

# Chapitre 3

## Apprentissage de la structure locale du squelette protéique

### 3.1 Objectif

L'objectif principal de cette thèse est l'élaboration et l'utilisation d'un alphabet structural dans une méthode de prédiction à partir de la séquence. Dans cette première partie de ma thèse, je développerai donc la méthode d'apprentissage des Blocs Protéiques (BPs). Ils permettent une description statistiquement représentative des conformations d'un ensemble de résidus consécutifs. Différent aspects de leur validation sur un plan structural seront détaillés. Dans le chapitre suivant, je développerai l'approche mise en place pour la prédiction locale de ces blocs à partir de la séquence.

Pour résumer ce chapitre, une nouvelle méthodologie a été mise au point pour obtenir un alphabet structural protéique. Elle se base sur les angles dièdres du squelette polypeptidique et procède par un apprentissage en deux étapes; une première non supervisée et une seconde qui tient compte des transitions qui existent entre blocs protéiques, d'une manière similaire à la technique des Chaînes de Markov Cachées (CMC) [149]. Ces blocs ont été conçus pour assurer une bonne approximation de la structure protéique. J'ai donc testé différentes séries ayant des nombres variables de blocs protéiques. La série qui a été conservée est celle dont le nombre de blocs protéiques répond au mieux à l'approximation de la structure protéique et à un taux de prédiction à partir de la séquence correct.

## 3.2 Méthode d'apprentissage

La méthode d'apprentissage que nous avons utilisée possède une certaine similitude avec celle de Schuchhardt et collaborateurs [175]. Il s'agit d'une méthode d'apprentissage proche des cartes auto-organisées de Kohonen [109].

Seront développés ici le type d'information prise en compte, avec la base de données et le codage des protéines en fonction des angles dièdres, la fonction de similitude entre deux fragments (pour savoir s'ils sont proches ou non), les deux étapes de la méthode d'apprentissage (non supervisée puis tenant compte des transitions) et enfin les critères permettant l'obtention d'un nombre différent de blocs protéiques pour chaque série.

### 3.2.1 Base de données

La base de données protéique est un critère primordial pour comprendre la relation existant entre la structure tridimensionnelle et la séquence (cf paragraphe 2.3, *Les alphabets structuraux*). Actuellement, la base de données des structures protéiques, la PDB (Protein Data Bank [6]) est composée de plus de 10 000 protéines. Conserver l'ensemble des protéines de la PDB reviendrait à travailler principalement sur quelques grands groupes, les plus "aisément" cristallisables (les plus nombreux dans la base alors), comme les lysozymes, ce qui biaiserait les analyses (en particulier statistique). Ne conserver que des protéines ayant une faible homologie de séquence permet de limiter ce type d'erreur. La base utilisée est composée de 342 protéines ayant moins de 25% d'identité de séquence [84, 83].

Les structures secondaires ont été assignées au préalable par une méthode consensuelle ([31] et cf. paragraphe 2.1.2.1) de trois algorithmes DSSP [100], P-CURVE [181] et DEFINE [154]. Ensuite, les protéines ont été classées suivant la nomenclature de Michie et collaborateurs [127] qui considèrent 4 classes de protéines : tout  $\alpha$ , tout  $\beta$ ,  $\alpha/\beta$  et non-classées.

### 3.2.2 Description des protéines

Nous avons choisi d'utiliser comme descripteur de la structure des protéines les angles dièdres  $\phi$  et  $\psi$ . (cf. figure 3.1). La définition des angles dièdres est donnée au paragraphe 2.1.2.

L'angle  $\omega$  n'a pas été pris en compte, car il reste particulièrement constant autour de  $180^\circ$ . Sur la PDB de juillet 2000, moins de 0,5 % des résidus étaient en dehors de l'intervalle classique

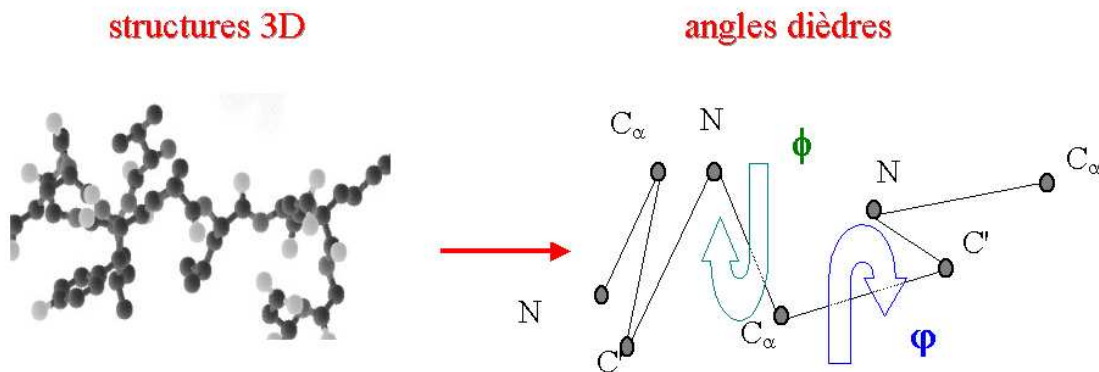


FIG. 3.1 – Schéma représentant le calcul des angles  $\phi$  et  $\psi$  à partir de la structure tridimensionnelle.

$[180^\circ-10^\circ; 180^\circ+10^\circ]$ . En outre, l'angle  $\omega$  suit directement les variations des angles  $\phi$  et  $\psi$  [125]. Selon la zone de Ramachandran  $[\phi, \psi]$ , l'angle  $\omega$  est statistiquement plus présent dans certaines zones. Leur utilisation aurait tendance à surestimer ce paramètre, 99,5 % ne jouant alors aucun rôle, les 0,5 % restant deviendraient trop important au regard de leur faible nombre.

La conformation des protéines a donc été décrite via leurs angles  $\phi$  et  $\psi$ . Une protéine de longueur  $L$  est ainsi traduite en un vecteur de longueur  $2 \times (L-2)$  (le premier  $\phi$  et le dernier  $\psi$  ne pouvant pas être calculés). L'intérêt d'utiliser des valeurs réelles des angles permet de ne pas avoir d'*a priori* particulier envers certaines zones préférentielles du diagramme de Ramachandran comme les zones répétitives et ainsi d'obtenir des blocs pertinents particulièrement au niveau des boucles.

### 3.2.3 Les fragments protéiques

Les 342 protéines de la base de données ont été ensuite découpées en fragment d'une longueur de  $M = 5$  résidus consécutifs. Cette taille est suffisante pour décrire les courtes hélices qui ont une taille de 4 résidus [113] ainsi que les petits feuillets de 3 acides aminés [31]. De plus, 5 résidus est une taille acceptable pour permettre l'établissement de liaisons hydrogènes qui sont importantes d'un point de vue structural (cf paragraphe 2.1.2.1). Il faut bien noter que les fragments sont chevauchants; ainsi une protéine de longueur  $L$  sera représentée par  $L-4$  fragments. La base de données comprend 86 628 fragments protéiques. Chaque fragment centré sur le carbone alpha ( $C\alpha_n$ ) représente donc les carbones alpha  $C\alpha_{n-2}, C\alpha_{n-1}, C\alpha_n, C\alpha_{n+1}$  et  $C\alpha_{n+2}$ . Lors de l'apprentissage, il sera représenté par un vecteur  $\mathbf{V}$  de  $2(M-1)$  résidus, soit 8 angles

dièdres ( $\psi_{n-2}, \phi_{n-1}, \psi_{n-1}, \phi_n, \psi_n, \phi_{n+1}, \psi_{n+1}, \phi_{n+2}$ ). Ainsi, un fragment (position  $i$ ) possède 4  $C_\alpha$  en commun avec le fragment suivant (position  $i + 1$ ), soit 6 angles dièdres en commun.

### 3.2.4 L'apprentissage

Les protéines ont été séparées en deux groupes distincts. Le premier, utilisé pour l'apprentissage, comprend 228 protéines, ce qui représente 2/3 de la base de données. Le second comprend les 114 protéines restantes qui serviront pour la validation finale. Le but de l'apprentissage est de définir une série de Blocs Protéiques (BPs) qui représentent au mieux les fragments protéiques de la base de données. Aussi, chaque bloc protéique a une longueur identique au fragment soit  $2(M-1)$  valeurs, 8 angles dièdres.

#### 3.2.4.1 Principe de l'apprentissage

La méthode utilise le principe des cartes auto-organisées de Kohonen (Kohonen Maps, ou Self-Organized Maps noté SOM [107, 109], cf. Annexe 1 et paragraphe 2.3.2.3). Elle procède par une lecture un certain nombre de fois (étape appelée cycle) de la base de donnée d'apprentissage. Dans la terminologie des SOMs, les neurones sont des classes d'objets. Leurs poids sont des informations moyennes associées à chaque neurone. Dans notre cas, il s'agit des blocs protéiques (*les neurones*) qui sont des vecteurs (*8 angles dièdres*).

Deux étapes d'apprentissage sont effectuées : la première consiste à apprendre les blocs protéiques seuls (par série de 5  $C_\alpha$ ); la seconde permet de renforcer les transitions entre les blocs comme dans les Chaînes de Markov Cachées (CMC, [149]). Cette seconde étape est effectuée pour tenir compte de l'architecture des protéines. Elle revient à optimiser la succession des blocs protéiques. Pour expliciter ceci d'un point de vue structural, cette méthode revient à faire qu'une partie centrale, régulière, d'une hélice  $\alpha$  assignée donc à un bloc "partie centrale  $\alpha$ " aille préférentiellement vers un bloc "sortie d'hélice  $\alpha$ ". Cette optique permet, en outre, de n'avoir qu'un nombre limité de blocs "sorties d'hélices  $\alpha$ ". D'un point de vue plus formel quand un fragment protéique (*exemple*) en position  $i$  dans une protéine est associé à un bloc protéique (*neurone*)  $j$ , le fragment (*exemple*) en position suivante  $i + 1$  devrait être associé à un bloc protéique (*neurone*)  $k$  qui n'est pas n'importe lequel des  $B$  blocs (*neurones*), mais un des quelques  $v$  blocs (*neurones*) qui suivent normalement le bloc (*neurone*)  $j$ . Cette approche

ne repose sur aucune hypothèse ou loi de distribution *a priori*, contrairement aux CMC.

### 3.2.4.2 Mesure de similitude

La première mesure à effectuer est de savoir si deux fragments, ou un fragment et un bloc, sont proches structuralement. Pour calculer cette similitude, j'ai utilisé la différence moyenne entre les valeurs angulaires (ou *RMSda* pour root mean square deviations on angular values [175]). Il s'agit d'une distance euclidienne entre deux vecteurs  $\mathbf{V}_1$  and  $\mathbf{V}_2$  :

$$RMSda(\mathbf{V}_1, \mathbf{V}_2) = \sqrt{\frac{\sum_{i=1}^{M-1} [\psi_i(\mathbf{V}_1) - \psi_i(\mathbf{V}_2)]^2 + [\phi_{i+1}(\mathbf{V}_1) - \phi_{i+1}(\mathbf{V}_2)]^2}{2(M-1)}}$$

avec  $\{\phi_i(\mathbf{V}_1), \psi_{i+1}(\mathbf{V}_1)\}$  (*resp.*  $\phi_i(\mathbf{V}_2), \psi_{i+1}(\mathbf{V}_2)$ ) les séries de  $(2M-1)$  angles dièdres pour  $\mathbf{V}_1$  (*resp.*  $\mathbf{V}_2$ ). Les différences entre les angles sont calculées modulo  $360^\circ$ .

Ainsi, durant la phase d'apprentissage, cette mesure est utilisée pour calculer la similitude entre un fragment et les différents blocs protéiques. La distance minimale permet d'associer un fragment à un bloc protéique.

### 3.2.4.3 Initialisation

Chaque bloc protéique  $PB_k$ , ou neurone, est initialisé par un vecteur  $\mathbf{W}(k)$  constitué de 8 angles dièdres choisis soit dans les zones denses du diagramme de Ramachandran (cf. figure 2.4b), soit tirés aléatoirement dans la base de données. Le nombre initial de blocs  $B$  a été fixé arbitrairement.

### 3.2.4.4 Apprentissage non supervisé

A chaque cycle, les fragments sont tirés aléatoirement. La figure 3.2 récapitule l'ensemble des deux étapes du processus d'apprentissage, la figure 3.3 pour la première et la figure 3.4 pour la seconde.

Dans un premier temps, l'apprentissage ne tient pas compte des transitions. (1) Un fragment est choisi aléatoirement dans la base de données d'apprentissage. (2) Son vecteur d'observation  $\mathbf{V}(m)$  est comparé à chacun des blocs protéiques  $PB_k$  à l'aide du critère du *RMSda*.  $B$  comparaisons sont donc calculées. (3) Le *RMSda* minimal correspond ainsi au bloc le plus proche

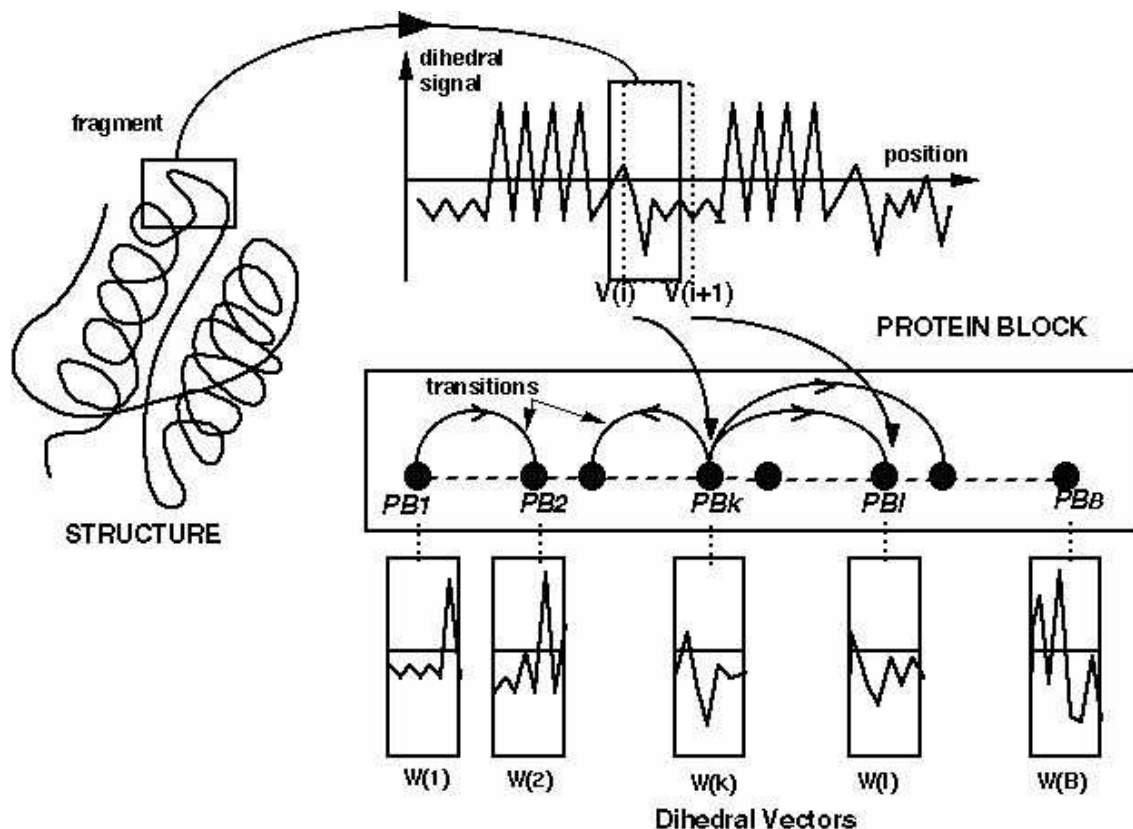


FIG. 3.2 – Schéma global sur la méthode d'apprentissage des blocs protéiques. En partant de la protéine, à gauche, sont représentés, la traduction de la structure protéique en angles dièdres, le tirage aléatoire d'un fragment protéique, puis le choix du bloc le plus proche structuralement, et enfin la deuxième phase où l'on recherche parmi les blocs ayant un RMSda faible, celui qui a le meilleur taux de transition par rapport au précédent fragment.

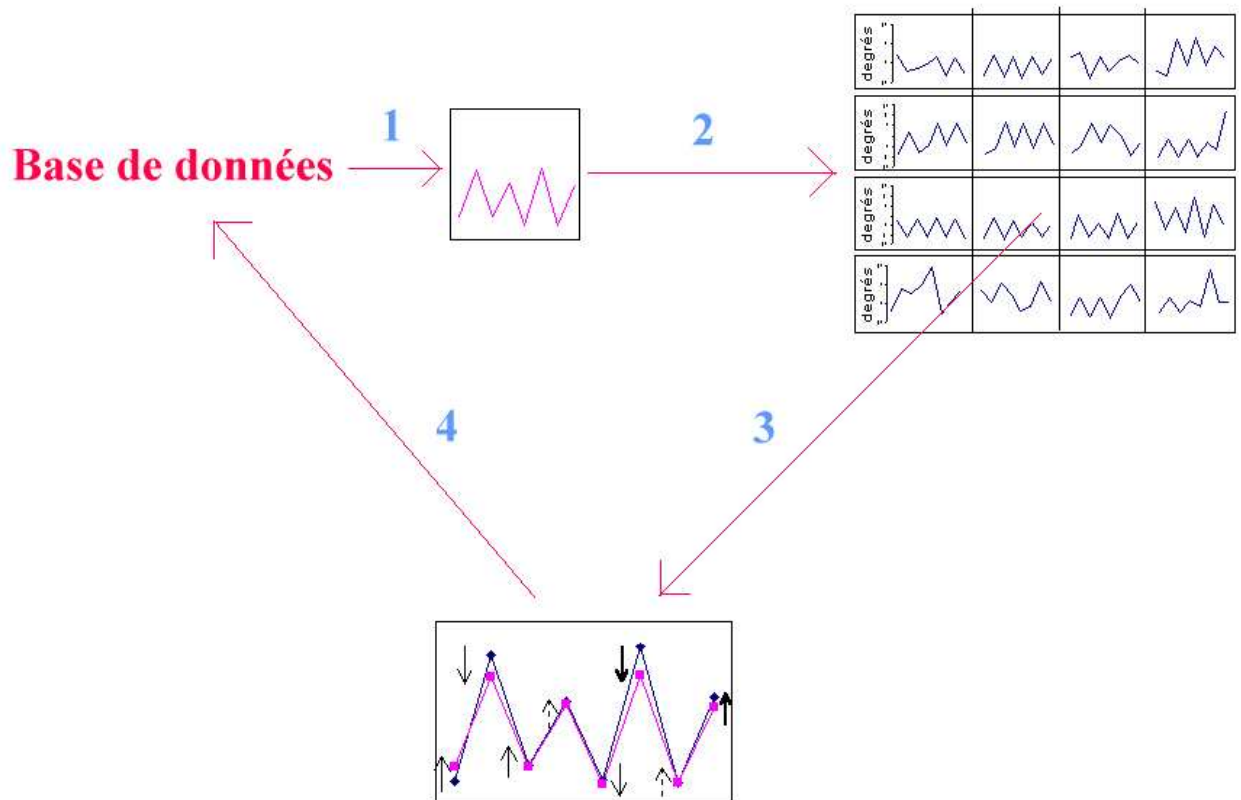


FIG. 3.3 – Première étape d'apprentissage. (1) Un fragment est choisi aléatoirement dans la base de données protéique. (2). Les B distances sont calculées avec les B blocs existants. (3). Le bloc le plus proche du fragment tiré aléatoirement est modifié légèrement pour lui ressembler. (4). Le processus recommence au (1).

structuralement du fragment. (4) Le vecteur  $\mathbf{W}(k)$  du bloc le plus proche est très faiblement modifié pour ressembler à celui présenté au réseau  $\mathbf{V}(m)$ :

$$\mathbf{W}(k) \leftarrow \mathbf{W}(k) + (\mathbf{V}(m) - \mathbf{W}(k)).\nu(c)$$

avec  $\nu(c)$  le coefficient d'apprentissage. Le symbole  $\leftarrow$  veut dire "la valeur de gauche est remplacée par celle calculée à droite". Ce coefficient, initialement pris faible, i.e.  $\nu_0 = 0.02$ , décroît pendant l'apprentissage.  $\nu(c)$  se modifie avec le nombre  $c$  de vecteurs d'observations déjà vus :

$$\nu(c) = \frac{\nu_0}{1 + \tau.c}$$

avec  $\tau$  fixé arbitrairement à  $1/N$ ,  $N$  représentant le nombre de vecteurs d'observations présents dans la base de donnée d'apprentissage. Dans notre cas,  $N = 56\,442$ . Ainsi,  $\nu(c)$  est divisé par deux après le premier passage complet de la base.

L'apprentissage est itératif, un certain nombre de cycles  $C$  ( $= 15$ ) est nécessaire pour définir des vecteurs optimaux  $\mathbf{W}$  associés aux blocs. Chaque fragment n'est utilisé qu'une seule fois par cycle.

### 3.2.4.5 Apprentissage tenant compte des transitions

Après  $C$  cycles d'apprentissage, une première série de blocs protéiques a été obtenue (cf. figure 3.4). Ces derniers sont utilisés pour recoder en terme de blocs protéiques l'ensemble de la base de données d'apprentissage. Ceci permet de calculer une matrice de transition entre les BPs en comptant les occurrences des paires de BPs consécutifs et en les transformant ensuite en fréquences.

Le principe est un peu similaire au précédent. (1) Une protéine est tirée au hasard. (2) En première position, la recherche du  $BP_r$ , le bloc réel est effectué à l'aide du critère du  $RMSda$  minimal, comme précédemment. Pour les autres fragments, le processus est un peu différent. Leur vecteur d'observation  $\mathbf{V}(m)$  est comparé à chacun des blocs protéiques  $PB_k$  à l'aide du critère du  $RMSda$  et  $B$  comparaisons sont donc calculées. Toutefois, le bloc vainqueur est choisi suivant des critères plus différents. (i)  $n$  blocs ayant un  $RMSda$  faible (inférieur à une valeur



donnée) sont sélectionnés. (ii) Le bloc protéique choisi est celui qui possède la fréquence de transition maximale entre le  $BP_r$  et  $BP_{r+1}$  pour les  $n$  blocs conservés. (4) Le processus est répété jusqu'au fragment C-terminal de la protéine. Ainsi, les transitions sont renforcées entre les différents blocs. À chaque cycle, la matrice de transition est réévaluée.

Au bout de  $C$  cycles d'apprentissage non supervisé, puis de  $C$  cycles d'apprentissage tenant compte des transitions, de nouveau  $C$  cycles d'apprentissage non supervisé sont effectués pour ne pas biaiser l'apprentissage.

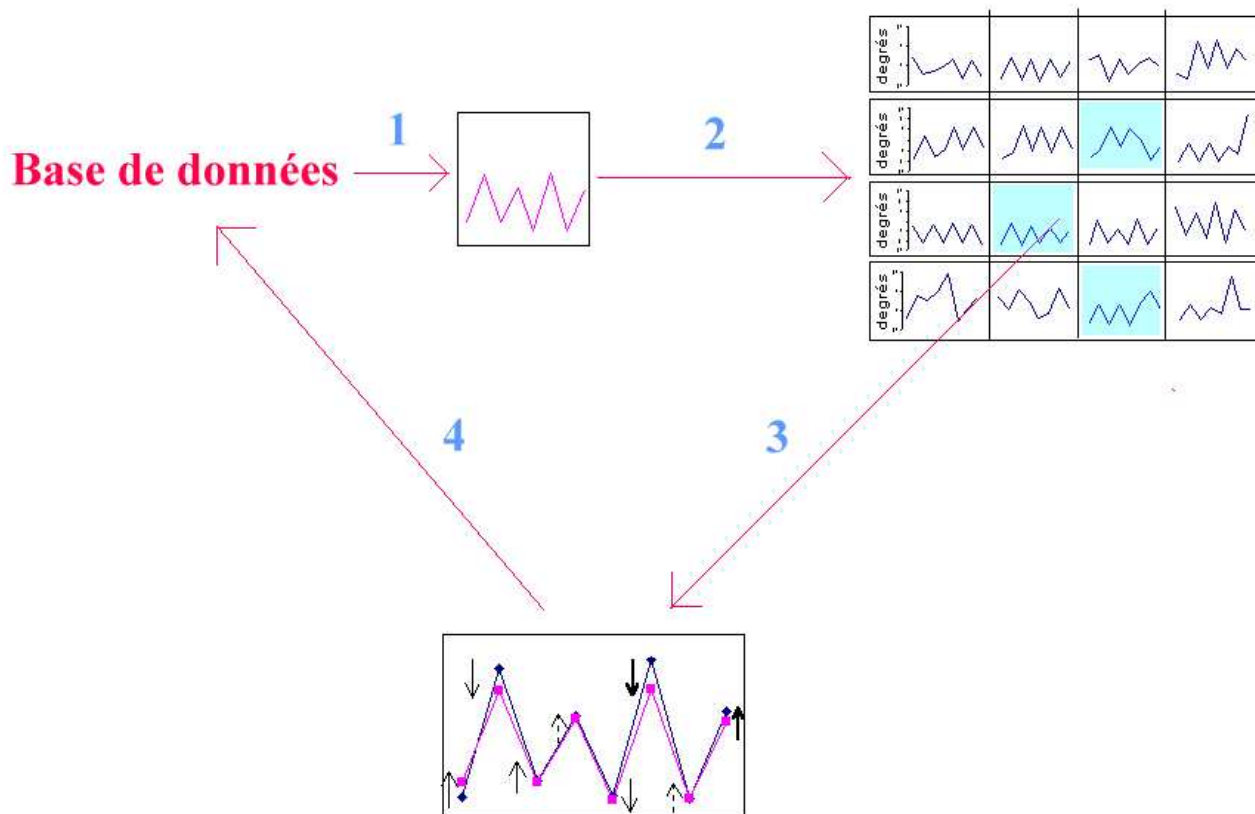


FIG. 3.4 – *Seconde étape d'apprentissage. (1) Un fragment est choisi aléatoirement dans la base de données protéique. (2). Les B distances sont calculées avec les B blocs existants. n blocs ayant un RMSda faible sont sélectionnés (les blocs mis en bleu clair). (3). Le bloc optimal est choisi parmi les n blocs, il représente la fréquence maximale de transition par rapport au bloc choisi précédemment (cf. le texte). Il est alors est modifié légèrement pour ressembler au fragment. (4). Le processus recommence au (1).*

### 3.2.4.6 Procédure de réduction du nombre de blocs protéiques

Partant d'un nombre particulièrement important de  $B$  ( $= 34$ ) blocs protéiques, nous avons décidé de réduire leur nombre pour posséder différentes séries ayant des nombres de blocs protéiques variables. Ces BPs seront utilisés pour la prédiction. Pour définir les blocs à éliminer,

nous avons décidé d'éliminer les blocs ayant de trop proches voisins. Pour définir le fait que 2 BPs sont proches, deux critères ont été utilisés. Tout d'abord, le fait qu'ils soient proches structuralement, ensuite que leurs transitions soient à peu près identiques. Pour cela, il suffit de calculer le *RMSda* des  $B$  PBs pris deux à deux, et de sélectionner les couples dont le *RMSda* est inférieur à une valeur définie. Ensuite, il suffit de faire une différence entre les fréquences de transition des deux PBs et si celle-ci est faible, cela veut dire qu'en effet les deux blocs sont pratiquement équivalents. Le bloc protéique le moins peuplé est alors éliminé.

Ayant réduit le nombre de blocs, il suffit de recommencer l'apprentissage avec le nouvel ensemble de blocs. Le processus s'arrête quand plus aucun bloc ne peut être éliminé avec les critères définis.

### 3.3 L'alphabet structural

Avec cette approche automatisée, le nombre de blocs passe de  $B = 34$  blocs à 22, puis 19, 16, 14, 12, 11 et enfin 10. Le *RMSda* moyen est au départ de  $25.4^\circ$ . Il augmente à  $28.5^\circ$  pour 22 blocs puis  $29.0^\circ$  pour 19 blocs et passe à  $30.0^\circ$  pour 16 blocs. Il se stabilise ensuite entre  $30^\circ$  et  $32^\circ$ . La première réduction est forte, plus d'une dizaine de blocs sont éliminés. Avec la diminution du nombre de blocs protéiques, l'approximation angulaire augmente fortement, puis se stabilise.

Toutefois, pour choisir le nombre optimal de BPs, il faut considérer en parallèle le pourcentage de bonne prédiction lié à chaque série de BPs. Celui-ci est discuté dans le paragraphe 4 (*"Prédiction de la structure locale en blocs protéiques"*).

#### 3.3.1 Description des blocs protéiques

La figure 3.5 montre le résultat de l'étape d'apprentissage avec les angles des 16 blocs. Le tableau 3.1 regroupe leurs valeurs exactes.

La figure 3.6 (visualisé avec le logiciel rasmol et sur la page de garde avec MOLSCRIPT [111]) montre pour chacun des 16 blocs protéiques, numérotés de  $a$  à  $p$ , des fragments superposés qui donnent ainsi une idée du type de repliement associé à chaque bloc. Les blocs ont été re-ordonnés suivant leurs fréquences de transition ainsi qu'en observant leur répartition dans les structures secondaires définies par la méthode de consensus [31]. Ces informations sont résumées dans le

| bloc | $\psi_{n-2}$ | $\phi_{n-1}$ | $\psi_{n-1}$ | $\phi_n$ | $\psi_n$ | $\phi_{n+1}$ | $\psi_{n+1}$ | $\phi_{n+2}$ |
|------|--------------|--------------|--------------|----------|----------|--------------|--------------|--------------|
| a    | 41,14        | 75,53        | 13,92        | -99,80   | 131,88   | -96,27       | 122,08       | -99,68       |
| b    | 108,24       | -90,12       | 119,54       | -92,21   | -18,06   | -128,93      | 147,04       | -99,90       |
| c    | -11,61       | -105,66      | 94,81        | -106,09  | 133,56   | -106,93      | 135,97       | -100,63      |
| d    | 141,98       | -112,79      | 132,20       | -114,79  | 140,11   | -111,05      | 139,54       | -103,16      |
| e    | 133,25       | -112,37      | 137,64       | -108,13  | 133,00   | -87,30       | 120,54       | 77,40        |
| f    | 116,40       | -105,53      | 129,32       | -96,68   | 140,72   | -74,19       | -26,65       | -94,51       |
| g    | 0,40         | -81,83       | 4,91         | -100,59  | 85,50    | -71,65       | 130,78       | 84,98        |
| h    | 119,14       | -102,58      | 130,83       | -67,91   | 121,55   | 76,25        | -2,95        | -90,88       |
| i    | 130,68       | -56,92       | 119,26       | 77,85    | 10,42    | -99,43       | 141,40       | -98,01       |
| j    | 114,32       | -121,47      | 118,14       | 82,88    | -150,05  | -83,81       | 23,35        | -85,82       |
| k    | 117,16       | -95,41       | 140,40       | -59,35   | -29,23   | -72,39       | -25,08       | -76,16       |
| l    | 139,20       | -55,96       | -32,70       | -68,51   | -26,09   | -74,44       | -22,60       | -71,74       |
| m    | -39,62       | -64,73       | -39,52       | -65,54   | -38,88   | -66,89       | -37,76       | -70,19       |
| n    | -35,34       | -65,03       | -38,12       | -66,34   | -29,51   | -89,10       | -2,91        | 77,90        |
| o    | -45,29       | -67,44       | -27,72       | -87,27   | 5,13     | 77,49        | 30,71        | -93,23       |
| p    | -27,09       | -86,14       | 0,30         | 59,85    | 21,51    | -96,30       | 132,67       | -92,91       |

TAB. 3.1 – Les 8 angles définissant les 16 blocs protéiques obtenus (notés en degrés).

tableau 3.3.

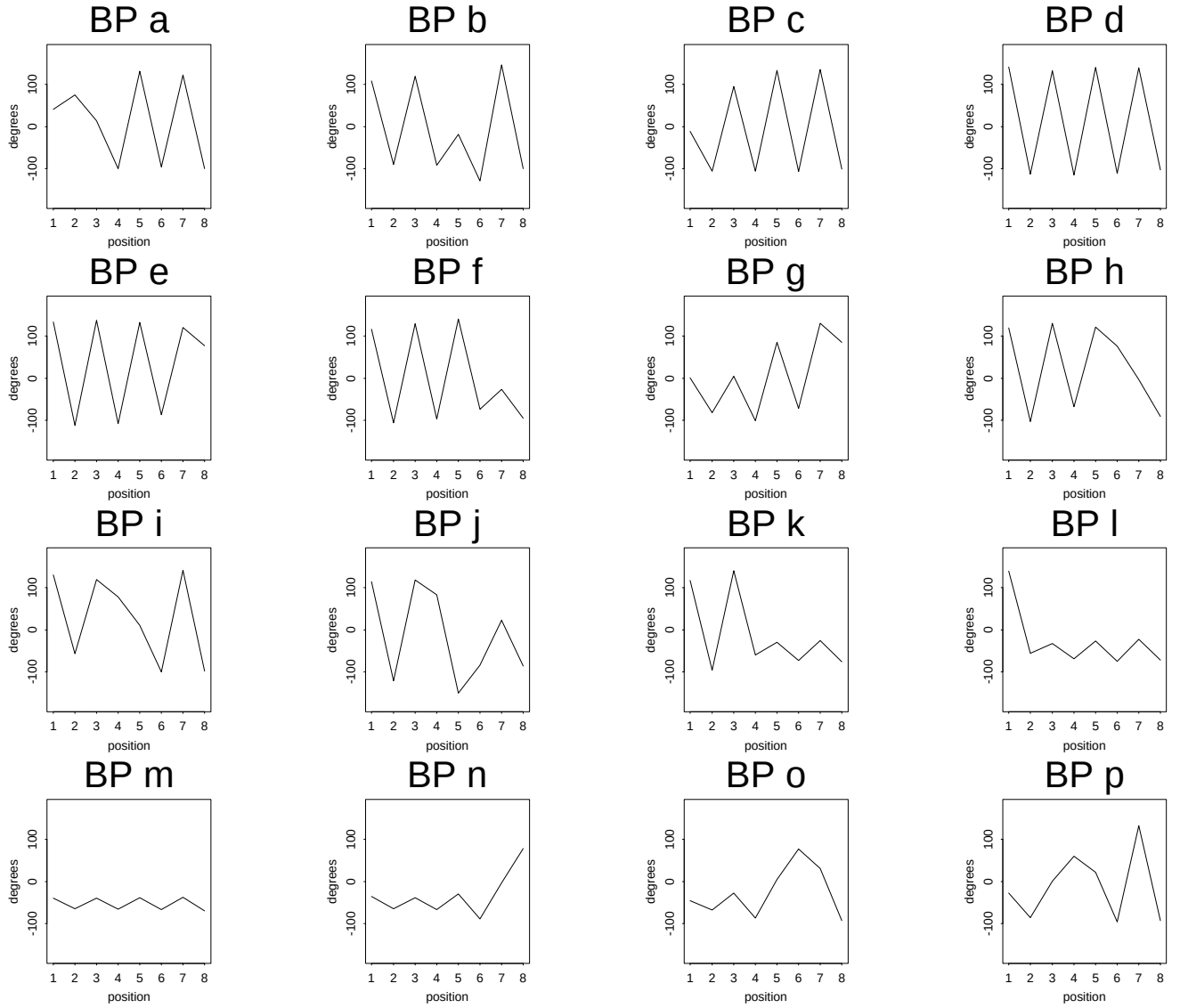


FIG. 3.5 – Angles des 16 Blocs Protéiques obtenus (en degrés).

### 3.3.2 Stabilité structurale des 16 blocs protéiques

#### 3.3.2.1 *RMSda* et *RMSd*

La qualité des blocs a été mesurée en utilisant les critères du *RMSd* et du *RMSda*. Pour le premier, chaque fragment protéique assigné à un bloc donné a été comparé à tous les autres fragments associés à ce même bloc. La moyenne des valeurs observées est récapitulée dans la seconde colonne du tableau 3.3. Ainsi, pour la totalité des blocs protéiques, excepté pour BPj, le *RMSd* est inférieur à 0,74 Å, sur 5 C $_{\alpha}$ . Il est intéressant de noter que les blocs caractéristiques des structures secondaires, BPm (partie centrale des hélices  $\alpha$ ) et BPd (pour les feuillets  $\beta$ ), ne

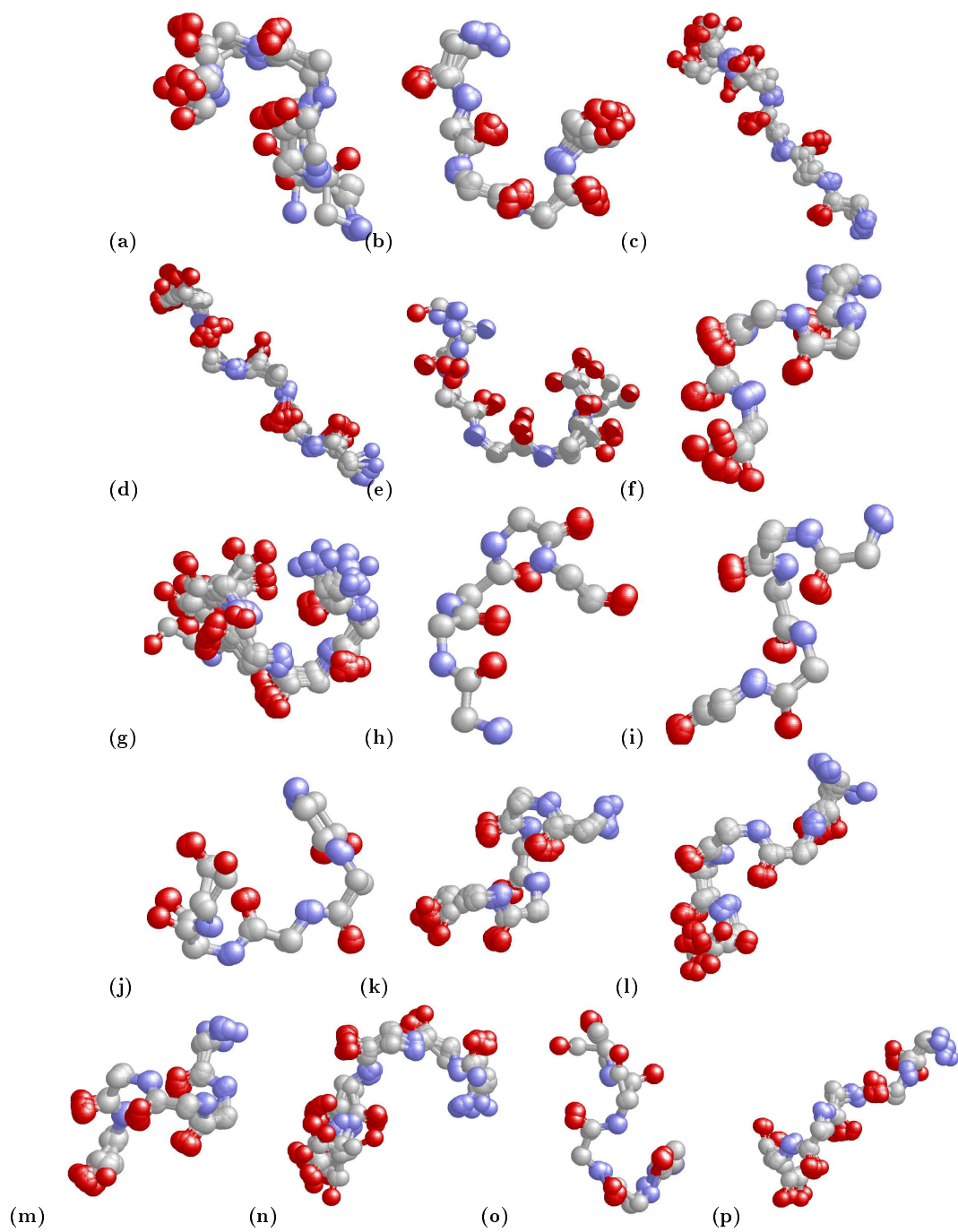


FIG. 3.6 – *Superposition de fragments pour les 16 blocs protéiques.*

sont pas les seuls à être bien approximatés, comme le bloc BP $p$  (0.46 Å) et de nombreux autres dont le *RMSd* est inférieur à 0,5 Å.

La moyenne des *RMSda* est de 30.0° pour l'ensemble des BPs. Seul le BP $m$  a une moyenne individuelle largement inférieure (15°), les autres sont tous compris entre 25° et 32°. Toutefois, il convient de noter la différence qui existe entre la définition de la moyenne des *RMSda* et de l'approximation angulaire des BPs. La moyenne des *RMSda* est la moyenne de chaque fragment protéique avec le bloc protéique qui lui est associé, or en chaque position (excepté pour les extrémités), chaque angle dièdre est codé non pas par 1 mais par 4 blocs protéiques, les fragments étant chevauchants. Ainsi, en faisant simplement la moyenne des angles des blocs protéiques présents en chaque position, 50 % des angles sont approximatés à moins de 21 % de la réalité. Ce point est par ailleurs discuté dans le paragraphe 3.3.2.6, où une nouvelle stratégie est proposée. En outre, la valeur du *RMSda* moyen est particulièrement touchée par quelques valeurs extrêmes qui ont un rôle important dans le calcul d'une moyenne.

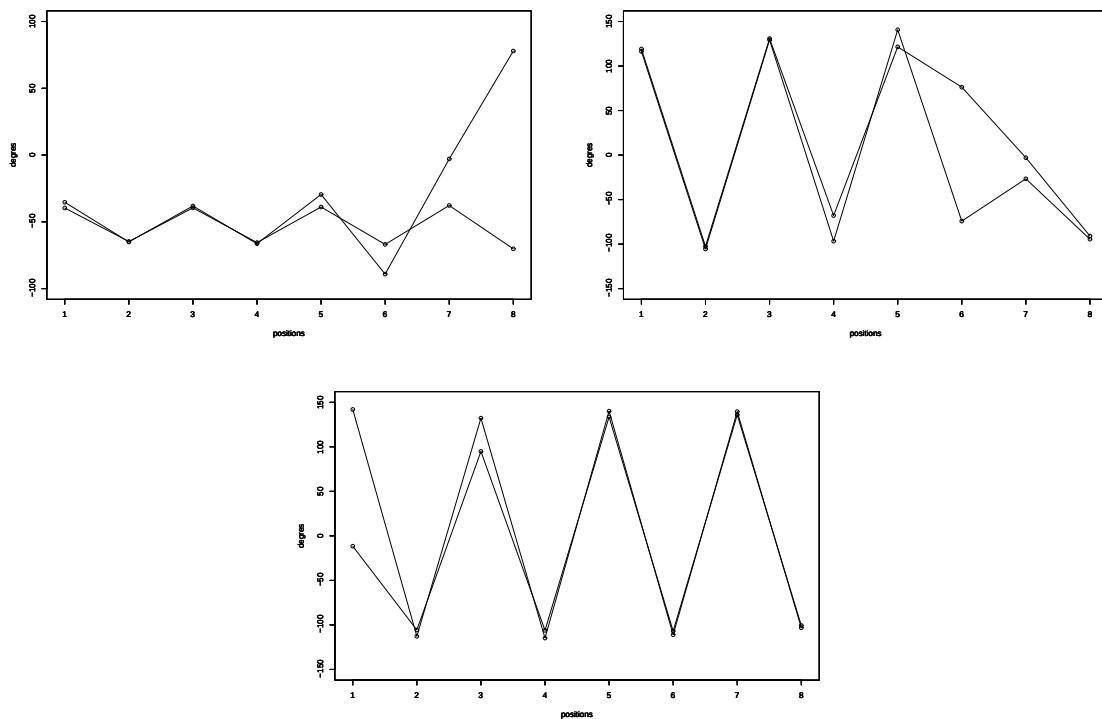


FIG. 3.7 – *Superposition des angles des blocs protéiques ayant le plus faible RMSda avec les blocs m et n (19,2°), f et h (19,5°) et c et d (19,8°)*

### 3.3.2.2 Différence structurale entre les blocs protéiques

Un des points importants est de s'assurer que les blocs sont bien distincts d'un point de vue structural. Les valeurs du *RMSd* calculé entre les blocs vont de 0,21 à 2,07 Å. Les blocs les plus proches sont les blocs protéiques *m* et *n* (0,21 Å), *f* et *h* (0,23 Å), *n* et *o* (0,24 Å), *c* et *d* (0,25 Å). Ces blocs sont répartis dans les même types de structures secondaires (cf. tableau 3.3). Le calcul du *RMSda* donne les même associations avec les blocs *m* et *n* (19,2°), *f* et *h* (19,5°) and *c* et *d* (19,8°). Toutefois ces *RMSd* faibles ne traduisent pas les différences réelles. Ainsi le bloc *n* va vers le BP *m* qui est une forme très répétitive; ils ont donc une partie commune importante, mais aussi une partie distincte. L'analyse individuelle des angles donne un meilleur éclairage sur ce point.

L'observation des différences angulaires entre ces paires de BPs (cf. figure 3.7) montrent la présence de 5 à 6 angles très proches (moins de 10° de différence), ce qui induit le *RMSda* et le *RMSd* faible; mais aussi 1 à 3 angles totalement différents (plus de 100° d'écart). Ce sont ces angles qui donnent leur spécificité au bloc. Il faut noter que le *RMSda* donne ici une information plus précise sur la distinction entre les blocs protéiques, ce qui corrobore les exemples donnés par Unger et collaborateurs (cf. paragraphe 2.3.2.1).

### 3.3.2.3 Sensibilité de l'assignation par le *RMSda*

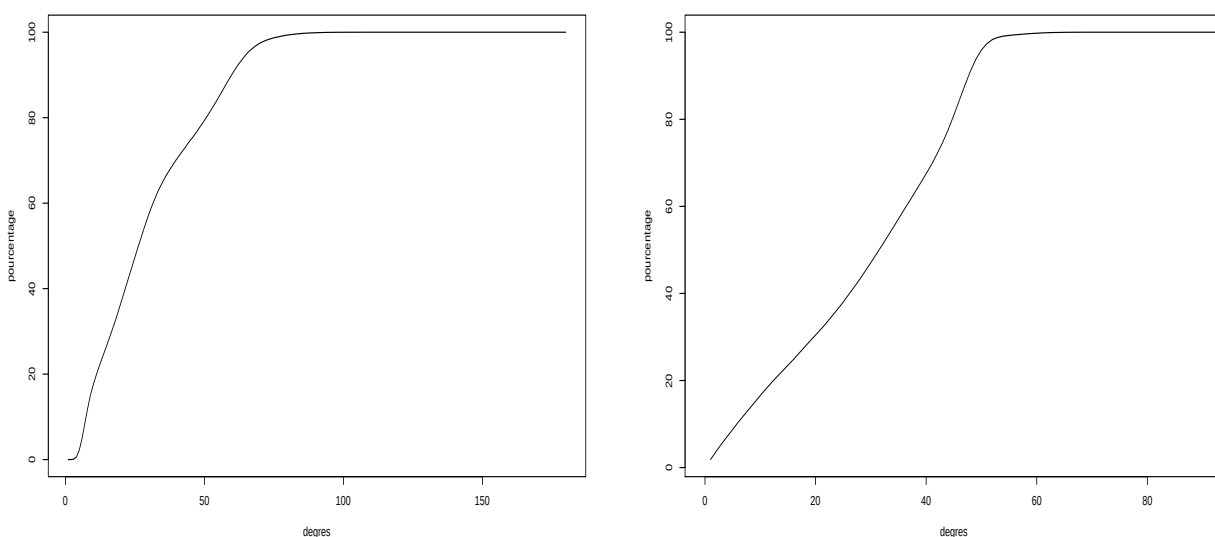


FIG. 3.8 – (a) Valeur du *RMSda* pour le premier bloc sélectionné. (b) Différence de valeur de *RMSda* entre les second et premier blocs sélectionnés.

Ayant vu que les blocs sont bien distincts entre eux et possèdent des spécificités angulaires importantes, un second point à observer est la qualité de l'assignation d'un fragment à un bloc. En effet, il faut vérifier que pour une majorité de fragments leur assignation à un bloc protéique est unique. En d'autre terme, il convient d'avoir le second bloc protéique le plus proche, le plus "lointain" possible. Pour évaluer ce critère, il convient donc d'analyser la répartition des valeurs des *RMSda* pour les blocs sélectionnés (blocs réels) et ensuite pour les blocs ayant le second *RMSda* le plus faible. La figure 3.8 récapitule (a) les distributions cumulées des valeurs d'assignations des fragments protéiques à leur bloc et (b) la différence entre le *RMSda* le plus faible (bloc réel) et le second plus faible. Ainsi, la distribution des *RMSda* associés au bloc réel montre que le *RMSda* moyen est faible. Une bonne différence de *RMSda* entre les deux meilleurs *RMSda* est observée. La confusion entre les blocs est donc faible. La différence moyenne est d'une valeur de 40°. Plus de 80% des blocs en second sont à plus de 20° de *RMSda*. La distinction entre le bloc "réel" et le second bloc le plus proche est donc bonne, et donc, ceci implique un codage en blocs protéiques de la structure protéique sans ambiguïté.

Pour évaluer la sensibilité de l'assignation des fragments par le critère du *RMSda*, une étude de dynamique moléculaire a été menée. Pour cela, des variations du squelette protéique de l'ubiquitine ont été simulées dans les gammes de facteurs de températures classiques (modifications des longueurs inter-atomiques et des angles de valence). Aucun changement d'assignation n'a été observé. L'assignation par le critère du *RMSda* est donc particulièrement stable.

### 3.3.2.4 Exemple de recodage

La figure 3.9 montre la protéine de liaison à l'ubiquitine (code PDB: 1aak) représentée en 3D par le logiciel MOLSCRIPT [111]. La figure 3.10a la montre recodée en blocs protéiques avec les variations de *RMSds* en chaque position (cf. 3.10b). Cette protéine est une protéine ( $\alpha/\beta$ ). Il est aisé de voir les successions de *BPd* et *BPm*. On peut aussi remarquer quatre séries de blocs *cfkl* associées à des boucles amenant à des hélices  $\alpha$ , ainsi que quatre séries *dfk*, deux *ehia* et deux *bccd* entre des feuillets  $\beta$ , ainsi que deux séries *opacd* localisées dans les boucles. Le codage global de la protéine est assez correct, seules 4 positions ont un *RMSd* supérieur à 1,0 Å. Les structures répétitives sont bien approximées, mais ne sont pas les seules (cf. paragraphe



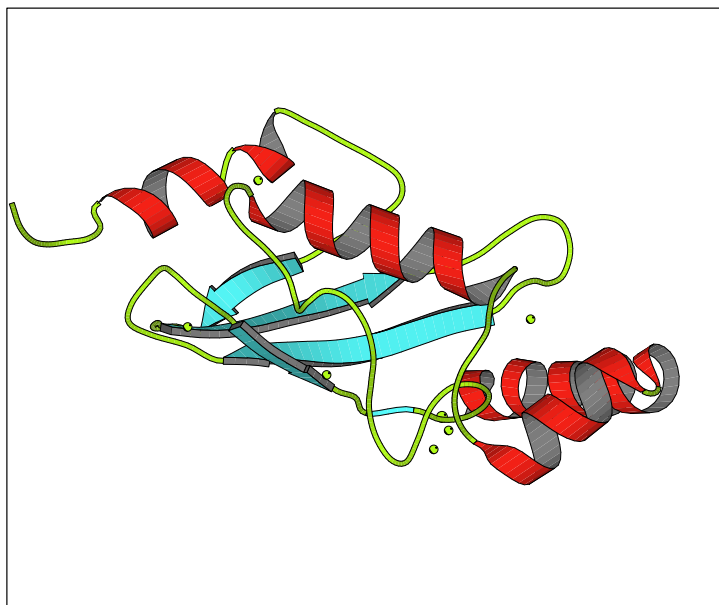


FIG. 3.9 – Exemple de la protéine 1aak représentée avec le logiciel MOLSCRIPT [111]

| motif               | nombre d'observations | <i>RMSd</i> moyen (Å) |
|---------------------|-----------------------|-----------------------|
| <i>mm(cc)dd</i>     | 30                    | 0,70                  |
| <i>dd(fkl)mm</i>    | 414                   | 1,26                  |
| <i>dd(fbc)dd</i>    | 121                   | 1,43                  |
| <i>mm(nopac)dd</i>  | 215                   | 0,76                  |
| <i>dd(fkopac)dd</i> | 64                    | 1,05                  |

TAB. 3.2 – Motifs représentatifs de longueur 2 à 6 entre deux structures secondaires de type BPm et/ou BPd avec le nombre d'observations pour chaque série et le *RMSd* moyen associé.

3.3.2.1).

### 3.3.2.5 Série de blocs : sensibilité structurale

Les autres alphabets structuraux décrits ont souvent des tailles supérieures [162, 53, 19]. Pour analyser la qualité du recodage, des séries de blocs de différentes tailles ont été extraites de la base de données structurales. J'ai regardé spécifiquement des séries de deux structures secondaires répétitives définies par les blocs *m* et/ou *d*. Des motifs de tailles 1 à 6 ont ainsi été examinés. Les exemples les plus représentatifs de chaque longueur sont réunis dans le tableau 3.2 avec le *RMSd* moyen associé. Pour expliciter les exemples, *mm(xyz)dd* est un motif *xyz* qui relie deux BPm et deux PBd.

Pour des motif courts d'un ou deux blocs, les occurrences sont faibles (moins de 40 observations). Pour des motifs plus longs, le nombre global de combinaisons de blocs protéiques augmente fortement. Par exemple, pour une longueur de 4 (respectivement 5 et 6), un nombre moyen de 20 motifs différents est trouvé (respectivement 22 et 30 motifs). Le nombre et le

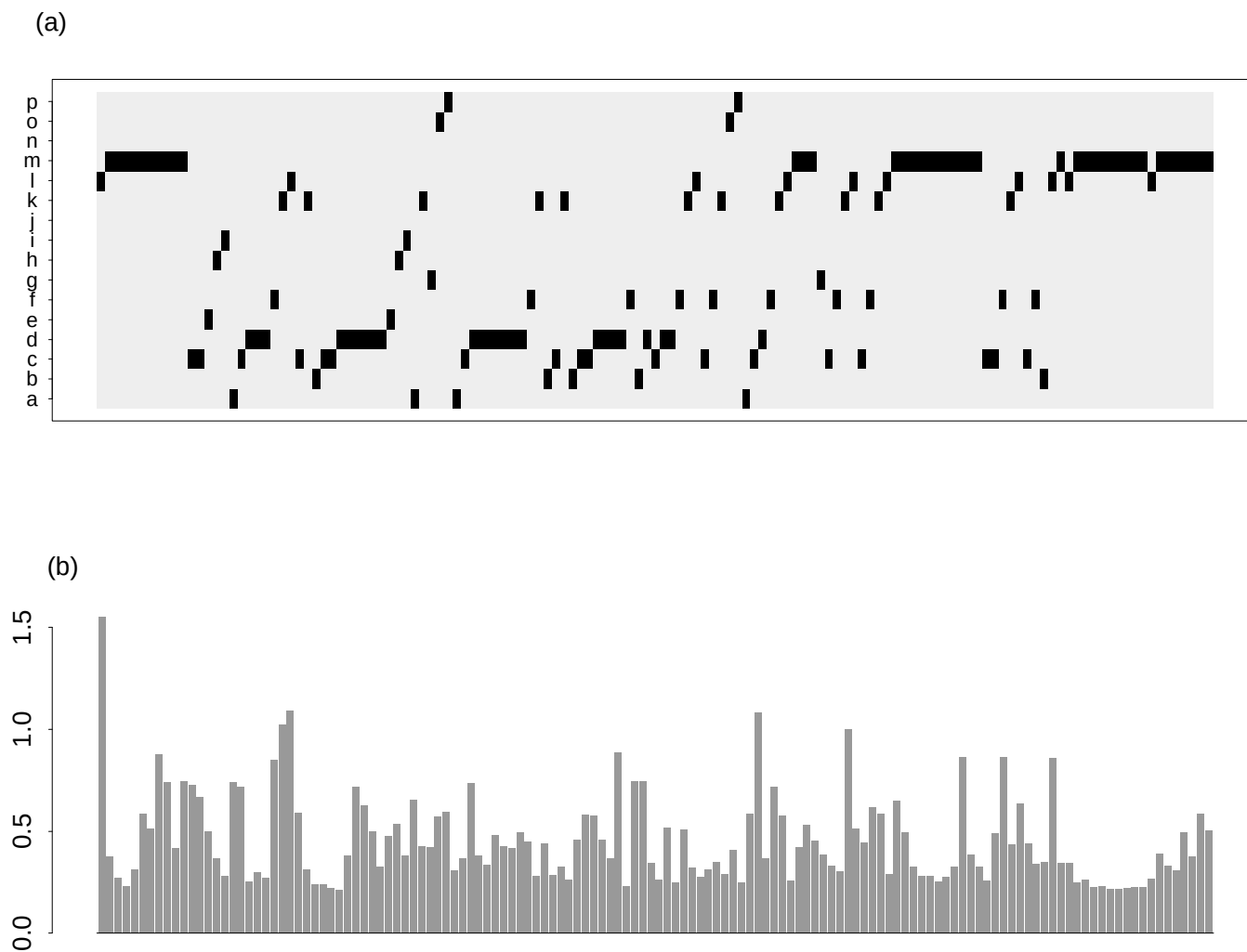


FIG. 3.10 – (a) Exemple de la protéine de liaison à l'ubiquitine (code PDB: 1aak) représentée en termes de blocs protéiques et (b) les RMSd entre chaque fragment de 5 C $\alpha$  de cette protéine et le bloc protéique auquel il est assigné.

type de motifs sont importants et dépendent principalement du type de structures secondaires localisées aux extrémités du motif. Pour un motif de longueur 3 ayant en extrémité N-terminale *dd* et finissant par les blocs *mm*, le motif *fkl* représente 98 % des occurrences trouvées; et en reliant *mm* à *mm*, le motif *nop* représente 82 % des occurrences trouvées avec 24 occurrences présentes. Dans les autres cas, et pour toutes les longueurs observées, il n'existe pas de motif représentant plus de 75 % des occurrences.

Les *RMSd* moyens calculés pour ces exemples montrent une forte stabilité structurale avec une moyenne proche de 1,0 Å. L'approximation structurale induite par l'utilisation des blocs protéiques reste correcte pour des motifs de taille importante. Cette notion couplée avec les propriétés des transitions existants entre les blocs, sera reprise dans le paragraphe 5.2 pour des fragments de tailles supérieures à 5 C<sub>α</sub>.

### 3.3.2.6 Reproduction de la structure : amélioration par l'utilisation d'une librairie

Ayant recodé les protéines à l'aide des blocs protéiques, il faut pouvoir ensuite approximer la structure réelle. Pour cela, une première approche simple a été mise au point. Les fragments sont chevauchants donc dans une succession de blocs *tuvw*, tous les PBs *t*, *u,v*, et *w* approximent les angles  $\phi_u$  et  $\psi_t$ . Ainsi, chaque site, à l'exception des résidus aux extrémités C- et N-terminales, est recodé par 4 blocs protéiques. Une moyenne entre leurs angles, pour le même site, permet un recodage simple. Il faut cependant faire attention aux effets de rotation des angles.

$$\phi_p = \frac{\sum_{i=-1}^{i=2} \phi_{p+i} \text{ modulo } (360^\circ)}{4}$$

$$\psi_p = \frac{\sum_{i=-2}^{i=1} \psi_{p+i} \text{ modulo } (360^\circ)}{4}$$

Ceci mène à une bonne approximation avec plus de 50% des angles approximés à moins de 21° et seulement 3 % de mauvaises approximations avec plus de 90° d'écart avec la réalité.

Pour améliorer ce recalcul des angles à partir du codage en terme de blocs protéiques, j'ai pris en compte le fait que les successions de blocs ne sont pas aléatoires (cf paragraphes 3.3.3 et 5.2). Seul un certain nombre de transitions est vu. Avec une nouvelle base de données de 553 protéines non redondantes [84, 83] et ayant moins de 25% d'identité de séquences (fournie par

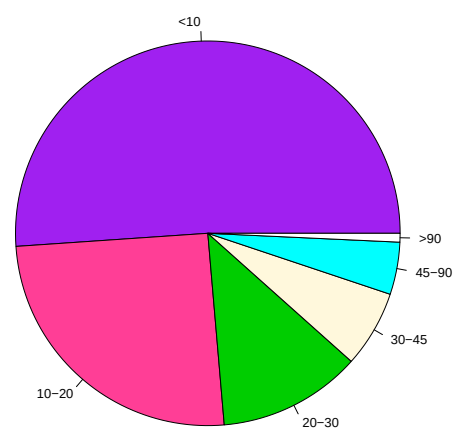
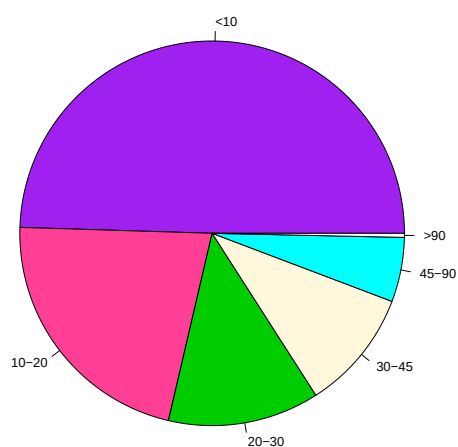
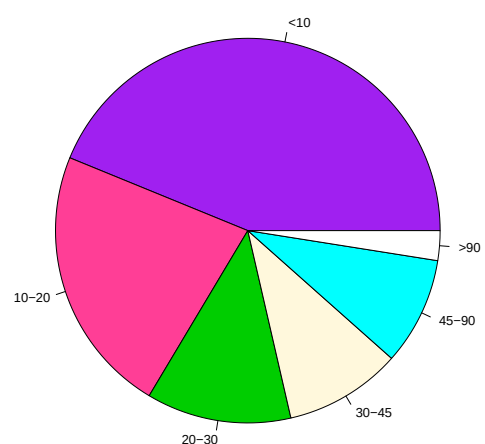
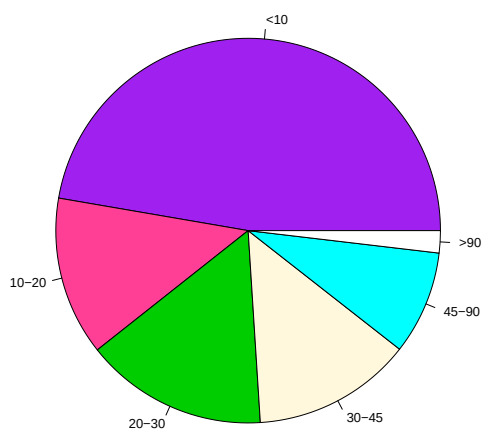


FIG. 3.11 – Calcul des angles  $\phi$  (respectivement (a) et (c)) et  $\psi$  (respectivement (b) et (d)) à partir des blocs protéiques, avec l'usage d'une moyenne des 4 PBs présents (respectivement (a) et (b)), et, avec l'utilisation d'une librairie (respectivement (c) et (d)).

R. Gautier), j'ai constitué pour chaque succession de 4 blocs indicés par  $t, u, v, w$  une librairie avec les angles  $\phi_u$  et  $\psi_t$  moyens qui leurs sont associés. Chaque fois que la succession est vue, au lieu de réutiliser la métrique moyenne, les angles moyens  $\phi_u$  et  $\psi_t$  associés à cette succession sont pris. Si la succession n'a jamais été vue, l'approximation se fait en réutilisant la métrique précédente. La figure 3.11 récapitule les deux approches. On peut ainsi observer que: (1) d'une manière générale les angles  $\psi$  sont les moins bien approximés et ceci avec les deux méthodes et (2) le nombre d'angles mal approximés chute fortement avec l'utilisation d'une librairie. Ainsi, il y a 4 fois moins d'angles à plus de  $90^\circ$  avec cette seconde méthode et 50% des angles sont approximés à moins de  $10^\circ$ .

Après avoir observé dans les derniers paragraphes la stabilité structurale de la série de 16 blocs protéiques, nous allons analyser dans les paragraphes suivants les blocs sur un plan structural en observant tout d'abord les transitions qui existent entre les blocs protéiques, puis ensuite la répartition des structures secondaires associées aux blocs et enfin pour mieux comprendre par la suite la prédiction locale, les différentes répartitions des acides aminés dans les blocs protéiques.

### 3.3.3 Transitions entre les blocs protéiques

Les transitions majoritaires d'un bloc vers d'autres blocs sont en nombre restreint. Elles sont récapitulées dans les cinquième à septième colonnes du tableau 3.3. Les fréquences de transition  $f_{xy}$  d'un bloc  $x$  vers un bloc  $y$  (différent de  $x$ ) à la position suivante sont calculées par la formule suivante :

$$f_{xy} = \frac{N_{xy}}{(\sum_{z=1}^{16} N_{xz}) - N_{xx}}$$

avec  $N_{xy}$ , le nombre de blocs  $x$  suivi par un bloc  $y$ , et  $N_{xx}$  suivi par lui-même.

Les trois transitions majoritaires des structures répétitives, en excluant les blocs correspondant aux parties centrales, représentent au minimum 76% de leurs transitions, sauf pour le bloc protéique  $j$ . Par exemple, pour le bloc protéique  $c$ , 92,2 % des transitions vont vers trois blocs: BP  $d$  (62.2 %), BP  $f$  (24.4 %) et BP  $e$  (5.6 %). De la même manière, moins de la moitié des transitions possibles apparaissent avec une fréquence de plus de 1,0 %. En règle générale, le nombre de transition par blocs dépassant 5 % est de 3, avec un maximum allant jusqu'à 5 pour

le bloc protéique  $j$ .

### 3.3.3.1 Relations avec les structures secondaires

Les blocs protéiques peuvent être caractérisés par leur composition en structures secondaires. Comme pour les autres alphabets structuraux [23], les blocs ne correspondent pas aux catégories classiques des structures secondaires (cf. tableau 3.3). Le bloc protéique  $m$  est une partie centrale d'hélice  $\alpha$  ( $\psi = -47^\circ$  et  $\phi = -57^\circ$ ). Le bloc  $d$  est un bloc "idéal" de feuillet  $\beta$  ( $\psi = 135^\circ$  et  $\phi = -139^\circ$ ).

En tenant compte de leurs compositions en structures secondaires et de leurs transitions, les blocs notés de  $a$  à  $c$  et  $d$  à  $f$  sont groupés autour du bloc  $d$  comme appartenant à des structures  $\beta$  (cf. tableau 3.3). Les blocs  $k$ ,  $l$  et  $n$ ,  $o$  et  $p$  sont des structures locales associées à des extrémités N- et C-terminale d'hélices  $\alpha$  et avec des taux élevés d'assignation dans des structures  $\alpha$ . Le dernier groupe est composé des blocs allant de  $g$  à  $j$ , et représente le groupe "boucles" (fréquence des boucles supérieur à 80 % pour le  $C_\alpha$  central).

La figure 3.12 permet de comparer visuellement l'assignation des blocs par rapport à une assignation par les structures secondaires classiques. Le patchwork est parfois difficilement lisible dans les parties boucles, mais montre bien la finesse de leur attribution. Dans l'hélice  $\alpha$  située en bas à gauche, l'utilisation des blocs protéiques permet de bien marquer deux résidus (en bleu) qui sont distincts d'une structure régulière d'hélice. Pour les feuillets, l'intérêt est encore plus grand au vue des différence visible entre le long feuillet  $\beta$  central et le plus court situé à sa gauche qui est plus tordu.

Un autre paramètre a été analysé, le nombre moyen de répétition sur soi-même noté  $nmr$  (cf. la quatrième colonne du tableau 3.3). Cette valeur est quantifiée par :

$$nmr = \frac{1}{1 - \frac{N_{xx}}{\sum_{z=1}^{z=16} N_{xz}}}$$

avec  $N_{xx} / \sum_{z=1}^{z=16} N_{xz}$  la fréquence de transition du bloc sur lui même,  $N_{xx}$  étant le nombre de fois où le BP $x$  était suivi par lui-même et  $N_{xz}$  où il est suivi par BP $z$ . Cette notion permet de bien caractériser les structures répétitives classiques :

- (i) BP $m$  possède un taux de transition sur lui même de  $p_{mm}=85,2\%$ , soit un  $nmr$  de 6,74

| BP       | f. occ. | RMSdm | nmr  | Transitions      |                  |                  | $\alpha$ | S2      |         | description                       |
|----------|---------|-------|------|------------------|------------------|------------------|----------|---------|---------|-----------------------------------|
|          |         |       |      | 1 <sup>er</sup>  | 2 <sup>eme</sup> | 3 <sup>eme</sup> |          | boucles | $\beta$ |                                   |
| <i>a</i> | 3,93    | 0,52  | 1,01 | 54,8( <i>c</i> ) | 16,5( <i>f</i> ) | 8,0( <i>b</i> )  | 0,1      | 76,7    | 23,3    | N-ter $\beta$                     |
| <i>b</i> | 4,58    | 0,51  | 1,00 | 44,4( <i>d</i> ) | 17,9( <i>c</i> ) | 13,7( <i>f</i> ) | 0,2      | 86,7    | 13,1    | N-ter $\beta$                     |
| <i>c</i> | 8,63    | 0,51  | 1,28 | 62,2( <i>d</i> ) | 24,4( <i>f</i> ) | 5,6( <i>e</i> )  | 0,1      | 58,2    | 41,7    | N-ter $\beta$                     |
| <i>d</i> | 18,84   | 0,48  | 2,74 | 51,9( <i>f</i> ) | 25,6( <i>c</i> ) | 19,2( <i>e</i> ) | 0,0      | 28,4    | 71,6    | $\beta$                           |
| <i>e</i> | 2,31    | 0,54  | 1,11 | 80,4( <i>h</i> ) | 9,1( <i>d</i> )  |                  | 0,0      | 49,8    | 50,2    | C-ter $\beta$                     |
| <i>f</i> | 6,72    | 0,50  | 1,00 | 60,7( <i>k</i> ) | 36,3( <i>b</i> ) |                  | 0,0      | 72,5    | 27,5    | C-ter $\beta$                     |
| <i>g</i> | 1,28    | 0,74  | 1,05 | 37,5( <i>h</i> ) | 28,0( <i>c</i> ) | 19,1( <i>o</i> ) | 6,9      | 83,8    | 9,3     | principalement boucles            |
| <i>h</i> | 2,35    | 0,62  | 1,04 | 62,4( <i>i</i> ) | 18,1( <i>j</i> ) | 10,2( <i>k</i> ) | 0,0      | 81,5    | 18,5    | principalement boucles            |
| <i>i</i> | 1,62    | 0,56  | 1,01 | 87,7( <i>a</i> ) |                  |                  | 0,0      | 94,5    | 5,5     | principalement boucles            |
| <i>j</i> | 0,96    | 1,03  | 1,01 | 17,0( <i>a</i> ) | 16,6( <i>b</i> ) | 16,1( <i>l</i> ) | 3,7      | 87,9    | 8,4     | principalement boucles            |
| <i>k</i> | 5,46    | 0,59  | 1,00 | 76,2( <i>l</i> ) | 13,6( <i>b</i> ) |                  | 35,1     | 64,2    | 0,7     | N-ter $\alpha$                    |
| <i>l</i> | 5,35    | 0,63  | 1,01 | 68,5( <i>m</i> ) | 9,2( <i>p</i> )  | 7,0( <i>c</i> )  | 44,4     | 54,9    | 0,7     | N-ter $\alpha$                    |
| <i>m</i> | 30,04   | 0,43  | 6,74 | 33,8( <i>n</i> ) | 18,5( <i>p</i> ) | 9,7( <i>b</i> )  | 86,7     | 13,2    | 0,1     | $\alpha$                          |
| <i>n</i> | 1,93    | 0,61  | 1,03 | 90,9( <i>o</i> ) |                  |                  | 68,4     | 31,3    | 0,3     | C-ter $\alpha$                    |
| <i>o</i> | 2,60    | 0,60  | 1,02 | 74,7( <i>p</i> ) | 8,3( <i>m</i> )  |                  | 43,1     | 56,8    | 0,1     | C-ter $\alpha$                    |
| <i>p</i> | 3,41    | 0,46  | 1,00 | 58,1( <i>a</i> ) | 22,7( <i>c</i> ) | 11,1( <i>m</i> ) | 11,2     | 87,5    | 1,3     | C-ter $\alpha$ vers N-ter $\beta$ |

TAB. 3.3 – Pour chaque bloc protéique (noté de BPa à BPp), sont donnés sa fréquence dans la base de donnée (f. occ.), son RMSd moyen (RMSdm), le nombre moyen de répétition sur lui même (nmr), ses 3 transitions principales, sa répartition en structure secondaire (hélice- $\alpha$ , feuillet  $\beta$  et boucles) pour le résidu central (S2) et une caractérisation grossière du bloc.

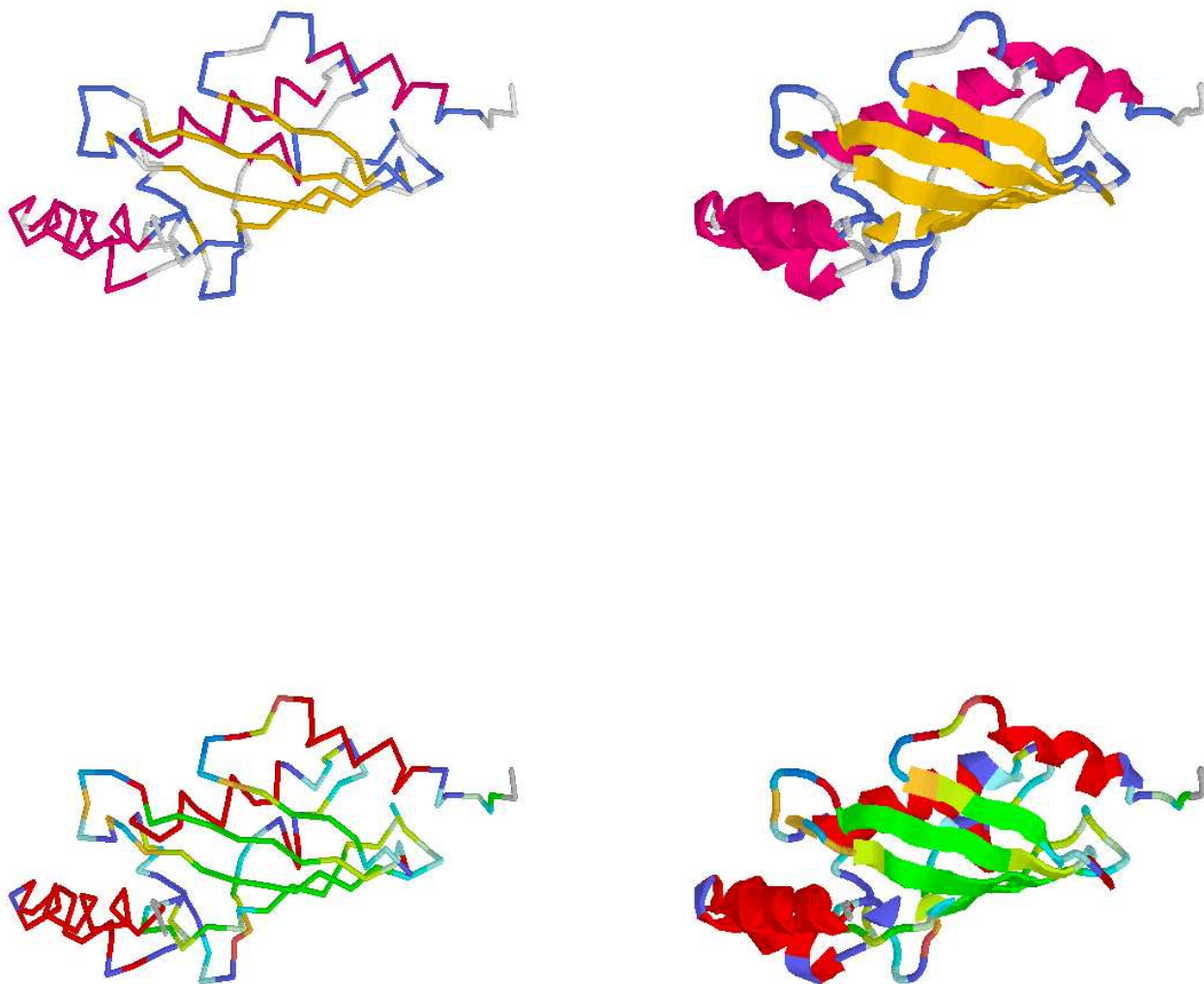


FIG. 3.12 – Colorisation de la protéine 1aak avec le logiciel rasmol. En haut est représentée la protéine 1aak coloriée avec les structures secondaires, en bas, les même représentations, avec les blocs protéiques.



blocs ce qui correspond bien à une hélice régulière. De la même manière, 78,1 % des fragments de 5  $C_\alpha$  de la base de données dont le troisième  $C_\alpha$  est considéré en hélice  $\alpha$  appartiennent à ce bloc et 86,7 % des fragments de ce bloc ont un troisième  $C_\alpha$  en hélice  $\alpha$ .

- (ii)  $BPd$  (avec  $p_{dd} = 63,5$  %) caractérise un feuillet  $\beta$  avec une longueur moyenne de 2,74 blocs.
- (iii)  $BPe$  et  $BPc$  ont un  $nmr$  supérieur à 1,1. Ils correspondent à des entrées ou des sorties de feuillets, structures de feuillets  $\beta$  déformées ou à des extrémités C- ou N-terminales de régions étendues du squelette protéique.

La dernière colonne du tableau 3.3 fait une analogie entre les blocs protéiques et les structures secondaires, toutefois cette nomenclature est rapide. Les structures secondaires n'ont que 3 états distincts: hélices  $\alpha$ , feuillets  $\beta$  et boucles (ou non- $\alpha$  / non- $\beta$ ), l'équivalence n'est donc que partielle avec un alphabet structural à 16 états. Par exemple, le bloc protéique  $b$  mis dans une catégorie "extrémité N-terminale de feuillet  $\beta$ ", va vers le bloc  $d$  principalement, bloc  $\beta$  par excellence, mais aussi directement vers une "extrémité C-terminale de feuillet  $\beta$ ", le bloc protéique  $f$  avec un taux de 13,7 %. De la même manière, le bloc  $m$  prototype de l'hélice  $\alpha$  va vers le bloc protéique  $b$ , une "extrémité N-terminale de feuillet  $\beta$ " avec une fréquence de 9,2 %. La flexibilité de l'alphabet est supérieur au simple label donné dans le tableau.

### 3.3.4 Les acides aminés présents dans blocs protéiques

Après la phase d'apprentissage, l'ensemble des protéines est codé à l'aide des blocs protéiques (pour chacune des séries de  $B$  blocs obtenus), en utilisant toujours comme critère d'attribution le  $RMSda$  minimal. Chaque bloc est donc associé avec un ensemble de séquences protéiques. Les matrices d'occurrence des différents acides aminés pour chaque position ont donc été calculées. Ainsi, chaque bloc protéique d'une longueur de  $M=5$  résidus est représenté par une matrice de taille  $M \times 20$ , pour chaque type d'acides aminés. Pour améliorer la spécificité des matrices, le calcul n'a pas été fait sur  $M$  résidus, mais pour une fenêtre plus étendue entre  $-w$  et  $+w$  autour du résidu central. Nous avons testé différentes longueurs  $w$  (avec  $w > 2$ ). De manière plus formelle, le nombre d'occurrences  $n_{ij}^k$  pour un type d'acide aminé donné (indexé par  $i =$

1,2,...,20) localisé à une position  $j$  ( $j$  variant dans l'intervalle  $[-w, +w]$ ) de la fenêtre est calculé. Ensuite, nous en déduisons la probabilité d'avoir ce type d'acide aminé en cette position pour le bloc,  $P(a_i \text{ en } j / PB_k)$  à l'aide du rapport  $n_{ij}^k / N_k$  où  $N_k$  est le nombre de  $PB_k$  de la base de donnée d'apprentissage.  $P(a_i \text{ en } j / PB_k)$  est la probabilité conditionnelle pour l'acide aminé  $a_i$  en position  $j$  pour le bloc  $PB_k$ . Nous avons défini une fenêtre de longueur 15 ( $w = 7$ ). Ce choix est discuté ultérieurement car il dépend des résultats de la méthode de prédiction.

### 3.3.5 Analyse des matrices d'occurrences des blocs

Pour analyser les spécificités des répartitions des acides aminés suivant le type de bloc impliqué, il est possible :

- de quantifier la spécificité de chaque position dans la fenêtre.
- de déterminer quels acides aminés en quelles positions ont une distribution spécifique dans le bloc considéré.

#### 3.3.5.1 La mesure de divergence asymétrique de Kullback-Leibler (KLd)

Pour aborder le premier point, l'entropie relative ou mesure de divergence asymétrique de Kullback-Leibler (KLd, [112]) est particulièrement utile :

$$K(\mathbf{p}, \mathbf{q}) = \sum_i p_i \ln \left( \frac{p_i}{q_i} \right)$$

Ce terme  $K$  quantifie la différence existante entre la distribution des acides aminés dans le bloc  $\mathbf{p}$ :  $\{p_i\}_{i=1,...,20}$  et celle attendue au vue de l'occurrence du bloc dans la base de données si tout était aléatoire. Le terme  $K_k(\mathbf{p}_j, \mathbf{q})$  a été calculé en chaque position  $j$  pour observer la différence existant entre la distribution observée en acides aminés  $\mathbf{p}_j$  et la distribution de référence de la base de donnée  $\mathbf{q}$  ajustée pour  $PB_k$ .

Cette mesure de divergence, notée KLd, permet de détecter des positions "informatives" en observant les positions  $j$  dans l'intervalle  $[-w; +w]$ .

L'entropie relative  $K(\mathbf{p}, \mathbf{q})$  est une valeur toujours supérieure ou égale à zéro. Elle suit, multipliée par  $2N$  ( $N$  étant le nombre d'observations dans la base de données), une loi classique du  $\chi^2$  à 19 degrés de liberté (ou ddl, car défini sur les 20 types d'acides aminés). La valeur seuil

a été choisie pour un risque  $\alpha$  de premier ordre de  $10^{-5}$ . Toute valeur supérieure à cette valeur seuil est donc fortement significative.

### 3.3.5.2 Z-scores

Pour évaluer localement l'influence de chaque type d'acide aminé, l'utilisation de la méthode des Z-scores est fort pertinente. Elle permet de mettre en évidence les sous- et sur-représentations. Ainsi, chaque matrice d'occurrence associée à chaque bloc a été normalisée comme suit :

$$Z_{ij}^k = \frac{(n_{ij}^k - n_{ib})}{\sqrt{n_{ib}}}$$

avec  $n_{ib}$  le nombre attendu du  $i$ ème acide aminé ( $n_{ib} = N_k \cdot f_i$  où  $N_k$  et  $f_i$  sont respectivement le nombre de  $PB_k$  et la fréquence observée de l'acide aminé  $i$  dans la base de données). Les Z-scores positifs (ou négatifs) correspondent, pour le bloc protéique  $k$ , à des sur-représentations d'acides aminés (ou sous-représentations). La valeur seuil a été prise égale à 4,4 ce qui représente une probabilité  $p$  inférieure à  $10^{-5}$ .

## 3.3.6 Relation entre blocs protéiques et séquences

Les relations entre blocs protéiques et distributions en acides aminés peuvent être analysées en observant les matrices d'occurrences associées à chaque BP.

### 3.3.6.1 Quatre Exemples

La figure 3.13 montre la structure tridimensionnelle de fragments de 5  $C_\alpha$  associés à 4 blocs protéiques caractéristiques. Ils sont visualisés à l'aide du logiciel Xmol développé par P. Tufféry [194]. Sont donc représentés les blocs protéiques  $p$ ,  $b$ ,  $d$  et  $m$ , ainsi que leurs matrices d'occurrences normalisées en Z-scores (cf. paragraphe 3.3.5.2) et le KLd associé (cf. paragraphe 3.3.5.1).

Les blocs protéiques  $m$ ,  $p$  et  $d$  ont une valeur de *RMSd* moyen assez faible (0,43 Å, 0,46 Å et 0,48 Å respectivement). Le PB  $b$  est légèrement plus variable avec un *RMSd* moyen de 0,51 Å. Cette variabilité est due principalement à une plus grande variabilité à ses extrémités.

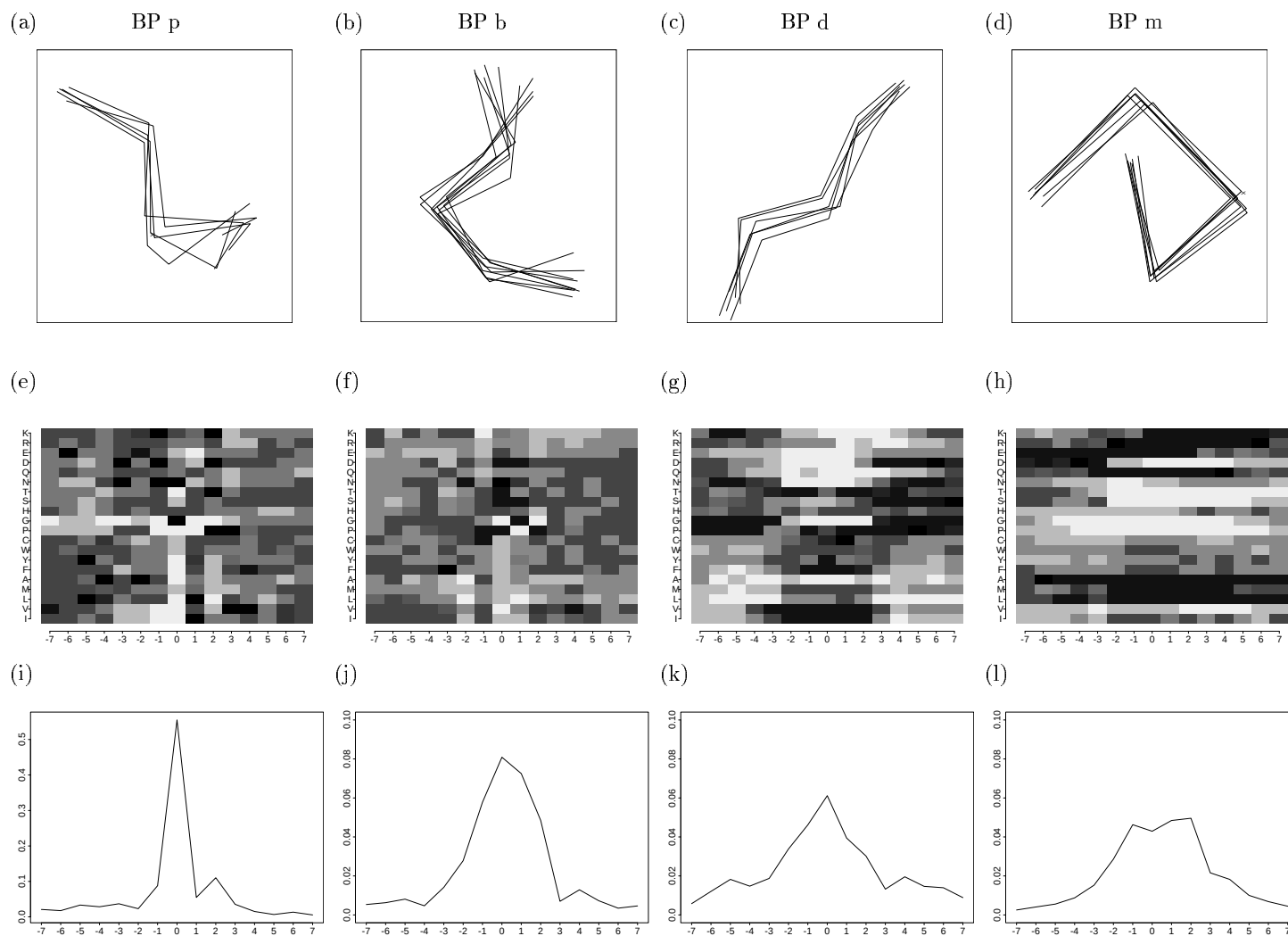


FIG. 3.13 – Exemple de 4 blocs protéiques. (a,e,i) BPp, (b,f,j) BPb, (c,g,k) BPd, (d,h,l), BPm. Figures (a-d) superposition de fragments, (e-h) matrices d'occurrences normalisées en Z-scores, en noir, Z-scores > 4,4, en blanc Z-scores < -4,4, en gris, Z-scores intermédiaires et (i-l) profils KLD associés à chaque type de blocs.

L'analyse de la répartition des acides aminés est simplifiée avec l'utilisation des Z-scores. Les matrices ont été seuillées. Ainsi les rectangles noirs indiquent un Z-score  $> 4,4$ ; inversement pour un rectangle blanc, un Z-score  $< -4,4$ . Ce seuil correspond à une erreur de type I avec une probabilité  $p$  inférieure à  $10^{-5}$ . Les zones en gris correspondent à un Z-score intermédiaire.

De même, l'analyse du KLd qui exprime la dissimilarité entre la distribution présente dans chaque matrice d'occurrence et la distribution dans la base de données, permet de définir les positions les plus informatives dans un bloc donné. Les 4 blocs protéiques de la figure 3.13 sont représentatifs de l'ensemble des 16 blocs protéiques. Ils ont été ordonnés en fonction de leur valeur de KLd maximale.

Le bloc protéique  $p$  est caractérisé par sa position centrale avec une sur-représentation de Glycine et d'Asparagine. Les sous-représentations sont aussi importantes que les sur-représentations à cette position. Le KLd montre un profil très pointu (KLd = 0,55). Le bloc  $b$  possède lui un profil en cloche à fromage cinq fois moins accentué que le précédent avec un KLd maximum de 0,08. En contre-partie de cette diminution de l'importance d'un site, d'autres sites deviennent informatifs dans l'ensemble de l'intervalle  $[-2; +2]$  et plus seulement sur la seule position centrale. En parallèle de la diminution des valeurs de KLd, la spécificité en acides aminés devient plus complexe à voir, on peut observer des sur- et sous-représentations alternées de Proline et de Glycine.

Le bloc protéique  $d$  correspond à un feuillet  $\beta$  régulier. Le profil du KLd est différent des précédents avec un KLd maximum de 0,06 sur le résidu central et une décroissance symétrique des deux côtés. Une forte différence dans les valeurs de Z-scores entre les Z-scores à l'intérieur et à l'extérieur du bloc structural  $[-2; +2]$  est observée. Les sur-représentations concernent principalement certains résidus hydrophobes Isoleucine et Valine et dans une moindre mesure la Phénylalanine, la Tyrosine, le Tryptophane et la Thréonine dans le bloc structural. Les sous-représentations concernent principalement des résidus polaires comme la Lysine, l'Arginine, l'Aspartate, la Glutamine et l'Asparagine; et à l'extérieur du bloc structural la Glycine et la Proline, ce qui est classique [90].

Le bloc protéique  $m$ , comme le bloc  $d$ , a un profil de KLd particulier. La majorité de la spécificité de la séquence se trouve dans la partie centrale et correspond aux 5 C $\alpha$  des positions -2 à +2 avec un KLd maximum de 0,05. Nous observons une sous-représentation des résidus aliphatiques, tels que la Leucine, la Méthionine, l'Alanine, et des résidus polaires, tels que

la Glutamine, le Glutamate, l'Arginine, l'Aspartate et la Lysine. La sous-représentation des casseurs d'hélices  $\alpha$  que sont la Proline et la Glycine est importante à l'intérieur de la fenêtre structurale. Il en est de même pour l'Histidine et l'Asparagine sur l'ensemble de la fenêtre de 15 résidus [154, 145, 113].

Un autre type de profil existant est un profil bimodal caractéristique de certains blocs, le  $BP_c$  (maximum en positions -2 et +2),  $BP_e$  (positions -1 et +1),  $BP_l$  (positions -2 et 0),  $BP_k$  (positions -1 et +1) et  $BP_p$  (positions 0 et +2). Ces positions sont principalement des casseurs de structures répétitives.

### 3.3.6.2 Utilisation des Z-scores pour déterminer les acides aminés sur- et sous-représentés dans chaque bloc

Le tableau 3.4 mentionne pour chaque bloc protéique, en chaque position, de la fenêtre allant de -4 à +4 autour de la position centrale, les acides aminés sur- et sous- représentés (Z-score  $>4.4$  noté +, ou,  $<-4.4$  noté -). Un grand nombre d'acides aminés par position ont un Z-score important.

Pour se focaliser sur les plus spécifiques des résidus, la valeur de 4,4 comme valeur seuil a été choisie comme pour la figure 3.13. Seule la zone allant de -4 à +4 est représentée. En dehors de cette zone, une seule position présentait encore un intérêt. Les positions informatives ont été retenues sur la base d'une valeur de Kld supérieure à  $300 / 2N_k$ , avec  $N_k$  le nombre de  $PB_k$  observé.

Les transitions principales entre les blocs protéiques (cf. tableau 3.3) sont retrouvées dans les compositions en acides aminés. Par exemple, les sur-représentations des Glycines et des Asparagines sont observées pour les blocs protéiques  $n$ ,  $o$ ,  $p$  et  $a$  en positions (+2), (+1), (0) et (-1). De la même manière, les blocs protéiques  $e$  et  $g$  en position +2, vont vers le  $BPh$  (en position +1) and  $BP_i$  (en position 0). Cependant des différences existent, comme pour le  $BP_d$  qui va vers le  $BP_f$  avec une fréquence de 51,9%, sa position (+1) a une sous-représentation différente par rapport au bloc  $f$  en position (0) où la Valine et l'Isoleucine sont sous-représentées pour le premier et sur-représentées pour le second. En fait, les proportions des acides aminés ne sont pas toujours conditionnées par les transitions préférentielles entre blocs. Ainsi, la sur-représentation de la Proline en position (+1) du  $BP_b$  est remplacée par une sous-représentation dans le bloc  $c$  alors que la fréquence de transition du  $BP_b$  au  $c$  est de 17,9 %.

| BP       | Zone            | -4 | -3  | -2      | -1           | 0          | +1          | +2          | +3         | +4     |
|----------|-----------------|----|-----|---------|--------------|------------|-------------|-------------|------------|--------|
| <i>a</i> | [-2;-1],[+1,+2] | +  | K   | DNP     | GN           | K          | PT          | PV          | V          |        |
|          |                 | -  |     | GV      | AEFILMPSTVWY | GP         | AL          |             |            |        |
| <i>b</i> | [-1;+2]         | +  |     | N       | CP           | DST        | DGS         | P           |            |        |
|          |                 | -  |     | AL      | K            | GV         | LP          | G           |            |        |
| <i>c</i> | (-3),           | +  | NPS | DGN     | HN           | IPV        | IPV         | IPV         | P          |        |
|          | [-1;+2]         | -  |     | AFILMVY | LP           | DG         | AEG         | AG          | K          | K      |
| <i>d</i> | [-6;+6]         | +  | GP  | GV      | ITV          | FIPTV      | FIVY        | FIPTV       | IPTV       | DPS    |
|          |                 | -  | AL  | AL      | DENQ         | ADEGN      | ADEGKNQ     | ADEGKNQR    | GKQ        | AKLQR  |
| <i>e</i> | (+2)            | +  |     | V       | V            |            | P           | GN          | G          |        |
|          |                 | -  |     |         |              |            | L           | AEFILPQRTVY |            |        |
| <i>f</i> | [-1;+1]         | +  |     | G       | FIPVY        | DNPST      | P           | DENS        | DE         |        |
|          | ,(+3)           | -  |     |         | ADG          | AEFGIKLQRV | G           | ILMPV       | L          |        |
| <i>g</i> | (+2)            | +  |     |         |              |            | P           | G           |            |        |
|          |                 | -  |     |         |              |            |             | AEILV       |            |        |
| <i>h</i> | [0;+1]          | +  |     |         | P            | NP         | GN          |             | T          |        |
|          |                 | -  |     |         |              | L          | AEFIKLMPTVY | P           |            |        |
| <i>i</i> | [-1;0]          | +  |     |         | NP           | GN         |             | PT          |            |        |
|          |                 | -  |     |         | L            | AEILPSTV   | P           |             |            |        |
| <i>j</i> | (0)             | +  |     | G       |              | G          |             |             | G          |        |
|          |                 | -  |     |         |              | AILPTV     |             |             |            |        |
| <i>k</i> | [-2;+3]         | +  | G   | FILVY   | DNPST        | EP         | DENS        | DEG         | L          |        |
|          |                 | -  |     | DES     | AEFIKLMQRVY  | GIV        | GILPV       | ILP         | E          |        |
| <i>l</i> | [-1;+2]         | +  | G   | V       | DNPST        | P          | DES         | DE          | L          | K      |
|          |                 | -  |     |         | AEFIKLRV     | I          | ILPV        | IP          | GP         | G      |
| <i>m</i> | [-4;+6]         | +  | AEL | ADEQ    | AELMQ        | AELMQR     | AEKLMQR     | AEKLMQR     | AEKLMQR    | AKLMQR |
|          |                 | -  | GP  | GPV     | GNPST        | GNPST      | DGPST       | DGPST       | DGPSTV     | DGPSTV |
| <i>n</i> | [0;+2]          | +  | E   | L       | AL           | KR         | AEK         | L           | GN         | P      |
|          |                 | -  |     | GP      |              |            |             | GPV         | AEFILPSTVY |        |
| <i>o</i> | [-1;+1]         | +  | Y   | AL      | K            | AEK        |             | GN          | K          | P      |
|          |                 | -  | P   |         |              |            | GPV         | AEFILPSTVY  |            | V      |
| <i>p</i> | [-1;+2]         | +  | A   | D       |              | DKN        | GN          | I           | DP         | PV     |
|          |                 | -  | G   |         | GIPV         | AFILPTVY   | EGP         | GL          |            |        |

TAB. 3.4 – Les acides aminés les plus importants dans chaque bloc protéique. Pour chaque position (numérotée de -4 à +4), avec les sur-représentations (Z-score > +4.4) et les sous-représentations (Z-score < -4.4) notés (+) et (-).

Les sous- et sur-représentations sont concentrées principalement autour de la position centrale, dans une zone recouvrant le bloc structural  $[-2;+2]$ . Le tableau 3.4 montre l'importance de ces positions.

On peut noter que les structures répétitives montrent bien des représentations classiques avec des sur-expressions de [AEL] et des sous-représentations de [GPST] pour le bloc protéique  $BP_m$ , prototype de la partie centrale de l'hélice  $\alpha$ , et des sur-expressions de [IV] et des sous-représentations de [ADEGN] pour  $BP_d$  [90, 145, 113]. De même, la sur-représentation de Glycine est le plus souvent associée à celle d'Asparagine dans les boucles. La flexibilité due à la Glycine permet une torsion importante, et la fonction amine de l'Asparagine permet une liaison directe avec le squelette peptidique assurant ainsi un brusque changement de conformation [47].

## 3.4 Comparaison avec les autres alphabets structuraux

Une des difficultés majeures de l'évaluation de cet alphabet structural à 16 états est la comparaison avec les autres alphabets existants. En effet, il existe fort peu d'alphabets structuraux accessibles.

Aussi, ce paragraphe présente les différentes études que j'ai pu réaliser. Après une rapide étude avec les structures définies par DSSP, une première comparaison a été effectuée avec les blocs définis par Rooman et collaborateurs [162], que j'ai caractérisé en partant des figures de leur article. La deuxième comparaison a porté sur une base de données recodées par Fetrow et collaborateurs [53], avec leur 6 super-structures secondaires accessibles sur leur site. Et enfin, le dernier alphabet comparé est celui développé par Camproux et collaborateurs [23, 24]; Anne-Claude Camproux m'ayant fourni une centaine de protéines recodées avec son alphabet.

### 3.4.1 Comparaison avec les attributions de DSSP

Le tableau 3.5 récapitule les sur- et sous-représentations des différentes classes définies par DSSP en fonction des 16 BPs. Les Z-scores sont explicités au paragraphe 3.3.5.2. Les correspondances vues au paragraphe 3.3.3.1 sont retrouvées avec une concentration des hélices  $\alpha$  dans les BPs  $l$  à  $n$  et une sous représentation dans les BPs  $o$  à  $j$ . Les hélices  $3_{10}$  montrent déjà une plus grande flexibilité d'attribution, elles sont retrouvées avec les BPs  $g$ ,  $k$ ,  $l$ , et  $n$  à  $p$ , et surtout jamais avec le BP "hélice centrale"  $m$ , ce qui montre encore la sensibilité de l'assignation. Pour



|   | hélice<br>$\alpha$ | hélice<br>$3_{10}$ | hélice<br>$\pi$ | pont<br>isolé | coudes | boucles | feuillet<br>$\beta$ | conformation<br>étendue $\beta$ |
|---|--------------------|--------------------|-----------------|---------------|--------|---------|---------------------|---------------------------------|
| a | -                  | -                  |                 | +             | -      | +       | -                   | +                               |
| b | -                  | -                  |                 | -             |        | +       | -                   | +                               |
| c | -                  | -                  |                 | +             | -      | +       | +                   | +                               |
| d | -                  | -                  |                 | +             | -      | +       | +                   | -                               |
| e | -                  | -                  |                 | +             | -      | +       | +                   |                                 |
| f | -                  | -                  |                 | +             | -      | +       | +                   | -                               |
| g | -                  | +                  |                 | +             |        | +       | -                   | +                               |
| h | -                  |                    |                 |               | +      | -       | -                   | +                               |
| i | -                  |                    |                 | -             | +      | -       | -                   | +                               |
| j | -                  |                    |                 |               | +      |         | -                   | +                               |
| k |                    | +                  |                 | -             | +      | -       | -                   |                                 |
| l | +                  | +                  |                 | -             | +      | -       | -                   |                                 |
| m | +                  |                    |                 | -             | -      | -       | -                   | -                               |
| n | +                  | +                  |                 | -             | +      | -       | -                   | -                               |
| o | -                  | +                  |                 | -             | +      | -       | -                   |                                 |
| p | -                  | +                  |                 | -             | +      |         | -                   | +                               |

TAB. 3.5 – Correspondance entre les attributions des structures secondaires par le logiciel DSSP [100] et les 16 PBs, avec (+) Z-scores  $> 4,4$ , (-) Z-scores  $< -4,4$  et () Z-scores intermédiaires. La base de données utilisée contient 906 chaînes protéiques possédant moins de 50 % d'identité de séquences.

les hélices  $\pi$ , aucune n'ayant été observée, aucune corrélation n'est détectable. Ce fait montre aussi l'intérêt d'un alphabet structural qui ne possède pas de structures "rares". Les coudes se trouvent préférentiellement associés aux PBs associés aux hélices  $\alpha$ , sauf le PB  $m$ , sa partie centrale, ce qui semble logique, la plupart des coudes ayant des liaisons hydrogènes proches de ces structures.

Enfin, les boucles et formes  $\beta$  ne se retrouvent pas dans les PBs  $k$  à  $o$ , elles sont associées à des PBs différents. Par exemple, les PBs  $a$  et  $b$  ne se retrouvent pas dans le feuillet, mais bien dans les boucles et en conformation étendue.

Ce tableau montre bien l'intérêt et la richesse d'analyse possible avec un alphabet structural qui distingue bien les différentes formes classiques entre elles.

### 3.4.2 Comparaison avec les blocs structuraux de Rooman et collaborateurs

Rooman et collaborateurs ont décrit 4 types de blocs avec 4 longueurs différentes, ici nous n'analyserons que les résultats obtenus pour les blocs de longueurs 5  $C_\alpha$  et 6  $C_\alpha$ . En quelques

|     | SBB          | Blocs Protéiques |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
|-----|--------------|------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|     |              | <i>a</i>         | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | <i>f</i> | <i>g</i> | <i>h</i> | <i>i</i> | <i>j</i> | <i>k</i> | <i>l</i> | <i>m</i> | <i>n</i> | <i>o</i> | <i>p</i> |
| (a) |              |                  |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
|     | $\eta_5$     | .                | .        | .        |          |          | .        | .        |          |          |          | +        | +        | +        | .        | .        | .        |
|     | $\epsilon_5$ |                  |          | .        |          |          |          |          |          |          |          |          |          | .        |          |          |          |
|     | $\zeta_5$    |                  |          |          |          |          |          |          |          |          |          |          | .        |          |          |          |          |
|     | $\lambda_5$  |                  |          |          |          |          |          |          |          | .        | .        |          | .        | .        |          |          | +        |
| (b) |              |                  |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
|     | $\eta_6$     | .                | .        | .        |          |          | .        | .        | .        |          |          | +        | +        | +        | .        | .        | .        |
|     | $\epsilon_6$ |                  | .        |          |          |          |          |          |          |          |          |          |          | .        |          |          |          |
|     | $\zeta_6$    |                  |          |          |          |          |          |          |          |          |          |          | .        | +        |          |          |          |
|     | $\lambda_6$  |                  |          |          |          |          |          |          |          |          |          |          | .        | .        | .        |          |          |
| (c) |              |                  |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
|     | $\eta_6$     | .                | .        | .        |          |          | .        |          |          |          |          | +        | +        | +        | .        | .        | .        |
|     | $\epsilon_6$ |                  |          |          |          | .        |          |          |          |          |          |          |          |          |          |          |          |
|     | $\zeta_6$    |                  |          |          |          |          |          |          |          |          |          |          | .        | +        | .        | .        |          |
|     | $\lambda_6$  |                  |          |          |          |          |          |          |          |          |          |          |          | .        | +        | .        |          |

TAB. 3.6 – *Adéquation entre les blocs structuraux définis par Rooman et collaborateurs[162] et les 16 PBs. Le critère utilisé est celui du RMSda, avec comme symboles (+) RMSda < 60°, (.) 60° < RMSda < 75°, et ( ) RMSda > 75°. (a) Blocs structuraux de longueur 5  $C_\alpha$ , (b) blocs structuraux de longueur 6  $C_\alpha$ , sur les 5 premiers  $C_\alpha$  des blocs, (c) blocs structuraux de longueur 6  $C_\alpha$ , sur les 5 derniers  $C_\alpha$ .*

mots, la famille  $\eta_5$  ( $\alpha$ ) est proche des blocs  $\alpha$  PBs  $k$  à  $m$ . La plupart des blocs protéiques associés aux boucles et aux feuillets en sont éloignés. Seule la famille  $\lambda_5$  (boucles) a une forte association avec le BP  $p$ . Les famille  $\epsilon_5$  et  $\zeta_5$  ne montrent qu'un fort éloignement par rapport aux BPs.

Le bloc structural de longueur 6  $C_\alpha$  a été analysé en deux parties pour les 5 premiers et les 5 derniers  $C_\alpha$ . Comme précédemment, peu de lien existe entre un bloc ou ensemble de BPs et l'un des 4 blocs structuraux. La famille  $\eta_6$  correspond à une hélice  $\alpha$  possédant une extrémité N-terminale, alors que la famille  $\zeta_6$  est synonyme de la partie centrale de l'hélice.  $\lambda_6$  correspond bien à des boucles, et est proche du BP  $n$ . La famille  $\epsilon_6$  ( $\beta$ ) n'est pas aussi liée aux blocs  $\beta$  que cela.

De cette simple comparaison, on peut conclure qu'il n'y a pas de correspondance directe entre ces deux types d'alphabet. Une première raison est la différence dans le nombre de blocs utilisés (4 contre 4 fois plus). Une autre cause est sûrement le faible nombre de protéines utilisées dans la première étude. Il semble que ce sont certains repliements qui ont été appris, ainsi la figure 2.13b montre clairement que pour  $\zeta_6$ , censée être caractéristique des boucles et des feuillets  $\beta$ , le bloc le plus proche est le bloc  $m$  soit une hélice  $\alpha$ . Cette famille est donc caractéristique non pas de boucles et des feuillets  $\beta$  simples, mais de boucles et des feuillets  $\beta$  allant vers une

| blocs    | Blocs Protéiques |          |          |          |          |          |          |          |          |          |          |          |          |          |          |          |
|----------|------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|          | <i>a</i>         | <i>b</i> | <i>c</i> | <i>d</i> | <i>e</i> | <i>f</i> | <i>g</i> | <i>h</i> | <i>i</i> | <i>j</i> | <i>k</i> | <i>l</i> | <i>m</i> | <i>n</i> | <i>o</i> | <i>p</i> |
| $\eta$   |                  | 5,7      |          |          |          |          |          | 11,4     | 1,6      | 2,2      | 11,9     | 9,2      | 23,1     | 15,4     | 11,7     | 2,7      |
| $\alpha$ |                  |          |          |          |          |          |          |          |          |          |          | 10,6     | 86,1     |          |          |          |
| $\tau$   | 7,1              | 6,4      | 10,2     | 16,0     |          | 2,2      | 5,0      |          | 12,7     |          |          | 5,2      | 7,3      | 1,2      | 8,2      | 16,4     |
| $\zeta$  | 2,0              |          | 1,5      | 1,3      | 9,4      | 34,9     | 1,9      | 5,1      |          | 2,1      | 27,2     | 2,7      | 9,0      |          |          | 1,1      |
| $\iota$  | 7,1              | 10,6     | 15,8     | 40,4     | 6,5      | 3,7      | 2,0      | 1,9      | 1,5      | 1,1      | 1,0      | 1,0      | 1,2      |          |          | 4,9      |
| $\beta$  | 7,2              | 7,1      | 18,3     | 53,2     | 2,4      | 6,1      |          | 1,6      |          |          |          |          |          |          |          | 1,9      |

TAB. 3.7 – Correspondance des 6 blocs définis par Fetrow et collaborateurs [53] avec les 16 BPs (seules les fréquences supérieures à 1,0 % sont notées, celles en gras sont supérieures à 10,0 %).

hélice  $\alpha$ . Ayant fait des apprentissages avec un nombre aussi limité de blocs, le résultat est surprenant. Pour avoir ce type de blocs caractéristiques, il faudrait au moins deux fois plus de blocs. La seule possibilité est que les protéines utilisées aient été particulièrement riches en certains repliements particuliers.

### 3.4.3 Comparaison avec les super-structures secondaires de Fetrow et collaborateurs

Le tableau 3.7 montre la correspondance entre les 6 blocs structuraux décrits par Fetrow et collaborateurs [53] et les 16 blocs protéiques. La comparaison est faite en observant la correspondance entre les protéines codées avec les 6 blocs structuraux (partie centrale, 5  $C_\alpha$  de 7 utilisés) et ces protéines codées en 16 blocs protéiques.

Le bloc structural  $\alpha$  est concentré sur les blocs protéiques  $l$  et  $m$  (96,7 %) et correspond donc à une hélice  $\alpha$  et une extrémité N-terminale d'hélice  $\alpha$ . Les autres blocs structuraux sont moins bien caractérisés. Le bloc structural  $\eta$  est plus flou et représente surtout un bloc non- $\beta$  avec une certaine tendance hélice  $\alpha$ . A l'inverse, le bloc  $\beta$  est surtout un bloc non- $\alpha$  complet. Le bloc  $\iota$  possède une forte similitude avec ce dernier. Toutefois, il est moins régulier car il a une plus forte propension de BPs "boucles" (PB  $g$  à  $j$ ) et surtout il a l'entrée  $\beta$  N-terminale PB  $b$ . Le bloc structural  $\zeta$  est encore un bloc lié au feuillet  $\beta$ , moins caractéristique. Seul le bloc structural  $\tau$  représente bien les "boucles".

Tout comme pour le précédent exemple, aucune correspondance directe n'est envisageable, du fait de la différence du nombre de blocs et des tailles utilisées. Cependant, il semble que cet alphabet soit assez pauvre, mettant le poids principalement sur les structures répétitives et limitant ainsi le rôle des boucles. Une explication possible est l'utilisation de différents types

| PBs | SBBs        | $\alpha_1$  | $\alpha_2$  | $\alpha'$ | $\alpha'_+$ | $\alpha'_-$ | $\gamma_{\alpha\beta}$ | $\gamma_{\beta\alpha}$ | $\gamma_1$  | $\gamma_2$  | $\gamma_\beta$ | $\beta_2$   | $\beta_1$   |
|-----|-------------|-------------|-------------|-----------|-------------|-------------|------------------------|------------------------|-------------|-------------|----------------|-------------|-------------|
| a   |             |             | 0,2         |           | 2,1         | 0,2         | 11,4                   | 0,4                    | 3,5         | 0,6         | 3,9            | 0,6         | <b>77,0</b> |
| b   |             |             | 6,2         | 2,3       | <b>47,6</b> | 0,4         | 8,8                    | <b>26,1</b>            | 4,9         |             | 0,8            | 1,5         | 1,5         |
| c   |             |             | 0           | 0,5       | 2,6         | 0,2         | 1,7                    | 2,2                    | 0,7         | <b>21,9</b> | 9,4            | <b>29,1</b> | <b>31,7</b> |
| d   |             |             | 0,1         |           | 1,3         | 0,1         | 0,3                    | 1,1                    | 0,9         | 16,4        | 9,3            | <b>55,8</b> | 14,7        |
| e   |             |             | 0           |           | 7,7         |             | 0,4                    | 3,5                    | 2,1         | 18,3        | 15,5           | <b>45,1</b> | 7,4         |
| f   |             |             | 0           | 0,1       | 4,3         |             | 0,6                    | 5,6                    | 1,2         | <b>36,8</b> | 8,4            | <b>27,6</b> | 15,3        |
| g   |             | 10,7        | 19,7        | 4,1       | 1,6         | 8,2         | 3,3                    | 1,6                    | 2,5         | 1,6         | 5,7            | <b>41,0</b> |             |
| h   |             |             | 0,7         |           | 6,8         | 3,9         | 0,7                    | 2,5                    | 11,1        | <b>22,9</b> | <b>23,3</b>    | 17,6        | 10,4        |
| i   |             |             |             | 0,9       |             | 18,9        | 0,5                    | 0,9                    | <b>76,5</b> |             | 0,9            |             | 1,4         |
| j   |             |             |             |           |             | 2,8         |                        |                        | 10,4        | 2,8         | <b>54,7</b>    | 0,9         | <b>28,3</b> |
| k   |             |             | 0,2         |           | <b>31,1</b> | 0,3         | 2,3                    | <b>55,0</b>            | 9,8         | 0,2         | 0,2            | 0,5         | 0,6         |
| l   | 10,7        | <b>64,1</b> | 0,8         | 1,8       | 0,9         | 9,2         |                        | 5,4                    |             |             | 0              |             | 7,2         |
| m   | <b>65,1</b> | 18,4        | 6,6         | 1,1       | 0,1         | 4,4         | 0,4                    | 0,9                    |             |             | 0              | 0,1         | 2,9         |
| n   | <b>35,3</b> | <b>27,5</b> | 14,9        | 3         | 1,5         | 7,1         | 2,2                    | 4,1                    | 0,4         | 0,4         |                |             | 3,7         |
| o   | <b>24,8</b> | <b>31,9</b> | <b>20,5</b> | 1,4       | 2,8         | 5,7         | 0,9                    | 6                      | 0,3         |             |                |             | 5,7         |
| p   |             | 15,9        | 7,6         | 2,2       | <b>55,4</b> | 4           | 0,4                    | 4,3                    | 0,7         | 0,2         | 0,7            | 8,5         |             |

TAB. 3.8 – Correspondance entre les 12 blocs protéiques (SBBs) définis par des chaînes de Markov cachées par Camproux et collaborateurs [23] (centré sur le troisième  $C_\alpha$ ) et les 16 blocs protéiques (centrés sur le troisième  $C_\alpha$ ). Les fréquences données sont relatives à chacun des 16 blocs.

d'informations.

### 3.4.4 Comparaison avec les blocs structuraux de Camproux et collaborateurs

Les deux précédents exemples montrent bien l'intérêt d'un alphabet plus complet avec un nombre d'états supérieurs à une dizaine. La comparaison avec l'alphabet à 12 états mis au point par Anne-Claude Camproux qui a utilisée la méthode des Chaînes de Markov Cachées est donc particulièrement intéressant. L'analyse est en deux parties. La première étant simplement la recherche de l'équivalence entre le bloc protéique et le bloc structural. La seconde tient compte du fait que les blocs structuraux (SSBs) font 4  $C_\alpha$ , et donc un bloc protéique est représenté par une série de deux blocs structuraux.

Les répartitions des blocs sont nettement plus déterministes qu'auparavant. Par exemple, le bloc structural  $\alpha_1$  ne correspond qu'aux blocs associés aux hélices (PBs  $l$  à  $o$ ). Plus de 80% du BP  $m$  se retrouve en  $\alpha_1$  et en  $\alpha_2$ , et inversement, le BP  $d$  ne leur est jamais associé. Toutefois, pour les blocs correspondant de part et d'autres aux boucles, les situations sont nettement plus

diverses. Le bloc structural  $\gamma_1$  est principalement lié aux PBs  $h$  à  $k$ , alors que le bloc structural  $\gamma_{\alpha\beta}$  est réparti dans un nombre nettement plus grand de blocs protéiques. De même, les blocs associés aux feuillettes sont particulièrement contrastés. Ainsi, le PB  $a$  est lié principalement au bloc structural  $\beta_1$ , mais le bloc structural  $\beta_1$  est lui présent dans de nombreux PBs distincts (PBs  $a$ ,  $c$ ,  $d$ ,  $f$  et  $j$ ).

Les blocs protéiques associés aux hélices  $\alpha$  sont les blocs les plus directement reliables aux blocs structuraux avec un nombre limité de successions préférentielles et surtout quelques unes fort importantes. Le PB  $m$  est totalement compris dans les blocs structuraux  $\alpha_1$  et  $\alpha_2$ .

Cette étude montre qu'il y a une convergence certaine entre ces deux approches malgré des informations et la méthode différentes, les alphabets obtenus possèdent une grande spécificité. Leur utilisation conjointe serait d'un grand intérêt.

## 3.5 Conclusion

Pour clore cette première partie sur l'alphabet structural, plusieurs points ont été vus. Tout d'abord sur la méthode utilisée qui tient compte d'un aspect statistique simple avec une prise en compte de l'aspect séquentiel des protéines.

Ayant testé un nombre important de série de blocs, la série de 16 blocs protéiques a été conservée. Elle montre une approximation particulièrement correcte de la structure tridimensionnelle des protéines. De plus, les blocs sont bien distincts entre eux et les fragments protéiques qui leurs sont associés ne sont proches réellement que d'un seul bloc. Ils permettent un recodage particulièrement efficace des structures 3D, avec plus de 50 % des angles approximés à moins de  $10^\circ$ . L'utilisation du *RMSda* comme critère de sélection a aussi démontré son efficacité. La comparaison avec les autres alphabets met en évidence la pertinence d'un choix de plus d'une dizaine de blocs.

En outre, la grande variété des BPs permet de mieux analyser la structure protéique que la structure secondaire et d'observer des distributions informatives en acides aminés. La quantification de cette information montre la spécificité de nombreux blocs, nous allons donc utiliser cette information dans le chapitre suivant avec une méthode de prédiction de la structure locale à partir de la séquence.