THESIS / THÈSE

MASTER EN SCIENCES BIOLOGIQUES

Alignement de structures tridimensionnelles de protéines par caractérisation des formes

Gengler, Christophe

Award date:
1994

Awarding institution:
Universite de Namur

Link to publication

FACULTES UNIVERSITAIRES N.-D. DE LA PAIX
NAMUR
FACULTE DES SCIENCES

ALIGNEMENT DE STRUCTURES TRIDIMENSIONNELLES DE PROTÉINES PAR CARACTÉRISATION DES FORMES

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

GENGLER Christophe

Décembre 1994

Facultés Universitaires Notre-Dame de la Paix
FACULTE DES SCIENCES
Rue de Bruxelles 61 - 5000 NAMUR
Tél. 081/72.41.11 - Telex 59222 Facnam-b - Telefax 081/72.44.20

Alignement de structures tridimensionnelles de protéines pararactérisation des formes.

GENGLER Christophe

Résumé

Dans l'optique des méthodes de modélisation de structures protéiques par comparaison de séquences (Knowledge-based modelling), les techniques d'alignement ont pris une importance capitale. En outre, vu le nombre toujours grandissant de séquences et de structures disponibles, ces techniques doivent être de plus en plus rapides.
La méthode SHOEBOX (Depiereux E. et Feytmans E., 1991, Biométrie - Praximétrie 34 13-34) propose une approche tout à fait originale. Elle est basée sur la comparaison de forme des segments protéiques et offre la possibilité de tenir compte ou non des chaînes latérales. De plus, elle est rapide, automatique et permet la réalisation d'alignements multiples pouvant compter jusqu'à 50 structures (40 000 résidus).
Notre travail a consisté en une évaluation des capacités de cette méthode. Il en est ressorti que cette mesure de distance ne paraît pas être adaptée à la réalisation d'alignements complets. Lorsque l'on ne considère que les atomes du squelette, il est possible de détecter les segments de conformation identique (de même structure secondaire). Mais, même si on tient compte des chaînes latérales, l'utilisation de la méthode SHOEBOX ne permet de sélectionner l'appariement correct que dans le cas où les SCR (régions structurellement conservées) comparées ont des séquences conservées et donc des chaînes latérales semblables.

Mémoire de licence en Sciences Biologiques
Décembre 1994
Promoteur : E. Feytmans

A l'issue de ce travail, je tiens à remercier M. le Professeur Feytmans de m'avoir accueilli dans son laboratoire. Son expérience et son savoir-faire m'ont beaucoup aidé.

Je tiens également à remercier M. Depiereux de m'avoir patiemment guidé tout au long de cette année.

Merci à Carla, Xavier et Guy pour leur aide et leurs conseils avisés.

Mes remerciements vont aussi à Mlle A. Tibor et à MM J. Remacle et D. Delforge qui ont accepté de lire ce mémoire.

Je ne voudrais pas oublier Jean-Yves, Damien, Fabrice et Etienne pour leur bonne humeur permanente.

Enfin, je tiens à associer à ce travail mes parents qui m'ont soutenu et encouragé, qui m'ont remonté le moral dans les moments de doute.

Je tiens encore à remercier Isabelle qui fut à mes côtés à tout instant.

Abréviations

Å	Ångström
ADM	average distance map
ADN	Acide désoxyribonucléique
COX	cholestérol oxydase
DFK	Depiereux-Feytmans-Kidera
DS	mesure de distance entre deux segments de structure
$FADH_2$	flavine adénine dinucléotide réduit
FCB	flavocytochrome β_2
FNR	ferredoxine
GOX	glycolate oxydase
GRS	glutathione oxydase
HCA	hydrophobic cluster analysis
MSP	maximum segment pair
MPM	matrice des probabilités de mutations
PAM	point accepted mutation
PDB	protein data bank
PHH	p-hydroxybenzoate hydrolase
pSCR	predicted structurally conserved regions
RMS	root mean square
SCC	structure-derived correlation coeficients
SCM	structure-derived matrix
SCR	structurally conserved regions
TPR	trypanothione reductase
TRB	thioredoxine réductase
VR	variable region

Table des Matières

I. INTRODUCTION.

Il est unanimement reconnu que la fonction d'une protéine est déterminée par sa structure. L'étude de cette dernière apparaît donc comme essentielle. Si les techniques de séquençage sont rapides et automatiques, la résolution d'une structure protéique fait appel à des techniques physiques laborieuses et difficiles à mettre en oeuvre. De plus, s'il est possible à partir de l'information contenue dans l'ADN de déduire la séquence d'une protéine, les mécanismes de repliement de cette séquence en structure tridimensionnelle sont inconnus.

Différentes approches existent pour déterminer la structure d'une protéine à partir de sa seule séquence. la première idée vient de la thermodynamique. D'après les lois de la thermodynamique, si la même conformation est toujours retrouvée, c'est parce que son énergie libre est minimale. Il en découle qu'en minimisant l'énergie d'une molécule par des méthodes numériques, il est possible de retrouver l'état de la protéine où son énergie est minimale. Malheureusement, les méthodes de minimisation convergent généralement vers des minima locaux.

Une seconde idée est la minimisation par comparaison de séquences. Elle se base sur un principe simple : deux séquences similaires doivent partager une même structure et une même fonction. Il faut donc à chaque protéine nouvellement séquencée associer une séquence (ou famille de séquences) similaire dont la structure est connue. Dans cette optique, les méthodes d'alignements de séquences et de structures ont une importance capitale.

Le programme MATCH-BOX (Depiereux et Feytmans, 1991, 1992, 1994) permet la réalisation d'alignements multiples de séquences et de structures protéiques.

Le centre de ce travail est constitué par l'analyse d'une mesure de distance originale utilisée par MATCH-BOX : la méthode SHOEBOX. Celle-ci permet la comparaison de segments de structures en se basant sur leur forme.

La première partie de ce mémoire a pour objectif, outre le rappel des notions de structure protéiques, de passer en revue les principales méthodes développées pour la comparaison de séquences et de structures protéiques.

Le chapitre « matériel et méthodes » consiste en la description des principales procédures de MATCH-BOX utilisées dans ce travail.

Enfin, une ultime partie rend compte du comportement de SHOEBOX et des raisons pouvant dicter ce comportement.

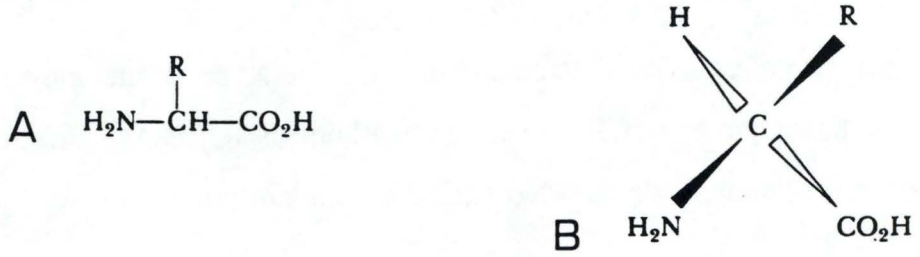


Figure I.1 : **a. Structure générale des acides aminés.**
b. Le Cα, atome central, est toujours sous la forme L.

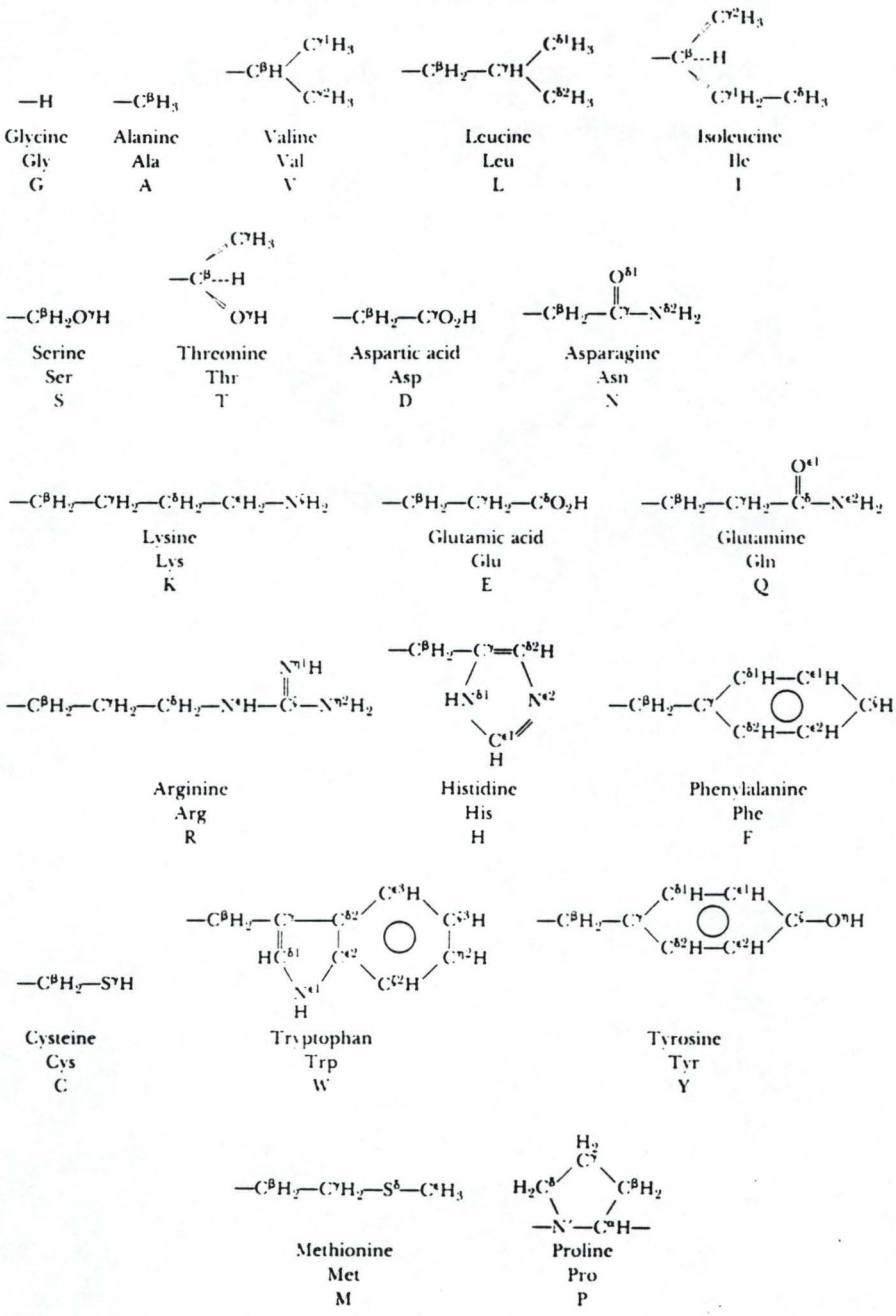


Figure I.2 : **Liste des acides aminés naturels et de leur chaîne latérale.**

I.1. STRUCTURE DES PROTÉINES.

Parmi les quatre principaux types de biopolymères, acides nucléiques, polysaccharides, assemblage de lipides et protéines, ces dernières sont sans doute les mieux connues : leur structure bien définie est adaptée à l'analyse.

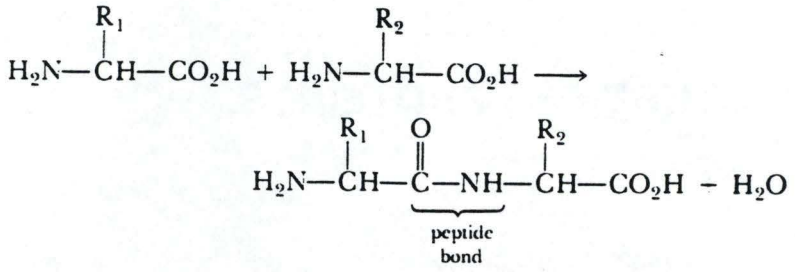
Les protéines stockent et transportent diverses particules (des électrons aux macromolécules), catalysent de nombreuses réactions, transmettent l'information, participent aux systèmes de défense de l'organisme, contrôlent l'expression des gènes, permettent l'activité musculaire et, de façon plus générale, forment l'architecture de toute cellule. Pratiquement, toute propriété caractérisant un organisme est effectuée par des protéines, ce qui fait d'elles une cible privilégiée pour les biologistes.

Les recherches sur les protéines furent mises à l'avant-plan il y a plus d'un siècle lorsque Hope-Seyler (1864) obtint des cristaux d'hémoglobine et lorsque Kühne (1876) purifia et caractérisa la trypsine. Depuis lors, de nombreuses techniques ont été développées dans le but de mieux connaître leurs propriétés et, en particulier, pour étudier leur fonction. L'impressionnante diversité des propriétés fonctionnelles des protéines ne peut se comprendre que par la relation avec la structure tridimensionnelle de ces dernières.

I.1.1. Structure primaire des protéines.

Les protéines sont des chaînes non ramifiées d'acides aminés unis par liens covalents (Fig. I.1). Les acides aminés communément rencontrés sont au nombre de vingt (Fig. I.2). Ils sont construits suivant la même charpente, portant chacun un atome d'hydrogène, une fonction carboxyle (COOH) et une fonction amine (NH_2), mais diffèrent les uns des autres par la nature et donc les propriétés de leur radical R (aussi

A



B

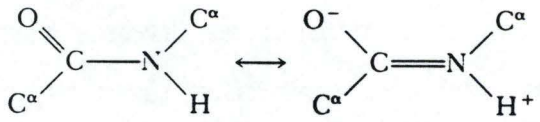


Figure I.3 : a. Le lien peptidique est formé par une réaction de condensation entre 2 acides aminés.
b. Un phénomène de résonance procure un caractère partiellement double au lien peptidique.

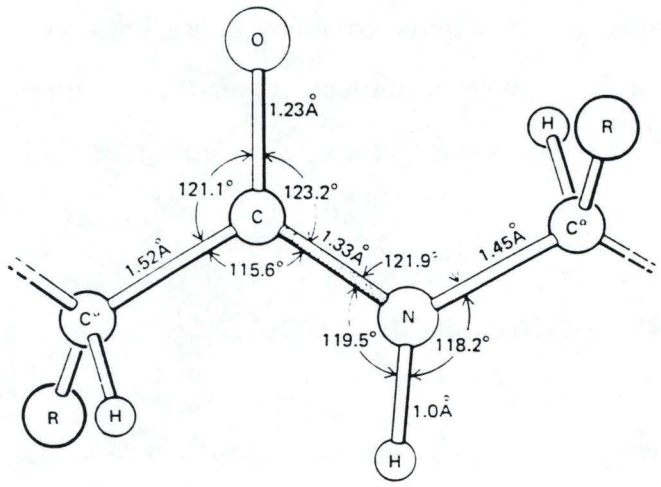


Figure I.4 : Géométrie du squelette peptidique. Les dimensions données sont des moyennes de mesures réalisées en cristallographie.

appelé chaîne latérale). Une seule exception à cette structure générale, la proline, pour laquelle on observe la formation d'un cycle par branchement de la chaîne latérale sur l'atome d'azote.

A l'exception de la glycine, pour laquelle le radical est uniquement constitué d'un hydrogène, l'atome central d'un acide aminé est un carbone asymétrique toujours présent sous la forme L (lévogyre).

Chaque type de molécule protéique possède une composition, une séquence et un poids moléculaire spécifiques codés sur l'ADN. Généralement, les chaînes formées comptent entre 50 et 3000 résidus.

N.B. On parlera de squelette, ou backbone, pour le polypeptide considéré sans ses chaînes latérales (il sera donc constitué d'une séquence répétée de trois atomes : N, C_α et C). Par convention, les résidus sont numérotés en commençant par l'extrémité N-terminale de la chaîne.

I.1.2. Structure secondaire des protéines.

Lors de la réaction de polymérisation, la perte d'une molécule d'eau permet la formation d'un lien amide partiellement double (40%). Il est stabilisé par un phénomène de résonance, les électrons de l'orbitale π étant partagés entre les liaisons C–O et C–N (Pauling *et al.*, 1951). Cela se traduit par un lien de 1,33Å, plus court que la liaison C–N classique (1,45Å) et plus long que la C=N (1,25Å), mais surtout par un lien rigide empêchant toute rotation (Fig. I.3 et I.4).

L'enchaînement des acides aminés entraîne l'établissement de nombreuses interactions entre chaînes latérales (forces électrostatiques, interactions de Van Der Waals, ponts Hydrogène, interactions hydrophobes,...). La protéine doit donc trouver un état, une conformation énergétiquement stable en réalisant des rotations autour des liaisons laissées libres.

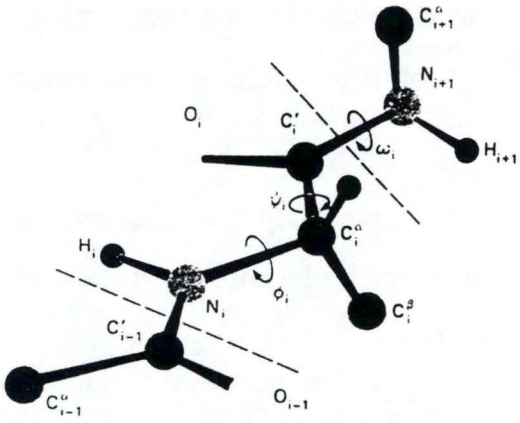


Figure I.5 : Localisation des différents angles de torsion le long de la chaîne peptidique.

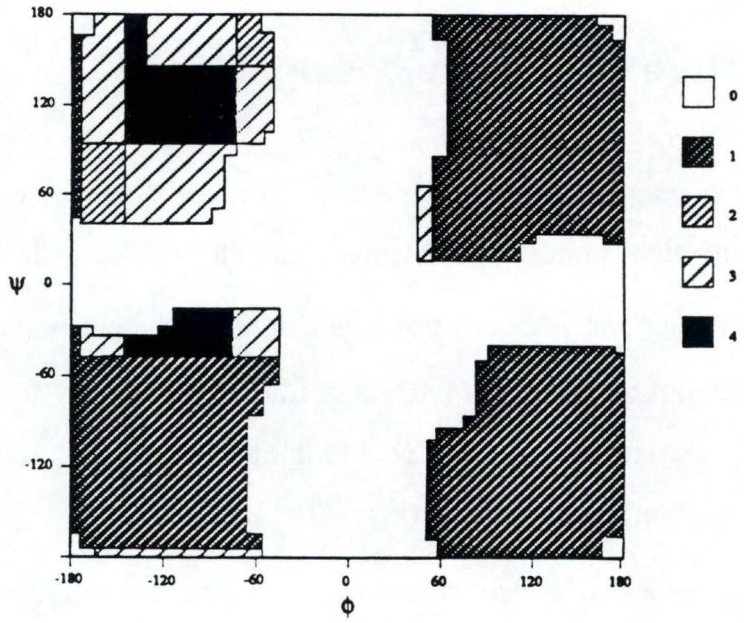


Figure I.6 : Graphe de Ramachandran indiquant la répartition théorique des angles de torsion φ et ψ pour les différents acides aminés.

- la glycine occupe les régions 1 à 4;
- l'alanine, les régions 2 à 4;
- les acides aminés à longues chaînes sont cantonnés aux régions 3 et 4;
- la valine et l'isoleucine ne se retrouvent que dans la région 4.

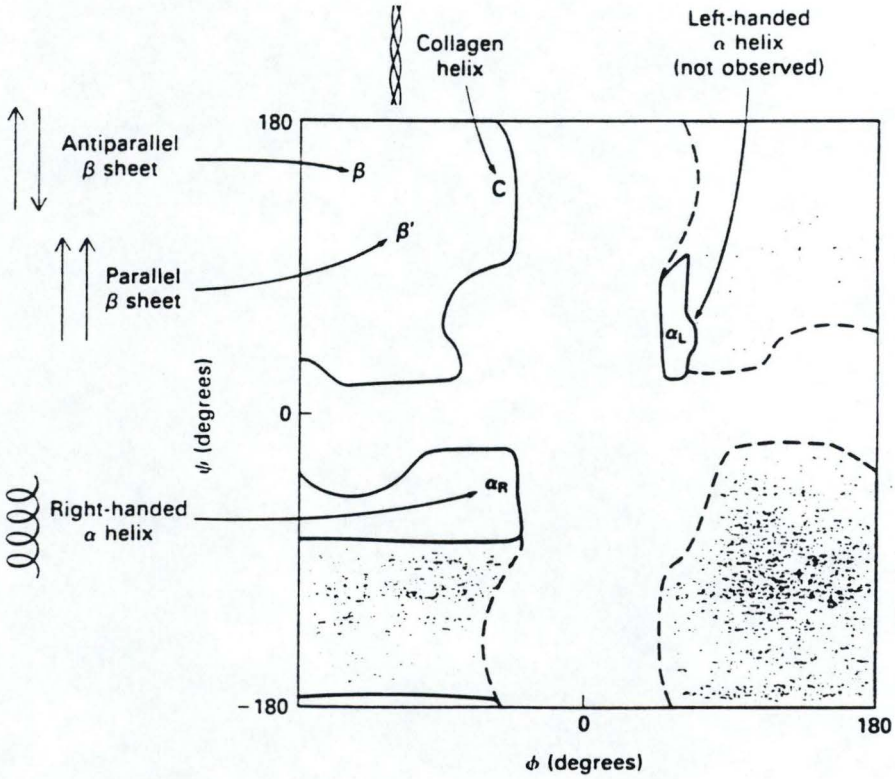


Figure I.7 : localisation des valeurs φ et ψ observées dans les différentes structures secondaires régulières.

AMINO ACID	α-HELIX (P_α)	β-SHEET (P_β)	REVERSE TURN (P_t)
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.28
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88

Figure I.8 : Table des fréquences d'apparition des 2 acides aminés naturels dans les principales structures secondaires régulières.

Par facilité, la convention fut prise de nommer les angles de torsion correspondant aux différents liens : ω pour le lien peptidique C–N, ϕ pour la liaison N–C_α et ψ pour la C_α–C. Les angles pour les liaisons des chaînes latérales sont désignés par χ_j où j indique la position de la liaison par rapport à la chaîne principale (Fig. I.5).

Les valeurs prises par ces angles sont limitées :

- ω : seules deux conformations sont possibles, l'une trans (180°), l'autre cis (0°). La seconde est nettement plus rare (un pour mille), sans doute par la proximité du carbone α et des chaînes latérales.

- ϕ et ψ : Certaines valeurs, pour lesquelles l'encombrement stérique est trop important, sont impossibles.

Les valeurs permises pour ϕ et ψ peuvent être reportées sur un diagramme de dispersion appelé plot de Ramachandran (Fig. I.6). Sur ce graphe on peut observer des régions où se regroupent les valeurs d'angles de torsion caractéristiques des différentes structures secondaires (Fig. I.7). Ces structures secondaires sont des arrangements locaux du polypeptide favorisés par les interactions entre chaînes latérales. Chaque acide aminé a donc, selon la nature de son radical, une préférence pour l'une ou l'autre structure secondaire (Fig. I.8).

I.1.2.1. L'hélice α (α helix).

L'hélice α, initialement décrite par Pauling *et al.* (1951), est l'élément de structure le plus classique et le mieux connu (Fig. I.9). C'est une hélice droite comptant 3,6 résidus par tour et pouvant atteindre jusqu'à 35 résidus de long. Elle est stabilisée par des ponts H s'établissant entre le C=O du résidu n et le N–H du n+4. Tous ces ponts ayant la même direction, ils font apparaître le long de l'axe un dipôle.

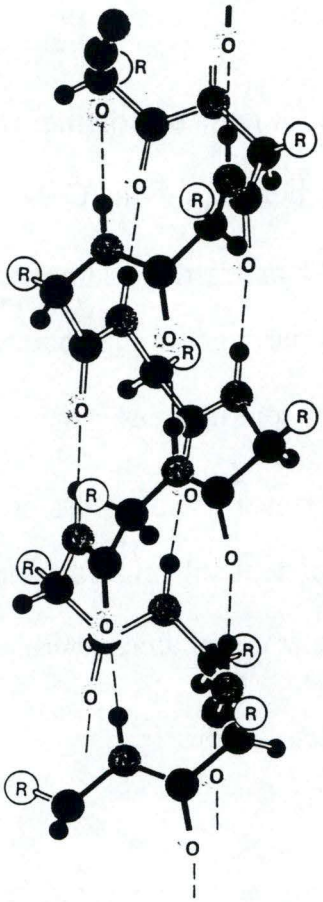


Figure I.9 : Représentation schématique de l'hélice α classique (3.6_{13}).

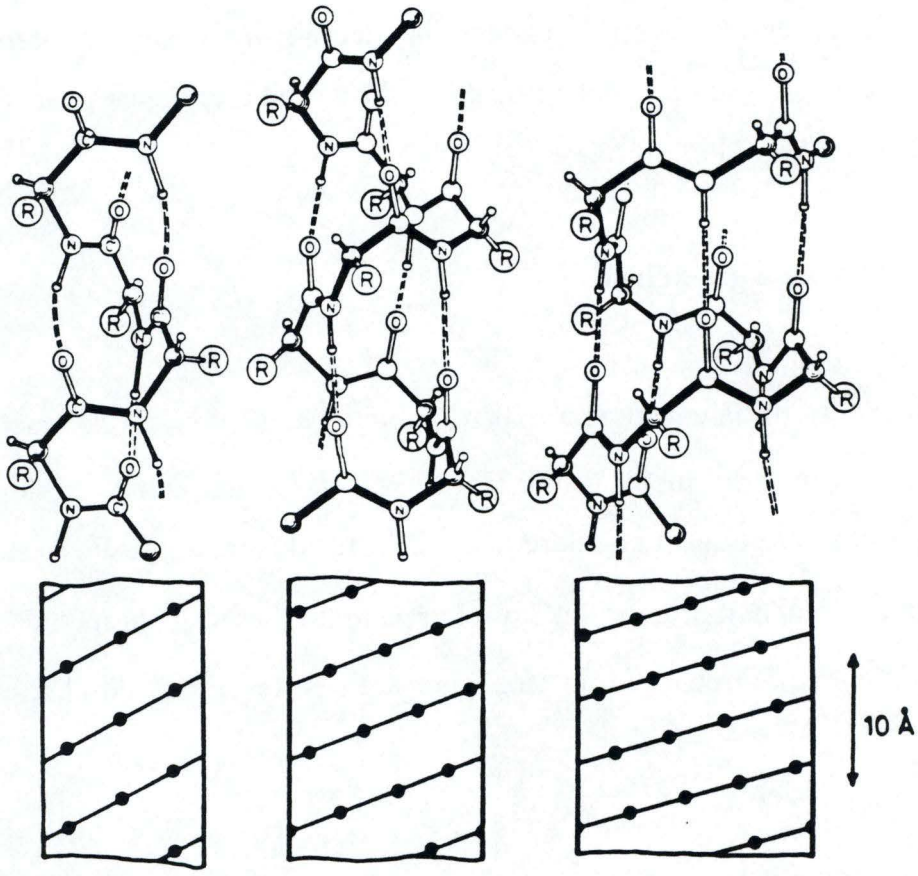


Figure I.10 : Représentation schématique des trois types d'hélice (3_{10}, 3.6_{13} et π)

Cette structure est aussi appelée hélice $3,6_{13}$, le 13 indiquant le nombre de résidus impliqués dans la boucle fermée par le pont H (···O–C–N–Ca–C–N–Ca–C–N–Ca–C–N–H···).

Pour cette structure, les valeurs de ϕ et ψ se situent autour de -60°. Certains acides aminés se retrouvent souvent dans les hélices, tels l'alanine, le glutamate, la leucine. D'autres comme la proline, la glycine, la tyrosine et la sérine y sont très rarement.

Il existe d'autres types moins fréquents d'hélices (Fig. I.10). L'hélice 3_{10} plus fine, plus compacte (3 résidus par tour, ponts H entre les résidus n et n+3) a des valeurs de ϕ et ψ égales respectivement à -60° et -30°. Cette structure se retrouve assez régulièrement mais n'est jamais très longue. Deux résidus en conformation 3_{10} forment un turn, de petites hélices 3_{10} existent dans de nombreux cas aux extrémités C-terminales des hélices α. 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 été observée dans les protéines.

I.1.2.2. Le plan β (β sheet).

Ce deuxième élément de structure secondaire est constitué de brins β, des chaînes polypeptidiques presque complètement étirées pouvant atteindre quinze résidus de long. Les valeurs de ϕ et ψ se situent dans le coin supérieur gauche du graphe de Ramachandran, elles sont respectivement de -150° et +135°. Ces brins peuvent interagir par des ponts H et ainsi former les feuillets β. On trouve deux types de plans β, les parallèles et les antiparallèles, différant par l'agencement des brins et le patron de ponts H (Pauling *et al.*, 1951). Les plans parallèles sont constitués de brins orientés dans le même sens et reliés par d'autres éléments de structure secondaire. Les

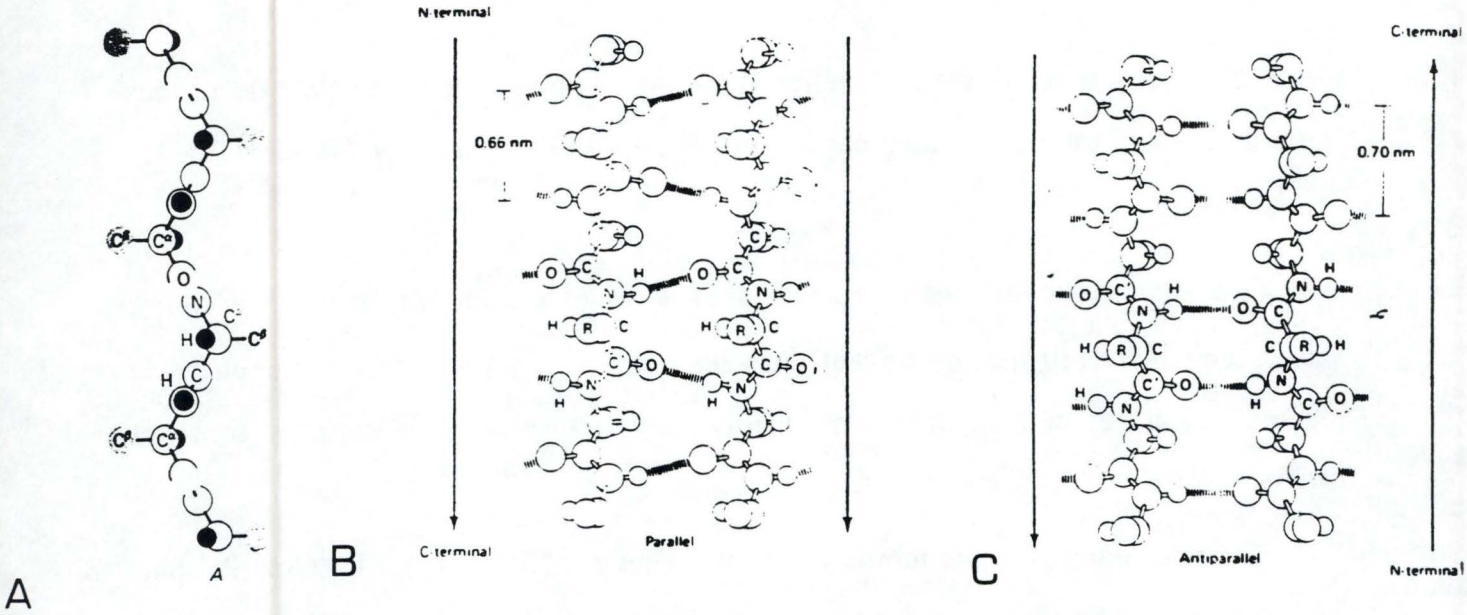


Figure I.11 : **Représentation schématique d'un brin β (a), d'un feuillet β parallèle (b) et d'un feuillet β antiparallèle (c).**

BEND TYPE	DIHEDRAL ANGLES OF TWO CENTRAL RESIDUES (°)[a]				NUMBER OF OBSERVED BENDS			
	ϕ_2	ψ_2	ϕ_3	ψ_3	Ideal bends[b]	Nonideal bends	Total bends	H-bonded bends[c]
I	−60	−30	−90	0	130	46	176	99
I'	60	30	90	0	8	5	13	10
II	−60	120	80	0	41	23	64	43
II'	60	−120	−80	0	15	5	20	16
III	−60	−30	−60	−30	66	11	77	45
III'	60	30	60	30	11	2	13	7
IV	A bend with two or more angles differing by at least 40° from those given above				0	35	35	5
V	−80	80	80	−80	1	2	3	0
V'	80	−80	−80	80	0	4	4	2
VI	A *cis* Pro at position 3				8	0	8	6
VII	A kink in the protein chain created by $\psi_2 \approx 180°$ and $\lvert\phi_4\rvert < 60°$ or $\lvert\psi_2\rvert < 60°$ and $\phi_3 \approx 180°$				8	0	8	1
Total					288	133	421	234

[a]The two central residues of a tetrapeptide β-turn are the $i + 1$ and $i + 2$ or 2nd and 3rd residues characterized by $(\phi, \psi)_2$ and $(\phi, \psi)_3$ given above.

[b]Bends that do not have any angle differing by more than 50° from the $(\phi, \psi)_2$ and $(\phi, \psi)_3$ for a particular bend type are considered as ideal.

[c]Bends with $O_{(i)}$ to $N_{(i+4)}$ distances < 3.5 Å were considered as hydrogen-bonded.

Figure I.12 : **Classification des différents turns et de leur fréquence d'occurrence relevées dans 26 protéines.**

antiparallèles présentent des brins en sens alternés entre lesquels les connections sont assurées par des boucles plus ou moins longues (Fig. I.11).

La plupart des plans observés dans la nature présentent une torsion dans le sens horlogique se traduisant par des valeurs ϕ et ψ plus positives (Chothia 1973). Cela semble dû aux tendances intrinsèques du squelette polypeptidique et aux interactions entre chaînes latérales.

I.1.2.3. Le coude (turn).

Les turns sont des éléments de structure plus courts que les deux premiers. Leur rôle est de permettre un changement de direction important de la chaîne polypeptidique. C'est, par exemple, le cas au niveau des feuillets β antiparallèles. Pour cette structure, et contrairement aux deux précédentes où les angles de torsions étaient répétitifs, la succession des valeurs ϕ et ψ a son importance (Fig. I.12). Les turns sont composés de trois ou quatre résidus. Ils peuvent être stabilisés par un pont H entre le C–O du résidu n et le N–H du n+3. Les résidus n+1 et n+2 sont responsables du turn proprement dit (Creighton, 1984).

Il existe trois types principaux de turns (Venkatachalam, 1968). Le type I possède une proline en position 3. Le type II nécessite des glycines en position 2 et 3. Le type III possède des valeurs ϕ et ψ répétitives (-60°, -30°) identiques à l'hélice 3_{10}.

I.1.2.4. La boucle (loop).

Les loops sont de longueurs et de conformations tout à fait irrégulières. Ils ont pour rôle d'effectuer la liaison entre les différents éléments de structure secondaire évoqués ci-dessus et de les maintenir au coeur de la protéine.

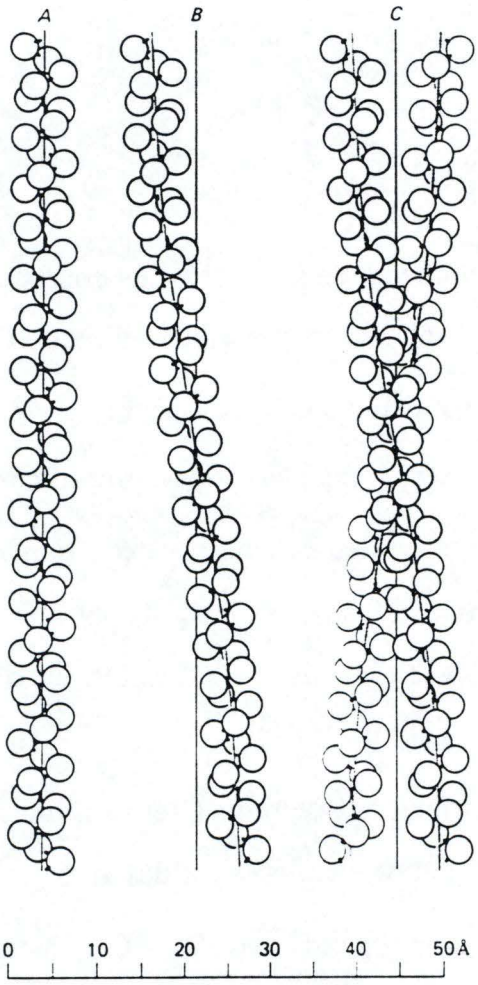


Figure I.13 : Représentation schématique de la coiled-coil-α-helix.

Du fait de leur situation en surface de la protéine, les boucles sont souvent riches en acides aminés hydrophiles et chargés pouvant interagir avec le solvant. On a pu remarquer que les variations de séquence et de structure apparues au cours de l'évolution intervenaient surtout au niveau des loops et non des autres éléments de structure secondaire. Il existe pourtant des exceptions dans le cas ou les boucles ont un rôle fonctionnel (fixation de substrat, de cofacteur,...), elles sont alors invariables.

I.1.3. Superstructure secondaire.

Les structures secondaires peuvent s'enchaîner et prendre une conformation particulière fréquemment rencontrée dans les protéines. Ce sont les superstructures secondaires dont voici quelques exemples.

I.1.3.1. Coiled-coil α-hélix.

Dans cette structure, deux hélices α s'enroulent l'une autour de l'autre pour former une « super-hélice » gauche (Fig. 13). Des coiled-coil α-helix ont été trouvées dans des protéines fibreuses (α-kératine, topomyosine,...) et dans des protéines globuleuses (protéine du manteau du virus de la mosaïque du tabac, bactériorhodopsine,...).

I.1.3.2. Hélice α-loop-héliceα.

Cette structure est fréquente notamment dans les protéines fixant le Ca^{++} et l'ADN.

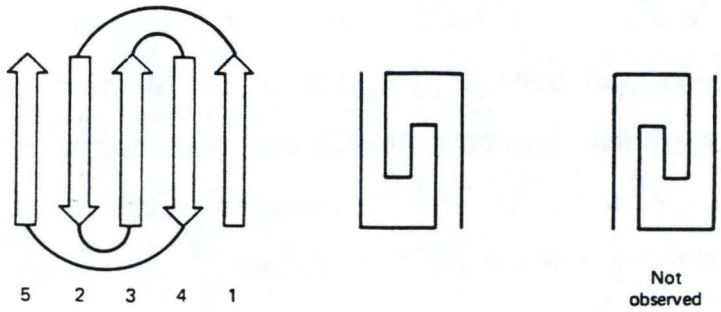


Figure I.14 : **Représentation schématique du motif en clé grecque.**

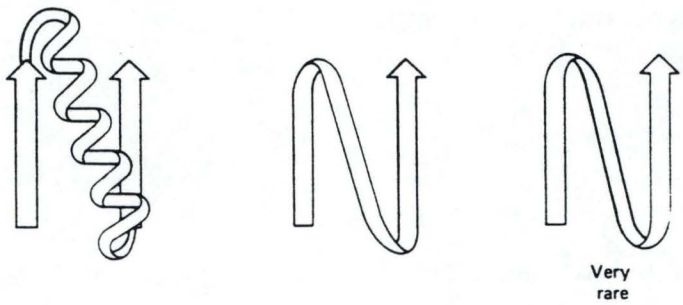


Figure I.15 : **Représentation schématique du motif βαβ.**

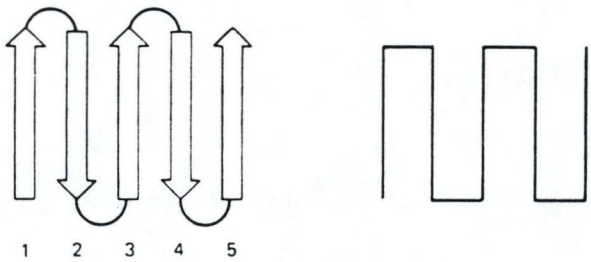


Figure I.16 : **Représentation schématique du motif β-β dans un feuillet antiparallèle.**

I.1.3.3. Clé grecque.

Ce motif, composé de quatre brins β antiparallèles liés par des loops, rappelle une décoration fréquemment utilisée dans l'architecture grecque (Fig. I.14). Elle est souvent rencontrée dans les feuillets β antiparallèles.

I.1.3.4. Motif βξβ (βxβ).

Cette structure est constituée de deux brins β reliés par une chaîne irrégulière (βcβ), par une hélice (βαβ, Fig. I.15), ou par un autre brin β (βββ). Ces éléments peuvent s'enchaîner comme dans le cas du Rossmann-fold, un motif βαβαβ (Rao et Rossmann, 1973).

I.1.3.5. Motif β-β.

Pour former cette superstructure secondaire, un loop joint deux brins β antiparallèles. On la retrouve fréquemment dans les plans β (Fig. I.16).

I.1.4. Structure tertiaire.

Les structures et superstructures secondaires sont, dans la protéine, agencées de manière compacte afin de former un ou plusieurs domaines. Ces domaines sont chacun responsables d'une fonction particulière (fixation de substrat, de coenzyme, site catalytique,...). Son repliement sur elle-même permet à la chaîne polypeptidique de positionner les résidus importants pour sa fonction. Ceux-ci peuvent être très distants dans la séquence.

I.1.5. Structure quaternaire.

Les protéines multimériques sont constituées de plusieurs structures tertiaires, ou domaines, associées en structure quaternaire. Le nombre de sous-unités varie d'une protéine à l'autre. Ces monomères peuvent soit être identiques (formation d'homodimère, d'homotétramère,...), soit être de nature et donc de fonction différentes (formation d'hétérodimère, d'hétérotétramère,...). Les interactions apparaissant au niveau des surfaces de contact entre domaines sont similaires à celles existant à l'intérieur des protéines (interactions hydrophobes, ponts H, ponts disulfures,...). Les principaux types de zones de contact sont : les compactages hélice α-hélice α, les compactages feuillet β-feuillet β, les feuillets β étendus, les interactions loops-loops.

I.2. PRÉDICTION DE LA STRUCTURE DES PROTÉINES.

Il est accepté, à l'heure actuelle, que la structure d'une protéine est responsable de sa fonction. Il apparaît donc essentiel de connaître cette dernière.

En 1993, le nombre de structures résolues s'élevait à 1000 alors que plus ou moins 50000 séquences étaient connues. Ce déséquilibre ne fait qu'augmenter : en effet, le taux de détermination des séquences est au moins 50 fois plus élevé que celui des structures (Bowie *et al.*, 1991). La résolution d'une structure fait appel à des techniques physiques laborieuses et difficiles à mettre en oeuvre (résonance magnétique nucléaire, cristallographie) et peut donc prendre plusieurs années. Au contraire, les grands projets de séquençage de génome permettent, en un court laps de temps, de déterminer un grand nombre de séquences.

Dans cette optique, les méthodes de prédiction de structures ont une importance énorme.

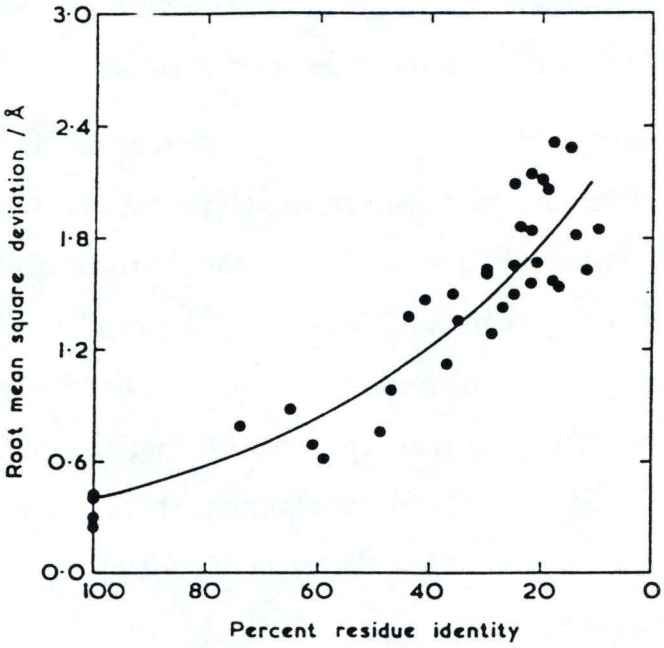
I.2.1. Méthodes cinétiques et thermodynamiques.

Ces méthodes reposent sur l'hypothèse thermodynamique selon laquelle une protéine se replie de façon à atteindre la plus faible énergie libre (Afinsen et Scheraga, 1975; Némethy et Scheraga, 1977). Toute l'information nécessaire à ce replis est donc contenue dans la séquence (Afinsen et Haber, 1973). Cette hypothèse n'a, à l'heure actuelle, jamais été prouvée. Elle est pourtant rendue très probable par la stabilité de nombreuses protéines et par le fait que certaines protéines dénaturées peuvent se renaturer après retour aux conditions normales. Cette renaturation permet de retrouver la structure native de la protéine et donc un rétablissement de la fonction.

Il existe cependant certaines observations mettant cette hypothèse en doute : le repliement d'une protéine peut changer selon les conditions du milieu (Scheraga *et al.*, 1984; Montelione et Scheraga, 1989) et il peut, dans certains cas, être dépendant de certaines protéines appelées chaperones (Ellis et Hemmingsen, 1989). Cela pourrait expliquer l'impossibilité, pour certains systèmes biologiques, de produire une protéine fonctionnelle après incorporation du génome d'un autre organisme. Certaines étapes indispensables au folding dépenderaient de ces chaperones.

Le but de cette méthode est donc la recherche de la conformation la plus énergétiquement stable. Si l'on considère un polypeptide de 100 résidus, il peut exister sous 10^{100} conformations différentes (en moyenne 10 par résidus). Sachant que la protéine passe d'une conformation à l'autre en 10^{-13} secondes, il lui faudra 10^{77} années pour essayer toutes les combinaisons. Les protéines nécessitant, *in vitro* et *in vivo*, entre 10^{-1} et 10^{3} secondes pour leur repliement, il faut en conclure qu'elles n'essaient pas toutes les possibilités (Sternberg et Thornton, 1978). La structure native d'une protéine est donc située au niveau d'énergie le plus bas parmi les conformations permises par la cinétique.

Plusieurs représentations algorithmiques ont été développées pour calculer empiriquement des potentiels d'énergie et en rechercher le minimum. Les champs de force les plus répandus sont AMBER (Weiner *et al.*, 1984, 1986), CHARMM (Brooks *et al.*, 1983), DISCOVER (Dauber-Osguthorpe *et al.*, 1988)et ECEPP (Momany *et al.*, 1975; Némethy *et al.*, 1983). Les fonctions utilisées pour les différentes composantes de cette énergie reposent sur des concepts physiques concernant la nature des composants. Elles contiennent aussi de nombreuses constantes décrivant la géométrie moléculaire (longueur des liaisons, angles de torsions, interactions entre atomes,...). Ce genre de calcul ne tient compte ni du milieu, ni de l'entropie du système. De plus, il est impossible de parcourir tout l'espace conformationnel (ensemble des conformations potentiellement attribuables à un polypeptide) et on risque donc de passer à côté du minimum global.



The relation of residue identity and the r.m.s. deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (see Table II).

Figure I.17 : Relation entre le pourcentage d'identité des noyaux de 32 protéines homologues et leur distance RMS.

Il existe différentes solutions à ce problème. La plus évidente est de commencer le processus à partir d'une conformation proche du minimum global. Une autre consiste à déterminer les structures secondaires, sites de nucléation autour desquels le reste de la protéine se replie. Il existe différentes méthodes de prédiction de structures secondaires (par exemple, Chou et Fasman, 1974; Lim, 1974; Sternberg et Thornton, 1978; Garnier *et al.*, 1978; Kabsch et Sander, 1983), mais leur taux de réussite n'atteint guère plus de 70% (Rost *et al.*, 1993). De plus, passé ce stade, il faut réaliser l'assemblage des éléments de structure secondaire, étape encore très délicate (Benner, 1992).

I.2.2. Modélisation par comparaison de séquence.

La comparaison de la structure tertiaire de protéines homologues montre que celle-ci est mieux conservée, dans l'évolution, que la structure primaire et surtout que la séquence d'ADN. Dans des molécules récemment divergentes, l'arrangement des éléments de structure secondaire est topologiquement comparable. Les remplacements d'acides aminés apparaissent le plus souvent en surface de la protéine, de sorte que la conformation de la chaîne principale est peu affectée. En accord avec ces observations, Chothia et Lesk (1986) ont pu montrer qu'il existait une relation entre le degré de divergence de la composition en acides aminés des régions du noyau et la divergence de structure correspondante. Cette relation est exponentielle (Fig. I.17) car les protéines acceptent beaucoup plus facilement les mutations en surface que les mutations qui concernent les résidus enfouis dans le noyau. En toute généralité, les protéines très apparentées diffèrent seulement par leurs résidus de surface alors que les protéines éloignées diffèrent en plus par leurs résidus constitutifs du noyau.

Différente des méthodes *ab initio*, l'approche « knowledge-based » de la prédiction de structure protéique tente de relier des séquences à des structures tridimensionnelles connues. En effet, si une séquence présente des similarités avec une

autre protéine résolue, on peut supposer qu'elles ont même structure et même fonction. Cette voie de recherches ouvertes par Browne et ses collaborateurs (1969) repose sur différentes observations :

- Les protéines sont groupées en familles de structures (Richardson, 1981).
- Le nombre de familles est limité, il est estimé à 1000 (Chothia, 1992).
- Le nombre de familles de topologies différentes se situe entre 500 et 700 (Blundell et Johnson, 1993).
- Approximativement 50% des structures nouvellement résolues sont en fait liées à des folds connus (Blundell et Doolittle, 1992).

Le but de cette méthode est donc de trouver des correspondances entre les résidus de différentes séquences, et ce afin de déterminer les régions structurellement conservées (SCR). Dans cette optique, les méthodes d'alignement de protéines ont une importance capitale.

Il est important de se rendre compte que les structures connues risquent de ne représenter qu'un échantillon non représentatif de l'ensemble des protéines. En effet, pour les étudier, elles doivent pouvoir être surproduites et cristallisées ce qui n'est pas le cas, par exemple des protéines trans-membranaires.

I.3. Comparaison de séquence.

A l'heure actuelle, les méthodes de comparaisons de séquences jouent un rôle capital dans la biologie moléculaire. L'étude et la classification d'une banque de séquences, la recherche dans une banque de séquences, la détection de résidus invariants (et donc ayant une importance au niveau fonctionnel), l'alignement de deux ou plusieurs séquences en vue de modélisation,... autant de problèmes rencontrés de plus en plus souvent et qui nécessitent une solution à la fois correcte et rapide.

Les points suivants auront pour but de faire une présentation de ces méthodes de comparaisons de séquences. Cette présentation, plus précisément centrée sur la problématique de l'alignement, n'est pas exhaustive mais évoque des techniques ayant marqués un tournant dans l'histoire de l'alignement de séquences.

I.3.1. Les matrices de scores.

Toute méthode d'alignement de séquences repose sur une série de scores caractérisant la similarité ou la distance entre chacune des 210 paires d'acides aminés possibles. La représentation la plus communément utilisée est celle des matrices de scores, matrices symétriques 20X20 où seuls 210 coefficients sont originaux (la paire A-I est considérée comme équivalente à la paire I-A).

On distingue deux types de matrices de scores :

- les matrices de similarité où des acides aminés de caractéristiques identiques ou similaires reçoivent un score élevé par rapport à ceux de caractères différents.
- les matrices de distance où la situation est inversée, plus les acides aminés sont différents, plus leur score est élevé.

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

Figure I.18 : Matrice d'identité.

I.3.1.1. Les matrices de similarité.

I.3.1.1.1. Matrice identité.

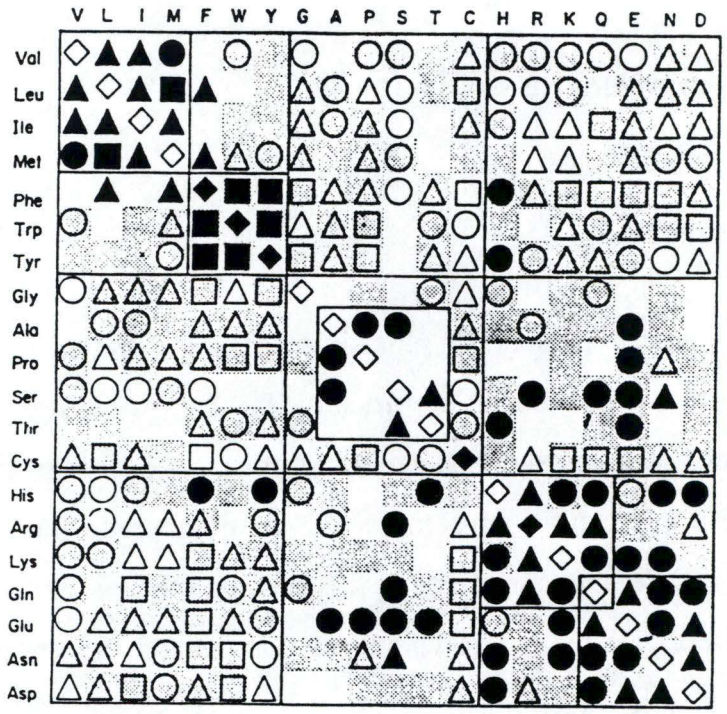
C'est la matrice de score la plus simple; les paires d'acides aminés sont classés en deux catégories : identiques et non-identiques. Les paires non-identiques reçoivent un score de 0, alors que les autres se voient attribuer un score positif, généralement 1 (Fig. I.18). Cette matrice ne tient compte que du nombre de résidus identiques entre les deux séquences, de plus elle donne la même importance à toutes les substitutions. Bien qu'étant limitée, elle permet le calcul du pourcentage d'identité, valeur fréquemment utilisée pour évaluer la similarité globale entre deux séquences.

I.3.1.1.2. Matrices de substitution.

I.3.1.1.2.1. Matrice de Mc Lachlan (1971).

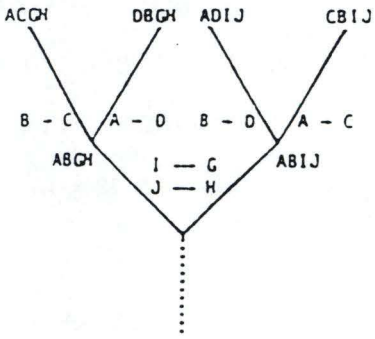
Mc Lachlan a travaillé sur 17 familles de séquences actuelles dans le but de calculer les fréquences de substitution d'un résidu par un autre. Selon lui, celles-ci rendent compte de la ressemblance entre résidus. Les substitutions fréquemment observées sont celles qui n'ont pas défavorisé la protéine, ce sont donc celles survenues entre deux acides aminés proches point de vue physico-chimique. Au contraire, un remplacement par un acide aminé tout à fait différent va affecter la conformation du site et donc rendre la protéine moins fonctionnelle.

La première étape consiste à compter, dans les alignements réalisés, le nombre de fois que le résidu i est remplacé par j ou vice et versa ($N_{ij}=N_{ji}$). Cela revient à compter le nombre de paires d'acides aminés ij rencontrées. Ce nombre est divisé par le nombre total de remplacements observés pour toutes les paires d'acides aminés (N_1). Ces deux termes sont ensuite divisés par le produit des occurrences des acides aminés i et j



Relative substitution frequencies $f(i,j)$ in homologous proteins.
●, ■, ▲, Values which are higher than average: ■, over 3·0; ▲, 3·0 to 1·75; ●, 1·74 to 1·30; blank squares, 1·29 to 0·80. ○, □, △, Values below average: ○, 0·79 to 0·60; △, 0·59 to 0·33; □, below 0·33. Shaded squares show mutations which are forbidden because they require more than one base change.

Figure I.19 : Fréquences relatives de substitutions des acides aminés f(i,j) observées par McLachlan (1972) dans 17 familles de protéines homologues.



A

	A	B	C	D	G	H	I	J
A			1	1				
B			1	1				
C	1	1						
D	1	1						
G							1	
H								1
I					1			
J						1		

B

Figure I.20 : a. Arbre phylogénétique simplifié (Dayhoff, 1971) Les séquences ancestrales inférées à partir des séquences actuelles sont placées dans la partie inférieur de la figure tandis que les séquences actuelles sont placées dans la partie supérieure. Les substitutions sont indiquées à côté des branches.

b. Matrice des mutations ponctuelles acceptées dérivées de l'examen de l'arbre phylogénétique.

Nous pouvons remarquer que l'acide aminé A est substitué soit en D soit en C. Une comparaison des séquences finales aurait entraîné la prise en compte de la substitution de D en C.

($n_i \times n_j$), lui-même divisé par la somme sur tous les acides aminés de ce produit $n_i \times n_j$ (N_2).

$$f_{ij} = \frac{\frac{N_{ij}}{N_1}}{\frac{n_i n_j}{N_2}} = \frac{N_{ij}}{N_1} \times \frac{N_2}{n_i n_j}$$

La figure I.19 montre les classes de fréquences de substitutions relatives, observées pour les différentes paires de résidus.

Sur base de ces résultats, une matrice de scores peut être calculée de la façon suivante :

- un score de 0, 1 ou 2 est attribué aux substitutions les plus rares;
- un score de 3 caractérise les substitutions neutres;
- un score de 4, 5 ou 6 est attribué aux substitutions les plus fréquentes;
- pour les identités, le score est fixé à 8, sauf dans le cas des identités Phe, Tyr, Trp et Cys où il est de 9.

I.3.1.1.2.2. Matrice de Dayhoff (1972).

L'idée développée par Dayhoff est semblable à celle de Mc Lachlan, deux résidus peuvent s'échanger s'ils ont des propriétés physico-chimiques proches. Pourtant il existe entre les deux méthodes de très importantes différences au niveau de la démarche. Dayhoff, contrairement à Mc Lachlan, suit l'apparition de substitutions dans des familles de protéines homologues classées sur des arbres phylogénétiques. La fréquence d'une substitution entre une séquence ancestrale et une séquence actuelle indique si elle est plus ou moins bien acceptée par la sélection naturelle au cours de l'évolution. De plus ce calcul évite une surestimation des remplacements (Fig. I.20). Dayhoff stocke dans une matrice appelée matrice de mutation (mutation data) le nombre de substitutions pour chaque paire d'acides aminés.

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.21 : Matrice de probabilité de mutation de Dayhoff (1971) sur une distance évolutive de 2 PAM. Un élément M_{ij} donne la probabilité qu'un acide aminé de la colonne j soit remplacé par un acide aminé de la ligne i après un intervalle évolutif de 2 mutations acceptables par 100 acides aminés. Toutes les valeurs sont multipliées par 10 000.

Ensuite le calcul de la probabilité de remplacement d'un acide aminé i par un acide aminé j peut être estimé par l'expression suivante :

$$M_{ij} = \frac{N_{ij}.N_i.100}{n_i.N_1}$$

avec : N_{ij}, la fréquence de remplacement de l'acide aminé i par l'acide aminé j;

N_i, la fréquence de remplacement de l'acide aminé i;

N_1, la fréquence de remplacement totale;

n_i, la fréquence de l'acide aminé i.

Cette fraction M_{ij} permet d'exprimer une fréquence de substitution normalisée selon la longueur de la séquence, la fréquence des acides aminés, leur exposition aux différents remplacements. 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, pendant l'intervalle évolutif au cours duquel une seule substitution est acceptée. Ces valeurs permettent donc la construction d'une matrice des probabilités de mutations (MPM) où chaque élément ij donne la probabilité que i soit remplacé par j. L'unité d'évolution représentée par cette matrice correspond donc à une mutation acceptée pour 100 résidus (1 PAM : Point Accepted Mutation).

Il est possible, par produit matriciel, de calculer des matrices de probabilités sur des intervalles évolutifs plus grands. La matrice 2 PAM (Fig. I.21) est le produit de la matrice 1 PAM par elle-même, la matrice 3 PAM est obtenue en multipliant la 1 PAM deux fois par elle-même et ainsi de suite. Ces opérations permettent d'obtenir toute une famille de matrices de similarités situées entre deux matrices théoriques extrêmes 0 PAM (matrice unité) pour laquelle aucun remplacement n'a le temps d'être observé et ∞ PAM (matrice dans laquelle chaque ligne donne la fréquence d'occurrence d'un acide aminé) permettant le remplacement de tout acide aminé par n'importe quel autre.

C Cys	12	C - sulfhydryl																			
S Ser	0	2																			
T Thr	-2	1	3																		
P Pro	-3	1	0	6	S - small hydrophilic																
A Ala	-2	1	1	1	2																
G Gly	-3	1	0	-1	1	5															
N Asn	-4	1	0	-1	0	0	2														
D Asp	-5	0	0	-1	0	1	2	4													
E Glu	-5	0	0	-1	0	0	1	3	4	N - acid, acid amide, hydrophilic											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4											
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6										
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6	H - basic								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	V - small hydrophobic					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				
F Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	F - aromatic	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	3	-3	-4	-5	-2	-6	0	0	17	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp	

Figure I.22 : Matrice de scores de Dayhoff pour une distance évolutive de 250 PAM. Un score positif indique une substitution plus fréquente dans les séquences proches que dans des séquences prises au hasard. Le score neutre est 0. Un score négatif indique une substitution moins fréquente dans des séquences proches que ce qu'on aurait trouvé par hasard (George *et al.*, 1990).

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

Figure I.23 : Matrice recommandée par Gonnet *et al.* (1992) pour l'alignement initial de séquences protéiques. Chaque élément (i,j) de cette matrice est le logarithme en base 10 du rapport entre les probabilités que i et j soient alignés et la probabilité de les trouver face à face par hasard. Ces valeurs sont normalisées pour deux protéines séparées par 250 PAM.

Une dernière modification est faite pour arriver à une matrice finale R, la « relatedness odds matrix » dont chaque terme représente la probabilité de remplacement d'un acide aminé i par un acide aminé j par occurrence de i par occurrence de j. . En effet, même si le résidu i apparaît dans une séquence au cours de l'évolution par remplacement d'un résidu j dans cette séquence, il se peut que ce résidu se trouve par chance à la même position sans pour autant qu'il existe un lien évolutif entre les deux séquences comparées. Les éléments R_{ij} sont obtenus de la manière suivante :

$$R_{ij} = \frac{M_{ij}}{f_i}$$

où f_i, la fréquence normalisée, est la probabilité que l'acide aminé i se trouve, par chance, dans la seconde séquence

La matrice de Dayhoff, particulièrement la 250 PAM (Fig. I.22), est souvent utilisée. Cependant, Risler *et al.* (1988) ont montré que cette matrice souffrait du fait qu'elle n'a pas été remise à jour depuis 1978. En effet, les substitutions ont été comptées sur un nombre limité de protéines (depuis lors de nombreuses séquences ont été élucidées) qui, de plus, appartenaient à des familles très proches.

I.3.1.1.2.3. Matrice de Gonnet et al.(1992).

Gonnet et ses collaborateurs ont réalisés un travail semblable à celui de Dayhoff, à la différence près que le nombre de séquences déterminées a considérablement augmenté, de même que le nombre de substitutions. Les nouveaux scores obtenus (Fig. I.23) sont donc plus représentatifs, plus significatifs et plus fiables que ceux calculés par Dayhoff. Selon les auteurs, leur matrice est plus adaptée que celle de Dayhoff pour la comparaison de séquences de similarité moyenne.

I.3.1.1.2.4. Matrice de Henikoff (1992, 1993).

Henikoff va, pour la construction de sa matrice, utiliser des fréquences d'occurrence observées dans des alignements locaux constitués de blocs de séquences sans gap. Chaque bloc représente donc une région conservée dans un groupe de protéines. Ces travaux réalisés sur plusieurs centaines de protéines ont permis de composer une banque de plus de 2000 blocs. Par la suite, la fréquence de chacune des 210 paires d'acides aminés est calculée et placée dans une table. La probabilité observée de chaque paire de résidus i et j (q_{ij}) peut donc être calculé de la manière suivante :

$$q_{ij} = \frac{f_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} f_{ij}}$$

Avec :

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 (p_i) et la probabilité d'occurrence de la paire ij (e_{ij}) :

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

$$e_{ij} = p_i p_j \text{ si } i = j$$
$$2 p_i p_j \text{ si } i \neq j$$

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Figure I.24 : Matrice BLOSUM 62 (Henikoff, 1992).

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

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

où S_{ij} est égale à 0, plus petite que 0 ou plus grande que 0, selon que les fréquences observées sont égales, plus petites ou plus grandes que ce qui est attendu. La matrice obtenue est appelée BLOSUM pour blocks substitution matrix.

Une dernière amélioration a encore été apportée à cette matrice pour éviter que les membres de familles de protéines fortement apparentées n'aient trop de poids. Prenons le cas où, à la position x d'un alignement de 10 séquences, on retrouve 9 alanines et une sérine. La paire AA pèsera donc beaucoup plus que la paire AS. La solution proposée est de regrouper les protéines ayant un certain pourcentage d'identité et de faire la moyenne de leurs contributions au point de vue des fréquences d'occurrence des paires d'acides aminés.

Selon la limite de pourcentage choisie, on obtiendra donc des matrices accordant plus ou moins de poids aux segments de similarité importante. Une matrice pour laquelle la limite est de 62% est appelée BLOSUM 62 (Fig. I.24).

I.3.1.1.2.5. Matrice de Johnson et Overington(1993).

Contrairement aux quatre précédentes, cette matrice est calculée sur base de fréquences de substitutions observées dans des alignements structuraux impliquant 235 protéines. La première étape consista à classer les 235 structures en familles homologues, puis, au sein de ces familles de réaliser les alignements. Au total, plus ou moins 200.000 substitutions furent observées et accumulées dans une table de fréquences.

Structure-based amino acid scoring table (upper triangle); all positive table (lower triangle); all values have been multiplied by 10

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
	6·0	−3·4	−1·6	−0·7	−3·2	−0·5	−3·1	−2·2	−0·9	−3·3	−1·5	−1·4	−1·0	−0·6	−1·6	0·0	−0·8	−0·5	−5·8	−4·0	A
		16·1	−9·7	−6·9	−4·4	−8·2	−8·2	−7·7	−8·7	−8·7	−4·4	−7·6	−8·9	−6·9	−5·6	−7·7	−6·0	−4·8	−9·1	−7·7	C
A	15·8		8·5	2·4	−7·0	−2·1	−0·7	−4·8	−1·5	−8·0	−5·9	2·6	−1·0	−1·1	−3·4	−0·2	−1·8	−5·2	−6·0	−3·8	D
C	6·4	25·9		8·6	−6·4	−2·5	−2·3	−4·8	1·1	−5·6	−2·8	−0·7	−1·5	2·4	−0·2	−2·2	−0·5	−4·2	−7·6	−3·7	E
D	8·2	0·1	18·3		10·4	−8·6	−1·7	0·5	−5·6	1·8	−0·6	−3·8	−5·0	−6·4	−6·0	−4·8	−5·0	−1·3	3·4	3·4	F
E	9·1	2·9	12·2	18·4		8·0	−3·2	−5·5	−3·5	−7·2	−5·2	−1·4	−2·5	−2·8	−2·8	−1·3	−3·8	−5·6	−6·3	−5·4	G
F	6·6	5·4	2·8	3·4	20·2		12·7	−5·1	0·1	−4·2	−2·3	1·7	−4·3	1·4	0·1	−2·6	−3·0	−3·9	−4·0	−0·4	H
G	9·3	1·6	7·7	7·3	1·2	17·8		8·1	−4·7	2·6	2·6	−4·7	−5·7	−7·0	−5·4	−4·7	−3·2	3·9	−3·3	−2·5	I
H	6·7	1·6	9·1	7·5	8·1	6·6	22·5		7·6	−3·4	−1·9	0·1	−0·6	1·1	3·2	−1·5	−0·2	−3·7	−5·4	−3·7	K
I	7·6	2·1	5·0	5·0	10·3	4·3	4·7	17·9		7·3	4·4	−4·8	−2·8	−4·4	−3·7	−5·2	−4·6	1·8	−1·0	−2·4	L
K	8·9	1·1	8·3	10·9	4·2	6·3	9·9	5·1	17·4		11·2	−3·7	−9·8	−0·6	−4·2	−4·8	−3·2	0·7	−0·9	−1·3	M
L	6·5	1·1	1·8	4·2	11·6	2·6	5·6	12·4	6·4	17·1		8·0	−2·4	−0·8	−1·5	1·0	0·1	−5·7	−6·1	−1·3	N
M	8·3	5·4	3·9	7·0	9·2	4·6	7·5	12·4	7·9	14·1	20·9		10·3	−3·6	−3·6	−1·0	−2·0	−5·2	−7·4	−7·0	P
N	8·4	2·2	12·4	9·1	6·0	8·4	11·5	5·1	9·9	5·0	6·1	17·8		9·0	2·1	−1·2	−0·4	−3·6	−8·2	−5·1	Q
P	8·8	0·9	8·8	8·3	4·8	7·3	5·5	4·1	9·2	7·0	0·0	7·4	20·1		10·0	−0·6	−1·4	−4·9	−3·8	−2·1	R
Q	9·2	2·9	8·7	12·2	3·4	7·0	11·2	2·8	10·9	5·4	9·2	9·0	6·2	18·8		5·8	2·0	−4·3	−6·2	−3·4	S
R	8·2	4·2	6·4	9·6	3·8	7·0	9·9	4·1	13·0	6·1	5·6	8·3	6·2	11·9	19·8		6·8	−1·9	−9·3	−2·7	T
S	9·8	2·1	9·6	7·6	5·0	8·5	7·2	5·1	8·3	4·6	5·0	10·8	8·8	8·6	9·2	15·6		7·0	−4·9	−1·8	V
T	9·0	3·8	8·0	9·3	4·8	6·0	6·8	6·6	9·6	5·2	6·6	9·7	7·8	9·4	8·4	11·8	16·6		15·2	2·3	W
V	9·3	5·0	4·6	5·6	8·5	4·2	5·9	13·7	6·1	11·6	10·5	4·1	4·6	6·2	4·9	5·5	7·9	16·8		10·5	Y
W	4·0	0·7	3·8	2·2	13·2	3·5	5·8	6·5	4·4	8·8	8·9	3·7	2·4	1·6	0·0	3·6	0·5	4·9	25·0		
Y	5·8	2·1	6·0	6·1	13·2	4·4	9·4	7·3	6·1	7·4	8·5	8·5	2·8	4·7	7·7	6·4	7·1	8·0	12·1	20·3	
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Figure I.25 : Matrice de Johnson (triangle supérieur) et matrice positive de Johnson (triangle inférieur) obtenue par l'ajout de 9,8 à toutes les valeurs.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
A	0																				A
C	135	0																			C
D	95	110	0																		D
E	30	134	93	0																	E
F	118	105	131	124	0																F
G	140	88	92	145	105	0															G
H	118	104	101	114	83	94	0														H
I	125	115	128	114	73	123	99	0													I
K	45	125	87	50	116	134	123	117	0												K
L	49	127	99	43	108	154	109	75	69	0											L
M	61	125	109	64	102	129	108	85	95	59	0										M
N	129	96	70	124	113	63	96	106	119	123	113	0									N
P	90	87	98	108	128	90	132	153	96	125	131	110	0								P
Q	68	116	89	61	125	128	92	113	84	71	78	114	118	0							Q
R	91	117	108	91	85	118	85	95	86	100	110	110	118	95	0						R
S	138	69	98	141	113	72	105	134	118	161	148	96	65	123	110	0					S
T	159	88	118	150	81	90	90	70	133	131	131	99	121	121	112	67	0				T
V	123	110	122	109	87	124	111	33	113	77	92	115	142	103	103	125	65	0			V
W	111	100	119	112	85	110	101	97	98	112	119	121	103	117	100	95	93	104	0		W
Y	159	75	119	157	76	81	87	87	143	146	131	95	113	128	101	66	54	85	94	0	Y
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	

Figure I.26 : Matrice des coefficients dérivés de la structure (SCM). Un score élevé dans cette matrice traduit une forte similarité (c'est-à-dire une conservation des angles de torsion) entre les résidus concernés (Niefind et Schomburg, 1991).

Ces fréquences sont ensuite transformées en probabilités:

$$P_{ij} = \frac{f_{ij}}{\sum_{j=1}^{20} f_{ij}}$$

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

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

On obtient une matrice (Fig. I.25) 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-aspartates ou méthionine-prolines) et 26 (pour la conservation de la cystéine). La matrice Birkbeck97 est une actualisation de la matrice de Johnson basée sur 97 familles de protéines.

I.3.1.1.3. Matrices à caractère structural.

I.3.1.1.3.1. Les scores de similarité de Niefind et Schomburg (1991).

La conformation de la chaîne principale d'un acide aminé est entièrement définie par la valeur des angles ϕ et ψ. Il existe donc une corrélation entre le type de résidu et la conformation de la charpente protéique à chaque position de la structure. Cependant, cette corrélation étant trop faible, elle ne permet pas la prédiction des valeurs des angles de torsion. à partir de la seule connaissance du type de résidu présent à chaque position. Il est par contre possible, en se basant sur la distribution des angles ϕ et ψ, de calculer des coefficients de similarité entre résidus (Fig. I.26).

Substitution probability table for α residues[a]

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.355	0.007	0.090	0.100	0.050	0.177	0.037	0.077	0.096	0.056	0.081	0.103	0.106	0.090	0.088	0.163	0.120	0.098	0.065	0.036	0.252
C	0.001	0.901	0.000	0.000	0.000	0.000	0.000	0.004	0.001	0.000	0.000	0.003	0.000	0.006	0.006	0.004	0.002	0.000	0.007	0.000	0.000
D	0.038	0.000	0.315	0.109	0.006	0.041	0.027	0.009	0.033	0.004	0.009	0.088	0.051	0.089	0.023	0.065	0.048	0.013	0.012	0.011	0.009
E	0.044	0.011	0.111	0.305	0.011	0.048	0.026	0.011	0.059	0.013	0.009	0.068	0.069	0.086	0.053	0.033	0.045	0.017	0.012	0.018	0.000
F	0.017	0.000	0.005	0.007	0.415	0.004	0.009	0.039	0.025	0.097	0.042	0.013	0.006	0.011	0.009	0.009	0.014	0.041	0.053	0.085	0.009
G	0.065	0.000	0.070	0.042	0.006	0.370	0.017	0.022	0.029	0.013	0.015	0.036	0.043	0.031	0.013	0.068	0.049	0.014	0.009	0.021	0.045
H	0.010	0.000	0.012	0.011	0.010	0.007	0.571	0.003	0.022	0.005	0.015	0.043	0.006	0.035	0.021	0.016	0.008	0.017	0.009	0.037	0.009
I	0.029	0.014	0.009	0.008	0.048	0.021	0.004	0.325	0.017	0.076	0.107	0.018	0.007	0.007	0.015	0.014	0.033	0.112	0.016	0.030	0.018
K	0.053	0.007	0.044	0.081	0.020	0.041	0.044	0.026	0.336	0.029	0.059	0.073	0.045	0.094	0.163	0.041	0.054	0.026	0.041	0.028	0.036
L	0.038	0.000	0.006	0.018	0.210	0.019	0.004	0.139	0.033	0.415	0.225	0.033	0.016	0.041	0.028	0.029	0.026	0.133	0.037	0.057	0.036
M	0.013	0.000	0.004	0.003	0.016	0.007	0.000	0.043	0.014	0.053	0.197	0.010	0.000	0.018	0.004	0.003	0.010	0.018	0.021	0.021	0.018
N	0.031	0.007	0.057	0.035	0.010	0.026	0.054	0.012	0.034	0.012	0.013	0.195	0.015	0.066	0.026	0.037	0.046	0.012	0.002	0.048	0.000
P	0.022	0.000	0.036	0.035	0.005	0.026	0.011	0.009	0.020	0.006	0.000	0.013	0.424	0.013	0.016	0.039	0.011	0.009	0.002	0.000	0.000
Q	0.025	0.011	0.045	0.039	0.011	0.021	0.031	0.004	0.045	0.015	0.035	0.059	0.015	0.183	0.029	0.030	0.030	0.008	0.007	0.025	0.009
R	0.019	0.011	0.012	0.023	0.005	0.008	0.019	0.010	0.067	0.009	0.004	0.018	0.013	0.028	0.348	0.030	0.019	0.005	0.007	0.018	0.018
S	0.086	0.021	0.075	0.047	0.012	0.079	0.033	0.020	0.041	0.020	0.009	0.089	0.082	0.069	0.063	0.264	0.096	0.028	0.005	0.020	0.054
T	0.043	0.007	0.039	0.033	0.020	0.038	0.014	0.026	0.032	0.015	0.026	0.057	0.028	0.046	0.035	0.065	0.266	0.037	0.016	0.034	0.000
V	0.055	0.000	0.018	0.021	0.069	0.022	0.044	0.178	0.025	0.111	0.016	0.018	0.025	0.017	0.015	0.029	0.060	0.350	0.012	0.043	0.162
W	0.009	0.000	0.003	0.004	0.022	0.004	0.007	0.006	0.012	0.006	0.020	0.001	0.001	0.006	0.004	0.002	0.007	0.003	0.588	0.064	0.000
Y	0.009	0.000	0.006	0.006	0.046	0.006	0.029	0.014	0.007	0.013	0.031	0.033	0.003	0.020	0.010	0.007	0.017	0.016	0.078	0.377	0.027
J	0.009	0.000	0.001	0.000	0.001	0.004	0.001	0.002	0.002	0.002	0.004	0.000	0.000	0.004	0.003	0.006	0.004	0.010	0.000	0.005	0.297
–	0.028	0.004	0.041	0.074	0.010	0.029	0.017	0.022	0.050	0.031	0.033	0.031	0.045	0.039	0.028	0.047	0.034	0.032	0.002	0.021	0.000

[a] The standard one-letter amino acid code is used with the exception of C for cystine (the disulfide-bonded form) and J for cysteine (the free thiol form). The values in the table give the probability of a substitution of a residue at the top of a column, by all other residues or at the site of an insertion/deletion; thus, the columns sum to 1.0.

A

Substitution probability table for β residues

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	J
A	0.275	0.000	0.025	0.047	0.023	0.086	0.007	0.029	0.036	0.031	0.074	0.041	0.035	0.050	0.050	0.057	0.055	0.065	0.014	0.031	0.080
C	0.000	0.910	0.000	0.016	0.014	0.000	0.000	0.003	0.008	0.000	0.000	0.000	0.000	0.008	0.015	0.002	0.001	0.000	0.000	0.000	0.020
D	0.008	0.000	0.350	0.059	0.008	0.011	0.014	0.017	0.018	0.006	0.000	0.095	0.040	0.020	0.010	0.026	0.020	0.013	0.006	0.012	0.000
E	0.018	0.016	0.054	0.192	0.004	0.015	0.021	0.012	0.071	0.009	0.037	0.039	0.028	0.056	0.053	0.018	0.036	0.018	0.002	0.015	0.000
F	0.020	0.022	0.013	0.005	0.398	0.008	0.021	0.049	0.017	0.046	0.023	0.006	0.012	0.006	0.015	0.020	0.012	0.021	0.071	0.096	0.020
G	0.092	0.000	0.021	0.033	0.015	0.623	0.007	0.016	0.018	0.016	0.042	0.019	0.017	0.033	0.017	0.036	0.028	0.013	0.049	0.021	0.020
H	0.003	0.000	0.006	0.010	0.008	0.002	0.332	0.006	0.022	0.004	0.000	0.035	0.014	0.021	0.023	0.009	0.010	0.009	0.000	0.008	0.020
I	0.040	0.010	0.044	0.021	0.089	0.020	0.017	0.358	0.022	0.105	0.077	0.025	0.012	0.010	0.015	0.021	0.026	0.119	0.041	0.034	0.020
K	0.024	0.011	0.021	0.099	0.015	0.007	0.070	0.013	0.299	0.017	0.012	0.060	0.052	0.060	0.139	0.026	0.051	0.010	0.004	0.017	0.040
L	0.057	0.000	0.027	0.031	0.111	0.023	0.031	0.143	0.051	0.459	0.169	0.031	0.038	0.026	0.031	0.025	0.028	0.119	0.096	0.044	0.060
M	0.026	0.000	0.006	0.021	0.011	0.013	0.021	0.021	0.007	0.031	0.244	0.010	0.002	0.031	0.002	0.007	0.013	0.019	0.006	0.006	0.000
N	0.016	0.000	0.092	0.044	0.003	0.018	0.073	0.008	0.038	0.008	0.019	0.261	0.026	0.016	0.011	0.036	0.032	0.004	0.012	0.017	0.000
P	0.014	0.000	0.027	0.020	0.001	0.005	0.021	0.007	0.029	0.019	0.002	0.017	0.504	0.011	0.023	0.010	0.021	0.009	0.018	0.003	0.000
Q	0.038	0.014	0.033	0.098	0.006	0.020	0.073	0.008	0.071	0.014	0.077	0.037	0.019	0.414	0.118	0.025	0.045	0.015	0.002	0.013	0.000
R	0.022	0.011	0.006	0.042	0.007	0.010	0.038	0.008	0.079	0.008	0.002	0.012	0.031	0.055	0.214	0.015	0.027	0.010	0.014	0.017	0.000
S	0.078	0.003	0.085	0.041	0.029	0.056	0.052	0.024	0.048	0.022	0.030	0.112	0.040	0.042	0.065	0.403	0.140	0.028	0.014	0.040	0.040
T	0.081	0.002	0.075	0.111	0.021	0.037	0.052	0.027	0.095	0.022	0.049	0.110	0.073	0.078	0.092	0.153	0.363	0.044	0.008	0.037	0.020
V	0.141	0.000	0.058	0.065	0.074	0.027	0.070	0.202	0.034	0.145	0.123	0.019	0.033	0.039	0.046	0.043	0.062	0.446	0.027	0.059	0.040
W	0.005	0.000	0.008	0.002	0.048	0.000	0.000	0.013	0.001	0.019	0.005	0.015	0.012	0.001	0.013	0.003	0.002	0.005	0.559	0.017	0.000
Y	0.026	0.000	0.027	0.037	0.112	0.011	0.049	0.024	0.020	0.018	0.014	0.033	0.005	0.013	0.032	0.026	0.022	0.023	0.051	0.505	0.000
J	0.003	0.002	0.000	0.000	0.001	0.001	0.007	0.001	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.001	0.000	0.000	0.620
–	0.012	0.000	0.021	0.007	0.002	0.006	0.021	0.012	0.013	0.004	0.002	0.021	0.007	0.008	0.015	0.038	0.007	0.009	0.004	0.009	0.000

B

Figure I.27 : Tables de probabilités de substitutions d'Overington *et al.* (1992) pour les résidus appartenant à une hélice α (a) et à une structure β (b).

Ces coefficients étant calculés à partir des données relatives aux conformations, ils sont appelés « structure-derived correlation coefficients » (SCC) et font partie de la « structure-derived matrix » (SCM).

I.3.1.1.3.2. Tables de substitutions d'Overington et al. (1992).

La recherche menée par l'équipe d'Overington avait pour but l'étude des substitutions intervenant dans des environnements structuraux semblables. Cette orientation repose sur le fait que l'environnement d'un résidu détermine l'acceptabilité d'une mutation à cette position. Les différents environnements sont définis selon les critères suivant : type de résidu, type de structure secondaire, accessibilité de la chaîne latérale, interactions entre chaînes latérales. Les fréquences de substitutions observées sont converties en probabilités et consignées dans des tables de substitutions propres au type d'environnement dans lequel elles s'opèrent. Dans ces tables, chaque valeur indique la probabilité que le résidus i, situé dans un environnement particulier, soit remplacé par le résidu j (Fig. I.27).

I.3.1.2 Les matrices de distance.

I.3.1.2.1. Matrices de code génétique.

Alors que la matrice identité considérait toute transition d'acide aminé sur un même pied, la matrice de code génétique introduite par Fitch (1966) considère le nombre de changements de base nécessaires à cette transition. Un remplacement ne nécessitant qu'un changement de base sera plus fréquent que ceux pour lesquels il faut deux ou trois changements. Feng *et al.* (1985) ont reprit cette idée en combinant des informations sur les caractéristiques structurelles des acides aminés et la redondance du code génétique.

Amino acid	Factor* 1	2	3	4	5	6	7	8	9	10
ALA	−1.56	−1.67	−0.97	−0.27	−0.93	−0.78	−0.20	−0.08	0.21	−0.48
ARG	0.22	1.27	1.37	1.87	−1.70	0.46	0.92	−0.39	0.23	0.93
ASN	1.14	−0.07	−0.12	0.81	0.18	0.37	−0.09	1.23	1.10	−1.73
ASP	0.58	−0.22	−1.58	0.81	−0.92	0.15	−1.52	0.47	0.76	0.70
CYS	0.12	−0.89	0.45	−1.05	−0.71	2.41	1.52	−0.69	1.13	1.10
GLN	−0.47	0.24	0.07	1.10	1.10	0.59	0.84	−0.71	−0.03	−2.33
GLU	−1.45	0.19	−1.61	1.17	−1.31	0.40	0.04	0.38	−0.35	−0.12
GLY	1.46	−1.96	−0.23	−0.16	0.10	−0.11	1.32	2.36	−1.66	0.46
HIS	−0.41	0.52	−0.28	0.28	1.61	1.01	−1.85	0.47	1.13	1.63
ILE	−0.73	−0.16	1.79	−0.77	−0.54	0.03	−0.83	0.51	0.66	−1.78
LEU	−1.04	0.00	−0.24	−1.10	−0.55	−2.05	0.96	−0.76	0.45	0.93
LYS	−0.34	0.82	−0.23	1.70	1.54	−1.62	1.15	−0.08	−0.48	0.60
MET	−1.40	0.18	−0.42	−0.73	2.00	1.52	0.26	0.11	−1.27	0.27
PHE	−0.21	0.98	−0.36	−1.43	0.22	−0.81	0.67	1.10	1.71	−0.44
PRO	2.06	−0.33	−1.15	−0.75	0.88	−0.45	0.30	−2.30	0.74	−0.28
SER	0.81	−1.08	0.16	0.42	−0.21	−0.43	−1.89	−1.15	−0.97	−0.23
THR	0.26	−0.70	1.21	0.63	−0.10	0.21	0.24	−1.15	−0.56	0.19
TRP	0.30	2.10	−0.72	−1.57	−1.16	0.57	−0.48	−0.40	−2.30	−0.60
TYR	1.38	1.48	0.80	−0.56	−0.00	−0.68	−0.31	1.03	−0.05	0.53
VAL	−0.74	−0.71	2.04	−0.40	0.50	−0.81	−1.07	0.06	−0.46	0.65

Figure I.28 : Table des 10 facteurs de Kidera *et al.* (1985).

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	
0	114	64	41	81	119	124	71	111	49	104	92	120	93	136	69	69	76	139	124	A
	0	121	121	111	143	124	121	179	116	105	123	117	132	100	145	71	127	164	128	C
		0	44	90	126	58	111	113	104	128	61	102	125	99	57	82	106	125	75	D
			0	97	133	107	106	90	88	91	83	152	82	81	93	80	118	99	111	E
				0	138	99	67	98	54	108	60	105	99	129	140	108	100	125	49	F
					0	180	156	130	148	128	102	169	159	168	127	103	123	172	102	G
						0	122	110	134	73	102	137	134	135	100	98	81	169	83	H
							0	147	104	116	57	150	73	114	82	65	48	124	76	I
								0	75	95	104	118	75	93	112	70	93	163	81	K
									0	121	136	98	129	112	107	74	73	130	84	L
										0	123	140	77	159	116	81	88	104	112	M
											0	102	49	105	88	76	109	144	64	N
												0	101	161	77	77	141	143	105	P
													0	112	97	58	113	134	118	Q
														0	118	51	108	139	77	R
															0	34	52	107	85	S
																0	37	111	69	T
																	0	140	64	V
																		0	80	W
																			0	Y

Figure I.29 : Matrice de Depiereux et Feytmans (1991).

I.3.1.2.2. Matrices physico-chimique : la matrice DFK (Depiereux et Feytmans, 1991).

Cette matrice est basée sur une étude réalisée par Kidera *et al.* (1985) qui réalisa une analyse factorielle de 188 propriétés chimiques et physiques des 20 acides aminés. Il en ressortit que toute l'information initialement contenue dans les 188 propriétés pouvait être condensée dans 10 facteurs indépendants (Fig. I.28).

Ces 10 facteurs physico-chimiques vont être utilisés par Depiereux et Feytmans pour définir une mesure de distance entre deux acides aminés i et j :

$$d_{ijk} = \left| z_{ik} - z_{jk} \right|$$

où z_{ik} est la valeur du facteur k pour le résidu i et z_{jk}, la valeur de ce même facteur k pour le résidu j.

Si on considère les séquences générées au hasard à partir des 20 acides aminés, la distribution de probabilité de $d_{ijk}/\sqrt{2}$ est approximativement normale (moyenne = 0, variance = 1).

Cette distance mise au carré suit approximativement une distribution de chi carré avec un degré de liberté.

Pour un ensemble de p facteurs, on peut alors calculer le carré de la distance entre deux acides aminés :

$$D^2 = \sum_{i=1}^{p} \sum_{j=1}^{w} \frac{d^2_{ij}}{2}$$

Cette valeur de distance est calculée pour chacune des 190 paires de résidus possibles et permet la construction d'une matrice (Fig. I.29). Les 20 couples identités

constituant la diagonale principale de la matrice se voient attribuer la valeur minimale de distance, à savoir 0.

I.3.1.3. Utilisation des matrices de scores.

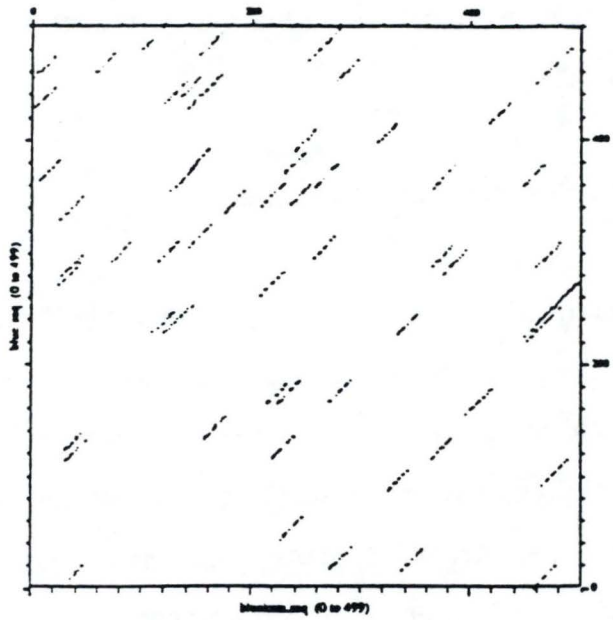
Une fois la matrice élaborée, l'étape suivante est la construction d'un alignement de séquences cohérent. Cela revient à établir une correspondance entre les résidus les plus similaires dans le but de pouvoir délimiter les régions structurellement conservées. Tout alignement est caractérisé par un score égal à la somme de tous les scores associés aux paires de résidus qui se font face dans l'alignement. Un alignement est optimal si son score est maximal. Il peut, dans certains cas, être nécessaire d'introduire des espacements dans l'une ou l'autre séquence pour permettre aux zones les plus similaires de se faire face.

N.B. 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.

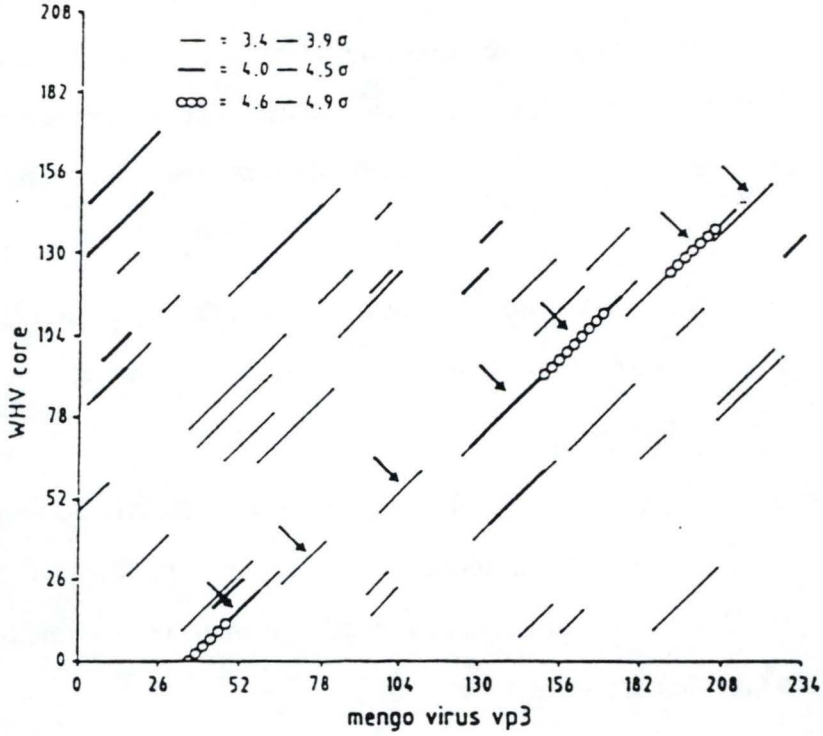
Pour détecter la meilleure correspondance possible, deux approches ont généralement été prises dépendant du nombre d'acides aminés pris en compte par comparaison.

La première considère, dans chacune des séquences, un acide aminé à la fois. Prenons le cas de l'alignement de deux séquences, la distance (ou la similarité) est donc calculée entre des paires de résidus.

La seconde, reposant 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 résidus se faisant face dans les deux fenêtres.



A



B

Figure I.30 : a. Matrice de comparaison dans laquelle un point apparaît si les résidus qu'il représente sont considérés comme similaires (ils sont caractérisés par une distance inférieure au seuil).

b. Matrice de comparaison où chaque classe de similarité est représentée par un trait différent.

I.3.2. Les alignements pairés.

I.3.2.1. Les matrices de comparaison ou « dot plot ».

Dans cette méthode, sans doute la plus simple, toutes les fenêtres possibles de A (séquence de longueur m) sont comparées avec toutes celles de B (séquence de longueur n). Les résultats obtenus pour ces comparaisons sont placés dans une matrice R ou chaque élément r_{ij} représente la similarité entre la $i^{ème}$ fenêtre de A et la $j^{ème}$ de B. Le nombre de comparaisons à réaliser est de l'ordre de $m \times n$.

Ces similarités sont calculées à partir des matrices de scores présentées plus haut. Par exemple, si on considère la comparaison de deux fenêtres de 7 résidus; ALGAWDE et ALATWDE en utilisant la matrice identité. Le score total sera de : 1+1+0+0+1+1+1=5.

Cette matrice de comparaison permet de visualiser les régions d'identité (ou de similarité) entre les séquences. Ces régions se présentent sous la forme de diagonales où les scores sont élevés.

La technique du « dot-plot » permet une visualisation plus aisée et une quantification des ressemblances entre les séquences. Elle permet de fixer un seuil de similarité ou de déterminer des classes de similarité. Dans le premier cas, les régions considérées comme similaires sont pointées; dans le second, chaque classe est représentée par un symbole différent (Fig. I.30).

Cette méthode bien qu'étant très simple à mettre en oeuvre, ne permet que de visualiser les régions semblables, elle ne permet pas la réalisation d'un alignement complet.

	A	C	F	G	S	T	V	I	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

Figure I.31 : Matrice de comparaison construite pour les séquences A et B.

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

Figure I.32 : 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	1

Figure I.33 : Matrice Z complète.

	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	1

Figure I.34 : 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

Figure I.35 : 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.

I.3.2.2. Méthode de programmation dynamique de Needleman et Wunsch (1970).

Needleman et Wunsch ont, au départ d'une matrice de comparaison, mis au point une technique analytique permettant la construction d'un alignement pairé. Cette méthode détecte les identités, les similarités et les distances, selon la matrice de scores utilisée. L'exemple présenté ci-dessous ne tient compte que des identités.

Soit deux séquences protéiques A et B ayant respectivement n_A et n_B pour longueur et $Y(n_A \times n_B)$, leur matrice de comparaison (Fig. I.31) remplie de 0 (si $A_i \neq B_j$) et de 1 (si $A_i = B_j$). Chaque élément y_{ij} de la matrice est modifié pour devenir un élément z_{ij} de la matrice Z d'identités cumulées. Pour ce faire, il faut se placer à la fin de deux séquences ($i=n_{A-1}$ et $j=n_{B-1}$) et remonter vers le début des séquences ($i=1$ et $j=1$) en se déplaçant toujours de droite à gauche. A chaque pas, y_{ij} est modifié : la valeur la plus élevée est recherchée au niveau de la ligne (i+1) qui se trouve à droite de y_{ij} et de la colonne (j+1) en dessous de y_{ij}. Cette valeur est additionnée à y_{ij} pour donner un score modifié z_{ij} (Fig. I.32).

Une fois la matrice Z complète (Fig. I.33.), chacun de ses 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.

L'ultime étape permet de trouver l'alignement optimal. Cela revient à trouver, dans la matrice Z, le tracé pour lequel la somme des scores z_{ij} des cases traversées est maximale. Il part de l'élément z_{ij} le plus grand et se caractérise par des indices i et j toujours supérieurs à ceux de la case précédente (Fig. I.34). Si la progression se fait toujours de z_{ij} vers $z_{i+1,j+1}$, on forme une diagonale complète. Dans tous les autres cas, le tracé « saute » d'une diagonale à l'autre. Cela se traduit, dans l'alignement, par l'introduction d'un espacement dans une des deux séquences. Ces « gaps » permettent de disposer face à face les zones les plus similaires (Fig. I.35).

L'alignement réalisé, il est possible d'en calculer le score global de l'alignement, somme des scores obtenus pour la comparaison des résidus placés face à face dans l'alignement. Pour éviter d'avoir des gaps trop nombreux ou trop long, ceux-ci peuvent être pénalisés par un facteur -g. Malgré cela le problème du gap reste entier, en effet on introduit avec ces gaps un paramètre arbitraire qui n'a aucun fondement biologique ou physico-chimique.

I.3.2.3. Méthode de Wilbur et Lipman (1983).

La méthode de Needleman et Wunsch est très rigoureuse et très fiable, mais dans le cas d'une recherche dans une banque de séquences, par exemple, elle devient très longue à mettre en oeuvre. Wilbur et Lipman ont, dans cette optique, mis au point un algorithme plus rapide.

Cette nouvelle technique recherche, dans les deux séquences, les zones identiques de w résidus de long. Cependant, cette recherche ne se fait plus sur toute la longueur des séquences mais se limite à une région de m résidus. Dans un premier temps, l'algorithme va parcourir l'entièreté des deux séquences pour y lire toutes les zones de w résidus. Ensuite, il va comparer chaque segment de la première séquence avec tous les segments de la deuxième séquence compris dans la région de taille m. Une fois les régions identiques détectées, la dernière étape est, comme pour la méthode de Needleman et Wunsch, la construction de l'alignement.

La vitesse d'exécution du programme dépendra de deux paramètres : w, la taille des zones comparées et m, la taille des régions de comparaisons. Si w=1 et m est très grand, cet algorithme est celui de Needleman et Wunsch.

I.3.2.4. Méthode de Lipman et Pearson (1985, 1988).

L'algorithme de Lipman et Pearson, aussi appelé FASTP, a pour principal objectif la comparaison d'une séquence cible à toute une banque de séquences. Sur base d'un alignement des zones identiques entre les deux séquences, cette méthode va réaliser un calcul de la similarité. Dans ce calcul, seules les régions similaires entrent en compte.

La première étape, la recherche des identités sur une zone de m résidus, est réalisée par l'algorithme de Wilbur et Lipman. Les positions des régions possédant les scores les plus élevés sont mémorisées. Ce sont donc les régions où le pourcentage d'identité est le plus élevé. La matrice de substitution 250 PAM de Dayhoff est ensuite utilisée pour calculer , dans ces cinq régions, un score de similarité. Seul le score le plus élevé est gardé pour caractérisé la similarité entre les deux séquences. Chaque séquence de la banque est donc associée à un score de similarité par rapport à la séquence cible. L'ensemble des score est représenté sous forme d'histogramme.

Le programme FASTA (Pearson et Lipman, 1988) apporte une amélioration au précédent : alors que FASTP ne considérait qu'une région initiale, FASTA recherche si d'autres régions initiales peuvent être prises en considération. Selon la localisation des régions initiales, leur score respectif et une pénalité jointure, FASTA calcule un alignement optimal égal à la combinaison des régions de score maximum compatibles entre elles.

I.3.2.5. Méthode de Altschul et al. (1990).

Cette méthode, appelée BLAST est très similaire à celle de Lipman et Pearson. Elle recherche la « Maximum Segment Pair » ou MSP, la paire de segments de séquence qui possède le score de similarité le plus élevé (il est ici question de segments continus et donc sans gap). Les limites d'une MSP sont choisies afin de

maximiser le score de similarité. Des scores positifs étant attribués aux identités ou aux remplacements conservatifs et des scores négatifs pénalisant les substitutions.

BLAST peut aussi être utilisé pour la recherche de paire de segments de longueur m dont le score est supérieur à un seuil T. Ce seuil T doit être choisi pour être suffisamment discriminatif (plus T est élevé, plus les segments sélectionnés sont semblables) et pour minimiser le temps de travail (plus T est bas, plus une paire de segments est facile à trouver). Le juste compromis peut être trouvé par simulation aléatoire.

I.3.3. Alignements multiples.

Le nombre de séquences actuellement disponibles est énorme. De plus il est en augmentation permanente. Partant du fait qu'une similarité de séquences a plus de signification si elle est partagée par plusieurs séquences, de nombreux auteurs ont développé des méthodes d'alignement multiple de séquences.

I. 3. 3. 1. Alignements multiples progressifs.

Le point commun de ces méthodes est de réaliser d'abord un alignement pairé auquel seront ajoutées ensuite d'autres séquences. Le principal inconvénient de la plupart de ces méthodes est que l'alignement obtenu est fortement dépendant des deux premières séquences comparées.

I.3.3.1.1. Méthode de Feng et Doolittle (1987).

La première étape de cet alignement consiste à mesurer les similarités entre toutes les paires possibles de séquences, et ce, grâce à l'algorithme de Needleman et Wunsch.

Ces scores de similarité sont ensuite transformés en scores de distance par la relation suivante :

$$D = -\ln\left(\frac{(S_v - S_r)}{(S_i - S_r)}\right) \times 100$$

avec :

S_v, le score de l'alignement observé;

S_r, le score obtenu par des séquences aléatoires de même composition;

S_i, le score moyen de tous les alignements des séquences entre elles.

Plus la distance D est petite, plus la similarité entre les deux séquences est grande. Cette mesure permet un classement des séquences par distances mutuelles croissantes c'est-à-dire en fonction de leurs liens phylogénétiques.

Les deux séquences les plus proches sont alignées en premier lieu, de façon à obtenir un alignement de score minimal. Ensuite, les deux séquences sont alignées avec celle qui leur est la plus proche. Et ainsi de suite, les séquences sont incorporées à l'alignement dans l'ordre de leur classement par distance. Des espaces peuvent être introduits dans l'alignement mais n'influence pas le score global. De plus, une fois ces espaces introduits, ils ne peuvent être modifiés. La troisième séquence doit s'aligner aux deux premières en fonction des espaces préexistants.

I.3.3.1.2. Méthode de Corpet (1988).

Cette méthode est divisée en deux étapes : la classification des séquences et la comparaison des séquences ou groupes de séquences.

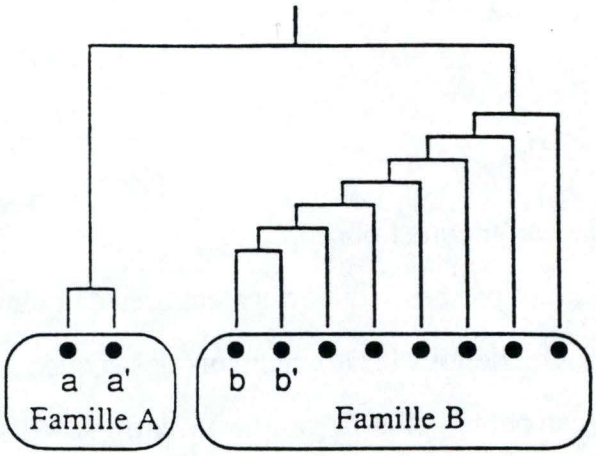


Figure I.36 : **Dendrogramme obtenu par la méthode de Corpet (1988).**

I.3.3.1.2.1. Classification des séquences.

Au départ des N séquences à aligner, on construit un tableau T_1 de dimension N×N. Chaque élément t_{1ij} est une mesure de la ressemblance entre les séquences i et j, cette mesure étant en fait le score global de l'alignement réalisé par la méthode de Lipman et Pearson. Dans un deuxième temps, les 2 séquences les plus proches sont alignées et forment un groupe qui prend la place de la séquence i. La séquence j est supprimée. Il faut donc recalculer un nouveau tableau T_2 de dimension N-1×N-1 tenant compte du nouveau groupe. Les nouvelles valeurs sont les suivantes :

t_{2jk} et t_{2kj} sont supprimées, pour toute séquence k;

t_{2kl} sont inchangées si k et l sont différents de i et j;

t_{2ik} et t_{2ki} sont identiques et égales à la moyenne des scores t_{1ik} et t_{1jk}.

Cette opération est recommencée jusqu'à ce qu'il n'y ait plus qu'un groupe (Fig. I.36).

I.3.3.1.2.2. Comparaison des séquences.

A chaque comparaison de 2 séquences ou de 2 groupes de séquences, une matrice de comparaison est construite. Selon la méthode de Needleman et Wunsch pour 2 séquences et selon une méthode développée par Corpet pour deux groupes de séquences.

Soit A_r, un groupe A de P séquences et B_s , un groupe B de Q séquences.

Soit i, la position d'un résidu aligné dans le groupe A et j, la position d'un résidu aligné dans le groupe B.

L'élément c_{ij} de la matrice de Corpet est calculé de la façon suivante :

$$c_{ij} = \frac{1}{P.Q} \sum_{r=1}^{P} \sum_{s=1}^{Q} \mathrm{Score}(A_r(i) \,.\, \mathbf{B}_s(j))$$

Une fois cette matrice construite, l'alignement de score maximal est trouvé par la méthode de Needleman et Wunsch.

I.3.3.1.3 Méthode de Subbiah et Harrison (1989).

Subbiah et Harrison ont développé un algorithme basé sur celui de Needleman et Wunsch et conçu pour aligner au maximum 10 séquences. L'alignement multiple ainsi généré résulte d'une succession de comparaisons pairées.

Prenons l'exemple de 3 séquences A, B et C à aligner. En premier lieu, toutes les comparaisons pairées possibles sont réalisées par la programmation dynamique de Needleman et Wunsch. Chaque alignement est caractérisé par sa signification statistique (un alignement est valable si son score est significativement supérieur à celui obtenu par hasard) et toutes les séquences sont classées par ordre décroissant de leur similarité respective. Une fois les séquences A et B alignées, elles sont notées A' et B' pour indiquer l'existence d'espacements nécessaires à l'alignement. Il en est de même pour toutes les autres séquences. On obtient donc trois alignements pairés : (A',B'),(B',C') et (A',C').

Dans un deuxième temps, l'alignement le plus significatif est gelé, c'est-à-dire que les espaces qu'il contient sont temporairement immuables. Cet alignement, par exemple (A',B'), est ensuite aligné à la troisième séquence, dans notre cas C. Les espaces à créer en vue de l'alignement optimal doivent l'être uniquement dans la séquence C. Le résultat est un triplet noté (A'B',C').

Lors de la troisième étape, les séquences B' et C' sont à leur tour gelées puis comparées à A. Il en résulte l'alignement (A'',B'C') plus précis que (A'B',C'). De la même façon, l'algorithme réaligne (A'',C') avec B, ce qui donne l'alignement final (A''C'',B'').

Plus le nombre de séquences augmente, plus le nombre de comparaisons augmente. Par exemple, l'alignement de six séquences nécessite des comparaisons

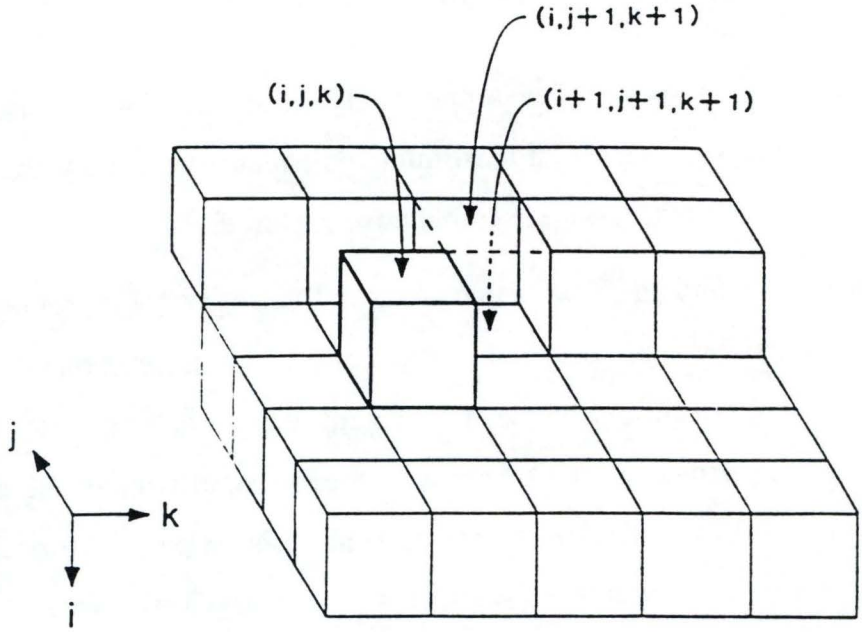


Figure I.37 : Tableau à trois dimensions construit par l'alignement de trois séquences par la méthode de Murata (1985). A chaque case représentant un triplet d'acides aminés ijk est associé un score. Il s'agit de retrouver, à travers ce cube, un tracé pour lequel le score cumulé de chacune des cases soit minimum.

entre tous les ensembles de 5-tuplets gelés/séquence originale, tous les ensembles de 4-tuplets/paires gelées,...

En résumé, les espaces nécessaires à l'alignement optimal sont créés au fur et à mesure de l'incorporation de nouvelles séquences. D'autre part, la pénalité associée aux espaces peut être modulée par l'utilisateur.

Peu de temps auparavant, un algorithme semblable avait été proposé par Barton et Sternberg (1987). Il différait pourtant en deux points :

- toutes les séquences sont d'abord alignées avant de recommencer des comparaisons entre groupes de séquences;
- la pénalité associée aux espaces est constante.

1. 3. 3. 2. Alignement multiple simultané, la méthode de Murata (1985).

L'algorithme développé par Murata est une simple extension de celui de Needleman et Wunsch et est destiné à la comparaison de trois séquences. En effet, il est basé sur la construction d'un tableau à trois dimensions où chaque case Y_{ijk} est la somme des scores associés aux trois paires de résidus superposés. Comme pour Needleman et Wunsch, l'algorithme va rechercher le tracé optimal lui permettant d'obtenir un score global maximal. Lorsque deux cases successives ne sont pas adjacentes, il faut introduire un gap, comme dans le cas de l'alignement de deux séquences (Fig. I.37).

L'extension de la méthode de programmation dynamique est limitée par la complexité exponentielle du processus et, de ce fait, la durée du temps calcul.

I.4. Comparaison de structures.

Les méthodes d'alignement de structures ont, dans le cadre de la biologie moléculaire actuelle, un rôle tout aussi important que les alignements de séquences. En effet, même si le nombre de protéines résolues est encore limité, il est nécessaire d'établir entre elles des classifications. Ces classification seront utiles dans la modélisation par comparaison de séquences. Grâce aux alignements de structures, il est aussi possible de calculer de nouvelles matrices de scores, en calculant, à des points topologiquement équivalents, les fréquences de substitution.

Actuellement, la gamme de méthodes est assez vaste. Dans la suite de ce chapitre, quelques-unes d'entre elles seront présentées. Ces techniques ont été choisies selon leur efficacité et leur originalité, afin de donner une idée générale de ce qui se fait.

I.4.1. Méthodes basées sur les acides aminés et leurs propriétés

I.4.1.1. Analyse par regroupement de résidus hydrophobes.

Cette méthode, développée par Henrissart *et al.* (1989), est à la limite entre les alignements de séquences et de structures puisqu'elle n'utilise pas les coordonnées tridimensionnelles des acides aminés mais est uniquement basée sur leur nature. Elle réalise des alignements en se basant sur la taille, la forme et l'orientation des clusters hydrophobes. La comparaison de séquences par analyse des regroupements de résidus hydrophobes (Hydrophobic cluster analysis, HCA) permet de détecter des similarités

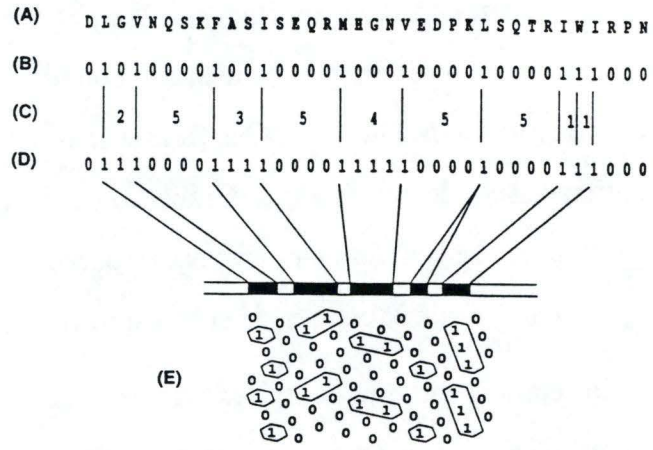


Figure I.38 : Construction de la représentation hélicoïdale des clusters hydrophobes :

a. séquence d'acides aminés

b. translation directe en un code à deux états (1 = hydrophobe, 0 = hydrophile).

c. évaluation des distances entre acides aminés hydrophobes

d. construction des clusters pour une distance de connectivité de 4

e. représentation hélicoïdale.

structurales entre des protéines de faible pourcentage d'identité. Elle repose sur deux observations :

- il existe une corrélation positive (clairement au dessus du bruit de fond) entre les groupes d'éléments hydrophobes et les éléments de structure secondaire;
- les groupes de résidus hydrophobes sont centrés sur les éléments réguliers de structure secondaire.

La HCA fait appel à une représentation particulière préalable des acides aminés d'une séquence. La séquence en acides aminés va d'abord être reportée sur une hélice α classique (3,6 résidus par tour). Les étapes parcourues sont brièvement présentées ci-dessous.

En premier lieu, les clusters sont représentés en une seule dimension, la séquence d'acides aminés étant convertie en une séquence binaire. Les résidus hydrophobes sont indiqués par un « 1 » alors que les hydrophiles sont remplacé par un « 0 ». Certains résidus comme la proline, la glycine, la cystéine, la sérine et la thréonine, sont symbolisés par des signes typiques. Les clusters sont ensuite construits en « connectant » les résidus hydrophobes séparés par une certaine distance. Cette distance, conformément à la représentation hélicoïdale, a été fixée à 4. En effet, sur une hélice α, deux résidus appartenant au même cluster sont séparés par 3 résidus. Un changement de cette « distance de connectivité » revient à choisir une autre représentation de la séquence (Fig. I.38).

La comparaison de séquence s'effectue via la comparaison des groupes de résidus hydrophobes et des symboles utilisés.

Cette méthode d'étude des clusters hydrophobes a aussi été appliquée à la détermination de la structure secondaire d'une séquence protéique (Woodcock *et al.*, 1992). Les groupes de résidus hydrophobes correspondent à des éléments réguliers de structure secondaire constituant le noyau protéique. Quand les segments hydrophiles, situés entre les éléments de structure secondaire, ne sont pas conservés ou que leur

taille diffère, ils sont considérés comme des boucles qui n'interviennent pas de façon déterminante dans le repliement protéique.

Cette méthode est rapide et permet de réaliser des alignements de 30 à 40 séquences. Mais, malheureusement, elle part d'une simplification énorme de la séquence. Toute protéine n'est plus représentée que par une suite de résidus soit hydrophobes, soit hydrophiles et perd donc de l'information utile.

I.4.1.2. Alignements utilisant la structure secondaire.

De nombreuses méthodes réalisent des comparaisons de protéines en se basant sur les structures secondaires, il ne faut pas oublier néanmoins que la détermination de ces structures secondaires est elle-même peu fiable.

Pour Abagayan et Maiorov (1988), chaque hélice α ou feuillet β est symbolisé par un vecteur tracé le long de son axe. D'autres vecteurs sont utilisés pour représenter les connections entre structures secondaires. La topologie de la structure peut donc être décrite par la longueur des vecteurs et les angles dihédraux existant entre ces derniers. Un algorithme de Mc Lachlan est utilisé pour superposer deux représentations vectorielles.

Sheridan *et al.* (1985) décrivent une méthode de comparaison de structure basée sur l'analyse de l'arrangement des structures secondaires (soit connues, soit prédites).

Rawlings *et al.* (1985) utilisent un langage logique de programmation pour réaliser une description des éléments de structure secondaire, de leurs caractéristiques et de leurs positions relatives. Cette base de données peut être parcourue pour retrouver les motifs topologiques d'intérêt. De la même façon, Schulze-Kremer et King (1992) ont développé une base de données, IPSA, contenant des représentations symboliques d'informations (géométriques, topologiques, chimiques et physiques) sur les structures secondaires de chaque protéine.

Fischel-Ghodsian *et al.* (1990) ont développé une méthode employant l'algorithme de programmation dynamique. Mais, contrairement à Needleman et Wunsch, à chaque acide aminé est associé un symbole indiquant l'élément de structure secondaire auquel il appartient. Une séquence codée en deux lettres (Xx) est donc générée à partir de la séquence en acides aminés. X représente un des 20 acides aminés et x la structure secondaire (hélice α, feuillet β, turn, loop).

Si on considère deux protéines $A = A_1a_1, A_2a_2,...A_ma_m$ et $B = B_1b_1, B_2b_2,...B_nb_n$., la mesure de similarité entre les résidus est une combinaison de deux matrices de similarité, une pour les acides aminés, l'autre entre structures :

$$s(A_{iai}, B_{jbj}) = k1S(A_i, B_j) + k2S(a_i, b_j)$$

où k_1 et k_2 sont des constantes.

Une matrice de similarité H(m×n) est ensuite générée en utilisant l'approche classique de la programmation dynamique.

$$H_{ij} = \max\{H_{i-1,j-1} + S(A_ia_i, B_jb_j), \max_k[H_{i,j-k} - W_k], \max_k[H_{i-k,j} - W_k], 0\}$$

où W_k est une pénalité pour l'insertion et la délétion de k résidus calculée selon la fonction linéaire :

$$W_k = w_1 + kw_2$$

Les régions de similarité maximale sont finalement trouvées en recherchant la plus grande valeur de similarité et en parcourant la matrice jusqu'à la valeur zéro.

```
LSADQISTVQASFDKVKGDPV--GILYAVFKADPSIMAKFTQFAGKDLESIKGTAP-FETHANRIVGFFSKIIGELPNIE
aaaa     aaaaa           aaaaaaaaaaaaaaaaaa           a aa           aaaaaa
      bbbb             bbbbbbbb                                  bbbbbb
                tttt   t                     tttt ..          tttt

LTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEVTG
*   *   *  *                          **      *         *    **             *


ADVNTFVASHKPRGVTHDQLNNFRAGFVSYMKAHTDFAGAEAAWGATLDTFFGMIFSKM
    aaaa                      aaaaaaaaaaaaaaaaaa
 bbb                     bbb                       bbbbb
         ..                                                ..
VVVSDATLKNLGSVHVSKGVADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDEL
  *                        *      *      * *   *
```

Figure I.39 : Alignement de deux globines grâce à l'utilisation des structures primaire et secondaire (a = hélice α, b = plan β, t = turn).

Some of the features that can be used in the comparison of protein structures

Residues	Segments
Properties	
Identity	Secondary structure type
Residue type properties	Amphipathicity
Local conformation	Improper dihedral angle
Distance from gravity centre	Distance from gravity centre
Side-chain orientation	Orientation relative to gravity centre
Main-chain orientation	
Solvent accessibility	Solvent accessibility
Position in space	Position in space
	Orientation in space
Relations	
Hydrogen bond	
Distances to 1 or more nearest neighbours	Distances to 1 or more nearest neighbours
Disulphide bond	Relative orientation of 2 or more segments
Ionic bond	
Hydrophobic cluster	

Figure I.40 : Liste de quelques caractéristiques pouvant être utilisées pour la comparaison de structures protéiques.

Cette technique permet d'améliorer le pourcentage d'alignement correct obtenu avec les techniques de programmation dynamique classique (pour lesquelles seule la séquence est prise en compte). Elle possède cependant les mêmes caractéristiques que ces dernières; bien que rigoureuse et précise, elle est difficilement applicable à plus de deux protéines (Fig. I.39).

Les auteurs utilisent aussi leur méthode pour aligner des protéines de structure inconnue. Dans ce cas, les structures secondaires associées aux acides aminés sont déterminées par la méthode de Chou et Fasman. De ce fait, la matrice de comparaison de structure secondaire compte quatre entrées supplémentaires (combinaisons possibles entre hélice, plan et turn) et permet ainsi le chevauchement de structures dans les cas où la prédiction n'est pas précise.

I.4.1.3. Méthodes basées sur l'utilisation de propriétés structurelles.

Sali et Blundell (1990) définissent une protéine comme une hiérarchie de structures : séquence, structure secondaire, superstructure secondaire, motif, domaine et protéine entière. A chaque niveau, correspond une série d'éléments indexés (acides aminés, éléments de structure secondaire,...), chacun étant associé à une série de propriétés et engagé dans un certain nombre de relations avec d'autres éléments du même niveau. Quelques unes de ces propriétés et relations sont présentées dans la Figure I.40.

On peut donc calculer une différence normalisée, w_{fij}, pour une caractéristique f, entre deux résidus i et j appartenant respectivement à la première et à la seconde protéine.

A partir de ces différences, il est possible de calculer une somme pondérée W_{ij} :

$$W_{ij} = \sum_{l} \left(\sum_{p} \rho^{p\,n} w_{ij}^{p} + \sum_{r} \rho^{r\,n} w_{ij}^{r} \right).$$

avec :

l, le nombre de niveau de hiérarchie;

p, le nombre de propriétés;

r, le nombre de relations;

ρ, un facteur permettant une pondération selon l'importance de la propriété ou de la relation.

Cette somme parcourant tous les niveaux de structure permet de constituer une matrice de comparaison W.

Cette matrice peut être utilisée dans une procédure de programmation dynamique et permettre ainsi l'alignement de paires de protéines. Elle peut aussi servir à un alignement de plus de deux séquences. Dans ce cas, la procédure suivie est semblable à celle de Feng et Doolittle.

Zhu *et al.* (1992) apporta une amélioration à cette méthode en introduisant une pénalité de gap comportant trois termes. Les deux premières composantes dépendent des caractéristiques structurales (structure secondaire et accessibilité des chaînes latérales) des deux protéines au niveau des régions concernées. La troisième est une constante indépendante de la longueur.

Dans cette optique, Flores *et al.* (1993) ont réalisé une étude intéressante permettant de suivre l'évolution de différents paramètres structuraux dans des séquences ayant des pourcentages d'identité décroissants. Pour des caractéristiques comme l'accessibilité, les « ooi numbers » (nombre de C_α compris dans une sphère entourant le C_α d'un résidu), la conservation des structures secondaires, la conservation

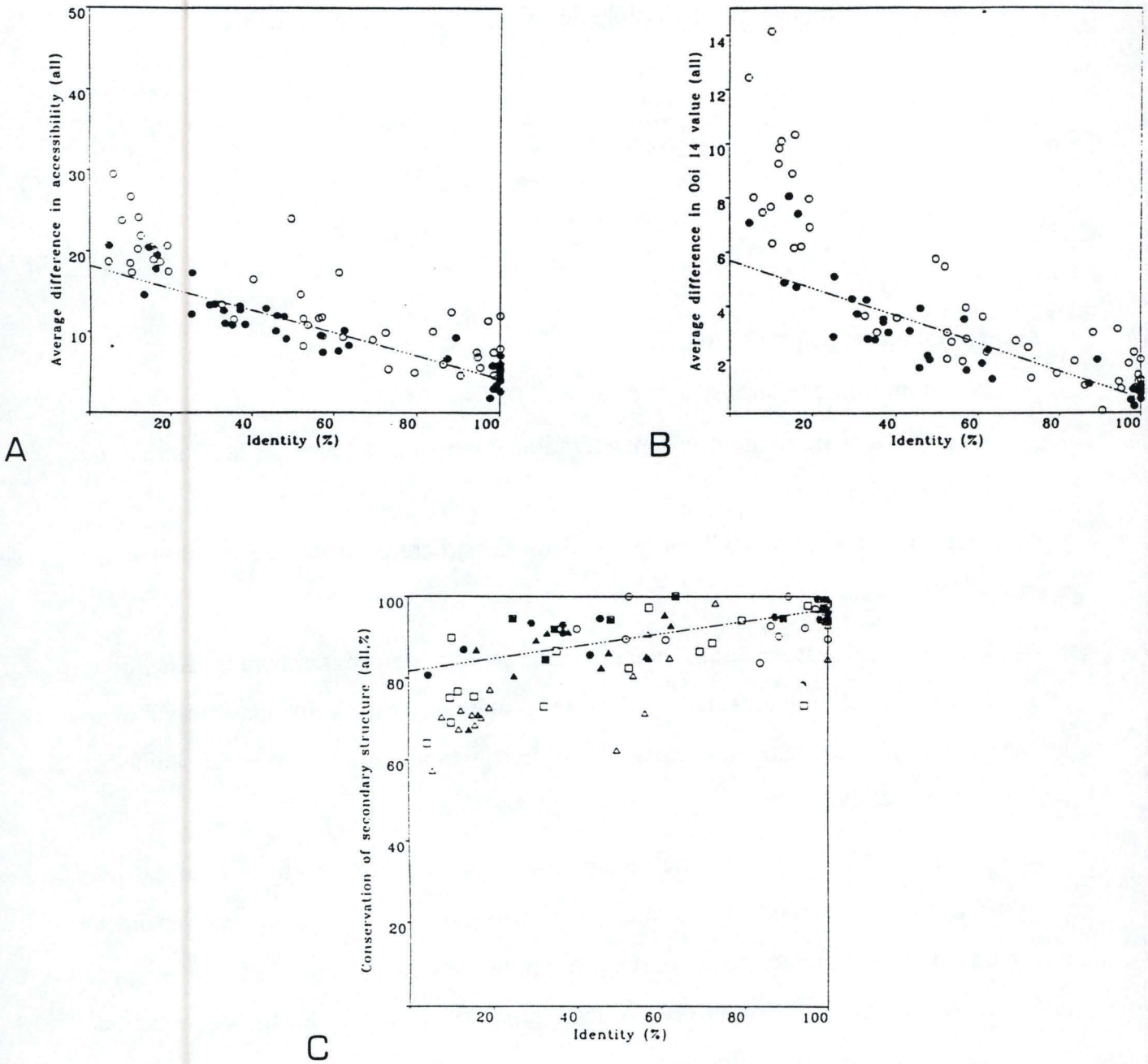


Figure I.41 : Relation entre le pourcentage d'identité et l'accessibilité (a), les ooi numbers (b) et la conservation des structures secondaires (c).

des conformations des chaînes latérales,... ils obtiennent des relations linéaires. Quant à la relation existant entre le pourcentage d'identité et la distance RMS (Root Mean Square ou carré moyen résiduel), elle est de type exponentiel. Ce sont donc autant de caractéristiques qui peuvent être utilisées pour la comparaison de protéines (Fig. I.41).

I.4.2. Superposition par minimisation des distances.

I.4.2.1. Superposition de deux structures.

I.4.2.1.1 Superposition par rigid body transformation.

Développée entre autres par Greer (1981) et Unger *et al.* (1989), cette technique est une des plus anciennes et une des plus utilisées encore à l'heure actuelle. Elle consiste à superposer des segments de structure connue en minimisant les distances qui les séparent. Les atomes du squelette (avec ou sans les atomes d'oxygène) ou les carbones α sont superposés et la distance moyenne entre les deux segments est calculée de la façon suivante :

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

avec :

a, le nombre d'atomes considérés par résidu;

w, le nombre de résidus de chaque segment;

x_{ij}, y_{ij} et z_{ij}, les coordonnées cartésiennes d'un atome i dans un segment j.

De façon à minimiser cette distance, un des deux segments est déplacé par translation et rotation, ce qui n'affecte en rien la forme générale de la structure, mais

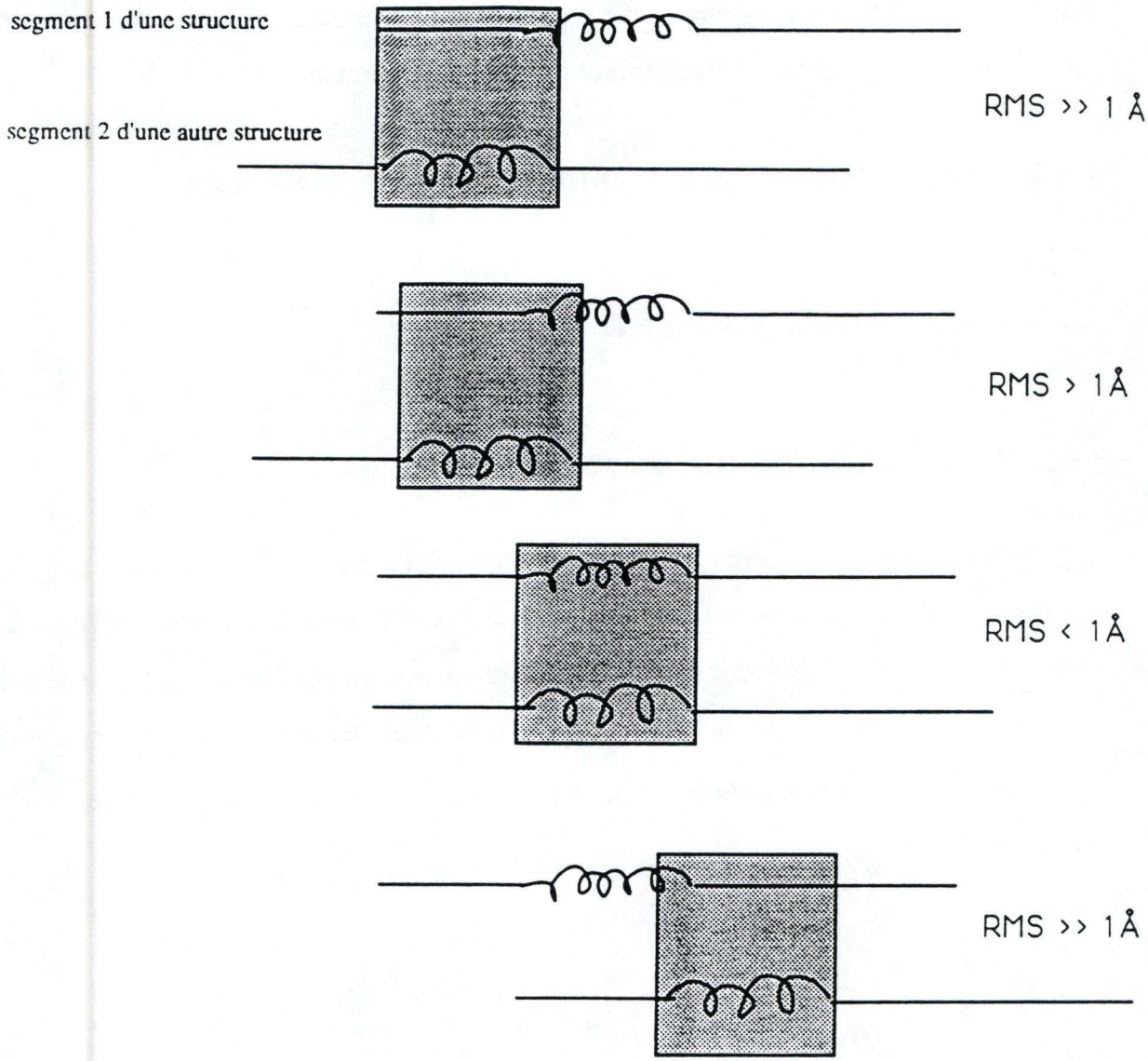


Figure I.42 : Schéma de la superposition de deux segments de structures. Le carré grisé représente le calcul de la distance entre ces deux segments.

uniquement les positions relatives des segments (Fig. I.42). Le RMS cumulé pour tous les atomes considérés est minimisé par un algorithme de moindres carrés non linéaire. Selon Unger et ses collaborateurs, la limite de 1Å permet généralement de distinguer les segments pris au hasard dans des protéines de structure connue. Cette limite est utilisée pour estimer la similarité structurelle de deux segments. Une fois les structures superposées, les régions qui se chevauchent très bien sont appelées « Structurally Conserved Regions » ou SCR. Entre ces segments fort semblables se trouvent des régions qui diffèrent beaucoup et appelées régions variables, VR.

D'autres techniques pour la superposition de deux structures ont été proposées. Pour information, on peut citer notamment celles de Rossmann et Argos (1977), de Mc Lachlan (1972, 1979, 1982), de Remington et Matthews (1980), Lesk (1986),...

La plupart des méthodes ont pourtant une ou plusieurs des limitations suivantes :

- de longues insertions ou délétions sont difficiles à détecter;
- seuls les alignements séquentiels peuvent être réalisés : des domaines similaires mais disposés différemment dans la séquence ne pourront être alignés;
- l'occurrence de multiples copies d'un motif n'est pas détectable;
- une variation des liaisons entre domaines ne peut être détectée;
- un alignement manuel initial est nécessaire;
- le temps calcul exigé est prohibitif.

Diverses solutions ont été proposées pour résoudre ces problèmes, en voici quelques-unes.

I.4.2.1.2. Méthode de Vriend et Sander (1991).

Les auteurs ont mis au point un algorithme, appelé SUPPOS, basé sur le critère des distances atomiques entre des fragments de longueurs différentes. Cet algorithme comporte trois étapes et permet de superposer des structures ayant des connections

topologiquement différentes ou de repérer des ressemblances entre des brins de directions inverses.

La première étape consiste à sélectionner toutes les paires de fragments qui se superposent bien (par une méthode classique de superposition optimale).

La seconde étape permet de regrouper certaines paires de fragments pour former des unités structurelles plus importantes.

Le critère utilisé pour l'extension d'un groupe est basé sur une comparaison des matrices de rotation. Il n'a donc rien de séquentiel et permet donc de repérer des motifs structuraux. En effet, tous les fragments appartenant à une substructure doivent obéir aux mêmes transformations rotationnelles et translationnelles. La troisième et dernière étape permet un raffinement de la superposition des résidus équivalents. Lors de ce procédé, seul le cluster le plus grand est conservé.

Cette méthode est assez rapide pour permettre un alignement pairé de 154 structures représentatives en 1 jour de temps calcul.

I.4.2.1.3. Méthode de Sander et Tuparev (non publiée).

Cette méthode utilise la technique de la programmation dynamique employée dans les alignements de séquences. L'alignement optimal est réalisé en trouvant le meilleur chemin (séquentiel) dans une matrice de similarité entre toutes les paires de résidus possibles. Le meilleur chemin est celui présentant le score le plus grand; l'introduction de gaps est contrôlée par une pénalité. La différence avec les alignements de séquences, réside dans l'expression de la similarité entre deux résidus. Le score de similarité est ici fonction de la distance entre les résidus.

Cet algorithme permet de détecter les changements dans les liaisons entre domaines. Il est plus lent que le précédent, réalisant les alignements des 154 protéines en 3 jours de temps calcul.

I.4.2.1.4. Méthode de Mezei (1994).

Une faible distance RMS garantit une forte ressemblance entre deux structures, mais le contraire n'est pas nécessairement vrai. Si on considère deux protéines composées de deux substructures identiques connectées par un lien ayant des angles de torsion différents, malgré leur ressemblance évidente, il est impossible de les superposer correctement. Il est donc intéressant de mettre au point une technique capable d'identifier des substructures semblables.

La méthode de Mezei permet de déterminer les liens charnières entre les domaines conservés. Pour ce faire, chaque liaison est définie par les deux atomes la formant et par un certain nombre de leurs voisins (soit, dans la plupart des cas, 8 atomes ou moins). La distance RMS minimale entre les paires de liens équivalents est ensuite calculée et comparée à une valeur limite. Si la distance est inférieure à RMS_{lim}, le lien est considéré comme conservé; dans le cas contraire, il devient un lien charnière. Une fois ces liens charnières déterminés, ils sont « délétés » et les superpositions entre domaines sont effectuées.

Cette méthode souffre aussi d'un énorme désavantage : elle doit établir des comparaisons entre liaisons équivalentes et donc nécessite la réalisation préalable d'une superposition des deux structures. Si cette superposition n'est pas correcte, la méthode de Mezei risque de localiser des faux positifs.

I.4.2.2. Superposition de plus de deux structures.

L'approche la plus simple pour la superposition de plusieurs protéines est l'utilisation d'une structure cible sur laquelle seront superposées toutes les autres.

Ces techniques utilisent donc, après superposition des centres de masses, un algorithme classique de comparaison de deux protéines dont le but est la minimisation de :

$$E = \frac{\sum_{j}^{N_{MOL}} \sum_{i}^{N_{ATOM}} w_{ij} \left| X_i - R_j Y_{ij} \right|^2}{\sum_{j} \sum_{i} w_{ij}}$$

avec :

N_{MOL}, le nombre de molécules concernées;

N_{ATOM}, le nombre d'atomes dans les molécules;

X_i, les cordonnées de l'atome i de la protéine cible;

Y_{ij}, les coordonnées de l'atome i dans la protéine j;

R_j, la matrice de rotation superposant la protéine j sur la protéine cible;

w_{ij}, une constante de pondération pour l'atome i de la molécule j.

Il est clair que l'utilisation d'une protéine comme base de la superposition de toutes les autres va biaiser la superposition. Prenons le cas de trois protéines A, B et C. Si A et B sont tour à tour superposées à C, en général, les structures A et B ne sont pas superposées de façon optimale. La superposition de A et B est correcte dans le seul cas où deux des structures sont identiques ou fort semblables au point de vue de leur forme.

Les méthodes présentées ci-dessous sont des méthodes de superpositions multiples simultanées pour lesquelles aucune protéine ne sert de base à l'alignement.

I.4.2.2.1. Méthode de Sutcliffe *et al.* (1987).

La méthode développée par Sutcliffe et ses collaborateurs a pour point de départ la construction d'un modèle dont chaque position est définie comme une moyenne des

positions topologiquement équivalentes dans les différentes protéines à aligner. L'algorithme utilisé comporte cinq étapes :

(1) Choisir au hasard une des structures : elle servira de première approximation pour le modèle. Aligner toutes les autres protéines au modèle.

(2) Déterminer le nouveau modèle grâce à la formule suivante :

$$ {}^{k}\underline{F}_i = \frac{1}{NMOL} \sum_{j=1}^{NMOL} \underline{Z}_{ij} $$

avec :

${}^{k}F_i$, les cordonnées du point i du modèle après la $k^{ème}$ itération;

N_{MOL}, le nombre de molécules;

$Z_{ij} = R_j Y_{ij}$, les coordonnées du point i de toutes les protéines superposées au modèle ${}^{k-1}F_i$.

(3) Si la distance RMS entre les deux modèles consécutifs (${}^{k-1}F$ et ${}^{k}F$) est inférieur à 10^{-5}Å, arrêter le procédé.

(4) Superposer toutes les protéines sur le nouveau modèle ${}^{k}F$(par alignement pairé).

(5) Recommencer au point (1) et incrémenter le nombre d'itération k d'une unité.

Une fois ce modèle construit, toutes les protéines lui sont superposées, donnant un alignement qu'aucune des structures n'influence.

I.4.2.2.2. Méthode de Diamond (1988, 1993).

L'approche réalisée ici est tout à fait différente. Diamond ne calcule pas de structure moyenne invoquant le fait que deux structures identiques avec des orientations différentes ont une structure moyenne qui ne leur est pas superposable. Sa

méthode permet la réorientation simultanée des n structures à superposer via un procédé qui, si on omet l'initialisation, est d'ordre n.

Ce traitement est basé sur une représentation matricielle des rotations :

$$P = \begin{pmatrix} l\sin\left(\theta/2\right) \\ m\sin\left(\theta/2\right) \\ n\sin\left(\theta/2\right) \\ \cos\left(\theta/2\right) \end{pmatrix}$$

avec :

l, m, n, indiquant la direction de l'axe de rotation;

θ, l'angle de rotation.

Notons E_{AB}, la distance entre les protéines A et B. Si on considère n protéines à superposer, il est nécessaire de calculer n(n-1)/2 interactions semblables à E_{AB} et de minimiser leur somme. Diamond a pu montrer que la somme des carrés des différences de coordonnées entre une structure B et une structure A subissant une rotation est définie par :

$$E_{AB} = E_{OAB} - 2\rho_A^T P_{AB} \rho_A$$

avec :

E_{0AB}, la distance entre B et A, avant rotation;

ρ_A, la rotation à appliquer à A;

P_{AB}, une matrice 4×4 réelle symétrique bilinéaire sur A et B, permettant le contrôle de la rotation de A sur B.

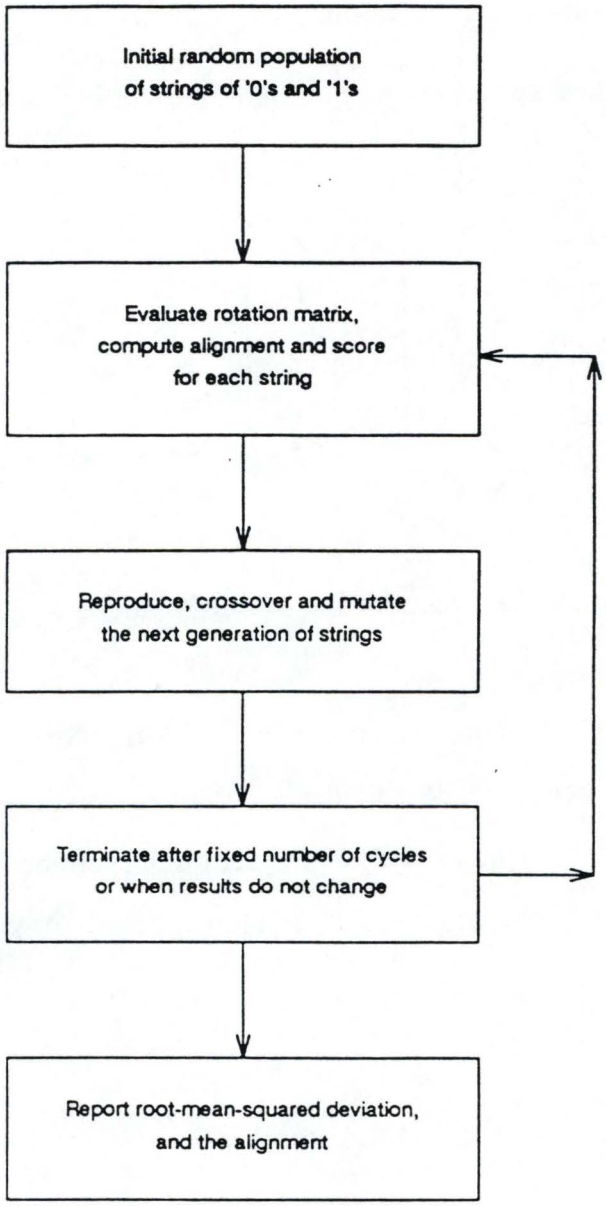


Figure I.43 : Schéma de l'algorithme génétique utilisé par Johnson *et al.* (1994) pour la comparaison de structures tridimensionnelles.

Il apparaît évident que E_{AB} est minimisé lorsque le deuxième terme de la différence est maximal. Cette condition est réalisée lorsque ρ_A est le premier vecteur propre de P_{AB}. La recherche de la rotation optimale revient donc au calcul de ce vecteur propre.

I.4.2.2.3. Méthode de Johnson *et al.* (1994).

Johnson et ses collaborateurs ont développé un algorithme génétique pour aligner des, structures de protéines. Selon eux, le désavantage de bon nombre de méthodes de superposition est la nécessité de connaître, au préalable, un ensemble de positions équivalentes communes à toutes les structures. La méthode développée est inspirée par les mécanismes de sélection naturelle. Cette méthode en cours de développement n'est pour le moment fonctionnelle que pour l'alignement de deux protéines (Fig. I.43).

L'algorithme utilise une représentation « chromosomale » du problème : une série de 56 bits au total permet d'encoder les transformations subies par la seconde structure. 24 bits (3×8) codent pour les angles de rotation α, β et γ. 24 autres bits codent pour les translations le long des axes x, y et z. Les 8 derniers codent pour la valeur assignée à la pénalité pour un gap.

Dans un premier temps, le centre de gravité de chaque structure est calculé et translaté vers l'origine des axes. Une structure est alors fixée et toutes les opérations ultérieures (rotations et translations) ne sont appliquées qu'à la seconde structure.

L'algorithme choisit ensuite une population de série de 56 bits, chacune définissant une rotation. Ces différentes rotations sont appliquées à la structure mobile et permettent donc l'obtention de plusieurs alignement des deux protéines. Pour chacun, un score est calculé par la programmation dynamique.

L'étape suivante est la constitution d'une nouvelle population de série de 56 bits. Les différentes séries seront plus ou moins impliquées dans les trois opérations génétiques selon le score obtenu pour la superposition (principe de la sélection).

Les séries les plus performantes seront donc reproduites (elles seront présentes dans la génération suivante), mutées (un de leur bit sera changé) et pourront même subir des crossing-over (des morceaux d'information seront échangés entre parents).

Une fois la nouvelle génération constituée, les rotations et translations définies par chaque série sont réalisées et évaluées. Ce procédé se poursuit pendant un nombre fixé de cycles et se termine par une minimisation du RMS visant à optimiser l'alignement obtenu.

I.4.3. Alignements de matrices de distances.

Soit A, une protéine de structure tridimensionnelle connue possédant, si on exclut les hydrogènes, $N_{(A)}$ atomes. Il est possible de calculer une matrice contenant les $N_{(A)} \times N_{(A)}$ distances existant entre les différents atomes. L'élément $D_{A(ij)}$ de cette matrice représente la distance entre les $i^{ème}$ et $j^{ème}$ atomes de la protéine A. En appliquant ce même procédé à une autre protéine B, on obtient deux matrices D_A et D_B permettant une comparaison entre les structures des deux protéines.

I.4.3.1. Mesure de la similarité globale (Pepperrell et Willett, 1991).

Cette première mesure permet d'évaluer la topographie générale, la forme d'une structure par la distribution des distances inter-atomiques. L'algorithme permet le calcul des $N_{(A)} \times (N_{(A)}-1)$ distances distinctes et leur répartition dans des classes de fréquences. Les distributions $F_{(A)}$ et $F_{(B)}$ peuvent ensuite être comparées pour obtenir une idée globale de la ressemblance entre les deux protéines.

La mesure présentée ci-dessus attribue la même importance à toutes les distances. Or elle tient compte des liaisons carbone-carbone, carbone-hétéroatome et hétéroatome-hétéroatome qui influencent différemment la structure.

Un deuxième procédé permet l'établissement de distributions différentes selon le type de distance. La somme de ces distributions est pondérée par des facteurs relativisant l'importance de chacune afin de donner la distribution totale.

I.4.3.2. Alignement par comparaison de matrice de distance, méthode de Kikuchi (1992).

L'étape initiale de cette technique est la construction de la matrice de distance. Contrairement à ce qui a été vu dans le point précédent, elle ne tient plus compte que de la distance entre les N×N C_α de chaque résidu. De plus, différentes échelles interviennent pour la mesure de ces distances, ces échelles variant selon le nombre de résidus (k) séparant les C_α considérés. La distance est multipliée par un facteur c dont la valeur varie :

c = 1 si k est inférieur à 9;

c = 2 si k est compris entre 9 et 20;

c = 3 si k est compris entre 21 et 30;

c = 4 si k est compris entre 31 et 40, et ainsi de suite.

L'étape suivante est la construction de l'Average Distance Map (ADM), un graphe de dimension N×N où la présence d'un point en ij signifie que cette distance est inférieure à une valeur seuil fixée à l'avance par l'utilisateur. Ce point indique la présence d'une « contact pair ».

Si on considère une paire de protéines A et B ($N_B \leq N_A$), une superposition de leur ADM va permettre de comparer leur structure. Dans un premier temps, $ADM_{(B)}$ est superposé à $ADM_{(A)}$ de façon à faire coïncider leur coin supérieur. Ensuite $ADM_{(B)}$ est glissé d'un résidu le long de la diagonale, vers le coin inférieur droit. Cette étape est répétée sur plusieurs résidus puis la même procédure est appliquée pour une translation vers le coin supérieur gauche.

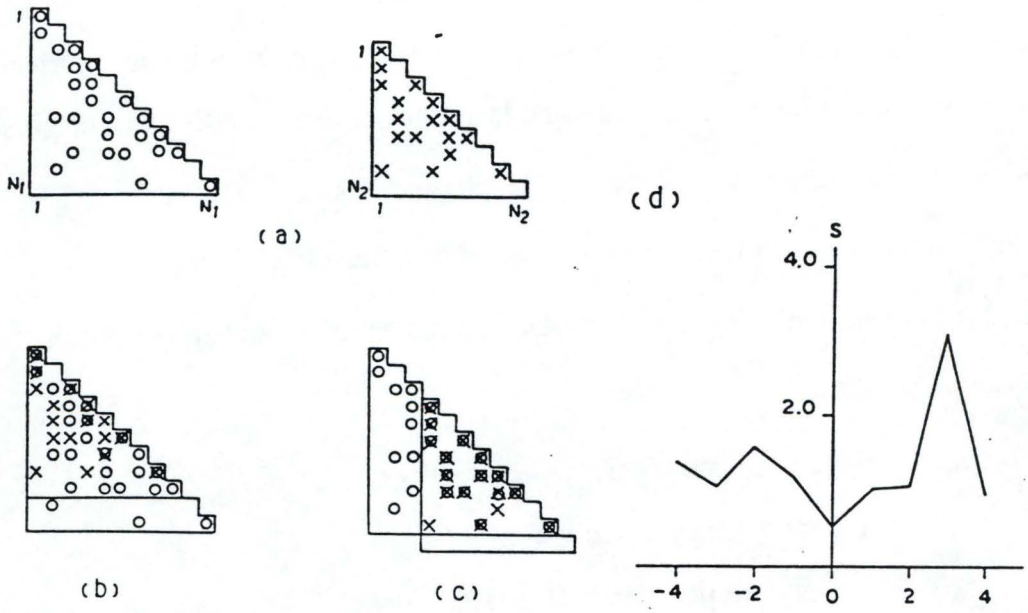


Figure I. 44 : a. Contact maps hypothétiques de deux protéines 1 et 2 fictives.
b. Superposition des deux protéines afin de faire correspondre les bords supérieurs des deux cartes.
c. Superposition optimale des deux protéines.
d. Variation du score S lors des différentes translations réalisées pour superposer les protéines 1 et 2.

A chaque position, un score de similarité S est calculé par le rapport entre la fréquence de superposition de « contact pairs » observées (Q_1) et la fréquence auquel on pourrait s'attendre par hasard (Q_2).

$$Q_1 = \frac{m}{(M_A + M_B - m)}$$

avec

m, le nombre de coïncidences de « contact pairs » dans la région superposée;

M_A et M_B, les nombres de « contact pairs » dans les régions superposées de A et B.

$$Q_2 = \frac{p}{M_A + M_B - p}$$

$$p = \frac{2M_A M_B}{(L(L+1))}$$

avec

p, le nombre de coïncidences dues au hasard;

L, le nombre de résidus superposés.

Lors de la translation de $ADM_{(B)}$ sur $ADM_{(A)}$, la recherche du score maximal va permettre de localiser l'alignement optimal (Fig. I.44).

Selon les auteurs, leur méthode est très bien adaptée à la comparaison de structures semblables ou la recherche de domaines communs. Par contre, l'alignement global de protéines topologiquement dissemblables est moins évident. La fluctuation des coïncidences dues au hasard est trop importante par rapport au pic indiquant les similarités.

Il existe d'autres méthodes utilisant les matrices de distances, on peut citer, par exemple, celle de Holm et Sander (1993). Leur algorithme, DALI, crée, dans un premier temps, une matrice de distances C_α-C_α. Cette matrice est ensuite décomposée en sous-matrices comparant les distances entre hexapeptides. Les sous-matrices calculées pour les deux protéines sont comparées et combinées en vue d'obtenir un ensemble cohérent de paires le plus grand possible. Un algorithme de Monte-Carlo est utilisé pour l'optimalisation de l'alignement.

Une méthode développée par Taylor et Orengo (1989), utilise aussi une comparaison des distances résidu-résidu pour aligner deux protéines. Elle est pourtant différente. En effet, elle considère non plus les distances entre C_α, mais entre C_β porteurs de plus d'informations structurelles. Par exemple, dans le cas d'un plan β, les C_β indiquent le côté où se trouvent les chaînes latérales.

Une troisième méthode développée par Richards et Kundrot (1988) utilise des matrices de distance où chaque case indique les relations existant entre les éléments de structure secondaire.

I.4.3.3. Méthode utilisant la représentation des protéines par des « labelled graph ».

L'algorithme utilisé est dérivé d'une méthode développée à l'origine en vue de l'étude de petites molécules chimiques. Selon cette méthode, les protéines sont représentées par une « table de connexion » contenant la liste de tous les atomes non-hydrogènes et les informations relatives aux liaisons existant entre ces atomes. Une caractéristique importante des tables de connexion est qu'elles peuvent être vues comme un graphe, une construction mathématique décrivant un set d'objets, appelés « nodes » et les relations , appelées « edges », existant entre eux. On peut définir un graphe G, constitué d'une série de « nodes », V et de « edges », E, réalisant les

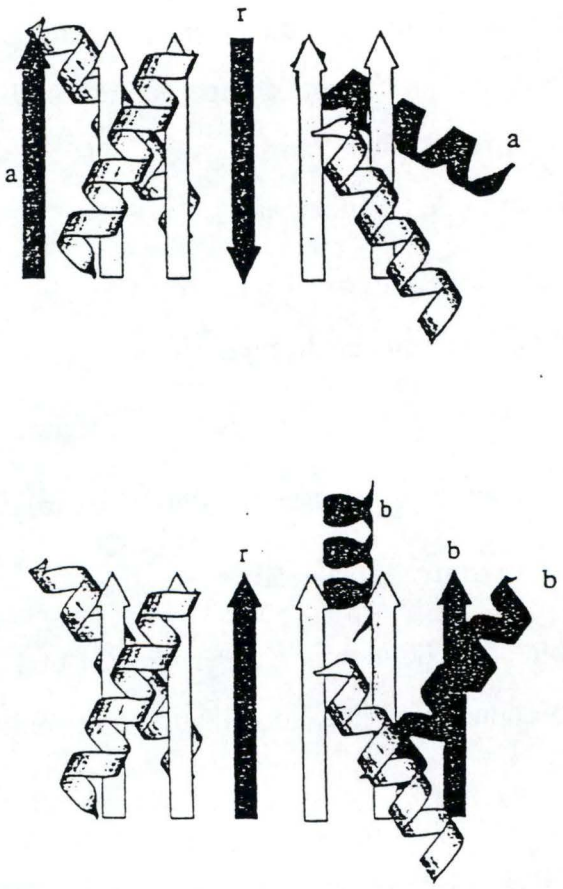


Figure I. 45 : Illustration du « maximal common subgraph » entre deux protéines A et B. Les éléments de structure secondaire spatialement équivalents sont représentés en clair. Les autres sont soit uniquement présents dans A (notés a) ou B (notés b), soit présents dans les deux protéines mais dans des orientations opposées (notés r).

connections entre les « nodes » ($E \subseteq N \times N$). Deux « nodes » sont considérés comme adjacents s'ils sont connectés par un « edge ». Un sous-graphe de G est un sous-ensemble, P, des « nodes » de G lié à un sous-ensemble, F, des « edges » reliant les paires de « nodes » de P (avec $P \subseteq V$ et $F \subseteq P \times P$).

Deux graphes A et B sont dit isomorphes s'ils ont la même structure. Un sous-graphe commun à deux graphes A et B consiste en un sous-graphe a de A et un sous-graphe b de B de sorte que a et b soient isomorphes. Le sous-graphe commun maximal est le plus grand de ces sous-graphes communs. Dans ce contexte, il existe trois types différents d'algorithmes visant à trouver les graphes isomorphes, les sous-graphes isomorphes et le plus grand sous-graphe isomorphe.

Pour la représentation d'une molécule biologique, les « nodes » représentent les atomes et les « edges », les distances inter-atomiques. Ces graphes donnent donc une bonne représentation des structures secondaires et de la position relative entre les éléments de structures. Mais la recherche de ressemblances entre ces graphes demande un temps calcul considérable. Pour en donner une idée, Subbarao et Haneef (1991) ont estimé à 26 heures le temps nécessaire pour localiser les similarités entre une petite protéine et une banque de 107 structures.

Dans cette optique, une autre représentation a été développée. Les hélices et les plans d'une protéine sont représentées par un vecteur suivant leur axe principal. Dans ce cas les « nodes » du graph sont les vecteurs en question et les « edges » sont les distances et les angles existant entre ceux-ci (Fig. I.45).

Nous n'allons pas décrire les algorithmes utilisés, simplement, il est intéressant de savoir qu'il en existe de deux types différents (Grindley *et al.*, 1993).

Le premier localise un atome commun à toutes les structures puis ajoute à ce dernier un atome à la fois. Et ce, jusqu'à ce que la structure 3D commune ne puisse plus être étendue. Le deuxième travaille différemment. Il recherche les cliques

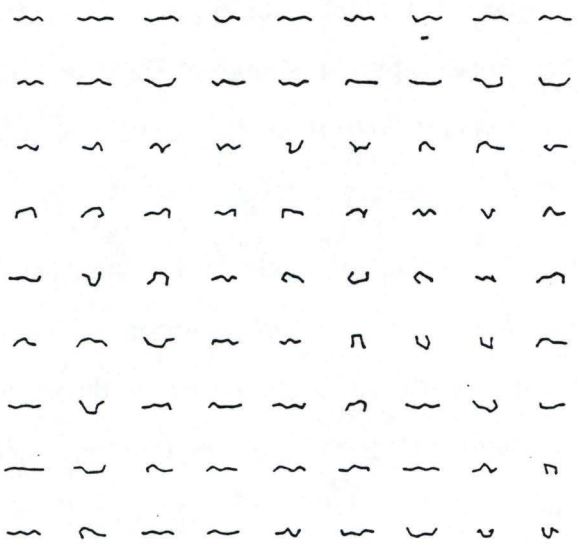


Figure I.46 : Représentation schématique des 81 blocs de construction (hexamères) rencontrés fréquemment dans les protéines.

présentes dans un graphe. Une clique est un sous-graphe où tous les « nodes » sont connectés, ce sous-graphe n'étant pas contenu dans un autre ayant ces propriétés. Cet algorithme réalise la recherche de la plus grande clique possible. Cette recherche est réalisée sur un nouveau graphe où chaque « node » représente une structure secondaire commune aux deux structures comparées et où les « edges » existent si les distances et les angles existant entre ces structures sont constants (les limites de variation sont de ±30° pour les angles et ±5Å pour les distances).

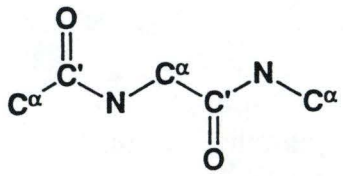
Selon Mitchell *et al.* (1990), cette méthode procure de bons résultats. De plus, elle est assez rapide pour permettre des recherches dans une banque de données. Un des avantages de ces méthodes est qu'elles sont capables de retrouver un motif choisi par l'utilisateur.

I.4.4. Méthode utilisant les motifs structuraux.

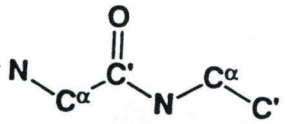
Lorsqu'on compare les structures de plusieurs protéines, on peut remarquer que certains motifs se retrouvent régulièrement. Les protéines utilisent un certain nombre de blocs de construction.

Efimov (1993) a réalisé des études sur ces motifs. Il définit trois niveaux de structures standards : les motifs courts (présents dans les régions irrégulières, les turns, ...), les motifs formés de l'assemblage de 2 éléments de structures secondaire (α-α hairpin, α-α corner, β-β hairpin, ...) et, enfin, les motifs formés de plus de 2 éléments de structures secondaires (différents plans β, ...). Le but de ces travaux est l'établissement d'une classification des différentes structures standards rencontrées.

Unger et Susman (1993), quant à eux, se sont concentrés sur des motifs structuraux de courtes longueurs, en fait 6 acides aminés. Ils ont pu dénombrer 81 blocs de 6 résidus qui revenaient régulièrement dans les structures connues (Fig; I.46).



Definition of the (ϕ, ψ) fragment.



Definition of the (ψ, ϕ) fragment.

Figure I.47 : Définition des fragments (φ, ψ) et (ψ, φ).

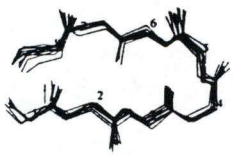
One-letter code and boundary ellipsoids to characterize each cluster

Cluster	One-letter code	Number of principal components	Fractional distribution [a] (%)	R [b]
α_1	a	6	92.0	5.5
α_2	b	6	96.6	3.1
β	c	4	92.6	3.5
γ_1	d	4	90.3	4.5
γ_2	e	4	92.4	2.2
A	A	5	91.6	4.0
B_1	B	4	93.5	4.0
Γ_1	C	4	90.5	4.0
Γ_2	D	5	94.2	4.0
Γ_3	E	5	91.8	2.1
Δ	F	5	92.2	4.0

[a] Fraction of distribution in percent of conformational points in each cluster that can be accounted for by the first several principal components whose number is listed as the number of principal components.

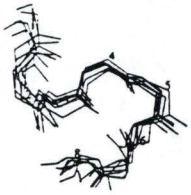
[b] The size of the ellipsoid is taken to be R times the standard deviation along each principal component.

Figure I.48 : Table des 5 groupes de fragments (φ, ψ) et des 6 groupes de fragments de fragments (ψ, φ).



(a)

(b)

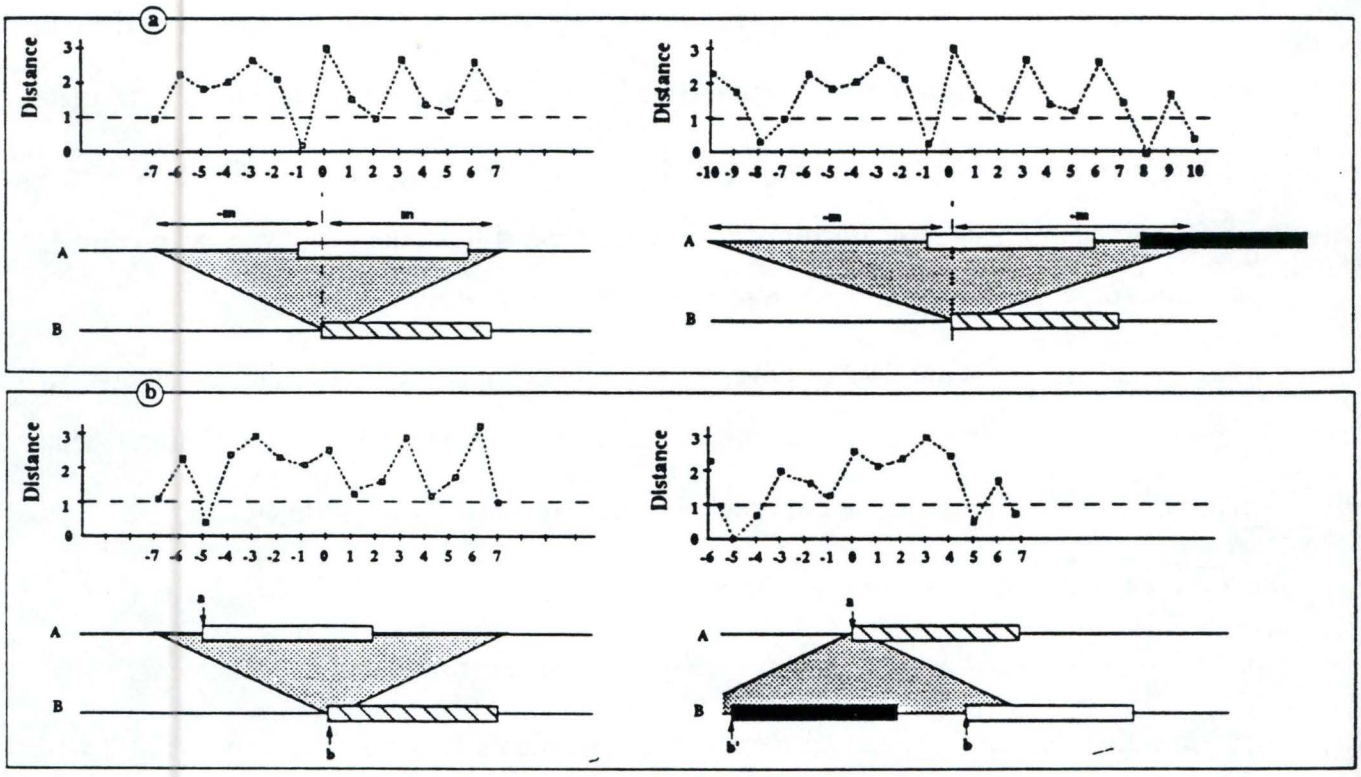


(c)

Figure I.49 : Exemples de superpositions de courts fragments de structures présentant des codes identiques. Il s'agit d'un β-ladder (a), d'un turn (b) et d'un loop αβ (c).

A eux seuls, ces 81 blocs représentaient 76% de tous les hexamères trouvés dans la banque si la déviation permise était de ±1Å. Ce chiffre montait à 92% si cette déviation était de ±1,25Å.. Leur travail se poursuivit par l'observation des acides aminés présents à chaque position de chaque type d'hexamère, en vue de prédire la structure prise par un segment de 6 résidus pris au hasard.

Takahashi et Go (1993) développèrent une méthode d'alignement basée sur l'étude de motifs structuraux de longueur réduite. Dans ce but, ils définirent deux types d'unités structurelles se chevauchant, les fragments (ϕ,ψ) et (ψ,ϕ) (Fig. I.47). Ils correspondent respectivement aux segments ($C_{\alpha i}$-C'O-N-$C_{\alpha i+1}$-C'O-N-$C_{\alpha i+2}$) et (N-$C_{\alpha i}$-C'O-N-$C_{\alpha i+1}$-C'). Par une analyse en composante principale, les fragments (ϕ,ψ) furent regroupés en 5 groupes distincts et les fragments (ψ,ϕ) en 6 groupes (Fig. I.48). A chaque groupe est attribué un symbole en une lettre servant à coder une structure protéique. Si deux structures présentent des séries de lettres (codant pour des conformations) identiques, on peut en déduire qu'elles sont similaires. Des tests ont été réalisés sur de courts fragments de structures (Fig. I.49). Bien que cette méthode semble plus fiable dans le cas de structures de type hélice et turn que dans le cas de structures étendues comme les feuillets, elle est intéressante pour la détection de similarité locale entre des structures protéiques.



Variation d'une mesure de distance au cours d'un balayage. Une fenêtre initiale (rectangle hachuré) est située en position 0 dans une séquence. L'axe X correspond au déplacement d'une fenêtre mobile de -m à + m résidus dans l'autre séquence. L'axe Y représente une mesure de distance hypothétique, chaque point (carré blanc) indiquant la valeur de cette distance pour l'appariement de la fenêtre initiale et de la fenêtre mobile située à la position correspondante. Le seuil en dessous duquel les fenêtres sont appariées est représenté par une ligne discontinue (Y=1). L'appariement caractérisé par la distance la plus faible est sélectionné. Un rectangle blanc symbolise l'appariement attendu. En cas d'échec, un rectangle noir symbolise l'appariement obtenu. ***a.*** *En augmentant la longueur du balayage, un succès (gauche) se transforme en échec (droite).* ***b.*** *L'appariement attendu est sélectionné dans la séquence A lorsque la fenêtre initiale se trouve dans la séquence B (gauche). Par contre, ce même appariement n'est pas sélectionné dans la séquence B lorsque la fenêtre initiale se trouve dans la séquence A.*

Figure I.50 : Schéma du balayage réalisé entre deux protéines A et B et de la variation d'une distance hypothétique le long de ce balayage.

I.5. MATCH-BOX : programme d'alignement multiple de séquences et de structures protéiques.

Le programme MATCH-BOX mis au point par Depiereux et Feytmans (1991,1992, 1994) a pour but de rechercher les segments similaires dans un ensemble de protéines et de les faire correspondre dans un alignement. Un alignement optimal doit être celui que l'on trouverait en réalisant la superposition des structures.

Le processus d'alignement peut être subdivisé en plusieurs étapes présentées ci-dessous.

I.5.1. Balayage et comparaison des séquences.

Initialement, les r protéines à aligner sont juxtaposées de façon à mettre face à face le premier résidu de chacune d'elles. Les comparaisons entre les séquences sont effectuées par l'intermédiaire de segments de w résidus appelés fenêtres. Chaque fenêtre d'une séquence (fenêtre fixe) est comparée à toutes les fenêtres de même longueur (fenêtres mobiles) situées dans une région plus ou moins étendue en amont et en aval de la position correspondante dans les (r-1) autres séquences. Ces comparaisons permettent d'apparier les segments les plus similaires d'une protéine à l'autre dans un certain voisinage. La taille w des fenêtres est fixée arbitrairement le plus souvent à 7 résidus, nombre suffisant pour être représentatif des diverses structures tout en permettant de cerner les régions conservées sur une faible distance. Par définition, le balayage désigne l'ensemble des comparaisons effectuées entre une fenêtre fixe et toutes les fenêtres mobiles d'une autre séquence (Fig. I.50). Étant donnée la position d'une fenêtre fixe, les fenêtres mobiles peuvent être définies entre deux positions extrêmes sur les autres protéines : m résidus en amont et m en aval par

rapport à la position de la fenêtre fixe. Par convention, nous dirons que m représente la largeur du balayage des r séquences, soient (2m+1) comparaisons entre une fenêtre fixe et l'ensemble des fenêtres mobiles. Cette longueur de balayage m peut être modifiée pour étendre ou restreindre la zone de projection. Si m est supérieur à L_x, la longueur de la plus longue séquence, toutes les comparaisons possibles seront effectuées.

La comparaison de deux fenêtres a pour but de détecter la fenêtre mobile la plus semblable à la fenêtre initiale et de déterminer si la similarité de cette paire est meilleure que celle que l'on peut s'attendre à observer par hasard. Les différentes mesures de similarité considérées pour répondre à ce problème seront présentées dans les points suivants.. De façon générale, deux fenêtres sont considérées comme significativement similaires si leur distance est inférieure à un seuil (=cutoff) défini par l'expérimentateur. Pour le cas des alignements de séquences, ce seuil est déterminé, de manière empirique, par l'étude de la distribution des fréquences des distances pour le set de protéines de départ et pour ce même set dont les séquences ont été réarrangées aléatoirement. Pour les alignements de structures, le choix des cutoffs est plus délicat. Une valeur est choisie soit parce que les résultats qu'elle permet d'obtenir sont bons, soit par corrélation avec les valeurs de distances entre séquences.

I.5.2. Définition des mesures de distances entre fragments de structure.

Comme nous l'avons déjà mentionné, il existe trois critères de distances possibles pour évaluer la ressemblance entre des segments de structures.

I.5.2.1. Utilisation de matrices de scores entre acides aminés.

Différentes matrices ont été utilisées, une tenant compte des distances physico-chimiques entre résidus (DFK), et les trois autres reposants sur des fréquences de

substitution observées. Parmi ces trois dernières, une a pour point de départ des alignements de séquences complets (GONNET), la seconde part d'alignements de blocs de séquences ne contenant pas de gap (HENIKOFF) et la troisième est basée sur des alignements de structures (BIRKBECK97).

I.5.2.2. Distance RMS.

La distance RMS est typiquement la mesure de distance de référence entre deux structures données. Elle correspond à la distance euclidienne moyenne minimale entre les atomes des deux squelettes protéiques impliqués dans la comparaison.

D'un point de vue quantitatif, la distance RMS moyenne entre deux fenêtres de w résidus est notée :

$$RMS = \min\left[\sum_{i=1}^{aw}\frac{(x_{i1}-x_{i2})^2+(y_{i1}-y_{i2})^2+(z_{i1}-z_{i2})^2}{aw}\right]^{0.5}$$

avec :

a, le nombre d'atomes considérés par résidu;

x_{ij}, y_{ij}, z_{ij}, les coordonnées cartésiennes de l'atome i dans la fenêtre j (j=1,2).

Bien que le nombre d'atomes considérés pour la comparaison ne soit pas fixe, nous le limitons généralement à ceux du squelette protéique en y incluant l'oxygène du carbonyle (N, C_α, C, O). Malgré son utilité, la mesure de la distance RMS ne tient pas compte de la forme des fragments; elle n'exprime que la proximité physique entre les atomes. Il faut également définir un seuil de distance RMS en dessous duquel deux segments comparés sont suffisamment similaires pour être appariés.

La superposition de structure de protéines permet de constater que les régions similaires sont faiblement distantes. D'après Unger *et al.* (1989), le seuil de distance RMS à considérer pour différencier les segments structurellement similaires de ceux qui ne le sont pas est de 1Å et représente les appariements entre fragments superposables. L'autre pic de fréquence caractérise les appariements entre fragments non superposables comme le seraient des fragments aléatoires.

I.5.2.3. Mesure de la forme des segments.

Comme la mesure de la distance RMS néglige la forme des segments, il est nécessaire de compléter la caractérisation structurale des segments en comparant leur forme. Un moyen d'y parvenir consiste à déterminer pour chaque segment les dimensions moyennes d'une boîte qui le contiendrait. La comparaison des dimensions des boîtes permet de classer les segments correspondants selon leur forme globale. Cette mesure de la forme est appelée SHOEBOX, par analogie à des boîtes à chaussures dont la comparaison de forme revient à comparer la taille du contenu. Dans ce qui suit le SHOEBOX est défini de façon quantitative.

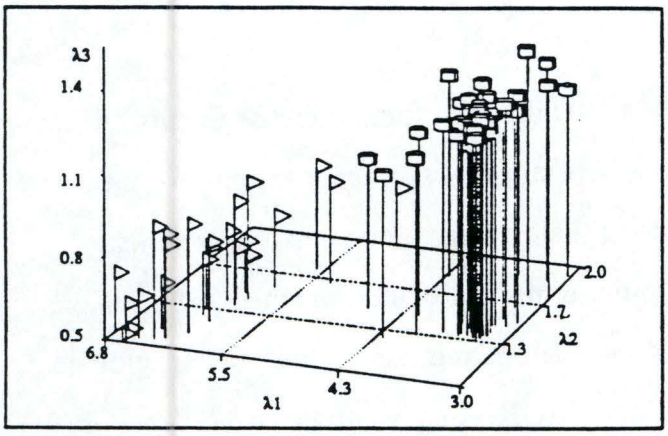
La mesure de la forme des segments est basée sur un certain nombre n d'atomes considérés par résidu. Les coordonnées cartésiennes (X, Y et Z, exprimées en Å) de ces atomes sont stockées dans une matrice X n×3. Soit H une matrice carrée n×n dont chaque élément est une constante 1/n.

Par définition, si

$$A = X - HX$$

et

$$V = \frac{1}{nA'A}$$



Représentation des dimensions des "boîtes" contenant des structures secondaires typiques. Les segments d'hélices alpha sont symbolisés par des cylindres, les segments de brins bêta par des triangles. λ1, λ2 et λ3 sont les trois valeurs propres, classées par ordre décroissant. Elles sont proportionnelles aux trois dimensions d'une boîte englobant la plupart des atomes du segment. Les deux types de structures correspondent à des formes de "boîtes" différentes et sont bien discriminés.

Figure I.51 : Distribution de la forme des "boîtes" contenant des segments d'hélice α (cylindres) et de feuillet β (triangles).

alors V est une matrice 3×3 et représente la matrice variance-covariance du nuage de points. Les trois valeurs propres de cette matrice V, λ_1, λ_2 et λ_3 sont déterminées en résolvant le système d'équations défini par le déterminant

$$\left| V - \lambda . I \right| = 0$$

Ces valeurs (exprimées en Å²) correspondent chacune à la plus grande variance qu'il est possible de trouver dans des directions orthogonales. Les valeurs représentées par les racines carrées de λ_i sont proportionnelles aux trois dimensions de la boîte qui contient, en moyenne, les points représentant les atomes du segment considéré. Quand la mesure est appliquée à des structures connues, il apparaît par exemple que les boîtes contenant les brins β sont plus longues et plus plates que celles contenant les hélices α (Fig. I.51).

La différence de forme entre une fenêtre fixe et une fenêtre mobile est mesurée par le carré de la distance entre les dimensions correspondantes :

$$D_\lambda = \sum_{i=1}^{3} (\lambda_{i1} - \lambda_{i2})^2$$

Le seuil associé à la mesure de cette distance est estimée *a posteriori* par corrélation avec les autres mesures de distances.

I.5.2.4. Combinaison des différentes mesures.

Les trois critères définis ci-dessus sont combinés pour le calcul de la distance globale entre deux segments :

$$DS = \frac{Ds(d)}{S1} \times \frac{D_\lambda}{S2} \times \frac{r.m.s.}{S3}$$

avec S_1, S_2 et S_3, les seuils respectifs de chaque distance considérée. Idéalement, deux segments similaires sont caractérisés par trois distances, toutes inférieures à leur seuil. Dans ce cas, DS est un produit de valeur inférieure à l'unité. C'est la raison pour laquelle l'algorithme d'alignement apparie les segments pour lesquels DS<1. Comme la mesure de DS est un produit de trois facteurs, il n'est pas indispensable que chacune des distances soit inférieure à son seuil pour que DS soit inférieur à 1. Certaines mesures peuvent se compenser. Il est évident que dans le cas où seulement un ou deux critères sont utilisés, seuls les termes les concernant sont gardés dans l'expression de DS.

I.5.3. Recherche des appariements complets.

A ce stade de la procédure d'alignement multiple, il existe une collection de segments de w résidus appariés en fonction des critères cités ci-dessus. Il est nécessaire de trier tous ces appariements de façon à pouvoir regrouper, dans l'alignement final, ceux qui traduisent une similarité optimale.

Soit un groupe de trois segments A, B et C correspondant par exemple respectivement à une fenêtre initiale (A) et à deux fenêtres mobiles (B et C) qui lui sont similaires. Il est important de rappeler que les mesures de similarité ne sont pas transitives, ce qui implique que les deux fenêtres B et C ne sont pas nécessairement similaires. Dés lors, en toute généralité, deux types de relations sont possibles entre les

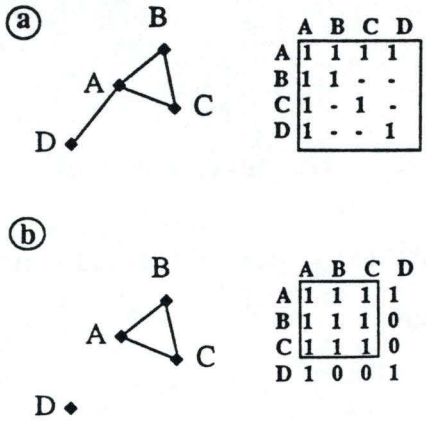


Figure I.52 : **Représentation de la procédure de regroupement : appariements simples (a) et complets (b).**

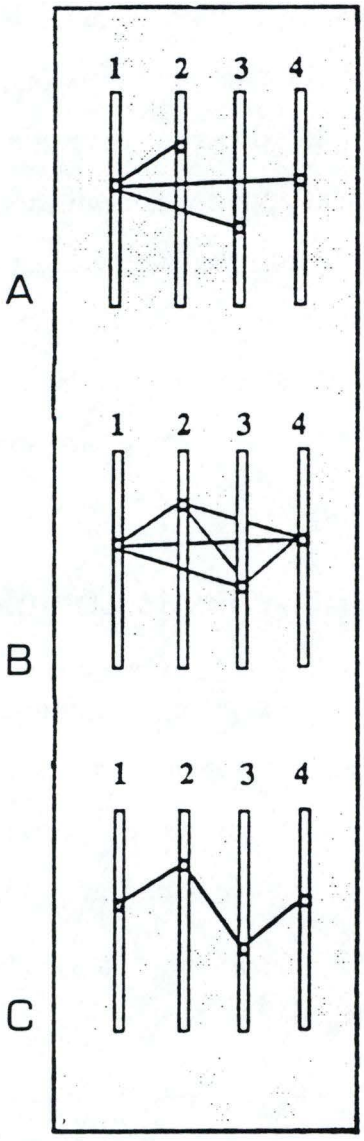


Figure I.53 : Recherche des appariements complets par la méthode de la distance minimale.

a. Le meilleur appariement est sélectionné dans chaque séquence.
b. Un test vérifie que cette sélection forme bien un appariement complet.
c. Un appariement complet est formé s'il en existe un.

trois segments A, B et C. Il est question d'un appariement simple lorsque A est similaire à B et C, ou lorsque B est similaire à A et C, ou encore lorsque C est similaire à A et B. Lorsque A est similaire à B et C, et lorsque B est similaire à C, tous trois forment un appariement complet (Fig. I.52). Ces définitions restent valables quel que soit le nombre de segments considérés. Un quatrième segment (D) peut être ajouté au groupe précédent et, s'il est similaire à au moins un autre, sa liaison au groupe est dite simple. D peut aussi être similaire au trois autres segments et, par définition, former alors un appariement complet. C'est la recherche de cette dernière situation qui est accomplie par l'algorithme.

Puisqu'il n'est pas possible d'envisager l'ensemble des résultats qui satisfont aux critères d'appariement, nous devons limiter la recherche au meilleur appariement rencontré, celui dont la distance est la plus faible (Fig. I.53).

Cette sélection est accomplie selon trois critères, cités par ordre décroissant de leur importance :

- le plus grand nombre de résidus identiques;
- la plus petite distance entre les segments (selon les critères cités plus haut);
- la plus petite différence entre la position des fenêtres dans les deux séquences.

La seconde étape revient à calculer une matrice de vérité R associée à une fenêtre fixe et tous les appariements sélectionnés dans les autres séquences. Dans le cas où R=1, un appariement complet est sélectionné, et dans le cas contraire, aucun appariement complet n'est sélectionné. Cette méthode a l'avantage d'éviter de devoir rechercher toutes les cliques possibles puisque seuls les segments de distance minimale sont pris en compte.

Néanmoins, le nombre d'appariements complets potentiels est fort réduit et les critères d'appariement ne sont pas infaillibles.

L'alignement final des protéines est réalisé dès que cette opération est terminée.

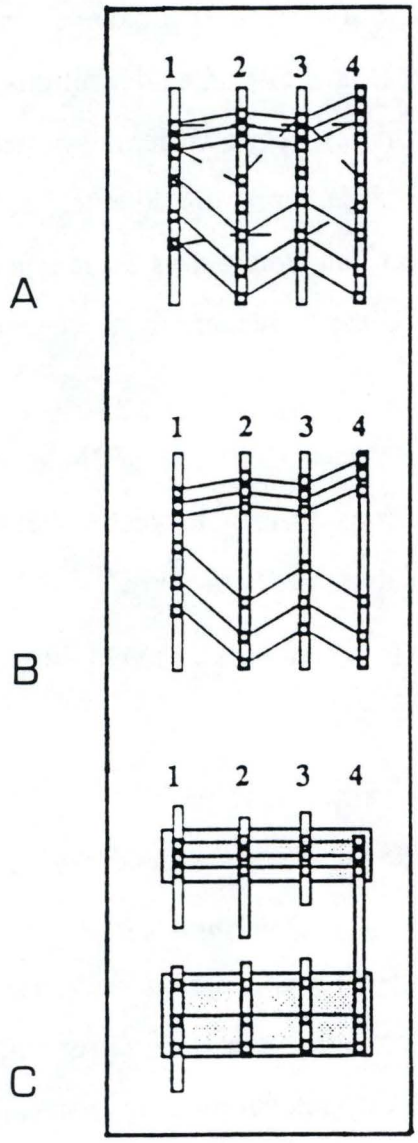


Figure I.54 : Alignement final des séquences :

a. Les appariements significatifs sont présentés par des lignes continues, le « bruit de fond » par des lignes discontinues.

b. Les appariements significatifs sont sélectionnés.

c. Les espaces sont introduits de façon à aligner les régions concernées.

I.5.4. Caractéristique d'un alignement de protéine.

La réalisation d'un alignement de séquences permet de mettre en évidence les régions suffisamment similaires pour prédire l'existence de régions conservées d'un point de vue structural. Un alignement de structures, quant à lui, permet de mettre en correspondance des segments qui sont tous superposables deux à deux. Dans les deux cas, les régions les plus similaires sont alignées de façon prioritaire et pour ce faire, les séquences sont éventuellement interrompues par des espaces. Ces espaces peuvent être courts, par exemple ceux qui témoignent de l'insertion ou de la disparition de quelques résidus au cours de l'évolution. Ils peuvent aussi être très longs, lorsque l'une des séquences étudiées ne correspond qu'à une portion de l'autre. L'insertion de ces espaces entraîne un décalage progressif des résidus d'une séquence par rapport à l'autre. Notons qu'ici, contrairement à bon nombre d'algorithme d'alignement multiple, la création d'espaces n'influence pas la mesure de la similarité entre protéine.

Les régions similaires sont constituées par le regroupement des segments similaires d'une séquence à l'autre dans un certain voisinage. Il est important de se rendre compte que dans un alignement, les régions similaires sont généralement situées entre des régions variables. Les régions similaires sont mises en évidence dans des boîtes qui englobent les appariements complets définis préalablement.

Par définition, des appariements complets sont significatifs s'ils peuvent être disposés simultanément dans l'alignement (Fig. I.54). Les appariements complets qui chevauchent les appariements significatifs et qui ne peuvent être alignés en respectant les espaces définis préalablement constituent le bruit de fond. Un ensemble d'appariements complets associés simultanément selon un ensemble d'espaces forme une boîte. La taille d'une boîte est définie par le nombre de résidus par séquence qu'elle inclut. Dans un alignement, la présence de boîtes indique que les protéines partagent des similarités communes, au moins sur une bonne partie de leur longueur.

I.6. But du travail

De nombreuses méthodes pour l'alignement de structures protéiques sont actuellement disponibles. Mais, parmi celles-ci, peu permettent de réaliser des alignements multiples simultanés en un court laps de temps.

La méthode SHOEBOX, basée sur la comparaison de forme des segments structuraux est une méthode rapide, automatique et peu exigeante au niveau du matériel. Elle est donc particulièrement bien adaptée pour une étude systématique des banques de structures. De plus, contrairement à de nombreuses autres méthodes comme la RMS, SHOEBOX permet de prendre en compte les atomes des chaînes latérales. Cette approche semble donc intéressante.

Les expériences réalisées dans le cadre de ce travail consistent en une évaluation de comportement de SHOEBOX lors de la comparaison de structures. En outre, il nous paraît plus intéressant de nous attarder sur les problèmes rencontrés par la méthode au cours de son utilisation afin d'essayer de déterminer ses limites et ses qualités.

II. MATÉRIEL ET MÉTHODES.

II.1. Matériel.

II.1.1. Support informatique.

Ce travail a nécessité l'utilisation de trois types d'ordinateurs :

- un Vax 6620 fonctionnant sous le système d'exploitation VMS. Les programmes du logiciel MATCH-BOX (Depiereux et Feytmans, 1991, 1992, 1994) tournent sur ce Vax et ont été développés en Fortran 77.
- un Silicon Graphics Iris Personal Computer 4D/20G travaillant avec le système d'exploitation Unix version V. Cette machine permet l'utilisation des logiciels de Biosym technologies (San Diego) tels que INSIGHT et HOMOLOGY.
- un Macintosh permet l'édition des résultats et la mise en graphique essentiellement via le programme EXCEL v4.0.

II.1.2. Banque de structures.

Les structures utilisées dans ce travail sont extraites de la Protein Data Bank (PDB) du Brookhaven National Laboratory (Cambridge, USA). Elle contient les coordonnées cartésiennes des atomes des protéines de conformation déterminées par cristallographie (Bernstein, 1977). Chaque structure de cette banque reçoit un nom de code.

En tout, 8 structures de flavoprotéines furent extraites de cette banque sur base d'un seul critère, leur utilisation comme cofacteur du $FADH_2$. Il s'agit du flavocytochrome β2 (FCB), de la cholestérol oxydase (COX), de la ferredoxine (FNR), de la glycolate oxydase (GOX), de la p-hydroxybenzoate hydroxylase (PHH), de la thiorédoxine réductase (TRB), de la trypanothione réductase (TPR) et de la gluthatione réductase (GRS).

II.2. Méthodes.

Comme nous l'avons déjà signalé, les alignements de protéines (point de vue séquence ou structure) occupent dans la biologie actuelle une place très importante. Dès qu'une nouvelle protéine est séquencée la comparaison avec d'autres séquences connues permet de déduire de nombreuses informations structurelles et fonctionnelles. De la même façon, l'étude d'une protéine nouvellement résolue passe par une étape de classification; la ranger dans une famille permet une caractérisation plus rapide.

Toute méthode d'alignement est basée sur une mesure de similarité ou de distance; la conversion entre les deux est très facile à réaliser. Tout au long de ce travail, nous utiliserons un logiciel, Match-Box, permettant l'alignement de séquences et de structures et cela en utilisant trois notions de distance :

- une matrice de distance définie *a priori* et permettant la comparaison de l'information contenue dans la séquence. Quatre matrices sont utilisées : une basée sur des distances physico-chimiques, les trois autres basées sur des substitutions observées dans des alignements de séquences ou de structures,
- la distance physique moyenne minimum entre les atomes constituant le squelette des fragments de structures. Cette mesure appelée « Root Mean Square » (RMS) est très utile mais ne tient pas compte de la forme des fragments comparés,

Square » (RMS) est très utile mais ne tient pas compte de la forme des fragments comparés,

- une nouvelle mesure tenant compte de la différence de taille et de forme des fragments comparés, puisque la mesure du RMS s'intéresse uniquement à la position des atomes comparés.

La combinaison de ces trois mesures permet un alignement de structures basé sur des informations de différents types, à la fois structurelles et séquentielles. Ces trois mesures de distances sont utilisées ensemble ou séparément par le programme MATCH-BOX.

Les point suivants visent à présenter d'une façon plus « pratique » les différents outils utilisés tout au long de ce travail

II.2.1 Superposition de structures.

Les programmes utilisés pour réaliser les alignements structuraux qui nous serviront de référence sont INSIGHT et HOMOLOGY (Biosym, San Diego). Ces programmes ne permettent pas de superposer plus de deux structures simultanément. Les comparaisons se font donc deux à deux.

II.2.1.1. INSIGHT.

INSIGHT réalise des modèles tridimensionnels à partir des coordonnées cristallographiques de PDB. Il sert d'interface graphique entre la banque de données, l'utilisateur et les modules de calcul comme HOMOLOGY. Les modèles peuvent être comparés et alignés à l'aide du module HOMOLOGY.

II.2.1.2. HOMOLOGY.

HOMOLOGY est un logiciel permettant l'alignement de portions de deux structures et de leurs séquences.

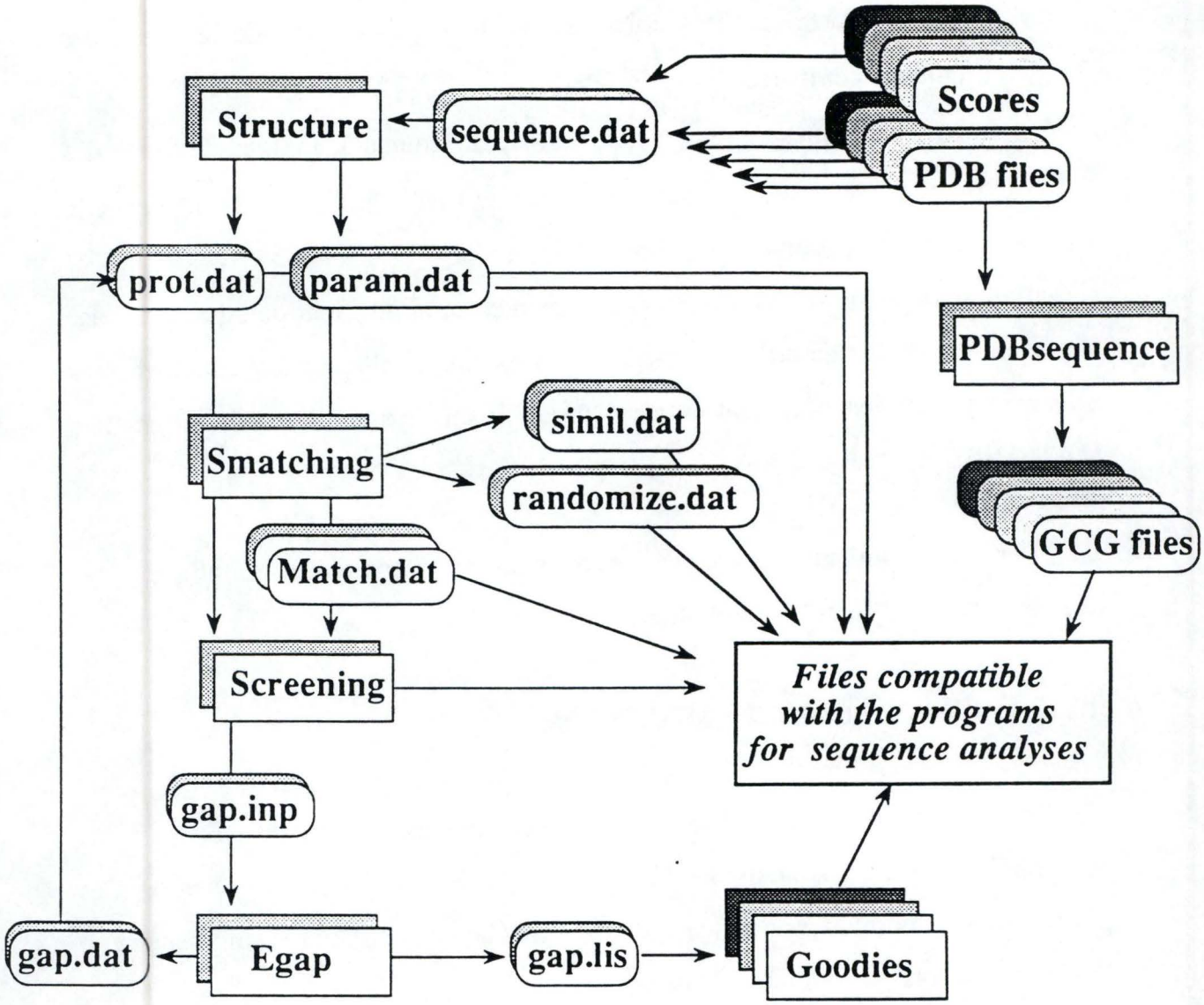


Figure II.1 : Schéma général de la réalisation d'un alignement de structure par le programme MATCH-BOX.

La superposition de structure est simple : la distance moyenne (Å) entre deux segments est calculée par la méthode du « Root Mean Square » (point 1.4.2). Cette distance est proportionnelle à la différence de forme qui existe entre les deux segments. Plus la distance est petite, plus les structures des deux segments de structures se ressemblent. La limite à partir de laquelle deux segments sont considérés comme étant de structure similaire est en général fixée à 1Å. Cette limite est arbitraire et peut être modifiée (Unger et al, 1989).

II.2.2. Le programme MATCH-BOX pour l'alignement de protéines.

Ce programme développé par Depiereux et Feytmans (1991,1992) a été présenté au point 1.5 de l'introduction. Ce qui suit vise à évoquer les différentes procédures du programme utilisées au cours de ce travail (Fig. II.1).

II.2.2.1. Fichiers de départ.

Deux types de fichiers sont nécessaires pour commencer la procédure d'alignement de protéines : d'une part ceux contenant les coordonnées de tous les atomes de chaque protéine (fichiers PDB), et d'autre part la matrice de score utilisée pour une comparaison éventuelle des séquences. Les noms de ces fichiers doivent être indiqué dans le fichier sequence.dat pour signaler à MATCH-BOX ceux qu'il doit utiliser.

La procédure PDBSEQUENCE est chargée de vérifier les fichiers PDB. La (les) séquence(s) incluse(s) dans ces fichiers est(sont) convertie(s) en fichiers séparés au format GCG (les différentes séquences d'un même fichier PDB sont identifiées par une extension). Si plusieurs séquences sont contenues dans le même fichier PDB, celle à utiliser doit être précisée, dans sequence.dat, par son extension.

La procédure STRUCTURE permet d'extraire des fichiers PDB les coordonnées XYZ et décrire les fichiers de données utilisés pour l'alignement proprement dit : les fichiers prot.dat et param.dat.

Le fichier prot.dat contient la liste originale de tous les résidus des séquences accolées dans l'ordre de leur citation dans sequence.dat. Pour chacun, sa position dans la séquence (en tenant compte des gaps, des shifts ou des boîtes gelées éventuels) et les coordonnées XYZ de ses atomes sont précisées.

Le fichier param.dat, comme son nom l'indique, contient la liste de tous les paramètres utilisés pour l'alignement : la longueur de la fenêtre, le nombre d'atomes pris en compte par SHOEBOX et par RMS, la largeur de balayage, le nombre d'identités requis et les différents cutoffs utilisés (la valeur 9999 inactive la méthode correspondante).

II.2.2.2. Procédures SMATCHING, SCREENING et EGAP.

La procédure SMATCHING permet d'effectuer la sélection des appariements de segments en fonction des critères contenus dans param.dat. Le résultat est contenu dans le fichier match.dat.

SCREENING réalise lui l'étape suivante qui consiste à trier les appariements de façon à former des appariements significatifs.

Enfin, les boîtes sélectionnées sont disposées de façon optimale dans un alignement par le programme EGAP.

Les résultats stockés dans le fichier gap.lis peuvent être, dans un dernier temps, mis en forme pour être présenté à l'écran (TY80) ou éditer en format Excel (TYXL).

II.2.2.3. Procédures RANDOMIZE et DISTRIB.

Le programme RANDOMIZE réalise, dans un premier temps, le même travail que SMATCHING à savoir qu'il réalise le balayage et recherche les appariements

répondant aux critères de distances définis dans param.dat. Dans ce cas, les distances entre segments ne sont calculées que sur base de la matrice de score utilisée. Ensuite la distribution de fréquence de ces distances est calculée.

Dans un deuxième temps, les séquences sont mélangées (d'où l'impossibilité d'utiliser ce programme avec les structures) et le même travail est recommencé.

La procédure DISTRIB recalcule les distributions de fréquences cumulées pour n'importe quel set de protéines issues du set initial. Ces distributions (randomisée et non randomisée) sont stockées dans le fichier distrib.lis et peuvent être portées en graphique et utilisées pour déterminer le cutoff à utiliser.

Ce processus de détermination du cutoff doit donc être recommencée lors d'un changement de la matrice de score ou de la longueur de la fenêtre initiale.

II.2.2.4. Procédure FACTOR.

Le lancement de SMATCHING ou de RANDOMIZE entraîne l'écriture d'un fichier simil.dat. Ce fichier contient une matrice de similarité n×n (n étant le nombre de protéines utilisées) dont chaque élément (i,j) représente le nombre de fenêtres initiales de la séquence j appariées à une fenêtre mobile de la séquence i. Le programme FACTOR va, à partir de cette matrice, réaliser une analyse en composantes principales permettant une classification des protéines.

Cette matrice S est asymétrique et va, dans un premier temps, être réduite en une matrice R symétrique où chaque élément r_{ij} est calculé par :

$$r_{ij} = r_{ji} = \frac{\min(S_{ij}, S_{ji})}{\sqrt{S_{ii} . S_{jj}}}$$

Dans un second temps, l'analyse en coordonnées principales proprement dite est réalisée et permet l'écriture d'un fichier factor.lis où chacune des colonnes représente les coefficients des vecteurs propres classés par ordre décroissant de leurs valeurs propres.

	pSCR	SCR
A	A L **G** A W D E Y G L **A** T W D E G	A L **G** A W D E Y G L **A** T W D E G
B	A L **G** A W D E Y G L **A** T W D E G	A L **G** A W D E Y G L **A** T W D E G
C	A L G A W D E **Y** G L A T W D E **G**	A L G A W D E **Y** G L A T W D E **G**
D	A **L** G A W D E Y G **L** A T W D E G	A **L** G A W D E Y G **L** A T W D E G

Figure II.2 : Différentes situations observées lors de la comparaison d'alignements hypothétiques (à gauche) avec l'alignement optimal (à droite)

a. les résidus G et A sont dans la pSCR et la SCR et sont correctement alignés.

b. les résidus G et A sont dans la pSCR et la SCR mais ne sont pas correctement alignés.

c. les résidus Y et G se retrouvent dans la pSCR mais ne sont pas dans la SCR.

d. les résidus L et L ne se retrouvent pas dans la pSCR.

vecteurs propres du fichier factor.lis) sur deux axes, X et Y, représentant chacun une dimension de l'espace recalculé.

II.2.2.5. Procédure SSCANNING.

Ce programme permet d'obtenir un fichier scan.lis donnant pour chaque fenêtre initiale les distances mesurées le long du balayage (si le balayage est de m, chaque fenêtre initiale est comparée à 2m+1 fenêtres mobiles). La distance est calculée pour chaque méthode (DFK, RMS et SHOEBOX) et pour la combinaison des trois.

Une mise en graphique de ces balayages permet d'observer de plus près le comportement des mesures de distances et, en particulier, de voir si l'appariement correct est bien trouvé.

II.2.3. Qualité d'un alignement.

Pour évaluer la qualité des alignements réalisés par MATCH-BOX, il est nécessaire de les comparer à l'alignement de référence. Différents cas de figures peuvent être relevés :

- 2 résidus se retrouvent dans les pSCR et les SCR et sont correctement alignés (fig. II.2.a),
- 2 résidus se retrouvent dans les pSCR et dans les SCR mais ne sont pas correctement alignés (fig. II.2.b),
- 2 résidus se retrouvent dans les pSCR mais ne sont pas dans les SCR (fig. II.2.c),
- 2 résidus ne se retrouvent pas dans les pSCR alors qu'ils font partie des SCR (fig. II.2.d).

Le premier cas représente un succès, les deux suivants des échecs et le quatrième une sous-estimation.

Pour critiquer chaque alignement, deux valeurs sont calculées :

- le pourcentage de succès : il représente le rapport entre le nombre de résidus correctement alignés et le nombre de résidus inclus dans les SCR (c'est-à-dire qui idéalement auraient dû être alignés),
- le pourcentage d'échecs : il représente le rapport entre le nombre d'échecs (de 2 types) et la somme des échecs et des succès (c'est-à-dire le nombre total d'appariements détectés grâce à SHOEBOX).

Ces deux valeurs permettent de se faire une opinion correcte de la qualité des appariements. La première indique dans quelle mesure MATCH-BOX permet de détecter toutes les SCR alors que la seconde signale si la sélection de ces appariements corrects n'est pas associée à un nombre trop élevé d'échecs.

Une troisième mesure permet d'évaluer le bruit de fond : le pourcentage de paires sélectionnées par MATCH-BOX qui seront consistantes avec l'alignement optimal.

Remarques : le pourcentage de succès et le pourcentage d'échecs sont 2 rapports où les dénominateurs sont différents (le nombre de résidus dans les SCR pour le premier et le nombre de résidus dans le pSCR pour le second) La somme des deux ne représente donc pas 100%

III. Résultats

Dans cette partie du travail, seront présentés les résultats obtenus lors de l'évaluation de l'efficacité de la méthode SHOEBOX. Ils seront regroupés en trois points. Dans un premier temps, le groupe de protéines de départ a été classifié et les alignements de référence ont été réalisés en superposant les structures. Dans un second temps, la technique de comparaison des formes de segments a été testée seule en tenant compte ou non des coordonnées des atomes des chaînes latérales. Enfin, dans une dernière partie, le comportement de SHOEBOX a été évalué lorsqu'il était utilisé avec d'autres mesures de similarité. Cela a permis de tirer des conclusions quant à l'apport de chaque méthode et, en particulier, quant aux capacités de SHOEBOX.

Remarque : Dans la suite de ce travail, de nombreux cas seront pris en exemple pour illustrer les résultats obtenus, ils ont tous été choisis parce qu'ils étaient représentatifs des situations observées. Cette sélection était nécessaire, il était impossible de présenter en particulier chaque cas traité.

III.1. Réalisation de l'alignement de référence.

Pour tester l'efficacité d'une méthode d'alignement quelle qu'elle soit, il est nécessaire de pouvoir la comparer à un alignement optimal, un alignement de référence.

Cette comparaison permettra de déterminer si les pSCR (régions prédites comme étant structurellement conservées) trouvées correspondent bien aux SCR (régions structurellement conservées) attendues.

	Largeur de la fenêtre	Mesure(s) de similarité utilisée(s)	cutoff(s) utilisé(s)
1	7 résidus	DFK	400
2			200
3		SHOEBOX	100000
4			50000
5		DFK et SHOEBOX	400 et 100000
6			300 et 75000
7	9 résidus	DFK	600
8			400
9		SHOEBOX	150000
10			100000
11		DFK et SHOEBOX	600 et 150000
12			400 et 100000
13	11 résidus	DFK	800
14			600
15		SHOEBOX	200000
16			150000
17		DFK et SHOEBOX	800 et 100000
18			600 et 150000

Table III.1 : Différents paramètres utilisés lors des analyses factorielles. Ces 18 analyses ont été réalisées pour deux valeurs différentes de balayage

La première difficulté réside dans le choix d'un groupe de structures à utiliser. Nous avons dû trouver des protéines possédant une structure suffisamment proche mais ne présentant pas de similarité de séquences trop importantes. En effet, il existe de nombreuses méthodes grâce auxquelles il est possible d'aligner correctement des séquences proches mais qui, lorsque la similarité diminue, ne sont plus si efficaces. C'est dans ce cas précis que le programme SHOEBOX peut être utile.

Une seconde difficulté apparait lorsqu'il faut superposer les différentes structures. En effet, il n'est pas possible de réaliser cette opération avec le programme HOMOLOGY seul et la plus grande partie des alignements doit donc être construite à l'oeil.

III.1.1. Analyse factorielle.

Cette première étape consiste en la réalisation d'une classification préalable au sein de notre groupe de structures. L'information ainsi obtenue doit, en outre, nous faciliter la tâche pour la phase ultérieure de superposition de structures.

Pour ce faire, la procédure FACTOR a été utilisée. Rappelons qu'elle permet la réalisation d'une analyse en coordonnées principales. Lors de ces expériences, il est possible de modifier divers paramètres : la taille de la fenêtre (w), l'amplitude du balayage (m), la (ou les) mesure(s) de similarité utilisée(s) et les cutoffs fixés pour ces dernières (table III.1).

Le fait de faire varier une série de paramètres doit nous permettre d'observer leur influence sur l'analyse. Intuitivement, on peut s'attendre aux résultats suivants :

- un cutoff plus strict va permettre la mise en évidence des ressemblances plus fortes entre les protéines.
- l'utilisation d'une fenêtre plus grande sera plus discriminatoire. En effet, il est plus probable d'avoir une ressemblance due au hasard entre deux fenêtres de 7 résidus qu'entre deux de 11 résidus.

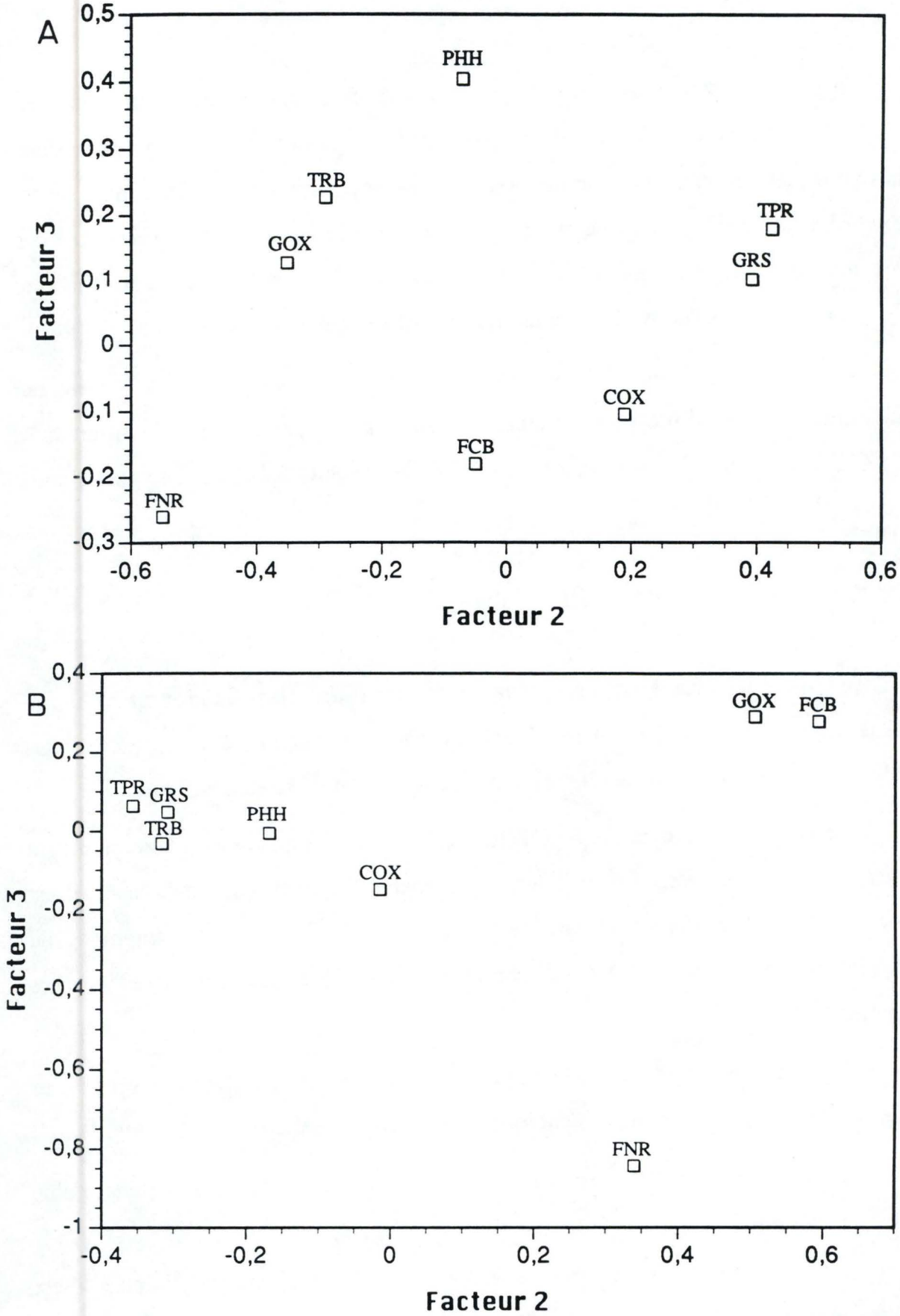


Figure III.1 : Analyse factorielle réalisée avec les méthodes SHOEBOX (cutoff 100 000) et DFK (cutoff 400) pour une fenêtre de 7 résidus.

a. présentation du plan des facteurs 2 et 3 pour un balayage de 25.

b. présentation du plan des facteurs 2 et 3 pour un balayage de 500.

- un balayage plus petit permet l'établissement de groupes plus significatifs. Prenons le cas extrême où le balayage est de 0. Si l'analyse factorielle met en évidence l'existence d'un groupe, il est évident que les séquences le constituant sont très proches.

La lecture des différents graphes nous permet déjà de conclure à certaines ressemblances ou dissemblances. En tout, 36 analyses factorielles ont été réalisées, ce qui correspond à un total de 108 graphes (chaque analyse donnant les coefficients des trois premiers vecteurs propres, un graphe étant réalisé pour toutes les combinaisons possibles de ces trois coefficients : 1*2, 2*3 et 1*3). Ces 108 graphes furent analysés et parmi ceux-ci, les deux plus significatifs sont choisis pour illustrer les conclusions présentées ci-dessous(fig. III.1).

- TPR et GRS semblent très proches. Quels que soient les paramètres utilisés, on les retrouve dans la même région du graphe, y compris dans les conditions citées ci-dessus où les cutoffs sont les plus stricts, la fenêtre est la plus grande et le balayage est le plus petit.
- GOX et FCB sont très rapprochées à condition que le balayage soit suffisamment grand.
- GRS, TPR, TRB, PHH forment un groupe plus ou moins compact si la même condition est respectée.
- COX occupe, sur chaque graphe, une position isolée. Elle semble pourtant être plus proche du groupe précédent.

III.1.2. Superposition des structures.

Les premières structures superposées sont celles de TPR et GRS. Le programme HOMOLOGY permet la réalisation d'une ébauche d'alignement. En effet, avec celui-ci, on ne peut retrouver que quelques zones où les ressemblances structurelles sont très importantes. Il est donc nécessaire, dans un second temps, de parcourir la totalité des structures et de détecter sur l'écran les zones à faire correspondre. L'alignement de référence ainsi obtenu semble être proche de l'alignement optimal.

Figure III.2 : Superposition des structures TPR (en noir) et GRS (en gris).

```
                10        20        30        40        50        60        70        80
            +         +         +         +         +         +         +         +
1 SRAYDLVVIGAGSGGLEAGWNAASLHKKRVAVIDLQKHHGPPHYAALGGTCVNVGCVPKKLMVTGANYMDTIRESAGFGW
2 VASYDYLVIGGGSGGLASARRAAELGARAAVVESHK---------LGGTCVNVGCVPKKVMWNTAVHSEFMHDHADYG-

                90       100       110       120       130       140       150       160
            +         +         +         +         +         +         +         +
ELDRESVRPNWKALIAAKNKAVSGINDSYEGMFADTEGLTFHQGFGALQDNHTVLVRESADPNSAVLETLDTEYILLATG
-FPSCEGKFNWRVIKEKRDAYVSRLNAIYQNNLTK-SHIEIIRGHAAF-------TSDPKPTIEVSGKKYTAPHILIATG

               170       180       190       200       210       220       230       240
            +         +         +         +         +         +         +         +
SWPQHLG---IEGDDLCITSNEAFYLDEAPKRALCVGGGYISIEFAGIFNAYKARGGQVDLAYRGDMILRGFDSELRKQL
GMPSTPHESQIPGASLGITSDGFFQLEELPGRSVIVGAGYIAVEMAGILSA---LGSKTSLMIRHDKVLRSFDSMISTNC

               250       260       270       280       290       300       310       320
            +         +         +         +         +         +         +         +
TEQLRANGINVRTHENPAKVTKNADGTRHVVF--------ESGAEADYDVVMLAIGRVPRSQTLQLEKAGVEVAKNGAIK
TEELENAGVEVLKFSQVKEVKKTLSGLEVSMVTAVPGRLPVMTMIPDVDCLLWAIGRVPNTKDLSLNKLGIQTDDKGHII

               330       340       350       360       370       380       390       400
            +         +         +         +         +         +         +         +
VDAYSKTNVDNIYAIGDVTDRVMLTPVAINEGAAFVDTVFANK-PRATDHTKVACAVFSIPPMGVCGYVEEDAAKKY--D
VDEFQNTNVKGIYAVGDVCGKALLTPVAIAAGRKLAHRLFEYKEDSKLDYNNIPTVVFSHPPIGTVGLTEDEAIHKYGIE

               410       420       430       440       450       460       470       480
            +         +         +         +         +         +         +         +
QVAVYESSFTPLMHNISGSTYKKFMVRIVTNHADGEVLGVHMLGDSSPEIIQSVAICLKMGAKISDVYNTIGVHPTSAEE
NVKTYSTSFTP-MYHAVTKRKTKCVMKMVCANKEEKVVGIHMQGLGCDEMLQGFAVAVKMGATKADFDNTVAIHPTSSEE

               490       500       510       520       530       540       550       560
            +         +         +         +         +         +         +         +
LCSMRTPAYFYEKGKR
LVTLR
```

Figure III.3 : Alignement de référence réalisé pour les séquences de TPR (1) et GRS (2).

La seule incertitude réside dans les effets de bords : le choix de la fin d'un box est toujours plus ou moins arbitraire.

Dans cet alignement, 389 résidus font partie des SCR pour un total de 490 résidus pour la plus grande séquence (TPR). Le pourcentage d'identité entre les deux séquences est de 34%. Le cas de ces deux protéines est très intéressant puisque ne possédant pas un pourcentage d'identité trop grand, elles présentent pourtant de très importantes ressemblances au point de vue de leur structure (fig. III.2 et III.3). Ces ressemblances avaient pu être détectées lors de l'analyse factorielle.

Deux autres alignements sont construits : un, regroupant 5 structures; l'autre, 2 structures. Mais, dans ces deux cas, l'utilisation du programme HOMOLOGY ne permet pas de détecter les zones à superposer. Il faut donc réaliser cette opération par observation des structures sur l'écran : les parcourir segment par segment et essayer de superposer les zones se ressemblant.

Le premier alignement a permis de mettre en évidence un domaine partagé par les structures COX, GRS, PHH, TPR et TRB. Le domaine, constitué de 5 segments, totalise 68 résidus dans des SCR. Pour parvenir à superposer ces 5 régions, il est nécessaire d'introduire dans les séquences GRS, PHH, TPR et TRB des gaps d'une centaine de résidus juste avant le 3e box. Les éléments nécessaires au domaine sont conservés contrairement aux loops les reliant. Ce groupe avait pu être repéré par l'analyse factorielle (m = 500). Seule COX est plus éloignée. Ce phénomène peut s'expliquer par le fait qu'elle semble la protéine la plus dissemblable du groupe.

Le second alignement implique les protéines FCB et GOX. Sur les 369 résidus de GOX, 290 peuvent être superposés à des résidus de FCB. Pour ce faire, il est nécessaire d'ajouter un gap de 119 résidus au début de la séquence GOX. Ceci explique le fait que la ressemblance entre ces deux protéines n'a pu être observée lors de l'analyse factorielle que pour un balayage de 500.

Ces trois alignements procurent trois terrains différents d'investigation pour l'évaluation des capacités de la méthode SHOEBOX.

	Largeur de la fenêtre	cutoffs utilisés (DFK, SHOEBOX, RMS)	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400, 100000, 1	0	98	10
2		300, 75000, 1	0	98	9
3		200, 50000, 1	0	94	6
4	9 résidus	600, 150000, 1	0	98	10
5		500, 125000, 1	0	98	9
6		400, 100000, 1	0	99	7
7	11 résidus	800, 200000, 1	0	98	7
8		700, 175000, 1	0	98	7
9		600, 150000, 1	0	97	6

Table III.2 : Résultats obtenus pour les différents paramètres utilisés pour la vérification de l'alignement de référence. Le balayage est de 0 et le fichier de données contient les gaps trouvé dans HOMOLOGY.

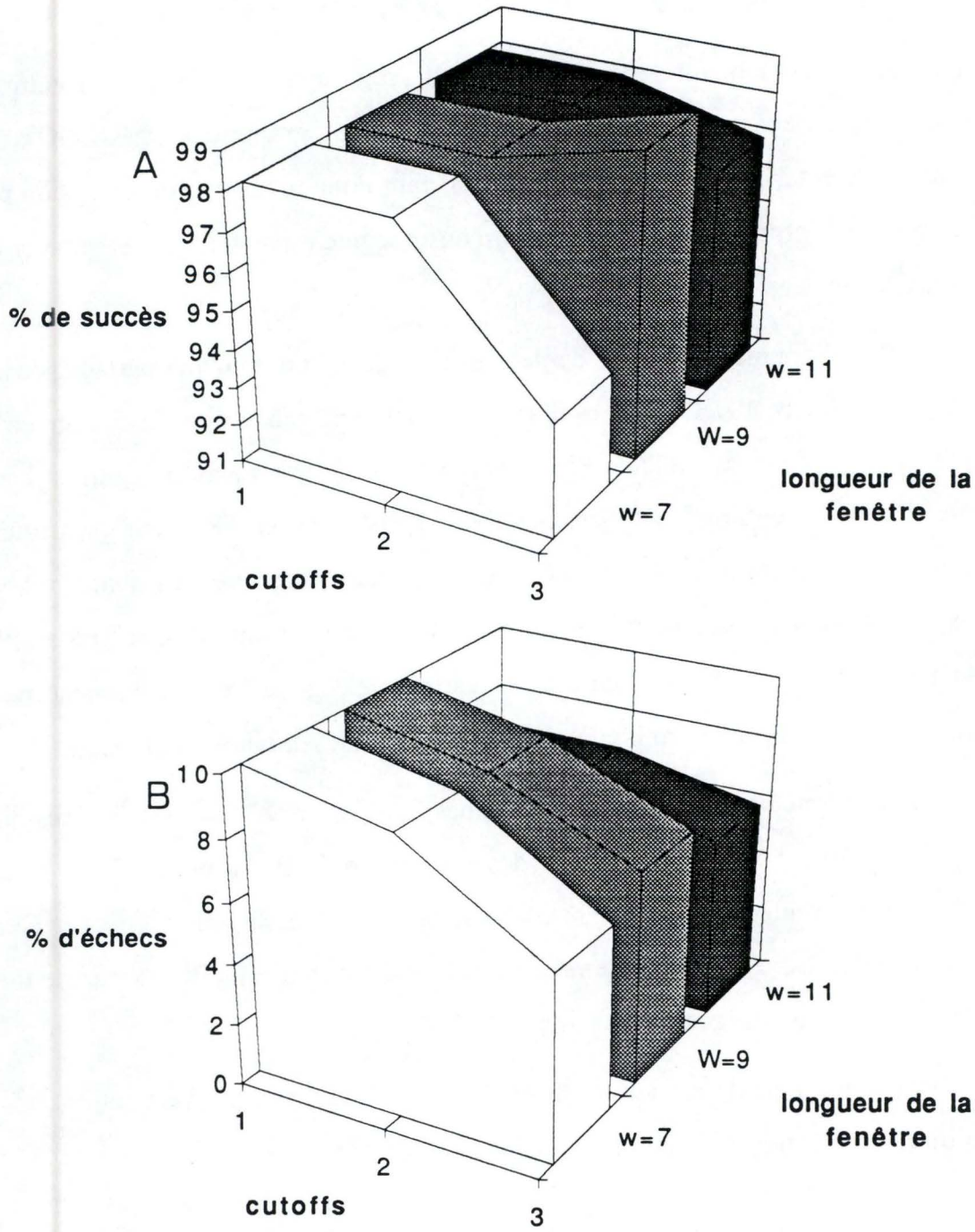


Figure III.4 : pourcentage de succès (a) et d'échecs (b) pour la vérification de l'alignement de référence. Pour chaque longueur de fenêtres, trois jeux de cutoffs sont utilisés (le jeu 3 étant le plus strict).

```
                10        20        30        40        50        60        70        80
                +         +         +         +         +         +         +         +
1 srayd l vvigagsggleagwnaasLHkkrvavidlqKHHGPPHYAAlggtcvnvgcvpkklmvtganymdtiresagfgW
2 vasydylviggggsgglasarraaeL-garaavveshK---------lggtcvnvgcvpkkvmwntavhsefmhdhadyg-

                90       100       110       120       130       140       150       160
                +         +         +         +         +         +         +         +
ELDRESvrpnwkaliaaknkavsgindsyegmfadtEGltfhqgfgalqDNHTVLVRESADPNSAVLEtldteyillatg
-FPSCEgkfnwrvikekrdayvsrlnaiyqnnltk-sHieiirghaaf-------tSDPKPTIEVSGKkytaphiliatg

               170       180       190       200       210       220       230       240
                +         +         +         +         +         +         +         +
swpqhlg---iegddlcitsneafyldeapkralcvgggyisiefagifnaykARGGqvdlayrgdmilrgfdselrkql
gmpstphesQipgaslgitsdgffqleelpgrsvivgagyiavemagilsa---lgSktslmirhdkvlrsfdsmistnc

               250       260       270       280       290       300       310       320
                +         +         +         +         +         +         +         +
teqlrangi nvrthenpakvtknadgtrHVVF--------ESGAEADydvvmlaigrvprsqtlqlekagvevakngaik
teelenagvevlkfsqvkevkktlsgleVSMVTAVPGRLPVMTMIPDvdcllwaigrvpntkdlslnklgiqtddkghii

               330       340       350       360       370       380       390       400
                +         +         +         +         +         +         +         +
vdayskt nvdniyaigdvtdrvmltpvainegaafvdtvfank-pratdhtkvacavfsippmgvcgyveedaakky--d
vdefqnt nvkgiyavgdvcgkalltpvaiaagrklahrlfeykeDskldynniptvvfshppigtvgltedeaihkygIE

               410       420       430       440       450       460       470       480
                +         +         +         +         +         +         +         +
qvavyessftplmhniSGSTYkkfmvrivtnhadgevlgvhmlgdsspeiiqsvaiclkmgakisdvyntigvhptsaee
nvktystsftp-myhavTKRKtkcvmkmvcankeekvvgihmqglgcdemlqgfavavkmgatkadfdntvaihptssee

               490       500       510       520       530       540       550       560
                +         +         +         +         +         +         +         +
lcsmRTPAYFYEKGKR
lvtlR-----------
```

Figure III.5 : Alignement pour une fenêtre de 9 résidus avec DFK, SHOEBOX et RMS (cutoffs 400, 100 000 et 1). Le

La suite de ce travail reposera uniquement sur l'alignement entre les protéines GRS et TPR. En effet, il semblait intéressant, outre d'évaluer les posiblités offertes par l'utilisation de la méthode SHOEBOX, de comprendre les résultats qu'elle permet d'obtenir.

III.1.3. Vérification de l'alignement optimal.

Une fois l'alignement de référence construit, et avant d'aller plus loin, il est nécessaire de vérifier que l'utilisation du programme MATCH-BOX permet bien de le retrouver. Pour cela, les 3 méthodes disponibles pour la mesure de la similarité sont utilisées et un alignement où le balayage est nul est réalisé. Les données utilisées pour cette expérience sont directement issues de l'alignement de référence : elles contiennent non seulement les séquences des protéines, mais aussi les gaps trouvés dans HOMOLOGY. Différents alignements sont lancés en faisant varier la longueur des fenêtres et les cutoffs (table III.2)

En comparant les différents résultats, on peut voir que l'alignement, réalisé avec une fenêtre de 9 résidus et des cutoffs de 400 pour DFK, de 1Å pour RMS et de 100 000 pour SHOEBOX, est le plus proche de la référence avec un pourcentage de succès de 99% pour un pourcentage d'échecs de 7% (fig. III.4). Pour tous les cas observés, le bruit de fond est nul puisque le balayage est fixé à 0.

L'alignement élaboré grâce à la méthode MATCH-BOX (fig. III.5), aux effets de bords près, est proche de celui de référence (cela veut dire que cet alignement est acceptable, que les appariements qu'il implique sont corrects; cela ne signifie pas que, à partir d'une situation initiale quelconque, cet alignement sera retrouvé).

III.2. Évaluation de l'efficacité de SHOEBOX seul.

III.2.1. Sens de la distance calculée par SHOEBOX.

Comme nous l'avons vu précédemment (point 1.4.5), la méthode SHOEBOX est basée sur le calcul de la forme de la boîte qui contiendrait, en moyenne, les points représentant les atomes du segment considéré. Trois valeurs – λ_1, λ_2 et λ_3 – correspondent chacune à la plus grande variance qu'il est possible de trouver dans des directions orthogonales. Les racines carrées de ces valeurs, les écarts types, sont proportionnelles aux dimensions de la boîte. λ est exprimé en $Å^2$ et donc $\lambda^{0,5}$ en Å.

Prenons la mesure de la distance entre 2 boîtes 1 et 2 :

$$D_\lambda = \sum_{i=1}^{3} (\lambda_{i1} - \lambda_{i2})^2$$

Elle est exprimée en $Å^4$ et indique un changement de forme plus ou moins important de la forme de la boîte, dans le sens de la longueur, de la largeur et de la hauteur.

La détermination du seuil pour SHOEBOX se fait soit par l'étude de la corrélation avec la distance physico-chimique, soit de manière empirique. Le seuil communément accepté est de $10^5 Å^4$.

III.2.2. Comportement de SHOEBOX lors de la comparaison de la forme du squelette peptidique.

Dans cette première série d'expériences, la mesure de la forme du segment ne fait entrer en ligne de compte que 4 atomes par résidu (N, C_α, C et O). Seules les coordonnées de ces atomes sont donc utilisées, celles des atomes des chaînes latérales sont ignorées. Cette mesure ne se base que sur la conformation du squelette (backbone) protéique.

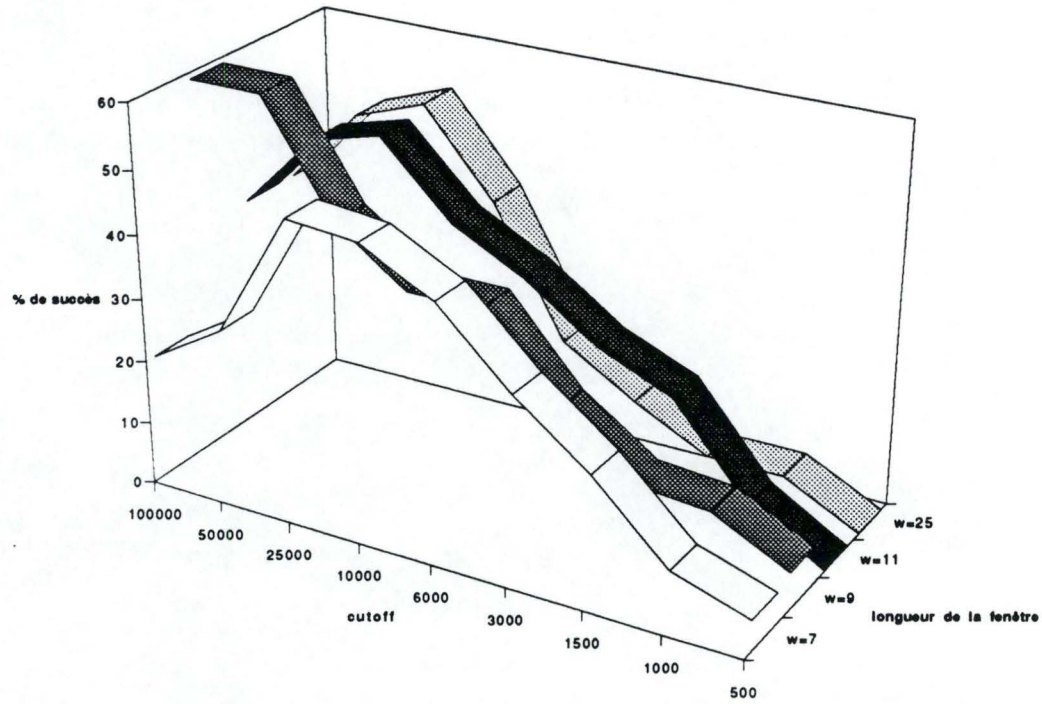
A	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	6	5	62
2	1000	36	9	52
3	1500	58	20	19
4	3000	69	29	34
5	6000	73	40	31
6	10000	73	46	35
7	25000	73	47	43
8	50000	74	27	65
9	100000	76	20	73

B	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	20	6	53
2	1000	47	12	38
3	1500	59	13	36
4	3000	65	21	36
5	6000	68	34	35
6	10000	69	34	38
7	25000	69	42	42
8	50000	69	59	34
9	100000	71	59	38

C	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	30	0	0
2	1000	58	7	53
3	1500	62	21	12
4	3000	65	26	11
5	6000	66	34	29
6	10000	69	40	21
7	25000	69	51	20
8	50000	67	48	25
9	100000	69	36	59

D	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	0	0	0
2	1000	0	6	0
3	1500	0	6	66
4	3000	8	12	50
5	6000	23	19	57
6	10000	22	39	36
7	25000	23	52	50
8	50000	36	48	39
9	100000	45	36	45

Table III.3 : Résultats obtenus lors de l'utilisation de SHOEBOX seul. Les chaînes latérales ne sont pas prises en compte et le balayage est fixé à 25. 4 longueurs de fenêtres sont considérées : 7 (a), 9 (b), 11 (c), et 25 (d).



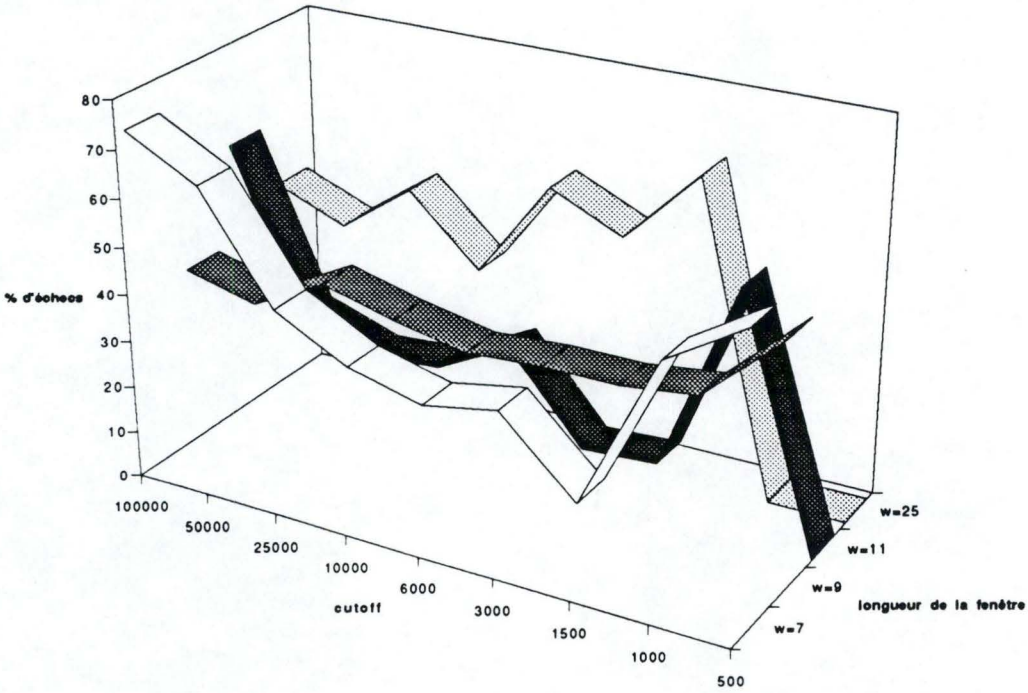


Figure III.6 : Pourcentage de succès et d'échecs pour les alignements réalisés avec SHOEBOX sans les chaînes latérales. 4 longueurs de fenêtres sont utilisées avec 9 valeurs de cutoffs différentes.

III.2.2.1. Alignements.

La première étape de l'évaluation de l'efficacité de SHOEBOX seul consiste à effectuer une série d'alignements en faisant varier les différents paramètres. Ces expériences sont réalisées avec un balayage réduit (m = 25). Dans notre cas, un balayage plus important ne ferait qu'augmenter le bruit de fond.

Différents cutoffs sont considérés allant d'une valeur stricte (500) à une valeur beaucoup plus large (100 000). 4 longueurs différentes de fenêtre sont employées : 7, 9, 11 et 25 (table III.3).

Comme on peut le voir dans la figure III.6, dès les cutoffs très bas, l'utilisation de SHOEBOX entraîne la sélection d'appariements incorrects. Pour toutes les longueurs de fenêtres, ces mêmes erreurs sont commises, avec une exception, celle où le segment de comparaison est de 25 résidus. Mais, même dans ce cas, dès l'augmentation du cutoff, le pourcentage d'échecs augmente. Dans les quatre séries d'alignements réalisées, celle pour laquelle la longueur de fenêtre était fixée à 9 résidus donne les meilleurs résultats avec un pourcentage d'échecs se situant autour des 40%, même pour un cutoff de 100.000. Les alignements où la fenêtre était de 11 résidus donnent également de bons résultats, mais l'augmentation du cutoff se traduit par un pourcentage d'échecs plus important (60% au cutoff 100.000).

On peut aussi remarquer que plus le cutoff est grand, plus le bruit de fond et le pourcentage d'échecs augmentent. Ce phénomène était prévisible. En effet, prenons le cas d'un cutoff très strict, la probabilité de trouver un appariement pour lequel la distance est, par hasard, inférieure à ce cutoff est faible. Dès qu'on augmente la valeur du cutoff, cette probabilité augmente à son tour.

Enfin, on peut voir qu'une augmentation du pourcentage de succès accompagne celle du cutoff. Cette augmentation est tout à fait logique dans la mesure où, en relevant le cutoff, on permet la sélection de plus en plus d'appariements et donc, entre autres, des appariements corrects. Dans certains cas, ce pourcentage de succès passe

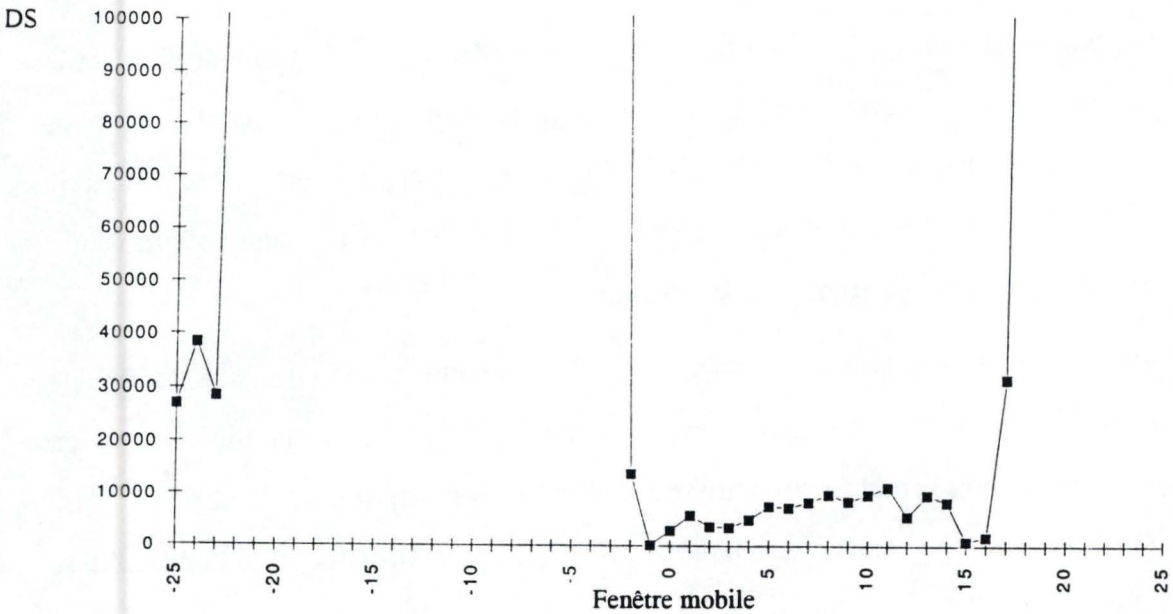


Figure III.7 : Balayage réalisé autour de la fenêtre 90. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX sans les chaînes latérales.

par un maximum puis diminue. Un bruit de fond trop important pourrait expliquer ce phénomène.

III.2.2.2. Comportement de SHOEBOX lors du balayage de séquence.

Pour se faire une idée plus précise et, éventuellement, trouver la raison expliquant les erreurs commises, nous avons regardé plus en détail les distances mesurées. Le fichier scan.lis nous procure pour chaque fenêtre initiale, les distances entre cette dernière et toutes les fenêtres considérées lors du balayage. En mettant ces données en graphe, il est possible de repérer le minimum trouvé et de vérifier s'il correspond bien à celui qu'on attend (représenté par la valeur 0 en abscisse).

Chaque procédure d'alignement implique des centaines de milliers d'appariements dont certains seront correctement sélectionnés alors que d'autres ne le seront pas. Il faut donc, dans un premier temps, parcourir le fichier et observer les résultats obtenus par la méthode SHOEBOX dans chaque cas. Il n'est évidemment pas possible de présenter en détail tous les cas observés, ceux présentés ci-dessous ont été choisis car ils étaient les plus fréquents et les plus typiques.

Les exemples illustrant ce travail sont ceux pour lesquels un mauvais appariement est choisi ou ceux permettant d'essayer de comprendre ses erreurs. Il est évident, comme l'indique les pourcentages d'échecs, que l'utilisation de la méthode SHOEBOX rend possible la détection d'une série d'appariements. Il semblait néanmoins plus constructif de traiter les cas d'erreurs.

Le premier balayage observé est celui réalisé pour la fenêtre initiale numéro 90 (fig. III.7) se situant dans une hélice α. Sur le graphe, on peut voir que les distances mesurées pour les fenêtres appartenant à cette même hélice sont de loin plus faibles que les autres. On se rend aussi compte que par la méthode SHOEBOX, on ne peut pas discerner, dans cette série de fenêtres appartenant à l'hélice, celle qui correspond exactement à la fenêtre initiale. Trois mesures de distances sont inférieures à celle de

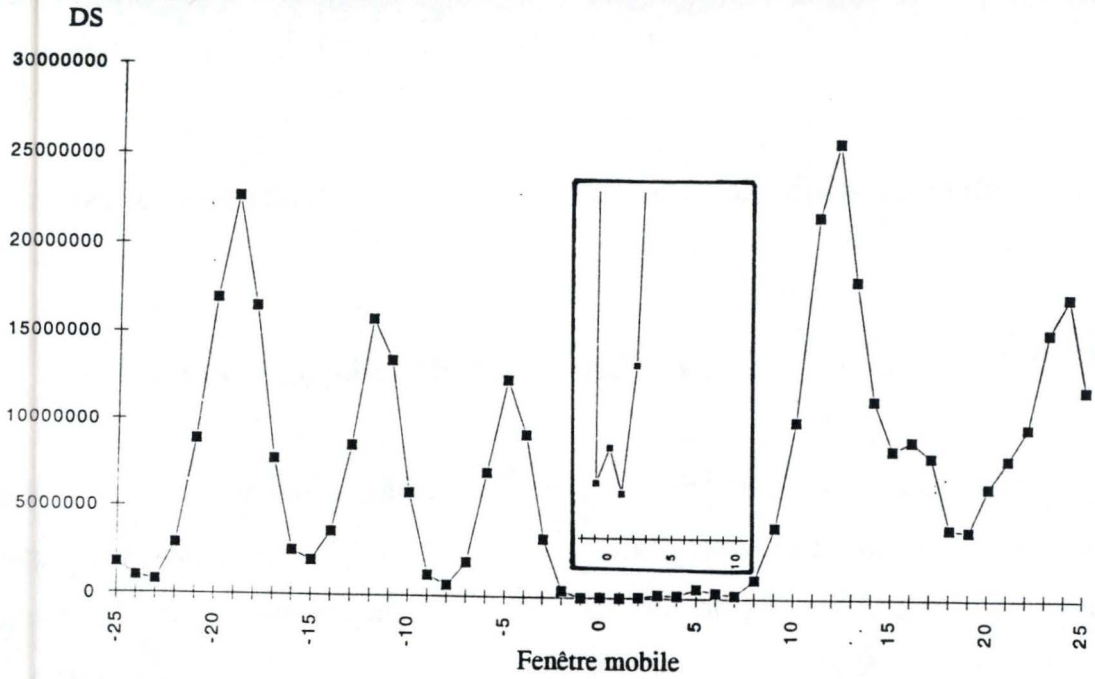


Figure III.8 : Balayage réalisé autour de la fenêtre 154. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX sans les chaînes latérales. L'encadré propose une vue rapprochée permettant de localiser le minimum.

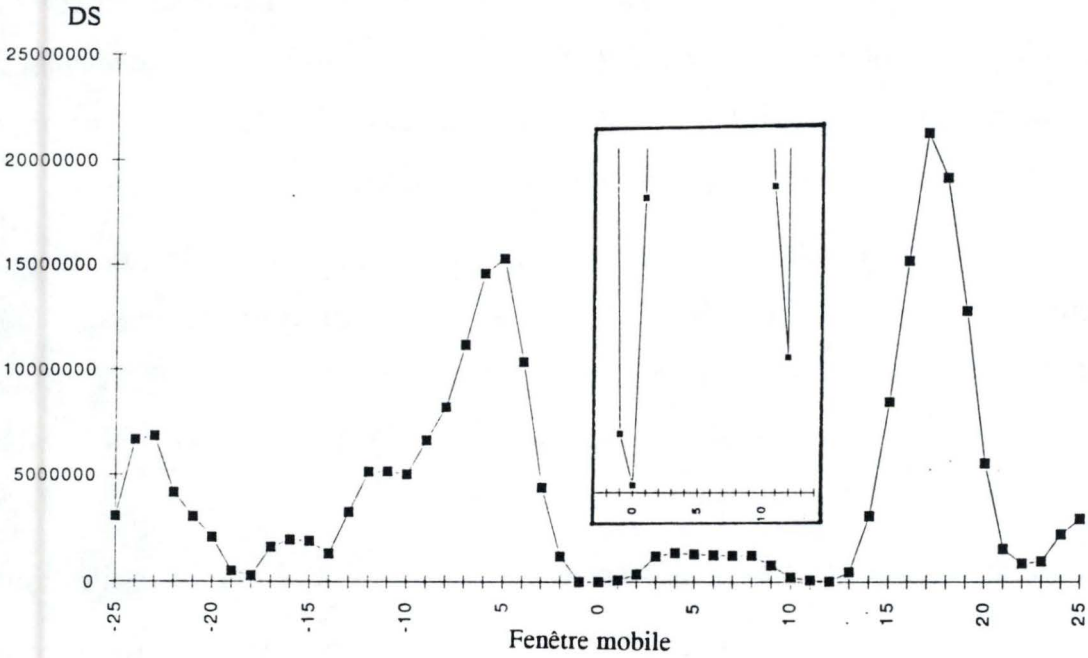


Figure III.9 : Balayage réalisé autour de la fenêtre 193. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX sans les chaînes latérales. L'encadré propose une vue rapprochée permettant de localiser le minimum.

l'appariement correct. D'autres exemples non illustrés montrent que ce n'est pas la seule fois que, face à une telle situation, l'utilisation de SHOEBOX cause des erreurs.

Le second cas choisi est celui de la fenêtre 154 (fig. III.8), se trouvant dans la région d'un plan β. Le graphe nous montre que les fenêtres appartenant aux différents brins composant le feuillet β sont caractérisées par des distances plus faibles. En revanche, en regardant de plus près, on remarque que l'appariement correct n'est toujours pas trouvé. Ce problème est à rapprocher du précédent; dans les deux cas, le programme permet de repèrer les segments ayant la même conformation, mais celui pour laquelle la différence de forme est la plus petite ne correspond pas à l'appariement correct.

Un troisième cas a été pris en exemple (fig. III.9) , non parce qu'il montrait une erreur, mais parce qu'il illustrait le fait que deux segments ayant même structure secondaire sont très proches quant à leur forme. Sur les 9 résidus de la fenêtre 193, 6 appartiennent à une hélice α et 2 à un brin β. Le balayage réalisé avec SHOEBOX permet la détection de l'appariement correct. Il existe pourtant un autre minimum local, 12 résidus plus loin. Si on observe la situation de plus près, l'existence de ce 2e minimum est facilement explicable. Du début de la fenêtre 0 à la fin de la 12, la séquence est la suivante :

V G A G Y A V E M A G I L S A L G S K T

β β – α α α α α α α α α α α α α α – – β β

0 12

Au niveau de la forme, ces segments sont fort semblables. Même si dans un cas les deux premiers résidus sont dans une conformation de brin β et les 6 derniers dans une conformation d'hélice α et que dans l'autre cas la situation est inverse, la boîte contenant ces points aura la même forme. La seule différence résidera dans la distribution des atomes dans cette boîte.

III.2.2.3. Discussion.

Au regard des résultats obtenus avec la méthode SHOEBOX utilisée seule dans les procédures d'alignement et des balayages observés, il est déjà possible de tirer certaines conclusions.

Les alignements nous ont montré que, même à des cutoffs très bas, l'appariement correct n'étaient pas toujours détecté. L'observation de ces mêmes alignements et de différents balayages réalisés avec SHOEBOX nous ont permis de préciser ce qui semblait poser problème.

La comparaison de la forme des segments du squelette protéique permet de déterminer ceux qui ont une même structure secondaire et donc une même conformation. Rappelons que les acides aminés présents dans un élément de structure secondaire ont des valeurs d'angles de torsions propres à cet élément (par exemple, pour l'hélice x, ϕ et ψ sont égaux en moyenne à -60°). Quel que soit le résidu, il aura donc toujours, dans un même élément de structure secondaire, une conformation identique ou du moins fort semblable. Il apparaît donc normal que, si on ne tient pas compte des chaînes latérales, la comparaison des formes de segments n'amène pas toujours à la détection de l'appariement correct : qu'est-ce qui ressemble plus à un segment d'hélice α qu'un autre segment d'hélice α?

Pour tenter de résoudre ce problème, un travail identique au précédent a été réalisé. Mais en plus des atomes du squelette, le calcul de la forme du segment tiendra compte pour chaque résidu, des atomes de la chaîne latérale.

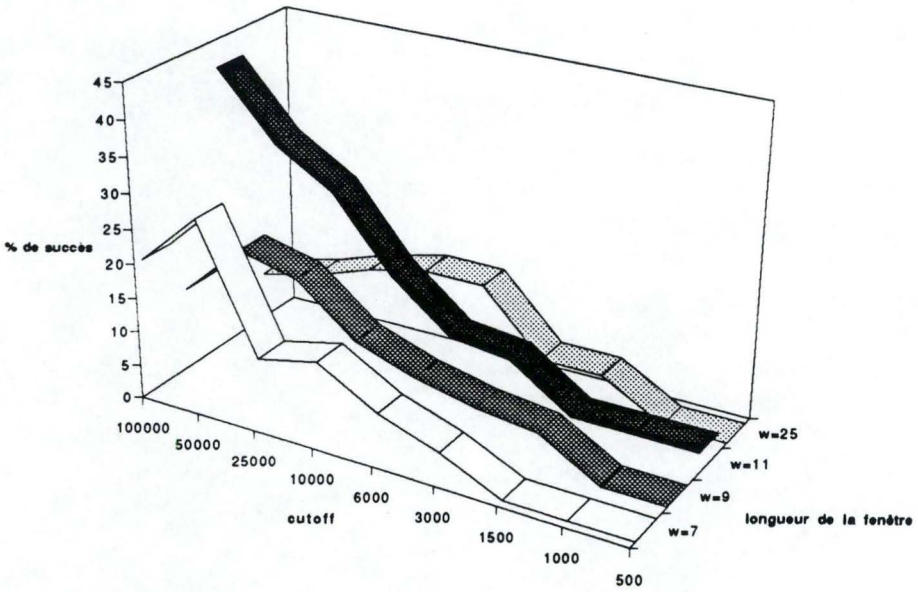
A	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	0	0	0
2	1000	0	0	0
3	1500	0	0	100
4	3000	0	4	57
5	6000	0	7	66
6	10000	8	12	66
7	25000	19	10	79
8	50000	33	28	57
9	100000	51	20	73

B	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	0	0	100
2	1000	0	0	100
3	1500	0	5	59
4	3000	2	6	62
5	6000	3	8	62
6	10000	10	11	70
7	25000	15	18	65
8	50000	33	20	69
9	100000	53	12	81

C	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	0	3	0
2	1000	0	3	0
3	1500	0	3	0
4	3000	0	9	0
5	6000	5	10	0
6	10000	7	17	41
7	25000	18	27	42
8	50000	35	32	42
9	100000	51	41	46

D	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	500	0	0	0
2	1000	0	0	0
3	1500	0	5	20
4	3000	0	5	20
5	6000	0	14	8
6	10000	10	14	58
7	25000	12	12	81
8	50000	25	9	85
9	100000	42	7	85

Table III.4 : Résultats obtenus lors de l'utilisation de SHOEBOX seul. Les chaînes latérales sont prises en compte et le balayage est fixé à 25. 4 longueurs de fenêtres sont considérées : 7 (a), 9 (b), 11 (c), et 25 (d).



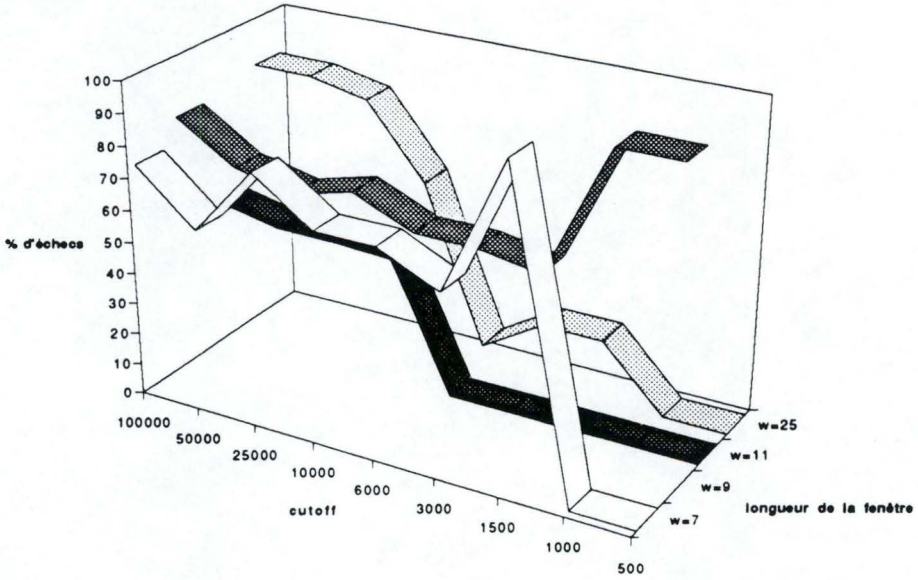


Figure III.10 : Pourcentage de succès et d'échecs pour les alignements réalisés avec SHOEBOX avec les chaînes latérales. 4 longueurs de fenêtres sont utilisées avec 9 valeurs de cutoffs différentes.

III.2.3. Comportement de SHOEBOX lors de la comparaison de la forme du peptide complet.

La méthode du RMS permet, par superposition, de comparer la forme de deux segments. Mais la nature même de cette technique empêche de tenir compte des chaînes latérales. Le but de cette technique étant de minimiser la distance entre atomes correspondants, elle reste cantonnée à l'utilisation des atomes N, C_α, C et O. En effet, comment mesurer une distance entre les atomes des chaînes secondaires de la glycine (-H) et de l'alanine (CH_3).

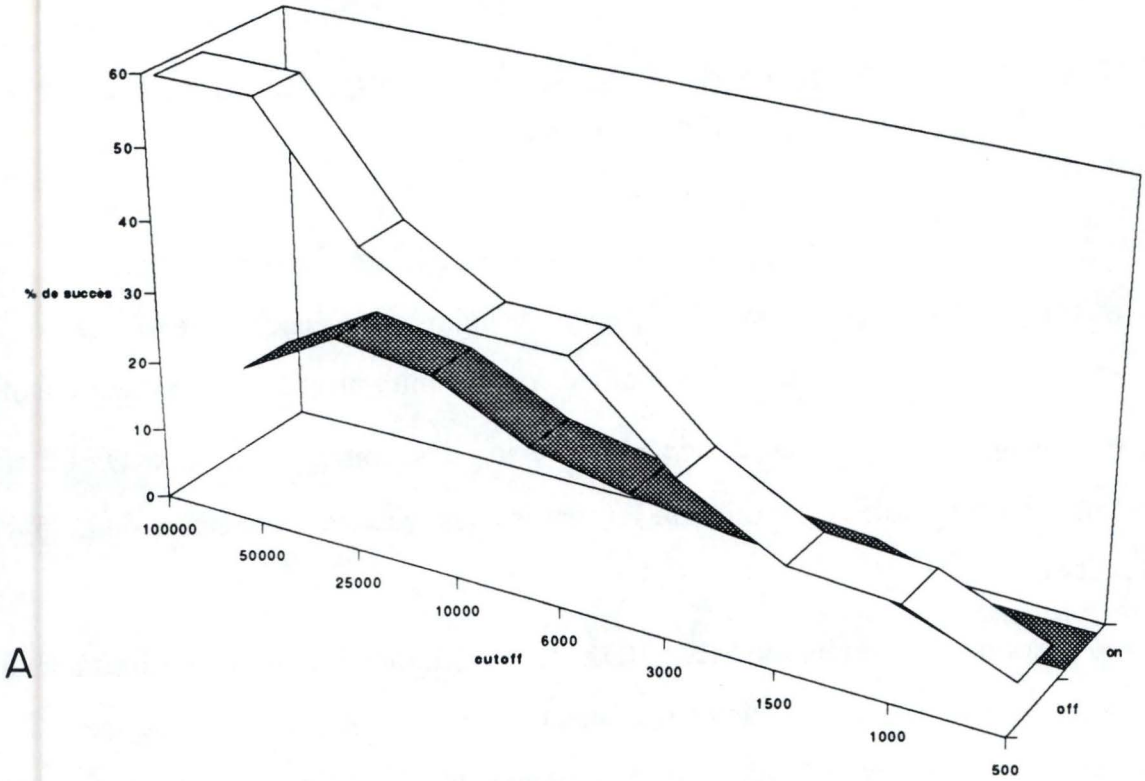
Au contraire, la méthode SHOEBOX permet l'utilisation de ces chaînes latérales. Étant basée sur une méthode de comparaison de la forme globale d'un segment, il n'y a besoin, à aucun moment, de réaliser des mesures de distances individuelles entre atomes. Deux segments d'hélice α se ressemblant peuvent donc être distingués par la nature et donc par l'espace occupé par leurs chaînes latérales.

III.2.3.1. Alignements.

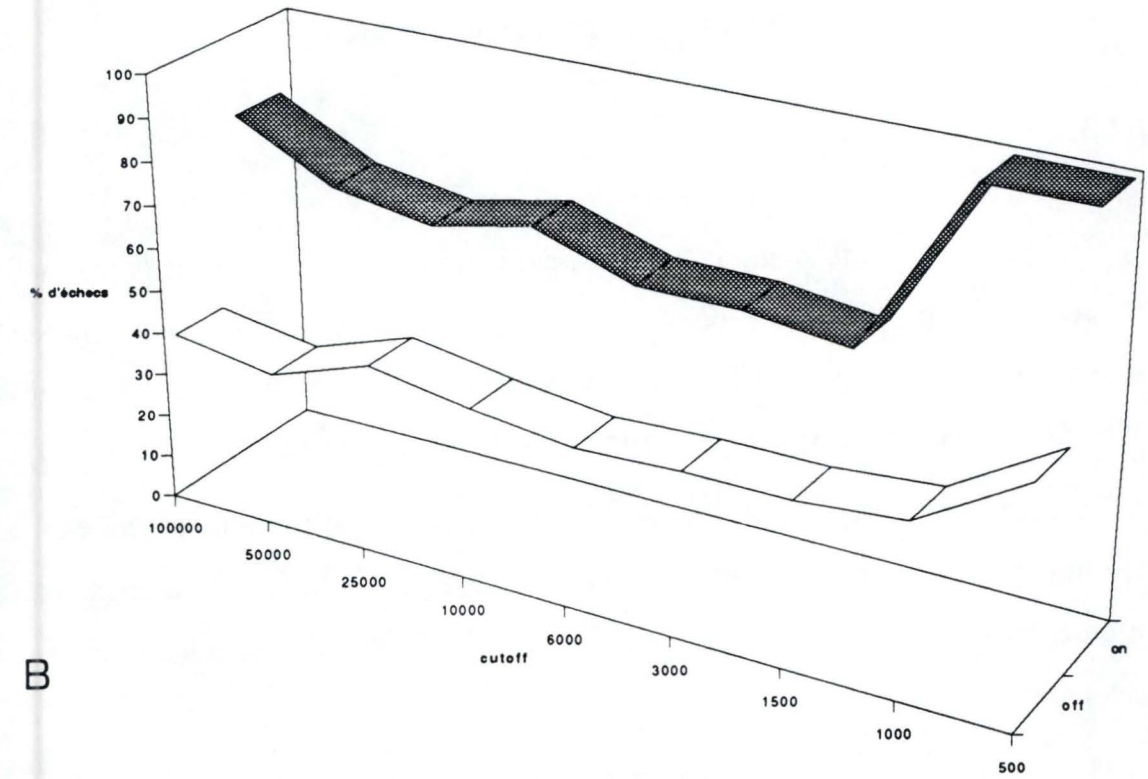
Une deuxième série d'alignements est donc lancée, elle est en tout point pareille à la première si ce n'est que, pour chaque résidu, tous les atomes sont considérés (Table III.4). Alors qu'on s'attendait à augmenter le pouvoir discriminatoire de SHOEBOX, on remarque que la situation ne s'est pas ou peu améliorée (fig. III.10).

Le pourcentage de succès diminue. Ce phénomène peut néanmoins être expliqué par le fait que, pour deux mêmes segments, leur différence de forme va augmenter si on ajoute dans le calcul les chaînes latérales. Un même cutoff sera donc plus strict si on tient compte des chaînes latérales.

Cette explication est également valable pour expliquer la diminution du bruit de fond : le fait que le cutoff soit plus strict permet une sélection moindre d'appariements dus au hasard.



A



B

Figure III.11 : Différence des pourcentage de succès (a) et d'échecs (b) lorsque la méthode SHOEBOX est utilisée avec (ON) ou sans (OFF) chaînes latérales. L'exemple présenté ici est celui de la fenêtre de 9 résidus, le balayage est toujours fixé à 25.

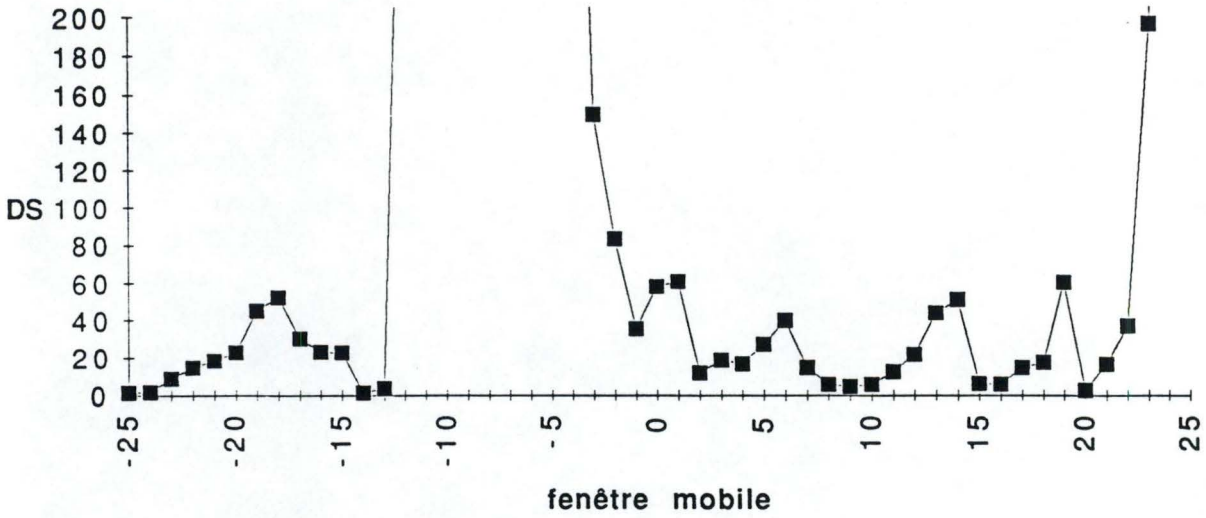


Figure III.12 : Balayage réalisé autour de la fenêtre 90. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX avec les chaînes latérales.

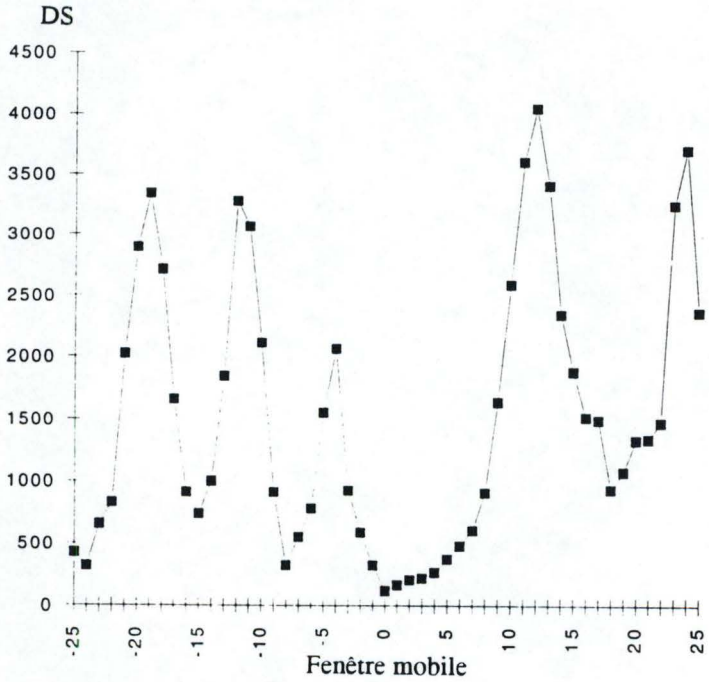


Figure III.13 : Balayage réalisé autour de la fenêtre 154. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX avec les chaînes latérales.

On se rend compte que, pour aucune des longueurs de fenêtre, la situation ne s'est améliorée mais qu'au contraire, elle semble s'être aggravée. En effet, le pourcentage d'échecs est plus important que dans la série d'alignements précédents. Ce qui signifie que la prise en compte des chaînes latérales cause des erreurs supplémentaires dans le choix des appariements. La figure III.11 nous permet de voir les différences de taux de succès et d'échecs lorsque SHOEBOX est utilisé avec ou sans les chaînes latérales.

III.2.3.2. Comportement de SHOEBOX lors du balayage des séquences.

Comme précédemment, l'étape suivante consiste à observer les résultats de différents balayage afin de voir ce que l'utilisation des chaînes latérales amène comme modification. De nouveau, il est nécessaire d'illustrer nos observations par différents exemples tirés du fichier scan.lis.

La première fenêtre considérée est, comme dans le point III.2.2.2., la fenêtre 90 (fig. III.12), faisant partie d'une hélice α. Sur le graphe représentant le balayage on peut, dans un premier temps, remarquer que, grâce à SHOEBOX, nous sommes toujours à même de détecter les fenêtres ayant même structure secondaire que la fenêtre initiale (les régions d'hélice α correspondant sur le graphe aux différents puits). Ensuite, en regardant de plus près, on peut voir que l'appariement correct n'est toujours pas détecté. La situation s'est même légèrement aggravée puisque, sur un balayage de 25 (c'est-à-dire 51 fenêtres), 34 appariements ont une distance inférieure à celle de l'appariement correct (pour 3 lorsque les chaînes latérales n'intervenaient pas).

Le second cas considéré est celui de la fenêtre 154 (fig. III.13). Rappelons que l'utilisation de SHOEBOX sans tenir compte des chaînes latérales, entraînait, pour cette fenêtre, la sélection d'un appariement incorrect. Ici, contrairement au cas précédent, l'utilisation de tous les atomes du segment permet d'augmenter le pouvoir discriminatoire. En effet, on peut voir sur le graphe que, avec SHOEBOX, il est

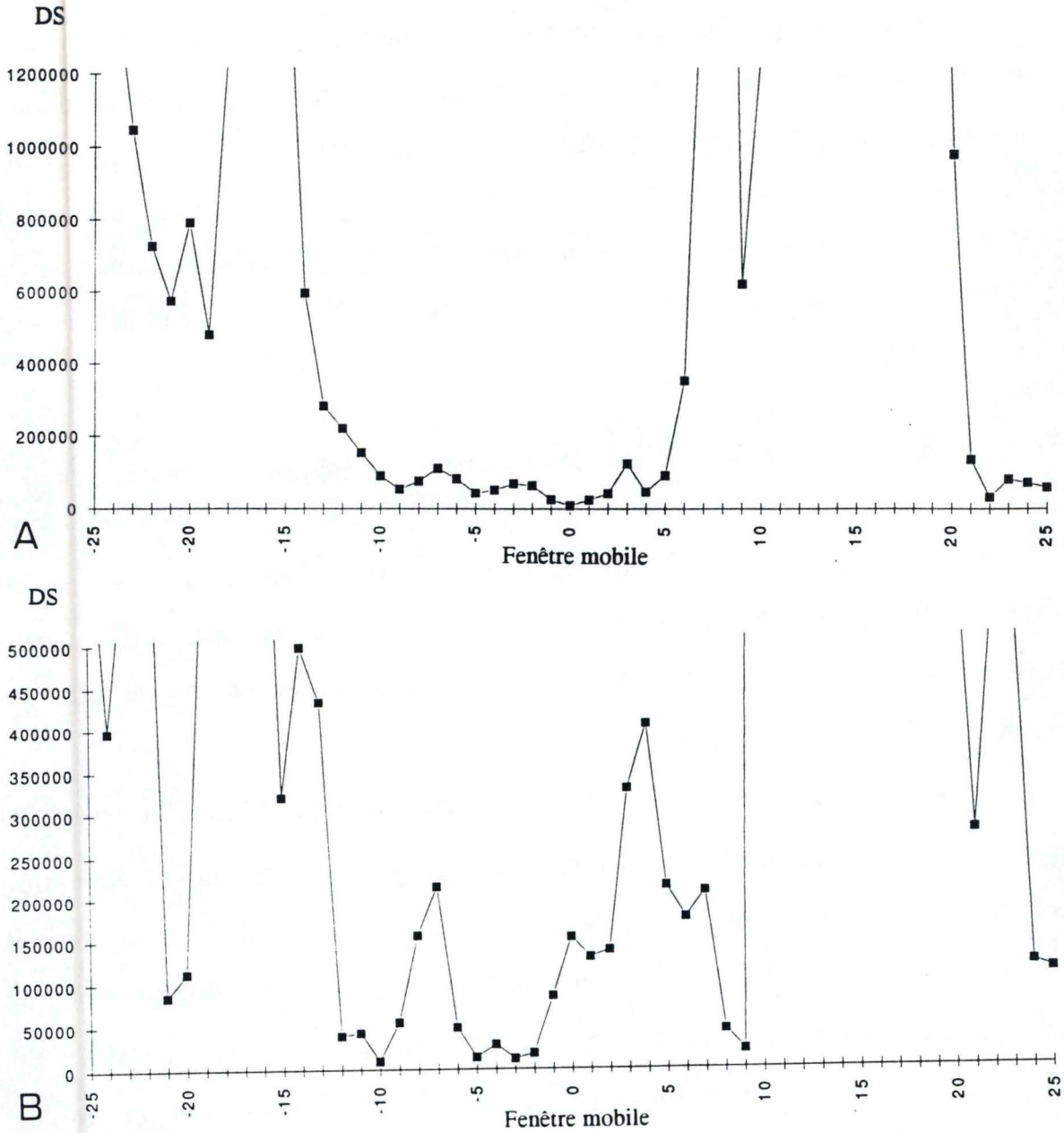


Figure III.14 : Balayage réalisé autour de la fenêtre 66. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées par la méthode SHOEBOX sans (a) et avec (b) les chaînes latérales.

possible de détecter différents brins β présents dans cette région mais que, de plus, l'appariement correct est retrouvé.

Le comportement de SHOEBOX a encore été observé dans le cas d'une troisième fenêtre, la 66, située dans une région d'hélice α (fig. III.14). Sur la figure représentant le balayage réalisé avec SHOEBOX sans les chaînes latérales, on remarque que les régions d'hélice α sont à nouveau détectées et que le minimum de distance trouvé correspond bien à l'appariement correct. Par contre lorsque les chaînes latérales entrent en ligne de compte, cet appariement n'est plus détecté.

Une explication de ce phénomène pourrait être que l'utilisation des chaînes latérales a masqué la ressemblance qui existait entre le squelette des fenêtres 66 de TPR et 56 de GRS.

III.2.3.3. Discussion.

Les exemples présentés ci-dessus ont été choisis parce qu'ils reflétaient bien l'image de ce que pouvait réaliser SHOEBOX lorsqu'on tenait compte des chaînes latérales. En effet, les différents résultats ont pu montrer que, dans certain cas (comme par exemple la fenêtre 154), l'utilisation des chaînes latérales permettait d'améliorer le pouvoir discriminatoire de SHOEBOX. Malheureusement, ils ont aussi permis de voir que dans d'autres cas, cela ne changeait rien (fenêtre 99) ou même que tenir compte des chaînes latérales entraînait des erreurs supplémentaires (fenêtre 66).

Une observation plus minutieuse nous montre que la plupart des cas où la situation est améliorée mettent en présence des paires de segments ayant un nombre plus ou moins important d'identités (et donc de chaînes latérales identiques) ou des résidus aux chaînes latérales semblables (par exemple le remplacement d'une isoleucine par une leucine ou d'une tyrosine par un histidine).

Par contre, lorsque l'utilisation des chaînes latérales amène à des erreurs, c'est, la plupart du temps, quand les substitutions impliquées mettent en jeu des résidus ayant

des chaînes latérales de nature différente (par exemple une alanine remplacée par une lysine, une thréonine remplacée par une phénylalanine).

Les observations évoquées ci-dessus sont issues d'expériences limitées à deux séquences. Donc, même si cette explication semble plausible, il est impossible de tirer des conclusions générales sur ce point. D'autres études seraient nécessaires pour déterminer pourquoi la détection de certains appariements est favorisée par la prise en compte des chaînes latérales, alors que pour d'autres, elle est défavorisée.

III.2.4. Conclusion.

Toutes les expériences réalisées ont permis de montrer que la méthode SHOEBOX utilisée seule ne nous permet pas la réalisation de l'alignement complet des deux structures GRS et TPR. Pourtant, nous avons pu mettre en évidence que la comparaison de la forme des segments (en tenant compte ou non des chaînes latérales) permet de détecter les structures secondaires identiques. Par contre, nous avons aussi vu que, au sein même de ces structures, l'appariement correct n'est pas toujours détecté.

Le fait de tenir compte des chaînes latérales était supposé augmenter le pouvoir discriminatoire de SHOEBOX. Malheureusement si cela se vérife dans certains cas (fenêtre 154), d'autres cas ne semblaient pas être améliorés (fenêtre 90). De plus, l'apport des chaînes latérales entraîne parfois des erreurs qui n'étaient pas commises auparavant (fenêtre 66).

Grâce aux résultats obtenus, nous avons pu déduire que la méthode SHOEBOX semblait tout à fait adaptée au cas où les SCR sont formés d'acides aminés ayant des chaînes latérales semblables. Des évaluations réalisées précédemment ont d'ailleurs pu le démontrer.

Par contre, si les SCR sont formés d'acides aminés possédant des chaînes latérales de natures différentes, SHOEBOX perd son pouvoir discriminatoire. Ce phénomène est facilement explicable par la nature même de la mesure de distance définie par SHOEBOX : la différence de forme entre deux segments de structures.

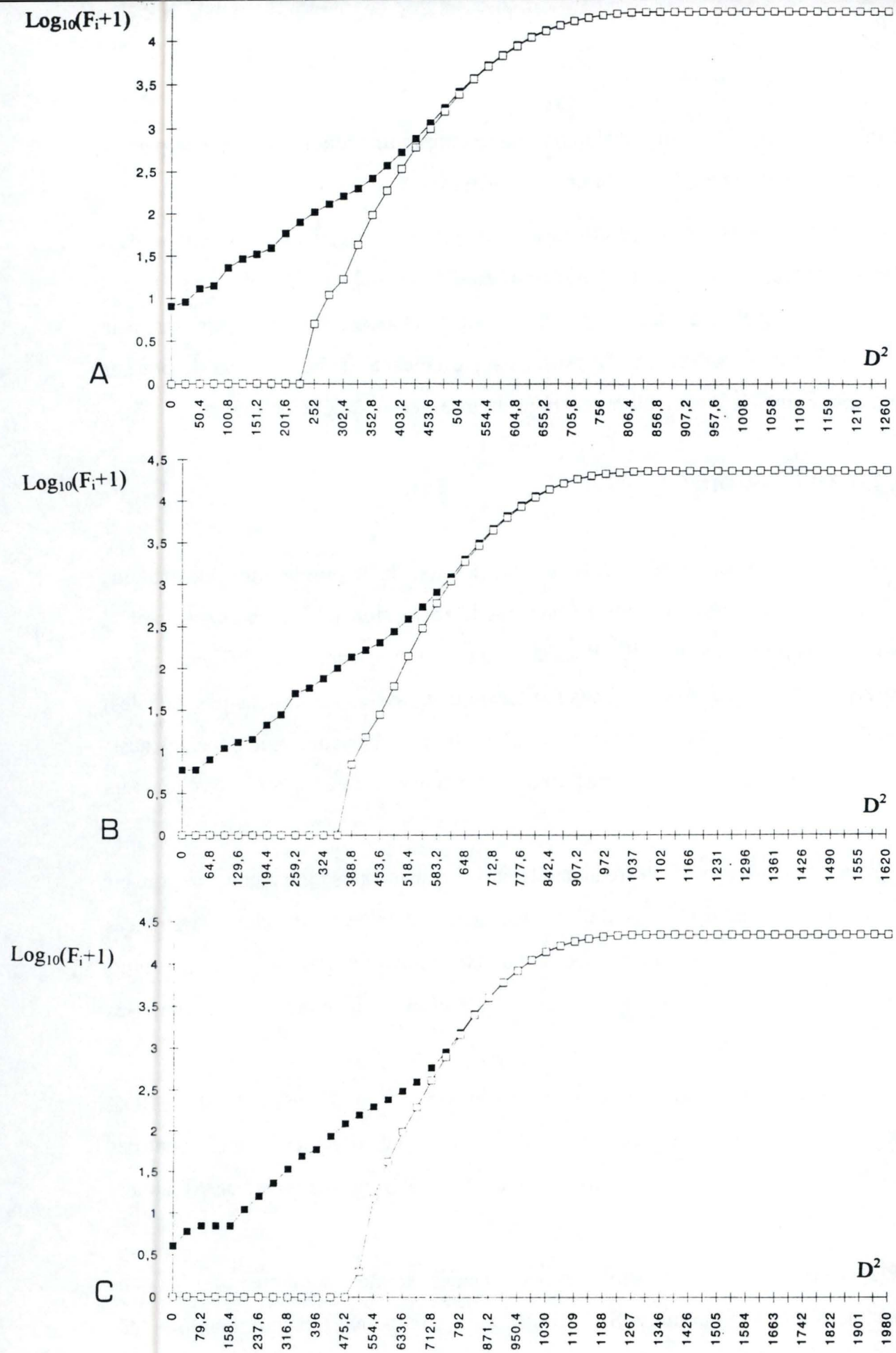


Figure III.15 : Graphes du logarithme des fréquences cumulées des distances (calculées à partir de la matrice DFK) observées dans les protéines réelles et dans des séquences générées aléatoirement. Trois longueurs de fenêtres sont utilisées : 7 (a), 9 (b) et 11 (c).

Une substitution impliquant deux acides aminés de volumes différents entraînera une modification de la forme des segments correspondants.

La prise en compte des chaînes latérales n'ayant pas eu l'effet escompté, il est nécessaire de trouver un autre moyen permettant de choisir l'appariement correct dans une région de même structure secondaire. Le programme MATCH-BOX offre la possibilité d'utiliser en même temps plusieurs mesures de similarité. Une fois combinées, elles peuvent être employées pour la recherche de l'appariement optimal. Utiliser SHOEBOX avec d'autres mesures de similarité de séquences ou de structures semblait pouvoir solutionner notre problème.

III.3. Utilisation de SHOEBOX en combinaison avec deux autres mesures de similarité.

III.3.1. SHOEBOX en collaboration avec DFK et RMS.

III.3.1.1. Les différentes mesures utilisées et leur cutoff respectif.

DFK : La valeur de cutoff utilisée pour la matrice de score DFK est déterminée par les procédures RANDOMIZE et DISTRIB présentées dans la partie « matériel et méthodes » de ce travail(fig. III.15).

Pour chaque longueur de fenêtre, nous utiliserons trois cutoffs différents allant d'une valeur stricte à une plus large. La plus stricte se traduira par une sélection moindre d'appariements mais également par un bruit de fond plus faible. La plus large, quant à elle se traduira par des caractéristiques contraires, à savoir, plus d'appariements sélectionnés pour un bruit de fond plus important.

RMS : Selon Unger et al (1989), la valeur de 1Å comme seuil de la distance RMS permet une bonne discrimination entre les segments de structures semblables et les autres.

	Longueur de la fenêtre	cutoffs utilisés (DFK, SHOEBOX, RMS)	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400, 100000, 1	80	24	64
2		300, 75000, 1	78	31	55
3		200, 50000, 1	78	61	26
4	9 résidus	600, 150000, 1	75	84	20
5		500, 125000, 1	74	86	18
6		400, 100000, 1	72	88	15
7	11 résidus	800, 200000, 1	74	68	28
8		700, 175000, 1	73	70	25
9		600, 150000, 1	70	71	24

Table III.5 : Résultats obtenus lors de l'utilisation des trois mesures de distances (DFK, SHOEBOX et RMS). Le balayage est fixé à 25 et, pour chaque longueur de fenêtres, trois jeux de cutoffs sont utilisés.

SHOEBOX : La valeur communément utilisée pour SHOEBOX est fixée à 10^5, lorsque le cutoff de DFK est de 400. Cette relation sera donc appliquée pour calculer le cutoff de SHOEBOX : il sera toujours égal à 250 fois celui de DFK. Lorsqu'il est combiné à d'autres mesures de distance, SHOEBOX est utilisé sans les chaînes latérales.

III.3.1.2. Réalisation d'alignements avec DFK, RMS et SHOEBOX utilisés simultanément.

Une premiere série d'alignements est réalisée en utilisant les trois mesures de similarité combinées (table III.5). Le calcul du score global utilisé a été présenté au point 1.5.2.4 de l'introduction.

Au vu des résultats, la combinaison de paramètres où la fenêtre compte 9 résidus de long et où le jeu de cutoffs est le plus strict, apparaît comme la plus performante avec un pourcentage de succès de 88% et un pourcentage d'échecs se situant à 15%.

On peut également se rendre compte que les résultats obtenus par cette comparaison des scores combinés sont de loin supérieurs à ceux obtenus avec la méthode SHOEBOX lorsqu'elle était utilisée seule.

La question que l'on peut se poser à ce niveau est de savoir quelle(s) composante(s) du score global permet(tent) d'obtenir de si bons résultats. Le point suivant consistera donc à évaluer les possibilités de chaque technique prise séparément.

A	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	1	81	91	15
2	9 résidus	1	76	89	18
3	11 résidus	1	69	94	10

B	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400	24	59	23
2		300	1	59	7
3		200	0	25	1
4	9 résidus	600	45	66	31
5		500	7	65	18
6		400	0	62	4
7	11 résidus	800	60	60	34
8		700	19	68	24
9		600	1	80	6

C	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	100000	83	20	73
2		75000	82	25	66
3		50000	80	27	65
4	9 résidus	150000	81	66	35
5		125000	79	61	37
6		100000	77	58	39
7	11 résidus	200000	80	55	47
8		175000	79	54	47
9		150000	78	46	53

Table III.6 : Résultats obtenus lors de l'utilisation des trois mesures de distances séparément : DFK (a), SHOEBOX (b) et RMS (c). La longueur du balayage est de 25.

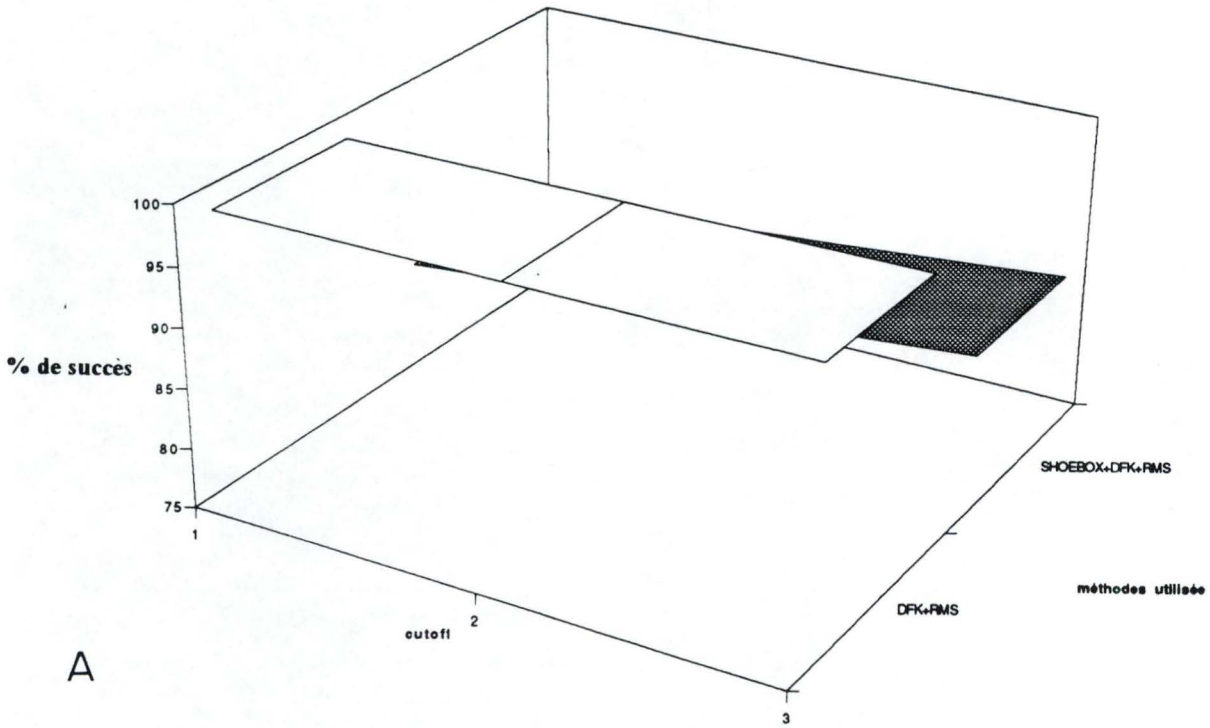
III.3.1.3. Réalisation d'alignements avec DFK, RMS et SHOEBOX utilisés séparément.

La table III.6 présentée ci-contre nous permet certaines observations :

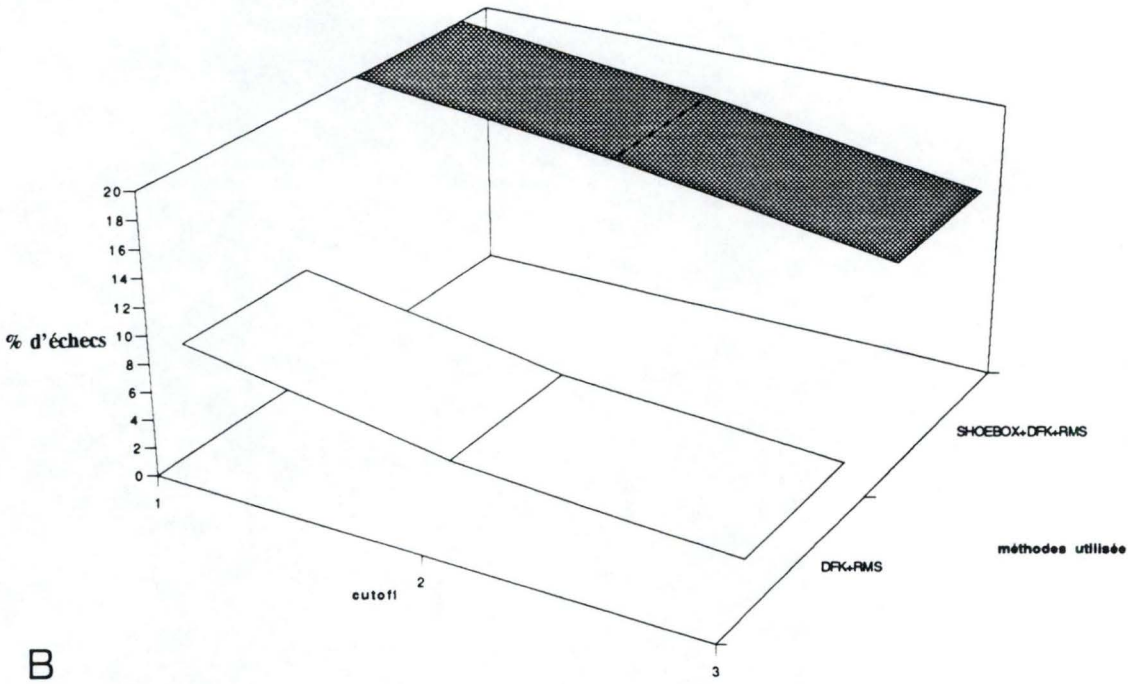
- la méthode la plus efficace semble être celle du calcul de la distance RMS. En effet, le pourcentage d'échecs mesuré dans ce cas est en moyenne de 18% alors que le pourcentage de succès s'élève jusqu'à 94% pour une fenêtre de 11 résidus. Cette bonne performance peut sans doute s'expliquer en partie par le fait que l'alignement de référence a été réalisé par superposition de structure et donc par minimisation de la distance RMS. Il paraît donc logique de retrouver, à peu de chose près, les mêmes résultats.
- les résultats obtenus pour la comparaison des séquences par la matrice DFK sont aussi très bons. De nouveau, les meilleures performances sont obtenues pour la fenêtre de 11 résidus avec un pourcentage de succès de 80% pour un pourcentage d'échecs de seulement 6%. On peut également remarquer que, pour des cutoffs stricts, le pourcentage d'échecs se situe toujours à un niveau très bas, avec une valeur record de 1% pour une fenêtre de 7 résidus. Ce pourcentage d'échecs est pourtant associé à un pourcentage de succès peu élevé (25%). L'augmentation du cutoff n'améliore pas cette situation puisque, même si elle permet plus de succès, elle entraîne aussi plus d'échecs.
- quant à la méthode SHOEBOX, ses résultats sont bien évidemment semblables à ceux obtenus précédemment. Un phénomène généralement observé est un taux d'échecs se situant à un niveau toujours très élevé par rapport aux autres méthodes. La valeur minimale obtenue pour une fenêtre de 9 résidus atteint 35%. Dans le meilleur des cas, donc, SHOEBOX provoque une fois sur trois l'erreur dans le choix des appariements.

	Longueur de la fenêtre	cutoffs utilisés (DFK, RMS)	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400, 1	76	92	12
2		300, 1	73	91	8
3		200, 1	68	86	6
4	9 résidus	600, 1	70	98	8
5		500, 1	68	98	5
6		400, 1	65	98	4
7	11 résidus	800, 1	63	98	9
8		700, 1	61	70	33
9		600, 1	58	90	10

Table III.7 : Résultats obtenus lors de l'utilisation des mesures de distances DFK et RMS. Le balayage est fixé à 25 et, pour chaque longueur de fenêtres, trois jeux de cutoffs sont utilisés.



A



B

Figure III.16 : Différences des pourcentages de succès (a) et d'échecs (b) pour les alignements réalisés avec DFK et RMS et avec les trois mesures de distances. Le balayage est de 25 et la fenêtre de 9 résidus.

III.3.1.4. Évaluation du rôle joué par SHOEBOX dans les alignements réalisés avec DFK, RMS et SHOEBOX utilisés simultanément.

Dans le point précédent, nous avons vu que SHOEBOX paraissait, parmi les trois méthodes utilisées, être la moins efficace. Il était donc intéressant d'observer les résultats d'un alignement pour lequel seule la comparaison des séquences par DFK et la distance RMS étaient utilisées (table III.7). En les comparant avec ceux obtenus pour l'alignement où les trois mesures étaient prises en compte, l'apport de SHOEBOX pouvait être précisé.

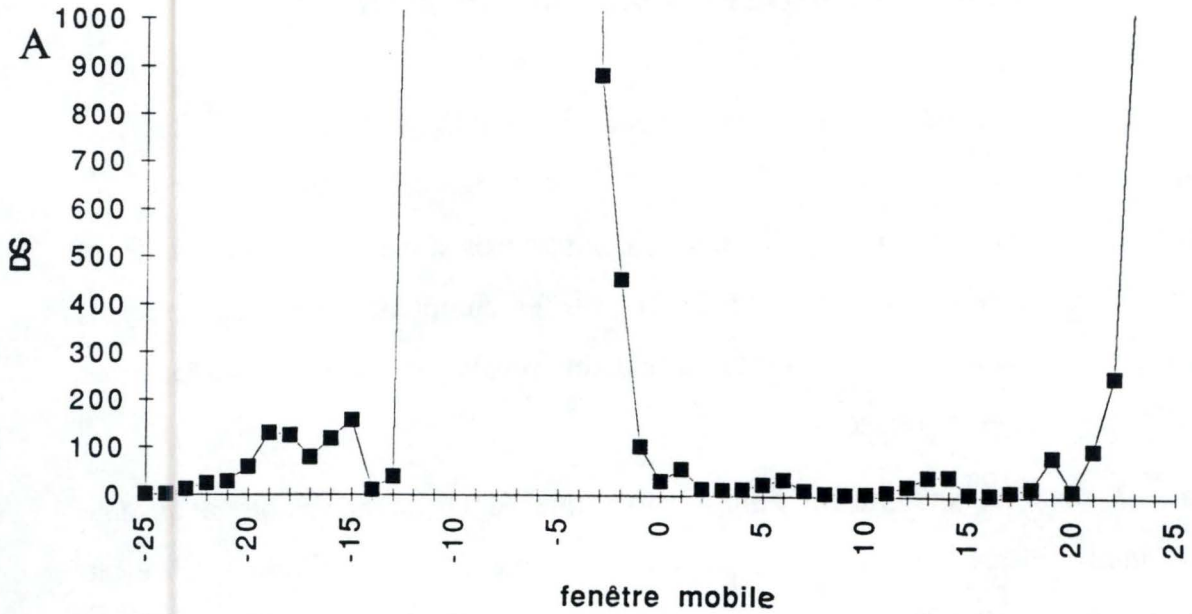
Face à ces résultats, on se rend compte que SHOEBOX semble avoir une efficacité moindre que les deux autres méthodes de mesure de similarité. En effet, prenons le cas d'une fenêtre de 9 résidus et d'un jeu de cutoffs stricts (fig. III.16), le fait de ne pas faire intervenir la méthode SHOEBOX entraîne une augmentation du pourcentage de succès (de 88% à 98%) conjuguée à une diminution du pourcentage d'échecs (de 15% à 4%). Il semble donc clair que SHOEBOX a un pouvoir discriminatoire plus faible que le RMS ou que DFK et que, de plus, lorsqu'il est utilisé en collaboration avec ces derniers, il entraîne une diminution de leurs performances.

Dans certains cas, l'utilisation de DFK et RMS permet de détecter les appariements corrects (aux effets de bords près). Par contre, si les trois mesures de similarité sont prises en compte, des appariements incorrects sont sélectionnés.

Pour essayer de comprendre ce phénomène, il est une nouvelle fois très utile de suivre la variation des distances lors du balayage réalisé par les trois méthodes.

III.3.1.5. Observation du comportement de DFK, RMS et SHOEBOX lors d'un balayage.

Comme nous l'avons vu précédemment, une des régions où l'apport de SHOEBOX semble être négatif est celle de l'hélice α dont fait partie, entre autre, la fenêtre numéro 90. Nous allons donc nous pencher sur le cas du balayage réalisé pour



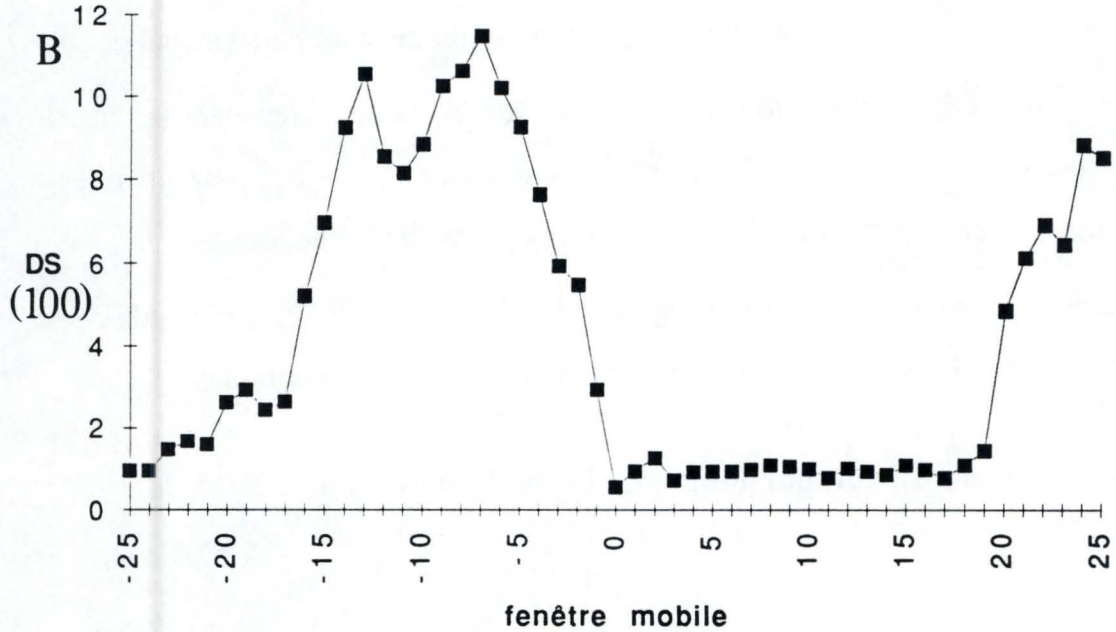


Figure III.17 : Balayage réalisé autour la fenêtre 90. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées avec DFK et RMS (a) et avec les trois mesures de distances (b).

	Longueur de la fenêtre	cutoffs utilisés (DFK, SHOEBOX, RMS)	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400, 100000, 1	80	24	64
2		300, 75000, 1	78	31	55
3		200, 50000, 1	78	61	26
4	9 résidus	600, 150000, 1	75	84	20
5		500, 125000, 1	74	86	18
6		400, 100000, 1	72	88	15
7	11 résidus	800, 200000, 1	74	68	28
8		700, 175000, 1	73	70	25
9		600, 150000, 1	70	71	24

Table III.8 : Distances observées par chaque mesure lors du balayage autour de la fenêtre 90. La fenêtre est de 9 résidus et le balayage de 25.

cette fenêtre particulièrement représentative du problème posé par l'utilisation de SHOEBOX (fig. III.17).

Le balayage réalisé grâce à MATCH-BOX lorsque seules les distances physico-chimiques et RMS sont prises en compte montre que l'appariement correct est localisé (le minimum trouvé correspond bien à l'abscisse 0). Par contre dès que l'on ajoute la mesure de la forme du segment dans le calcul du score global, ce minimum n'est plus détecté. C'est donc bien le juste reflet de la situation observée dans les alignements.

Le problème posé par SHOEBOX est le suivant : une fois utilisé avec d'autres mesures de similarité, il semble avoir plus de poids que les autres méthodes dans le calcul du score global.

Cela peut facilement s'expliquer lorsqu'on regarde les distances mesurées, au cours d'un balayage, par DFK, RMS et SHOEBOX (table II.8).

En effet, dans le cas présenté ci-contre, on peut remarquer que la plus petite distance mesurée avec la méthode SHOEBOX est de 11300 alors que la plus grande est de 18498900. C'est-à-dire un rapport de plus de 1000 entre ces deux valeurs. Pour DFK ce rapport est de 2 ou 3 et pour RMS il est de l'ordre de 15.

Rappelons que le mode de calcul du score global est le suivant :

$$DS = \frac{Ds(d)}{S1} \times \frac{D_\lambda}{S2} \times \frac{r.m.s.}{S3}$$

avec S_1, S_2 et S_3, les seuils respectifs de DFK, SHOEBOX et RMS.

Dans cette expression, il est évident que les distances calculées par SHOEBOX auront un poids beaucoup plus important que les autres.

Pour mieux comprendre cette situation, prenons à nouveau l'exemple de la fenêtre 90. Le minimum de distance trouvé par SHOEBOX est de 11300 et se situe 14 résidus en amont de l'appariement correct pour lequel la distance mesurée est, elle, de 580600. Il y a donc entre ces deux valeurs un rapport de plus ou moins 50, ou, en d'autres termes, le score pour le segment -14 sera 50 fois plus petit que celui du segment à apparier (rappelons qu'on mesure une distance et donc que les segments

proches se traduisent par un score plus petit). Cela signifie que le produit des rapports pour les deux autres mesures doit être supérieur à 50 pour pouvoir redresser la situation. Plus clairement, le produit suivant :

$$\frac{Ds(d)}{S1} \times \frac{r.m.s.}{S3}$$

doit être 50 fois plus petit pour l'appariement correct que pour le « faux » minimum trouvé par SHOEBOX. C'est évidemment impossible. La prépondérance de SHOEBOX par rapport aux deux autres méthodes est trop marquée..

Une première idée pour remédier à ce problème est de modifier les valeurs de cutoffs fixées pour SHOEBOX, ce qui revient à pondérer le terme relatif à la forme des segments dans l'expression du score global. Les résultats obtenus après ce changement étaient en tous points pareils aux précédents. La seule différence lors du balayage résidait dans un changement d'échelle.

On pouvait s'attendre à ces résultats puisque le fait de diviser les valeurs de distances par 1000 ou par 100000 ne change rien aux rapports existants entre elles. De par là même, l'importance prise par SHOEBOX dans le score global est conservée.

En ce qui concerne les alignements réalisés avec un cutoff plus strict pour SHOEBOX, le pourcentage d'échecs était évidemment plus petit. Un cutoff plus strict permet de détecter moins d'appariements dus au hasard. Mais, le pourcentage de succès est très faible et pour l'augmenter, il est nécessaire de remonter le cutoff. La situation n'est donc pas améliorée et montre une nouvelle fois que le pouvoir discriminatoire de SHOEBOX semble trop faible.

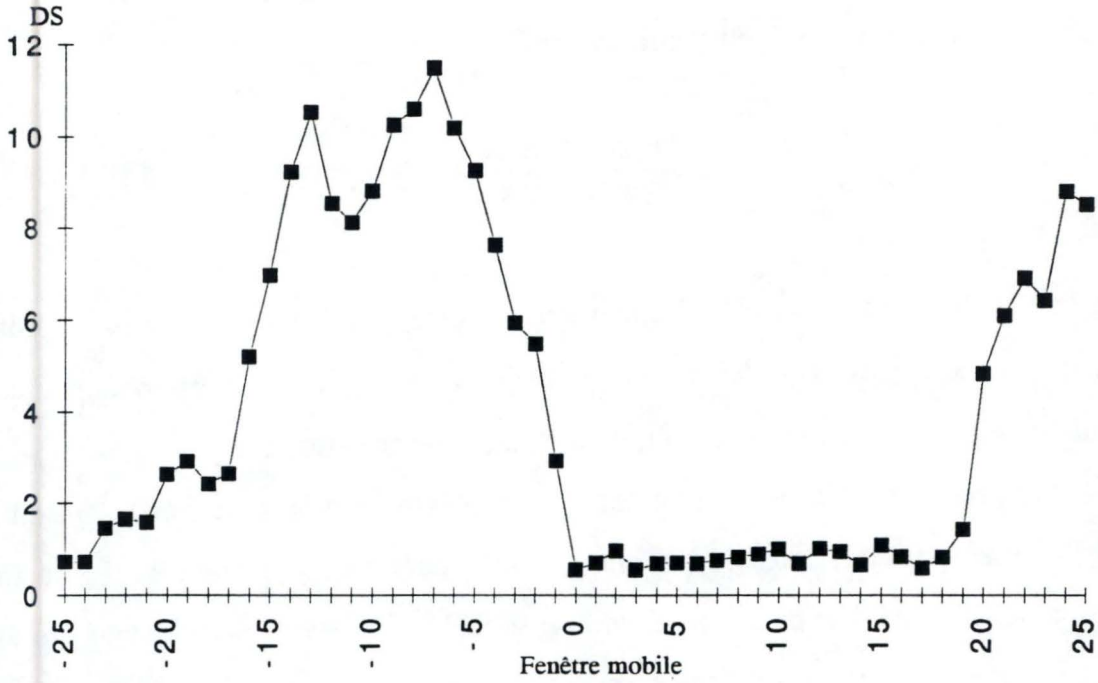


Figure III.18 : Balayage réalisé autour la fenêtre 90. La longueur du balayage est de 25 et la longueur des fenêtres est de 9 résidus. Les distances (DS) sont celles mesurées avec les trois mesures de distances dont SHOEBOX borné.

III.3.1.6. Fixation de bornes pour limiter l'apport de SHOEBOX.

Finalement, pour diminuer l'importance prise par SHOEBOX dans le score global, un seul moyen paraît efficace : fixer des bornes. En clair, cela signifie que l'on va limiter la variation de la mesure de distance entre une valeur inférieure et une valeur supérieure. Tout rapport distance/cutoff (le rapport D_λ/S_2 de l'expression du score global) plus petit que la borne inférieure se verra automatiquement attribuer la valeur de celle-ci et, identiquement, un rapport trop grand sera remplacé par la valeur de la borne supérieure (fig. III.18).

Dans notre cas, cela revient à empêcher SHOEBOX de s'exprimer. En effet, pour la région de la fenêtre 90, il est nécessaire de fixer les bornes à 0,75 pour l'inférieure et à 1 pour la supérieure pour éviter que SHOEBOX n'impose ses erreurs dans le score global. Ces bornes sont telles que SHOEBOX n'a plus aucun poids dans le calcul du score global. Il est donc plus facile, pour obtenir un résultat plus proche de l'alignement de référence, de retirer SHOEBOX.

III.4. Utilisation de SHOEBOX en collaboration avec différentes matrices de scores.

Cependant le problème doit aussi être abordé sous un autre point de vue : même si la méthode RMS est plus performante, elle est aussi beaucoup plus exigeante en temps calcul que SHOEBOX. Par exemple, pour la recherche des appariements entre nos deux structures avec une fenêtre de 9 résidus et un balayage de 500, l'utilisation conjointe de DFK et RMS nécessite 2 h 48 min de cpu alors que celle de de DFK et SHOEBOX n'en demande que 53,71 secondes.

Il est donc intéressant, dans un dernier point, d'étudier le comportement de SHOEBOX lors de son utilisation avec différentes matrices de scores sélectionnées pour leurs bonnes performances.

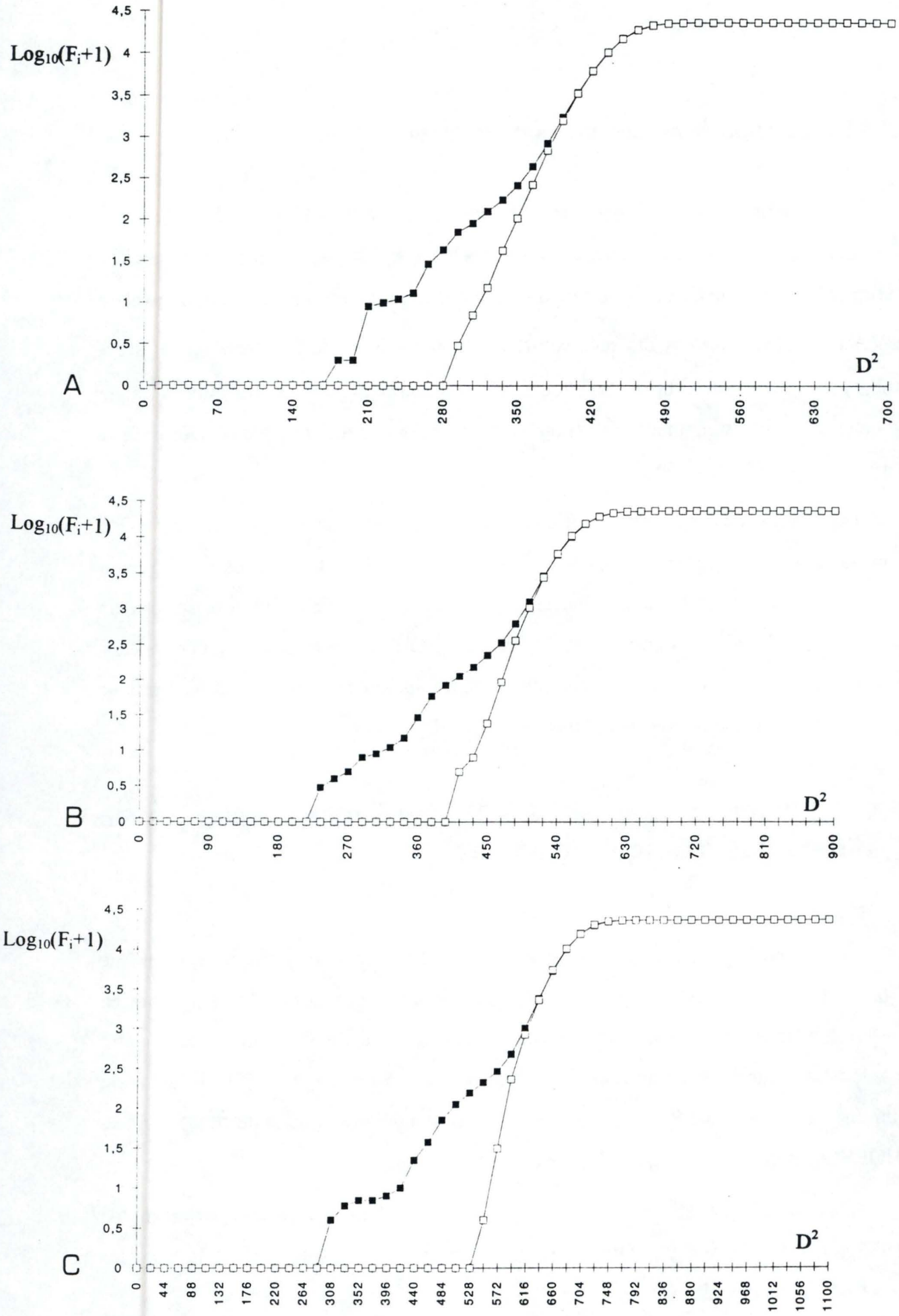


Figure III.19 : Graphes du logarithme des fréquences cumulées des distances (calculées à partir de la matrice BIRKBECK97) observées dans les protéines réelles et dans des séquences générées aléatoirement. Trois longueurs de fenêtres sont utilisées : 7 (a), 9 (b) et 11 (c).

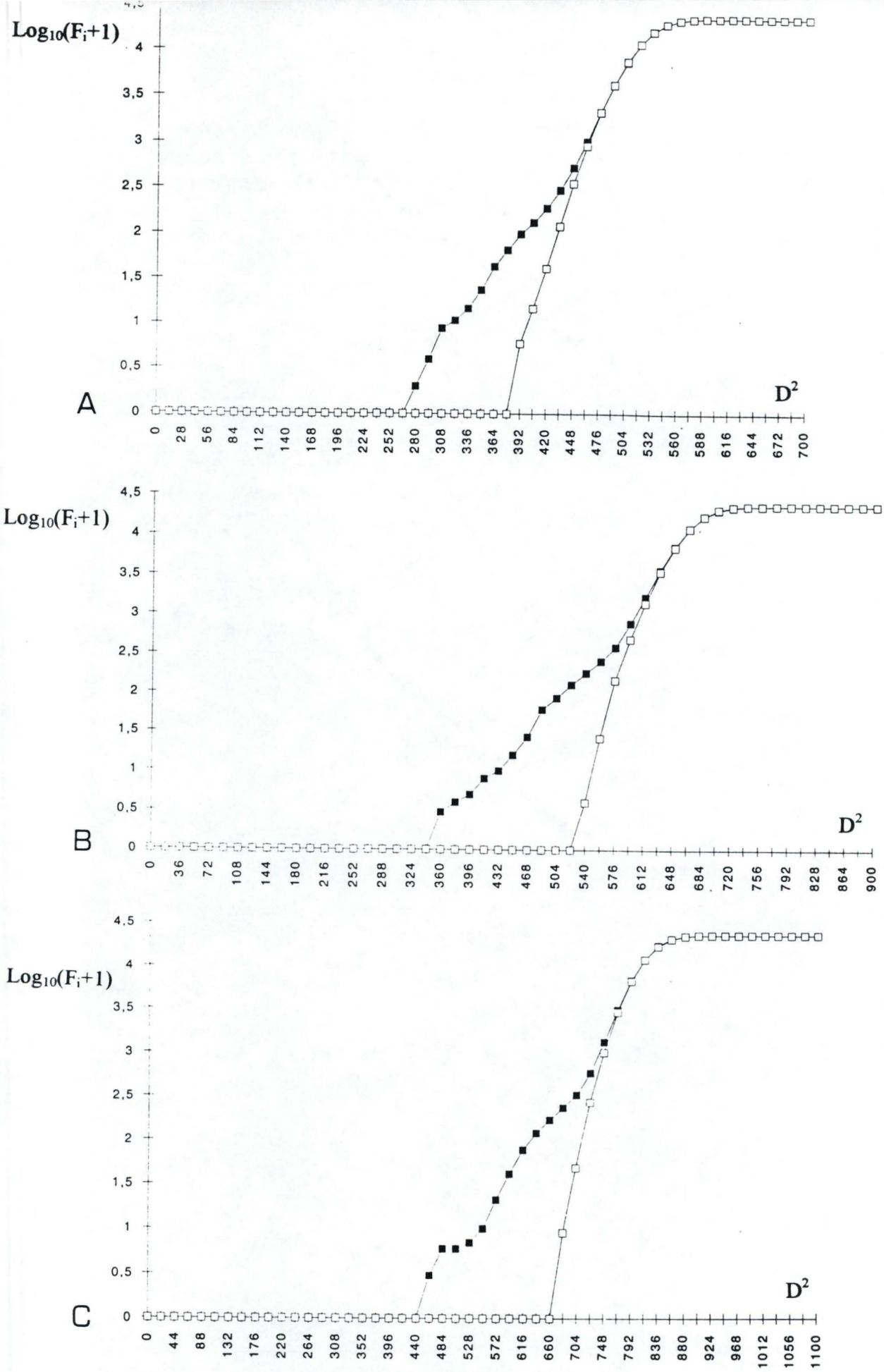


Figure III.20 : Graphes du logarithme des fréquences cumulées des distances (calculées à partir de la matrice GONNET) observées dans les protéines réelles et dans des séquences générées aléatoirement. Trois longueurs de fenêtres sont utilisées : 7 (a), 9 (b) et 11 (c).

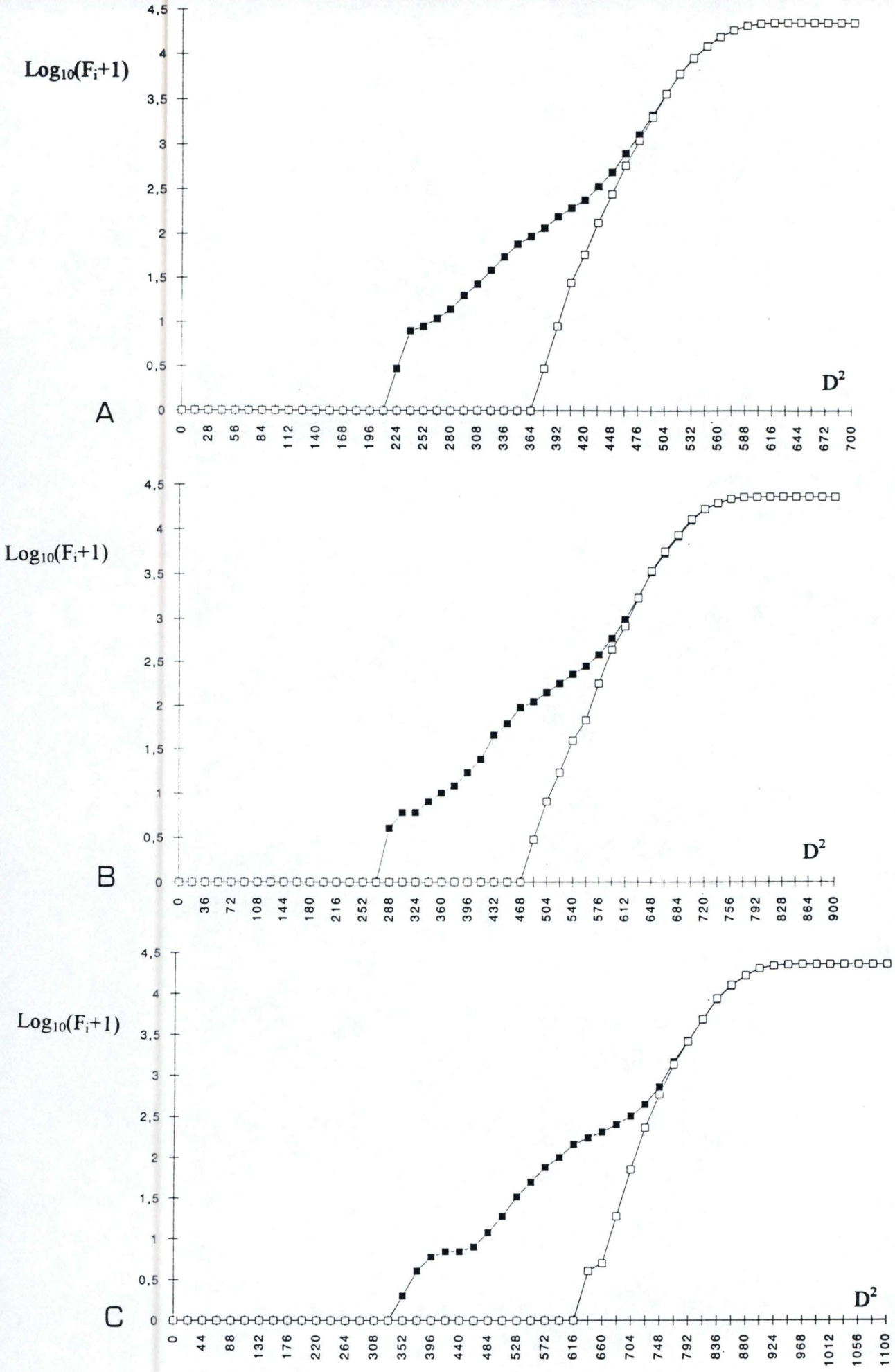


Figure III.21 : Graphes du logarithme des fréquences cumulées des distances (calculées à partir de la matrice HENIKOFF) observées dans les protéines réelles et dans des séquences générées aléatoirement. Trois longueurs de fenêtres sont utilisées : 7 (a), 9 (b) et 11 (c).

A	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400,100000	74	16	79
2		300,75000	72	14	77
3		200,50000	71	37	55
4	9 résidus	600,150000	72	50	37
5		500,125000	69	57	36
6		400,100000	67	59	34
7	11 résidus	800,200000	73	53	47
8		700,175000	71	47	43
9		600,150000	68	50	37

B	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400,100000	76	13	82
2		300,75000	73	16	78
3		200,50000	72	38	50
4	9 résidus	600,150000	75	39	60
5		500,125000	72	57	41
6		400,100000	69	57	39
7	11 résidus	800,200000	75	48	50
8		700,175000	73	48	49
9		600,150000	71	49	47

C	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400,100000	75	9	87
2		300,75000	73	12	83
3		200,50000	72	34	55
4	9 résidus	600,150000	74	26	70
5		500,125000	70	56	43
6		400,100000	68	56	38
7	11 résidus	800,200000	78	49	5
8		700,175000	72	51	47
9		600,150000	69	60	35

D	Longueur de la fenêtre	cutoffs utilisés	résultats : % de bruit de fond	% de succès	% d'échecs
1	7 résidus	400,100000	75	17	77
2		300,75000	73	18	75
3		200,50000	72	39	42
4	9 résidus	600,150000	73	36	57
5		500,125000	70	57	41
6		400,100000	68	58	37
7	11 résidus	800,200000	74	48	48
8		700,175000	72	53	43
9		600,150000	69	48	45

Table III.9 : Résultats obtenus lors de l'utilisation de SHOEBOX avec les différentes matrices de scores : DFK (a), BIRKBECK97 (b), GONNET (c) et HENIKOFF (d). Le balayage est fixé à 25 et, pour chaque longueur de fenêtres, trois jeux de cutoffs sont utilisés.

III.4.1. Les différentes matrices de scores utilisées et leurs cutoffs respectifs.

Outre la matrice DFK mesurant les distances physico-chimiques, trois autres matrices de scores furent sélectionnées : BIRKBECK97, GONNET et HENIKOFF. Nous avons vu au point I.3.1 de l'introduction que la première était basée sur des calculs de substitutions observées lors de comparaisons de structures alors que, en ce qui concerne les deux autres, seuls des alignements de séquences avaient été utlisisés. Des tests réalisés sur ces trois matrices par Johnson et Overington (1993) ont pu montrer qu'elles permettaient d'obtenir de très bons résultats.

La recherche des cutoffs à utiliser pour ces matrices est comme précedemment réalisée grâce aux procédures RANDOMIZE et DISTRIB. A la lecture des graphes (Fig. III.19, III.20 et III.21), nous pouvons voir que les valeurs déjà utilisées pour la matrice DFK peuvent être conservées pour les trois autres. Les distributions de distances sont légèrement différentes pour les quatre matrices, mais l'utilisation de trois valeurs de cutoffs par longueur de fenêtre nous permet, pour chaque matrice, d'avoir au moins un cutoff adapté.

III.4.2. Réalisation des alignements avec SHOEBOX utilisé en collaboration avec les différentes matrices de scores.

Pour chaque matrice de scores, une série de quatre alignements est entreprise, toujours avec un balayage fixé à 25. Les résultats sont présentés dans la table III.9

On peut tout de suite se rendre compte que la situation ne s'est toujours pas améliorée. Pour rappel, lorsque SHOEBOX était utilisé seul, la meilleure combinaison de paramètres permet d'obtenir un pourcentage de succès de 66% pour un pourcentage d'échecs de 35%.

Si on prend les meilleurs résultats obtenus lorsque SHOEBOX est utilisé avec une matrice de scores, ce pourcentage de succès n'atteint, dans le meilleur des cas, que 60% pour un pourcentage d'échecs se situant à 35%.

Il est évident que l'explication évoquée au point III.3.1.5 reste toujours valable : les distances mesurées grâce à SHOEBOX auront plus de poids dans le calcul du score global que les distances mesurées grâce aux matrices.

III.4.3. Conclusion.

Dans cette troisième partie des résultats, nous avons essayé d'associer SHOEBOX à d'autres mesures de similarité. Nous avons ainsi pu voir que, comme nous l'avions évoqué dans le point III.2, la méthode SHOEBOX semble avoir un pouvoir discriminatoire trop faible pour détecter certains appariements. Les résultats des alignements prenant en compte DFK et RMS avec ou sans SHOEBOX ont montré que ce dernier semblait « imposer ses erreurs ».

Ce phénomène a pu être expliqué facilement par le comportement des distances calculées avec la méthode SHOEBOX. En effet, elles peuvent varier dans un rapport allant de 1 à 1000, ce qui donne beaucoup plus de poids à la méthode SHOEBOX qu'aux autres.

Jusque maintenant, la seule solution à ce problème réside dans la fixation de bornes. Mais comme nous l'avons vu et à cause des moindres performances de SHOEBOX par rapport aux autres mesures de distances, ces bornes semblent uniquement destinées à éviter de tenir compte des distances mesurées par SHOEBOX.

Les résultats obtenus lorsque SHOEBOX était allié à différentes matrices de scores n'ont fait que corroborer cette idée.

IV. Conclusions générales et perspectives.

La modélisation par comparaison de séquences semblent être aujourd'hui une alternative intéressante aux techniques physiques de résolution de structures. Elle permet d'associer à une protéine nouvellement séquencée une autre protéine (ou famille de protéines) de structure connue. Ces associations sont basées sur des similarités de séquences traduisant une ressemblance structurelle (Chothia et Lesk, 1986).

Pour mettre en oeuvre cette méthode de modélisation, il est nécessaire de disposer de programmes de comparaisons de structures, rapides et performants. Or, il existe peu de techniques capables de réaliser un alignement correct de plusieurs structures en un court laps de temps.

Le programme HOMOLOGY, par exemple, utilisé pour effectuer la superposition des structures lors de la réalisation de l'alignement de référence, est un logiciel complexe pour lequel une utilisation correcte demande un apprentissage important. Ce programme ne réalise que des superpositions de structures deux à deux et exige un contrôle visuel, ce qui introduit une part de subjectivité dans l'alignement final. De plus, HOMOLOGY est un logiciel coûteux, nécessitant un ordinateur graphique puissant.

Quant à lui, SHOEBOX, développé au sein du programme MATCH-BOX, est une méthode originale basée sur une comparaison de forme de segments protéiques. C'est une technique rapide, automatique et peu exigeante au point de vue du matériel. SHOEBOX réalise des alignements multiples pouvant impliquer jusqu'à 50 structures (40 000 résidus), il semblait donc pouvoir apporter une solution intéressante au problème de l'étude systématique de banques de structures.

Nous avons d'abord pu évaluer l'efficacité de la mesure de la forme du segment lorsque seuls les atomes de la chaîne principale étaient considérés. Il apparaît que l'utilisation de SHOEBOX dans cette configuration permet de détecter les segments dont le squelette possède une configuration identique. Cela signifie que tous les segments appartenant à une même structure secondaire seront caractérisés par une forme semblable. Nous avons également vu que, si on excepte le cas des courts motifs structuraux, la méthode SHOEBOX ne permet pas de retrouver l'appariement correct dans une région de même conformation. Lorsqu'on incorpore les atomes des chaînes latérales au calcul de la forme du segment, on n'observe pas d'augmentation du pouvoir discriminatoire de SHOEBOX. En effet, si dans certains cas, l'apport des chaînes latérales est positif, dans d'autres, il apparaît comme neutre ou même négatif. Une observation plus minutieuse de la variation des distances mesurées au cours des balayages nous ont permis de comprendre ces différentes situations :

- dans le cas où les séquences des segments comparées sont particulièrement conservées (c'est-à-dire que la nature des chaînes latérales est fort semblable), il est logique que ces segments aient des formes très proches.
- dans le cas contraire où des substitutions importantes ont eu lieu (c'est-à-dire que la nature des chaînes latérales change fortement), les segments ont évidemment des formes différentes. L'utilisation de ces chaînes latérales dans le calcul de la forme du segment ne permet donc pas de sélectionner l'appariement correct. Pour les courts motifs structuraux, des chaînes latérales non conservées peuvent même « masquer » les ressemblances du squelette.

La méthode SHOEBOX utilisée seule ne permettant pas de réaliser un alignement correct, l'étape suivante vise à l'utiliser avec d'autres mesures de distances. Le but recherché était d'allier à SHOEBOX une mesure permettant de choisir l'appariement correct parmi les régions de conformation semblables. Nous avons dû à nouveau constater le faible pouvoir discriminatoire de SHOEBOX. De plus, il apparaît que la distance calculée par SHOEBOX a une importance prépondérante dans le calcul du score global. Ce qui signifie que des erreurs dues à l'utilisation de SHOEBOX se répercuteront le plus souvent sur le score global.

Pour résoudre ce problème, la seule solution semble être la fixation de bornes pour limiter la variation de la distance entre une valeur inférieure et une supérieure. Cependant, fixer des bornes rend négligeable l'apport des distances mesurées par la méthode SHOEBOX.

Les différentes expériences réalisées avec SHOEBOX ont pu montrer quelles semblaient être les limites de cette méthode. En effet, si elle permet d'aligner correctement les régions pour lesquelles la séquence est conservée, les appariements impliquant des chaînes latérales trop différentes ne peuvent être détectés.

Dans le cadre de ce travail, nous avons essayé de proposer une explication aux résultats peu satisfaisants obtenus par l'utilisation de la méthode SHOEBOX. Il serait également nécessaire de vérifier ces conclusions par la réalisation d'expériences supplémentaires sur d'autres groupes de protéines.

Un autre champ d'application éventuel de la méthode SHOEBOX peut être celui de la classification des structures secondaires mais il sera nécessaire de déterminer si la méthode SHOEBOX peut apporter des résultats concluants dans ce domaine.

V. Bibliographie.

Abagyan R.A. and Maiorov V.N.
A simple qualitative representation ofpolypeptide chain folds : comparison of protein tertiary structures.
J. Biomol. Struct. Dynam. 5 (1988) 1267-1279.

Afinsen C.B.
Principals that govern the folding of protein chains.
Science 181 (1973) 223-230.

Afinsen C.B. and Sheraga H.A.
Experimental and theoritical aspects of protein folding.
Adv.Protein Chem. 29 (1975) 239-352.

Altschul S.F. and Lipman D.J.
Protein database searches for multiple alignment.
Proc. Natl. Acad. Sci. USA 87 (1990) 5509-5513.

Barton S.F. and Sternberg M.J.E.
A strategy for the rapid multiple alignment of protein sequences-confidence levels from tertiary structure comparisons.
J. Mol. Biol. 198 (1987) 327-337.

Benner S.A.
Predicting de novo the folded structure of proteins.
Curr. Opin. Struct. Biol. 2 (1992) 402-412.

Bernstein F.
The protein data bank : a computer-based archival file for macrmolecular structures.
J. Mol. Biol. 112 (1977) 535-542.

Blundell T.L. and Doolittle R.F.
Sequences and topology-An inverse approach to the old folding problem.
Curr. Opin. Struct. Biol. 2 (1992) 381-383.

Blundell T.L. and Johnson M.S.
Catching a common fold.
Prot. Science 2 (1993) 877-883.

Bowie J.U., Lüthy R. and Eisenberg D.
A method to identify protein sequences that fold into a known three-dimensional structure.
Science 235 (1991) 164-170.

Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan S. and Karplus M.
CHARMM : a program for macromolecular energy, minimization and dynamics calculations.
J. Comp. Chem. 4 (1993) 187-217.

Browne W.J., North A.C.T., Phillips D.C., Brew K., Vanaman T.C. and Hill R.L.
A possible three-dimensional structure of bovine alpha-albumin based on that of hen's egg white lysozyme.
J. Mol. Biol. 42 (1969) 65-86.

Chothia C.
Conformation of twisted b-pleated sheets in proteins.
J. Mol. Biol. 75 (1973) 295-302.

Chothia C.
One thousand families for the molecular biologist.
Nature 357 (1992) 543-544.

Chothia C. and Lesk A.M.
The relation between the divergence of sequence and structure in proteins.
EMBO J. 5 (1986) 823-826.

Chou P.Y. and Fasman G.D.
Conformational parameters for amino acids in helical, b-sheet and random coil regions calculated from proteins.
Biochem. 13 (1974) 211-221.

Corpet F.
Multiple sequence alignment with hierarchical clustering.
Nucl. Ac. Res. 16 (1988) 10881-10890.

Creighton T.E.
in : Proteins, W. H. Freeman, New York (1984)

Dauber-Osguthorpe P., Roberts V.A., Osguthorpe D.J., Wolff J., Genest M. and Hagler A.T.
Structure and energetics of ligand binding to proteins : *Eschericia coli* dihydrofolate reductase-timethoprim, a drug-receptor system.
Proteins Struct. Funct. and Genet. 4 (1988) 31-47.

Dayhoff M.O., Eck R.V. and Park C.M.
9. A model of evolutionary change in proteins.
Atlas of protein seq. struct. 5 (1972) 89-99.

Depiereux E. and Feytmans E.
Simultanuous and multivariatealignment of protein sequences-correspondence between physicochemical profiles and structurally conserved regions (SCR).
Protein Engng 4 (1991) 603-613.

Depiereux E. and Feytmans E.
MATCH-BOX : a fundamentally new algorithm for the simultaneous alignment of several protein sequences.
CABIOS 8 (1992) 501-509.

Depiereux E. and Feytmans E.
Elaboration de matrices de scores pour l'alignement de séquences de protéines sur base de similarités évaluées dans des structures semblables.
Biométrie-Praximétrie 34-1 (1994) 13-34.

Diamond R.
A note on the rational superposition.
Acta Cristallogr. A44 (1988) 211-216.

Diamond R.
On the multiple simultaneous superposition of molecular structures by rigid body transformations.
Protein Sci. 1 (1992) 1279-1287.

Efinov A.V.
Standard structures in proteins.
Prop. Biophys. Molec. Biol. 60 (1993) 201-239.

Ellis R.J. and Hemmingoen S.M.
Molecular chaperones : proteins essential for the biogenesis of some macromolecular structures.
Trends Biochem. Sci. 14 (1989) 339-342.

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

Feng D.F., Johnson M.S. and Doolittle R.F.
Aligning amino acid sequences : comparison of commonly used methods.
J. Mol. Evol. 21 (1985) 112-125.

Fischel-Ghodsian F., Mathiowitz G. and Smith T.F.
Alignment of protein sequences using secondary structure : a modified dynamic programming method.
Protein Engng 3 (1990) 577-281.

Fitch W.M. and Margoliash E.
Construction of phylogenetic trees.
Csience 15 (1967) 279-284.

Flores T.P., Orengo C.A., Moss D.S. and Thornton J.M.
Comparison of conformational characteristics in structurally similar protein pairs.
Protein Sci. 2 (1993) 1811-1826.

Garnier J., Osguthorpe D.J. and Robson B.
Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.
J. Mol. Biol. 120 (1978) 97-120.

George D.G., Barker W.C. and Hunt L.T.
Mutation data matrix and its uses.
Methods Enzymol. 183 (1990) 333-351.

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

Greer J.
Comparative model building of the mammalian serine proteases.
J. Mol. Biol. 153 (19814) 1027-1042.

Grindley H.M., Artymiuk P.J., Rice D.W. and Willet P.
Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm.
J. Mol. Biol. 229 (1993) 707-721.

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

Henikoff S. and Henikoff J.G.
Performance evaluation of amino acid substitution matrices.
Proteins Struc. Funct. Genet. 17 (1993) 49-61.

Henrissat B., Claeyssens M., Tomme P., Lemesle L. and Mormon J.-P.
Cellulase families revealed by hydrophobic cluster analysis.
Gene 8 (1989) 81-95.

Holm L. and Sander C.
protein structure comparison by alignement of distance matrices.
J. Mol. Biol. 233 (1993) 123-138.

Hoppe-Seyler F.
Über die chemischen und optischen eigenschaften des blutfarbstoffs.
Virchows Arch. 29 (1864) 233-235.

Johnson M.S., Overington J.P., Edwards Y., May A.C.W. and Rodionov M.A.
The comparison of structures and sequences : alignment, searching and detection of common fold.
Proc. 27th Ann. Hawaii Intern. Conf. System Sci. (1994) 296-305.

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

Kabsch W. and Sander C.
Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical feautures.
Biopolym. 22 (1983) 2577-2637.

Kidera A., Konishi Y., Oka M., Aoi T. and Scheraga H.A.
Statistical analysis of the physical properties of the 20 naturally occuring amino acids.
J. Prot. Chem. 4 (1985) 23-53.

Kikuchi T.
Similarity between average distance maps of structurally homologous proteins.
J. Protein Chem. 11 (1992) 305-320.

Kühne W.
Über das verhalten verschiedener organisirter und sogenannter ungeformter fermente. Über das trypsin (enzym des pankreas).
FEBS Lett. 62 (1976) E3-E7.

Lesk A.M.
A toolkit for computational molecular biology. II. On the optimal superposition of two sets of coordinates.
Acta Crystallogr. A42 (1986) 110-113.

Lim V.I.
Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure.
J. Mol. Biol. 88 (1974) 857-872.

Lipman D.J., Altschul S.F. and Kececioglu J.D.
A tool for multiple sequence alignment.
Proc. Natl. Acad. Sci. USA 86 (1989) 4412-4415.

Lipman D.J. and Pearson W.R.
Rapid and sensitive protein similarity searches.
Science 227 (1985) 1435-1441.

McLachlan A.D.
Test for comparing related amino-acid sequences. Cytochrome c and c_{551}.
J. Mol. Biol. 61 (1971) 409-424.

McLachlan A.D.
A mathematical procedure for superimposing atomic coordinates of proteins.
Acta Crystallogr. A28 (1972) 656-657.

McLachlan A.D.
Gene duplication in the structuralevolution of chymotrypsin.
J Mol. Biol. 128 (1979) 97-102.

McLachlan A.D.
Rapid comparison of protein structures.
Acta Crystallogr. A38 (1982) 871-873.

Mezei M.
A heuristic procedure for the detection of locally similar substructure of two equivalent structures.
Proyein Engng 7 (1994) 331-333.

Mitchell E.M., Artymink P.J., Rice D.W. and Willet P.
Use of techniques derived from graph theory to compare secondary structure motifs in proteins.
J. Mol. Biol. 212 (1990) 151-168.

Momany F.A., McGuire R.F., Burgess A.W. and Scheraga H.A.
Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occuring amino acids.
J. Phys. Chem. 79 (1975) 2361-2381.

Monteglione G.T. and Scherada H.A.
Formation of local structuresin protein folding.
Acts Chem. Res. 22 (1989) 70-76.

Murata M., Richardson J.S. and Sussman J.L.
Simultaneous comparison of three protein sequences.
Proc. Natl. Acad. Sci. USA 82 (1985) 3073-3077.

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

Némethy G., Pottle M.S. and Scherada H.A.
Energy parameters in polypeptides. 9. Udating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occuring amino acids.
J. Phys. Chem. 87 (1983) 1883-1887.

Némethy G. and Sheraga H.A.
Protein folding.
Qt. Rev. Biophys. 10 (1977) 239-352.

Niefind K. and Schomburg D.
Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles.
J. Mol. Biol. 219 (1991) 481-497.

Pauling L., Grey R.B. and Brauson H.R.
Configuration of polypeptide chains with favored orientations around single-bonds : two new pleated sheet.
Proc. Natl. Acad. Sci. USA 37 (1951) 205-211.

Pepperell C.A. and Willett P.
Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances.
J. Comput.-Aided mol. Design 5 (1991) 455-474.

Rao S.T. and Rossman M.G.
Comparison of super-secondary structures in protein.
J. Mol. Biol. 76 (1973) 241-256.

Rawlings C.J., Taylor W.R., Nyakairu J., Fox J. and Sternberg M.J.E.
reasoning about protein topology using the logic programming language PROLOG.
J. Mol. Graph 3 (1985) 151-157.

Remington S.J. and Matthews B.W.
A systematic approach to the comparison of protein structures.
J. Mol. Biol. 140 (1980) 77-99.

Richards F.M. and Kundrot C.E.
Identification of structural motifs from protein coordinate data : Secondary structure and first level supersecondary structure.
Proteins : Struct. Funct. Genet. 3 (1988) 71-84.

Richardson J.S.
The anatomy and toxonomy of protein structure.
Adv. Protein Chem. 34 (1981) 167-339.

Risler J.L., Delorme M.O., Delacroix H. and Henaut A.
Amino acid substitutions in stucturally related proteins : a pattern recognition approach. Determination of a new efficient scoring matrix.
J. Mol. Biol. 204 (1988) 1019-1029.

Rossman M.G. and Argos P.
Exploring structural homology of proteins.
J. Mol. Biol. 105 (1976) 75-95.

Rost B., Schneider R. and Sander C.
Progress in protein structure prediction?
Trends Biochem. Sci. 18 (1993) 120-123.

Sali A. and Blundell T.L.
Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming.
J. Mol. Biol. 212 (1990) 403-428.

Scherada H.A., Konishi Y. and Ooi T.
Multiple pathways for regenerating ribonuclease A.
Adv. Biophys. 18 (1984) 21-41.

Schultze-Kremer S. and king R.D.
IPSA-Inductive protein Structure Analysis.
protein Engn 5 (1992) 377-390.

Sheridan R.P., Dixon J.S. and Venkataraghavan R.
Generating plausible protein folds by secondary structure similarity.
Int. J. Pept. Protein Res. 25 (1985) 132-143.

Sternberg M.J.E. and Thornton J.M.
Prediction of protein structures from amino acid sequence.
Nature 271 (1978) 12-20.

Subbaroo N. and Haneef I.
Defining topological equivalences in macromolecules.
Protein Engng 4 (1991) 877-884.

Subbiah S. and Harisson S.
A method for multiple sequence alignment witn gaps.
J. Mol. Biol. 209 (1989) 539-548.

Sutcliffe M.J., Haneef I., Carney D. and Blundell T.L.
Knowledge based modeling of homologous proteins, part I : three-dimensional framework derived from the simultaneous superposition of multiple structures.
Protein Engng 1 (1987) 377-384.

Takahashi K. and Gô N.
Conformational classification of short backbone fragments in globular proteins and its use for coding backbone conformations.
Biophys. Chem. 47 (1993) 163-178.

Taylor W. and Orengo C.
Protein structure alignement.
J. Mol. Biol. 208 (1989) 1-22.

Unger R., Harel D., Wherland S. and Sussman J.L.
A 3D building blocks approach to analysing and predicting structure of proteins.
Proteins 5 (1989) 355-373.

Unger R. and Sussman J.L.
The importance of short structural motifs in protein structure analysis.
J. Comp. Mol. Design 7 (1993) 457-472.

Venkatachalam C.M.
Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units.
Biopolymers 6 (1968) 1425-1436.

Vriend G. and Sander C.
Detection of common three-dimensional substructures in proteins.
Proteins 11 (1991) 52-58.

Weiner S.J., Kollman P.A., Case D.A., Singh U.C., Ghio C., Alagona G., Profeta S.J. and Weiner P.
A new force field for molecular mechanical simulation of nucleic acids and proteins.
J; Am. Chem. Soc. 106 (1984) 765-784.

Weiner S.J., Kollman P.A., Nguyen D.T. and Case D.A.
An all atom force field for simulations of proteins and nucleic acids.
J. Comp. Chem. 7 (1986) 230-252.

Wilbur W.R. and Lipman D.J.
Rapid similarity searches of nucleic acid and protein data banks.
Proc. Natl. Acad. Sci. USA 80 (1983) 526-535.

Woodcock S., Mornon J. and Henrissat B.
Detection of secondary structure elements in proteins by hydrophobic cluster analysis.
Protein Engng 5 (1992) 629-635.

Zhu Z., Sali A. and Blundell T.L.
A variable gap penalty function and feature weights for protein 3-D structure comparisons.
Protein Engng 5 (1992) 43-51.