Chapitre 6

Compactage des structures tridimensionnelles des protéines

6.1 Objectif

Dans ce chapitre, nous présentons une nouvelle approche de compactage des structures 3D des protéines, labellée méthode de la protéine hybride (MPH). L'objectif de MPH répond à un problème lié aux principes même des méthodes de classification. Dans une méthode de classification par partition type nuées dynamiques ou carte de Kohonen, le nombre de classes est fixé, et ces classes sont disjointes. La proximité entre les classes obtenues ne peut être vu qu'*a posteriori*. Les chaînes de Markov Cachées [149] sont une alternative à ce type de problème, toutefois, elles se basent sur la donnée des lois *a priori* des observations, ce qui est génant quand les lois appropriées ne sont pas connues. Appliqué à notre recherche, cela revient à considérer indépendant chacun des blocs, alors qu'une continuité existe entre un bloc *y* et un bloc *y'*. Le principal intérêt de MPH sera de mettre directement en relation les groupes pour tenir compte de la continuité. Le réseau discontinu (cf. chapitre 5) nous a démontré que des séries de blocs protéiques existent, avec MPH, nous allons observer des séries "floues" de blocs protéiques et ainsi analyser si des séries telles que, par exemple, *uvwxyz* et *uv'wxy'z* sont réelleement distinctes. Avoir un alphabet structural plus long (des séries de blocs protéiques) et plus "flou" (*uvwxyz==uv'wxy'z*), nous permetterait en effet d'avoir des séries équivalentes plus nombreuses, donc de réduire par la suite le nombre de fragments à utiliser pour des méthodes de recherche d'homologie structurale ou d'enfilage.

6.2 Principe général de la méthode de la protéine hybride

Dès mon DEA, avec le Pr. Hazout, nous avons travaillé sur cette problématique. Notre méthode se base sur une observation: "dans les protéines, s'il y a une entrée dans un feuillet, il y a une partie centrale puis une sortie qui va aller vers une boucle, et ainsi de suite jusqu'à la fin de la protéine". Il y a donc une continuité logique, une architecture locale forte. Ainsi, quand on fait une liste des séries de blocs, on observe des successions préférentielles, ceci permet même la construction d'un graphe, d'un réseau exprimant ces transitions (cf paragraphe 5.2).

En conséquence, il est possible de concevoir qu'une série X_i d'observations, peut aller vers une série X_{i+1}, et celle-ci ira vers une série X_{i+2}, et ainsi de suite. De même, la série X_i va aller vers une série X'_{i+1} qui ira vers X_{i+2} ensuite; X_{i+1} et X'_{i+1} se différenciant fort peu. Dans notre cas, par exemple, cela peut correspondre à des sorties de feuillet qui sont légèrement différentes. Avec une méthode classique de partitionnement, il est difficile de ne pas surdécouper le problème et ainsi considérer X_{i+1} et X'_{i+1} distincts.

Aussi, pour appréhender ce problème, nous avons élaboré une méthode dite "Méthode de la Protéine Hybride" (en francais, MPH, en anglais, "Hybrid Protein Model" ou HPM [36, 37]). La protéine hybride est une matrice de longueur L et de dimension l, cette dernière étant la dimension des observations (par exemple, 20 si l'on travaille sur les acides aminés ou 16 pour les blocs protéiques). En chaque position j, au départ, pour chaque type d'observation, une valeur moyenne est mise (la fréquence des acides aminés ou des blocs par exemple), ce qui représente une valeur attendue aléatoirement.

Le principe de l'apprentissage consiste à tirer des observations de longueur p aléatoirement dans la base de données, et, à les placer au mieux dans la matrice par une technique proche des cartes auto-organisées. Ainsi, petit à petit les tendances se focaliseront. Avec des paramètres d'apprentissage correctement choisis, à la fin de l'apprentissage, toutes les observations ressemblant à X_i seront en un site donné j_i et celles ressemblant à X_{i+1} et X'_{i+1} seront en j_{i+1}. En conséquence, il y a un regroupement local qui permet, si un apprentissage a été effectué avec des observations de longueurs p, d'avoir des successions qui permettent une analyse d'un ordre supérieur à p.

Cette approche a été appliquée aux séries de blocs protéiques (cf. paragraphe 6.3) et aussi lors

d'un apprentissage commun entre la séquence et la structure à l'aide de descripteurs physico-chimiques et d'angles dièdres (cf. parapgraphe 6.5). Cette approche a été appliquée à l'analyse de données génomiques par un Chromosome Hybride et a donné des résultats intéressants sur les régions subtélomériques des chromosomes de la levure, dont les réarrangements sont particulièrement étudiés [16, 1].

6.3 Application au compactage des structures

6.3.1 Objectif

Prédire la structure tridimensionnelle des protéines à partir de la séquence n'est pas une tâche aisée. Pour des protéines ayant des taux d'identité de séquences supérieurs à 30%, diverses techniques existent, elles se basent principalement sur des modélisations avec contraintes spatiales, physico-chimiques et des méthodes statistiques [15, 169, 93]. L'absence de protéines proches implique l'utilisation de méthodes différentes comme la prédiction *ab initio* (cf. paragraphe 2.2.2.3), qui toutefois est limitée à des protéines de petite taille [39, 137]. D'autres approches existent comme l'enfilage (ou "threading", cf. paragraphe 2.2.2.2) qui consiste en une recherche de compatibilité entre une séquence et un grand nombre de fragments issus de structures de protéines connues [126, 153, 7, 110]. Toutes ces techniques nécessitent l'utilisation d'une base de données protéiques non redondante (cf. paragraphe 2.2.1.4). Actuellement sont disponibles des bases de données reposant sur un critère de similitude de séquence [84, 83], et de similitude structurale [126, 87]. Pour cette première approche, MPH va donc servir à la compression d'une base de données structurales en une protéine hybride, ce qui pemettra dl'obtention d'un nombre limité de conformations locales, de repliements. Réduire le nombre de conformations locales peut être intéressant pour des recherches comme l'enfilage ou la recherche d'homologie structurale.

6.3.2 Principe

Ce chapitre va donc mettre en avant l'utilisation de MPH dans la compression d'une base de données structurales en une protéine hybride. Dans l'étape d'apprentissage, pour pouvoir compacter la structure, nous avons utilisé la traduction des protéines de cette base structurale

en suite de blocs protéiques. Chaque bloc représente toujours 5 C_α consécutifs, l'utilisation de séries de blocs va permettre la concaténation de différents repliements.

La compaction des structures générera des structures locales avec des parties communes et d'autres moins déterminées. Cette utilisation de séries de blocs dans la protéine hybride est à mettre en relation avec le fait que le choix de la longueur et le nombre des blocs a toujours été une question ardue donnant lieu à de nombreuses réponses : 6 blocs longs de 7 résidus pour Fetrow et collaborateurs [53], 4 à 7 blocs pour une longueur de 4 résidus pour Rooman et collaborateurs [162], 12 blocs ayant 4 résidus pour Camproux et collaborateurs [23, 24], 13 blocs pour des longueurs variables pour Bystroff et Baker [19], et une centaine d'hexamères pour Unger et collaborateurs [198] et Schuchhardt *et al.* et collaborateurs. Les structures locales comprises dans la protéine hybride nous permetterons de passer au-dessus de cet écueil en proposant des fragments longs, mais composés de blocs plus courts.

6.3.3 Application du Modèle de la Protéine Hybride aux séries de blocs protéiques

Pour compacter notre base de données structurales, nous avons donc utilisé la méthode de la protéine hybride (cf. paragraphe 6.2). La protéine hybride est une protéine chimérique composée de N sites où chaque position i n'est pas définie par un seul bloc, mais par une loi de probabilité $f_i(b_x)$, b_x étant un des 16 BPs (x=1, 2,..., 16).

La figure 6.1 récapitule les différentes étapes de l'apprentissage. Dans un premier temps, la base de données de 553 protéines ayant moins de 25% de similitude de séquence (cf. paragraphe 3.3.2.6) établie par Romain Gautier a été traduite en terme de blocs protéiques.

Ensuite, l'apprentissage proprement dit consiste en une recherche de la position optimale de chaque fragment dans la protéine hybride (cf. figure 6.1c à 6.1f). L'apprentissage est proche des cartes auto-organisées (cf. Annexe 1), mais sans processus de diffusion. Ainsi, la protéine hybride correspond à des séries de structures locales "floues".

L'intérêt principal de la méthode est de maintenir la séquentialité pour créer des séries de structures locales successives ayant un fort taux de recouvrement. Cette stratégie diffère donc fortement d'une classification classique où les groupes sont indépendants.

Un fragment $\boldsymbol{F}$ est tiré aléatoirement dans la base de données (cf. figure 6.1c). Chaque

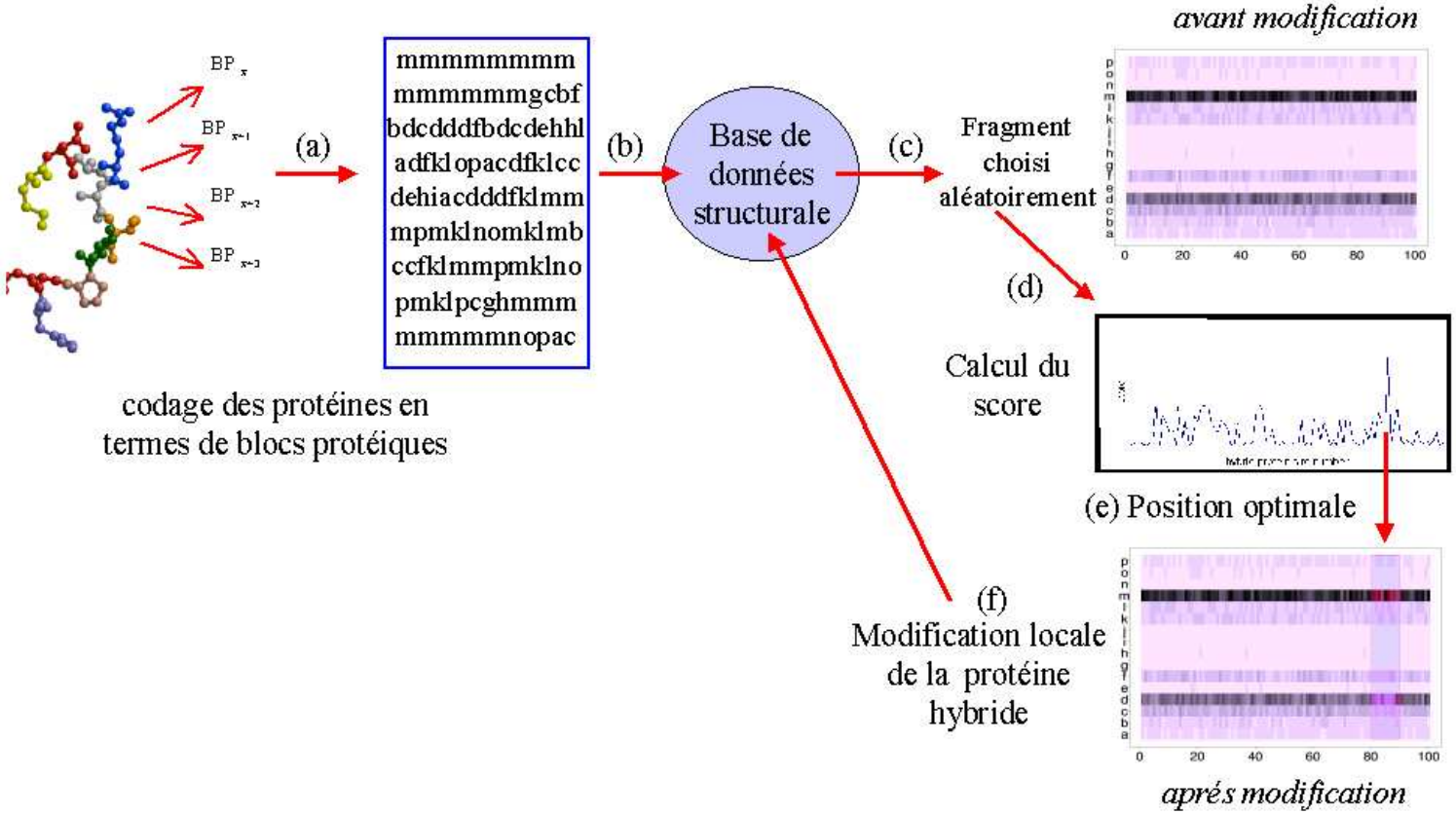

FIG. 6.1 – *Principe de l'apprentissage de la protéine hybride des blocs protéiques.(a) Codage des structures tridimensionnelles des protéines en blocs protéiques . (b) Constitution d'une base de données structurales. (c) Tirage aléatoire d'un fragment. (d) Calcul d'un score de similitude en chaque position de la protéine. (e) Recherche de la position optimale. (f) Modification de cette position et reprise à partir de (c) jusqu'à la stabilisation du système.*

structure locale $\boldsymbol{F}$ est définie par L blocs consécutifs $\{b_1^*, b_2^*, ..., b_L^*\}$ (ici L=10 BPs, soit 14 C_α).

Localisation de la structure locale dans la protéine hybride. Un score S_i est calculé en chaque position i de la protéine hybride :

$$S_i = \sum_{k=1}^{k=L} ln[f_{i+k-1}(b_k^*)]$$

avec k=1,2,...,L. Ainsi le score S_i (i.e. le logarithme de la vraisemblance d'observation de la structure locale $\boldsymbol{F}$ en un site donné i) mesure la similitude entre la structure locale et une région donnée de même dimension dans la protéine hybride (cf. figure 6.1d). La région la plus proche de la structure locale possède le moins de différence avec celle-ci et donc son score est maximal. Ainsi cette position i_0 est définie comme (cf. figure 6.1e):

$$i_0 = \ argmax[S_i]$$

Modification locale de la protéine hybride. Ayant trouvé une zone de plus forte ressemblance, cette zone va être légèrement modifiée pour ressembler à la structure locale $\boldsymbol{F}$. Les positions i_0 à i_0+L-1 seront ainsi modifiées (cf. 6.1f):

si $b = b_k^*$ (i.e. le bloc protéique présent à la position k dans la structure locale), alors

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b) + \alpha}{1+\alpha}$$

si $b \neq b_k^*$ (i.e. les autres blocs protéiques présents à la position k), alors

$$f_{i_0+k-1}(b) \leftarrow \frac{f_{i_0+k-1}(b)}{1+\alpha}$$

Le symbole $\leftarrow$ signifie que la valeur calculée remplace la valeur précédente. Le coefficient d'apprentissage α est encore égal à $\alpha_0/(1 + t/\nu)$, avec α_0 le taux initial d'apprentissage (ici $\alpha_0 = 0.1$), t le nombre de structures locales de L blocs déjà utilisées dans l'apprentissage et ν le nombre de structures locales présentes dans la base de données. L'apprentissage est progressif, ainsi l'ensemble de la base de données a été examiné entièrement C fois, ici C = 15 (phases: figure 6.1c à la figure 6.1f pour chaque structure locale de la base de données). Un passage complet de la base de données est appelé cycle. Pendant le premier cycle, le coefficient

d'apprentissage α est maintenu constant ($\alpha = \alpha_0$) pour modifier de façon importante la protéine hybride et ainsi faire moins jouer l'initialisation [109].

Initialisation. La protéine hybride est initialement définie par une série de N distributions sur les blocs $f_i(b_x)$ quasiment identiques :

$$f_i(b_x) = f(b_x).(1 + \epsilon_i)$$

avec $f(b_x)$ la fréquence du bloc protéique b_x dans la base de données, ϵ_i est une valeur aléatoire tirée dans l'intervalle [-τ;+τ] (τ a été fixé à 0,20). En chaque position la somme des probabilités des blocs a été réajustée à 1, en recalculant :

$$f_i(b_x) \leftarrow \frac{f_i(b_x)}{\sum_{i=1}^{i=16} f_i(b_x)}$$

Il faut noter que la protéine hybride est "fermée" dans le sens où le Nème site est contigu avec le premier, ainsi, il n'existe pas d'effet de bord.

6.3.4 Résultats

6.3.4.1 La protéine hybide

La figure 6.3 donne les résultats de l'apprentissage de la protéine hybride après $C = 15$ cycles à partir d'une protéine hybride initiale avec un coefficient d'initialisation $\tau = 0{,}20$ (cf . figure 6.2) et un coefficient d'apprentissage $\alpha_0 = 0{,}10$. La figure 6.3a montre la composition en blocs le long de la protéine hybride. L'analyse de la protéine hybride montre que les structures secondaires répétitives sont clairement détectables (les hélices α avec le bloc m et les feuillets β avec le bloc d). Trois types d'hélices α sont distinguables par leurs tailles : 2 à 4 BPs (positions [38:41]), 7 BPs [3:9] et 10 BPs [82:91], et 4 pour les feuillets β: 4 BPs [66:69], 5 BPs [74:78], 8 BPs [51:58], 9 BPs [15:23].

Différentes transitions entre structures secondaires régulières sont trouvées:

- hélice α à hélice α en positions [93:96],
- hélice α à feuillet β [9:14],

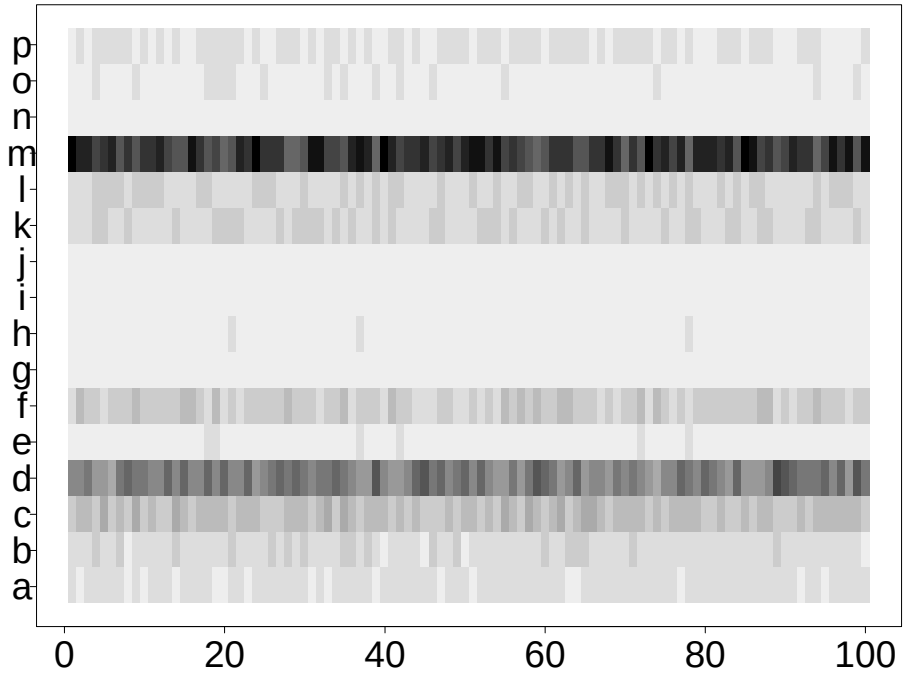

FIG. 6.2 – *Protéine hybride initiale. En abscisse les N sites, en ordonnée les 16 BPs, en chaque position se trouve la fréquence du BP correspondant avec une variation autorisée de τ = 0,20. L'ensemble des fréquences par site est renormalisé à 1,0.*

- feuillet β à hélice α [33:37] et [99:1],

- feuillet β à feuillet β [23:28], [58:65] et [69:71].

La spécificité des sites est nette, en chaque site seuls deux ou trois types de blocs sont présents. De plus, la majorité des structures locales sont consécutives. Ainsi, quand une structure locale $\boldsymbol{F}_j$ est localisée en position i_0 sur la protéine hybride, alors la structure locale $\boldsymbol{F}_{j+1}$, qui suit la structure locale $\boldsymbol{F}_j$, est localisée en la position i_0+1 sur la protéine hybride avec une probabilité de 81 %. La continuité est uniformément présente, aucune zone de cassure particulière n'est visible. Les structures locales sont réparties de manière assez uniforme autour d'une moyenne de 1115 observations par site (cf. Figure 6.3b), seules les positions des hélices α régulières sont largement au-dessus.

Il faut bien noter que chaque site i de la protéine hybride est un ensemble de structures locales de longueur L, ainsi les sites *i-1* et i ont *L-1* distributions de blocs en commun.

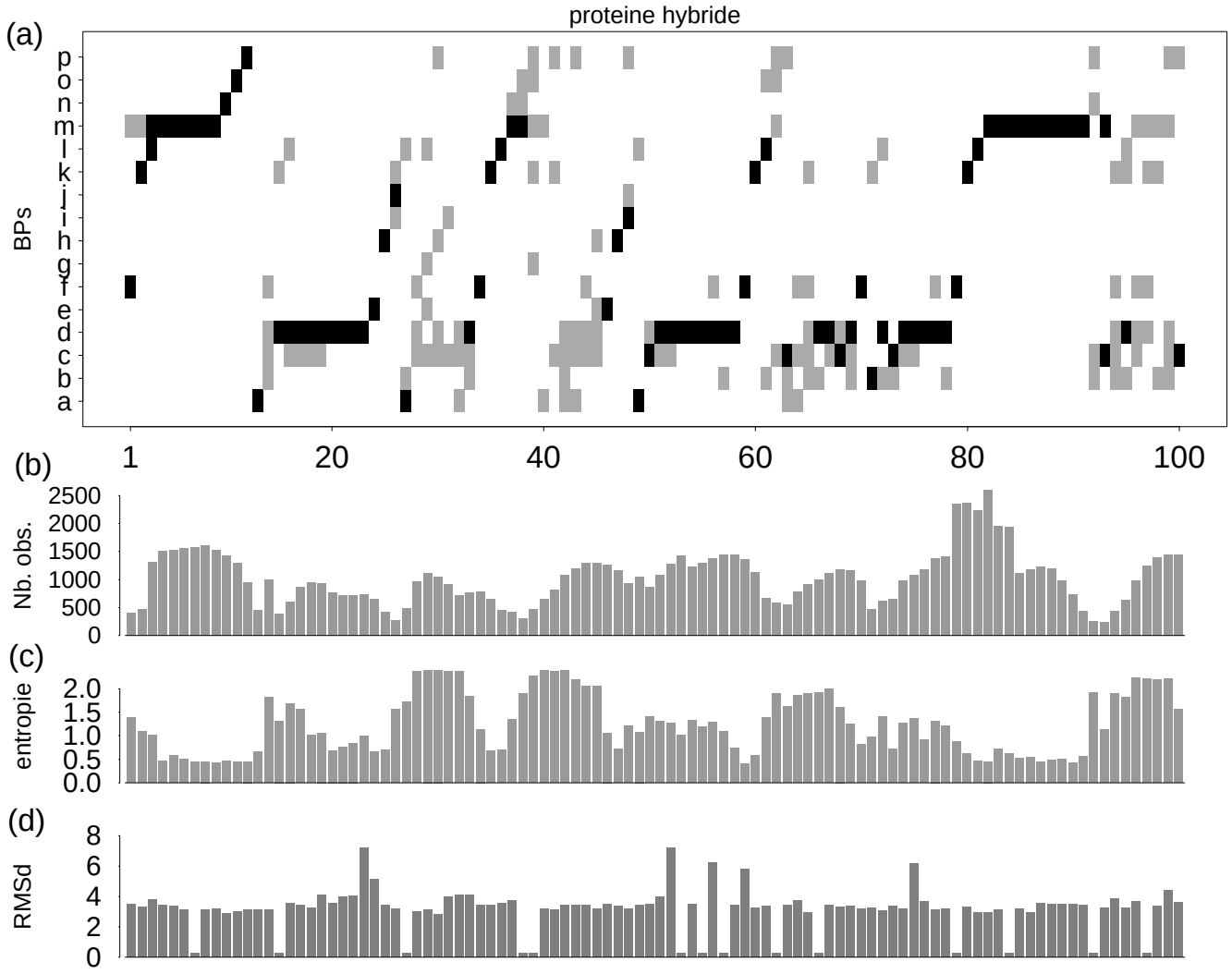

FIG. 6.3 – *Résultats de l'apprentissage. (a) La protéine hybride avec en chaque position, en noir, les fréquences des blocs supérieures à 0,35; en gris, ceux compris entre 0,35 et 0,10; en blanc, ceux inférieurs à 0,10. (b) Le nombre d'observations, (c) l'entropie relative et (d) le* RMSd *moyen de chaque site.*

6.3.4.2 Calcul de l'entropie des distributions des blocs protéiques

Chaque site de la protéine hybride est défini par une loi de probabilité des blocs. Une entropie H_i peut donc être calculée pour quantifier la diversité des blocs en chaque site :

$$H_i = -\sum_{b=1}^{16} f_i(b).ln[f_i(b)]$$

avec i la position du site, b un type de blocs protéiques et f_i sa distribution correspondante. Ainsi, une entropie faible correspond à un site avec peu de blocs hautement probables et donc hautement déterminés.

6.3.4.3 Analyse de l'entropie

L'entropie calculée le long de la protéine hybride est représentée dans la figure 6.3c et montre la haute spécificité de chaque site avec une valeur maximale de 2,40 et minimale de 0,41, 40% des positions ont une entropie inférieure à 1,0. Trois catégories de sites peuvent être distinguées par leur entropie.

Un premier groupe dont l'entropie est inférieure à 1,0. Il contient les sites les moins variables comme les hélices α. Ainsi, pour les positions [4:13] une hélice α centrale est suivie par une sortie C-terminale, les structures locales commencent principalement avec 6 BPm suivis par BPn, BPo, BPp et BPa aux sites [10:13], ceci est alors noté m_6nopa. Les structures locales en positions [79:91] sont des hélices α ayant une extrémité N-terminale $fklm_{10}$. Comme l'apprentissage s'effectue avec des structures locales de 10 blocs protéiques, ce motif contient en réalité les structures locales de type $fklm_7$, klm_8, lm_9 et m_{10}. La protéine hybride permet aussi la création de feuillet β avec extension N-terminale en [20:25] avec un motif d_4eh et des transitions courtes, telles des extrémités C-terminales de feuillet β dfk [58:60] et fk [35:36].

Le second groupe correspond à des zones intermédiaires avec une entropie comprise entre 1,0 et 1,5. Les plus longs correspondent à des feuillets β étendus ayant une extrémité N-terminale en positions [48:57], des structures locales de type iac_xd_{7-x},(x=1,2,3). Ensuite, les deux zones les plus représentatives correspondent à une extrémite N-terminale d'une hélice α en [1:3], fkl, et un feuillet β aux sites [77:78], d_2.

Le dernier groupe comprend les zones ayant une entropie élevée (supérieure à 1,5). Quatre zones sont plus variables et correspondent à des zones de boucles longues ou des feuillets allongés

en positions [26:34] et [38:45], des coudes entre deux feuillets [62:68] ou des boucles entre hélices α avec des feuillet β [94:100].

6.3.5 La stabilité structurale de la protéine hybride

La protéine hybride a donc bien appris l'ensemble des fragments, toutefois, ayant utilisé des structures locales de longueur égale à 10 blocs protéiques (la valeur de L), une question se pose quant à la qualité de l'approximation structurale de cet apprentissage.

Le *RMSd* moyen des fragments de chaque site a donc été calculé (cf. figure 6.3d). Le *RMSd* moyen est de 3,14 Å. Seuls 6 sites sont plus variables avec un *RMSd* moyen supérieur à 5 Å, et 14 sites ont un *RMSd* moyen inférieur à 1,0 Å.

Les zones structurellement variables sont principalement associées aux feuillets β. Par exemple, les structures locales en [19:28] correspondent à deux populations différentes: (i) un feuillet β court (type d_2) allant à un autre feuillet β, (ii) un feuillet β (de type d_4) allant vers une hélice α. De la même manière, les structures locales [48:57] et [52:61] sont associées avec des feuillets β de tailles différentes. Le site 59 correspond à des structures locales comprises entre [55:64] et composées de feuillet β de type $d_4fkopac$, $d_4fklpac$ ou c_2d_2fkopa. Le site 75 [71:80] est principalement composé d'une population de feuillet β $bccd_6f$ et d'une seconde population commençant par $bdcd_3$. En réalité, les zones les plus variables sont des feuillets β avec extrémité N- et C-terminale. Ces structures locales sont associées avec des séries de blocs plus complexes et diversifiées. Elles possèdent localement des séries communes et des parties distinctes, ce qui crée des zones particulièrement floues qui entraînent un *RMSd* moyen plus élevé.

Les structures locales ayant une variabilité faible sont des hélices α (voir tableau 6.1, 5 premières lignes) et quelques structures β (voir tableau 6.1, 5 dernières lignes).

En conclusion, les sites les mieux définis sont associés avec les structures régulières, mais pas exclusivement, comme avec le site 66 qui est une transition entre deux feuillets β.

6.3.5.1 Exemples de sites caractéristiques

Méthode d'analyse Pour quantifier l'importance de chaque type d'acides aminés en chaque position, les occurrences des acides aminés ont été calculées, en comptant pour chaque position

site	motif	caractérisation
7	$m_8$no	hélice α avec N- et C-ter
38	$fklm_8$	courte hélice α
79	d_4fklm_3	transition rapide entre un feuillet β et une hélice α
84	klm_8	une hélice α avec N-ter
92	m_5cmcd_2	hélice α allant à un feuillet β
15	opa	β N-ter
27	$ehia$	entrée en N-cap β
53	$acddd$	feuillet β C-ter
55	$acddd$—-f	feuillet β C-ter
57	$dddd$–f	feuillet β C-ter
62	opa	transition entre deux feuillets β
66	opa	transition entre deux feuillets β
97	cd —- f	feuillet β

TAB. 6.1 – *Sites de la protéine hybride les plus stables suivant le critère du* RMSd *moyen, avec la position des sites (centré sur le cinquième C_α), le motif associé localement et une description en termes de structures secondaires.*

s chaque série de L blocs protéiques. Elles sont ensuite normalisées en Z-scores :

$$Z_j^i = \frac{\left(n_j^i - n_a^i\right)}{\sqrt{n_a^i}}$$

où n_j^i est le nombre observé de résidus a en position j et n_a^i son nombre attendu. Ce dernier est le produit :

$$n_a^i = N_i.\mathbf{q}_a$$

avec N_i et $\mathbf{q}_a$ respectivement le nombre d'observations à cette position et la fréquence du résidu a dans la base de données. Ainsi, les Z-scores positifs (inversement négatifs) correspondent à des sur-représentations (inversement sous-représentations) de ce résidu en cette position (cf. paragraphe 3.3.5.2).

Exemples La figure 6.4 montre pour trois sites d'intérêt la superposition de fragments de longueur 10 C_α appartenant à chaque site et les matrices d'occurrence associées; ces matrices ont été normalisées en Z-scores. Ces sites sont dans l'ordre:

- le site 7 qui correspond aux structures locales [3:12], il a un *RMSd* moyen de 0,3 Å associé à une entropie de 0,43 (cf. figure 6.4a).

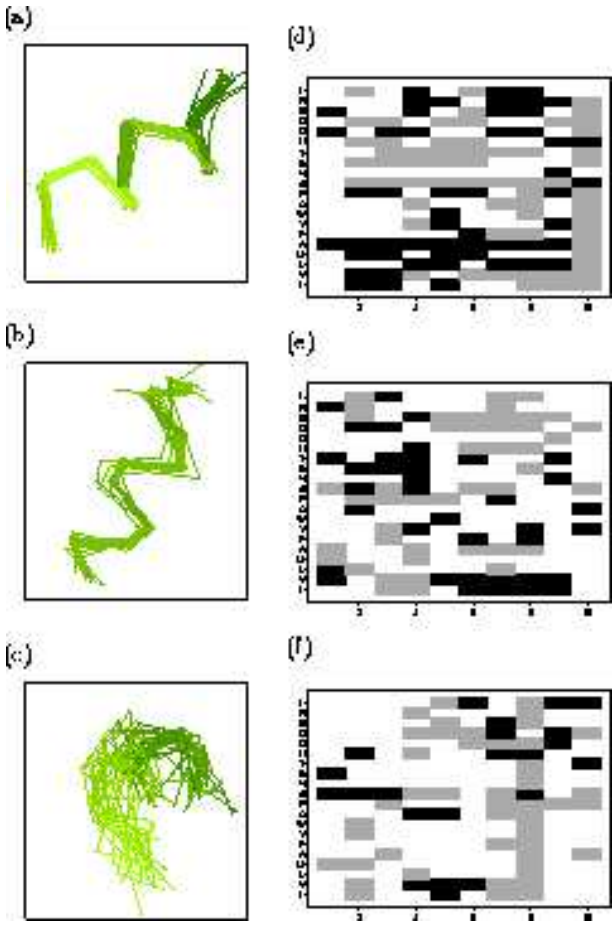

FIG. 6.4 – *Les sites caractéristiques en positions 7, 73 et 56, avec (a-c) les superpositions des fragments associés aux sites, visualisés par XmMol [194] et (d-e) les matrices d'occurrences associées à ces sites, normalisées en Z-scores (noir : Z-scores < 4,4, gris: Z-scores > -4,4 et blanc: Z-scores intermédiaires). Les acides aminés sont classés de bas en haut comme suit: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K.*

- le site 73 [69:78] qui a une entropie de 0,72 et un *RMSd* moyen de 3,3 Å (cf. figure 6.4b).

- le site 56 [52:61], entropie de 1,39 et un *RMSd* moyen de 6,21 Å (cf. figure 6.4c).

Ces trois sites correspondent aux trois principales catégories de sites. Les figures 6.4d à 6.4f montrent les matrices d'occurrences associées normalisées en Z-scores. Le premier site (figures 6.4a et 6.4d) représente une partie centrale d'hélice α avec des sur-représentations en Alanine et des résidus non-polaires. La présence des résidus chargés tels la Lysine, l'Arginine et l'Acide Glutamique en positions 7 et 8 est classiquement associée avec la présence de l'extrémité C-terminale [154], et celles de Glycine et d'Asparagine en position 10 avec un élément de rupture structural. Le second site (figures 6.4b et6.4e) est quant à lui associé avec la présence d'acides aminés aliphatiques caractéristiques des feuillets β. L'extrémité C-terminale des feuillets β est vue avec des sous-représentations de résidus polaires en position 4 et en Isoleucine et Valine aux sites 5 à 9. Le dernier motif (figures 6.4c et6.4f) est moins déterminé, toutefois le repliement des structures locales est globalement similaire. En parallèle du manque de spécificité structurale, la composition en acides aminés est moins informative que précédemment. Seule la position 8 est hautement spécifique avec une sur-représentation de Glycine et d'Asparagine et une sous-représentation de 17 autres acides aminés.

6.3.6 Relation avec la séquence

6.3.7 Analyse de la répartition des acides aminés par l'utilisation des Z-scores

Ayant examiné précédemment trois sites caractéristiques, nous allons dans cette partie analyser la distribution des acides aminés sur l'ensemble de la protéine hybride pour caractériser la pertinence des sites en termes d'acides aminés et ensuite pour analyser leur relation avec les structures locales.

La figure 6.5 montre les acides aminés normalisés en Z-scores de la protéine hybride. Pour obtenir cette information, il suffit de compter en chaque position centrale de chaque fragment de L blocs protéiques le nombre d'occurrences de chaque acide aminé. Les proportions classiques des acides aminés des structures secondaires répétitives sont retrouvées avec les sur-représentations d'Alanine (positions [2:10] et [81:90]) pour les hélices α, des résidus chargés (Lysine, Arginine

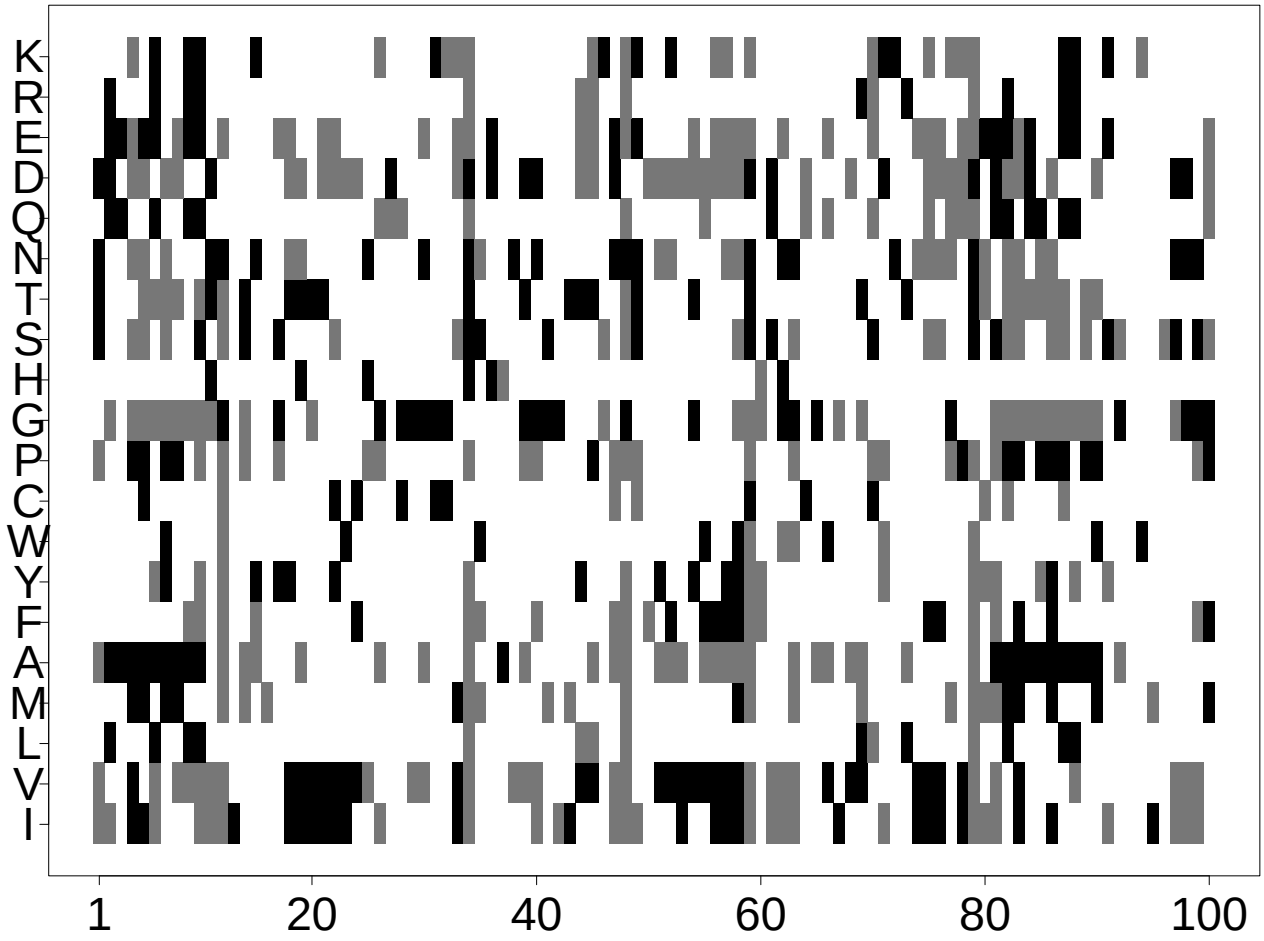

FIG. 6.5 – *Matrice d'occurrences des acides aminés centraux, normalisés en Z-scores (noir: Z-scores < 4,4, gris: Z-scores > -4,4 et blanc intermédiaire), avec en abcisse les 100 sites de la protéine hybride, en ordonnée les acides aminés.*

et Acide Glutamique) pour leur extrémité N-terminale (positions [9:10] et [87:88]), des résidus aliphatiques pour les feuillets β (positions [18:24] et [51:58]) et des Glycines (positions [28:32] et [39:42]) dans les boucles. Quelques spécificités intéressantes peuvent être observées comme la présence de Phénylalanine en extrémité C-terminale des feuillet β en [55:58], mais absente des autres extrémités C-terminales des feuillets β ([21:23], [69:70] et [76:78]).

6.3.7.1 KLd pour quantifier la spécificité des acide aminés

Les répartions des acides aminés attendues ont été retrouvées. Toutefois, elles ne donnent pas d'idées quant à la spécificité de chaque site d'un point de vue statistique. Aussi pour l'évaluer, les données précédentes ont été analysées á l'aide du KLd. Le calcul de l'entropie relative ou mesure de la divergence assymétrique de Kullback-Leibler (noté KLd [112]) a été explicitée dans le paragraphe 3.3.5.1. La spécificité des résidus a été calculée par rapport à ceux de la base de données. La quantité $N_i.K(\mathbf{p}_i,\mathbf{q})$ suit une distribtion de type χ^2 , avec N_i le nombre d'observations (ici, le nombre de structures locales associées au site i). Ainsi, les positions ayant une forte spécificité seront associées à des valeurs élevées.

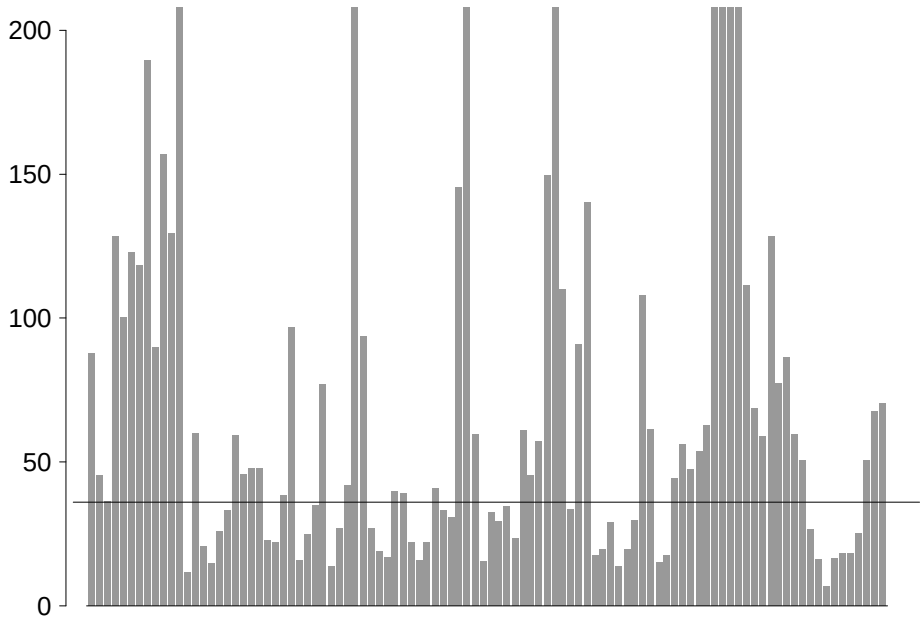

FIG. 6.6 – *KLd associé à la répartition en acides aminés (cf. figure 6.5) en chaque position de la protéine hybride, avec en abcisse les sites de la protéine hybride et en ordonnée la valeur de $N_i \times$ KLd, la valeur seuil est de 36 et représente une erreur de premier ordre de 0,05 avec 19 ddl.*

Un premier résultat visuel net apparaît: les valeurs obtenues sont nettement supérieures à celles calculées précédemment avec les blocs protéiques seuls (cf. Figure 6.6). La valeur seuil utilisée est de 36 soit un risque de première espèce (α) de 0,05 pour 19 degrés de liberté (les 20 types d'acides aminés). Les hélices α et les feuillets β sont présents, mais d'autres positions en dehors des structures répétitives présentent une structure stable comme des boucles en [59:60] incluant la série de blocs *kl*.

6.3.7.2 Similitude des sites de la protéine hybride en termes d'acides aminés

L'ensemble des sites ayant été précédemment caractérisé du point de vue de leur composition en acides aminés, se pose la question de la similitude existante entre ces sites. Pour voir si des sites ayant le même type de distribution du point de vue des acides aminés sont associés aux mêmes structures locales ou non, les sites ont été regroupés avec la méthode de classification *k-means* [79]. Cette méthode permet de regrouper dans un même groupe des observations proches (cf. Annexe 1).

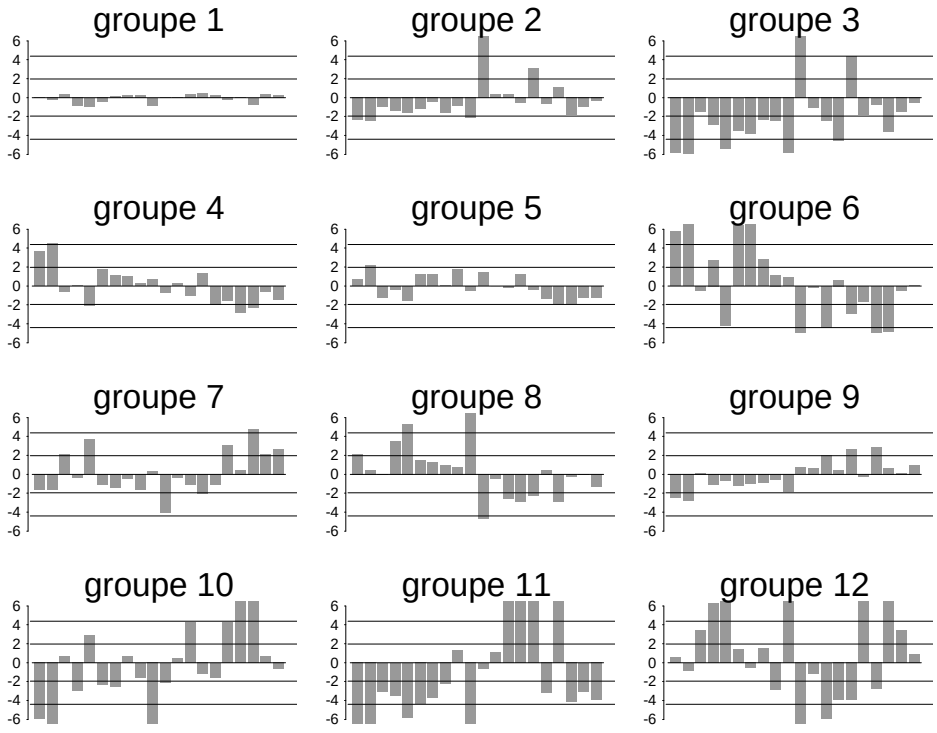

FIG. 6.7 – *Classification par la méthode des k-means des positions des acides aminés transformés en Z-scores (cf. figure 6.6). Les acides aminés sont classés de gauche à droite comme suit: I, V, L, M, A, F, Y, W, C, P, G, H, S, T, N, Q, D, E, R, K.*

Les 100 sites mis en Z-scores (cf. paragraphe 6.3.7) ont été classés en 12 groupes (logiciel R,

librairie *mva*, méthode *kmeans*) . La figure 6.7 présente les 12 groupes obtenus avec les valeurs pour ces 100 sites.

Le groupe 1 (25 sites) ne montre aucune spécificité en acides aminés et possède un nombre très divers de blocs pour chaque site, les 16 types de blocs protéiques se retrouvent associés à ce groupe. Le groupe 2 (5 sites) montre une forte sur-représentation en Glycine et Asparagine, cependant des blocs fort distincts lui sont associés comme les blocs d, j ou m. Le groupe 3 possède le même type de sur-représentation. Cependant il est fortement éloigné du précédent, car il est associé avec des sous-représentations plus fortes et ne correspond, par ailleurs, qu'à des blocs i, j et k qui correspondent à des zones de ruptures de structures régulières.

Les groupes 4 à 6 sont associés à des feuillets β, et les groupes 7 et 8 à des hélices α. Le groupe 4 (17 sites) est associé à une sur-représentation d'acides aminés aliphatiques tels l'Isoleucine et la Valine correspondant à des feuillets β centraux et des extrémités N-terminales de feuillet β. Le groupe 5 (12 sites) est composé de PB d et d'autres blocs associés aux extrémités N- et C-terminales des feuillets β (de PB a au PB f) et lié à une sur-représentation de Valine et une sous-représentation de résidus chargés. Le groupe 6 (1 site en position 58) est un feuillet β régulier avec une forte sur-représentation d'acides aminés non-polaires et une sous-représentation d'acides aminés polaires. Le groupe 7 (10 sites) correspond à une hélice α, plus spécifiquement, une extrémité C-terminale d'hélice α. Une sous-représentation de Leucine, Méthionine et de résidus polaires est retrouvée. Le groupe 8 (8 sites) montre une sur-représentation d'Alanine, de Méthionine, d'Isoleucine et de Proline et une sous-représentation de Glycine; ce groupe ne comprend que les parties centrales des hélices α.

Le groupe 9 (15 sites) avec une sur-représentation de petits acides aminés polaires est caractéristique des changements brusques de structures avec des PBs comme les PBs f, h, a, l et o. Le groupe 10 (1 site en position 81) est associé au PB h; il a une sur-représentation de résidus aliphatiques et de Proline, ainsi qu'une forte sous-représentation des résidus polaires. Le groupe 11 (3 sites) montre une sur-représentation de petits résidus polaires et des sous-représentation de résidus non-polaires et de Proline. Le groupe 12 (1 site en position 70) montre une sur-représentation de Glycine et une sous-représentation des résidus non-polaires. Pour les quatres sites qui composent les deux derniers groupes le bloc protéique le plus important est toujours le PBf, avec cependant une différence notable; dans le groupe 11, le bloc f est dans une série *fkl* alors que dans le groupe 12, il est dans une succession *fbd*. Ce détail est d'importance car,

comme l'a montré le graphe (cf. paragraphe 5.2), le premier fait partie d'une série allant vers des hélices alors que le second est plus présent dans les transitions vers des feuillets β.

6.3.8 Etude de l'influence des paramètres de l'apprentissage

L'influence des différents paramètres utilisés lors de l'apprentissage a été analysée :

(i) la longueur de la protéine hybride $N = 100$ a été choisie pour obtenir une caractérisation correcte des structures locales. Avec $N > 100$, certains sites auraient été fort peu peuplés, et pour $N < 100$, le nombre de structures locales mal approximées aurait augmenté fortement.

(ii) le coefficient d'apprentissage α_0 contrôle à la fois la qualité et la vitesse de l'apprentissage. α_0 pris égal à 0,10 permet de diminuer l'importance de l'initialisation, le système étant bousculé fortement pendant les premiers cycles d'apprentissage. L'utilisation d'une valeur plus faible et plus classique permet un apprentissage plus rapide mais moins sûr.

(iii) Le résultat de la protéine hybride n'est pas fortement modifié par le tirage des fluctuations aléatoires ϵ_i et du coefficient τ. Un léger décalage de la protéine hybride est souvent observé.

(iv) La valeur du nombre de cycles C est définie par l'utilisateur. En pratique, ce nombre est déterminé en se basant sur le fait que plus aucun changement notable n'est observé pendant l'apprentissage au delà d'un certain nombre de cycles. Dans la pratique, les 5 premiers cycles déterminent l'apprentissage dans les grandes lignes.

En conclusion, l'influence des paramètres est relativement mineur. La séquentialité qu'implique les blocs protéiques et la présence de structures comme les structures répétives et leurs entrées et sorties permettent d'obtenir des résultats stables.

6.3.9 Conclusion de l'apprentissage

MPH est donc une nouvelle approche qui permet le compactage d'une base de données de structures de protéines. L'étape d'apprentissage permet de conserver une excellente séquentialité entre les fragments (81% sont contigus). L'entropie montre que la caractérisation des

sites est correcte avec un maximum de trois types de blocs protéiques différents par sites. Le déterminisme étant important, des structures similaires sont classées dans les mêmes régions de la protéine hybride.

Le calcul des *RMSd* moyens a montré la forte stabilité des structures locales et ceci malgré la grande hétérogénéité de ces structures. Les *RMSd*s les plus importants sont dus principalement à des sites contenant à la fois des feuillets β courts et d'autres longs. Un point important est, que malgré un niveau élevé de *RMSd*, les repliements associés à ces sites possèdent des formes similaires. En outre, les structures répétitives ne sont pas les seules qui soient bien approximées.

L'exemple donné à la figure 6.5 montre bien trois catégories caractéristiques de sites avec leurs compositions plus ou moins marquées en acides aminés. Cette caractérisation en acides aminés montrée dans la figure 6.7 et la classification réalisée à partir de ces données montrent que plus de 75 % des sites sont informatifs. Ceci corrobore le fait qu'une partie des séquences protéiques ne sont pas informatives [177]. De plus, ces distributions en acides aminés sont souvent associées avec des distributions caractéristiques en blocs protéiques, ce qui peut être pris en compte dans le cas d'une prédiction de la structure par la séquence.

En conclusion de cette première partie, nous avons vu que la méthode de compaction par MPH permet d'obtenir une protéine hybride qui représente bien la majorité des structures locales. De plus, la composition en acides aminés est fortement spécifique. L'ensemble de ces données en prenant en compte le fait que la continuité est assurée fait de cette protéine un outil intéressant pour l'avenir. Dans les paragraphes suivants, la protéine hybride va servir à mettre en place une méthode rapide de recherche d'homologie structurale entre deux protéines.

6.4 Application à la recherche d'homologie

6.4.1 Principe de la recherche

L'intérêt principal de MPH appliqué séries de blocs protéiques est bien sur la compaction d'une base de donnée structurale. Toutefois, l'utilisation de la protéine hybride n'est pas limitée qu'à ce seul point. En chaque position, un ensemble de fragments proches se trouve localisé. Donc, si l'on recherche des fragments similaires courts d'un point de vue structural il suffit de trouver la position correspondante sur la protéine hybride. Pour trouver des fragments

plus longs, le principe est identique. Des fragments longs et similaires structuralement seront théoriquement localisés aux mêmes positions sur la protéine hybride. En outre, la continuité étant forte, ces positions seront souvent consécutives.

Nous avons donc mis au point une méthode qui tient compte de ces concepts pour rechercher des fragments structuraux proches. La figure 6.8 explicite l'ensemble du processus. Une première étape consiste à recoder la structure tridimensionnelle des protéines en blocs protéiques. A ce niveau simple, la recherche d'homologie est difficile, les structures répétitives gênant particulièrement la recherche. Aussi, la méthode utilise directement la protéine hybride.

Les séries de blocs, structures locales sont recodées en position sur le protéine hybride. Ensuite, un dotplot est calculé. Un dotplot est une matrice de taille N_1 par N_2 avec N_1 longueur de la première protéine sur N_2 longueur de la seconde protéine. Cette matrice est remplie comme suit :

si les positions sont les mêmes sur la protéine hybride $\rightarrow 1$.

si les positions sont différentes sur la protéine hybride $\rightarrow 0$.

Dans un premier temps, le dotplot est une matrice composée de 0 et de 1. Cette information doit donc être travaillée pour ne conserver que les plus longues successions identiques. Ainsi, le dotplot est filtré en ne conservant que les diagonales de longueur G où un nombre H de points est égal à 1. Travaillant sur des protéines proches, H a été pris égal à G. Au départ G a été pris élevé pour extraire les diagonales les plus longues, puis progressivement a été réduit. Les deux protéines étant proches et ayant des longueurs comparables entre 400 et 450 acides aminés, l'extraction des diagonales n'a été effectuée que dans une zone médiane [$-\theta$,$+\theta$]. Cela veut dire que pour une position c d'une protéine, la recherche de similitude s'est focalisée dans une région [c-θ,c+θ] de la seconde protéine et inversement. θ était fixé à 50 résidus. En outre, quand une diagonale était sélectionnée, les zones correspondantes dans les deux protéines n'étaient plus réutilisées ce qui permet de construire des véritables séries de fragments similaires distincts.

Enfin, pour vérifier que le fait de passer par les blocs protéiques, puis le recodage par la protéine hybride n'entraîne pas de faux positifs, chaque couple de fragments similaires a été superposée et le *RMSd* résultant calculé. Cette approche permet ainsi la recherche de structures possédant un repliement proche.

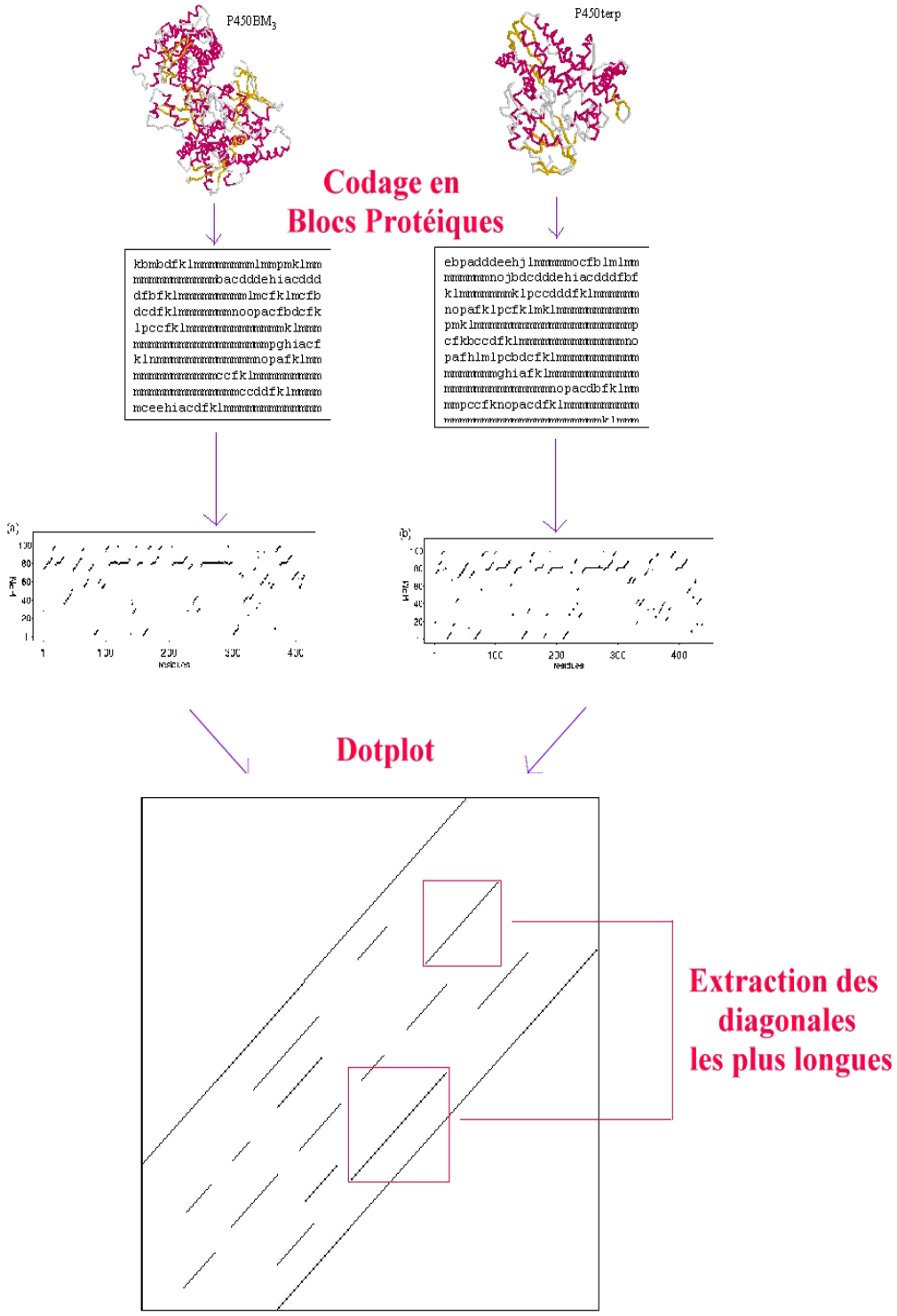

FIG. 6.8 – *Principe de la recherche d'homologues structuraux. Avec de bas en haut, le codage des deux structures tridimensionnelles en blocs protéiques, puis la localisation des séries de blocs protéiques sur la protéine hybride et enfin le dotplot des positions sur la protéine hybride des deux protéines et enfin la recherche des fragments les plus longs en commun.*

6.4.2 Homologie structurale pour deux cytochromes P450

Les cytochromes P450 ont été bien décrits. Ils ont une identité de séquence allant de 10 à 30%. Haseman et collaborateurs [80] les ont caractérisés en termes de fragments communs de structures secondaires (hélices α, 3_{10}-helix, π-helix, feuillet β et β-bugle). Jean et collaborateurs [94] ont déterminé des blocs structuraux communs (Common Structural blocs ou CSBs) entre differents cytochromes P450. Ils ont utilisés des comparaisons deux à deux pour ainsi aboutir à la modélisation d'un nouveau cytochrome P450. Nous avons utilisé deux cytochromes P450: $P450_{BM3}$ (code PDB: 2hpd[151]) et $P450_{terp}$ (code PDB: 1cpt[81]).

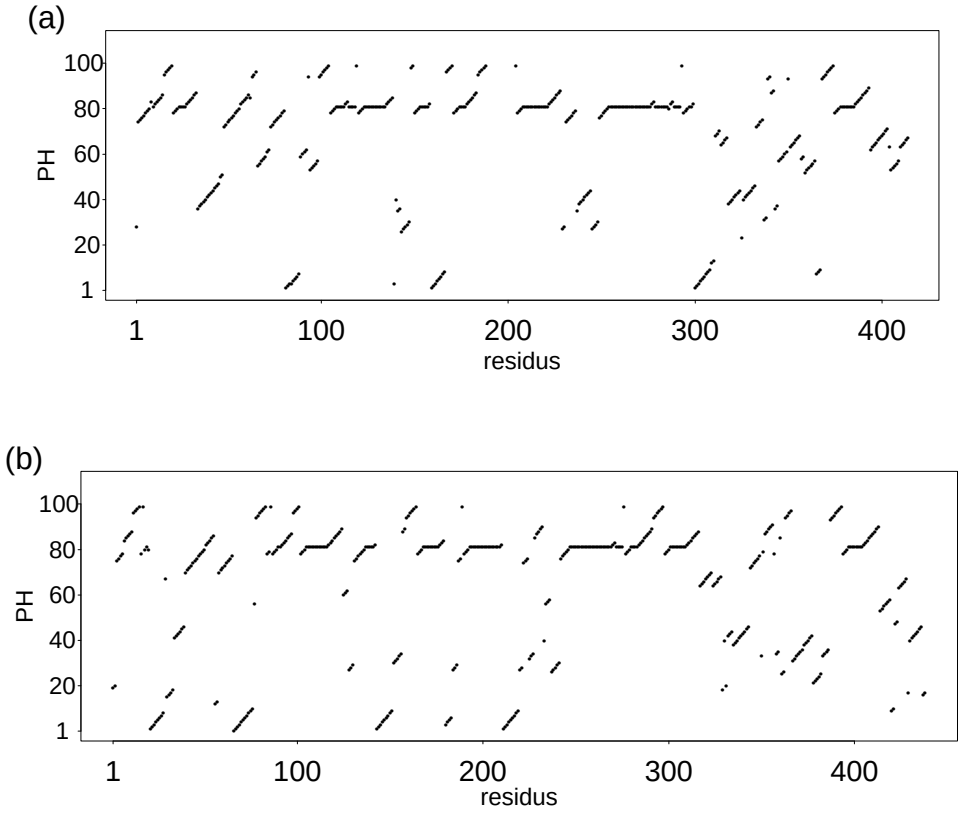

FIG. 6.9 – *Recodage (a) du cytochromes $P450_{BM3}$ et (b) du cytochromes $P450_{terp}$ sur la protéine hybride. Avec en abcisse les positions des résidus des cytochromes et en ordonnée de la protéine hybride.*

Dans un premier temps, les deux structures ont été codées en termes de blocs protéiques. Ensuite, utilisant des successions de 10 blocs, chaque protéine a été codée à partir de la protéine

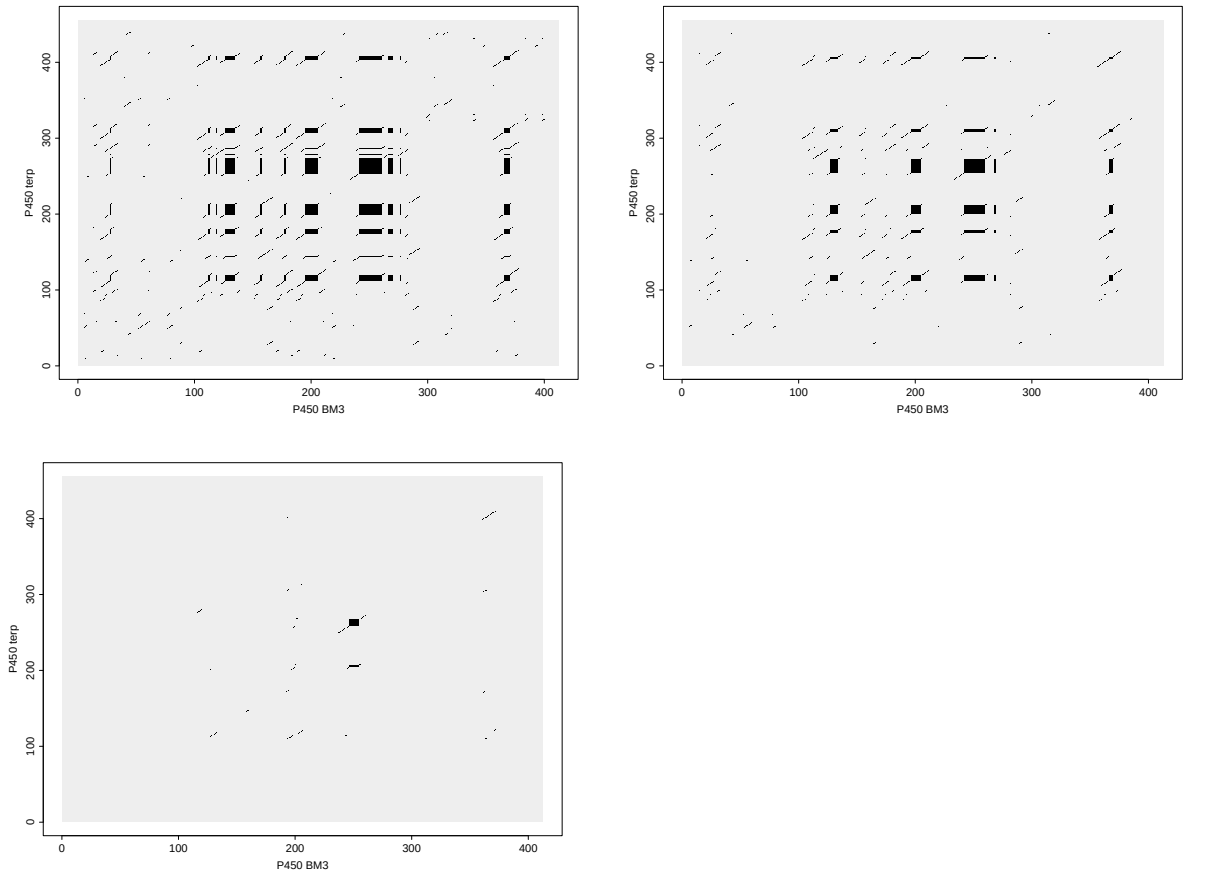

FIG. 6.10 – *Dotplot entre les positions recodées sur la protéine hybride entre $P450_{BM3}$ (en abcisse) et $P450_{terp}$ (en ordonnée) (cf. figure 6.9) avec un filtre (a) de taille 4, (b), de taille 6 et (c) de taille 15.*

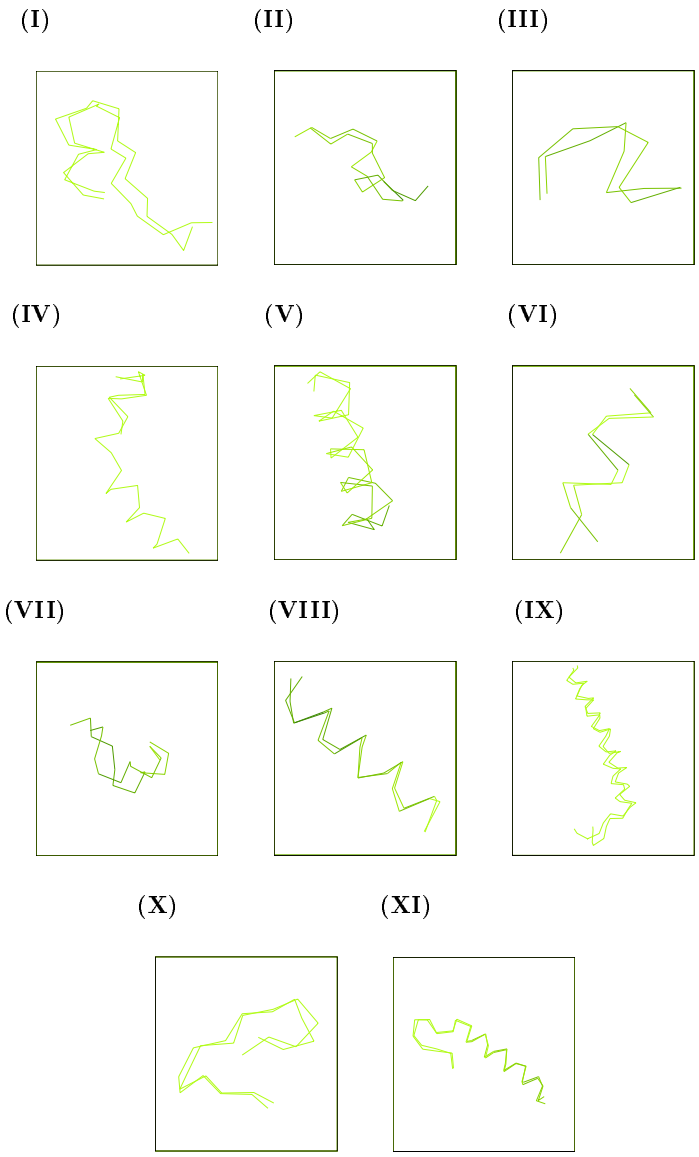

FIG. 6.11 – *Les 11 structures communes trouvées entre les cytochromes* $P450_{BM3}$ *et*$P450_{terp}$.

numéro	P450 $_{terp}$	P450$_{BM3}$	résidus	taux d'identité PBs %	taux d'identité acide aminé %	RMSD Å	CSB numéro	structures secondaires
I	47 - 64	45 - 62	18	88.8	0.0	1.6	1	β_{1-2}, α_B
II	98 - 109	81 - 92	12	58.3	8.0	2.6	/	3_{10} $b^{(a)}$
III	105 - 112	91 - 98	8	75.0	0.0	1.7	2A	α_C
IV	120 - 141	106 - 127	22	95.0	18.0	0.7	2B	$\alpha_{C'}{}^{(b)}$, 3_{10} $b^{(a)}$, 3_{10} $c^{(c)}$, α_D
V	151 - 169	139 - 157	19	89.5	0.0	1.8	3	β_{3-1}, $\pi_{E'}{}^{(b)}$, α_E
VI	175 - 183	178 - 186	9	88.8	0.0	1.1	$4^{(*)}$	3_{10} $c^{(c)}$, α_F
VII	169 - 180	192 - 203	12	75.0	25.0	3.6	$4^{(*)}$	3_{10} $c^{(c)}$, $\alpha_F{}^{(a)}$, $\alpha_G{}^{(a)}$
VIII	209 - 225	201 - 217	17	94.1	6.0	0.7	5	α_G
IX	246 - 285	243 - 282	40	90.0	30.0	1.9	7	β_{5-1}, β_{5-2}, α_I
X	325 - 339	340 - 354	15	73.3	20.0	1.8	$9^{(-)}$+10	β_{2-1},β_{2-2}, β_{1-3}
XI	369 - 396	392 - 419	28	92.8	25.0	0.7	12A +12B	β-bulges, α_L

TAB. 6.2 – *Les 11 structures locales communes avec leurs positions dans les deux cytochromes, leurs longueurs, le taux d'identité en blocs protéiques et en acides aminés, le RMSD entre les deux structures, l'équivalent dans la classification de Jean et collaborateurs [94] et Haseman et collaborateurs [80]. / désigne une structure non trouvée dans les CSBs,* $^{(*)}$ *deux structures locales recouvrent le même CSB 4,* $^{(-)}$ *le fragment* ***X*** *contient le CSB 10 et la fin du CSB 9,* $^{(a)}$ *le label n'est pas à la même place dans les deux cytochromes mais à proximité immédiate (moins de 50 résidus),* $^{(b)}$ *il n'existe pas dans un des deux cytochromes, et* $^{(c)}$ *il n'est pas à la même place dans les deux cytochromes et à une distance importante (plus de 50 résidus).*

hybride en utilisant la même métrique que précédemment. La figure 6.9 donne les positions des deux protéines sur la protéine hybride. Les positions sont le plus souvent continguës, effet de l'apprentissage de la protéine hybride. Un dotplot est ensuite calculé entre les deux protéines recodées par la protéine hybride. Les figures 6.10a à 6.10c montrent les dotplots obtenus pour des longueurs de fenêtres de filtrage G égal à 4, 6 et 15. Le tableau 6.2 récapitule les 11 couples de structures locales notées en chiffres romains avec leurs positions dans les deux cytochromes, leurs longueurs, leurs *RMSd*s respectifs, et, les correspondances de ces fragments par rapport aux travaux précédents de Haseman et collaborateurs [80] et de Jean et collaborateurs [94]. La figure 6.11 montre la superposition de ces 11 structures locales.

6.4.3 Analyse des résultats

Jean *et al.* avaient trouvé 15 CSBs notés de 1 à 13. Les CSB 2A-2B et 12A-12B étaient clairement liés mais n'avaient pas pu être déterminés automatiquement par leur méthode. Avec notre approche, 11 des 15 CSBs ont été retrouvés. Les CSBs 6, 11 et 13 n'ont pas été détectés du fait de leur faible taille (12, 8 et 9 résidus). Utilisant des séries de 10 blocs protéiques de longueur 5 C_α, soit 14 C_α au total, ce type de fragments est considéré comme trop petit. Le CSB 8 est plus long avec ses 17 résidus, mais n'a pas été détecté non plus. Ce dernier point est discuté plus bas. Les structures locales **I**, **IV**, **V**, **VI**,**VII** et **IX** ont la même longueur que leurs CSBs correspondants.

Quelques structures locales ont des différences notables en comparaison des précédents travaux. Ce sont les fragments **II**, **IV**, **V**, **VI** and **VII** par rapport au travail de Haseman et collaborateurs et les fragments **II**, **VII**, **X** et **XI** par rapport au travail de Jean et collaborateurs. Le taux de d'identité des blocs protéiques (cf. tableau 6.2, cinquième colonne) donne une idée de la proximité entre les structures locales. Ils dépassent globalement 73 %, sauf pour la structure locale **II** avec un taux de 58,3 %.

structure locale II: La structure locale **II** n'est pas retrouvée dans les CSBs et possède un *RMSd* important. Toutefois, ce type de structure locale est pratiquement équivalente entre les deux cytochromes, surtout dans la partie N-terminale. Un seul motif caractéristique est retrouvé dans ce fragment, une hélice 3_{10} helix, notée b, trouvée exclusivement dans le cytochrome $P450_{terp}$ [80].

structure locale IV : La structure locale **IV** est composée d'hélices α et 3_{10}, toutefois, la classification en structures secondaires note que l'hélice α C' n'existe pas dans le cytochrome $P450_{BM3}$ et les deux hélices 3_{10} *b* and *c* sont disposées à des positions distinctes. Cette différence de classification n'empêche pas un *RMSd* faible de 0,7 Å.

structure locale V : La structure locale commune **V** a un *RMSd* de 1.8 Å entre les cytochromes 450_{terp} et $P450_{BM3}$. Cette valeur plus importante est due à l'absence d'une hélice π, notée E', dans le cytochrome $P450_{BM3}$. Cette zone se retrouve incluse dans une hélice α, notée E. La différence est ponctuelle, mais fait augmenter le *RMSd*. Toutefois, il faut noter que ces structures locales possèdent le même type de repliement.

structure locale VI : Elle possède une hélice 3_{10}, notée *c*, qui se trouve incluse pour le cytochrome $P450_{BM3}$ dans le fragment **IV**.

structure locale VII : Elle correspond, comme la structure locale **VI**, au CSB 4. Toutefois sa localisation est distincte au niveau du cytochrome $P450_{BM3}$. Le *RMSd* est de 3,6 Å contre 1.1 Å précédemment. Les raisons de cette différence s'explique par le taux plus faible d'identité des blocs protéiques (75,0% contre 88,8% pour le fragment **VI**). Le CSB est composé principalement d'une hélice, notée F [80], qui présente des longueurs fort différentes selon les cytochromes, et surtout, c'est le seul CSB à avoir été sélectionné manuellement par Jean et collaborateurs.

structure locale X : La structure locale **X** inclus le CSB 10 et la fin du CSB 9.

structure locale XI : La structure locale **XI** inclus les CSBs 12A (une poche cysteique [80]) et 12B (une hélice α notée L [80]) qui avaient été décrits séparément, mais sont vraiment liés comme le montre le *RMSd* de 0,7 Å.

Ces résultats démontrent bien l'intérêt de la protéine hybride dans l'extraction de structures locales stucturellement similaires entre deux protéines. Un simple dotplot n'utilisant que les blocs protéiques ne permet pas une extraction aussi simple et rapide, du fait de la répétitivité des structures secondaires. Le bon résultat obtenu est dû au principe de l'apprentissage de la protéine hybride qui apprend des fragments protéiques et regroupent les fragments similaires au même site.

Un point à noter est que les 11 paires de structures locales n'ont jamais un taux d'identité de blocs protéiques de 100 %, mais varient entre 73,3 % et 95,0 % avec l'exception de la paire **II** avec 58.3%. Le taux d'identité de séquence en acide aminés est faible comme attendu (inférieur à 30%). Les différences majeures entre les différentes méthodes concernent les zones les plus

variables comme la structure locale **II**. Les divergences avec la classification des structures secondaires viennent du fait que cette classification place des petites structures un peu partout, sans tenir compte de la modularité des cytochromes. Par exemple, l'hélice 3_{10} *c* notée en position [173-176] pour le cytochrome $P450_{terp}$ est en position [109-114] pour le $P450_{BM3}$. De la même manière la structure locale **V** est composée d'un court feuillet β, une hélice π et d'une hélice α de 12 résidus pour le cytochrome $P450_{BM3}$. Pour le cytochrome $P450_{terp}$, le même feuillet β est présent, mais l'hélice π est alors incluse dans une hélice α de 12 résidus. Le CSB 8 n'a pas été trouvé alors qu'il est d'une taille conséquente et possède un *RMSd* correct de 1,2 Å. Cet oubli vient de la métrique du filtrage du dotplot qui est en tout ou rien. 3 sites du CSB 8 sur la protéine hybride sont différents entre le cytochrome $P450_{terp}$ et $P450_{BM3}$ et donc n'ont pas été pris en compte. Une amélioration du filtrage serait nécessaire.

6.4.4 Conclusion

Dans un premier temps, nous avons vu que la MPH permet l'obtention d'une protéine hybride qui approxime correctement un des fragments protéiques. A chaque site, un nombre important de structures similaires sont présentes. Ainsi, cette méthode permet de réduire la taille d'une base de données structurales protéiques et donc est utile dans une approche de type enfilage. Certaines améliorations sont possibles. Concernant les rares sites ayant un *RMSd* plus important que la moyenne, le rôle de la taille L des séries de blocs étant prépondérant, sa diminution pourrait améliorer la reconnaissance d'homologie structurale. Une autre optique possible est l'utilisation de fragments de tailles variables, comme l'ont fait Bystrofff et Baker[19], mais l'utilisation finale de la protéine hybride serait plus complexe. L'intérêt principal de cette méthode est le maintien de la continuité entre les fragments protéiques et permet d'observer des structures continues avec des compositions en acides aminés intéressant. L'exemple des deux cytochromes P450 montre un exemple d'application à la modélisation locale par homologie. Cette approche de la protéine hybride peut donc être fort utile dans une méthode de prédiction ou de modélisation moléculaire. De plus, la recherche de zones structuralement homologues est particulièrement simple et rapide car elle ne nécessite pas de méthode d'optimisation pour rechercher les homologies, le codage en blocs protéiques et le positionnement dans la protéine hybride suffisent.

6.5 Application à l'étude des relations séquence-structure

6.5.1 Objectif

Les protéines se replient suivant un nombre limité de conformations [70], toutefois la complexité et le nombre important de paramètres physico-chimiques, cinétiques, dynamiques et stèriques rentrant en jeu rendent la prédiction de la structure tridimensionnelle d'une protéine difficile. L'augmentation des bases de données génomiques rend la compréhension de ce problème encore plus crucial aujourd'hui.

Dans le second chapitre, nous avons utilisé une base de données structurales pour définir un alphabet structural, puis nous avons mis en relation les séquences et les blocs protéiques pour établir une stratégie de prédiction. Nous sommes donc passés de la structure à la séquence pour repasser à la prédiction soit en résumé:

$$\text{Structure} \rightarrow \text{Séquence} \rightarrow \text{Structure}$$

Avec la protéine hybride construit à partir des blocs protéiques (cf. paragraphe 6.3) nous avons caractérisé la "dépendance floue" entre la séquence et la structure locale du squelette protéique, mais d'abord en apprenant sur la structure et en déduisant ensuite les caractéristiques en termes de composition en acides amimés.

L'objectif de cette présente étude est d'analyser les dépendances entre l'information structurale et l'information séquentielle, cette dernière étant moins conservée au cours de l'évolution.

6.5.2 Principe

La protéine hybride présentée dans cette partie vise à apprendre à la fois la séquence et la structure des protéines, et donc permet d'étudier la répartition de l'information structure + séquence. Les données sur la séquence ont été recodées en fonction de critères physico-chimiques (volume, charge, hydrophobicité) et les données structurales avec l'aide des angles dièdres ϕ et ψ. L'analyse de la répartition le long de la protéine hybride en relation avec la nature du bloc structural permet d´affiner le rôle de certains acides aminés dans les structures secondaires et des régions flanquantes. L'étude aboutit ainsi à un concept de modèle flou entre la séquence et la structure, une séquence n'étant pas associée à un seul type locale de repliement, mais à plusieurs, et inversement un repliement local est catégorisé avec des séquences différentes [36].

6.5.3 données structures et données séquences

La base de données est celle préalablement utilisée, composée de 342 protéines ayant moins de 25% d'identités [84, 83]. Les protéines ont été découpées en fragments de $M = 5$ acides aminés consécutifs, ce qui donnent un base de données contenant 86980 fragments. Chaque acide aminé a été recodé selon trois variables :

* l'hydrophobicité suivant l'échelle de Kyte et Doolitle [115].
* le volume de la chaîne latèrale suivant l'échelle de Zamyatin [207].
* la charge avec trois niveaux : -0.5 attribué à K, R et H, +0.5 à D et E, et 0 aux autres acides aminés.

Les deux premières variables ont été normalisées entre -1,0 et +1,0 et sont représentées sur la figure 6.12. Les grandes catégories sont retrouvées (cf. figure 2.2). Les valeurs des trois échelles sont données dans le tableau 6.3

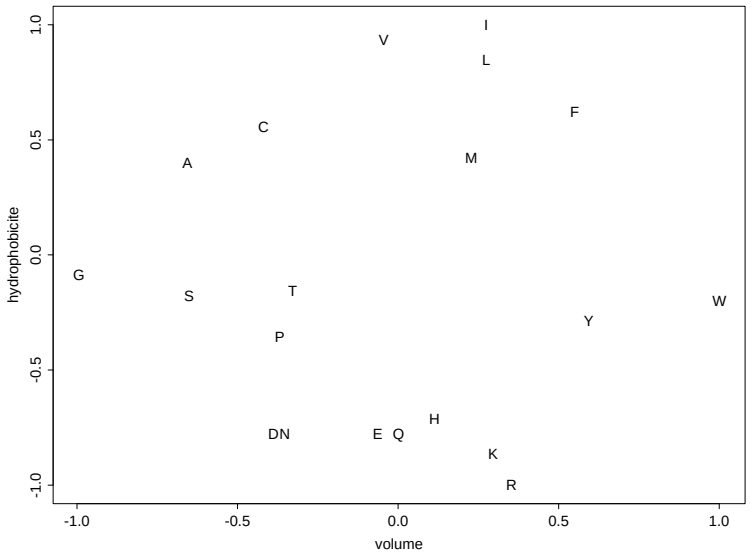

FIG. 6.12 – *Représentation de l'échelle des volumes des acides aminé de Zamyatin [207] en fonction de celle de Kyte et Doolittle [115]. Les deux échelles sont normalisées entre -1,00 et +1,00.*

La structure trimensionnelle de la chaîne carbonée associée à 5 résidus consécutifs (le résidu central étant en position s dans la séquence protéique) est caractérisée par 8 angles dièdres ψ_{s-2}, ϕ_{s-1}, ψ_{s-1}, ϕ_s, ψ_s, ϕ_{s+1}, ψ_{s+1}, ϕ_{s+2}) qui ont été normalisés entre -1,0 et +1,0. Ils ont été préalablement décalés pour les angles ϕ supérieurs à 120° de -360° et pour les angles ϕ inférieurs

acide aminé	hydrophobicité	volume	polarité
A	-0,66	+0,40	0,00
R	+0,36	-1,00	-0,50
D	-0,39	-0,78	+0,50
N	-0,36	-0,78	+0,50
C	-0,42	+0,56	0,00
E	-0,06	-0,78	0,00
Q	0,00	-0,78	0,00
G	-1,00	-0,09	0,00
H	+0,11	-0,71	-0,50
I	+0,27	+1,00	0,00
L	+0,27	+0,84	0,00
K	+0,30	-0,87	-0,50
M	+0,23	+0,42	0,00
F	+0,55	+0,62	0,00
P	-0,37	-0,36	0,00
S	-0,65	-0,18	0,00
T	-0,33	-0,16	0,00
W	+1,00	-0,20	0,00
Y	+0,59	-0,29	0,00
V	-0,04	+0,93	0,00

TAB. 6.3 – *Ensemble des variables avec les échelles normalisées de l'hydrophobicité [115], du volume [207] et de la polarité.*

à -120^o de +360^o. Chaque fragment de 5 résidus est donc défini par un vecteur $\mathbf{V}$ ayant $m = 23$ composantes (15 pour la séquence et 8 pour la structure), toutes comprises dans l'intervalle [-1,0:+1,0].

6.5.4 Apprentissage de la protéine hybride

Dans notre étude, la protéine hybride correspond à une succession de L fragments d'une longueur $M = 5$ résidus, chacun caractérisé en termes séquence-structure par un vecteur de m composantes (ici, $m = 23$). Elle est donc symbolisée par une matrice de dimension L x m. Le principe de base de MPH est d'apprendre "au mieux" l'ensemble de la base de vecteurs (au nombre de 86 980) par cet hybride de L vecteurs. L'apprentissage est similaire à celui d'une carte auto-organisée de Kohonen ("Self-Organizing Map" ou SOM [108, 109]). Cependant, dans notre cas, l'apprentissage est monodimensionnel et la diffusion de l'information le long de l'hybride n'est pas réalisée artificiellement. Elle est implicite car plusieurs vecteurs successifs sont utilisés à la même étape de l'apprentissage. Ce point est distinct de celui de la précédente protéine

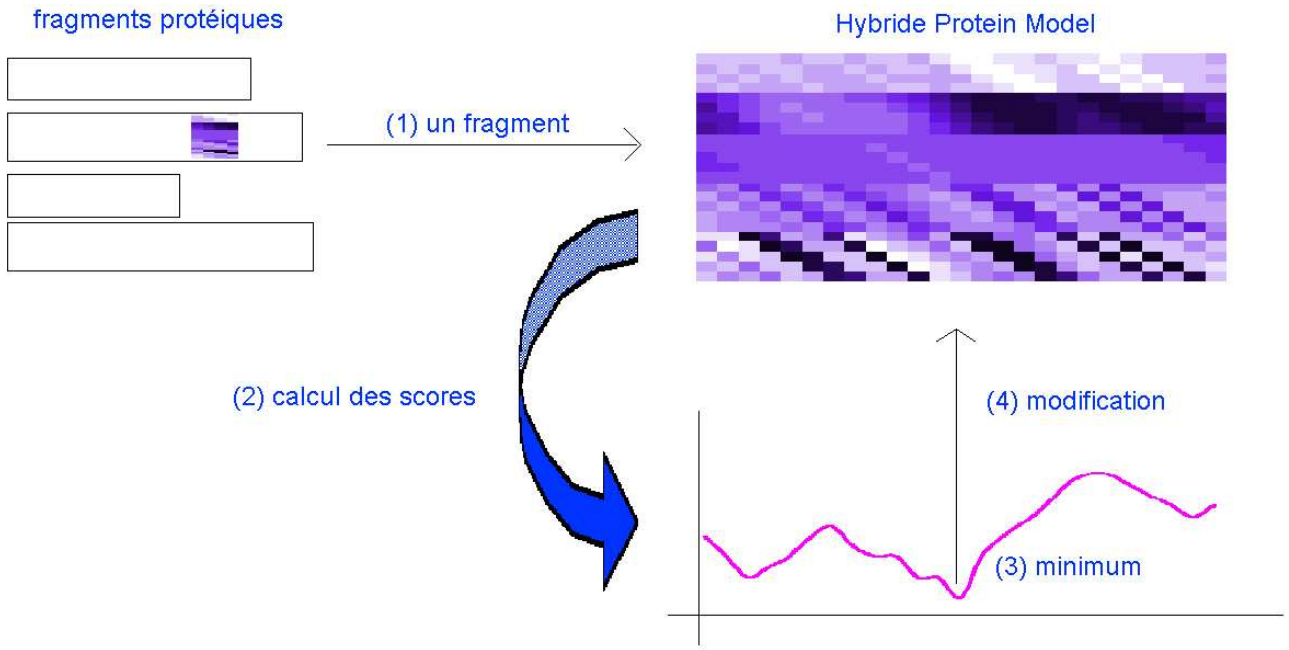

FIG. 6.13 – *Schéma représentant le principe d'apprentissage des informations séquence-structure par la protéine hybride. (1) Sélection d'un fragment défini par une sous-matrice d'observations. (2) Calcul d'un score local entre ce fragment et une région de l'hybride de même taille. (3) Détermination de la position optimale dans l'hybride en recherchant le score minimal. (4) Modification de l'information locale dans l'hybride.*

hybride sur les séries de blocs protéiques où la continuité était comprise directement dans la série de blocs protéiques. Un vecteur d'observation était présenté avec un seul type de blocs pour chaque position alors qu'ici une sous-matrice de f vecteurs est utilisée ($f = 5$ dans cette étude). En effet, nous présentons f vecteurs consécutifs à l'hybride pour effectuer un apprentissage en continu à la fois de la séquence et de la structure. La méthode schématisée dans la figure 6.13 se décompose en trois étapes :

(i) Initialisation de la protéine hybride : on effectue un tirage au hasard de L vecteurs dans les protéines codées.

(ii) Apprentissage séquentiel des matrices d'observations :

(1) on tire un fragment avec son environnement de taille f de la base de données. Il est défini par une sous-matrice $\boldsymbol{V}$ de f vecteurs de taille m.

(2) Pour chacune position p de l'hybride un score de dissemblance $\boldsymbol{S}(p)$ (une distance euclidienne) est calculé entre la sous-matrice $\boldsymbol{V}$ et celle $\boldsymbol{W}(p)$ de même taille prise dans l'hybride. Ainsi un profil de scores est établi le long de l'hybride. (3) On recherche alors le score minimum S_{min} et la position $p^* = argmin\{\boldsymbol{S}(p)\}$ associé à la plus forte ressemblance entre le fragment observé dans une protéine et celui dans la protéine hybride. (4) Ayant localisé le fragment, on va donc modifier légèrement le contenu de la sous-matrice $\boldsymbol{W}(p)$ pour qu'elle ressemble davantage à celle présentée, soit $\mathbf{V}$. La transformation est définie par l'équation :

$$\boldsymbol{W}(p) \rightarrow \boldsymbol{W}(p) + (\boldsymbol{V} - \boldsymbol{W}(p)).\alpha(n)$$

et

$$\alpha(n) = \alpha_0/(1 + n/N)$$

n désignant le nombre de sous-matrices présentées à l'hybride, N le nombre total de sous-matrices de la base de données et α_0 le coefficient d'apprentissage initial. Le coefficient d'apprentissage α(n) décroit au cours de l'apprentissage. Ayant modifié l'hybride, on passe au fragment suivant jusqu'à traiter complétement la base.

(iii) Renforçage de l'apprentissage : on effectue un certain nombre C de cycles d'apprentissage de la base en recommençant l'étape (ii). Cette relecture des informations permet de renforcer l'apprentissage en regroupant progressivement les blocs semblables.

6.5.5 Résultats

La figure 6.14 donne le résultat de l'apprentissage effectué avec un coefficient d'apprentissage $\alpha_0 = 0{,}03$ pour $C = 20$ cycles pour une protéine hybride de longueur $L = 25$. La séquentialité des fragments est visible, toutefois le vecteur caractéristique du fragment en position p ne présente pas exactement le même vecteur décalé du fragment en position *(p - 1)*. D'après les variations en niveau de gris, les variables hydrophobicité et angle dièdre ϕ semblent jouer un rôle important, et à un moindre niveau, le volume et l'angle dièdre ϕ. La charge joue un rôle mineur, cela peut s'expliquer par le nombre faible de chargés par rapport au nombre total de résidus ou par un problème d'étalonnage des variables. Les 25 positions présentent globalement des caractéristiques différentes. Cependant une classification des 25 positions par une classification hiérarchique (paquage *hclust* du logiciel *S-Plus*) montre la constitution de deux groupes homogènes distincts ayant comme frontières les deuxième et treizième positions. Après apprentissage, le nombre de fois où une position de l'hybride est sélectionnée a été calculée. La répartition des observations est relativement homogéne le long de l'hybride, les nombres variant entre 2950 et 4500 observations.

6.5.6 Correspondance entre la protéine hybride et les blocs structuraux

La figure 6.15 donne la composition en acides aminés du résidu central des 25 fragments (ce n'est qu'une information partielle) sur sa partie gauche, et les fréquences relatives des fragments pour chaque bloc structural. Afin de simplifier cette figure, seuls les groupes dont le nombre de fragments attribués était supérieur à 100 et dont la fréquence dans le bloc protéique était supérieure à 4 % (i.e. 1/L) ont été mis. Il s'avère que seulement 15,6 % des fragments de la base n'ont pu être attribués. Les constations déduites de l'étude des figures sont les suivantes :

(i) une dépendance forte entre les fragments de l'hybride et les blocs structuraux; des positions 2 à 13, les blocs qui concernent les hélices α et leurs régions flanquantes sont

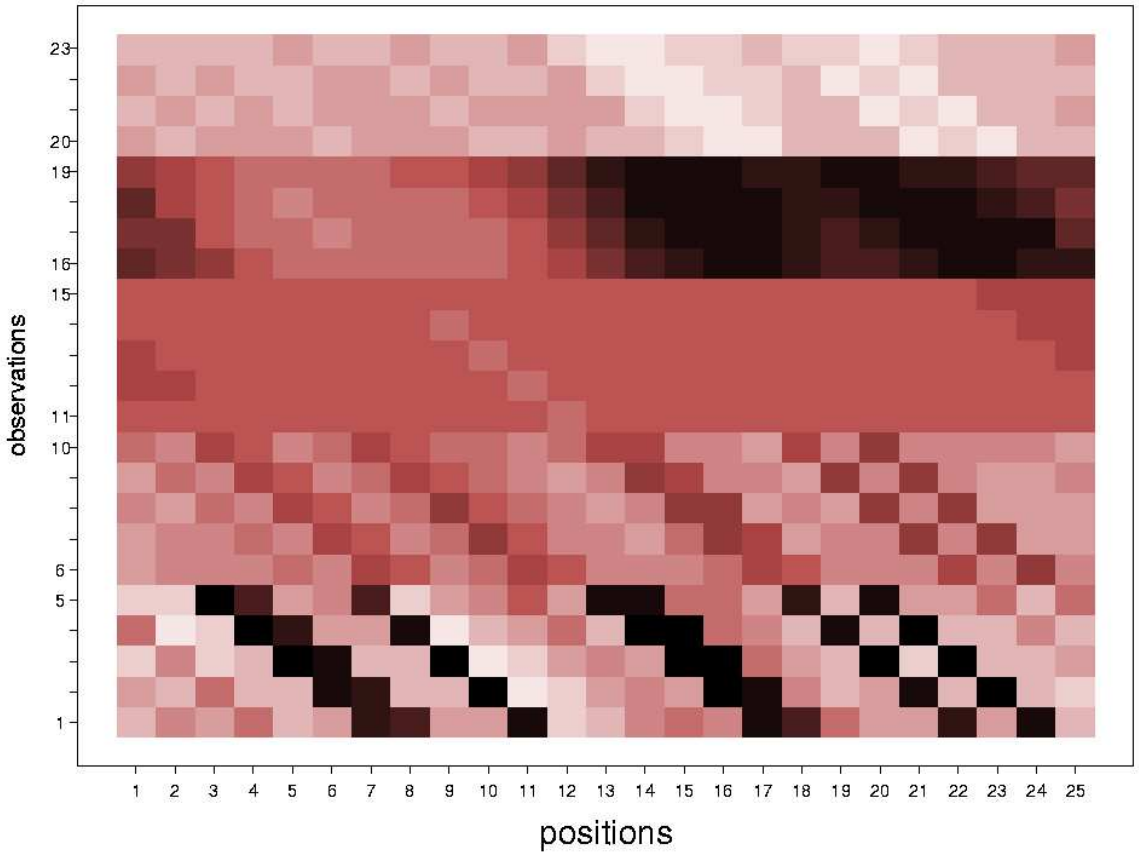

FIG. 6.14 – *La protéine hybride après apprentissage. La protéine hybride est composée de 25 fragments de 5 résidus, l'ensemble des vecteurs de 23 observations ont été déterminés par l'apprentissage. En ordonnée, pour les 5 acides aminés, sont présents de bas en haut [1:5] l'hydrophobicité, [6:10] le volume, [11:15] la charge, les angles dièdres [16:19]* ψ *et [20:23]* ϕ*.*

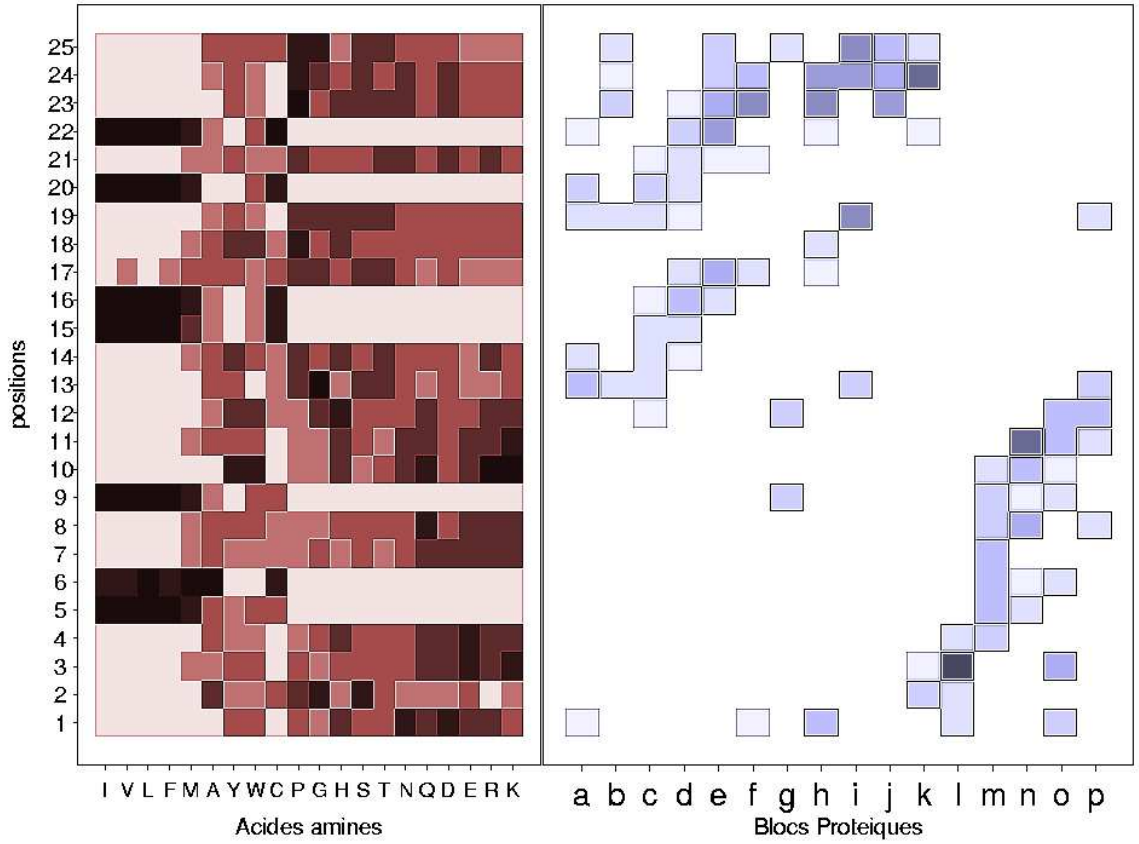

FIG. 6.15 – *Correspondances entre la protéine hybride et les blocs structuraux. (a) Figure de gauche : Répartition des acides aminés le long de la protéine hybride uniquement pour l'élément central des fragments. (b) Figure de droite : Répartition des fragments dans les blocs structuraux.*

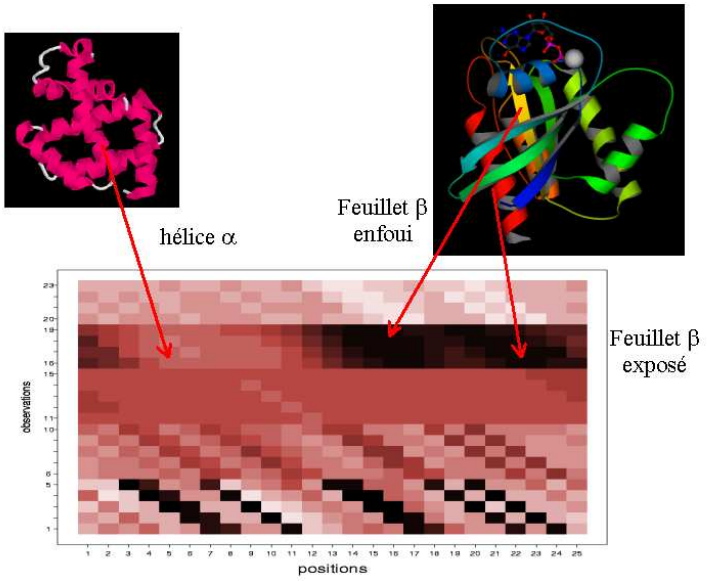

FIG. 6.16 – *Positionement des hélices et des feuillets dans la protéines hybrides.*

retrouvées, et de 13 à 25 (cf. figure 6.16), les feuillets β, leurs régions flanquantes et les boucles. Ces dernières sont localisées dans les positions 1, 12-13, 17-19 et principalement 22-25. On note des sur-représentations des Glycines et des Prolines dans ces zones en association avec les blocs protéiques h, i et j.

(ii) la séquentialité des blocs structuraux et des fragments dans l'hybride se retrouvent dans la figure de droite. On observe deux tendances dans les feuillets β, l'une aux positions de 13 à 19, et l'autre de 20 à 25. Le BPd contient deux hydrophobes consécutifs dans le premier type qui peut correspondre à une hélice enfouie, et deux hydrophobes séparés par un hydrophile dans le second type qui correspond à un feuillet dont l'un des brins est enfoui et l'autre exposé (cf. figure 6.16).

(iii) certains blocs protéiques se retrouvent majoritairement en certaines positions de l'hybride. Par exemple, BP l (extrémité N-terminale d'une hélice α) est localisé en positions [1:4] où apparaissent des résidus chargés, BPm (région centrale d'une hélice α) en positions [4:11] dans lesquelles des résidus hydrophobes (Isoleucine, Valine, Leucine, Phénilalanine, Méthionine et partiellement Alanine) et où un motif hydrophobe est retrouvé (i, i+1, i+4), une partie de l'hélice est enfouie.

(iv) les positions centrales (5, 6, 15, 16, 20 et 22) présentent un caractère hydrophobe très marqué. La Cystéine occupe aussi les mêmes positions, cependant à un degré moindre

dans d'autres positions,

(v) le rôle des résidus chargés est moins précis alors qu'ils devraient intervenir dans les régions flanquantes des hélices et des feuillets.

6.5.7 Conclusion

La répartition des Blocs Protéiques montre ainsi une forme de regroupement que l'on pourrait simplement comparé à l'alphabet structural classique à 3 états (hélices, feuillets et boucles). Un simple comptage des 3 types de structures donne une information évidemment corrélée à celle obtenue dans l'hybride. Toutefois, cette information ne tient aucun compte de l'aspect séquentiel. La correspondance avec l'alphabet structural à 16 états permet au niveau de la structure d'avoir une information bien plus précise. On peut ainsi voir la spécificité de l'entrée dans l'hélice α (autour de la position 3 principalement) sur le plan de la composition en acides aminés, de même en sortie d'hélice (positions 8-11). L'hybride permet de tenir compte de l'hétérogénéité de longueur inhérente aux structures répétitives. De plus l'utilisation des structures secondaires se heurte à l'établissement toujours arbitraire de règle d'attribution à l'une des 3 classes [31, 33]. Avec les feuillets β, l'exemple est encore plus frappant, l'apprentissage ayant permis d'obtenir deux types distincts de feuillets (positions 14-17 et positions 19-23), de longueurs et surtout de compositions fort distinctes.

La protéine hybride permet, en apprenant structure et séquence de concert, d'avoir une compression de ces données en un nombre fini d'états, où les deux types d'information sont combinés. L'utilisation des 16 BPs, outil intéressant pour analyser les structures, permet de distinguer de facon fine les divers états de la structure. En conclusion, la correspondance entre la succession des fragments de 5 résidus dans la protéine hybride et les différents types de Blocs Protéiques d'autre part a permis de mettre en relief le concept de relation séquence-structure "floue". C'est à dire qu'une chaîne d'acides aminés est associée à une distribution de patterns structuraux (i.e. les BPs) et inversement. Cela implique que sur le plan d'une prédiction de la séquence vers la structure, certains blocs protéiques doivent être considérés comme équivalents.

Un tel concept devrait être pris en compte dans la construction de modèles structuraux protéiques que ce soit par une stratégie d'enfilage (threading) ou une modèlisation *ab initio*. Cette méthode peut aussi de servir de validation lors de l'élaboration de modèles structuraux

complets, à la manière de ProCheck avec la carte de Ramachandran. La figure 6.15 de gauche pourrait aussi servir à l'analyse des relations entre séquence et structure à l'intérieur de familles de protéines. Toutefois, des améliorations restent à effectuer, comme par exemple, changer d'échelles pour décrire les acides aminés.

6.6 Comparaison entre les cartes de Kohonen et MPH

Juste avant de conclure ce dernier chapitre, je désire récapituler les différences et les ressemblances méthodologiques existantes dans la méthode que nous avons mis au point et les cartes de Kohonen.

Les cartes de Kohonen sont des outils puissants d'analyse (cf. Annexe 1), elles sont donc le plus souvent bidimensionnelles, MPH est lui unidimensionel. L'apprentissage s'effectue dans les deux cas par un processus itératif, avec un coefficient d'apprentissage faible qui permet une modification locale limitée décroissant avec le nombre d'observation présenté.

La plus grande distinction entre les deux méthodes concerne la diffusion de l'information. Dans une carte de Kohonen, une diffusion contrôlée sur les neurones voisins existe. Ainsi, quand un neurone U_{i-x} est modifié de α %, les neurones contigus U_{i-x} et U_{i+x} seront modifiés de α/τ % ($\tau > 1$). Cette diffusion de l'information permet de rassembler dans une zone des neurones proches. Pour MPH, aucune diffusion n'est effectivement effectuée, elle est remplacée par un processus d'empilage. Les neurones U_{i-x} et U_{i+x} vont apprendre une partie de l'information, un décalage s'effectue naturellement, les observations prises avec leur environnement et les neurones contigus au vainqueur vont apprendre cet environnement. Ainsi, une continuité s'intaure entre les neurones.

Les neurones obtenus ne sont donc pas indépendant comme dans les SOMs mais "séquentiels", un neurone U_i est conditonné par ses neurones voisins U_{i-x} et U_{i+x}.

6.7 Conclusion

La méthode de la protéine hybride (MPH) est un modèle flou de compaction de la structure locale des protéines. La protéine hybide lors de la première étude a permis l'apprentissage d'une base de données structurales recodées en blocs protéiques. Elle est donc composée d'une série

de lois de probabilité d'observation des blocs. L'apprentisage est un processus qui consiste pour chaque série de blocs à trouver un site qui lui est très proche et à le modifier faiblement. Ce processus nous a permis de cosntruire une protéine hybride qui possède en chaque position une série de blocs bien déterminée, et, ces séries quoique parfois différentes possèdent entre elles une déviation quadratique moyenne trés acceptable de 3,14 Å pour des longueurs de 10 C_α. En plus de l'analyse des structures et des acides aminés qui leurs sont associés, l'utilisation de la protéine hybride dans la recherche de zone d'homologie structurale entre deux cytochromes P450 montre une efficacité évidente. L'utilisation conjointe de cette information structurale avec une recherche sur la séquence dans une méthode de modélisation par homologie devrait-être envisagée. De même, la compaction de la base de données en une centaine de prototypes peut permettre une diminution du nombre de candidats à tester dans une approche de type enfilage. Enfin, MPH a montré ses capacités dans l'établissement des liens entre la séquence et la structure. Cette dernière approche peut servir à de nombreuses méthodes comme validation de la compatibilité entre un modèle construit et la séquence, ou encore dans l'analyse d'incompatibilité de repliements dans une méthode de repliements *ab initio*.