MASTER EN SCIENCES BIOLOGIQUES DES ORGANISMES ET ÉCOLOGIE

Etude de la sensibilite et de la spécificité des appariements de segments de séquences protéiques en vue de l'amélioration des performances d'un alignement multiple

Van Campenhout, Jean-Marc

Award date:
1999

FACULTES UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR

Faculté des Sciences

ETUDE DE LA SENSIBILITE ET DE LA SPECIFICITE DES APPARIEMENTS DE SEGMENTS DE SEQUENCES PROTEIQUES EN VUE DE L'AMELIORATION DES PERFORMANCES D'UN ALIGNEMENT MULTIPLE

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

Jean-Marc VAN CAMPENHOUT

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

Etude de la sensibilité et de la spécificité des appariements de segments de séquences protéiques en vue de l'amélioration des performances d'un d'alignement multiple

VAN CAMPENHOUT Jean-Marc

Résumé

Au cours de ce travail, nous avons tenté d'améliorer les performances de Match-Box, un programme d'alignement multiple de séquences protéiques. Notre investigation s'est principalement concentrée dans l'étape du *matching* qui réalise l'appariement des segments similaires communs au jeu de protéines, dans le but de délimiter les segments conservés structurellement. Au cours de cette étape, l'algorithme fait référence à une matrice de similarité pour pondérer le remplacement d'un résidu par un autre dans une paire alignée. Ce mémoire trouve son point de départ au sein même d'un paradoxe puisque jusqu'à présent, la plupart des méthodes d'alignement ont toujours été limitées à l'usage simultané d'une seule de ces matrices, alors que la pondération d'un acide aminé par un autre est forcément différente suivant son emplacement dans la protéine repliée (dans une hélice exposée, dans un brin enfoui ou dans un coude intégré...). Nous avons testé les performances de 134 de matrices de similarité sur base d'un jeu de 78 familles de protéines de référence, dont l'alignement est considéré comme correct. Les performances de ce programme ont été évaluées selon deux critères : la puissance et la confiance. Nous avons réalisé une étude de faisabilité pour estimer les limites d'efficacité maximales que Match-Box pourrait atteindre. Notre challenge a été d'associer des caractéristiques de séquences protéiques aux matrices, en vue d'implémenter au sein du logiciel, le choix de la matrice de similarité la mieux adaptée localement, conférant à Match-Box les meilleures performances.

Mémoire de licence en Sciences biologiques

Juin 1999

Promoteur: E. Depiereux

Des mots pour remercier...Si peu de choses à côté de toutes celles que j'ai vécues, et ce grâce à vous tous, professeurs, chercheurs, amis et parents, vous à qui je dois tout...

A l'issue de ce travail, je tiens à remercier chaleureusement le Professeur Eric Depiereux pour m'avoir ouvert les portes de son laboratoire et m'avoir guidé patiemment tout au long de mon mémoire. Son expérience et son savoir-faire dans de nombreux domaines m'ont beaucoup aidé.

Toute ma gratitude revient à Christophe Lambert qui m'a permis de réaliser ce travail par son aide précieuse, sa gentillesse et ses talents de programmeur.

Mes remerciements vont également à Bernard, Carlos, Christophe, Etienne, Isabelle, Katalin, Nicolas, Philippe, Thynna pour l'excellente ambiance qu'ils font régner dans le laboratoire.

Mes remerciements s'adressent tout particulièrement à B. Masereel, M. Rooman, D. Vercauteren et S. Wattiaux qui ont accepté de lire ce mémoire.

Enfin, je tiens à associer à ce travail, ma famille et tout spécialement mes parents que je remercie de tout cœur pour m'avoir permis de réaliser ces études et m'avoir toujours manifesté leur profond intérêt durant ces années.

Table des matières

Chapitre I. Introduction *4*

I.1. Contexte biologique de l'alignement de séquences **4**

I.2. Les différents niveaux de structures protéiques **7**

- Structure primaire : 8
- Structure secondaire : 8
  - L'hélice α : 9
  - Le plan β : 11
  - Les Coils : 12
- Les super-structures secondaires 13
- Les structures tertiaires et quaternaires 13
- Prédiction de structure secondaire : 14
- Modélisation par comparaison de séquence : 15

I.3. Les matrices de scores **16**

I.4. Méthode d'alignement de deux séquences **22**

I.5. Méthode d'alignement multiple **26**

I.6. Logiciels d'alignement de séquences **28**

- ClustalW 28
- BlockMaker : 29
- Gibbs : 30
- Match-Box : 31
  - Étape de paramétrisation, scanning. 33
  - Étape de constitution de boîte, le matching 33
  - Étape de triage, le screening 34
- Concept de puissance-confiance 35
- Comparaison de ClustalW et de Match-Box : 36

Chapitre II. Matériels et outils bioinformatiques *38*

II.1 Support informatique **38**

- Station de travail Silicon Graphics Octane : 38

ABREVIATIONS

ClW	ClustalW
DSSP	Dictionary of Secondary Structure Protein
MB	Match-Box
MHz	Mega Hertz
Mips	Multy initial processors
MSP	Maximum Segment Pair
PAM	Point Accepted Mutation
PDB	Protein Data Bank
RAM	Random Accessible Memory
RMS	Root Means Square
SAPS	Statistical Analysis Protein Sequences
SAS	Statistical Analysis System
SCR	structurally conserved regions
SMTP	Simple Mail Transfer Protocol
SS	Structure Secondaire
TCP/IP	Transmission Control Protocol/Internet Protocol
VR	Variables Regions
WWW	World Wide Web

- Serveur Digital Alpha 4400: 38
- Langage de programmation : 39
- Réseau Internet : 39
- PDB (Banque de structures) : 40
- DSSP : 41
- PHD : 41
- ALIGN : 43
- SAPS : 43
- MWCALC : 43
- SAS : 44
- STATISTICA : 44

II.2 Support protéique 45

- Les cas-tests : 45

II.3 Les matrices de score 46

Chapitre III. Objectif du mémoire 47

Chapitre IV. Recherche du meilleur incrément pour une batterie de matrices de scores 48

IV.1 Méthode 48

IV.2 Conclusion 51

Chapitre V. Etude de faisabilité sur les performances de Match-Box en changeant de matrices de scores en fonction du cas-test 53

V.1. Sur trente-trois cas-tests 53

- méthode 53
- Conclusion 56

V.2. Développement d'une banque de septante-huit cas-tests 56

V.3. Travail réalisé sur septante-huit cas-tests : 59

- Méthode 59
- Conclusion 62

Chapitre VI. Caractérisation des cas tests 63

VI.1. Méthode 63

Procédure multivariée 65

VI.3 Conclusion 66

Chapitre VII. Caractérisation de fenêtres *68*

VII.1 Méthode **68**

VII.2. Test réalisé avec 42 matrices de scores différentes **71**

VII.3. Test réalisé avec 134 matrices de scores différentes : **74**

VII.4. Implémentation des observations. **77**

Chapitre VIII. Conclusions générales et perspectives *80*

Chapitre IX. Bibliographie *83*

Chapitre I. Introduction

I.1. Contexte biologique de l'alignement de séquences

Avec le rôle déterminant qu'occupe désormais l'informatique dans la recherche, il est indispensable de procurer aux biologistes l'accès aux banques de données les plus à jour et un choix de programmes d'analyse le plus vaste possible. Depuis la dernière décennie, ces données et ces programmes constituent un environnement de travail qui devient chaque jour plus complexe et plus volumineux. La taille déjà considérable de ces banques double encore chaque année, ce qui impose le recours à des méthodes perfectionnées et à des moyens techniques importants pour les exploiter. Ce boom est largement dû aux nombreux génomes qui sont complètement séquencés *(E. coli, S. cerevisae, Arabidopsis thaliana...)*. On prévoit, pour 2005, le séquençage complet du génome humain : loin d'être l'aboutissement d'une recherche, cette base de donnée sera le point de départ de bien d'autres questions.

Permettre aux chercheurs d'utiliser l'environnement de logiciels et de données, suppose une activité de mise à jour permanente consacrée tant aux outils informatiques en général qu'aux logiciels biologiques et aux travaux théoriques qui en constituent le fondement. Pour exploiter efficacement cette ressource, nous assistons à l'émergence d'une science nouvelle joignant la biologie à l'informatique : « la bioinformatique ». Ce terme évoqué pour la première fois en 1991, peut être défini comme la recherche liée à l'exploitation des bases de données (Boguski, 1998). Cette véritable science concerne tous les champs de la biologie et permet ainsi aux biologistes de mobiliser rapidement et aisément l'ensemble des connaissances souhaitées. La quantité d'information est telle qu'il est pratiquement impossible de tout réunir sur un site

propre. C'est pourquoi Internet devient le support idéal pour un réseau de communication rapide entre les biologistes du monde entier. La création de logiciels destinés à la biologie moléculaire n'est pas achevée : c'est un domaine de recherche fleurissant. La recherche de nouveaux algorithmes et le développement des logiciels qui les mettent en œuvre sont des activités indispensables. Il faut également développer pour ces logiciels des interfaces qui les rendront utilisables par un biologiste non informaticien.

En soi, la gestion des banques représente le premier problème essentiel. À l'heure actuelle, les fragments séquencés sont accumulés dans les banques sans que l'on parvienne réellement à assurer la fiabilité des données : les erreurs de séquençage affectent 5 % des protéines, mais surtout, les annotations qui s'y trouvent sont inférées par homologie entre séquences.

Dans ce procédé, il n'est pas possible de savoir si la protéine de référence est validée par une évidence expérimentale ou, si elle est elle-même annotée par similarité. La similarité n'étant pas transitive, une dérive rapide peut affecter fortement la fiabilité des annotations. L'utilisation efficace de ces banques est donc compromise et nécessite une expérience considérable, d'où l'intérêt de développer des logiciels efficaces pour en tirer le meilleur parti.

On peut se demander comment il est possible de déterminer la structure d'une protéine à partir de sa séquence. Cette idée trouve son origine dans l'observation de Richardson, qui observe que les 400 structures connues par cristallographie aux rayons X, ne sont pas fondamentalement différentes (Richardson, 1981). Donc des séquences très éloignées peuvent correspondre à des structures très proches. Par contre, jamais encore on a observé deux protéines présentant un grand nombre d'identités, mais ayant des structures fondamentalement différentes (Blundell *et al.*, 1987). En sachant cela, si une protéine de structure connue présente un pourcentage d'identité élevé avec une protéine de structure inconnue, on peut imaginer que ces dernières sont structurellement similaires.

La connaissance de la structure d'une protéine ne permet pas en soi de préciser sa fonction. C'est toutefois un élément d'information important car il permet de comparer cette structure avec des éléments structuraux appartenant à d'autres protéines, ou de proposer très précisément des expériences destinées à modifier la fonction, donc à pouvoir observer la perturbation si phénotype il y a.

L'alignement de séquences est une étape cruciale dans la recherche de fonction d'une protéine inconnue : il est impliqué dès le départ dans la recherche de protéines similaires dans les banques. Le rôle des alignements multiples de séquences protéiques est ensuite de délimiter des régions où le niveau de similarité est suffisant pour inférer que ces régions sont sans doute structurellement et donc fonctionnellement conservées. Ces résultats fournissent enfin des paramètres essentiels aux méthodes de prédiction de structure et permettent de cibler les régions cruciales pour l'expérimentation (clonage, mutagenèse dirigée).

En règle générale, lorsque deux séquences présentent au moins 25 % d'identité, elles ont alors une très forte probabilité d'adopter la même conformation générale (Doolittle, 1981). Certains acides aminés non identiques sont plus proches entre eux que d'autres, et leur remplacement est mieux toléré. Il y a donc des séquences homologues qui présentent des résidus physico-chimiquement similaires qui se substituent plus souvent les uns aux autres. (ex: Lys-Arg , Asp-Glu , Ser-Thr...) La notion de similarité entre les différents acides aminés est exprimée par diverses échelles de scores établies par plusieurs auteurs et qui expriment la ressemblance ou la différence entre ces acides aminés sur base de leurs propriétés physico-chimiques ou les variations de résidus observés au cours de l'évolution ou encore sur base de caractéristiques plus structurelles. Ces résultats sont placés dans un tableau appelé matrice de scores, qui pondère chaque paire d'acides aminés dans un alignement. En règle générale, lorsque deux séquences présentent au moins 25 % d'identité, elles ont alors une très forte probabilité d'adopter la même conformation générale (Doolittle,

1981). Si la plupart des acides aminés sont identiques, l'alignement de séquences est assez trivial et peut même être réalisé visuellement. Par contre quand deux séquences de protéines n'ont pas exactement la même longueur, leur alignement exige l'introduction d'un espacement dans au moins une des deux séquences. Ici se présente un problème que l'on pourrait intituler : « *le paradoxe d'insertions et de délétions* ». En effet, ces « *indels* », également appelés *gaps*, permettent dans un alignement de disposer face à face les zones les plus similaires. Cependant l'utilisation d'*indels,* introduit dans l'alignement un paramètre arbitraire qui n'a aucun fondement biologique ou même physico-chimique. Le fait de placer des *indels* augmente le pourcentage d'identité entre deux séquences choisies au hasard. Prenons l'exemple de deux séquences dans lesquelles on insère beaucoup d'*indels*. Ces deux séquences vont présenter un pourcentage d'identité très élevé sans pour autant être homologues. Bon nombre de méthodes d'alignement, notamment celle de programmation dynamique, utilisent le poids des *indels* comme paramètre pour réaliser ces alignements. Or les *indels* sont aussi le résultat attendu d'un alignement pour que les zones très similaires s'apparient et créent ainsi des zones vides. Un programme d'alignement devra justement détecter ces régions suffisamment similaires.

I.2. Les différents niveaux de structures protéiques

Il est accepté, à l'heure actuelle, que la structure d'une protéine est responsable de sa fonction. Il est donc essentiel de connaître cette structure. On peut déterminer cette structure expérimentalement (diffraction aux rayons X, R.M.N.) ou encore théoriquement (modélisation par homologie (Vinals *et al.*, 1995) L'inconvénient majeur est que la similarité des structures n'est pas toujours détectable en terme d'homologie de séquence. Pour éviter cela, prédire les structures secondaires d'une protéine à partir de sa séquence semble se présenter comme une bonne alternative.

Décrivons brièvement les différents niveaux de structures des protéines ainsi que les types de structures secondaires principalement utilisées dans ce travail.

- **Structure primaire :**

Les protéines sont des chaînes non ramifiées d'acides aminés unis par un lien covalent appelé liaison peptidique. Ces longs enchaînements linéaires sont constitués sur base de vingt types d'acides aminés qui diffèrent par la nature de leur chaîne latérale . En effet, chaque acide aminé possède une base commune (c'est-à-dire un carbone α tétraédrique (Cα) lié à un groupe aminé (-NH_2), à un groupe carboxyle (-COOH),et à un groupe hydrogène (-H)) et une partie variable (-R), chaîne latérale. La structure primaire est l'ordre des acides aminés dans la séquence. Connaissant seulement la séquence d'une protéine, nous en savons en fait très peu à son sujet. A ce niveau de structure, nous ne savons par exemple encore rien de sa conformation spatiale qui conférera à la molécule ses propriétés spécifiques. Certains essais prometteurs essaient également de prédire ce repliement par des potentiels dérivés de structures connues. Toutefois, des indications fonctionnelles peuvent venir suite à un alignement multiple entre une séquence primaire d'une protéine inconnue et des séquences primaires de protéines connues.

- **Structure secondaire :**

L'enchaînement des résidus entraîne de nombreuses interactions entre les chaînes latérales. La protéine doit donc trouver un état stable en réalisant des rotations autour de liaisons laissées libres. Ces relations sont définies par des angles de torsions, tels que ω pour la liaison peptidique C—N, ϕ pour la liaison N—Cα et ψ

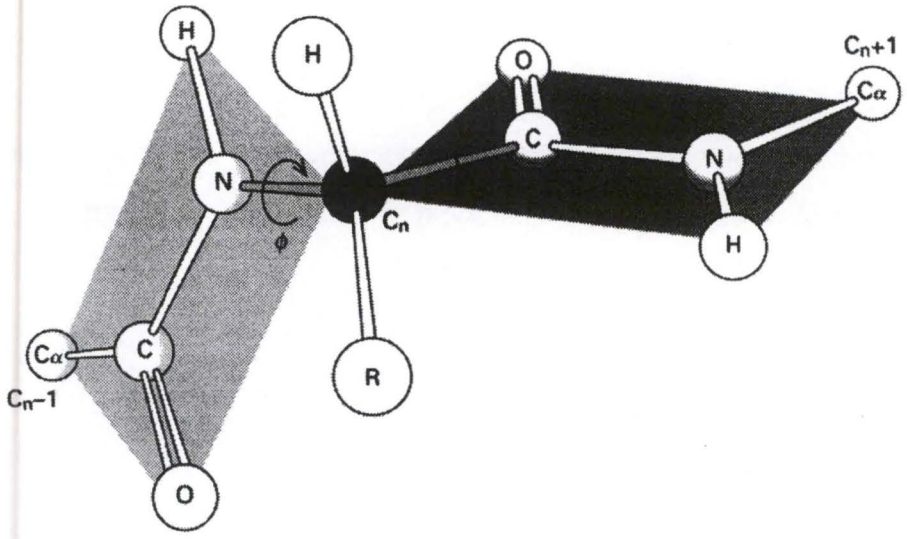


Figure.I.2.1. Localisation des différents angles de torsion le long de la chaîne peptidique.

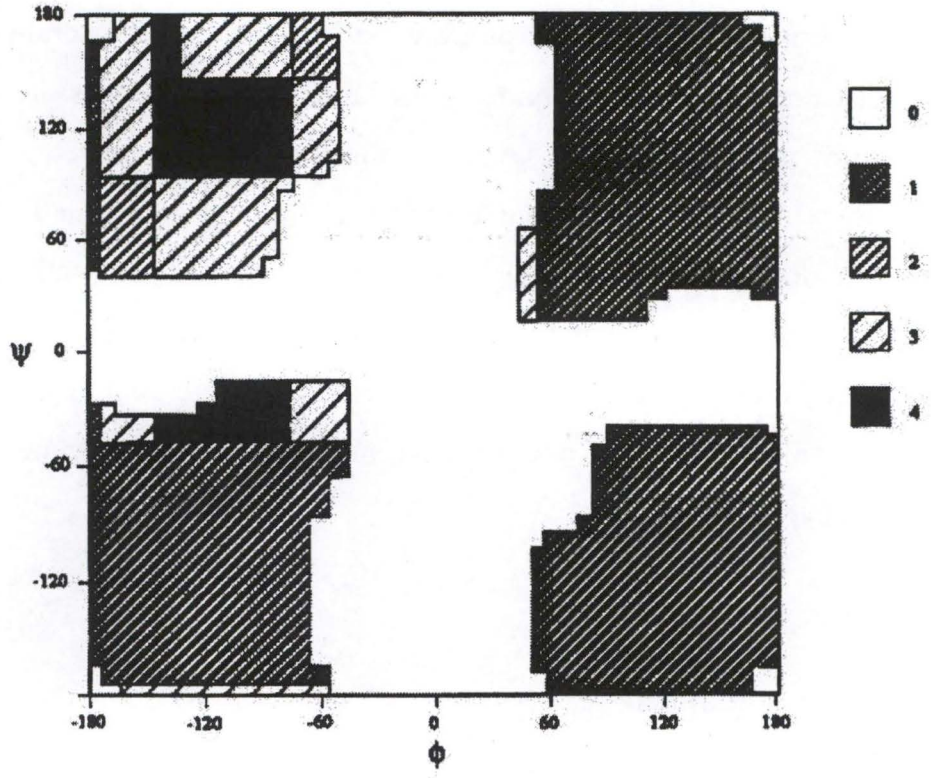


Figure I.2.2. Représentation d'un graphe de Ramachandran indiquant la répartition théorique des angles de torsion ϕ et ψ pour les acides aminé. Gly occupe les régions de 1 à 4 – Ala de 2 à 4 – Val et Ile ne se retrouvent que dans la région 4. Les autres acides aminés possédant une grande chaînes se situent tous dans les régions 3 et 4.

pour le lien Cα—C. Les angles des chaînes latérales sont quant à eux désignés $X_{(j)}$, où l'indice $_{(j)}$ indique la position de la liaison par rapport à la chaîne principale. Les valeurs prisent par ces angles, sont limitées *(cfr. Figure I.2.1)* :

* ω : seulement deux possibilités conformationnelles nous sont offertes, l'une en *cis* (angle = 0°), l'autre en *trans* (angle = 180°). On observe un rapport des conformations *trans/cis* de plus ou moins mille. Ceci est peut-être expliqué par la délocalisation de la charge de l'azote sur l'oxygène qui stabilise la conformation *trans* en plus la proximité du carbone asymétrique et des chaînes latérales défavorable en *cis* (sauf pour la proline).

* **ϕ et ψ** : toutes les valeurs ne sont pas possibles à cause de l'encombrement stérique. Les valeurs permises pour ϕ et ψ peuvent être rapportées sur un diagramme de dispersion appelé plot de Ramachandran. Sur ce graphe, on peut facilement visualiser des zones dans lesquelles se regroupent les valeurs permises des angles caractéristiques des différentes structures secondaires *(cfr. Figure 1.2.2 & 1.2.3)*.

Nous allons nous focaliser sur trois types de structures secondaires. Ce sont des arrangements locaux qui sont favorisés par les interactions entre les chaînes latérales. Chaque acide aminé a, selon la nature de son radical, une préférence pour l'une ou l'autre structure secondaire, soit dans notre cas l'hélice α, le feuillet β et les coils.

– *L'hélice α :*

Une structure de type hélice α fut pour la première fois envisagée par Pauling et Correy (1975). À l'heure actuelle, c'est sans doute l'élément de structure le plus

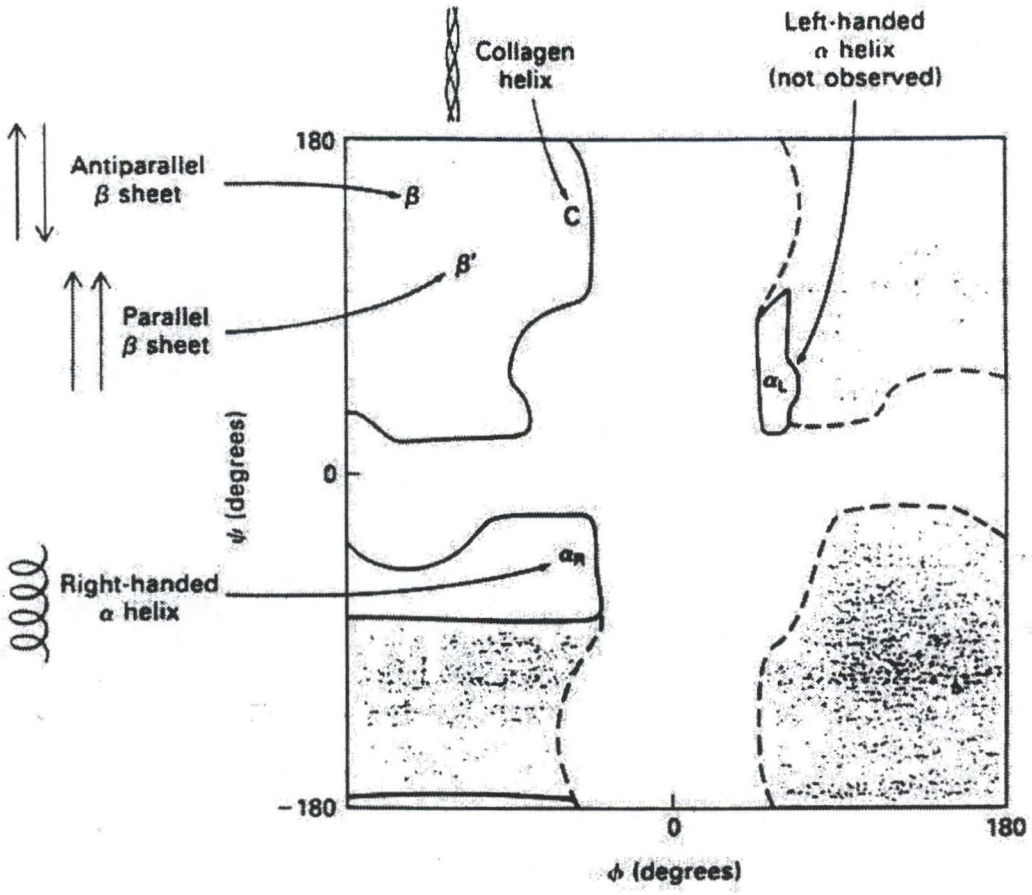


Figure I.2.3. Représentation de la localisation des valeurs de ϕ et ψ observées dans les différentes structure secondaires régulières.

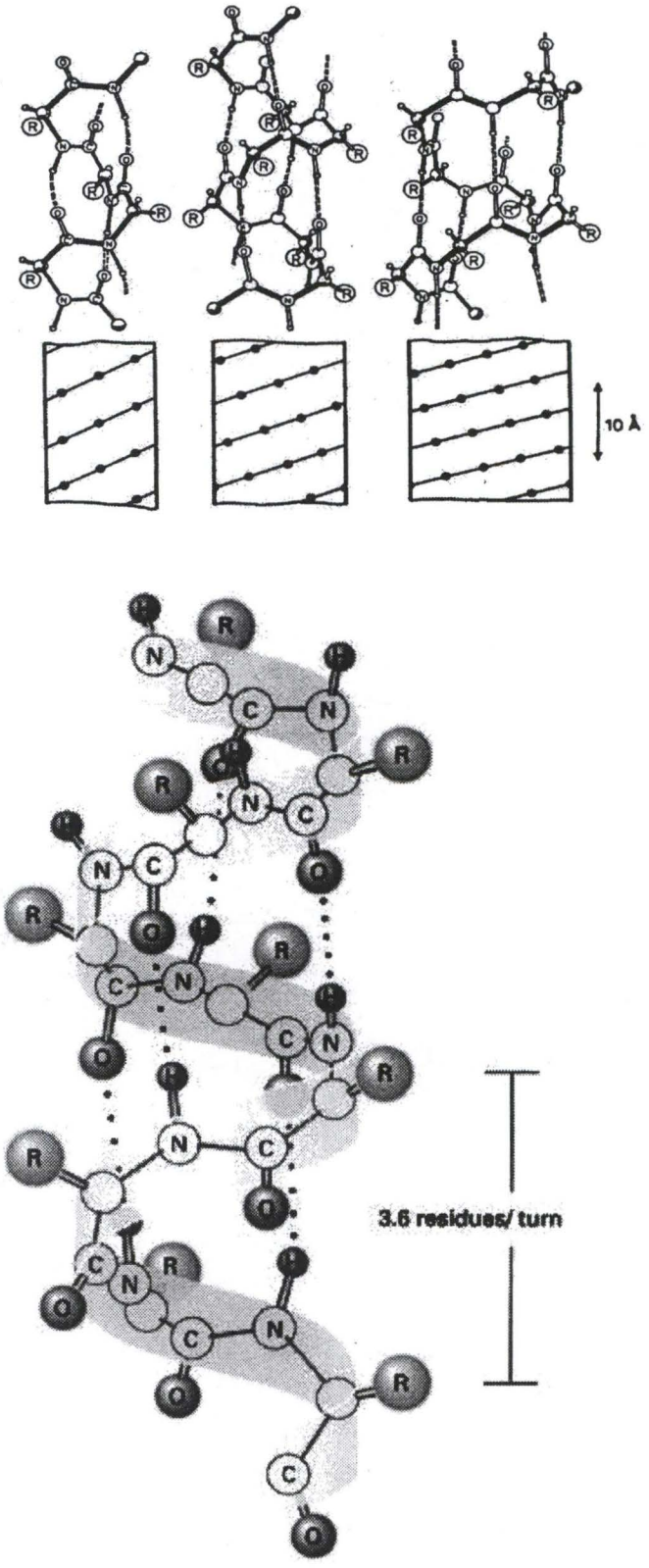


Figure I.2.4. Représentation shématique des trois types d'hélice (3_{10} - $3,6_{13}$ - l'hélice α_{π}) Seconde représentation de l'hélice α $(3,6_{13})$ classique.

classique et le mieux connu. Une hélice peut-être de pas gauche ou de pas droit. Dans les structures protéiques, seules les hélices α droites sont présentes car elles sont plus stables que les gauches, ceci étant dû à une composition en acides aminés L (lévogyre). Dans un squelette polypeptidique enroulé en hélice, chaque oxygène des groupes carbonyles est en liaison hydrogène avec l'hydrogène de l'amide appartenant au quatrième résidu situé en aval (c'est-à-dire du côté C- terminal). Ces liaisons hydrogène sont quasiment parallèles à l'axe longitudinal de l'hélice. Par contre les N, H et O de la liaison hydrogène sont presque colinéaires. Les résidus sont distants de 0,15 nm le long de l'axe de l'hélice. Un tour de spire complet compte 3,6 résidus, et représente 0,54 nm sur l'axe de l'hélice, C'est cette distance qu'on appelle « pas » de l'hélice. Dans ce genre de structure, on observe une plus grande fréquence de certains types d'acides aminés, plus aptes à former une telle conformation. Pour les hélices ce sont par exemple, l'alanine, le glutamate et la leucine. D'autres, comme la proline (dû à son noyau pyrrolidone pentagonal ne pouvant pas se plier aux exigences de la conformation générale), la glycine, la tyrosine et la sérine y sont très rarement présentés. Ce phénomène est notamment à la base de la prédiction de structure secondaire à partir de la séquence. Si, dans une région de la séquence, il y a prédominance de résidus polaires, on sait déjà que l'on ne se trouve pas dans une hélice, car ces résidus auraient tendance à se repousser et à déstabiliser l'hélice α .

Il existe d'autres types moins fréquents d'hélices *(cfr. figure 1.2.4)* :

L'hélice 3_{10} plus fine et plus compacte (3 résidus par tour, ponts H entre les résidus n et n+3) possède des valeurs de ϕ et ψ égales respectivement à –60° et –30°. Cette structure se retrouve assez régulièrement, sous forme de petites hélices 3_{10} existant dans de nombreux cas aux extrémités C terminales des hélices α, mais ne sont jamais très longues. Deux résidus également en conformation 3_{10} forment un coude. L'hélice α_{π} est plus lâche et est stabilisée par des ponts H entre les résidus n et n+5 Cette hélice, n'a jamais encore été observée dans les protéines (Lodish, 1997).

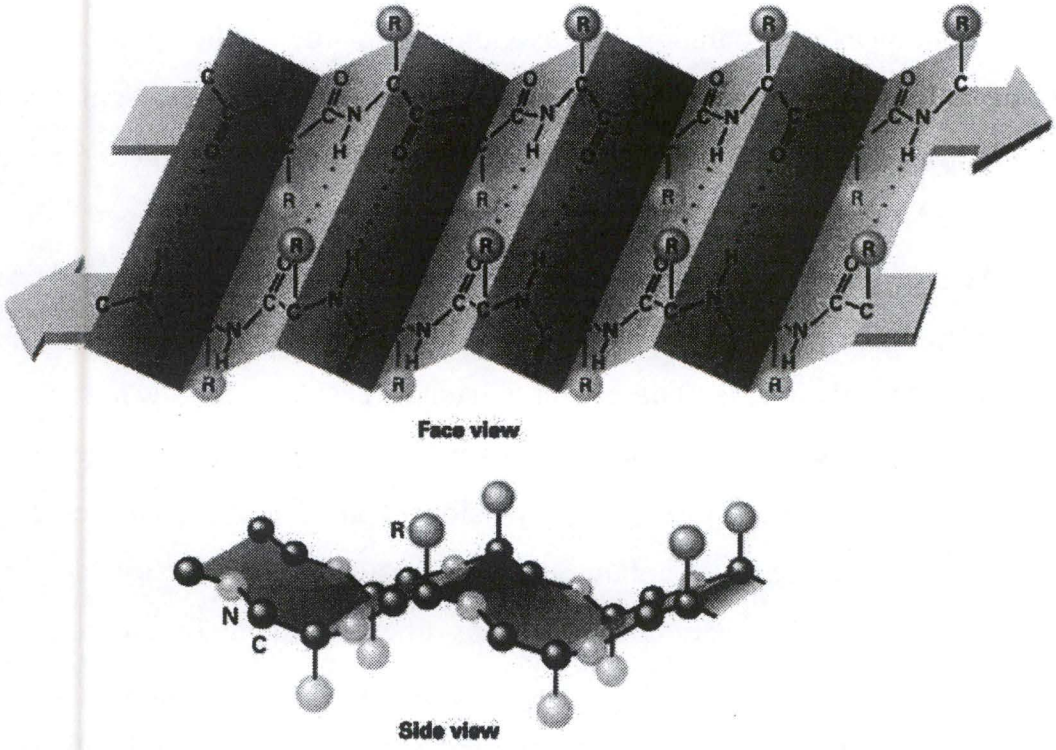
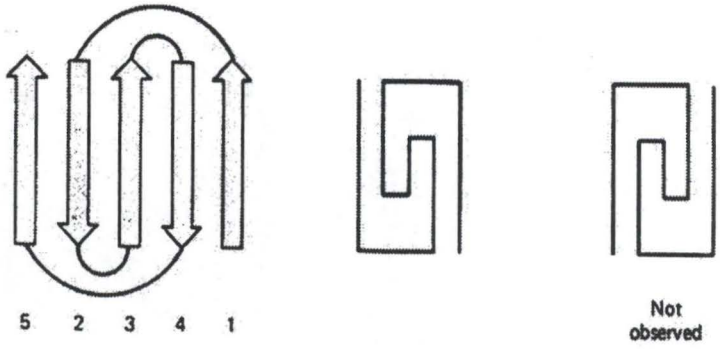
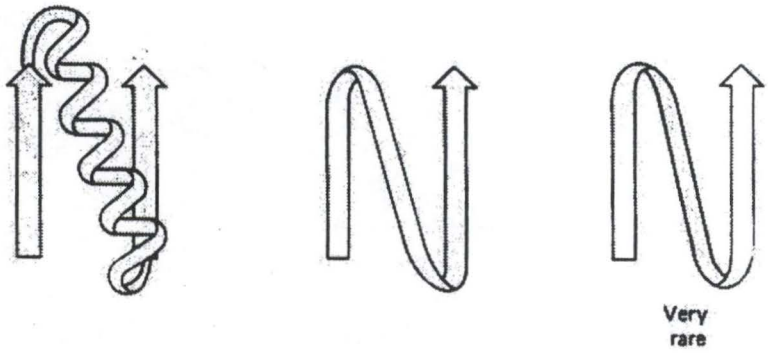


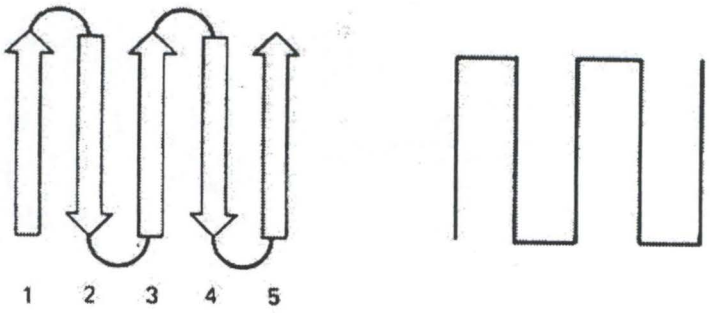
Figure I.2.5. Représentation schématique d'un feuillet β antiparallèle.



Représentation schématique du motif clef grecque

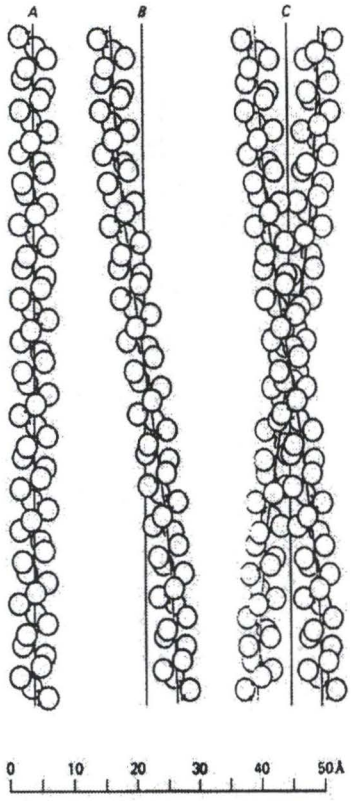


Représentation schématique du motif βαβ

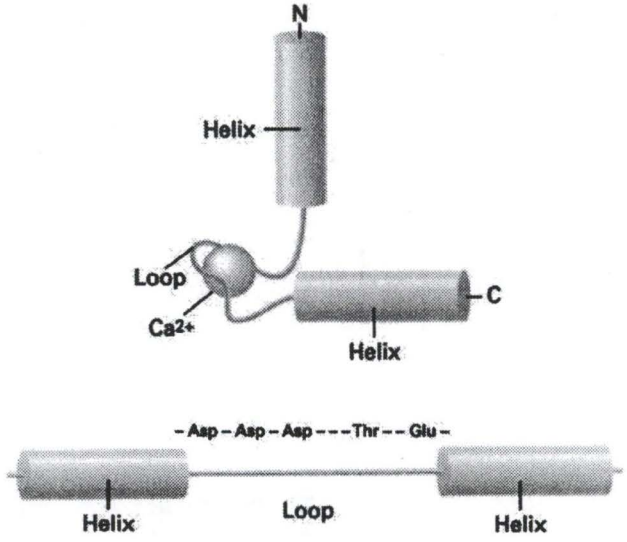


Représentation schématique du motif ββ

Figures I.2.6. Ensemble de représentations schématiques de quelques super-structures secondaires rencontrées dans les protéines repliées.



Représentation schématique d'un coiled coil α-helix



Représentation schématique d'une hélice α - loop - helice α .

- *Le plan β :*

Ce second élément de structure secondaire est formé de brins β, chaînes polypeptidiques presque complètement étirées pouvant atteindre une quinzaine de résidus de long. Les liaisons de type pont hydrogène entre l'oxygène du groupement carbonyle et l'hydrogène de l'amide, sont quasiment perpendiculaires à l'axe principal de la chaîne polypeptidique. C'est la formation de tels ponts hydrogène entre deux ou plusieurs chaînes adjacentes, qui permet la formation d'un feuillet β. Une analyse plus approfondie de cette structure permet de montrer que cette conformation n'est pas plane, mais très légèrement plissée. Ceci est sans doute dû aux angles de liaison adoptés par la chaîne polypeptidique. Il existe des feuillets parallèles et les antiparallèles, ils diffèrent par l'agencement des brins et le patron de ponts H (Pauling L., Grey R.B. & Brauson H.R., 1951). Les plans parallèles sont constitués de brins orientés dans le même sens et reliés par d'autres éléments de structure secondaire. En ce qui concerne les feuillets antiparallèles, ils présentent des brins en sens alternés entre lesquels les connections sont assurées par des boucles plus ou moins longues *(cfr. Figure 1.2.5)*. Pour les chaînes latérales des différents acides aminés constituant le plan, elles sont dirigées alternativement vers le haut, ensuite vers le bas du plan. Ce type de structure secondaire n'a pas non plus sa composition en acides aminés en fréquence égale. Dans le plan β, la présence de proline est permise malgré son effet disloquant de feuillets. Comme pour les hélices α, les résidus de grandes tailles et chargés restent rares.

Certaines microfibrilles, comme la fibroïne de la soie, sont construites par un empilement de feuillets répondant à une séquence consensus $(ser\text{-}gly\text{-}ala\text{-}gly)_n$, cet empilement étant maintenu par les seules interactions de Van der Waals entre les chaînes latérales (Lodish, 1997).

- *Les Coils :*

Dans un souci de simplification, un « coil » représente à peu près tout ce qui n'est ni hélice α, ni plan β c'est-à-dire les structures secondaires qui permettent la flexibilité de la chaîne polypeptidique autorisant des structures moins régulières. (ex : coudes, boucles…)

- **Les coudes** :

Ce sont de courts éléments de structure secondaire en forme de coude dont la fonction est de permettre un changement de direction important de la chaîne polypeptidique. Selon Venkatachalam, il existe trois principaux types de coudes. Celui de type I qui a une proline en position 3. Le type II qui nécessite des glycines en position 2 et 3. Et enfin celui de type III qui possède des valeurs ϕ et ψ répétitives (-60°, -30°) semblables à l'hélice 3_{10}. Nous pouvons par exemple voir de tels coudes au niveau des feuillets β antiparallèles. Ces derniers sont alors constitués de seulement de trois ou quatre résidus pouvant être stabilisé par un pont H entre le C==O du résidu n et le N—H du résidu n+3. Le coude au sens strict, est formé par les résidus n+1 et n+2 (Creighton, 1984). Pour cette structure, contrairement aux deux précédentes où les angles de torsions étaient répétitifs, la succession des valeurs ϕ et ψ a son importance.

- **Les boucles** :

Ce sont des fragments polypeptidiques de longueurs et de conformations tout à fait irrégulières. Ils ont pour rôle d'effectuer la liaison entre les différents éléments de structures secondaires. 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. On a pu voir qu'au cours de l'évolution, les régions au niveau des boucles ne sont pas conservées mais très variables. Des exceptions à cette règle sont les cas où ces boucles ont un rôle fonctionnel (ex : cofacteur, fixation d'un substrat...).

- **Les super-structures secondaires**

Les structures secondaires peuvent se combiner et prendre une conformation particulière souvent rencontrée dans les protéines. Ce sont les super-structures secondaires appelées motifs ou encore « folds ».

Quelques exemples *(cfr. Figure 1.2.6)* :

-Les coiled coil α-helix ; deux hélices α s'enroulent l'une autour de l'autre pour former une « super hélice » gauche.

-Hélice-boucle-hélice ; motif très fréquent dans les protéines impliquées dans la fixation du calcium et dans la fixation de l'ADN.

-Clé grecque ; composée de quatre brins β antiparallèles liés par des boucles. On la rencontre souvent dans des feuillets β antiparallèles.

-Motif βαβ ; deux brins β reliés par une chaîne irrégulière ou par une hélice α.

-Motif ββ : Une boucle joint deux brins β antiparallèles.

- **Les structures tertiaires et quaternaires**

On entend par structure tertiaire l'agencement des structures secondaires et des supers structures formant ainsi un (ou plusieurs) domaine (s) responsable (s) d'une fonction particulière. Il faudrait s'étendre plus longuement sur ce phénomène de repliement des

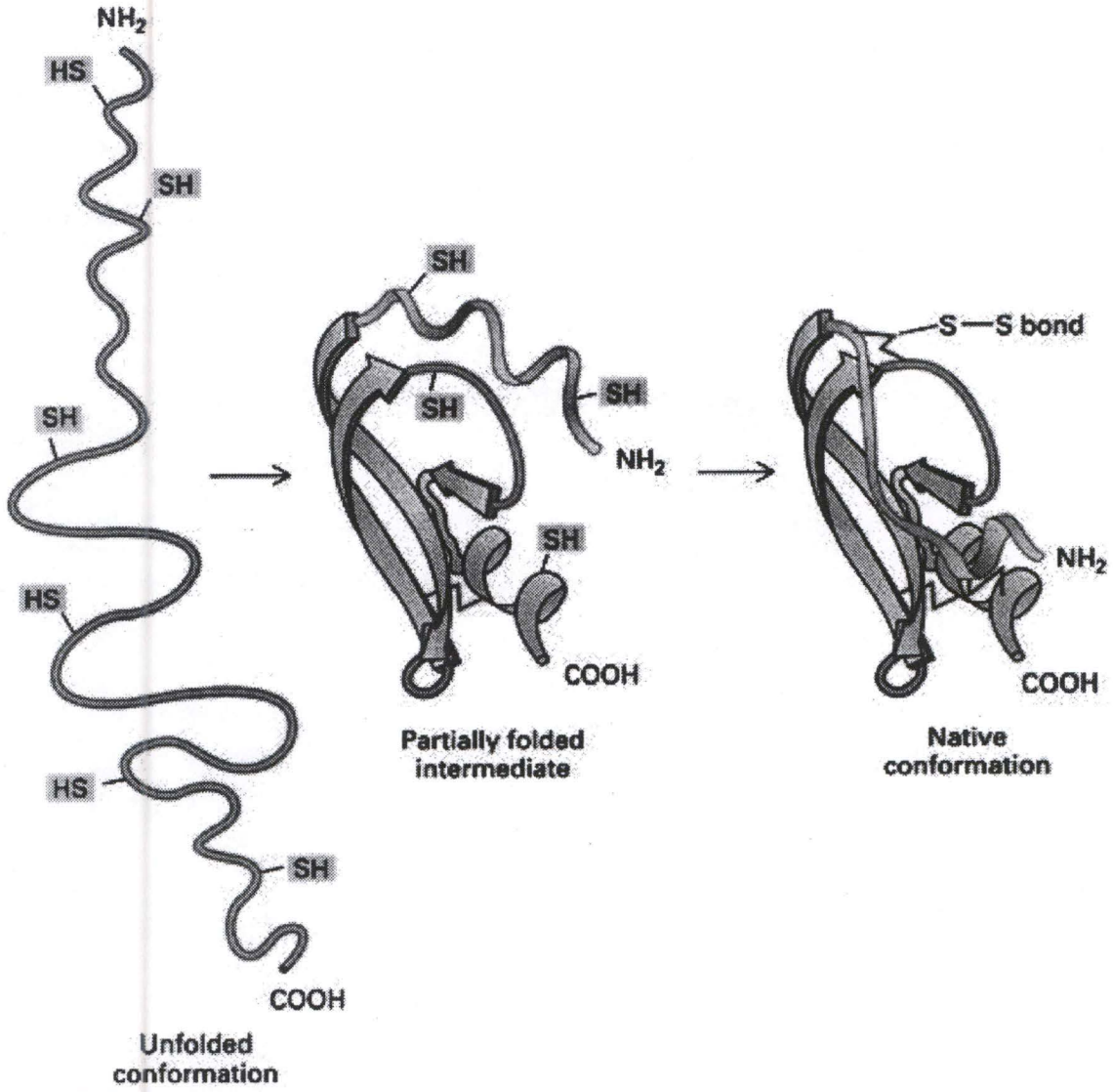


Figure1.2.7. Représentation schématique du repliement d'une protéine. On remarque que deux résidus éloignés au sein de la séquence primaire peuvent interagir une fois la protéine repliée.

protéines. Pour en comprendre les mécanismes, qui permettent de positionner correctement les résidus importants, afin d'assurer les fonctions de ces protéines. Même si des résidus étaient distants au niveau de la structure primaire, ils peuvent se retrouver très proches au point d'interagir ensemble dans la conformation native de la protéine.*(cfr figure.1.2.7)*. La structure quaternaire est le dernier niveau d'assemblage de plusieurs chaînes polypeptidiques. Une telle protéine serait constituée de sous-unités ou protomères ayant la faculté de s'associer via des interactions entre les chaînes latérales des résidus appartenant aux différents monomères. Selon la nature de ces monomères et le fait qu'ils soient identiques ou non, on distingue des homodimères ou hétérodimères. Les contacts sont établis entre ces monomères au niveau des hélices α, hélices β ou encore feuillets β.

- **Prédiction de structure secondaire :**

Un alignement de séquences trouvé à partir de la similarité des résidus permet d'inférer une similarité structurale probable uniquement de certaines régions où le pourcentage d'identité est élevé. De plus la similarité structurale n'est pas toujours détectable en terme de similarité de séquence. Pour pallier à ceci, il a été proposé de prédire les structures secondaires. Pour une telle détermination, il existe diverses méthodes.

La première d'entre elles est une méthode statistique basée sur des paramètres empiriques déterminés à partir de protéines résolues.

La deuxième se base sur des considérations physico-chimiques, elle tient compte des facteurs stéréochimiques, ou encore de l'hydrophobicité des résidus composant la protéine.

Il existe aussi une troisième méthode utilisant des réseaux neuronaux, constitués de différentes unités intégrant leurs propres données et les transformant en réponses qui sont transmises aux autres unités connectées en parallèles. Ces neurones sont entraînés

sur un jeu de protéines-test, et s'ensuit une extrapolation sur un jeu de protéines de structures inconnues. (*cfr. Support informatique PHD)*

- **Modélisation par comparaison de séquence :**

La résolution d'une structure de protéines fait appel à des techniques laborieuses à mettre en œuvre comme la cristallographie ou la résonance magnétique nucléaire. C'est en tenant compte de cela que l'approche théorique via des méthodes de prédiction de structures joue un rôle non négligeable. Surtout quand on pense aux applications possibles telles que la conception rationnelle de molécules thérapeutiques ou à l'ingénierie des protéines. La comparaison de structures tertiaires de protéines homologues montre que celles-ci sont mieux conservées, dans l'évolution, que les séquences protéiques, elles-mêmes bien mieux conservées que les séquences nucléiques.

Diverses méthodes de prédiction de structure protéique tentent de relier des séquences à des structures tridimensionnelles connues. S'il existe une similarité importante entre une séquence de structure inconnue et une séquence de structure connue, il y a possibilité de déterminer la structure tridimensionnelle de la première séquence par homologie avec la seconde.

Cette optique de recherche fut abordée en premier lieu par Browne et ses collaborateurs (1969) et trouve son origine dans diverses observations :

Les protéines forment des groupes de familles de structures (Richardson, 1981).

Le nombre de familles est assez limité. Il serait estimé à 1000 (Chothia, 1992).

Le nombre de familles de topologies différentes est de l'ordre de 500 à 700 (Blundell and Johnson, 1993).

La difficulté de la modélisation par homologie dépend du taux d'identité entre les séquences de structures connues et celle d'intérêt dont la structure est inconnue. Dans ce genre d'approche, nous comprenons de suite l'importance de la qualité des méthodes de comparaison de séquences. L'alignement multiple jouera un rôle essentiel afin de déterminer les régions dont la structure est conservée.

Grâce à l'outil Internet, nous pouvons rechercher, à travers des banques de données, un maximum de protéines. Plus il y aura de protéines intégrées dans l'alignement, plus les régions conservées qui sortiront de l'alignement seront fiables. Il est néanmoins indispensable d'avoir des protéines de structures connue pour réaliser une modélisation par homologie.

Ensuite, le modèle structural sera élaboré par des logiciels spécialisés qui identifieront les régions variables de celles conservées. Après quoi ce modèle brut peut-être affiné en utilisant la mécanique et la dynamique moléculaire ainsi que l'analyse des champs de forces pour étudier exactement le positionnement des chaînes latérales et permettre la modélisation des régions variables.

Il faut rester critique vis-à-vis de cette méthode :

- Les structures connues ne représentent encore qu'un échantillon de l'ensemble des protéines.
- Une séquence donnée peut-être similaire à plusieurs protéines différentes.
- Le résultat final de la modélisation dépendra fortement du taux de similarité entre les séquences. Des seuils de similarité ont été établis, pour 80 % de similarité ou plus, il est facile de faire de la modélisation par homologie. Par contre cette modélisation devient plus ardue et moins fiable lorsqu'on est confronté à des pourcentages de similarité de l'ordre 50 à 25 %. En dessous de ce seuil, la faisabilité de l'alignement se pose. De plus, la fiabilité du modèle établi sera largement remise en doute.

I.3. Les matrices de scores

Toute méthode d'alignement de séquences repose sur une série de scores caractérisant la similarité ou la distance entre les paires d'acides aminés. L'amélioration des performances est donc largement dépendante de l'optimisation de ces scores. La représentation la plus communément utilisée est celle des matrices de scores, tableaux

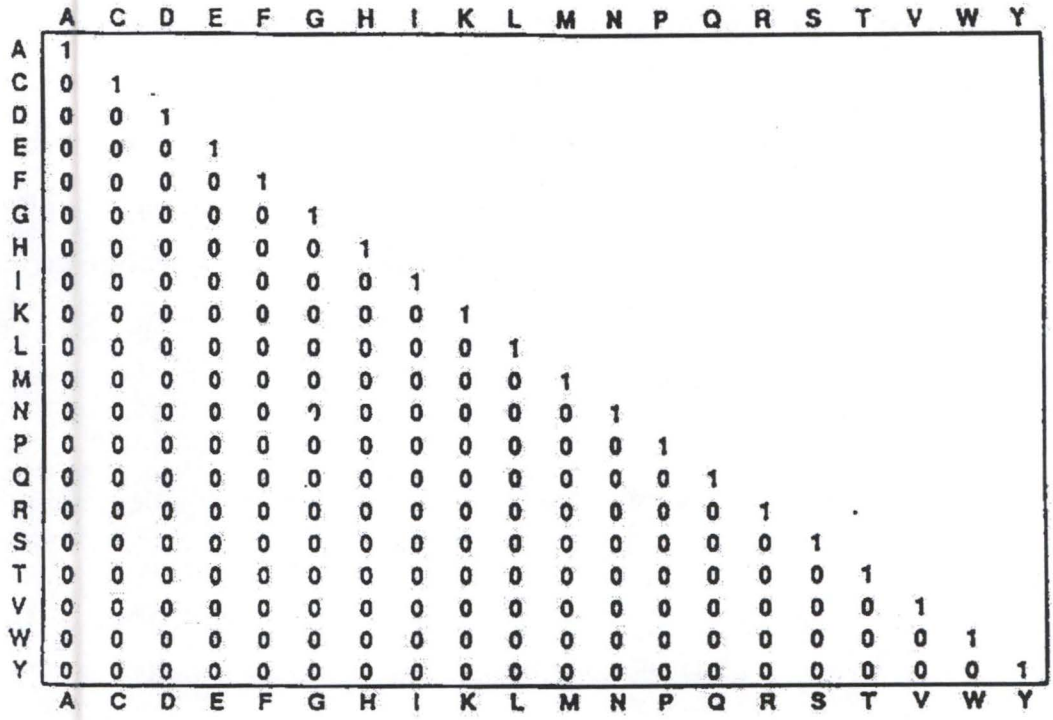


Figure I.3.1 Représentation de la matrice identité

symétriques 20 X 20 où seuls 210 coefficients sont originaux (la paire des résidus A-I étant considérée comme équivalente à la paire des résidus I-A). Il existe bon nombre de matrices de scores dont la conception diffère fondamentalement. Certaines sont construites sur base d'échelles de caractéristiques physico-chimiques des divers acides aminés, d'autres sont établies sur le taux de mutations, de substitutions. C'est ce que nous allons découvrir au travers de divers exemples.

** On distingue deux représentations de matrices de scores :*

- Les matrices de similarité, où des paires de résidus de caractéristiques identiques (ou très proches), reçoivent un score élevé par rapport à celles présentant des caractères différents ou éloignés.
- Les matrices de distance, où la situation est inverse puisque ici plus les résidus des paires sont différents plus leurs scores sont élevés.

Rem : Toute matrice de similarité peut être transformée en matrice de distance en retirant à toutes les valeurs le score maximum et en changeant de signe.

** Exemples de matrices de similarité :*

- La matrice identité est la matrice de similarité la plus simple *(cfr. figure 1.3.1)*. Les paires d'acides aminés sont classées en deux catégories : identiques et non identiques. Les paires non-identiques reçoivent un score de zéro, alors que les autres se voient attribuer un score positif, généralement de un. Cette matrice d'identité tient uniquement compte du nombre de résidus identiques entre deux séquences. De plus elle donne la même importance à toutes les substitutions. Bien que très limitée, elle nous permet néanmoins de calculer un pourcentage d'identité, cette valeur fréquemment utilisée pour évaluer la similarité globale entre deux séquences.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	Ala	9730	0	31	24	5	34	37	42	5	3	5	18	19	5	54	99	45	0	0	32
R	Arg	0	9881	5	0	0	13	0	0	17	0	0	23	18	2	0	1	0	0	0	0
N	Asn	14	7	9701	36	0	20	7	10	24	4	2	19	1	0	10	51	17	0	0	4
D	Asp	13	0	45	9757	0	27	96	8	6	0	2	8	1	0	1	26	2	0	0	4
C	Cys	1	0	0	0	9928	0	0	1	0	2	0	0	11	0	0	12	3	0	0	6
Q	Gln	12	14	15	16	0	9736	24	4	14	4	2	9	11	0	11	13	10	0	0	5
E	Glu	21	0	9	95	0	40	9726	13	4	4	4	13	1	0	17	15	12	0	0	7
G	Gly	40	0	22	13	3	11	22	9870	1	0	2	5	0	0	17	42	8	0	0	7
H	His	2	19	20	4	0	15	3	0	9865	4	3	6	0	3	0	10	5	11	4	1
I	Ile	1	0	3	0	3	4	3	0	4	9703	22	4	22	14	2	3	14	0	0	70
L	Leu	4	0	3	3	0	4	7	2	6	52	9899	6	99	19	0	5	7	0	0	24
K	Lys	17	65	37	13	0	23	21	5	14	9	6	9845	11	0	6	22	14	0	4	13
M	Met	2	7	0	0	5	4	0	0	0	7	14	2	9672	5	0	5	2	0	0	12
F	Phe	2	3	0	0	0	0	0	0	4	18	10	0	18	9879	0	5	2	30	74	2
P	Pro	23	0	9	1	0	13	13	7	0	3	0	3	0	0	9850	11	5	0	0	4
S	Ser	59	2	67	28	27	22	16	26	17	4	3	14	23	6	15	9598	69	0	0	7
T	Thr	30	0	25	3	8	20	14	6	8	24	5	10	11	3	8	76	9759	0	0	20
W	Trp	0	0	0	0	0	0	0	0	4	0	0	0	0	8	0	0	0	9941	7	0
Y	Tyr	0	0	0	0	0	0	0	0	4	0	0	2	0	51	0	0	0	17	9909	0
V	Val	27	0	7	5	18	12	10	6	3	156	22	12	82	3	8	9	25	0	0	9783

Figure I.3.2 Représentation de la matrice de probabilité de mutation de Dayhoff (1971)

- La matrice de similarité de Dayhoff est l'une des plus connue. L'idée développée par M. Dayhoff est que deux résidus peuvent se permuter s'ils ont des propriétés physico-chimiques proches. M. Dayhoff suit l'apparition de substitutions dans des familles de protéines homologues classées sur des arbres phylogénétiques (Dayhoff M. O., 1972), quantifiant la fréquence de substitution entre une séquence ancestrale et une séquence actuelle. Cette valeur de fréquence indique si, au cours de l'évolution, la substitution est plus ou moins bien acceptée par la sélection naturelle. Ce nombre de substitutions pour chaque paire d'acides aminés est stocké dans une matrice appelée matrice de mutation « mutation data » ; la probabilité de remplacement d'un acide aminé i par un acide aminé j est calculée. Cette valeur peut être estimé par l'expression suivante :

$$M_{ij} = \frac{N_{ij} N_i 100}{n_i N_1} \qquad (1.1)$$

- $\mathbf{N_{ij}}$ = la fréquence de remplacement de l'acide aminé i par l'acide aminé j
- $\mathbf{N_i}$ = la fréquence de remplacement de l'acide aminé i
- $\mathbf{N_1}$ = la fréquence de remplacement totale
- $\mathbf{n_i}$ = la fréquence de l'acide aminé i

La fraction M_{ij} permet d'exprimer une fréquence de substitution normalisée selon la longueur de la séquence, la fréquence des acides aminés et leur exposition aux divers remplacements. Ce M_{ij} est donc la probabilité que la substitution acceptée soit celle de l'acide aminé i par l'acide aminé j, dans une séquence de 100 résidus, cela pendant l'intervalle évolutif au cours duquel une seule substitution ait été autorisée. Ces valeurs permettent de construire une matrice des probabilités de mutations (MPM) où chaque élément ij donne la probabilité que i soit remplacé par j *(cfr. Figure 1.3.2).* L'unité d'évolution représentée par cette matrice correspond donc à une mutation

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	11.5																			
S	0.1	2.2																		
T	-0.5	1.5	2.5																	
P	-3.1	0.4	0.1	7.6																
A	0.5	1.1	0.6	0.3	2.4															
G	-2.0	0.4	-1.1	-1.6	0.5	6.6														
N	-1.8	0.9	0.5	-0.9	-0.3	0.4	3.8													
D	-3.2	0.5	0.0	-0.7	-0.3	0.1	2.2	4.7												
E	-3.0	0.2	-0.1	-0.5	0.0	-0.8	0.9	2.7	3.6											
Q	-2.4	0.2	0.0	-0.2	-0.2	-1.0	0.7	0.9	1.7	2.7										
H	-1.3	-0.2	-0.3	-1.1	-0.8	-1.4	1.2	0.4	0.4	1.2	6.0									
R	-2.2	-0.2	-0.2	-0.9	-0.6	-1.0	0.3	-0.3	0.4	1.5	0.6	4.7								
K	-2.8	0.1	0.1	-0.6	-0.4	-1.1	0.8	0.5	1.2	1.5	0.6	2.7	3.2							
M	-0.9	-1.4	-0.6	-2.4	-0.7	-3.5	-2.2	-3.0	-2.0	-1.0	-1.3	-1.7	-1.4	4.3						
I	-1.1	-1.8	-0.6	-2.6	-0.8	-4.5	-2.8	-3.8	-2.7	-1.9	-2.2	-2.4	-2.1	2.5	4.0					
L	-1.5	-2.1	-1.3	-2.3	-1.2	-4.4	-3.0	-4.0	-2.8	-1.6	-1.9	-2.2	-2.1	2.8	2.8	4.0				
V	0.0	-1.0	0.0	-1.8	0.1	-3.3	-2.2	-2.9	-1.9	-1.5	-2.0	-2.0	-1.7	1.6	3.1	1.8	3.4			
F	-0.8	-2.8	-2.2	-3.8	-2.3	-5.2	-3.1	-4.5	-3.9	-2.6	-0.1	-3.2	-3.3	1.6	1.0	2.0	0.1	7.0		
Y	-0.5	-1.9	-1.9	-3.1	-2.2	-4.0	-1.4	-2.8	-2.7	-1.7	2.2	-1.8	-2.1	-0.2	-0.7	0.0	-1.1	5.1	7.8	
W	-1.0	-3.3	-3.5	-5.0	-3.6	-4.0	-3.6	-5.2	-4.3	-2.7	-0.8	-1.6	-3.5	-1.0	-1.8	-0.7	-2.6	3.6	4.1	14.2

Figure I.3.3 Représentation de la matrice de Gonnet

acceptée pour 100 résidus (1 PAM : Point Accepted Mutation). Il est possible, par produit matriciel, d'obtenir toute une famille de matrices de similarités situées entre deux matrices théoriques extrêmes. C'est la famille de matrices de scores PAM qui reprend toute une série de matrices dans lesquelles chaque ligne donne la fréquence d'occurrence d'un acide aminé, pondérant ainsi le remplacement de tout résidu par n'importe quel autre. On remarque dans la littérature que la matrice PAM 250 par exemple est particulièrement souvent utilisée (Dayhoff M. O., 1983). Cette matrice a su rester une des matrices classiques alors que, depuis sa création, de nombreuses matrices ont été construites sur base d'un plus grand nombre de protéines connues.

- La matrice de Gonnet, établie d'une manière semblable à celle de Dayhoff, est représentative et donc plus fiable, car elle se base sur un nombre de séquences et donc de substitutions beaucoup plus élevé (Gonnet *et al.*, 1992). Selon les auteurs, leur matrice est mieux adaptée que celle de Dayhoff pour la comparaison de séquences de similarité moyenne *(cfr. Figure 1.3.3)*.

- La matrice de Henikoff est générée par utilisation de fréquences d'occurrence observées dans des alignements locaux, constitués de blocs de séquences sans *indel*. Chaque bloc représente une région conservée dans un groupe de protéines (Henikoff and Henikoff, 1992; Henikoff and Henikoff, 1993). Ces recherches ont été réalisées sur plusieurs centaines de protéines, ce qui a permis de composer une banque de plus de deux mille blocs. Ensuite, pour chacune des 210 paires de résidus, la fréquence est calculée et placée dans un tableau. La probabilité observée de chaque paire de résidus i et j (q_{ij}) peut donc être calculé de la manière suivante :

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}} \qquad (1.2)$$

Avec :

$\mathbf{f_{ij}}$, la fréquence observée pour la paire **ij**.

Il est ensuite possible de calculer la probabilité d'occurrence de l'acide aminé **i** dans la paire **ij** ($\mathbf{p_i}$) et la probabilité d'occurrence de la paire **ij** ($\mathbf{e_{ij}}$) :

$$p_i = q_{ii} + \sum_{j \neq i} \frac{q_{ij}}{2} \tag{1.3}$$

$$\begin{aligned} e_{ij} &= p_i p_j \quad ;si \quad i = j \\ &= 2 p_i p_j \;;si \quad i \neq j \end{aligned} \tag{1.4}$$

La dernière étape permet le calcul d'une matrice « log-odds » où chaque entrée est définie comme suit :

$$S_{ij} = \log_2 \left(\frac{q_{ij}}{e_{ij}} \right) \tag{1.5}$$

où $\mathbf{S_{ij}}$ est égale à zéro, plus petite que zéro ou plus grande que zéro, selon que les fréquences observées sont égales, plus petites ou plus grandes que ce qui est attendu. La matrice obtenue est appelée *BLOSUM* (blocks substitution matrix). Une dernière amélioration de cette matrice consiste à éviter que les membres de familles de protéines fortement apparentées aient trop de poids. Prenons le cas où, à la position x d'un alignement de vingt séquences, on retrouve dix-huit alanines et deux sérines. La paire AA pèsera donc beaucoup plus que la paire AS. Une solution serait de regrouper les protéines ayant un certain pourcentage d'identité et ensuite de faire la moyenne de leurs contributions au point de vue des fréquences d'occurrence des paires de résidus. Selon la limite de pourcentage

											Henikoff		BLOSUM62							
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	53																			
C	26	86																		
D	13	6	66																	
E	20	0	40	60																
F	13	13	6	6	66															
G	26	6	20	13	6	66														
H	13	6	20	26	20	13	80													
I	20	20	6	6	26	0	6	53												
K	20	6	20	33	6	13	20	6	60											
L	20	20	0	6	26	0	6	40	13	53										
M	20	20	6	13	26	6	13	33	20	40	60									
N	13	6	33	26	6	26	33	6	26	6	13	66								
P	20	6	20	20	0	13	13	6	20	6	13	13	73							
Q	20	6	26	40	6	13	26	6	33	13	26	26	20	60						
R	20	6	13	26	6	13	26	6	40	13	20	26	13	33	60					
S	33	20	26	26	13	26	20	13	26	13	20	33	20	26	20	53				
T	26	20	20	20	13	13	13	20	20	20	20	26	20	20	20	33	60			
V	26	20	6	13	20	6	6	46	13	33	33	6	13	13	6	13	26	53		
W	6	13	0	6	33	13	13	6	6	13	20	0	0	13	6	6	13	6	100	
Y	13	13	6	13	46	6	40	20	13	20	20	13	6	20	13	13	13	20	40	73

Figure I.3.4 Représentation de la matrice BLOSUM 62 (Henikoff 1992)

que l'on a choisi, on obtiendra des matrices pondérant les segments de similarité importante. Une matrice pour laquelle la limite est de soixante-deux pour cent est appelée BLOSUM62 *(cfr. figure 1.3.4)*. Cette matrice de scores est utilisée par défaut par le programme d'alignement multiples Match-Box décrit au point I.5 sur les logiciels d'alignement de séquences.

- Les matrices de la famille de Johnson et Overington, contrairement aux quatre précédentes, sont calculées sur base de fréquences de substitutions observées dans des alignements structuraux, impliquant deux cents trente-cinq protéines. La première étape a consisté à classer les deux cent trente-cinq structures en familles homologues, puis, au sein de ces familles de réaliser les alignements (Johnson and Overington, 1993). Au total, plus ou moins deux cents mille substitutions furent observées et accumulées dans une table de fréquences. Ces fréquences ont ensuite été transformées en probabilité selon la formule :

$$p_{ij} = \frac{f_{ij}}{\sum_{j=1}^{20} f_{ij}} \qquad (1.6)$$

Finalement, la matrice de probabilité est convertie en matrice de « log-odds » :

$$O_{ij} = P_{ij}\left[\frac{\sum_i f_{ij}}{\sum_i \sum_j f_{ij}}\right] \qquad (1.7)$$

On obtient une matrice contenant des valeurs s'échelonnant entre -9,8 et 16,1. Après addition de 9,8 à toutes les valeurs, les scores varient entre 0 (pour les substitutions cystéine-aspartate ou méthionine-proline) et 26 (pour la conservation de la cystéine).

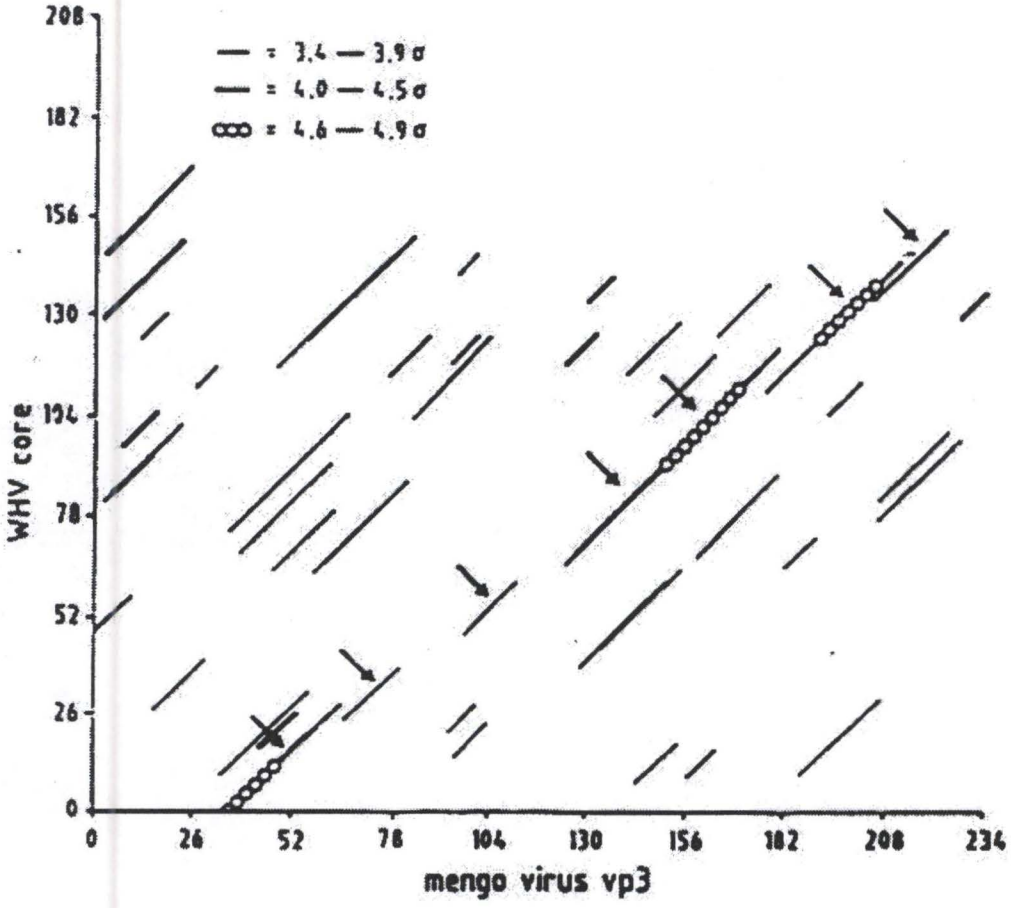


Figure I.4.1 Représentation d'un « Dot Plot »,table de comparaison entre deux séquences protéiques où chaque point représente des résidus considérés comme similaires. Les classes de similarités sont représentée par un trait différent.

* Utilisation des matrices de scores lors de l'alignement :

Nous utiliserons ces matrices dans l'étape de construction d'alignement de séquences afin de rendre ceux-ci les plus cohérents possible. Le rôle des matrices consiste à établir une correspondance entre les acides aminés les plus similaires pour pouvoir délimiter les régions structurellement conservées (SCR). Pour détecter la meilleure correspondance, deux méthodes sont généralement utilisées. Elles dépendent toutes deux du nombre de résidus pris en compte lors d'une comparaison. La première méthode considère un résidu à la fois, dans chacune des séquences. Prenons le cas de l'alignement pairé, où la similarité est calculée entre des paires de résidus. La seconde méthode, basée sur l'observation d'une distribution non-uniforme des résidus similaires ou identiques, réalise des comparaisons entre fenêtres de résidus. Dans ce cas, les distances sont calculées en additionnant les distances entre les acides aminés se faisant face dans les deux fenêtres.

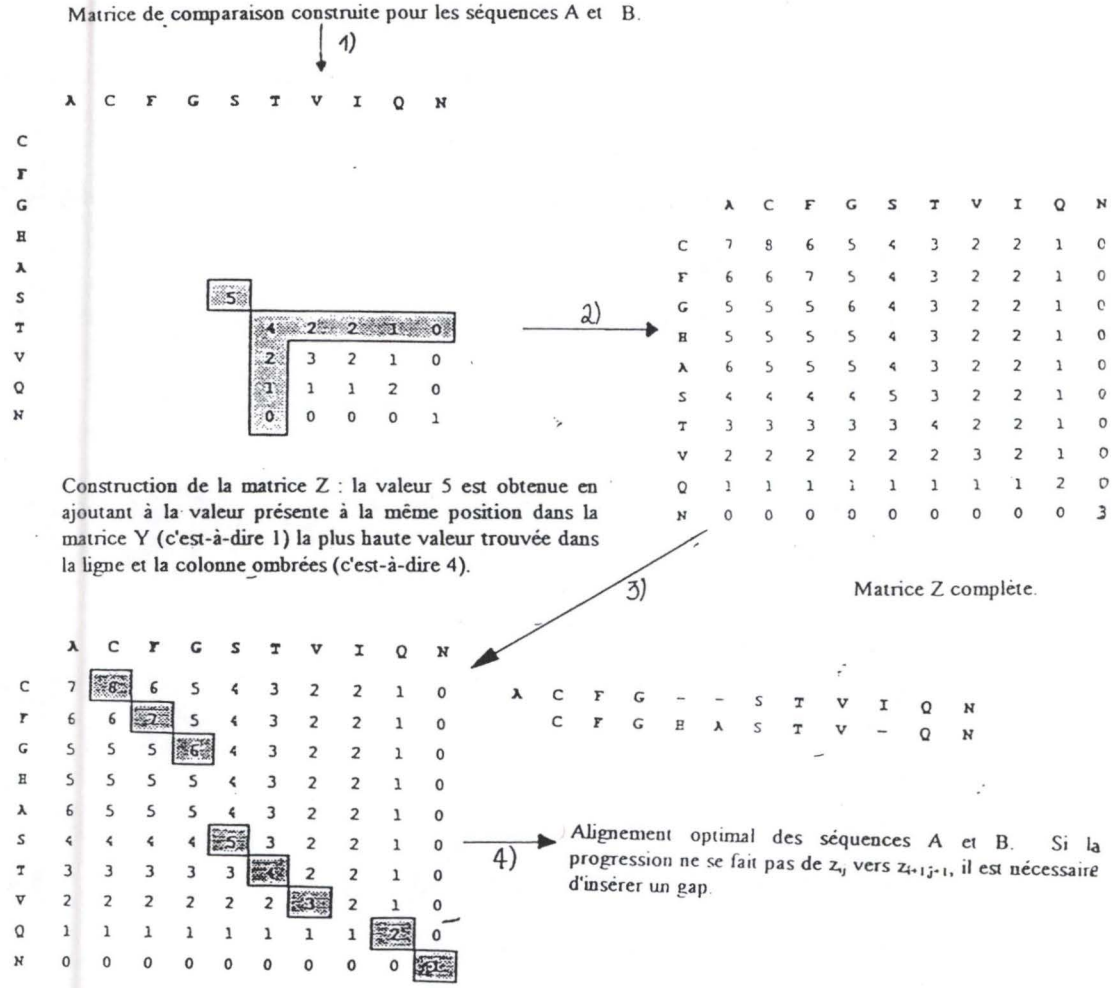
I.4. Méthode d'alignement de deux séquences

Différentes méthodes de comparaison de deux séquences ont été mises au point pour y déceler des segments identiques ou similaires, tout en essayant de réduire le temps de calcul et l'espace disque utilisé. Chaque alignement reçoit un score qui est la somme des scores pour chaque paire d'acides aminés se trouvant dans l'alignement. Si on utilise une matrice de similarité, l'alignement le plus probable sera celui qui possède le plus grand score.

Le « *dot plot* » est une méthode de comparaison de séquence utilisant une matrice de similarité pour pondérer chaque paire de résidus dans un alignement. Dans une telle analyse de séquence, toutes les fenêtres possibles d'une séquence A (de longueur m), sont comparées avec toutes celles d'une autre séquence B (de longueur n). On place les résultats de ces comparaisons dans une matrice R où chaque élément de score r_{ij}

	A	C	F	G	S	T	V	T	Q	N
C	0	1	0	0	0	0	0	0	0	0
F	0	0	1	0	0	0	0	0	0	0
G	0	0	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	0
S	0	0	0	0	1	0	0	0	0	0
T	0	0	0	0	0	1	0	0	0	0
V	0	0	0	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	1	0
N	0	0	0	0	0	0	0	0	0	1

Matrice de comparaison construite pour les séquences A et B.



Construction de la matrice Z : la valeur 5 est obtenue en ajoutant à la valeur présente à la même position dans la matrice Y (c'est-à-dire 1) la plus haute valeur trouvée dans la ligne et la colonne ombrées (c'est-à-dire 4).

	A	C	F	G	S	T	V	I	Q	N
C	7	8	6	5	4	3	2	2	1	0
F	6	6	7	5	4	3	2	2	1	0
G	5	5	5	6	4	3	2	2	1	0
H	5	5	5	5	4	3	2	2	1	0
A	6	5	5	5	4	3	2	2	1	0
S	4	4	4	4	5	3	2	2	1	0
T	3	3	3	3	3	4	2	2	1	0
V	2	2	2	2	2	2	3	2	1	0
Q	1	1	1	1	1	1	1	1	2	0
N	0	0	0	0	0	0	0	0	0	3

Matrice Z complète.

Tracé optimal dans la matrice Z

A	C	F	G	–	–	S	T	V	I	Q	N
C	F	G	H	A	S	T	V	–	–	Q	N

Alignement optimal des séquences A et B. Si la progression ne se fait pas de z_{ij} vers $z_{i+1,j+1}$, il est nécessaire d'insérer un gap.

Figure I.4.2 Représentation des étape successives de la méthode de programmation dynamique de Needleman et Wunsh consistant en un alignement entre deux séquences protéiques.

représentent la similarité entre la $i^{ème}$ fenêtre de A et la $j^{ème}$ fenêtre de B. L'homologie apparaît sous la forme d'une diagonale de points, les *indels* correspondent aux césures *(cfr.figure I.4.1)*.

L'avantage est que l'ensemble des comparaisons possibles est réalisé tout en permettant de fixer un seuil de similarité qui quantifie les ressemblances entre les séquences ou de déterminer des classes de similarité. Par contre cette technique ne met en œuvre qu'une comparaison pairée, elle est donc véritablement contrainte à comparer uniquement deux séquences entre elles. Ce genre de comparaison entraîne un risque énorme de faux positifs (risque d'erreur α important). Le *dot plot* ne permet qu'une visualisation plus aisée et une quantification des ressemblances entre des régions de séquences. Cette méthode ne peut pas mettre en œuvre un alignement pairé complet. Mais ce principe de la diagonale la plus significative est à la base de la plupart des programmes d'alignement pairé repose sur ce type de technique. Il existe plusieurs méthodes d'alignement pairé a proprement parlé, nous nous contenterons de parler des plus connues.

- Méthode de programmation dynamique de Needleman et Wunsh (Needleman S. B., 1970). Les auteurs ont mis au point une technique analytique construisant un alignement pairé sur base d'une matrice de scores. Pour comprendre cette méthode, prenons un exemple où le point de départ est la matrice identité *(figure 1.4.2)*. Soit deux séquences A et B de longueur n_A et n_B . Leur matrice de comparaison est la matrice identité Y (n_A x n_B), où la valeur de score vaut zéro si A_i est différent de B_j et la valeur de score vaut 1 si A_i est égal à B_j. Chaque score y_{ij} de la matrice est transformée pour devenir un élément z_{ij} de la matrice d'identités cumulées, nommée Z. Cette matrice est obtenue en se positionnant à la fin de deux séquences ($i= n_{A-1}$ et $j= n_{B-1}$) et en remontant vers l'amont des séquences (i=1 et j=1) , avec un déplacement de droite à gauche. À chaque pas, y_{ij} est modifié, la valeur maximale est recherchée au niveau de la ligne (i+1) se situant à droite de y_{ij} et de la colonne (j+1) située en dessous de y_{ij} . Cette valeur est cumulée à y_{ij} pour donner un score modifié. Une fois la matrice Z achevée, chacun des éléments z_{ij} représente

le nombre d'identités obtenu lorsqu'on aligne les segments des séquences en aval de i pour l'une et de j pour l'autre. La dernière étape consiste à trouver l'alignement optimal, c'est-à-dire dans la matrice Z le tracé pour lequel la somme des scores z_{ij} des traversées est maximale Si la diagonale se fait toujours de z_{ij} vers $z_{i+1\ j+1}$, on trace une diagonale complète. Dans tous les autres cas, le tracé saute d'une diagonale à l'autre. Dans l'alignement, ce sont les *indels* dont nous avons déjà parlé et qui permettent de disposer face à face les zones les plus similaires. Une fois l'alignement réalisé, il est possible d'en calculer un score global, somme des scores obtenus pour la comparaison des résidus placés face à face dans l'alignement. Pour éviter d'avoir trop d'*indels*, on peut sanctionner ceux-ci par un facteur de pénalité. Cependant le problème des *indels* reste arbitraire sans aucun fondement biologique. Cette méthode est très rigoureuse, mais l'algorithme est trop lent pour la mettre en œuvre dans l'optique d'une recherche dans une banque de séquences.

- La méthode de Wilbur et Lipman, recherche dans deux séquences, les zones identiques de w résidus de long (Wilbur W. J., 1983). Cette recherche ne se fait plus sur toute la longueur des séquences mais uniquement sur un fragment de m résidus. Cet algorithme beaucoup plus rapide que celui de Needleman et Wunsh va dans un premier temps parcourir l'entièreté des deux séquences pour y lire toutes les régions de w résidus. Il va ensuite, comparer chaque segment de la première séquence avec chaque segment de la seconde, compris dans le fragment de m résidus. Une fois les régions identiques détectées, comme dans le cas de Needleman et Wunsh, la dernière étape consiste en la construction de l'alignement. La vitesse d'exécution de l'algorithme dépendra de deux critères : de la taille des zones comparées (w) et de la taille des régions de comparaison (m). Dans le cas où w=1 et que m est très grand, on retrouve alors l'algorithme de Needleman et Wunsh.

- L'algorithme mis au point par Lipman et Pearson est celui qu'exécute le programme *FASTA* qui a pour principal objectif la comparaison d'une séquence cible par rapport à toute une banque de séquences. Sur base d'un alignement des régions identiques entre deux séquences, *FASTA* va réaliser un calcul de similarité où seul les régions similaires entrent en compte (Pearson and Lipman, 1988). Cet algorithme provient de *FASTP*, également mis au point par Lipman et Pearson. *FASTP* recherche les identités sur une zone de m résidus, tout comme le faisait celui de Wilbur et Lipman (Lipman and Pearson, 1985). Les positions des régions possédant les scores les plus élevés sont mémorisées (là où le pourcentage d'identité est maximum). Ensuite la matrice de scores *PAM250* est utilisée pour calculer un score de similarité dans ces régions. Seul le score maximum est conservé pour caractériser la similarité entre les deux séquences. Pour chaque séquence de la banque, il y a association d'un score par rapport à la séquence cible. L'ensemble des scores est alors représenté sous la forme d'un histogramme. *FASTA* apporte une amélioration dans le sens où *FASTP* ne recherchait qu'une région initiale, alors que *FASTA* recherche s'il n'y a pas d'autres régions initiales à considérer. Ensuite il calcule un alignement optimal égal à la combinaison des régions de score maximum compatibles entre elles.

- La méthode de Altschul et de ses collaborateurs est très semblable à celle décrite ci-dessus. Cette méthode, plus connue sans doute sous le nom de *BLAST (Basic Local Alignment Search Tool),* recherche la paire de segments de séquences qui possède le score de similarité le plus élevé, « Maximum Segment Pair » (M.S.P.). Pour *BLAST*, il est question de segments continus, et donc il n'y a plus d'*indels*. Les limites d'une M.S.P. sont établies de façon à maximiser le score de similarité. Des scores positifs sont attribués aux identités ainsi qu'aux remplacements conservatifs. Par contre, des scores négatifs pénalisent les substitutions.

I.5. Méthode d'alignement multiple

Une méthode d'alignement traitant plusieurs séquences est nécessaire pour éviter le problème de faux positifs évoqué plus haut. L'alignement pairé n'est pas délaissé pour autant, puisque de nos jours c'est le principal moteur de nombreux logiciels comme *BLAST,* s'exécutant sur l'Internet.

L'analyse de similarité utilisant un alignement multiple est devenue vraiment indispensable. L'avantage certain de ce type d'alignement non pairé, appelé multiple, est qu'une région similaire entre plusieurs séquences a beaucoup plus de poids qu'une région similaire uniquement partagée entre deux séquences. Ici, il n'est plus question de réaliser un alignement à l'œil, sauf pour de rares cas où les séquences impliquées dans l'alignement présenteraient un pourcentage d'identité supérieur à 90 %.

Il existe deux types d'alignement multiple, un dit progressif, l'autre dit simultané. L'alignement multiple dit progressif, est en réalité un alignement pairé auquel on vient rajouter toutes les séquences impliquées. Le désavantage de ce type d'alignement est qu'il est très lié aux deux premières séquences alignées. Néanmoins, un motif sera considéré comme très intéressant à la condition d'être ubiquiste dans plusieurs séquences. Ce simple fait limite déjà fortement le risque de faux positifs. Certains de ces algorithmes célèbres comme celui de Feng et Doolittle ou encore Corpet, sont représentés ci-dessous.

- La méthode de Feng et Doolittle mesure la similarité existant entre toutes les paires de résidus appartenant à des séquences différentes. Cette méthode transforme la similarité en une valeur numérique représentative de la distance entre les paires, de résidus qui porte le nom de score. Plus le score est faible, plus la similarité est grande. La méthode aligne les deux séquences présentant le score minimal, ensuite ces deux séquences sont alignées avec la séquence suivante, celle qui présente la

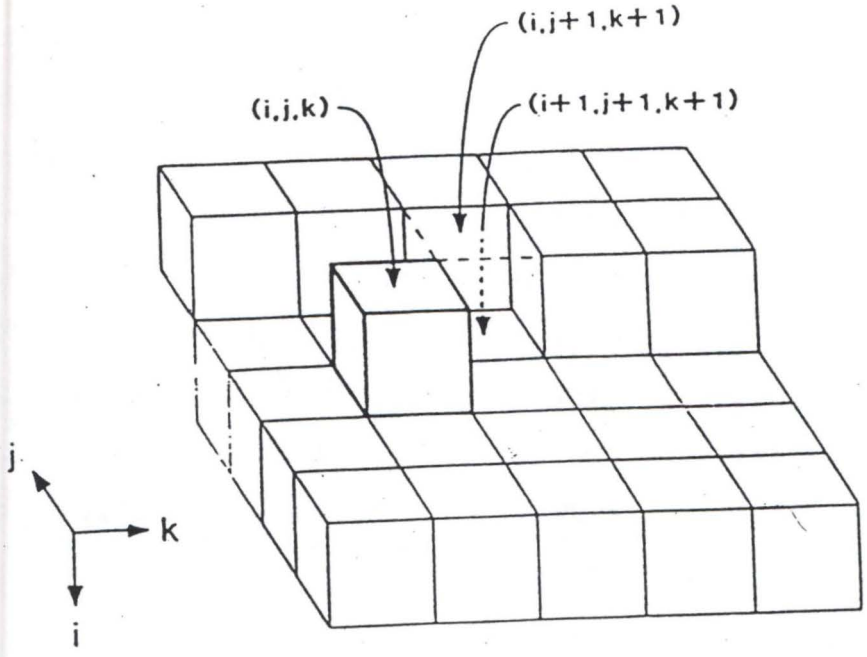


Figure I.5.1. Tableau à trois dimensions construit par l'alignement de trois séquences par la méthode de Murata (1985). A chaque case représentant un triplet d'acides aminés ijk est associé un score. Il s'agit de retrouver, à travers ce cube, un tracé pour lequel le score cumulé de chacune des cases soit minimum.

distance la plus faible parmi toutes les séquences restant à aligner. Et ainsi de suite, jusqu'à ce que toutes les séquences impliquées soient alignées (Feng D. F., 1987).

- La méthode de Corpet se base sur un tableau reprenant les mesures de ressemblance entre les diverses séquences à aligne. Les deux séquences les plus proches sont alignées : on sélectionne en quelque sorte l'alignement pairé le plus significatif et l'on forme ainsi un premiers groupe prenant la place des séquences alignées. Cela engendre la formation d'un second tableau tenant compte du premier appariement de séquences, et l'algorithme recherche quelle est la séquence qui s'apparie le mieux avec les séquences restantes ou éventuellement avec le premier groupe formé. On recommence jusqu'à ce que toutes les séquences se soient incorporées. À ce moment-là, un seul groupe subsiste, celui déterminant l'alignement optimum (Corpet, 1988).

Le second type d'alignement multiple, dit simultané, est plus connu sous le nom de l'algorithme de Murata (Murata M., 1985). Cette méthode est en fait une extension de l'algorithme de Needleman et Wunsh. La différence entre l'alignement multiple progressif et l'alignement multiple est qu'il est cette fois, réellement question de comparer trois séquences à la fois, cherchant ainsi un tracé optimal permettant d'obtenir un score maximal.

- Cette méthode de Murata est basée sur une comparaison entre trois séquences. L'algorithme génère un espace en trois dimensions où la case Y (i,j,k) correspond à la somme des scores associés à trois paires d'acides aminés appartenant respectivement à la première, à la seconde et à la dernière séquence. La méthode recherche le tracé optimal qui lui permettra d'obtenir un score global maximum. L'inconvénient majeur de cette méthode est que l'algorithme utilise la pondération des *indels* directement dans son calcul de similarité *(cfr. figure I.5.1).*

I.6. Logiciels d'alignement de séquences

Le but ici n'est pas d'être exhaustif au sujet des divers programmes d'alignement à disposition sur le marché. Quelques logiciels, les plus largement utilisés, sont simplement présentés de façon à rendre compte de l'éventail de programmes existant.

- ClustalW

ClustalW est un logiciel d'alignement multiple de séquences nucléiques ou protéiques basé sur la phylogénie (méthode Neighbour Joining). C'est sans doute l'un des logiciels plus utilisé à cet escient. Il s'agit d'une mise à jour de l'ancienne version de ClustalV mis au point par Higgins *et al.* (http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html). Plusieurs points de cette version ont améliorés, notamment la sensibilité progressive au cours de l'alignement de séquences (Thompson *et al.*, 1994). ClustalW utilise une méthode de programmation dynamique établissant un score global sur les séquences alignées, pour construire ses alignements. Il pondère chaque séquence protéique se trouvant dans un alignement partiel, ce qui lui permet de minimiser le poids des séquences qui sont dupliquées sous forme native. De cette façon, il augmente l'importance au niveau des séquences les plus divergentes. Il offre la possibilité d'utiliser des matrices de scores différentes en se basant sur deux familles de matrices de scores. (soit la famille des *BLOSUM*, soit dans celle des *PAM*) L'utilisation de l'une ou l'autre famille est fonction du pourcentage d'identité entre les différentes protéines à aligner. Pour éviter d'introduire des *indels* trop nombreux ou trop longs, ClustalW pénalise les scores sur base de la similarité entre les résidus, c'est

ce que l'on nomme le « *gap penalties* ». Ce « *gap penalties* » est indispensable car l'insertion d'un *indel* augmenterait de trop le pourcentage de similarité entre les séquences alignées. La différence principale entre la version ClustalV et ClustalW réside en une réduction locale dans les régions hydrophiles favorisant la formation de nouveaux *indels* dans les structures secondaires de type « boucle » plutôt que dans des structures secondaires plus régulières de type hélice α ou plan β. On suppose que les structures secondaires régulières les mieux conservées ont un rôle plus fonctionnel que les autres. Lorsque ces *indels* sont découverts précocement, au niveau de régions moins régulières, ils reçoivent une réduction de la pénalité. Ce qui permet de découvrir de nouveaux gaps aux alentours de ces positions.

- **BlockMaker :**

BlockMaker est un programme d'alignement de séquences 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 segments de la séquence. Lorsqu'on lui soumet plusieurs séquences, il aligne les séquences en construisant de nouveaux blocs toujours constitués de régions fortement conservées. Il génère ces blocs entre les séquences homologues, ensuite choisit lesquels assembler et auto-valide les blocs au sein de l'alignement. BlockMaker est capable de proposer des points d'ancrage pour des alignements de séquences présentant un niveau très faible d'homologie (Henikoff *et al.*, 1995). C'est le programme d'alignement multiple qui, par sa philosophie, se rapproche le plus de certains aspects de Match-Box.

(http://www.cs.auc.dk/~claus/block.html)

- Gibbs :

Ce programme trouve le meilleur score via des méthodes statistiques de probabilité en se basant sur la conservation des acides aminés au sein d'une famille (Lawrence *et al.*, 1993). Le programme Probe qui en découle 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 un effort de 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 obtenir une différence entre le meilleur alignement trouvé et les candidats obtenus par le biais du hasard. Par conséquent, il est souvent utile de lancer l'analyse d'un échantillon plusieurs fois pour voir si les résultats convergent vers le même alignement pour chaque essai. Si ce n'est pas le cas, (par exemple, pour des motifs de séquences très subtils), il faut nécessairement accomplir un grand nombre de recherches indépendantes avec en même temps un nombre suffisant d'échantillons. Le problème de segments de séquences répétées 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 entre chaque séquence (Neuwald *et al.*, 1997). De par la philosophie de son algorithme, Gibbs aura des résultats complètement imprédictibles. Par exemple, il alignera parfois mieux des séquences difficiles à aligner, alors que ses performances chuteront pour des cas très simples . (http://bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html).

- **Match-Box :**

Le logiciel Match-Box proposé par E. Depiereux et E. Feytmans est une méthode d'alignement multiple par comparaison de segments de 7 à 9 résidus de long (Depiereux and Feytmans, 1991). Il a pour but de rechercher des segments similaires dans un ensemble de protéines et de les faire correspondre dans un alignement. L'originalité de ce programme est qu'il travaille avec des segments de séquences plutôt qu'avec l'entièreté des séquences. De ce fait, Match-Box n'utilise pas d'*indels* pour paramétrer ses alignements. Match-Box établit lui-même son paramétrage par une analyse statistique préliminaire, qui évalue le bruit en vue de le maîtriser.
Match-Box est constitué de trois étapes successives et complémentaires que nous décrirons dans la suite de ce chapitre. La première étape consiste en une paramétrisation ayant pour but d'établir des seuils statistiques (étape du *scanning*). La seconde étape constitue une série de boîtes de segment de séquences établies sur base d'un indice de similarité et qui sont susceptibles de se retrouver dans l'alignement final (étape du *matching*). La troisième étape est un triage de toutes ces boîtes, afin de réaliser l'alignement le plus cohérent possible (étape du *screening*).

Une opération préliminaire appelée EXPLORE, détermine 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 routine de Match-Box compare toutes les séquences protéiques et aboutit à l'attribution d'un score.
*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.
*Une distribution de la fréquence d'apparition de tous les scores est comparée à celle qui est obtenue lorsque les résidus des deux séquences sont mélangés aléatoirement

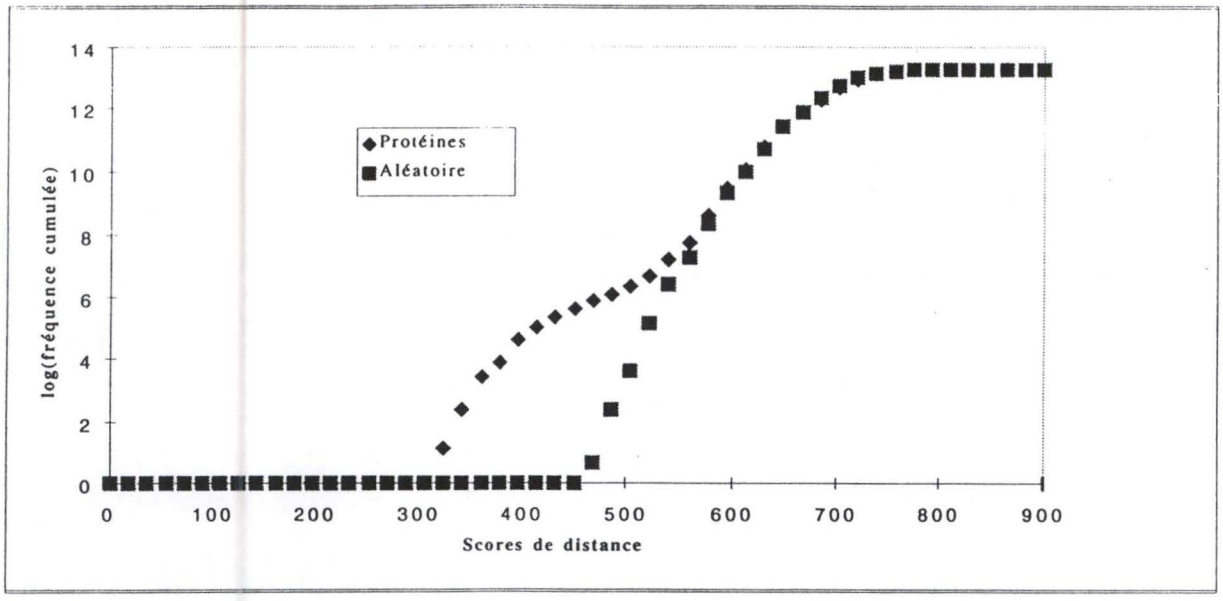


Figure I.6.1.

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.

(cfr. figure .I.6.1). Si la première est significativement différente de la seconde, on en déduit que la similarité entre les deux séquences considérées se différencie de ce que le hasard aurait pu produire. Dans le cas contraire, l'utilisateur est averti du risque d'obtenir des résultats qui sont liés au hasard.

Au cours des étapes de *scanning* et de *matching*, une fenêtre (segment de longueur m constante) se positionne au premier acide aminé de la première séquence. Le segment recouvert par cette fenêtre (la fenêtre « fixe ») est comparé à, une seconde fenêtre (la fenêtre « mobile ») qui balaye systématiquement toutes les séquences du début à la fin. Chacune de ces positions fait donc l'objet d'une comparaison pairée avec la fenêtre fixe.

Lorsque toutes ces comparaisons sont effectuées, la fenêtre fixe se déplace d'un acide aminé sur la première séquence et un nouveau 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.

L'alignement multiple repose en fait sur une comparaison pairée des résidus occupant la même position dans chacun des deux type de fenêtres. Mais l'alignement est multiple puisque la comparaison est faite avec toutes les séquences impliquées. Lors de chaque comparaison pairée, un score est attribué en fonction de la similarité entre les deux acides aminés. Les scores de similarité sont repris dans les matrices de scores. 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 (ce score va de 0 à maximum 900 puisque la comparaison entre deux résidus se base sur une échelle allant de zéro à 100 et que l'on a une fenêtre de 9 acides aminés).

La taille de la fenêtre est déterminée par l'utilisateur et reste constante tout au long de la procédure. Tout ce que Match-Box est capable de « voir » est compris dans de cette fenêtre, ce qui implique qu'elle 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.

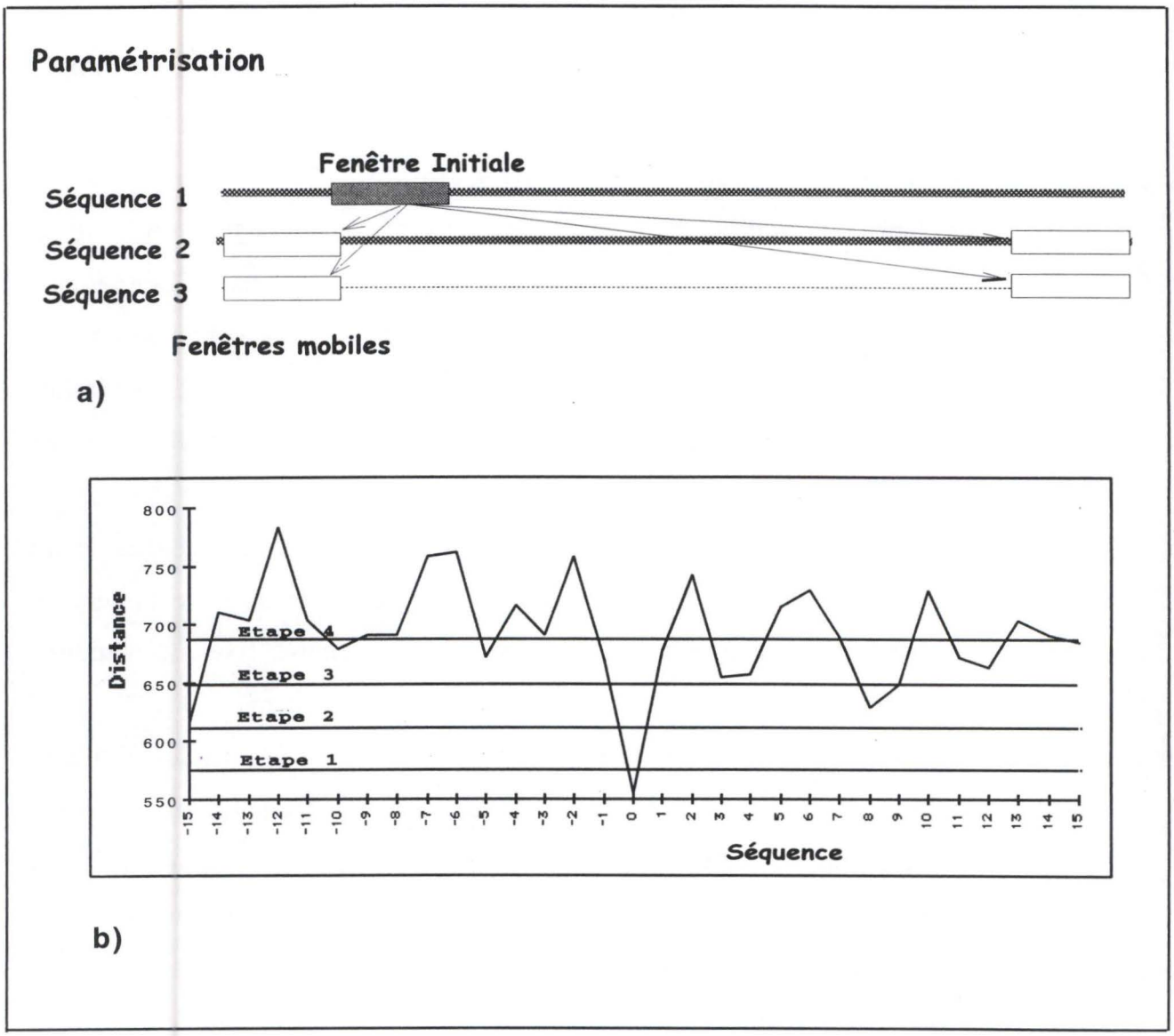


Figure I.6.2. :

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

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

Un serveur Match-Box se trouve sur le réseau de communication Internet, il est destiné à recevoir des séquences protéiques en vue de les aligner (Depiereux *et al.*, 1997). Ce serveur se trouve à l'adresse suivante :
http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html

- *Étape de paramétrisation, scanning.*

Le *scanning* est une étape de paramétrisation, précédant tout alignement de séquence, qui détermine la fréquence d'apparition des scores lors des comparaisons pairées des séquences à aligner *(cfr. figure I.6.2.a)*. La distribution de ces scores permet la paramétrisation du système en déterminant quatre valeurs seuils distinguant de façon optimal le signal (score faible correspondant à une similarité significativement élevée) du bruit (score élevé) *(cfr. figure I.6.2.b)*. Le seuil appelé *cut-off*, va ainsi limiter le bruit de fond en servant de filtre et déterminer l'endroit à partir duquel on va commencer à rechercher du signal. L'algorithme est capable de calculer ce bruit à partir de n'importe quel jeu de séquences et donc de déterminer statistiquement des caractéristiques optimales qui seront utilisées pour les étapes ultérieures de Match-Box Ces valeurs seuils seront utilisées dans l'étape suivante, celle du *matching*.

- *Étape de constitution de boîte, le matching*

Le *matching* a pour but est de constituer des boîtes de segments similaires. La fenêtre initiale se trouve toujours sur la première séquence, et le premier balayage de la fenêtre mobile se déplace successivement sur toute la longueur des séquences avec un premier seuil statistique très sévère. Pour chaque séquence, il détermine un score minimum correspondant au meilleur appariement (*best match*), c'est-à-dire la région la plus similaire de chaque séquence à la fenêtre initiale de la première séquence. Ensuite d'autres balayages auront lieu aux environs de cette zone, en relâchant progressivement le seuil de sévérité. Ici pour chaque étape du balayage sur ces régions,

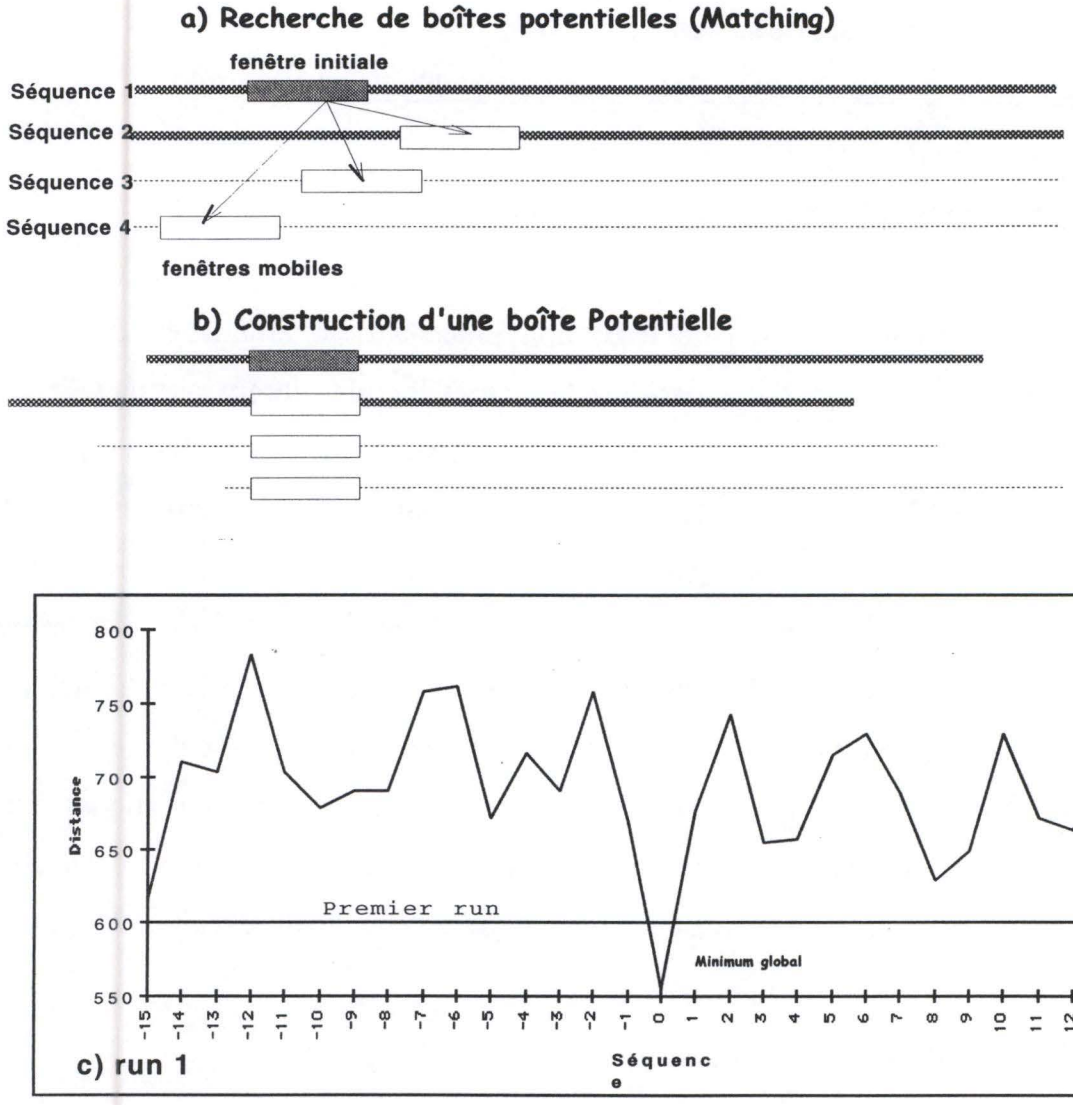


Figure I.6.3.

a)Représentation shématique des différentes comparaisons entre segments protéiques s'éffectuant dans le matching avec détermination du meilleur appariement.
b) Représentation shématique d'une boîte potentielle par rassemblement des meilleurs appariements
c) Shématisation du premier run du matching où l'on observe l'élimination de certaines boîtes potentielles en fonction des seuils établis lors du scanning.

on vérifie si le meilleur appariement trouvé répond positivement aux conditions élaborées lors de l'étape de paramétrisation. Si les conditions sont vérifiées, on obtient alors un appariement complet *(cfr. figure I.6.3)*. Après avoir défini l'ensemble des boîtes de segments similaires, ces boîtes potentielles sont classées par ordre de positionnement en allant de l'amont vers l'aval. Le *matching* constitue ainsi une collection d'appariements complets ayant passés les cribles statistiques et destinés maintenant à être triés de façon à pouvoir regrouper, dans l'alignement final, celles traduisant une similarité optimale.

- *Étape de triage, le screening*

Le *screening*, ultime étape de Match-Box, constitue une banque de boîtes à partir des appariements complets trouvés lors dès étapes antérieures. Cette banque est constituée de segments corrects et de segments incorrects. Cette sélection est accomplie selon trois critères :

* le plus grand nombre de résidus identiques
* la plus petite distance entre les segments
* la plus petite différence entre la position des fenêtres dans les deux séquences

Dans l'ensemble des boîtes la plus grande d'entre elles est considérée comme étant *a priori* la meilleure *(cfr. figure I.6.5)*. Le *screening* cherche toutes les boîtes compatibles avec la plus longue. Les boîtes compatibles sont des boîtes ou l'alignement de l'une n'empêche pas l'alignement de l'autre, et à l'inverse, l'incompatibilité entre deux boîtes est définie comme l'impossibilité d'aligner simultanément une boîte avec quelconque portion de l'autre *(cfr. figure I.6.4)*. Parmi toutes les boîtes compatibles trouvées, le *screening* choisit de nouveau la plus grande.

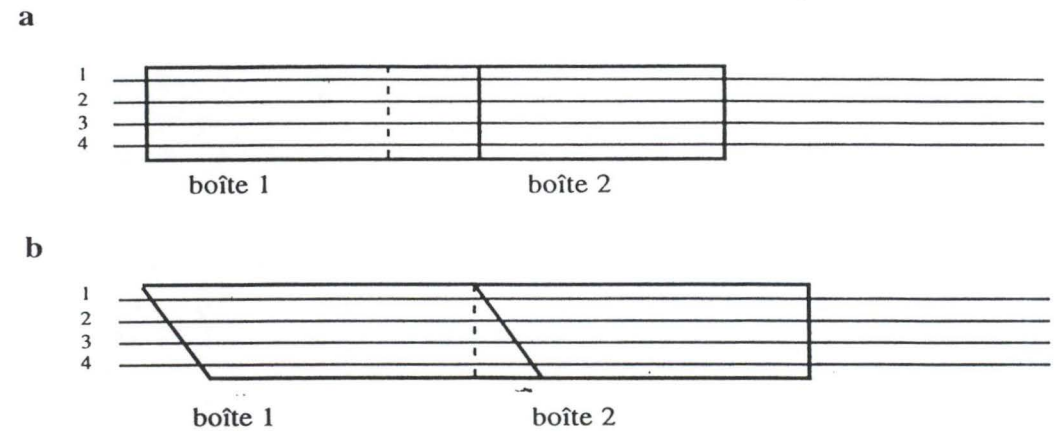


Figure: I.6.4.

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.*

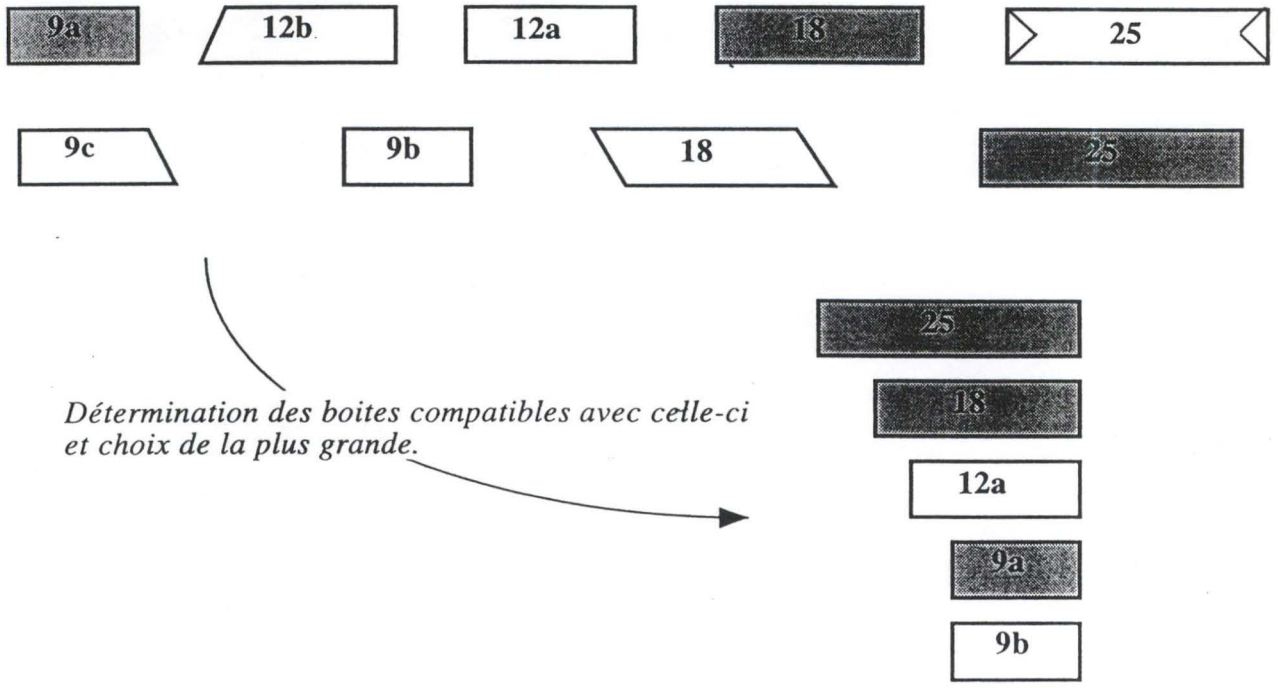


Figure: I.6.5

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.

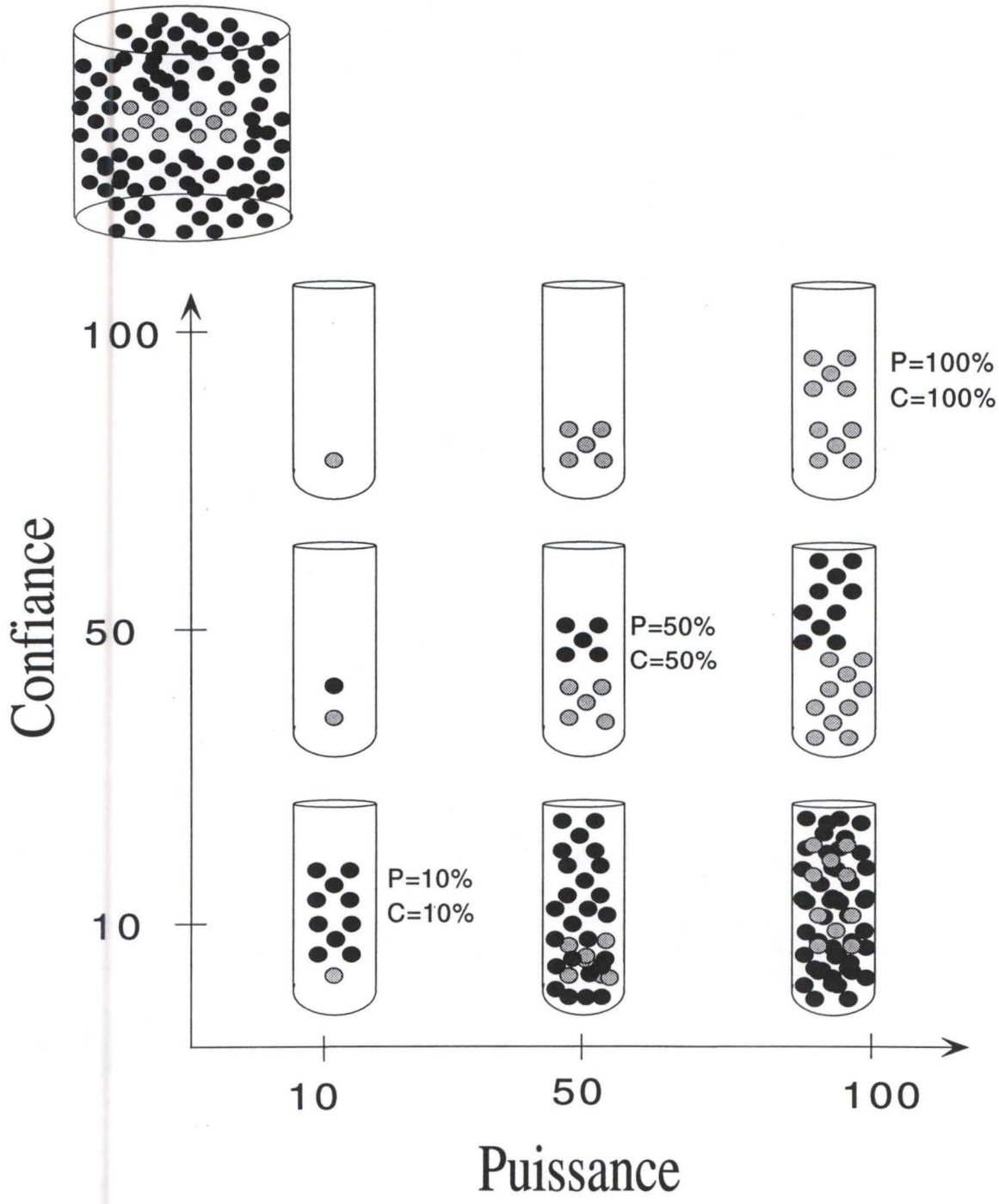


Figure I.6.6. *Diagramme représentant le concept de puissance et de confiance*

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. La puissance est représenteé par 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, est représentée par 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.

Il rassemble ces boîtes en les incluant entre elles, cette procédure recommencera et ainsi de suite. En d'autres termes, cette dernière étape permet l'agencement final des boîtes dans un alignement et conditionne la qualité du résultat final.

- Concept de puissance-confiance

La qualité des différentes méthodes d'alignement doit impérativement être évaluée sur base de deux critères : la puissance et la confiance. On ne peut pas dissocier ces deux critères pour décrire les performances d'un logiciel d'alignement multiple. La puissance fait référence à la capacité d'aligner le plus grand nombre de boîtes correctes dans un ensemble de protéines. La confiance est vue comme la proportion entre le nombre de positions correctement alignées dans les séquences et le nombre total de résidus alignés par le programme dans ces séquences. Parmi l'éventail de programmes d'alignement multiple, beaucoup délaissent la confiance pour bénéficier d'une puissance plus importante. On observe entre ces deux critères un phénomène de vase communiquant, c'est-à-dire qu'au plus un programme sera puissant, au plus il le paiera en confiance. Un programme ayant une grande confiance, alignera moins de résidus, mais ce qu'il alignera sera proportionnellement plus correct.
Sur un graphe de puissance en abscisse et de confiance en ordonnée, on représente l'efficacité d'une méthode d'alignement par un point appelé score d'efficacité. Ce score représente la distance entre ce point et le point idéal de valeur maximum (100,100) *(cfr. figure I.6.6)*. Ce score d'efficacité est calculé sur base de la puissance et de la confiance par la formule suivante :

$$\sqrt{(100 - puissance)^2} + \sqrt{(100 - confiance)^2}$$

Rem :

Jusqu'à présent aucune amélioration n'a permis à Match-Box d'augmenter simultanément sa puissance et sa confiance. Lorsque Match-Box gagnait uniquement en terme de puissance, on le payait dans la confiance de ses prédictions.

Attention qu'une augmentation de performance se marquerait sur le graphe puissance-confiance par une diminution du score d'efficacité puisque, la distance par rapport au point 100,100 diminuerait.

- **Comparaison de ClustalW et de Match-Box :**

* Avantages de ClustalW par rapport à Match Box :

En plus de sa rapidité (due à un degré d'optimisation très élevé), ClustalW est plus puissant que Match-Box pour réaliser un alignement multiple global de séquences protéiques.

ClustalW a la possibilité de fusionner plusieurs alignements créant ainsi des arbres phylogénétiques qui permettent de classer les protéines à aligner. Ces résultats ne sont pas présentés aussi clairement dans Match-Box.

Il peut aligner jusqu'à 300 séquences avec une longueur pouvant aller jusqu'à 5000 résidus, contrairement à la version serveur de Match-Box qui a été limitée actuellement à 50 séquences d'une longueur moyenne de 2000 résidus, cependant l'algorithme qu'il utilise peut en aligner davantage.

Clustal W n'utilise pas toujours la même matrice par défaut. Selon le pourcentage de similarité entre les séquences, il opte soit pour la famille des PAM ou soit celle des BLOSUM.

* Avantage de Match Box par rapport à ClustalW :

Match-Box aligne aisément des séquences présentant des homologies locales.

Globalement, il est moins puissant que ClustalW, mais jouit d'une meilleure confiance, c'est-à-dire que ce qu'il alignera sera proportionnellement plus correct que ce qui est aligné par ClustalW. Cette baisse de puissance et cette augmentation de

confiance proviennent du fait que Match-Box délimite des boîtes indiquant le niveau d'homologie entre les séquences : il évalue la fiabilité de toutes boîtes alignées, pour ne sélectionner que les meilleures d'entre elles. ClustalW ne donne quant à lui aucune indication sur le degré d'homologie existant entre les séquences, ni même sur la confiance de l'alignement proposé.

De plus, Match-Box ne se sert pas des *indels* comme un paramètre pour réaliser ses alignements, il propose plutôt un décalage entre les séquences en se basant sur des zones d'homologie découvertes entre certaines régions. Pour Match-Box, les *indels* sont un résultat inévitable de l'alignement. ClustalW introduit ces *indels* pour aligner, ensuite les sanctionne dans l'alignement. Or, ces décalages ne peuvent pas être justifiés et peuvent même le mener à réaliser des alignements tout à fait erronés lorsque la taille entre les séquences est très différente.

Chapitre II. Matériels et outils bioinformatiques

II.1 Support informatique

- Station de travail Silicon Graphics Octane :

Nos manipulations ont principalement été exécutées sur est une station de travail *Silicon Graphics* Octane. Ce type de station, développée par SGI, possède une architecture radicalement différente des autres ordinateurs et des stations classiques. Elle tourne sous le système d'exploitation UNIX (IRIX 6.5) qui est un système multi-utilisateurs et multi-tâches. Cette station de travail combine deux processeurs MIPS® R10000, 21 giga d'espace disque et un système graphique puissant. Les capacités multiprocesseurs cadencées à 225 MHz. de cette architecture vont permettre d'accélérer considérablement les temps calcul.

Pour éditer et traiter nos données chiffrées, nous avons principalement utilisé le logiciel de tableur Excel 98 développé par Microsoft, conçu pour des architectures plus classiques (Macintosh).

- Serveur Digital Alpha 4400:

Certains de nos tests ont été réalisés sur une station de travail de type Digital AlphaServer 4100 5/300 (TIMOUR centre de calcul FUNDP). Cette architecture

regroupe deux processeurs cadencés chacun à du 300 MHz et possède un espace disque de 24 giga. La station Digital Alpha tourne sous le système d'exploitation UNIX 4.0.

- Langage de programmation :

L'ensemble des scriptes et des procédures nous permettant de tester les performances de Match-Box, ont été programmés par Christophe Lambert, pour la quasi majorité dans le langage FORTRAN. Les initiales FORTRAN viennent, de « The IBM Mathematical FORmula TRANslation System ». Il a été initialement conçu pour simplifier la programmation de calculs numériques sur les plates-formes IBM 704 (Sammet J., 1954). La première version officielle du FORTRAN n'est apparue qu'au début de l'année 1957 et même si les programmes obtenus à partir de code FORTRAN étaient plus lents que ceux obtenus à partir de codes en langage machine, le FORTRAN s'est imposé auprès de la communauté scientifique : il était bien plus facile à écrire que le langage machine. Très vite, il est devenu possible de réutiliser des codes FORTRAN sur des plates-formes autres que celles d'IBM. En 1978, les spécifications du FORTRAN 77 furent adoptées.

- Réseau Internet :

Le développement des nouvelles technologies de l'information et des communications est en train de changer le paysage scientifique de notre société. Un exemple est l'utilisation d'Internet accélérant le processus de ces changements. Internet est un réseau de réseaux initialement destiné aux besoins d'échanges et de communication entre centres de l'armée, centres de recherche et universités américaines. Il a été constitué d'infrastructures fournies gratuitement par le département de la défense américaine et par les universités. Ces réseaux dispersés dans le monde entier et fédérés

sur la base d'un protocole commun de transmission des données : TCP/IP (Transmission Control Protocol / Internet Protocol) Sans être une norme à proprement parler, Ce protocole de transmission est un standard qui fournit un langage commun pour l'interopérabilité des différents types de réseaux locaux. De plus comme nous l'avons déjà signalé, vu la taille sans cesse croissante des banques de données, il devient impossible de contenir l'ensemble de ces données sur un site propre. Internet s'impose comme l'outil de communication entre les communautés scientifiques, offrant à n'importe quel chercheur, des possibilités de communication immenses.

1. *transfert de fichiers* :copier des informations depuis un ordinateur vers un autre en utilisant le réseau comme support de transmission ;
2. *partage de fichiers* : permettre l'utilisation d'un fichier stocké sur une machine distante
3. *messagerie électronique* : système de courrier informatisé beaucoup plus rapide et moins coûteux que le courrier postal, les messages électroniques peuvent en outre contenir des éléments multimédia (sons, images vidéo…)
4. *accès à l'information* : le couplage à des systèmes d'indexation et de recherche, facilite la collecte d'information ;
5. *impression* : le partage d'imprimante en réseau permet d'imprimer un document à distance ;
6. *exécution de commandes à distance* : un logiciel sur la machine du client peut utiliser la capacité de calcul d'une machine connectée au réseau afin de lui faire exécuter des opérations (qu'elle ne peut pas faire elle même faute de puissance ou de l'interface logicielle nécessaire)

- PDB (Banque de structures) :

PDB (*Protein Data Bank*) est une bibliothèque comprenant les déterminations expérimentales de structures de macromolécules biologiques. Ce service a été créé par

le *Brookhaven National Laboratory (Cambridge USA)* qui contient les coordonnées cartésiennes des atomes de protéines résolues par cristallographie, par résonance magnétique nucléaire et par modélisation. http://www2.ebi.ac.uk/pdb/

- DSSP :

Le programme *DSSP (Dictionary of Protein Secondary Structure)* est utilisé pour définir la structure secondaire d'une protéine à partir de ses coordonnées dans l'espace. Ce programme a été écrit par Kabsh et Sander, qui insistent sur le fait que leur produit n'est pas un logiciel de prédiction de structures. *DSSP* présente les caractéristiques géométriques, la structure secondaire ainsi que l'accessibilité au solvant d'une protéine en se basant sur les coordonnées spatiales des atomes composant la protéine (Kabsch, W. & Sander C., 1983).
http://swift.embl-heidelberg.de/dssp/

- PHD :

PHD est un serveur consacré à la recherche automatisée dans des banques de données protéiques. C'est un logiciel de prédiction de structure secondaire, d'accessibilité au solvant, d'hélice trans-membranaire et de leur topologie, sur base d'un alignement multiple. *PHD* est constitué d'un ensemble de sous programmes, ils en font un « super–outil » très puissant qui dans notre cas, établi rapidement un ensemble de caractéristiques sur des protéines à partir de leurs séquences. Les auteurs de *PHD* sont Rost B. et Sander C. Des structures de *DSSP* ont servi à définir celles qui devront être prédites par *PHD*. Son utilisation est très simple puisqu'il suffit à l'utilisateur de soumettre une séquence d'acides aminés ainsi que son adresse électronique plus

quelques paramètres. *PHD,* envoie un fichier contenant la structure secondaire ainsi que d'autres renseignements tirés à partir de la séquence fournie. Les différents sous-programmes inclus à *PHD* utilisent tous un système de réseaux neuronaux pour établir leurs prédictions.

La structure secondaire est prédite par *PHDsec* pour trois états (hélices, brin et *loop*) avec une précision de 72 % (Rost and Sander, 1993). Il fut évalué comme étant plus performant de 10 % que les méthodes n'utilisant seulement que l'information basée sur l'ordre des résidus.

PHD permet également de prédire l'accessibilité au solvant (*PHDacc*). Cette prédiction permet de savoir si un résidu se trouve plutôt exposé en surface ou au contraire confiné à l'intérieur de la protéine, en déterminant un coefficient de corrélation (corrélation entre l'accessibilité relative du solvant expérimentalement observée et prédit) (Rost and Sander, 1994). En sortie de *PHDacc,* on retrouve 10 états d'accessibilité relative. Ces états sont exprimés en unités de différence entre la prédiction par *l'homology modeling* (la meilleure méthode) et la prédiction au hasard (la plus mauvaise méthode). *PHDacc* offre quand même la possibilité de présenter ses résultats sous la forme résumée de trois états.

PHDhtm prédit les d'hélices trans-membranaires par un algorithme de programmation dynamique qui retrouve correctement quasi toutes les hélices trans-membranaires d'une séquence. Le problème de *PHDhtm* est qu'il prédit souvent des hélices trop longues. Celles-ci sont donc coupées par un filtre empirique. Selon Rost et ses collaborateurs, la prédiction finale serait exacte à environ de 95%. Le taux de faux positifs, c'est-à-dire, de protéines globulaires dans la réalité, qui sont prédites en tant qu'hélices trans-membranaires, est de l'ordre de 2% . http://www.embl-heidelberg.de/Services/sander/predictprotein/

- **ALIGN :**

C'est un programme d'alignement pairé global et optimal. Il fonctionne selon un algorithme décrit par E. Myers et W. Miller, et utilise par défaut la matrice de scores BLOSUM50 (Myers and Miller, 1988).
http://vega.crbm.cnrs-mop.fr/bin/align-guess.cgi

- **SAPS :**

SAPS, Statistical Analysis Protein Sequences, est un autre programme d'analyse de séquences qui évalue selon des critères statistiques un grand nombre de propriétés d'une ou plusieurs séquences protéiques notamment, la composition en acides aminés, la distribution des charges, les structures répétitives et la périodicité (Brendel *et al.*, 1992). Des dispositifs statistiques sont mis en œuvre dans SAPS pour trouver des valeurs significatives pour toute une série de caractéristiques, qui peuvent suggérer des régions prometteuses pour la recherche expérimentale. Le programme trouve également son application dans la description des familles de protéines ainsi que pour expliquer la diversification des groupements de protéines. Le programme SAPS a été développé dans le groupe du prof. Samuel Karlin à l'université de Stanford
http://www.infobiogen.fr/services/analyseq/cgi-bin/saps_in.pl

- **MWCALC :**

MWCALC est un logiciel également destiné à l'analyse de séquences protéiques, calculant pour une séquence protéique donnée, la composition en nombre et en masse en acides aminés, la masse moléculaire totale, l'indice de polarité (pourcentage

d'acides aminés hydrophiles), le point isoélectrique théorique, la densité optique à 260 et 280 nm et la composition atomique. Un ensemble de valeurs de référence sont utilisées par le programme pour établir les caractéristiques d'une protéine donnée.
http://www.infobiogen.fr/services/analyseq/cgi-bin/mwcalc_in.pl

- SAS :

SAS est l'abréviation de « Statistical Analysis System ». Ce logiciel, développé par la SAS Institute Inc., est un ensemble de routines complémentaires intégrés dans un même environnement. *SAS* offre la possibilité de faire, en une seule étape, toute série de traitement statistique suite à l'encodage d'un scripte. Ce logiciel nécessite une bonne connaissance de ces scriptes pour pouvoir profiter au maximum de son potentiel d'analyse. Cependant ce programme devient incontournable lorsqu'il est question de traiter des tableaux de données numériques d'une taille très importante.

- STATISTICA :

STATISTICA est un logiciel développé par le groupe Starsoft, spécialisé dans le domaine des statistiques. Il offre une puissance et une vitesse très élevée pour traiter de vastes fichiers de données. Il comprend une série de procédures statistiques complètes, totalement intégrées avec des graphiques de la plus grande qualité (exemple : graphe de dispersion en 3D). *STATISTICA* nous servira comme outil de visualisation et d'interprétation des données.

II.2 Support protéique

- Les cas-tests :

Les cas-tests sont des familles de protéines dont les séquences et les structures sont connues. Ces familles constituent un outil privilégié pour les prédictions de régions structurellement conservées et peuvent permettre d'évaluer une méthode d'alignement à partir de la capacité de la méthode à retrouver des motifs structuraux partagés dans toutes les protéines constituant un cas-test, mais en tenant compte uniquement de l'information qui réside dans les séquences primaires. Ces cas-tests possèdent un certain nombre de protéines alignées entre elles. Nous intéresserons uniquement à ceux présentant au minimum trois séquences, puisque Match-Box réalise des alignements non pairés. Les cas tests ont été sélectionnés, notamment au niveau des structures secondaires, pour représenter au mieux l'ensemble des problèmes rencontrés dans une procédure d'alignement. Il existe donc une grande variabilité au sein d'un ensemble de cas-tests. Cela signifie qu'il existe des cas tests plus facile à aligner que d'autres qui le seront très difficilement. La difficulté à réaliser un alignement dépend principalement du taux de similarité entre les protéines d'une famille. Les cas-tests les plus simples à aligner possèdent plus de 35% de similarité. Par contre certaines de ces familles ne se verront alignées que par quelques programmes d'alignement très performants. C'est le cas pour les cas-tests présentant moins de 25% de similarité entre eux mais possédant néanmoins des régions structurellement alignées. La réalisation d'un alignement de séquences permet donc de mettre en évidence les régions suffisamment similaires pour prédire l'existence de régions conservées d'un point de vue structural. Pour nos cas-tests, nous connaissons les coordonnées spatiales de leurs atomes, mais le principe fondamental de leur alignement reste le même que celui des alignements de séquences. Ici, la connaissance des coordonnées spatiales des atomes des résidus élargit simplement le nombre de critères potentiels de comparaison entre des fragments de

structures. Aligner des protéines de structures connues revient à aligner leurs séquences sur base de critères de similarités structurales.

II.3 Les matrices de score

Les familles les plus utilisées ont été celles de PAM, BLOSUM, GONNET, JOHNSON très généralistes.*(cfr. Introduction, I.2 «Les matrices de scores»)*. À côté de cela, il existe toute une série de matrices de scores à haute spécificité locale, établies sur base de facteurs spatiaux. Certaines de ces matrices sont spécifiques pour l'alignement d'hélices, d'autres pour l'alignement de brins ou encore sensibles aux résidus se trouvant dans le noyau inaccessible au solvant ou au contraire aux résidus exposés où l'accessibilité au solvant est élevée.

Chapitre III. Objectif du mémoire

Ce travail a pour objectif d'augmenter les performances d'un programme d'alignement multiple, Match-Box et plus précisément l'étape du *matching* au cours de laquelle l'algorithme attribue un score à la comparaison de deux acides aminés en fonction de leur similarité en faisant référence à une matrice de scores (*cfr. description de Match-Box)*. Seule une librairie de vérité et un ensemble de matrices de scores localisées dans un support informatique constitue notre terrain d'investigation pour améliorer le *matching*. Ce mémoire trouve son point de départ au sein d'un paradoxe. En effet, l'ensemble des méthodes d'alignement de séquences protéiques ont toujours été limitées à l'usage simultané d'une seule matrice de scores, alors que le « prix » du remplacement d'un acide aminé par un autre, est différente suivant son emplacement dans la protéine repliée : exposé, enfoui, intégré dans une hélice, un brin, un coude…

Notre challenge sera de faire progresser Match-Box en accroissant sa puissance et simultanément sa confiance, ce qui n'a encore jamais été réalisé auparavant puisque chaque amélioration tentée jusqu'ici apportait au mieux un gain d'un des deux critères.

Au cours de ce travail nous essayerons notamment d'associer des matrices de scores à une circonstance spatiale particulière, ceci en vue d'implémenter à Match-Box un nouveau *matching* utilisant de façon autonome plusieurs matrices à la fois au cours de cette étape.

En améliorant ainsi cette méthode d'alignement multiple, on fournira des paramètres essentiels aux méthodes ayant recours à ce type d'alignement. Par exemple, les techniques de prédiction de structures secondaires qui permettront de mieux cibler les régions cruciales destinées à une utilisation plus expérimentale (clonage, mutagenèse dirigée…).

Chapitre IV. Recherche du meilleur incrément pour une batterie de matrices de scores

IV.1 Méthode

Notre investigation essayant d'augmenter les performances du *matching* a consisté en l'optimisation d'une valeur de paramétrage propre à chaque matrice utilisée lors du *matching*. Ce paramètre appelé incrément, joue un rôle dans la recherche du meilleur appariement pour chaque séquence impliquée dans l'alignement. Il est bon de rappeler que le *matching* fonctionne en quatre étape (*run*) dont le premier balayage traite toute la longueur d'une séquence avec un seuil très sévère, ensuite la zone du balayage se restreint en relâchant le seuil. Ce seuil déterminé lors d'une analyse statistique préliminaire, a pour rôle de limiter le bruit de fond par rapport au signal.
L'incrément étudié, intervient dans le relâchement du seuil au sein d'un même *run*, il dimensionne le pallier qui est utilisé lors de ce relâchement. Cette valeur n'a pas encore fait d'une investigation afin de l'optimiser et jusqu'à présent, Match-Box utilise une valeur fixée arbitrairement à 10 pour la matrice de score BLOSUM62. Choix de 10, puisque empiriquement, cette valeur semblait être un bon compromis entre le temps de calcul et la détection du meilleur incrément. Nous nous sommes demandés, si dix était réellement la valeur optimale et sinon dans quelles mesures son optimisation augmenterait-elle les performances du *matching*.

Nous avons réalisé l'interprétation d'une étude préliminaire effectuée au laboratoire (communication personnelle ; Christophe Lambert). Sur une batterie de trente-trois cas-tests, Match-Box a été exécuté avec 47 matrices de scores différentes, avec pour

chacune d'entre elles une valeur d'incrément allant de 3,0 à 18,0 par pas de 0,1. À l'issue de ce programme qui nécessita cent jours de calcul sur TIMOUR (cfr. Digital Alpha Server 4100 5/300), nous disposions d'un fichier informatique reprenant le pourcentage de puissance et de confiance pour l'ensemble des situations investiguées. Nous avons calculé une valeur de score d'efficacité en moyenne sur 33 cas tests pour chaque valeur d'incrément *(cfr tableau IV.1.1)*.

Incrément	Blosum30	Blosum40	Blosum62	PAM440	Gonnet
3,0	81,4	75,9	71,9	86,4	72,8
3,1	81,5	76,0	71,7	86,4	72,5
3,2	81,4	76,0	71,1	86,4	72,6
3,3	81,0	76,2	71,4	86,4	72,5
3,4	81,2	75,9	71,9	86,4	72,3
3,5	80,7	75,8	71,5	86,4	72,2
3,6	80,2	75,6	70,8	86,4	71,9
3,7	80,6	75,7	70,8	86,5	72,0
3,8	80,9	74,8	71,2	86,5	71,7
3,9	81,2	74,4	71,1	86,5	72,5
4,0	80,8	74,5	71,2	86,5	72,4
:	:	:	:	:	:
:	:	:	:	:	:
8,1	.	.	.	83,3	.
:	:	:	:	:	:
:	:	:	:	:	:
14,2	:	:	62,7	:	:
:	.	.	.	.	.
:	:	:	:	:	:
16,1	70,6	65,6	64,6	84,51	64,3
16,2	70,6	65,4	64,3	85,00	64,4
16,3	70,7	65,8	64,1	84,93	64,1
16,4	71,0	65,8	64,1	84,86	63,4
16,5	71,0	65,6	64,1	85,78	63,4
16,6	70,8	65,8	64,3	85,71	63,7
16,7	70,8	65,8	64,3	85,64	63,7
16,8	70,5	65,8	64,3	85,64	63,6
16,9	70,5	65,8	64,3	85,14	63,7
17,0	70,6	65,8	64,3	84,72	63,6
17,1	70,7	65,5	64,0	84,72	63,1
17,2	70,7	65,5	64,0	84,65	63,1
17,3	70,3	64,8	64,0	84,57	65,0
17,4	70,3	64,8	64,0	84,57	65,0
17,5	70,6	64,8	64,3	84,57	65,4
17,6	70,4	65,0	64,3	84,50	65,4
17,7	70,4	65,0	64,3	84,50	65,4
17,8	71,0	64,7	64,4	83,94	66,0
17,9	71,0	64,7	64,4	83,94	66,0
18,0	71,0	64,7	64,4	83,94	66,0

Tableau IV.1.1

À titre d'exemple, ce tableau reprend différentes valeurs de score d'efficacité pour cinq (quarante-sept au total) matrices de scores. Le chiffre ombré représente le meilleur score d'efficacité (distance la plus faible) obtenu à une certaine valeur d'incrément selon le type de matrice de scores observée.

Comme nous pouvons le constater, la valeur optimale de l'incrément pour l'ensemble des matrices étudiées, est comprise entre les bornes 8,1 pour la matrice *PAM440* à 18,0 pour plusieurs matrices dont BLOSUM40. Manifestement, cette valeur d'incrément est très variable pour toutes les matrices de scores utilisées. Chaque matrice possède une valeur d'incrément optimale qui lui est propre.

Examinons en particulier, la matrice BLOSUM62 utilisée par défaut par Match-Box avec un incrément arbitraire fixé à 10. On observe que la valeur optimale est de 14,2 et non de 10 *(cfr. tableau IV.1.1).* On peut suivre l'évolution du score d'efficacité, pour cette matrice (*cfr. figure IV.1.1*).

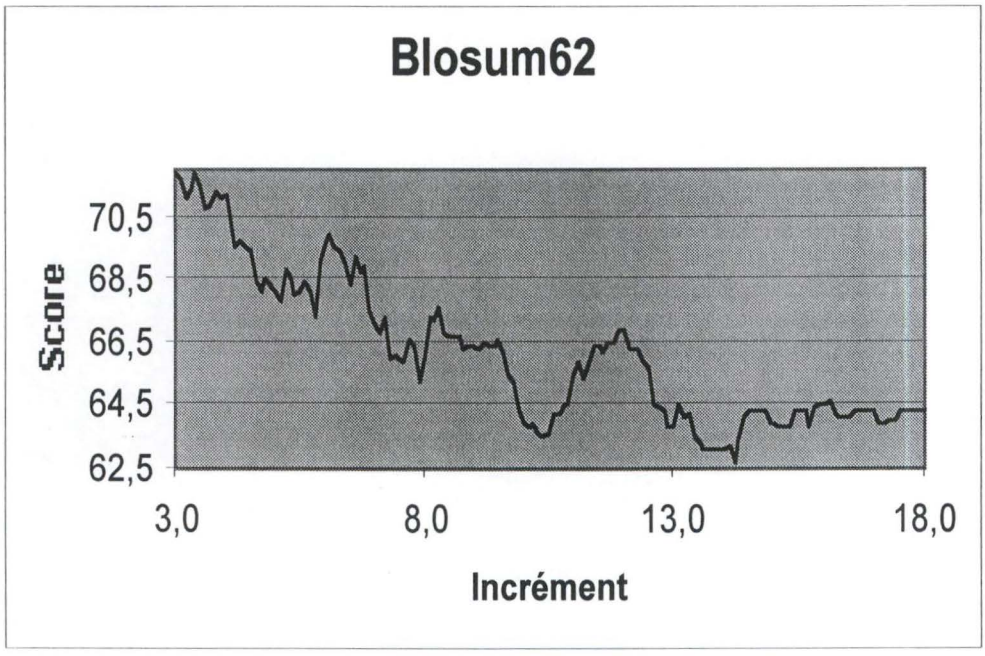


Figure IV.1.1

Ce tableau représente l'évolution du score d'efficacité de la matrice BLOSUM62 en fonction de valeurs d'incrément allant de 3,0 à 18,0. On observe une diminution du score d'efficacité pour un incrément pour 14,2 (le minimum) ainsi qu'aux alentours de 10 (celui-ci choisit arbitrairement)

On remarque que pour la matrice de score BLOSUM62, au cours de l'augmentation de l'incrément, sa valeur est optimale à 14,2 ce qui se traduit par une diminution du score à cette valeur. Nous remarquons également qu'aux environs de la valeur d'incrément de 10 le score d'efficacité est bas. Observons maintenant, les performances de Match-Box obtenues pour ces deux valeurs d'incrément *(cfr.tableau IV.1.2)*.

	Incrément	Puissance	Confiance	Score
MB avec Inc.=10	10,0	56,3	53,4	63,9
MB avec Inc.Optimum	14,2	54,7	56,6	62,7

Tableau IV.3

Ce tableau illustre les performances atteintes par Match-Box selon les deux critères puissance et confiance calculés en moyenne sur 33 cas tests pour la matrice BLOSUM62 avec la valeur d'incrément de 10 et de 14,2. Ce tableau illustre également l'efficacité en terme de distance le séparant du maximum de puissance et de confiance simultané (score d'efficacité).

Au vu de ce tableau, les performances de Match-Box augmentent sensiblement avec ce paramètre optimisé pour *BLOSUM62* . On observe une perte de 1,6% en puissance contre un léger gain de 3,2% en confiance. Ce qui améliore globalement le score d'efficacité de 1,2% .

IV.2 Conclusion

Cette méthode à permis d'améliorer légèrement les performances de Match-Box, qui utilisera dorénavant pour chaque matrice sa valeur d'incrément optimisée et non plus 10 pour comme il le faisait pour toutes les matrices qu'on lui fournissait. Pour BLOSUM62, la valeur 10 utilisée par défaut par Match-Box n'était pas le choix optimal mais néanmoins un choix arbitraire empirique très correct.

Pour les autres matrices de scores et même celles dont la valeur optimale d'incrément s'avère plus éloignée de 10, la tendance au gain mineur de performances est généralisée. (de l'ordre de 2% sur les 33 cas tests). Par contre comme nous l'avons déjà signalé, on retrouve à nouveau ce phénomène de vases communiquant entre la puissance et la confiance. Lors de ce test, le moindre bénéfice en confiance, si minime soit-il, parvient encore à être corrélé par une perte en puissance. Nous allons maintenant chercher une autre stratégie pour augmenter d'avantage les performances de Match-Box et aller à l'encontre de ce phénomène de vase communicant.

Chapitre V. Etude de faisabilité sur les performances de Match-Box en changeant de matrices de scores en fonction du cas-test

V.1. Sur trente-trois cas-tests

- méthode

Nous sommes partis d'une observation très simple, Match-Box comme tous les programmes d'alignement multiple, utilise de façon simultanée une seule matrice scores pour réaliser un alignement donné. Pourtant, il existe une vaste gamme de matrices de similarité. Dès lors, nous avons cherché si Match-Box obtenait toujours les meilleures performances avec la matrice *BLOSUM62* et sinon dans quelles mesures, le choix ciblé de la meilleure matrice augmenterait les performances de Match-Box.

Nous sommes repartis du fichier généré par la première méthode, dans lequel nous avons sélectionné pour chacun des trente-trois cas tests, la matrice donnant le score d'efficacité (calculé) le plus petit, c'est-à-dire la matrice offrant les meilleures performances *(cfr. tableau V.1.1)*.

CAS TEST	MATRICE	Puissance	Confiance	Score
ace.test	blosum62	71	76,5	37,3
adk.test	pam120	71,1	45,4	61,8
amg.test	blosum30	61,3	53,5	60,5
aprot2l.test	johnson92	17,3	37,5	103,7
aprotease.test	johnson92	72,5	85,8	30,9
cys.test	pam110	29,5	34,9	96
cytc.test	blosum60	84,9	74,5	29,6
fabp.test	blosum90	58,3	63,2	55,6
flav.test	pam250	84,8	68,5	35
glo.test	johnson92	53,3	66	57,8
hip.test	blosum100	88,4	71,7	30,6
hth.test	blosum100	0		
immuc.test	gonnet	79,7	58,6	46,1
immuv.test	gonnet	94,8	71,6	28,9
lipoc.test	blosum90	61	56,5	58,4
ltn.test	blosum40	95	82,7	18
lyzlac.test	blosum65	90,3	89,5	14,3
maldh.test	blosum70	86,3	69,6	33,3
peroxydase.test	pam230	89,8	68,1	33,5
phycoc.test	pam140	98,1	98,1	2,7
pkinasest.test	blosum65	39,1	92,6	61,3
pkinaset.test	pam190	90	97,1	10,4
plasto.test	blosum80	72,6	77,6	35,4
ricinb.test	gonnet	64	53,3	59
serbact.test	pam240	86,5	78,8	25,1
sh3.test	blosum40	97,9	80,7	19,4
sprotm.test	blosum62	89,1	85	18,5
sprotmb.test	blosum85	44,5	69,9	63,1
subtilisin.test	blosum55	86,5	79,3	24,7
super.test	pam100	99,3	98,6	1,6
vcoatpp.test	blosum55	56,5	77,2	49,1
vcoatppr.test	blosum100	0		
vcoatpr.test	blosum40	14,1	17,8	118,9
MOYENNE		**67,5**	**66,5**	**42,6**

Tableau V.1

Ce tableau reprend pour les 33 cas tests la matrice la plus adaptée à chacun d'eux et les performances obtenues.

Nous constatons la grande diversité des matrices utilisées pour obtenir systématiquement les meilleurs résultats.

Nous allons comparer les performances de Match-Box obtenues avec ces diverses matrices par rapport à celles obtenues lorsqu'il utilise systématiquement BLOSUM62. Nous connaîtrons à l'issue de cette analyse dans quelles mesures les performances augmenteraient-elles *(cfr. figure V.1.1).*

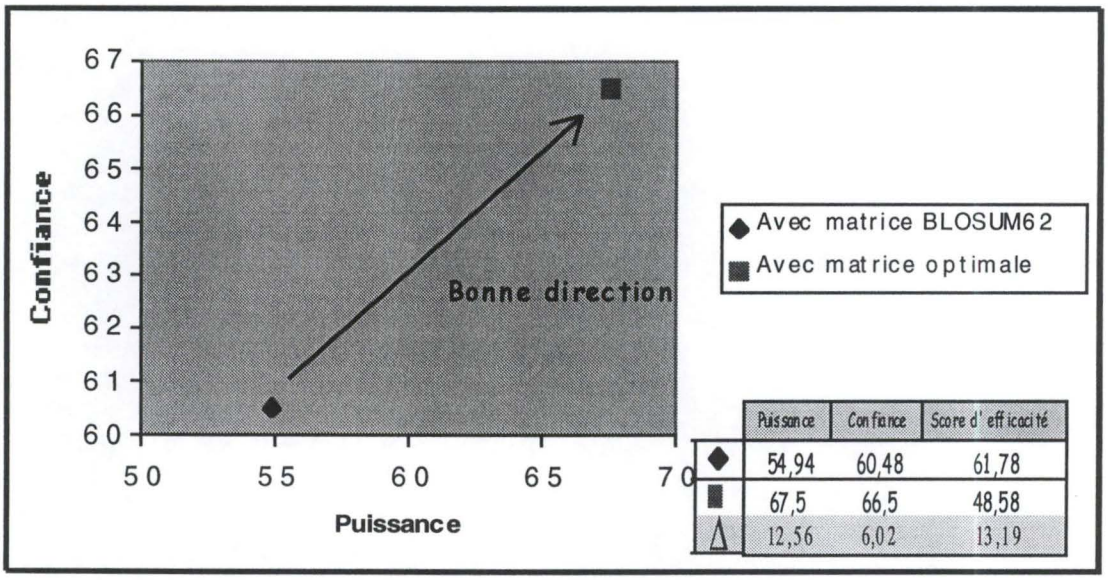


Figure V.1.1

Figure illustrant un graphe de puissance - confiance représente l'évolution des perforances entre deux ver - sion de Match-Box pour 33 cas tests. Une utilisant la matrice de scores BLOSUM62 et une autre hypothé - tique dans laquelle il utilise la matrice donnant le meilleur résultat.

Nous constatons pour la première fois une augmentation simultanée de puissance et de confiance. Lorsque Match-Box utilise la matrice la plus appropriée à chaque cas-test, on a augmentation de 12,6% en puissance et de 6,02 en. confiance. Les deux critères augmentent simultanément ce qui rapproche le score d'efficacité de 13% du point maximum (100% confiance, 100% puissance). À notre agréable surprise, le phénomène de vase communiquant entre la puissance et la confiance a disparu.

- Conclusion

Cette étude de faisabilité nous a permis de voir le gain éventuel de puissance et de confiance si Match-Box connaissait a priori la meilleure matrice a utiliser pour chaque cas test. Ce gain de performances est important et concerne puissance et confiance en même temps.

Jusqu'à présent, ce travail a été effectué *a posteriori* . Nous allons donc concentrer notre effort de recherche sur la possibilité de déterminer ce choix de la meilleure matrice de similarité à partir de caractéristiques tirées *a priori* sur base des séquences de cas tests. Pour réunir un ensemble de conditions suffisamment représentatives il est indispensable d'élargir notre batterie de cas tests.

V.2. Développement d'une banque de septante-huit cas-tests

Pour dresser un inventaire des alignements considérés comme corrects il faut disposer d'une série d'alignement de référence. Il existe plusieurs critères de distance envisageables pour évaluer la ressemblance entre des segments de structures (SS). La mesure de référence la plus connue et la plus fiable est la distance r.m.s. (root means square). Des segments de structures connues peuvent être superposés en minimisant les distances qui les séparent (Greer, 1981). La mesure r.m.s. correspond à la distance euclidienne moyenne minimale entre les atomes des deux squelettes protéiques impliqués dans la comparaison. Les atomes du squelette (avec ou sans les atomes d'oxygène) ou les carbones α sont superposés et la distance moyenne entre deux segments est calculée de la façon suivante :

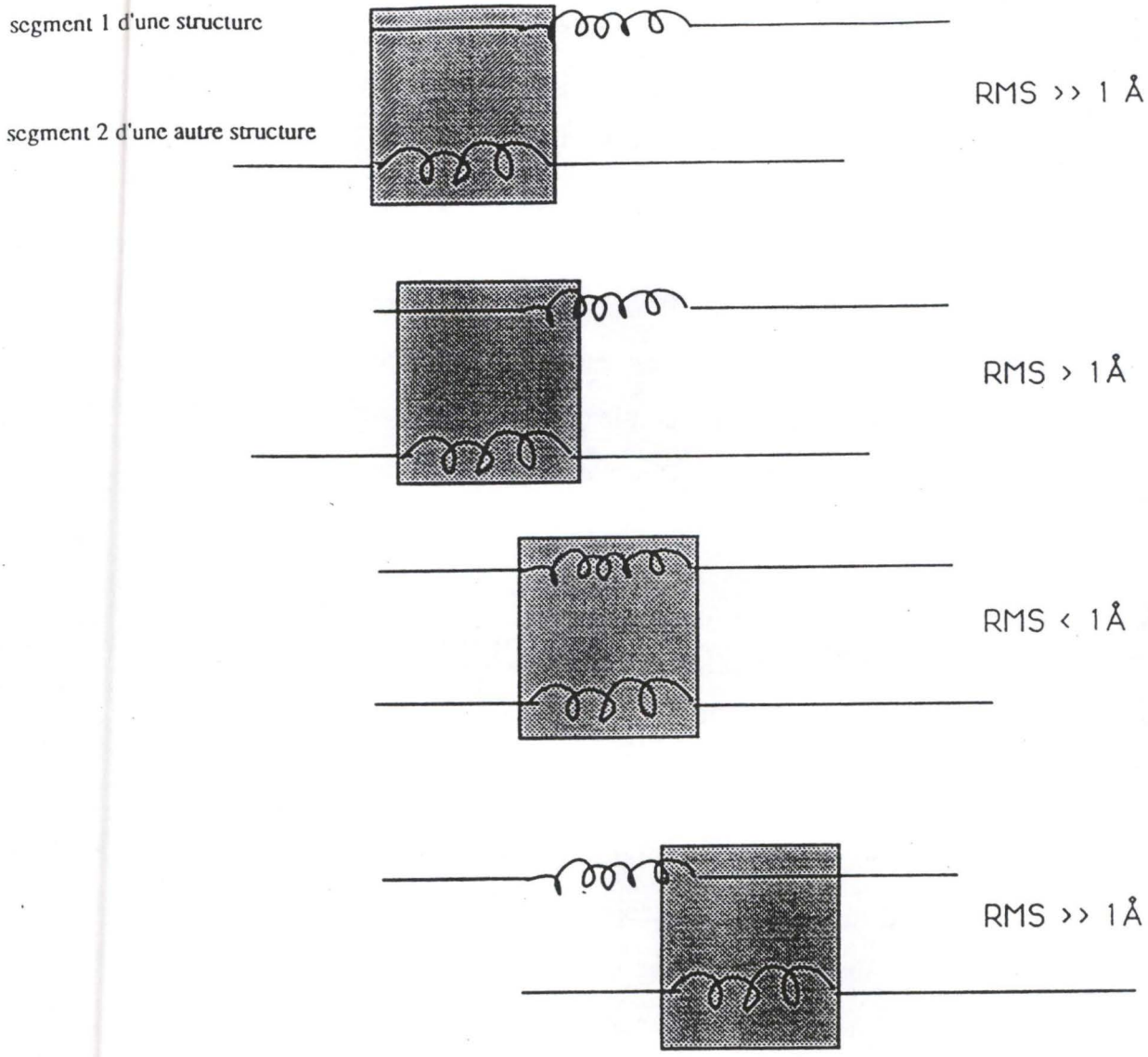


Figure V.2.1.

Schéma de la superposition de deux segments de deux structures. Le carré grisé représente le calcul de la distance entre les deux bouts encadrés des segments.

$$RMS = \min \sqrt{\sum_{i=1}^{aw} \frac{(x_{i1} - x_{i2})^2 + (y_{i1} - y_{i2})^2 + (z_{i1} - z_{i2})^2}{aw}}$$

Où : **a** est le nombre d'atomes par résidu

w est le nombre de résidus dans chaque segment

x_{ij}, y_{ij} , z_{ij} , sont les coordonnées cartésiennes d'un atome i dans un segment **j** (J = 1,2)

De manière à minimiser les distances entre deux segments, un des segments est déplacé par translation et/ou rotation, ce qui n'affecte pas la forme générale de la structure mais uniquement les positions relatives aux deux segments. Le r.m.s. cumulé pour l'ensemble des atomes considérés est minimisé par un algorithme des moindres carrés non linéaire. Selon Unger (Unger et al. 1989), la limite de 1 Angstrom permet généralement de distinguer les segments possédant une similarité structurelle très rarement atteinte, pour ne pas dire jamais, par des segments pris au hasard dans des protéines de structure connue. Cette limite est utilisée pour estimer la similarité structurelle de deux segments *(cfr.figure V.2.1)*. Quand les structures sont correctement superposées, les régions qui se chevauchent très bien sont appelées "*Structurally Conserved Régions*": SCR. Entre ces segments fort semblables se trouvent des régions qui diffèrent beaucoup plus et sont appelées régions variables (VR). La détermination de ce qui est considéré comme correctement aligné dans l'alignement en se basant sur la distance r.m.s. semble de loin la meilleure méthode, mais elle est également la plus longue et fastidieuse à mettre en œuvre, nécessitant un raffinement à l'œil sur les écrans.

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 V.2.2. : *Représentation en gras de ce qui est aligné sans gap.*

Figure V.2.3 : *Représentation en gras des zones où la conservation de la structure secondaire est de 100%.*

1) La première batterie de 20 cas-tests dont nous disposions avait été établie selon la mesure de distance r.m.s . L'encodage des références de ces vingt familles de protéines a été réalisé au laboratoire (Briffeuil *et al.*, 1998).

2) Une seconde batterie de cas-tests a été utilisée. Elle est constituée des vingt premiers cas tests définis par la distance r.m.s., ainsi que de treize cas tests supplémentaires tirés de la littérature. Les premiers tests ont été établis sur ces trente-trois cas-tests.

3) Une troisième batterie fut constituée à partir de nonante-six familles tests ayant été structurellement alignées (Overington. *et al.*, 1996). Ces alignements ont été résolus par une technique de programmation dynamique. Nous avons éliminé de cette batterie tous les cas tests comptabilisant moins de trois séquences (Match-Box réalise un alignement multiple). Nous ne disposions pas de la mesure de la distance r.m.s. pour cette batterie, et l'établir pour cette nouvelle batterie aurait été beaucoup trop long. Dans le cadre de notre travail, il a donc fallu se référer à autre chose que la distance r.m.s., comme référence de vérité. Il faudra obligatoirement s'en remettre à une autre référence de vérité tenant compte des structures secondaires.

La première alternative serait de prendre comme référence de vérité uniquement ce qui dans les alignements de structures est aligné sans *gap (cfr. figure V.2.2)*. Nous surestimerons alors légèrement ce qui serait considéré comme aligné correctement par la distance r.m.s, cette vérité est donc plus lâche, puisque tout ce qu'il est possible de prendre est sélectionné.

Une alternative beaucoup plus stricte cette fois serait de prendre comme vrai dans les alignements de structures les régions correspondant à 100 % de conservation *(cfr. figure V.2.3)*.

Les divers tests se baseront sur l'une ou l'autre de ces deux types de vérités, en espérant que les résultats que l'on pourrait escompter sur base de la vérité r.m.s. seront intermédiaires entre ces deux extrêmes.

Cas Test	Matrice choisie	%Hélice	%Brin	%Boucle	%Enfoui	%Exposé	%Intermed.	Poids Mol. Moyen	ex. de polarité	Point Isoélectrique	Minimum d'identité	Maximum d'identité	Moyenne d'identité	% Ala	% Gly	% Met	% Ser	% Cys	% His	% Asn	% Thr	% Asp	% Ile	% Pro	% Val	% Gln	% Lys	% Glu	% Trp	% Phe	% Leu	% Arg	% Tyr
aa.test	pam230	0.26	0.19	0.55	0.50	0.45	0.04	53706.56	46.096	5.08	24.90	66.60	40.68	8.4	8.3	1.9	8.2	1.2	1.8	7.6	8.4	7.9	5.6	3.8	5.9	2.6	3.7	3.2	1.9	4.1	6.8	2.8	6
ace.test	blosum50	0.27	0.15	0.58	0.51	0.46	0.03	58691.365	42.55	5.58	28.00	42.10	33.50	7	9.4	2.6	7.9	1	1.9	6.4	5.3	5.9	4.7	6	6.2	4.4	4.2	3.1	1.5	6	9.2	3.6	3.9
adk.test	pam130	0.47	0.14	0.39	0.46	0.50	0.04	22529.309	48.39	8.16	18.60	20.90	26.75	6.4	8.3	2.2	5.3	0.5	1.4	2.7	6.4	6.3	5.7	4.5	7.8	8.2	8.9	3.3	0.5	3.5	8.9	5.8	3.3
annexin.test	pam500	0.74	0.03	0.23	0.42	0.54	0.04	106732.32	49.89	5.42	76.70	77.40	81.67	8.3	6.8	2.4	6	0.4	1.2	2.3	7.4	7.6	5.8	1.5	5.3	8.6	6.8	3.5	0.3	4.1	11.7	6.5	3.5
asp.test	blosum80	0.05	0.47	0.48	0.56	0.41	0.02	35308.267	43.12	4.6	23.20	40.40	37.14	6.1	11.1	1.3	10.5	1.5	1	4	8.4	6.7	6.1	4.6	7.3	3.5	3.2	4.3	1.2	5.4	7.5	1.5	4.9
az.test	pam100	0.09	0.42	0.49	0.51	0.46	0.03	11667.594	43.23	5.23	13.50	27.30	32.05	9.1	9.4	3.4	6.3	1.4	2.9	4.9	6	6.4	5	4.9	10	6.4	7.2	2.3	0.5	4.4	5.6	0.8	3.1
bowman.test	pam500	0.03	0.07	0.90	0.32	0.63	0.05	6292.1706	52.98	6.62	39.70	56.30	46.10	2.4	0.6	3	13.7	21.4	1.2	3.6	3.6	9.5	2.4	8.9	0.6	4.2	7.7	4.2	0	2.4	2.4	5.4	3
cbp.test	pam120	0.58	0.02	0.40	0.40	0.58	0.02	18360.819	51.66	4.43	16.80	20.40	32.97	6.9	7.3	5.7	4.2	0.6	0.6	4.8	6.2	11.6	5.2	1.8	4.9	11.3	6.3	3.1	0.9	6.7	6.8	3.6	1.6
cryst.test	pam290	0.04	0.39	0.57	0.46	0.46	0.08	16224.709	48.71	6.93	18.40	21.50	42.52	3.1	8	2.4	8.1	2.6	2.7	4.6	2.9	6.6	3.9	5	5.8	5.9	2.9	5.9	1.9	4.5	7.2	8.8	7
cys.test	blosum45	0.16	0.25	0.59	0.49	0.47	0.04	2448.455	41.31	7.55	27.50	28.80	41.27	6.6	13.1	0.7	6.8	3.9	1.7	4.9	5.1	4.2	5.9	5	7.9	5.1	5.3	4	2.6	2.1	4.5	4.1	6.5
cyt3.test	pam140	0.12	0.12	0.76	0.35	0.62	0.03	11822.55	52.16	8.85	29.90	45.40	45.75	10	8	1.8	5.2	7.3	8	2.3	6.2	7.1	0.7	5.2	6.8	4.6	16.9	1.6	0	1.8	3.9	0.5	2.3
cyt5.test	blosum85	0.47	0.04	0.50	0.47	0.50	0.04	8434.38	38.53	6.08	24.20	26.40	50.40	15.5	11	2.1	5.2	3	1.8	3.4	2.7	6.1	3.7	7	6.4	5.8	9.5	3.4	2.1	1.5	7.9	0.9	0.9
cytc.test	blosum30	0.32	0.04	0.64	0.44	0.53	0.03	12089.232	48.49	9.28	26.20	54.10	43.16	9	10.8	1.9	3.8	1.9	2	5.1	7	5.1	4.3	4.2	4.6	6.6	14.5	2.5	1	3.9	5.7	1.7	4.1
cyto.test	pam450	0.77	0.00	0.23	0.55	0.40	0.05	16791.9	38.7	6.64	74.70	76.40	77.80	10.4	7.2	2.4	6.5	3	3.3	0.9	4.3	3	2.6	5.2	3.3	5.2	2	9.1	0.9	4.1	20.4	4.3	1.7
dhfr.test	pam220	0.15	0.29	0.55	0.48	0.49	0.03	19641.984	46.59	6.79	26.30	73.70	35.42	6.2	6.1	2.6	5.2	0.6	2.8	4.1	4.9	6.7	5.5	5.8	7.8	7	7.4	3.8	2.2	4.5	8.7	4.8	3.3
egf.test	pam460	0.00	0.23	0.77	0.38	0.58	0.04	5297.908	47.57	5.31	21.60	28.40	29.24	3.1	10.4	1	6.6	12.5	3.5	5.9	5.2	6.2	2.1	3.8	5.2	7.6	5.6	3.8	1	4.2	5.2	3.1	3.8
fer2.test	blosum85	0.13	0.32	0.55	0.47	0.49	0.04	10470.73	47.84	4.18	48.00	51.00	60.21	8	7.4	0.4	7	4.9	1.4	1.4	8.5	10.3	6.2	3.7	6.4	10.1	3.9	3.9	0	2.7	7.8	1.2	4.5
fer4.test	pam500	0.08	0.14	0.79	0.52	0.46	0.02	7597.23	38.98	3.95	15.90	23.50	28.42	10.5	5.1	1.4	3.7	9.9	0.6	4	2.3	12.7	8.8	7.9	8.8	9.6	3.4	2	0.6	2.5	2.8	0.8	2.8
flav.test	pam120	0.40	0.20	0.40	0.48	0.49	0.03	17442.677	45.91	4.38	21.20	36.40	33.38	7	10.7	1.1	5.8	1.4	0.4	4.4	5.2	10.6	7	2.5	6.5	8.2	5.4	3.8	2	4.7	7.4	2.3	3.6
glob.test	blosum85	0.79	0.01	0.21	0.52	0.44	0.04	16044.643	43.54	8.76	14.90	23.20	28.04	12.6	7.8	1.9	5.8	0.4	4.7	3.1	4.2	5.5	4.2	3.4	7.4	5.5	9.4	3	1.2	5.2	10.5	2.3	1.9
gluts.test	gonnet	0.61	0.06	0.33	0.44	0.50	0.06	24284.825	43.65	8.13	24.70	24.70	45.05	6.6	6.3	2.4	5.1	1.4	1.6	4.2	3.4	5.9	4.8	5.7	5	6.2	7.5	4.6	1	4.2	13.5	5.1	5.4
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
toxin.test	pam500	0.00	0.30	0.70	0.36	0.60	0.04	7061.9712	47.4	8.55	22.40	75.00	38.98	2.8	6.7	1.6	6.2	13.5	2.1	5.9	9.1	4	4.8	7.9	4.6	3.3	8.9	2.9	1.3	1.7	4.2	5.1	3.5
xia.test	pam360	0.50	0.08	0.42	0.48	0.47	0.05	43073.87	43.26	5.26	64.60	66.50	66.27	11.9	9.8	1.7	3.2	0.1	2.9	3.5	4.9	8.6	3.4	4.6	4.9	6.9	4	2.7	1.4	6.1	10.3	6.4	2.6
cf_CCHH.test	pam500	0.24	0.08	0.68	0.35	0.60	0.05	3468.6051	61.25	9.25	13.90	45.20	29.76	3.1	2.4	1.4	8	7.6	8.7	3.1	7.6	2.8	4.5	2.8	4.2	5.5	12.5	2.1	0	5.2	3.1	11.1	4.5

Tableau VI.1.3
A titre d'exemple ce tableau reprend une partie des 78 cas tests pour lesquels nous avons cherché une valeur pour chaque caractéristique prédite a priori à partir de leurs séquences.

V.3. Travail réalisé sur septante-huit cas-tests :

- **Méthode**

Nous avons exécuté le programme Match-Box sur la batterie des septante-huit cas-tests, en utilisant successivement une des quarante-sept matrices de scores. Nous avons ciblé pour chacun des septante-huit cas-tests, la matrice procurant le meilleur score d'efficacité. Nous avons calculé, à partir de la puissance et de la confiances, le score d'efficacité obtenus pour l'ensemble des meilleures situations. Nous les avons comparé à la moyenne de ceux obtenus par Match-Box utilisant uniquement la matrice *BLOSUM62* (*cfr. tableau V.3.1*).

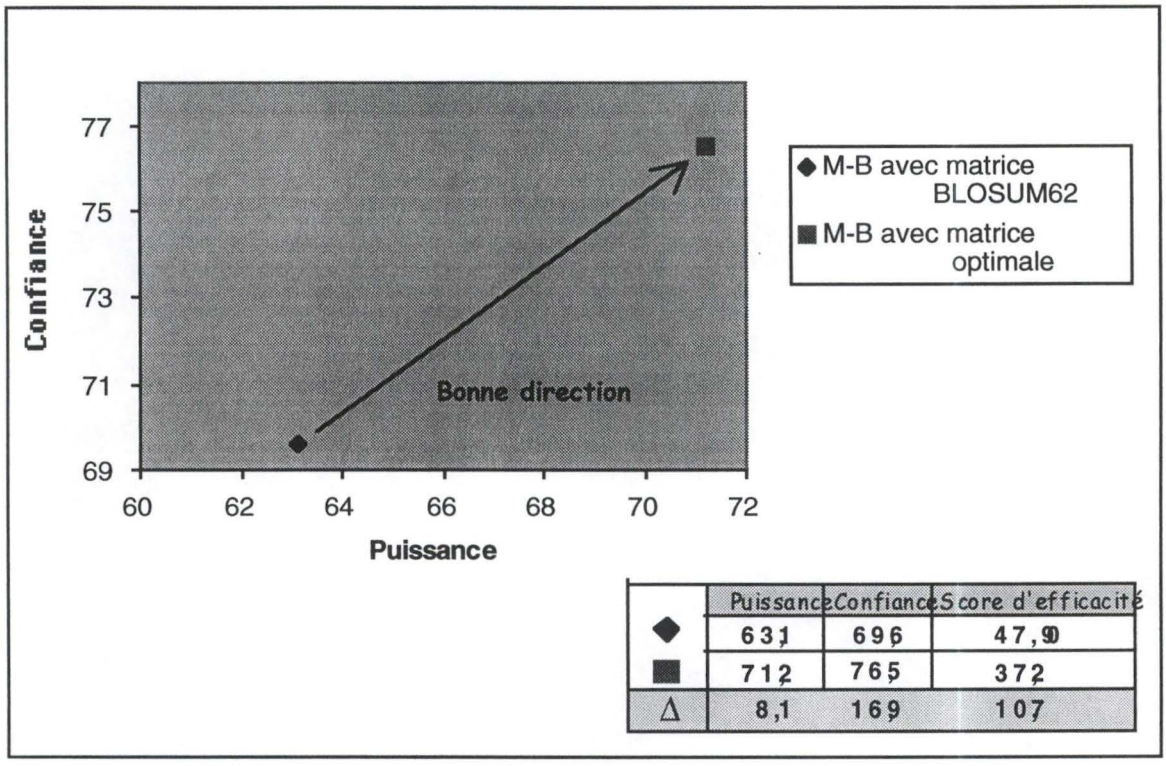
	Cas Test	Matrice choisie	score d'éfficacité		Cas Test	Matrice choisie	score d'éfficacité
1	aa.test	pam230	46	40	lipo.test	pam500	
2	ace.test	blosum50	45,8	41	ltn.test	pam440	36,9
3	adk.test	pam130	73,8	42	lys.test	blosum85	10,4
4	annexin.test	pam500	0	43	mmp.test	pam500	1,8
5	asp.test	blosum80	41,1	44	mthina.test	pam500	0
6	az.test	pam100	85,9	45	mthinb.test	pam460	9,4
7	bowman.test	pam500	6,6	46	mycin.test	blosum30	10,6
8	cbp.test	pam120	85,4	47	ndk.test	pam500	0
9	cryst.test	pam290	81,3	48	neur.test	pam370	30,9
10	cys.test	blosum45	39,2	49	p450.test	blosum45	90,4
11	cyt3.test	pam140	43,5	50	parv.test	pam500	0
12	cyt5.test	blosum85	76,1	51	phoslip.test	pam280	18,7
13	cytc.test	blosum30	21,8	52	rep.test	pam200	60,6
14	cyto.test	pam450	0	53	rhv.test	blosum45	43,8
15	dhfr.test	pam220	25,8	54	ricin.test	pam100	28,9
16	egf.test	pam460	47,1	55	rnasebact.test	pam500	1,4
17	fer2.test	blosum85	3	56	rnasemam.test	blosum30	16,4
18	fer4.test	pam500		57	rnh.test	blosum85	28,2
19	flav.test	pam120	63,5	58	rub.test	pam120	9,6
20	glob.test	blosum85	102,7	59	rubisco.test	blosum75	33,2
21	gluts.test	gonnet	13,9	60	rvp.test	blosum62	45,6
22	gpdh.test	pam310	10,7	61	seatoxin.test	pam500	74,3
23	grs.test	pam200	73	62	serbact.test	pam240	24,6
24	hip.test	pam100	31,7	63	sermam.test	blosum70	43,2
25	hla.test	pam500	0	64	serpin.test	pam230	29,2
26	hom.test	pam490	0	65	sh2.test	blosum85	30,5
27	hpr.test	pam500	0	66	sh3.test	pam490	19,1
28	igC1.test	pam400	30,3	67	sodcu.test	pam140	1,5
29	igcon.test	pam500		68	sodfe.test	pam120	40
30	igV.test	blosum50	46	69	squash.test	pam500	26,2
31	igvar_h.test	blosum75	4	70	subt.test	blosum100	23,3
32	igvar_l.test	blosum70	7,5	71	sugbp.test	blosum90	66,8
33	il8.test	pam320	39,4	72	thiored.test	blosum30	133,4
34	ins.test	pam150	2,3	73	tim.test	blosum80	23,5
35	intb.test	pam450	86,6	74	tln.test	pam160	41,5
36	kazal.test	pam430	17,2	75	tms.test	pam330	8,7
37	kringle.test	pam250	21	76	toxin.test	pam500	49,8
38	kunitz.test	pam320	10,5	77	xia.test	pam360	0,4
39	ldh.test	blosum75	60,2	78	zf_CCHH.test	pam500	54,8

Tableau V.3.1

Ce tableau reprend pour la batterie de 78 cas tests, les performances obtenues avec Match-Box(pour la meilleure matrice à utiliser).

Nous constatons, comme dans le cas des trente-trois cas tests, une utilisation d'une gamme diversifiée de matrices scores pour obtenir les meilleurs résultats. Remarquons également que toutes les matrices de scores de la famille des *JOHNSON* n'entrent plus en ligne de compte puisque ces matrices ont été construites sur base des alignements de structures de ces mêmes cas-test. Notre but n'étant pas d'améliorer le *matching* uniquement sur ces 78 cas tests et pour ne pas fausser nos résultats nous avons soustrait les deux matrices de la famille *JOHNSON* . Pour cela nous avons créé un palmarès d'efficacité des matrices par cas tests avec à chaque fois le score d'efficacité s'y rapportant. Lorsque la matrice de *JOHNSON92* ou *JOHNSON96* était la meilleure (pour information 21 fois sur les 78 cas) nous prenions la seconde meilleure matrice. Remarquons qu'après cela la matrice *PAM500* ressort 16 fois comme en tête, ce qui en fait la matrice qui fonctionne la plus généraliste sur les 78 cas tests.

Comme dans le cas de la batterie de 33 cas tests, nous allons comparer les performances de Match-Box obtenues avec ces diverses matrices par rapport à celles obtenues lorsqu'il utilise systématiquement BLOSUM62. Nous connaîtrons à l'issue de cette analyse dans quelles mesures les performances augmenteraient-elles *(cfr. figure V.3.1)*



	Puissance	Confiance	Score d'efficacité
◆	63,1	69,6	47,9
■	71,2	76,5	37,2
Δ	8,1	16,9	10,7

Figure V.3.1

Figure illustrant un graphe de puissance - confiance représente l'évolution des perforances entre deux version de Match-Box pour 78 cas tests. Une utilisant la matrice de scores BLOSUM62 et une autre hypothétique dans laquelle il utilise la matrice donnant le meilleur résultat.

Cette méthode nous indique les performances maximales que l'on pourrait atteindre, en sélectionnant la meilleure matrice de scores pour chaque famille à aligner. La tendance observée est la même que sur les trente-trois cas tests, Le gain de puissance (8,1%) est simultané au gain de confiance (6,9%).

Nous constatons une amélioration du score d'efficacité d'environ 11% sur les 78 cas tests, nous en avions 13% dans le cas de la batterie de 33 cas tests. C'est du même ordre de grandeur dans les deux cas (la différence de 2% provient du fait qu'il y sans doute un ou deux cas tests plus difficile à aligner).

- **Conclusion**

L'élargissement de la batterie de familles de référence confirme bien que si Match-Box était capable de choisir la matrice optimum à utiliser, les performances augmenteraient. Maintenant que nous possédons assez d'outil en main, nous allons pouvoir concentrer notre effort de recherche sur la possibilité de déterminer le choix de la meilleure matrice de similarité. Cette prédiction de matrice sera établie à partir de caractéristiques qu'il faudra déterminer *a priori* sur base de la séquence des cas tests.
Ce point constitue la perspective de toute la suite de ce travail.

Chapitre VI. Caractérisation des cas tests

VI.1. Méthode

On se trouve confronté à une situation comparable à une situation expérimentale où l'on devrait déterminer des caractéristiques communes de familles de protéines utilisant les mêmes matrices de scores. Nous allons chercher à décrire et à comprendre le choix d'une matrice de scores par un programme d'alignement. Certaines familles tests sont mieux alignées lorsque Match-Box utilise certaines matrices de scores. C'est un phénomène dans lequel intervient sans doute un grand nombre de variables, parmi lesquelles, il nous semble *a priori* impossible de déterminer celles pouvant jouer un rôle prépondérant. Il est cependant vraisemblable que ce qui guide ce choix se trouve quelque part caché dans les séquences de nos cas tests.

Malheureusement, nous ne savons pas *a priori* quelles caractéristiques (établies sur base de ces séquences) il pourrait s'agir. Nous allons réaliser toute une série de mesures prédictibles à partir des séquences de nos cas tests pour générer un tableau contenant un grand nombre de données sans avoir d'hypothèse particulière à vérifier. Pour générer ce tableau, nous avons soumis toutes les séquences (435) de nos cas tests à divers programmes bioinformatiques destinés à les analyser. Nous avons commencé par exécuter le programme *PHD* nous donnant la structure secondaire ainsi que l'accessibilité au solvant de séquences (*crf. Matériel PHDSS & PHDACC*).

Après avoir regroupé toutes les informations, nous avons pu calculer un pourcentage par cas test de résidus se trouvant dans une hélice α , dans un brin β , encore dans une structure irrégulière par simplification désigné par le terme boucle (*loop)*. De la même manière, un pourcentage de résidus se trouvant plutôt en surface de la protéine, au cœur de la protéine ou encore en position intermédiaire.

À ce stade, nous avons pour chaque cas test, six valeurs prédites sans savoir si ces caractéristiques influenceront ou pas le ciblage des matrices. Nous avons continué à collecter diverses informations, en soumettant nos séquences à d'autres logiciels (*MWCALC, SAPS, ALIGN)*. Au terme de cette étape nous avons donc rassemblé pour chaque cas-test les caractéristiques reprises dans le tableau VI.1.1.

Caractéristique	Programme utilisé
le poids moléculaire moyen	MWCALC
l'index de polarité	MWCALC
le point isoélectrique	MWCALC
la composition de chacun des vingt acides aminés	SAPS
le minimum d'identité	ALIGN
le maximum d'identité	ALIGN
la moyenne d'identité	ALIGN
le pourcentage de résidus se trouvant en hélice, en brin en loop	PHDsec
le pourcentage de résidus se trouvant au cœur, en surface, en intermédiaire	PHDacc

Tableau VI.1.1.

Ce tableau nous illustre les différentes caractéristiques prédites sur base de la séquence des 78 cas tests ainsi que les programmes utilisés pour les réunir.

Nous sommes en possession d'un tableau de données à deux entrées, par convention générale désigné par X. Les caractéristiques qui le composent seront indicées de la façon suivante (*cfr. tableau VI.1.2*).

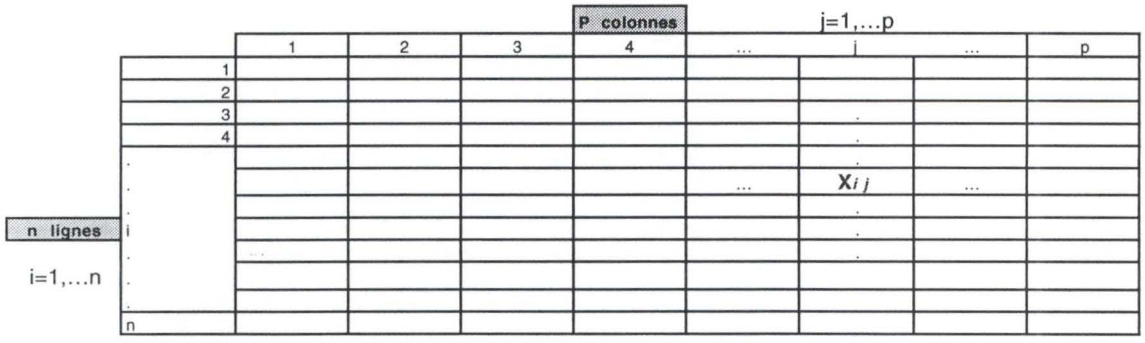


Tableau VI.1.2

*Ce tableau sera défini de dimensions n*p . Une dimension du tableau est fixée a priori, la seconde est fonction de l'échantillonnage. Le nombre de cas-test (objet étudier) est défini (78). Chaque objet est caractérisé par une série de descripteurs qui par convention correspondent aux colonnes du tableau.*

Nous allons analyser un tableau complet, dont les descripteurs sont subdivisés en plusieurs sous groupes (exemple : le groupe relatif aux structures secondaires est composé de trois sous groupe à savoir %H, %E, %L) *(cfr. tableau VI.1.3)*. Si le type de matrices utilisée par cas tests, est prise comme étant une variable dépendante et les diverses caractéristiques comme des variables indépendantes, notre but serait de prédire le choix matriciel, en fonction des données caractérisant les cas tests.
Seule une procédure d'analyse multivariée pourra nous aider à analyser un tel tableau. C'est ce que nous avons réalisé grâce au logiciel *STATISTICA* qui nous a permis de représenter les résultats des observations dans un espace géométrique multidimensionnel. Avant ce type de démarche était longue et difficile à mettre en œuvre vu l'importance du nombre d'opération sous-jacente à ce type d'approche. Néanmoins, afin de mieux comprendre certains points de la philosophie de ce travail, il est bon de décrire brièvement en quoi consiste une telle analyse et quelle en est le but.

Procédure multivariée

La variance totale d'un échantillon peut être décrite en plusieurs dimensions. Le but à atteindre est de réduire les dimensions de la matrice originale en perdant le moins possible d'information. Les dimensions qui subsistent sont des combinaisons linéaires des variables originales, qui apparaissent successivement comme modèles interdépendants de variation de l'échantillon. Dans la recherche d'une association éventuelle entre une matrice et une caractéristique prédites, nous nous intéresserons aux intercorrélations entre les variables plutôt qu'a la variance totale. Il y aura donc distinction entre variance commune et variance spécifique. Cette variance commune sera décrite par un certain nombre de dimensions (appelé facteur) calculée automatiquement par *STATISTICA*. Mais l'espace des facteurs communs n'est plus une simple projection orthogonale de la configuration originale multidimensionnelle.

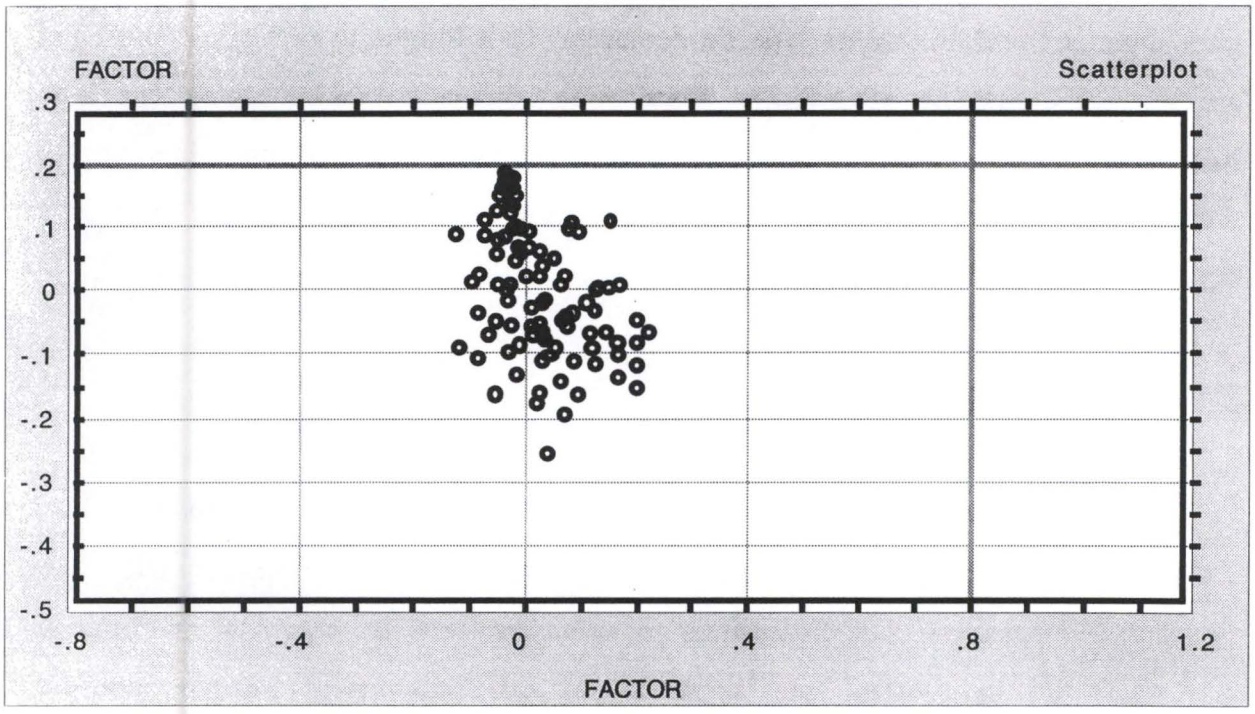


Figure VI.1.1.

Représentation ce diagramme de dispersion établi à partir de la procédure multivariée. Les matrices de scores apparaissent toutes corrélées entre elle et ne permettent pas de déterminer une caractéristique précise permettant choisir la meilleure matrice par cas test.

La représentation des caractéristiques prédites dans un espace géométrique multidimensionnel peut se faire en prenant comme axes de référence l'ensemble des variables directement mesurées. Pour notre tableau cette option semble très difficile à mettre en œuvre, vu le nombre de caractéristiques étudiées. Le type d'alternative qui a été envisagée a été d'utilisant le module d'analyse factorielle du logiciel *STATISTICA*. Ce module va choisir comme nouveaux axes du système des facteurs (appelées aussi composantes principales) qu'il calcule à partir de l'ensemble des variables. Pour ne pas avoir autant de composantes principales que de variables, il va y avoir élimination des axes considérés comme négligeables, c'est-à-dire dont l'élimination ne diminue pas l'information. On transforme donc un système d'axes x1,...,xn en un autre dont le choix des axes résulte de façon à ce que la contribution des facteurs sélectionnés explique la variabilité de façon maximale. L'objectivité de cette méthode fait qu'elle a souvent été préférée aux autres. L'inconvénient est la masse des calculs nécessaires (transformations de matrice...). L'utilisation du logiciel STATISTICA nous a donc été quasiment indispensable.

Au vu de cette analyse statistique, il ne semble pas pouvoir discriminer les caractéristiques influençant ou pas le ciblage des matrices.

Les résultats observés ont été rassemblés dans la figure VI.1.1.

VI.3 Conclusion

Il a été montré que Match-Box pourrait augmenter ces performances en choisissant parmi un bouquet de matrices celle qui est la mieux adaptée au cas test, cependant à ce niveau-là, rien ne permet de prédire sans ambiguïté ce choix. La première constatation étonnante est la faible dispersion des familles de matrices alors que la philosophie sur laquelle repose leur pondération complètement différente (cfr. Matrice de scores *PAM*

et BLOSUM) Nous n'avons donc pas pu associer sur base d'une simple analyse statistique une corrélation nette entre une caractéristique clé d'un cas test et une matrice de score précise. Une caractérisation globale sur un ensemble de protéines n'est pas discriminante pour matcher une matrice de similarité à un environnement préalablement définit par une série de caractéristiques. Nous avons changé notre support de caractérisation pour qu'il devienne moins global, en ciblant des segments de séquences à la place de l'entièreté du cas test. C'est ce que nous allons mettre au point dans le chapitre suivant. Maintenant, le choix d'une matrice adaptée dépend peut-être aussi d'une spécificité non reprise lors de cette analyse.

Chapitre VII. Caractérisation de fenêtres

VII.1 Méthode

L'idée développée sera fondamentalement la même que celle de la caractérisation des cas tests, mais à un niveau plus proche de la séquence. Puisque caractériser globalement un cas test semble trop vaste pour pouvoir tirer une information précise quant aux circonstances d'utilisation de matrices de similarité. Nous avons dû restreindre cette caractérisation globale à une caractérisation beaucoup plus locale. Nous allons détecter les matrices de similarité qui identifient correctement des segments conformément alignés en références aux deux types de vérité décrits précédemment (*cfr. développement d'une batterie de 78 cas tests*).

Chacun de ces segments sera caractérisé par une série de variables de son environnement établies sur base du programme *PHD*. Dans un second temps, nous mettrons en relation ces variables de l'environnement avec les matrices de similarité qui sont localement les plus performantes.

Ces segments sont définis par des fenêtres initiales de neuf résidus situées dans un segment correctement aligné et ne seront comptabilisées que celles situées dans un meilleur appariement.

Pour chacune de ces fenêtres, nous possédons déjà l'information sur la structure secondaire et sur l'accessibilité au solvant, prédite par *PHD,* mais n'est pas directement exploitable faute de localisation ces fenêtres dispersées dans l'ensemble des 435 séquences des 78 cas tests.

Un programme conçu au laboratoire va lire les critères de vérités, pour associer à chaque fenêtre répertoriée, l'information concernant ces microenvironnements. Le programme Match-Box a été modifié pour identifier les matrices de scores retrouvant

chacun des appariements corrects et en évaluer ensuite l'efficacité de détection de ce signal.

A partir d'un jeu de séquences parfaitement alignées, on peut facilement définir deux critères d'efficacité pour une matrice de score donnée.

* Un premier critère booléen de type succès–échec, nous renseigne si une matrice de score retrouve une fenêtre correctement alignée. Il y aura succès lorsque le score global correspondra à l'appariement correct entre fenêtre fixe et fenêtre mobile. Dans le cas inverse, on parlera d'échec. Ce critère est donc un indicateur de l'exactitude de détection des similarités par une matrice de score donnée.

* Une seconde mesure quantitative identifie l'intensité du signal par rapport au bruit. On caractérisera ce comportement par une statistique de type *t* de Student :

$$t = \frac{x - \bar{x}}{s}$$

Où : X est le score minimum global du balayage.
X est la moyenne des scores globaux (sauf le minimum) du balayage.
S correspond à l'écart type des scores globaux (sauf le minimum) du balayage.

Pour nos tests, plus la valeur de *t* sera faible plus score minimum global se détachera des autres scores considérés *(cfr.figureVII.1.1)*. Ce deuxième critère est donc un indicateur de l'intensité de la réponse de la matrice de scores lors de la détection des appariements.

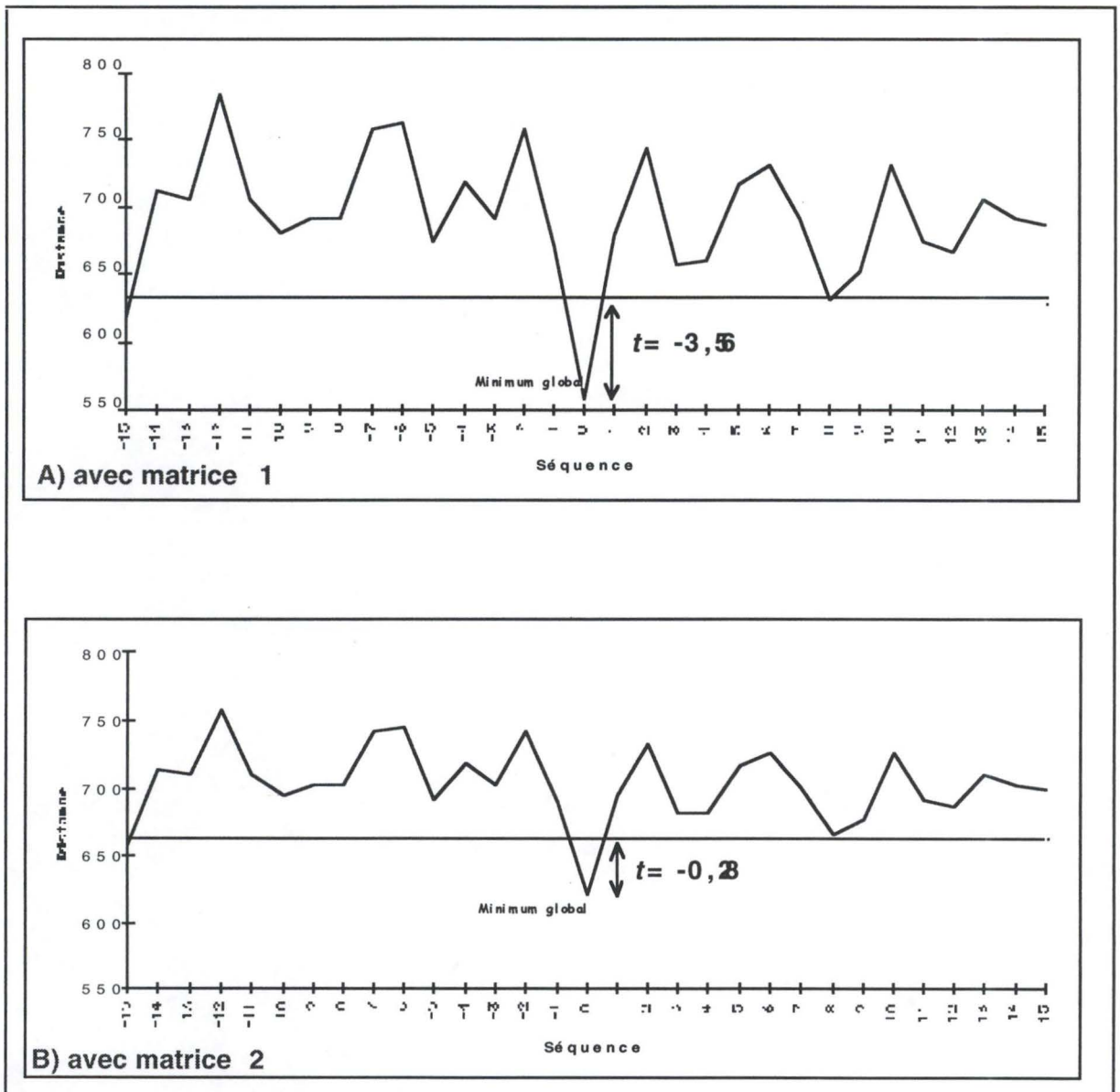


Figure VII.1.1 Cette figure représente la distribution des scores globaux au cours d'un balayage en utilisant successivement la matrice de score n°1 et la n°2. Avec ces deux matrices, le meilleur appariement est détecté à la position attendue dès lors toutes deux sont en situations de succès. Cependant, avec la matrice n°1 la valeur de t est plus faible, ce qui indique que l'intensité de la réponse de la matrice de score n°1 est plus forte lors de la détection des appariements.

VII.2. Test réalisé avec 42 matrices de scores différentes

Nous avons mis en œuvre la méthode décrite au point précédent avec une batterie de 42 matrices de similarité que nous allons maintenant essayer d'associer aux segments décrits par les variables d'environnement.

L'exécution du programme sur la station de travail S*ilicon Graphics* nous transmet à sa sortie un tableau reprenant l'ensemble de nos fenêtres initiales de neufs résidus avec leur composition (pourcentage) respective en résidus exposés, partiellement ou totalement enfouis incorporés dans une hélice, dans un brin ou dans une boucle. Notons également à partir de quel critère de vérité l'appariement est considéré comme correcte. Le programme trouve approximativement 48000 de fenêtres. Ce tableau très volumineux nécessite un pré-traitement dans l'environnement *SAS (cfr. Matériel SAS)*, pour nous permettre par la suite d'analyser nos données plus facilement. Dans *SAS*, nous avons traité le fichier brut de sortie du programme pour en dresser un tableau. Ce tableau reprend par fenêtre initiale impliquée, la valeur numérique des six variables d'environnement ainsi que les valeurs de *t* pour les 42 matrices de scores

Nous avons fait une analyse factorielle à l'aide du programme *STATISTICA*, où un graphe de dispersion en deux dimensions nous a permis de visualiser la position des six états *(brin, hélice, boucle, enfoui, exposé, integré)* et de voir la distribution des 42 matrices de scores vis-à-vis de ces états *(cfr.tableau VII.2.1 figure VII.2.1 et VII.2.2)*

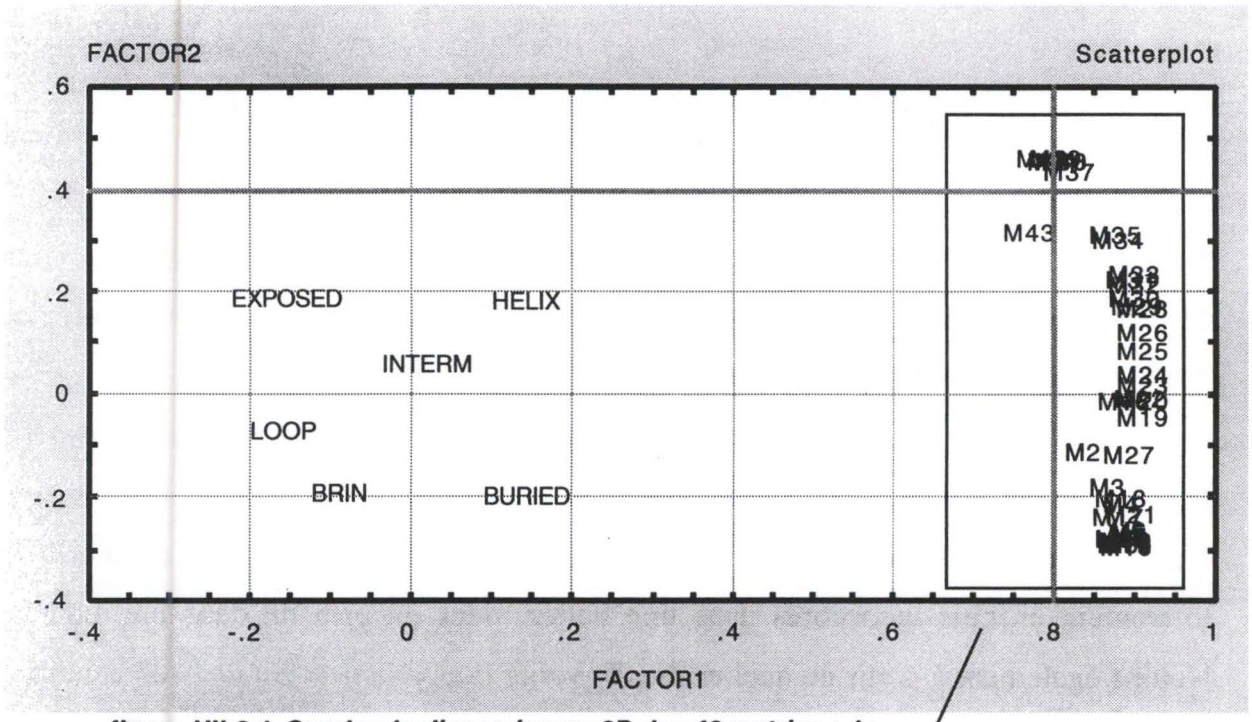


figure VII.2.1. ***Graphe de dispersion en 2D des 42 matrices de similarité . On peut voir la répartition de ces matrices par rapport aux six états caractéristiques.***

ZOOM

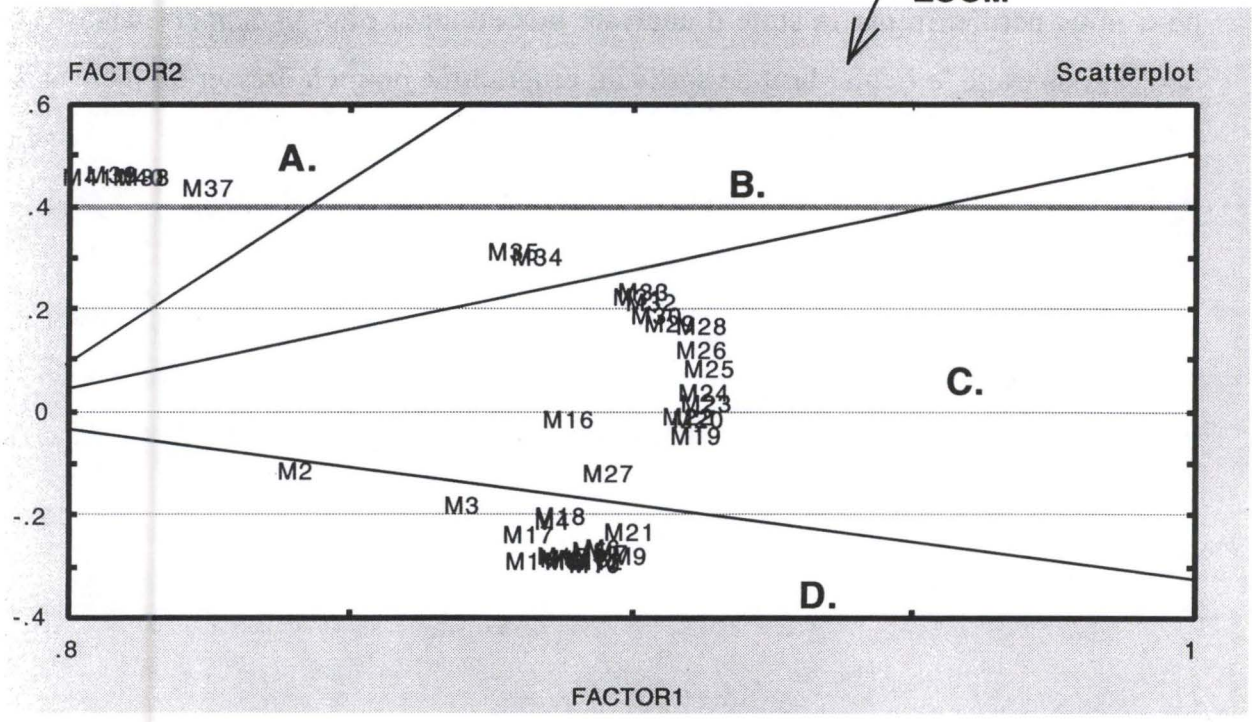


figure VII.2.2. ***Graphe de dispersion en 2D ciblé sur la distribution des 42 matrices de similarité. On remarque que ces matrices se distribuent sur un croissant, dont les deux extrémités tendent vers l'environnement hélice exposée pour l'une et brin enfoui pour l'autre. On peut mettre en évidence quatre zones (A,B,C,D) de niveau de spécificité différente. Le corps du croissant, reprend une série de matrices fortement corrélées entre elles, dont la spécificité locale est plus faible.***

n°Matrice fig. VI.1 & VI.2	Matrices	Environnement	Spécificité Matrice	
M39	PAM440	HELICE EXPOSEE	Zone à plus grande spécificité	A.
M42	PAM490			
M38	PAM430			
M40	PAM450			
M41	PAM460			
M37	PAM400			
M43	PAM500			
M35	PAM290			
M34	PAM280			
M33	PAM260			
M31	PAM240			
M32	PAM250			
M30	PAM230			B.
M29	PAM220			
M28	PAM210			
M26	PAM190			
M25	PAM160			
M24	PAM150			
M23	PAM140			
M22	PAM130			
M20	PAM110			
M16	GONNET			
M19	PAM100			
M2	BLOSUM30			
M27	PAM200	Intermédiaire		
M3	BLOSUM35		Zone à plus faible spécificité	C.
M18	JOHNSON96			
M4	BLOSUM40			
M21	PAM120			
M17	JOHNSON92	Boucle		
M6	BLOSUM50			
M5	BLOSUM45			
M7	BLOSUM55			
M9	BLOSUM62			
M15	BLOSUM90			
M14	BLOSUM85			
M13	BLOSUM80			
M11	BLOSUM70			
M1	BLOSUM100			
M12	BLOSUM75		Zone à plus grande spécificité	D.
M8	BLOSUM60	Brin enfoui		
M10	BLOSUM65			

Tableau VII.2.1.

Ce tableau indique le nom des matrices correspondant aux n° des matrices dans les deux graphes de dispersion (cfr. figures VII.1 et VII.2). Il représente la tendance à l'association des familles de matrices à un type d'environnement donné. Cette information est reprise sous forme des lettres A, B, C, D. Ce tableau illustre également une échelle de spécificité des matrices de scores utilisées.

L'analyse des graphes de dispersion révèle de façon stupéfiante que les matrices se sont rassemblées par famille. Les familles BLOSUM et PAM sortent pratiquement dans l'ordre. Dans partie gauche de la figure VII.2.1, on observe que certains microenvironnements sont plus proches les uns des autres. Ainsi la position exposée serait plus volontiers incorporée dans une structure secondaire de type hélice et la position enfouie dans une structure secondaire de type brin, les environnements boucle et intermédiaire s'intercalant au milieu des deux autres.

En modifiant l'échelle du graphe de dispersion, on remarque que le peloton de matrices appartenant à la famille des *PAM* tend plutôt vers l'état hélice - exposée et à l'inverse, tout le groupe de matrices des *BLOSUM* tend vers l'état brin - enfoui. Cette observation est encourageante dans le sens où, à partir d'une prédiction de structures secondaires tirée d'une séquence de protéine on obtient une tendance à l'utilisation préférentielle de l'une ou l'autre famille de matrices ou d'autre. Avec la distribution de ces matrices, nous pouvons établir une sorte de gradient de spécificité, où l'on retrouve effectivement dans la zone à faible spécificité, les matrices réputées les plus généralistes (*cfr. tableau VII.2.1*).

Rem :

Cette tendance ne se visualisait pas lors de la caractérisation au niveau global du cas test, où la composition des séquences d'un cas test est composée de plusieurs fenêtres, qui prises localement sont associées à des catégories différentes d'environnement. La spécificité des environnements était atténuée. Remarquons qu'en exécutant Match-Box sur les 78 cas-tests avec la famille de matrices de *JOHNSON*, les performances sont plus élevées que celles obtenues avec les familles *BLOSUM et PAM.* La famille des matrices de *JOHNSON* a été construite et optimisée sur cette même banque de 78 cas-tests. C'est la famille la plus performante car la plus généraliste. Ceci appuie également notre observation.

En conclusion, l'hypothèse selon laquelle les matrices de similarité peuvent être spécifiques à un microenvironnement est vérifiée. Ce test a été réalisé avec un ensemble de matrices réputées au départ généralistes, il faudra donc élargir la batterie de matrices en y incluant des matrices à plus grande spécificité locale. L'hypothèse doit être poursuivie pour accroître d'avantage cette tendance d'association d'une matrice spécifique à un environnement local. Ceci constituera le point suivant de notre travail.

VII.3. Test réalisé avec 134 matrices de scores différentes :

L'hypothèse a été investiguée en rassemblant un maximum de matrices de scores qui historiquement avaient été un peu délaissées du fait de leur faible potentiel à aligner une protéine entière. Nous avons réuni une batterie comprenant cent trente-quatre matrices de scores différentes. Cet ensemble comprend une série de matrices à haute spécificité, établies sur base de variables spatiales. D'après la littérature, certaines de ces matrices spécifiques ont été réalisées sur base de substitution entre résidus situés dans une hélice, dans un brin, dans le cœur des protéines, inaccessible au solvant ou au contraire pour des résidus exposés en surface où l'accessibilité au solvant est élevée. Nous avons exécuté le même programme qu'au point précédent en utilisant quatre vingt-neuf matrices de scores supplémentaires. Le tableau généré par le programme fut traité de la même manière, c'est-à-dire par une série de manipulations successives, nous ramenant de l'environnement *SAS* à celui de *STATISTICA ; celui-ci* permettant un mode graphique beaucoup plus raffiné *(cfr. étapes au point précédent)*.
Dans *STATISTICA,* nous avons réalisé une analyse factorielle pour construire un graphe de dispersion en trois dimensions, permettant une meilleure perception visuelle de la distribution des différentes matrices et des différents environnements. Le but

Type d'environnement	Meilleure matrice	Origine de la matrice	Auteur
Si Brin	OVEJ920102	"Environment-spécific amino acids substitution matrix for beta residues"	(Overington *et al* ., 1992)
Si Hélice	GEOD900101	"Hydrophobicity scoring matrix"	(George et al.,1990)
Si Boucle	JOND940101	"The 250 PAM transmembrane protein exchange matrix"	(Luthy et al., 1991)
Si Enfouis	LUTR910104	"Structure-based comparison table for inside alpha class"	(Luthy et al., 1991)
Si Exposé	LUTR910103	"Structure-based comparison table for outside alpha class"	(Luthy et al., 1991)
Si Interm.	AZAE970102	"Substitution matrix derived from spatially conserved motifs"	(Risler et al., 1988)

Tableau VII.3.1

Ce tableau établit la correspondance entre un type d'environnements particuliers et la matrice qui lui est le mieux adaptée. Ces matrices furent donc trouvée sur base de la distance les séparant de leur environnement.

On observe que notre méthode retrouve les environnements pour lesquels les matrices ont été conçues.

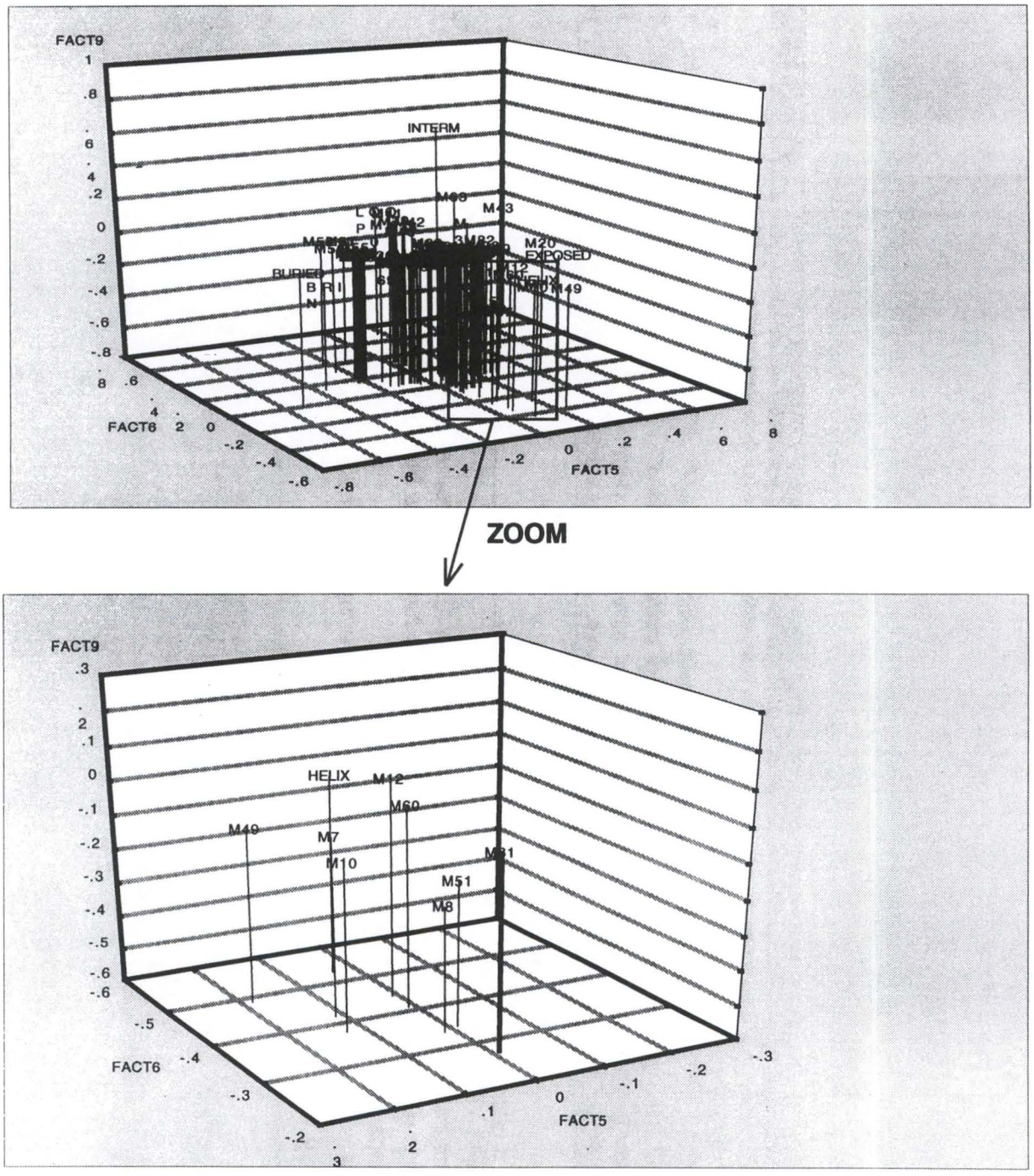


figure VII.3.1 ***Cette figure représente un graphe de dispersion en trois dimensions établi à partir des facteurs de l'analyse factorielle. Chaque points dans l'espace représente une matrice de similarité ou un type d'environnement précis. La vue d'ensemble ne nous apporte rien dans le sens où un zoom est indispenssable pour associer une matrice à un environnement. Un "point matrice" s'associe au "point environnement" le plus proche.***

étant ici d'associer à un environnement local une matrice lui étant le plus spécifique possible. Géométriquement, cela signifie que chaque type d'environnement et chaque type de matrice est représentés par un point dans un espace dont les trois dimensions ont été définies à partir des facteurs provenant de l'analyse factorielle. Chaque point représentant un microenvironnement sera associé au point le plus proche représentant une matrice. En se focalisant successivement dans l'espace avoisinant les six points représentant chaque type d'environnement (hélice, brin, boucle, exposé, enfoui, intermédiaire), nous avons pu leur associer les matrices les plus proches *(cfr. tableau VII.3.1)*. Nous avons donc identifié certaines matrices susceptibles d'être plus performantes pour aligner chaque type d'environnement. Cette démarche, principalement visuelle, est représentée dans la figure VII.3.1.

Maintenant, dans le cas d'une fenêtre dont la totalité de ses résidus sont exposés tout en étant incorporés dans une hélice, quelle sera la matrice à utiliser ? Celle pour l'hélice, celle pour la position enfouie ou une autre …
Nous ne sommes pas encore en mesure de dire laquelle à choisir. Nous sommes ici confronté à un problème de distances entre environnements et matrices de scores.
Intuitivement, nous avons établi l'ensemble des combinaisons qu'il est possible d'obtenir sur base des trois états de structures secondaires et des trois états de localisations dans la protéine. Nous avons neuf combinaisons d'environnement de structures assemblées à une localisation :

Hélice Enfouie	Hélice Intermédiaire	Hélice Exposée
Brin Enfoui	Brin Intermédiaire	Brin Exposé
Boucle Enfouie	Boucle Intermédiaire	Boucle Exposée

Nous allons, pour chaque combinaison, tenter de les représenter dans le graphe de dispersion en trois dimensions. Pour cela nous avons créé neuf points virtuels se trouvant tous exactement dans l'espace à mi-distance entre chaque type d'états, pris

deux à deux. Pour inférer ces points, nous nous sommes servis de l'analyse factorielle pour définir neufs facteurs pour chaque variable environnementale et pour chaque matrice. Sur base de ces facteurs, des coordonnées ont été définies pour l'ensemble des situations envisageables par une équation de droite à neuf dimensions. Chaque combinaison d'environnement est définie dans neuf dimensions.

Par exemple pour trouver le point virtuel situé entre l'état brin et l'état enfoui, nous appliquons la formule suivante :

$$\sqrt{(F_{1brin}-F_{1enfoui})^2+(F_{2brin}-F_{2enfoui})^2+(F_{3brin}-F_{3enfoui})^2+(F_{4brin}-F_{4enfoui})^2+(F_{5brin}-F_{5enfoui})^2+(F_{6brin}-F_{6enfoui})^2+(F_{7brin}-F_{7enfoui})^2+(F_{8brin}-F_{8enfoui})^2+(F_{9brin}-F_{9enfoui})^2}$$

En calculant les coordonnées spatiales pour chaque situation et en les replaçant dans un tableau, on obtient véritablement une « matrice » de distance symétrique entre les différents états et les diverses combinaisons de ceux-ci *(cfr. tableau VII.3.2)* .

	BRIN	HELIX	LOOP	BURIED	EXPOSED	INTERM	Brin-Buried	Brin-Exposed	Brin-Interm.	Hélix-Buried	Hélix-Exposed	hélix-Interm.	Loop-Buried	Loop-Exposed	Loop-Interm.
BRIN	**0,000**	0,941	0,857	0,317	1,096	1,036	0,704	0,852	0,741	0,690	0,696	0,809	0,689	0,707	0,790
HELIX	0,941	**0,000**	1,254	0,910	0,847	1,005	0,797	0,673	0,875	0,907	0,979	0,686	0,773	0,840	0,742
LOOP	0,857	1,254	**0,000**	1,083	0,790	0,921	0,849	0,900	0,884	0,868	0,877	1,060	1,015	0,905	1,031
BURIED	0,317	0,910	1,083	**0,000**	1,279	1,027	0,871	0,901	0,846	0,781	0,785	0,933	0,860	0,787	0,937
EXPOSED	1,096	0,847	0,790	1,279	**0,000**	1,034	0,716	0,754	0,886	0,951	1,016	0,752	0,800	0,920	0,750
INTERM	1,036	1,005	0,921	1,027	1,034	**0,000**	1,016	0,652	0,985	0,987	1,017	0,970	1,079	0,990	0,996
Brin-Buried	0,704	0,797	0,849	0,871	0,716	1,016	**0,000**	0,934	0,434	0,518	0,618	0,649	0,639	0,369	0,622
Brin-Exposed	0,852	0,673	0,900	0,901	0,754	0,652	0,934	**0,000**	0,999	1,004	1,035	0,642	0,710	1,026	0,684
Brin-Interm.	0,741	0,875	0,884	0,846	0,886	0,985	0,434	0,999	**0,000**	0,131	0,231	0,864	0,861	0,258	0,837
Hélix-Buried	0,690	0,907	0,868	0,781	0,951	0,987	0,518	1,004	0,131	**0,000**	0,120	0,927	0,909	0,283	0,902
Hélix-Exposed	0,696	0,979	0,877	0,785	1,016	1,017	0,618	1,035	0,231	0,120	**0,000**	0,983	0,953	0,396	0,955
hélix-Interm.	0,809	0,686	1,060	0,933	0,752	0,970	0,649	0,642	0,864	0,927	0,983	**0,000**	0,219	0,904	0,084
Loop-Buried	0,689	0,773	1,015	0,860	0,800	1,079	0,639	0,710	0,861	0,909	0,953	0,219	**0,000**	0,900	0,179
Loop-Exposed	0,707	0,840	0,905	0,787	0,920	0,990	0,369	1,026	0,258	0,283	0,396	0,904	0,900	**0,000**	0,889
Loop-Interm.	0,790	0,742	1,031	0,937	0,750	0,996	0,622	0,684	0,837	0,902	0,955	0,084	0,179	0,889	**0,000**

Tableau VII.3.2. Représentation de la distance entre les différents environnements. Cette distance a été calculée à partir des neufs facteurs provenant de l'analyse factorielle et servant d'axes de dimensions. Cette matrice de distance est symétrique et compte 120 coefficients originaux.

Nous allons maintenant pouvoir associer à tous points environnementaux virtuels ou non, la matrice s'y rapprochant le plus dans l'espace.

Type d'environnement	Meilleure matrice	Origine de la matrice	Auteur
Si Brin	OVEJ920102	"Environment-spécific amino acids substitution matrix for beta residues"	(Overington *et al* ., 1992)
Si Hélice	GEOD900101	"Hydrophobicity scoring matrix"	(George et al.,1990)
Si Boucle	JOND940101	"The 250 PAM transmembrane protein exchange matrix"	(Luthy et al., 1991)
Si Enfouis	LUTR910104	"Structure-based comparison table for inside alpha class"	(Luthy et al., 1991)
Si Exposé	LUTR910103	"Structure-based comparison table for outside alpha class"	(Luthy et al., 1991)
Si Interm.	AZAE970102	"Substitution matrix derived from spatially conserved motifs"	(Risler et al., 1988)
Si Brin-enfouis	OVEJ920104	"Environment-spécific amino acids substitution matrix for inaccessible residues"	(Overington *et al* ., 1992)
Si Brin-exposé	LEVJ860101	"The secondary structure similarity matrix"	(Levin et al., 1986)
Si Brin-interm.	AZAE970102	"Substitution matrix derived from spatially conserved motifs"	(Luthy et al., 1991)
Si Hélice-enfouis	QU_C930102	"Cross-correlation coefficients of factor"	(Overington *et al* ., 1992)
Si Hélice-exposée	KOSJ950113	"Context-dependent optimal substitution matrices for exposed residues"	(Koshi-Goldstein, 1995)
Si Hélice-interm	AZAE970102	"Substitution matrix derived from spatially conserved motifs"	(Overington *et al* ., 1992)
Si Boucle-enfouis	OVEJ920104	"Environment-spécific amino acids substitution matrix for inaccible residues"	(Overington *et al* ., 1992)
Si Boucle-exposé	KOSJ950104	"Context-dependent optimal substitution matrices for exposed coils"	(Koshi-Goldstein, 1995)
Si Boucle-interm.	AZAE970102	"Substitution matrix derived from spatially conserved motifs"	(Overington *et al* ., 1992)

Tableau VII.3.3.

Ce tableau établit la correspondance entre une combinaison d'environnements particuliers et la matrice qui lui est la mieux adaptée. Ces matrices furent trouvées sur base de la distance les séparant des points virtuels situés entre chaque type d'environnement.
On observe que notre méthode retrouve souvent les environnements pour lesquels les matrices sélectionnées ont été conçues.

Prenons l'exemple où nous cherchons à connaître la matrice à utiliser pour un segment exposé qui serait incorporé dans une boucle. On classe les matrices par écart croissant de distance par rapport au point virtuel situé entre l'environnement boucle et exposé. La première matrice du classement sera la plus susceptible de pondérer le mieux les résidus de ce type d'environnement. Cette démarche sera effectuée pour l'ensemble des situations. Cela nous amène donc à dresser un tableau reprenant pour chaque type d'environnement une matrice à spécificité locale *(cfr. tableau VII.3.3)*.

A la suite d'une recherche bibliographique sur les matrices sélectionnées, nous avons constaté que notre méthode retrouvait le type d'environnement sur lequel était basé leur conception. Ce résultat positif permet donc de valider notre méthode.
Reste maintenant à créer un fichier reprenant l'ensemble de ces directives qui seront lues lors de la procédure *matching*. Cette étape d'implémentation constitue le point suivant de notre investigation.

VII.4. Implémentation des observations.

Nous allons observer les performances du *matching* sur les 78 cas tests, en utilisant plusieurs matrices à spécificité locales. La procédure du *matching* a été retravaillée pour permettre à Match-Box de fonctionner avec un jeu de plusieurs matrices différentes.
Les performances seront comparées à celles obtenues lors de l'utilisation d'une même matrice très généraliste. Dès lors, lorsque Match-Box rencontrera une fenêtre à aligner, il lira les informations environnementales s'y rapportant et sélectionnera de façon autonome la meilleure matrice à utiliser. La procédure du *matching* choisira, parmi une série de matrices reprises dans un fichier qu'on lui aura implémenté, celle qui sera la mieux adaptée à l'environnement. Ce fichier inclura toutes les conditions environnementales qui guideront le choix de la marice à utiliser.

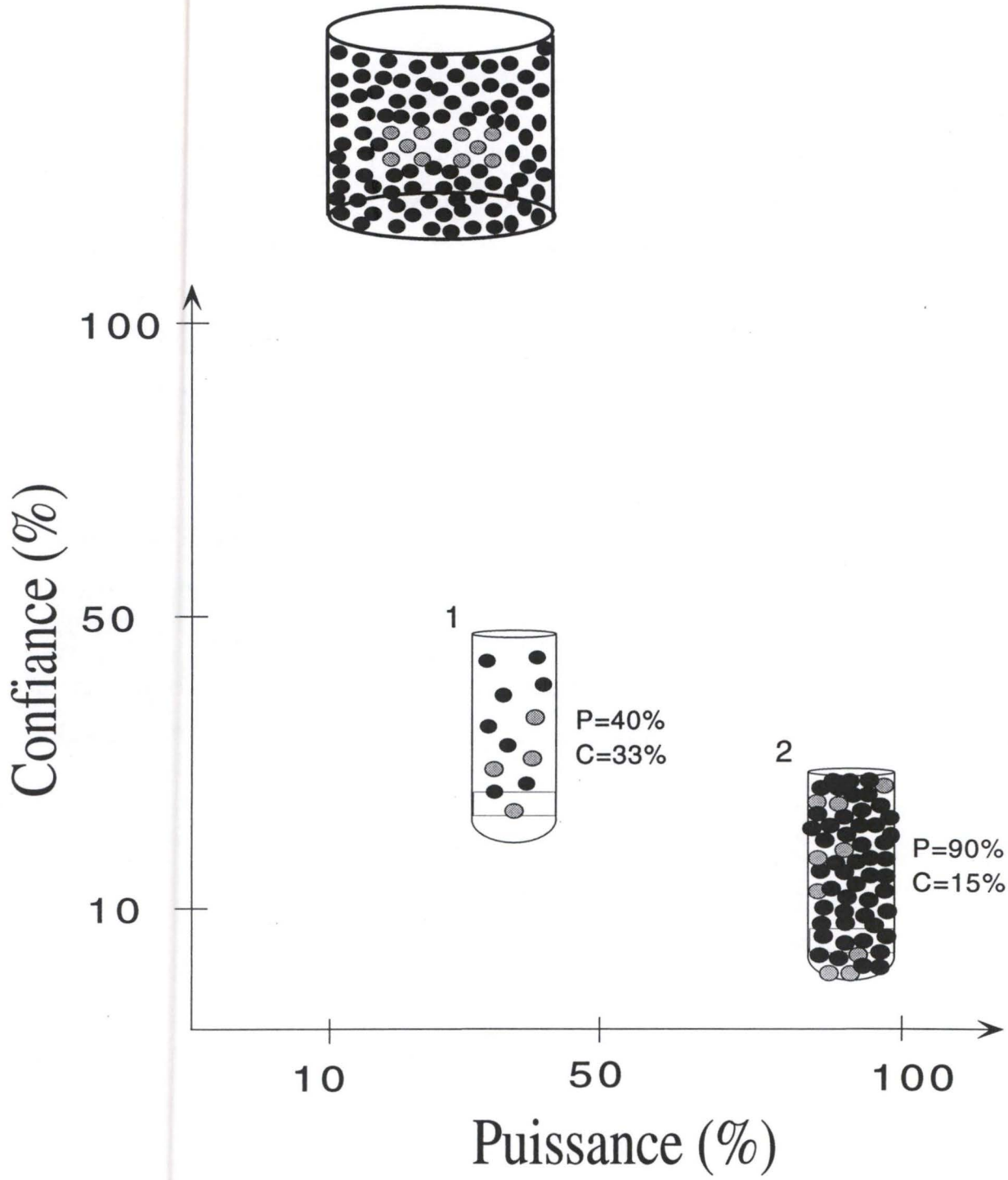


Figure: VII.4.1

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 le filtre. On constate que le retrait du filtre assure une augmentation en puissance mais provoque une diminution en confiance.

Notre but étant avant tout d'améliorer la procédure du *matching,* les tests ont été effectués uniquement sur cet algorithme. Si nous avions exécuté Match-Box entierèrement, nous aurions dû établir une valeur d'incrément propre à chaque matrice de la nouvelle batterie *(cfr. Tets sur incrément).* Ce *matching* fonctionne donc sans filtre, ce qui provoque une augmentation en puissance et provoque une diminution en confiance (on a plus de bruit puisque pas de seuil) (cfr. figure VII.4.1). Néanmoins, nous comparerons notre méthode à la « classique » dans les mêmes conditions *(cfr.tableau VII.4.1)* .

En conservant une confiance comparable	Puissance
matching avec JOHNSON92	70,8
matching avec notre fichier implémenté	71,8

Tableau VII.4.1

Représentation des performances entre deux matching fonctionnant sans filtre. Le premier utilisant la méthode classique (une matrice généraliste) basée sur JOHNSON92 et notre méthode utilisant plusieurs matrices (notre fichier) à spécificité locale. La confiance n'est pas indiquée puisque il n'y a pas de screening. (Le programme nous renseigne quand même sur la confiance et elle est comparable pour les deux situations).

Sur la batterie de 78 cas tests, les performances obtenues par notre méthode n'augmentent que très légèrement (1%). Néanmoins ce gain bien que mineur est substantiel par rapport à la méthode « classique » qui fait l'objet d'une optimisation depuis plusieurs années.

La famille de matrice généraliste utilisée était celle de *JOHNSON* qui, rappelons-le, a justement été construite et affinée sur base des cas tests sur lesquels s'est basée notre implémentation. Ceci signifie que sur un jeux de protéines inconnus, notre méthode pourrait être plus performante que l'utilisation de cette matrice très généraliste.

Remarquons également que lorsqu'une seule de ces matrices à spécificité locale est utilisée pour aligner l'ensemble des 78 cas tests, la performance du *matching* est faible. Ceci prouve bien que ces matrices sont très spécifiques et que c'est leur coopération qui leur permet d'avoir des résultats comparables à ceux obtenus avec la matrice de *JOHNSON*.

Cette nouvelle voie n'a cependant pas encore été optimisée. Pour cela il faudrait associer une matrice pour chaque proportion d'environnement d'une fenêtre. C'est-à-dire augmenter le nombres de matrice à spécificité locale à implémenter.

Chapitre VIII. Conclusions générales et perspectives

Au terme de ce travail, nous sommes parvenus à ouvrir une voie permettant un gain de performances selon les deux critères d'efficacité d'une méthode d'alignement multiple. C'est la première fois qu'une technique propose simultanément une amélioration en puissance et en confiance. Auparavant, toute tentative d'amélioration implantée à Match-Box permettait au mieux l'augmentation d'un seul des deux critères. Chaque gain de performance d'un critère était immédiatement corrélé par une perte dans l'autre. Ce phénomène de vases communicants a également été constaté lors de notre démarche d'optimisation d'un paramètre propre à chaque matrice de similarité.

Selon notre l'étude de faisabilité réalisée, la voie la plus prometteuse en matière de gain simultané de performances serait d'utiliser un éventail de matrices lors d'une même procédure d'alignement. Au vu de l'augmentation importante des performances que l'on pourrait escompter atteindre, on comprend immédiatement l'enjeu sous-jacent de cette hypothèse. Dès lors, nous avons cherché la possibilité de déterminer ce qui, dans les séquences protéiques, guide le choix préférentiel pour l'une ou l'autre matrice de similarité.

Au cours de nos manipulations, nous avons constaté qu'il était plus discriminant de caractériser *a priori* une séquence à un niveau local plutôt que globalement sur l'ensemble d'un jeu à aligner. Nous avons réussi à associer pour chaque catégorie de segments, définis majoritairement par un environnement particulier, une matrice à grande spécificité locale. En faisant travailler ensemble la quinzaine de matrices spécifiques ayant été sélectionnées, nous n'avons pu mettre en évidence qu'un gain

mineur de performances. Néanmoins, cela mérite un approfondissement puisque notre méthode parvient à améliorer légèrement les performances de Match-Box par rapport à une méthode faisant l'objet d'une optimisation permanente depuis plusieurs années. En effet, nos résultats furent comparés à ceux de Match-Box lorsqu'il utilise systématiquement la même matrice généraliste construite et affinée sur base des mêmes cas tests que ceux utilisés pour notre implémentation. De plus, l'affinage de cette méthode est encore loin d'être terminé puisque nous avons associé une matrice pour seulement une quinzaine d'environnements types, alors que la distribution du pourcentage des variables environnementales d'un segment est très élevée.

Il faut de même signaler que ces diverses catégories ont été établies sur base de caractéristiques structurales prédites et non des caractéristiques structurales réelles. Puisque la structure des cas tests est connue, nous aurions pu également établir les catégories sur base de la réalité. Cependant, lorsqu'on soumet des séquences protéiques à Match-Box, on possède évidemment très rarement ces informations. En optant pour les caractéristiques prédites, nous rendons la méthode reproductible à partir de n'importe quel jeu de séquences, mais nous introduisons inévitablement des imprécisions provenant de la prédiction de structures secondaires de *PHD* limitée à 70%.

L'idéal aurait été de disposer d'un nombre de cas tests beaucoup plus élevé afin d'en composer deux batteries. La première nous servant alors pour pouvoir réunir un ensemble de conditions suffisamment représentatives et mettre notre méthode en œuvre. La seconde pour tester cette méthode.

Les techniques d'alignement de séquences sont encore essentiellement limitées par l'information qu'il est possible d'exprimer à travers les matrices de similarité. Bien que les performances finales de notre méthode ne soient pas encore optimales, il est néanmoins remarquable que notre approche, bien que fondamentalement différente, produise des performances comparables à celle obtenues avec la matrice la plus généraliste sur un nombre encore assez restreint de cas tests.

Nous pensons donc que l'évolution de cette méthode est liée à l'élaboration de nos propres matrices de comparaison de séquences qui seront plus spécifiques encore que celles existantes.

Chapitre IX. Bibliographie

Blundell, T.L. and Johnson, M.S. (1993) Catching a common fold. *Protein Sci*, **2**, 877-883.

Blundell, T.L., Sibanda, B.L., Sternberg, M.J. and Thornton, J.M. (1987) Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, **326**, 347-352.

Boguski, M.S. (1998) Bioinformatics a new era. *trends guide to bioinformatique*, 1-3.

Brendel, V., Bucher, P., Nourbakhsh, I.R., Blaisdell, B.E. and Karlin, S. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A*, **89**, 2002-2006.

Briffeuil, P., Baudoux, G., Regeinster, 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.

Chothia, C. (1992) Proteins. One thousand families for the molecular biologist [news]. *Nature*, **357**, 543-544.

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

Dayhoff M. O., B.W.C., Hunt L. T. (1983) Etablishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524-545.

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

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.

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

Feng D. F., D.R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**, 351-360.

Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992) Exhaustive matching of the entire protein sequence database [see comments]. *Science*, **256**, 1443-1445.

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.

Henikoff, S., Henikoff, J.G., Alford, W.J. and Pietrokovski, S. (1995) Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene-COMBIS*, **163**, 17-26.

Johnson, M.S. and Overington, J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol*, **233**, 716-738.

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.

Lipman, D.J. and Pearson, W.R. (1985) Rapid and Sensitive Protein Similarity Searches. *Science*, **227**, 1435-1441.

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

Murata M., R.J.S., Sussman J.L. (1985) Simultaneus comparison of three protein sequences. *Proc. Natl. Acad. Sci. USA*, **82**, 443-453.

Myers, E. and Miller, W. (1988) Optimal Alignments in Linear Space. *CABIOS*, **4**, 11-17.

Needleman S. B., W.C.D. (1970) A general method applicable to search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, **48**, 443-453.

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.

Pearson, W.R. and Lipman, D.J. (1988) Improved Tools for Biological Sequence Analysis. *Proc. Natl. Acad. Sci. USA*, **85**, 2444-2448.

Richardson, J.S. (1981) The anatomy and taxonomy of protein structure. *Advances in prot. chemistry, Academic press inc.*, **34**, 167-369.

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

Rost, B. and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55-72.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTALw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acid Research*, **22**, 4673-4680.

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

Wilbur W. J., L.D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*, **80**, 726-730.

ANNEXE I. Source de 78 cas tests

aa.fas	glycosyl hydrolase family 13
ace.fas	alpha beta-hydrolase
adk.fas	nucleotide kinase
annexin	fas annexin
asp.fas	aspartic proteinase
az.fas	azurin/plastocyanin
bowman.fas	serine proteinase inhibitor
cbp.fas	calcium-binding protein
cryst.fas	crystallin
cys.fas	cysteine proteinase
cyt3.fas	cytochrome-c3
cyt5.fas	cytochrome-c5
cytc.fas	cytochrome-c
cyto.fas	cytokine
dhfr.fas	dihydrofolate reductase
egf.fas	EGF-like domain
fer2.fas	ferredoxin (2Fe-2S)
fer4.fas	ferredoxin (4Fe-4S)
flav.fas	flavodoxin
glob.fas	globin
gluts.fas	glutathione S-transferase
gpdh.fas	glyceraldehyde phosphate dehydrogenase
grs.fas	disulphide oxidoreductase
hip.fas	high potential iron protein
hla.fas	histocompatibility antigen-binding domain
hom.fas	DNA-binding homeodomain
hpr.fas	histidine carrier protein
igC1.fas	immunoglobulin domain
igV.fas	immunoglobulin domain
igcon.fas	immunoglobulin domain
igvar_h.fas	immunoglobulin domain
igvar_l.fas	immunoglobulin domain
il8.fas	interleukin 8-like protein
ins.fas	insulin
intb.fas	interleukin
kazal.fas	serine proteinase inhibitor
kringle.fas	kringle domain
kunitz.fas	serine proteinase inhibitor
ldh.fas	lactate/malate dehydrogenase
lipo.fas	lipocalin
ltn.fas	plant lectin
lys.fas	glycosyl hydrolase family 22
mmp.fas	matrix metalloproteinase
mthina.fas	metallothionein
mthinb.fas	metallothionein
mycin.fas	antibacterial protein
ndk.fas	nucleotide diphosphate kinase
neur.fas	glycosyl hydrolase family 34
p450.fas	cytochrome p450
parv.fas	calcium-binding protein
phoslip.fas	phospholipase A2
rep.fas	DNA-binding repressor
rhv.fas	picornavirus coat proteins
ricin.fas	ricin-like protein
rnasebact.fas	ribonuclease
rnasemam.fas	pancreatic ribonuclease
rnh.fas	ribonuclease H
rub.fas	rubredoxin
rubisco.fas	ribulose-1,5-biphosphate carboxylase/oxygenase
rvp.fas	retroviral proteinase
seatoxin.fas	sea anemone toxin
serbact.fas	serine proteinase
sermam.fas	serine proteinase
serpin.fas	serine proteinase inhibitor
sh2.fas	SH2 domain
sh3.fas	SH3 domain
sodcu.fas	Cu/Zn superoxide dismutase
sodfe.fas	Fe/Mn superoxide dismutase
squash.fas	serine proteinase inhibitor
subt.fas	subtilase
sugbp.fas	Sugar-binding-protein
thiored.fas	thioredoxin
tim.fas	triose phosphate isomerase
tln.fas	zinc metalloproteinase
tms.fas	thymidylate synthase
toxin.fas	snake toxin
xia.fas	xylose isomerase
zf_CCHH.fas	zinc finger