonction objectif. De manière à augmenter la qualité des alignements multiples, Cédric Notredame a conçu une fonction objectif appelée COFFEE (*Consistency based Objective Function For alignmEnt Evaluation*) (Notredame *et al*., 1998; Notredame *et al*., 2000). Le score COFFEE reflète le niveau de compatibilité entre un alignement multiple et une librairie contenant des alignements pairés de ces mêmes séquences. Le score peut être utilisé comme indice de fiabilité pour des alignements multiples de séquences. Un autre algorithme génétique a été développé par Zhang (Zhang and Wong, 1997) avec une fonction objectif différente.

I.2.4.6. Régions fiables dans un alignement multiple

Le principal problème d'interprétation d'un alignement multiple est de trouver ses régions fiables. Ce problème est bien connu (Smith and Waterman, 1981; Depiereux and Feytmans, 1992; Morgenstern *et al*., 1998) dans les méthodes d'alignement local. Deux systèmes peuvent être utilisés pour déterminer si des régions d'un alignement sont pertinentes ou non: (i) calculer un score et (ii) effectuer une analyse statistique.

I.2.4.6.1. Calcul d'un score

Parmi les systèmes de score présentés par Gotoh (Figure 11) (Gotoh, 1999), certains utilisent les relations phylogénétiques entre les séquences à aligner (arbre, somme pondérée des paires et maximum de vraisemblance) alors que d'autres systèmes ne les utilisent pas (somme des paires, entropie minimum, étoile et consensus). L'idée de base de l'alignement local par programmation dynamique est le calcul d'un score basé sur des pénalités de *gaps* et une matrice de scores (Smith and Waterman, 1981). D'autres stratégies ont été utilisées. Par exemple, Morgenstern (Morgenstern *et al*., 1998) utilise un système pondéré pour calculer des indices de fiabilité et Notredame (Notredame *et al*., 1998; Notredame *et al*., 2000) décrit un système de score reflétant le niveau de compatibilité entre un alignement

multiple et une librairie contenant les alignements pairés des mêmes séquences.

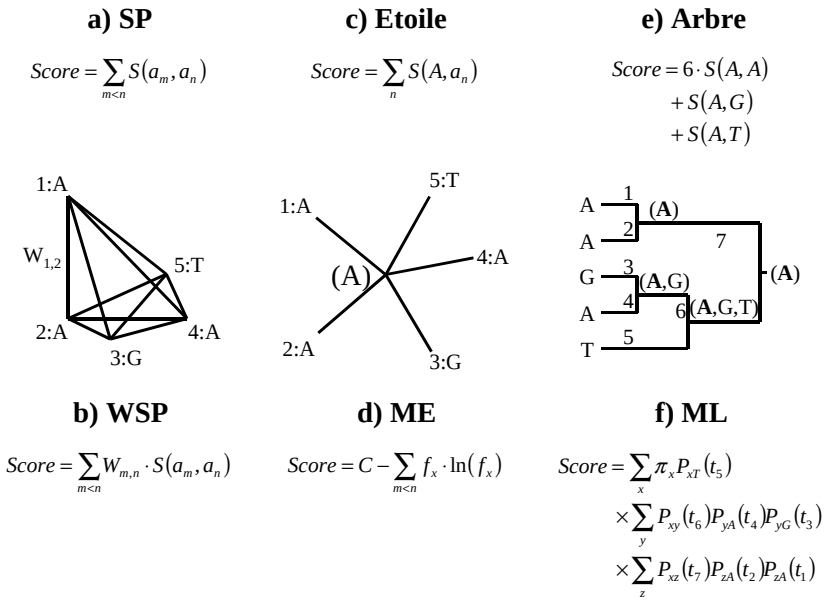


Figure 11: Six systèmes pour attribuer un score à un alignement multiple. (a) Somme des paires (SP). (b) Somme des paires pondérée. (c) Etoile ou score consensus. Dans l'exemple, le résidu le plus fréquent est placé au centre de l'arbre en forme d'étoile. (d) Minimum d'entropie (ME), où C est l'entropie de base calculée à partir de l'abondance moyenne des résidus. (e) Score d'alignement en arbre. Le(s) résidu(s) déduit(s) pour résider à chaque nœud interne est (sont) représenté(s) entre parenthèses. Les résidus en gras sont ceux impliqués dans les changements évolutifs les plus parcimonieux. (f) Score de maximum de vraisemblance. π, représente la distribution *a priori* des résidus des types *x*, et $P_w(t_b)$ indique la probabilité de transition du type de résidus *x* vers *y* pendant la période de t_b.

I.2.4.6.2. Analyse statistique

Le principe de l'évaluation de fiabilité locale d'un alignement par analyse statistique est de recommencer certaines étapes de l'alignement multiple en utilisant des séquences de même longueur et de même composition mais avec un ordre des acides aminés aléatoire. Une analyse statistique est effectuée pour déterminer les positions de l'alignement qui s'écartent hautement du hasard (Goad and Kanehisa, 1982; Gotoh, 1987).

Depiereux (Depiereux *et al*., 1997) effectue une analyse statistique et calibre sa méthode sur un ensemble d'alignements de structures de référence. Les indices assignés par l'algorithme prédisent la confiance associée à chaque position de l'alignement.

Neuwald (Neuwald *et al*., 1997) et Lawrence (Lawrence *et al*., 1993) utilisent les statistiques bayesiennes pour calculer la distribution postérieure des acides aminés dans chaque position alignée et déduisent le logarithme de la probabilité que ces acides aminés soient alignés. Le programme SOAP (Loytynoja and Milinkovitch, 2001) permet de tester la stabilité d'un alignement de séquences réalisé avec ClustalW en fonction des pénalités des *gaps*. Plus récemment, Loytynoja (Loytynoja and Milinkovitch, 2003) utilise les HMMs et les statistiques bayesiennes pour évaluer la fiabilité de chaque colonne d'un alignement de séquences *a posteriori*.

I.2.5. EVALUATION DES PERFORMANCES DES PROGRAMMES D'ALIGNEMENT MULTIPLE

Publication présentée (voir Annexe 2):

P. Briffeuil, G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans and E. Depiereux
Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance predictions reliability
Bioinformatics **14**(4):357-366 (1998).

Peu de chercheurs ont essayé d'évaluer les performances des programmes d'alignement de séquences, en dépit du fait que cette évaluation est nécessaire pour pouvoir mesurer les gains de performance d'une version à l'autre d'un programme ou d'une méthode à l'autre. De plus, les utilisateurs exigent généralement d'être informés des performances qu'ils peuvent attendre de diverses méthodes.

I.2.5.1. Critères d'évaluation des méthodes d'alignement de séquences

McClure (McClure *et al*., 1994) a publié une évaluation des méthodes d'alignement de séquences décrites à l'époque. Elle distinguait deux critères pour l'évaluation de la qualité d'un alignement, et donc, implicitement, d'une méthode d'alignement: l'un était de pouvoir y retrouver des motifs homologues et l'autre, de retrouver les résidus équivalents du point de vue fonctionnel ou structural. Son étude s'était focalisée sur la capacité d'une série de programmes à identifier correctement de courts motifs présents dans quatre familles de protéines homologues. D'autres auteurs ont évalué la sensibilité, pourcentage de colonnes correctes retrouvées par des méthodes d'alignement multiple, en utilisant comme alignements réputés "vrais", des alignements de structures de protéines homologues (Barton and Sternberg, 1987; Barton and Sternberg, 1987; Subbiah and Harrison, 1989; Gotoh, 1996; Morgenstern *et al*., 1998;

Thompson *et al*., 1999; Sauder *et al*., 2000). L'étude la plus récente et probablement la plus complète est celle de Julie Thompson en 1999 (Thompson *et al*., 1999) qui utilise pour son évaluation, la BAliBASE (Thompson *et al*., 1999), une banque de données de 138 alignements réputés corrects.

Dans une toute autre approche, certains auteurs ont évalué la qualité des programmes d'alignement par leur capacité à identifier des nouveaux membres de familles de protéines par des recherches en banque de données (Tatusov *et al*., 1994; Henikoff and Henikoff, 1997; Rychlewski *et al*., 2000).

L'examen détaillé de toutes ces publications montre qu'un critère a principalement été utilisé jusqu'à présent pour évaluer les performances des programmes d'alignement de séquences: la sensibilité. Ce terme est également connu en statistique sous le nom de puissance, c'est-à-dire la quantité de vérité qui a été retrouvée par une méthode de prédiction (Briffeuil *et al*., 1998). Cependant, en statistique, un autre terme tout aussi important permet de juger la pertinence d'une méthode. Il s'agit de la confiance ou sélectivité, qui est la fraction de la prédiction qui est vraie (Briffeuil *et al*., 1998).

Il nous est apparu que la sélectivité des programmes d'alignement de séquences n'avait pas encore été évaluée en dépit du fait qu'une faible sélectivité génère des alignements contenant un nombre appréciable de résidus incorrectement alignés. Ces alignements peuvent alors entraîner des hypothèses erronées et peuvent sérieusement affecter le travail expérimental qui en découle (Briffeuil *et al*., 1998). En 2000, Sauder (Sauder *et al*., 2000) a publié un test de méthodes d'alignement de séquences en utilisant, sans le préciser, les deux critères que nous venons de présenter. En effet, il utilisait deux indices f_D et f_W qui correspondent respectivement au point de vue du développeur (sensibilité) et à celui de l'utilisateur (sélectivité).

Pour les programmes d'alignement de séquences, les critères de sensibilité et de sélectivité ne peuvent être définis que si on dispose des "vrais" alignements. Il faut donc disposer d'une définition standard de ce qu'est un alignement "vrai" ou correct, ce qui fait l'objet de la section suivante. Pour évaluer la sensibilité et la sélectivité d'un programme d'alignement de séquences, on compare les colonnes d'acides aminés prédites par le programme à celle se trouvant dans l'alignement correct.

Pour définir de manière plus formelle les concepts de sensibilité ou de sélectivité, il faut faire appel à une table de confusion comme celle du Tableau 1. Le nombre de colonnes se trouvant aussi bien dans l'alignement de séquences à évaluer et dans l'alignement vrai y est appelé *a*, le nombre de vrais positifs. Le nombre des colonnes de l'alignement à évaluer ne se trouvant pas dans l'alignement vrai est noté *b*, le nombre de faux positifs. Le

nombre des colonnes de l'alignement vrai ne se trouvant pas dans l'alignement à évaluer est noté *c*, le nombre de faux négatifs. Enfin, de par la définition de l'alignement vrai, il n'est pas possible d'y retrouver des acides aminés "mal alignés". Par conséquent, lors de l'évaluation des programmes d'alignement de séquences, les vrais négatifs n'existent pas (*d*=0).

Tableau 1: Les quatre différentes combinaisons de statut réel et de résultat d'un test.

	Statut réel	
Résultat du test	**Positif**	**Négatif**
Positif	*a*	*b*
Négatif	*c*	*d*

Sur base des quantités *a*, *b*, *c* et *d* qui viennent d'être définies, il est maintenant possible de proposer une définition rigoureuse de la sensibilité et de la sélectivité. C'est ce qui est fait dans le Tableau 2. L'intérêt pour le développeur est de pouvoir retrouver un maximum de colonnes correctement alignées (sensibilité élevée). L'intérêt pour l'utilisateur est que l'alignement (ensemble de la prédiction) contienne une proportion élevée de colonnes correctement alignées (sélectivité élevée).

Nous verrons que dans le cas des programmes d'alignement de séquences, sensibilité et sélectivité sont généralement liées. De plus, il sera montré que ces deux paramètres dépendent principalement du pourcentage d'identité entre les séquences à aligner.

Tableau 2: Définitions des quatre paramètres d'évaluation d'un test et les formules pour les calculer.

Paramètre	**Formule**	**Définition**
Sensibilité	$a/(a+c)$	Capacité du test d'identifier les cas réellement positifs
Spécificité	$d/(d+b)$	Capacité du test d'identifier les cas réellement négatifs
sélectivité ou valeur prédictive positive	$a/(a+b)$	Probabilité qu'un résultat positif soit réellement positif
Valeur prédictive négative	$d/(c+d)$	Probabilité qu'un résultat négatif soit réellement négatif

I.2.5.2. Définition des zones fiables d'un alignement de structures

L'évaluation de la sélectivité et de la sensibilité d'un programme d'alignement de séquences ne peut se faire que si l'on se réfère à un ensemble d'alignements réputés corrects sur base d'une certaine définition de ce qui est « correctement aligné ».

La plupart des auteurs (Barton and Sternberg, 1987; Barton and Sternberg, 1987; Subbiah and Harrison, 1989; Gotoh, 1996; Briffeuil *et al*., 1998; Morgenstern *et al*., 1998; Thompson *et al*., 1999; Sauder *et al*., 2000) utilisent des alignements de structures de protéines comme référence. Cependant, les programmes permettant de calculer ces alignements (Sali and Blundell, 1990; Russell and Barton, 1992; Holm and Sander, 1993; Sali and Blundell, 1993; Shindyalov and Bourne, 1998; Ortiz *et al*., 2002) fournissent presque toujours des résultats légèrement différents. C'est pourquoi il est préférable d'utiliser le consensus de plusieurs programmes d'alignement de structures pour établir un alignement de séquences de référence. Récemment, c'est le consensus de deux programmes (CE (Shindyalov and Bourne, 1998) et MAMMOTH (Ortiz *et al*., 2002)) qui a servi à définir les alignements de référence au récent congrès CASP5 (Asilomar, CA, USA, 2002). Ces deux programmes présentent l'avantage d'utiliser aussi bien la similarité globale des structures comparées que la similarité structurale locale, ce qui améliore grandement la qualité de l'alignement final (Shindyalov and Bourne, 1998).

Les programmes d'alignement de structures qui viennent d'être décrits ont servi à constituer des banques de données d'alignements de structures. Citons, 3D-ali (Pascarella and Argos, 1992; Pascarella *et al*., 1996) qui utilise des procédures décrites par Argos et Rossmann (Rossmann and Argos, 1976; Argos and Rossmann, 1979), CAMPASS (Sowdhamini *et al*., 1996) qui utilise SEA (Rufino and Blundell, 1994), FSSP (Holm and Sander, 1996; Holm and Sander, 1998) qui utilise DALI (Holm and Sander, 1993), HOMSTRAD (Mizuguchi *et al*., 1998) qui utilise COMPARER (Sali and Blundell, 1990) et DBAli (Marti-Renom *et al*., 2001) qui utilise MODELLER (Sali and Blundell, 1993). Par contre, BAliBASE (Thompson *et al*., 1999) est construite à partir des différentes banques citées ci-dessus, ainsi que des alignements de référence publiés dans la littérature.

Tous les travaux qui viennent d'être évoqués doivent cependant être examinés d'une façon assez critique. En effet, la définition de ce qui est correctement aligné, structuralement parlant, est loin d'être claire. Suite à la superposition optimale de deux structures, on peut se rendre compte que certaines régions des structures sont plus similaires que d'autres. Dans les régions les moins similaires, la distance moyenne entre les deux structures ou *root mean square deviation* (RMSD) pourra être assez importante. Cette

distance entre deux segments de *w* résidus est calculée par la formule suivante:

$$RMSD = \min\left[\sum_{i=1}^{a \bullet w} \frac{(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2 + (z_{i1} - z_{i2})^2}{a \bullet w}\right]^{0,5}$$

où *a* est le nombre d'atomes considérés par résidu.

x_{ij}, y_{ij}, z_{ij}, sont les coordonnées cartésiennes de l'atome *i* dans le segment *j* (*j*=1,2).

Puisque le nombre d'atomes considérés pour la comparaison n'est pas fixe, la comparaison est généralement limitée aux squelettes protéiques. Malgré son utilité indiscutable, le RMSD ne tient pas compte de la structure tridimensionnelle prise par les fragments, elle n'exprime que la proximité physique entre les atomes. Il faut donc choisir une distance à partir de laquelle deux régions ne seront plus considérées comme suffisamment similaires. Plusieurs auteurs ont déjà abordé le sujet (Unger *et al*., 1989; Briffeuil *et al*., 1998; Sauder *et al*., 2000) mais nous recommandons la distance seuil de 3 Å mise en évidence en 2000 par Sauder (Sauder *et al*., 2000). Il a déduit cette distance d'une vaste comparaison de structures de protéines. Par contre, Briffeuil *et al*. (Briffeuil *et al*., 1998) ont choisi la distance de 1.8 Å pour des segments de 9 acides aminés et Unger (Unger *et al*., 1989), 1 Å pour des segments de 6 résidus. L'utilisation d'une distance seuil est la méthode la plus correcte mais cependant la plus difficile à mettre en œuvre. C'est pourquoi, dans la suite de notre travail, nous avons utilisé deux autres définitions de ce qui est correct dans un alignement de structure (Figure 12). Celles-ci sont imparfaites théoriquement mais présentent l'avantage d'être applicables à grande échelle:

- La première consiste à prendre toutes les régions de l'alignement de structures ne contenant pas de *gap* (ASG: Aligné Sans *Gap*). Dans ce cas, nous surestimons la vérité par rapport aux régions de l'alignement validées par le RMSD seuil de 3 Å.
- La seconde est un sous-ensemble de la première ne contenant que les régions où la structure secondaire (définie par le programme DSSP (Kabsch and Sander, 1983)) est identique dans toutes les structures alignées (CSS: Conservation de la Structure Secondaire). Dans ce cas, nous sous-estimons la vérité par rapport aux régions de l'alignement validées par le RMSD seuil de 3 Å.

Lors de l'évaluation des qualités d'un alignement de séquences donné, nous aurons les relations suivantes:

$$ncorr_{ASG} > ncorr_{RMSD} > ncorr_{CSS}$$

où $ncorr_{ASG}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences à évaluer en utilisant le critère ASG

$ncorr_{RMSD}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences à évaluer en utilisant le critère du RMSD local

$ncorr_{CSS}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences à évaluer en utilisant le critère CSS

$$nref_{ASG} > nref_{RMSD} > nref_{CSS}$$

où $nref_{ASG}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences de référence en utilisant le critère ASG

$nref_{RMSD}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences de référence en utilisant le critère du RMSD local

$nref_{CSS}$ est le nombre d'acides aminés correctement alignés dans l'alignement de séquences de référence en utilisant le critère CSS

Or, la sélectivité est le rapport entre le nombre d'acides aminés correctement alignés dans l'alignent de séquences à évaluer (ncorr) et le nombre d'acides aminés alignés dans l'alignement de séquences à évaluer. Ce dernier étant constant, quel que soit le critère utilisé pour définir les acides aminés correctement alignés, nous aurons la relation suivante:

$$sélectivité_{ASG} > sélectivité_{RMSD} > sélectivité_{CSS}$$

De plus, la sensibilité est le rapport entre le nombre d'acides aminés correctement alignés dans l'alignent de séquences à évaluer (ncorr) et le nombre d'acides aminés alignés dans l'alignement de séquences de référence (nref). D'après les relations précédentes nous aurons donc l'inégalité suivante:

$$sensibilité_{ASG} \neq sensibilité_{RMSD} \neq sensibilité_{CSS}$$

et, généralement, de par la plus grande facilité d'aligner correctement les acides aminés situés dans des régions où la structure secondaire est conservée, nous aurons la relation suivante entre les sensibilités calculées suivant les différents critères:

$$sensibilité_{ASG} < sensibilité_{RMSD} < sensibilité_{CSS}$$

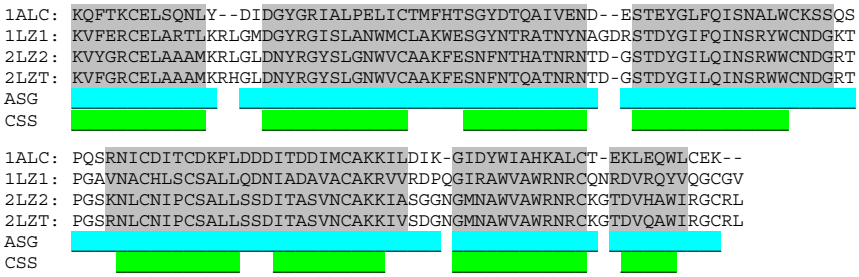


Figure 12: Alignement de séquences de lysozymes de *Homo sapiens* (1LZ1), de *Meleagris gallopavo* (2LZ2) et de *Gallus gallus* (2LZT), et d'une alpha-lactalbumine (1ALC) de *Papio hamadryas cynocephalus*. Les zones correspondant au critère du RMSD local sont colorées en gris, et les zones correspondant aux critères ASG et CSS sont surlignées respectivement en bleu turquoise et en vert.

I.2.5.3. Performances des programmes d'alignement de séquences

Un grand nombre de programmes d'alignement de séquences ont été développés jusqu'à présent et ils n'offrent pas tous les mêmes performances d'alignement. Il est donc nécessaire de comparer leurs performances pour sélectionner le meilleur programme pour un type de problème donné.

Thompson (Thompson *et al.*, 1999) a comparé les programmes PRRP (Gotoh, 1996), ClustalW (Thompson *et al.*, 1994), SAGA (Notredame and Higgins, 1996), Multalign (Barton and Sternberg, 1987), PileUp (Wisconsin Package v.8; Accelrys Inc., San Diego, CA, USA), Multal (Taylor, 1988), hmmt (Eddy, 1995), DIALIGN (Morgenstern *et al.*, 1998) et PIMA (Smith and Smith, 1992). Elle a étudié la sensibilité de ces programmes en fonction de différentes caractéristiques des séquences à aligner: séquences très similaires, séquences peu similaires, séquences présentant de larges insertions ou délétions. Ses conclusions rejoignent en partie celles publiées précédemment (Barton and Sternberg, 1987; Barton and Sternberg, 1987; Subbiah and Harrison, 1989; McClure *et al.*, 1994; Gotoh, 1996; Morgenstern *et al.*, 1998; Sauder *et al.*, 2000):

- Aucun programme n'est significativement meilleur que les autres. Par contre, pour un ensemble de séquences donné, un programme fournira un meilleur alignement que les autres.
- La qualité de l'alignement dépend du niveau de similarité des séquences à aligner, mais aussi de leurs longueurs ou de la taille des insertions présentes dans certaines séquences.

- Si les séquences partagent entre 20 et 30% d'identité, l'alignement devient en grande partie incorrect.
- Les programmes d'alignement global (PRRP, ClustalW et SAGA) obtiennent de meilleurs résultats si la similarité entre les séquences est répartie sur toute leur longueur. Par contre, les programmes produisant un alignement local (DIALIGN et PIMA) sont meilleurs si les séquences possèdent de larges insertions ou délétions ou si la similarité est concentrée sur quelques segments.

Quant à Briffeuil *et al.* (Briffeuil *et al*., 1998), ils ont testé sept programmes d'alignement de séquences disponibles sous forme de serveur *web*. Ces programmes étaient PIMA (Smith and Smith, 1992), ClustalW (Thompson *et al*., 1994), MAP (Huang, 1994), BlockMaker-Motif (Henikoff *et al*., 1995), BlockMaker-Gibbs (Henikoff *et al*., 1995), Match-Box (Depiereux and Feytmans, 1992; Depiereux *et al*., 1997) et MEME (Grundy *et al*., 1996). Les conclusions principales de Briffeuil *et al.* étaient les suivantes:

- Les programmes d'alignement global (PIMA, ClustalW et MAP) ont une sensibilité variant de 75% (PIMA) à 81% (MAP) et celle-ci est liée de manière linéaire à la sélectivité qui varie de 65% (PIMA) à 73% (MAP).
- Les programmes d'alignement local (deux versions de BlockMaker et MEME) ont une sensibilité faible, variant de 24% (BlockMaker-Motif) à 40% (MEME) malgré une sélectivité variant de 70% (MEME) à 75% (BlockMaker-Gibbs). Dans ce cas, la relation linéaire entre sélectivité en sensibilité est moins évidente et dépend d'un cas à l'autre.
- Le programme Match-Box peut fournir différents niveaux de sélectivité suivant les indices de confiance des positions prises en compte (Figure 16, page 41). En prenant toutes les positions prédites par ce programme, on obtient une sélectivité supérieure à celle des autres programmes (77%) et une sensibilité de 68%. Malgré ses performances comparables aux autres programmes d'alignement multiple, il semble que la relation entre sensibilité et sélectivité pour Match-Box varie, comme pour les autres programmes d'alignement local, d'un cas à l'autre sans être clairement linéaire.
- Le consensus de plusieurs méthodes d'alignement de séquences bien choisies permet d'améliorer sensiblement la sélectivité sans réelle perte de sensibilité.

I.2.6. FONCTIONNEMENT DU PROGRAMME MATCH-BOX

Match-Box (Depiereux and Feytmans, 1991) est un programme d'alignement de séquences développé au sein de notre laboratoire. Le fonctionnement de ce programme se déroule en deux étapes. La première, appelée EXPLORE, effectue une analyse statistique sur les séquences pour permettre à l'utilisateur de déterminer leur similarité globale et, ainsi, de prédire la fiabilité de l'alignement de séquences final. Ces résultats n'étant pas utilisés pour réaliser l'alignement des séquences, nous ne décrirons pas davantage cette procédure. La deuxième partie, appelée ALIGN, réalise l'alignement des séquences. Son fonctionnement se déroule en trois étapes principales (Figure 13): le SCANNING analyse les séquences et calcule de manière statistique les meilleures valeurs des paramètres utilisés dans les étapes qui suivent; le MATCHING recherche des segments de longueur fixe conservés dans toutes les séquences et génère des alignements multiples des segments désignés par les auteurs sous le nom de "boîtes", et qui correspondent aux "*blocks*" de BLOCK-MAKER; finalement, le SCREENING sélectionne les boîtes les plus appropriées pour la réalisation de l'alignement final. Celui-ci est calculé en quatre itérations *matching-screening* pendant lesquelles on augmente progressivement un seuil statistique ($T_{i,j}$ voir section I.2.6.2, description du *matching*). Ce seuil statistique est utilisé pour déterminer un indice de confiance pour chaque position alignée.

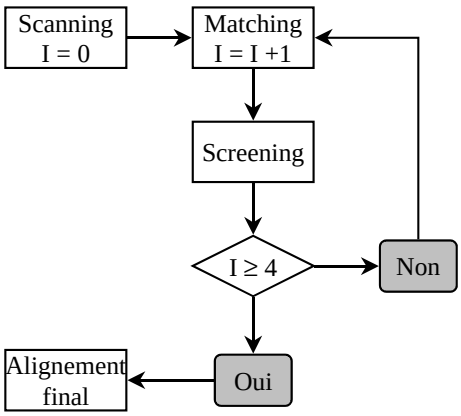


Figure 13: Schéma général du fonctionnement de Match-Box.

I.2.6.1. Algorithme de « SCANNING »

Un segment de *W* résidus consécutifs est défini à partir de sa position *j* dans une séquence *i*. Chaque segment (segment d'analyse) de chaque séquence est comparé avec tous les segments de même longueur dans chaque autre séquence. La distance entre deux segments est calculée par la formule suivante:

$$D_{\mathrm{i,j,l,m}} = \sum_{k=0}^{W-1} y_{i,\,j+k,l,m+k}$$

où *i* est le numéro de la séquence contenant le premier segment
j est la position du premier acide aminé de ce segment
l est le numéro de la séquence contenant le deuxième segment
m est la position du premier acide aminé du deuxième segment
$y_{i,j+k,l,m+k}$ est la distance séparant les acides aminés aux positions $j+k$ de la séquence *i* et $m+k$ de la séquence *l*. *y* peut prendre 210 valeurs qui sont fournies par la matrice de scores choisie exprimée en distance et non en similarité.
$D_{i,j,l,m}$ est la distance séparant les deux segments

Le minimum global des $D_{i,j,l,m}$ est défini comme le meilleur appariement, $D_{i,j,l}$. Un seuil peut être fixé sur $D_{i,j,l,m}$ pour déterminer quand les segments sont considérés comme significativement similaires. Pour déterminer au préalable ce seuil, l'ordre des acides aminés des séquences est modifié de manière aléatoire. La distribution des $D_{i,j,l,m}$ dans les séquences originales est comparée à la distribution des $D_{i,j,l,m}$ dans les séquences aléatoires. Un exemple de distribution est repris à la Figure 14 dans laquelle on peut distinguer cinq zones.

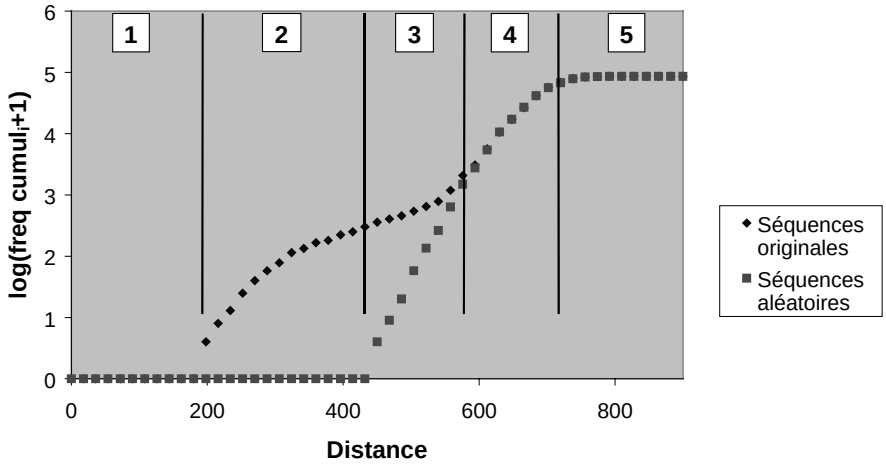


Figure 14: Distribution des fréquences cumulées des $D_{i,j,l,m}$ dans les séquences originales (losanges) et les séquences aléatoires (carrés), exprimées sous une échelle logarithmique. Le diagramme a été découpé en cinq régions correspondant aux quatre seuils statistiques définis dans le texte.

Dans la première région, il n'y a pas d'appariement de segments, ni dans les séquences originales, ni dans les séquences aléatoires. Dans la deuxième région, il n'y a des appariements que pour les séquences originales. La distance à laquelle apparaît le premier appariement entre les séquences originales est un seuil qui sera utilisé plus tard, dans le *matching*. On le note $S_{i,j,1}$, où i est le numéro de la première séquence, j celui de la deuxième séquence et 1 est le numéro du seuil. Dans la troisième région, il y a apparition d'appariements dans les séquences aléatoires et donc, le « signal » venant des séquences originales commence à contenir un bruit de fond. La distance à laquelle le premier appariement apparaît dans les séquences aléatoires est notée $S_{i,j,2}$. Dans la quatrième région, il n'est plus possible de distinguer les deux distributions d'un point de vue statistique: la distance à partir de laquelle ce phénomène se produit est notée $S_{i,j,3}$. Enfin, dans la cinquième partie, il n'y a plus d'appariement entre les séquences originales et la distance déterminant le début de cette partie est notée $S_{i,j,4}$. Le seuil permettant de déterminer quand deux segments sont considérés comme significativement similaires a été défini par les auteurs de Match-Box suivant la formule suivante:

$$S_{i,j} = \frac{S_{i,j,2} + S_{i,j,3}}{2}$$

I.2.6.2. Algorithme de « MATCHING »

Un segment de *W* résidus à la position *j* d'une séquence *i* est comparé aux segments de *W* résidus de chaque autre séquence *l* (Figure 15 A). Dans chaque séquence, on retient les segments correspondant à la distance minimum, $D_{i,j,l}$. Si plusieurs segments correspondent à cette distance, seul le premier segment trouvé est retenu. La boîte formée par les segments ainsi retenus est alors analysée par un filtre statistique: tous les segments de la boîte sont comparés entre eux (Figure 15 B). La boîte est rejetée si la distance entre deux segments est supérieure à un certain seuil $T_{i,j}$ dont la valeur initiale est $S_{i,j,k}$ où *k* est le numéro de l'itération. Si après avoir effectué ces opérations pour tous les segments de toutes les séquences, toutes les boîtes sont rejetées, on recommence le *matching* après avoir augmenté le seuil $T_{i,j}$. Sinon, les boîtes sont traitées dans le *screening*.

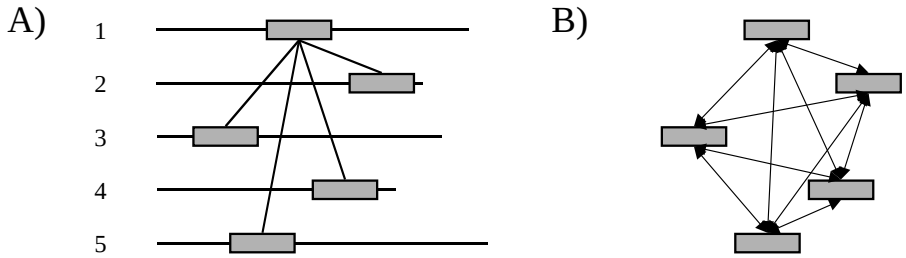


Figure 15: A) Comparaison d'un segment de la séquence 1 à tous les segments des autres séquences. Le segment correspondant à la distance minimale ($D_{i,j,l}$) est sélectionné dans chaque séquence (rectangles gris) pour former une boîte. B) Comparaison de tous les segments de la boîte deux à deux.

Pour bien comprendre les travaux sur le *matching* qui seront présentés dans la section IV.3, il faut préciser que l'algorithme qui vient d'être décrit porte le nom de *matching_MB*. Lorsque le filtre est utilisé lors de la phase de *matching_MB*, celui-ci s'arrête dès qu'au moins une boîte est retenue. La sensibilité de la recherche est donc extrêmement réduite alors que la sélectivité est très importante. Les capacités de sélection du *screening* ne seront alors pas vraiment utilisées.

D'autre part, si le *matching* est effectué sans filtre statistique, il portera le nom de *matching_SF*. Dans celui-ci, toutes les boîtes sont automatiquement sélectionnées. La sensibilité est alors importante et la sélectivité est très faible. La sélection des boîtes « correctes » repose donc entièrement sur la phase de *screening*.

I.2.6.3. Algorithme de « SCREENING »

Au cours du *screening*, les boîtes provenant du *matching* sont d'abord comparées entre elles pour tenter de former des boîtes plus longues: si la différence entre les positions des premiers résidus de chaque segment de la première boîte et chaque segment de la deuxième boîte est constante et inférieure à *W*, il est possible de remplacer les deux boîtes par une seule boîte plus longue. Lorsque cette comparaison est terminée, on obtient un ensemble de boîtes de longueurs différentes mais supérieures ou égales à *W*.

Ensuite, la génération d'une ébauche de l'alignement final peut commencer. La boîte la plus longue est sélectionnée car on la considère *a priori* comme la meilleure (celle qui a le moins de chance d'être le fruit du hasard). L'algorithme cherche ensuite toutes les autres boîtes compatibles (deux boîtes, A et B, sont compatibles si les acides aminés de la boîte A sont situés tous à gauche - ou à droite – de ceux de la boîte B) avec la première et sélectionne celle qui maximise le score empirique *F* suivant:

$$F = \frac{L}{\sqrt{1 + SCE\Delta_g}}$$

où *L* est la longueur de la boîte recherchée

SCEΔg est la somme des carrés des écarts des différents décalages Δ_g entre les boîtes.

Les boîtes incompatibles avec cette boîte sont éliminées et la recherche est recommencée jusqu'à ce que toutes les boîtes aient été sélectionnées ou éliminées.

Les boîtes sélectionnées par le *screening* sont utilisées comme points d'ancrage pour la construction de l'alignement final. Elles sont donc figées pendant les étapes de *matching* des itérations suivantes. Celles-ci seront alors limitées aux segments constituant les zones non-alignées et sera effectuée avec un seuil $T_{i,j}$ moins stringent ($T_{i,j}$ plus élevé) pour rechercher des boîtes moins fiables du point de vue statistique.

Lors de la présentation de nos travaux concernant le *screening* (section IV.4), nous désignerons l'algorithme de *screening* de Match-Box par *screening_MB*.

I.2.6.4. Indice de confiance

Match-Box calcule un indice de confiance *I* pour chaque boîte sur base du seuil $T_{i,j}$ utilisé pour sélectionner la boîte. Cet indice *I* est calculé par la formule suivante:

$$
\begin{aligned}
I = \quad & 1, \text{ si } T_{i,j} < X_1 \\
& 2, \text{ si } (T_{i,j} > X_1) \text{ et } (T_{i,j} < X_2) \\
& 3, \text{ si } (T_{i,j} > X_2) \text{ et } (T_{i,j} < X_3) \\
& 4, \text{ si } (T_{i,j} > X_3) \text{ et } (T_{i,j} < X_4) \\
& 5, \text{ si } (T_{i,j} > X_4) \text{ et } (T_{i,j} < X_5) \\
& 6, \text{ si } (T_{i,j} > X_5) \text{ et } (T_{i,j} < X_6) \\
& 7, \text{ si } (T_{i,j} > X_6) \text{ et } (T_{i,j} < X_7) \\
& 8, \text{ si } (T_{i,j} > X_7) \text{ et } (T_{i,j} < X_8) \\
& 9, \text{ si } T_{i,j} > X_8
\end{aligned}
$$

où $T_{i,j}$ est le seuil utilisé pour sélectionner la boîte

I et l'indice de confiance obtenu

X_i sont des seuils spécifiques à la matrice de scores utilisée, et déterminés par un processus d'optimisation au cours duquel on a tenté d'obtenir, pour un ensemble d'alignements de références, une relation linéaire entre l'indice *I* calculé et la sélectivité observée

La valeur de l'indice de confiance de Match-Box s'étend sur une échelle allant de 1 (haute sélectivité, faible taux de faux positifs) à 9 (faible sélectivité, taux de faux positifs important). La relation entre l'indice de confiance de Match-Box et la sélectivité observée dans des alignements de référence est quasi linéaire (Figure 16): on peut donc convertir facilement la valeur de l'indice *I* en sélectivité, pour chaque position alignée. Les résultats présentés ont été obtenus pour les 4552 positions alignées d'un ensemble de 33 alignements de référence (voir section IV.1).

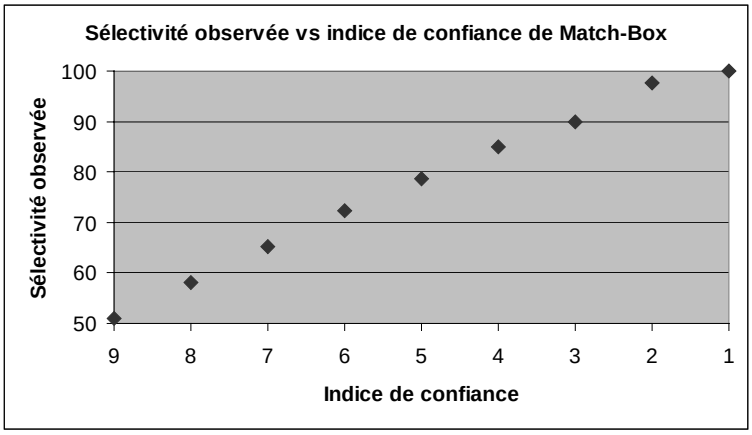


Figure 16: Sélectivité observée dans les résultats du programme Match-Box en fonction de l'indice de confiance fourni par ce programme.

Un alignement réalisé par Match-Box (Figure 17) peut être découpé en zones correspondant à des valeurs de l'indice de confiance. On peut, par exemple, isoler les colonnes de l'alignement possédant un indice de confiance inférieur à 5, et qui auront 50% de chance ou plus d'être correctement alignées.

```
               10        20        30        40        50        60        70
                +         +         +         +         +         +         +
1ALC  kqftkcelsqnlyd--idgygrialpelictmfhtsgydtqai--vendesteyglfqisnalwckssqs
1LZ1  kvfercelartlkrLGmdgyrgislanwmclakwesgyntratNYnagdrstdygifqinsrywcndgkt
2LZ2  kvygrcelaaamkrLGldnyrgyslgnwvcaakfesnfnthatN-rntdgstdygilqinsrwwcndgrt
2LZT  kvfgrcelaaamkrHGldnyrgyslgnwvcaakfesnfntqatN-rntdgstdygilqinsrwwcndgrt

      33333333333337  333333333777777773333333333  311111111111111111111111

               80        90       100       110       120       130
                +         +         +         +         +         +
1ALC  pqsrnicditcdkflddditddimcakkil-dikgidywiahkalctekleqwlcek---
1LZ1  pgavnachlscsallqdniadavacakrvvRdpqgirawvawrnrcqnrdvrqyvqgCGV
2LZ2  pgsknlcnipcsallssditasvncakkiaSggngmnawvawrnrckgtdvhawirgCRL
2LZT  pgsrnlcnipcsallssditasvncakkivSdgngmnawvawrnrckgtdvqawirgCRL

      113333333333333333333333333333 333333333333333337799999999
```

Figure 17: Alignement de séquences de lysozymes de *Homo sapiens* (1LZ1), de *Meleagris gallopavo* (2LZ2) et de *Gallus gallus* (2LZT), et d'une alpha-lactalbumine (1ALC) de *Papio hamadryas cynocephalus* réalisé avec le programme Match-Box. Les positions alignées par le programme sont reprises en minuscule. Un indice de confiance variant de 1 à 9 est calculé pour chaque position alignée.

I.3. Réseaux neuronaux

Le modèle des réseaux neuronaux est très utilisé en bioinformatique et trouve son origine dans les efforts réalisés pour modéliser les opérations qui prennent place dans le système nerveux. Ce modèle est constitué de deux éléments essentiels:

- Un ensemble d'unités, ou neurones, qui fonctionnent indépendamment et en parallèle et qui reçoivent des signaux des unités auxquelles elles sont connectées.
- Le motif de connexion entre les neurones qui relie les neurones entre eux. Ce motif définit la structure ou topologie du réseau neuronal. Les signaux sont modulés par une force de connexion entre les unités. Les unités combinent ces signaux et génèrent un résultat. Ce résultat est ensuite propagé aux autres unités du réseau.

Il a été montré (Hebb, 1949; Rosenblatt, 1958) qu'un apprentissage peut prendre place dans ce type de modèle au travers d'une modification adaptative des forces de connexion et des seuils (appelés collectivement les poids du réseau neuronal). Les réseaux neuronaux tels que décrits ici ne correspondent pas à un processus biologique. Néanmoins, ils peuvent être vus comme un outil mathématique capable d'apprentissage automatique, à partir d'exemples d'entraînement, pour reproduire une correspondance entre un ensemble de données d'entrée et un ensemble de résultats de sortie. Quand le réseau a été entraîné (optimisé), il peut effectuer des prédictions sur des données nouvelles. Les performances de ces prédictions dépendent de la topologie du réseau, de l'encodage des entrées et sorties, et de la structure sous-jacente du problème dont les exemples d'entraînement et de test sont tirés.

I.3.1. TOPOLOGIE ET CALCUL

Un exemple de réseau neuronal est donné dans la Figure 18 (programme PHD (Rost and Sander, 1993)). Dans ce cas, il prédit la structure secondaire d'une protéine dans trois états particuliers: hélice α, brin β et boucle. Les rectangles représentent les unités de calcul et les lignes reliant ces unités sont des connexions au travers desquelles les signaux de sortie d'une unité passent en signaux d'entrée d'une autre unité. Chacun des 13 rectangles montrés dans la couche d'entrée représente un groupe d'unités utilisées pour encoder l'acide aminé de la fenêtre à la position correspondante. Chaque groupe consiste en 21 unités d'entrée, une pour chaque acide aminé possible à cette position et une entrée nulle utilisée quand la fenêtre de 13 acides aminés est au début ou à la fin de la séquence

protéique. Donc, pour une fenêtre de 13 acides aminés, 13 des 273 entrées du réseau seront mises à 1 et les autres seront mises à 0. Un bloc de la couche cachée ou de la couche de sortie représente une seule unité. La prédiction est effectuée pour l'acide aminé central de la fenêtre d'entrée.

Les 7 neurones de la couche cachée traitent et propagent l'information venant des neurones de la couche d'entrée vers les 3 neurones de sortie. Ceux-ci fournissent, pour l'acide aminé central, la probabilité qu'il soit dans une hélice α, dans un brin β ou dans une boucle.

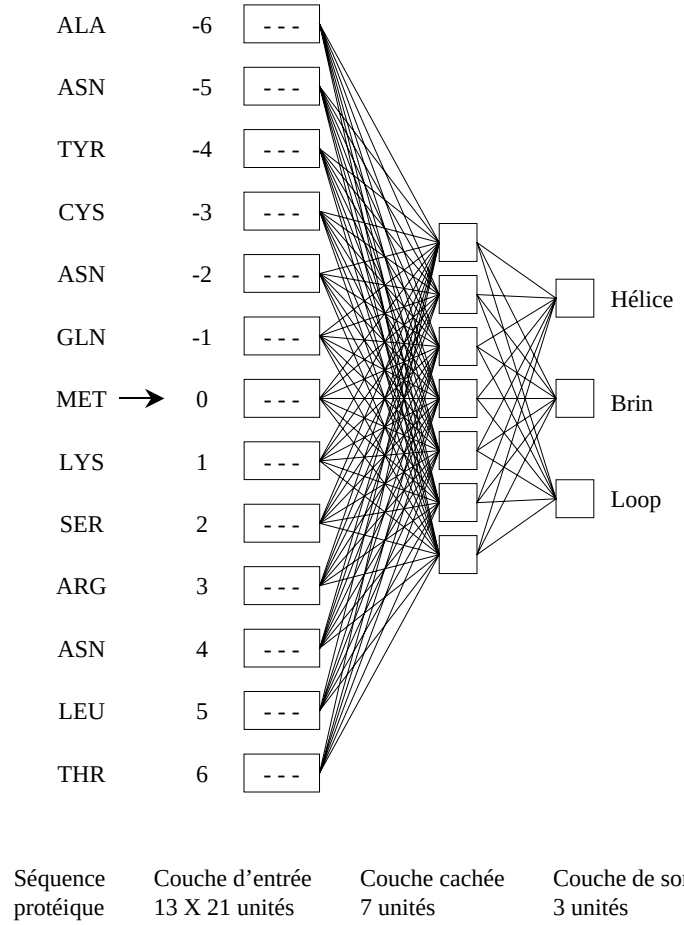


Figure 18: Topologie du réseau neuronal du programme PHD qui prédit la structure secondaire des protéines.

Généralement, les connexions entre les unités peuvent se faire dans n'importe quelle direction entre n'importe quelle paire de neurones. Dans cette thèse, nous ne nous intéresserons qu'aux réseaux dont la topologie est constituée de trois couches: une couche d'entrée, une couche cachée et une couche de sortie. Dans ce cas, les connexions entre les neurones ne doivent

s'effectuer qu'entre un neurone d'une couche et un neurone de la couche directement supérieure, comme représenté dans la Figure 18. La structure des couches d'entrée et de sortie est imposée par le système d'encodage choisi par le problème. Le nombre optimum et la taille des couches cachées doivent être déterminés empiriquement pour chaque problème modélisé.

Les calculs qui se déroulent dans chaque neurone sont représentés dans la Figure 19. Lors de la propagation des signaux entre chaque couche du réseau neuronal, le calcul se déroule dans chaque unité de la couche cachée et chaque unité de la couche de sortie. Les sorties de la couche *i* précédente, Y_i, avec les forces de connexion positives ou négatives entre les unités *i* et *k*, W_{ik}, sont additionnées sur toutes les entrées de l'unité et ajustées avec un seuil, b_k. Le résultat Y_k de l'unité *k* est alors généré d'après la formule donnée ci-dessus et propagé à la couche suivante du réseau. Vu la formule de calcul de Y_k (tangente hyperbolique), ce résultat est toujours compris entre 0 et 1. Cette fonction sigmoïde fournit un comportement similaire à celui d'un commutateur quand les unités passent du statut inhibiteur (négatif) à celui d'excitateur (positif).

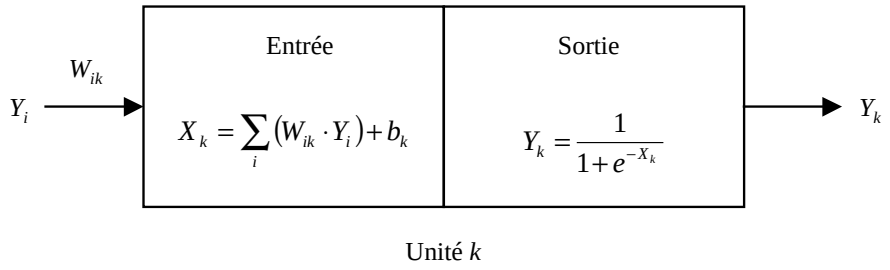


Figure 19: Traitement des données dans un neurone.

I.3.2. ENTRAÎNEMENT DU RÉSEAU NEURONAL

Après avoir effectué les choix de topologie et de l'encodage des entrées et sorties, un réseau neuronal subit un entraînement. Dans cette phase, un ensemble d'entraînement de paires entrée/sortie est présenté au réseau et des ajustements des différents paramètres W_{ik} et b_k sont effectués pour reproduire au mieux ces correspondances. L'entraînement du réseau neuronal peut être vu comme un problème d'optimisation. En effet, il consiste à minimiser l'erreur totale (typiquement, la somme des carrés des écarts entre les valeurs calculées par les unités de sortie et les valeurs réelles) en agissant sur les poids du réseau neuronal. Lorsqu'on évalue les performances du réseau à chaque itération de minimisation de l'erreur totale sur un ensemble de test, on se rend compte (Figure 20) que l'erreur commence par diminuer puis augmente. Ce phénomène est connu sous le nom de « surentraînement »: le réseau est devenu trop spécifique de

l'ensemble d'entraînement. Il faut donc arrêter l'entraînement juste avant que l'erreur totale sur l'ensemble de test n'augmente. Dans un cas réel d'entraînement sur un ensemble de données, on utilise la technique d'entraînement par validation croisée sur, par exemple, 6 parties. Pour cela, on divise l'ensemble des données en 6 sous-ensembles de tailles plus ou moins identiques. On entraîne le réseau sur 4 ensembles, on contrôle l'évolution des performances sur le cinquième et on évalue les performances sur le dernier. Ces opérations sont effectuées 6 fois en permutant à chaque fois les ensembles d'entraînement, de test et d'évaluation. On obtient ainsi 6 réseaux qui seront tous utilisés pour effectuer les prédictions, le résultat final étant soit la moyenne des résultats des 6 réseaux, soit le résultat majoritaire de ces 6 réseaux. Cette technique permet d'éviter certains problèmes liés à l'optimisation d'un seul réseau et permet d'avoir un outil de prédiction entraîné globalement sur toutes les données disponibles. La performance moyenne du système est calculée comme étant la moyenne des performances de chacun des réseaux sur les ensembles d'évaluation.

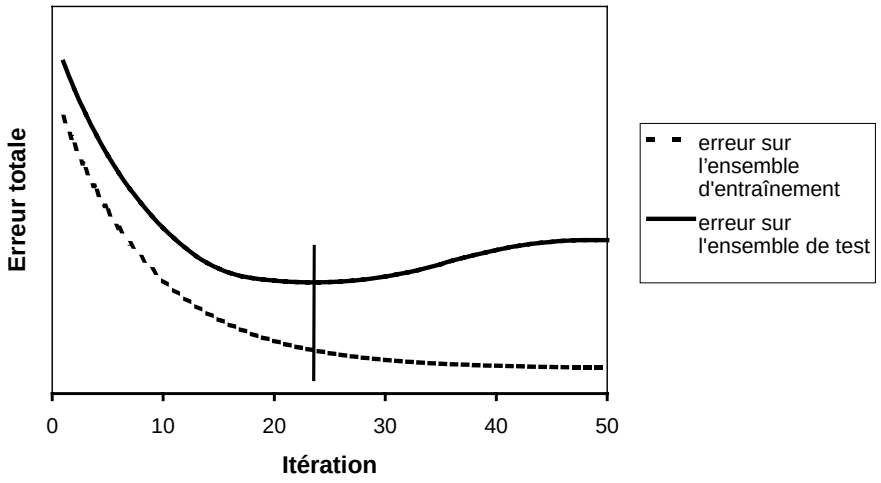


Figure 20: Evolution des erreurs totales sur l'ensemble d'entraînement et sur l'ensemble de test en fonction du nombre d'itérations de la minimisation. Après la 24ème itération, le réseau neuronal devient trop spécifique de l'ensemble d'entraînement.

I.4. Prédiction de structure secondaire

Le but des méthodes de prédiction de structure secondaire est de prédire la localisation des structures secondaires sur la séquence des protéines en se basant uniquement sur des informations contenues dans cette séquence. Selon l'exposé de Rost (Rost and Sander, 2000), ces méthodes peuvent être regroupées en trois types:

- Les méthodes de première génération (entre 1960 et 1980) qui sont basées sur la tendance naturelle de certains acides aminés à se trouver dans certaines 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 secondaires de 15 protéines non homologues de structures connues. Mais cette méthode est très peu fiable (précision < 50%) (Kabsch and Sander, 1983). De plus, le nombre de séquences sur lesquelles ils se sont basés était trop faible pour couvrir un nombre suffisamment large d'observations. D'autres méthodes (Lim, 1974; Garnier *et al*., 1978) se basent sur le même principe mais tiennent compte, en outre, des interactions entre les résidus et leur environnement local via la théorie de l'information.
- Les méthodes de seconde génération, qui ont dominé la scène jusqu'au début des années 1990, utilisent les tendances naturelles de segments allant de 3 à 51 résidus adjacents à se trouver dans certaines structures secondaires (Ptitsyn and Finkelstein, 1983; Levin *et al*., 1986; Gibrat *et al*., 1987; Biou *et al*., 1988; Levin and Garnier, 1988; Zhang *et al*., 1992). Cependant, ces types de méthodes, qui ne reposent que sur l'information contenue dans une seule séquence, ne prédisent les structures secondaires correctes que pour à peu près 60% des résidus.
- Les méthodes de troisième génération ont bénéficié d'améliorations de différentes sources. Dans un premier temps, Dickerson *et al.* (Dickerson *et al*., 1976) se sont rendu compte que l'information contenue dans les alignements de séquences pouvait améliorer les prédictions. Zvelebil *et al.* (Zvelebil *et al*., 1987) ont incorporé ce concept dans une méthode de prédiction automatique de la structure secondaire des protéines. Cependant, les méthodes de troisième génération ont pu atteindre des exactitudes de prédiction supérieures à 70% grâce à l'utilisation combinée de grandes banques de données de séquences protéiques et des algorithmes plus complexes (Rost and Sander, 1993; Rost and Sander, 2000). Un élément majeur de ces nouvelles méthodes est l'utilisation de

l'information phylogénétique contenue dans des profils spécifiques de la position (PSSM: *Position Specific Scoring Matrix*) décrivant quel résidu peut être échangé avec un autre et à quelle position. Ces profils contiennent des informations importantes à propos de la structure de la protéine (Rost, 2001). La mise au point d'outils de recherche tels que PSI-BLAST (Altschul *et al*., 1997), SAM-T99 (Karplus *et al*., 1998) et HMMsearch (Eddy, 1998) a permis de générer des PSSMs plus fiables pour des protéines peu similaires (Rost, 2001). Les performances continuent de croître grâce à l'utilisation des réseaux neuronaux ou des HMMs. Ainsi, d'après Eyrich *et al*. (Eyrich *et al*., 2001), les méthodes suivantes ont une précision supérieure à 70%: HMMSTR (Bystroff *et al*., 2000), PHD (Rost *et al*., 1994), PSIPRED (Jones, 1999), PROF (Rost, 2000), GORV (Kloczkowski *et al*., 2002), JPred2 (Cuff and Barton, 2000), SAM-T99sec (Karplus *et al*., 1998) et SSpro (Baldi *et al*., 1999).

Les informations sur la structure secondaire des protéines permettent d'améliorer la précision des programmes d'alignement de structures (Shindyalov and Bourne, 1998; Ortiz *et al*., 2002), d'alignement de séquences (Thompson *et al*., 1994; Jennings *et al*., 2001; Shi *et al*., 2001), de recherche en banque de données de séquences (Rost, 1995; Fischer and Eisenberg, 1996; Russell *et al*., 1996) et des programmes de recherche du repliement des protéines (Rost, 1995; Fischer and Eisenberg, 1996; Russell *et al*., 1996; Ayers *et al*., 1999; de la Cruz and Thornton, 1999; Di Francesco *et al*., 1999; Jones, 1999; Jones *et al*., 1999; Kelley *et al*., 2000); toutes ces applications pouvant améliorer la prédiction de la structure 3D des protéines (Fiser *et al*., 2000; Shi *et al*., 2001).

I.5. Prédiction de structure tertiaire

I.5.1. INTRODUCTION

Certaines caractéristiques fonctionnelles des protéines ne peuvent pas être détectées par un simple examen de la séquence ou par une étude d'un alignement avec des homologues, mais nécessite une étude de sa structure 3D. En effet, bien que l'on observe une étroite relation entre la similarité des séquences et la conservation de leur structure (Chothia and Lesk, 1986), la fonction n'est pas toujours conservée (Devos and Valencia, 2000). On peut ainsi découvrir qu'un site actif est très conservé au niveau structural, alors que certains des résidus qui le constituent, et qui sont donc spatialement proches, sont très éloignés dans la séquence. Par exemple, il est possible qu'un site actif soit très conservé au niveau structural mais que certains résidus qui y sont très proches spatialement soient très éloignés dans la séquence d'intérêt, du fait de l'insertion d'une boucle ou du rapprochement de régions distantes par le repliement de la structure. Dans d'autres cas, la connaissance de la structure protéique permet d'infirmer certaines hypothèses émises 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 masqués par un élément de structure que l'on ne pouvait pas détecter par similarité.

D'autre part, la connaissance de la structure 3D d'une protéine permet de planifier rapidement des expériences de mutagenèse dirigée en ciblant les résidus directement impliqués dans le site actif ou dans l'interaction avec le substrat. Ces mutations permettent de montrer si un ou plusieurs résidus sont impliqués dans la fonction. Elles permettent également de pratiquer l'ingénierie de peptides ou de protéines, en modifiant leur solubilité, leur stabilité ou leur activité.

Enfin, si la structure 3D est très précise, son étude détaillée peut aider à la conception de molécules (pharmacophores, médicaments, …) se fixant spécifiquement à la protéine cible. Elle permet également d'étudier les interactions protéines-ligand, protéines-protéines, protéines-ADN et de mieux comprendre la fonction d'une protéine, 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 *et al*., 2000).

La connaissance de la structure 3D d'une protéine se révèle donc être d'une importance capitale. Cependant, le nombre de structures protéiques connues est relativement faible par rapport au nombre de séquences disponibles dans les banques de données de séquences (*nr* du NCBI, par exemple), et ce nombre de séquences est en croissance exponentielle (un doublement environ tous les 14 mois (Baxevanis, 2003)). En effet, les principales techniques expérimentales de détermination des structures de protéines, à savoir la diffraction des rayons X et la RMN, ne peuvent être appliquées à n'importe quelle protéine:

- La diffraction des rayons X, bien qu'étant la technique la plus précise, exige la préparation de monocristaux de taille supérieure ou égale à 0,2 mm 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.
- La RMN ne se prête qu'à des protéines dont le poids moléculaire est inférieur à environ 30 kDa.

Pour illustrer ces limitations, le Tableau 3 présente la fraction des protéines du génome de *Helicobacter pylori* qui ont franchi chacune des étapes de la détermination de la structure 3D par diffraction des rayons X ou par RMN. On y remarque que la structure 3D a pu être établie pour 11% des protéines initialement sélectionnées.

Tableau 3: Fraction de protéines du génome de *Helicobacter pylori* ayant franchi chacune des étapes de la détermination expérimentale de leur structure 3D par diffraction de rayons X ou par RMN (origine: http://s2f.umbi.umd.edu).

Etape	Nombre protéines	Fraction
Protéines sélectionnées	335	100%
Protéines clonées	308	92%
Protéines surexprimées	172	51%
Protéines purifiées	113	34%
Protéines cristallisées	45	13%
Données de diffraction de rayons X	34	10%
Données de RMN	6	2%
Chaînes tracées	38	11%
Modèles raffinés	36	11%
Structures déposées	35	10%

C'est pour faire face à cette situation que différentes méthodes de prédiction de structures protéiques ont été développées. Ces méthodes ont, en effet, l'avantage d'être beaucoup plus rapides que la détermination de structures par diffraction des rayons X ou par RMN.

Dans les sections qui suivent, nous commencerons par présenter la problématique du repliement des protéines. Ensuite, nous détaillerons la prédiction de structure 3D par la méthode de modélisation par homologie, qui est le point central de cette thèse. Enfin, les techniques de reconnaissance de *fold* et de modélisation *de novo* seront rapidement abordées.

I.5.2. PROBLÉMATIQUE DU REPLIEMENT DES PROTÉINES

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 inscrite dans sa structure primaire.

Ce principe est à l'origine des méthodes dites *ab initio* ou *de novo* (Sternberg *et al*., 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 (*folding*) des protéines sont loin d'être complètement élucidées. En effet, l'hypothèse de Anfinsen n'a été vérifiée que pour une faible proportion de protéines. Par exemple, les protéines multimériques et multidomaines ne peuvent pas se replier correctement après dénaturation: des régions hydrophobes ont souvent tendance à s'agréger, ce qui provoque la formation de structures ne correspondant plus à la structure native.

De plus, le repliement de la majorité des protéines qui intéressent les biologistes moléculaires et cellulaires fait intervenir, outre la séquence, d'autres facteurs:

- Les protéines chaperonnes (qui font partie des protéines impliquées dans la protection contre les chocs thermiques: *Heat Shock Proteins* ou HSP) jouent un rôle important aussi bien chez les eucaryotes que chez les procaryotes. A l'heure actuelle, la proportion de protéines qui se replient avec l'aide de chaperonnes n'est pas déterminée mais elle semble être importante.
- Les cofacteurs (tel l'hème dans les cytochromes) ou les ions (comme le zinc dans les motifs "en doigt de zinc" liant l'ADN, présents dans certains facteurs de transcription) influencent 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 force ionique, le solvant, la température…

Ces facteurs extérieurs intervenant dans le repliement des protéines montrent que le postulat d'Anfinsen n'est pas universellement applicable. Dès lors, il est souvent impossible de prédire la structure d'une protéine uniquement à partir de sa séquence. Des méthodes de prédiction fondées sur d'autres postulats ont dès lors été imaginées: la modélisation par homologie, la reconnaissance de *fold* et des méthodes de prédictions *de novo*.

I.5.3. MODÉLISATION PAR HOMOLOGIE

I.5.3.1. Principe et utilité

Le principe de la modélisation par homologie est de construire un modèle tridimensionnel d'une protéine que l'on appellera protéine cible ou séquence cible (en anglais ***target*** ou ***query***) en se basant sur son alignement avec une ou plusieurs protéines de structure connue, que l'on nommera *template(s)* (Blundell *et al*., 1987; Greer, 1990; Bajorath *et al*., 1993; Johnson *et al*., 1994; Sali, 1995; Sanchez and Sali, 1997; Fiser *et al*., 2000). Cette méthode de modélisation est la première à avoir été utilisée, dès 1969, pour modéliser l'α-lactalbumine bovine (Browne *et al*., 1969).

La condition *sine qua non* pour l'utilisation d'une telle méthode est le partage d'un pourcentage d'identité 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 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.

Actuellement, plus d'un tiers des séquences protéiques peuvent être reliées à au moins une protéine de structure connue (Huynen *et al*., 1998; Rychlewski *et al*., 1998; Sanchez and Sali, 1998; Jones, 1999). Comme le nombre actuel de séquences protéiques connues est approximativement de 1.500.000, la modélisation par homologie pourrait s'appliquer à plus de 500.000 séquences, soit 20 fois plus que le nombre de structures disponibles aujourd'hui. De plus, il est important de souligner qu'il s'agit, actuellement, de la méthode de modélisation la plus précise (Mosimann *et al*., 1995; Martin *et al*., 1997; Alwyn Jones and Kleywegt, 1999; Tramontano *et al*.,

2001). L'utilité de la modélisation par homologie se renforce avec le temps parce que le nombre de structures déterminées expérimentalement augmente exponentiellement (Holm and Sander, 1996) alors que le nombre de repliements adoptés par les protéines est limité (Chothia, 1992; Zhang, 1997). Cette méthode est donc la plus prometteuse puisqu'on prédit que, dans 10 ans, au moins un exemplaire de la plupart des repliements sera connu, rendant la modélisation par homologie applicable à la plupart des séquences protéiques (Sanchez and Sali, 1997).

I.5.3.2. Les étapes

Les méthodes de modélisation par homologie se composent de quatre étapes essentielles (Eisenhaber *et al*., 1995; Sali, 1995; Sanchez and Sali, 1997). La première consiste à identifier les protéines de structure 3D connue reliées à la séquence cible. Au cours de la seconde étape, on les aligne à la séquence cible et on sélectionne celles qui seront utilisées comme *templates*. La troisième étape est la construction du modèle de la séquence cible sur base de son alignement avec le ou les *template(s)*. Enfin, dans la quatrième et dernière étape, le modèle est évalué en utilisant divers critères. Les différences entre les méthodes de modélisation par homologie existantes résident dans l'implémentation de ces quatre étapes. Dans certaines méthodes, les étapes de sélection des *templates*, alignements et construction du modèle sont réitérées jusqu'à obtention d'un modèle satisfaisant.

I.5.3.3. Recherche de séquences homologues à la séquence cible

I.5.3.3.1. Banques de données

La toute première étape de la modélisation par homologie 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 également un maximum de séquences homologues, de façon à obtenir un alignement le plus fiable possible. La séquence cible peut être comparée à des banques de données de séquences comme PIR (Wu *et al*., 2002), GenBank (Benson *et al*., 2003), TrEMBL/Swiss-Prot (Bairoch and Apweiler, 1999) ou des banques non redondantes comme la banque *nr* du NCBI et/ou à des banques de données de structures comme la Protein Data Bank (PDB) (Abola *et al*., 1987; Berman, 1999; Berman *et al*., 2000), SCOP(Murzin *et al*., 1995; Hubbard *et al*., 1999), DALI (Holm and Sander, 1999) et CATH (Orengo *et al*., 1997; Orengo *et al*., 1999). Actuellement, la probabilité de trouver une séquence de structure connue similaire à une séquence prise au hasard dans un génome varie de 20% à 50% (Huynen *et al*., 1998; Rychlewski *et al*., 1998; Sanchez and Sali, 1998; Jones, 1999).

Le nombre élevé de banques de données disponibles sur l'Internet pose certains problèmes de choix: quelle est la plus adéquate, la plus précise, la mieux mise à jour ou encore dans quel format les données y sont-elles disponibles? Parmi les banques de données protéiques citées plus haut, NRL-3D (la banque de séquences de PDB) a l'avantage d'être directement reliée à l'information structurale mais, réciproquement, 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. Swiss-Prot, quant à elle, fournit d'excellentes annotations mais recense moins de séquences que PIR. La banque *nr* (*non redundant*) est le résultat de la compilation de diverses autres banques de données protéiques. Cependant, les séquences identiques, ou ne différant que pour quelques résidus, ont été retirées. C'est pour cela qu'elle porte le nom de *non redundant*. Cette banque permet ainsi, indirectement, d'exécuter des recherches dans plusieurs autres banques, et de le faire plus rapidement.

I.5.3.3.2. Identification des templates potentiels

On distingue trois catégories principales de programmes de recherche en base de données pour l'identification du *template* de la séquence cible (Fiser *et al*., 2000).

La première catégorie compare de manière indépendante la séquence cible à chaque séquence de la banque de données, en utilisant des méthodes d'alignement pairé (Apostolico and Giancarlo, 1998). La performance de ces méthodes dans la recherche de séquences (Pearson, 1995) et dans l'assignation du *template* a été évaluée de manière exhaustive (Brenner *et al*., 1998). Les programmes les plus populaires de cette catégorie sont FASTA (Pearson and Lipman, 1988; Pearson, 1990) et BLAST2 (Altschul *et al*., 1997).

On classe dans la deuxième catégorie, des méthodes qui utilisent directement ou indirectement des alignements multiples (ou des profils) pour détecter des similarités entre séquence cible et *templates*, pour améliorer la sensibilité de la recherche (Gribskov, 1994; Henikoff and Henikoff, 1994; Krogh *et al*., 1994; Altschul *et al*., 1997; Rychlewski *et al*., 1998). Le programme le plus connu dans cette catégorie est PSI-BLAST (Altschul *et al*., 1997). Une autre approche similaire qui semble plus performante que PSI-BLAST a aussi été décrite: FFAS (Rychlewski *et al*., 1998). Elle commence par trouver toutes les séquences de la banque de séquences qui sont clairement liées à la cible et qui peuvent lui être facilement alignées. L'alignement multiple de ces séquences devient le profil cible. Des profils similaires sont construits pour chaque *template* potentiel. Les *templates* sont alors trouvés en comparant le profil cible avec chaque profil de *template* potentiel en utilisant une méthode de programmation dynamique locale

basée sur la matrice de scores BLOSUM62 (Henikoff and Henikoff, 1994). D'autres techniques utilisant des processus de recherche par séquences intermédiaires (ISS ou *Intermediate Sequence Search*) (Rychlewski *et al.*, 2000) ou utilisant des HMM (Karplus *et al.*, 1997) ont également été développées. Ces techniques plus sensibles sont très intéressantes pour trouver des relations structurales significatives quand le pourcentage d'identité entre la séquence cible et les *templates* potentiels est inférieur à 25%. En fait, cette catégorie de méthodes qui se basent sur les alignements multiples de séquences apparaît actuellement comme étant l'approche complètement automatique la plus sensible pour détecter des relations séquence-structure lointaines (Altschul *et al.*, 1997; Huynen *et al.*, 1998; Jaroszewski *et al.*, 1998; Zhang *et al.*, 1998).

Dans la troisième catégorie sont classées les méthodes qui réalisent une comparaison pairée entre une séquence et une structure de protéine en utilisant de l'information structurale (un potentiel statistique, voir section I.5.3.8). La séquence cible est comparée à une librairie de profils 3D ou enfilée (*threaded*) au travers d'une librairie de *folds*, le *fold* de la séquence cible étant celui qui a l'énergie la plus basse. Ces méthodes sont aussi appelées reconnaissance de *fold* (*fold recognition*), *threading* ou appariement de *template* 3D (3D *template matching*) (voir section I.5.4). Elles sont décrites dans (Jones, 1997; Smith *et al.*, 1997; Torda, 1997) et évaluées dans (Levitt, 1997). Ces méthodes sont spécialement utiles quand il n'est pas possible de construire un profil pour la séquence cible parce qu'il n'existe pas assez de séquences similaires connues pour la séquence cible ou pour les *templates*. On comprend donc que si la modélisation bénéficie largement de l'augmentation du nombre de structures connues, elle bénéficie aussi de l'augmentation des banques de données de séquences protéiques (Larson *et al.*, 2003).

I.5.3.4. Alignement de séquences

Quand le ou les *template(s)* sont identifiés, la deuxième étape est de calculer un alignement multiple de la séquence cible avec tous ces *templates* potentiels (Holm and Sander, 1996; Taylor, 1996; Baxevanis, 1998; Briffeuil *et al.*, 1998). Cette étape permet d'aligner les zones similaires entre la séquence cible et le ou les *template(s)* dans le but de prédire les régions structuralement conservées (pSCR: *predicted Structurally Conserved Regions*) (Vinals *et al.*, 1995; de Fays *et al.*, 1999).

En général, l'alignement fourni par un programme de recherche en base de données n'est pas optimal (Venclovas *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(s) *template(s)* sélectionné(s) à la séquence cible.

Si le pourcentage d'identité entre les deux séquences dépasse approximativement 50%, un alignement pairé des 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 s'accentuent avec la baisse de la similarité (Johnson and Overington, 1993). L'alignement multiple se base sur le fait qu'une similarité de séquences est plus 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 lors d'un alignement pairé (Venclovas *et al*., 1999).

La qualité de l'alignement influence fortement la fiabilité du modèle (Vinals *et al*., 1995) et est actuellement considérée comme le principal facteur déterminant la qualité du modèle 3D final (Tramontano *et al*., 2001). Un effort particulier doit donc être entrepris pour obtenir les meilleurs alignements possibles car les étapes ultérieures de la modélisation par homologie ne pourront pas corriger des erreurs d'alignement (Fiser *et al*., 2000).

La plupart des erreurs sont dues à la position incorrecte d'insertions et de délétions qui ont pour effet de décaler l'alignement. Il est possible, surtout lorsque le pourcentage d'identité entre la séquence cible et le(s) *template(s)* 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. (Briffeuil *et al*., 1998 ; de Fays *et al*., 1999; Thompson *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* sont des informations précieuses qui peuvent être prises en compte pour l'optimisation de l'alignement final (Kabsch and Sander, 1983; Jennings *et al*., 2001; Shi *et al*., 2001; Lambert *et al*., 2003).

I.5.3.5. Construction du modèle tridimensionnel de la protéine cible

La troisième étape de la modélisation par homologie se fait en attribuant, aux régions conservées entre la séquence cible et le(s) *template(s)*, la structure des régions équivalentes dans le(s) *template(s)*. Cette étape fournit un ensemble de coordonnées spatiales pour la séquence cible, généralement des C_α dans les zones conservées.

A ce stade, on obtient donc un modèle partiel du squelette de la protéine, car les régions variables correspondant le plus souvent aux *loops* ne sont pas encore modélisées. Trois méthodes ont été proposées (Fiser *et al*., 2000):

1. La modélisation par assemblage de corps rigides (Browne *et al*., 1969; Blundell *et al*., 1987; Greer, 1990) se base sur la séparation de la structure des protéines en régions de cœur conservées, des *loops* variables qui les connectent et des chaînes latérales branchées sur le squelette (*backbone*). Elle se déroule en six phases: (1) superposition des *templates*; (2) calcul des coordonnées moyennes des segments de structure conservés dans les *templates*; (3) assignation des coordonnées de la chaîne principale en superposant les segments de structure aux segments de la séquence cible les plus similaires (du point de vue de la séquence); (4) construction des *loops* en scannant la banque PDB pour trouver des régions variables qui s'adaptent aux régions modélisées; (5) modélisation de chaînes latérales en se basant sur une banque de rotamères; et (6) minimisation d'énergie ou dynamique moléculaire pour ajuster la stéréochimie de la structure prédite. Un exemple bien connu de ce type d'approche est le programme COMPOSER (Sutcliffe *et al*., 1987).

2. La modélisation par appariement de segments ou reconstruction des coordonnées se base sur le fait que la plupart des segments de 6 acides aminés (hexapeptides) peuvent être regroupés en 100 classes structurales différentes (Unger *et al*., 1989). Ces segments peuvent être obtenus soit à partir de toutes les structures existantes (Claessens *et al*., 1989; Holm and Sander, 1991), soit à partir d'une recherche conformationnelle contrainte par une fonction d'énergie (Bassolino-Klimas and Bruccoleri, 1992; van Gelder *et al*., 1994). Les modèles peuvent donc être construits en utilisant un sous-ensemble des coordonnées atomiques des structures du (des) *template(s)* comme « guides » et en identifiant, puis assemblant les segments courts correspondant à ces guides. Un exemple est le programme SegMod (Levitt, 1983; Levitt, 1992).

3. La modélisation par satisfaction de contraintes spatiales (Havel and Snow, 1991) commence par générer toute une série de contraintes sur la structure de la séquence cible en utilisant son alignement au(x) *template(s)*. Ces contraintes sont généralement obtenues en supposant que les distances entre les résidus alignés dans le(s) *template(s)* et la séquence cible sont similaires. A ces contraintes, sont également ajoutées des contraintes stéréochimiques sur les longueurs des liaisons, les angles de liaison, les contacts inter-atomiques, … qui sont obtenues à partir de champs de forces moléculaires. Le modèle est alors calculé en minimisant les violations de ces contraintes. Les programmes les plus connus dans cette catégorie sont MODELLER (Sali and Blundell, 1990; Sali and Blundell, 1993; Sali and Overington, 1994; Sali *et al*., 1997; Sanchez and Sali, 1997) et ProMod (Peitsch, 1995).

I.5.3.6. Prédiction des *loops*

Dans l'alignement entre séquence cible et *template(s)*, on observe fréquemment des groupes d'acides aminés de la cible qui n'ont pas d'équivalent dans le(s) *template(s)*. Pour ceux-ci, aucune information structurale provenant des *templates* ne peut donc être utilisée pour modéliser ces acides aminés. Ces régions correspondent généralement à des *loops* positionnés à la surface de la protéine. La précision de la modélisation des *loops* doit être élevée car c'est un facteur majeur influençant l'utilité des modèles 3D pour des applications comme le *docking* (prédiction de l'interaction entre deux protéines ou entre une protéine et un ligand).

Certaines techniques de modélisation des *loops* utilisent des banques de données de structures de *loops* extraits de structures cristallographiques (Jones and Thirup, 1986; Chothia and Lesk, 1987; Bates and Sternberg, 1999). Parmi ces fragments de structure, celui qui est le plus énergétiquement favorable et qui permet de joindre 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.

D'autres méthodes recherchent la meilleure conformation que pourrait adopter le fragment de séquence en utilisant la dynamique moléculaire et la minimisation d'énergie (Fine *et al*., 1986; Moult and James, 1986; Bruccoleri and Karplus, 1987; Dudek and Scheraga, 1990). Certaines méthodes combinent les deux approches précédentes (Chothia *et al*., 1986; Martin *et al*., 1989; van Vlijmen and Karplus, 1997).

Malgré le nombre important de méthodes décrites (Chothia *et al*., 1986; Fine *et al*., 1986; Jones and Thirup, 1986; Moult and James, 1986; Bruccoleri and Karplus, 1987; Chothia and Lesk, 1987; Martin *et al*., 1989; Bruccoleri and Karplus, 1990; Dudek and Scheraga, 1990; Higo *et al*., 1992; Collura *et al*., 1993; Zheng *et al*., 1993; Abagyan and Totrov, 1994; Ring and Cohen, 1994; Zheng *et al*., 1994; Koehl and Delarue, 1995; Rosenbach and Rosenfeld, 1995; van Vlijmen and Karplus, 1997; Samudrala and Moult, 1998), il n'est pas encore possible de modéliser correctement des *loops* plus grands que 6 à 10 résidus (Mosimann *et al*., 1995; Sanchez and Sali, 1997; van Vlijmen and Karplus, 1997).

I.5.3.7. Positionnement des chaînes latérales

Pour des raisons stériques, les angles de torsion des chaînes latérales prennent des valeurs limitées: les chaînes latérales adoptent des conformations préférentielles appelées rotamères stables (Vasquez, 1996). Des graphes reprennent les distributions d'angles retrouvées pour chacune des chaînes latérales des 20 acides aminés dans les structures

cristallographiques obtenues avec la meilleure résolution. Le positionnement des chaînes latérales se fait en leur attribuant les conformations prises par les chaînes latérales équivalentes dans le ou les *template(s)* et, si ce n'est pas possible, en se basant sur les conformations préférentielles qu'elles peuvent prendre (Sali, 1995; Vasquez, 1996; Bates and Sternberg, 1999; Dunbrack, 1999). Cependant, des états non rotamériques (i.e. non énergétiquement favorables) sont systématiquement observés dans les structures cristallographiques des protéines, généralement à cause des interactions tridimensionnelles avec d'autres chaînes latérales (Schrauber *et al.*, 1993; Lee, 1995; Bower *et al.*, 1997). Dès lors, s'il est basé sur les rotamères stables, le positionnement des chaînes latérales n'est pas optimal (Sali *et al.*, 1995; Huang *et al.*, 1998). De plus, la qualité de ce positionnement dépend fortement de la qualité de l'alignement et de la structure du(des) *template(s)* (Venclovas *et al.*, 1999).

I.5.3.8. Optimisation du modèle

Selon l'hypothèse thermodynamique du repliement protéique, la conformation native correspond au minimum d'énergie libre de Gibbs (*G*). Cette énergie est difficilement calculable puisqu'on ne connaît pas nécessairement les conditions de pression (*P*) et de température (*T*), le volume (*V*) ni surtout l'entropie du système (*S*). Par contre, on peut la relier à l'énergie interne de la protéine (*U*): $G = U - T \cdot S + P \cdot V$. Cependant, le calcul de l'énergie interne de telles molécules est actuellement irréalisable en pratique si on veut utiliser des méthodes très précises (mécanique quantique).

Néanmoins, on peut estimer cette énergie grâce à des fonctions empiriques beaucoup plus simples portant le nom de "champs de forces" (*force fields*) tels que CHARMM (Brooks *et al.*, 1983), AMBER (Cornell *et al.*, 1995), GROMOS (van Gunsteren and Berendsen, 1990) ou CVFF (Dauber-Osguthorpe *et al.*, 1988). Celles-ci décrivent l'ensemble des interactions subies par chaque atome d'une protéine (les interactions de van der Waals et électrostatiques, et les liaisons covalentes). L'énergie interne totale est calculée en sommant l'ensemble des énergies potentielles correspondant à chaque interaction (ou chaque force). La construction des champs de forces extrapole les données d'énergies internes dérivées de systèmes simples aux systèmes macromoléculaires que sont les protéines. Cette extrapolation 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, l'approche des «potentiels statistiques» permet de calculer l'énergie libre de Gibbs (G) et non plus un U, comme dans les champs de forces. Suivant cette approche, les structures connues de

protéines sont prises comme seule source d'information pour extraire les forces et potentiels qui stabilisent les protéines. 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. Dès lors, la loi inverse de Boltzmann permet de retrouver l'énergie libre d'une molécule à partir des fréquences d'occurrence des différents états énergétiques possibles:

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

Où *r* est l'ensemble des coordonnées des atomes ou des acides aminés

G(r) est l'énergie libre de la conformation *r* donnée

f(r) est une fonction des distances entre atomes ou acides aminés dans la conformation *r* donnée

k est la constante de Boltzmann

T est la température absolue en Kelvin

Ceci peut s'appliquer aux protéines: on peut, par exemple, calculer un potentiel statistique à partir des fréquences d'occurrence des distances entre chaque acide aminé et tous les autres acides aminés dans les protéines de structure connue. Ce potentiel est calculé sur base des conformations d'énergie minimum des protéines puisqu'elles sont à leur minimum énergétique à l'état natif d'après l'hypothèse d'Anfinsen. Pour calculer l'énergie libre d'un modèle, il suffit d'additionner les énergies de chaque résidu calculées à partir de ce potentiel. Si cette énergie est positive, la protéine étudiée n'est pas dans une conformation stable. Cela signifie que le modèle n'est probablement pas correct.

Parmi les méthodes d'optimisation du modèle, la minimisation d'énergie a pour but de ramener la fonction énergétique globale à un minimum par rapport aux coordonnées atomiques. 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. La minimisation d'énergie présente l'inconvénient de ne pas pouvoir franchir les barrières énergétiques. Or, une protéine possède un grand nombre de conformations correspondant à des minima locaux d'énergie. Une simple minimisation d'énergie a, dès lors, pour effet d'amener la conformation initiale du modèle à un minimum d'énergie qui ne correspond pas obligatoirement au minimum absolu,

supposé correspondre à la conformation « native ». En conséquence, la conformation calculée ne s'écarte pas beaucoup du modèle initial.

La dynamique moléculaire, pour sa part, simule l'agitation thermique de la protéine et lui permet de franchir les barrières énergétiques en apportant virtuellement de l'énergie cinétique à la protéine. Cette simulation produit toute une série de conformations différentes.

Pratiquement, l'exploration de l'espace conformationnel, en vue de trouver la structure la plus stable, se fait par itérations, avec alternance de dynamique moléculaire et de minimisation d'énergie (celle-ci permettant d'obtenir une conformation correspondant à un minimum énergétique local). Après un certain nombre d'itérations, on se retrouve avec un ensemble de conformations d'énergie minimum dont celle ayant l'énergie la plus basse est considérée comme la structure native de la protéine.

Remarquons que ce protocole ne conduit pas nécessairement à l'obtention d'un modèle 3D meilleur. En effet, bien qu'étant plus énergétiquement favorable, la conformation du modèle « optimisé » peut s'écarter de la structure native (Tramontano *et al*., 2001). De plus, les minimisations et dynamiques moléculaires ne permettent pas de corriger certaines erreurs du modèle (voir section I.5.3.10). Par conséquent, ces techniques peuvent aussi être source d'erreurs (Tramontano *et al*., 2001).

I.5.3.9. Evaluation du modèle *a priori*

Avant de pouvoir utiliser un modèle 3D dans le cadre de travaux théoriques ou expérimentaux, il faut estimer sa précision, globalement et localement. Les modèles 3D sont généralement évalués en se référant aux préférences géométriques des acides aminés ou des atomes qui sont dérivées des structures protéiques connues. Des relations empiriques entre les erreurs dans les modèles et le pourcentage d'identité entre la séquence cible et le *template* peuvent aussi être utilisées. Il existe plusieurs programmes ou serveurs permettant d'effectuer cette évaluation (Laskowski *et al*., 1998) (Wilson *et al*., 1998). Une bonne stratégie est d'évaluer les modèles en utilisant plusieurs méthodes et en identifiant le consensus entre elles.

Il est intéressant d'avoir une approche hiérarchique de l'évaluation d'un modèle (Sanchez and Sali, 1998). Il faut d'abord vérifier si le modèle a un *fold* correct. Ce sera le cas si le *template* correct a été utilisé et si le *template* est aligné au moins approximativement avec la séquence cible. Une fois que le *fold* du modèle est confirmé, une évaluation plus détaillée de la qualité globale du modèle peut être effectuée en se basant sur la similarité entre la séquence cible et le *template*. Finalement, une variété de profils d'erreurs peut être construite pour quantifier les erreurs vraisemblables dans différentes régions du modèle.

Une condition de base pour avoir un bon modèle est d'avoir une bonne stéréochimie. Les programmes les plus utilisés pour l'évaluation de la stéréochimie sont WHAT-CHECK (Vriend and Sander, 1993), PROCHECK (Laskowski *et al*., 1993; Laskowski *et al*., 1996), AQUA (Laskowski *et al*., 1996) et SQUID (Oldfield, 1992). Les caractéristiques vérifiées incluent les angles de torsion et de valence, la chiralité, la longueur des liaisons, et les conflits stériques entre paires d'atomes non liés. En plus d'une bonne stéréochimie, l'énergie potentielle totale, calculée sur base d'un champ de forces comme CHARMM22 (Brooks *et al*., 1983), doit être assez basse.

Les distributions d'un grand nombre de caractéristiques spatiales ont été compilées à partir de structures de protéines de haute résolution et une large déviation des valeurs les plus vraisemblables est interprétée comme un bon indicateur d'erreur dans le modèle. Ces caractéristiques incluent la formation d'un cœur hydrophobe (Bryant and Amzel, 1987), les accessibilités au solvant (Baumann *et al*., 1989; Chiche *et al*., 1990; Vila *et al*., 1991; Holm and Sander, 1992; Koehl and Delarue, 1994) la distribution spatiale des groupes chargés (Bryant and Lawrence, 1991), la distribution des distances inter-atomiques (Colovos and Yeates, 1993), les volumes atomiques et les liaisons par pont hydrogène de la chaîne principale (Laskowski *et al*., 1993).

Cependant, même si ces contraintes sont respectées, le modèle n'est pas pour autant toujours correct (Novotny *et al*., 1984; Novotny *et al*., 1988). En effet, il est 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'avèrent capitales pour la fonction d'une protéine.

Certains programmes d'évaluation des modèles 3D prenant implicitement en compte beaucoup de critères cités précédemment, utilisent des profils 3D et des potentiels statistiques (décrits plus haut) (Sippl, 1990; Luthy *et al*., 1992). Ces méthodes évaluent l'environnent de chaque résidu dans le modèle par rapport à l'environnement moyen, trouvé dans des structures de haute résolution déterminées par rayons X. Les programmes utilisant cette approche sont Verify 3D (Luthy *et al*., 1992), PROSA (Sippl, 1993), HARMONY (Topham *et al*., 1994) et ANOLEA (Melo and Feytmans, 1998). Ces fonctions d'énergie sont, en général, conçues pour fonctionner à un certain niveau de détail et ne sont pas appropriées pour juger le modèle à un niveau plus fin (atomique) ou même plus grossier (au niveau des acides aminés) (Park *et al*., 1997).

Outre leur rôle d'évaluation, ces méthodes peuvent aider les procédures actuelles de modélisation en incorporant les critères de qualité dans une fonction qu'il suffira d'optimiser pour dériver le meilleur modèle. Ces méthodes permettent de détecter des erreurs qui font que le modèle ou

certaines régions du modèle ne ressemblent pas à une structure 3D déjà observée. Souvent, elles ne permettent pas de détecter les erreurs de modélisation décrites dans la section suivante.

I.5.3.10. Erreurs de modélisation par homologie

Les erreurs dans la modélisation par homologie peuvent être divisées en 5 catégories (Sanchez and Sali, 1997): (1) erreurs dans l'orientation des chaînes latérales, (2) distorsions ou déplacements d'une région qui est correctement alignée avec la structure du *template*, (3) distorsions ou déplacements d'une région qui n'a pas de segment équivalent dans la structure du *template*, (4) distorsions ou déplacements d'une région qui n'est pas correctement alignée avec la structure du *template* et (5) une structure incorrecte provenant d'un mauvais choix du *template*.

Les erreurs (3) à (5) sont peu fréquentes quand la séquence cible et les *templates* partagent plus de 40% d'identité. Dans cette zone de similarité, l'alignement est relativement simple à construire, il n'y a pas beaucoup de *gaps* et les différences structurales entre les protéines sont généralement limitées aux boucles et aux chaînes latérales.

Quand le pourcentage d'identité est situé entre 30 et 40%, les différences structurales deviennent plus importantes et les *gaps* dans l'alignement sont plus fréquents et plus longs. Il en résulte que les parties communes sont relativement bien modélisées mais que les autres résidus sont modélisés avec des erreurs importantes car les distorsions de structures restent difficiles à prédire, et qu'il est encore impossible de corriger automatiquement les erreurs d'alignement.

En dessous de 40% d'identité, les erreurs d'alignement et les insertions dans la séquence cible deviennent les problèmes majeurs. Des insertions de plus de 6 à 10 résidus ne peuvent être modélisées précisément alors que les boucles plus courtes sont fréquemment modélisées avec succès (van Vlijmen and Karplus, 1997; Samudrala and Moult, 1998).

Quand le pourcentage d'identité descend en dessous de 30%, le problème principal commence à devenir l'identification d'un *template* et l'alignement global entre celui-ci et la séquence cible. En général, on peut estimer qu'environ 20% des résidus seront de toute façon mal alignés et donc mal modélisés. Il en résulte un RMSD supérieur à 3Å entre le modèle de la séquence cible et sa structure réelle (Johnson and Overington, 1993). Ceci est un problème sérieux pour la modélisation par homologie puisqu'il apparaît que lorsqu'on aligne des protéines homologues, dans plus de 50% des cas, on observe un pourcentage d'identité inférieur à 30% (Rost, 1997; Sanchez and Sali, 1998).

Il faut néanmoins garder à l'esprit que la précision des modèles obtenus est, dans la plupart des cas, comparable à la précision des méthodes expérimentales. Ainsi, si on observe un RMSD de 1 Å entre un modèle d'une protéine et sa structure déterminée par DRX, on peut estimer que ce modèle a une résolution comparable à celle des meilleures structures déterminées par DRX. De même, si un RMSD de 2,5 Å est observé entre une structure expérimentale et un modèle, on peut penser que sa résolution est similaire à celle des structures déterminées par DRX à faible résolution (Ohlendorf, 1994). Ce RMSD de 2,5 Å est encore semblable à celui qui est mesuré entre une structure déterminée par DRX et par RMN.

I.5.3.11. Qualité et utilité d'un modèle prédit par homologie

La modélisation par homologie permet d'obtenir des informations du plus haut intérêt sur la protéine modélisée. Par exemple, les modèles 3D peuvent être intéressants pour concevoir des mutants permettant de tester des hypothèses à propos de la fonction de la protéine (Boissel *et al.*, 1993; Wu *et al*., 1999), pour identifier les sites actifs et/ou de liaison (Sheng *et al.*, 1996), pour chercher à améliorer les ligands pour un site de liaison donné (Ring *et al*., 1993), pour modéliser la spécificité d'un substrat (Xu *et al*., 1996), pour prédire des épitopes d'antigènes (Sali *et al*., 1993), pour simuler des interactions protéine-protéine (Vakser, 1997), pour faciliter le remplacement moléculaire dans la détermination des structures par rayons X (Howell *et al*., 1992), pour raffiner des modèles basés sur des contraintes en RMN (Modi *et al*., 1996), pour tester et améliorer l'alignement séquence-structure (Wolf *et al*., 1998) et pour confirmer des relations structurales lointaines (Guenther *et al*., 1997).

Heureusement, la qualité d'un modèle ne doit pas être absolument parfaite pour qu'il soit utile en biologie. Cette utilité sera tout de même fonction de sa qualité. Celle-ci dépend du pourcentage d'identité partagé entre la séquence cible et son (ses) *template(s)* (Sanchez and Sali, 1997) et de la qualité du *template*.

Du point de vue de leur utilité, on peut classer les modèles en trois catégories (Peitsch, 1996):

- Les modèles basés sur des alignements en partie incorrects entre la séquence cible et le(s) *template(s)*. Lorsque les erreurs ne se localisent pas dans les régions bien conservées telles que les sites actifs des enzymes, de tels modèles restent toutefois utiles pour donner une idée de la structure 3D globale de la protéine et permettent de localiser la mutation d'un acide aminé et, parfois, de proposer des hypothèses pour expliquer certains phénotypes observés.

- 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 la fixation de ligands.
- Les modèles de protéines partageant un fort taux d'identité (>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 qualité peuvent apporter des réponses à des problèmes biologiques. En effet, la plupart des caractéristiques fonctionnelles peuvent souvent être suggérées à partir du seul *fold* du modèle (Tramontano, 1998; Fiser *et al.*, 2000).

I.5.3.12. Evaluation des performances des méthodes de modélisation par homologie

Les CASP (*Critical Assessment of techniques for Structure Prediction*) sont des congrès bisannuels qui existent depuis 1994. Leur but est d'évaluer les méthodes de prédiction de structures de protéines. Avant la session, des séquences dont la structure 3D est connue mais pas encore publiée, sont soumises à la communauté scientifique. Les participants, qui ne connaissent alors pas la structure réelle de ces protéines, sont tenus de les modéliser et de soumettre leurs prédictions à la direction du CASP. Celle-ci compare les modèles proposés aux structures réelles en utilisant divers critères qui ont varié au cours du temps, mais qui sont actuellement: le RMSD (voir définition page 30), le score LGA_Q, le score GDT_TS, le score AL0, le score AL4 et le score AL4+. Ces critères sont définis ci-dessous.

Lors du congrès proprement dit, les structures qui devaient être prédites sont révélées, et des classements des participants à la compétition sont présentés. Parmi ces classements, citons celui des méthodes de modélisation par homologie. L'intérêt de tels congrès est de fournir une évaluation indépendante des diverses méthodes de modélisation, de faire apparaître les diverses nouveautés dans le domaine et de définir les points faibles qu'il faudra améliorer pour le futur.

En 1998, une nouvelle compétition bisannuelle a été créée et liée au CASP: le CAFASP (***C**ritical **A**ssessment of **F**ully **A**utomated **S**tructure **P**rediction*), qui vise à comparer les performances des différents serveurs automatiques de modélisation. Ensuite, la multiplication des serveurs

automatiques a conduit à la création de systèmes d'évaluation continue comme par exemple le font EVA (Eyrich *et al*., 2001), LiveBench (Bujnicki *et al*., 2001; Bujnicki *et al*., 2001) ou, plus récemment, PDB-CAFASP.

I.5.3.12.1. GDT (Global Distance Test)

Le GDT calcule le plus grand ensemble de résidus qu'il est possible de superposer entre un modèle 3D et la structure réelle de la protéine de telle manière que la distance entre les résidus de chaque paire soit inférieure à un certain seuil de distance, exprimé en Angström.

Le GDT_P-*n* estime le pourcentage de résidus pour un seuil de distance de *n* Å.

Le GDT_TS (GDT *Total Score*) est obtenu par la formule suivante:

GDT_TS = (GDT_P-1 + GDT_P-2 + GDT_P-4 + GDT_P-8)/4.0

On peut considérer qu'il représente le pourcentage de la protéine qui est correctement modélisé.

I.5.3.12.2. LGA_Q

Le score LGA_Q est calculé par la formule suivante:

LGA_Q = 0.1*N/(0.1+RMSD)

où N est le nombre total de résidus superposés en dessous d'un seuil de distance de 5 Å

RMSD est le RMSD calculé sur les N résidus superposés sous le seuil de 5 Å

Ce score mesure la qualité de la modélisation des résidus les mieux modélisés. Il est très sensible aux variations de RMSD, surtout pour de faibles valeurs de celui-ci.

I.5.3.12.3. Mesures de la qualité de l'alignement (AL0, AL4 et AL4+)

Les mesures de la qualité de l'alignement sont basées sur la superposition optimale du modèle proposé pour la SI et de sa structure observée expérimentalement, quand une telle superposition a pu être identifiée par le programme LGA (Zemla, 2000). Pour chaque résidu du modèle, le résidu le plus proche dans la structure de la séquence cible est identifié.

AL0 est le nombre de résidus du modèle pour lesquels le résidu le plus proche dans la structure de la séquence cible (1) est le résidu correct et (2) est à une distance inférieure à 3.8 Å. AL0 représente donc le nombre de

résidus correctement alignés. Il peut également être exprimé en pourcentage de la longueur de la séquence cible.

AL4 est le nombre de résidus du modèle pour lesquels le résidu le plus proche dans la structure de la séquence cible (1) est distant de maximum 4 résidus du résidu correct dans la séquence cible et (2) est à une distance inférieure à 3.8 Å. AL4 représente le nombre de résidus alignés à quatre résidus près du résidu correct. Il peut également être exprimé en pourcentage de la longueur de la séquence cible.

AL4+ est le nombre de résidus du modèle pour lesquelles le résidu correct dans la structure de la séquence cible est à une distance inférieure à 3.8 Å. Il peut également être exprimé en pourcentage de la longueur de la séquence cible.

I.5.4. RECONNAISSANCE DE *FOLD*

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* (voir section I.1.3) très similaires sans pour autant avoir une similarité de séquence ou de fonction. De plus, on pense que le nombre de *folds* serait limité, en particulier par les contraintes physico-chimiques, et pourrait être compris entre 1000 et 5000 (Chothia and Murzin, 1993; Holm and Sander, 1996). Dès lors, prédire la structure d'une protéine revient à se demander si la séquence considérée n'adopterait pas un des *folds* déjà connus.

On distingue deux types de méthodes de reconnaissance de *fold*:

- le ***threading*** se base sur des considérations énergétiques (via l'utilisation de potentiels statistiques, voir section I.5.3.8). La séquence d'intérêt est alignée à chacune des structures d'une librairie de *folds* et le programme recherche l'alignement séquence-structure qui est le plus énergétiquement favorable. Cette méthode est utilisée par des programmes tels ProFIT (Sippl, 1993), THREADER (Jones *et al.*, 1992) ou PROSPECT (Xu and Xu, 2000).
- le ***pseudo-threading*** quant à lui, décrit chaque *fold* connu sous forme d'une succession (un "profil") de propriétés associées à chaque résidu dans la structure: la structure secondaire locale, l'accessibilité au solvant et le degré de polarité des atomes. En effet, ces propriétés sont généralement plus conservées que les résidus eux-mêmes. Les méthodes qualifiées de "*pseudo-threading*" vont donc tenter d'aligner ces profils à

la séquence à modéliser: le meilleur alignement devrait ainsi permettre de prédire la structure 3D de la séquence d'intérêt.

Notons que les méthodes de reconnaissance de *fold* ne sont pas très sensibles ni sélectives: environ 50% des *folds* sont assignés correctement. Toutefois, ce pourcentage n'est valable que si l'on ne tient compte que de la similarité jugée la plus significative par le programme utilisé. 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 (Ginalski *et al*., 2003)
- examiner les 10 premiers *folds* de chaque programme.
- vérifier si le *fold* choisi est plausible en collectant un maximum d'informations structurales et expérimentales sur la famille de séquences à laquelle il correspond et en s'assurant qu'il concorde avec les prédictions de structure secondaire.

I.5.5. MÉTHODES DE PRÉDICTION *DE NOVO*

La modélisation par homologie et le *fold recognition* sont limités par le besoin absolu d'une structure *template*. Pour près de la moitié des protéines, un *template* correspondant ne peut être détecté ou n'est pas encore connu (Sanchez and Sali, 1998; Jones, 1999). Dans ces cas, les méthodes de prédiction *de novo* qui tentent de prédire la structure 3D à partir de la seule séquence (Friesner and Gunn, 1996; Jones, 1997; Levitt *et al*., 1999), sont la seule alternative.

Ces méthodes de prédiction reposent sur les hypothèses thermodynamiques du repliement des protéines (Anfinsen, 1973). Ces hypothèses suggèrent que la structure native de la protéine corresponde à son état d'énergie minimale globale. Suivant cette idée, la plupart des méthodes de prédiction *de novo* peuvent être décrites comme des méthodes d'optimisation qui recherchent le minimum d'énergie global sur une hyper-surface d'énergie de la protéine. Alors que le repliement de petites protéines a été simulé au niveau atomique (Boczko and Brooks, 1995; Duan and Kollman, 1998), il est nécessaire de simplifier la représentation des protéines en utilisant, par exemple, un ou plusieurs centres par résidu (Levitt, 1976) ou en représentant la protéine sous forme de treillis en trois dimensions (Skolnick and Kolinski, 1990). Certaines méthodes partent d'un modèle simplifié pour arriver finalement à un modèle atomique détaillé (Hirst *et al*., 1996).

Les fonctions d'énergie pour les simulations de repliement incluent des champs de forces (voir section I.5.3.8) de programmes de mécanique

moléculaire comme CHARMM (Brooks *et al*., 1983; Roterman *et al*., 1989), AMBER (Cornell *et al*., 1995) et ECEPP (Nemethy *et al*., 1992), les derniers potentiels statistiques (voir section I.5.3.8) dérivés d'un grand nombre de structures de protéines (Vajda *et al*., 1997) et des potentiels basés sur des caractéristiques chimiques (Callaway, 1994; Hinds and Levitt, 1994; Huang *et al*., 1995; Yue and Dill, 1996). Certaines méthodes incorporent aussi des contraintes spatiales obtenues à partir d'alignements multiples de séquences et d'autres considérations pour réduire la taille de l'espace conformationnel à explorer (Aszodi *et al*., 1995; Mumenthaler and Braun, 1995; Sun *et al*., 1995; Ortiz *et al*., 1998; Standley *et al*., 1998).

Plusieurs méthodes d'optimisation ont été appliquées au problème du repliement des protéines (Vasquez *et al*., 1994; Berne and Straub, 1997). On trouve parmi ces méthodes des simulations de dynamique moléculaire (Levitt, 1983; Wilson and Doniach, 1989), l'échantillonnage Monte Carlo (Covell, 1992; Monge *et al*., 1994; Ortiz *et al*., 1998), des méthodes d'équation de diffusion (Kostrowicki and Scheraga, 1992) ou d'algorithmes génétiques (Sun, 1993; Dandekar and Argos, 1994; Cui *et al*., 1998).

Une approche qui a connu récemment un véritable succès assemble le modèle de la protéine à partir de blocs de construction relativement courts (Vasquez and Scheraga, 1988; Simons *et al*., 1997; Bystroff and Baker, 1998). Ces blocs de construction sont des structures tridimensionnelles qui sont associées à de petits peptides de la séquence à modéliser, par une recherche par similarité de séquences dans une banque de données de peptides de structure connue (cette banque de données ayant été élaborée à partir de la PDB). Le modèle final est alors assemblé à partir de toutes ces pièces par l'optimisation d'une fonction statistique d'énergie par une méthode de type *Monte Carlo* (Simons *et al*., 1997).

Même si les méthodes de prédiction *de novo* doivent être encore grandement améliorées, il y a un intérêt à les combiner aux méthodes de modélisation par homologie. En effet, la modélisation des *loops* peut être vue comme un problème de repliement à petite échelle. Il semble donc que ces méthodes puissent aider à corriger certaines limitations de la modélisation par homologie. Les bonnes performances du programme ROBETTA (Simons *et al*., 1997) aux derniers CASP en sont une illustration (Chivian *et al*., 2003).

I.6. Brucella

I.6.1. HISTORIQUE

En 1860, le premier cas d'une maladie provoquant de fortes fièvres atypiques fut décrit chez des soldats anglais en fonction sur l'île de Malte. En 1887, David Bruce isola l'agent responsable de cette maladie à partir de la rate d'un patient décédé et l'appela *Micrococcus melitensis*. Par la suite, cette maladie déjà connue sous le nom de « fièvre de Malte » ou « fièvre ondulante » fut également appelée « brucellose ».

En parallèle, aux Etats-Unis, une bactérie responsable d'avortement chez les bovins fut décrite par Bernhard Bang et nommée *Bacillus abortus*. En 1918, les deux espèces ont été regroupées sous le nom générique de *Brucella*, en l'honneur de David Bruce.

I.6.2. GÉNÉRALITÉS

Les bactéries du genre *Brucella* sont des coccobacilles d'environ 0.5 µm sur 1 µm, intracellulaires facultatifs à Gram négatif n'ayant ni capsule, ni forme de résistance décrite. Ces bactéries sont pathogènes pour l'homme et l'animal (Corbel, 1997; Corbel, 1997; Boschiroli *et al*., 2001). Le genre *Brucella* appartient à la subdivision α-2 des protéobactéries et est composé de huit espèces en fonction de l'hôte parasité: *Brucella melitensis* pour les ovins et les caprins, *Brucella abortus* pour les bovins, *Brucella suis* pour les porcs, *Brucella canis* pour les chiens, *Brucella ovis* pour les ovins, *Brucella cetaceaes* pour les cétacés, *Brucella pinnipediae* pour les pinnipèdes et *Brucella neotomae* pour les rongeurs. Parmi ces sept espèces, trois sont virulentes chez l'homme, il s'agit de *Brucella melitensis, Brucella abortus* et *Brucella suis*. A coté de ces animaux d'élevage, on a aussi mis en évidence la présence de *Brucella* dans de nombreuses espèces sauvages comme les bisons, les rennes, les ours, les phoques, les dauphins et les baleines (Cloeckaert *et al*., 2001; Rhyan *et al*., 2001; Tryland *et al*., 2001).

Chez l'animal, la brucellose touche principalement les organes reproducteurs, ce qui peut provoquer la stérilité chez le mâle et l'avortement chez la femelle gestante (Samartino and Enright, 1993). Cette maladie est une des zoonoses les plus communes, et son impact économique est très important. L'argent investi dans le contrôle de la brucellose aux Etats-Unis correspond à 150 millions de dollars par an, uniquement pour le traitement des bêtes d'élevage. En Amérique latine, les pertes animales sont estimées à 600 millions de dollars par an (Boschiroli *et al*., 2001). En France et en

Belgique, la brucellose a été presque complètement éradiquée grâce à un programme sanitaire mené depuis une dizaine d'années.

Les principales causes de contamination chez l'homme sont l'ingestion de produits d'animaux contaminés (lait, fromage non pasteurisé) ou la proximité homme-animal (bergers, fermiers, vétérinaires,…). Dans les pays en voie de développement où la maladie n'est pas diagnostiquée ou traitée à temps, on peut assister à des complications pouvant entraîner la mort des patients. Bien que difficilement diagnosticable, car les symptômes sont variés et non spécifiques, la brucellose se traite bien par antibiothérapie (Ariza *et al*., 1985). Toutefois, les individus susceptibles de contracter la brucellose (fermiers, scientifiques,...) sont prévenus des différents symptômes, ce qui permet un diagnostic plus rapide.

Il n'existe aucun vaccin efficace pour l'homme, mais deux vaccins sont couramment utilisés chez l'animal. Ces vaccins consistent en des souches vivantes atténuées *Brucella abortus* B19 et *Brucella melitensis* Rev1 et présentent trois inconvénients: 1) il est impossible de discriminer un animal infecté d'un animal vacciné en utilisant les tests sérologiques classiques, 2) les vaccins sont toujours infectieux pour l'homme et 3) ces vaccins provoquent l'avortement chez la femelle gestante.

I.6.3. GÉNOME DE BRUCELLA MELITENSIS

Le génome de *Brucella melitensis* 16M fut complètement séquencé en 2001 (DelVecchio *et al*., 2002). Il contient 3294935 paires de bases (bp) réparties sur deux chromosomes circulaires de 2117146 bp et 1177789 bp contenant 3197 pCDS (*predicted CoDing Sequences* ou séquences codantes prédites). En utilisant la suite ERGO (Integrated Genomics Inc., Chicago, USA), une fonction a pu être assignée à 2487 pCDS (78%). Le contenu moyen en G et C des pCDS est de 57%. Le

Tableau 4 récapitule les caractéristiques générales du génome de *Brucella melitensis* au moment du séquençage (DelVecchio *et al*., 2002).

Tableau 4: Caractéristiques générales du génome de *Brucella melitensis* (DelVecchio *et al*., 2002)

Nombre de chromosomes	2	
Séquence ADN totale	3296953 bp	100%
Taille du chromosome I	2118216 bp	64%
Taille du chromosome II	1178737 bp	36%
Séquences ADN codantes	2874027 bp	87%
Contenu en GC		57%
Plasmides	Aucun	
Nombre total de pCDS	3197	100%
Nombre de pCDS sur le chromosome I	2059	64%
Nombre de pCDS sur le chromosome II	1138	36%
pCDS avec une fonction assignée	2487	78%
pCDS sans fonction assignée	710	22%
pCDS sans fonction ou similarité	228	7%
pCDS sans fonction mais avec similarité	482	15%

La séquence complète du génome de *Brucella melitensis* a fourni des informations significatives sur les éléments essentiels de sa physiologie comme le métabolisme, les mécanismes de sécrétion et d'adhésion, les transporteurs et certaines caractéristiques de la paroi cellulaire (DelVecchio *et al*., 2002). Les génomes de bactéries proches de *Brucella melitensis* du point de vue évolutif sont également séquencés ou en cours de séquençage: *Brucella suis* (Paulsen *et al*., 2002), *Brucella abortus*, *Sinorhizobium meliloti* (Galibert *et al*., 2001), *Agrobacterium tumefaciens* C58 (Goodner *et al*., 2001; Wood *et al*., 2001), *Mesorhizobium loti* (Kaneko *et al*., 2000; Kaneko *et al*., 2000) et *Caulobacter crescentus* (Nierman *et al*., 2001). La comparaison des génomes de ces bactéries devrait permettre d'identifier des facteurs de virulence.

II. Objectifs

La structure 3D est une information capitale pour la compréhension de la fonction des protéines, de leurs interactions avec d'autres molécules et des effets phénotypiques des mutations de leurs acides aminés. La méthode la plus précise pour la prédiction de cette structure 3D est actuellement la modélisation par homologie. Elle pourra s'appliquer à un nombre toujours croissant de protéines vu l'augmentation exponentielle du nombre de structures déterminées expérimentalement. Mais cette technique de modélisation n'est pas parfaite, et il a été reconnu (Tramontano *et al*., 2001) que sa principale source d'erreurs réside dans la mauvaise qualité de l'alignement entre la séquence cible et son (ou ses) *template(s)*.

Simultanément, les biologistes moléculaires demandent à avoir accès à un nombre de plus en plus élevé d'informations théoriques et expérimentales sur les gènes et génomes qu'ils étudient. Ces informations se retrouvent bien souvent dans des banques de données dont le principal problème est le contrôle de la qualité de l'information qui y est stockée.

C'est dans ce contexte que trois objectifs ont été poursuivis au cours de ce travail: (1) l'amélioration d'une méthode d'alignement multiple de séquences protéiques, (2) la réalisation d'un système automatique de modélisation par homologie très efficace et (3) la construction d'une banque de données structurales et fonctionnelles de haute qualité pour les pCDS du génome de *Brucella melitensis*.

III. Ordinateurs et langages de programmation

Une partie des recherches a été effectuée sur des stations de travail Silicon Graphics Octane duo et INDIGO2 dont les fiches techniques sont:

1) Silicon Graphics Octane duo

2 processeurs 64 bits R10000 MIPS cadencés à 225MHz

512 MB de RAM

2 cartes graphiques ESI

2 HD SCSI de 9 GB et 1 HD SCSI de 4 GB

OS: IRIX 6.5

2) Silicon Graphics INDIGO2

1 processeur 32 bits R4400 MIPS cadencé à 150MHz

64 MB de RAM

1 carte graphique ESI

1 HD SCSI de 4 GB et 1 HD SCSI de 2 GB

OS: IRIX 5.3

Une autre partie des recherches a été réalisée sur un cluster de PC Priminfo constitué de:

1) Serveur Priminfo Xeon

2 processeurs 32 bits Intel Pentium 4 cadencés à 2.2 GHz

2 GB de RAM

3 HD SCSI de 36 GB et 2 HD EIDE de 120 GB

OS: Red Hat Linux 7.3

2) 6 nœuds de calcul Priminfo Xeon

2 processeurs 32 bits Intel Pentium 4 cadencés à 2.2 GHz

2 GB de RAM

1 HD EIDE de 80 GB

OS: Red Hat Linux 7.3

Les programmes développés dans le cadre de ce travail ont été écrits en langage Fortran 77 (ANSI, 1978), C (ANSI,), C++ (Stroustrup, 1999) et

Perl (Wall *et al*., 2000). Les versions des compilateurs correspondant sont SGI MIPSpro C/C++ 7.3, GNU GCC 3.04, SGI MIPSpro F77 7.3, interpréteur Perl 5.6.1.

IV. Amélioration du logiciel Match-Box

Dans notre unité de recherche, un programme d'alignement multiple, Match-Box, avait été développé en 1992 (Depiereux and Feytmans, 1992). Nous avons cherché à améliorer ses performances et ce travail été effectué en six étapes:

- Définition d'un ensemble d'alignements de référence permettant l'évaluation de la sensibilité et de la sélectivité d'un programme d'alignement de séquences.
- Réalisation d'un programme d'alignement hybride combinant la haute sélectivité de Match-Box et la haute sensibilité du programme ClustalW.
- Amélioration du *matching*, première étape de Match-Box, dans la perspective de tenter d'améliorer la sensibilité et la sélectivité de ce programme. Cette amélioration a été tentée d'abord en sélectionnant une matrice de scores plus performante, puis en choisissant la matrice de scores en fonction de chaque environnement local prédit et, finalement, en construisant, pour chaque séquence, une matrice de scores spécifique de la position (PSSM) via une recherche dans une banque de données de séquences.
- Amélioration du *screening*, deuxième étape de Match-Box. Celle-ci a été effectuée de deux manières. La première visait à améliorer la recherche du meilleur alignement et la seconde, à utiliser les prédictions de structure secondaire pour améliorer la construction de l'alignement.
- Evaluation de nos améliorations et évaluation d'autres programmes d'alignement de séquences.
- Développement d'une approche hybride utilisant toute une série de programmes d'alignement de séquences dans le *matching* et en développant un nouvel algorithme de *screening* utilisant des réseaux neuronaux.

Nous conclurons notre travail en discutant les diverses améliorations de la qualité de l'alignement de séquences et leur apport dans des applications utilisant des alignements de séquences.

IV.1. Construction d'ensembles d'alignements de référence

Au cours de nos recherches, quatre ensembles d'alignements de référence ont été utilisés. Ces ensembles diffèrent par le nombre d'alignements, le nombre de séquences qu'ils contiennent, le pourcentage d'identité moyen des séquences et la définition des parties supposées correctement alignées.

Le premier ensemble est composé des 20 alignements utilisés par Briffeuil *et al*. (Briffeuil *et al*., 1998) pour évaluer les performances de 7 programmes d'alignement de séquences. Les parties correctement alignées ont été déterminées manuellement après alignement des différentes structures des protéines composant chaque alignement. Les parties correctement alignées sont celles pour lesquelles le RMSD entre chaque segment aligné de 9 acides aminés ne dépasse pas 1.7 Å (alignements de structures réalisés avec InsightII).

Le deuxième ensemble comporte 33 alignements: aux 20 premiers alignements du premier ensemble sont venus s'ajouter 13 alignements de référence décrits dans la littérature (Tableau 5). Les parties correctement alignées sont celles décrites dans la littérature.

Tableau 5: Les 13 familles de séquences ajoutées à l'ensemble des 20 alignements de séquences de référence de Briffeuil *et al.* (Briffeuil *et al*., 1998).

Famille de protéines	Nombre de séquences
Les deux lobes des protéinases aspartiques (Blundell *et al*., 1991)	10
Globines (Bashford *et al*., 1987)	7
Motif Helix-Turn-Helix (Lawrence *et al*., 1993)	30
Immunoglobulines, domaine constant (Cohen *et al*., 1981)	4
Immunoglobulines, domaine variable (Cohen *et al*., 1981)	4
Lipocalines (Flower *et al*., 1993)	5
Protéines Sérine-Thréonine Kinases (Hanks *et al*., 1988)	13
Protéines Tyrosine Kinases (Hanks *et al*., 1988)	8
Domaines de la Ricine B (Rutenber *et al*., 1987)	6
Sérine protéinases de bactéries et de mammifères (Ding *et al*., 1994)	3

Protéines de l'enveloppe de virus de plantes (Carrington *et al*., 1987)	4
Protéines de l'enveloppe de virus de plantes et de Rhinovirus (Arnold and Rossmann, 1990)	4
Protéines de l'enveloppe de Rhinovirus (Arnold and Rossmann, 1990)	3

Pour le troisième ensemble, afin d'améliorer la qualité de nos évaluations, une nouvelle collection d'alignements de référence a été élaborée. Elle est extraite de la collection d'alignements multiples de structures décrite par Overington (Overington *et al*., 1992). Cette collection a été construite de la manière suivante: 96 familles de structures 3D (c'est-à-dire au total 443 structures) déterminées expérimentalement ont été soumises aux programmes d'alignement de structures MYNFIT (Sutcliffe *et al*., 1987) et COMPARER (Sali and Blundell, 1990). Parmi ces 96 familles de protéines, nous avons sélectionné celles qui possédaient au moins 3 séquences (alignement multiple), ce qui a réduit leur nombre à 78 (voir la liste dans l'Annexe 3). Les deux critères de vérité (ASG et CSS, voir section I.2.5.2) ont été appliqués pour définir les positions correctement alignées. En calculant la sensibilité et la sélectivité à l'aide de ces deux critères, il est possible d'établir des hypothèses sur les performances réelles déterminées en utilisant la comparaison des segments locaux par la technique du RMSD qui est lourde à déterminer.

Enfin, un quatrième ensemble d'alignements de séquences a été développé pour entraîner un réseau neuronal et évaluer les performances de notre méthode d'alignement pairé. Notre choix s'est porté sur la banque de données d'alignements de familles de structures 3D, HOMSTRAD. Les alignements de cette banque sont construits de la même manière que celle décrite par Overington (Overington *et al*., 1992). Nous avons utilisé la version de cette banque datant du mois de décembre 2002 dont seuls les 588 alignements pairés ont été retenus. Ce nombre étant néanmoins trop élevé que pour pouvoir réaliser nos tests dans un temps raisonnable, nous avons sélectionné de manière aléatoire 420 alignements. La liste de ces 420 familles de séquences est reprise à l'Annexe 4. Comme pour les 78 alignements de référence précédents, les deux critères extrêmes de ce qui est considéré comme correctement aligné (ASG et CSS) ont été utilisés.

IV.2. Match-Tal (Combinaison Match-Box/ClustalW)

IV.2.1. DESCRIPTION DE LA MÉTHODE

Notre premier essai d'amélioration du logiciel d'alignement de séquences Match-Box fut la réalisation d'un logiciel d'alignement hybride combinant les programmes Match-Box et ClustalW, d'où son nom Match-Tal. Ce logiciel avait comme objectif de combiner la sensibilité élevée de ClustalW et la sélectivité importante de Match-Box (propriétés décrites dans la section I.2.5.3).

Le fonctionnement de Match-Tal comporte quatre étapes (Figure 21):

1) L'alignement des séquences par Match-Box fournit un certain nombre de « boîtes » renfermant chacune un nombre déterminé de colonnes alignées. Chaque colonne alignée est caractérisée par un indice de confiance (Figure 21 a). Ces boîtes seront désignées par la suite par BOMB, Boîtes Originales de Match-Box.

2) Match-Tal sélectionne, dans les BOMB, les colonnes ayant une confiance élevée, suivant une valeur seuil introduite par l'utilisateur (indice $\leq$ 5, dans l'exemple). Les colonnes sélectionnées forment de nouvelles « boîtes » (BSMT, Boîtes Sélectionnées par Match-Tal), sous-ensemble de celui trouvé par Match-Box. Dans notre exemple (Figure 21 b), on peut distinguer ces BSMT (c'est-à-dire les parties 2, 4 et 6) du reste des protéines (parties 1, 3, 5 et 7).

3) Les parties non sélectionnées (1, 3, 5 et 7) sont ensuite soumises à ClustalW. Pour augmenter la qualité du travail d'alignement de ClustalW sur les segments de séquence bordant les BSMT, nous avons remarqué qu'il était indispensable de les accompagner des BSMT amont et aval (Figure 21 d, cadre A pour la partie 1, B pour le 3, C pour le 5, D pour le 7). Pour éviter que ClustalW ne modifie l'alignement des BSMT, nous avons remplacé les résidus de chaque colonne alignée des BSMT par le résidu y apparaissant le plus fréquemment (Figure 21 c), formant ainsi des boîtes modifiées par Match-Tal (BMMT) dans lesquelles les résidus sont conservés à 100% et dont ClustalW respectera l'alignement.

4) Lorsque le travail de ClustalW est terminé, l'alignement final est constitué par remplacement des BMMT par les BSMT (Figure 21 e).

a) Alignement de Match-Box.

```
Séq. 1  -------krrgaqlarlefne---nrylterssvlg-----lnewfqnkrakikks--------
Séq. 2  ----MRKrgrqtqtlelefhf---nryltrrahala-----ltewfqnrrmkwkkeNKTKGEPG
Séq. 3  MRKWQQIlfqaynpskeetlvEECnraeciqsqaqgLGSNLvtewfanrrkeeafrH-------

Ind. Conf.      99888555555588     555555588888       533333333333555
```

b) Choix de l'indice de confiance <= 5 pour Match-Tal.

```
Séq. 1  -------KRRGAqlarlefNE---nryltersSVLG-----lnewfqnkrakikks--------
Séq. 2  ----MRKRGRQTqtlelefHF---nryltrrAHALA-----ltewfqnrrmkwkkeNKTKGEPG
Séq. 3  MRKWQQILFQAYnpskeetLVEECnraeciqSQAQGLGSNLvtewfanrrkeeafrH-------

Ind. Conf.           5555555       5555555           533333333333555

Fragment     1          2      3      4         5           6            7
```

c) Remplacement des a.a. par les a.a. les plus fréquents dans la colonne

```
Séq. 1  -------KRRGAqlarlefNE---nryltersSVLG-----ltewfqnrrakikks--------
Séq. 2  ----MRKRGRQTqlarlefHF---nryltrrAHALA-----ltewfqnrrakikksNKTKGEPG
Séq. 3  MRKWQQILFQAYqlarlefLVEECnryltersQAQGLGSNLltewfqnrrakikksH-------

Fragment     1          2'     3      4'        5           6'           7
```

d) Soumission des boîtes A, B, C et D à ClustalW

```
                A            B             C             D
Séq. 1  -------KRRGAqlarlefNE---nryltersSVLG-----ltewfqnrrakikks--------
Séq. 2  ----MRKRGRQTqlarlefHF---nryltrrAHALA-----ltewfqnrrakikksNKTKGEPG
Séq. 3  MRKWQQILFQAYqlarlefLVEECnryltersQAQGLGSNLltewfqnrrakikksH-------

Fragment     1          2'     3      4'        5           6'           7
```

e) Assemblage de l'alignement final et remplacement des boîtes 2', 4' et 6' par les boîtes 2, 4 et 6

```
Séq. 1  -KR----RG--AqlarlefNE---nryltersSS--VLG---lnewfqnkrakikks--------
Séq. 2  MRK----RGRQTqtlelefHF---nryltrrAH--ALA---ltewfqnrrmkwkkeNKTKGEPG
Séq. 3  MRKWQQILFQAYnpskeetLVEECnraeciqSQAQGLGSNLvtewfanrrkeeafrH-------

Ind. Conf.          5555555     5555555          533333333333555

Fragment     1'        2     3'    4        5'          6           7'
```

Figure 21: Fonctionnement du programme Match-Tal.

IV.2.2. RÉSULTATS

Cette méthode a été évaluée sur base de l'ensemble des 33 alignements de référence: chaque jeu de séquences a été aligné en utilisant 8 matrices de scores différentes (BLOSUM45, BLOSUM62 et BLOSUM80 (Henikoff and Henikoff, 1992; Henikoff and Henikoff, 1993), les matrices JOHNSON92 et JOHNSON96 (Johnson and Overington, 1993) et les matrices de PAM120, PAM200, PAM250 (Dayhoff *et al.*, 1978)) et 9 valeurs seuil différentes pour l'indice de confiance.

Les résultats obtenus sont présentés dans l'Annexe 5. Pour plus de facilité et étant donné que les conclusions tirées pour les différentes matrices de scores sont similaires, nous avons choisi d'analyser en détail les résultats obtenus uniquement pour la matrice PAM120 (Figure 22) en fonction des 9 valeurs seuil de l'indice de confiance. Avec cette matrice, Match-Box a une sensibilité de 56,5% et une sélectivité de 58,9% alors que ClustalW a une sensibilité de 63,8% et une sélectivité de 53,3%. La sensibilité de Match-Tal (utilisant PAM120) n'est jamais supérieure à celle de ClustalW, et sa sélectivité est toujours inférieure à celle de Match-Box, quelle que soit la valeur seuil de l'indice de confiance choisie par l'utilisateur de Match-Tal. Les performances obtenues pour une valeur seuil de 3 ne peuvent pas être atteintes avec les autres matrices de scores.

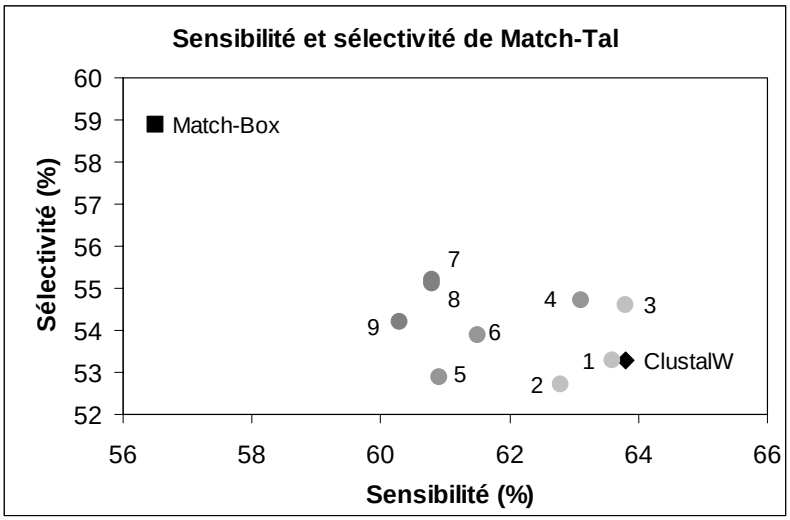


Figure 22: Sensibilité de Match-Tal (disques gris) en fonction de sa sélectivité pour différents seuils d'indice de confiance de Match-Box. On n'atteint jamais la sélectivité de Match-Box (carré noir) tout en gardant la sensibilité de ClustalW (losange noir).

IV.2.3. CONCLUSIONS

Les résultats obtenus avec cette technique montrent qu'il n'est pas possible d'améliorer les performances de l'alignement en suivant cette voie. En effet, non seulement il n'a pas été possible de profiter de la sensibilité de ClustalW, mais en plus, la sélectivité de cette approche est inférieure à celle de Match-Box.

Ces mauvais résultats peuvent s'expliquer par le fait que ClustalW doit aligner des segments peu similaires dans leur centre et très similaires (identiques) à leur extrémité. Ce sont ces extrémités qui influencent sans

doute le choix de la matrice de scores utilisée par ClustalW, celle-ci ne convenant probablement pas pour l'alignement des zones peu similaires.

IV.3. Amélioration de l'algorithme de « matching »

Le *matching* est la deuxième étape de la méthode Match-Box. C'est au cours de cette phase que les segments similaires sont identifiés et c'est donc des performances du *matching* dans la détection des « bonnes » similarités que dépendront d'abord les performances de Match-Box. En effet, si le *matching* ne permet de retrouver que, par exemple, 60% de positions correctement alignées, il ne sera pas possible de dépasser cette limite après l'exécution de la phase de *screening*.

Le but du *matching* étant de retrouver un maximum de positions correctement alignées, nous nous sommes intéressés uniquement à la sensibilité du *matching_SF* (voir description de ce *matching*, page 38). En effet, l'application du filtre statistique dans le *matching_MB* ne permet pas d'utiliser complètement les capacités du *screening*. De plus, nous avons remarqué lors de tests préliminaires que les positions correctement alignées retrouvées par Match-Box se trouvaient, dans plus de 95% des cas dans le résultat du *matching_SF*. Nous avons donc décidé de nous focaliser sur la réalisation d'un nouveau programme d'alignement fonctionnant non pas en 4 itérations comme dans la version originelle de Match-Box, mais en un seul cycle *matching-screening*. Dans ce seul cycle, il est par conséquent nécessaire d'avoir la plus haute sensibilité dans l'étape de *matching*.

Dans un premier temps, nous avons donc tenté de déterminer la sensibilité de la méthode actuelle, en fonction de matrices de scores choisies. Ensuite, des prédictions de la structure secondaire et de l'accessibilité au solvant ont été utilisées pour améliorer la recherche de segments similaires. Enfin, des matrices de scores spécifiques de chaque position de chaque séquence ont été utilisées pour améliorer les performances du *matching_SF*.

IV.3.1. PERFORMANCE EN FONCTION DE LA MATRICE DE SCORES

La sensibilité du *matching_SF* a été évaluée pour 52 matrices de scores en utilisant le critère ASG (voir section I.2.5.2) et notre banque de 78 alignements de référence. L'analyse n'a pas été effectuée avec le critère CSS (voir section I.2.5.2) car celui-ci est beaucoup trop restrictif et une amélioration de ce seul critère ne garantirait pas une amélioration du critère de référence utilisant les RMSD.

Les résultats complets de cette évaluation sont repris dans l'Annexe 6 et le Tableau 6 reprend les 14 matrices ayant la plus haute sensibilité ainsi que les deux ayant la plus faible sensibilité. La meilleure matrice est la matrice JOHNSON92. Cela n'est pas une surprise puisque cette matrice a été

construite en utilisant en partie notre banque de 78 alignements de référence (Johnson and Overington, 1993). Nous pouvons remarquer que, à part pour PAM120 et PAM200, les matrices de la famille BLOSUM obtiennent de meilleurs résultats que celles de la famille PAM. Cependant, nous devons constater que les 20 meilleures matrices de scores ont une sensibilité fort similaire, variant dans une tranche de 2% seulement.

Tableau 6: Sensibilité de l'algorithme de *matching_SF* en fonction de la matrice de scores utilisée.

Matrice	Sensibilité (%)	Matrice	Sensibilité (%)
JOHNSON92	70,7	BLOSUM45	70,0
PAM120	70,5	BLOSUM60	70,0
GONNET	70,5	BLOSUM85	70,0
PAM200	70,4	BLOSUM55	70,0
JOHNSON96	70,3	BLOSUM70	69,9
BLOSUM100	70,2	BLOSUM75	69,8
BLOSUM90	70,0	BLOSUM62	69,8
...	...	...	...
PAM460	62,5	PAM490	62,2

Les performances des programmes d'alignement de séquences peuvent varier considérablement suivant la matrice de scores utilisée. De plus, suivant le jeu de séquences à aligner, la qualité des résultats dépend également de la matrice de scores utilisée. Les performances du *matching_SF* peuvent donc être améliorées en utilisant la matrice de scores fournissant la meilleure sensibilité. En conclusion de cette évaluation, la matrice BLOSUM62, qui est celle par défaut de Match-Box, pourrait être remplacée par la matrice PAM120 ou GONNET.

IV.3.2. MATRICE DE SCORES SPÉCIFIQUE DE L'ENVIRONNEMENT

Pour améliorer les performances de l'algorithme de *matching_SF*, nous avons imaginé d'utiliser des matrices de scores différentes en fonction de deux caractéristiques prédites de chaque segment de séquence: la structure secondaire (brin β, hélice α ou boucle) et l'accessibilité au solvant (enfoui, exposé ou intermédiaire). La combinaison de ces deux caractéristiques nous a permis de définir 15 états différents pour chaque segment suivant leurs portions de résidus dans une hélice, un brin ou une

boucle et exposé, intermédiaire ou enfoui (voir Tableau 8). Nous avons tenté de savoir quelles étaient, parmi les 134 matrices de scores trouvées dans la littérature (Annexe 7), celles qui devaient être utilisées pour chaque état. Pour cela, nous avons appliqué à chacun des 78 alignements de référence, et pour chacune des 134 matrices de scores, les opérations décrites ci-après.

Chaque séquence est soumise au programme PHD (Rost *et al*., 1994) qui prédit la structure secondaire et l'accessibilité au solvant de chaque résidu. Ensuite, toutes les comparaisons pairées de segments sont effectuées suivant l'algorithme de *matching*, et chaque fois qu'une paire de segments correspondant à la distance minimum $D_{i,j,l}$ (voir section I.2.6.2) est observée, on vérifie si cette paire de segments est réellement alignée suivant le critère ASG (voir section I.2.5.2), dans l'alignement de structures.

S'ils le sont, la composition en hélice/brin/boucle et exposé/intermédiaire/enfoui est retenue et l'écart entre la distance minimum $D_{i,j,l}$ et la moyenne des distances calculées pour les autres paires de segments, $\overline{D_{i,j,l}}$, est calculée via une statistique de type *t* de Student:

$$t_{i,j,l,p} = \frac{\overline{D_{i,j,l}} - D_{i,j,l,p}}{s_{i,j,l}}$$

$$\overline{D_{i,j,l}} = \frac{1}{L_l - W} \cdot \sum_{\substack{m=1 \\ m \neq p}}^{L_l - W + 1} D_{i,j,l,m}$$

Où L_l est la longueur de la séquence l

W est la taille des segments comparés

i est le numéro de la séquence contenant le premier segment

j est la position du premier acide aminé de ce segment

l est le numéro de la séquence contenant le deuxième segment

m est la position du premier acide aminé du deuxième segment

p est la position du premier acide aminé du deuxième segment lorsque celui-ci est à la distance minimum $D_{i,j,l}$ du premier segment

$s_{i,j,l}$ est l'écart type des distances $D_{i,j,l,m}$ avec $m \neq p$

Plus la valeur de $t_{i,j,l,p}$ est élevée, plus la distance $D_{i,j,l,p}$ se détachera statistiquement des autres distances considérées. Ce critère est donc un indicateur de l'intensité de la réponse de la matrice de scores lors de la détection des appariements.

Les quelques 48000 paires de segments ainsi obtenues sont donc caractérisées par leurs compositions dans les six environnements locaux et leurs valeurs *t* de Student pour les 134 matrices de scores. Le programme de traitement statistique STATISTICA a ensuite été utilisé pour traiter toutes ces données par une analyse factorielle. Celle-ci fournit un graphe de dispersion en trois dimensions: chaque environnement local et chaque matrice sont représentés par un point dans un espace dont les trois dimensions ont été définies à partir des facteurs principaux provenant de l'analyse factorielle.

Le but étant d'associer à un environnement local la matrice de scores la plus spécifique possible, chaque point représentant un environnement local a été associé au point le plus proche représentant une matrice de scores. Les six matrices de scores les plus spécifiques pour chaque environnement local sont reprises dans le Tableau 7.

Tableau 7: Matrice de scores la plus spécifique de chaque environnement local avec sa description dans la littérature.

Environnement local	Meilleure matrice	Description de la matrice
Brin	OVEJ920102	*Environment-specific amino acids substitution for beta residues* (Overington *et al*., 1992)
Hélice	GEOD900101	*Hydrophobicity scoring matrix* (George *et al*., 1990)
Boucle	JOND940101	*The 250 PAM transmembrane protein exchange matrix* (Jones *et al*., 1994)
Enfoui	LUTR910104	*Structure-based comparison table for inside alpha class* (Luthy *et al*., 1991)
Exposé	LUTR910103	*Structure-based comparison table for inside alpha class* (Luthy *et al*., 1991)
Intermédiaire	AZAE970102	*Substitution matrix derived from spatially conserved motifs* (Azarya-Sprinzak *et al*., 1997)

Pour identifier les matrices les plus spécifiques des 9 combinaisons d'environnements, un point a été créé pour chaque combinaison. Ce point se situe exactement à mi-distance entre le point de structure secondaire et celui d'accessibilité au solvant dans l'espace factoriel. La matrice de scores la plus spécifique de chaque combinaison est celle dont le point dans l'espace factoriel est le plus proche. Le résultat de cette opération est repris dans le Tableau 8.

Tableau 8: Matrice de scores la plus spécifique de chaque combinaison d'environnement local avec sa description dans la littérature.

Environnement local	Meilleure matrice	Description de la matrice
Brin-enfoui	OVEJ920104	*Environment-specific amino acid substitution matrix for inaccessible residues* (Overington *et al*., 1992)
Brin-exposé	LEVJ860101	*The secondary structure similarity matrix* (Levin *et al*., 1986)
Brin-intermédiaire	AZAE970102	*The substitution matrix derived from spatially conserved motifs* (Azarya-Sprinzak *et al*., 1997)
Hélice-enfoui	QU_C930102	*Cross-correlation coefficients of preference factors side chain* (Qu *et al*., 1993)
Hélice-exposé	KOSJ950113	*Context-dependent optimal substitution matrices for exposed residues* (Koshi and Goldstein, 1995)
Hélice-intermédiaire	AZAE970102	*The substitution matrix derived from spatially conserved motifs* (Azarya-Sprinzak *et al*., 1997)
Boucle-enfoui	OVEJ920104	*Environment-specific amino acid substitution matrix for inaccessible residues* (Overington *et al*., 1992)
Boucle-exposé	KOSJ950104	*Context-dependent optimal substitution matrices for exposed coil* (Koshi and Goldstein, 1995)
Boucle-intermédiaire	AZAE970102	*The substitution matrix derived from spatially conserved motifs* (Azarya-Sprinzak *et al*., 1997)

Les matrices retrouvées par notre méthode correspondent souvent au type d'environnement local sur lequel est basée leur conception, ce qui valide en quelque sorte notre méthode.

Après avoir sélectionné les matrices de scores permettant une meilleure identification des paires de segments en utilisant leur environnement local prédit, l'algorithme de *matching_SF* a été modifié pour tenir compte de ces choix. La nouvelle version de cet algorithme sera désignée par *matching_EL*. Celui-ci lit d'abord un fichier reprenant les accessibilités au solvant et les structures secondaires prédites par le programme PHD. Ensuite, un fichier reprenant les compositions des différents états et la matrice de scores associée est lu. Suivant le type

d'environnement prédit, la matrice de scores la plus adéquate est donc utilisée.

L'évaluation sur notre ensemble de 78 alignements de référence donne une sensibilité de 71.8% soit une amélioration de 1.2% par rapport à la sensibilité observée pour la matrice de scores la plus performante sur cet ensemble: PAM120. La matrice JOHNSON92 ne peut pas être considérée puisqu'elle a été construite à partir d'alignements de cet ensemble.

En conclusion, nous avons montré que cette approche pouvait améliorer légèrement la sensibilité de l'étape de *matching*, mais pas de façon intéressante. Néanmoins, cette amélioration nous montre une voie à suivre pour améliorer les performances du *matching_SF*. Nous avons donc voulu généraliser cette technique en utilisant des matrices de scores spécifiques de chaque position et non plus d'un segment.

IV.3.3. MATRICE DE SCORES SPÉCIFIQUE DE LA POSITION

Pour améliorer leurs performances, toute une série de techniques de prédiction se basent sur l'information contenue dans la variabilité des séquences protéiques. Cette information peut être retrouvée par une recherche de séquences similaires dans une banque de séquences. Citons par exemple les techniques de reconnaissance de *fold* (Jones, 1999), de prédiction de la structure secondaire (Jones, 1999; Cuff and Barton, 2000), de prédiction de l'accessibilité au solvant (Rost and Sander, 1995). Pour l'alignement multiple de séquences, cette information a surtout été utilisée dans les méthodes d'alignement progressives et par affinement itératif (voir sections I.2.4.3 et I.2.4.4). Cependant, ces techniques n'utilisent que la variabilité des séquences à aligner et non celle issue d'une recherche de similarité dans une banque de séquences.

Nous avons donc imaginé une stratégie de *matching* où la matrice (réduite à un vecteur) de scores est différente d'une position à l'autre de chaque séquence. Cette matrice de scores spécifique de la position (PSSM) peut être générée par le programme PSI-BLAST (voir section I.2.3.4) par une recherche dans la banque de données *nr* du NCBI. Le nombre maximum d'itérations de PSI-BLAST est de 3 et le seuil d'*expected value* (*E-value*) de 0.001. Chaque séquence à aligner possède dès lors sa propre PSSM. Le *matching_SF* a été modifié pour utiliser les PSSM générées par PSI-BLAST. Ce *matching_PSSM* détecte les segments de 9 résidus les plus similaires, de la même manière que celle décrite dans la section I.2.6.2.

La distance entre deux segments est obtenue par la formule suivante:

$$D_{\mathrm{i,j,l,m}} = \sum_{k=0}^{W-1} P_{i,\,j+k}\left(AA_{l,m+k}\right)$$

où *i* est le numéro de la séquence contenant le premier segment

j est la position du premier acide aminé de ce segment

l est le numéro de la séquence contenant le deuxième segment

m est la position du premier acide aminé du deuxième segment

$P_{i,j+k}(AA_{l,m+k})$ est la distance séparant les acides aminés aux positions $j+k$ de la séquence *i* et $m+k$ de la séquence *l*. $P_{i,j+k}$ peut prendre 20 valeurs qui sont fournies par la PSSM de la séquence *i* exprimée en distance et non en similarité.

$D_{i,j,l,m}$ est la distance séparant les deux segments

La sensibilité de cette approche a été testée sur notre ensemble de 78 alignements de référence en utilisant le critère ASG et vaut 80.1%, soit 9.6% de plus qu'en utilisant la meilleure matrice PAM120. Pour les mêmes raisons que celles évoquées au point IV.3.1, le critère CSS n'a pas été utilisé.

Après avoir amélioré de façon assez importante la sensibilité de l'étape de *matching*, il est nécessaire d'améliorer l'étape de *screening* pour qu'elle discrimine au mieux les boîtes « correctes » (contenant des segments correctement alignés), des boîtes dites « incorrectes ».

IV.4. Amélioration de l'algorithme de « screening »

Le *screening* est la troisième étape du fonctionnement de Match-Box. Nous avons d'abord évalué ses performances, et ensuite nous avons tenté de les faire progresser. Pour cela, deux types de modification du *screening* ont été imaginées: la première vise à améliorer l'algorithme de recherche du meilleur alignement, et la seconde à utiliser les prédictions de structure secondaire pour améliorer la construction de l'alignement. Les performances de ces deux approches ont été évaluées et comparées à celles de la méthode originale.

IV.4.1. EVALUATION DES PERFORMANCES DE L'ALGORITHME DE *SCREENING* DE MATCH-BOX

IV.4.1.1. Description de la méthode

La qualité du *screening* a été évaluée sur l'ensemble de 78 alignements de référence en utilisant les deux critères de vérité expliqués précédemment (ASG et CSS, voir section I.2.5.2). L'évaluation a été réalisée avec deux objectifs.

Premièrement, nous mesurons l'efficacité du *screening* de Match-Box (nommé *screening_MB*). Cette mesure représente la capacité du *screening* à retrouver les boîtes correctes fournies par l'étape de *matching*. Entre d'autres termes, il s'agit du rapport entre la sensibilité d'un cycle *matching_screening* et celle de l'étape de *matching*. Les efficacités des améliorations de l'étape de *screening* pourront ainsi être rapportées à celle de l'algorithme original.

Deuxièmement, les qualités d'alignement d'un cycle complet *matching_SF-screening_MB* seront analysées pour établir une référence à laquelle seront comparées les futures améliorations. Lors du *matching_SF*, la matrice de scores utilisée était BLOSUM62 (Henikoff and Henikoff, 1992; Henikoff and Henikoff, 1993) et la longueur des segments variait de 5 à 21.

IV.4.1.2. Résultats

Les Figure 23 et Figure 24 reprennent les résultats de l'évaluation effectuée en utilisant les deux critères définis ci-dessus et en faisant varier la taille du segment d'analyse (voir section I.2.6.1) utilisée dans le *matching_SF*. En effet, la taille du segment d'analyse dans Match-Box a été

fixée à 9 résidus de manière empirique et nous voulions vérifier si ce choix était le meilleur.

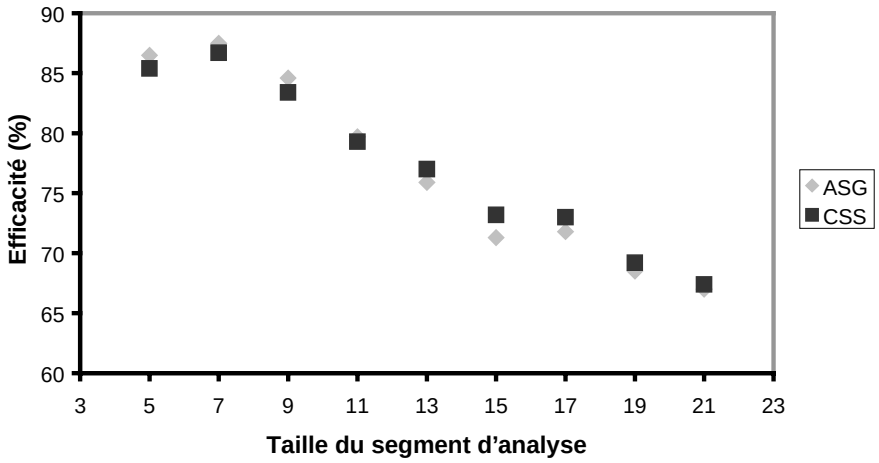


Figure 23: Efficacité du *screening de Match-Box* en fonction de la taille du segment d'analyse pour les deux critères de vérité ASG (losanges gris) et CSS (carrés noirs).

L'analyse de l'efficacité du *screening* (Figure 23) montre que pour les deux critères ASG et CSS, celle-ci avoisine les 85% lorsqu'on utilise des segments de 5, 7 et 9 résidus, avec un maximum de 87,5% pour une taille de segment de 7 résidus en utilisant le critère ASG. Par contre, la tendance générale veut que plus la taille des segments augmente (11 à 21), plus l'efficacité diminue.

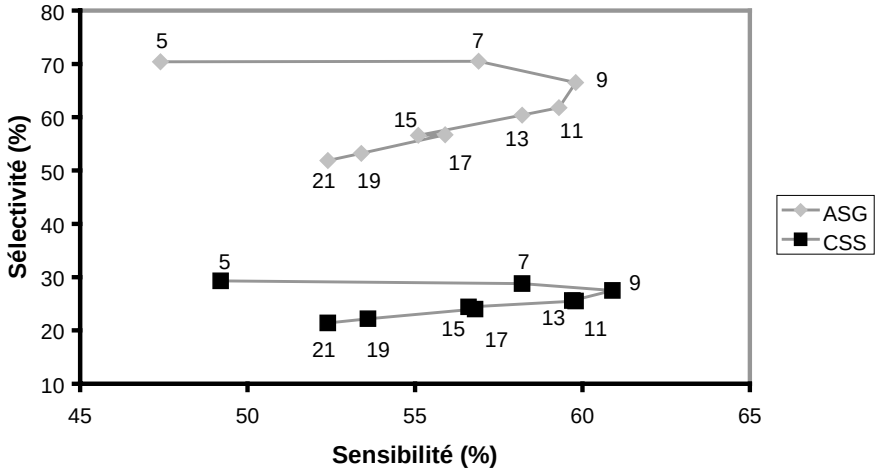


Figure 24: Sélectivité de l'alignement en fonction de la sensibilité pour un cycle *matching-screening*, en utilisant les critères de vérité ASG (losanges gris) et CSS (carrés noirs). Les valeurs associées à chaque point représentent les tailles des segments.

Pour ce qui est de l'évolution de la sélectivité en fonction de la taille du segment d'analyse (Figure 24), les mêmes observations peuvent être faites que pour l'évolution de l'efficacité. En effet, nous observons une sélectivité élevée, de l'ordre de 70% (ASG) ou 30% (CSS), pour des longueurs de segments de 5 et 7 et puis, de manière générale, celle-ci diminue en fonction de l'augmentation de la taille du segment.

De son côté, la sensibilité augmente jusqu'à un maximum d'environ 60% pour des segments de 9 résidus avant de diminuer constamment avec l'augmentation de la taille du segment d'analyse. Un résultat de cette analyse qui n'est pas montré dans cette thèse est que la sensibilité du *matching_SF* augmente constamment, quel que soit le critère (ASG ou CSS), passant d'environ 55% pour des segments de 5 résidus à environ 75% pour des segments de 11 résidus. La sensibilité du *matching_SF* plafonne ensuite à environ 77% pour les tailles de segment supérieures. Par conséquent, une taille de 9 résidus semble être un bon compromis entre l'augmentation de la sensibilité du *matching_SF* et la diminution de l'efficacité du *screening_MB*.

La diminution des performances lorsqu'on augmente la taille du segment d'analyse peut s'expliquer de la manière suivante: à cause des effets de bord, le nombre de résidus mal alignés augmente lorsqu'on utilise des segments d'analyse de taille élevée, ce qui entraîne un phénomène d'incompatibilité entre boîtes de plus en plus important. Par conséquent, le nombre de boîtes choisies par le *screening_MB* est plus réduit (perte de sensibilité).

IV.4.1.3. Conclusions

Pour les deux critères de vérité considérés (ASG et CSS), l'efficacité du *screening_MB* n'atteint jamais 100%. Néanmoins, un maximum de 87,5% est atteint pour le critère ASG et un maximum de 86,7% est obtenu avec le critère CSS. Une amélioration du *screening_MB* reste donc possible.

Le *screening* de Match-Box (*screening_MB*) est plus efficace lorsque la taille des segments est de 7. Cependant, ce qui nous intéresse *in fine* pour l'amélioration de l'alignement, c'est la sensibilité et la sélectivité de l'ensemble *matching-screening*. Or, ces taux sont les plus élevés lorsque la taille des segments est de 9 résidus, c'est-à-dire celle qui est choisie empiriquement par le concepteur de Match-Box.

Ces conclusions montrent qu'il est possible d'améliorer l'étape de *screening* de Match-Box. Deux approches seront tentées: la première se base sur l'extension de l'algorithme actuel et la seconde consiste à améliorer le système de score des boîtes en tenant compte des structures secondaires prédites.

IV.4.2. DÉVELOPPEMENT D'UNE NOUVELLE STRATÉGIE DE *SCREENING*

IV.4.2.1. Description de la méthode

Avant de décrire un nouvel algorithme de *screening*, il semble utile de rappeler brièvement le fonctionnement du *screening* originel de Match-Box (voir section I.2.6.3). Le *screening* sélectionne d'abord la boîte la plus longue pour la construction de l'alignement final. Ensuite, les opérations suivantes sont répétées:

a. inventaire des boîtes compatibles avec la ou les boîtes déjà sélectionnées,

b. sélection de la boîte la plus longue de l'inventaire.

Ces itérations sont arrêtées quand il n'y a plus de boîte compatible avec celles déjà sélectionnées.

La nouvelle stratégie de *screening* (*screening_NS*) teste tous les arrangements possibles de toutes les boîtes compatibles entre elles et recherche l'alignement ayant la longueur la plus grande possible.

Nous avons évalué les performances de ce nouvel algorithme de *screening* sur base de la collection de 78 alignements de référence et en utilisant les critères ASG et CSS. Au cours des exécutions de cette version

modifiée de Match-Box, nous avons utilisé la matrice de scores par défaut, BLOSUM62, et nous avons fait varier le segment d'analyse de 5 à 21 résidus pour vérifier si la longueur de 9 reste toujours optimale malgré le changement d'algorithme.

IV.4.2.2. Résultats

La Figure 25 et la Figure 26 montrent que, sauf pour des tailles de fenêtre d'analyse de 5 et 7, les performances du cycle *matching_SF-screening_NS* sont, en moyenne, meilleures que celles obtenues pour un cycle *matching_SF-screening_MB* et ce, quel que soit le critère de vérité considéré. Un examen plus attentif de ces deux graphiques montre qu'une fenêtre d'analyse de 9 résidus présente un meilleur couple sélectivité-sensibilité (se rapprochant plus des 100% de sélectivité et de sensibilité) qu'avec une autre taille de fenêtre d'analyse, lorsque nous travaillons avec la nouvelle procédure de *screening*.

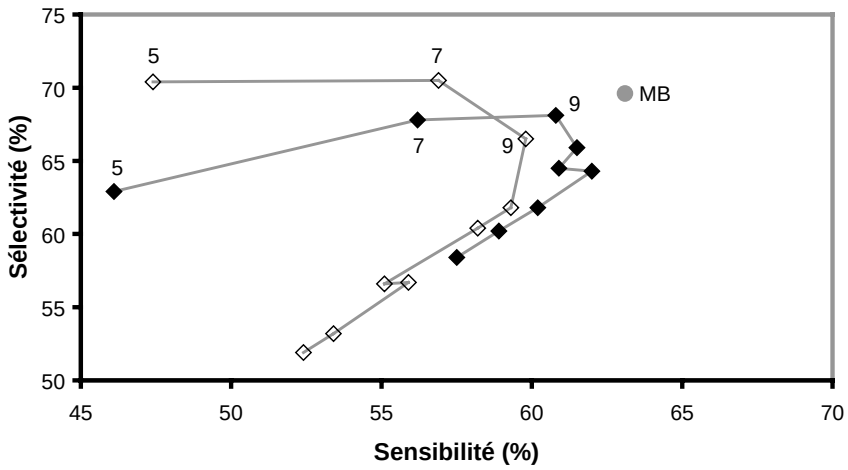


Figure 25: Evolution de la sensibilité et de la sélectivité d'un cycle *matching_SF-screening_NS* (losanges noirs) et du *matching_SF-screening_MB* (losanges blancs), en fonction de la taille du segment d'analyse pour le critère de vérité ASG. Le disque gris représente les performances du programme Match-Box originel.

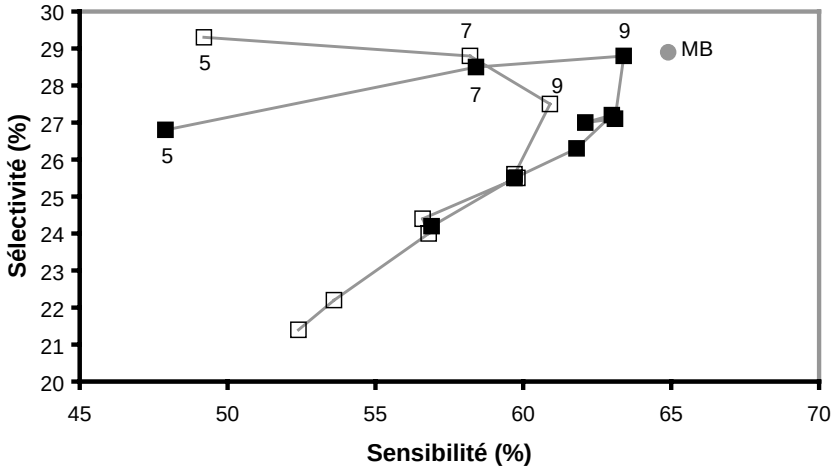


Figure 26: Evolution de la sensibilité et de la sélectivité du nouvel algorithme de *screening* (carrés noirs) et du *screening* originel (carrés blancs), en fonction de la taille du segment d'analyse pour le critère de vérité CSS. Le disque gris représente les performances du programme Match-Box.

Un segment de 9 résidus permet d'obtenir de meilleurs performances avec l'ensemble *matching_SF-screening_NS*. La comparaison entre les valeurs de sélectivité et de sensibilité de ce nouveau programme d'alignement et celles de Match-Box montre que la nouvelle procédure d'alignement permet d'obtenir, en un seul cycle d'exécution, des résultats proches de ceux de Match-Box.

L'évolution de l'efficacité du *screening_NS* en fonction de la taille de la fenêtre d'analyse est représentée dans la Figure 27. Cette efficacité est moins bonne que celle du *screening_MB* pour des longueurs de segment de 5 et 7 résidus alors qu'elle est plus élevée pour des tailles plus importantes, quel que soit le critère de vérité utilisé.

La meilleure efficacité du *screening_NS* pour les deux critères de vérité est observée avec une fenêtre d'analyse de 9 résidus, c'est-à-dire la taille utilisée dans la version originelle de Match-Box. Les améliorations de l'efficacité *screening_NS* par rapport au *screening_MB* pour les critères ASG et CSS sont respectivement de 1.4% et 3.4% pour une fenêtre de 9 résidus.

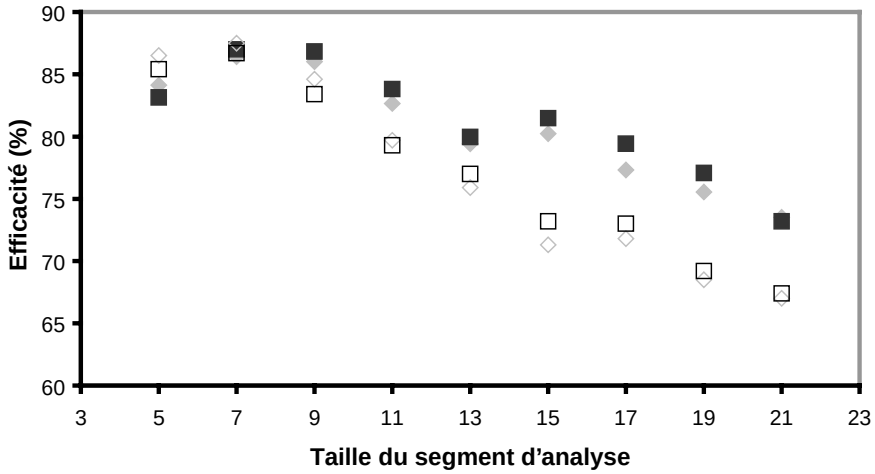


Figure 27: Evolution de l'efficacité du *screening_MB* (quadrilatères blancs) et du *screening_NS* (quadrilatères pleins) pour les deux critères de vérité ASG (losanges gris) et CSS (carrés noirs) en fonction de la taille du segment d'analyse.

Le *screening_NS* a été intégré dans Match-Box mais les performances n'ont pas été modifiées, ce qui est normal car on sait que le *screening* ne travaille pas de manière optimale lorsque le *matching* utilise un filtre.

IV.4.2.3. Conclusions

Le nouvel algorithme de *screening* est plus efficace et permet au programme *matching_SF-screening_NS* d'obtenir, de façon générale, une meilleure sélectivité et une meilleure sensibilité que celles du programme *matching_SF-screening_MB* et ce, quel que soit le critère de vérité choisi (ASG ou CSS).

Les résultats obtenus par notre nouveau programme en un seul cycle sont très proches de ceux obtenus avec Match-Box, en quatre cycles *matching_MB-screening_MB.*

IV.4.3. PRISE EN COMPTE DE LA STRUCTURE SECONDAIRE DANS L'ALGORITHME DE *SCREENING*

IV.4.3.1. Etude de faisabilité

IV.4.3.1.1. Description de la méthode

Avant de procéder à l'utilisation des prédictions de structure secondaire dans l'étape de *screening*, il fallait d'abord vérifier si la conservation de la structure secondaire dans les boîtes permettait de discriminer les boîtes correctes des boîtes incorrectes. La conservation de la structure secondaire dans une boîte se calcule de la manière suivante: pour chaque colonne de la boîte, on calcule la fréquence relative des acides aminés observés dans chacun des deux types de structures secondaires régulières (hélice α et brin β). Ensuite, on calcule la moyenne des fréquences les plus importantes observée dans chaque colonne de la boîte.

Nous avons donc étudié, pour le *matching_SF*, l'évolution de la sélectivité (pourcentage moyen de colonnes correctement alignées dans les boîtes) en fonction de la conservation de la structure secondaire et de la longueur des boîtes. Le *matching_SF*, employant la matrice BLOSUM62 a été appliqué sur notre banque de 78 alignements de référence en faisant varier la taille des segments d'analyse de 5 à 21 résidus. Les critères ASG et CSS ont servi à déterminer les zones correctement alignées.

De plus, la structure secondaire de chaque protéine de notre banque fut obtenue de deux manières. D'une part, les coordonnées des atomes de toutes les protéines, extraites de la banque de données PDB, furent soumises au programme DSSP (Kabsch and Sander, 1983) qui fournit la structure secondaire observée de la protéine. D'autre part, la structure secondaire de chacune des protéines a été prédite en soumettant leur séquence au logiciel PHD: nous avons ainsi pu évaluer la sélectivité des boîtes sortant du *matching_SF* dans la situation où seule la structure secondaire prédite est utilisable.

IV.4.3.1.2. Résultats

Les résultats de l'évaluation sont représentés sur quatre figures reprises dans l'Annexe 8. Ces quatre diagrammes étant similaires, un seul d'entre eux est présenté dans le texte (Figure 28). On observe que plus la conservation de la structure secondaire dans une boîte est élevée et plus cette boîte est longue, plus la sélectivité est élevée. De manière systématique, lorsque la structure secondaire est prédite par PHD, la sélectivité est plus faible que lorsqu'on se base sur la structure secondaire réelle. Ce qui est

explicable par le fait que la qualité de la prédiction des structures secondaires ne dépasse pas 75%, en moyenne. De même, la sélectivité obtenue pour le critère ASG est systématiquement plus élevée que pour le critère CSS. Ceci est normal puisque le critère CSS est plus strict.

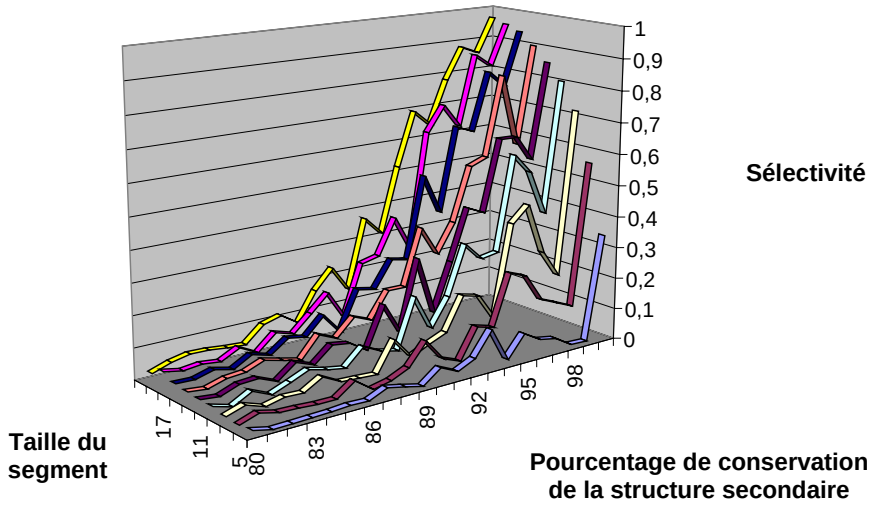


Figure 28: Evolution de la sélectivité en fonction de la conservation de la structure secondaire dans les boîtes et de la taille du segment d'analyse pour le critère ASG, en utilisant PHD pour prédire la structure secondaire.

IV.4.3.1.3. Conclusion

Suite à l'analyse de la Figure 28 et des résultats présentés à l'Annexe 8, nous pouvons conclure que plus la structure secondaire est conservée dans une boîte et plus cette boîte est longue, plus le nombre de résidus correctement alignés dans une boîte est élevé.

Par conséquent, la conservation de la structure secondaire dans une boîte représente un complément d'information susceptible d'améliorer les performances du *screening*.

IV.4.3.2. Implémentation de la méthode

IV.4.3.2.1. Description de la méthode

Nous proposons une approche (*screening_SS*) où le critère (score) utilisé pour sélectionner les différentes boîtes ne dépend plus uniquement de la longueur des différentes boîtes et de la somme des carrés des écarts (SCE) des différents décalages entre ces boîtes (voir section I.2.6.3). Nous ajoutons

à ce critère la conservation en structures secondaires prédites par PHD dans les diverses boîtes. Ce nouveau critère s'exprime sous forme d'un score (*S*) calculé par la formule empirique suivante:

$$S = \frac{SS^{a} \cdot L^{b}}{SCE\Delta_{g}^{c}}$$

où *S* est le score

SS est la conservation de la structure secondaire

L est la longueur de la boîte

$SCE\Delta_g$ est la somme des carrés des écarts des différents décalages entre les boîtes

a, *b* et *c* sont des exposants qu'il faut déterminer de manière empirique

Les valeurs des exposants *a*, *b* et *c* ont été optimisées par la méthode *steepest descent* de manière à obtenir le meilleur couple sélectivité-sensibilité pour un seul cycle *matching_SF-screening_SS*, pour notre banque de 78 alignements de séquences et la BLOSUM62 et un segment d'analyse de 9 résidus. Ensuite, nous avons réalisé une nouvelle version de Match-Box en remplaçant le *screening_MB* par le nouveau *screening_SS*. Cette version de Match-Box (Match-Box_SS) fonctionnant en quatre itérations *matching_MB-screening_SS*, a été évaluée en utilisant la banque de 78 alignements de référence et les deux critères de vérité ASG et CSS. Pour cette évaluation, nous avons utilisé, dans l'étape de *matching_MB*, la matrice BLOSUM62 et un segment d'analyse de 9 résidus.

IV.4.3.2.2. Résultats

Les meilleures performances de l'ensemble *matching_SF-screening_SS* ont été obtenues avec des valeurs pour les exposants *a*, *b* et *c* de 4.50, 2.12 et 0.145, respectivement. Cela signifie par exemple qu'une boîte de 9 résidus dont la structure secondaire est conservée à 100% sera prioritaire par rapport à une boîte de 14 résidus ayant une conservation de structures secondaires inférieure à 90% (Tableau 9).

Tableau 9: Longueurs et conservations de la structure secondaire des boîtes obtenant un score identique à une boîte de 9 résidus ayant une conservation de la structure secondaire de 100%. La somme des carrés des écarts des différents décalages entre les boîtes est supposée constante.

Longueur de la boîte (acides aminés)	Conservation de Structure Secondaire (%)
9	100,0
10	97,7
11	95,6
12	93,8
13	92,1
14	90,6
15	89,2

Les performances de l'ensemble *matching_SF-screening_SS* sur la banque de 78 alignements de référence sont reprises dans le Tableau 10. Les résultats obtenus expriment une nette amélioration de la sélectivité et de la sensibilité par rapport à l'ensemble *matching_SF-screening_MB*, quel que soit le critère de vérité utilisé. Le *screening_SS* a une efficacité de 93% pour le critère ASG et de 91,3% pour le critère CSS, soit une amélioration de 8,4% et 7,9% respectivement, par rapport au *screening_MB*. Ceci nous conduit à préciser que la limite d'efficacité de 100% n'est plus si éloignée et qu'il sera difficile de s'en rapprocher davantage.

Tableau 10: Performances d'un cycle *matching_SF-screening_MB* et d'un *matching_SF-screening_SS* déterminées avec les critères de vérité ASG et CSS. Les améliorations dues au *screening_SS* sont également rapportées.

Programme	Sensibilité (%)	Sélectivité (%)	Efficacité *screening* (%)
Matching_SF-screening_MB, ASG	59.8	66.5	84.6
Matching_SF-screening_SS, ASG	65.7	69.9	93.0
Amélioration, ASG	**5.9**	**3.4**	**8.4**
Matching_SF-screening_MB, CSS	60.9	27.5	83.4
Matching_SF-screening_SS, CSS	66.7	28.9	91.3
Amélioration, CSS	**5.8**	**1.4**	**7.9**

Les performances de notre méthode d'alignement fonctionnant en un seul cycle *matching_SF-screening_SS* sont également meilleures que celles du programme Match-Box (Tableau 11). En effet, par rapport au programme Match-Box, la sensibilité de l'algorithme *matching_SF-screening_SS* est meilleure de 2,6% (ASG) et 0,6% (CSS) alors que la sélectivité est meilleure de 0,6% (ASG) ou identique (CSS).

Tableau 11: Performances de Match-Box et de l'exécution d'un seul cycle *matching_SF-screening_MB* pour les critères de vérité ASG et CSS. Les améliorations des performances gagnées en utilisant un seul cycle *matching_SF-screening_MB* sont également rapportées.

Programme	Sensibilité (%)	Sélectivité (%)
Match-Box, ASG	63.1	69.6
Matching_SF-screening_SS, ASG	65.7	69.9
Amélioration, ASG	**2.6**	**0.3**
Match-Box, CSS	64.9	28.9
Matching_SF-screening_SS, CSS	65.5	28.9
Amélioration, CSS	**0.6**	**0.0**

L'intégration du *screening_SS* dans le programme Match-Box_SS permet d'obtenir une sélectivité et une sensibilité légèrement plus élevées que celles de Match-Box utilisant le *screening_MB* (Tableau 12). Cependant, les valeurs de sélectivité et sensibilité ne sont pas aussi élevées que celles obtenues avec un seul cycle *matching_SF-screening_SS* (voir Tableau 11). Ceci montre que, comme nous l'avions déjà précisé dans l'introduction (section I.2.6.2), le filtre du *matching_MB* dans Match-Box limite l'amélioration des performances de ce programme.

Tableau 12: Performances de Match-Box (avec *screening_MB*) et de Match-Box_SS (avec *screening_SS*) pour les critères de vérité ASG et CSS. Les améliorations dues au *screening_SS* sont également rapportées.

Programme	Sensibilité (%)	Sélectivité (%)
Match-Box, ASG	63.1	69.6
Match-Box_SS, ASG	63.7	69.6
Amélioration	**0.6**	**0.0**
Match-Box, CSS	64.9	28.9
Match-Box_SS, CSS	65.5	28.9
Amélioration	**0.6**	**0.0**

IV.4.3.3. Conclusions

Trois conclusions peuvent être tirées de l'utilisation des prédictions de structure secondaire dans l'étape de *screening*:

- L'efficacité du *screening_SS* est nettement meilleure que celle du *screening_MB*, ce qui améliore la sélectivité et la sensibilité lors de l'exécution d'un seul cycle *matching_SF-screening_SS* par rapport à l'exécution d'un seul cycle *matching_SF-screening_MB*. C'est donc la deuxième fois, en tenant compte des résultats obtenus avec le *screening_NS*, que l'efficacité de l'étape de *screening* est améliorée.
- La sélectivité et la sensibilité de notre approche *matching_SF-screening_SS* en un seul cycle sont meilleures que celles du programme Match-Box. C'est la première fois depuis la dernière publication concernant Match-Box que les performances de cette approche ont été améliorées, en tenant compte du contenu des boîtes, c'est-à-dire de la conservation des structures secondaires.
- L'étape de *screening* ne fonctionne pas de manière efficace dans Match-Box à cause du filtre statistique du *matching_MB*. Cependant, l'utilisation du *screening_SS* dans Match-Box_SS permet d'obtenir une sensibilité légèrement plus élevée tout en ne perdant pas en sélectivité. Néanmoins, notre approche en un seul cycle *matching_SF-screening_SS* reste plus performante.

IV.5. Conclusions de l'amélioration de Match-Box et évaluation des méthodes d'alignement de séquences

Nous venons de voir comment il était possible d'améliorer les performances d'alignement de Match-Box en modifiant ses deux parties principales que sont le *matching* et le *screening*, et en ne réalisant qu'un seul cycle *matching-screening*. Sur notre banque de 78 alignements de référence, en utilisant le critère ASG, l'utilisation des PSSM a permis d'améliorer les performances de l'étape de *matching* de 9.6%. Le développement et l'utilisation d'un nouvel algorithme de *screening* (*screening_NS*) a permis d'augmenter son efficacité de 1.4% et l'utilisation des structures secondaires (*screening_SS*) a permis d'améliorer son efficacité de 8.4%. La prochaine étape logique serait la combinaison de ces trois améliorations.

Une première évaluation a été réalisée en combinant l'utilisation des PSSM et le nouvel algorithme de *screening* sur la banque de 78 alignements de référence et en utilisant le critère ASG avec des segments de 9 résidus. Celle-ci n'a pas été concluante puisque les performances de l'approche *matching_PSSM-screening_NS* étaient de 61.7% pour la sensibilité et de 69.7% pour la sélectivité. Ces résultats ne correspondent respectivement qu'à une amélioration de 1.0% et 1.6% par rapport à l'approche *matching_SF-screening_NS*, ce qui reste assez faible par rapport au gain de 9.6% du *matching_PSSM*.

Par manque de temps, toutes les combinaisons n'ont pas été testées et il serait peut-être nécessaire d'étudier les performances de ces combinaisons en fonction de la taille des segments utilisés. De plus, d'autres systèmes de score pourraient être utilisés dans le *screening_NS* pour rechercher l'alignement optimum. Néanmoins, il semble que les diverses améliorations que nous pourrions apporter à Match-Box ou à notre stratégie en un seul cycle *matching-screening* ne permettraient que de gagner quelques pourcents en performance. Même si cela reste intéressant, c'est insuffisant pour arriver au même niveau de performance que les meilleurs programmes d'alignement de séquences.

Six ans après l'évaluation de Briffeuil *et al.* sur la sensibilité et la sélectivité de différents programmes d'alignement multiple de séquences, nous avons décidé d'effectuer une mise à jour avec des programmes développés depuis lors. Comme dans la publication Briffeuil *et al.*, nous avons utilisé l'ensemble de 20 alignements de séquences et le critère des RMSD locaux. Les résultats de cette mise à jour sont repris dans la Figure 29. Les versions des programmes Match-Box (Depiereux and Feytmans, 1992), ClustalW (Thompson *et al.*, 1994) et Map (Huang, 1994) sont les

mêmes que dans l'évaluation de Briffeuil *et al.*. Le programme Multalin existait à l'époque mais n'avait pas été évalué, et les programmes Dialign2 (Morgenstern, 1999), SAGA (Notredame and Higgins, 1996), PRRP (Gotoh, 1996), DBClustal (Thompson *et al.*, 2000) et T-COFFEE (Notredame *et al.*, 2000) n'existaient pas il y a six ans. Nous pouvons ainsi constater que peu de progrès ont été réalisés dans le domaine de l'alignement multiple de séquences depuis six ans.

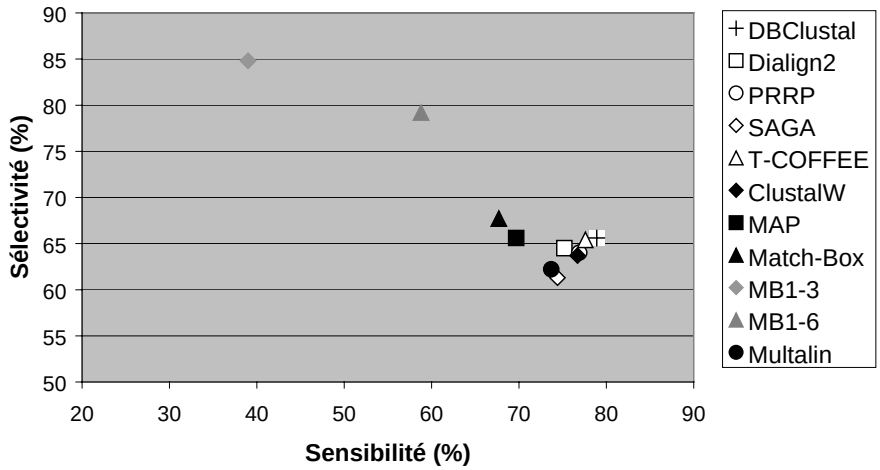


Figure 29: Evaluation de la sensibilité et de la sélectivité de différents programmes d'alignement de séquences sur la banque de 20 alignements de référence, en utilisant le critère des RMSD locaux. Le programme Match-Box est représenté par trois points suivant que l'on regarde uniquement les colonnes ayant un indice de confiance allant de 1 à 3 (MB 1-3), de 1 à 6 (MB 1-6) ou de 1 à 9 (Match-Box, alignement complet).

Nous pourrions conclure de cette évaluation que les méthodes d'alignement de séquences ont atteint leur maximum de performance et qu'il n'y a plus de progrès à réaliser dans ce domaine. Cependant, s'il semble que cela soit vrai pour les méthodes n'utilisant que l'information contenue dans les séquences à aligner, nous allons montrer dans ce travail qu'il est encore possible d'obtenir de meilleures performances en combinant les résultats de plusieurs programmes et en recherchant de l'information supplémentaire dans les banques de données biologiques.

IV.6. Stratégie "consensus"

Dans le but d'obtenir des alignements de séquences de meilleure qualité, beaucoup d'efforts ont été consentis pour améliorer les performances du programme Match-Box. Dans toutes les tentatives, il a seulement été possible d'améliorer très légèrement les performances de ce programme. Entre temps, d'après la littérature et selon nos évaluations, aucune amélioration significative n'a été publiée par d'autres groupes. De plus, aucune méthode d'alignement ne peut être considérée comme la plus fiable. En effet, les évaluations (Briffeuil *et al*., 1998; Thompson *et al*., 1999) ont montré que les performances des programmes d'alignement de séquences dépendent fortement du jeu de séquences à aligner.

Cependant, dans beaucoup de domaines de la bioinformatique, l'utilisation de l'information contenue dans la variabilité des séquences biologiques (Rost and Sander, 1993; Rost and Sander, 1995; Jones, 1999; Jones, 1999) et la combinaison de plusieurs programmes permettent d'augmenter la sensibilité et la sélectivité des méthodes (Cuff *et al*., 1998; Pazos *et al*., 1999; Ginalski *et al*., 2003; Ginalski and Rychlewski, 2003).

Déjà Briffeuil *et al.* en 1998 (Briffeuil *et al*., 1998) avaient montré que la combinaison de plusieurs programmes d'alignement de séquences permettait d'obtenir une meilleure sélectivité. Nous avons avancé l'idée (Lambert *et al*., 2003) qu'il pourrait en être de même pour la sensibilité. Nous avons donc développé un programme d'alignement pairé basé sur la combinaison de résultats de plusieurs autres programmes. C'est faute de temps que nous n'avons pas pu développer, suivant les mêmes principes, un programme d'alignement multiple, mais, telle quelle, notre méthode s'est avérée très utile dans la modélisation par homologie.

Notre méthode, ESyPAli, fonctionne en cinq étapes qui seront détaillées par la suite, mais dont les deux plus importantes, *matching* et *screening*, sont similaires à Match-Box. L'algorithme de *matching* rassemble les résultats de différents programmes d'alignement et l'algorithme de *screening* produit l'alignement pairé final. Cette méthode d'alignement est hybride dans le sens où elle combine, de façon la plus adéquate possible, les résultats de plusieurs autres programmes. Ensuite, les performances de notre approche ont été évaluées et comparées à celles des différents programmes d'alignement utilisés. Enfin, l'étape de *screening* a été améliorée en utilisant un réseau de neurones pour améliorer le choix des positions correctement alignées.

IV.6.1. DESCRIPTION

Le logiciel d'alignement pairé nommé ESyPAli (*Expert System for Pairwise Alignment*) que nous avons mis au point fonctionne en cinq étapes successives (Figure 30).

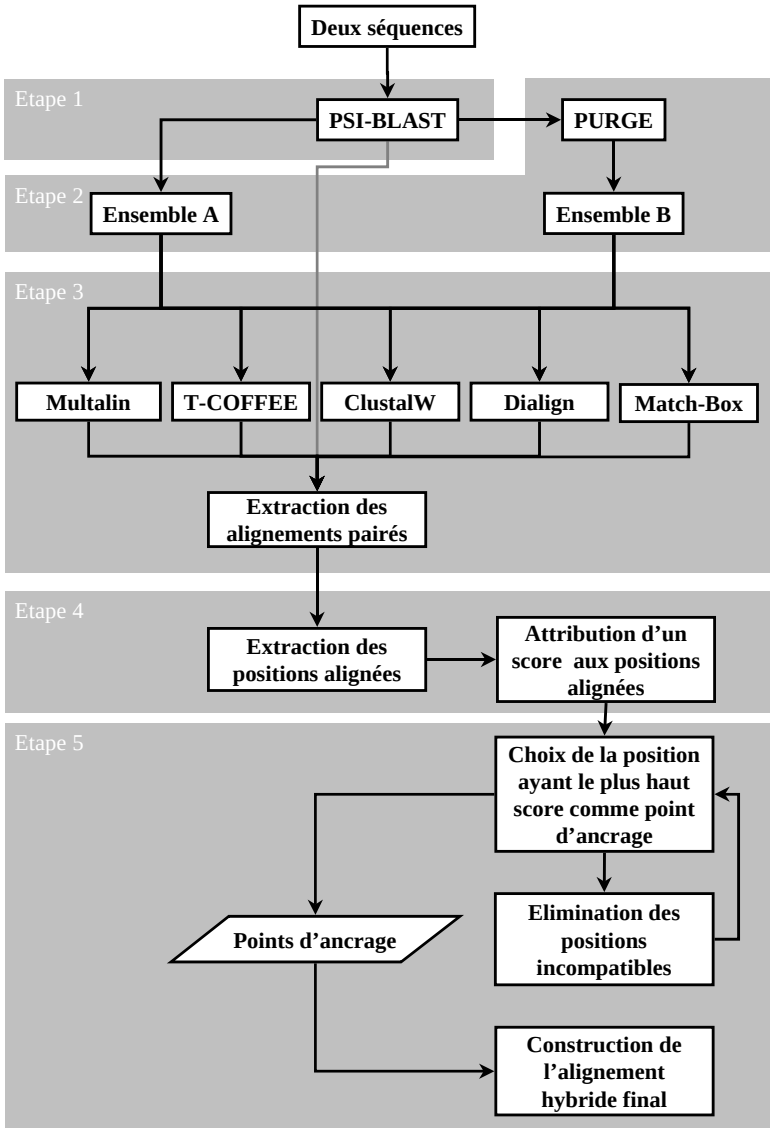


Figure 30: Schéma de fonctionnement de ESyPAli.

La première étape consiste à rechercher des séquences similaires aux deux protéines à aligner. Cette recherche est effectuée par le programme PSI-BLAST (Altschul *et al.*, 1997) qui est exécuté en utilisant la dernière

version disponible de la banque *nr* du NCBI. Le résultat est obtenu après quatre itérations de PSI-BLAST, le seuil de similarité (*E-value*) pour la recherche étant de 0.001.

Dans la deuxième étape, toutes les séquences similaires retrouvées par PSI-BLAST sont récupérées. Suivant le constat de Thompson *et al*. (Thompson *et al*., 1999), la qualité des alignements de séquences dépend très fortement de la composition de l'ensemble de séquences soumis à un alignement multiple. Deux ensembles de séquences (A et B) sont donc générés de manière à créer deux points de départ différents pour exécuter les programmes d'alignement multiple. L'ensemble A contient les 50 séquences les plus similaires (d'après PSI-BLAST) incluant les deux séquences à aligner (le nombre de séquences est limité à 50 pour réduire le temps de calcul). L'ensemble B est un sous-ensemble d'au moins 7 séquences non redondantes incluant les deux séquences à aligner. Il est généré en utilisant le programme PURGE (fourni avec le programme GIBBS (Lawrence *et al*., 1993)) sur l'ensemble des séquences retrouvées par PSI-BLAST. Le score BLAST choisi pour sélectionner ou éliminer les séquences au moyen de PURGE vaut 250. La construction de l'alignement pairé proprement dit est réalisée dans les étapes qui suivent.

Dans la troisième étape appelée *matching*, les deux ensembles de séquences (A et B) sont alignés par cinq programmes d'alignement de séquences qui ont obtenu les meilleurs résultats d'après deux évaluations de la littérature (Briffeuil *et al*., 1998; Thompson *et al*., 1999): ClustalW (Thompson *et al*., 1994), Dialign2 (Morgenstern, 1999), Match-Box (Depiereux *et al*., 1997), Multalin (Corpet, 1988) et T-COFFEE (Notredame *et al*., 2000). Trois programmes performants ont dû être éliminés: MAP (Huang, 1994), SAGA (Notredame and Higgins, 1996) et PRRP (Gotoh, 1996). En effet, le programme MAP n'aligne pas les séquences si elles sont trop peu similaires, et SAGA s'arrête de façon aléatoire en cours d'exécution. Ce problème a été transmis à l'auteur qui nous a conseillé l'utilisation de T-COFFEE. Enfin, PRRP, bien que très performant, demande un temps de calcul difficilement prédictible vu qu'il dépend du nombre d'itérations nécessaires pour atteindre la convergence de son alignement (des temps de calcul de plusieurs jours ont été observés plusieurs fois). Dix alignements multiples sont ainsi générés, chacun incluant les deux séquences à aligner. Les alignements pairés de ces deux séquences sont alors extraits, conduisant à dix alignements pairés différents auxquels est ajouté l'alignement pairé fourni par PSI-BLAST.

Dans la quatrième étape, pour chacun des alignements entre les deux séquences, toutes les positions où un acide aminé de la première séquence et un acide aminé de la seconde se correspondent sont enregistrées dans une banque de données. Ensuite, un score est attribué à chaque position alignée: c'est la fréquence de chaque position alignée dans la banque. Ce score reflète

la confiance de la prédiction. Par exemple, un score de 4 signifie que 4 programmes d'alignement multiple ont prédit la même paire d'acides aminés alignés (Figure 31).

target	...VQADL..IIYLRTSPEVAYERIRQRARSEES..C..VPL..KYLQELHE
Ali_1	LGALPEDR..HIDRLAKRQRPGERLDLAMLAAIR..R..VYGLLANTVRYLQ
Ali_2	...LPGTN..IVLGALPEDRHIDRLAKRQRPGER..L..D.....
Ali_3	...IVLGA..LPEDRHI.....DRLAKRQRPGER..L..DLA...MLAAIRR
Ali_4	...VYVPEPMTYWRVLGASETIANIYTTQHRLDQGEISAGDA..AVVMTSAQ
Ali_5	.....GTN..IVLGALPEDRHIDRLAKRQRPGER.....LDL..AMLAAIRR
Ali_6	....PGTN..IVLGALPEDRHIDRLAKRQRPGERLDL..AML..AAIRRVYG
ESyPAli	LPGTN IVLGALPEDRHIDRLAKRQRPGER L DLA MLAAIRR
Score	12333 33333333333444444444444 2 211 2222222

Figure 31: Exemple de calcul du score pour les positions alignées majoritairement par 6 méthodes d'alignement pairé entre un fragment de la première séquence (*target*: deoxyribonucléotide kinase de *Drosophilia*) et les fragments correspondants de la deuxième séquence (Ali_1 à Ali_6: thymidine kinase de *Herpes Simplex Virus* Type 1). Les acides aminés repris dans l'alignement final sont grisés.

La cinquième et dernière étape, appelée *screening*, est la construction de l'alignement hybride final. La position retrouvée avec la plus haute fréquence est prise d'abord comme point d'ancrage pour construire l'alignement final. Dans le cas où les scores de deux positions sont identiques, une des positions est choisie de manière aléatoire. Les résultats incompatibles, alignant des régions localisées avant et après le point d'ancrage (voir Figure 32), sont effacés de la banque de données. Le processus continue, déterminant de nouveaux points d'ancrage ou éliminant des régions incompatibles, jusqu'à ce que toutes les positions soient sélectionnées ou éliminées. L'alignement final est donc composé des positions les plus fréquemment alignées et compatibles.

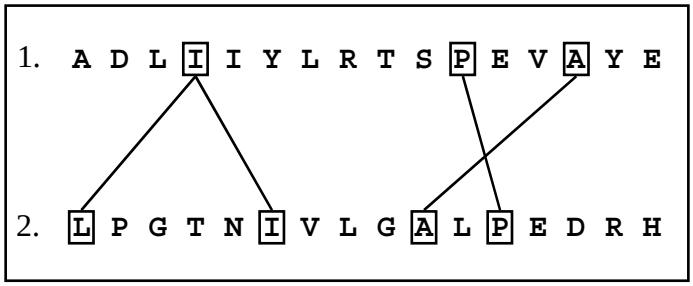


Figure 32: Exemple de résultats compatibles et incompatibles sur deux séquences hypothétiques. Trois cas sont présentés. (a) Les alignements I-I et I-L ne sont pas compatibles car le même acide aminé dans la séquence 1 est aligné à deux acides aminés dans la séquence 2. (b) Les alignements P-P et A-A ne sont pas compatibles. P dans la séquence 1 est à la droite de A mais P dans la seconde séquence est à la gauche de A. (c) Les alignements I-I et P-P sont compatibles. Les prolines (P) sont toutes les deux à droite des isoleucines (I).

IV.6.2. EVALUATION DES PERFORMANCES DE ESYPALI

Les performances de ESyPAli ont été évaluées sur l'ensemble de 420 alignements de référence (quatrième ensemble) en utilisant les critères ASG (Tableau 13) et CSS (Tableau 14). Ces performances ont également été comparées aux performances de chacun des programmes intervenant dans la construction de l'alignement hybride final.

Les performances ont été mesurées en prenant en compte toutes les colonnes alignées par les différents programmes d'alignement. Cependant, nous n'avons pas tenu compte du fait que le programme Match-Box n'aligne pas tous les acides aminés des séquences soumises. Ainsi, nous aurons les deux relations suivantes:

$$n_{corr}E \approx n_{corr} \qquad \text{et} \qquad n_{col}E > n_{col}$$

où $n_{col}E$ est le nombre de colonnes alignées par Match-Box (alignement total, tel que considéré dans notre évaluation)

n_{col} est le nombre de colonnes prédites comme alignées par Match-Box (indice de confiance compris entre 1 et 9 inclus)

$n_{corr}E$ est le nombre de colonnes correctement alignées, parmi les $n_{col}E$ colonnes

n_{corr} est le nombre de colonnes correctement alignées, parmi les n_{col} colonnes

Ces deux relations impliquent que la sélectivité du programme Match-Box, mesurée lors de l'évaluation ($n_{corr}E/n_{col}E$), sera inférieure à la

sélectivité réelle du programme (n_{corr}/n_{col}). De plus, nous verrons plus loin que l'évaluation de la sélectivité est problématique.

Nous pouvons observer dans le Tableau 13 et le Tableau 14 que, quel que soit le critère de vérité considéré, la sensibilité de notre approche, ESyPAli, est meilleure que celle des autres programmes. Pour la banque de 420 alignements pairés de référence, la sensibilité a été améliorée de 3% environ (critère ASG et CSS) par rapport au meilleur programme d'alignement qui ressort de notre évaluation: T-COFFEE. Le programme Match-Box a une sensibilité plus faible d'environ 10% par rapport aux autres programmes d'alignement, mais il est surtout caractérisé par la confiance des résultats proposés (que nous savons sous-estimée dans notre évaluation).

Nous pouvons également mesurer l'efficacité de l'étape de *screening* (étape 5), c'est-à-dire, tout comme dans le chapitre précédent, le rapport entre la sensibilité de l'approche complète (ESyPALi) et celle de l'étape de *matching*. En d'autres termes, c'est la capacité du *screening* à retrouver les positions correctement alignées prédites par l'étape de *matching*. Imaginons un instant que notre procédure de *screening* soit parfaite et aille, à chaque fois, retrouver la position correctement alignée parmi celles proposées par les différents programmes d'alignement de séquences. La sensibilité de cette approche pourrait être prise comme un maximum que nous ne pourrions pas dépasser par manque d'information dans les alignements utilisés pour la construction de l'alignement hybride. La sensibilité maximum obtenue serait de 92,7% pour le critère ASG et de 95,0% pour le critère CSS. Les sensibilités obtenues par ESyPAli, 88,7% et 92,2% respectivement selon les critères ASG et CSS, se rapprochent donc de 3 à 4% de la limite « théorique » du *screening*.

L'interprétation des résultats concernant la sélectivité pose un problème pour l'alignement pairé de séquences. En effet, si on veut aligner deux séquences, le nombre de positions alignées diminuera avec l'augmentation du nombre des *gaps* dans l'alignement. On arrive donc à un paradoxe dans notre évaluation, où le programme d'alignement le plus sélectif (PSI-BLAST) est celui qui place le plus de *gaps* dans l'alignement (ou qui propose le moins de positions alignées). Malgré ces problèmes d'interprétation, ESyPAli obtient une sélectivité classée en troisième position sur les 7 programmes évalués, derrière PSI-BLAST et T-COFFEE.

Tableau 13: Performances de ESyPAli et des six programmes utilisés pour la construction de l'alignement hybride, calculées avec le critère ASG. Le nombre de positions alignées par chaque programme est également repris.

	Sensibilité	Sélectivité	Pos. Ali.
Référence	100%	100%	94583
ClustalW	83,6%	83,6%	94669
Dialign	83,7%	87,2%	90731
Match-Box	71,9%	78,6%	86539
Multalin	80,5%	88,1%	86418
PSI-BLAST	82,0%	93,2%	83210
T-COFFEE	85,6%	91,4%	88526
ESyPAli	88,7%	88,8%	94441

Tableau 14: Performances de ESyPAli et des six programmes utilisés pour la construction de l'alignement hybride, calculées avec le critère CSS. Le nombre de positions alignées par chaque programme est également repris.

	Sensibilité	Sélectivité	Pos. Ali.
Référence	100%	100%	49984
ClustalW	87,5%	46,2%	94669
Dialign	87,5%	48,2%	90731
Match-Box	74,3%	42,9%	86539
Multalin	84,2%	48,7%	86418
PSI-BLAST	85,3%	51,3%	83210
T-COFFEE	89,4%	50,5%	88526
ESyPAli	92,2%	48,8%	94441

En combinant les résultats de différents programmes d'alignement de séquences, la sensibilité a pu être augmentée de 3% par rapport à T-COFFEE pour une perte de 1,7 à 2,6% en sélectivité, sur notre banque de 420 alignements de séquences de référence. Pour tenter d'améliorer encore la sensibilité de notre approche, il nous a semblé judicieux de changer le système de score utilisé dans ESyPAli: c'est l'objet de la section suivante.

IV.6.3. DÉVELOPPEMENT D'UN RÉSEAU DE NEURONES

Nous avons montré ci-dessus que les programmes d'alignement de séquences ont des sensibilités différentes, ce qui est conforme aux résultats des différentes évaluations décrites dans la littérature (Briffeuil *et al*., 1998; Thompson *et al*., 1999). Néanmoins, pour ESyPAli, chaque programme a un poids identique dans la construction de l'alignement pairé final.

Nous avons décidé de développer un second programme d'alignement, où le score de chaque position alignée serait une somme de scores attribués par le réseau neuronal à chaque paire d'acides aminés alignés par chaque méthode d'alignement. Ce nouveau programme a été nommé ESyPAliNN (NN pour *Neural Networks*).

Pour entraîner le réseau neuronal de ESyPAliNN, il était nécessaire de posséder une banque de données de positions correctement alignées et de positions prédites par différents programmes d'alignement de séquences. Les positions correctement alignées ont été extraites de notre banque de 420 alignements de référence et les positions prédites ont été extraites des alignements pairés fournis par chaque programme pour chacune des 420 paires de séquences de la banque.

Comme il a été dit dans la section de l'introduction traitant des réseaux neuronaux (page 42), l'entraînement par validation croisée sur 6 parties a été utilisé. Les 420 alignements de référence ont été divisés en 6 ensembles de 70 séquences. Le réseau a été entraîné sur 4 parties, l'évolution des performances a été contrôlée sur la cinquième et les performances ont été évaluées sur la dernière. Ces opérations ont été effectuées 6 fois en permutant à chaque fois les ensembles d'entraînement, de test et d'évaluation. Six réseaux ont été ainsi obtenus. Ils ont tous été utilisés pour effectuer le calcul des scores de chaque paire d'acides aminés alignés, le résultat final étant la moyenne des résultats des 6 réseaux.

Dans ce réseau neuronal, le codage des entrées se fait en deux étapes. Dans la première, pour chaque acide aminé de la première séquence, les acides aminés de la deuxième séquence, alignés par les différents programmes, sont remplacés par leur position absolue. En cas de *gap*, la position indiquée est -1. Pour le cas fictif du Tableau 15a, on obtient une table similaire au Tableau 15b.

Dans la seconde étape (Tableau 15c), une matrice triangulaire composée uniquement de 0 et de 1 est construite de la manière suivante pour chaque acide aminé de la première séquence: les positions proposées par chaque programme sont comparées à celles proposées par tous les autres programmes. Si la position prédite par les deux programmes d'alignement est la même, la cellule vaut 1, sinon elle vaut 0. Si l'acide aminé de la

première séquence est aligné à un *gap*, le score attribué vaut 0. Par exemple, la matrice triangulaire correspondant à la troisième ligne du Tableau 15b est donnée dans le Tableau 15c.

Tableau 15: a) Propositions d'alignement de la deuxième séquence par les différents programmes. b) Première étape du codage des entrées (voir explications dans le texte). c) Deuxième étape du codage des entrées pour l'acide aminé Y (voir explications dans le texte). CL: ClustalW, DI: Dialign2, MB: Match-Box, MU: Multalin, PB: PSI-BLAST et TC: T-COFFEE.

a.

		Acides aminés dans la séquence 2											
		Ensemble A						Ensemble B					
		CL	DI	MB	MU	PB	TC	CL	DI	MB	MU	PB	TC
Acides aminés de la séquence 1	A	L	V	L	G	A	L	A	L	A	A	A	L
	G	A	L	A	V	W	A	W	A	G	G	W	A
	Y	W	-	W	F	-	W	-	W	L	L	W	W
	W	W	A	W	Y	W	W	W	W	A	A	-	W

b.

		Positions des acides aminés dans la séquence 2											
		Ensemble A						Ensemble B					
		CL	DI	MB	MU	PB	TC	CL	DI	MB	MU	PB	TC
Acides aminés de la séquence 1	A	5	4	5	10	6	5	6	5	3	3	6	5
	G	6	5	6	11	7	6	7	6	4	4	7	6
	Y	7	-1	7	12	-1	7	-1	7	5	5	8	7
	W	8	6	8	13	8	8	8	8	6	6	-1	8

c.

		Ensemble A						Ensemble B					
		CL	DI	MB	MU	PB	TC	CL	DI	MB	MU	PB	TC
Ensemble A	CL	1											
	DI	0	0										
	MB	1	0	1									
	MU	0	0	0	1								
	PB	0	0	0	0	0							
	TC	1	0	1	0	0	1						
Ensemble B	CL	0	0	0	0	0	0	0					
	DI	1	0	1	0	0	1	0	1				
	MB	0	0	0	0	0	0	0	0	1			
	MU	0	0	0	0	0	0	0	0	1	1		
	PB	0	0	0	0	0	0	0	0	0	0	1	
	TC	1	0	1	0	0	1	0	1	0	0	0	1

Chaque cellule de la matrice triangulaire correspond ainsi à un neurone d'entrée. Le nombre de neurones d'entrée est donc de 78 (12*(12+1)/2), le nombre des neurones de sortie est de 12 puisque nous désirons attribuer un score à chaque prédiction proposée par chaque programme d'alignement. La topologie de notre réseau neuronal est celle

décrite dans la section traitant des réseaux neuronaux (page 42). Le résultat des prédictions du réseau neuronal pour notre exemple du Tableau 15a est repris dans le Tableau 16a. Le score final de chaque proposition est la somme des scores attribués par le réseau neuronal. Pour notre exemple du Tableau 15a, les scores des différentes propositions sont repris dans le Tableau 16b.

Après différents essais, nous avons pu déterminer que le nombre de neurones cachés optimum était de 99. Avec un nombre de neurones inférieur ou supérieur dans la couche cachée, la sensibilité de ESyPAliNN obtenue en utilisant ces réseaux neuronaux était plus faible.

Tableau 16: Obtention d'un score pour les prédictions des programmes d'alignement de séquences. a) Résultats du réseau neuronal. b) Score final des trois meilleures propositions. c) Alignement entre les deux séquences fictives, le fragment étudié dans l'exemple est grisé. CL: ClustalW, DI: Dialign2, MB: Match-Box, MU: Multalin, PB: PSI-BLAST et TC: T-COFFEE. Pos.: Position de l'a.a. de la deuxième séquence.

a.

		Scores des différentes prédictions attribuées par le réseau neuronal											
		Ensemble A						Ensemble B					
		CL	DI	MB	MU	PB	TC	CL	DI	MB	MU	PB	TC
Acides aminés de la séquence 1	A	0,8	0,2	0,7	0,1	0,4	0,9	0,5	0,5	0,4	0,3	0,4	1,0
	G	0,7	0,2	0,8	0,1	0,8	1,0	0,6	0,7	0,3	0,4	0,8	0,9
	Y	0,9	0,0	0,9	0,0	0,0	1,0	0,0	0,7	0,2	0,2	0,1	1,0
	W	1,0	0,4	0,9	0,1	0,8	1,0	0,9	0,9	0,5	0,1	0,0	1,0

b.

		Scores des trois meilleures propositions					
		Pos./a.a.	Score	Pos./a.a.	Score	Pos./a.a.	Score
Acides aminés de la séquence 1	A	5 / L	3,9	6 / A	1,3	3 / A	0,7
	G	6 / A	4,1	7 / W	2,2	4 / G	0,7
	Y	7 / W	4,5	5 / L	0,4	8 / W	0,1
	W	8 / W	6,5	6 / A	1,0	13 / Y	0,1

c.

Séquence 1	…	A	G	A	G	Y	W	T	-	I	Y	Y	…
Séquence 2	…	A	G	L	A	W	W	S	G	V	F	W	…
Scores	…	5,2	4,8	3,9	4,1	4,5	6,5	5,1		4,2	7,6	6,8	…

IV.6.4. PERFORMANCES DE ESYPALINN

Les sensibilités de ESyPAliNN pour les 6 ensembles de 70 séquences, après l'entraînement des réseaux neuronaux, sont reprises dans le Tableau 17 pour les critères ASG et CSS. Nous y avons également rappelé,

pour chaque ensemble de référence, la sensibilité limite expliquée dans la section consacrée à l'évaluation des performances de ESyPAli. La sensibilité totale ainsi que la sélectivité sont également présentées.

Tableau 17: Performances de ESyPAliNN sur les 6 ensembles de d'évaluation.

	ASG		CSS	
	Sensibilité	**Limite**	**Sensibilité**	**Limite**
Ensemble 1	88,6%	89,8%	93,3%	92,8%
Ensemble 2	84,1%	88,6%	86,7%	90,5%
Ensemble 3	91,6%	94,5%	93,8%	95,8%
Ensemble 4	91,4%	95,0%	95,7%	97,4%
Ensemble 5	89,7%	93,6%	93,4%	95,6%
Ensemble 6	92,0%	94,7%	95,6%	97,3%
Total	89,6%	92,7%	93,2%	95,0%
Sélectivité	90,0%		49,4%	

On observe dans le Tableau 17 que la sensibilité obtenue par ESyPAliNN est proche de la limite de sensibilité que nous pourrions obtenir. Néanmoins, on peut également observer pour l'ensemble 1 et pour le critère CSS, que la sensibilité de notre méthode dépasse la sensibilité limite telle que nous l'avions définie. Ce comportement est dû simplement à la génération de positions correctement alignées après la sélection des positions ayant un score maximum. En effet, après avoir sélectionné ces positions, l'alignement n'est généralement pas complet puisque la procédure de *screening* élimine les propositions des différents programmes qui sont incompatibles avec le choix opéré: il est donc nécessaire de compléter l'alignement. Cette opération est effectuée en ajoutant les acides aminés qui manquent dans chacune des séquences, en essayant d'aligner, tant que possible, un acide aminé de la première séquence à un acide aminé de la deuxième séquence. La Figure 33 contient un exemple permettant d'expliquer en partie ce qui vient d'être énoncé.

1) Alignement correct	2) Différentes propositions	3) Construction initiale
séq 1 ASD	séq 1 --A-S-D--	ASD
séq 2 VTE	p1 --V----TE	V-E
	p2 VT----E--	
	p3 --V---TE-	**4) Ajout de l'acide aminé manquant**
	p4 -VT---E--	ASD
	cons --V---E--	VTE
	score --2---2--	

Figure 33: Exemple de génération de positions correctement alignées après l'exécution du *screening*. (1) Alignement correct de deux séquences (séq 1 et séq 2). (2) Propositions de quatre programmes d'alignement (p1, p2, p3 et p4). Le système de score choisi pour les positions alignées est ici leur fréquence. L'alignement consensus final (cons) ne contient que deux positions avec le score maximum qui est 2. (3) Construction initiale de l'alignement final, à partit du consensus des différents programmes. (4) Lors de la construction de l'alignement final après sélection des meilleures solutions, la séquence 2 est complétée et la thréonine est alignée à la sérine.

Si on compare le Tableau 13, le Tableau 14 et le Tableau 17, on remarque que l'utilisation des réseaux neuronaux a permis une amélioration d'environ 1% par rapport à ESyPAli, aussi bien pour la sensibilité que pour la sélectivité, quel que soit le critère de référence utilisé.

IV.6.5. CONCLUSIONS

Un programme d'alignement pairé de séquences (ESyPAli) a été développé puis amélioré en utilisant des réseaux neuronaux (ESyPAliNN). Sa sensibilité est meilleure que celles des programmes les plus performants évalués et sa sélectivité est proche de celles des programmes ayant la plus haute sélectivité.

Nous avons montré que notre méthode ne parvenait pas à retrouver toutes les positions correctement alignées par les différents programmes, mais en revanche, elle est capable de générer d'elle-même des positions correctement alignées supplémentaires.

Les performances de notre méthode sont principalement limitées par celles des programmes utilisés dans la partie *matching*. A l'avenir, toute amélioration de l'un de ces programmes ou l'incorporation d'autres

programmes améliorerait la qualité finale des alignements de ESyPAli et ESyPAliNN.

Les performances de nos deux programmes pourraient être encore augmentées par l'utilisation d'un meilleur système de score pour les positions alignées. Nous pourrions aussi, par exemple, analyser les alignements par fenêtre d'acides aminés au lieu de le faire position par position, comme actuellement. Enfin, une méthode d'alignement utilisant la programmation dynamique pourrait être utilisée pour compléter l'alignement final provenant de la procédure de *screening*.

IV.7. Application de ESyPAliNN à la modélisation de la monoamine oxydase A humaine (MAO A)

Publication présentée (voir Annexe 9):

N. Léonard, C. Lambert, E. Depiereux and J. Wouters
Modeling of Human Monoamine Oxidase A: From Low Resolution Threading Models to Accurate Comparative Models Based on Crystal Structures
NeuroToxicology in press

Les monoamine oxydases A et B (MAO A et B) sont des cibles d'intérêt pharmacologique étant donné leur rôle dans le métabolisme de neurotransmetteurs comme la sérotonine et la dopamine. Une activité altérée des deux MAOs a été trouvée dans de nombreuses maladies neuro-psychiatriques et les inhibiteurs de ces enzymes sont utilisés dans le traitement de plusieurs maladies. En particulier, les inhibiteurs de la MAO A sont utilisés comme anti-dépresseurs et ceux de la MAO B sont utilisés dans le traitement de la maladie de Parkinson.

L'intérêt de la détermination de la structure 3D de la MAO A et B est dès lors évident puisqu'elle permettrait une meilleure connaissance de la structure globale de la protéine. La connaissance de la structure du site actif serait particulièrement utile faciliter la conception de nouveaux inhibiteurs puissants et sélectifs.

La modélisation de la structure de la MAO A par homologie a été réalisée en deux temps:

1. Avant la détermination de la structure 3D de la MAO B (publiée en 2002), nous avions construit un modèle de la MAO A que nous avons soumis à la PDB (PDB ID: 1h8q). Ce modèle a été généré à partir des alignements réalisés avec ESyPAli entre la séquence de chaque MAO et deux *templates*: la polyamine oxydase de *Zea mays* (PAO) et la L-amino acid oxydase de *Calloselasma rhodostoma* (LAAO). Le pourcentage d'identité entre la MAO A et les séquences des PAO et LAAO était faible (19,4% et 20,0%). Une technique spécifique a donc dû être utilisée, en combinant les alignements avec les deux *templates* et en corrigeant l'alignement sur base des données de structure secondaire prédite.
2. Suite à la détermination de la structure 3D de la MAO B, nous avons pu modéliser la MAO A de manière plus précise en utilisant la structure de la MAO B comme *template* (plus de 70% d'identité). L'alignement de séquences fut réalisé grâce au programme ESyPAliNN.

Avant la détermination de la structure cristallographique de la MAO B, aucune protéine de structure connue ne présentait une homologie de séquence élevée avec la MAO A, empêchant ainsi une modélisation par homologie aisée de cette enzyme. Seules deux flavoprotéines (la polyamine oxydase et la L-amino acid oxydase) présentaient une faible homologie de séquence avec la MAO A.

Malgré le faible pourcentage d'identité entre la MAO A et ses *templates* (moins de 20%), un premier modèle a été construit en utilisant une méthodologie originale combinant des alignements de séquences et de structures et des prédictions de structures secondaires. Ces modèles ont permis de mettre de nouvelles caractéristiques structurales en évidence: en particulier, l'existence possible d'un sandwich d'acides aminés aromatiques formé par deux tyrosines localisées près du noyau flavine et formant une partie du site actif. L'implication possible de la lysine en position 335 pendant la catalyse a également été révélée par ce modèle. En raison du faible pourcentage d'identité entre la MAO A et ses *templates*, certaines régions du modèle étaient modélisées moins précisément. Ainsi, les caractéristiques générales des flavoprotéines sont conservées et bien représentées dans le modèle, alors que des caractéristiques spécifiques à la MAO A n'ont pas pu être bien prédites.

L'obtention de la structure cristalline de la MAO B humaine a permis une modélisation précise de la structure de la MAO A. Ce dernier modèle a été comparé au précédent et les différences ont été discutées. Ainsi, les différences principales se retrouvent au niveau des boucles ou des zones plus flexibles présentes en surface de la protéine. Certaines différences ont également été observées dans le site catalytique. Ces modèles de la MAO A ouvrent des perspectives intéressantes pour la conception rationnelle d'inhibiteurs ayant un intérêt thérapeutique potentiel.

IV.8. Conclusions et perspectives

Nous avons tenté de réaliser un logiciel d'alignement de séquences très performant en combinant deux méthodes performantes, Match-Box et ClustalW. Cette expérience a montré que la simple combinaison des résultats de ces deux programmes ne pouvait pas améliorer la qualité de l'alignement multiple de séquences. Ensuite, nous avons essayé d'améliorer toutes les étapes d'alignement du programme Match-Box en utilisant une meilleure matrice de scores, en utilisant des matrices de scores spécifiques de l'environnement, en utilisant des matrices de scores spécifiques de la position, en utilisant les prédictions de structure secondaire et en développant un nouvel algorithme de *screening*. Ces améliorations ont chaque fois, au mieux, augmenté légèrement la qualité de l'alignement de séquences final.

Après avoir constaté qu'une des limites principales de Match-Box était de ne pas retrouver un nombre suffisant de positions correctement alignées lors de son étape de *matching*, nous avons décidé de combiner les résultats de différents logiciels; ce qui a conduit à la réalisation de ESyPAli. Ce programme a ensuite été amélioré en utilisant des réseaux neuronaux, ce qui a donné ESyPAliNN qui est plus performant que le meilleur programme utilisé pour réaliser un alignement pairé et sera bientôt disponible sous la forme d'un serveur web. La seule limite de notre approche est le temps de calcul supérieur de trois ordres de grandeur par rapport à un algorithme rapide (une seconde pour ALIGN (Myers and Miller, 1989) contre une heure pour ESyPAliNN). Cette limite exclut son utilisation pour la recherche par similarité dans des banques de données de séquences.

Ce programme d'alignement pairé pourrait encore être amélioré en utilisant de meilleures méthodes d'alignement de séquences dans son étape de *matching* ou en développant une meilleure fonction pour attribuer un score aux diverses positions alignées prédites.

A l'heure actuelle, le développement d'un programme d'alignement multiple utilisant la même technique est compliqué. Deux voies peuvent être envisagées:

- la première, et la plus simple, consiste à aligner toutes les séquences par rapport à une seule qui sert de guide pour construire l'alignement multiple final. Cette technique est utilisée, par exemple, dans PSI-BLAST et dans les programmes utilisant les HMMs (où on aligne toutes les séquences au HMM),
- la seconde consiste à combiner toutes les positions pairées alignées et à rechercher l'alignement qui maximise une fonction qui mesure la qualité intrinsèque de l'alignement. De telles fonctions ont déjà été développées:

COFFEE (Notredame *et al*., 1998) ou NORMD (Thompson *et al*., 2001). Un algorithme génétique pourrait être utilisé pour combiner au mieux les résultats des différents programmes.

Lorsque le programme d'alignement multiple sera développé, d'autres applications pourraient en bénéficier: la phylogénie, la prédiction de structure secondaire, la prédiction d'accessibilité au solvant et toutes les méthodes se basant sur un alignement le plus fiable possible, sous réserve qu'elles puissent accepter un temps calcul plus important.

Toutefois, dans son état actuel, ESyPAliNN peut être utilisé avec fruit en modélisation par homologie, dont l'étape critique est l'alignement de séquences. Nous l'avons d'ailleurs utilisé pour modéliser avec succès une protéine d'intérêt pharmacologique, la MAO A. L'utilisation de notre méthode d'alignement pairé pour la modélisation par homologie sera détaillée dans le chapitre suivant.

V. Développement d'un serveur de prédiction de structure protéique par homologie: ESyPred3D

Comme indiqué dans l'introduction, l'étape la plus critique de la modélisation par homologie est l'alignement entre la séquence cible et le(s) *template*(*s*). Dans le chapitre précédent, nous avons décrit deux méthodes (ESyPAli et ESyPAliNN) qui calculent des alignements pairés de très bonne qualité. Nous avons utilisé ces deux méthodes dans la réalisation d'un serveur de modélisation par homologie, nommé ESyPred3D. Dans ce chapitre, nous allons donc:

1. décrire les principes de fonctionnement de ce serveur de modélisation,
2. présenter des évaluations des performances de ce serveur quand il utilise la méthode d'alignement ESyPAli,
3. présenter d'autres évaluations de performance quand la méthode d'alignement ESyPAliNN est employée,
4. discuter des performances observées et proposer des améliorations possibles de notre serveur.

V.1. Description

Comme indiqué dans l'introduction, la modélisation par homologie se déroule en quatre étapes (voir page 51):

1. Recherche en banque de données d'une séquence de structure connue (SSC) similaire à la séquence d'intérêt (SI).
2. Alignement entre SI et SSC.
3. Attribution des coordonnées 3D des résidus de la SSC à ceux de la SI dans les régions alignées et élaboration d'une structure plausible pour les régions non alignées: un modèle 3D de la SI est ainsi obtenu.
4. Evaluation du modèle 3D sur base de critères géométriques et énergétiques.

Pour développer notre serveur de modélisation, nous avons tenté d'utiliser le meilleur logiciel permettant de réaliser chacune d'entre elles.

Pour la première étape, l'analyse de la littérature (Altschul *et al*., 1997; Park *et al*., 1998; Sauder *et al*., 2000) indique que le programme SAM-T99 (Hughey and Krogh, 1996) retrouve plus de séquences homologues (sensibilité élevée) que les autres programmes de recherche en banque de données de séquences: PSI-BLAST (Altschul *et al*., 1997), FASTA (Pearson and Lipman, 1988; Pearson, 1990), BLAST2 (Altschul *et al*., 1997), Bic_SW (Smith and Waterman, 1981) et ISS (Park *et al*., 1997). Cependant, pour la modélisation par homologie, PSI-BLAST a une sensibilité proche de SAM-T99 (Shi *et al*., 2001) et, d'après nos tests, il produit moins de faux positifs que SAM-T99. De plus, il est considéré par certains auteurs (Dunbrack, 1999) comme le meilleur programme d'alignement pour la modélisation par homologie. Ces trois arguments nous ont fait préférer PSI-BLAST pour le développement de notre serveur.

Rappelons (voir page 18) que PSI-BLAST construit, par recherche itérative, dans une banque de données (ici *nr*, décrite page 52), un profil spécifique de la famille de séquences à laquelle appartient la séquence d'intérêt. Cette technique itérative permet de détecter des similarités faibles entre séquences. Comme la base de données *nr* contient la banque PDB, on peut espérer détecter, avec PSI-BLAST, une similarité faible entre la SI et une SSC. La SSC ayant l'*E-value* la plus faible est utilisée comme *template* pour la modélisation.

L'étape d'alignement de séquences est effectuée par l'une de nos deux techniques ESyPAli ou ESyPAliNN (suivant le choix de l'utilisateur).

L'alignement PSI-BLAST entre SI et SSC utilisé par celles-ci provient de la recherche dans la banque de données *nr* décrite ci-dessus.

Le calcul d'un modèle de la structure 3D à partir d'alignements pairés entre la SI et la SSC peut être réalisé par différents programmes (MODELLER (Sali and Blundell, 1993), COMPOSER (Sutcliffe *et al.*, 1987; Sutcliffe *et al.*, 1987), ProMod (Peitsch, 1996)). Nous avons choisi l'assignation des coordonnées et la modélisation des boucles effectuées par le programme MODELLER qui a donné les meilleurs résultats aux CASP3 (Burke *et al.*, 1999) , CASP4 (Venclovas, 2001) et CASP5 (Proteins Suppl *in press*).

Pour l'évaluation finale de la structure prédite, deux groupes de méthodes peuvent être utilisés (voir page 60): des méthodes de vérification géométrique et des méthodes de vérification énergétique. Wilson *et al.* (Wilson *et al.*, 1998) ont comparé plusieurs programmes de vérification géométrique des modèles sans vraiment déterminer le meilleur d'entre eux. En conséquence, nous avons choisi arbitrairement d'utiliser le programme Procheck (Laskowski *et al.*, 1993). Bien qu'il soit judicieux d'effectuer un test énergétique, nous n'en ferons pas pour le moment, car nous n'avons pas pu trouver de publication comparant les méthodes de vérification énergétique existantes. Néanmoins, cette lacune devrait être comblée à l'avenir par l'utilisation du programme Verify3D (Luthy *et al.*, 1992) ou du programme ANOLEA (Melo and Feytmans, 1997).

La méthode décrite ci-dessus a été nommée ESyPred3D (***E**xpert **Sy**stem for **Pred**icting **3D** protein structures*). Son schéma de fonctionnement est repris dans la Figure 34. Les différents programmes développés pour ESyPred3D ont été écrits en C++ (traitement des séquences et calculs d'alignement) et PERL (script général pour mettre en marche les différents programmes et gérer le serveur web) et ont été compilés pour IRIX 6.5, RedHat Linux 7.3 et MacOSX 10.2.

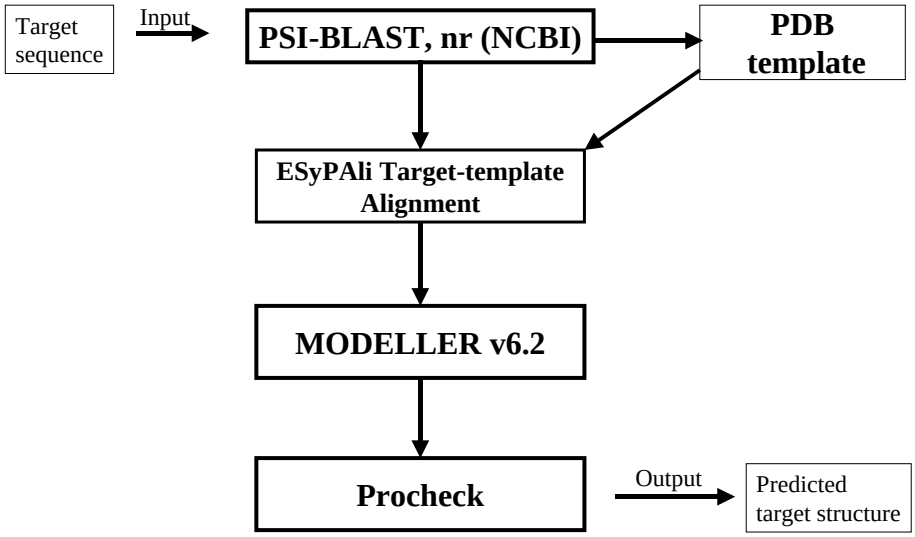


Figure 34: Schéma de fonctionnement de ESyPred3D.

Un site web a été créé et est accessible aux chercheurs à l'adresse http://www.sciences.fundp.ac.be/urbm/bioinfo/esypred. Les calculs de ce serveur de modélisation sont effectués, actuellement, sur un cluster de PC Priminfo (description technique voir page 75). En cas de panne de ce cluster, trois stations SGI Octane (description technique voir page 75) peuvent prendre le relais. Pour soumettre une requête au serveur web, l'utilisateur effectue un "copier-coller" de la séquence d'intérêt dans le cadre adéquat de la page d'accueil. Après avoir cliqué sur "*submit*", la séquence et les différents paramètres choisis par l'utilisateur sont envoyés au serveur de calcul qui renvoie le modèle 3D de la SI par courrier électronique à l'utilisateur.

Actuellement, le nombre de modélisations effectuées est en moyenne de plus de 1500 par mois avec des pics de plus de 150 modélisations par jour (voir Figure 35).

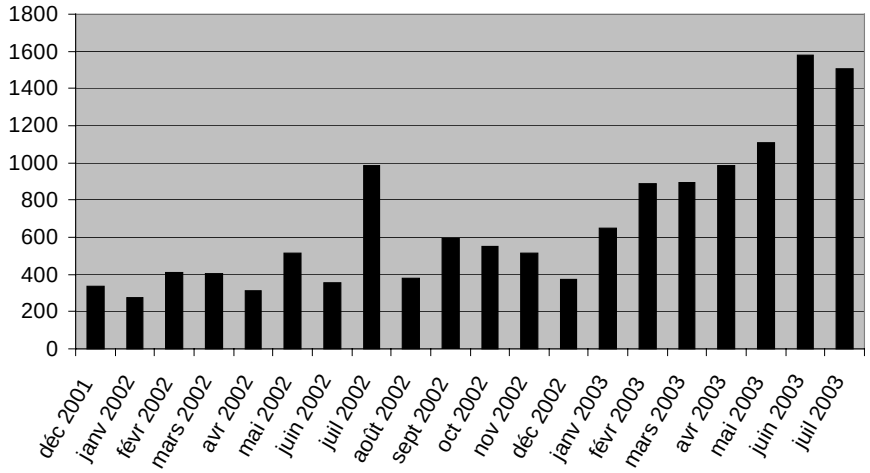


Figure 35: Nombre de modélisations par mois réalisées par le serveur ESyPred3D depuis sa mise en accès public en décembre 2001.

V.2. Performances de ESyPred3D

V.2.1. EVALUATION DE ESYPRED3D UTILISANT ESYPALI SUR 9 PROTÉINES-TEST

L'objectif de ESyPred3D étant la modélisation fiable de protéines partageant un faible pourcentage d'identité avec une SSC, la sélection des séquences à modéliser pour tester le logiciel a été réalisée de la manière suivante:

1. Alignement de toutes les séquences de la banque PDB (version de janvier 2000) deux à deux par le programme ALIGN (Myers and Miller, 1989) (Gap penalty -12/-2, matrice de scores BLOSUM50).
2. Sélection des séquences partageant un pourcentage d'identité de 30% maximum avec toutes les autres séquences, soit 715 séquences. Nous avons limité la sélection à des séquences de plus de 120 résidus, contenant un domaine fonctionnel complet. Dans la mesure où 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, nous ne voulions pas aborder la problématique de la prédiction de la position relative des domaines structuraux. Ces critères réduisent la sélection à 541 séquences.
3. Pour des raisons de temps calcul et pour pouvoir analyser en détail toutes les modélisations, nous avons choisi 9 protéines au hasard parmi cette sélection. Leurs caractéristiques sont reprises dans le Tableau 18 sous leur identifiant dans la banque de données PDB.

Tableau 18: Caractéristiques des 9 structures cristallographiques dont les séquences ont été utilisées pour tester les performances de ESyPred3D. Les caractéristiques des *templates* sélectionnés sont reprises en dessous de chaque protéine test.

PDB ID	Rés. (Å)	R free	Fonction	Taille (a.a.)
1AMY	2,8	0,153	α-1,4 glucan-4-glucanhydrolase (α-amylase) de *Hordeum vulgare*	403
template: 1JAE	1,65	0,206	alpha-amylase de *Tenebrio molitor*	471
1BMT A	3,0	0,170	Méthionine synthase (domaines se liant à la vitamine B12) de *Escherischia coli*	246

template: 5REQ A	2,2	0,292	methylmalonyl-coa mutase de *Propionibacterium freudenreichii shermanii*	727
1D2F	2,5	0,201	aminotransférase probable, enzyme qui dégrade l'inducteur du système du maltose chez *Escherischia coli*	390
template: 1BKG A	2,6	0,336	aspartate aminotransferase de *Thermus thermophilus*	385
1DUP A	1,9	0,150	déoxyuridine 5'-triphosphate nucléotide synthase (DUTPase) de *Escherischia coli*	152
template: 1DUT A	1,9	0,249	déoxyuridine 5'-triphosphate pyrophosphatase du virus d'immunodéficience féline	133
1LXA	2,6	0,180	ADP N-acétyl glucosamine acétyltransférase de *Pseudomonas aeruginosa*	262
template: 1XAT	3,2	0,259	Chloramphénicol acétyltransférase de *Pseudomonas aeruginosa*	212
1NEC	1,95	0,170	NAD(P)H nitroréductase insensible à l'O_2 de *Entrerobacter cloacae*	216
template: 1VFR A	1,8	0,213	NAD(P)H / flavine mononucléotide oxydoréductase de *Vibrio fischeri*	218
1OXA	2,1	0,196	cytochrome P450 de *Saccharopylospora erythraea*	403
template: 1CMN	1,7	0,228	cytochrome P450 55A1 de *Fusarium oxysporum*	402
1QOR A	2,2	0,140	quinone oxydoréductase complexée au NADPH de *Escherischia coli*	327
template: 1TEH A	2,7	0,283	alcool déshydrogénase de *Homo sapiens*	373
3PTE	1,6	0,148	D-alanyl-D-alanine carboxypeptidase (transpeptidase) de *Streptomyces* sp R161	349
template: 1GCE A	1,8	0,232	béta-lactamase de *Enterobacter cloacae*	364

Les 9 protéines-test ont ensuite été retirées de la banque de données de séquences *nr* et PDB pour simuler les conditions de modélisation réelles. Ensuite, trois modèles ont été réalisés pour chacune des 9 protéines, dans le

but de tester l'influence de la qualité de l'alignement sur la modélisation proprement dite:

- le premier, en utilisant notre méthode ESyPAli mais sans tenir compte de l'alignement réalisé par PSI-BLAST pour générer l'alignement pairé SI-SSC final (modèle I)
- le deuxième, en utilisant l'alignement fourni par ESyPAli en tenant compte de l'alignement réalisé par PSI-BLAST (modèle II)
- le troisième, en utilisant uniquement l'alignement généré par PSI-BLAST (modèle III)

Pour chacun des modèles, des RMSD globaux et locaux ont été calculés par rapport aux structures réelles sur base des carbones α, en utilisant le module Homology de InsightII (Accelrys Inc., San Diego). Pour les modèles I, II et III de chacune des 9 protéines, les RMSD globaux sont présentés dans le Tableau 19.

Tableau 19: Comparaison des modèles prédits pour chaque séquence à leur structure réelle, en calculant le RMSD global entre les deux. Le meilleur modèle généré pour chaque SI, sur base du RMSD global, est repris dans la dernière colonne. %id.: pourcentage d'identité.

	identifiant PDB	*template*	%id. cible *template*	MODELE I RMSD (Å) global	MODELE II RMSD (Å) global	MODELE III RMSD (Å) global	meilleur modèle sur base du RMSD global
moins de 20% id.	1BMTA	5REQ	9,9%	21,2	pas calculé	19,6	modèle III
	1LXA	1XAT	18,2%	16,1	17,1	13,5	modèle III
	1AMY	1JAE	18,4%	11,9	11,8	12,3	modèle II
plus de 20% id.	3PTE	1GCE	20,0%	6,8	6,6	4,9	modèle III
	1D2F	1BKG	20,7%	3,4	2,9	2,8	modèle III
	1QORA	1TEH	23,5%	3,8	3,3	4,3	modèle II
	1DUPA	1DUT	29,0%	3,0	3,0	4,5	modèle II/modèle I
	1OXA	1CMN	30,3%	3,9	3,8	3,9	modèle II
	1NEC	1VFRA	33,0%	2,1	pas calculé	2,7	modèle I

Pour analyser les RMSD présentés dans ce tableau, nous avons considéré qu'un modèle est de bonne qualité si le RMSD global par rapport à la structure observée vaut moins de 7 Å. Cette valeur est supérieure à celle utilisée aux CASP (3,5 Å) (Hubbard, 1999; Lackner *et al.*, 1999) car nous avons constaté qu'une petite erreur de modélisation dans un modèle globalement très bon peut faire croître le RMSD de façon considérable.

Sur base de ce critère de 7 Å, on constate, dans le Tableau 19, que si le pourcentage d'identité entre SI et SSC est supérieur à 20%, les modèles obtenus sont tous acceptables. De plus, les modèles de type I et II (élaborés sur un alignement consensus calculé suivant notre méthode) sont en majorité les meilleurs (4/6). Pour un pourcentage d'identité de moins de 20% entre SI et SSC, les modèles (3) sont tous mauvais.

Enfin, l'analyse du Tableau 19 montre que, sur base du RMSD global, la qualité des modèles II (incluant l'alignement PSI-BLAST dans

ESyPAli) est souvent meilleure (5/7) ou identique (1/7) à celle des modèles I.

Afin d'analyser plus finement les modèles générés, nous avons appliqué une procédure basée sur les RMSD locaux. Une approche similaire a été utilisée lors du CASP4 avec la fonction LCS du programme LGA (Zemla, 2000). Cette méthode est moins sensible aux erreurs locales de modélisation et permet de détecter facilement les portions correctes et incorrectes du modèle:

- La structure du modèle et la structure réelle sont alignées sur base de leur séquence par le programme ClustalW du module Homology de InsightII (Accelrys Inc., San Diego).
- On associe à chaque acide aminé du modèle un RMSD par rapport à la structure réelle. Ce RMSD est calculé sur une fenêtre de 9 positions de l'alignement décrit ci-dessus et centrée sur l'acide aminé étudié. Le RMSD est calculé par le module Homology du package InsightII (Accelrys Inc., San Diego).
- Sur base de l'ensemble des RMSD calculés, les fréquences relatives des valeurs inférieures ou égales à 1 Å et des valeurs supérieures à 2 Å sont calculées. La première fréquence calculée donne la fraction du modèle très similaire à la structure observée (Unger *et al.*, 1989), et la seconde la fraction du modèle peu similaire à cette même structure (Briffeuil *et al.*, 1998). Pour les modèles I, II et III de chacune des 9 protéines, ces fractions sont présentées dans le Tableau 20.

Tableau 20: Comparaison des modèles à la structure réelle en calculant les pourcentages de fenêtres ayant un RMSD inférieur à 1 Å ou supérieur à 2 Å.

	identifiant PDB	*template*	%id. cible *template*	MODELE I		MODELE II		MODELE III		meilleur modèle selon les fréquences RMSD local
				Fréq. RMSD < 1 Å	Fréq. RMSD > 2 Å	Fréq. RMSD < 1 Å	Fréq. RMSD > 2 Å	Fréq. RMSD < 1 Å	Fréq. RMSD > 2 Å	
moins de 20% id.	1BMTA	5REQ	9,9%	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
	1LXA	1XAT	18,2%	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
	1AMY	1JAE	18,4%	21,8	67,3	33,0	56,7	26,7	61,4	modèle II
plus de 20% id.	3PTE	1GCE	20,0%	40,6	41,2	44,6	39,0	39,2	43,9	modèle II
	1D2F	1BKG	20,7%	52,4	31,9	54,4	16,3	61,5	12,5	modèle III
	1QORA	1TEH	23,5%	44,5	28,1	49,4	21,1	53,2	23,2	modèle II/modèle III
	1DUPA	1DUT	29,0%	51,7	12,9	50,0	20,2	44,2	30,0	modèle I
	1OXA	1CMN	30,3%	65,6	16,3	66,9	16,0	64,3	16,3	modèle II
	1NEC	1VFRA	33,0%	69,7	14,4	N.A.	N.A.	63,9	17,6	modèle I

L'analyse du Tableau 20 montre que si la SI et la SSC partagent plus de 20% d'identités, les modèles sont globalement bons: dans ce cas, en moyenne 77% de chaque modèle est bien modélisé suivant le critère de 2 Å. De plus, on constate que les modèles I et II (élaborés selon ESyPAli sans et avec l'alignement PSI-BLAST) sont majoritairement les meilleurs (5/6) (critères 1 Å et 2 Å). Enfin, la qualité des modèles II (incluant l'alignement PSI-BLAST dans ESyPAli) est souvent meilleure (5/6) que celle des modèles I, quel que soit le critère utilisé.

Nous pouvons donc conclure que, sur ces 9 protéines, ESyPAli fournit majoritairement des modèles meilleurs que ceux réalisés à partir de l'alignement SI-SSC de PSI-BLAST. De plus, nous avons montré l'utilité de l'alignement SI-SSC fourni par PSI-BLAST dans ESyPAli: les modèles II (ESyPAli avec PSI-BLAST) sont majoritairement meilleurs que les modèles I (ESyPAli sans PSI-BLAST).

V.2.2. EVALUATION DE ESYPRED3D AU CASP4

Publication présentée (voir Annexe 10):

C. Lambert, N. Léonard, X. DeBolle and E. Depiereux
ESyPred3D: Prediction of proteins 3D structures
Bioinformatics **18**(9):1250-1256 (2002)

Après la première évaluation de la section précédente, nous avons voulu comparer la qualité de nos modèles par rapport à celle des modèles des meilleurs experts dans le domaine de la modélisation par homologie.

Nous avons donc caractérisé nos modèles, et ceux présentés lors des CASP 1, 2 et 3, par deux valeurs: le pourcentage d'identité entre SI et SSC et le RMSD global par rapport à la structure réelle. Pour les modèles des CASP, le RMSD choisi était celui du meilleur modèle obtenu pour chaque séquence cible. Pour nos modèles réalisés en utilisant l'alignement de ESyPAli (modèles II), le RMSD est le RMSD global tel que calculé dans la section précédente. L'ensemble des modèles caractérisés par deux valeurs (pourcentage d'identité et RMSD global) a été présenté dans un graphe bi-dimensionnel (Figure 36). Dans ce graphe, les meilleurs modèles devaient évidemment être localisés à des valeurs faibles du RMSD. On y remarque que les modèles réalisés par ESyPred3D (modèle II, carrés blancs) ont une qualité similaire, pour la plupart, à celle des meilleurs modèles des CASP 1, 2 et 3 (losanges noirs). Néanmoins, lorsque le pourcentage d'identité SI-SSC est inférieur à 20%, les RMSD des modèles générés par ESyPred3D par rapport à leur structure réelle sont plus élevés que ceux de modèles réalisés par les meilleurs experts internationaux en modélisation.

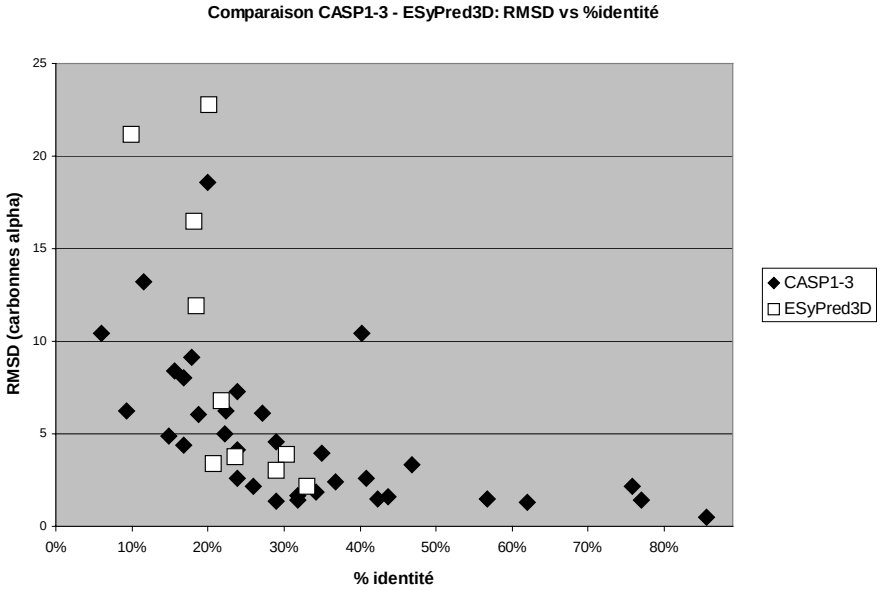


Figure 36: Evolution du RMSD global d'un modèle par rapport à sa structure observée en fonction de son pourcentage d'identité SI-SSC. Les carrés blancs représentent les modèles que nous avons réalisés (modèle II), et les losanges noirs les meilleurs modèles des CASP 1, 2 et 3 obtenus par modélisation par homologie.

D'après ces observations, nous pourrions conclure que notre méthode permet d'obtenir des modèles dont la qualité est comparable à celle des modèles réalisés par les meilleurs experts dans le domaine. Cependant, cette conclusion a été tirée sur un ensemble de 9 protéines, ce qui n'est pas statistiquement significatif.

Au vu de la fiabilité de notre méthode, nous avons alors décidé de participer au CASP4 pour comparer les performances de notre méthode à celles d'autres groupes. Notre groupe, sous le nom "LAMBERT-CHRISTOPHE" (groupe numéro 218), a soumis deux modèles pour chacune des 13 protéines décrites dans le Tableau 21:

- Le premier modèle fut construit en utilisant ESyPred3D avec la méthode d'alignement ESyPAli (modèle désigné T0xxxTS218_1)
- Le deuxième, en remplaçant l'alignement consensus par l'alignement brut fourni par PSI-BLAST (modèle désigné T0xxxTS218_2)

Tableau 21: Nom, pourcentage d'identité avec le *template* et fonction des 13 SI modélisées par ESyPred3D lors du CASP4.

SI	% id. avec *template*	Fonction
T0090	17,6	ADP-ribose pyrophosphatase, *E. coli*
T0092	18,6	Hypothetical protein HI0319, *H. influenzae*
T0099	30,5	No description
T0103	22,3	Pepstatin insensitive carboxyl proteinase, *Pseudomonas sp*.
T0111	50,0	Enolase, *E. coli*
T0112	23,3	Ketose Reductase / Sorbitol Dehydrogenase, *B. argentifolii*
T0113	37,3	Short chain 3-hydroxyacyl-coa dehydrogenase, rat
T0117	17,9	Deoxyribonucleoside kinase, *D. melanogaster*
T0121	22,0	MalK, *T. litoralis*
T0122	32,3	Tryptophan Synthase alpha subunit, *P. furiosus*
T0123	18,6	Beta-lactoglobulin, pig
T0125	18,6	Sp18 protein, *H. fulgens*
T0128	54,6	Manganese superoxide dismutase homolog, *P. aerophilum*

L'analyse des modèles réalisés par ESyPred3D au CASP4 (voir publication (Lambert *et al*., 2002) présentée à l'Annexe 10) montre que:

- Les modèles réalisés par ESyPred3D sont parfois très proches (critères LGA, AL0 et GDT_TS; voir description page 64) des meilleurs modèles reçus par les organisateurs du CASP4. 3 modèles sont classés second (SI T0103, T0121 et T0122) et 1 modèle est classé cinquième (SI T0117), sur plus de 200 modèles élaborés pour chaque cible, en moyenne.
- 7 modèles sur 13 (54%) sont classés dans le premier quart du classement des modèles soumis.
- Si on compare les modèles fournis par ESyPred3D aux modèles du CASP4 qui ont été construits à partir de la même SSC (de manière à tester la qualité de l'alignement SI-SSC), on observe que les modèles fournis par ESyPred3D sont parmi les 4 meilleurs modèles dans 10 cas sur 13 (77%).
- ESyPred3D a été classé 28ème sur 123 groupes participants et 9ème sur les 24 serveurs participants (Tramontano *et al*., 2001).

- 9 modèles sur 13 (67%) sont meilleurs que les modèles réalisés à partir de l'alignement SI-SSC fourni par PSI-BLAST.
- Aucun groupe n'obtient systématiquement de meilleurs résultats que ESyPred3D.

ESyPred3D se trouve donc parmi les meilleures méthodes de modélisation du CASP4. Nos modèles réalisés à partir de l'alignement SI-SSC de PSI-BLAST sont moins bons, ce qui montre que les performances de ESyPred3D sont essentiellement dues à notre méthode d'alignement, ESyPAli.

V.2.3. EVALUATION DE ESYPRED3D UTILISANT ESYPALINN AUX CASP5 ET CAFASP3

Après avoir développé ESyPAliNN (décrit à la page 113), nous l'avons intégré dans ESyPred3D. Nous avons ensuite participé au CASP5 (groupe n°35, LAMBERT-Christophe) et la nouvelle version du serveur ESyPred3D a été enregistrée dans la session CAFASP3 (groupe n°34, ESyPred3D), réservée aux serveurs automatiques. 31 protéines ont ainsi été modélisées (voir Tableau 37 à l'Annexe 11).

Pour le CASP5, les modèles envoyés ont été réalisés de façon semi-automatique en fournissant à ESyPred3D le *template* sélectionné par le programme 3D-Jury (Ginalski *et al*., 2003). Ce serveur combine les propositions de toute une série de programmes de reconnaissance de *fold* et en sélectionne le meilleur *template a priori*.

D'autre part, l'évaluation de ESyPred3D par le CAFASP3 a été réalisée de façon automatique par des programmes installés sur le serveur bioinfo.pl. Ce sont ces programmes qui ont soumis les séquences cibles et ont récupéré et évalué les modèles réalisés par ESyPred3D. De plus, les modèles envoyés au CAFASP3 ont été évalués par les assesseurs du CASP5 suivant les mêmes critères que les modèles des autres groupes de modélisation non automatique.

La comparaison des modèles soumis au CASP5 et au CAFASP3 a permis de comparer la sélection du *template* de ESyPred3D à celle du 3D-Jury (Ginalski *et al*., 2003).

L'évaluation officielle des organisateurs du CASP5 n'étant pas encore publiée, nous avons comparé nos résultats à ceux des autres groupes/serveurs en effectuant un classement des 172 groupes ayant soumis des modèles dans la catégorie "modélisation par homologie". Ce classement a été réalisé de la manière suivante:

- Six caractéristiques ont été reprises pour chaque modèle: GDT_TS, AL0, AL4, AL4+, RMSD et LGA_Q (voir définitions page 64).
- Pour chaque SI et pour tous les modèles de celle-ci, chaque caractéristique a été réexprimée en nombre d'écarts type par rapport à la moyenne de cette caractéristique pour tous les modèles soumis pour la SI.
- Pour chaque groupe, un score a été calculé. Il correspond à la moyenne du nombre d'écarts type pour toutes les caractéristiques de chaque modèle.

Le classement obtenu par cette méthode est repris dans l'Annexe 12 pour les 40 premiers groupes. D'emblée, il nous faut signaler que ce classement n'est qu'indicatif car l'ordre est fonction de la méthode de classification. En effet, il est difficile de déterminer un ordre précis et indiscutable puisque le nombre de séquences modélisées est différent pour chaque groupe.

Malgré ses limites, notre classement permet de dessiner les grandes tendances de nos résultats au CASP5:

- Les résultats obtenus par notre méthode totalement automatique, groupe n°34, ($40^{ème}$ position sur 172) sont en moyenne moins bons que ceux obtenus en utilisant ESyPred3D avec le *template* sélectionné par 3D-Jury ($29^{ème}$ position).
- Nos deux groupes (n°34 et 35) sont situés dans les 25% des meilleurs groupes de modélisation et le serveur automatique ESyPred3D est meilleur qu'un grand nombre de groupes de modélisation manuelle.
- Les meilleurs serveurs sont très proches des meilleurs groupes de modélisation manuelle (Fischer and Rychlewski, 2003).

Pour le CAFASP3, nous pouvons faire les observations suivantes:

- Les méta-serveurs obtiennent les meilleurs résultats au CAFASP3 (Fischer and Rychlewski, 2003).
- Le serveur ESyPred3D arrive à la $17^{ème}$ position sur 55 serveurs dans notre classement et dans le classement effectué par les organisateurs du CAFASP3 (Proteins 2003, *in press*) sur base d'une seule mesure de la qualité des modèles (MaxSub (Siew *et al*., 2000)). Si on distingue les serveurs des méta-serveurs, le serveur ESyPred3D arrive à la $10^{ème}$ position sur 55.
- En ce qui concerne les cibles "faciles" (*template* identifiable par PSI-BLAST), ESyPred3D arrive à la $3^{ème}$ position sur 55 et est considéré comme le meilleur serveur de modélisation par homologie (Proteins

2003, *in press*)(voir site CAFASP3, http://www.cs.bgu.ac.il/~dfischer/CAFASP3/ALevaluation/index.html).

- Une autre analyse des résultats du CAFASP3 publiée sur le site http://cubic.bioc.columbia.edu/eva/cafasp/fr_cafasp_hom/ en utilisant 8 critères de qualité des modèles générés montre que ESyPred3D arrive en moyenne à la 8ème position sur 55 serveurs. De plus, ESypred3D obtient, en moyenne, des résultats significativement meilleurs que plus de 50% des serveurs participant au CAFASP3 et seuls trois meta-serveurs sont significativement meilleurs que ESyPred3D, mais pour deux critères seulement.

La participation au CAFASP3 a permis de situer ESyPred3D parmi les meilleurs serveurs de modélisation par homologie. Il a même été qualifié de meilleur serveur de modélisation par homologie dans la mesure où les serveurs ayant obtenu de meilleures performances sont en fait des méta-serveurs qui combinent les modèles reçus de plusieurs serveurs de modélisation, soit pour fournir le meilleur modèle 3D parmi ceux proposés (Ginalski *et al*., 2003), soit pour réaliser un modèle hybride combinant les meilleures caractéristiques des différents modèles (Bates *et al*., 2001). Cette façon de procéder permet d'obtenir de meilleurs modèles mais, tout comme pour notre approche d'alignement pairé, les méta-serveurs ont besoin de serveurs performants pour leur fournir les meilleurs résultats possibles (Valencia, 2003). Ainsi, ce n'est qu'en améliorant les serveurs de modélisation que les meta-serveurs seront plus performants.

Au CASP5, les meilleurs serveurs ou groupes de modélisation combinaient les résultats de plusieurs programmes. De plus, la comparaison des modèles des groupes 34 (sélection du *template* par PSI-BLAST) et 35 (sélection du *template* par 3D-Jury) a montré que la sélection du *template* par 3D-Jury est meilleure que celle effectuée dans ESyPred3D en utilisant PSI-BLAST.

V.2.4. EVALUATION DE ESYPRED3D UTILISANT ESYPALINN PAR EVA

EVA (http://eva.salilab.org/~eva/cm) (Eyrich *et al*., 2001) est un système d'évaluation continue qui envoie systématiquement à chaque serveur de modélisation participant toutes les nouvelles structures insérées dans la banque de données PDB, quelques heures avant que celles-ci ne soient rendues publiques. De cette façon, il n'est pas possible aux différents programmes de modélisation de disposer de la structure réelle de la SI pour effectuer leur prédiction.

Notre serveur ESyPred3D a été inscrit à EVA dès sa mise en service en décembre 2001. Ce système a donc analysé les versions successives de ESyPred3D et les résultats obtenus nous ont parfois permis de corriger certaines erreurs de programmation.

Pour comparer au mieux les trois serveurs participant à EVA (3D-JIGSAW (Bates *et al*., 2001), Swiss-Model (Peitsch, 1996) et ESyPred3D (Lambert *et al*., 2002)), les résultats de comparaison de tous les modèles réalisés de janvier 2003 (dernière mise à jour de ESyPred3D) à juin 2003 avec leur structure réelle ont été téléchargés directement du site de EVA. Parmi tous les modèles réalisés, nous avons analysé ceux correspondant aux mêmes séquences cibles, soit 853 modèles pour chaque serveur. La comparaison a été faite en utilisant les quatre principaux indicateurs de performances de EVA (voir site EVA: http://eva.salilab.org):

- Le pourcentage de "couverture", qui représente la portion de la SI qui a été modélisée.
- Le pourcentage de positions de C_α équivalentes (distantes de maximum 3.5 Angstroms) entre le modèle et la structure réelle de la séquence cible, superposés de façon optimale par le programme CE (Shindyalov and Bourne, 1998). Plus il est élevé, plus le modèle et la structure réelle se ressemblent. C'est une mesure de l'identification du repliement correct.
- Le pourcentage de résidus correctement alignés (dans ce cas, identiques) est calculé dans l'alignement de séquences extrait de la superposition optimale entre le modèle et la structure réelle de la séquence cible. C'est une mesure de la qualité de l'alignement de séquences.
- Le RMSD global (calculé avec le programme CE (Shindyalov and Bourne, 1998)) entre le modèle et la structure réelle après superposition optimale des C_α. Plus il est faible, meilleur est le modèle.

Les résultats sont repris dans le Tableau 22, assortis d'un test statistique (*t* de Student) qui vérifie si la différence moyenne observée pour les différents indicateurs est significative: probabilité inférieure à 5% d'être due au hasard.

Tableau 22: Différences moyennes entre ESyPred3D, 3D-Jigsaw et Swiss-Model pour les quatre indicateurs principaux de performances de EVA. Les cellules grisées surlignent des différences en faveur de ESyPred3D et les valeurs de *t* significatives au seuil 5% (n=853).

	Différence ESyPred3D / 3D-Jigsaw		Différence ESyPred3D / Swiss-Model	
	Moyenne	t-student	Moyenne	t-student
Couverture (%)	0,581	2,229	1,571	4,174
Positions équivalentes (%)	1,878	3,740	0,643	1,470
Résidus correctement alignés (%)	4,632	4,823	4,218	4,740
RMSD (Å)	-0,460	-4,674	-0,789	-7,154

On peut observer dans le Tableau 22 que la qualité des modèles réalisés par ESyPred3D est, en moyenne, toujours meilleure que celle de 3D-Jigsaw et de Swiss-Model. L'avantage de performance pour le nombre de résidus correctement alignés est statistiquement hautement significatif ($P<1.3\ 10^{-6}$), ce qui met en évidence la qualité de notre méthode d'alignement ESyPAliNN.

Sur les 853 protéines modélisées par 3D-JIGSAW, Swiss-Model et ESyPred3D depuis janvier 2003, ESyPred3D est donc toujours meilleur en moyenne que les autres, et le plus souvent de façon statistiquement significative.

V.3. Conclusions et perspectives

Un serveur de modélisation par homologie (ESyPred3D) a été développé. A différentes étapes de ce développement, des tests rigoureux de performances ont été effectués (comparaison aux résultats des CASP1, 2 et 3, participation aux CASP4 et 5, ainsi qu'au CAFASP3 et évaluation continue via EVA). L'analyse des résultats de ces tests montre que ESyPred3D est actuellement l'un des meilleurs serveurs de modélisation par homologie et ses performances sont essentiellement dues à la qualité des alignements de séquences calculés par ESyPAliNN.

ESyPred3D peut néanmoins être amélioré à chacune des quatre étapes (voir page 51) du processus de modélisation par homologie.

Pour la première étape, la sélection d'une SSC plus adéquate lorsque la similarité entre la SI et les SSC les plus proches est faible, pourrait être envisagée de deux manières: (1) utilisation d'une autre méthode de recherche par similarité de séquence telle SAM-T2002 (Kevin Karplus, unpublished) et (2) réalisation d'un consensus de programmes de reconnaissance de *fold* comme le fait, par exemple, 3D-Jury (Ginalski *et al*., 2003). Cette approche serait surtout utile lorsque la similarité de séquence entre SI et SSC est très faible.

Au niveau de la seconde étape, il faudrait pouvoir aligner plusieurs *templates* à la SI car l'utilisation de plusieurs *templates* pourrait permettre d'obtenir un meilleur modèle 3D. Ceci nécessiterait la transformation de ESyPAliNN en une méthode d'alignement multiple. Ainsi, nous alignerions toutes les SSC, la SI et les séquences homologues à la SI avec différents programmes. Ensuite, un alignement multiple consensus entre la SI et les SSC serait extrait des différents alignements multiples. Néanmoins, il a été montré lors du CASP4 que l'utilisation de plusieurs *templates* n'améliorait pas toujours la qualité du modèle 3D généré, à cause des erreurs d'alignement et de la sélection des *templates* (Bates *et al*., 2001; Tramontano *et al*., 2001; Venclovas, 2001). Une approche assez performante, 3D-JIGSAW (Contreras-Moreira *et al*., 2003), combine plusieurs *templates* et alignements en utilisant un algorithme génétique pour construire le modèle 3D final.

Dans la troisième étape, la modélisation des boucles a été menée jusqu'à présent en mode automatique par MODELLER. Il serait notamment possible de modéliser les boucles, via MODELLER, par dynamique moléculaire. Les résultats publiés sont très intéressants si la partie à modéliser contient moins de 15 résidus. L'orientation des chaînes latérales pourrait également être améliorée par l'utilisation du programme SCWRL

(Canutescu *et al*., 2003) qui semble être le meilleur programme pour réaliser cette tâche.

Enfin, nous pourrions modifier la quatrième étape suivant deux voies bien distinctes:

1. En faire une étape de sélection du meilleur modèle parmi un ensemble de structures qui auraient été calculées au préalable lors des trois premières étapes (structures résultant d'alternatives possibles au niveau du *template*, de l'alignement ou encore du calcul du modèle 3D). C'est une approche déjà envisagée par Shi *et al*. dans le programme FUGUE (Shi *et al*., 2001) qui réalise plusieurs modèles et choisit le meilleur sur base de critères énergétiques ou géométriques.
2. En faire une étape de correction de l'alignement SI-SSC sur base des erreurs détectées dans le modèle calculé. Ces corrections pourraient alors permettre de calculer un nouveau modèle, qui serait à son tour évalué. On aurait ainsi un processus d'alignement et de modélisation itératif. Un système de ce genre vient d'être décrit par John *et al*. (John and Sali, 2003).

Cependant, les performances reconnues de notre serveur nous autorisent à l'utiliser dans sa forme actuelle pour la modélisation automatique des protéines d'un génome bactérien complet. Ce travail fera l'objet du chapitre suivant.

VI. Développement d'une base de données structurales et fonctionnelles pour le génome de *Brucella melitensis*

La création, la validation, l'exploitation et la mise à jour des bases données sont parmi les défis majeurs de la bioinformatique. Jusqu'aux environs de 1995, les universités et un certain nombre d'entreprises mettaient à jour localement des bases de données comme par exemple GenBank, et la PDB. La mise à jour était bimensuelle et l'espace disque nécessaire au stockage de ces données croissait de façon exponentielle.

Un tournant a été pris depuis lors, le monde académique s'orientant massivement vers l'utilisation des versions « en ligne » disponibles sur Internet et mises à jour en temps réel, tandis que le privé développait des solutions internes pour maintenir à jour localement ces mêmes banques de données et s'orientait pour des raisons de confidentialité vers la constitution de bases de données spécialisées.

Parallèlement, la vitesse d'accumulation des données a augmenté fortement sous l'effet des programmes de séquençage de génomes et, rapidement, le problème de la qualité des données stockées est devenu critique. En effet, les homonymies et synonymies des mots clés, associées aux erreurs humaines dans l'encodage, ont provoqué un effet « tour de Babel » rendant inopérantes les recherches que l'on désirait y réaliser. Les annotations basées sur des preuves expérimentales n'étant pas souvent clairement distinguées de celles inférées par similarité, ces dernières pouvaient générer une nouvelle inférence, et ainsi de suite. La similarité des séquences n'étant pas transitive, il en est résulté une grande quantité d'annotations douteuses voire complètement erronées (Apweiler *et al*., 2001; Karp *et al*., 2001).

Plusieurs initiatives ont été prises pour remédier à cette situation, entre autres, la banque de données RefSeq (Pruitt *et al*., 2000; Pruitt and Maglott, 2001) du NCBI et la remarquable banque de données Swiss-Prot (Bairoch and Apweiler, 1999). Récemment, face à l'ampleur du problème, les grandes banques de données Swiss-Prot, GenBank et PIR se sont unies pour créer une banque de données unique: UniProt (www.uniprot.org). Néanmoins, la tendance actuelle des universités et des sociétés biotechnologiques est à la constitution de banques de données spécialisées, spécifiques d'une problématique, et à très haute valeur ajoutée (Lambert *et al*., 2003).

C'est dans cette direction que nous nous sommes engagés. En effet, notre unité de recherche s'intéresse depuis plusieurs années à *Brucella*

melitensis (voir Introduction page 69). Une banque de données structurales et fonctionnelles contenant toute une série d'informations sur les pCDS de *Brucella melitensis* et, plus généralement, sur les génomes des différentes espèces de *Brucella* a donc été développée. Dans cette banque de données, nous mettrons l'accent sur la fiabilité des informations.

De plus, après avoir développé un serveur automatique fiable de modélisation par homologie, nous avons décidé de prédire les structures 3D d'un maximum de protéines déduites du génome de *Brucella melitensis* et les rendre accessibles à la communauté scientifique via cette banque de données.

La constitution de cet ensemble de modèles 3D et sa mise à jour régulière est motivée par trois objectifs principaux:

Le premier objectif est de fournir directement aux utilisateurs une structure 3D la plus fiable possible qui leur permettra de sélectionner les mutations à effectuer par mutagenèse dirigée et d'expliquer leurs effets phénotypiques. L'utilisateur pourra également utiliser ce modèle 3D pour toute une série d'applications telles que le *docking* avec un ligand, la recherche des résidus impliqués dans le site actif, ... (voir Prédiction de structure tertiaire, page 48).

Le deuxième objectif est de réaliser dans un avenir proche des recherches structurales sur les sites actifs ou plus généralement des sites de complexation des protéines (Yao *et al*., 2003). Plusieurs systèmes de ce type existent et pourraient être appliqués à notre banque de modèles 3D. Ainsi, la Relibase (Hendlich *et al*., 2003) permet de rechercher dans les protéines de cette banque toutes les conformations répondant aux contraintes fixées par l'utilisateur (Bergner *et al*., 2001). Autre exemple, LigBase (Stuart *et al*., 2002) est une base de données de sites de liaison et permet la recherche de sites suivant différents critères dans chaque protéine pour toutes les structures de PDB. De plus, ProCAT (Wallace *et al*., 1996; Wallace *et al*., 1997) permet de retrouver les sites actifs potentiels dans une structure de protéine donnée. Enfin, McLaughin *et al.* (McLaughlin and Berman, 2003) ont développé une méthode de recherche des domaines de liaison à l'ADN du type hélice-coude-hélice et, récemment, Jambon *et al*. (Jambon *et al*., 2003) ont proposé un logiciel de recherche de sites communs à plusieurs protéines.

Le troisième objectif est de réaliser le *docking* de petites molécules dans les structures prédites de protéines pour chercher, par exemple, des inhibiteurs de leur activité enzymatique. L'information obtenue (où se fixe la molécule, quelles sont ses interactions avec la protéine) pourrait être utilisée pour la conception de nouvelles molécules ou dans l'amélioration de molécules existantes. Cependant, lorsque l'on doit travailler avec un grand nombre de protéines ou de molécules, il est également possible d'utiliser des

méthodes de *screening* utilisant des programmes de *docking* (Glen and Allen, 2003) même si ceux-ci ont toutefois des limites (Pospisil *et al*., 2002). Récemment, il a été montré que cette approche pouvait être utilisée pour la recherche d'inhibiteurs de GPCRs (*G Protein-Coupled Receptors*) dans des modèles 3D réalisés à partir de la rhodopsine bovine (Bissantz *et al*., 2003).

Dans ce chapitre, nous présenterons les données disponibles pour la constitution de notre banque de données, puis nous décrirons la structure de celle-ci. Ensuite, nous détaillerons le fonctionnement du système d'interrogation de la banque de données, ainsi que les résultats obtenus en mettant l'accent sur les statistiques d'utilisation et la prédiction de la structure 3D des protéines. Enfin, nous donnerons quelques perspectives et améliorations qui pourraient être apportées à notre banque de données.

VI.1. Données disponibles sur le génome de Brucella melitensis et méthodes d'analyse utilisées

Le génome de *Brucella melitensis* a été séquencé et annoté automatiquement par la firme Integrated Genomics Inc. (DelVecchio *et al.*, 2002) (voir Introduction, page 70). Il contient 3197 pCDS codant pour autant de protéines déduites. Cependant, de nombreuses erreurs ont été constatées dans les prédictions de la position du codon *start* de ces pCDS. La valeur de notre banque de données reposant sur la fiabilité de l'information contenue, nous avons entrepris un projet de correction de ces positions, réalisé au sein de notre unité de recherche et en collaboration avec quatre équipes de recherche européennes membres du consortium de recherche européen COST845 (VLA Weybridge, UK; U. Navarra, Espagne; U. Cantabria, Espagne; Inserm U 431, France). L'ensemble des pCDS du génome de *Brucella melitensis* a été divisé en 5 parties, 640 pCDS pour chaque groupe participant au processus de correction. Les propositions de correction ont été proposées via l'interface de consultation de notre banque de données. Cette interface est décrite page 153.

Le protocole de correction pour chaque pCDS était le suivant:

1. Télécharger les deux fichiers (chromosomes) du génome de *Brucella melitensis* (nommés NC_003318.gbk et NC_003317.gbk) en format GenBank "FLAT" du NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Brucella_melitensis), et utiliser le logiciel Artemis (Rutherford *et al.*, 2000) (http://www.sanger.ac.uk/Software/Artemis/) pour visualiser les annotations.
2. Détecter les mauvais codons *start*, en effectuant les trois analyses suivantes:
   (1) Examiner la nature du codon *start*: si c'est un TTG, le codon *start* est généralement mauvais.
   (2) Voir si la pCDS en recouvre une autre: si oui, le codon *start* sera généralement mauvais.
   (3) Analyser les alignements réalisés par le programme BLASTP contre la banque *nr* du NCBI pour cette pCDS. Si la longueur des séquences similaires dans des organismes phylogénétiquement proches est plus longue ou plus courte, le codon *start* est probablement mauvais.

3. Rechercher un nouveau codon *start*, à l'aide du programme Artemis, en utilisant l'ordre de préférence suivant: ATG puis GTG puis TTG. En effet, le codon *start* ATG est utilisé plus fréquemment que le codon GTG qui est lui-même utilisé plus fréquemment que le codon TTG.
4. Choisir la séquence codante la plus longue s'il y a un RBS (*Ribosome Binding Site*) en amont du nouveau codon *start* (Salgado *et al*., 2000). Si la pCDS est à l'intérieur d'un opéron, les positions -1 et -4 du codon *start*, par rapport au codon stop de la pCDS précédente, sont préférées. Par exemple, ATGA, où les codons ATG et TGA se recouvrent ("-4" est très fréquent) ou TGATG, où il y a aussi ATG et TGA ("-1" est fréquent).
5. S'il y a plusieurs possibilités de position *start*, sans opéron identifié ni RBS, la pCDS ayant une taille proche de celle des séquences homologues détectées dans *nr* a été sélectionnée.
6. Si la situation 5. est rencontrée mais sans homologues, la pCDS la plus longue avec un ATG ou un GTG a été choisie.

Les propositions de corrections ont été collectées et analysées. Les corrections les plus vraisemblables ont ensuite été apportées à notre banque de données qui a acquis ainsi une haute qualité dans la définition des séquences des pCDS de *Brucella melitensis*.

Ces pCDS ont été traduites en séquences protéiques (protéines déduites ou *Deduced Protein*, DP), et la structure 3D de ces dernières est prédite tous les mois à l'aide du programme ESyPred3D. Ces calculs durent chaque fois trois jours sur le cluster de 7 PC Priminfo de notre unité de recherche (voir fiche technique, page 75). Ces prédictions structurales sont d'autant plus précieuses qu'aucune détermination expérimentale de structure 3D de protéine de *Brucella melitensis* n'a été publiée (à notre connaissance).

D'autres prédictions basées sur les pCDS sont présentées dans notre banque de données, voir Tableau 23, page 149.

VI.2. Structure de la banque de données

La banque de données actuelle de *Brucella melitensis* est organisée en répertoires correspondant à chaque type d'information stockée. La description de l'information contenue dans les répertoires est reprise dans le Tableau 23, page 149. Ces répertoires contiennent chacun 3197 fichiers (un fichier par pCDS du génome) à l'exception des répertoires ARCHIVES et SEARCH. En effet, le répertoire ARCHIVES contient des anciens fichiers d'information qui ont été mis à jour tandis que le répertoire SEARCH contient des tables des données stockées dans les autres répertoires du système. Ces tables permettent de retrouver rapidement des informations recherchées.

Au total, la banque de données contient donc environ 80.000 fichiers et occupe environ 3,5 GB sur disque.

Tableau 23: Description de la l'information contenue dans les différents répertoires de la banque de données, périodicité de mise à jour de ces informations et types de recherche pouvant être effectuées sur ces données. Lorsqu'une caractéristique n'est pas applicable, NA (Non Applicable) est inscrit dans la cellule. Le type de recherche "motif" permet d'utiliser des expressions régulières pour effectuer la recherche d'information.

Répertoire	Description de l'information	Méthode de mise à jour et périodicité	Type de recherche	Taille du répertoire (KB)
ARCHIVES	Répertoire contenant les anciens fichiers de données.	Automatique	NA	3.564
BLASTP	Résultat de la recherche de séquences similaires à la DP dans la banque de données *nr* du NCBI avec le programme BLASTP.	Automatique Hebdomadaire	Texte	1.035.548
BLASTPSP	Résultat de la recherche de séquences similaires à la DP dans la banque de données Swiss-Prot (Boeckmann *et al.*, 2003) avec le programme BLASTP.	Automatique Hebdomadaire	Texte	301.160
COMMENT	Commentaires envoyés par les utilisateurs de la banque de données. Ces commentaires sont révisés par le gestionnaire avant publication.	Manuelle	Texte	7.220
CROSSREF	Liens hypertextes concernant la DP vers les banques de données KEGG *Metabolic Pathways* et TrEMBL.	NA	NA	13.860
FASNUC	Séquence nucléotidique de la pCDS.	NA	Motif	16.104
FASPEP	Séquence protéique déduite de la pCDS.	NA	Motif	15.196
FREECOMMENT	Commentaires envoyés par les utilisateurs de la banque de	Manuelle	NA	8

	données, mais pas révisés par le gestionnaire avant publication.			
FUNCTION	Fonction de la DP selon l'annotation de GenBank et établie par Integrated Genomics Inc. (DelVecchio *et al*., 2002).	NA	Texte	14.144
GCCONTENT	Fréquences relatives en G+C et A+T de la séquence nucléotidique de la pCDS. Contient également le rapport (G+C)/(A+T).	NA	Nombre	16.556
INFOPRED3D	Alignement de la DP avec le *template* utilisé lors de la prédiction de la structure 3D par ESyPred3D. Contient l'identifiant PDB du *template*, le pourcentage d'identité entre SI et *template* dans la zone modélisée et la longueur de la zone modélisée.	Automatique Mensuelle	Texte (*template*) ou Nombre	13.960
METABOLICPATHWAY	Voies métaboliques dans lesquelles est impliquée la DP. La mise à jour est mensuelle et est effectuée à partir de la banque de données KEGG *Metabolic Pathways* (Kanehisa, 2002).	Manuelle	NA	257.484
OBSFUNCTION	Fonction de la DP déterminée expérimentalement. Celle-ci doit avoir fait l'objet d'une publication scientifique et doit être introduite par les utilisateurs de notre banque de données.	Manuelle	Texte	8
OBSINTERACTIONS	Interactions déterminées expérimentalement entre la DP et d'autres protéines. Celles-ci doivent avoir fait l'objet d'une publication scientifique et doivent être introduites par les utilisateurs de notre banque de données.	Manuelle	NA	0
OBSLOCALIZATION	Localisation cellulaire de la DP déterminée	Manuelle	Texte	0

	expérimentalement. Celle-ci doit avoir fait l'objet d'une publication scientifique et doit être introduite par les utilisateurs de notre banque de données.			
PFAM	Alignements entre la DP et des domaines PFAM (Bateman *et al.*, 2002) possédant une similarité significative (*E-value*<0.01).	Automatique Trimestrielle	Texte	21.276
PFAMSUM	Identifiants des domaines PFAM (Bateman *et al.*, 2002) possédant une similarité significative (*E-value*<0.01) avec la DP.	Automatique Trimestrielle	Texte	12.980
PI_MW	Point isoélectrique et masse moléculaire de la DP prédits par le programme "Compute pI / Mw" (Wilkins *et al.*, 1998).	NA	Nombre	12.968
POS_CHROMOSOME	Nom du chromosome où se trouve la pCDS ainsi que les positions de début et de fin de la pCDS sur celui-ci. On peut en déduire la longueur des séquences peptidiques et nucléotidiques.	NA	Nombre	14.420
PRED3D	Modèle 3D prédit par ESyPred3D pour la DP.	Automatique Mensuelle	NA	277.232
PREDLOCALIZATION	Résumé de la prédiction de localisation cellulaire effectuée par le programme PSORT (Nakai and Horton, 1999) pour la DP.	NA	Texte	436
PSORT	Résultat complet de la prédiction de localisation cellulaire effectuée par le programme PSORT (Nakai and Horton, 1999) sur la DP.	NA	NA	18.412
REFERENCES	Références bibliographiques concernant la pCDS,	Manuelle	NA	0

	introduites par les utilisateurs.			
SIMILORG	Organismes proches de *Brucella melitensis* du point de vue phylogénétique, caractérisés par l' *E-value* de leur fragment d'ADN le plus similaire à la pCDS. La similarité est déterminée par le programme BLASTN.	Manuelle	Nombre	18.328
SPFUNCTION	Fonction de la DP de la pCDS, déterminée par une recherche par similarité dans la banque de données Swiss-Prot (Boeckmann *et al.*, 2003), en utilisant BLASTP. La fonction de la séquence la plus similaire (*E-value* la plus faible) est attribuée à la pCDS.	Automatique Hebdomadaire	Texte	204
SSPRED	Résultat de la prédiction de la structure secondaire réalisée avec le programme PSIPRED2 (McGuffin *et al.*, 2000) sur la DP.	Automatique Trimestrielle	NA	13.632
TMHMM	Prédiction des hélices transmembranaires de la DP, exécutée par le programme TMHMM (Krogh *et al.*, 2001).	NA	Nombre	13.212
SEARCH	Répertoire contenant les tables des données stockées dans les autres répertoires du système. Ces tables permettent de retrouver rapidement des informations recherchées.	Automatique Hebdomadaire	NA	1.335.148

VI.3. Fonctionnement du système d'interrogation de la banque de données

Le site *web* de la banque de données des génomes de *Brucella sp.* est accessible à l'URL suivante:

http://serine.urbm.fundp.ac.be/~seqbruce/GENOMES

La page d'accueil générale de la banque de données (Figure 37) permet de télécharger les fichiers en format GenBank "FLAT" contenant les génomes des différentes espèces de *Brucella* déjà séquencées (*Brucella melitensis* et *Brucella suis* 1330) ainsi qu'un fichier contenant les corrections des positions *start* des pCDS de *Brucella melitensis* proposées par les différentes équipes européennes.

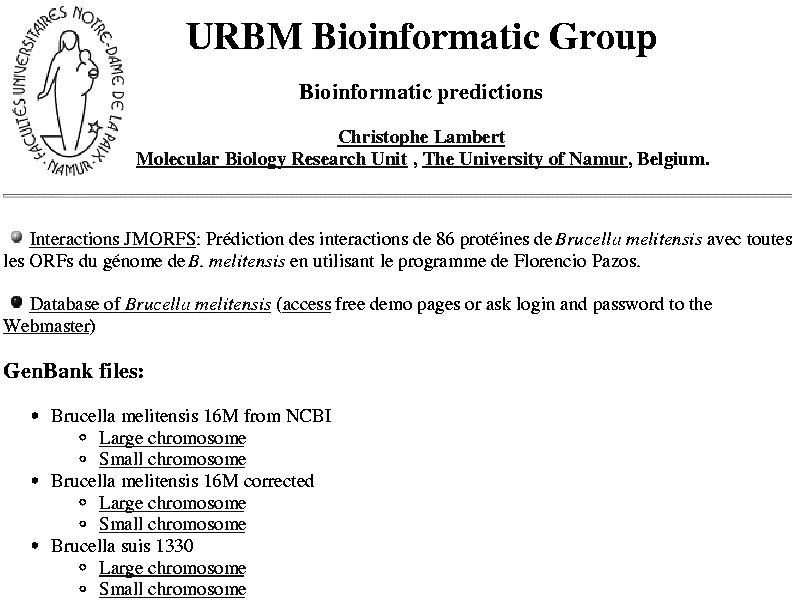


Figure 37: Page d'accueil générale de la banque de données *Brucella sp*.

La page d'accueil de notre banque de données concernant le génome de *Brucella melitensis* est divisée en 5 parties (Figure 38):

- ***Advanced tools*** est un lien hypertexte vers des outils permettant de réaliser des requêtes complexes dans la banque de données. Actuellement, un seul outil a été développé. Il permet de rechercher parmi toutes les pCDS et DPs, celles contenant une séquence constituée de la répétition d'un motif défini par l'utilisateur.

- ***Access an ORF*** permet d'accéder à la page HTML reprenant toutes les informations concernant la pCDS choisie en introduisant son identifiant GenBank (BMEIxxxx ou BMEIIxxxx) ou l'identifiant donné par *Integrated Genomics* (RBMExxxxx). Il est également possible d'obtenir la liste de toutes les pCDS en cliquant sur « ***List all ORFs*** ».
- ***Search text in the database*** permet à l'utilisateur de rechercher, à l'aide d'expressions régulières, des informations dans les 11 descripteurs des pCDS (repris au Tableau 23, page 149) qui acceptent une recherche textuelle.
- ***Advanced search*** est un lien hypertexte vers une page de recherche avancée (Figure 39 et Figure 40) qui permet d'effectuer des recherches de pCDS en combinant toutes les informations contenues dans la banque. Dans la première partie de la page (Figure 39), on peut sélectionner le(s) chromosome(s) sur le(s)quel(s) se positionnent les pCDS recherchées et effectuer la recherche de textes ou motifs contenus dans les informations des pCDS décrites dans le Tableau 23. Cette recherche de textes ou motifs peut être combinée à celle de pCDS dont les données numériques (reprises dans le Tableau 23, page 149) sont comprises dans un intervalle fixé par l'utilisateur.
  Dans la deuxième partie de cette page de recherche (Figure 40), il est possible de combiner les critères précédents à la recherche de pCDS partageant une certaine similarité avec des fragments de génomes d'organismes proches de *Brucella melitensis* (α-protéo bactéries) ou dans un organisme de référence, *Escherichia coli*. L'utilisateur choisit le niveau de similarité attendue par une valeur d'*E-value*. Ce type de recherche permet, par exemple, de retrouver toutes les pCDS présentes chez *Brucella melitensis* mais absentes dans les génomes d'organismes choisis par l'utilisateur.
  La recherche est lancée en cliquant sur ***Search***. Le résultat de la recherche est une liste de pCDS associées à un des descripteurs du Tableau 23 choisi par l'utilisateur.
- ***BLAST the database*** permet d'effectuer des recherches par similarité de séquences dans les génomes des *Brucella* ou dans leurs protéomes, via le logiciel BLAST2.
- ***Search the database to find motives*** permet une recherche de motifs dans les séquences nucléotidiques ou peptidiques, à l'aide d'expressions régulières dont la syntaxe est décrite en bas de la page (voir Figure 38).

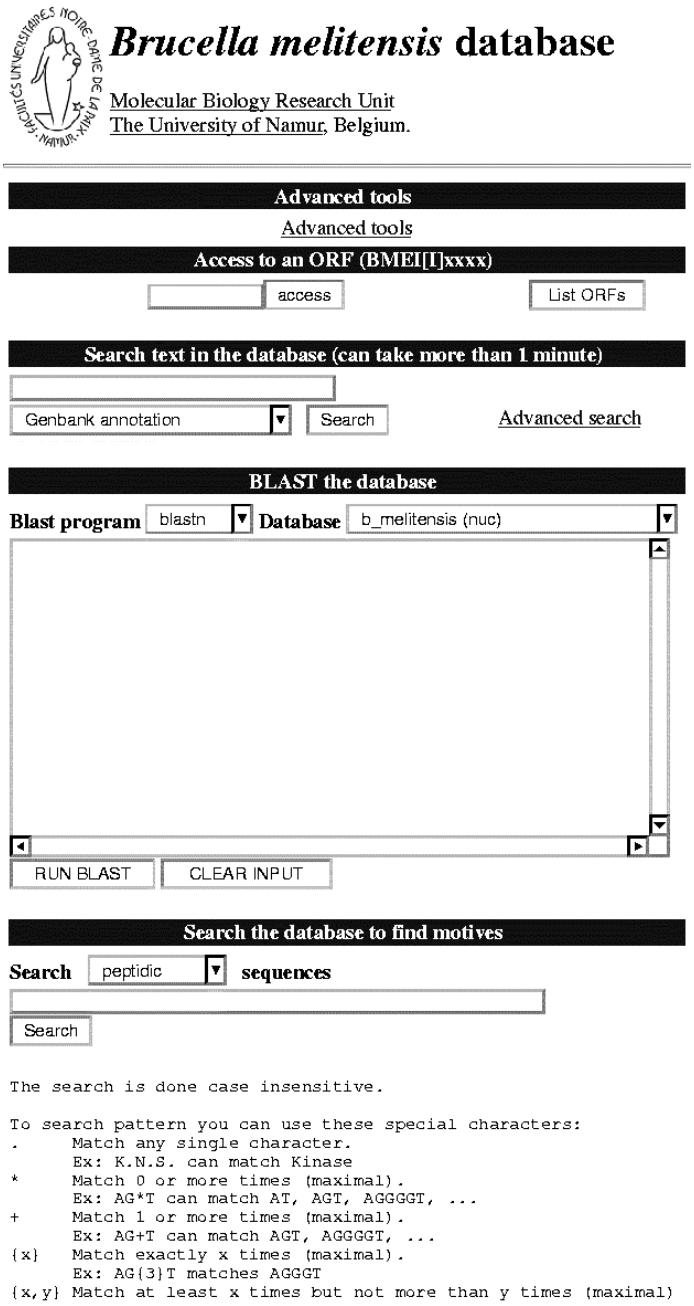


Figure 38: Page d'accueil de la banque de données concernant le génome de *Brucella melitensis*.

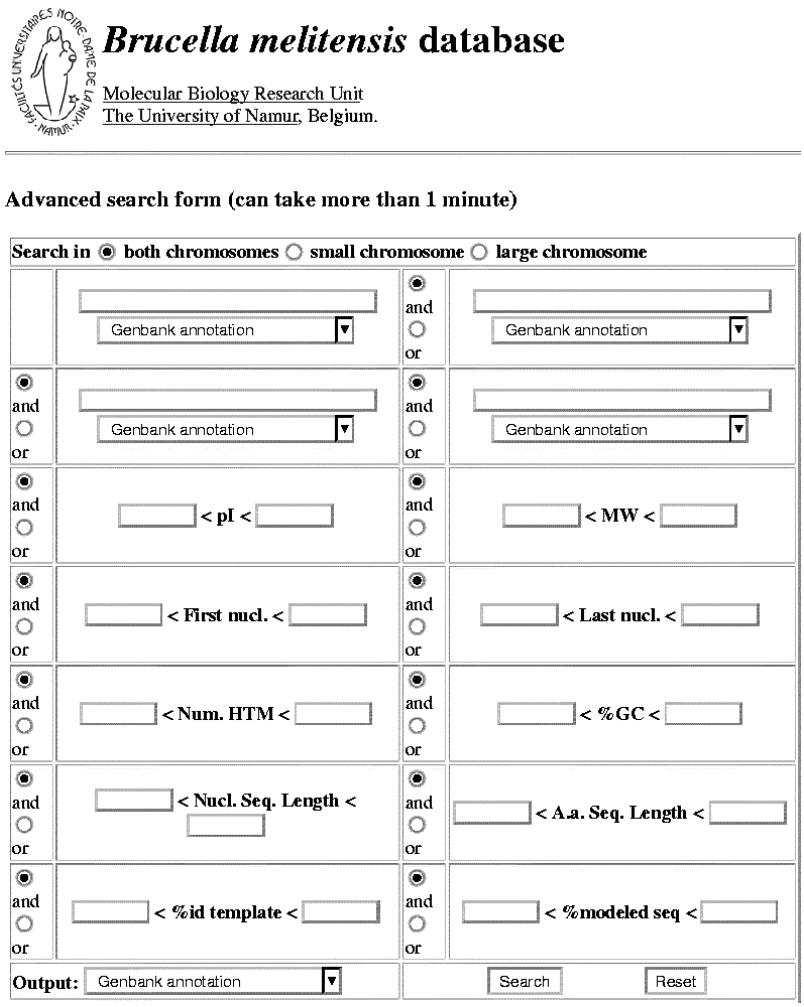


Figure 39: Première partie du formulaire de recherche avancée.

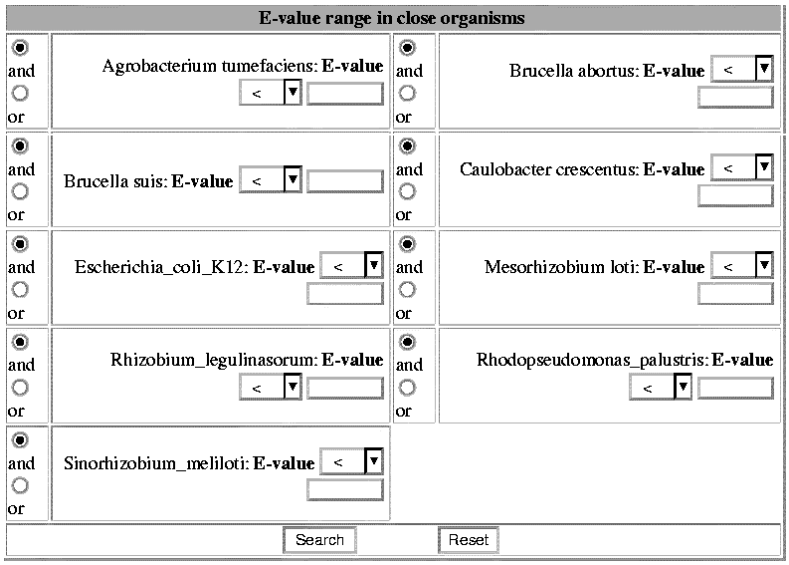


Figure 40: Deuxième partie du formulaire de recherche avancée.

La page HTML présentant la "carte d'identité" d'une pCDS (Figure 41) commence par un rappel de l'identifiant de la pCDS et des identifiants des pCDS suivantes et précédentes dans le génome. Ensuite, la page contient sept zones de données.

La première zone concerne la fonction de la pCDS. Ainsi, on peut voir dans la Figure 41 que l'annotation de la pCDS dans GenBank peut être facilement comparée à celle de la protéine la plus similaire dans Swiss-Prot, aux domaines fonctionnels prédits par Pfam et à la fonction réelle de la protéine codée par la pCDS. Les résultats des recherches dans les banques de données *nr* et Swiss-Prot par le programme BLASTP sont également accessibles par un lien hypertexte.

Summary for BMEI0019

BMEI0018 << >> BMEI0020

Function

Genbank annotation

LACI-FAMILY TRANSCRIPTION REGULATOR

BLASTP result against nr (NCBI) (*last update 7/4/2003*)

SwissProt similarity

(E= 2e-39) gi|1168844|sp|P46828|CCPA_BACME
GLUCOSE-RESISTANCE AMYLASE REGULATOR (CATABOLITE CONTROL PROTEIN)

BLASTP result against Swiss-Prot (*last update 7/4/2003*)

Pfam summary

Model	Description	Score	E-value	N
Peripla_BP_like	Periplasmic binding proteins and suga	68.3	7.9e-20	1
lacI	Bacterial regulatory proteins, lacI f	30.8	3.2e-08	1

PFAM result (*last update 2/3/2003*)

Observed function

Modify Send Modification

Figure 41: Première zone de la carte d'identité d'une pCDS.

La deuxième zone de la page HTML reprend (Figure 42) des liens hypertextes vers les cartes des voies métaboliques prédites par le KEGG ainsi qu'une brève description des voies métaboliques dans lesquelles est impliquée la pCDS.

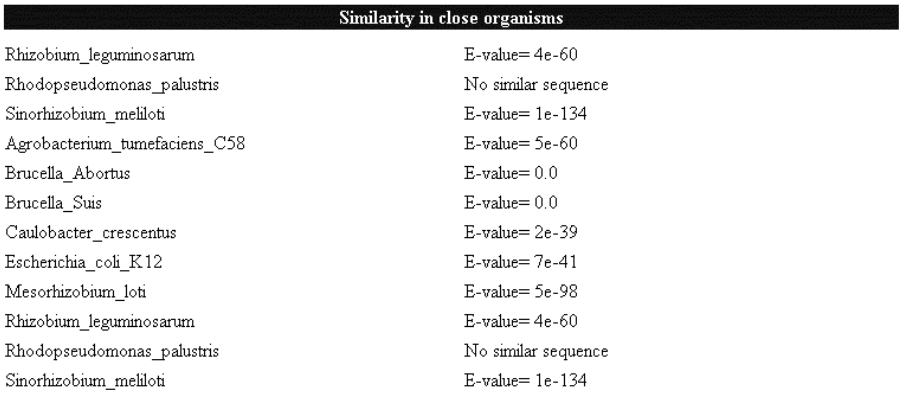
Metabolic and regulatory pathway from KEGG

16.1 **ABC transporters, prokaryotic** [PATH:bme02010]
BMEI0019 LacI-family transcription regulator

Last updated: Feb 27, 2003

Figure 42: Deuxième zone de la carte d'identité d'une pCDS.

La troisième zone (Figure 43) montre la similarité de la pCDS avec des fragments d'ADN de différents organismes proches de *Brucella melitensis* du point de vue évolutif. La présence d'*Escherichia coli* dans la liste permet de signaler l'existence éventuelle d'une forte similarité de séquence avec le génome de cet organisme de référence.



Similarity in close organisms	
Rhizobium_leguminosarum	E-value= 4e-60
Rhodopseudomonas_palustris	No similar sequence
Sinorhizobium_meliloti	E-value= 1e-134
Agrobacterium_tumefaciens_C58	E-value= 5e-60
Brucella_Abortus	E-value= 0.0
Brucella_Suis	E-value= 0.0
Caulobacter_crescentus	E-value= 2e-39
Escherichia_coli_K12	E-value= 7e-41
Mesorhizobium_loti	E-value= 5e-98
Rhizobium_leguminosarum	E-value= 4e-60
Rhodopseudomonas_palustris	No similar sequence
Sinorhizobium_meliloti	E-value= 1e-134

Figure 43: Troisième zone de la carte d'identité d'une pCDS.

La quatrième zone permet aux utilisateurs d'envoyer des commentaires au gestionnaire de la banque de données pour proposer des corrections ou des mises à jour de l'information contenue dans la banque. Dans la Figure 44, nous pouvons voir, par exemple, qu'un utilisateur demande de changer la taille de la pCDS pour qu'elle soit plus similaire à celle des pCDS dans les organismes phylogénétiquement proches.

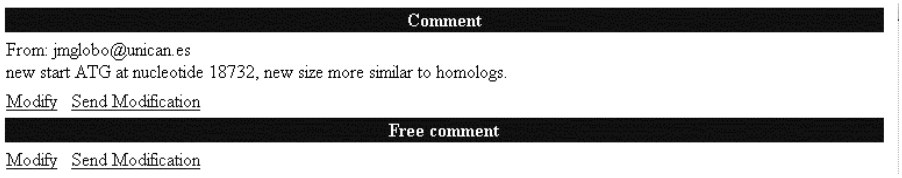


Figure 44: Quatrième zone de la carte d'identité d'une pCDS.

Dans la cinquième zone (Figure 45), consacrée à la localisation cellulaire, on peut comparer la localisation cellulaire observée et la localisation cellulaire prédite, ainsi que la présence éventuelle d'hélices transmembranaires. Cette dernière information permet d'infirmer ou de confirmer la localisation prédite lorsqu'il s'agit d'une protéine en membrane interne.

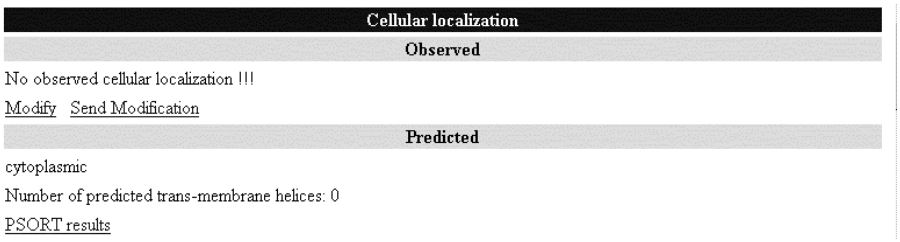


Figure 45: Cinquième zone de la carte d'identité d'une pCDS.

La sixième zone (Figure 46) fournit les positions de début et de fin de la pCDS sur un des deux chromosomes de *Brucella melitensis* ainsi que les longueurs des séquences nucléotidique et peptidique. D'autres informations sont également disponibles, telles que le contenu en GC, le point isoélectrique prédit, la masse moléculaire prédite, les interactions physico-chimiques prédites ou déterminées expérimentalement avec d'autres pCDS, des références bibliographiques concernant la pCDS, des liens avec d'autres banques de données et, finalement, la structure secondaire prédite.

Figure 46: Sixième zone de la carte d'identité d'une pCDS.

Enfin, la page HTML (Figure 47) présente la structure 3D prédite par ESyPred3D et quelques caractéristiques de ce modèle: l'alignement entre la séquence cible et le *template*, le pourcentage d'identité entre la DP et le *template* dans la région modélisée et la fraction de la séquence modélisée. Un exemple de structure est montrée (Figure 48) pour BMEI0008, une

protéine prédite comme la méthyltransférase gidB (*Glucose inhibited division protein B*).Nous discuterons de l'utilité de cette prédiction dans la prochaine section. Pour terminer, la page HTML reprend les séquences nucléotidique de la pCDS et peptidique de la DP et un bouton permettant de retrouver les séquences nucléotidiques en amont et en aval de la pCDS, sur le chromosome.

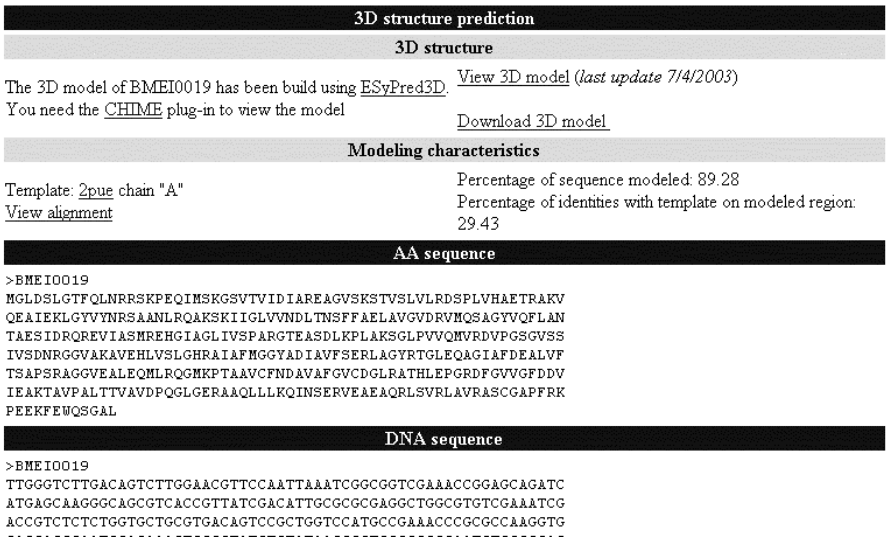


Figure 47: Septième zone de la carte d'identité d'une pCDS.

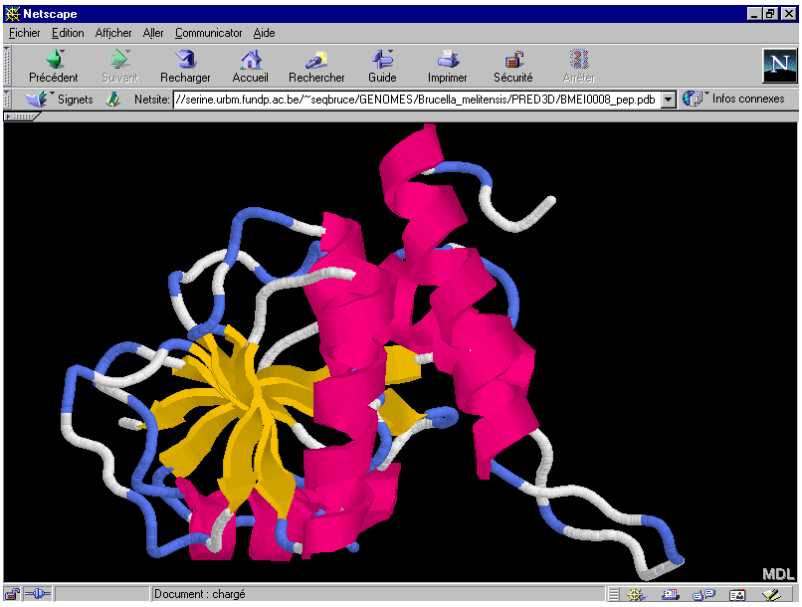


Figure 48: Structure de BMEI0008, une protéine prédite comme la méthyltransférase gidB (*Glucose inhibited division protein B*).

VI.4. Résultats

VI.4.1. STATISTIQUES D'UTILISATION

Depuis sa mise en ligne en janvier 2002, notre banque de données a fait l'objet de 140.000 requêtes (voir Figure 49) et reçu 2620 commentaires d'utilisateurs. Le nombre de connexions à la banque de données est passé d'un peu moins de 2000 en janvier 2002 à plus de 8000 pour le seul mois de juin 2003. Des pics de consultation ont été observés en février 2002, janvier 2003 et février 2003. Ils correspondent à l'indexation systématique de toute la banque de données par le moteur de recherche des FUNDP (février 2003) et à la vérification des corrections proposées par les groupes de recherche participant à la redéfinition de la position des codons *start* des pCDS (janvier 2003 et février 2003).

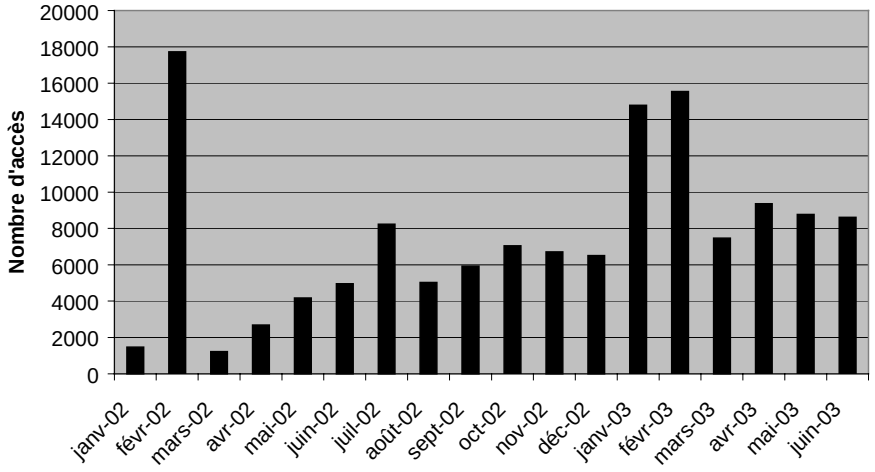


Figure 49: Nombre d'accès par mois de la banque de données sur *Brucella melitensis* depuis sa mise en accès public en janvier 2002.

D'autre part, nous avons dressé les statistiques d'accès suivant le type d'information ou de fonctionnalité demandé. Celles-ci sont reprises dans le Tableau 24. Nous pouvons y constater que les quatre informations les plus consultées sont la « carte d'identité » des pCDS, le résultat de la recherche en similarité dans la banque de données *nr*, le résultat de la prédiction de localisation cellulaire et les cartes des voies métaboliques dans lesquelles sont impliquées les DP. D'autre part, les quatre fonctionnalités les moins consultées sont l'obtention de la liste de toutes les pCDS, la recherche de motifs, l'utilisation des outils avancés et la consultation des archives.

Tableau 24: Fréquence d'utilisation des différentes fonctions et informations de la banque de données de janvier 2002 à juillet 2003.

Information ou fonctionnalité	Nombre d'accès
Consultation de la « carte d'identité » des pCDS	75312
Résultat complet de la recherche en similarité dans la banque *nr* pour la DP	13909
Résultat complet de la prédiction de localisation cellulaire	5322
Consultation des cartes des voies métaboliques	3477
Modification des pCDS par le gestionnaire	3149
Envoi de modifications par un utilisateur	2579
Résultat complet de la recherche en similarité dans la banque Swiss-Prot pour la DP	1523
Fonction « recherche des séquences amont et aval »	1010
Fonction « recherche avancée »	950
Téléchargement des fichiers des chromosomes en format GenBank	852
Résultat complet de la prédiction de domaines conservés	637
Fonction « visualisation de la structure 3D »	611
Utilisation de la suite de programmes BLAST	573
Fonction « recherche de motifs »	342
Fonction « outils avancés »	123
Consultation des archives	36

VI.4.2. CORRECTION DES POSITIONS *START* DES PCDS

La banque de données et les différents outils de recherche ont permis de faciliter et d'accélérer le travail de correction des définitions des positions *start* des pCDS. Pendant le processus de correction, nous avons reçu 2490 commentaires concernant 1781 pCDS. L'analyse de tous ces commentaires a conduit à la modification de 908 pCDS. La différence moyenne de position est de 69 nucléotides ou 23 acides aminés. 565 pCDS ont été raccourcies et 334 ont été allongées (voir la distribution des décalages sur la Figure 50).

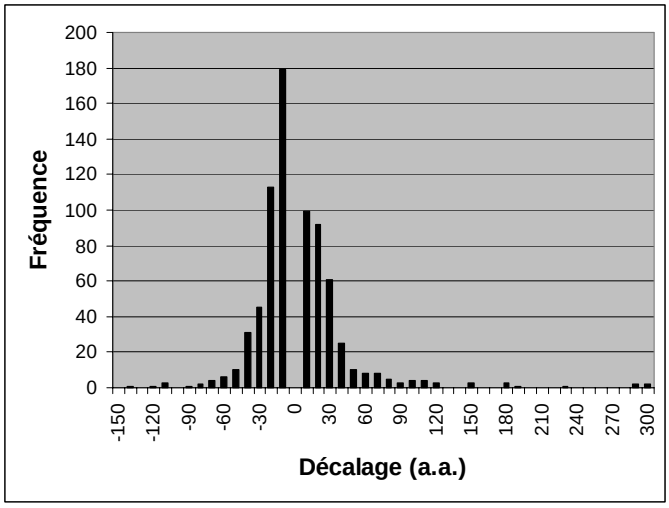


Figure 50: Distribution des décalages dans les positions *start* des pCDS entre avant et après la correction des 908 pCDS.

VI.4.3. PRÉDICTION DES STRUCTURES 3D DES PCDS

De par l'accroissement du nombre de protéines de structure connue et le nombre de protéines recensées dans les banques de données, le nombre de protéines déduites pour lesquelles il est possible de construire un modèle 3D augmente régulièrement. Cette évolution est reprise dans la Figure 51.

Entre décembre 2002 et juillet 2003, le nombre de modèles est passé de 1621 à 1720 et la fraction de protéines modélisées, au moins en partie, est ainsi passée de 50,7% à 53,8%.

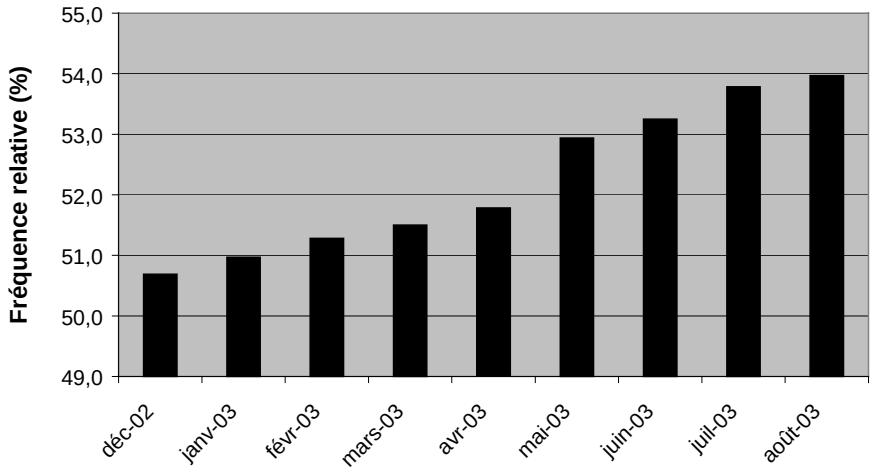


Figure 51: Evolution mensuelle de la fraction des protéines déduites du génome de *Brucella melitensis* possédant un modèle 3D prédit, le nombre total de DP étant de 3197.

Pour évaluer la qualité des modèles 3D prédits, nous avons utilisé les deux critères suivants:

1. le pourcentage d'identité dans la zone modélisée doit être supérieur à 20%,
2. la portion de la protéine qui est modélisée doit être supérieure à 80% de sa longueur.

L'évolution des fractions des protéines modélisées de manière fiable ou non fiable (suivant les deux critères précédents) est présentée dans la Figure 52. Pour le mois de juillet 2003, 891 modèles (sur un total de 1720) avaient été décrits comme fiables.

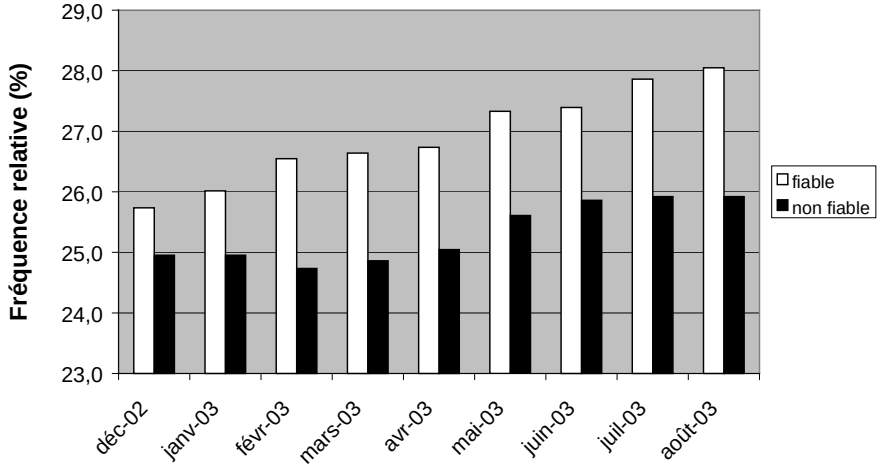


Figure 52: Evolution mensuelle des fractions des protéines déduites du génome de *Brucella melitensis* possédant des modèles 3D prédits fiables (blanc) et non fiables (noir), le nombre total de DP étant de 3197.

D'autre part, certains auteurs (Sander and Schneider, 1991; Rost, 1999) ont établi des critères plus précis pour prédire la qualité d'un modèle. Le dernier en date (Rost, 1999) a effectué une vaste comparaison des pourcentages d'identité entre protéines appartenant à une même famille structurale. Son étude a permis d'établir la formule suivante, reliant le nombre de résidus alignés entre SI et SSC d'une même famille structurale et le pourcentage d'identité à partir duquel les modèles construits pour les SI à partir des SSC peuvent être considérés comme fiables:

$$p^I(L) = 480 \cdot L^{-0.32 \cdot \left(1+e^{-L/1000}\right)}$$

où $p^I(L)$ est le seuil de pourcentage d'identité au-dessus duquel les modèles sont considérés comme valables

L est le nombre d'acides aminés alignés

Nous pouvons donc conclure que si le pourcentage d'identité, p, entre SI et SSC est supérieur au seuil, les modèles peuvent être considérés comme fiables. En d'autres termes, si $p > p^I(L)$, le modèle calculé sur base d'un alignement contenant L résidus alignés et de pourcentage d'identité p, est considéré comme fiable.

Nous n'avons malheureusement pu dresser des statistiques basées sur cet indice que pour la dernière version de notre ensemble de modèles 3D (version de juillet 2003). Celles-ci sont présentées dans la Figure 53. On y constate que 989 modèles (sur un total de 1720) sont considérés comme fiables.

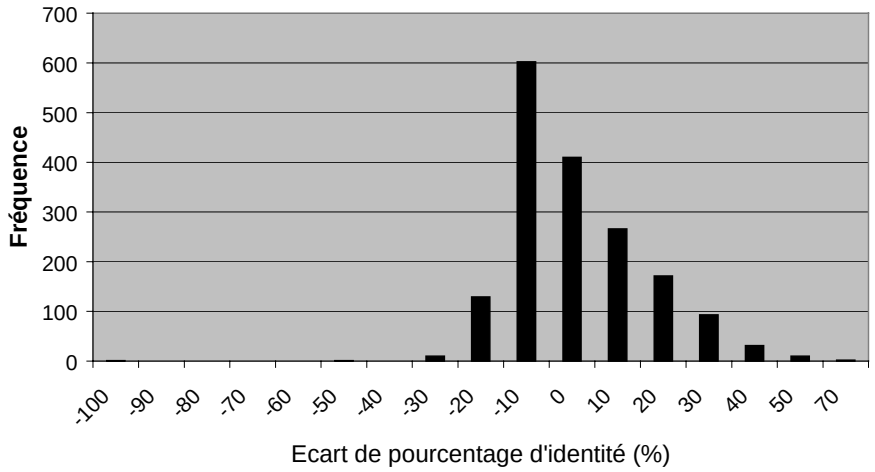


Figure 53: Distribution des écarts de pourcentage d'identité entre le pourcentage d'identité entre SI et SSC et le seuil $p^I(L)$.

La méthode que nous utilisions précédemment pour la détermination de la fiabilité des modèles permet donc seulement d'approcher la formule plus précise décrite par Rost *et al*. puisque nous ne nous intéressions qu'aux modèles représentant au moins 80% de la longueur des protéines déduites. Nous pouvons donc dès à présent prédire la qualité d'un modèle 3D en utilisant les caractéristiques de sa prédiction.

VI.5. Conclusions et perspectives

Nous avons réalisé une banque de données permettant de stocker de façon structurée des informations sur le génome et le protéome de la bactérie *Brucella melitensis*. Cette banque, accessible via le *web*, contient des données expérimentales publiées dans la littérature scientifique, ainsi que des résultats de prédictions. Pour ces dernières informations, le caractère théorique est clairement indiqué à l'utilisateur, et un indice de fiabilité leur est associé. De plus, les possibilités d'interrogation de notre banque de données permettent d'effectuer des requêtes qui ne sont pas possibles dans d'autres banques de données. Elle va donc nous permettre de mieux caractériser le génome de *Brucella melitensis* par une série de propriétés des pCDS et de DP. Par exemple, la prédiction des points isoélectriques et masses moléculaires de toutes les DP permettra une interprétation plus facile des gels 2D. Parmi les prédictions proposées à l'utilisateur, nous avons souhaité inclure des modèles 3D de protéines.

Le premier objectif que nous nous étions fixé (voir page 143) pour la prédiction de la structure 3D des protéines est déjà atteint. En effet, les utilisateurs de la banque de données peuvent télécharger des modèles 3D et les examiner avec CHIME (MDL Information Systems Inc, San Leandro, CA, USA) ou avec leur outil de visualisation favori. Le deuxième (recherche structurale) et le troisième objectif (*docking* de petites molécules) seront atteints dans un délai d'un an vu l'existence de méthodes bien documentées.

Néanmoins, dans son état actuel, cette banque de données est déjà d'une grande utilité pour la communauté des chercheurs étudiant *Brucella melitensis*. Nous en tenons pour preuve le nombre d'accès mensuels, et le nombre de propositions de modification des données. Ce succès est très probablement dû à la structure de l'interface de consultation qui permet un examen rapide mais approfondi des pCDS.

Cependant, il faut considérer cette banque de données comme une première étape vers la réalisation d'un système de stockage et de dissémination d'informations sur les génomes et protéomes de toutes les souches de *Brucella*. Nous comptons atteindre cet objectif par deux types de développements:

1. développements techniques, par l'utilisation d'un système de gestion de base de données (SGBD), qui permettra de mieux structurer l'information, créer une collection de données non redondantes, faciliter l'intégration de nouvelles propriétés, optimiser l'espace de stockage physique, accélérer l'accès aux données et permettre l'exécution de requêtes complexes. Un prototype de base de données réalisée avec le SGBD MySQL (Widenius *et al.*, 2002) a montré que les requêtes sont

beaucoup plus rapides qu'avec le système actuel et que l'espace disque utilisé était réduit de plus de 60%. De plus, il est possible d'interroger cette base de données prototype directement en langage SQL. Le nombre de requêtes possibles n'est dès lors plus limité par des formulaires prédéfinis mais seulement par le langage SQL et la structure de la banque de données relationnelle.

2. développements scientifiques, par l'introduction de méthodes de recherche de sites actifs dans des modèles 3D, ou encore de *docking* rapide d'un très grand nombre de ligands potentiels avec des récepteurs.

VII. Conclusion générale

Au cours de ce travail, nous avons développé une méthode d'alignement pairé, ESyPAliNN, qui s'est avérée être d'une très grande fiabilité. Celle-ci nous a permis de développer un système automatique de prédiction par homologie de la structure 3D des protéines, ESyPred3D. Cette méthode, classée parmi les meilleures méthodes de modélisation au CASP5, a permis de prédire une structure 3D fiable pour près de 28% des protéines déduites du génome de *Brucella melitensis*. Ces modèles ont été stockés dans une banque de données qui a été développée à cet effet, et qui contient également toute une série d'autres informations essentielles pour une meilleure caractérisation des pCDS et DP.

Le travail qui vient d'être résumé ici était organisé autour de trois thèmes de recherche:

(1) l'alignement des séquences de protéines

(2) la prédiction de la structure 3D des protéines

(3) la caractérisation du génome de *Brucella melitensis* et de ces protéines déduites.

Nous allons maintenant terminer cette thèse en présentant, pour chacun de ces trois thèmes, les conclusions générales que nous pouvons tirer de nos recherches et les perspectives qui s'ouvrent.

VII.1. Alignement de séquences

Dans la partie consacrée à l'amélioration de la qualité des alignements de séquences, toute augmentation de la qualité de l'alignement passait par un apport d'informations supplémentaires, que ce soit en utilisant plusieurs matrices de scores, en utilisant les prédictions de structures secondaires ou d'accessibilité au solvant, ou encore, en recherchant l'information dans les PSSM. Néanmoins, l'amélioration de la qualité de l'alignement de séquences était souvent très faible. Nous avons développé une méthode d'alignement pairé, ESyPAliNN, qui combine les résultats d'alignements d'autres méthodes, dont Match-Box, à l'aide de réseaux neuronaux. Nous avons montré que ESyPAliNN fournit de meilleurs résultats que chacun des programmes d'alignement utilisés pour générer l'alignement final. Cette combinaison de plusieurs programmes pour fournir un meilleur résultat est de plus en plus utilisée en bioinformatique pour améliorer les prédictions. Par exemple, cette approche a été utilisée dans la prédiction de structure secondaire (Geourjon and Deleage, 1995; Cuff *et al.*, 1998). Plus récemment, elle a été utilisée dans la prédiction de structures 3D (Arne Elofsson, unpublished) et a été, pour ainsi dire, « consacrée » aux CASP5 et CAFASP3 où les méta-serveurs ont obtenu les meilleurs résultats.

Notre méthode pourrait, nous semble-t-il, être encore améliorée en travaillant dans deux directions:

- Utilisation d'un nombre plus important de programmes d'alignement de séquences dans la partie *matching* et amélioration des programmes existants.
- Mise au point d'un meilleur algorithme pour générer l'alignement optimal. Par exemple, nous pourrions utiliser un algorithme génétique qui trouvera plus efficacement la meilleure solution que notre méthode exhaustive.

D'autre part, la transformation de notre méthode d'alignement pairé en une technique d'alignement multiple peut être envisagée suivant deux voies:

- la première, et la plus simple, consiste à aligner toutes les séquences par rapport à une seule séquence servant de guide pour construire l'alignement multiple final. Cette technique est utilisée, par exemple, dans PSI-BLAST et dans les programmes utilisant les HMMs (on aligne alors toutes les séquences au HMM),
- la seconde consiste à combiner toutes les paires d'acides aminés alignées et à rechercher l'alignement multiple qui maximise une fonction évaluant la qualité intrinsèque de l'alignement. De telles fonctions ont

déjà été développées: COFFEE (Notredame *et al*., 1998) ou NORMD (Thompson *et al*., 2001). Pour accélérer le temps calcul de notre méthode, un algorithme génétique pourrait alors être utilisé pour combiner au mieux les résultats des différents programmes.

Nous pensons que le développement d'un programme d'alignement multiple basé sur les idées qui viennent d'être présentées pourrait permettre de faire progresser d'autres domaines de recherche: la phylogénie, la prédiction de structure secondaire, la prédiction d'accessibilité au solvant, ainsi que toutes les méthodes devant utiliser des alignements multiples très fiables.

VII.2. Prédiction de structure 3D de protéines

La fiabilité de notre méthode d'alignement pairé, ESyPAliNN, explique en grande partie les bonnes performances de notre système de modélisation par homologie, ESyPred3D. Toute amélioration apportée au premier devrait avoir une répercussion positive sur le fonctionnement du second.

Les limites de ESyPred3D pourraient être repoussées grâce à deux types d'amélioration: des améliorations dépendantes de la méthode d'alignement et des améliorations indépendantes cette dernière.

VII.2.1. AMÉLIORATIONS DÉPENDANTES DE LA MÉTHODE D'ALIGNEMENT

La qualité des modèles générés par ESyPred3D est actuellement limitée par un seuil d'environ 20% d'identité entre SI et SSC. Cette limite est principalement due aux programmes d'alignement utilisés par ESyPAliNN qui ne peuvent aligner correctement des séquences partageant ces faibles taux d'identité. Néanmoins, à ces faibles taux d'identité, une autre méthode de sélection du *template* pourrait être envisagée.

Pour repousser cette limite de 20% d'identité et pour améliorer la sélection du meilleur *template*, il nous semble judicieux d'utiliser des programmes utilisant d'autres techniques de construction et d'évaluation de l'alignement. Nous pourrions utiliser plusieurs programmes de reconnaissance de *fold* tels que SAM-T02, FFAS ou PROSPECT et en déterminer le *fold* consensus. Cette méthode de consensus est utilisée dans les méta-serveurs qui ont obtenu les meilleurs performances aux CAFASP3 et CASP5. Nous pourrions également générer un alignement consensus en traitant les résultats de ces programmes avec notre méthode ESyPAliNN. Certains groupes du CASP5 utilisent déjà une technique similaire (Prasad *et al.*, 2003).

VII.2.2. AMÉLIORATIONS INDÉPENDANTES DE L'ALIGNEMENT DE SÉQUENCES

La qualité des modèles générés par ESyPred3D pourrait être améliorée de quatre manières différentes:

1. par une meilleure prédiction de l'orientation des chaînes latérales. Ceci pourrait être obtenue en appliquant le programme SCWRL (Canutescu *et*

al., 2003) qui semble être actuellement le meilleur programme dans ce domaine.

2. par une meilleure prédiction des *loops*. Deux solutions peuvent être envisagées. La première est de mieux utiliser les différentes options de prédiction du programme Modeller qui obtient de bonnes performances dans la prédiction des *loops*. La deuxième, à l'instar de ROBETTA (Chivian *et al*., 2003), est de combiner une approche de modélisation par homologie et une approche de modélisation *de novo*.
3. par la génération de plusieurs modèles à partir de *templates* différents et la sélection du meilleur modèle. Cette sélection peut être effectuée après évaluation énergétique ou sur base du score global de l'alignement de séquences. C'est une technique utilisée, par exemple, par le programme FUGUE (Shi *et al*., 2001).
4. par l'itération des étapes de modélisation, évaluation et modification de l'alignement. Cette technique a été décrite récemment par John *et al*. (John and Sali, 2003).

VII.3. Base de données

La réalisation d'un système performant de prédiction de structure par homologie nous a permis de modéliser une bonne partie des protéines déduites du génome de *Brucella melitensis*. La diffusion de ces prédictions dans la communauté scientifique internationale a nécessité la constitution d'une banque de données structurales et fonctionnelles du génome et du protéome de *Brucella melitensis*, ainsi qu'une interface conviviale de consultation.

Cette banque de données consacrée au génome et au protéome de *Brucella melitensis*, contient des données expérimentales et des prédictions bioinformatiques soigneusement validées. Elle est ainsi devenue un outil de travail indispensable pour les chercheurs européens travaillant sur *Brucella*.

Cette banque de données est actuellement une collection de fichiers et doit être transformée le plus rapidement possible en une base de données relationnelle qui structurera mieux l'information, qui permettra des recherches plus complexes et qui diminuera l'espace disque utilisé par la banque de données actuelle.

Nous avons prêté une grande attention à la fiabilité de l’information contenue dans notre banque de données. C'est pourquoi nous avons corrigé les positions des codons *start* des CDS prédites par la société Integrated Genomics Inc. Nous avons également, chaque fois que c'était possible, attribué un indice de fiabilité aux prédictions proposées.

De par le nombre important de requêtes possibles et d'informations fiables disponibles, nous pouvons dire que notre banque est l’une des meilleures banques de données consacrée à *Brucella melitensis*.

Néanmoins, cette banque pourrait être améliorée en y ajoutant les génomes des organismes proches de *Brucella* et en facilitant les comparaisons transversales telles que la comparaison des opérons dans différentes bactéries.

Notre banque de données pourrait être développée pour la transformer en un système de collecte, de validation, de stockage et de distribution d'information au sujet du génome et du protéome de cette bactérie. Nous pourrions par exemple:

1. ajouter des méthodes théoriques et faire des prédictions sur tout le génome. Les résultats de ces méthodes pourraient être comparés aux résultats expérimentaux connus
2. proposer des expériences sur base des résultats théoriques

3. susciter plus d'annotations de la part des utilisateurs et nommer des "referee" pour valider les annotations
4. envoyer un journal aux utilisateurs pour parler des nouveautés de la banque de données

Enfin, l'ensemble des modèles 3D des pCDS pourrait être mieux exploité par le développement d'un système de *screening* à grande échelle (*docking* de petites molécules contre les protéines du génome) ou par un outil de recherche de sites actifs. Ces modèles de bonne qualité pourraient également être utilisés pour prédire les réseaux d'interactions protéine-protéine dans un génome complet grâce à l'utilisation de programmes de *docking* protéine-protéine. Cette dernière application n'en est qu'à ses débuts mais certains auteurs l'ont déjà réalisée (Tovchigrechko *et al*., 2002; Lu *et al*., 2003). Une telle approche ne pourra être appliquée que lorsque les méthodes de *docking* protéine-protéine seront plus précises ou plus fiables.

VIII. Discussion

Nos travaux en bioinformatique nous ont permis d'acquérir une connaissance importante de l'état des recherches dans une grande variété de problématiques (alignement de séquences, prédiction de structure secondaire, modélisation par homologie, *threading*, recherche par similarité en base de données, etc ...). Dans certains de ces domaines de recherche, on peut dire que les développements de nouvelles méthodes est en stagnation. En effet, on constate que, faute de mieux, on développe des méthodes consensus pour le *threading*, la prédiction de structure secondaire, ou encore, ... l'alignement de séquences.

De même, on peut parler de retard inquiétant, en ce qui concerne la gestion de l'information biologique. Dans ce domaine, le nombre de séquences protéiques et nucléotidiques connues croît de façon exponentielle. Ce qui oblige d'effectuer des annotations par similarité sans que cela soit nécessairement clairement spécifié. Les banques du futur devront indiquer clairement l'origine de l'information contenue ainsi que la date à laquelle elles ont été introduites ou mises à jour. L'avenir passe probablement par une banque de données générale où toutes les informations sont stockées. On pourrait imaginer, par exemple, que l'information sur chaque génome soit gérée par des spécialistes de l'organisme en question. Il y a besoin d'un réel travail de nettoyage de l'information contenue dans les banques de données et d'une meilleure répartition de la gestion de cette information. Ce travail doit faire l'objet d'une collaboration mondiale pour une meilleure qualité de l'information et pour une croissance plus rapide de la connaissance des mécanismes qui se déroulent dans les êtres vivants.

Cependant, on peut observer des développements majeurs et soudains dans certains autres domaines. L'exemple le plus frappant est la mise au point de Rosetta, la méthode de prédiction de structure 3D *de novo* de Baker et ses collaborateurs.

Les chercheurs en bioinformatique arriveront-ils à sortir de cette période de stagnation que nous venons d'évoquer? L'examen du succès de la méthode "Rosetta" est instructif: on y décèle d'abord une grande créativité intellectuelle, soutenue par l'utilisation d'une puissance de calcul très importante. Ce seront deux conditions des succès futurs. Mais l'analyse des difficultés actuelles en matière, par exemple, de prédiction de structure secondaire ou d'alignement de séquences est également porteuse de leçons. L'absence de définition précise de ce qu'est un alignement (et surtout, un alignement vrai) montre le manque de rigueur de la bioinformatique. De plus, une analyse précise de la problématique de la prédiction de structure

secondaire pourrait mettre en évidence des limitations internes qui interdisent peut-être de nouveaux progrès.

Bref, créativité intellectuelle, rigueur et grande puissance de calcul seront trois conditions nécessaires des futurs développements de la bioinformatique.

IX. Annexes

Annexe 1: Review of common sequence alignment methods: clues to enhance reliability, Current Genomics **4**(2): 131-146 (2003)

Annexe 2: Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance predictions reliability, Bioinformatics **14**(4):357-366 (1998)

Annexe 3: Liste des familles de protéines de la banque de 78 alignements de référence

Annexe 4: Liste des familles de protéines de la banque de 420 alignements pairés de référence

Annexe 5: Evaluation de Match-Tal

Annexe 6: Sensibilité de l'algorithme de matching en fonction de la matrice de scores utilisée

Annexe 7: Liste des 134 matrices de scores tirées de la littérature

Annexe 8: Evolution de la sélectivité en fonction de la taille des segments, de la conservation de la structure secondaire, du critère de vérité et de la méthode utilisée

Annexe 9: Modeling of Human Monoamine Oxidase A: From Low Resolution Threading Models to Accurate Comparative Models Based on Crystal Structures, NeuroToxicology in press (2003)

Annexe 10: ESyPred3D: Prediction of proteins 3D structures, Bioinformatics **18**(9):1250-1256 (2002)

Annexe 11: Liste des protéines modélisées aux CASP5 et CAFASP3

Annexe 12: Classement obtenu au CASP5

Annexe 1: Review of common sequence alignment methods: clues to enhance reliability, Current Genomics 4(2): 131-146 (2003)

Annexe 2: Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance predictions reliability, Bioinformatics 14(4):357-366 (1998)

Annexe 3: Liste des familles de protéines de la banque de 78 alignements de référence

Tableau 25: Liste des familles de protéines de la banque de 78 alignements de référence. Le nombre de séquences de chaque famille est donné dans la colonne de droite.

Famille de protéines	Nombre de séquences
alpha beta-hydrolase	3
annexin	3
antibacterial protein	4
aspartic proteinase	11
azurin/plastocyanin	8
calcium-binding protein	5
calcium-binding protein	4
crystallin	5
Cu/Zn superoxide dismutase	4
cysteine proteinase	4
cytochrome p450	3
cytochrome-c	9
cytochrome-c3	4
cytochrome-c5	4
cytokine	3
dihydrofolate reductase	4
disulphide oxidoreductase	8
DNA-binding homeodomain	3
DNA-binding repressor	3
EGF-like domain	6
Fe/Mn superoxide dismutase	5
ferredoxin (2Fe-2S)	5
ferredoxin (4Fe-4S)	5
flavodoxin	5

globin	19
glutathione S-transferase	5
glyceraldehyde phosphate dehydrogenase	4
glycosyl hydrolase family 13	4
glycosyl hydrolase family 22	7
glycosyl hydrolase family 34	3
high potential iron protein	4
histidine carrier protein	3
histocompatibility antigen-binding domain	5
immunoglobulin domain	5
immunoglobulin domain	11
immunoglobulin domain	20
immunoglobulin domain	23
immunoglobulin domain	4
insulin	5
interleukin	5
interleukin 8-like protein	5
kringle domain	6
lactate/malate dehydrogenase	9
lipocalin	8
matrix metalloproteinase	3
metallothionein	3
metallothionein	3
nucleotide diphosphate kinase	3
nucleotide kinase	4
pancreatic ribonuclease	3
phospholipase A2	7
picornavirus coat proteins	6
plant lectin	5
retroviral proteinase	3
ribonuclease	3

ribonuclease H	3
ribulose-1,5-biphosphate carboxylase/oxygenase	3
ricin-like protein	3
rubredoxin	5
sea anemone toxin	3
serine proteinase	3
serine proteinase	14
serine proteinase inhibitor	8
serine proteinase inhibitor	6
serine proteinase inhibitor	3
serine proteinase inhibitor	3
serine proteinase inhibitor	6
SH2 domain	4
SH3 domain	5
snake toxin	13
subtilase	7
Sugar-binding-protein	3
thioredoxin	4
thymidylate synthase	3
triose phosphate isomerase	4
xylose isomerase	3
zinc finger	10
zinc metalloproteinase	3

Annexe 4: Liste des familles de protéines de la banque de 420 alignements pairés de référence

Tableau 26: Liste des familles de protéines de la banque de 420 alignements pairés de référence. Le pourcentage d'identité entre les protéines est repris dans la colonne de droite.

Famille de protéines	Pourcentage d'identité
3,4-dihydroxy-2-butanone 4-phosphate synthase	46
5'-3' exonuclease	22
6-phosphogluconate dehydrogenases	33
7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)	56
7kD DNA-binding domain	84
7S seed storage protein	50
Activin types I and II receptor domain	18
Acyl CoA binding protein	27
acylphosphatase	54
Adenylyl- / guanylyl cyclase, catalytic domain	29
ADP-ribosyl cyclase	33
Alanine dehydrogenase/pyridine nucleotide transhydrogenase	30
Aldehyde ferredoxin oxidoreductase	40
Alkaline phosphatase	32
Alpha adaptin AP2, C-terminal domain of C-terminal region	18
Alpha adaptin AP2, C-terminal region (consists of 2 domains)	16
Alpha-2-macroglobulin family A	66
Alpha-2-macroglobulin family B	84
Alphavirus core protein	68
Amidase	28
Amidinotransferase	39
Amino acid kinase family	47

Aminotransferase class IV	27
Aminotransferases class-V	46
AMP-binding enzyme	19
Anaphase-promoting complex, subunit 10 (APC10)	31
Anaphylatoxin homologous domain	68
animal haem peroxidase	22
AP endonuclease family 1	28
Apocytochrome F	71
Arginase family	43
arginine repressor, C-terminal domain	30
Arginosuccinate synthase	29
ARID/BRIGHT DNA binding domain	26
Arrestin (or S-antigen)	59
Arthropod defensin	39
Asp/Glu/Hydontoin racemase	18
Asparagine synthase	21
Aspartate carbamoyltransferase regulatory chain	84
Astacin (Peptidase family M12A)	32
ATP synthase, gamma subunit	24
ATP-sulfurylase	56
avidin	32
Bacterial DNA recombination protein, RuvA	33
Bacterial extracellular solute-binding proteins, family 5	25
bacterial lipase	84
Bacterial regulatory helix-turn-helix proteins, araC family, the two structural repeats	46
Bacterial regulatory proteins, merR family, N-terminal domain	29
Bacterial RNA polymerase, alpha chain C-terminal domain	31
Bacteriochlorophyll A protein	79
Bacterioferritin	47

BAG domain	81
Beta-eliminating lyase	49
beta-lactamase class B	34
Biopterin-dependent aromatic amino acid hydroxylase	65
BRCA1 C Terminus (BRCT) domain	30
C1q domain	48
C-5 cytosine-specific DNA methylase	31
Calpain family cysteine protease, catalytic domain (domain II)	67
Capsid protein (F protein)	67
Caspase recruitment domain	14
CAT RNA binding domain	41
Cdc48-like, domains 1 and 2	26
Cellulose binding domain	37
Cellulose binding domain family 2	26
Chaperonin 10 kD subunit	43
Chitinase class I	66
Chorismate mutase	15
CIDE-N domain	29
Class II Aldolase	25
CoA-ligase	45
Coenzyme A transferase	23
C-terminal tandem repeated domains in type 4 procollagen	64
Cutinase	21
Cyclopropane-fatty-acyl-phospholipid synthase	57
Cystine-knot domain	36
Cytidylyltransferase A	27
Cytidylyltransferase B	15
Cytochrome b	72
Cytochrome C and Quinol oxidase polypeptide I	54
Cytochrome C1 family	58

Cytochrome C1 family, N-terminal domain	61
Cytochrome c552	77
cytochrome cd1-nitrite reductase	67
D-alanine--D-alanine ligase	31
DBL homology domain	20
Defensin/corticostatin	25
DegT/DnrJ/EryC1/StrS family	21
Dehydratase large subunit	71
Dehydratase medium subunit	61
Dehydratase small subunit	56
Dehydrogenase E1 component	38
Dehydroquinase class II	43
delta endotoxin	35
Delta-5-3-ketosteroid isomerase, steroid delta-isomerase, KSI	34
Delta-aminolevulinic acid dehydratase	37
Deoxynucleoside kinase	38
Di-haem cytochrome c peroxidase	48
Dihydrodipicolinate synthetase family	24
Dihydroneopterin aldolase	20
Dihydropteroate synthase	38
DNA gyrase B, 2 N-terminal domains	48
DNA methylase	23
DNA mismatch repair protein, N-terminal domain and second domain	25
DNA photolyase	38
DNA polymerase X family	22
DnaJ molecular chaperone homology domain	47
Domain of unknown function DUF28	45
Dps protein family	26
DSBA oxidoreductase	40
dTDP-4-dehydrorhamnose 3,5-epimerase	49

Dual specificity phosphatase, catalytic domain	34
Dwarfin	49
E2 (early) protein, N terminal	47
E3-binding domain	32
Egg lysin (Sperm-lysin)	68
eIF1-like	32
eIF-6 family	33
Electron transfer flavoprotein alpha subunit	55
Electron transfer flavoprotein beta subunit	59
Elicitin	87
Endonuclease III	20
Endothelin	67
enolase	64
Envelope glycoprotein GP120	88
Envelope Polyprotein GP41	37
EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)	22
Eukaryotic DNA topoisomerase I	53
Eukaryotic DNA topoisomerase I, catalytic core	18
Eukaryotic initiation factor 1A	32
Eukaryotic initiation factor 4E	28
Eukaryotic initiation factor 5A hypusine (eIF-5A)	45
F5/8 type C domain	42
FAD dependent oxidoreductase	29
Ferrochelatase	27
FHA domain	19
Fibronectin type 2 domain	40
Flavin containing amine oxidase	19
Flavin reductase like domain	15
Flavoprotein	28
formate/glycerate dehydrogenases	27

Frataxin-like domain	25
Fructose-bisphosphate aldolase class-II	24
Fusion glycoprotein F0	13
FYVE zinc finger	45
Fz domain	46
GA module	52
gag gene protein p17 (matrix protein)	53
GHMP kinases putative ATP-binding protein	15
glucagon	90
Glucosamine-6-phosphate isomerases/6-phosphogluconolactonase	58
glucose-6-phosphate 1-dehydrogenase	32
Glutamine synthetase	52
glycosyl hydrolase family 28	19
glycosyl hydrolase family 46 (chitosanase)	23
Glycosyl hydrolase family 47	42
Glycosyl hydrolase family 48	61
glycosyl hydrolases family 15	36
glycosyl hydrolases family 16	77
glycosyl hydrolases family 17	52
Glycosyl transferase family 8	19
glycosyl-asparaginase	36
Glycosyltransferase family 6	44
Gram-negative pili assembly chaperone	33
Green/red fluorescent protein	23
GTP cyclohydrolase I	36
GTPase-activator protein for Rho-like GTPases	17
haemocyanin	34
haloacid dehalogenase-like hydrolase	45
Ham1 family	32
heat shock protein 90, N-terminal domain	70

hemerythrin	46
Hint (Hedgehog/Intein) domain	15
Histidine biosynthesis protein	21
Histone-like transcription factors (CBF/NF-Y) and archaeal histones	84
HIT family	20
holliday junction resolvase	31
HSF-type DNA-binding domain	43
Hsp20/alpha crystallin family	27
Hydroxymethylglutaryl-coenzyme A reductase	22
inosine-uridine preferring nucleoside hydrolase	79
Insect virus proteins	50
Insulinase (Peptidase family M16)	22
Integrase DNA binding domain	22
Integrase Zinc binding domain	54
Interferon alpha, beta and delta.	35
Interferon gamma	60
Interferon regulatory factor transcription factor	56
interleukin-10 (IL-10)	92
Iron dependent repressor, N-terminal DNA binding domain	85
Iron only hydrogenase large subunit, C-terminal region	44
Iron/Ascorbate oxidoreductase	20
Iron-containing alcohol dehydrogenase	48
Isoamylase and glycosyltrehalose trehalohydrolase	18
Isochorismatase family	15
Isocitrate lyase family	40
Jacalin-like lectin domain	80
KH-domain	24
KU domain	17
LEM domain	40
leucine rich repeats in splicesomal U2A' protein and	19

internalin B	
LIF / OSM family	23
light-harvesting complex II	33
linker histone H1 and H5 family	42
Lipoprotein family 6	67
Lipoxygenase	73
Major intrinsic protein	29
Major spike protein (G protein)	41
Malic enzyme	53
maltogenic amylase	46
MDM2 oncoprotein	73
methane monooxygenase component A alpha chain	82
methane monooxygenase component A beta chain	60
methane monooxygenase component A gamma chain	51
Methanol dehydrogenase beta subunit	74
methyl-accepting chemotaxis protein II	68
Methylamine dehydrogenase, L chain	79
Methyl-CpG binding domain	39
MGS-like domain	16
Microtubule associated protein 1A/1B, light chain 3	58
MIF4G domain	14
Molybdate-binding protein, ModA	26
Molybdenum-binding protein C-terminal domain	35
MSP (Major sperm protein) domain	82
Mu DNA-binding domain	39
Multicopper oxidase	29
MutS family	43
MutT-like domain	22
Myo-inositol-1-phosphate synthase	19
Myosin, Large ATPases	50
Myristoyl-CoA	57

N-acetylmuramoyl-L-alanine amidase	28
N-acetyltransferase	34
NAD(P) transhydrogenase beta subunit	46
NAD-dependent DNA ligase adenylation domain	44
Neurohypophysial hormones, C-terminal domain	85
Neurotransmitter-gated ion-channel transmembrane region	52
Ni-containing carbon monoxide dehydrogenase	57
Nitrile hydratase beta subunit	34
Nitrile hydratase, alpha chain	45
Nitrogen regulatory protein P-II	63
nitrogenase iron protein	70
Nitrophorin	46
Nucleotidyltransferase, N-terminal domain and domain 2	44
NusB family	33
O-methyltransferase	28
Orotidine 5'-phosphate decarboxylases	22
Oxidoreductase family	19
P53	90
PAC motif	31
Paired amphipathic helix repeat	63
Paired box domain	61
PAN domain	13
Parathyroid	8
Parvovirus coat protein VP2	53
Pathogenesis-related protein Bet v I family	59
PBP/GOBP family	16
Pectinesterase	34
Peptidase C1-like family	40
Peptidase family M10A	57
Peptidase family S24	33

Peptide methionine sulfoxide reductase	60
PEP-utilizing enzyme	57
periplasmic binding protein - branched-chain amino acid	79
Periplasmic solute binding protein family	31
Pertussis toxin, subunit 2 and 3	71
PHBH-like	17
Phosphatidylethanolamine-binding protein	29
phosphatidylinositol-specific phospholipase C - bacterial	31
Phosphoenolpyruvate carboxykinase	44
phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 2	23
phosphofructokinase	55
Phosphoglucomutase/phosphomannomutase	54
Phosphoglucose isomerase	22
Phospholipase/Carboxylesterase	33
Phosphoribulokinase / Uridine kinase family	17
Phosphotriesterase-like	30
Photosystem I psaA/psaB protein	48
Phycoerythrin, alpha/beta chain	64
Plant proteinase inhibitors	69
Platelet activating factor acetylhydrolase	66
Poly-adenylate binding protein, unique domain	54
Polypeptide deformylase	31
Polysaccharide lyase family 8	21
PotD/PotF	36
PPIC-type PPIASE domain	38
Prismane	67
Proliferating cell nuclear antigen	36
prolyl aminopeptidase	57
Protease inhibitor Inh	36
Protein of unknown function DUF101	21

Protein prenyltransferase alpha subunit repeat	22
protocatechuate-3,4-dioxygenase, alpha and beta chains	28
Putative esterase	73
Putative undecaprenyl diphosphate synthase	38
PX domain	25
Pyridoxal-dependent decarboxylase	67
Pyridoxal-dependent decarboxylase, pyridoxal binding domain	69
Pyridoxal-phosphate dependent enzymes	21
Pyridoxamine 5'-phosphate oxidase	42
Pyroglutamyl peptidase	36
Quinolinate phosphoribosyl transferase	42
RanBP1 domain	63
Ras association (RalGDS/AF-6) domain	46
recA bacterial DNA recombination protein	63
Rel homology domain (RHD).	59
Repeats in polycystic kidney disease 1 (PKD1) and other proteins	16
retroviral integrase	24
Reverse transcriptase (RNA-dependent DNA polymerase)	26
RHO protein GDP dissociation inhibitor	76
Ribokinase-like	23
Ribonuclease P	49
ribonuclease T2	29
Ribonuclease U2	87
Ribosomal L18p/L5e family	29
Ribosomal protein L14p/L23e	42
Ribosomal protein L15	28
Ribosomal protein L22p/L17e	23
Ribosomal protein L30p/L7e	19
Ribosomal protein L4/L1 family	24
Ribosomal protein L6	24

Ribosomal protein L7/L12 C-terminal domain	71
Ribosomal protein S15	60
Ribosomal protein S1-like RNA-binding domain	21
Ribosomal protein S5	55
Ribosomal protein S7	56
Ribosomal protein S8	62
Ribosomal RNA adenine dimethylases	50
Rieske iron-sulfur protein (ISP), watersoluble domain	33
RNA polymerases H / 23 kDa subunit	44
RNA polymerases K / 14 to 18 kDa subunit	71
RNA polymerases N / 8 kDa subunit	46
Rop protein	81
S-adenosylmethionine synthetase	58
SAICAR synthetase	25
SCP-2 sterol transfer family	40
SCP-like extracellular protein	32
Sec1 family	68
Sec7 domain	86
Serine hydroxymethyltransferase	47
serine/threonine protein phosphatase	39
SET domain	42
Shiga-like toxin beta subunit	62
Shikimate kinase	34
Sigma-54 transcription factors	26
Signal peptide-binding domain	30
Similarity to lectin domain of ricin beta-chain, 3 copies.	60
Single-strand binding proteins	36
Sir2 family	29
Skp1 family	36
SNAP-25 family	63
Somatotropin hormone family	89

Spectrin repeats	21
Spermine/spermidine synthase	35
spider toxin	71
SRP19 protein	27
SRP54-type protein	35
Staphylokinase/Streptokinase family	13
START domain	22
STAS domain	31
STAT protein, domains 2-4	56
Sterile alpha motif (SAM)/Pointed domain	36
Succinate dehydrogenase/fumarate reductase	33
Sulfatase	29
Syntaxin	16
Tautomerase enzyme	76
TAZ zinc finger	29
T-box	52
TBP-associated factors	20
T-cell prolymphocytic leukemia oncogenes	40
Tetrahydrofolate dehydrogenase/cyclohydrolase	44
Tetramerization domain of potassium channels	37
Thioesterase domain	19
Thiolase/Beta-ketoacyl synthases	17
Thymidine and pyrimidine-nucleoside phosphorylases	43
thymidylate kinase	29
TIR domain	54
Tissue inhibitor of metalloproteinase family	43
TonB-dependent receptor proteins	17
Transaldolase	59
Transcription factor S-II (TFIIS)	40
Transcription factor TFIIB repeat	32
Transcriptional regulatory protein, C-terminal	34

Transglutaminase-like superfamily	36
Transglycosylase SLT domain	22
Translation initiation factor IF-3	48
TrkA-N domain	21
tRNA synthetases class I (R)	24
tRNA synthetases class I (W and Y)	12
Trypanosome variant surface glycoprotein	17
Tryptophan RNA-binding attenuator protein	82
Tryptophan synthase alpha chain	33
Tubulin	40
Tubulin binding cofactor A	31
two-domain cytochrome c	27
Type II DNA topoisomerase, domains 2-4	22
Ubiquinol-cytochrome C reductase complex 14kD subunit	36
Ubiquitin carboxyl-terminal hydrolase, family 1	35
UBX domain	26
Uncharacterized protein family UPF0033	24
Universal stress protein family	20
Urease beta subunit	54
Urease, alpha subunit	64
Urease, gamma subunit	74
UreE urease accessory protein	21
Uroporphyrinogen decarboxylase (URO-D)	34
Uteroglobin	56
UvrD/REP helicase	43
vanadium-dependent haloperoxidase	35
viral capsid protein (n-terminal domain)	23
Viral proteases	89
Voltage gated chloride channel	81
VPR/VPX protein	15
WD domain, G-beta repeat	22

WH1 domain	73
WHEP-TRS domain	63
XPG	53
yrdC domain	25
Zinc dependent phospholipase C	26
zinc finger -- CCHC-type	47

Annexe 5: Evaluation de Match-Tal

Tableau 27: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores Johnson92, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: Johnson92		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,6	53,2
Match-Tal (*I*≤2)	62,8	52,7
Match-Tal (*I*≤3)	62,7	53,6
Match-Tal (*I*≤4)	63,0	54,3
Match-Tal (*I*≤5)	63,3	54,5
Match-Tal (*I*≤6)	61,1	55,4
Match-Tal (*I*≤7)	59,9	54,8
Match-Tal (*I*≤8)	59,9	54,7
Match-Tal (*I*≤9)	59,4	53,6
Match-Box	57,3	57,7

Tableau 28: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores Johnson96, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: Johnson96		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,6	53,2
Match-Tal (*I*≤2)	62,8	52,9
Match-Tal (*I*≤3)	61,8	52,9
Match-Tal (*I*≤4)	60,7	52,9
Match-Tal (*I*≤5)	61,3	53,7
Match-Tal (*I*≤6)	61,6	53,6
Match-Tal (*I*≤7)	58,2	52,2
Match-Tal (*I*≤8)	58,0	52,5
Match-Tal (*I*≤9)	57,5	51,3
Match-Box	55,2	54,6

Tableau 29: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores Blosum45, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: Blosum45		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,5	53,2
Match-Tal (*I*≤2)	63,5	53,2
Match-Tal (*I*≤3)	60,4	50,6
Match-Tal (*I*≤4)	61,5	53,1
Match-Tal (*I*≤5)	57,2	49,3
Match-Tal (*I*≤6)	57,3	49,4
Match-Tal (*I*≤7)	56,5	49,9
Match-Tal (*I*≤8)	56,4	50,2
Match-Tal (*I*≤9)	56,0	50,3
Match-Box	53,6	53,8

Tableau 30: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores Blosum62, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: Blosum62		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,3	53,0
Match-Tal (*I*≤2)	61,3	51,8
Match-Tal (*I*≤3)	62,2	53,2
Match-Tal (*I*≤4)	64,0	54,7
Match-Tal (*I*≤5)	62,2	53,8
Match-Tal (*I*≤6)	62,2	54,3
Match-Tal (*I*≤7)	59,3	53,2
Match-Tal (*I*≤8)	58,9	52,5
Match-Tal (*I*≤9)	58,3	52,2
Match-Box	54,7	56,4

Tableau 31: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores Blosum80, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: Blosum80		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,6	53,3
Match-Tal (*I*≤2)	61,4	51,9
Match-Tal (*I*≤3)	62,6	53,4
Match-Tal (*I*≤4)	63,3	54,2
Match-Tal (*I*≤5)	62,4	54,3
Match-Tal (*I*≤6)	62,7	54,5
Match-Tal (*I*≤7)	60,0	54,4
Match-Tal (*I*≤8)	60,4	54,9
Match-Tal (*I*≤9)	58,9	52,5
Match-Box	58,1	54,9

Tableau 32: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores PAM120, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: PAM120		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,6	53,3
Match-Tal (*I*≤2)	62,8	52,7
Match-Tal (*I*≤3)	63,8	54,6
Match-Tal (*I*≤4)	63,1	54,7
Match-Tal (*I*≤5)	60,9	52,9
Match-Tal (*I*≤6)	61,5	53,9
Match-Tal (*I*≤7)	60,8	55,2
Match-Tal (*I*≤8)	60,8	55,1
Match-Tal (*I*≤9)	60,3	54,2
Match-Box	56,5	58,9

Tableau 33: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores PAM200, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: PAM200		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,7	53,3
Match-Tal (*I*≤2)	61,8	52,4
Match-Tal (*I*≤3)	62,6	53,9
Match-Tal (*I*≤4)	60,0	52,4
Match-Tal (*I*≤5)	59,3	51,8
Match-Tal (*I*≤6)	59,2	51,8
Match-Tal (*I*≤7)	58,6	52,6
Match-Tal (*I*≤8)	58,6	53,1
Match-Tal (*I*≤9)	58,3	52,5
Match-Box	56,3	56,6

Tableau 34: Evaluation de la sensibilité et de la sélectivité du programme Match-Tal pour différentes valeurs seuil de l'indice de confiance de Match-Box, pour la matrice de scores PAM250, et en utilisant la banque de 33 alignements de référence. Par comparaison, les valeurs de sensibilité et de sélectivité des programmes Match-Box et ClustalW sont indiquées.

Matrice de scores: PAM250		
	Sensibilité (%)	**Sélectivité (%)**
ClustalW	63,8	53,3
Match-Tal (*I*≤1)	63,6	53,2
Match-Tal (*I*≤2)	63,5	53,2
Match-Tal (*I*≤3)	63,9	53,6
Match-Tal (*I*≤4)	61,5	52,7
Match-Tal (*I*≤5)	58,2	51,4
Match-Tal (*I*≤6)	55,6	50,0
Match-Tal (*I*≤7)	55,6	50,7
Match-Tal (*I*≤8)	55,0	49,8
Match-Tal (*I*≤9)	55,0	49,4
Match-Box	53,6	51,4

Annexe 6: Sensibilité de l'algorithme de matching en fonction de la matrice de scores utilisée

Tableau 35: Sensibilité de l'algorithme de *matching_SF* en fonction de la matrice de scores utilisée.

Matrice	Sensibilité	Matrice	Sensibilité
JOHNSON92	70,7	PAM160	68,1
PAM120	70,5	PAM230	67,3
GONNET	70,5	PAM220	67,3
PAM200	70,4	PAM280	67,3
JOHNSON96	70,3	PAM250	67,2
BLOSUM100	70,2	PAM260	67,2
BLOSUM90	70,0	BLOSUM30	67,2
BLOSUM45	70,0	PAM240	67,1
BLOSUM60	70,0	PAM290	67,0
BLOSUM85	70,0	PAM300	66,7
BLOSUM55	70,0	PAM310	66,7
BLOSUM70	69,9	PAM330	66,4
BLOSUM75	69,8	PAM500	66,2
BLOSUM65	69,8	PAM340	65,9
BLOSUM62	69,8	PAM320	65,8
BLOSUM80	69,8	PAM370	64,9
BLOSUM35	69,6	PAM360	64,9
BLOSUM50	69,5	PAM380	64,5
BLOSUM40	69,4	PAM390	64,4
PAM150	68,7	PAM400	64,0
PAM110	68,7	PAM430	63,6
PAM100	68,6	PAM450	63,5
PAM190	68,4	PAM440	63,3
PAM210	68,3	PAM460	62,5

PAM130	68,3	PAM490	62,2
PAM140	68,2		

Annexe 7: Liste des 134 matrices de scores tirées de la littérature

Tableau 36: Liste des 134 matrices de scores tirées de la littérature pour l'amélioration des performances du *matching*.

Matrice	Description
ALTS910101	*The PAM-120 matrix* (Altschul, 1991)
AZAE970101	*The single residue substitution matrix from interchanges of spatially neighbouring residues* (Azarya-Sprinzak *et al*., 1997)
AZAE970102	*The substitution matrix derived from spatially conserved motifs* (Azarya-Sprinzak *et al*., 1997)
BENS940101	*Log-odds scoring matrix collected in 6.4-8.7 PAM* (Benner *et al*., 1994)
BENS940102	*Log-odds scoring matrix collected in 22-29 PAM* (Benner *et al*., 1994)
BENS940103	*Log-odds scoring matrix collected in 74-100 PAM* (Benner *et al*., 1994)
BENS940104	*Genetic code matrix* (Benner *et al*., 1994)
BLOSUM30	*BLOSUM30 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM35	*BLOSUM35 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM40	*BLOSUM40 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM45	*BLOSUM45 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM50	*BLOSUM50 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM55	*BLOSUM55 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM60	*BLOSUM60 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM62	*BLOSUM62 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM65	*BLOSUM65 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM70	*BLOSUM70 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM75	*BLOSUM75 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM80	*BLOSUM80 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM85	*BLOSUM85 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM90	*BLOSUM90 substitution matrix* (Henikoff and Henikoff, 1992)
BLOSUM100	*BLOSUM100 substitution matrix* (Henikoff and Henikoff, 1992)

CSEM940101	*Residue replace ability matrix* (Cserzo *et al*., 1994)
DAYM780301	*Log odds matrix for 250 PAMs* (Dayhoff *et al*., 1978)
DFK	(Depiereux and Feytmans, 1994)
FEND850101	*Structure-Genetic matrix* (Feng *et al*., 1984)
FITW660101	*Mutation values for the interconversion of amino acid pairs* (Fitch, 1966)
GEOD900101	*Hydrophobicity scoring matrix* (George *et al*., 1990)
GONG920101	*The mutation matrix for initially aligning* (Gonnet *et al*., 1992)
GONNET	*The mutation matrix for initially aligning* (Gonnet *et al*., 1992)
GRAR740104	*Amino acid difference formula to help explain protein evolution* (Grantham, 1974)
HENS920101	*BLOSUM45 substitution matrix* (Henikoff and Henikoff, 1992)
HENS920102	*BLOSUM62 substitution matrix* (Henikoff and Henikoff, 1992)
HENS920103	*BLOSUM80 substitution matrix* (Henikoff and Henikoff, 1992)
IDENT	Matrice identité
JOHM930101	*Structure-based amino acid scoring table* (Johnson and Overington, 1993)
JOHNSON92	*Structure-based amino acid scoring table* (Johnson and Overington, 1993)
JOHNSON96	*STR matrix from structure-based alignments* (Overington *et al*., 1992)
JOND920103	*The 250 PAM PET91 matrix* (Jones *et al*., 1992)
JOND940101	*The 250 PAM transmembrane protein exchange matrix* (Jones *et al*., 1994)
KOLA920101	*Conformational similarity weight matrix* (Kolaskar and Kulkarni-Kale, 1992)
KOSJ950101	*Context-dependent optimal substitution matrices for exposed helix* (Koshi and Goldstein, 1995)
KOSJ950102	*Context-dependent optimal substitution matrices for exposed beta* (Koshi and Goldstein, 1995)
KOSJ950103	*Context-dependent optimal substitution matrices for exposed turn* (Koshi and Goldstein, 1995)
KOSJ950104	*Context-dependent optimal substitution matrices for exposed coil* (Koshi and Goldstein, 1995)
KOSJ950105	*Context-dependent optimal substitution matrices for buried helix*

	(Koshi and Goldstein, 1995)
KOSJ950106	*Context-dependent optimal substitution matrices for buried beta* (Koshi and Goldstein, 1995)
KOSJ950107	*Context-dependent optimal substitution matrices for buried turn* (Koshi and Goldstein, 1995)
KOSJ950108	*Context-dependent optimal substitution matrices for buried coil* (Koshi and Goldstein, 1995)
KOSJ950109	*Context-dependent optimal substitution matrices for alpha helix* (Koshi and Goldstein, 1995)
KOSJ950110	*Context-dependent optimal substitution matrices for beta sheet* (Koshi and Goldstein, 1995)
KOSJ950111	*Context-dependent optimal substitution matrices for turn* (Koshi and Goldstein, 1995)
KOSJ950112	*Context-dependent optimal substitution matrices for coil* (Koshi and Goldstein, 1995)
KOSJ950113	*Context-dependent optimal substitution matrices for exposed residues* (Koshi and Goldstein, 1995)
KOSJ950114	*Context-dependent optimal substitution matrices for buried residues* (Koshi and Goldstein, 1995)
KOSJ950115	*Context-dependent optimal substitution matrices for all residues* (Koshi and Goldstein, 1995)
LEVJ860101	*The secondary structure similarity matrix* (Levin *et al*., 1986)
LUTR910101	*Structure-based comparison table for outside other class* (Luthy *et al*., 1991)
LUTR910102	*Structure-based comparison table for inside other class* (Luthy *et al*., 1991)
LUTR910103	*Structure-based comparison table for outside alpha class* (Luthy *et al*., 1991)
LUTR910104	*Structure-based comparison table for inside alpha class* (Luthy *et al*., 1991)
LUTR910105	*Structure-based comparison table for outside beta class* (Luthy *et al*., 1991)
LUTR910106	*Structure-based comparison table for inside beta class* (Luthy *et al*., 1991)
LUTR910107	*Structure-based comparison table for other class* (Luthy *et al*., 1991)
LUTR910108	*Structure-based comparison table for alpha helix class* (Luthy *et al*., 1991)

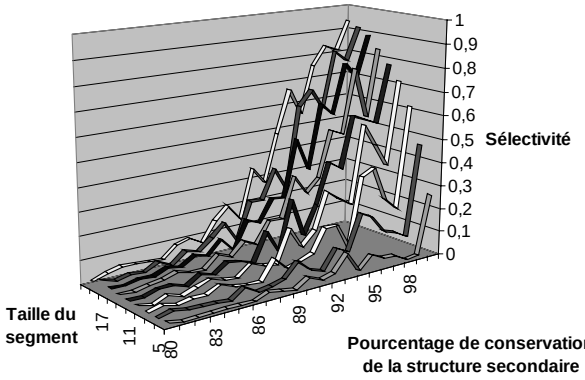
LUTR910109	*Structure-based comparison table for beta strand class* (Luthy *et al*., 1991)
MCLA710101	*The similarity of pairs of amino acids* (McLachlan, 1971)
MCLA720101	*Chemical similarity scores* (McLachlan, 1972)
MIYS930101	*Base-substitution-protein-stability matrix* (Miyazawa and Jernigan, 1993)
MIYT790101	*Amino acid pair distance* (Miyata *et al*., 1979)
MOHR870101	*EMPAR matrix* (Mohana Rao, 1987)
NIEK910101	*Structure-derived correlation matrix 1* (Niefind and Schomburg, 1991)
NIEK910102	*Structure-derived correlation matrix 2* (Niefind and Schomburg, 1991)
OVEJ920101_1	*STR matrix from structure-based alignments* (Overington *et al*., 1992)
OVEJ920101	*Environment-specific amino acid substitution matrix for alpha residues* (Overington *et al*., 1992)
OVEJ920102	*Environment-specific amino acid substitution matrix for beta residues* (Overington *et al*., 1992)
OVEJ920103	*Environment-specific amino acid substitution matrix for accessible residues* (Overington *et al*., 1992)
OVEJ920104	*Environment-specific amino acid substitution matrix for inaccessible residues* (Overington *et al*., 1992)
PAM10	*Log odds matrix for 10 PAMs* (Dayhoff *et al*., 1978)
PAM20	*Log odds matrix for 20 PAMs* (Dayhoff *et al*., 1978)
PAM30	*Log odds matrix for 30 PAMs* (Dayhoff *et al*., 1978)
PAM40	*Log odds matrix for 40 PAMs* (Dayhoff *et al*., 1978)
PAM50	*Log odds matrix for 50 PAMs* (Dayhoff *et al*., 1978)
PAM60	*Log odds matrix for 60 PAMs* (Dayhoff *et al*., 1978)
PAM70	*Log odds matrix for 70 PAMs* (Dayhoff *et al*., 1978)
PAM80	*Log odds matrix for 80 PAMs* (Dayhoff *et al*., 1978)
PAM90	*Log odds matrix for 90 PAMs* (Dayhoff *et al*., 1978)
PAM100	*Log odds matrix for 100 PAMs* (Dayhoff *et al*., 1978)
PAM110	*Log odds matrix for 110 PAMs* (Dayhoff *et al*., 1978)
PAM120	*Log odds matrix for 120 PAMs* (Dayhoff *et al*., 1978)

PAM130	*Log odds matrix for 130 PAMs* (Dayhoff *et al*., 1978)
PAM140	*Log odds matrix for 140 PAMs* (Dayhoff *et al*., 1978)
PAM150	*Log odds matrix for 150 PAMs* (Dayhoff *et al*., 1978)
PAM160	*Log odds matrix for 160 PAMs* (Dayhoff *et al*., 1978)
PAM170	*Log odds matrix for 170 PAMs* (Dayhoff *et al*., 1978)
PAM180	*Log odds matrix for 180 PAMs* (Dayhoff *et al*., 1978)
PAM190	*Log odds matrix for 190 PAMs* (Dayhoff *et al*., 1978)
PAM200	*Log odds matrix for 200 PAMs* (Dayhoff *et al*., 1978)
PAM210	*Log odds matrix for 210 PAMs* (Dayhoff *et al*., 1978)
PAM220	*Log odds matrix for 220 PAMs* (Dayhoff *et al*., 1978)
PAM230	*Log odds matrix for 230 PAMs* (Dayhoff *et al*., 1978)
PAM240	*Log odds matrix for 240 PAMs* (Dayhoff *et al*., 1978)
PAM250	*Log odds matrix for 250 PAMs* (Dayhoff *et al*., 1978)
PAM260	*Log odds matrix for 260 PAMs* (Dayhoff *et al*., 1978)
PAM270	*Log odds matrix for 270 PAMs* (Dayhoff *et al*., 1978)
PAM280	*Log odds matrix for 280 PAMs* (Dayhoff *et al*., 1978)
PAM290	*Log odds matrix for 290 PAMs* (Dayhoff *et al*., 1978)
PAM300	*Log odds matrix for 300 PAMs* (Dayhoff *et al*., 1978)
PAM310	*Log odds matrix for 310 PAMs* (Dayhoff *et al*., 1978)
PAM320	*Log odds matrix for 320 PAMs* (Dayhoff *et al*., 1978)
PAM330	*Log odds matrix for 330 PAMs* (Dayhoff *et al*., 1978)
PAM340	*Log odds matrix for 340 PAMs* (Dayhoff *et al*., 1978)
PAM350	*Log odds matrix for 350 PAMs* (Dayhoff *et al*., 1978)
PAM360	*Log odds matrix for 360 PAMs* (Dayhoff *et al*., 1978)
PAM370	*Log odds matrix for 370 PAMs* (Dayhoff *et al*., 1978)
PAM380	*Log odds matrix for 380 PAMs* (Dayhoff *et al*., 1978)
PAM390	*Log odds matrix for 390 PAMs* (Dayhoff *et al*., 1978)
PAM400	*Log odds matrix for 400 PAMs* (Dayhoff *et al*., 1978)
PAM410	*Log odds matrix for 410 PAMs* (Dayhoff *et al*., 1978)
PAM420	*Log odds matrix for 420 PAMs* (Dayhoff *et al*., 1978)
PAM430	*Log odds matrix for 430 PAMs* (Dayhoff *et al*., 1978)

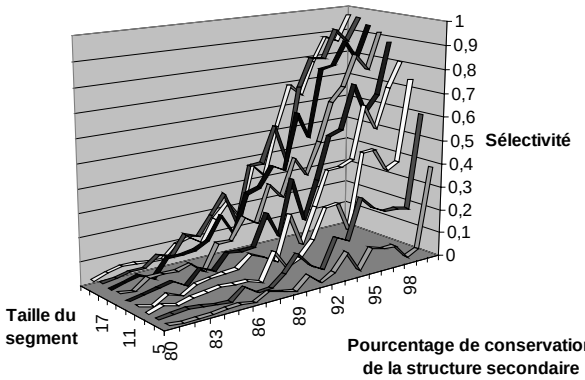
PAM440	*Log odds matrix for 440 PAMs* (Dayhoff *et al*., 1978)
PAM450	*Log odds matrix for 450 PAMs* (Dayhoff *et al*., 1978)
PAM460	*Log odds matrix for 460 PAMs* (Dayhoff *et al*., 1978)
PAM470	*Log odds matrix for 470 PAMs* (Dayhoff *et al*., 1978)
PAM480	*Log odds matrix for 480 PAMs* (Dayhoff *et al*., 1978)
PAM490	*Log odds matrix for 490 PAMs* (Dayhoff *et al*., 1978)
PAM500	*Log odds matrix for 500 PAMs* (Dayhoff *et al*., 1978)
QU_C930101	*Cross-correlation coefficients of preference factors main chain* (Qu *et al*., 1993)
QU_C930102	*Cross-correlation coefficients of preference factors side chain* (Qu *et al*., 1993)
QU_C930103	*The mutant distance based on spatial preference factor* (Qu *et al*., 1993)
RIER950101	*Hydrophobicity scoring matrix* (Riek *et al*., 1995)
RISJ880101	*Scoring matrix* (Risler *et al*., 1988)
TUDE900101	*Isomorphicity of replacements* (Tudos *et al*., 1990)
WEIL970101	*WAC matrix constructed from amino acid comparative profiles* (Wei *et al*., 1997)
WEIL970102	*Difference matrix obtained by subtracting the BLOSUM62 from the WAC matrix* (Wei *et al*., 1997)

Annexe 8: Evolution de la sélectivité en fonction de la taille des segments, de la conservation de la structure secondaire, du critère de vérité et de la méthode utilisée

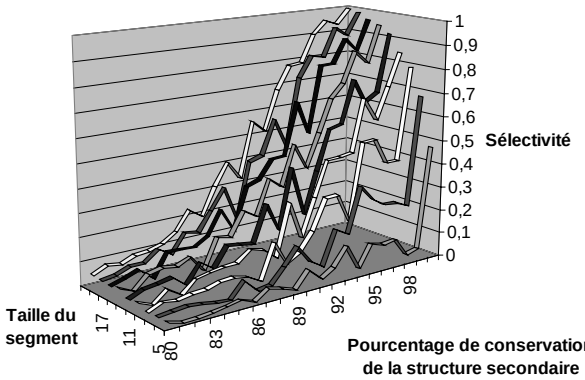
PHD - Evolution de la sélectivité en fonction de la conservation de la SS et de la taille du segment d'analyse pour le critère CSS



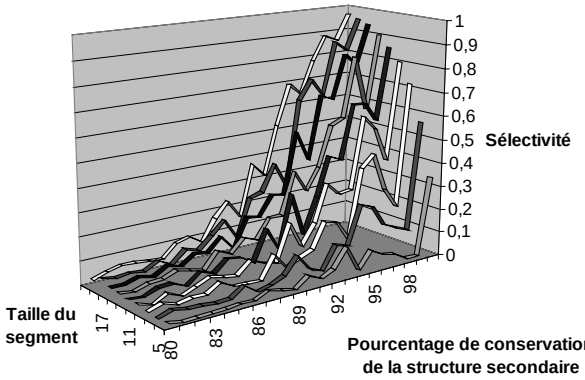
DSSP - Evolution de la sélectivité en fonction de la conservation de la SS et de la taille du segment d'analyse pour le critère CSS



DSSP - Evolution de la sélectivité en fonction de la conservation de la SS et de la taille du segment d'analyse pour le critère ASG



PHD - Evolution de la sélectivité en fonction de la conservation de la SS et de la taille du segment d'analyse pour le critère ASG



Annexe 9: Modeling of Human Monoamine Oxidase A: From Low Resolution Threading Models to Accurate Comparative Models Based on Crystal Structures, NeuroToxicology in press (2003)

Annexe 10: ESyPred3D: Prediction of proteins 3D structures, Bioinformatics 18(9):1250-1256 (2002)

Annexe 11: Liste des protéines modélisées aux CASP5 et CAFASP3

Tableau 37: Nom, pourcentage d'identité avec le meilleur *template* et fonction des 31 protéines cibles modélisées par ESyPred3D lors du CASP5 et du CAFASP3.

Cible	%id. avec *template*	Fonction
T0133	15,1	HIP1R N-terminal domain from *Rattus norvegicus*
T0137	42,1	Fatty acid binding protein FABP1 from *Echinococcus granulosus* (PDB code: 1O8V)
T0140	41,7	1b11, synthetic protein
T0141	21,9	AmpD protein from *Citrobacter freundii* (PDB code: 1IYA)
T0142	24,4	Nitrophorin from *Cimex lectularius*
T0143	24,9	V8 protease from *Staphylococcus aureus*
T0150	30,2	Ribososmal protein L30E from *Thermococcus celer* (PDB code: 1H7M)
T0151	29,1	Single-strand binding protein (SSB) from *Mycobacterium tuberculosis* H37Rv
T0152	18,6	Hypothetical protein Rv1347c from *Mycobacterium tuberculosis* H37Rv
T0153	35,2	Deoxyuridine 5'-triphosphate nucleotidohydrolase (dUTPase) from *Mycobacterium tuberculosis* (PDB code: 1MQ7)
T0154	40,4	Pantothenate synthetase from *Mycobacterium tuberculosis* (PDB code: 1MOP)
T0155	29,9	Probable dihydroneopterin aldolase (DHNA) from *Mycobacterium tuberculosis*
T0160	21,1	VAP-A protein from *Rattus norvegicus*
T0165	15,3	Cephalosporin C deacetylase from *Bacillus subtilis* (PDB code: 1L7A)
T0167	36,0	Hypothetical Cytosolic Protein yckF from *Bacillus subtilis* (PDB code: 1M3S)
T0169	17,9	Hypothetical Protein yqjY from *Bacillus subtilis* (PDB code: 1MK4)
T0176	20,4	Hypothetical protein yggU from *Escherichia coli* o157:h7

		edl933 (PDB code: 1N91)
T0177	31,1	Hypothetical protein HP0162 from *Helicobacter pylori* (PDB code: 1MW7)
T0178	26,1	Deoxyribose-phosphate aldolase from *Aquifex aeolicus* (PDB code: 1MZH)
T0179	42,2	Spermidine synthase homolog from *Bacillus subtilis* (PDB code: 1IY9)
T0182	40,9	Methionine Aminopeptidase (TM1478) from *Thermotoga maritima* (PDB code: 1O0X)
T0183	27,9	Deoxyribose-Phosphate Aldolase (TM1559) from *Thermotoga maritima* (PDB code: 1O0Y)
T0184	17,8	Ribonuclease III (TM1102) from *Thermotoga maritima* (PDB code: 1O0W)
T0185	23,3	Udp-N-Acetylmuramate--Alanine Ligase (TM0231) from *Thermotoga maritima* (PDB code: 1J6U)
T0186	20,9	N-Acetylglucosamine-6-Phosphate Deacetylase (TM0814) from *Thermotoga maritima* (PDB code: 1O12)
T0188	28,1	Hypothetical Protein (TM1816) from *Thermotoga maritima* (PDB code: 1O13)
T0189	17,5	Sugar Kinase (TM0828) from *Thermotoga maritima* (PDB code: 1O14)
T0190	27,6	Transthyretin-related protein from *Escherichia coli*
T0191	18,6	Shikimate 5-dehydrogenase, *Methanococcus jannaschii* (PDB code: 1NVT)
T0192	20,5	Spermidime/Spermine Acetyltransferase (SSAT) from *Homo sapiens*
T0195	20,5	Hypothetical esterase in SMC3-MRPL8 intergenic region from *Saccharomyces cerevisiae*

Annexe 12: Classement obtenu au CASP5

Tableau 38: Classement obtenu, pays d'origine, institution et score des 40 premiers groupes de modélisation du CASP5 sur les 172 groupes participants. Les noms de groupe en gras sont des serveurs automatiques. Ceux en gras italique souligné sont des méta-serveurs.

Ordre	Pays	Université/Société	Nom Groupe	Score
1	Israël	Ben-Gurion University	Sasson-Iris	1,022
2	Pologne	IIMCB	GeneSilico	1,008
3	Israël	Ben-Gurion University	Fischer	1,000
4	Japon	Kitasato University	Chimera	0,985
5	Suède	Stockholm University	***PMODEL***	0,972
6	USA	University of Michigan	Lomize-Andrei	0,934
7	Pologne	IIMCB	Bujnicki-Janusz	0,931
8	Pologne	BioInfoBank Institute	Ginalski	0,931
9	USA	Celltech Inc.	Celltech	0,900
10	Pologne	BioInfoBank Institute	BIOINFO.PL	0,894
11	USA	Oak Ridge National Laboratory	ORNL-PROSPECT	0,881
12	Israël	Ben-Gurion University	**INBGU**	0,874
13	USA	University of Washington	***BAKER-ROBETTA***	0,874
14	UK	University of Cambridge	**FUGUE3**	0,859
15	Pologne	Warsaw University	Skolnick-Kolinski	0,857
16	Suède	Stockholm University	***PMODEL3***	0,853
17	Russie	Russian Academy of Sciences of Pushchino	Pushchino	0,847
18	Israël	Ben-Gurion University	***3DSN-INBGU***	0,838
19	UK	Imperial College of Science Technology and Medicine	Sternberg	0,835
20	USA	University of California, San Francisco	Friesner	0,830
21	USA	University of California, Santa Cruz	**SAM-T99**	0,819

22	Pologne	BioInfoBank Institute	**BIOINFO.PL-BASICB**	0,803
23	Pologne	IIMCB	GeneSilico.pl-servers-only	0,801
24	Suède	Stockholm University	***PCONS3***	0,801
25	Pologne	BioInfoBank Institute	**BIOINFO.PL-ORFBLAST**	0,797
26	USA	Columbia University	Honig Lab	0,797
27	Japon	Kitasato University	**FAMSD**	0,795
28	Pologne	BioInfoBank Institute	**BIOINFO.pl-PDB-BLAST**	0,791
29	Belgique	University of Namur	Lambert-Christophe	0,787
30	UK	University College London	**GENTHREADER**	0,786
31	Canada	University of Waterloo	RAPTOR	0,784
32	Japon	Kitasato University	Chimerax	0,784
33	USA	University of California, San Diego	Schulten-Wolynes	0,757
34	USA	Eidogen, Inc.	Bionomix	0,744
35	Japon	National Institute of Advanced Industrial Science and Technology	**FORTE1**	0,728
36	Israël	Ben-Gurion University	***3D-SHOTGUN-3DS5***	0,727
37	USA	Strubix Inc.	Sbi	0,724
38	UK	NIMR	Taylor	0,721
39	Japon	Nagoya University	**ALAX**	0,713
40	Belgique	University of Namur	**ESyPred3D**	0,706

X. Bibliographie

Abagyan, R. and M. Totrov (1994). "Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins." *J Mol Biol* **235**(3): 983-1002.

Abola, E. E., F. C. Bernstein, S. H. Bryant, T. F. Koetzle and J. Weng (1987). Protein data bank. Crystallographic databases - Information, content, software systems, scientific applications. F. H. Allen, G. Bergerhoff and R. Sievers. Bonn/Cambridge/Chester, Data Commission of the International Union of Crystallography**:** 107-132.

Altschul, S. F. (1989). "Gap costs for multiple sequence alignment." *J Theor Biol* **138**(3): 297-309.

Altschul, S. F. (1991). "Amino acid substitution matrices from an information theoretic perspective." *J Mol Biol* **219**(3): 555-65.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic Local Alignment Search Tool." *Journal of Molecular Biology* **215**: 403-410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleid Acids Research* **25**(17): 3389-3402.

Alwyn Jones, T. and G. J. Kleywegt (1999). "CASP3 comparative modeling evaluation." *Proteins* **Suppl 3**: 30-46.

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

ANSI, X. Standard - The C Langage. X3J11/90-013. ISO Standard ISO/IEC 9899. Washington, DC, USA, Computer and Business Equippment Manufacturers Association.

ANSI, X. (1978). American National Standard Programming Langage FORTRAN. Washington, DC, USA, Computer and Business Equippment Manufacturers Association.

Apostolico, A. and R. Giancarlo (1998). "Sequence alignment in molecular biology." *J Comput Biol* **5**(2): 173-96.

Apweiler, R., P. Kersey, V. Junker and A. Bairoch (2001). "Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp *et al*." *Bioinformatics* **17**(6): 533-534.

Argos, P. and M. G. Rossmann (1979). "Structural comparisons of heme binding proteins." *Biochemistry* **18**(22): 4951-60.

Ariza, J., F. Gudiol, R. Pallares, G. Rufi and P. Fernandez-Viladrich (1985). "Comparative trial of rifampin-doxycycline versus tetracycline- streptomycin in the therapy of human brucellosis." *Antimicrob Agents Chemother* **28**(4): 548-51.

Arnold, E. and M. G. Rossmann (1990). "Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 A." *J. Mol. Biol.* **211**(4): 763-801.

Aszodi, A., M. J. Gradwell and W. R. Taylor (1995). "Global fold determination from a small number of distance restraints." *J Mol Biol* **251**(2): 308-26.

Ayers, D. J., P. R. Gooley, A. Widmer-Cooper and A. E. Torda (1999). "Enhanced protein fold recognition using secondary structure information from NMR." *Protein Sci* **8**(5): 1127-33.

Azarya-Sprinzak, E., D. Naor, H. J. Wolfson and R. Nussinov (1997). "Interchanges of spatially neighbouring residues in structurally conserved environments." *Protein Eng* **10**(10): 1109-22.

Bairoch, A. and R. Apweiler (1999). "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999." *Nucleic Acids Res* **27**(1): 49-54.

Bajorath, J., R. Stenkamp and A. Aruffo (1993). "Knowledge-based model building of proteins: concepts and examples." *Protein Sci* **2**(11): 1798-810.

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

Baldi, P., Y. Chauvin, T. Hunkapiller and M. A. McClure (1994). "Hidden Markov models of biological primary sequence information." *Proc Natl Acad Sci U S A* **91**(3): 1059-63.

Barre, S., A. S. Greenberg, M. F. Flajnik and C. Chothia (1994). "Structural conservation of hypervariable regions in immunoglobulins evolution." *Nat Struct Biol* **1**(12): 915-20.

Barton, G. J. and M. J. Sternberg (1987). "Evaluation and improvements in the automatic alignment of protein sequences." *Protein Eng* **1**(2): 89-94.

Barton, G. J. and M. J. E. Sternberg (1987). "A Strategy for the Rapid Multiple Alignment of Protein Sequences. Confidence Levels from Tertiary Structure Comparisons." *J Mol Biol* **198**: 327-337.

Bashford, D., C. Chothia and A. M. Lesk (1987). "Determinants of a protein fold. Unique features of globin amino acid sequences." *J. Mol. Biol.* **196**(1): 199-216.

Bassolino-Klimas, D. and R. E. Bruccoleri (1992). "Application of a directed conformational search for generating 3-D coordinates for protein structures from alpha-carbon coordinates." *Proteins* **14**(4): 465-74.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer (2002). "The Pfam protein families database." *Nucleic Acids Res* **30**(1): 276-80.

Bates, P. A., L. A. Kelley, R. M. MacCallum and M. J. Sternberg (2001). "Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM." *Proteins* **45**(Suppl 5): 39-46.

Bates, P. A. and M. J. Sternberg (1999). “Model building by comparison at CASP3: using expert knowledge and computer automation.” *Proteins* **Suppl**(3): 47-54.

Baumann, G., C. Frommel and C. Sander (1989). “Polarity as a criterion in protein design.” *Protein Eng* **2**(5): 329-34.

Baxevanis, A. D. (1998). “Practical aspects of multiple sequence alignment.” *Methods Biochem Anal* **39**: 172-88.

Baxevanis, A. D. (2003). “The Molecular Biology Database Collection: 2003 update.” *Nucleic Acids Res* **31**(1): 1-12.

Bellman, R. E. (1957). Dynamic Programming. Princeton, N.J. (USA), Princeton Univ. Press.

Benner, S. A., M. A. Cohen and G. H. Gonnet (1994). “Amino acid substitution during functionally constrained divergent evolution of protein sequences.” *Protein Eng* **7**(11): 1323-32.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler (2003). “GenBank.” *Nucleic Acids Res* **31**(1): 23-7.

Berger, M. P. and P. J. Munson (1991). “A novel randomized iterative strategy for aligning multiple protein sequences.” *Comput Appl Biosci* **7**(4): 479-84.

Bergner, A., J. Gunther, M. Hendlich, G. Klebe and M. Verdonk (2001). “Use of Relibase for retrieving complex three-dimensional interaction patterns including crystallographic packing effects.” *Biopolymers* **61**(2): 99-110.

Berman, H. M. (1999). “The past and future of structure databases.” *Curr Opin Biotechnol* **10**(1): 76-80.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne (2000). “The Protein Data Bank.” *Nucleic Acids Res* **28**(1): 235-42.

Berne, B. J. and J. E. Straub (1997). “Novel methods of sampling phase space in the simulation of biological systems.” *Curr Opin Struct Biol* **7**(2): 181-9.

Biou, V., J. F. Gibrat, J. M. Levin, B. Robson and J. Garnier (1988). “Secondary structure prediction: combination of three different methods.” *Protein Eng* **2**(3): 185-91.

Bissantz, C., P. Bernard, M. Hibert and D. Rognan (2003). “Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets?” *Proteins* **50**(1): 5-25.

Blundell, T. L., J. B. Cooper, A. Sali and Z. Y. Zhu (1991). “Comparison of the sequences, 3-D structures and mechanisms of pepsin-like and retroviral aspartic proteinases.” *Adv. Exp. Med. Biol.* **306**: 443-453.

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

Boczko, E. M. and C. L. Brooks, 3rd (1995). "First-principles calculation of the folding free energy of a three- helix bundle protein." *Science* **269**(5222): 393-6.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic Acids Res* **31**(1): 365-70.

Boissel, J. P., W. R. Lee, S. R. Presnell, F. E. Cohen and H. F. Bunn (1993). "Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure." *J Biol Chem* **268**(21): 15983-93.

Bonneau, R., C. E. Strauss and D. Baker (2001). "Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation." *Proteins* **43**(1): 1-11.

Boschiroli, M. L., V. Foulongne and D. O'Callaghan (2001). "Brucellosis: a worldwide zoonosis." *Curr Opin Microbiol* **4**(1): 58-64.

Bower, M. J., F. E. Cohen and R. L. Dunbrack (1997). "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool." *J Mol Biol* **267**(5): 1268-82.

Brenner, S. E., C. Chothia and T. J. Hubbard (1998). "Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships." *Proc Natl Acad Sci U S A* **95**(11): 6073-8.

Briffeuil, P., G. Baudoux, C. Lambert, X. De Bolle, C. Vinals, E. Feytmans and E. Depiereux (1998). "Comparative analysis of seven multiple protein sequence alignment servers: clues to enhance predictions reliability." *Bioinformatics* **14**(4): 357-366.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus (1983). "CHARMM: A program for macromolecular energy minimization and thermodynamics calculations." *J Comp Chem* **4**: 187-217.

Browne, W. J., A. C. North and D. C. Phillips (1969). "A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme." *J Mol Biol* **42**(1): 65-86.

Bruccoleri, R. E. and M. Karplus (1987). "Prediction of the folding of short polypeptide segments by uniform conformational sampling." *Biopolymers* **26**(1): 137-68.

Bruccoleri, R. E. and M. Karplus (1990). "Conformational sampling using high-temperature molecular dynamics." *Biopolymers* **29**(14): 1847-62.

Bryant, S. H. and L. M. Amzel (1987). "Correctly folded proteins make twice as many hydrophobic contacts." *Int J Pept Protein Res* **29**(1): 46-52.

Bryant, S. H. and C. E. Lawrence (1991). "The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential: a statistical model for nonbonded interactions." *Proteins* **9**(2): 108-19.

Bujnicki, J. M., A. Elofsson, D. Fischer and L. Rychlewski (2001). “LiveBench-1: continuous benchmarking of protein structure prediction servers.” *Protein Sci* **10**(2): 352-61.

Bujnicki, J. M., A. Elofsson, D. Fischer and L. Rychlewski (2001). “LiveBench-2: large-scale automated evaluation of protein structure prediction servers.” *Proteins* **Suppl 5**: 184-91.

Burke, D. F., C. M. Deane, H. A. Nagarajaram, N. Campillo, M. Martin-Martinez, J. Mendes, F. Molina, J. Perry, B. V. Reddy, C. M. Soares, R. E. Steward, M. Williams, M. A. Carrondo, T. L. Blundell and K. Mizuguchi (1999). “An iterative structure-assisted approach to sequence alignment and comparative modeling.” *Proteins* **Suppl**(3): 55-60.

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

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

Callaway, D. J. (1994). “Solvent-induced organization: a physical model of folding myoglobin.” *Proteins* **20**(2): 124-38.

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

Cardon, L. R. and G. D. Stormo (1992). “Expectation maximization algorithm for identifying protein-binding sites with variable lengths from unaligned DNA fragments.” *J Mol Biol* **223**(1): 159-70.

Carillo, H. and D. Lipman (1988). “The multiple sequence alignment problem in biology.” *SIAM J Appl Math* **48**(5): 1073-1082.

Carrington, J. C., T. J. Morris, P. G. Stockley and S. C. Harrison (1987). “Structure and assembly of turnip crinkle virus. IV. Analysis of th coat protein gene and implications of the subunit primary structure.” *J. Mol. Biol.* **194**(2): 265-276.

Chiche, L., L. M. Gregoret, F. E. Cohen and P. A. Kollman (1990). “Protein model structure evaluation using the solvation free energy of folding.” *Proc Natl Acad Sci U S A* **87**(8): 3240-3.

Chivian, D., D. E. Kim, L. Malmstrom, P. Bradley, T. Robertson, P. Murphy, C. E. M. Strauss, R. Bonneau, C. A. Rohl and D. Baker (2003). “Automated prediction of CASP-5 structures using the ROBETTA server.” *Proteins* **in press**.

Chothia, C. (1992). “One thousand families for the molecular biologist.” *Nature* **357**(6379): 543-4.

Chothia, C. and A. M. Lesk (1986). “The relation between the divergence of sequence and structure in proteins.” *Embo J* **5**(4): 823-6.

Chothia, C. and A. M. Lesk (1987). “Canonical structures for the hypervariable regions of immunoglobulins.” *J Mol Biol* **196**(4): 901-17.

Chothia, C., A. M. Lesk, M. Levitt, A. G. Amit, R. A. Mariuzza, S. E. Phillips and R. J. Poljak (1986). "The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure." *Science* **233**(4765): 755-8.

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

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.

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

Cloeckaert, A., J. M. Verger, M. Grayon, J. Y. Paquet, B. Garin-Bastuji, G. Foster and J. Godfroid (2001). "Classification of Brucella spp. isolated from marine mammals by DNA polymorphism at the omp2 locus." *Microbes Infect* **3**(9): 729-38.

Cohen, F. E., J. Novotny, M. J. Sternberg, D. G. Campbell and A. F. Williams (1981). "Analysis of structural similarities between brain Thy-1 antigen and immunoglobulin domains. Evidence for an evolutionary relationship and a hypothesis for its functional significance." *Biochem J.* **195**(1): 31-40.

Collura, V., J. Higo and J. Garnier (1993). "Modeling of protein loops by simulated annealing." *Protein Sci* **2**(9): 1502-10.

Colovos, C. and T. O. Yeates (1993). "Verification of protein structures: patterns of nonbonded atomic interactions." *Protein Sci* **2**(9): 1511-9.

Contreras-Moreira, B., P. W. Fitzjohn and P. A. Bates (2003). "In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling." *J Mol Biol* **328**(3): 593-608.

Corbel, M. J. (1997). "Brucellosis: an overview." *Emerg Infect Dis* **3**(2): 213-21.

Corbel, M. J. (1997). "Recent advances in brucellosis." *J Med Microbiol* **46**(2): 101-3.

Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. J. Merz, D. M. Fergusson, D. C. Spellmeyer, D. C. Fox, J. W. Caldwell and P. A. Kollman (1995). "A second generation force field for the simulation of proteins and nucleic acids." *J Am Chem Soc* **117**: 5179-5197.

Corpet, F. (1988). "Multiple sequence alignment with hierarchical clustering." *Nucleic Acids Res* **16**(22): 10881-90.

Covell, D. G. (1992). "Folding protein alpha-carbon chains into compact forms by Monte Carlo methods." *Proteins* **14**(3): 409-20.

Cserzo, M., J. M. Bernassau, I. Simon and B. Maigret (1994). "New alignment strategy for transmembrane proteins." *J Mol Biol* **243**(3): 388-96.

Cuff, J. A. and G. J. Barton (2000). "Application of multiple sequence alignment profiles to improve protein secondary structure prediction." *Proteins* **40**(3): 502-511.

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

Cui, Y., R. S. Chen and W. H. Wong (1998). "Protein folding simulation with genetic algorithm and supersecondary structure constraints." *Proteins* **31**(3): 247-57.

Dandekar, T. and P. Argos (1994). "Folding the main chain of small proteins with the genetic algorithm." *J Mol Biol* **236**(3): 844-61.

Dauber-Osguthorpe, P., V. A. Roberts, D. J. Osguthorpe, J. Wolff, M. Genest and A. T. Hagler (1988). "Structure and energetics of ligand binding to proteins: Escherichia coli dihydrofolate reductase-trimethoprim, a drug-receptor system." *Proteins* **4**(1): 31-47.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. M. O. Dayoff. Washington, D.C. (USA), National Biomedical Research Foundation. **5 Suppl. 3:** 352.

De Bolle, X., C. Vinals, D. Prozzi, J. Y. Paquet, R. Leplae, E. Depiereux, J. Vandenhaute and E. Feytmans (1995). "Identification of residues potentially involved in the interactions between subunits in yeast alcohol dehydrogenases." *Eur J Biochem* **231**(1): 214-9.

de Fays, K., A. Tibor, C. Lambert, C. Vinals, P. Denoel, X. De Bolle, J. Wouters, J. J. Letesson and E. Depiereux (1999). "Structure and function prediction of the Brucella abortus P39 protein by comparative modeling with marginal sequence similarities." *Protein Eng* **12**(3): 217-23.

de la Cruz, X. and J. M. Thornton (1999). "Factors limiting the performance of prediction-based fold recognition methods." *Protein Sci* **8**(4): 750-9.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg (1999). "Improved microbial gene identification with GLIMMER." *Nucleic Acids Res* **27**(23): 4636-41.

DelVecchio, V. G., V. Kapatral, R. J. Redkar, G. Patra, C. Mujer, T. Los, N. Ivanova, I. Anderson, A. Bhattacharyya, A. Lykidis, G. Reznik, L. Jablonski, N. Larsen, M. D'Souza, A. Bernal, M. Mazur, E. Goltsman, E. Selkov, P. H. Elzer, S. Hagius, D. O'Callaghan, J. J. Letesson, R. Haselkorn, N. Kyrpides and R. Overbeek (2002). "The genome sequence of the facultative intracellular pathogen Brucella melitensis." *Proc Natl Acad Sci U S A* **99**(1): 443-8.

Depiereux, E., G. Baudoux, P. Briffeuil, I. Reginster, X. De Bolle, C. Vinals and E. Feytmans (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 (1991). "Simultaneous and multivariate alignment of protein sequences: correspondence between physiochemical profiles and structurally conserved regions (SCR)." *Protein Engineering* **4**(6): 603-613.

Depiereux, E. and E. Feytmans (1992). "Match-Box: a fundamently new algorithm for simultaneous alignment of several protein sequences." *CABIOS* **8**(5): 501-509.

Depiereux, E. and E. Feytmans (1994). "Elaboration de matrices de scores pour l'alignement de séquences de protéines sur base de similarités évaluées dans des structures semblables." *Biom. Praxim.* **34**: 13-34.

Devos, D. and A. Valencia (2000). "Practical limits of function prediction." *Proteins* **41**(1): 98-107.

Di Francesco, V., P. J. Munson and J. Garnier (1999). "FORESST: fold recognition from secondary structure predictions of proteins." *Bioinformatics* **15**(2): 131-40.

Dickerson, R. E., R. Timkovich and R. J. Almassy (1976). "The cytochrome fold and the evolution of bacterial energy metabolism." *J Mol Biol* **100**(4): 473-91.

Ding, D. F., A. Sali and T. L. Blundell (1994). "A differential geometric treatment of protein structure comparison." *Bull. Math. Biol.* **56**(5): 923-943.

Doolitle, R. F. (1986). Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. Mill Valey, CA, USA, University Science Books.

Doolittle, R. F. (1981). "Similar amino acid sequences: chance or common ancestry?" *Science* **214**(4517): 149-59.

Duan, Y. and P. A. Kollman (1998). "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution." *Science* **282**(5389): 740-4.

Dudek, M. J. and H. A. Scheraga (1990). "Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops." *J Comp Chem* **11**: 121-151.

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

Dunbrack, R. L., Jr. (2002). "Rotamer libraries in the 21st century." *Curr Opin Struct Biol* **12**(4): 431-40.

Eddy, S. R. (1995). "Multiple alignment using hidden Markov models." *Proc Int Conf Intell Syst Mol Biol* **3**: 114-20.

Eddy, S. R. (1998). "Profile hidden Markov models." *Bioinformatics* **14**(9): 755-63.

Efimov, A. V. (1991). "Structure of alpha-alpha-hairpins with short connections." *Protein Eng* **4**(3): 245-50.

Efimov, A. V. (1991). "Structure of coiled beta-beta-hairpins and beta-beta-corners." *FEBS Lett* **284**(2): 288-92.

Eisenhaber, F., B. Persson and P. Argos (1995). "Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence." *Crit Rev Biochem Mol Biol* **30**(1): 1-94.

Eyrich, V. A., M. A. Marti-Renom, D. Przybylski, M. S. Madhusudhan, A. Fiser, F. Pazos, A. Valencia, A. Sali and B. Rost (2001). "EVA: continuous automatic evaluation of protein structure prediction servers." *Bioinformatics* **17**(12): 1242-1243.

Felsenstein, J. (1989). "PHYLIP -- Phylogeny Inference Package (Version 3.2)." *Cladistics* **5**: 164-166.

Feng, D. F. and R. F. Doolittle (1987). "Progressive sequence alignment as a prerequisite to correct phylogenetic trees." *J Mol Evol* **25**(4): 351-60.

Feng, D. F., M. S. Johnson and R. F. Doolittle (1984). "Aligning amino acid sequences: comparison of commonly used methods." *J Mol Evol* **21**(2): 112-25.

Fine, R. M., H. Wang, P. S. Shenkin, D. L. Yarmush and C. Levinthal (1986). "Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations." *Proteins* **1**(4): 342-62.

Fischer, D. and D. Eisenberg (1996). "Fold Recognition Using Sequence-Derived Predictions." *Protein Science* **5**: 947-955.

Fischer, D. and L. Rychlewski (2003). "The 2002 olympic games of protein structure prediction." *Protein Eng* **16**(3): 157-60.

Fiser, A., A. Sánchez, F. Melo and A. Sali (2000). Comparative protein structure modeling. Computational Biochemistry and Biophysics. M. Watanabe, B. Roux, A. MacKerell and O. Becker, Marcel Dekker.

Fitch, W. M. (1966). "An improved method of testing for evolutionary homology." *J Mol Biol* **16**(1): 9-16.

Flower, D. R., A. C. North and T. K. Attwood (1993). "Structure and sequence relationships in the lipocalins and related proteins." *Protein Sci.* **2**(5): 753-761.

Fredman, M. L. (1984). *Bull Math Biol* **46**: 553.

Friesner, R. A. and J. R. Gunn (1996). "Computational studies of protein folding." *Annu Rev Biophys Biomol Struct* **25**: 315-42.

Galibert, F., T. M. Finan, S. R. Long, A. Puhler, P. Abola, F. Ampe, F. Barloy-Hubler, M. J. Barnett, A. Becker, P. Boistard, G. Bothe, M. Boutry, L. Bowser, J. Buhrmester, E. Cadieu, D. Capela, P. Chain, A. Cowie, R. W. Davis, S. Dreano, N. A. Federspiel, R. F. Fisher, S. Gloux, T. Godrie, A. Goffeau, B. Golding, J. Gouzy, M. Gurjal, I. Hernandez-Lucas, A. Hong, L. Huizar, R. W. Hyman, T. Jones, D. Kahn, M. L. Kahn, S. Kalman, D. H. Keating, E. Kiss, C. Komp, V. Lelaure, D. Masuy, C. Palm, M. C. Peck, T. M. Pohl, D. Portetelle, B. Purnelle, U. Ramsperger, R. Surzycki, P. Thebault, M. Vandenbol, F. J. Vorholter, S. Weidner, D. H. Wells, K. Wong, K. C. Yeh and J. Batut (2001). "The composite genome of the legume symbiont Sinorhizobium meliloti." *Science* **293**(5530): 668-72.

Garnier, J., D. J. Osguthorpe and B. Robson (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.

George, D. G., W. C. Barker and L. T. Hunt (1990). "Mutation data matrix and its uses." *Methods Enzymol* **183**: 333-51.

Geourjon, C. and G. Deleage (1995). "SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments." *Comput. Appl. Biosci.* **11**(6): 681-684.

Gibrat, J. F., J. Garnier and B. Robson (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.

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

Ginalski, K. and L. Rychlewski (2003). "Detection of reliable and unexpected protein fold predictions using 3D-Jury." *Nucleic Acids Res* **31**(13): 3291-2.

Glen, R. C. and S. C. Allen (2003). "Ligand-Protein Docking: Cancer Research at the Interface between Biology and Chemistry." *Curr Med Chem* **10**(9): 763-7.

Goad, W. B. and M. I. Kanehisa (1982). "Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries." *Nucleic Acids Res* **10**(1): 247-63.

Gonnet, G. H., M. A. Cohen and S. A. Benner (1992). "Exhaustive matching of the entire protein sequence database." *Science* **256**(5062): 1443-5.

Goodman, L. A. (1974). *Biometrika* **61**: 215.

Goodner, B., G. Hinkle, S. Gattung, N. Miller, M. Blanchard, B. Qurollo, B. S. Goldman, Y. Cao, M. Askenazi, C. Halling, L. Mullin, K. Houmiel, J. Gordon, M. Vaudin, O. Iartchouk, A. Epp, F. Liu, C. Wollam, M. Allinger, D. Doughty, C. Scott, C. Lappas, B. Markelz, C. Flanagan, C. Crowell, J. Gurson, C. Lomo, C. Sear, G. Strub, C. Cielo and S. Slater (2001). "Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58." *Science* **294**(5550): 2323-8.

Gotoh, O. (1986). "Alignment of three biological sequences with an efficient traceback procedure." *J Theor Biol* **121**(3): 327-37.

Gotoh, O. (1987). "Pattern matching of biological sequences with limited storage." *Comput Appl Biosci* **3**(1): 17-20.

Gotoh, O. (1993). "Optimal alignment between groups of sequences and its application to multiple sequence alignment." *Comput Appl Biosci* **9**(3): 361-70.

Gotoh, O. (1996). "Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments." *J. Mol. Biol.* **264**: 823-838.

Gotoh, O. (1999). "Multiple sequence alignment: algorithms and applications." *Adv Biophys* **36**: 159-206.

Grantham, R. (1974). "Amino acid difference formula to help explain protein evolution." *Science* **185**(4154): 862-4.

Greer, J. (1990). "Comparative modeling methods: application to the family of the mammalian serine proteases." *Proteins* **7**(4): 317-334.

Gribskov, M. (1994). "Profile analysis." *Methods Mol Biol* **25**: 247-66.

Grundy, W. N., T. L. Bailey and C. P. Elkan (1996). "ParaMEME: a parallel implementation and a Web interface for a DNA and protein motif discovery tool." *CABIOS* **12**(4): 303-310.

Guenther, B., R. Onrust, A. Sali, M. O'Donnell and J. Kuriyan (1997). "Crystal structure of the delta' subunit of the clamp-loader complex of E. coli DNA polymerase III." *Cell* **91**(3): 335-45.

Gupta, S. K., J. D. Kececioglu and A. A. Schaffer (1995). "Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment." *J Comput Biol* **2**(3): 459-72.

Hanks, S. K., A. M. Quinn and T. Hunter (1988). "The protein kinase family: conserved features and deduced phylogeny of the catalytic domains." *Science* **241**: 42-52.

Hart, P. E., N. J. Nilson and B. Raphael (1968). *IEEE Trans Syst Sci Cybern* **4**: 100.

Havel, T. F. and M. E. Snow (1991). "A new method for building protein conformations from sequence alignments with homologues of known structure." *J Mol Biol* **217**(1): 1-7.

Hebb, D. O. (1949). The Organization of Behaviour. New York, John Wiley.

Hein, J. (1989). "A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given." *Mol Biol Evol* **6**(6): 649-68.

Hendlich, M., A. Bergner, J. Gunther and G. Klebe (2003). "Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions." *J Mol Biol* **326**(2): 607-20.

Henikoff, J. G. and S. Henikoff (1996). "BLOCKs database and its applications." *Methods in Enzymology* **266**: 88-105.

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

Henikoff, S. and J. G. Henikoff (1993). "Performance evaluation of amino acid substitution matrices." *Proteins* **17**(1): 49-61.

Henikoff, S. and J. G. Henikoff (1994). "Protein family classification based on searching a database of blocks." *Genomics* **19**: 97-107.

Henikoff, S. and J. G. Henikoff (1997). "Embedding strategies for effective use of information from multiple sequence alignments." *Protein Sci* **6**(3): 698-705.

Henikoff, S., J. G. Henikoff, W. J. Alford and S. Pietrokovski (1995). "Automated construction and graphical presentation of protein blocks from unaligned sequences." *Gene-COMBIS* **163**: 17-26.

Henikoff, S., J. G. Henikoff and S. Pietrokovski (1999). "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations." *Bioinformatics* **15**(6): 471-9.

Higgins, D. G. and P. M. Sharp (1989). "Fast and sensitive multiple sequence alignments on a microcomputer." *Comput Appl Biosci* **5**(2): 151-3.

Higo, J., V. Collura and J. Garnier (1992). "Development of an extended simulated annealing method: application to the modeling of complementary determining regions of immunoglobulins." *Biopolymers* **32**(1): 33-43.

Hinds, D. A. and M. Levitt (1994). "Exploring conformational space with a simple lattice model for protein structure." *J Mol Biol* **243**(4): 668-82.

Hirosawa, M., Y. Totoki, M. Hoshida and M. Ishikawa (1995). "Comprehensive study on iterative algorithms of multiple sequence alignment." *Comput Appl Biosci* **11**(1): 13-8.

Hirst, J. D., M. Vieth, J. Skolnick and C. L. Brooks, 3rd (1996). "Predicting leucine zipper structures from sequence." *Protein Eng* **9**(8): 657-62.

Hogeweg, P. and B. Hesper (1984). "The alignment of sets of sequences and the construction of phyletic trees: an integrated method." *J Mol Evol* **20**(2): 175-86.

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

Holm, L. and C. Sander (1992). "Evaluation of protein models by atomic solvation preference." *J Mol Biol* **225**(1): 93-105.

Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." *J Mol Biol* **233**(1): 123-38.

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

Holm, L. and C. Sander (1998). "Touring protein fold space with Dali/FSSP." *Nucleic Acids Res.* **26**(1): 316-319.

Holm, L. and C. Sander (1999). "Protein folds and families: sequence and structure alignments." *Nucleic Acids Res* **27**(1): 244-7.

Howell, P. L., S. C. Almo, M. R. Parsons, J. Hajdu and G. A. Petsko (1992). "Structure determination of turkey egg-white lysozyme using Laue diffraction data." *Acta Crystallogr B* **48**(Pt 2): 200-7.

Huang, E. S., P. Koehl, M. Levitt, R. V. Pappu and J. W. Ponder (1998). "Accuracy of side-chain prediction upon near-native protein backbones generated by Ab initio folding methods." *Proteins* **33**(2): 204-17.

Huang, E. S., S. Subbiah and M. Levitt (1995). "Recognizing native folds by the arrangement of hydrophobic and polar residues." *J Mol Biol* **252**(5): 709-20.

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

Hubbard, T. J. (1999). "RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions." *Proteins* **Suppl**(3): 15-21.

Hubbard, T. J., B. Ailey, S. E. Brenner, A. G. Murzin and C. Chothia (1999). "SCOP: a Structural Classification of Proteins database." *Nucleic Acids Res* **27**(1): 254-6.

Hughey, R. and A. Krogh (1996). "Hidden Markov models for sequence analysis: extension and analysis of the basic method." *Comput Appl Biosci* **12**(2): 95-107.

Huynen, M., T. Doerks, F. Eisenhaber, C. Orengo, S. Sunyaev, Y. Yuan and P. Bork (1998). "Homology-based fold predictions for Mycoplasma genitalium proteins." *J Mol Biol* **280**(3): 323-6.

Ishikawa, M., T. Toya, M. Hoshida, K. Nitta, A. Ogiwara and M. Kanehisa (1993). "Multiple sequence alignment by parallel simulated annealing." *Comput Appl Biosci* **9**(3): 267-73.

Jambon, M., A. Imberty, G. Deleage and C. Geourjon (2003). "A new bioinformatic approach to detect common 3D sites in protein structures." *Proteins* **52**(2): 137-45.

Jaroszewski, L., L. Rychlewski, B. Zhang and A. Godzik (1998). "Fold prediction by a hierarchy of sequence, threading, and modeling methods." *Protein Sci* **7**(6): 1431-40.

Jennings, A. J., C. M. Edge and M. J. Sternberg (2001). "An approach to improving multiple alignments of protein sequences using predicted secondary structure." *Protein Eng* **14**(4): 227-31.

John, B. and A. Sali (2003). "Comparative protein structure modeling by iterative alignment, model building and model assessment." *Nucleic Acids Res* **31**(14): 3982-92.

Johnson, M. S. and R. F. Doolittle (1986). "A method for the simultaneous alignment of three or more amino acid sequences." *J Mol Evol* **23**(3): 267-78.

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.

Johnson, M. S., N. Srinivasan, R. Sowdhamini and T. L. Blundell (1994). "Knowledge-based protein modeling." *Crit Rev Biochem Mol Biol* **29**(1): 1-68.

Jones, D. T. (1997). "Progress in protein structure prediction." *Curr Opin Struct Biol* **7**(3): 377-87.

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

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

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

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences." *Comput Appl Biosci* **8**(3): 275-82.

Jones, D. T., W. R. Taylor and J. M. Thornton (1994). "A mutation data matrix for transmembrane proteins." *FEBS Lett* **339**(3): 269-75.

Jones, D. T., M. Tress, K. Bryson and C. Hadley (1999). "Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure." *Proteins* **Suppl**(3): 104-11.

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

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* **22**(12): 2577-637.

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

Kanehisa, M. (2002). "The KEGG database." *Novartis Found Symp* **247**: 91-101.

Kaneko, T., Y. Nakamura, S. Sato, E. Asamizu, T. Kato, S. Sasamoto, A. Watanabe, K. Idesawa, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, Y. Mochizuki, S. Nakayama, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada and S. Tabata (2000). "Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti." *DNA Res* **7**(6): 331-8.

Kaneko, T., Y. Nakamura, S. Sato, E. Asamizu, T. Kato, S. Sasamoto, A. Watanabe, K. Idesawa, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, C. Kiyokawa, M. Kohara, M. Matsumoto, A. Matsuno, Y. Mochizuki, S. Nakayama, N. Nakazaki, S. Shimpo, M. Sugimoto, C. Takeuchi, M. Yamada and S. Tabata (2000). "Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti (supplement)." *DNA Res* **7**(6): 381-406.

Karp, P. D., S. Paley and J. Zhu (2001). "Database verification studies of SWISS-PROT and GenBank." *Bioinformatics* **17**(6): 526-32; discussion 533-4.

Karplus, K., C. Barrett and R. Hughey (1998). "Hidden Markov models for detecting remote protein homologies." *Bioinformatics* **14**(10): 846-56.

Karplus, K., K. Sjolander, C. Barrett, M. Cline, D. Haussler, R. Hughey, L. Holm and C. Sander (1997). "Predicting protein structure using hidden Markov models." *Proteins* **Suppl**(1): 134-9.

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

Kim, J., S. Pramanik and M. J. Chung (1994). "Multiple sequence alignment using simulated annealing." *Comput Appl Biosci* **10**(4): 419-26.

Kirkpatrick, S., G. C.D. and M. P. Vecchi (1983). "Optimization by simulated annealing." *SCIENCE* **220**(4598): 671-680.

Kloczkowski, A., K. L. Ting, R. L. Jernigan and J. Garnier (2002). "Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence." *Proteins* **49**(2): 154-66.

Koehl, P. and M. Delarue (1994). "Polar and nonpolar atomic environments in the protein core: implications for folding and binding." *Proteins* **20**(3): 264-78.

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

Kolaskar, A. S. and U. Kulkarni-Kale (1992). "Sequence alignment approach to pick up conformationally similar protein fragments." *J Mol Biol* **223**(4): 1053-61.

Koshi, J. M. and R. A. Goldstein (1995). "Context-dependent optimal substitution matrices." *Protein Eng* **8**(7): 641-5.

Kostrowicki, J. and H. A. Scheraga (1992). "Application of the diffusion equation method for global optimization to oligopeptides." *J Phys Chem* **18**: 7442-7449.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander and D. Haussler (1994). "Hidden Markov models in computational biology. Applications to protein modeling." *J Mol Biol* **235**(5): 1501-31.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." *J Mol Biol* **305**(3): 567-80.

Lackner, P., W. A. Koppensteiner, F. S. Domingues and M. J. Sippl (1999). "Automated large scale evaluation of protein structure predictions." *Proteins* **Suppl**(3): 7-14.

Lambert, C., N. Leonard, X. De Bolle and E. Depiereux (2002). "ESyPred3D: Prediction of proteins 3D structures." *Bioinformatics* **18**(9): 1250-6.

Lambert, C., J. M. Van Campenhout, X. De Bolle and E. Depiereux (2003). "Review of common sequence alignment methods: clues to enhance reliability." *Curr. Genomics* **4**: 131-146.

Lambert, C., J. Wouters, X. De Bolle and E. Depiereux (2003). "Biologie *in silico*: Point de Vue de Bioinformaticiens." *Chimie nouvelle* **in press**.

Larson, S. M., A. Garg, J. R. Desjarlais and V. S. Pande (2003). "Increased detection of structural templates using alignments of designed sequences." *Proteins* **51**(3): 390-6.

Laskowski, R. A., M. W. MacArthur and J. M. Thornton (1998). "Validation of protein models derived from experiment." *Curr Opin Struct Biol* **8**(5): 631-9.

Laskowski, R. A., M. W. McArthur, D. S. Moss and J. M. Thornton (1993). "PROCHECK: A program to check the stereochemical quality of protein structures." *J Appl Cryst* **26**: 283-291.

Laskowski, R. A., D. S. Moss and J. M. Thornton (1993). "Main-chain bond lengths and bond angles in protein structures." *Journal of Molecular Biology* **231**(4): 1049-1067.

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

Lawrence, C. E. and A. A. Reilly (1990). "An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences." *Proteins* **7**(1): 41-51.

Lawrence, E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald and J. C. Wooton (1993). "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment." *Science* **262**: 208-214.

Lee, C. (1995). "Testing homology modeling on mutant proteins: predicting structural and thermodynamic effects in the Ala98-->Val mutants of T4 lysozyme." *Fold Des* **1**(1): 1-12.

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.

Levin, J. M. and J. Garnier (1988). "Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool." *Biochim Biophys Acta* **955**(3): 283-95.

Levin, J. M., B. Robson and J. Garnier (1986). "An algorithm for secondary structure determination in proteins based on sequence similarity." *FEBS Lett.* **205**(2): 303-308.

Levitt, M. (1976). "A simplified representation of protein conformations for rapid simulation of protein folding." *J Mol Biol* **104**(1): 59-107.

Levitt, M. (1983). "Protein folding by restrained energy minimization and molecular dynamics." *J Mol Biol* **170**(3): 723-64.

Levitt, M. (1992). "Accurate modeling of protein conformation by automatic segment matching." *J Mol Biol* **226**(2): 507-33.

Levitt, M. (1997). "Competitive assessment of protein fold recognition and alignment accuracy." *Proteins* **Suppl**(1): 92-104.

Levitt, M., M. Gerstein, E. Huang, S. Subbiah and J. Tsai (1999). "Protein folding: the endgame." *Ann Rev Biochem* **66**: 1368-1372.

Lim, V. I. (1974). "Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure." *J Mol Biol* **88**(4): 857-72.

Lipman, D. J., S. F. Altschul and J. D. Kececioglu (1989). "A tool for multiple sequence alignment." *Proc Natl Acad Sci U S A* **86**(12): 4412-5.

Lodish, Baltimore, Berk, Zipursky, Matsudaira and Darnell (1997). Structure et fonction des protéines. Biologie moléculaire de la cellule, Deboek université.

Loytynoja, A. and M. C. Milinkovitch (2001). "SOAP, cleaning multiple alignments from unstable blocks." *Bioinformatics* **17**(6): 573-4.

Loytynoja, A. and M. C. Milinkovitch (2003). "A hidden Markov model for progressive multiple alignment." *Bioinformatics* **19**(12): 1505-13.

Lu, L., A. K. Arakaki, H. Lu and J. Skolnick (2003). "Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome." *Genome Res* **13**(6A): 1146-54.

Lukashin, A. V., J. Engelbrecht and S. Brunak (1992). "Multiple alignment using simulated annealing: branch point definition in human mRNA splicing." *Nucleic Acids Res* **20**(10): 2511-6.

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

Luthy, R., A. D. McLachlan and D. Eisenberg (1991). "Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities." *Proteins* **10**(3): 229-39.

Martin, A. C., J. C. Cheetham and A. R. Rees (1989). "Modeling antibody hypervariable loops: a combined algorithm." *Proc Natl Acad Sci U S A* **86**(23): 9268-72.

Martin, A. C., M. W. MacArthur and J. M. Thornton (1997). "Assessment of comparative modeling in CASP2." *Proteins* **Suppl**(1): 14-28.

Marti-Renom, M. A., V. A. Ilyin and A. Sali (2001). "DBAli: a database of protein structure alignments." *Bioinformatics* **17**(8): 746-7.

McClure, M. A., T. K. Vasi and W. M. Fitch (1994). "Comparative analysis of multiple protein-sequence alignment methods [published erratum appears in Mol Biol Evol 1994 Sep;11(5):811]." *Mol Biol Evol* **11**(4): 571-92.

McGuffin, L. J., K. Bryson and D. T. Jones (2000). "The PSIPRED protein structure prediction server." *Bioinformatics* **16**(4): 404-405.

McLachlan, A. D. (1971). "Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551." *J Mol Biol* **61**(2): 409-24.

McLachlan, A. D. (1972). "Repeating sequences and gene duplication in proteins." *J Mol Biol* **64**(2): 417-37.

McLaughlin, W. A. and H. M. Berman (2003). "Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif." *J Mol Biol* **330**(1): 43-55.

Melo, F. and E. Feytmans (1997). "Novel Knowledge-based Mean Force Potential at Atomic Level." *Journal of Molecular Biology* **267**: 207-222.

Melo, F. and E. Feytmans (1998). "Assessing protein structures with a non-local atomic interaction energy." *J Mol Biol* **277**(5): 1141-52.

Miyata, T., S. Miyazawa and T. Yasunaga (1979). "Two types of amino acid substitutions in protein evolution." *J Mol Evol* **12**(3): 219-36.

Miyazawa, S. and R. L. Jernigan (1993). "A new substitution matrix for protein sequence searches based on contact frequencies in protein structures." *Protein Eng* **6**(3): 267-78.

Mizuguchi, K., C. M. Deane, T. L. Blundell and J. P. Overington (1998). "HOMSTRAD: a database of protein structure alignments for homologous families." *Protein Sci* **7**(11): 2469-71.

Modi, S., M. J. Paine, M. J. Sutcliffe, L. Y. Lian, W. U. Primrose, C. R. Wolf and G. C. Roberts (1996). "A model for human cytochrome P450 2D6 based on homology modeling and NMR studies of substrate binding." *Biochemistry* **35**(14): 4540-50.

Mohana Rao, J. K. (1987). "New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters." *Int J Pept Protein Res* **29**(2): 276-81.

Monge, A., R. A. Friesner and B. Honig (1994). "An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure." *Proc Natl Acad Sci U S A* **91**(11): 5027-9.

Morgenstern, B. (1999). "DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment." *Bioinformatics* **15**(3): 211-8.

Morgenstern, B., K. Frech, A. Dress and T. Werner (1998). "DIALIGN: Finding local similarities by multiple sequence alignment." *Bioinformatics* **14**(3): 290-294.

Mosimann, S., R. Meleshko and M. N. James (1995). "A critical assessment of comparative molecular modeling of tertiary structures of proteins." *Proteins* **23**(3): 301-17.

Moult, J. and M. N. James (1986). "An algorithm for determining the conformation of polypeptide segments in proteins by systematic search." *Proteins* **1**(2): 146-63.

Mumenthaler, C. and W. Braun (1995). "Predicting the helix packing of globular proteins by self-correcting distance geometry." *Protein Sci* **4**(5): 863-71.

Murata, M., J. S. Richardson and J. L. Sussman (1985). "Simultaneous comparison of three protein sequences." *Proc Natl Acad Sci U S A* **82**(10): 3073-7.

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chotia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." *Journal of Molecular Biology* **247**(4): 536-540.

Myers, E. and W. Miller (1989). "Optimal Alignments in Linear Space." *CABIOS* **4**: 11-17.

Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." *Trends Biochem Sci* **24**(1): 34-6.

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.

Nemethy, G., K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey and H. A. Scheraga (1992). "Energy parameters in peptides improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides." *J Phys Chem* **96**(6472-6484): 6472-6484.

Neuwald, A. F., J. S. Liu, D. J. Lipman and C. E. Lawrence (1997). "Extracting protein alignment models from the sequence database." *Nucleic Acids Research* **25**(9): 1665-1677.

Niefind, K. and D. Schomburg (1991). "Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles." *J Mol Biol* **219**(3): 481-97.

Nierman, W. C., T. V. Feldblyum, M. T. Laub, I. T. Paulsen, K. E. Nelson, J. A. Eisen, J. F. Heidelberg, M. R. Alley, N. Ohta, J. R. Maddock, I. Potocka, W. C. Nelson, A. Newton, C. Stephens, N. D. Phadke, B. Ely, R. T. DeBoy, R. J. Dodson, A. S. Durkin, M. L. Gwinn, D. H. Haft, J. F. Kolonay, J. Smit, M. B. Craven, H. Khouri, J. Shetty, K. Berry, T. Utterback, K. Tran, A. Wolf, J. Vamathevan, M. Ermolaeva, O. White, S. L. Salzberg, J. C. Venter, L. Shapiro, C. M. Fraser and J. Eisen (2001). "Complete genome sequence of Caulobacter crescentus." *Proc Natl Acad Sci U S A* **98**(7): 4136-41.

Notredame, C. and D. G. Higgins (1996). "SAGA: sequence alignment by genetic algorithm." *Nucleic Acids Res* **24**(8): 1515-24.

Notredame, C., D. G. Higgins and J. Heringa (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." *J Mol Biol* **302**(1): 205-17.

Notredame, C., L. Holm and D. G. Higgins (1998). "COFFEE: an objective function for multiple sequence alignments." *Bioinformatics* **14**(5): 407-22.

Novotny, J., R. Bruccoleri and M. Karplus (1984). "An analysis of incorrectly folded protein models. Implications for structure predictions." *J Mol Biol* **177**(4): 787-818.

Novotny, J., A. A. Rashin and R. E. Bruccoleri (1988). "Criteria that discriminate between native proteins and incorrectly folded models." *Proteins* **4**(1): 19-30.

Ohlendorf (1994). "Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin 1Beta." *Acta Cryst* **D50**: 808-812.

Oldfield, T. J. (1992). "SQUID: a program for the analysis and display of data from crystallography and molecular dynamics." *J Mol Graph* **10**(4): 247-52.

Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton (1997). "CATH--a hierarchic classification of protein domain structures." *Structure* **5**(8): 1093-108.

Orengo, C. A., F. M. Pearl, J. E. Bray, A. E. Todd, A. C. Martin, L. Lo Conte and J. M. Thornton (1999). "The CATH Database provides insights into protein structure/function relationships." *Nucleic Acids Res* **27**(1): 275-9.

Ortiz, A. R., A. Kolinski and J. Skolnick (1998). "Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments." *J Mol Biol* **277**(2): 419-48.

Ortiz, A. R., C. E. Strauss and O. Olmea (2002). "MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison." *Protein Sci* **11**(11): 2606-21.

Overington, J., D. Donnelly, M. S. Johnon, A. Sali and T. L. Blundell (1992). "Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds." *Protein Sci.* **1**(2): 216-226.

Park, B. H., E. S. Huang and M. Levitt (1997). "Factors affecting the ability of energy functions to discriminate correct from incorrect folds." *J Mol Biol* **266**(4): 831-46.

Park, J., K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard and C. Chothia (1998). "Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods." *J Mol Biol* **284**(4): 1201-10.

Park, J., S. A. Teichmann, T. Hubbard and C. Chothia (1997). "Intermediate sequences increase the detection of homology between sequences." *J Mol Biol* **273**(1): 349-54.

Pascarella, S. and P. Argos (1992). "A data bank merging related protein structures and sequences." *Protein Eng* **5**(2): 121-37.

Pascarella, S., F. Milpetz and P. Argos (1996). "A databank (3D-ali) collecting related protein sequences and structures." *Protein Eng* **9**(3): 249-51.

Paulsen, I. T., R. Seshadri, K. E. Nelson, J. A. Eisen, J. F. Heidelberg, T. D. Read, R. J. Dodson, L. Umayam, L. M. Brinkac, M. J. Beanan, S. C. Daugherty, R. T. Deboy, A. S. Durkin, J. F. Kolonay, R. Madupu, W. C. Nelson, B. Ayodeji, M. Kraul, J. Shetty, J. Malek, S. E. Van Aken, S. Riedmuller, H. Tettelin, S. R. Gill, O. White, S. L. Salzberg, D. L. Hoover, L. E. Lindler, S. M. Halling, S. M. Boyle and C. M. Fraser (2002). "The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts." *Proc Natl Acad Sci U S A* **99**(20): 13148-53.

Pazos, F., B. Rost and A. Valencia (1999). "A platform for integrating threading results with protein family analyses." *Bioinformatics* **15**(12): 1062-3.

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

Pearson, W. R. (1995). "Comparison of methods for searching protein sequence databases." *Protein Sci.* **4**(6): 1145-1160.

Pearson, W. R. and D. J. Lipman (1988). "Improved Tools for Biological Sequence Analysis." *Proc. Natl. Acad. Sci. USA* **85**: 2444-2448.

Peitsch, M. C. (1995). "ProMod: automated knowledge-based protein modelling tool." *PDB Quarterly Newsletter* **72**: 4.

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

Pospisil, P., T. Kuoni, L. Scapozza and G. Folkers (2002). "Methodology and problems of protein-ligand docking: case study of dihydroorotate dehydrogenase, thymidine kinase, and phosphodiesterase 4." *J Recept Signal Transduct Res* **22**(1-4): 141-54.

Prasad, J. C., S. R. Comeau, S. Vajda and C. J. Camacho (2003). "Consensus alignment for reliable framework prediction in homology modeling." *Bioinformatics* **19**(13): 1682-91.

Pruitt, K. D., K. S. Katz, H. Sicotte and D. R. Maglott (2000). "Introducing RefSeq and LocusLink: curated human genome resources at the NCBI." *Trends Genet* **16**(1): 44-7.

Pruitt, K. D. and D. R. Maglott (2001). "RefSeq and LocusLink: NCBI gene-centered resources." *Nucleic Acids Res* **29**(1): 137-40.

Ptitsyn, O. B. and A. V. Finkelstein (1983). "Theory of protein secondary structure and algorithm of its prediction." *Biopolymers* **22**(1): 15-25.

Qu, C. X., L. H. Lai, X. J. Xu and Y. Q. Tang (1993). "Phyletic relationships of protein structures based on spatial preference of residues." *J Mol Evol* **36**(1): 67-78.

Rabiner, L. R. (1989). "A tutorial on hidden Markov-models and selected applications in speech recognition." *Proc IEEE* **77**(2): 257-286.

Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan (1963). "Stereochemistry of polypeptide chain configurations." *J. Mol. Biol.* **7**: 95-99.

Rhyan, J. C., T. Gidlewski, T. J. Roffe, K. Aune, L. M. Philo and D. R. Ewalt (2001). "Pathology of brucellosis in bison from Yellowstone National Park." *J Wildl Dis* **37**(1): 101-9.

Riek, R. P., M. D. Handschumacher, S. S. Sung, M. Tan, M. J. Glynias, M. D. Schluchter, J. Novotny and R. M. Graham (1995). "Evolutionary conservation of both the hydrophilic and hydrophobic nature of transmembrane residues." *J Theor Biol* **172**(3): 245-58.

Ring, C. S. and F. E. Cohen (1994). "Conformational sampling of loop structures using genetic algorithm." *Isr J Chem* **34**: 245-252.

Ring, C. S., E. Sun, J. H. McKerrow, G. K. Lee, P. J. Rosenthal, I. D. Kuntz and F. E. Cohen (1993). "Structure-based inhibitor design by using protein models for the development of antiparasitic agents." *Proc Natl Acad Sci U S A* **90**(8): 3583-7.

Risler, J. L., M. O. Delorme, H. Delacroix and A. Henaut (1988). "Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix." *J Mol Biol* **204**(4): 1019-29.

Rose, G., L. Gierasch and J. Smith (1985). "Turns in peptides and proteins." *Adv Protein Chem* **37**: 1-109.

Rosenbach, D. and R. Rosenfeld (1995). "Simultaneous modeling of multiple loops in proteins." *Protein Sci* **4**(3): 496-505.

Rosenblatt, F. (1958). "The Perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review* **65**: 386-408.

Rossmann, M. G. and P. Argos (1976). "Exploring structural homology of proteins." *J Mol Biol* **105**(1): 75-95.

Rost, B. (1995). TOPITS: Threading One-Dimensional Prediction Into Three-dimensional Structures. The third international conference on Intelligent Systems for Molecular Biology (ISMB). C. Rawling, D. Clark, R. Altman*et al*. Cambridge, AAAI Press**:** 314-321.

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

Rost, B. (1999). "Twilight zone of protein sequence alignments." *Protein Eng* **12**(2): 85-94.

Rost, B. (2000). Better secondary structure prediction through more data, Columbia University.

Rost, B. (2001). "Review: protein secondary structure prediction continues to rise." *J Struct Biol* **134**(2-3): 204-18.

Rost, B., P. Fariselli and R. Casadio (1996). "Topology prediction for helical transmembrane proteins at 86% accuracy." *Protein Science* **7**: 1704-1718.

Rost, B. and C. Sander (1993). "Improved prediction of protein secondary structure by use of sequence profiles and neural networks." *Proc. Natl. Acad. Sci. USA* **90**(16): 7558-7562.

Rost, B. and C. Sander (1993). "Prediction of secondary structure at better than 70% accuracy." *Journal of Molecular Biology* **232**(2): 584-599.

Rost, B. and C. Sander (1995). "Progress of 1D protein structure prediction at last." *Proteins* **23**(3): 295-300.

Rost, B. and C. Sander (2000). "Third generation prediction of secondary structures." *Methods Mol Biol* **143**: 71-95.

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

Roterman, I. K., M. H. Lambert, K. D. Gibson and H. A. Scheraga (1989). "A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. Phi-psi maps for N-acetyl alanine N'-methyl amide: comparisons, contrasts and simple experimental tests." *J Biomol Struct Dyn* **7**(3): 421-53.

Rufino, S. D. and T. L. Blundell (1994). "Structure-based identification and clustering of protein families and superfamilies." *J Computer-Aided Mol Des* **8**: 5-27.

Russell, R. B. and G. J. Barton (1992). "Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels." *Proteins* **14**(2): 309-23.

Russell, R. B., R. R. Copley and G. J. Barton (1996). "Protein fold recognition by mapping predicted secondary structures." *J Mol Biol* **259**(3): 349-65.

Rutenber, E., M. Ready and J. D. Robertus (1987). "Structure and evolution of ricin B chain." *Nature* **326**(6113): 624-6.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream and B. Barrell (2000). "Artemis: sequence visualization and annotation." *Bioinformatics* **16**(10): 944-5.

Rychlewski, L., L. Jaroszewski, W. Li and A. Godzik (2000). "Comparison of sequence profiles. Strategies for structural predictions using sequence information." *Protein Sci* **9**(2): 232-41.

Rychlewski, L., B. Zhang and A. Godzik (1998). "Fold and function predictions for Mycoplasma genitalium proteins." *Fold Des* **3**(4): 229-238.

Salgado, H., G. Moreno-Hagelsieb, T. F. Smith and J. Collado-Vides (2000). "Operons in Escherichia coli: genomic analyses and predictions." *Proc Natl Acad Sci U S A* **97**(12): 6652-7.

Sali, A. (1995). "Modeling mutations and homologous proteins." *Curr Opin Biotechnol* **6**(4): 437-51.

Sali, A. and T. L. Blundell (1990). "Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming." *J Mol Biol* **212**(2): 403.

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

Sali, A., R. Matsumoto, H. P. McNeil, M. Karplus and R. L. Stevens (1993). "Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes." *J Biol Chem* **268**(12): 9023-34.

Sali, A. and J. P. Overington (1994). "Derivation of rules for comparative protein modeling from a database of protein structure alignments." *Protein Sci* **3**(9): 1582-96.

Sali, A., L. Potterton, F. Yuan, H. van Vlijmen and M. Karplus (1995). "Evaluation of comparative protein modeling by MODELLER." *Proteins* **23**(3): 318-26.

Sali, A., R. Sanchez and A. Badretdinov (1997). "MODELLER: A Program for Protein Structure Modeling Release 4." .

Salzberg, S. L., A. L. Delcher, S. Kasif and O. White (1998). "Microbial gene identification using interpolated Markov models." *Nucleic Acids Res* **26**(2): 544-8.

Samartino, L. E. and F. M. Enright (1993). "Pathogenesis of abortion of bovine brucellosis." *Comp Immunol Microbiol Infect Dis* **16**(2): 95-101.

Samudrala, R. and J. Moult (1998). "A graph-theoretic algorithm for comparative modeling of protein structure." *J Mol Biol* **279**(1): 287-302.

Sanchez, R. and A. Sali (1997). "Advances in comparative protein-structure modelling." *Curr Opin Struct Biol* **7**(2): 206-14.

Sanchez, R. and A. Sali (1997). "Evaluation of comparative protein structure modeling by MODELLER-3." *Proteins* **Suppl**(1): 50-8.

Sanchez, R. and A. Sali (1998). "Large-scale protein structure modeling of the Saccharomyces cerevisiae genome." *Proc Natl Acad Sci U S A* **95**(23): 13597-602.

Sander, C. and R. Schneider (1991). "Database of homology-derived protein structures and the structural meaning of sequence alignment." *Proteins* **9**(1): 56-68.

Sankoff, D. (1975). *SIAM J. Appl. Math.* **27**: 35.

Sankoff, D., R. J. Cedergren and G. Lapalme (1976). "Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA." *J Mol Evol* **7**(2): 133-49.

Santibanez, M. and K. Rohde (1987). "A multiple alignment program for protein sequences." *Comput Appl Biosci* **3**(2): 111-4.

Sauder, J. M., J. W. Arthur and R. L. Dunbrack Jr (2000). "Large-scale comparison of protein sequence alignment algorithms with structure alignments [In Process Citation]." *Proteins* **40**(1): 6-22.

Schrauber, H., F. Eisenhaber and P. Argos (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.

Schuler, G. D., S. F. Altschul and D. J. Lipman (1991). "A workbench for multiple alignment construction and analysis." *Proteins* **9**(3): 180-90.

Sheng, Y., A. Sali, H. Herzog, J. Lahnstein and S. Krilis (1996). "Modelling, expression and site-directed mutagenesis of human Beta2-glycoprotein I: Identification of the major phospholipid binding site." *J Immunol* **157**: 3744-3751.

Shi, J., T. L. Blundell and K. Mizuguchi (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." *J Mol Biol* **310**(1): 243-57.

Shindyalov, I. N. and P. E. Bourne (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path." *Protein Eng* **11**(9): 739-47.

Sibanda, B. L. and J. M. Thornton (1985). "Beta-hairpin families in globular proteins." *Nature* **316**(6024): 170-4.

Siew, N., A. Elofsson, L. Rychlewski and D. Fischer (2000). "MaxSub: an automated measure for the assessment of protein structure prediction quality [In Process Citation]." *Bioinformatics* **16**(9): 776-85.

Simons, K. T., C. Kooperberg, E. Huang and D. Baker (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." *J Mol Biol* **268**(1): 209-25.

Simons, K. T., I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff and D. Baker (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins." *Proteins* **34**(1): 82-95.

Sippl, M. J. (1990). “Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.” *Journal of Molecular Biology* **213**: 859-883.

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

Skolnick, J. and A. Kolinski (1990). “Simulations of the folding of a globular protein.” *Science* **250**: 1121-1125.

Smith, H. O., T. M. Annau and S. Chandrasegaran (1990). “Finding sequence motifs in groups of functionally related proteins.” *Proc. Natl. Acad. Sci. ASU* **87**(2): 826-830.

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.

Smith, T. F., L. Lo Conte, J. Bienkowska, C. Gaitatzes, R. G. Rogers, Jr. and R. Lathrop (1997). “Current limitations to protein threading approaches.” *J Comput Biol* **4**(3): 217-25.

Smith, T. F. and M. S. Waterman (1981). “Identification of common molecular subsequences.” *J Mol Biol* **147**(1): 195-7.

Sobel, E. and H. M. Martinez (1986). “A multiple sequence alignment program.” *Nucleic Acids Res* **14**(1): 363-74.

Sowdhamini, R., S. D. Rufino and T. L. Blundell (1996). “A database of globular protein structural domains: clustering of representative family members into similar folds.” *Fold Des* **1**(3): 209-20.

Spouge, J. L. (1989). “Speeding up dynamic programming algorithms for finding optimal lattice paths.” *SIAM J Appl Math* **49**(5): 1552-1566.

Spouge, J. L. (1991). “Fast optimal alignment.” *Comput Appl Biosci* **7**(1): 1-7.

Standley, D. M., J. R. Gunn, R. A. Friesner and A. E. McDermott (1998). “Tertiary structure prediction of mixed alpha/beta proteins via energy minimization.” *Proteins* **33**(2): 240-52.

Sternberg, M. J., P. A. Bates, L. A. Kelley and R. M. MacCallum (1999). “Progress in protein structure prediction: assessment of CASP3.” *Curr Opin Struct Biol* **9**(3): 368-73.

Stoye, J., V. Moulton and A. W. Dress (1997). “DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment.” *Comput Appl Biosci* **13**(6): 625-6.

Stroustrup, B. (1999). The C++ Programming Langage. Reading, Mass USA, Addison-Wesley.

Stuart, A. C., V. A. Ilyin and A. Sali (2002). “LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures.” *Bioinformatics* **18**(1): 200-1.

Subbiah, S. and S. C. Harrison (1989). "A method for multiple sequence alignment with gaps." *Mol. Biol.* **209**(4): 539-548.

Sun, S. (1993). "Reduced representation model of protein structure prediction: statistical potential and genetic algorithms." *Protein Sci* **2**(5): 762-85.

Sun, S., P. D. Thomas and K. A. Dill (1995). "A simple protein folding algorithm using a binary code and secondary structure constraints." *Protein Eng* **8**(8): 769-78.

Sutcliffe, M. J., I. Haneef, D. Carney and T. L. Blundell (1987). "Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures." *Protein Eng* **1**(5): 377-84.

Sutcliffe, M. J., F. R. Hayes and T. L. Blundell (1987). "Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains." *Protein Eng* **1**(5): 385-92.

Tajima, K. (1993). . Genome Informatics Workshop. T. T. e. al. Yokahoma, Universal Academy. **4:** 183.

Tatusov, R. L., S. F. Altschul and E. V. Koonin (1994). "Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks." *Proc Natl Acad Sci U S A* **91**(25): 12091-5.

Taylor, W. R. (1988). "A flexible method to align large numbers of biological sequences." *J Mol Evol* **28**(1-2): 161-9.

Taylor, W. R. (1995). "An investigation of conservation-biased gap-penalties for multiple protein sequence alignment." *Gene* **165**(1): GC27-35.

Taylor, W. R. (1996). "Multiple protein sequence alignment: algorithms and gap insertion." *Methods Enzymol* **266**: 343-67.

Thompson, J. D., G. H. Desmond and T. J. Gibson (1994). "Improved sensitivity of profile searches through the use of sequence weights and gap excision." *CABIOS* **10**(1): 19-29.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTALw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucleic Acid Research* **22**(22): 4673-4680.

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

Thompson, J. D., F. Plewniak and O. Poch (1999). "A comprehensive comparison of multiple sequence alignment programs." *Nucleic Acids Res* **27**(13): 2682-2690.

Thompson, J. D., F. Plewniak, R. Ripp, J. C. Thierry and O. Poch (2001). "Towards a reliable objective function for multiple sequence alignments." *J Mol Biol* **314**(4): 937-51.

Thompson, J. D., F. Plewniak, J. Thierry and O. Poch (2000). "DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches [In Process Citation]." *Nucleic Acids Res* **28**(15): 2919-26.

Topham, C. M., N. Srinivasan, C. J. Thorpe, J. P. Overington and N. A. Kalsheker (1994). "Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table." *Protein Eng* **7**(7): 869-94.

Torda, A. E. (1997). "Perspectives in protein-fold recognition." *Curr Opin Struct Biol* **7**(2): 200-5.

Tovchigrechko, A., C. A. Wells and I. A. Vakser (2002). "Docking of protein models." *Protein Sci* **11**(8): 1888-96.

Tramontano, A. (1998). "Homology modeling with low sequence identity." *Methods* **14**(3): 293-300.

Tramontano, A., R. Leplae and V. Morea (2001). "Analysis and assessment of comparative modeling predictions in CASP4." *Proteins* **Suppl**(5): 22-38.

Tryland, M., A. E. Derocher, Y. Wiig and J. Godfroid (2001). "Brucella sp. antibodies in polar bears from Svalbard and the Barents Sea." *J Wildl Dis* **37**(3): 523-31.

Tudos, E., M. Cserzo and I. Simon (1990). "Predicting isomorphic residue replacements for protein design." *Int J Pept Protein Res* **36**(3): 236-9.

Tybo and Goupil (1996). Le guide du jeune couple en BD.

Unger, R., D. Harel, S. Wherland and J. L. Sussman (1989). "A 3D building blocks approach to analyzing and predicting structure of proteins." *Proteins* **5**(4): 355-73.

Vajda, S., M. Sippl and J. Novotny (1997). "Empirical potentials and functions for protein folding and binding." *Curr Opin Struct Biol* **7**(2): 222-8.

Vakser, I. A. (1997). "Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex." *Proteins* **Suppl**(1): 226-30.

Valencia, A. (2003). "Meta, meta(n) and cyber servers." *Bioinformatics* **19**(7): 795.

van Gelder, C. W., F. J. Leusen, J. A. Leunissen and J. H. Noordik (1994). "A molecular dynamics approach for the generation of complete protein structures from limited coordinate data." *Proteins* **18**(2): 174-85.

van Gunsteren, W. F. and H. J. C. Berendsen (1990). "Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry." *Angew. Chem. Int. Ed. Engl.* **29**: 992-1023.

van Vlijmen, H. W. and M. Karplus (1997). "PDB-based protein loop prediction: parameters for selection and methods for optimization." *J Mol Biol* **267**(4): 975-1001.

Vasquez, M. (1996). "Modeling side-chain conformation." *Curr Opin Struct Biol* **6**(2): 217-21.

Vasquez, M., G. Nemethy and H. A. Scheraga (1994). "Conformational energy calculations on polypeptides and proteins." *Chem Rev* **94**: 2183-2239.

Vasquez, M. and H. A. Scheraga (1988). "Calculation of protein conformation by the build-up procedure. Application to bovine pancreatic trypsin inhibitor using limited simulated nuclear magnetic resonance data." *J Biomol Struct Dyn* **5**(4): 705-55.

Venclovas, C. (2001). "Comparative modeling of CASP4 target proteins: combining results of sequence search with three-dimensional structure assessment." *Proteins* **Suppl 5**: 47-54.

Venclovas, C., K. Ginalski and K. Fidelis (1999). "Addressing the issue of sequence-to-structure alignments in comparative modeling of CASP3 target proteins." *Proteins* **Suppl**(3): 73-80.

Vila, J., R. L. Williams, M. Vasquez and H. A. Scheraga (1991). "Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor." *Proteins* **10**(3): 199-218.

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

Vingron, M. and P. Argos (1989). "A fast and sensitive multiple sequence alignment algorithm." *Comput Appl Biosci* **5**(2): 115-21.

Vingron, M. and P. Argos (1991). "Motif recognition and alignment for many sequences by comparison of dot- matrices." *J Mol Biol* **218**(1): 33-43.

Vingron, M. and P. A. Pevzner (1995). *Adv Appl Math* **16**: 1.

Viterbi, A. (1967). "Error Bound for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm." *IEEE Transactions on Information Theory* **13**(2): 260-269.

Vogt, G., T. Etzold and P. Argos (1995). "An Assessment of Amino Acid Exchange Matrices in Aligning Protein Sequences: The Twilight Zone Revisited." *Journal of Molecular Biology* **249**: 816-831.

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

Wall, L., T. Christiansen and J. Orwant (2000). Programming Perl. Sebastopol, CA, O'Reilly & Associates, Inc.

Wallace, A. C., N. Borkakoti and J. M. Thornton (1997). "TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites." *Protein Sci* **6**(11): 2308-23.

Wallace, A. C., R. A. Laskowski and J. M. Thornton (1996). "Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases." *Protein Sci* **5**(6): 1001-13.

Waterman, M. S. and M. Perlwitz (1984). *Bull Math Biol* **46**: 567.

Waterman, M. S., T. F. Smith and W. A. Beyer (1976). *Adv Math* **20**: 367.

Wei, L., R. B. Altman and J. T. Chang (1997). "Using the radial distributions of physical features to compare amino acid environments and align amino acid sequences." *Pac Symp Biocomput*: 465-76.

Westbrook, J., Z. Feng, L. Chen, H. Yang and H. M. Berman (2003). "The Protein Data Bank and structural genomics." *Nucleic Acids Res* **31**(1): 489-91.

Widenius, M. M., D. Axmark and M. AB (2002). MySQL Reference Manual, O'Reilly and Associates.

Wilkins, M. R., E. Gasteiger, A. Bairoch, J.-C. Sanchez, K. L. Williams, A. R.D. and H. D.F. (1998). Protein Identification and Analysis Tools in the ExPASy Server. 2-D Proteome Analysis Protocols. A. J. Link. Nashville, TN, USA, Humana Press.

Wilson, C. and S. Doniach (1989). "A computer model to dynamically simulate protein folding: studies with crambin." *Proteins* **6**(2): 193-209.

Wilson, K. S., Z. Dauter, V. S. Lamsin, M. Walsh, S. Wodack, J. Richelle, J. Pontius, A. Vaguine, R. W. W. Hooft, C. Sander, G. Vriend, J. M. Thornton, R. A. Laskowski, M. W. MacArthur, E. J. Dodson, G. Murshudov, T. J. Oldfield, R. Kaptein and J. A. C. Rullman (1998). "Who checks the checkers? Four validation tools applied to eight atomic resolution structures. EU 3-D Validation Network." *J Mol Biol* **276**(2): 417-436.

Wolf, E., A. Vassilev, Y. Makino, A. Sali, Y. Nakatani and S. K. Burley (1998). "Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase." *Cell* **94**(4): 439-49.

Wood, D. W., J. C. Setubal, R. Kaul, D. E. Monks, J. P. Kitajima, V. K. Okura, Y. Zhou, L. Chen, G. E. Wood, N. F. Almeida, Jr., L. Woo, Y. Chen, I. T. Paulsen, J. A. Eisen, P. D. Karp, D. Bovee, Sr., P. Chapman, J. Clendenning, G. Deatherage, W. Gillet, C. Grant, T. Kutyavin, R. Levy, M. J. Li, E. McClelland, A. Palmieri, C. Raymond, G. Rouse, C. Saenphimmachak, Z. Wu, P. Romero, D. Gordon, S. Zhang, H. Yoo, Y. Tao, P. Biddle, M. Jung, W. Krespan, M. Perry, B. Gordon-Kamm, L. Liao, S. Kim, C. Hendrick, Z. Y. Zhao, M. Dolan, F. Chumley, S. V. Tingey, J. F. Tomb, M. P. Gordon, M. V. Olson and E. W. Nester (2001). "The genome of the natural genetic engineer Agrobacterium tumefaciens C58." *Science* **294**(5550): 2317-23.

Wu, C. H., H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Z. Hu, R. S. Ledley, K. C. Lewis, H. W. Mewes, B. C. Orcutt, B. E. Suzek, A. Tsugita, C. R. Vinayaka, L. S. Yeh, J. Zhang and W. C. Barker (2002). "The Protein Information Resource: an integrated public resource of functional annotation of proteins." *Nucleic Acids Res* **30**(1): 35-7.

Wu, G., A. Fiser, B. ter Kuile, A. Sali and M. Muller (1999). "Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase." *Proc Natl Acad Sci U S A* **96**(11): 6285-90.

Xu, L. Z., R. Sanchez, A. Sali and N. Heintz (1996). "Ligand specificity of brain lipid-binding protein." *J Biol Chem* **271**(40): 24711-9.

Xu, Y. and D. Xu (2000). "Protein threading using PROSPECT: Design and evaluation." *Proteins* **40**(3): 343-354.

Yao, H., D. M. Kristensen, I. Mihalek, M. E. Sowa, C. Shaw, M. Kimmel, L. Kavraki and O. Lichtarge (2003). "An accurate, sensitive, and scalable method to identify functional sites in protein structures." *J Mol Biol* **326**(1): 255-61.

Yue, K. and K. A. Dill (1996). "Folding proteins with a simple energy function and extensive conformational searching." *Protein Sci* **5**(2): 254-61.

Zemla, A. (2000). "LGA program: A Method for Finding 3-D Similarities in Protein Structures." *Accessed at http://PredictionCenter.llnl.gov/local/lga*.

Zhang, B., L. Jaroszewski, L. Rychlewski and A. Godzik (1998). "Similarities and differences between nonhomologous proteins with similar folds: evaluation of threading strategies." *Fold Des* **2**: 307-317.

Zhang, C. and A. K. Wong (1997). "A genetic algorithm for multiple molecular sequence alignment." *Comput Appl Biosci* **13**(6): 565-81.

Zhang, C. T. (1997). "Relations of the numbers of protein sequences, families and folds." *Protein Eng* **10**(7): 757-61.

Zhang, X., J. P. Mesirov and D. L. Waltz (1992). "Hybrid system for protein secondary structure prediction." *J Mol Biol* **225**(4): 1049-63.

Zheng, Q., R. Rosenfeld, C. DeLisi and D. J. Kyle (1994). "Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral angle perturbations." *Protein Sci* **3**(3): 493-506.

Zheng, Q., R. Rosenfeld, S. Vajda and C. DeLisi (1993). "Determining protein loop conformation using scaling-relaxation techniques." *Protein Sci* **2**(8): 1242-8.

Zvelebil, M. J., G. J. Barton, W. R. Taylor and M. J. Sternberg (1987). "Prediction of protein secondary structure and active sites using the alignment of homologous sequences." *J Mol Biol* **195**(4): 957-61.