**MASTER EN SCIENCES BIOLOGIQUES DES ORGANISMES ET ÉCOLOGIE**

# Recherche de l'augmentation de la sensibilité d'une méthode d'alignement multiple de séquences protéiques, sans perte de spécificité

Noël, Nicolas

*Award date:*
1999


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

**Faculté des Sciences**

# RECHERCHE DE L'AUGMENTATION DE LA SENSIBILITE D'UNE METHODE D'ALIGNEMENT MULTIPLE DE SEQUENCES PROTEIQUES, SANS PERTE DE SPECIFICITE

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

Nicolas NOEL

Juin 1999

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

# Recherche de l'augmentation de la sensibilité d'une méthode d'alignement multiple de séquences protéiques, sans perte de spécificité

NOEL Nicolas

Résumé

La quantité de données biologiques, et notamment celle des séquences d'ADN, a augmenté considérablement ces dernières années. Suite à cette augmentation importante d'informations disponibles, les logiciels d'alignement de séquences deviennent donc des outils de plus en plus indispensables, puisqu'ils constituent une étape préliminaire indispensable dans la prédiction de structures protéiques par « Homology Modeling ».

Notre travail a pour but d'améliorer le logiciel d'alignement multiple de séquences « Match-Box », mis au point depuis quelques années par notre laboratoire, ceci afin d'augmenter la confiance et la puissance des résultats de son alignement.

La première partie du travail a consisté à élaborer différentes techniques en vue de choisir une batterie de groupes de protéines ou « cas-tests », indispensables à la quantification des performances du logiciel Match-Box.

Une fois ces cas-tests déterminés, notre travail s'est focalisé sur l'étude approfondie de la troisième étape de Match-Box (le *screening*) en vue de l'amélioration de ses performances. Il s'agit de l'algorithme qui effectue le tri entre les appariements fiables et les appariements parasites engendrés par l'étape précédente (le *matching*). Cette étude s'est effectuée à différents niveaux : en faisant varier la taille de la fenêtre d'analyse et en tenant compte de la conservation des structures secondaires dans les boîtes alignées. Une nouvelle procédure de screening a été également mise en œuvre au cours de ce mémoire.

Les résultats obtenus par ces différentes techniques montre une amélioration non négligeable des performances de Match-Box, se traduisant par une meilleure qualité de l'alignement multiple des familles de protéines testées.

Mémoire de licence en Sciences biologiques

Juin 1999

**Promoteur:** E. Depiereux

"Le commencement de toutes les sciences, c'est l'étonnement de ce que les choses sont ce qu'elles sont."

*Aristote*

Je tiens tout d'abord à remercier chaleureusement le Professeur Depiereux pour m'avoir ouvert les portes de son laboratoire et m'avoir suivi tout au long de ce mémoire.

Je remercie les membres du jury, D. Devos, S. Hamels, J. Messiaen, J. Wouters, d'avoir accepté de consacrer un peu de leur temps à la lecture de ce mémoire.

J'adresse toute ma reconnaissance à Christophe, pour tous les conseils et les encouragements reçus tout au long de ce travail, arrivant toujours au moment précis.

Mes remerciements vont également à Katalin, Isabelle, Thynna, Stéphanie, Christophe, Philippe, Etienne, Carlos, Jean-Yves, Bernard, sans qui ce laboratoire ne serait pas ce qu'il est, et à tous les membres de l'URBM pour les discussions fructueuses lors des réunions.

Tout grand merci à Natacha, Françoise, Nancy, Elisabeth, Benoît, Jean-Marc, Benjamin, Thierry, Pierre, François pour toutes les activités "extra-mémoire".

Enfin, ce mémoire n'aurait jamais vu le jour sans mes parents et Isabelle. Je les remercie du fond du cœur pour leur présence permanente à mes côtés, dans les bons moments comme dans les moins bons.

Si ce n'est déjà fait, je vous remercie vous, cher lecteur, qui allez parcourir ces pages relatant ce mémoire qui fut ma passion pendant l'année qui vient de s'écouler, et vous souhaite une agréable lecture.

## ABREVIATIONS

| | |
|---|---|
| aa | Acide aminé |
| ADN | Acide désoxyribonucléique |
| PDB | Protein data bank |
| DSSP | Dictionary of protein secondary structure |
| RMS | Root means square |
| SCE | Somme des carrés des écarts |
| PAM | Point accepted mutation |
| SCR | structurally conserved regions |
| RAM | Random accessible memory |
| MHz | Mega hertz |
| GB | Giga bits |
| Mips | Multy initial processor |
| TCP/IP | Transmission Control protocol/Internet protocol |
| WWW | world wide web |
| SMTP | Simple mail transfer protocol |
| MB | Match-Box |
| ClW | ClustalW |
| SS | Structure secondaire |

# PLAN DU TRAVAIL

# INTRODUCTION

# Avant-Propos

La quantité de données biologiques, et notamment celle des séquences d'ADN, a augmenté considérablement ces dernières années. De plus, aux alentours de l'an 2005, le génome humain sera complètement séquencé (Boguski, 1998). Nous pourrions considérer cet aboutissement non comme étant la fin d'une étape mais comme le commencement de beaucoup d'autres.

Pour pouvoir exploiter cette multitude de données, nous disposons d'une science qui lie à la fois la biologie et l'informatique, à savoir la bio-informatique (Boguski, 1998). Celle-ci possède la qualité d'aborder tous les domaines de la biologie en permettant à tous et toutes d'obtenir les informations souhaitées via des programmes spécialisés ou bien par Internet.

Trouver un gène dans une séquence d'ADN génomique n'est pas une tâche facile et encore moins lui assigner une fonction. Par la suite, une compréhension des interactions entre les gènes et leur produit doit être assurée. Cette compréhension ne doit pas se limiter aux contextes des voies métaboliques à l'intérieur et entre les cellules, mais doit aussi s'élargir en considérant l'évolution de familles de gènes au sein d'une espèce et entre différentes espèces.

La bio-informatique regroupe plusieurs domaines, en aval du séquençage. Ce domaine est en grand développement pour le moment. En effet, de nombreux génomes sont complètement séquencés ou en voie de séquençage, et tous les fragments sont stockés dans des banques. De ce fait, l'augmentation du nombre de séquences est exponentielle par opposition à l'augmentation de la connaissance de leurs fonctions et leurs structures, qui dépendent de l'expérimentation. En effet, ces fonctions sont en général inférées par similarité avec une fiabilité douteuse et non évaluées. De plus, les banques de stockage comme par exemple Genbank sont devenues de larges complexes d'archives de séquences redondantes qui requièrent une expérience considérable pour leur utilisation efficace (Boguski, 1998).

Face à ce problème, une grande quantité de nouveaux logiciels se développent pour maximiser l'utilisation de ces données. Ces outils informatiques devraient remplir plusieurs rôles. Citons-en quelques-uns :

- Eviter la redondance dans les banques de données ;
- intégrer les données de séquençage de génomes entiers avec les analyses expérimentales et automatiques de ces séquences comme par exemple la détection des ORF, les sites de régulation, les voies métaboliques... ;
- Disposer d'une échelle d'exactitude sur les analyses automatiques.

Parmi ces techniques, l'alignement de séquences a pour but d'étudier une correspondance entre les résidus les plus similaires afin de nous éclairer sur des régions de séquences conservées structurellement et/ou fonctionnellement. Les résultats qui en découleront pourront servir à une première modélisation de la structure tridimensionnelle des protéines. L'alignement sert aussi à déterminer des résidus potentiellement essentiels, à émettre des hypothèses quant à leur rôle qui sera validé par mutagenèse ... Enfin, l'alignement permet la recherche fonctionnelle par homologie et phylogénie, et accessoirement à la détermination d'oligonucléotides dégénérés permettant le clonage de gènes par amplification PCR.

Toutes les méthodes d'alignement dépendent d'un jeu de paramètres essentiel qui porte le nom de matrice de scores. L'amélioration des performances des méthodes d'alignement est donc en partie liée a l'optimisation de ces scores. L'optimisation des méthodes d'alignement dépend quant à elle de l'amélioration des matrices de scores. Celles-ci sont représentées dans des tableaux 20 X 20 permettant de comparer les différents acides aminés deux à deux en fonction de leur similarité ou de leur distance et de leur attribuer un score. Cependant, il existe beaucoup de matrices de scores différentes construites sur base des caractéristiques physico-chimiques des divers acides aminés. Nous pouvons distinguer deux types différents de matrices de scores. D'une part, les matrices de similarité mettent en évidence des acides aminés qui possèdent des caractéristiques similaires ou plus ou moins similaires et d'autre part, les matrices de distance sont des matrices de scores dans lesquelles plus les acides aminés sont différents, plus score est élevé. Il est possible de transformer une matrice

de similarité en matrice de distance, en retirant le score maximum à toutes les valeurs et en changeant de signe.

Ces matrices de scores se divisent en différents sous groupes. Les matrices d'identité, de substitution, de structure, de mutation, de code génétique et enfin les matrices physico-chimiques.

La réalisation d'alignement de séquences, peut se faire soit de manière pairée (alignement de deux séquences) soit de manière multiple (alignement de plus de deux séquences).

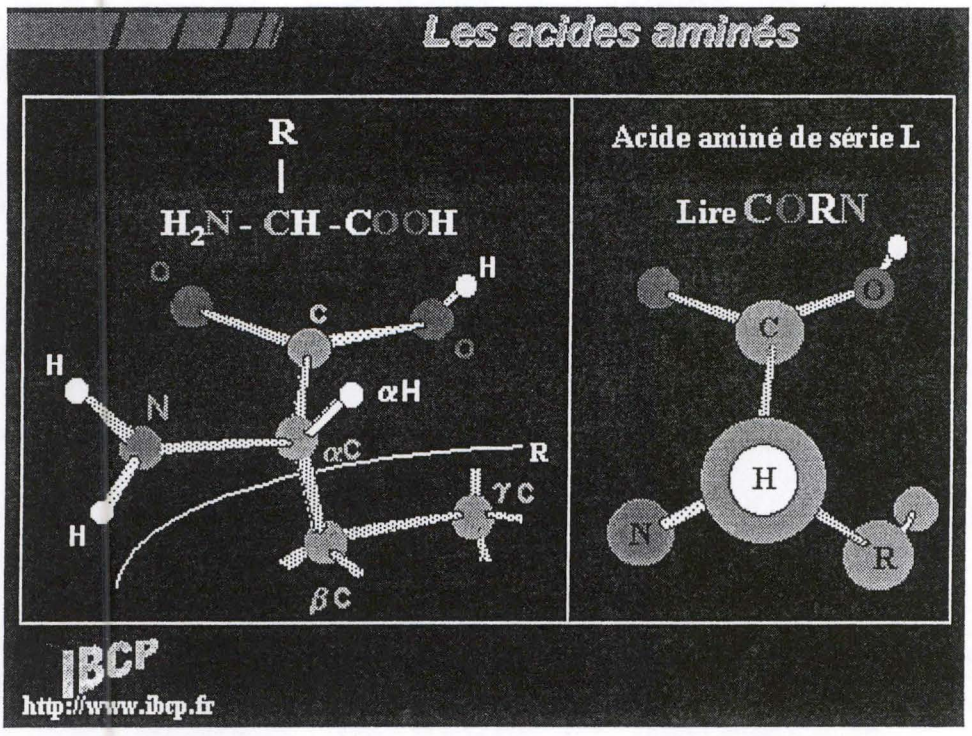



*Figure : 1.1.1.1*
*Représentation d'un acide aminé.*

# Chapitre 1 : Prédiction de structures protéiques

## 1.1 Rappel : les différents niveaux de structures protéiques

### 1.1.1 Structures primaires

Les protéines sont des chaînes non ramifiées d'acides aminés unis par liens covalents.

Les acides aminés sont au nombre de 20 différents et sont tous fabriqués suivant le même modèle : le carbone central (carbone) porte dans tous les cas un atome d'hydrogène, une fonction carboxyle, une fonction amine, et un radical R (ou chaîne latérale). Ce qui distingue les 20 acides aminés est la nature et les propriétés de cette chaîne latérale (*figure 1.1.1.1*).

En général, l'atome central d'un acide aminé est un carbone asymétrique toujours présent sous la forme L (lévogyre). La glycine et la proline font toutes deux exception à cette règle. La proline forme un cycle par un embranchement de la chaîne latérale sur l'atome d'azote. Le radical de la glycine est uniquement constitué d'un atome d'hydrogène ; en conséquence, son carbone central n'est pas asymétrique (Lodish *et al.*, 1997).

### 1.1.2 Structures secondaires

#### *1.1.2.1 L'hélice α*

Ce n'est qu'au début des années 50 que Linus Pauling s'aperçut que les segments polypeptidiques composés de certains acides aminés tendent à s'enrouler en spirale régulière, c'est-à-dire en hélice α (Lodish *et al.*, 1997). En fait, l'oxygène du

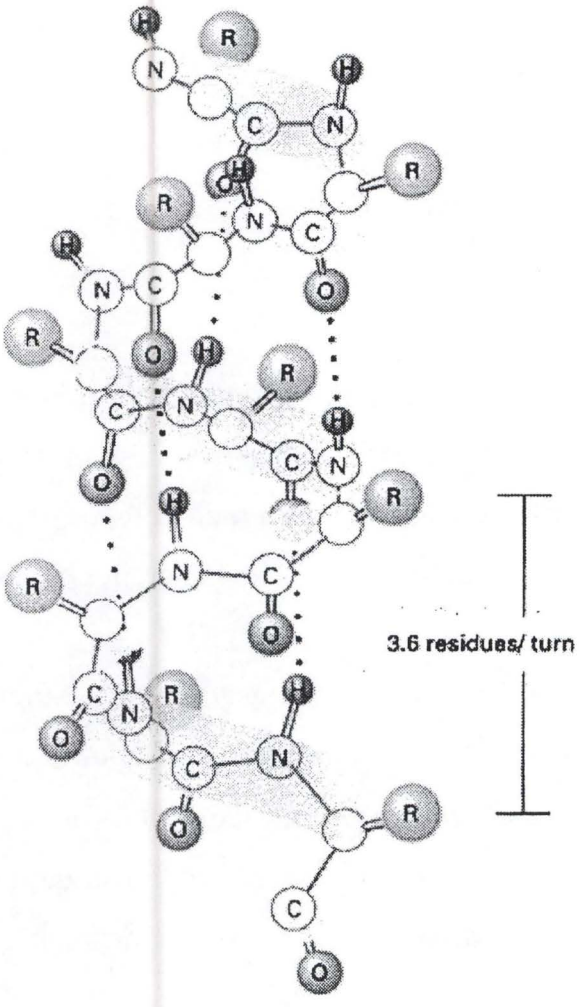



*Figure : 1.1.2.1.1*
*Représentation d'une hélice α.*

carbonyle de toute liaison peptidique (soit le n°1) y forme une liaison hydrogène avec l'hydrogène amide de la liaison peptidique n°4 qui la suit du côté carboxylique (*figure 1.1.2.1.1*). Le squelette peptidique s'enroule en une hélice de 3,6 résidus par tour et chaque résidu allonge cet « escalier en spirale » d'environ 0,15 nm dans l'axe de l'hélice. Le déplacement est donc de 5,4 nm le long de l'axe de l'hélice lorsque l'on effectue un tour de spire. C'est ce qu'on nomme le pas de l'hélice. Dans l'hélice α, cette disposition stable et rigide des acides aminés maintient la structure en forme de squelette allongé, d'où sortent les chaînes latérales. Le caractère hydrophile ou hydrophobe de l'hélice α dépend des chaînes latérales, puisque les groupes polaires du squelette peptidique sont déjà mutuellement engagés en liaisons hydrogène et ne peuvent plus contribuer au caractère hydrophobe ou hydrophile de l'hélice. Nous pouvons remarquer, sans en connaître la raison, que certaines séquences d'acides aminés s'organisent plus facilement que d'autres en hélice α et que certains facteurs agissent plutôt comme déstabilisateurs de l'hélice. Par exemple, la proline se trouve rarement dans ces structures secondaires de type hélice α parce que les angles de liaisons portés par son carbone α déformeraient le squelette hélicoïdal et que son azote ne pourrait participer à une liaison hydrogène. De même, et pour des raisons opposées, l'hélice α contient rarement de la glycine car son résidu, qui n'est qu'un atome d'hydrogène, augmente la flexibilité autour du carbone α, ce qui provoque une inflexion de l'hélice à ce niveau.

En tant que cylindre rigide, l'hélice α est un module structural qui participe à différentes fonctions biologiques. Par exemple une des fonctions purement structurale de l'hélice α est celle que l'on rencontre dans les protéines fixatrices de l'oxygène, à savoir la myoglobine et l'hémoglobine. Ces protéines sont soutenues par huit courts cylindres d'hélice α unis par des coudes. Par contre, dans d'autres protéines, l'hélice α sert de site d'interaction avec d'autres protéines ou de site de liaison à l'ADN.

### 1.1.2.2 Le plan ß

Le plan (ou feuillet) ß est formé de brins ß serrés côte à côte. Les brins β sont de

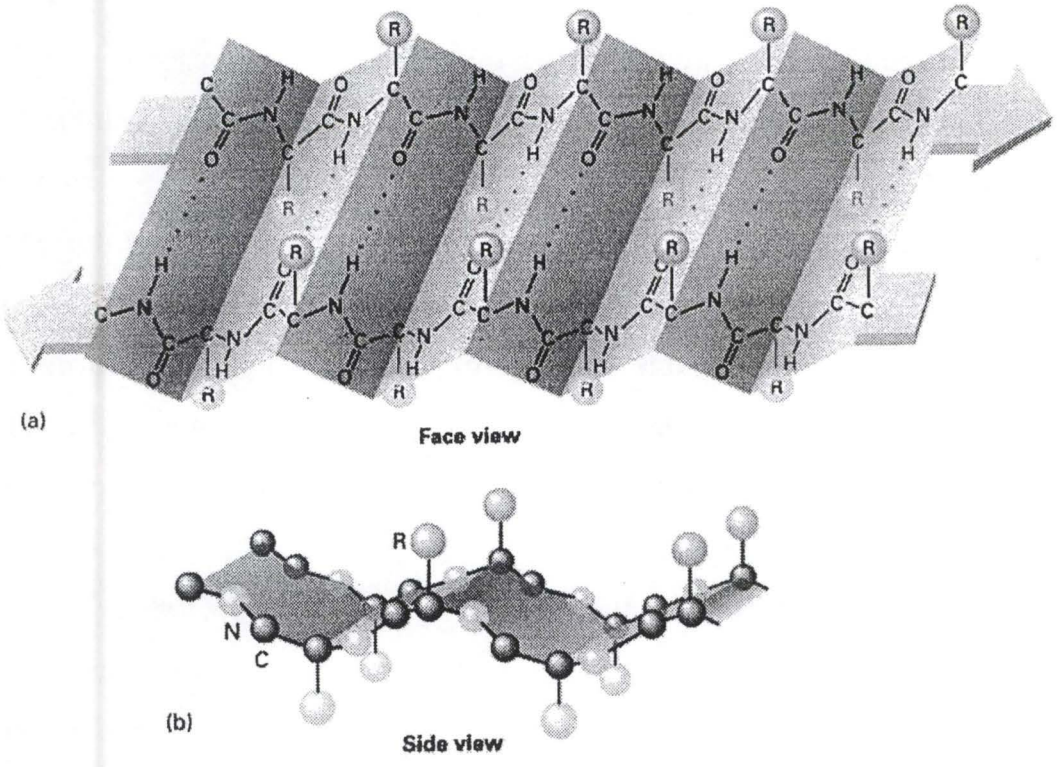



*Figure : 1.1.2.2.1*
*Représentation d'un plan ß.*

courtes chaînes polypeptidiques de 5 à 8 résidus presque entièrement étirés *(figure 1.1.2.2.1)*. En fait, les atomes du squelette du brin ß se lient en liaison hydrogène avec ceux d'une même ou d'une autre chaîne. Le plan ß possède aussi une polarité qui est régie par l'orientation de la liaison peptidique. Si on examine un feuillet plissé, les brins ß voisins sont soit parallèles soit antiparallèles. Cette différence se marque seulement par le fait que la suite des amino-carbonyles des liaisons peptidiques des chaînes adjacentes est soit dans le même sens, soit dans le sens inverse. Mais dans les deux cas, les chaînes latérales font saillie sur chaque face du feuillet (Lodish *et al.*, 1997).

Par opposition à l'hélice, la proline est parfois présente malgré que les groupements chargés disloquent les feuillets. Les gros groupements ou les groupements chargés quant à eux se retrouvent rarement dans les feuillets ß.

Comme pour l'hélice α, le feuillet ß peut intervenir de différentes façons dans les fonctions cellulaires. Par exemple, il peut servir de plancher à une poche de liaison. D'autre part, un empilement de feuillets ß confère la rigidité de nombreuses protéines de structure comme la fibroïne de la soie, qui n'est faite presque exclusivement que de feuillets antiparallèles. En fait, les fibres de la soie sont flexibles, car les feuillets peuvent glisser les uns sur les autres. Elles sont en outre résistantes à la traction, car le squelette peptidique court parallèlement à l'axe de la fibre. Par l'analyse de la structure secondaire de diverses protéines fibreuse de certaines espèces, on a remarqué que certaines d'entre elles sont presque entièrement étirées. Ceci est dû au fait que la liaison hydrogène entre l'oxygène du groupement carbonyle et l'hydrogène de l'amide est presque perpendiculaire au grand axe de la chaîne du polypeptide. Et c'est justement la formation de la liaison hydrogène entre deux ou plusieurs chaînes étirées adjacentes qui permet la formation du feuillet plissé.

L'analyse de cette structure permet de voir que celle-ci n'est pas tout à fait plane mais légèrement plissée. Nous constatons que cela est dû aux angles de liaison adoptés par la chaîne polypeptidique.

### *1.1.2.3 Les boucles*

Ces structures secondaires sont en forme de boucle connectant des structures

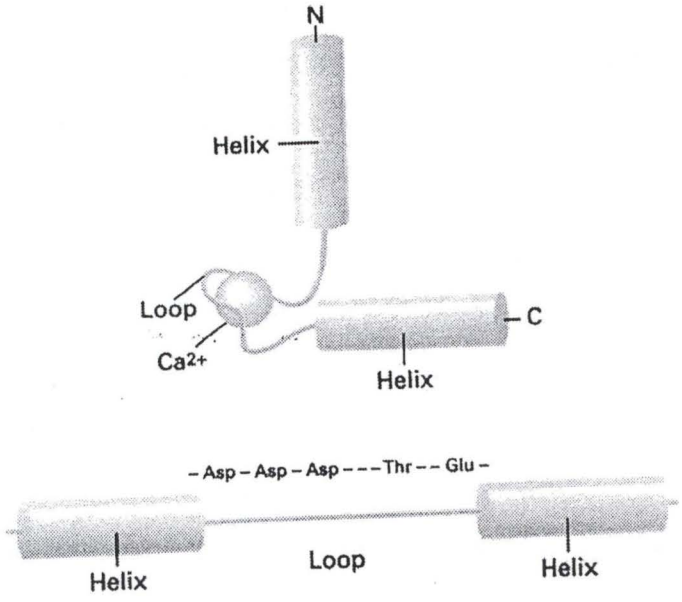



*Figure : 1.1.2.3.1*
*Représentation d'une boucle.*

secondaires *(figure 1.1.2.3.1)*. Ces fragments polypeptidiques sont de longueurs et de conformations tout à fait irrégulières. Ils ont pour rôle d'effectuer la liaison entre les différents éléments de structure et surtout de les maintenir au centre de la protéine. Du fait de leur situation en surface de la protéine, les boucles sont souvent riches en acides aminés hydrophiles et chargés pouvant interagir avec les solvants. L'évolution a montré que les régions au niveau des boucles ne sont pas conservées mais très variables. Cependant, il y a des exceptions à cette règle. En effet, il existe des cas où ces boucles auraient un rôle fonctionnel (ex. : cofacteur, fixation d'un substrat...) et seraient dès lors invariables. (Lodish *et al.*, 1997).

### *1.1.2.4 Le coude*

Formés de 3 ou 4 résidus, les coudes sont des structures secondaires en U maintenues par une liaison hydrogène entre les deux résidus situés aux extrémités du coude. Les coudes se trouvent à la surface des protéines, ce qui signifie que la plupart des acides aminés que l'on trouve en leur sein sont hydrophiles et chargés (Lodish *et al.*, 1997). En effet, ceux-ci doivent interagir avec le substrat et permettre le repliement du squelette polypeptidique vers l'intérieur. Les séquences des coudes comportent souvent de la glycine et de la proline. Une protéine dépourvue de coudes serait volumineuse, étirée et peu compacte. Leur rôle n'est pas d'induire un changement de direction mais de servir de lien entre les différentes structures secondaires.

Lors d'une réaction de polymérisation, un lien peptidique partiellement double se forma par perte d'une molécule d'eau. Ce lien est stabilisé par un phénomène de résonance établi par un partage d'électron entre les deux liaisons C-O et C-N. Lorsque les acides aminés s'additionnent les uns à la suite des autres, des interactions entre les chaînes latérales de ces acides aminés se produisent. Par conséquent, il faut que ces protéines trouvent une conformation stable au niveau énergétique par rotation autour des liaisons libres (Lodish *et al.*, 1997).

Ces liaisons sont : $\phi$ pour la liaison C-N

$\psi$ pour la liaison C-C

Ces liaisons libres peuvent tourner et adopter ainsi des angles de torsions différents

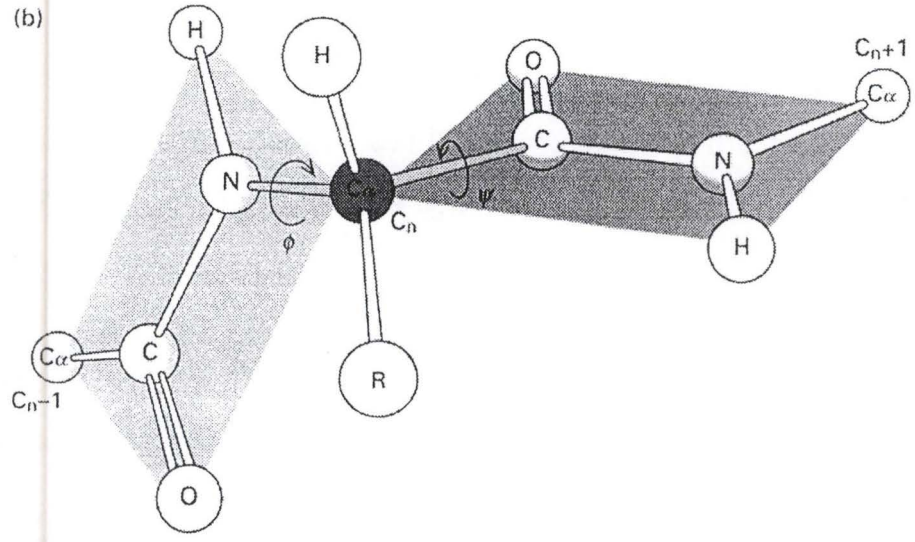



*Figure : 1.1.2.4.1*
*Localisation des différents angles de torsion le long de la chaîne peptidique.*

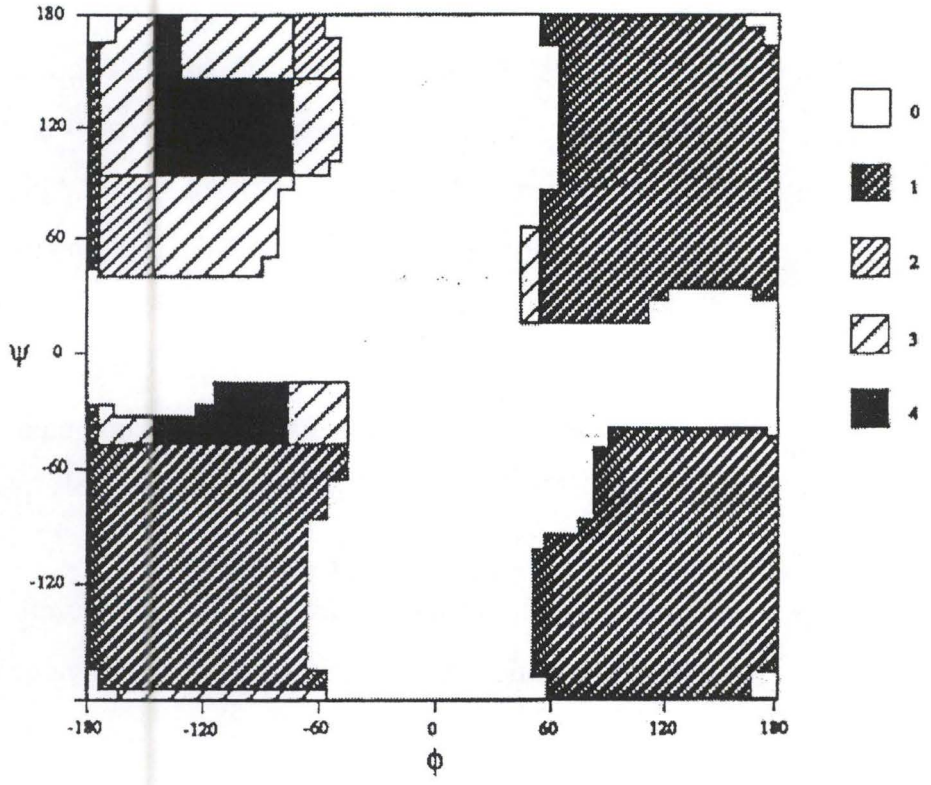



*Figure : 1.1.2.4.2*
*Graphe de Ramachandran indiquant la répartition théorique des angles de torsion $\phi$ et $\psi$ pour les acides aminés.*

*(figure 1.1.2.4.1)*. L'amplitude de ces angles est généralement caractérisée par rapport à la chaîne principale. Selon les diverses liaisons et selon les différents angles de torsions, nous constatons que les angles ne prennent pas toutes les valeurs possibles. En effet, selon l'idée du biochimiste indien G.N. Ramachandran, la conformation d'une chaîne polypeptidique peut être complètement définie quand on représente chaque résidu d'acide aminé de la chaîne par ses valeurs $\phi$ et $\psi$ sur un diagramme bidimensionnel *(figure 1.1.2.4.2)*. Ce diagramme de Ramachandran spécifie les conformations habituellement observées dans une chaîne polypeptidique. Les valeurs de $\psi$ et $\phi$ pour tous les résidus d'une protéine tombent le plus souvent dans les zones colorées limitées par les pointillés. Cependant, les paramètres de certains acides aminés tombent en dehors des zones spécifiées. On explique cela en admettant que la structure globale d'une protéine résulte d'un compromis par lequel une interaction défavorable est tolérée à un endroit pour en permettre une plus favorable ailleurs. En outre, on observe que si les résidus appartiennent à une hélice $\alpha$ ou à un brin $\beta$, les valeurs des angles observés ne sont pas aléatoires, mais caractéristiques de ces deux types de structures secondaires.

## 1.1.3 Superstructures secondaires

Ces structures sont fréquemment rencontrées dans les protéines. Elles résultent d'un enchaînement de structures secondaires qui adoptent une conformation particulière en relation avec une fonction particulière dans la cellule (Lodish *et al.*, 1997).
Les structures secondaires existent sous différentes formes dont voici quelques exemples.

### *1.1.3.1 Coiled-coil/hélice $\alpha$*

Cette superstructure secondaire correspond à deux hélices $\alpha$ qui s'enroulent l'une autour de l'autre *(figure 1.1.3.1.1)*. Nous avons observé ce genre de structure chez les protéines fibreuses et globuleuses (Lodish *et al.*, 1997).

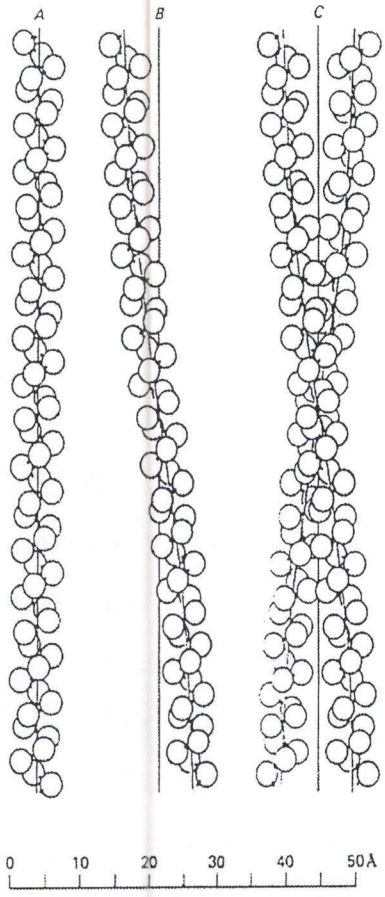



***Figure : 1.1.3.1.1***
*Représentation d'un Coiled-coil/hélice α.*

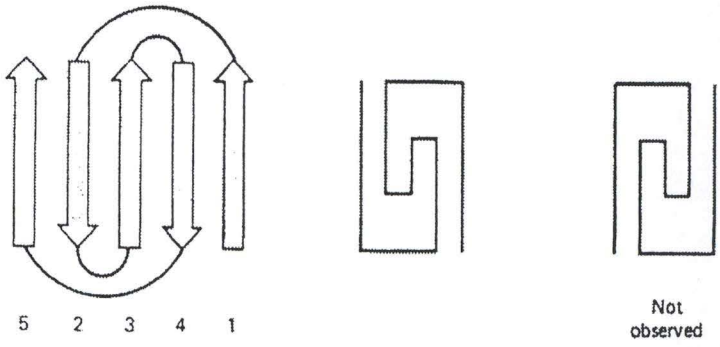



***Figure : 1.1.3.3.1***
*Représentation des clés grecques.*

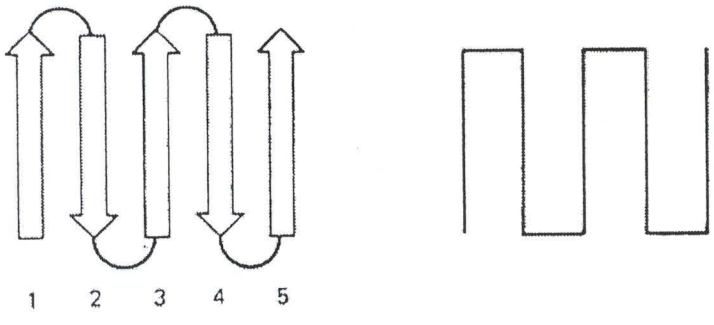



***Figure : 1.1.3.4.1***
*Représentation du motif ß/ß.*

### *1.1.3.2 Hélice α-loop-hélice α*

On trouve souvent ces structures chez les protéines qui fixent le Calcium et l'ADN *(figure 1.1.2.3.1)* (Lodish *et al.*, 1997).

### *1.1.3.3 Clés grecques*

Elle se compose de brins ß antiparallèles liés par des loops. On les observe dans les feuillets ß antiparallèles *(figure 1.1.3.3.1)* (Lodish *et al.*, 1997).

### *1.1.3.4 Motif ß/ß*

Pour former cette superstructure secondaire, un loop joint deux brins β antiparallèles *(figure 1.1.3.4.1)*. On les trouve souvent dans les plans ß (Lodish *et al.*, 1997).

## 1.1.4 Structure tertiaire

Les structures et superstructures secondaires sont agencées dans la protéine de façon compacte, dans le but de former un ou plusieurs domaines (Lodish *et al.*, 1997). Ceux-ci sont chacun responsables d'une fonction particulière comme la fixation de substrats ou de coenzymes par exemple. Le repliement de la protéine sur elle-même permet à la chaîne polypeptidique de positionner les résidus importants pour sa fonction, même s'ils sont très distants au niveau de la séquence.

## 1.1.5 Structure quaternaire

Les protéines multimériques sont constituées de plusieurs unités protéiques ou monomères (homodimère et hétérodimère) ou plusieurs structures tertiaires qui s'associent en une structure quaternaire. Les monomères de cette structure sont soit identiques soit de nature et donc de fonctions différentes. Au niveau de la surface de contact entre les domaines, il se crée des interactions qui sont de même nature que celles existant à l'intérieur de la protéine (Lodish *et al.*, 1997).

## 1.2 Prédiction de structures secondaires

Pour la prédiction des structures secondaires, diverses méthodes sont utilisées.

- La première méthode est une méthode statistique qui se base sur des paramètres empiriques déterminés à partir des structures connues des protéines ;
- la deuxième se base sur des considérations physico-chimiques, c'est-à-dire qu'elle tient compte des facteurs stéréochimiques ou de l'hydrophobicité des acides aminés de la protéine d'intérêt ;
- la troisième se nomme la méthode des réseaux neuronaux. Son rôle est de reproduire de manière artificielle ce qui se passe dans le cerveau. Les réseaux neuronaux sont constitués de différentes unités ou « neurones » qui intègrent leurs propres données et les transforment en réponses qui sont transmises aux autres unités, puisqu'elles sont connectées en parallèles. Ce système doit être entraîné sur un jeu de protéines tests et peut ensuite être extrapolé sur des protéines de structure inconnue.

## 1.3 Prédiction de structure tertiaire

Deux conditions sont nécessaires pour procéder à la modélisation par homologie d'une protéine :

1) Sa séquence primaire doit être connue ;
2) Il faut disposer d'une protéine homologue dont la structure a été déterminée de façon expérimentale, et qui dès lors servira de structure de référence.

La difficulté de la modélisation par homologie dépend du taux d'identité entre les séquences cible et la séquence d'intérêt. Ce pourcentage de similarité peut être caractérisé par différentes valeurs (C. Vinals, communication personnelle) :

- >80% et donc l'alignement de séquences est possible et facile ;
- 50 à 80% et donc l'alignement de séquences est possible ;

- 25 à 50% et donc l'alignement de séquences est possible mais difficile ;
- <25 à 30% et la question de la faisabilité de l'alignement se pose.

En d'autres termes, plus les résidus entre les deux protéines sont différents, plus la prédiction de la structure sera difficile. C'est pour cela que les méthodes de comparaison de séquences jouent un rôle capital en biologie moléculaire. Nous pouvons constater que de nombreux efforts ont été réalisés pour améliorer la qualité des différentes méthodes d'alignement, et ce surtout lorsque l'alignement de plusieurs séquences de protéines est considéré.

La modélisation par homologie se réalise par la succession d'une série d'étapes.

- La séquence est analysée. Ceci se fera par l'examination de la composition en acides aminés et par un découpage de la protéine en domaines distincts ;
- un maximum de protéines qui lui sont homologues seront recherchées. Il est important de rassembler le plus de séquences possibles. En effet, plus le nombre de protéines intégrées dans l'alignement est important, plus les régions conservées qui sortiront de l'alignement seront fiables. En outre, la présence de protéines dont la structure est connue est indispensable à la réalisation de la modélisation par homologie ;
- les protéines homologues sont ensuite alignées. La qualité des résultats dépend des performances du logiciel d'alignement multiple utilisé ;
- le modèle structural de la protéine est élaboré au moyen de programmes spécialisés identifiant les régions variables et les régions de séquences conservées. Ces programmes peuvent être utilisés soit par des serveurs souvent gratuits, soit sur le site Internet directement ;
- enfin, la méthode d'affinage utilise la mécanique et la dynamique moléculaires ainsi que l'analyse des champs de forces pour étudier le placement des chaînes latérales et permettre la modélisation des régions variables. L'affinage, contrairement aux méthodes décrites ci-dessus, a un temps de calcul élevé et requiert l'installation de logiciels spécifiques.

Les résultats obtenus dépendent de plusieurs critères :

- Le nombre de protéines homologues disponibles ;

- le pourcentage d'identité entre les séquences de la protéine d'intérêt et les protéines homologues sélectionnées ;
- la ressemblance structurale de la protéine d'intérêt et la protéine de référence;
- l'expertise de la personne qui réalise le programme.

# Chapitre 2 : Alignement de séquences protéiques

## 2.1 Principe des alignements de séquences protéiques

### 2.1.1 Alignement pairé

Lorsque nous décidons d'aligner deux séquences choisies de manière aléatoire et si nous considérons que la fréquence des différents acides aminés au sein de ces protéines est identique, nous pouvons affirmer que ces deux séquences juxtaposées, aléatoires et non alignées devraient avoir en moyenne 5% de résidus identiques (C. Vinals, communication personnelle). L'objectif de l'alignement pairé est d'avoir, en décalant les séquences les unes par rapport aux autres, un nombre élevé de résidus identiques qui se correspondent entre ces deux séquences.

Lorsque celles-ci possèdent une proportion importante d'acides aminés ou d'acides nucléiques similaires, l'alignement est plus facile à réaliser et peut presque se faire par inspection visuelle. Par contre, lorsque les différences entre les séquences sont relativement élevées, l'alignement se révèle plus difficile à effectuer.

En réalité, il est difficile d'aligner la totalité de deux segments avec une correspondance exacte entre résidus. Dès lors, l'amélioration de l'alignement s'effectue par l'insertion d'espaces appelés « indels » (insertion/délétion). Ces espaces permettent de compenser des résidus absents dans une séquence et présents dans l'autre séquence comprend. Les indels peuvent être soit longs si l'une des séquences ne correspond qu'à un domaine de l'autre séquence, soit courts si ceux-ci ont pour origine la disparition de quelques résidus au cours de l'évolution.

L'insertion d'indels est l'élément clé de la procédure d'alignement : la localisation des indels et leur longueur est le résultat attendu d'un algorithme d'alignement.

Par conséquent, on considère que deux séquences ayant le même nombre d'acides aminés et un nombre limité d'indels sont homologues lorsqu'il y a au moins 20% de résidus identiques (C. Vinals, communication personnelle). Une analyse plus précise

réalisée dans les séquences homologues a permis de montrer que des acides aminés chimiquement similaires se substituent plus souvent les uns aux autres.

L'alignement de deux séquences peut être réalisé par différentes méthodes. Le dot plot en est une.

***Remarque***

Dans une région protéique importante au niveau fonctionnel, nous constatons que le remplacement d'un seul résidu peut suffire à modifier radicalement les propriétés de la protéine.

Deux séquences dont les résidus sont identiques prennent des conformations identiques (à l'exception du modèle hypothétique du prion).

Des motifs structuraux qui n'ont pas de résidus communs peuvent être presque les mêmes.

### 2.1.2 Alignement multiple

L'alignement simultané de plusieurs séquences d'acides aminés est maintenant un outil essentiel en biologie moléculaire.

Avec l'évolution technologique augmentant considérablement le nombre de séquences disponibles, l'alignement multiple fut élaboré. Celui-ci est utilisé pour caractériser des familles de protéines, pour détecter ou démontrer une homologie entre de nouvelles séquences, pour aider à la prédiction de structures secondaires et tertiaires de nouvelles séquences.

L'alignement multiple se base sur le fait qu'une similarité de séquences est plus hautement significative si elle est partagée par plusieurs séquences (Depiereux and Feytmans, 1992).

L'alignement multiple peut se faire par plusieurs méthodes de type progressif ou simultané. L'alignement de type progressif est un alignement pairé auquel d'autres séquences sont alignées progressivement. Il est évident que son résultat dépend de l'alignement obtenu pour les deux séquences initiales, ce qui est son inconvénient majeur. Différentes méthodes recourent à cet algorithme :

- La méthode de Feng et Doolittle (Feng and Doolittle, 1987) mesure les similarités entre toutes les paires possibles de séquences et les transforme en scores de distance. Plus la distance entre les deux séquences est petite, plus leur similarité est grande. La méthode aligne les deux séquences les plus proches en premier lieu afin d'obtenir un alignement de score minimal. Ensuite, ces deux séquences sont alignées avec celle qui leur est la plus proche. Et ainsi de suite, les séquences sont incorporées à l'alignement dans l'ordre de leur alignement par distance.
- La méthode de Corpet (Corpet, 1988): un tableau $T_1$ de 20X20 collecte les valeurs $T_{1ij}$ qui mesurent la ressemblance entre les séquences i et j. Ensuite, les deux séquences les plus proches sont alignées et forment un groupe qui prend la place de la séquence. Ceci implique donc la formation d'un nouveau tableau $T_2$ de dimension de N-1 x N-1 et qui tient compte du nouveau groupe. L'opération est répétée jusqu'à ce qu'il n'y ait plus qu'un groupe. Dans un deuxième temps, à chaque comparaison de deux séquences ou de deux groupes, une matrice de comparaison est construite, ce qui permet de déterminer l'alignement de score maximal.
- La troisième méthode d'alignement progressif porte le nom de méthode de Murata (Murata *et al.*, 1985). Elle se base sur la comparaison de 3 séquences où l'algorithme se traduit par la construction d'un tableau en trois dimensions ou chaque case $Y_{ijk}$ est l'addition des scores associés aux trois paires de résidus superposés. L'algorithme cherche le tracé optimal qui lui permet d'avoir un score global maximum. Lorsque deux cases successives ne sont pas adjacentes, il faut alors introduire un indel comme dans le cas de l'alignement de deux séquences. L'inconvénient de cette méthode est une durée élevée du temps de calcul.

## 2.2 Description des logiciels d'alignement multiple simultané

### 2.2.1 ClustalW

ClustalW (Thompson *et al.*, 1994) est un logiciel d'alignement multiple entre des séquences protéiques. C'est sans doute l'un des logiciels les plus utilisés dans le monde à cet escient. Il s'agit d'une version mise à jour de ClustalV (ancienne version) mis au point par Higgins *et al.* (Higgins *et al.*, 1992). Cette nouvelle version a amélioré plusieurs points, notamment la sensibilité progressive au cours de l'alignement de séquence (Thompson *et al.*, 1994).

ClustalW est un logiciel qui utilise, pour construire ses alignements, une procédure d'opération dynamique établissant un score global sur les séquences alignées. ClustalW pondère chaque séquence protéique se retrouvant dans un alignement partiel, ce qui lui permet de minimiser le poids des séquences qui sont dupliquées telles quelles. À l'inverse, cette méthode augmente le poids des séquences les plus divergentes. ClustalW offre la possibilité d'utiliser différentes matrices de scores. Il se base sur deux familles de matrices. L'utilisation de l'une ou l'autre famille dépend de la similarité entre les protéines à aligner. ClustalW fonctionne suivant une méthode de « gap penalty ». Ceux-ci pénalise un score qui est proportionnel à la similarité observée entre les acides aminés. Ce score diminue avec l'introduction d'indels et leur élongation. Cependant, des indels de taille variable seront insérés par ClustalW lorsqu'ils parviennent à compenser la pénalité qui leur est attenante par une augmentation de la similarité des résidus qu'il aligne. En conclusion, il nous faut quand même préciser chez ClustalW que l'indel sert de paramètre mais aussi de résultat.

### 2.2.2 BlockMaker

BlockMaker est un logiciel dont le rôle est de constituer des blocs de segments alignés (sans indels) correspondant à des régions fortement conservées de familles de protéines. Lorsqu'une seule séquence est soumise à BlockMaker, il cherche dans sa

base de données des blocs qui peuvent englober un segment de la séquence. Par contre, lorsque l'on soumet plusieurs séquences à ce logiciel, il aligne les séquences entre elles en construisant de nouveaux blocs toujours constitués de régions fortement conservées, entre les séquences.

### 2.2.3 Gibbs

Ce programme, et plus récemment Probe qui en découle (Neuwald *et al.*, 1997), construit ses alignements pour trouver le meilleur score via des méthodes statistiques de probabilité en se basant sur les acides aminés conservés au sein d'une famille. En fait, Probe lance plusieurs fois Gibbs avec un certain nombre de paramètres aléatoires. Gibbs examine différents candidats possibles du meilleur alignement selon une méthode stochastique et fournit la recherche du meilleur alignement, mesuré à partir du taux maximum *a posteriori* de probabilité. Il est important de savoir que puisque c'est une méthode stochastique, il faudra qu'il y ait des variations entre le meilleur alignement trouvé et les candidats obtenus par hasard. Par conséquent, il est souvent utile de lancer l'analyse d'un échantillon plusieurs fois pour voir si cela converge vers le même alignement à chaque essai. Si ce n'est pas le cas, comme par exemple pour l'alignement de motifs très subtils, il faut nécessairement accomplir un grand nombre de recherches indépendantes avec en même temps un nombre suffisant d'échantillons qui ont un optimum proche de celui-ci. Le problème de segments de séquences est contourné par Gibbs en optimisant la longueur finale de l'alignement. Par contre, comme on utilise une probabilité pour chaque résidu sur chaque séquence, il faut une similarité homogène.

### 2.2.4 Match-Box

Match-Box est un logiciel conçu par E. Depiereux et E. Feytmans ayant pour objectif de réaliser des alignements multiples de protéines, dans le but de déterminer des régions qui seraient structurellement ou fonctionnellement conservées (Depiereux and Feytmans, 1991).

Le logiciel fonctionne selon un principe qui relie trois algorithmes successifs, à savoir le *scanning*, le *matching* et le *screening*. Ces trois algorithmes feront l'objet d'une description plus détaillée dans la suite de ce chapitre. Les deux premiers algorithmes de Match-Box, c'est-à-dire le *scanning* et le *matching*, se basent sur une multitude de comparaisons de segments protéiques deux à deux. Le troisième algorithme, à savoir le *screening*, annoncera les résultats.

En fait, au cours des étapes de *scanning* et de *matching*, Match-Box « regarde » au travers d'une fenêtre d'analyse ayant une taille déterminée. Cette taille correspond à celle des segments protéiques comparés deux à deux. Cette fenêtre se positionne au premier acide aminé de la première séquence. Le segment recouvert par cette fenêtre (la fenêtre dite « fixe ») est comparé à tous les segments possibles de toutes les autres protéines à aligner avec la première, comme si une seconde fenêtre (la fenêtre dite « mobile ») balayait systématiquement toutes les autres séquences du début à la fin et faisait l'objet, pour chacune de ses positions, d'une comparaison pairée avec la fenêtre fixe. Une fois toutes les comparaisons effectuées entre la fenêtre mobile et la fenêtre fixe, cette dernière se déplace d'un acide aminé sur la première séquence et le balayage reprend avec la fenêtre mobile. Cette opération se poursuit jusqu'à ce que la fenêtre fixe ait parcouru l'entièreté de la première séquence.
La comparaison de la fenêtre mobile avec la fenêtre fixe s'effectue en comparant deux à deux les résidus occupant la même position dans chacune des deux fenêtres. Un score est attribué à la comparaison de deux acides aminés en fonction de leur similarité. Ces scores de similarité sont repris dans des tableaux que l'on nomme matrices de scores (Cfr point 1.1). Le score attribué à la comparaison des deux fenêtres correspond à la somme des scores obtenus lors de la comparaison de chacun de leurs résidus deux à deux.

La taille de la fenêtre est déterminée par l'utilisateur et reste constante tout au long de la procédure d'alignement. La taille est un paramètre crucial pour permettre le bon déroulement des comparaisons lors des étapes de *scanning* et de *matching*. Si celle-ci est trop petite ou trop grande, la précision et l'efficacité de l'alignement risque d'être

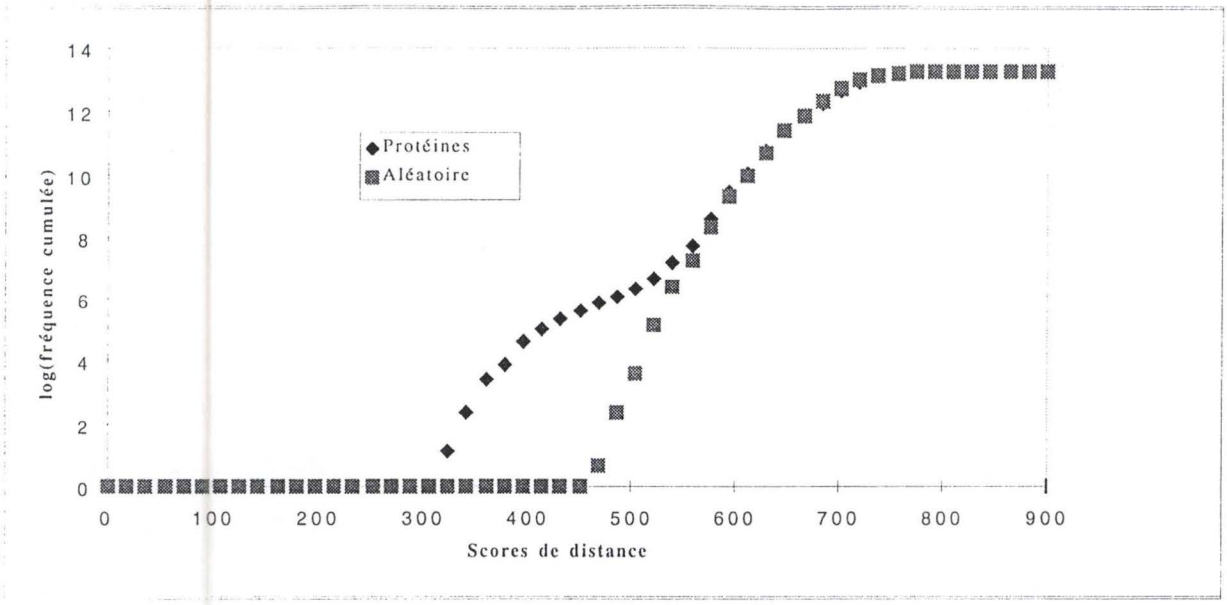



*Figure: 2.2.4.1*

*Résultat obtenu grâce à EXPLORE.*
*La courbe caractérisée par des losanges représente la fréquence d'apparition de tous les scores dans un alignement de deux séquences.*
*La courbe caractérisée par des carrés représente la fréquence d'apparition de tous les scores de l'alignement de ces mêmes séquences mais où les résidus ont été mélangés de façon aléatoire.*
*Si la première courbe est significativement différente de la seconde, on en déduit que l'alignement se distingue de ce que le hasard aurait pu produire.*

affecté. Par exemple, la fenêtre d'analyse doit être suffisamment grande pour pouvoir contenir une Région Structurellement Conservée (SCR : Structurally Conserved Region). Des études ont montré jusqu'à nouvel ordre, que la taille optimale de la fenêtre d'analyse de Match-Box est de 9 résidus (Depiereux *et al.*, 1997).

Une étape préliminaire au fonctionnement des 3 étapes de Match-Box est requise afin de déterminer si les protéines que l'on désire aligner ont un niveau de similarité suffisant pour que l'alignement se démarque des résultats attendus par le hasard. Cette opération est reprise sous le nom d'EXPLORE.
EXPLORE consiste à la comparaison de toutes les séquences protéiques deux à deux. Pour ce faire, le principe des fenêtres mobiles décrit ci avant est à nouveau d'application, avec une fenêtre mobile ne balayant ici qu'une seule séquence puisque les comparaisons sont pairées. La comparaison des fenêtres fixes et mobiles aboutit, de même que décrit précédemment, à l'attribution d'un score. A ce stade, deux analyses sont proposées :

1) Une matrice de similarité entre les séquences prises deux à deux est établie. Les similarités sont calculées en faisant le rapport entre le nombre de fenêtres considérées comme similaires (ayant un score inférieur à un seuil préétabli) et le nombre total de comparaisons possibles entre les fenêtres. Il est en outre possible de produire une représentation graphique de la similarité entre séquences.
2) La distribution de la fréquence d'apparition de tous les scores est comparée à celle qu'on obtient lorsque les résidus des deux séquences sont mélangés (« randomized sequences »). Si la première est significativement différente de la seconde, on en déduit que la similarité entre les deux séquences considérées se distingue de ce que le hasard aurait pu produire. On peut dès lors les soumettre au premier algorithme de Match-Box, à savoir le *scanning*. Dans le cas contraire, l'utilisateur est averti du risque qu'il encourt d'obtenir des résultats qui sont liés au hasard *(figure 2.2.4.1)*.

Le *scanning* *(figure 2.2.4.2)* consiste en un premier balayage des deux premières séquences avec les fenêtres fixe et mobile. Ceci aboutit à la détermination de la

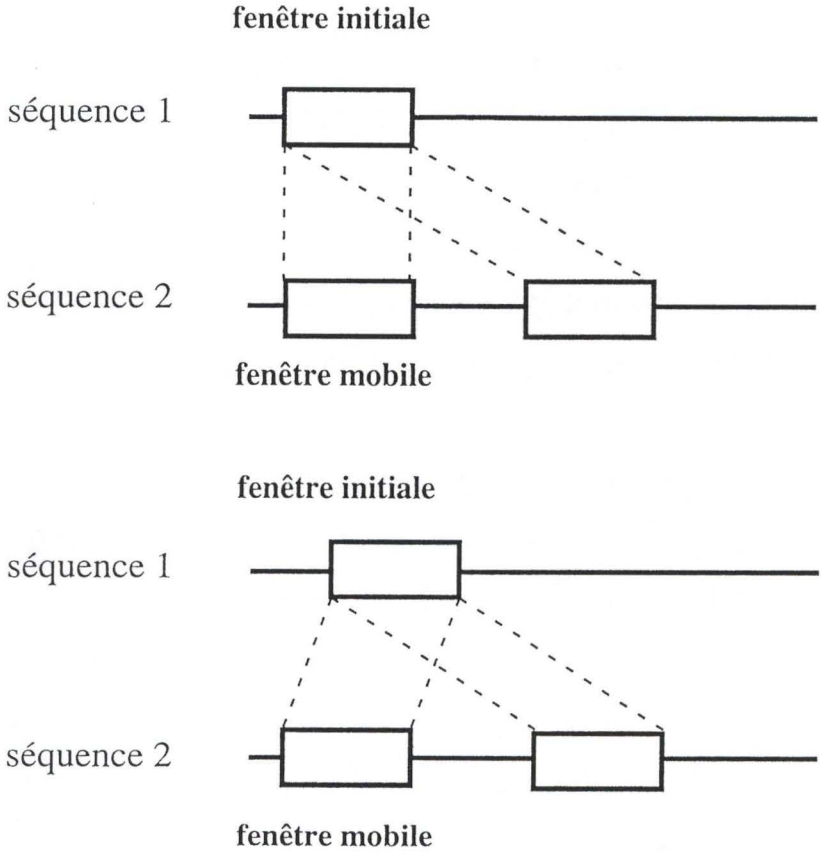



*Figure: 2.2.4.2*

*Représentation schématique des différentes comparaisons entre segments protéiques s'effectuant dans le scanning.*

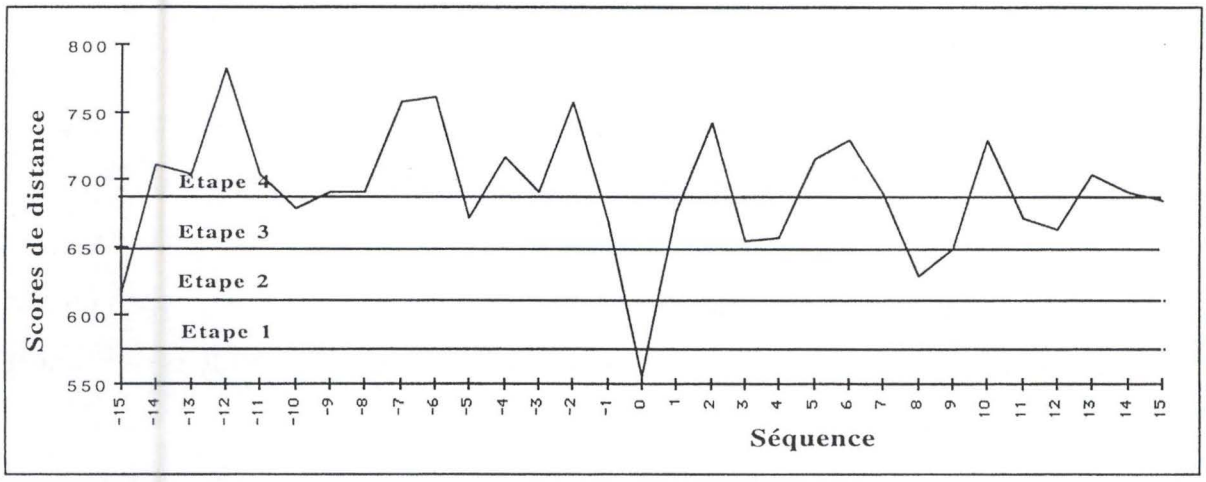



*Figure: 2.2.4.3*

*Représentation de la paramétrisation du système par l'établissement de 4 seuils permettant de distinguer le signal du bruit.*

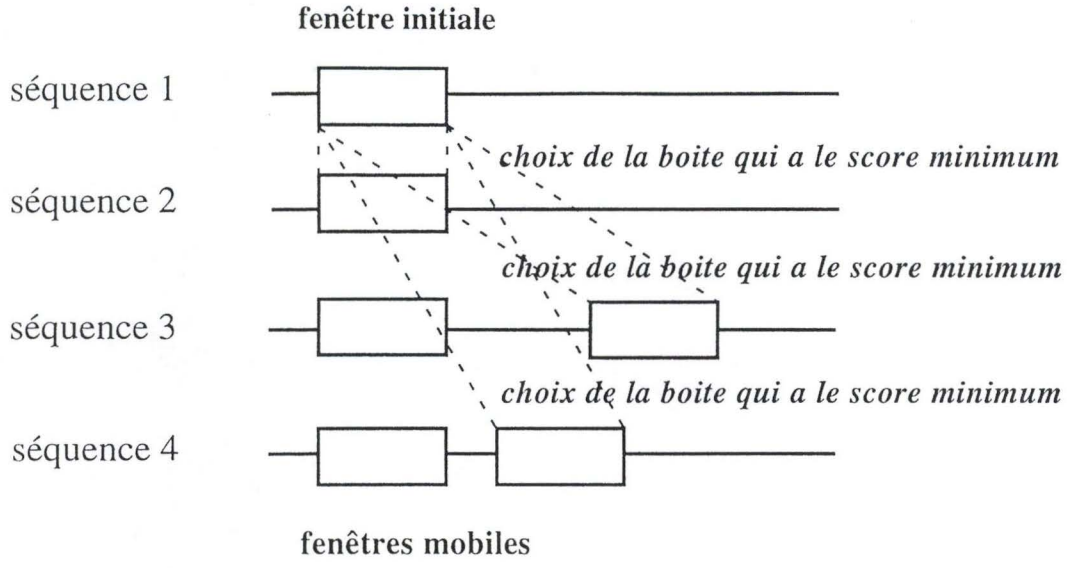



*Figure: 2.2.4.4*

*Représentation schématique des différentes comparaisons entre segments protéiques s'effectuant dans le matching avec détermination du meilleur appariement.*

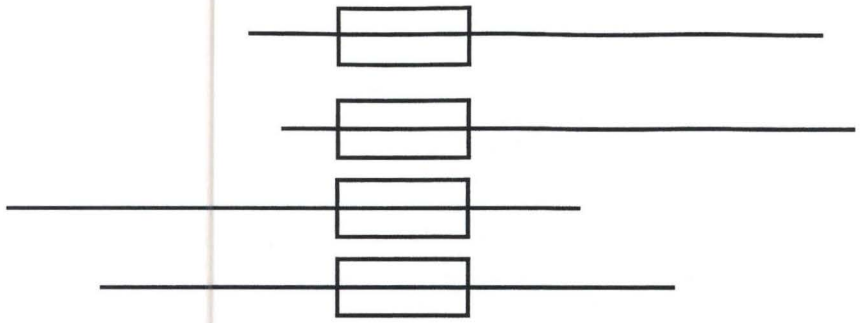


*Figure: 2.2.4.5*

*Représentation d'une boîte potentielle par rassemblement des meilleurs appariements.*

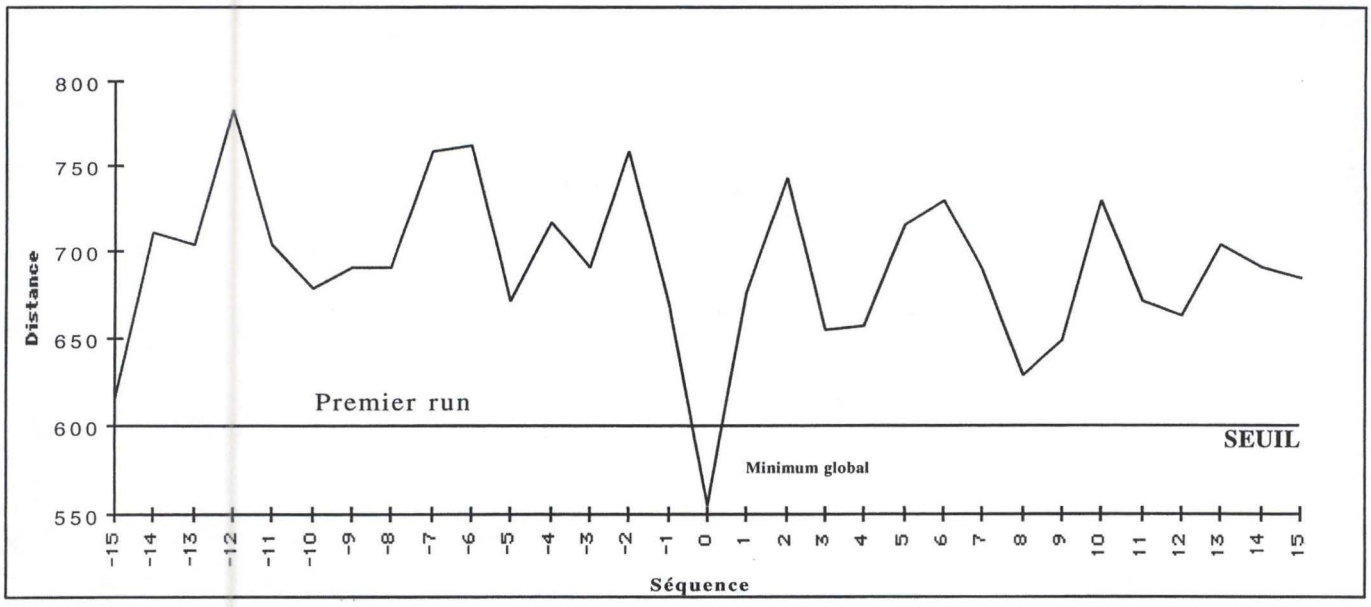



*Figure: 2.2.4.6*

*Schématisation du premier run du matching avec élimination de certaines boîtes potentielles en fonction des seuils établis lors du scanning.*

fréquence d'apparition des scores lors de la comparaison des séquences à aligner. La distribution de ces scores permet *in fine* la paramétrisation du système, *i.e.*, la détermination de quatre valeurs seuils distinguant de façon optimal le signal (score faible correspondant à une similarité significativement élevée) du bruit (score élevé) *(figure 2.2.4.3)*. Ces valeurs seuils seront utilisées dans l'étape suivante, celle du *matching*.

Dans le *matching*, la fenêtre initiale se trouve toujours sur la première séquence, et la fenêtre mobile se déplace successivement sur toutes les autres séquences à aligner. Au niveau de l'exécution de l'algorithme, les fenêtres mobiles de chaque séquence sont comparées les unes à la suite des autres à la fenêtre initiale *(figure 2.2.4.4)*. Le résultat de ces comparaisons donne, pour chaque séquence, la fenêtre caractérisée par le score minimum. C'est ce qu'on appelle le meilleur appariement. L'ensemble des segments sélectionnés sur l'ensemble des séquences comparées suggère l'existence d'une boîte potentielle *(figure 2.2.4.5)*. Cependant, le seul lien entre les segments sélectionnés est d'être identique à la fenêtre initiale. Il reste donc à tester la similarité entre eux en utilisant les seuils statistiques établis lors du *scanning (figure 2.2.4.6)*. Les alignements complets ayant passés ce crible pourront être soumis au troisième algorithme, le *screening*.

Le *screening* est la dernière étape algorithmique de Match-Box. Son premier rôle est de constituer une banque de boîtes à partir du travail d'appariement complet exécuté par le *scanning* et le *matching*. Cette banque est constituée de segments corrects et de segments incorrects. Une fois que cette banque est disponible, le deuxième rôle du *screening* est de trier les segments corrects parmi les segments incorrects. Parmi les boîtes ainsi obtenues, la plus grande d'entre elles est considérée comme étant *a priori* la meilleure. L'algorithme cherche ensuite toutes les boîtes compatibles avec celle-ci. Les boîtes compatibles sont des boîtes où l'alignement de l'une n'empêche pas l'alignement de l'autre. Par opposition, les boîtes incompatibles sont des boîtes où l'alignement avec une portion quelconque d'une autre boîte est impossible *(figure 2.2.4.7)*. Parmi toutes les boîtes compatibles trouvées, le *screening* choisit la plus

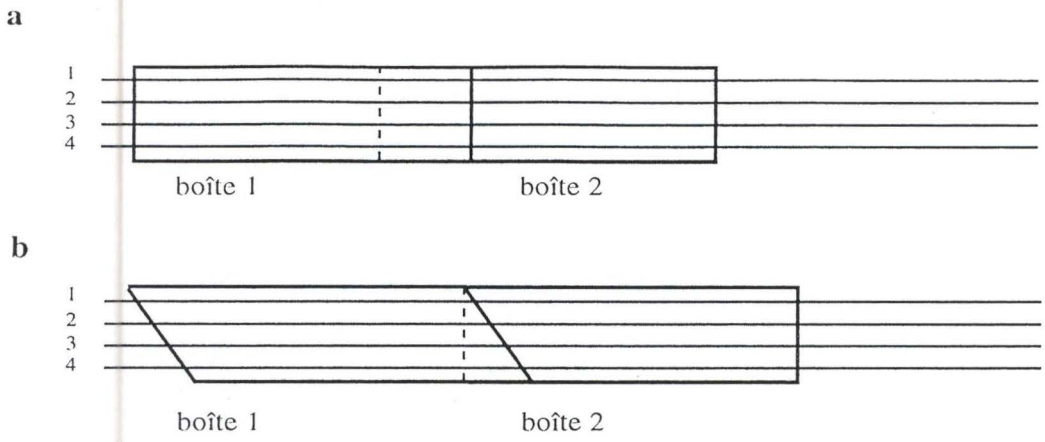



*Figure: 2.2.4.7*

*a) Représentation de deux boîtes compatibles.*
*b) Représentation de deux boîtes incompatibles.*

*Représentation d'une banque constituée de boîtes correctes (signal) et de boîtes incorrectes (bruit). La plus grande boite est* a priori *la plus fiable.*

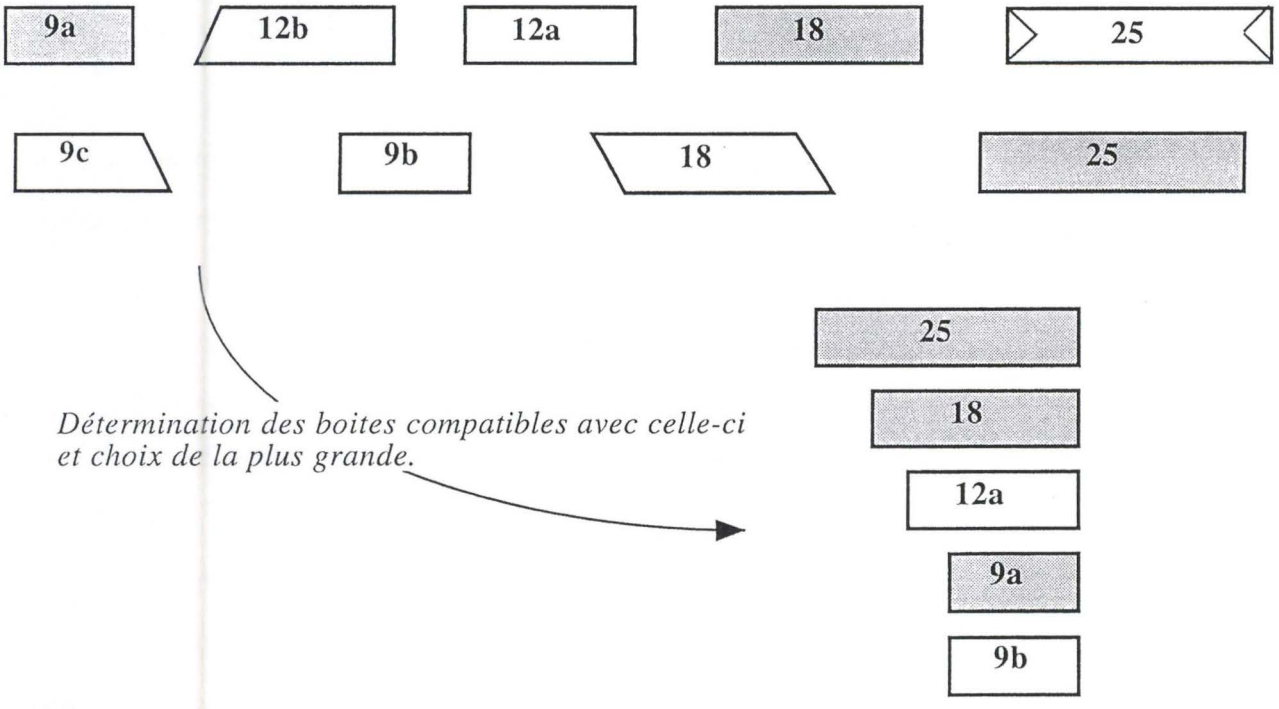



*Figure: 2.2.4.8*

*Représentation schématique du screening.*
*Les boîtes grisées représentent les boîtes correctes, les boîtes blanches les boîtes incorrectes.*
*La forme de la boîte symbolise un décalage dans la banque de boîtes.*

grande. Par la suite, de nouvelles boîtes compatibles avec cette dernière seront déterminées, la plus grande sera choisie et ainsi de suite *(figure 2.2.4.8)*. En d'autres termes, le *screening* permet l'agencement final des boîtes dans un alignement. Dès lors, sa fonction est de conditionner entièrement la qualité du résultat final et du degré d'automatisation de la méthode.

## 2.3 Comparaison de Match-Box avec les autres logiciels d'alignement multiple (Briffeuil *et al.*, 1998)

### 2.3.1 ClustalW *vs* Match-Box (Depiereux *et al.*, 1997)

Avantages de ClustalW envers Match-Box :

- ClustalW est rapide et automatique, avec la possibilité d'utiliser une procédure interactive ;
- Il reconnaît de façon automatique plusieurs formats proposés par les banques de données ;
- Il peut aboutir à la création d'un arbre phylogénétique servant au classement des protéines à aligner ;
- le choix du format de sortie des résultats est réalisable ;
- il permet l'alignement d'un nombre maximal de 300 séquences ayant une taille pouvant atteindre 5000 acides aminés alors que Match-Box est limité à 50 séquences d'une longueur maximale de 2000 résidus.

Avantages de Match-Box envers ClustalW :

- Match-Box propose un décalage entre séquences en se basant sur des zones d'homologies découvertes entre les séquences, et non en se basant sur les notions de pénalités établies a priori ;
- il délimite des boîtes ou zones qui indiquent le niveau d'homologie entre séquences.

Désavantages de ClustalW envers Match-Box :

- ClustalW pénalise les décalages qui apparaissent dans l'alignement alors que ces décalages ne peuvent être justifiés que sur une base biochimique et structurale ;
- ClustalW ne donne aucune information sur le degré d'homologie entre les séquences, ni même sur la confiance de l'alignement proposé.

### 2.3.2 BlockMaker *vs* Match-Box

Avantage de BlockMaker contre Match-Box :

- Pour des alignements de séquences de faible homologie, BlockMaker propose des points d'ancrages corrects.

### 2.3.3 Gibbs *vs* Match-Box.

Avantages de Gibbs :

- Gibbs offre une bonne stratégie car il utilise des probabilités par colonne et par blocs ;
- il se base sur une distribution pour chaque acide aminé pour chaque position dans les blocs ;
- vu la méthode utilisée, on optimise la longueur de l'alignement.

Désavantage de Gibbs :

- Etant donné que Gibbs se base sur une probabilité d'apparition de chaque acide aminé à tel ou tel endroit, il faut une similarité homogène entre les différentes séquences considérées.

## 2.4 Dilemme du concept confiance-puissance (Depiereux *et al.*, 1997)

Lorsque tout logiciel d'alignement multiple de séquences protéiques réalise son alignement, celui-ci sera évalué sur base de la confiance et de la puissance que l'on

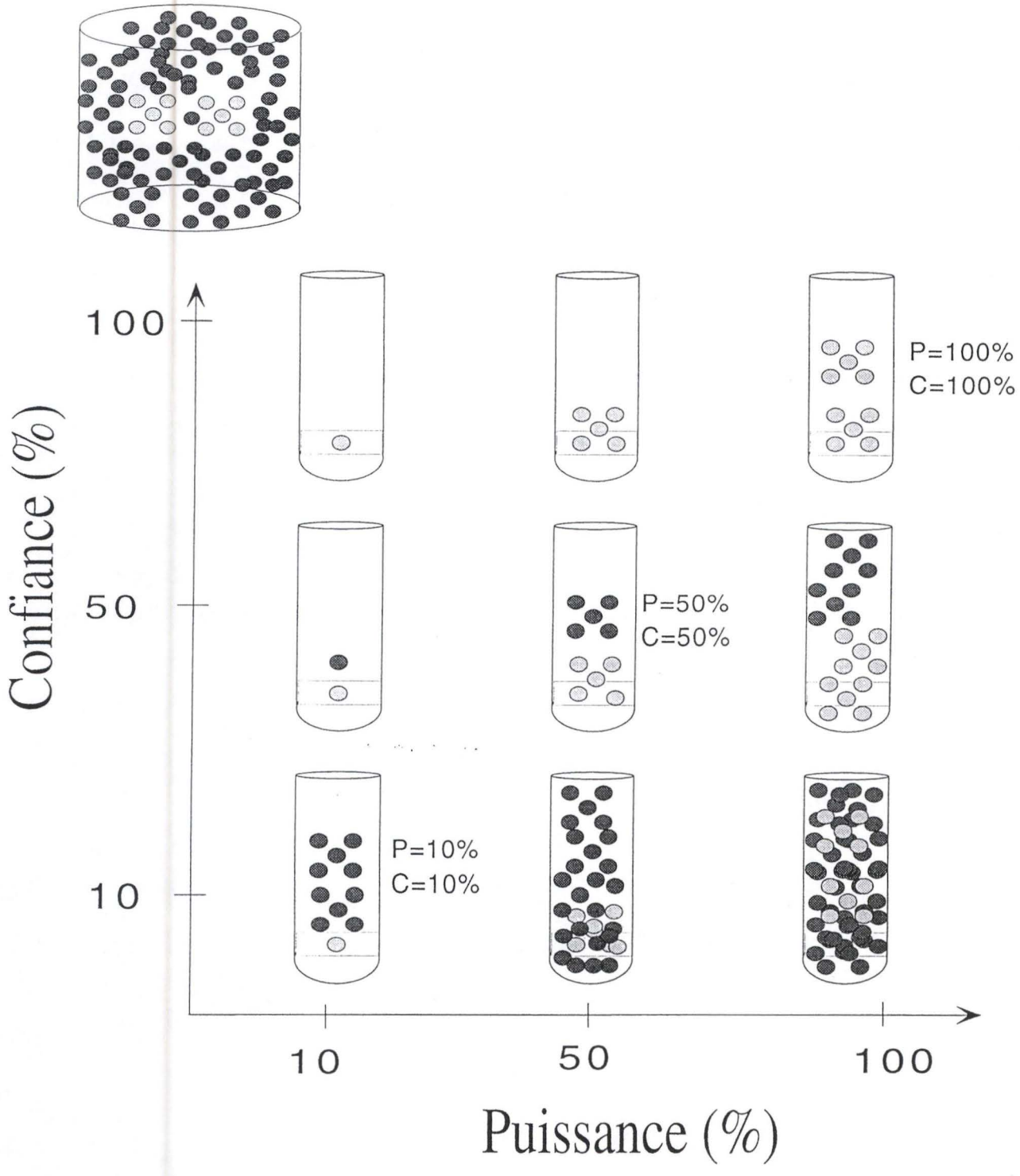



*Figure: 2.4.1*

*Diagramme représentant les notions de confiance et de puissance.*
*Considérons un cylindre contenant un ensemble de billes claires et foncées ; le but de l'expérience étant de les séparer et de ne garder que les claires. Différentes tentatives indépendantes ont été réalisées et les résultats ont été placés dans plusieurs éprouvettes. Nous considérons la puissance comme étant le rapport entre le nombre de billes claires trouvées dans le cylindre et le nombre total de billes claires qui auraient du être trouvées (10) dans ce même cylindre. La confiance, quant à elle, représente le rapport entre le nombre de billes claires trouvées dans le cylindre et le nombre total de billes (claires et foncées) prises dans ce même cylindre.*

peut accorder à cet alignement. Comme ces deux termes ne sont pas fortement explicites, nous allons les définir et expliquer leurs relations. La confiance représente le rapport entre le nombre de résidus correctement alignés dans les séquences et le nombre total de résidus alignés dans les séquences par la méthode. La puissance quant à elle représente le rapport entre le nombre de résidus correctement alignés dans la séquence et le nombre total de résidus qui auraient du être alignés dans ces mêmes séquences. Nous pouvons considérer la puissance comme étant la sensibilité du programme d'alignement et la confiance comme étant sa sélectivité. Dans l'état actuel des choses, la confiance et la puissance agissent un peu comme des vases communicants. En effet, il apparaît que toute augmentation de l'un s'accompagne d'une diminution de l'autre et *vice-versa (figure 2.4.1)*.

# MATÉRIEL

# 1 Les ordinateurs

La réalisation de ce mémoire a nécessité l'utilisation de différentes variétés d'ordinateurs.

Une bonne partie des recherches a été effectuée sur des stations de travail Silicon Graphics Octane duo et INDIGO2 fonctionnant sur les systèmes d'exploitation IRIX 6.5 et IRIX 5.3 respectivement. Ces ordinateurs fournissent aux utilisateurs un environnement agréable, convivial et facile pour la réalisation de différentes tâches. La majorité de nos programmes a été créée et testée sur ces stations de travail étant donné que notre logiciel d'utilisation, à savoir Match-Box, tournait sur ces deux machines.

Voici la fiche technique de ces ordinateurs :

**1) Silicon Graphics Octane duo.**

- 2 processeurs R10000 Mips cadencés à 225MHz
- 512 mégas octets de mémoire RAM
- 2 cartes graphiques ESI
- Un disque dur de 2X9 GB et 1X4 GB

**2) Silicon Graphics INDIGO2**

- 1 processeur R4400 Mips cadencé à 150MHz
- 64 mégas octets de mémoire RAM
- 1 carte graphique ESI
- Un disque dur de 1X4 GB et 1X2 GB

Un autre type d'ordinateur fut également indispensable pour la réalisation de ce travail. Nous avons utilisé des ordinateurs de marque MacIntosh fonctionnant sous le système d'exploitation Mac-OS 7.5 pour traiter facilement nos résultats. Ceux-ci ont été en grande partie synthétisés en tableaux sur le logiciel Microsoft Excel 98.

## 2 Langage de programmation

Tous les programmes écrits pour la réalisation de nos tests ont été écrits en Fortran 77 (Mazet, 1996). Le terme FORTRAN a pour origine le " IBM Mathematical Formula Translation System ". Il a été initialement conçu pour simplifier la programmation de calculs numériques sur les plateformes IBM 704. La première version du FORTRAN n'est apparue qu'au début de l'année 1957, et même si les programmes obtenus à partir du code FORTRAN étaient plus lents que ceux obtenus à partir des codes en langage machine, le FORTRAN s'est imposé auprès de la communauté scientifique : il était bien plus simple à écrire. Très rapidement, il a été possible de réutiliser des codes FORTRAN sur d'autres plateformes que celles d'IBM.

## 3 Internet

Il s'agit d'un système de télécommunication qui a été inventé pendant la Deuxième Guerre Mondiale, pour les besoins de l'armée américaine.

Internet est un réseau de réseaux informatiques couvrant toute la planète. On n'y retrouve aucune autorité centrale ni de serveur prédominant. Internet n'appartient à personne ou, plus précisément, il appartient à tout le monde.

Internet a recours au protocole TCP/IP (Transmission Control Protocol/Internet Protocol). Il s'agit d'un protocole mis au point à l'origine pour les ordinateurs tournant sous le système d'exploitation UNIX et qui leur permettait de communiquer entre eux sur de longues distances. Ce concept est au cœur d'Internet.

La colonne vertébrale d'Internet est constituée de mini-ordinateurs, tournant sous UNIX en général, reliés par des câbles en fibre optique offrant un meilleur débit lors du transfert de données. Des centaines d'autres réseaux, réseaux locaux et même des ordinateurs isolés peuvent s'y connecter. Enfin, il y a des points d'accès pour des ordinateurs utilisant d'autres protocoles que TCP/IP grâce à des convertisseurs ou

passerelles. Le résultat est un métaréseau unique ayant approximativement 150 millions d'utilisateurs, sans doute plus. Il est en effet impossible de donner une valeur précise du nombre d'utilisateurs d'Internet dans la mesure où de grands réseaux reliés à Internet sont comptés comme un simple nœud. Le succès d'Internet est dû sans doute à son protocole ouvert, contrairement aux protocoles " maison " des fabricants d'ordinateurs, repris par des dizaines d'industriels informatiques et par les éditeurs de logiciels. Les services que peut offrir Internet sont innombrables. Par exemple, le World Wide Web (WWW) est un des services les plus populaires d'Internet. Il est basé sur les techniques d'hypertexte. On recherche à l'aide de mots issus d'un index puis on " navigue " dans les documents grâce aux liens hypertexte. Le résultat d'une recherche est un document contenant des liens vers d'autres documents.

Internet permet aussi :

- La consultation de banques de données ou de bibliothèques.
- Une utilisation de la messagerie électronique. C'est certainement la fonction la plus utilisée grâce au protocole SMTP (Simple Mail Transfer Protocol). C'est un moyen simple et économique de communication entre personnes.
- Les conférences électroniques. Sous le terme de conférences électroniques, on regroupe les différents services permettant une communication interactive entre les utilisateurs. C'est une extension de la messagerie traditionnelle. Elles offrent un espace de communication basé sur des centres d'intérêts communs par l'échange de messages, la diffusion de lettres spécialisées, de répertoires etc.

http://urfist.univ-lyon1.fr/internet.html

http://www.microtec.net/fr/FAQ/howtoxs.html#top

# 4 Match-Box (Depiereux and Feytmans, 1991)

Comme décrit dans l'introduction, Match-Box est une méthode d'alignement de séquences protéiques conçue par E. Depiereux et E. Feytmans permettant de réaliser des alignements multiples de séquences ayant pour but de délimiter des régions qui

seraient structurellement ou fonctionnellement conservées. Ce serveur est disponible à l'adresse Internet suivante :
http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html

Lors de la réalisation de ce travail, nous avons essentiellement testé et amélioré la troisième étape de Match-Box à savoir, le *screening*.

## 5 PDB (Sussman *et al.*, 1998)

Une partie des données que nous utilisons pour la réalisation de ce travail a été puisée de PDB (Protein Data Bank) du Brookhaven National Laboratory (Cambridge USA). PDB est une sorte de bibliothèque contenant les structures tridimensionnelles connues de macromolécules biologiques, chacune étant répertoriée par un nom de code. Cette bibliothèque contient également les coordonnées atomiques, les citations bibliographiques ainsi que des informations sur les structures primaires et secondaires, comme par exemple les facteurs responsables de la structure cristallographique.
Cette bibliothèque est disponible à l'adresse Internet suivante :
http://www2.ebi.ac.uk/pdb/index.shtml

## 6 DSSP (Kabsch and Sander, 1983)

Le rôle de ce programme est de fournir la structure secondaire des protéines à partir de l'ensemble de leurs coordonnées tridimensionnelles. Ce programme ne prédit pas la structure tertiaire de la protéine, mais en fournit la structure secondaire, ses caractéristiques géométriques ainsi que l'exposition au solvant de la protéine.
Ce programme est disponible à l'adresse Internet suivante :
http://www.sander.embl-heidelberg.de/dssp/

## 7 PHD (Rost and Sander, 1993)

PHD est un serveur dédié à la recherche de données à propos de la protéine d'intérêt. Lorsqu'on envoie une séquence d'acide aminé à PredictProtein@EMBL-Heidelberg.DE, PHD nous renvoie différentes informations comme par exemple une prédiction de la structure secondaire (PHDsec) qui se réalise par une méthode de réseaux neuronaux ayant une exactitude de 72% sur la prédiction d'hélices, de brins, et de boucles (Rost and Sander, 1993). PHDsec procède en trois étapes successives. Son but initial est d'améliorer l'exactitude des prédictions en utilisant des alignements multiples de séquences protéiques. Ensuite, il procède à l'amélioration de la prédiction des brins. Enfin il tente d'améliorer la prédiction des segments de structures secondaires en utilisant des systèmes de niveaux multiples, impliquant la connaissance des informations décrites ci-après :

- Une prédiction de l'accessibilité au solvant des résidus de la protéine d'intérêt (PHDacc).
- Une prédiction de la tendance qu'a la protéine d'intérêt à se globulariser (GLOBE).
- Une prédiction sur la présence éventuelle d'hélices transmembranaires (PHDhtm), de régions éventuellement possédant des coiled-coils (COILS) et la reconnaissance de domaines (TOPITS).

Ce serveur est disponible à l'adresse Internet suivante :
http://www.embl-heidelberg.de/Services/sander/predictprotein/

## 8 Les cas-tests

L'alignement multiple de séquences protéiques est maintenant reconnu comme un outil indispensable dans l'étude de la biologie moléculaire. Il joue un rôle important

pour la prédiction de régions fonctionnelles ou structurelles partagées par une famille de protéines.

Les cas-tests sont des outils très utiles et très importants pour permettre une évaluation de l'efficacité des méthodes de prédictions. Ce sont en fait des familles de protéines qui constituent un échantillonnage relativement représentatif des divers problèmes rencontrés lors d'une procédure d'alignement de séquences. Ces problèmes s'étalent sur une échelle allant des cas les plus faciles et accessibles par n'importe quel programme (c'est-à-dire plus de 35% d'acides aminés qui se correspondent lors d'un alignement pairé) aux cas-tests beaucoup plus difficiles se composant d'un ensemble de séquences très peu similaires (moins de 25%) mais possédant quand même certaines régions structurellement conservées (C. Vinals, communication personnelle).

Toutes les familles sont composées de protéines dont la structure tridimensionnelle a été déterminée expérimentalement au préalable. Les similarités structurales entre ces protéines sont mises en évidence lors de l'alignement de structure de ces mêmes protéines. Tous les acides aminés qui occupent une position structurale égale dans l'alignement de structure sont alignés en colonnes. Par conséquent, les fragments de protéines structurellement conservés au sein d'une famille sont consignés dans des boîtes.

L'évaluation de l'efficacité d'une méthode d'alignement de séquences, résulte dans la capacité de cette dernière à retrouver des motifs structuraux conservés dans toutes les protéines appartenant à une même famille mais en ne tenant compte que de l'information contenue dans la séquence en acides aminés. Par conséquent, un programme d'alignement de séquence sera d'autant performant s'il aligne un maximum de résidus correspondant aux zones structurellement conservées (la puissance) et s'il ne prédit pas de correspondance structurale non observée au sein d'une même famille (la confiance).

Nous disposions au départ d'une batterie de 20 cas tests déterminée dans la littérature par P. Briffeuil (Briffeuil *et al.*, 1998). Cette batterie fut ensuite élargie à 33 cas-tests. Les 13 cas-tests supplémentaires proviennent eux aussi de la littérature et en voici les noms, leur famille respective, ainsi que leur référence :

- APROT2L: Les deux lobes des Aspartic proteinases (Blundell *et al.*, 1991)
- GLO: Globines (Bashford *et al.*, 1987)
- HTH: Motif Helix-Turn-Helix (Lawrence *et al.*, 1993)
- IMMUC: Immunoglobulines, Constant Domain (Cohen *et al.*, 1981)
- IMMUV: Immunoglobulines, domaine variable (Cohen *et al.*, 1981)
- LIPOC: Lipocalines (Flower *et al.*, 1993)
- PKINASEST: Protein Serine-Threonine Kinases (Hanks *et al.*, 1988)
- PKINASET: Protein Tyrosyine Kinases (Hanks *et al.*, 1988)
- RICINB: Domaines de la Ricine B (Rutenber *et al.*, 1987)
- SPROTMB: Sérines protéinases de bactéries et de mammifères (Ding *et al.*, 1994)
- VCOATPP: Virus Coat Protein of plants (Carrington *et al.*, 1987)
- VCOATPPR: Virus Coat Protein of Plants and Rhinovirus (Arnold and Rossmann, 1990)
- VCOATPR: Virus Coat Protein of Rhinovirus (Arnold and Rossmann, 1990)

## 9 SCOP (Brenner *et al.*, 1996)

La structure d'une protéine est en général une information utile pour la prédiction des fonctions qu'elle assure dans la cellule. L'information que l'on peut tirer de la structure peut soit être propre à la protéine d'intérêt, soit concerner d'autres protéines partageant la même histoire.

L'obtention de ces informations nécessite une bonne connaissance de la structure de la protéine concernée ainsi que des relations qu'elle établit avec ses partenaires. Ces deux aspects *in vivo* ne sont pas indépendants. En effet, la compréhension structurale d'une seule protéine requiert une connaissance générale des interactions que celle-ci établit dans la cellule. Inversément, la compréhension des relations des protéines avec leurs partenaires est renforcée par une information détaillée quant à la structure de toutes ces protéines. Heureusement, ce problème qui peut paraître complexe n'est

pas insurmontable et cela pour deux raisons. La première est que la compréhension des structures protéiques est facile car il n'y a qu'un nombre limité d'éléments de structures secondaires, et les arrangements de ces éléments sont restreints par les différentes lois de la physique et probablement aussi par celles de l'évolution. La deuxième est que les ressources aidant à la reconnaissance des relations entre structures protéiques sont maintenant disponibles. Ces informations sont disponibles dans SCOP. SCOP est un catalogue de conformations structurales contenant toutes les entrées PDB disponibles à ce jour. SCOP a été conçu pour faciliter la recherche détaillée de familles de protéines particulières et pour une navigation fructueuse et aisée dans toute la base de données.

Cette banque de données est disponible à l'adresse Internet suivante :

http://scop.mrc-lmb.cam.ac.uk/scop/

# RÉSULTATS

# Buts du travail

Notre travail a pour objectif une amélioration conjointe de la confiance et de la puissance du logiciel d'alignement multiple de séquences Match-Box au niveau des résultats de son alignement.

A cette fin, nous avons tenté deux approches différentes :
La première stratégie consiste en une combinaison de deux logiciels d'alignement multiple de séquences protéiques, à savoir ClustalW et Match-Box, le premier étant caractérisé par une puissance élevée aux dépens de la confiance, qui est trop faible, le second se caractérisant par une confiance très élevée, mais une puissance relativement faible. Cette opération, consistant à combiner les deux logiciels, a pour objectif d'améliorer la qualité de l'alignement en profitant de la puissance de l'un (ClustalW) tout en conservant la confiance de l'autre (Match-Box).
La deuxième approche s'intéresse plus précisément en une amélioration du troisième et dernier algorithme de Match-Box, c'est-à-dire l'étape de screening. Pour réaliser cette tâche, nous tenterons d'une part d'élaborer un nouvel algorithme de screening et d'autre part de tenir compte des données de structures secondaires dans l'étape de screening.

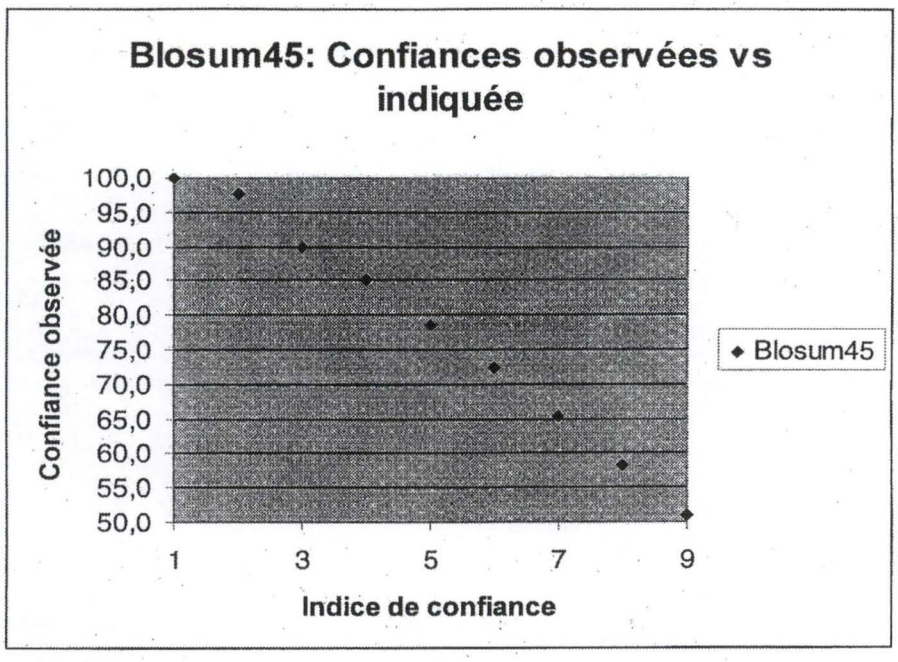
# Chapitre 1 : Match-Tal

## 1.1 Description de la méthode

La première technique que nous avons réalisée et testée se nomme Match-Tal. Cette technique est un logiciel d'alignement hybride combinant les procédures de Match-Box et ClustalW. Nous savons d'une part que le logiciel Match-Box dispose d'une confiance particulièrement élevée dans ce qu'il aligne, mais ceci toutefois aux dépens de sa puissance, qui est plus faible que celle de ClustalW. Lorsque nous analysons ClustalW, nous remarquons que la tendance est inversée : ClustalW fournit une confiance dans son alignement plus faible que Match-Box, mais sa puissance est plus grande. Match-Tal a comme objectif d'améliorer l'alignement en tentant de combiner la puissance élevée de ClustalW tout en gardant la confiance de Match-Box.

Avant de décrire le fonctionnement de Match-Tal, il est nécessaire de définir la notion de coefficient de confiance du logiciel Match-Box. Match-Box attribue à chaque colonne alignée un coefficient se rapportant au taux de faux positifs observés sur les différents cas-tests. Ce taux de faux positifs est inversément proportionnel à la confiance observée sur ces mêmes cas-tests. La valeur du coefficient de confiance de Match-Box s'étend sur une échelle allant de un (la meilleure valeur, correspondant à la confiance la plus élevée) à neuf (la moins bonne valeur, correspondant à la confiance la plus faible).

- Proportionnalité entre la confiance observée sur les cas-test et le coefficient de confiance donné par Match-Box




Ce graphique représente la confiance moyenne observée sur les 33 cas-tests avec la matrice de score Blosum45 en fonction de l'indice de confiance. Sur celui-ci, nous pouvons remarquer que plus l'indice de confiance augmente, plus la confiance observée est faible.

Match-Tal va chercher, parmi toutes les colonnes alignées par Match-Box, celles dont le coefficient de confiance est plus petit ou égal à une valeur fixée par l'utilisateur. Ensuite, Match-Tal fait aligner les résidus restants non alignés par ClustalW.

**Explication** ***(voir figures 1.1.1 a,b,c,d,e)*** **:**

1) L'alignement des séquences par Match-Box fournit un certain nombre de boîtes renfermant chacune un nombre déterminé de colonnes alignées. Chaque colonne

*Figure 1.1.1*

*Figure 1.1.1a : Alignement de Match-Box.*

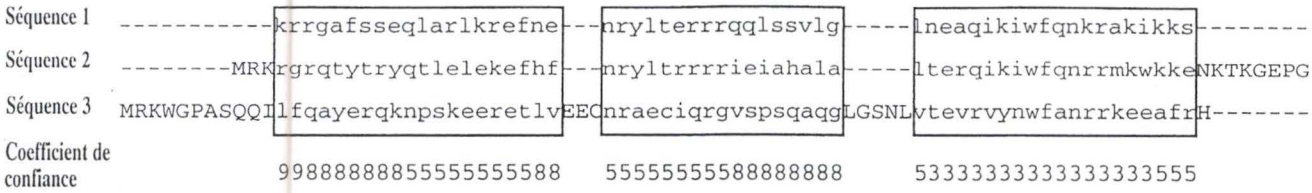



*Figure 1.1.1b : Choix du coefficient de confiance ≤5 pour Match-Tal.*

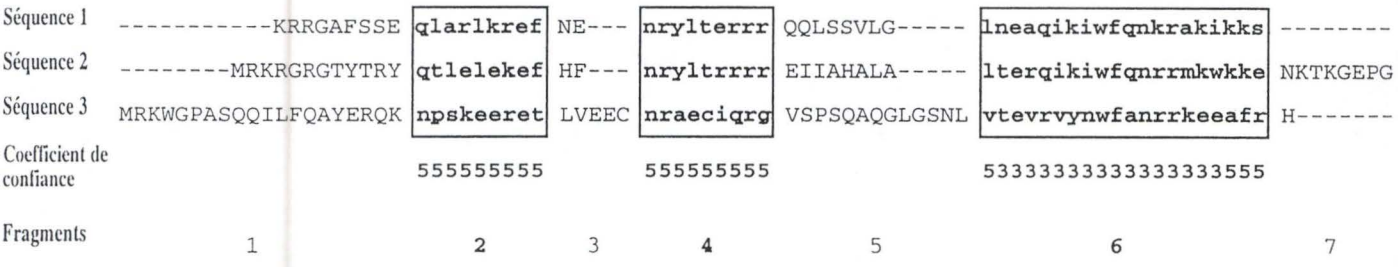



*Figure 1.1.1c : On soumet à ClustalW les régions non alignées encadrées des régions alignées.*

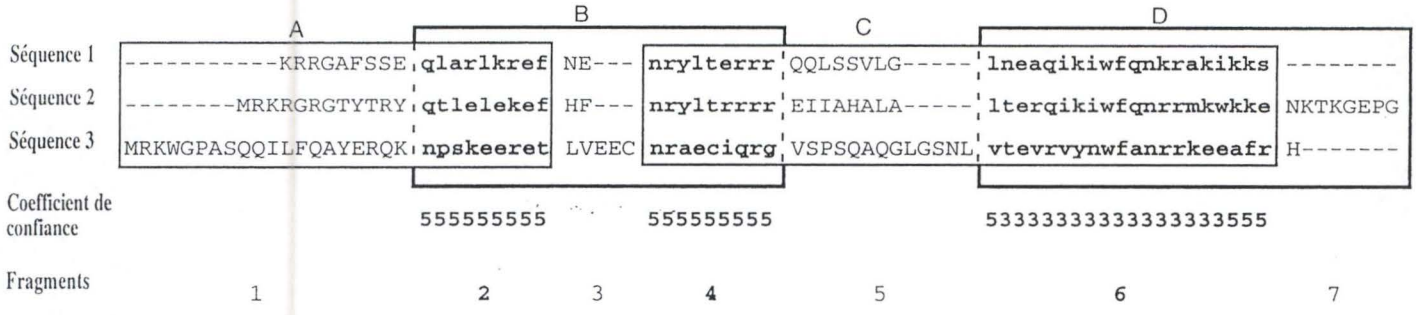



*Figure 1.1.1d : L'encadrement est renforcé par le remplacement des acides aminés par les acides aminés les plus fréquents.*

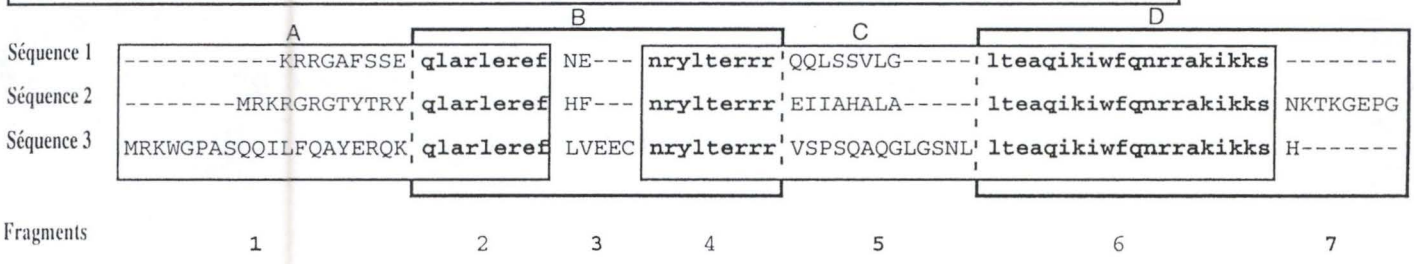
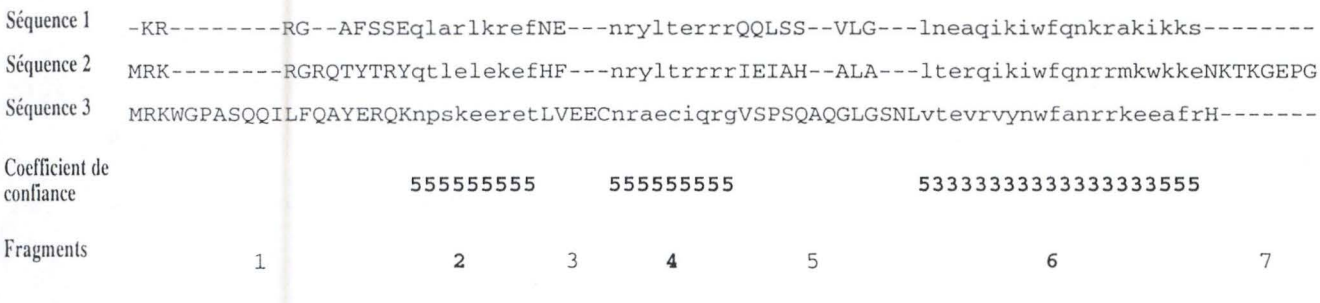



*Figure 1.1.1e: Résultat final.*



alignée dans les boîtes est caractérisée par un coefficient de confiance *(figure 1.1.1a)*. Les acides aminés qui ne sont pas dans les boîtes ne sont pas alignés.

2) Match-Tal sélectionne, dans l'alignement réalisé par Match-Box, les colonnes ayant un coefficient de confiance inférieur ou égal à celui donné par l'utilisateur (≤ 5, dans l'exemple). Les colonnes sélectionnées forment de nouvelles boîtes, possédant un nombre de colonnes inférieur ou égal à celles trouvées par Match-Box. Dans notre exemple *(figure 1.1.1b)*, on peut distinguer les nouvelles boîtes (c'est-à-dire les fragments 2, 4 et 6) du reste des protéines (fragments 1, 3, 5 et 7).

3) Les parties non sélectionnées (fragments 1, 3, 5 et 7) sont ensuite soumises par Match-Tal à ClustalW. Pour que le travail d'alignement de ClustalW sur les fragments de séquence bordant les boîtes isolées par Match-Box et sélectionnées par Match-Tal soit optimal, nous avons remarqué qu'il était indispensable de les accompagner de leurs cadres droit et gauche (cadres a,b,c,d). Ces cadres correspondent aux boîtes choisies par Match-Tal *(figure 1.1.1c, cadre A pour le fragment 1, B pour le 3, C pour le 5, D pour le 7)*. Il est évident que ClustalW ne doit pas modifier l'alignement des séquences sur lesquelles Match-Box a déjà travaillé. Ceci fut facilement résolu en remplaçant les résidus de chaque colonne alignée par le résidu y apparaissant le plus fréquemment *(figure 1.1.1d)*. Dès lors, puisque les fragments de séquence à aligner par ClustalW sont bordés par des séquences où les résidus sont conservés à 100%, ClustalW respectera l'alignement.

4) Lorsque le travail de ClustalW est terminé, l'alignement total est reconstitué par réinsertion, dans les boîtes « falsifiées », du résultat trouvé par Match-Box lors de la première étape de Match-Tal *(figure 1.1.1e)*.

Nous avons testé cette nouvelle méthode en lui soumettant la batterie de 33 cas-tests et en changeant à chaque soumission la matrice de score utilisée par la méthode. Ces différentes matrices sont Blosum 45, 62 et 80 (qui se basent sur des alignements venant de BlockMaker) (Henikoff and Henikoff, 1992; Henikoff and Henikoff, 1993), les matrices de Johnson 92 et 96 (qui se basent sur l'alignement de structures)

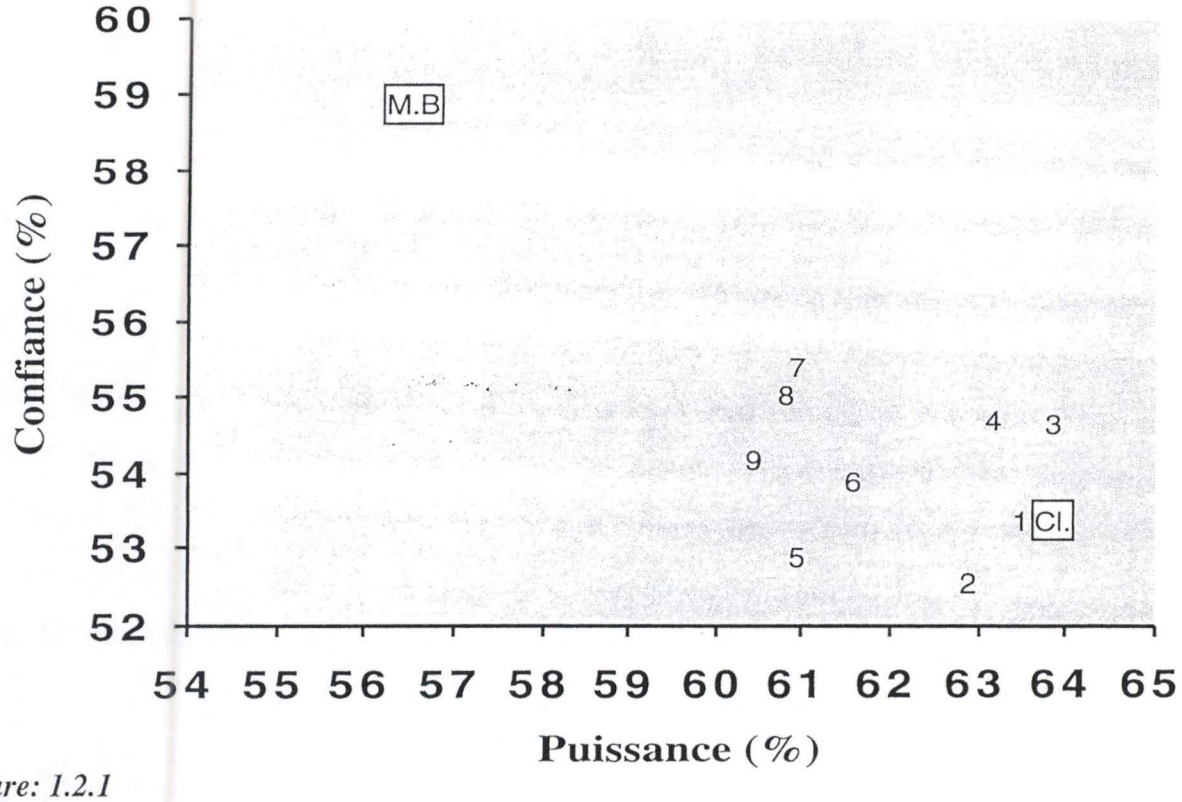



*Figure: 1.2.1*

*Représentation graphique des résultats se rapportant aux trois méthodes sur les 33 cas-tests pour la matrice de score Pam 120. Les résultats de Match-Box et ClustalW sont encadrés, les chiffres de 1 à 9 représentent les différents résultats obtenus par Match-Tal en fonction du seuil de coefficient de confiance choisi.*

***Tableau 1.2.1:*** *Résultats de Match-Tal en confiance et en puissance sur 33 cas-tests, diverses matrices de scores et des seuils de coefficient de confiance allant de 1 à 9. Moy M-T correspond à la moyenne des résultats de Match-Tal.*

*tableau 1.2.1a*

| Johnson 92 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,6 | 53,2 |
| 2 | 62,8 | 52,7 |
| 3 | 62,7 | 53,6 |
| 4 | 63 | 54,3 |
| 5 | 63,3 | 54,5 |
| 6 | 61,1 | 55,4 |
| 7 | 59,9 | 54,8 |
| 8 | 59,9 | 54,7 |
| 9 | 59,4 | 53,6 |
| Match-Box | 57,3 | 57,7 |
| moy M-T | 61,7 | 54,1 |

*tableau 1.2.1b*

| Johnson 96 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,6 | 53,2 |
| 2 | 62,8 | 52,9 |
| 3 | 61,8 | 52,9 |
| 4 | 60,7 | 52,9 |
| 5 | 61,3 | 53,7 |
| 6 | 61,6 | 53,6 |
| 7 | 58,2 | 52,2 |
| 8 | 58 | 52,5 |
| 9 | 57,5 | 51,3 |
| Match-Box | 55,2 | 54,6 |
| moy M-T | 60,6 | 52,8 |

*tableau 1.2.1c*

| Blosum 45 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,5 | 53,2 |
| 2 | 63,5 | 53,2 |
| 3 | 60,4 | 50,6 |
| 4 | 61,5 | 53,1 |
| 5 | 57,2 | 49,3 |
| 6 | 57,3 | 49,4 |
| 7 | 56,5 | 49,9 |
| 8 | 56,4 | 50,2 |
| 9 | 56 | 50,3 |
| Match-Box | 53,6 | 53,8 |
| moy M-T | 59,1 | 51,0 |

*tableau 1.2.1d*

| Blosum 62 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,3 | 53 |
| 2 | 61,3 | 51,8 |
| 3 | 62,2 | 53,2 |
| 4 | 64 | 54,7 |
| 5 | 62,2 | 53,8 |
| 6 | 62,2 | 54,3 |
| 7 | 59,3 | 53,2 |
| 8 | 58,9 | 52,5 |
| 9 | 58,3 | 52,2 |
| Match-Box | 54,7 | 56,4 |
| moy M-T | 61,3 | 53,2 |

***Tableau 1.2.2:*** *Résultats de Match-Tal en confiance et en puissance sur 33 cas-tests, diverses matrices de scores et des seuils de coefficient de confiance allant de 1 à 9. Moy M-T correspond à la moyenne des résultats de Match-Tal.*

*tableau 1.2.2a*

| Blosum 80 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,6 | 53,3 |
| 2 | 61,4 | 51,9 |
| 3 | 62,6 | 53,4 |
| 4 | 63,3 | 54,2 |
| 5 | 62,4 | 54,3 |
| 6 | 62,7 | 54,5 |
| 7 | 60 | 54,4 |
| 8 | 60,4 | 54,9 |
| 9 | 58,9 | 52,5 |
| Match-Box | 58,1 | 54,9 |
| moy M-T | 61,7 | 53,7 |

*tableau 1.2.2b*

| Pam120 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,6 | 53,3 |
| 2 | 62,8 | 52,7 |
| 3 | 63,8 | 54,6 |
| 4 | 63,1 | 54,7 |
| 5 | 60,9 | 52,9 |
| 6 | 61,5 | 53,9 |
| 7 | 60,8 | 55,2 |
| 8 | 60,8 | 55,1 |
| 9 | 60,3 | 54,2 |
| Match-Box | 56,5 | 58,9 |
| moy M-T | 62,0 | 54,1 |

*tableau 1.2.2c*

| Pam200 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,7 | 53,3 |
| 2 | 61,8 | 52,4 |
| 3 | 62,6 | 53,9 |
| 4 | 60 | 52,4 |
| 5 | 59,3 | 51,8 |
| 6 | 59,2 | 51,8 |
| 7 | 58,6 | 52,6 |
| 8 | 58,6 | 53,1 |
| 9 | 58,3 | 52,5 |
| Match-Box | 56,3 | 56,6 |
| moy M-T | 60,2 | 52,6 |

*tableau 1.2.2d*

| Pam250 | | |
|---|---|---|
| | Puissance (%) | Confiance (%) |
| ClustalW | 63,8 | 53,3 |
| 1 | 63,6 | 53,2 |
| 2 | 63,5 | 53,2 |
| 3 | 63,9 | 53,6 |
| 4 | 61,5 | 52,7 |
| 5 | 58,2 | 51,4 |
| 6 | 55,6 | 50 |
| 7 | 55,6 | 50,7 |
| 8 | 55 | 49,8 |
| 9 | 55 | 49,4 |
| Match-Box | 53,6 | 51,4 |
| moy M-T | 59,1 | 51,6 |

(Johnson and Overington, 1993) et les matrices de Pam 120, 200, 250 (qui se basent sur les mutations) (Dayhoff M. O., 1972).

## 1.2 Résultats

La soumission de notre batterie de 33 cas-tests à Match-Tal et ce pour les différentes matrices de scores utilisées et en utilisant les différentes valeurs de seuil de coefficient de confiance ne fournit pas les résultats espérés. En effet, non seulement nous ne pouvons profiter de la puissance de ClustalW, mais de plus nous observons une diminution de la confiance par rapport à Match-Box *(tableaux 1.2.1 et 1.2.2)*.
Pour plus de facilité, étant donné que les tests ont été réalisés sur huit matrices de scores différentes, nous avons choisi d'analyser en détail les résultats obtenus pour une seule matrice de scores. Nous prenons comme exemple la matrice Pam 120 *(tableau 1.2.2b)*. Dans les résultats obtenus avec cette matrice, nous remarquons que Match-Box obtient une puissance de 56,5% et une confiance de 58,9%. ClustalW, quant à lui, obtient une puissance de 63,8% pour une confiance de 53,3%. Nous pouvons aussi dans cet exemple rappeler les points forts de chacune des méthodes à savoir la confiance chez Match-Box et la puissance chez ClustalW. L'analyse des résultats de Match-Tal pour cette matrice ne montre jamais une puissance supérieure à celle de ClustalW mais surtout, les résultats obtenus en confiance sont inférieurs à ceux obtenus par Match-Box, et ce quel que soit le coefficient de confiance seuil choisi par l'utilisateur pour Match-Tal. La *figure 1.2.1* montre très bien ces résultats.
Les résultats réalisés avec les autres matrices de scores sont similaires à ceux obtenus avec la matrice Pam 120. Dès lors, les conclusions que l'on peut en tirer sont donc similaires. Ce phénomène s'explique par le fait que ClustalW est une méthode d'alignement de séquences qui se base sur la technique de « gap penalty » et qui choisit une matrice de substitutions d'acides aminés en fonction de la similarité entre séquences. Et justement dans notre cas, les fragments de séquences soumis à

*Tableau 1.4.1 : Représentation des meilleures méthodes à utiliser par cas-tests.*

| Cas test: | Choix: | Puissance (%) | Confiance (%) |
|---|---|---|---|
| ace.test | Match-Box | 65,7 | 79,2 |
| adk.test | Match-Box | 51,5 | 36 |
| amg.test | Match-Box | 51,1 | 50,4 |
| aprot2l.test | Match-Box | 17,3 | 32,1 |
| aprotease.test | Clustalw | 79,4 | 76,2 |
| cys.test | Match-Box | 28,4 | 31,8 |
| cytc.test | Match-Box | 83,7 | 73,5 |
| fabp.test | Match-Box | 52,2 | 60,6 |
| flav.test | Clustalw | 75,2 | 59,4 |
| glo.test | Clustalw | 93,3 | 91,8 |
| hip.test | Match-Box | 88,4 | 64,4 |
| hth.test | Clustalw | 0 | 0 |
| immuc.test | Match-Box | 76,6 | 57 |
| immuv.test | Clustalw | 93,5 | 71,3 |
| lipoc.test | Match-Tal | 58 | 52,7 |
| ltn.test | Match-Box | 97,8 | 81,9 |
| lyzlac.test | Clustalw | 99,1 | 91,8 |
| maldh.test | Match-Box | 76 | 72 |
| peroxydase.test | Match-Tal | 84,4 | 62,9 |
| phycoc.test | Clustalw | 94,4 | 94,4 |
| pkinasest.test | Match-Tal | 50,2 | 63,8 |
| pkinaset.test | Match-Tal | 90 | 97,5 |
| plasto.test | Match-Box | 77,4 | 68,6 |
| ricinb.test | Clustalw | 80 | 71,4 |
| serbact.test | Clustalw | 88,4 | 77,8 |
| sh3.test | Match-Tal | 97,9 | 82,1 |
| sprotm.test | Match-Box | 77,7 | 84,6 |
| sprotmb.test | Match-Box | 43,8 | 66,7 |
| subtilisin.test | Match-Box | 85,6 | 78,8 |
| super.test | Match-Tal | 99,3 | 97,3 |
| vcoatpp.test | Match-Box | 59,1 | 66,7 |
| vcoatppr.test | Clustalw | 0 | 0 |
| vcoatpr.test | Match-Box | 11,1 | 11,5 |
| | **moyenne** | **67,5** | **63,8** |

ClustalW sont fort similaires. En effet, il y a remplacement des acides aminés originaux par des acides aminés identiques.

## 1.3 Conclusions

Les résultats obtenus avec cette technique montre qu'il n'est pas possible d'améliorer les performances de l'alignement en suivant cette voie.

## 1.4 Perspectives

Suite à ces résultats peu concluants, dans le sens où l'alignement ne fut pas amélioré, une étude de faisabilité fut réalisée. Celle-ci visait à déterminer *a priori*, la meilleure méthode à utiliser (Match-Box, ClustalW ou Match-Tal) pour chacun des cas-tests. Les résultats obtenus dans cette étude de faisabilité sont assez prometteurs. En effet, les résultats moyens de confiance et de puissance sur les 33 cas-tests sont plus élevés que les précédents *( tableau 1.4.1)*. Ceux-ci montrent une confiance de 63,8% et une puissance de 67,5%. Une comparaison en moyenne avec la confiance de Match-Box utilisant la matrice de score Pam 120 sur les 33 cas-tests et la puissance de ClustalW sur ces mêmes cas-tests est intéressante puisque nous dépassons largement ces deux valeurs.

| | **Puissance (%)** | **Confiance (%)** |
|---|---|---|
| **Match-Box** | 56,5 | 58,9 |
| **ClustalW** | 63,8 | 53,3 |
| **Etude de faisabilité** | **67,5** | **63,8** |

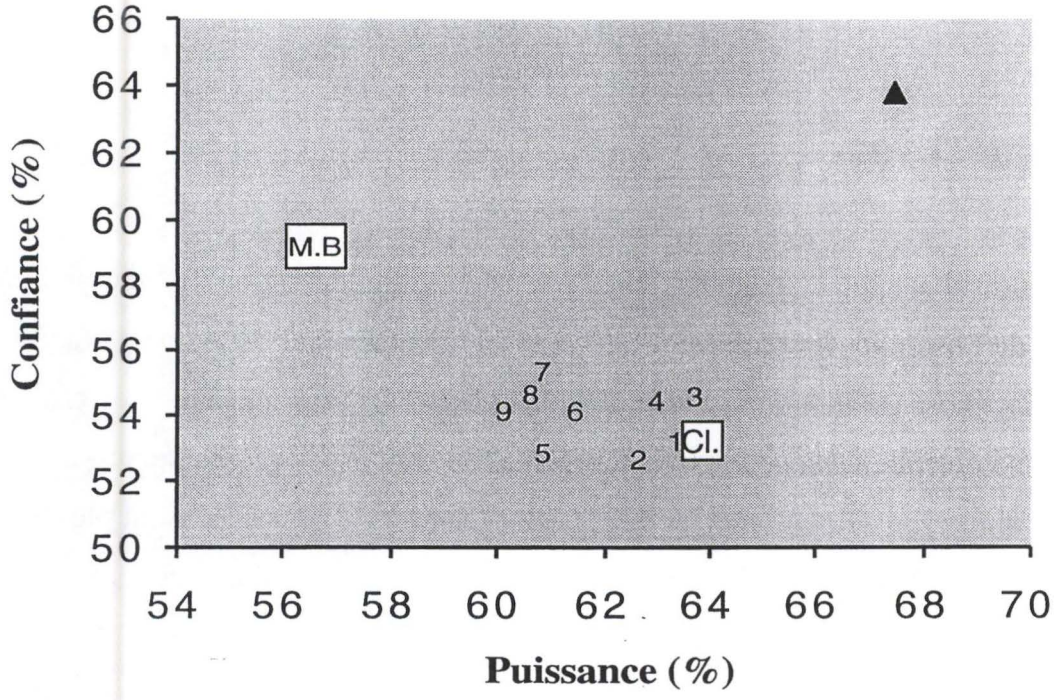



***Figure: 1.4.1***

*Représentation graphique des résultats se rapportant aux trois méthodes sur les 33 cas-tests pour la matrice de score Pam 120 et lorsque l'on choisit la meilleure méthode. Le résultat est représenté par le petit triangle noir. Les résultats de Match-Box et ClustalW sont encadrés, les chiffres de 1 à 9 représentent les différents résultats obtenus par Match-Tal en fonction du seuil de coefficient de confiance choisi.*

Le tableau ci-dessus est parlant car il montre bien un gain de 4,9% en confiance par comparaison avec celle de Match-Box et un gain de 3,7% en puissance par comparaison avec celle de ClustalW. De plus, il faut quand même signaler qu'une augmentation en puissance et en confiance est observée simultanément. Ceci permet de préciser que, pour la première fois, ces deux paramètres ne jouent plus le rôle de « vases communicants » *(figure 2.4.1)*, mais peuvent apparemment évoluer de façon indépendante.

La *figure 1.4.1* ajoute aux résultats obtenus avec la matrice Pam 120 un point (confiance-puissance) correspondant à l'amélioration maximale que nous pouvons espérer en sélectionnant *a priori* pour chaque cas-test la meilleure méthode correspondant à une matrice choisie. Le développement d'une technique permettant cette sélection n'a pu être réalisé durant ce mémoire, mais reste cependant une perspective très prometteuse.

# Chapitre 2 : Développement d'une banque de 78 cas-tests

## 2.1 Description de la méthode

Afin d'améliorer la qualité de nos tests et par conséquent la qualité de nos résultats, nous avons décidé d'utiliser une nouvelle batterie de familles de protéines. Celle-ci couvre 78 cas-tests totalement différents de la batterie de 33 cas-tests.

Cette batterie de 78 cas-tests provient d'une batterie de structures 3D décrite dans la littérature (Overington *et al.*, 1992).

Voici en résumé le principe utilisé pour le développement de cette batterie de 78 cas-tests : 96 familles de structures tridimensionnelles (c'est-à-dire au total 443 structures) déterminées expérimentalement ont été soumises par J. Overington à un alignement de structure (Overington *et al.*, 1992). Ces alignements ont été effectués par superposition des structures tridimensionnelles des protéines en utilisant une technique de programmation dynamique (Johnson *et al.*, 1996). Cette technique fait référence à deux programmes informatiques, à savoir MYNFIT (Sutcliffe *et al.*, 1987) et COMPARER (Johnson *et al.*, 1996) (Sali and Blundell, 1990). En outre, les résidus de chaque position structurellement alignée se trouvant dans toutes les structures d'une même famille furent classifiés par le programme informatique JOY (Overington *et al.*, 1992). Ce programme a permis de fournir les informations suivantes :

1) Il donne le type d'acides aminés dont le remplacement peut être toléré dans les positions alignées au niveau de l'alignement de structure.

***Tableau 2.1.1:*** *Représentation des différentes familles de protéines avec leurs cas-tests respectifs.*

| **Les différentes familles de protéines** | **les cas-tests des familles** |
|---|---|
| alpha beta-hydrolase | ace |
| annexin | annexin |
| antibacterial protein | mycin |
| aspartic proteinase | asp |
| azurin/plastocyanin | az |
| calcium-binding protein | cbp |
| calcium-binding protein | parv |
| crystallin | cryst |
| Cu/Zn superoxide dismutase | sodcu |
| cysteine proteinase | cys |
| cytochrome p450 | p450 |
| cytochrome-c | cytc |
| cytochrome-c3 | cyt3 |
| cytochrome-c5 | cyt5 |
| cytokine | cyto |
| dihydrofolate reductase | dhfr |
| disulphide oxidoreductase | grs |
| DNA-binding homeodomain | hom |
| DNA-binding repressor | rep |
| EGF-like domain | egf |
| Fe/Mn superoxide dismutase | sodfe |
| ferredoxin (2Fe-2S) | fer2 |
| ferredoxin (4Fe-4S) | fer4 |
| flavodoxin | flav |
| globin | glob |
| glutathione S-transferase | gluts |
| glyceraldehyde phosphate dehydrogenase | gpdh |
| glycosyl hydrolase family 13 | aa |
| glycosyl hydrolase family 22 | lys |
| glycosyl hydrolase family 34 | neur |
| high potential iron protein | hip |
| histidine carrier protein | hpr |
| histocompatibility antigen-binding domain | hla |
| immunoglobulin domain | igC1 |
| immunoglobulin domain | igcon |
| immunoglobulin domain | igvar_h |
| immunoglobulin domain | igvar_l |
| immunoglobulin domain | igV |
| insulin | ins |
| interleukin | intb |
| interleukin 8-like protein | il8 |
| kringle domain | kringle |
| lactate/malate dehydrogenase | ldh |
| lipocalin | lipo |
| matrix metalloproteinase | mmp |
| metallothionein | mthina |
| metallothionein | mthinb |
| nucleotide diphosphate kinase | ndk |
| nucleotide kinase | adk |
| pancreatic ribonuclease | rnasemam |
| phospholipase A2 | phoslip |
| picornavirus coat proteins | rhv |
| plant lectin | ltn |
| retroviral proteinase | rvp |
| ribonuclease | rnasebact |
| ribonuclease H | rnh |
| ribulose-1,5-biphosphate carboxylase/oxygenase | rubisco |
| ricin-like protein | ricin |
| rubredoxin | rub |
| sea anemone toxin | seatoxin |

| | |
|---|---|
| serine proteinase | serbact |
| serine proteinase | sermam |
| serine proteinase inhibitor | kunitz |
| serine proteinase inhibitor | serpin |
| serine proteinase inhibitor | squash |
| serine proteinase inhibitor | bowman |
| serine proteinase inhibitor | kazal |
| SH2 domain | sh2 |
| SH3 domain | sh3 |
| snake toxin | toxin |
| subtilase | subt |
| Sugar-binding-protein | sugbp |
| thioredoxin | thiored |
| thymidylate synthase | tms |
| triose phosphate isomerase | tim |
| xylose isomerase | xia |
| zinc finger | zf_CCHH |
| zinc metalloproteinase | tln |

2) Il donne la structure secondaire des résidus d'après les informations reçues de la part de DSSP. Ce dernier assigne à chaque acide aminé sa position dans des hélices α, des feuillets β, des structures irrégulières, ou met en évidence des résidus ayant un angle ϕ positif (Kabsch and Sander, 1983).

3) Il renseigne sur l'accessibilité des résidus au solvant (Lee and Richards, 1991). Pour cette caractéristique, un seuil de 7% est utilisé pour pouvoir faire une distinction entre les résidus où le noyau est inaccessible au solvant (valeur α inférieure ou égale à 7%) et les résidus où l'accessibilité au solvant est élevée (supérieure à 7%).

4) Il renseigne sur l'aptitude des résidus à participer à une liaison par pont Hydrogène. Cette aptitude est basée sur la distance entre les atomes donneurs et receveurs d'Hydrogène. Celle-ci doit être d'au moins 3,5 Å pour permettre la formation d'un pont Hydrogène.

Dans cette batterie de 96 familles de protéines, nous avons sélectionné les cas-tests qui possédaient au moins 3 séquences (alignement multiple), ce qui a conduit à réduire le nombre de cas-tests dans notre batterie à 78. Comme décrit plus haut, ces 78 cas-tests correspondent à des familles de protéines. L'utilisation du catalogue de conformation SCOP disponible sur Internet nous a permis de déterminer à quelles familles appartenaient ces 78 cas-tests *(tableau 2.1.1)*.
Maintenant que nous savons d'où proviennent ces 78 familles et comment elles ont été élaborées, il nous faut spécifier ce que nous considérerons par la suite comme étant correctement aligné dans l'alignement structural tridimensionnel de cette batterie de 78 familles de protéines. Par extension de langage, nous avons décidé d'appeler « vérité » ce qui est correctement aligné dans l'alignement de structure.
Cette vérité est idéalement définie par des zones dans l'alignement de structure où la distance RMS (Root Mean Square) entre les structures est faible (Briffeuil *et al.*, 1998). La distance RMS est la mesure de distance de référence entre deux structures données. Elle correspond à la distance euclidienne moyenne minimale entre les atomes des deux squelettes protéiques impliqués dans la comparaison.

Séquence : N°1 .....**MRIILLGAPGAGKGTQAQFIMEKY..GDMLRAAVKSGSELGKQAQDIMDAGKLVTD**

Séquence : N°2 .....R**LLRAIMGAPGSGKGTVSSRITKHF..GDLLRDNMLRGTEIGVLAKTFIDQGKLIPD**

Séquence : N°3 MEEKLKKS**KIIFVVGGPGSGKGTQCEKIVQKY..GDLLRAEVSSGSARGKMLSEIMEKGQLVPL**

Séquence : N°4 .....S**RPIVISGPSGTGKSTLLKKLFAEY**PD**TPRAGEVNGKDYNFVSVDEFKSMIKNNEFI**

***Figure 2.1.1 :*** *Représentation en gras de ce qui est aligné sans gap.*

Séquence : N°1 .....MRIILLGAPGAGKGTQAQFIMEKY..GDMLRAAVKSGSELGKQAQDIMDAGKLVTD
.....LE**EEEE**LLLLLL**HHHHHHHHHHHH**H..HHHHHHHHHHLLLLLL**HHHHHH**HHLLLLLH

Séquence : N°2 .....RLLRAIMGAPGSGKGTVSSRITKHF..GDLLRDNMLRGTEIGVLAKTFIDQGKLIPD
.....LLE**EEEE**LLLLLL**HHHHHHHHHHHH**L..HHHHHHHHLLLLHHHH**HHHHHH**HLLLLLLH

Séquence : N°3 MEEKLKKSKIIFVVGGPGSGKGTQCEKIVQKY..GDLLRAEVSSGSARGKMLSEIMEKGQLVPL
LHHHHHLLLE**EEEE**LLLLLL**HHHHHHHHHHHH**L..HHHHHHHHHHLLHHHH**HHHHHH**LLLLLLLH

Séquence : N°4 .....SRPIVISGPSGTGKSTLLKKLFAEYPDTPRAGEVNGKDYNFVSVDEFKSMIKNNEFI
.....LLL**EEEE**LLLLLL**HHHHHHHHHHHH**LLLLLLLLLLLLLLEEELL**HHHHHH**HHLLLEE

***Figure 2.1.2 :*** *Représentation en gras des zones où la conservation de la structure secondaire est de 100%.*

Si l'on regarde d'un point de vue quantitatif, la distance RMS moyenne entre deux fenêtres de W résidus est notée :

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

Avec :

- a : le nombre d'atomes considérés par résidu.
- $X_{ij}$, $Y_{ij}$, $Z_{ij}$, : les coordonnées cartésiennes de l'atome i dans la fenêtre j (j=1,2).

Puisque le nombre d'atomes considérés pour la comparaison n'est pas fixe, nous limitons généralement la comparaison à celle du squelette protéique en y incluant l'oxygène du carbonyle. Malgré son utilité indiscutable, la mesure du RMS ne tient pas compte de la structure tridimensionnelle prise par les fragments : elle n'exprime que la proximité physique entre les atomes. Il faut également définir un seuil RMS en dessous duquel deux segments comparés sont suffisamment proches spatialement pour être appariés. D'après Unger, le seuil de distance RMS à considérer pour différencier les segments structurellement similaires de ceux qui ne le sont pas est de 1 Å.

La détermination de la vérité par RMS est de loin la meilleure méthode, mais malheureusement nous n'avons les résultats de cette analyse que pour notre batterie de 20 cas-tests et une partie seulement pour celle des 33 cas-tests. L'obtention de l'analyse des valeurs de RMS pour la batterie de 78 familles de protéines serait fastidieuse, prendrait trop de temps et n'entre par conséquent pas dans le cadre de ce mémoire. Ceci nous oblige donc à déterminer la vérité d'une autre manière :

- Soit en prenant dans l'alignement de structure tout ce qui est aligné sans gap comme étant de la « vérité » *(figure 2.1.1)*. Dans ce cas, nous considérons tout ce qu'il nous est possible de prendre et dès lors, nous surestimons ce qui serait considéré comme vérité par la mesure du RMS.
- Soit en prenant dans l'alignement de structure les endroits où la structure secondaire est conservée à 100% *(figure 2.1.2)*. Ici, nous sous-estimons ce qui

serait considéré comme vérité par la mesure du RMS, car ce procédé est nettement plus stringent.

Nous espérons que ces deux façons de procéder fourniront des conclusions similaires à celles qui seraient obtenues pour la détermination par RMS.

## 2.2 Conclusion

La mise au point de ce nouvel ensemble de cas-tests nous permettra d'obtenir une meilleure distinction entre les méthodes d'alignement de séquences qui alignent beaucoup de résidus (c'est-à-dire puissantes, mais dont la confiance est plus faible) et les méthodes d'alignements qui alignent moins de résidus (moins puissantes) mais qui ont une confiance plus élevée dans ce qu'elles prédisent.

# Chapitre 3 : Amélioration du *screening*

## 3.1 Etude de faisabilité

### 3.1.1 Description de la méthode

L'objectif de ce travail est d'améliorer le logiciel d'alignement de séquences Match-Box. Comme décrit dans l'introduction, ce logiciel fonctionne avec trois algorithmes complémentaires et successifs : le *scanning*, suivi de quatre cycles de *matching-screening*. Nous allons focaliser notre travail sur l'étape de *screening*. Pour vérifier s'il est possible d'améliorer cette étape, nous avons analysé le premier cycle de *matching-screening* de Match-Box afin de déterminer la quantité de boîtes correctes que le *screening* va rechercher dans la banque créée par le *matching*. Nous avons réalisé ces tests en utilisant la matrice de score Blosum 62 (Henikoff and Henikoff, 1992; Henikoff and Henikoff, 1993), notre batterie de 78 cas-tests et des longueurs de fenêtres d'analyse allant de 5 à 21. De plus, pour tester la qualité de l'alignement nous avons utilisé les deux critères de vérité expliqués dans le chapitre précédent.

Dans les tests qui vont suivre, il nous est indispensable de modifier l'étape précédant le *screening*, à savoir le *matching*. En effet, le *matching* est constitué de deux parties : la première cherche les segments les plus similaires dans les séquences de manière à former des boîtes potentielles (recherche du « best match »), la deuxième élimine des boîtes sur base de critères statistiques établis lors de l'étape de *scanning* à savoir le filtre. Lorsque ce filtre est utilisé au moment du *matching*, le *screening* travaille peu car il ne fait que trier les boîtes que lui donne le *matching*. Par contre, quand ce filtre est enlevé au moment du *matching*, le *screening* doit rechercher l'information correcte dans ce qui vient du *matching*, pour ensuite arranger les boîtes afin d'obtenir un alignement correct. En d'autres termes, on remarque que la présence du filtre au niveau du *matching* provoque une augmentation de confiance au

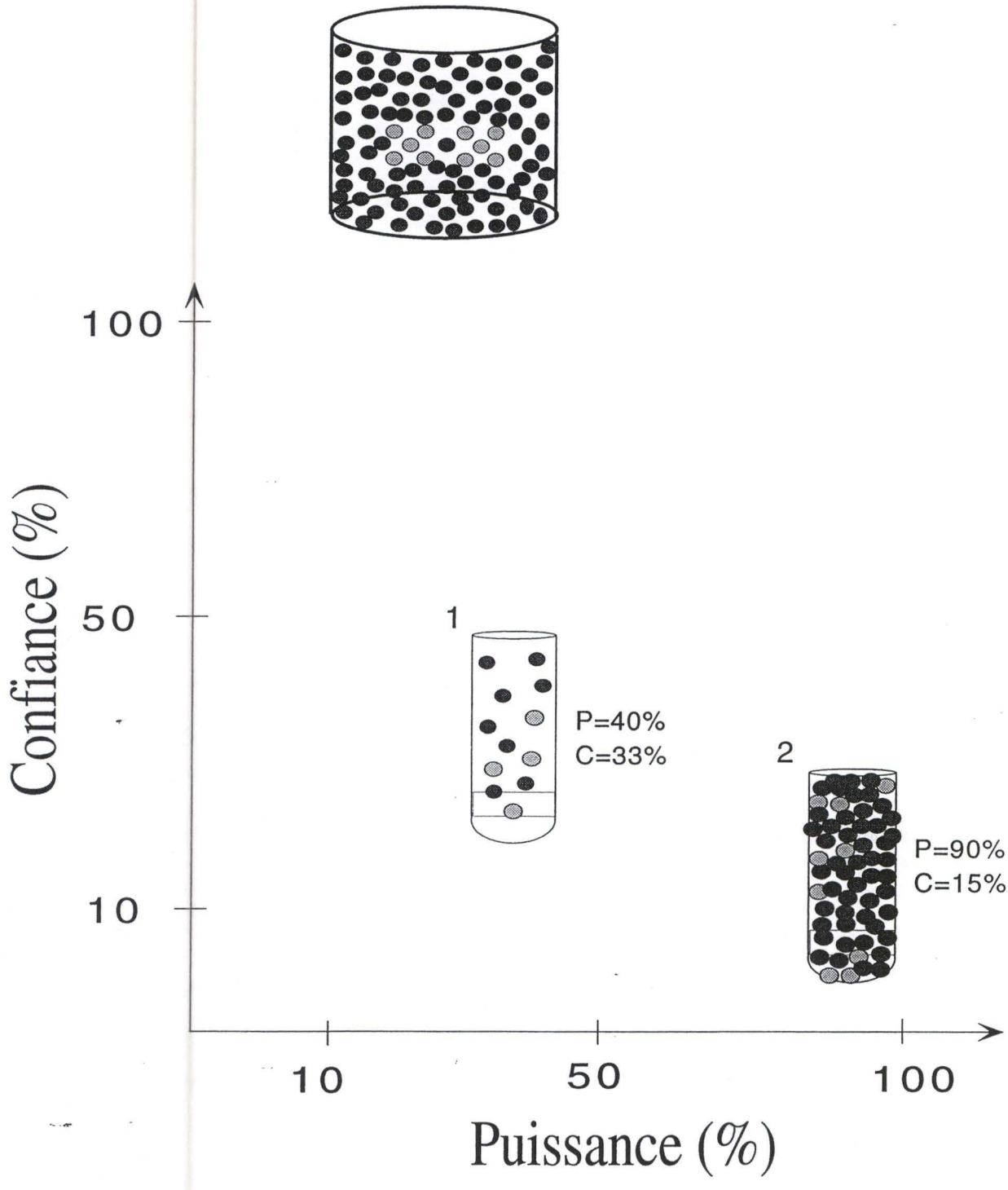



***Figure: 3.1.1.1***

*Comparaison de la puissance et de la confiance obtenue en présence et en absence des filtres au moment du matching.*
*Considérons un cylindre contenant un ensemble de billes claires et foncées ; le but de l'expérience étant de les séparer et de ne garder que les claires. Deux tentatives indépendantes ont été réalisées et les résultats ont été placés dans deux éprouvettes. La première (1) correspond aux résultats en confiance et en puissance du matching lorsque celui-ci possède un filtre. La deuxième (2) représente les résultats toujours en confiance et en puissance du matching mais lorsque celui-ci ne possède plus les filtres. On constate que le retrait des filtres assure une augmentation en puissance mais provoque une diminution en confiance.*

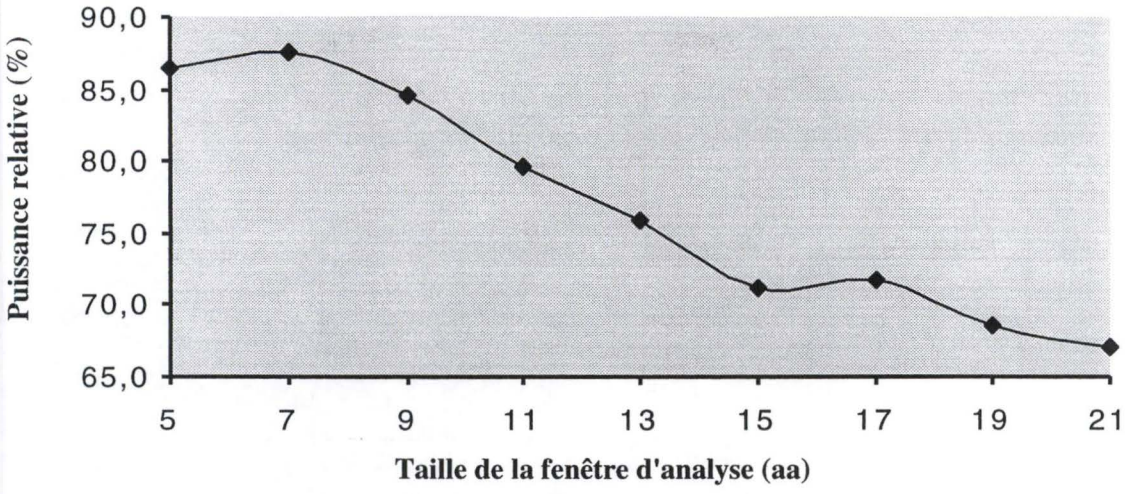



*Figure : 3.1.2.1*

*Variation de la puissance relative en fonction de l'augmentation de la taille de la fenêtre d'analyse en utilisant comme critère de vérité ce qui est aligné sans gap.*

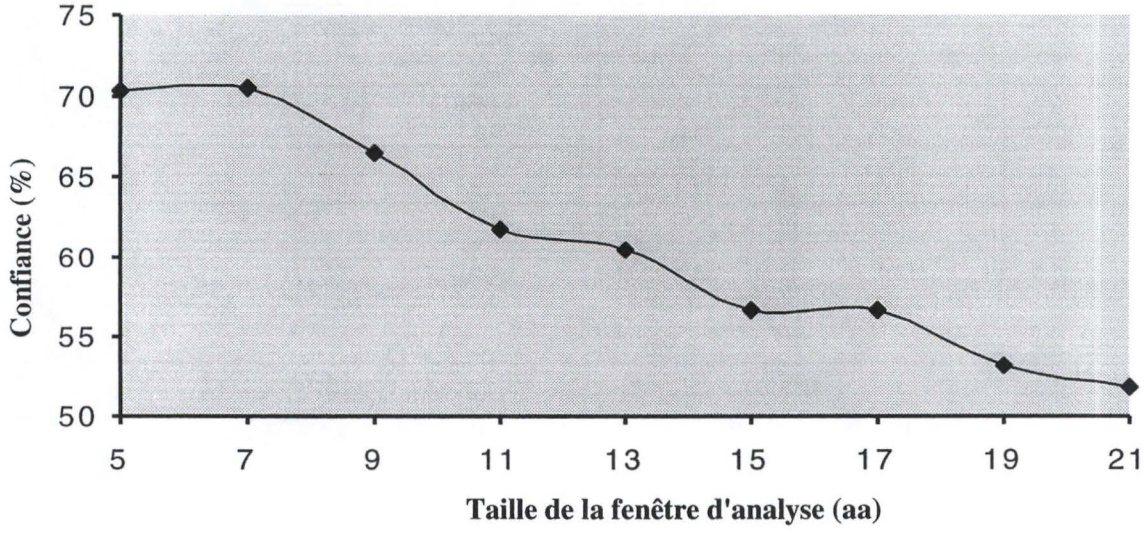



*Figure : 3.1.2.2*

*Variation de la confiance en fonction de l'augmentation de la taille de la fenêtre en utilisant comme critère de vérité ce qui est aligné sans gap.*

*Tableau 3.1.2.1: Résultats du screening de Match-Box sans filtre et sur un seul run avec une vérité définie comme étant ce qui est aligné sans gap.*

| Taille de la fenêtre (aa) | Puissance (%) | Confiance (%) | Puissance relative (%) |
|---|---|---|---|
| 5 | 47,4 | 70,4 | 86,5 |
| 7 | 56,9 | **70,5** | **87,5** |
| **9** | **59,8** | 66,5 | 84,6 |
| 11 | 59,3 | 61,8 | 79,7 |
| 13 | 58,2 | 60,4 | 75,9 |
| 15 | 55,1 | 56,6 | 71,3 |
| 17 | 55,9 | 56,7 | 71,8 |
| 19 | 53,4 | 53,2 | 68,5 |
| 21 | 52,4 | 51,9 | 67,0 |

*Tableau 3.1.2.2: Résultats du screening de Match-Box sans filtre et sur un seul run avec une vérité définie comme étant les zones où la conservation des structures secondaires est de 100%.*

| Taille de la fenêtre (aa) | Puissance (%) | Confiance (%) | Puissance relative (%) |
|---|---|---|---|
| 5 | 49,2 | **29,3** | 85,4 |
| 7 | 58,2 | 28,8 | **86,7** |
| **9** | **60,9** | 27,5 | 83,4 |
| 11 | 59,7 | 25,6 | 79,3 |
| 13 | 59,8 | 25,2 | 77,0 |
| 15 | 56,6 | 24,4 | 73,2 |
| 17 | 56,8 | 24 | 73,0 |
| 19 | 53,6 | 22,2 | 69,2 |
| 21 | 52,4 | 21,4 | 67,4 |

profit d'une diminution de la puissance *(figure 3.1.1.1)*. Dès lors, le retrait de ce filtre est effectué pour garantir une amélioration et pour évaluer au mieux les performances du *screening*.

### 3.1.2 Résultats

Avant de traiter les résultats ci-dessous, il est nécessaire de définir quelques termes utilisés :

- Nous appellerons la puissance du *screening* comme étant la quantité totale d'informations correctes dans les cas-tests ;
- Nous appellerons la puissance relative comme étant la puissance relative à la vérité disponible fournit par le *matching*.

Il est évident que les performances du *screening* dépendent de ce que le *matching* lui fournit comme information.

1) Analyse des résultats lorsque nous utilisons ce qui est aligné sans gap dans l'alignement de structure comme critère de vérité. La *figure 3.1.2.1* montre que la puissance relative avoisine les 85% lorsqu'on utilise des fenêtres d'analyse de 5, 7 et 9 résidus de longueur, avec un maximum de 87,5% pour une taille de fenêtre d'analyse égale à 7 résidus. Par contre, pour le reste des résultats de ce graphique, nous constatons que la tendance générale veut que plus la taille de la fenêtre d'analyse augmente (11 à 21), plus la puissance relative diminue. Pour ce qui est de l'évolution de la confiance en fonction de la taille de la fenêtre d'analyse exprimée par la *figure 3.1.2.2*, nous voyons des résultats plus ou moins similaires. En effet, nous observons une confiance élevée de l'ordre de 70% pour des longueurs de fenêtres de 5 et 7 et puis, de manière générale, celle-ci diminue en fonction de l'augmentation de la taille de la fenêtre (les données des différentes figures proviennent du *tableau 3.1.2.1*).

2) Analyse des résultats en prenant comme critère de vérité les zones de l'alignement de structure où la conservation de la structure secondaire est de 100%. Cette analyse fournit des résultats plus ou moins similaires. En effet, nous avons une puissance

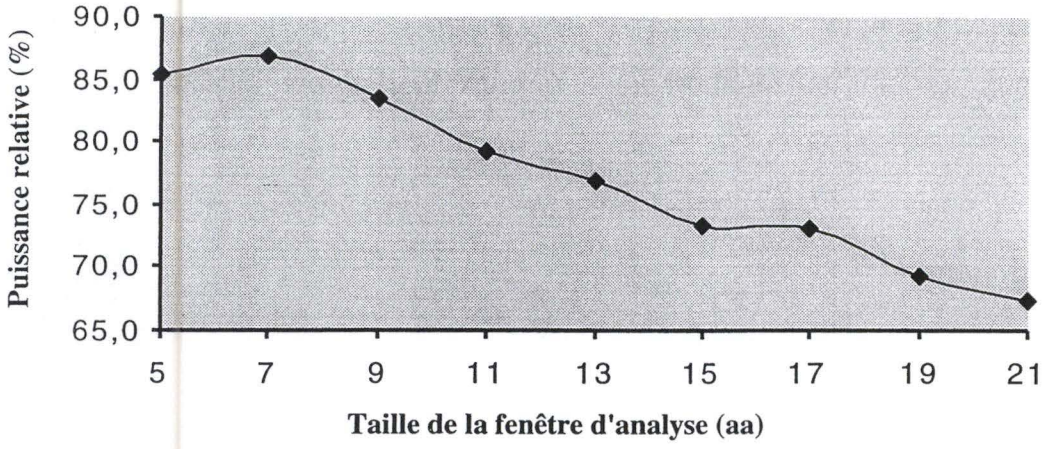



*Figure : 3.1.2.3*

*Variation de la puissance relative en fonction de l'augmentation de la taille de la fenêtre d'analyse en utilisant comme critère de vérité les zones où la conservation de la structure secondaire est de 100%.*

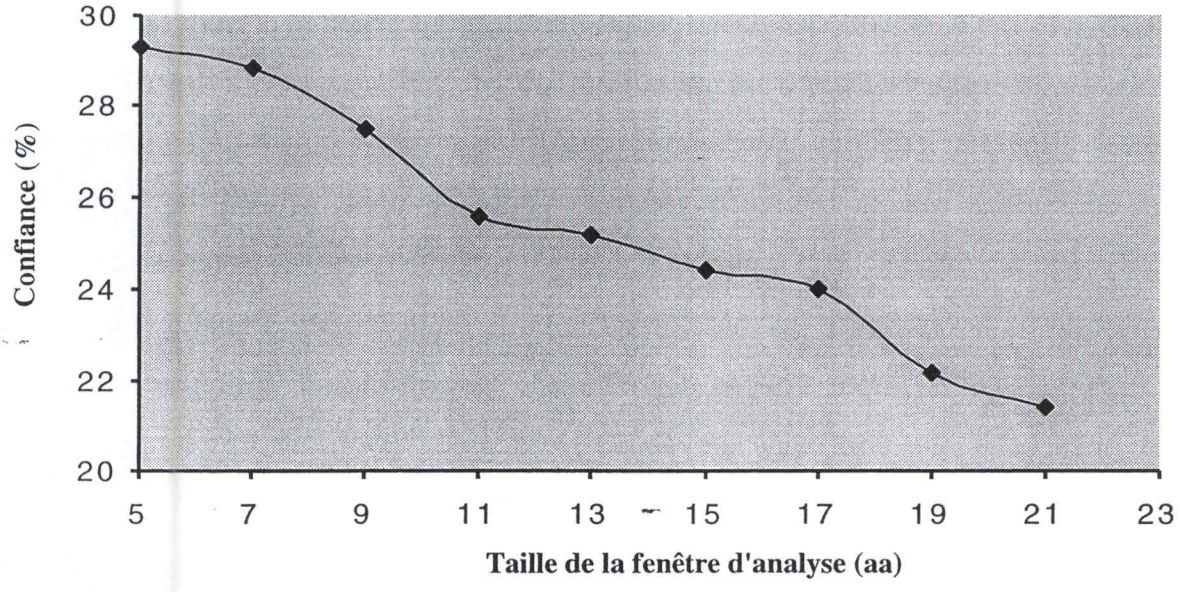



*Figure : 3.1.2.4*

*Variation de la confiance en fonction de l'augmentation de la taille de la fenêtre d'analyse en utilisant comme critère de vérité les zones où la conservation de la structure secondaire est de 100%.*

relative de 86,7% *(figure 3.1.2.3)* lorsqu'une fenêtre d'analyse de 7 résidus est utilisée. De plus, nous voyons dans la *figure 3.1.2.4* que la confiance diminue en fonction de l'augmentation de la taille de la fenêtre d'analyse (les données des différentes figures proviennent du *tableau 3.1.2.2*).

Cette tendance générale, c'est-à-dire la diminution des performances lorsqu'on augmente la taille de la fenêtre d'analyse, peuvent s'expliquer par deux observations. D'une part, l'utilisation de fenêtres d'analyse de taille élevée engendre un phénomène d'incompatibilité entre boîtes. Par conséquent, le nombre de boîtes choisies par le *screening* est diminué (perte de puissance). D'autre part, le nombre de résidus non alignés correctement augmente lorsqu'on utilise des fenêtres d'analyses de taille élevées.

### 3.1.3 Conclusions

Suite aux résultats décrits ci-dessus, nous pouvons déduire quelques conclusions :

- Pour les deux critères de vérité considérés (ce qui est aligné sans gap dans l'alignement de structure ou bien les zones où la conservation de la structure secondaire est de 100% comme critère de vérité), nous constatons que la puissance relative n'atteint jamais les 100%. Cela dit, nous n'en sommes pas si loin. En effet, un maximum de 87,5% *(tableau 3.1.2.1)* est atteint lorsque le critère de vérité correspond à ce qui est aligné sans gap et un maximum de 86,7% *(tableau 3.1.2.2)* est obtenu avec l'autre critère de vérité (zones où la conservation de la structure secondaire est de 100%). N'arrivant donc pas à 100% d'informations correctes trouvées par le *screening* à partir de la banque créée par le *matching*, nous pouvons en déduire que l'amélioration du *screening* reste possible.
- Quel que soit le critère de vérité utilisé, les conclusions sont semblables. Ceci est important car cela rejoint ce que nous avions espéré dans le chapitre 2. En effet, nous souhaitions dans ce chapitre que les trois façons de considérer la « vérité » (RMS, aligné sans gap, 100% de conservation de la structure secondaire) pourraient fournir des conclusions similaires malgré leurs caractéristiques divergentes du point de vue de leur stringence. Notre observation que les

conclusions sont similaires quel que soit le critère de vérité considéré nous permet de supposer que nos conclusions précitées seraient identiques si nous prenions aussi la détermination de la distance RMS comme critère de vérité.

- La puissance relative est la plus élevée avec une fenêtre d'analyse ayant une taille de 7 résidus *(tableau 3.1.2.1 et 3.1.2.2)*. Nous montrons ainsi que le *screening*, dans sa version actuelle, fonctionne de manière optimale par rapport au *matching* lorsque la taille de la fenêtre d'analyse est de 7. Cependant, ce qui nous intéresse *in fine* pour l'amélioration du logiciel Match-Box est la quantité totale d'informations correctes (puissance du *screening*) se trouvant à la sortie du *screening (tableaux 3.1.2.1 et 3.1.2.2)*. Il se trouve que ce taux est alors plus élevé lorsqu'une fenêtre d'analyse de 9 résidus est utilisée plutôt qu'une fenêtre d'analyse de 7 résidus.

### 3.1.4 Perspectives

Suite aux conclusions que nous avons tirées au point précédent, nous déduisons qu'il est possible d'améliorer encore l'étape de *screening* de Match-Box. Cette amélioration tentera d'être réalisée en poursuivant deux approches : la première se base sur la conception d'un nouvel algorithme, la deuxième consiste à améliorer le processus du *screening* actuel en tenant compte des structures secondaires.

## 3.2 Développement d'une nouvelle stratégie de *screening*

### 3.2.1 Description de la méthode

Avant de rentrer dans le vif du sujet, à savoir l'explication de cette nouvelle stratégie que nous nous proposons d'utiliser pour le *screening*, il faut peut-être rappeler brièvement le fonctionnement du *screening* actuel de Match-Box. Celui-ci prend en compte la longueur de la boîte pour déterminer quelle est la meilleure d'entre elles au sortir de l'étape de *matching*. Donc, lorsque les boîtes sortent de l'étape du *matching*, le *screening* sélectionne la plus longue d'entre elles comme étant la meilleure.

*Tableau 3.2.2.1a*

*Résultats obtenus sur les 78 cas-tests en confiance et en puissance en utilisant le screening actuel de Match-Box, ce qui est aligné sans gap comme critère de vérité et en variant la la longueur de la fenêtre d'analyse de 7 à 21 aa.*

| Taille (aa) | Puissance (%) | Confiance (%) |
|---|---|---|
| 5 | 47,4 | 70,4 |
| 7 | 56,9 | 70,5 |
| 9 | 59,8 | 66,5 |
| 11 | 59,3 | 61,8 |
| 13 | 58,2 | 60,4 |
| 15 | 55,1 | 56,6 |
| 17 | 55,9 | 56,7 |
| 19 | 53,4 | 53,2 |
| 21 | 52,4 | 51,6 |

*Tableau 3.2.2.1b*

*Résultats obtenus sur les 78 cas-tests en confiance et en puissance en utilisant le nouveau screening de Match-Box, ce qui est aligné sans gap comme critère de vérité et en variant la la longueur de la fenêtre d'analyse de 7 à 21 aa.*

| Taille (aa) | Puissance (%) | Confiance (%) |
|---|---|---|
| 5 | 46,1 | 62,9 |
| 7 | 56,2 | 67,8 |
| 9 | 60,8 | 68,1 |
| 11 | 61,5 | 65,9 |
| 13 | 60,9 | 64,5 |
| 15 | 62 | 64,3 |
| 17 | 60,2 | 61,8 |
| 19 | 58,9 | 60,2 |
| 21 | 57,5 | 58,4 |

*Tableau 3.2.2.1c*

*Résultats obtenus sur les 78 cas-tests en confiance et en puissance en utilisant le screening actuel de Match-Box, les zones où la conservation de la structure secondaire est de 100% comme critère de vérité et an variant la longueur de la fenêtre d'analyse de 7 à 21 aa.*

| Taille (aa) | Puissance (%) | Confiance (%) |
|---|---|---|
| 5 | 49,2 | 29,3 |
| 7 | 58,2 | 28,8 |
| 9 | 60,9 | 27,5 |
| 11 | 59,7 | 25,6 |
| 13 | 59,8 | 25,2 |
| 15 | 56,6 | 24,4 |
| 17 | 56,8 | 24 |
| 19 | 53,6 | 22,2 |
| 21 | 52,4 | 21,4 |

*Tableau 3.2.2.1d*

*Résultats obtenus sur les 78 cas-tests en confiance et en puissance en utilisant le nouveau screening de Match-Box, les zones où la conservation de la structure secondaire est de 100% comme critère de vérité et an variant la longueur de la fenêtre d'analyse de 7 à 21 aa.*

| Taille (aa) | Puissance (%) | Confiance (%) |
|---|---|---|
| 5 | 47,9 | 26,8 |
| 7 | 58,4 | 28,5 |
| 9 | 63,4 | 28,8 |
| 11 | 63,1 | 27,1 |
| 13 | 62,1 | 27 |
| 15 | 63 | 27,2 |
| 17 | 61,8 | 26,3 |
| 19 | 59,7 | 25,5 |
| 21 | 56,9 | 24,2 |

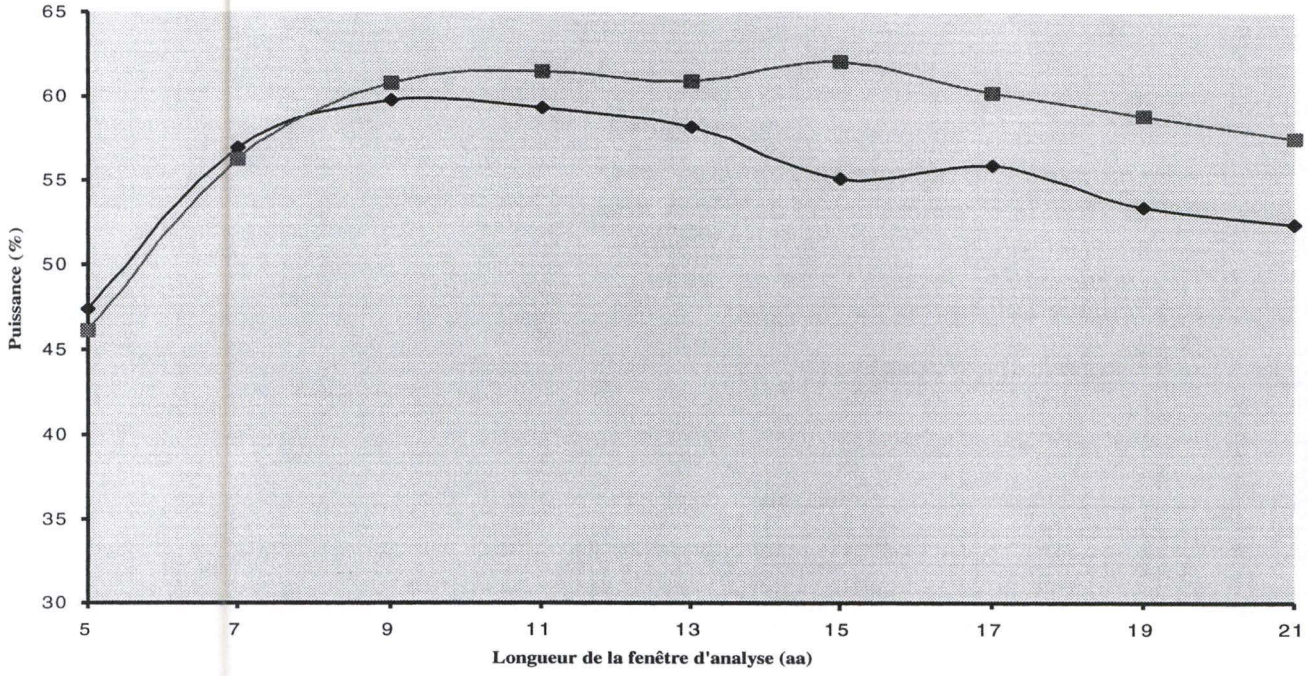



***Figure : 3.2.2.1***

*Evolution de la puissance à la sortie du screening en fonction de la longueur de la fenêtre d'analyse entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) en utilisant ce qui est aligné sans gap comme critère de vérité.*

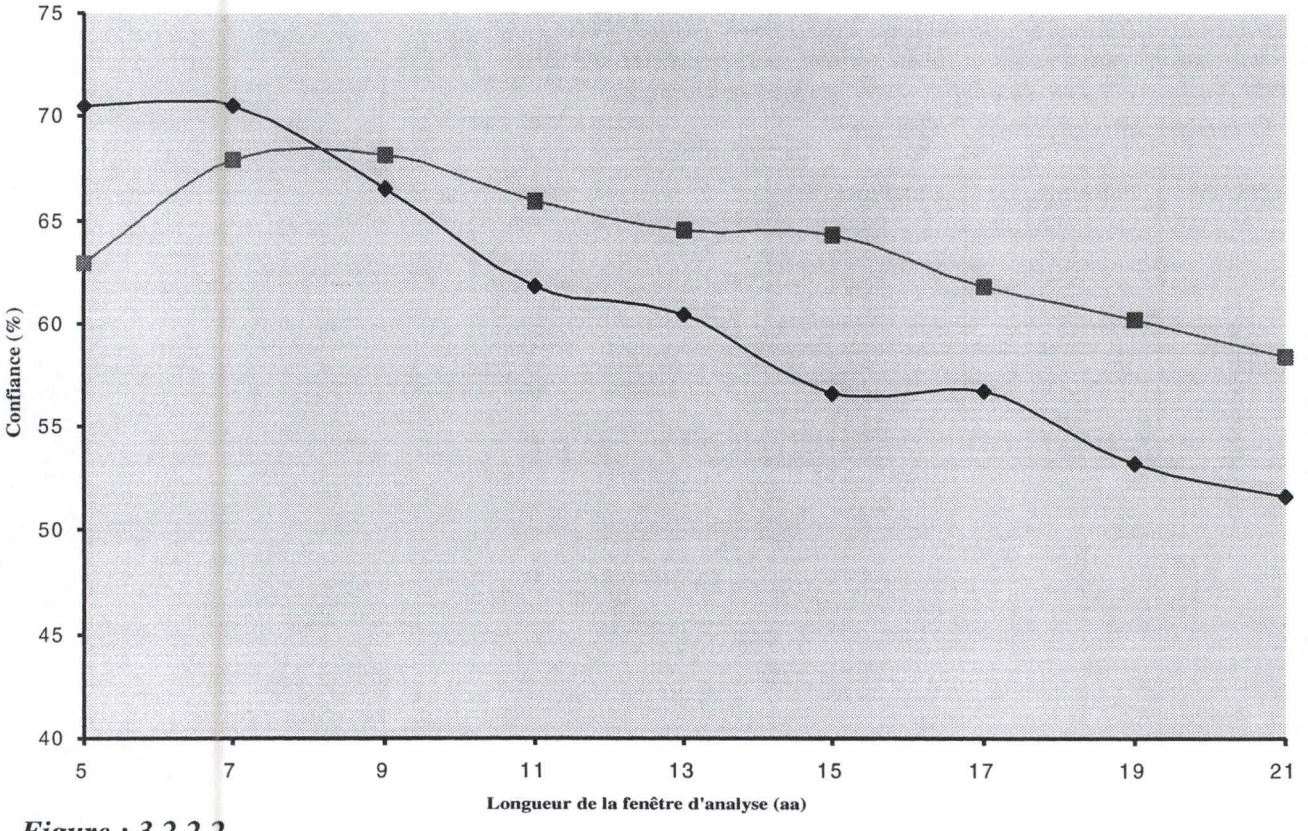



***Figure : 3.2.2.2***

*Evolution de la confiance du screening en fonction de la longueur de la fenêtre d'analyse entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) en utilisant ce qui est aligné sans gap comme critère de vérité.*

Ensuite, il cherche toutes celles qui sont compatibles avec cette dernière et choisit la plus longue d'entre elles. Ensuite, il prend la plus longue boîte compatible avec les deux premières, détermine les boîtes compatibles avec celle-ci, choisit la plus longue d'entre elles et ainsi de suite (voir introduction).

La nouvelle stratégie de *screening* que nous proposons d'implémenter dans Match-Box a toujours pour objectif d'augmenter au maximum la quantité d'information que le *screening* va chercher parmi les boîtes isolées au terme de l'étape de *matching*, mais ceci se ferait sans plus tenir compte de la longueur de la boîte. Ce nouveau *screening* s'intéresse plutôt à la longueur totale de l'alignement. Il fonctionne de la façon suivante : à la sortie du *matching*, le *screening* teste tous les arrangements possibles de toutes les boîtes compatibles entre elles afin d'obtenir l'alignement le plus long possible.

Nous avons soumis à cette nouvelle procédure de *screening* la batterie de 78 cas-tests, en utilisant la matrice de score que Match-Box possède par défaut qui est Blosum 62 et en faisant varier la longueur de la fenêtre d'analyse de 5 à 21. Nous avons exécuté ces tests d'abord avec ce qui est aligné sans gap comme critère de vérité mais également, par la suite, avec les zones où la conservation de la structure secondaire est conservée à 100% comme autre critère de vérité.

### 3.2.2 Résultats

Les différents résultats obtenus suite à cette nouvelle stratégie de *screening* sont repris sur les *tableaux 3.2.2.1 a,b,c et d.*

L'analyse de tous les résultats obtenus a été effectuée de la façon suivante :

- Nous avons exprimé sous forme de graphique l'évolution de la puissance en fonction de la variation de la longueur de la fenêtre d'analyse et ce pour les deux critères de vérité communément utilisés. Les résultats de ces tests sont présentés sur les *figures 3.2.2.1 et 3.2.2.3*. Ceux-ci montrent que les performances obtenues grâce au nouveau *screening* (puissance) sont en moyenne meilleures que celles obtenues avec le *screening* actuel de Match-Box, et ce quel que soit le critère de vérité considéré.

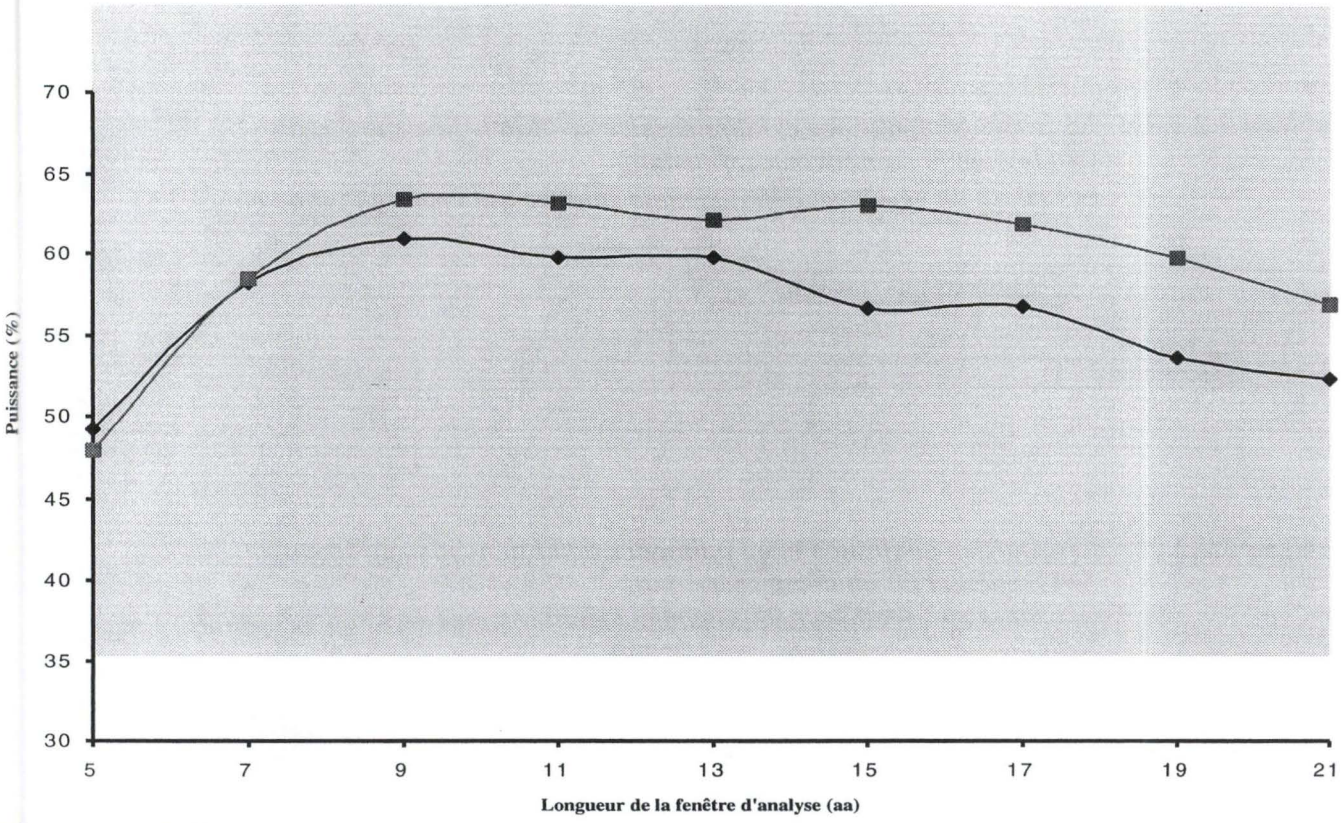



***Figure : 3.2.2.3***

*Evolution de la puissance à la sortie du screening en fonction de la longueur de la fenêtre d'analyse entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) en utilisant les zones où la conservation de la structure secondaire est de 100%.*

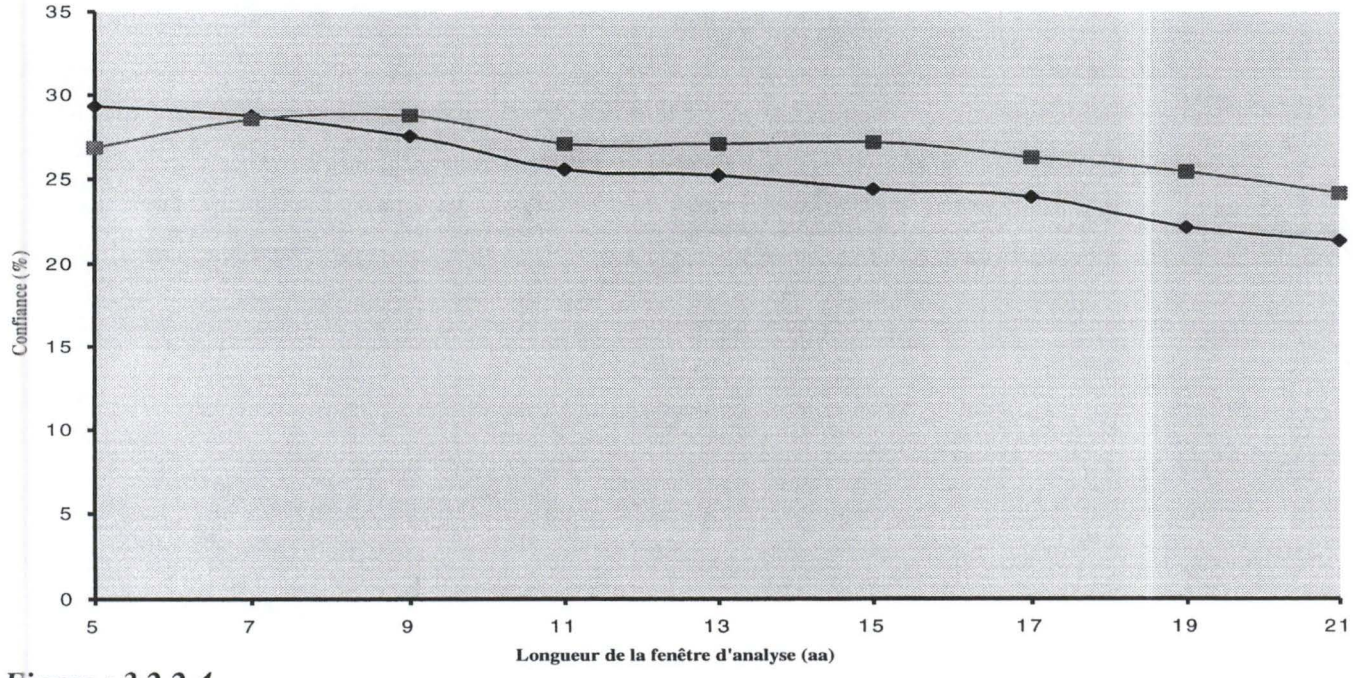



***Figure : 3.2.2.4***

*Evolution de la confiance du screening en fonction de la longueur de la fenêtre d'analyse entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) en utilisant les zones où la conservation de la structure secondaire est de 100%.*

*Tableau 3.2.2.2a: Résultats en confiance et en puissance de Match-Box sur 4 runs, de l'ancien et du nouveau screening. Le critère de vérité est défini comme étant ce qui est aligné sans gap.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Match-Box complet** | 63,1 | 69,6 |
| **Ancien screening** | 59,8 | 66,5 |
| **Nouveau screening** | 60,8 | 68,1 |

*Tableau 3.2.2.2b: Résultats en confiance et en puissance de Match-Box sur 4 runs, de l'ancien et du nouveau screening. Le critère de vérité est défini comme étant les zones où la conservation de la structure secondaire est de 100%.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Match-Box complet** | 64,9 | 28,9 |
| **Ancien screening** | 60,9 | 27,5 |
| **Nouveau screening** | 63,4 | 28,8 |

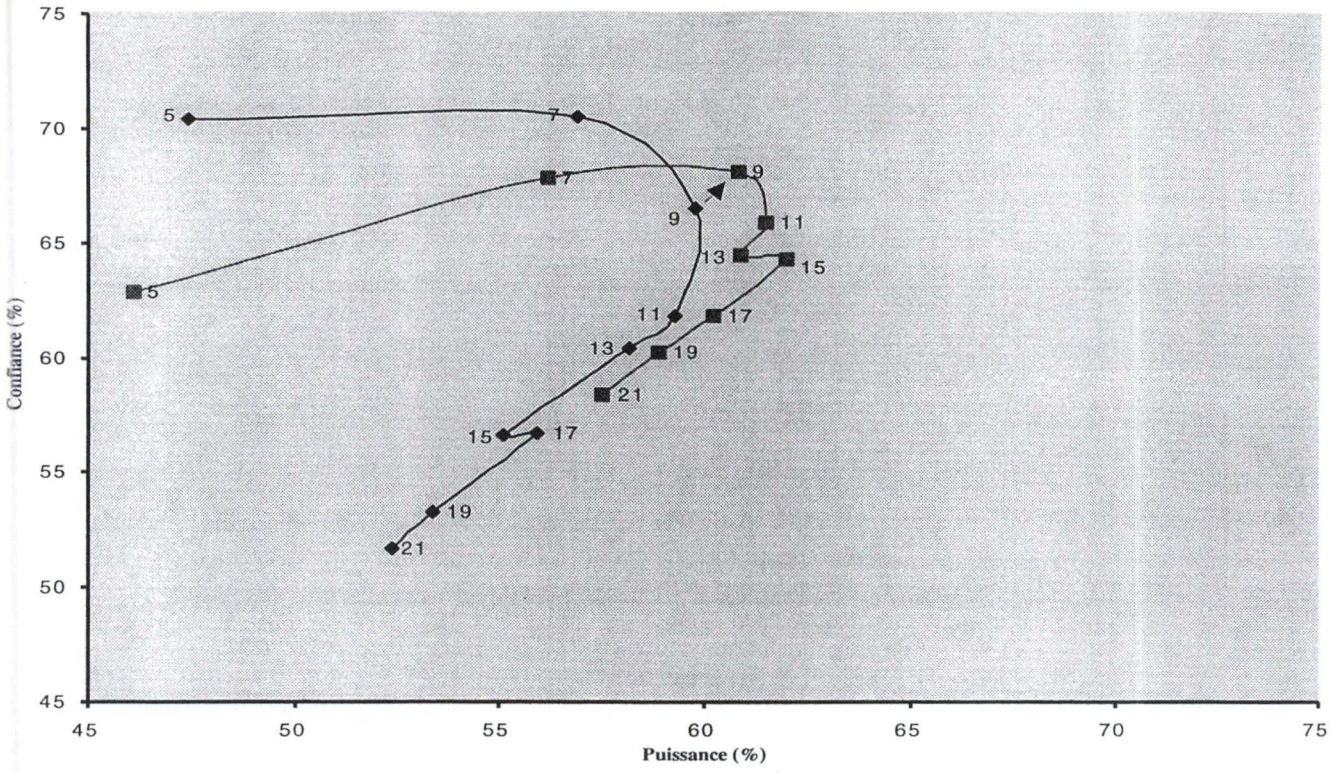



*Figure : 3.2.2.5*

*Comparaison entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) sur les 78 cas-tests en utilisant comme vérité ce qui est aligné sans gap et en variant la taille de la fenêtre d'analyse.*

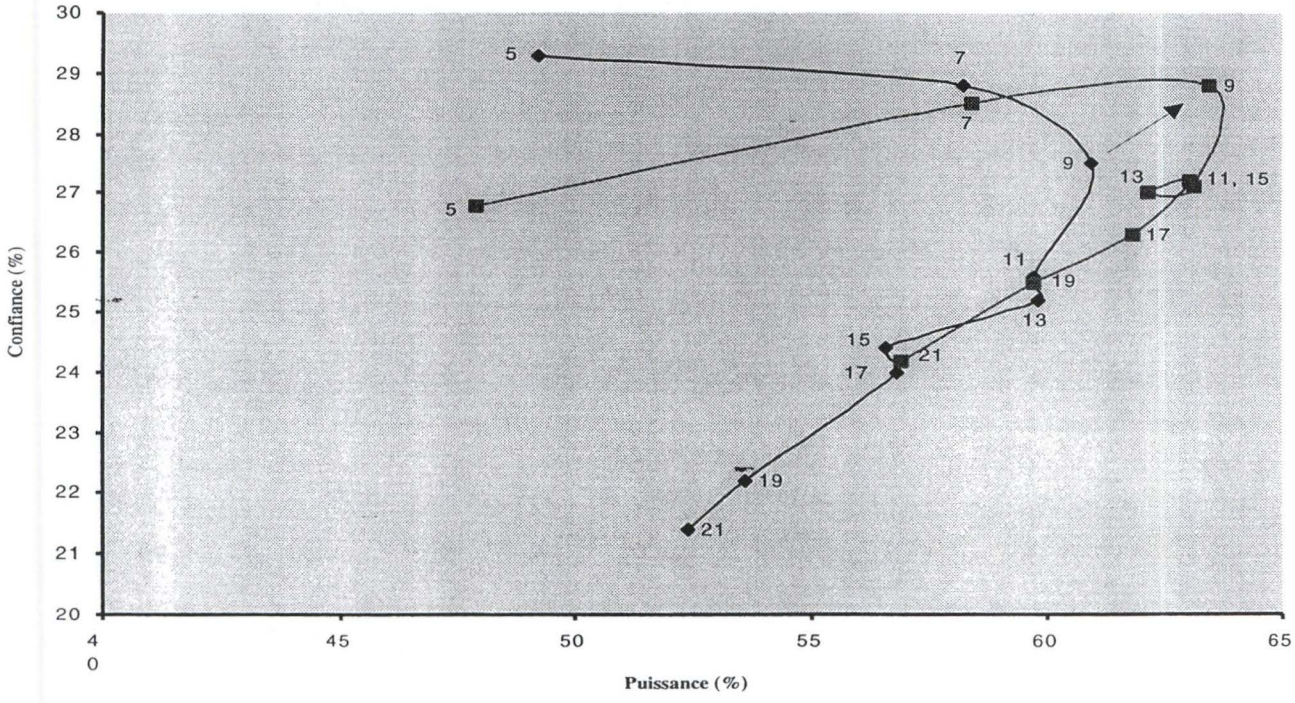



*Figure : 3.2.2.6*

*Comparaison entre l'ancien (courbe foncée) et le nouveau screening (courbe claire) sur les 78 cas-tests en utilisant comme vérité les zones où la conservation de la structure secondaire est de 100% et en variant la taille de la fenêtre d'analyse.*

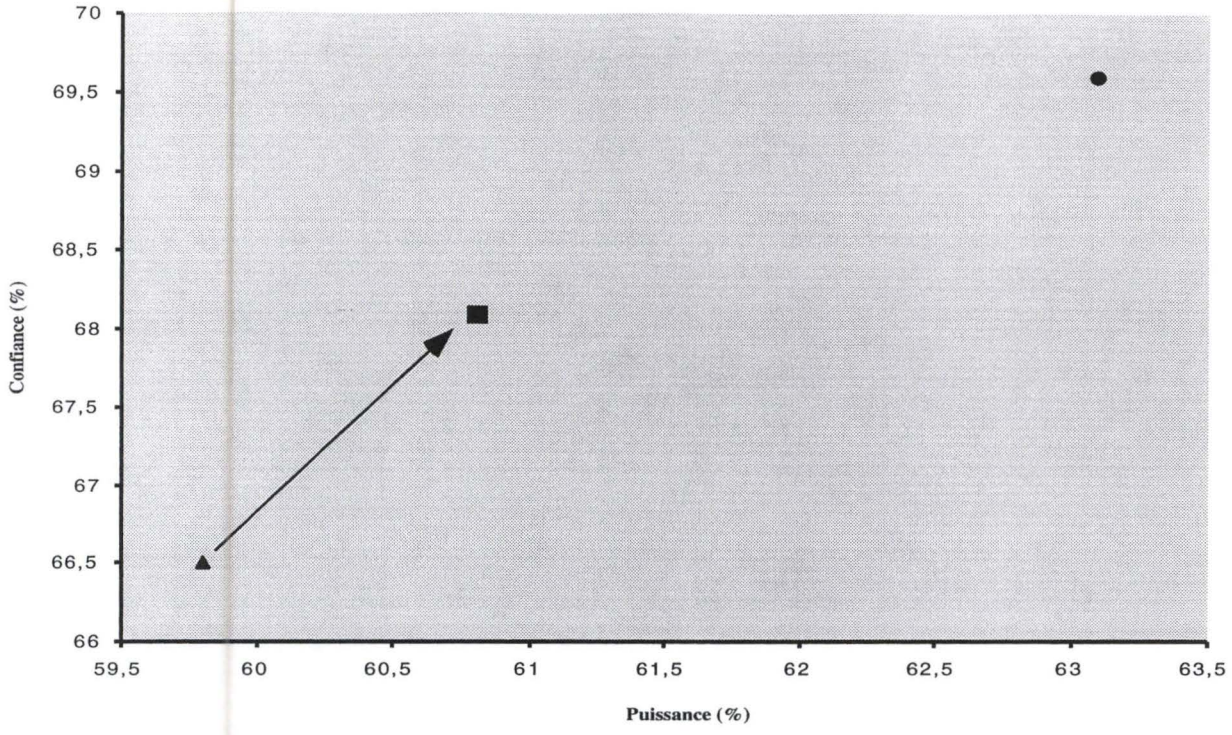



***Figure : 3.2.2.7***

*Comparaison entre l'ancien screening (petit triangle), le nouveau screening (petit carré) et Match-Box complet (4 runs) (petit rond) sur les 78 cas-tests en utilisant comme vérité ce qui est aligné sans gap et une fenêtre d'analyse de taille 9 aa.*

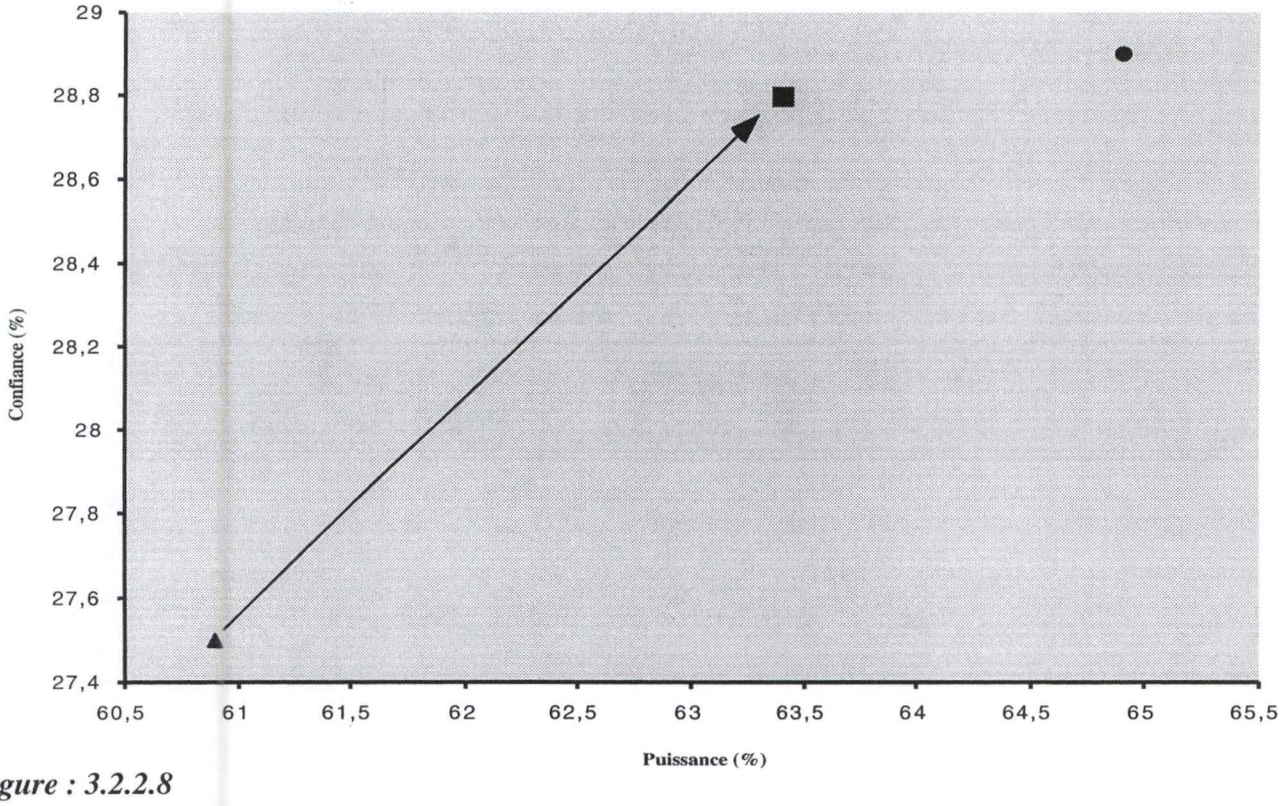



***Figure : 3.2.2.8***

*Comparaison entre l'ancien screening (petit triangle), le nouveau screening (petit carré) et Match-Box complet (4 runs) (petit rond) sur les 78 cas-tests en utilisant comme vérité les zones où la conservation de la structure secondaire est de 100% et une fenêtre d'analyse de taille 9 aa.*

- Nous avons ensuite analysé l'évolution de la confiance en fonction de la longueur de la fenêtre d'analyse et toujours suivant les deux mêmes critères de vérité. Ces résultats, exprimés dans les *figures 3.2.2.2 et 3.2.2.4*, fournissent les mêmes conclusions pour la confiance que celles déduites pour la puissance.
- Les *figures 3.2.2.5 et 3.2.2.6* représentent les résultats obtenus en confiance et en puissance pour l'ancien et le nouveau *screening*, en utilisant l'un ou l'autre des deux critères de vérité et en faisant varier la taille de la fenêtre d'analyse. Un examen plus précis de ces deux graphiques montre qu'une fenêtre d'analyse ayant une taille de 9 résidus offre un point confiance-puissance meilleur (se rapprochant plus des 100% de confiance et de puissance) qu'avec une autre taille de fenêtre d'analyse lorsque nous travaillons avec la nouvelle procédure de *screening*.
- Puisque nous venons de conclure qu'une fenêtre d'analyse ayant une longueur de 9 résidus permettait l'obtention de meilleurs résultats dans le nouveau *screening*, nous l'avons isolé sur un graphe. Les résultats que nous analysons maintenant proviennent des *tableaux 3.2.2.2 a et b*. Ils sont exprimés sur les *figures 3.2.2.7 et 3.2.2.8*. Sur la première figure sont représentées les deux valeurs de confiance-puissance pour le nouveau et l'ancien *screening* (sur un run, cfr. chapitre 2) et celui de Match-Box tel qu'il existe actuellement, c'est-à-dire fonctionnant sur 4 runs. Nous constatons qu'avec la nouvelle procédure de *screening*, fonctionnant sur un seul run et non quatre comme dans Match-Box, nous sommes très proches de la version actuelle Match-Box.

Nous n'avons pas testé la nouvelle stratégie de *screening* sur quatre runs comme dans la version actuelle de Match-Box, et ce pour plusieurs raisons :

- L'utilisation du nouveau *screening* dans la version actuelle de Match-Box fonctionnant par défaut avec un *matching* ayant un filtre aurait pris un temps de calcul trop important (quatre mois). De plus si le temps l'avait permis, l'augmentation des performances n'aurait pas été certaine car on sait que la nouvelle procédure de *screening* ne travaille pas de manière optimale quand le *matching* possède un filtre.

### 3.2.3 Conclusions

Suite à ces différents résultats, nous pouvons conclure que :

- Le nouveau *screening* est, de façon générale, meilleur en confiance et en puissance que celui utilisé par la version actuelle de Match-Box et ce quel que soit le critère de vérité choisi (ce qui est aligné sans gap ou les zones où la structure secondaire est conservée à 100%) ;
- les résultats obtenus par un seul run du nouveau *screening* sont très proches de ceux obtenus avec la version actuelle de Match-Box lorsque celui-ci fonctionne sur 4 runs ;
- l'optimum des résultats pour les deux critères de vérité s'observe avec une longueur de fenêtre d'analyse égale à 9 résidus, c'est-à-dire celle qui est utilisée dans la version actuelle de Match-Box.

### 3.2.4 Perspectives

- Optimiser la nouvelle procédure de *screening* pour que celui-ci fonctionne en plusieurs runs, comme dans la version actuelle de Match-Box ;
- Optimiser l'algorithme pour que le temps utilisé par le *screening* afin d'effectuer son travail soit réduit. Actuellement, le temps qu'utilise le nouveau *screening* varie de quelques secondes (pour quelques cas-tests) à plusieurs heures (pour la majorité des cas-tests) et parfois plusieurs mois (pour certains cas-tests), par opposition au *screening* actuel de Match-Box qui, lui, n'utilise que quelques secondes (pour chacun des cas-tests).

# 3.3 Prise en compte de la structure secondaire dans le *screening* actuel

## 3.3.1 Etude de faisabilité

### *3.3.1.1 Description de la méthode*

Avant de procéder à l'insertion de ce nouveau critère (la structure secondaire) dans l'étape *screening*, il nous faut d'abord vérifier que la conservation de la structure secondaire dans les boîtes est un facteur discriminant permettant de les distinguer. En d'autres termes, nous allons étudier la relation entre la conservation de la structure secondaire et la vérité dans les boîtes sortant du *matching*, et ce en analysant l'évolution de la confiance en fonction de la conservation de la structure secondaire.
Ces analyses ont été réalisées sur notre batterie de 78 cas-tests avec des tailles de fenêtres d'analyse allant de 5 à 21 résidus. Les résultats ont été obtenus en utilisant l'un des deux critères de vérité que nous avons déjà utilisé dans les tests qui ont précédé celui-ci, c'est à dire soit ce qui est aligné sans gap, soit les zones où la conservation de la structure secondaire est de 100%. La structure secondaire de chaque protéine de notre batterie de 78 cas-tests fut obtenue de deux manières :

- Soit en soumettant à DSSP le fichier PDB de chacune de ces protéines, renfermant les informations concernant leur structure tridimensionnelle. En retour, DSSP nous a donné la succession de structures secondaires présente dans chaque protéine ;
- soit en soumettant la séquence de chacune de ces protéines à PHD, qui nous donne alors en retour une structure secondaire prédite.

### *3.3.1.2 Résultats*

Les résultats que nous avons obtenus avec cette technique sont représentés sur 4 figures.

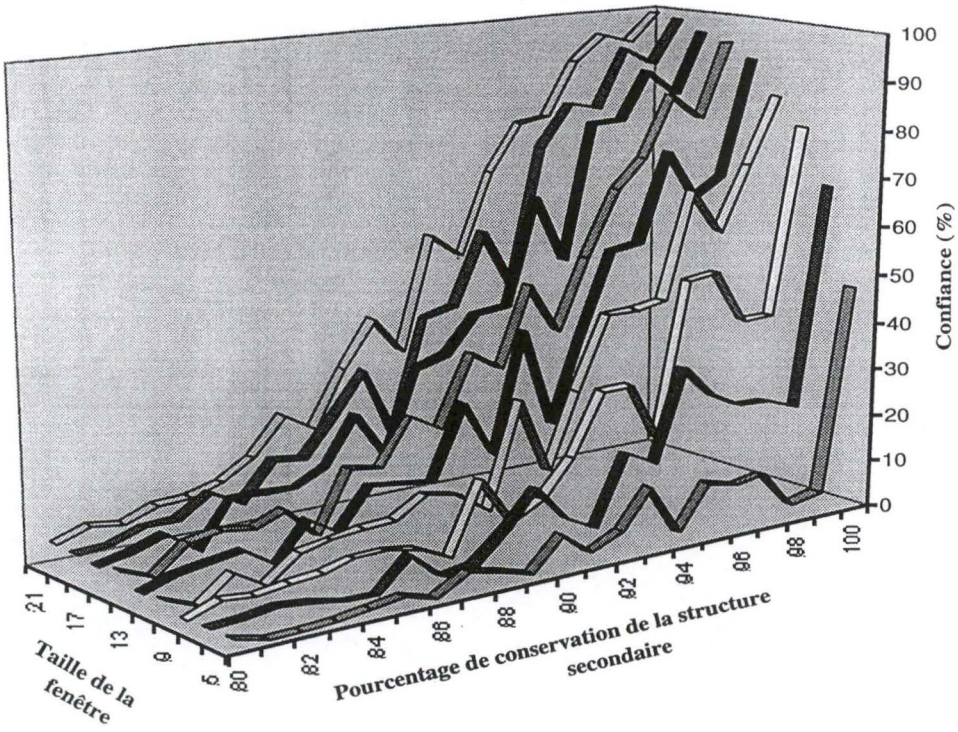



*Figure : 3.3.1.2.1*

*Evolution de la confiance en fonction de la conservation de la structure secondaire et de l'augmentation de la taille de la fenêtre d'analyse.*
*Le critère de vérité est représenté par ce qui est aligné sans gap et la structure secondaire est fournie par DSSP.*

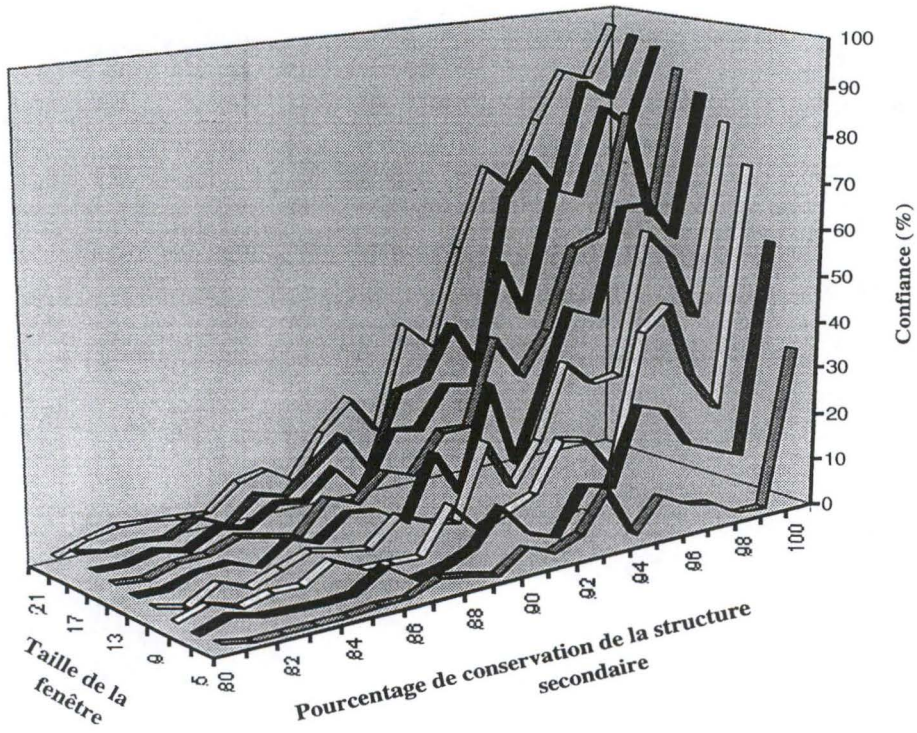



*Figure : 3.3.1.2.2*

*Evolution de la confiance en fonction de la conservation de la structure secondaire et de l'augmentation de la taille de la fenêtre d'analyse.*
*Le critère de vérité est représenté par ce qui est aligné sans gap et la structure secondaire est prédite par PHD.*

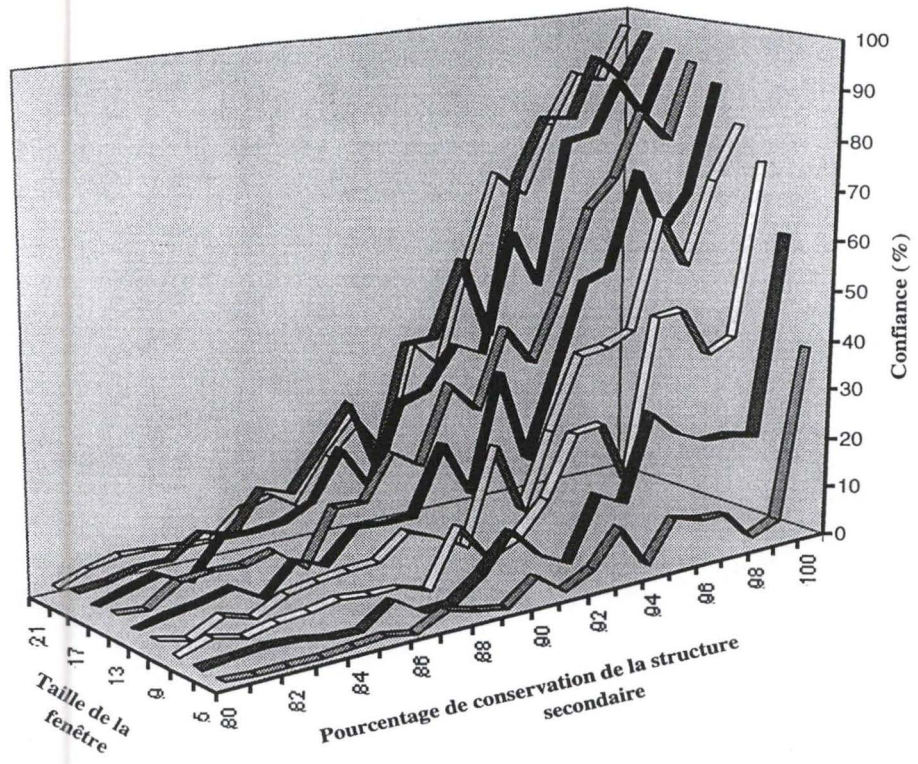



*Figure : 3.3.1.2.3*

*Evolution de la confiance en fonction de la conservation de la structure secondaire et de l'augmentation de la taille de la fenêtre d'analyse.*
*Le critère de vérité est représenté par les zones ou la conservation de la structure secondaire est de 100% et la structure secondaire est fournie par DSSP.*

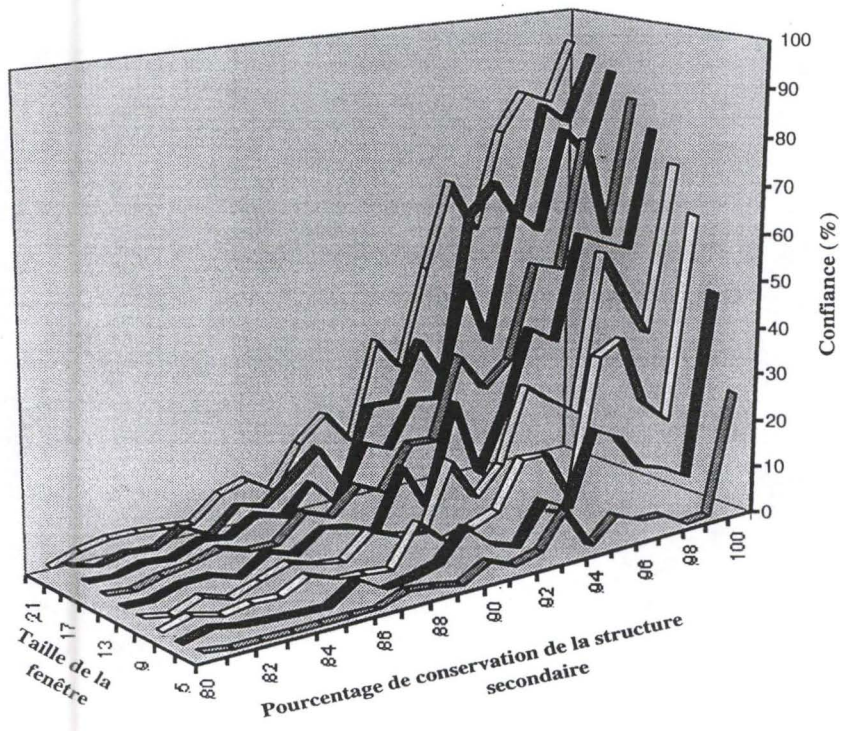



*Figure : 3.3.1.2.4*

*Evolution de la confiance en fonction de la conservation de la structure secondaire et de l'augmentation de la taille de la fenêtre d'analyse.*
*Le critère de vérité est représenté par les zones ou la conservation de la structure secondaire est de 100% et la structure secondaire est fournie par DSSP.*

- La première *figure (3.3.1.2.1)* représente l'évolution la confiance en fonction de la conservation de la structure secondaire et de la taille de la fenêtre d'analyse. Nous prenons dans ce cas ce qui est aligné sans gap comme critère de vérité ; la structure secondaire est fournie par DSSP. Les résultats de cette figure nous montrent que plus le pourcentage de conservation de la structure secondaire dans une boîte est élevé et plus cette boîte est longue, plus la confiance est élevée.
- La *figure 3.3.1.2.2* représente la même chose que la *figure 3.3.1.2.1*, exception faite de la méthode de prédiction des structures secondaires, qui a été réalisée par PHD. Les résultats obtenus au terme de cette analyse sont similaires. Cependant, nous observons que la confiance est relativement plus faible que lorsque la structure secondaire est fournie par DSSP, et ce quels que soient la taille de la boîte ou le pourcentage de conservation de la structure secondaire. Ceci peut être expliqué par le fait que la structure secondaire a été prédite et non observée en tant que « vérité ». Une prédiction n'étant pas une vérité, la confiance obtenue en utilisant une prédiction plutôt qu'une observation de la structure secondaire ne peut être que plus faible.
- Les *figures 3.3.1.2.3 et 3.3.1.2.4* représentent les mêmes relations que les deux premières mais cette fois, nous avons utilisé comme critère de vérité les zones où la conservation de la structure secondaire est de 100% plutôt que ce qui est aligné sans gap. L'analyse de ces deux figures montre une tendance similaire aux deux autres, avec aussi une confiance plus faible pour PHD que lorsque la structure secondaire est prédite par DSSP.
- Une comparaison entre les figures générées en utilisant les deux critères de vérité et où la structure secondaire a été prédite par PHD est intéressante *(figures 3.3.1.2.2 et 3.3.1.2.4)*. En effet, nous constatons que l'évolution de la confiance en fonction de la conservation de la structure secondaire et de la taille de la fenêtre d'analyse est meilleure lorsque ce qui est aligné sans gap est choisi comme critère de vérité plutôt que les zones où la conservation de la structure secondaire est de 100%.

#### 3.3.1.3 Conclusions

- Suite à l'analyse de ces résultats, nous pouvons tirer une conclusion générale sur les quatre figures : plus la structure secondaire est conservée dans une boîte et plus cette boîte est longue, plus l'alignement qui en résulte possède une confiance élevée.
- Par conséquent, nous montrons que la conservation de la structure secondaire dans une boîte est un critère important, que nous allons donc considérer dès à présent pour tenter d'améliorer les performances du *screening*.

### 3.3.2 Tests de la méthode

#### 3.3.2.1 Description de la méthode

Pour bien comprendre le travail qui a été effectué afin de pouvoir utiliser des données de conservation de structure secondaire dans l'étape de *screening*, il nous faut entrer dans plus de détails en ce qui concerne le fonctionnement du programme.

Comme décrit dans l'introduction, le *screening* est la dernière étape algorithmique de Match-Box. Pour rappel, son premier rôle est de constituer une banque de boîtes constituée de segments corrects et de segments incorrects. Une fois que cette banque est disponible, le deuxième rôle du *screening* est de trier les segments corrects des incorrects. Parmi les boîtes ainsi obtenues, la plus grande d'entre elles est considérée comme étant *à priori* la meilleure. L'algorithme cherche ensuite toutes les boîtes compatibles avec celle-ci. Parmi toutes les boîtes compatibles trouvées, le *screening* choisit la plus grande. Par la suite, de nouvelles boîtes compatibles avec cette dernière seront déterminées, la plus grande sera choisie et ainsi de suite.

Nous proposons dans ce travail une approche où le critère (score) utilisé pour sélectionner les différentes boîtes ne dépend plus uniquement, comme dans la version actuelle de Match-Box, de la longueur des différentes boîtes et de la somme des carrés des écarts (SCE) des différents décalages entre ces boîtes. En effet, nous

ajoutons à ce critère la conservation en structure secondaire prédite par PHD dans les diverses boîtes. Ce nouveau critère s'exprime par la formule suivante :

$$S = L^b \times SS^a \times \sqrt[c]{SCE\Delta_g} \quad (2)$$

avec :

S, le score

L, la longueur de la boîte

SS, la structure secondaire

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

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

Nous avons donc optimisé les valeurs des exposants a, b et c de manière à obtenir le meilleur couple confiance-puissance pour l'étape de *screening*. Comme expliqué dans l'étude de faisabilité, nous avons retiré le filtre du *matching* pour pouvoir optimiser au mieux les performances du *screening*.

Nous avons ensuite replacé le filtre du *matching* et testé les performances de Match-Box avec les paramètres optimaux dans le *screening*.

Tous les tests ont été effectués sur les 78 cas-tests en utilisant la matrice de scores Blosum 62, une taille de fenêtre d'analyse de 9 résidus, et les deux critères de vérité (ce qui est aligné sans gap ou les zones où le pourcentage de conservation de structure secondaire est de 100%).

### 3.3.2.2 Résultats

- Les meilleures performances du *screening* ont été obtenues avec des valeurs pour les exposants a, b et c de 4.50, 2.12 et -0.145, respectivement. Cela signifie qu'une boîte de 9 résidus dont la structure secondaire est conservée à 100% sera sélectionnée par le *screening* plutôt qu'une boîte de 10 résidus ayant une conservation de structures secondaires inférieure ou égale à 95%, ou même qu'une boîte de 11 résidus ayant une conservation de structures secondaires égale

*Tableau 3.3.2.2.1a: Résultats en confiance et en puissance entre le screening actuel (1 run) et le screening tenant compte des structures secondaires. Le critère de vérité est défini comme étant ce qui est aligné sans gap.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Screening actuel 1 run** | 59,8 | 66,5 |
| **Screening tenant compte des SS** | 65,7 | 69,9 |
| **Amélioration** | **5,9** | **3,4** |

*Tableau 3.3.2.2.1b: Résultats en confiance et en puissance entre le screening actuel (1 run) et le screening tenant compte des structures secondaires. Le critère de vérité est défini comme étant les zones où la conservation de la structure secondaire est de 100%.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Screening actuel 1 run** | 60,9 | 27,5 |
| **Screening tenant compte des SS** | 66,7 | 28,9 |
| **Amélioration** | **5,8** | **1,4** |

*Tableau 3.3.2.2.2a: Résultats en confiance et en puissance entre Match-Box (4 runs) et le screening tenant compte des structures secondaires dans MB 4 runs. Le critère de vérité est défini comme étant ce qui est aligné sans gap.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Match-Box 4 runs** | 63,1 | 69,6 |
| **Screening SS dans M.B 4 runs** | 63,7 | 69,6 |
| **Amélioration** | **0,6** | **0** |

*Tableau 3.3.2.2.2b: Résultats en confiance et en puissance entre Match-Box (4 runs) et le screening tenant compte des structures secondaires dans MB 4 runs. Le critère de vérité est défini comme étant les zones où la conservation de la structure secondaire est de 100%.*

| | Puissance (%) | Confiance (%) |
|---|---|---|
| **Match-Box 4 runs** | 64,9 | 28,9 |
| **Screening SS dans M.B 4 runs** | 65,5 | 28,9 |
| **Amélioration** | **0,6** | **0** |

ou inférieure à 91%, et ainsi de suite en fonction de la taille de la boîte *(voir tableau ci-dessous).*

| **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 du *screening* sur 78 cas-tests (en puissance et en confiance), lorsque nous tenons compte des structures secondaires, sont repris dans les *tableaux 3.3.2.2.1a et 3.3.2.2.1b* Les résultats obtenus expriment une nette amélioration de la confiance et de la puissance du *screening* lorsque celui-ci tient compte des structures secondaires, et ce quel que soit le critère de vérité utilisé *(figures 3.3.2.2.1 et 3.3.2.2.2).* Les données présentées dans ces deux figures ne donnent cependant qu'une information biaisée de l'amélioration réelle que nous venons d'obtenir en incluant les données de structures secondaires dans l'étape de *screening*. En effet, dans notre étude de faisabilité sur la possible amélioration du *screening* (cfr. point 3.1 et *tableaux 3.1.2.1 et 3.1.2.2*), nous montrions que celui-ci pouvait puiser dans le *matching* 84,6% de boîtes correctes pour ce qui est aligné sans gap comme critère de vérité et 83,4% de boîtes correctes pour l'autre critère de vérité (en utilisant une fenêtre de 9 résidus). Le *screening* prenant en compte les données de structures secondaires offre une puissance relative de 93% pour le premier critère de vérité et de 91,3% pour le deuxième critère. Soit une amélioration de 8,4% pour l'un et 7,9% pour l'autre. Ceci nous conduit à préciser que le fameux point de 100% d'informations correctes prise par le *screening* dans le *matching* n'est plus si éloigné que ça.

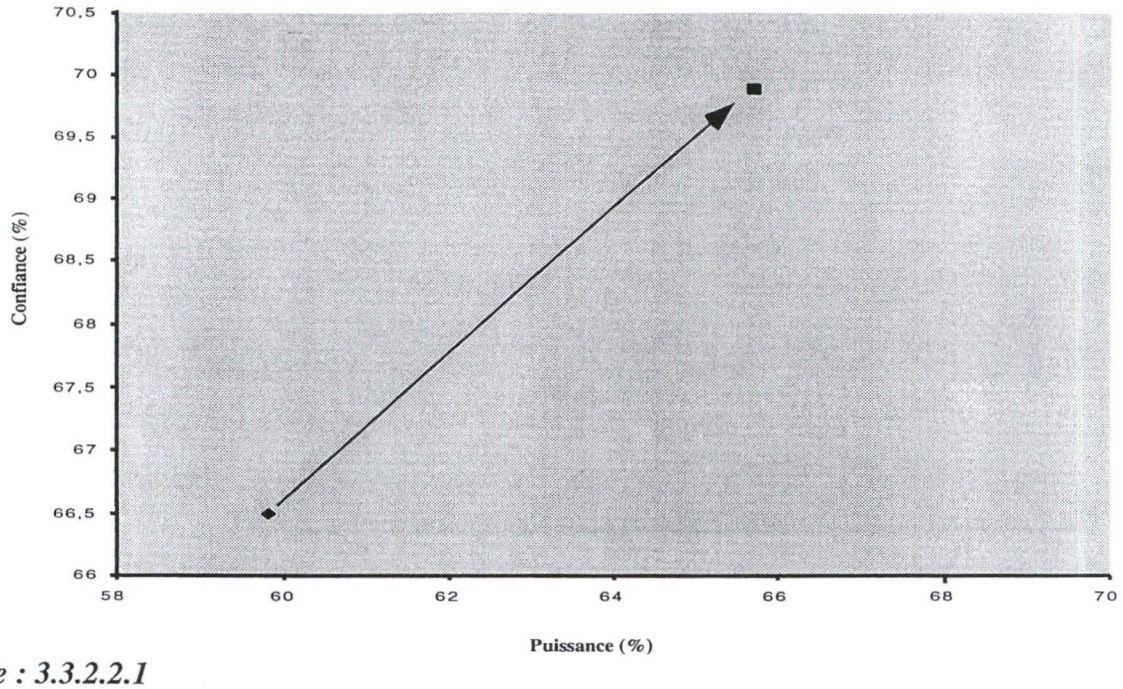



***Figure : 3.3.2.2.1***

*Comparaison des résultats confiance-puissance entre le screening actuel de Match-Box (losange noir) et le screening tenant compte des structures secondaires (carré noir).*
*Le critère de vérité choisit correspond à ce qui est aligné sans gap.*

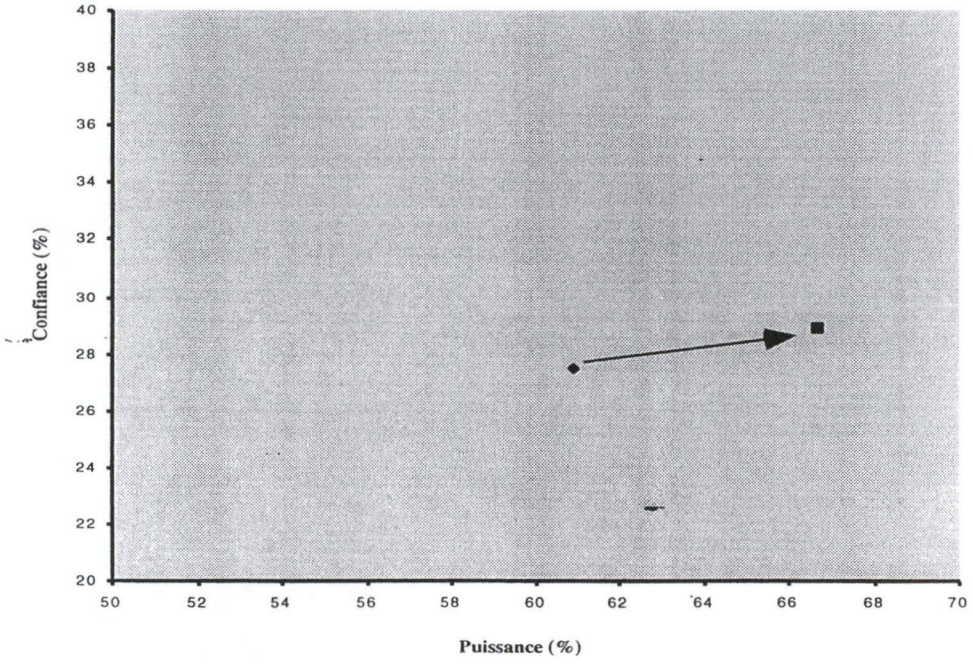



***Figure : 3.3.2.2.2***

*Comparaison des résultats confiance-puissance entre le screening actuel de Match-Box (losange noir) et le screening tenant compte des structures secondaires (carré noir).*
*Le critère de vérité correspond aux zones où la conservation de la structure secondaire est de 100%.*

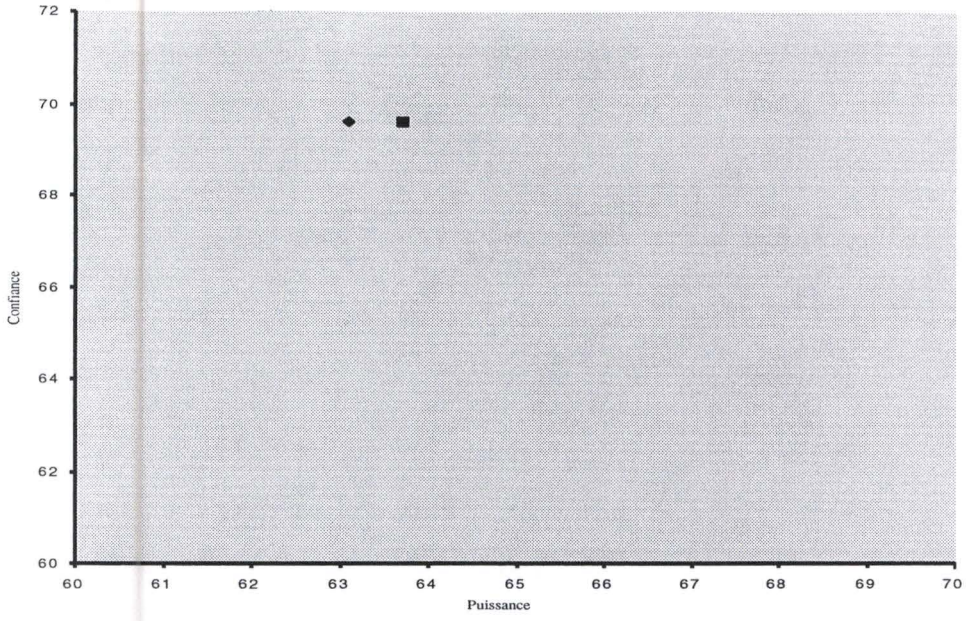



***Figure : 3.3.2.2.3***

*Comparaison des résultats confiance-puissance entre la version actuelle de Match-Box fonctionnant sur 4 runs (losange noir) et l'intégration du screening tenant compte des structures secondaires dans la version actuelle de Match-Box fonctionnant sur 4 runs (carré noir).*
*Le critère de vérité choisit correspond à ce qui est aligné sans gap.*

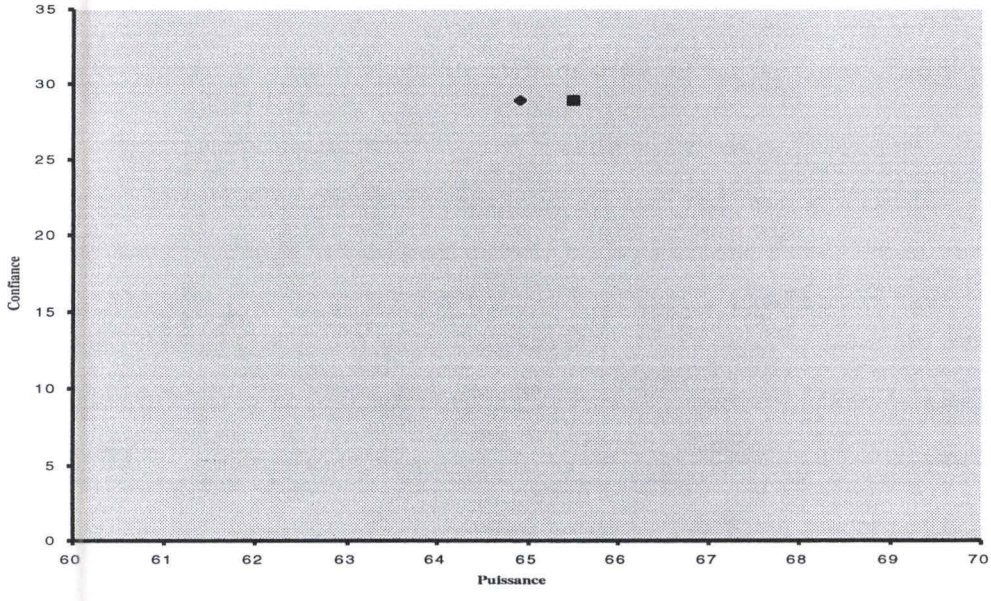



***Figure : 3.3.2.2.4***

*Comparaison des résultats confiance-puissance entre la version actuelle de Match-Box fonctionnant sur 4 runs (losange noir) et l'intégration du screening tenant compte des structures secondaires dans la version actuelle de Match-Box fonctionnant sur 4 runs (carré noir).*
*Le critère de vérité correspond aux zones où la conservation de la structure secondaire est de 100%.*

- L'intégration du *screening* tenant compte des données de structures secondaires dans la version actuelle de Match-Box, c'est-à-dire sur 4 runs et avec les filtres, fournit des résultats confiance-puissance plus élevés que Match-Box dans sa version actuelle (4 runs et filtre) *(tableaux 3.3.2.2.2a et 3.3.2.2.2b)*. Cependant, les valeurs confiance-puissance ne sont pas aussi élevées que celles obtenues lors de l'optimisation des paramètres du *screening* lorsque celui-ci fonctionne sur un seul run et à la suite d'un *matching* sans filtre *(figures 3.3.2.2.3 et 3.3.2.2.4)*.

### 3.3.2.3 Conclusions

- Le *screening* tenant compte des données de structures secondaires et fonctionnant sur un seul run donne de meilleurs résultats en confiance et en puissance que le *screening* de la version actuelle de Match-Box. Cela fait donc la deuxième fois, en tenant compte des résultats obtenus avec le nouveau *screening*, que nous augmentons les performances du *screening* en puissance ainsi qu'en confiance.
- Le *screening* ne fonctionne pas de manière optimale dans la version de Match-Box possédant un filtre au niveau du *matching*. Cependant, celui-ci permet quand même d'obtenir une puissance plus élevée tout en ne perdant pas en confiance. Il faut en outre remarquer que c'est la première fois qu'une augmentation en puissance ne s'accompagne pas d'une perte en confiance comme le postule la *figure 2.4.1*.
- Depuis tous les tests qui ont été effectués avant ce travail sur le *screening*, nous constatons que c'est aussi la première fois que les performances du *screening* sont améliorées en tenant compte du contenu des boîtes, c'est-à-dire de la conservation des structures secondaires.

# RÉSUMÉ, CONCLUSIONS ET PERSPECTIVES

Suite aux différents objectifs abordés plus haut, à savoir :

- L'élaboration d'une technique hybride (Match-Tal) reprenant deux logiciels d'alignement multiple (ClustalW et Match-Box), celle-ci ayant pour but d'améliorer la qualité de l'alignement en profitant de la puissance de ClustalW tout en gardant la confiance de Match-Box

et

- l'amélioration du dernier algorithme de Match-Box (le *screening*), en tentant d'implémenter une nouvelle stratégie de *screening* d'une part, et en tenant compte des structures secondaires dans le *screening* d'autre part,

Nous pouvons écrire les conclusions générales suivantes :

- Les résultats obtenus en testant les 33 cas-tests sur la technique hybride « Match-Tal » n'a pas permis l'amélioration de la qualité de l'alignement. En effet, nous ne profitons pas d'une part de la puissance du logiciel ClustalW, mais de plus nous perdons la confiance du logiciel Match-Box. Par contre, une perspective toujours à l'état embryonnaire fut testée. Celle-ci vise à déterminer *a priori* la meilleure méthode à utiliser (Match-Box, ClustalW ou Match-Tal) pour chacun des cas-tests. Les résultats de cette technique sont très parlants. En effet, les performances obtenues sur les 33 cas–tests améliore de façon non négligeable la qualité de l'alignement au niveau sa confiance et de sa puissance. Le développement d'une technique permettant la détermination *a priori* de la méthode d'alignement à utiliser n'a pas pu être réalisé durant ce mémoire mais reste cependant une perspective très prometteuse.
- L'élaboration du nouveau *screening* (fonctionnant sur un seul run) offre de manière générale une confiance et une puissance plus élevées que le *screening* utilisé dans la version actuelle de Match-Box, et ce quel que soit le critère de vérité choisi. De plus, les résultats obtenus par un seul run du nouveau *screening* sont très proches de ceux obtenus avec la version actuelle de Match-Box lorsque celui-ci fonctionne sur 4 runs. Le test impliquant une variation de la taille de la fenêtre d'analyse dans le nouveau *screening* montre des performances optimales lorsque celle-ci est égale à 9 résidus, et ce quel que soit le critère de vérité choisi.

Les améliorations à introduire dans ce nouveau *screening* seraient d'une part qu'il puisse fonctionner en plusieurs runs comme dans la version actuelle de Match-Box et d'autre part que l'on puisse réduire son temps de travail par optimisation de l'algorithme.

- La prise en compte des structures secondaires dans le *screening* lorsque celui-ci fonctionne sur un seul run améliore considérablement les performances de l'alignement. En effet, la confiance et la puissance obtenues par celui-ci sont significativement plus élevées que celles obtenues par le *screening* de la version actuelle de Match-Box lorsque celui-ci fonctionne aussi sur un run. Par contre, ce *screening* voit ses performances diminuées lorsqu'il se trouve dans la version actuelle de Match-Box c'est-à-dire ayant un *matching* disposant de ses filtres. Une perspective à cette technique serait d'améliorer la quantité d'informations correctes que fournit l'étape précédant le *screening* c'est-à-dire, le *matching*.

# BIBLIOGRAPHIE

Arnold, E. and Rossmann, M. G. (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:** 763-801.

Bashford, D., Chothia, C. and Lesk, A. M. (1987) Determinants of a protein fold. Unique features of globin amino acid sequences. *J. Mol. Biol.* **196:** 199-216.

Blundell, T. L., Cooper, J. B., Sali, A. and Zhu, Z. Y. (1991) Comparison of the sequences, 3-D structures and mechanisms of pepsin-like and retroviral aspartic proteinases. *Adv. Exp. Med. Biol.* **306:** 443-453.

Boguski, M. S. (1998) Bioinformatics, a new era. *Trends guide to bioinformatics*: 1-3.

Brenner, S. E., Chothia, C., Hubbard, T. J. and Murzin, A. G. (1996) Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* **266:** 635-43.

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

Carrington, J. C., Morris, T. J., Stockley, P. G. and Harrison, S. C. (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:** 265-276.

Cohen, F. E., Novotny, J., Sternberg, M. J., Campbell, D. G. and Williams, A. F. (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:** 31-40.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16:** 10881-90.

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

Depiereux, E., Baudoux, G., Briffeuil, P., Reginster, I., De Bolle, X., Vinals, C. and Feytmans, E. (1997) Match-Box server: a multiple sequence alignment tool placing emphasis on reliability. *Comput. Appl. Biosci.* **13:** 249-256.

Depiereux, E. and Feytmans, E. (1991) Simultaneous and multivariate alignment of protein sequences: correspondence between physiochemical profiles and structurally conserved regions (SCR). *Protein Engineering* **4:** 603-613.

Depiereux, E. and Feytmans, E. (1992) Match-Box: a fundamently new algorithm for simultaneous alignment of several protein sequences. *CABIOS* **8:** 501-509.

Ding, D. F., Qian, J. and Feng, Z. K. (1994) A differential geometric treatment of protein structure comparison. *Bull Math Biol.* **56:** 923-43.

Feng, D. F. and Doolittle, R. F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25:** 351-60.

Flower, D. R., North, A. C. and Attwood, T. K. (1993) Structure and sequence relationships in the lipocalins and related proteins. *Protein Sci.* **2:** 753-761.

Hanks, S. K., Quinn, A. M. and Hunter, T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241:** 42-52.

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

Henikoff, S. and Henikoff, J. G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17:** 49-61.

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

Johnson, M. S., May, A. C. W., Rodionov, M. A. and Overington, J. P. (1996) Discrimination of common protein folds: application of protein structure to sequence/structure comparaisons. *Methods Enzymol.* **266**: 577-598.

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

Johnson, M. S., Sali, A. and Blundell, T. L. *This series* **183:** 670.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-637.

Lawrence, E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wooton, J. C. (1993) Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* **262:** 208-214.

Lee, B. and Richards, F. M. (1991) *J. Mol. Biol* **55:** 379.

Lodish, Baltimore, Berk, Zipursky, Matsudaira and Darnell (1997). Structure et fonction des protéines in Université, D. B. (Ed), *Biologie moléculaire de la cellule*, pp. 51-100.

Mazet, L. (1996) Fortran 77 Langage d'un autre âge. *http://www-syscom.univ-mlv.fr/~mazet/fortran/fortran.html*.

Murata, M., Richardson, J. S. and Sussman, J. S. (1985) Simultaneous comparaison ofthree protein sequences. *Proc. Natl. Acad. Sci.* **82:** 3073-3077.

Neuwald, A. F., Liu, J. S., Lipman, D. J. and Lawrence, C. E. (1997) Extracting protein alignement models from the sequence database. *Nucleic Acids Res.* **25:** 1665-1677.

Overington, J., Donnelly, D., Johnon, M. S., Sali, A. and Blundell, T. L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1:** 216-226.

Rost, B. and Sander, C. (1993) Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232:** 584-599.

Rutenber, E., Ready, M. and Robertus, J. D. (1987) Structure and evolution of ricin B chain. *Nature* **326:** 624-6.

Sali, A. and Blundell, T. L. (1990) *J. Mol. Biol.* **212:** 403.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. and Abola, E. E. (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. D. Biol. Crystallogr.* **54:** 1078-84.

Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L. (1987) *Protein Eng.* **1:** 377.

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