Chapitre 4

Prédiction de la structure locale en blocs protéiques

4.1 Objectif

L' alphabet structural défini permet d'analyser de manière fine la structure tridimensionnelle des protéines (cf. paragraphe 3, *"Apprentissage de la structure locale du squelette protéique"*). La répartition en acides aminés dans chaque bloc montre une forte spécificité (cf. paragraphe 3.3.4); cette information paraissant pertinente, nous allons l'utiliser directement dans une méthode de prédiction de la structure locale à partir de la séquence.

La prédiction a été effectuée avec une méthode bayésienne proche de celle utilisée pour la prédiction des structures secondaires [193], de l'accessibilité au solvant [192] et de modèles plus théoriques d'analyse des modes biologiques [117].

Dans un second temps pour améliorer le taux de prédiction nous avons pris en compte le fait que différentes séquences peuvent être liées à un même type de repliement (*1 bloc protéique* $\rightarrow$ *n séquences*).

Enfin deux stratégies ont été développées pour proposer en fonction d'un taux de prédiction moyen un certain nombre de blocs potentiels. En effet, un type de séquence n'est pas toujours associé au même type de repliement (*1 séquence* $\rightarrow$ *n blocs protéiques*). Ces derniers travaux se basent principalement sur un indice entropique tenant compte de la qualité de la prédiction locale.

Il convient de noter que pour définir la méthode de prédiction, nous avons décidé d'utiliser une méthode statistique simple qui permet de comprendre quels acides aminés sont importants pour chaque type de structures. Il est certain que les réseau neuronaux donnent des résultats

meilleurs, mais en contrepartie, il est peu aisé de comprendre comment ils ont "appris". Aussi, une approche de type bayésienne est particulièrement apropriée.

4.2 Prédiction bayésienne simple

4.2.1 Méthodes

Pour chaque site s d'une protéine, qui comprend aussi bien la position centrale que l'ensemble de la fenêtre de la séquence [-w;+w] autour de cette position centrale, nous avons calculé pour une séquence d'acides aminés X_S, la probabilité d'observer cette séquence dans un bloc donné PB_k, notée $P(PB_k/X_S)$.

De cette probabilité conditionnelle préalablement définie, il est possible de calculer la probabilité d'avoir ce bloc connaissant la séquence, en utilisant le théorème de Bayes. Il accomplit, en effet, l'inversion de la séquence X_S et de la structure PB_k:

$$P\left(PB_k/X_S\right) = \frac{P\left(X_S/PB_k\right).P(PB_k)}{P\left(X_S\right)}$$

avec $P(PB_k)$, la probabilité d'observer PB_k dans la base de données et $P(X_S)$, la probabilité d'observer la séquence d'acides aminés X_S sans aucune information sur la structure. Cette dernière est égale au produit des fréquences des acides aminés dans la base de données. Une approche assez similaire a été utilisée par Thompson et Goldstein [193] pour la prédiction des structures secondaires.

Le terme $P(X_S/PB_k)$ est la probabilité conditionnelle d'observer une séquence donnée X_S (a_{-w},...,a_{+w}) pour un bloc PB_k. Il est calculé comme le produit des probabilités pour chaque acide aminé en position j dans la séquence dans le bloc (cf. figure 4.2). Ce qui amène à l'équation:

$$P(X_S/PB_k) = \prod_{j=-w}^{j=+w} P(a_j/PB_k)$$

Pour définir le bloc optimal PB^* pour une série d'acides aminés X_S en un site s d'une protéine, nous utilisons le ratio R_k (ou son logarithme) défini par:

$$R_k = \frac{P\left(PB_k/X_S\right)}{P(PB_k)} = \frac{P\left(X_S/PB_k\right)}{P(X_S)}$$

Du théorème de Bayes, R_k est défini par le ratio $P(X_S/PB_k)/P(\mathrm{X}_S)$ qui est calculé à partir des matrices d'occurrences. Grâce à ce ratio, la probabilité d'observer un bloc donné PB_k sachant la séquence X_S est comparé à la probabilité d'observer PB_k sans avoir d'information sur la séquence. Ainsi, quand $ln(R_k)$ est positif, la connaissance de la séquence X_S est favorisée par les occurrences de PB_k, et inversement quand il est négatif.

La règle pour définir le bloc optimal PB^* pour la séquence X_S revient à sélectionner, parmi les B blocs, le bloc PB pour lequel ce ratio R_k est maximum. Par conséquence, une liste des B blocs protéiques est définie selon leurs valeurs décroissantes de R_k, le bloc optimal étant le premier. Ainsi, nous pouvons calculer le pourcentage de bonne prédiction Q(1) au premier rang et and Q(r) quand le bloc réel est parmi les r premières solutions.

4.2.2 Résultats

4.2.2.1 Choix du nombre de blocs

La prédiction bayésienne a été effectuée pour chaque série de blocs obtenue, allant de $B =$ 34 blocs, à 22, puis 19, 16, 14, 12, 11 et enfin 10 blocs. La taille de la fenêtre de prédiction a été prise égale à 15 résidus, soit 5 de part et d'autre du bloc structural. La figure 4.3 récapitule les résultats obtenus pour la série de 16 blocs avec des tailles de fenêtre allant de 5 à 19 résidus. A partir de 15 résidus, il y a une saturation dans le gain du taux de prédiction, des résultats similaires ont été obtenus pour les autres séries. Comme attendu, plus le nombre B de blocs augmente, plus le taux de prédiction diminue (cf. figure 4.3).

Dans le choix du nombre de blocs conservés, deux séries (11 et 18 PBs) ont été enlevés car ils avaient un taux de prédiction inférieur à des séries ayant plus de blocs protéiques. En observant les différentes séries obtenues (cf. figure 4.4), on s'aperçoit qu'avec peu de bloc ($B = 10$), le taux de prédiction est bon (39 %), mais l'approximation structurale est alors plus faible (*RMSda* moyen de 32^o). Le choix de 16 est le plus approprié car le taux de prédiction est acceptable (34 %), le *RMSda* moyen reste correct (30^o). En outre, le bloc le moins représenté est égal à un pour cent de la base de données. Cette dernière remarque a son importance: pour la série précédente, les blocs les moins observés représentent moins de 0,5 % de la base de données et il est donc difficilement utilisable pour la prédiction (le nombre d'observations étant alors trop faible).

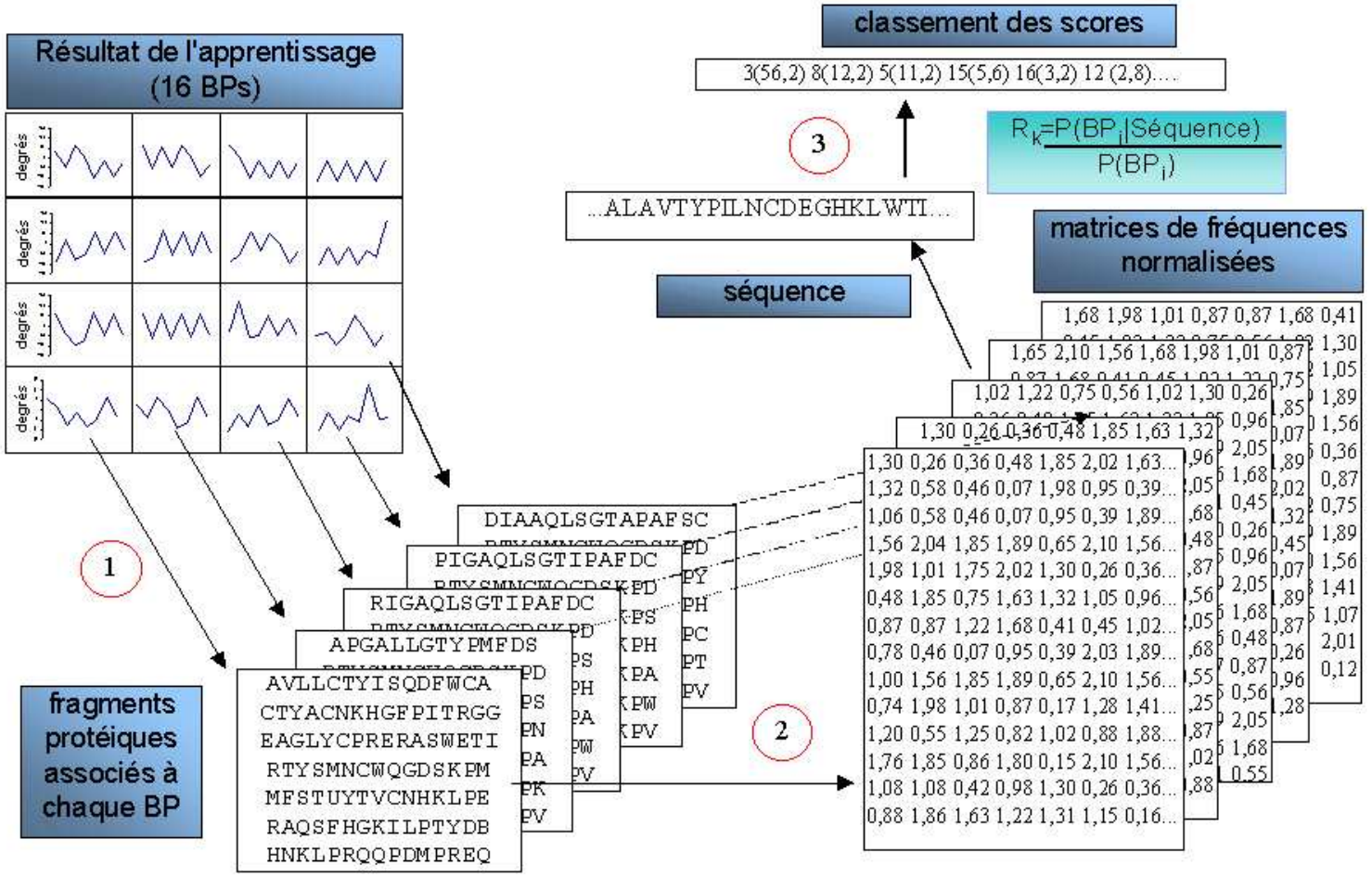

FIG. 4.1 – *Schéma de la prédiction bayésienne: (1) les fragments protéiques sont associés à leur bloc protéique, puis (2) la matrice d'occurrence 15 × 20 est normalisées en fonction de la fréquence de chaque type d'acides aminés, (3) Le score est calculé (cf. figure 4.2), les 16 scores sont classés par ordre décroissant.*

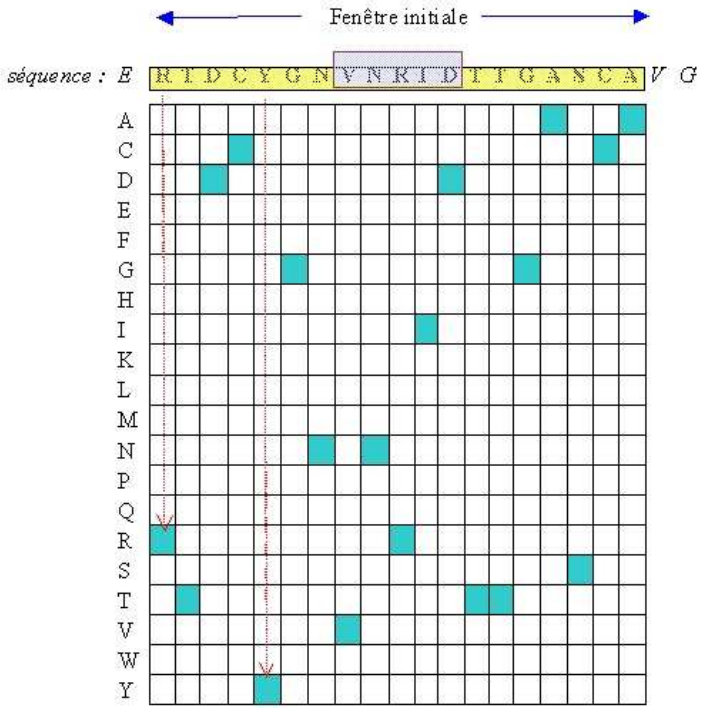

FIG. 4.2 – *Principe du calcul du score. En chaque position de la séquence, le produit des fréquences normalisées est effectué avec chaque matrice d'occurrence normalisée pour chaque BP.*

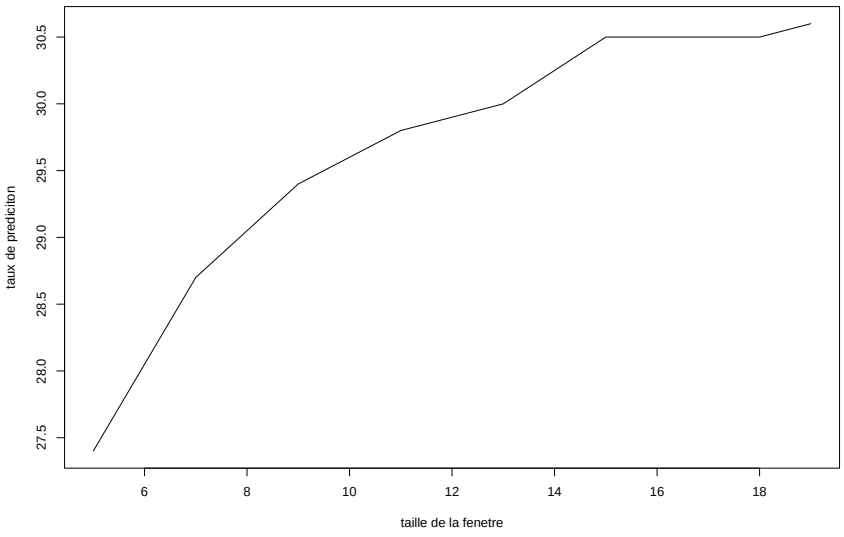

FIG. 4.3 – *Evolution du taux de prédiction en fonction de la longueur de la fenêtre pour la série de 16 blocs protéiques, avec la taille de la fenêtre de prédiction, en abcisses et le pourcentage de prédiction associé, en ordonnés.*

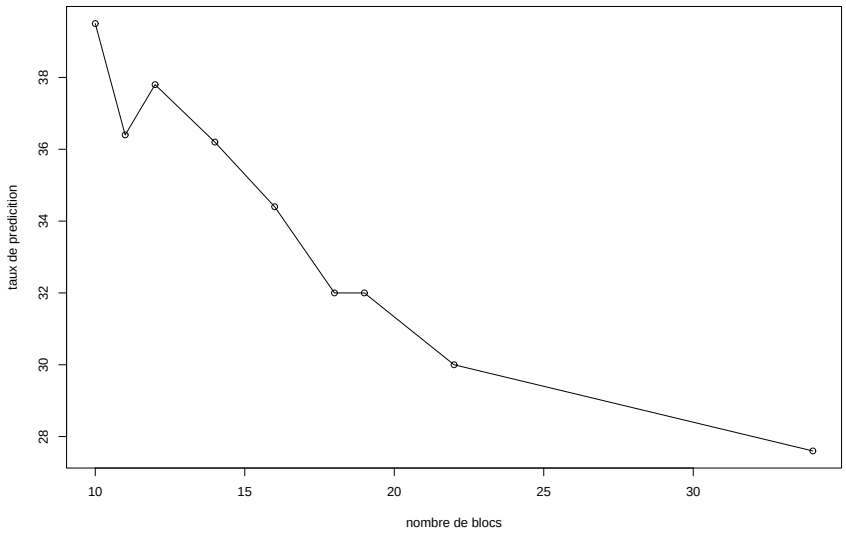

FIG. 4.4 – *Evolution du taux de prédiction en fonction du nombre de blocs, avec le nombre de blocs protéiques dans chaque série, en abcisses, et le pourcentage de prédiction associé, en ordonnés.*

4.2.2.2 La prédiction locale

En utilisant la stratégie bayésienne simple, avec une fenêtre de 5 résidus, correspondant donc au bloc structural, le taux de prédiction est de 30,0 %. Avec une fenêtre totale de 15 résidus, soit 5 de part et d'autre du bloc, le taux de prédiction passe à 34,4%. On peut noter qu'une recherche purement aléatoire donnerait un taux de prédiction de 7 %. L'ensemble des blocs protéiques gagnent en gain de prédiction, par exemple, BPb passe de 11,0 % à 13,5 %, BPe de 33,0 % à 43,2 %,BPi de 32,9 % à 42,2 %, et BPp de 26,9 % à 33,5 %.

Une certaine hétérogénéité est observée dans les taux de prédictions associés à chaque bloc protéique qui va de 13,3 % pour le bloc b à 60,3 % pour le BP a. La figure 4.5 récapitule les différents taux de prédiction par blocs, et, montre aussi le taux de prédiction obtenu quand un certain nombre de solutions sont conservées. En effet, en classant les blocs par score décroissant, le bloc réel se trouve assez réguliérement parmi les blocs les plus probables, ayant donc un score élevé. Ainsi, en conservant les 2 blocs les plus probables, le pourcentage d'avoir le bloc réel passe à 52,1 %, pour 4 solutions conservées, le taux est à 71,4 % et avec 8 solutions conservées sur les 16, soit la moitié, le taux passe à 90,4 %.

Cette distribution montre bien que l'utilisation d'une information séquentielle contient assez d'information pour conditionner la structure locale.

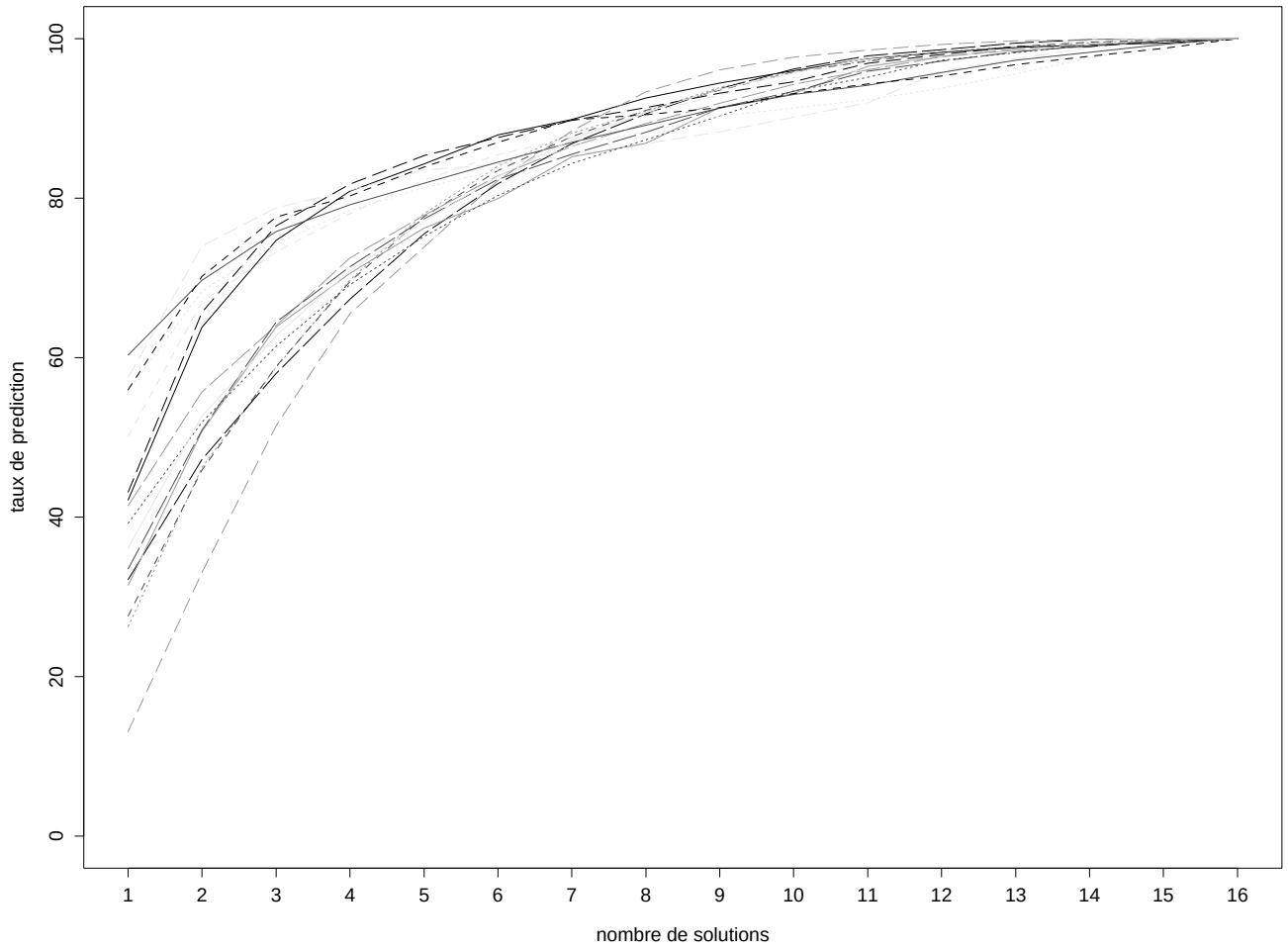

FIG. 4.5 – *Evolution du taux de prédiction pour les 16 blocs protéiques en fonction du nombre de blocs sélectionnés par l'approche bayésienne.*

4.3 Les familles séquentielles

4.3.1 Principe des familles séquentielles

La méthode de prédiction bayésienne implique pour chaque bloc l'utilisation d'une matrice d'occurrence qui lui est propre. Ainsi, si un bloc est composé de deux types de séquences distinctes, l'utilisation d'une matrice commune entraîne un phénomène de moyennage qui fait perdre de l'informativité à ce bloc. Aussi pour améliorer la prédiction, nous avons mis au point la méthode "des familles séquentielles". Elle consite en la génération, pour un bloc protéique x de f matrices distinctes contenant chacune une partie des fragments protéiques du BP x. Pour cela, nous avons procédé à une classification des fragments par une méthode proche des cartes topologiques de Kohonen [109].

Pour un bloc protéique x, f matrices d'occurrence sont créées. Chacune des f matrices est initialisée en mettant les fréquences des acides aminés de la matrice associée au bloc x avec une légère variation pour les individualiser. Toutes les fréquences ont été recalculées pour avoir une fréquence égale à 1 en chaque position.

Ensuite, chaque fragment protéique associé au BP x est alloué à la matrice qui lui ressemble le plus parmi les f matrices. Pour cela, la probabilité conditionnelle $P_l = P(X_S/BP_k^l)$ est calculé pour l allant de 1 à f, avec X_S la séquence en acide aminé correspondant au fragment. Ainsi, le score maximal $P_l^* = \max\{P_l\}$ permet de caractériser la matrice l^* correspondant au mieux au fragment X_S. Cette matrice va être donc légèrement modifiée pour ressembler un peu plus au fragment X_S. Chaque fréquence d'acides aminés f_{aa} en position k est modifiée.

- pour l'acide aminé o présenté à la position k du fragment X_S:

$$f_o^k \leftarrow \frac{f_o^k + \alpha}{1 + \alpha}$$

- pour les 19 autres types d'acides aminés à la même position :

$$f_{aa}^k \leftarrow \frac{f_{aa}^k}{1 + \alpha}$$

Cette transformation permet de conserver en chaque position une somme des fréquences

toujours égale à un. Le coefficient d'apprentissage α est égal à $\alpha_0/(1 + t/N_x)$, avec α_0 le taux initial d'apprentissage pris égal à 0,05, t représentant le nombre de fragments déjà vus et N_k, le nombre total de fragments associés au bloc protéique x. Le processus est itératif, l'ensemble des fragments est donc vu totalement à chaque cycle. Au bout d'un certain nombre de cycles, les fragments se focalisent sur une seule des f matrices. 5 cycles ont été utilisés dans cette apprentissage.

4.3.2 Construction des familles séquentielles

La figure 4.6 illustre le principe de la séparation en deux matrices du bloc protéique b en deux familles séquentielles. On peut observer sur les matrices normalisées en Z-scores une différence de localisation des sur- et des sous-représentations. Ceci se retrouve dans les zones de plus grande informativité obtenues par le profil KLd (cf. figure 4.7). Le Kld maximal est passé de 0,1 à 0,3. En n'observant que les valeurs supérieures à 0,08, on voit que pour la première famille séquentielle la zone d'intérêt se trouve dans l'intervalle -3 à +2, ainsi que pour les positions (-7) et (+4); pour le second, la zone est restreinte à l'intervalle [-2;+2]. Leurs modes aussi sont différents, respectivement en (-1) et (0).

En comparant les deux matrices associées, des différences nettes en composition en acides aminés sont visibles en la première et la seconde famille séquentielle, comme une sur-représentation en Alanine en position (-7), Acide Aspartique (-2), Proline en (-1), Histidine et Aspartate en (0), Proline en (+1) and Phénylalanine en (+6), ainsi que des sous-représentations en Lysine en position (-2), Glycine en (+1) et Cystéine en (+4). Les caractéristiques principales du bloc protéique b sont retrouvées dans ses deux familles séquentielles, comme la sous-représentation en Proline en position (+2).

Des essais pour tout les blocs ont été effectués en prenant un nombre de familles f compris entre 2 et 6. Les blocs divisés en familles séquentielles ont été choisis en prenant comme critère le taux de prédiction globale au premier rang, soit $\mathbf{Q}(1)$. Comme plusieurs matrices pour le même bloc protéique étaient utilisées, seule celle ayant le plus haut score est conservée pour la prédiction bayésienne. Un autre critère a été pris en compte, il s'agissait de rééquilibrer les taux de prédiction entre tous les blocs. Le tableau 4.1 récapitule le nombre de familles séquentielles conservées. Le taux de prédiction $\mathbf{Q}(1)$ est passé de 34,4 % à 40,7 % (gain de 6,3 %), avec les

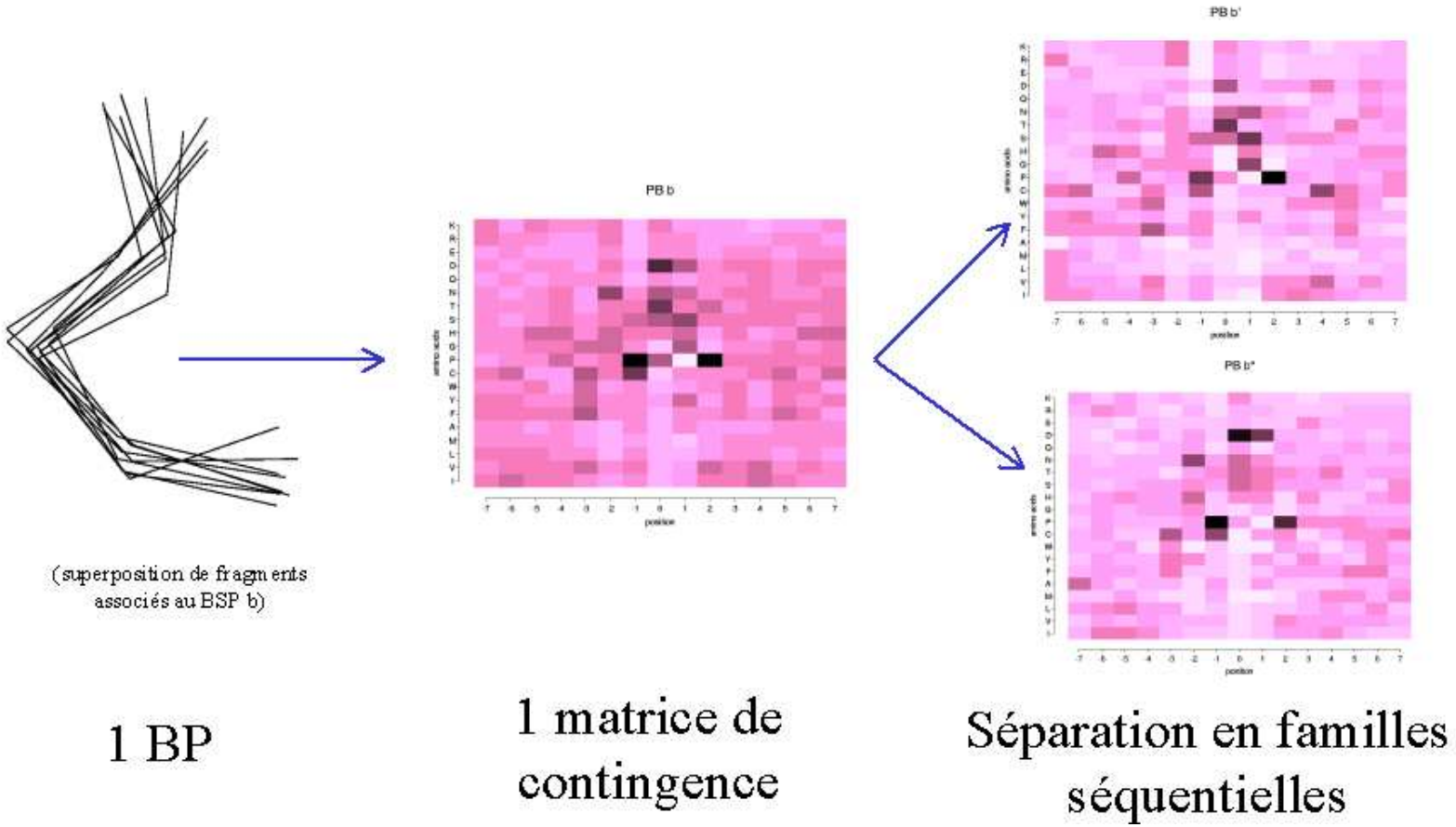

FIG. 4.6 – *Principe du découpage d'une matrice d'occurrence pour un bloc protéique donné en deux matrices d'occurrences associées à des familles séquentielles.*

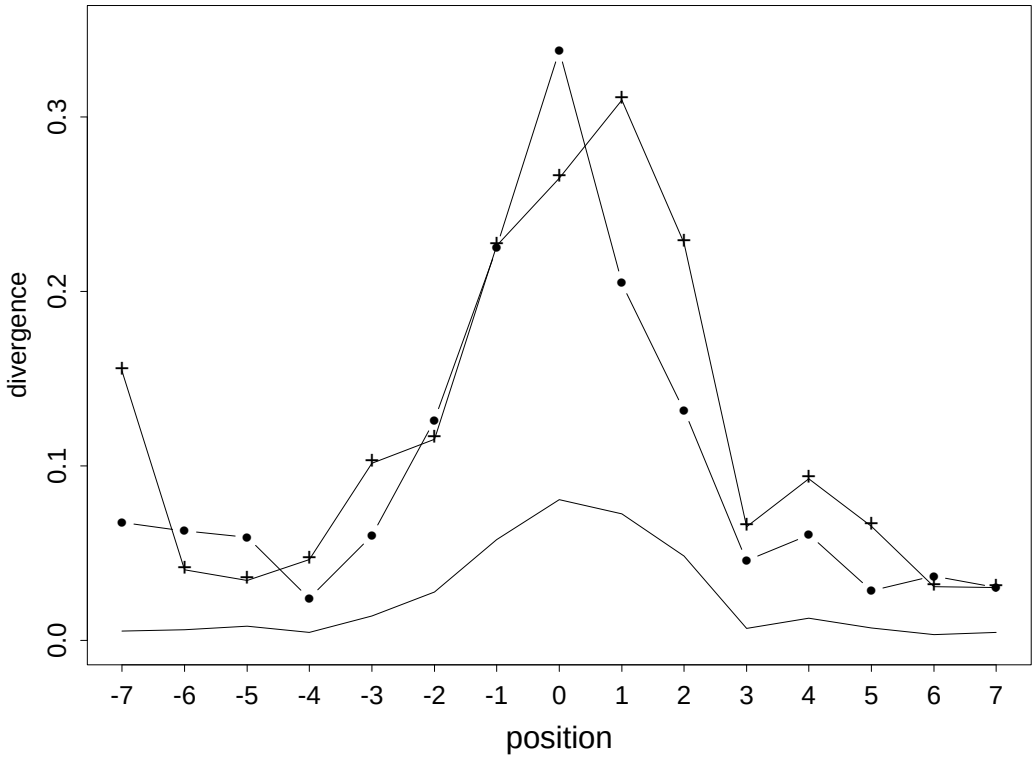

FIG. 4.7 – *Exemple de l'évolution du KLd dans le découpage en famille séquentielle pour le bloc protéique* b*, avec le profil KLd initial en ligne pleine et les profils des Kld des deux matrices issues de ce bloc en ligne avec des points et avec des ronds.*

bloc	nombre de familles séquentielles	taux de prédiction initial	taux de prédiction fam. séq.
a	1	60,3	53,5
b	2	13,1	27,0
c	2	26,3	32,9
d	3	27,6	34,8
e	1	43,1	35,9
f	2	32,2	36,2
g	1	31,4	35,1
h	1	52,1	42,7
i	1	42,1	41,0
j	1	57,5	47,2
k	1	41,4	35,2
l	1	39,2	32,1
m	6	36,1	50,8
n	1	56,0	44,7
o	1	55,8	45,8
p	1	33,5	33,9
Total	26	34,4	40,7

TAB. 4.1 – *Nombre de familles séquentielles par bloc, avec leur taux de prédiction pour l'approche initale et celle utilisant les familles séquentielles (fam. séq.).*

taux de prédiction initial et avec l'utilisation des familles séquentielles.

En résumé, ce sont les blocs les plus fréquents qui ont pu être subdivisés, la fréquence finale de leur famille séquentielle se rapproche d'ailleurs alors de la fréquence des autres blocs protéiques.

Par ailleurs, il a fallu vérifier que la création de ces nouvelles familles séquentielles construites sur le plan de la séquence n'ait pas eu de conséquence sur le plan de la structure. Aussi, pour chaque famille séquentielle, le vecteur d'observation moyen des 8 angles dièdres (cf. paragraphe 3.2) le caractérisant a été calculé à l'aide des fragments appartenant à chaque nouvelle matrice. Ces vecteurs ont été comparés au vecteur décrivant le bloc dont ils sont issus (cf. tableau 3.1) ainsi qu'à celui des autres familles séquentielles du même bloc. Un seul angle se trouve à plus de 3 degrés de différences. En conclusion, les familles séquentielles n'ont pas créé de "nouveau" bloc protéique.

4.3.3 Influence des familles séquentielles dans la prédiction

Avec les 26 matrices obtenues par l'utilisation des familles séquentielles, le taux de pédiction $\mathbf{Q}(1)^*$ est passé à 40,7 %, en conservant les deux solutions les plus probables ($\mathbf{Q}(2)^*$), le taux est de 57,5 %, soit 5,4 % de gain par rapport à l'approche Bayésienne simple, de même $\mathbf{Q}(4)^*$

= 75,8 % (gain de 4,4 %) et atteint 90,2 % pour $\mathbf{Q}(7)^*$, soit un gain d'un rang pour le même taux de probabilité.

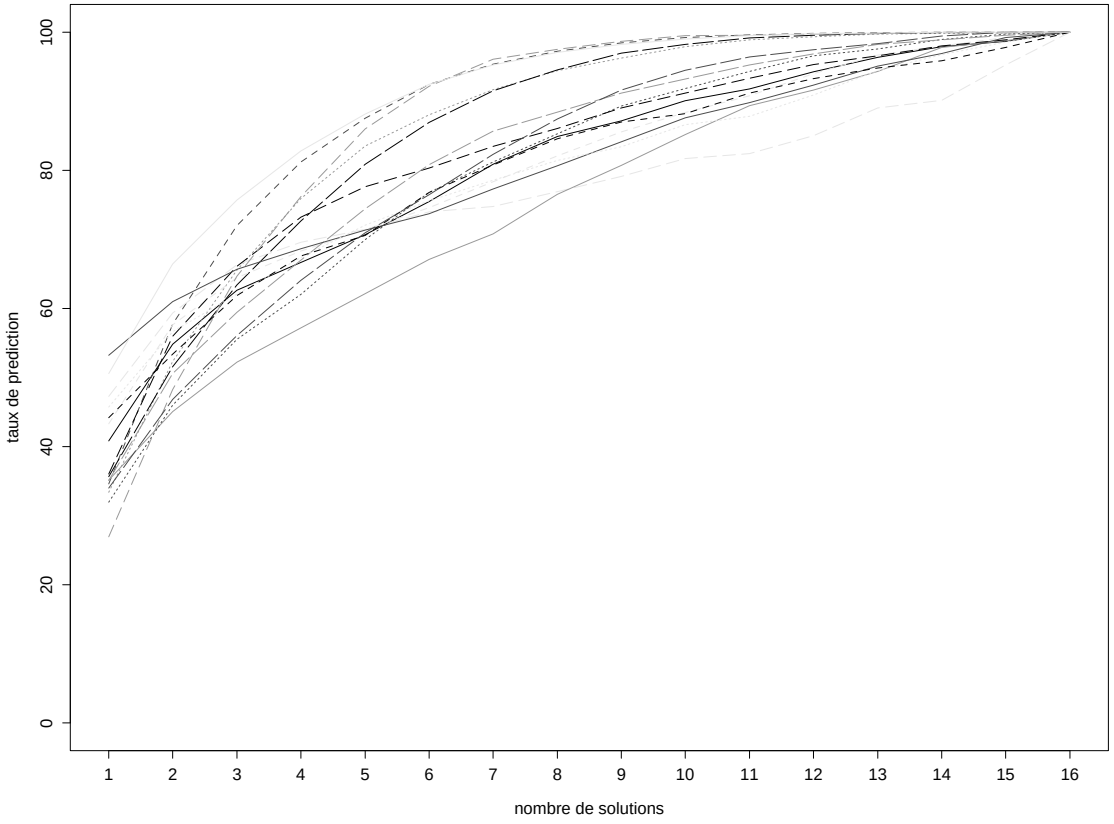

FIG. 4.8 – *Taux de prédiction pour chaque type de blocs protéiques en conservant de 1 à 16 solutions possibles avec utilisation des familles séquentielles*

La figure 4.8 montre, comme la figure 4.5 pour la prédiction bayésienne simple, l'évolution du taux de prédiction individuel des BPs en fonction du nombre de solutions conservées. L'effet de concentration des blocs dans un intervalle plus restreint est obtenu. L'écart entre le BP le meilleur et le plus mal prédit est passé de 47,2 % à 26,2 % avec une augmentation du taux de prédiction du BPb de 13,1 % à 27,0 % et une diminution de celui du BPa de 60,3 % à 53,2 %, ce dernier n'ayant pas été divisé.

La figure 4.9 montre la différence qui existe entre le taux de prédiction initial $\mathbf{Q}(1)$ indiqué en abscisse et la différence entre ce taux et $\mathbf{Q}(1)^*$ obtenu par les familles. Cette figure montre bien que le gain concerne la majorité des protéines, 95 % ont gagné en taux de prédiction. Maintenant 51,4 % des protéines ont un taux de prédiction supérieur à 40 % contre moins de 21 % auparavant. Ce gain n'est pas équivalent selon le type de protéine ainsi en moyenne les protéines tout-α ont un gain de 9,1 % (37,3% contre 46,4 %), les tout-β ont un gain plus faible

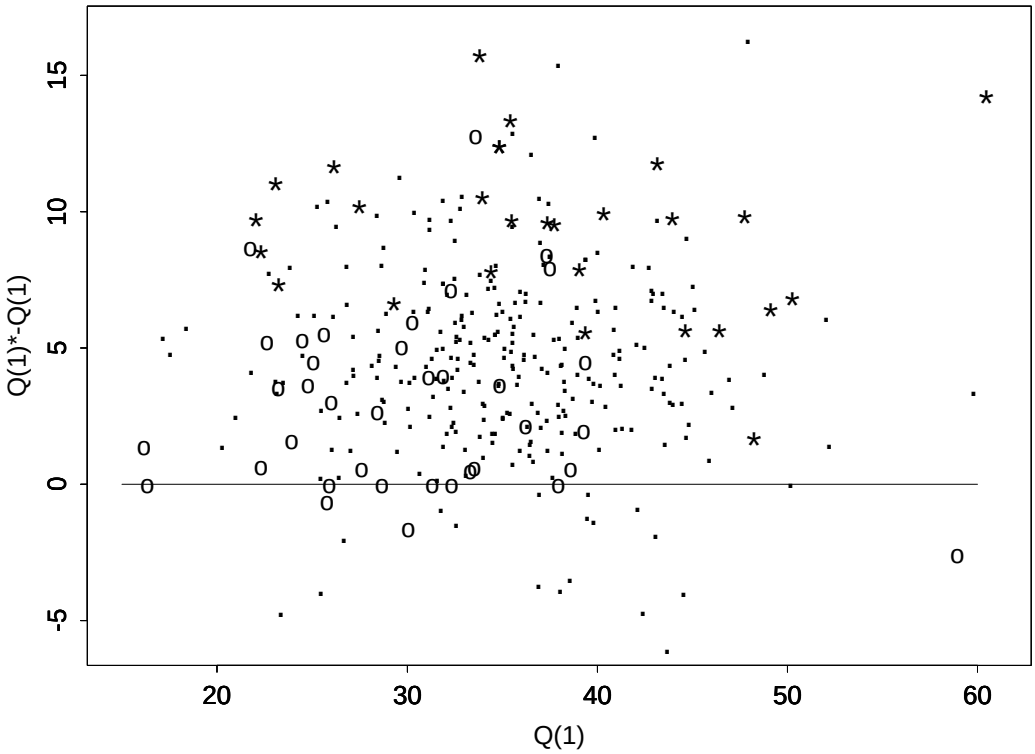

FIG. 4.9 – *Gain du taux de prédiction par l'utilisation des familles séquentielles, avec en abcisse le taux de prédiction* $\boldsymbol{Q}(1)$ *pour la méthode bayésienne simple et en ordonnée la différence entre ce taux* $\boldsymbol{Q}(1)$ *et celui obtenu avec les familles séquentielles* $\boldsymbol{Q}(1)^*$*. Les taux ont été donnés pour chaque protéine de la base de données: (*) les protéines tout-*α*, (o) les protéines tout-*β *et (.) les autres protéines; la classification suit la définition de Michie et collaborateurs [127].*

de 3%, (30,2 % contre 33,2 %), les $\alpha+\beta$ ont un gain de 4.9% (35,7 % - 40,6 %), et 4,8% pour les non-classifiées (33, 9% - 38,7%).

Globalement, on observe une augmentation des taux de prédiction pour les blocs ayant plusieurs familles séquentielles et en contre-partie une diminution pour les autres. Toutefois, cette diminution est faible. Il aurait été simple d'augmenter artificiellement les taux de prédiction des blocs les plus fréquents en les sub-divisant encore plus, jouant ainsi sur l'effet de l'importance numérique du bloc. Mais, alors les blocs moins fréquents auraient vu leur taux individuel décroître rapidement en dessous des 10 %, le taux global devenant par ailleurs largement supérieur à celui obtenu ici. La figure 4.8 montre que le nombre de familles séquentielles obtenues est raisonnable.

Nous pouvons noter que ce ne sont pas les blocs des structures répétitives les mieux prédits; le bloc protéique m (hélice α) a un taux de 50,6 % et le bloc protéique d (feuillet β) de 34,6 %, le PB a, une entrée en feuillet β, atteint lui un taux de 53,2 %.

4.4 Stratégies de prédiction

4.4.1 Le "nombre équivalent de blocs" (N_{eq}): un indice de confiance de prédiction

Après avoir prédit à partir de la séquence le (ou les) bloc protéique(s) le(s) plus probable(s) en appliquant le principe des familles séquentielles (*n séquences - 1 repliement local*). Nous avons décidé d'introduire un concept flou *1 séquence - n repliements* qui prend en compte le fait qu'une séquence peut être associée à plusieurs types de repliements, donc plusieurs types de blocs protéiques. En observant les résultats de la prédiction, le bloc réel est souvent le plus probable, mais il est surtout fort souvent parmi les plus probables. En conséquence, j'ai essayé de définir des stratégies pour sélectionner le nombre optimal de blocs r à prendre en compte en chaque site pour avoir un taux de prédiction donné.

Dans la suite de ce chapitre, deux types de stratégies distinctes vont être utilisées. Elles se basent toutes sur deux l'entropie de Shannon et sur le fait qu'une grande homogénéité de scores R_k en un site donné veut dire que l'informativité de la séquence X_S doit être faible. La prédiction associée localement est alors peu fiable. Inversement, un score élevé pour le bloc

protéique le plus probable doit être associé à un bon taux de prédiction. Dans le premier cas, il faudra choisir un nombre élevé de blocs, alors que dans le second, il en faudra moins. Pour quantifier cette incertitude, une entropie H a été calculée sur les scores R_k. Ces scores ont été dans un premier temps renormalisés en probabilités S_k :

$$S_k = \frac{R_k}{\sum_{l=1}^{l=B} R_l}$$

avec l d'ecrivant l'ensemble des blocs protéiques, avec $B = 16$ dans notre cas.

L'expression de l'entropie est alors :

$$H = -\sum_{l=1}^{l=B} S_l \ln(S_l)$$

Ensuite, l'entropie H est transformée en nombre équivalent de blocs noté N_{eq}:

$$N_{eq} = exp[H]$$

Cette quantité varie entre 1 quand un seul bloc est prédit, et, B quand les B blocs sont équiprobables. Les sites ayant un N_{eq} variant entre 1 et un N_{eq}^g (g allant de 1 à 8) ont été extraits de la base de données et le pourcentage de bonne prédiction Q_r a été ainsi calculé pour r rangs avec r variant entre 1 et 6, le bloc réel étant trouvé parmi les r rangs conservés. Les blocs sont tout d'abord classés par ordre de score décroissant.

De cette distribution, associée avec un intervalle de N_{eq} donné, nous déterminons le nombre rang optimal r pour assurer un taux de prédiction fixé. Cette étape a été effectuée pour tous les rangs possibles, allant de 1 à B.

Deux stratégies différentes ont ainsi été définies à partir des observations précédentes :

(i) *une approche globale* qui consiste à définir un nombre variable de blocs à conserver en chaque site s pour atteindre un taux global de prédiction Q_g fixé au préalable. Dans cette optique le nombre de blocs protéiques conservés varie le long de la séquence.

(ii) *une approche locale* qui tend à trouver les sites permettant pour un nombre fixé r de blocs conservés d'obtenir un taux de prédiction Q_l, lui aussi fixé au préalable. Dans cette approche, la prédiction est limitée à certaines régions des séquences protéiques.

Nous verrons donc dans une première partie, l'influence des familles séquentielles sur le N_{eq}, puis un exemple de prédiction sur un fragment de protéine pour voir l'évolution du N_{eq}, puis enfin successivement les deux stratégies.

4.4.2 Influence des familles séquentielles sur le N_{eq}

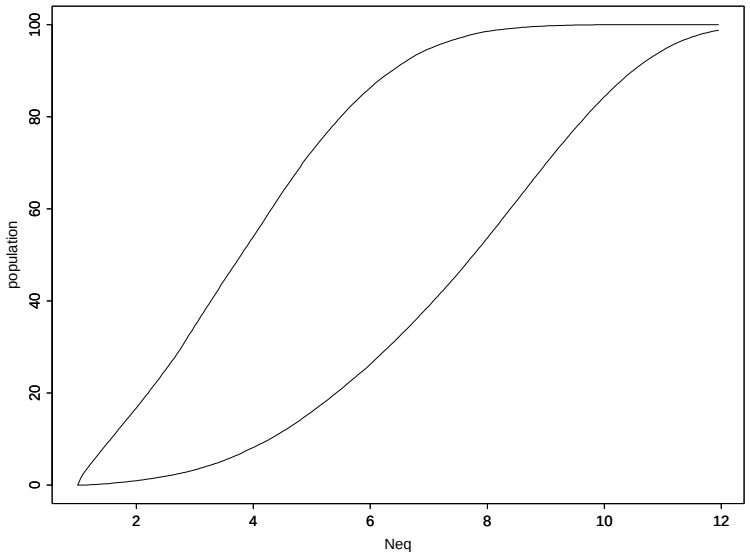

FIG. 4.10 – *Evolution du N_{eq} entre l'approche Bayésienne simple (courbe du bas) et l'utilisation des familles séquentielles (courbe supérieure).*

Par ailleurs, l'utilisation des familles séquentielles entraîne, du fait de la spécialisation séquentielle des blocs, une diminution du N_{eq} moyen comme reportée sur la figure 4.10 qui met en parallèle l'évolution du N_{eq} en fonction du nombre d'observation de la base de données, pour la prédiction par l'approche bayésienne simple et avec l'utilisation des familles séquentielles. Pour les blocs ayant plusieurs familles séquentielles, seul le score le plus élevé est conservé, donc le N_{eq} diminue; ceci change fortement l'allure de la courbe obtenue. Le N_{eq} moyen est passé de 7,6 à 4,2.

4.4.3 Exemple de prédiction

Pour mieux expliciter l'intérêt des stratégies, l'exemple suivant montre un exemple de la prédiction et le rôle du N_{eq}. Le tableau 4.2 donne les prédictions des 18 premières positions de la protéine de conjugaison à l'ubiquitine (cf. paragraphe 3.3.2.4), avec la fenêtre de 15 résidus correspondant au 5 C_α du bloc et aux 5 résidus présents de part et d'autre de cette fenêtre.

sequence			bloc réel	N_{eq}	BPs prédits		
gauche	centrale	droite			1^{er}	2^{nd}	3^{eme}
DMSTP	**ARKLM**	RDFKR	l	2,74	m(11,21)	$\underline{l}$(0,63)	d(0,36)
MSTPA	**RKLMR**	DFKRL	m	2,43	$\underline{m}$(17,51)	d(0,74)	f(0,43)
STPAR	**KLMRD**	FKRLQ	m	2,06	$\underline{m}$(33,68)	f(0,38)	d(0,18)
TPARK	**LMRDF**	KRLQQ	m	2,63	$\underline{m}$(11,05)	l(0,51)	k(0,36)
PARKL	**MRDFK**	RLQQD	m	2,48	$\underline{m}$(22,13)	f(1,25)	b(0,40)
ARKLM	**RDFKR**	LQQDP	m	3,78	$\underline{m}$(7,77)	k(1,90)	c(0,54)
RKLMR	**DFKRL**	QQDPP	m	2,92	$\underline{m}$(12,48)	b(0,94)	c(0,34)
KLMRD	**FKRLQ**	QDPPA	m	3,49	$\underline{m}$(12,98)	n(2,60)	p(0,73)
LMRDF	**KRLQQ**	DPPAG	m	6,32	$\underline{m}$(3,51)	n(0,55)	d(0,38)
MRDFK	**RLQQD**	PPAGI	m	8,61	p(2,02)	b(1,08)	$\underline{m}$(1,02)
RDFKR	**LQQDP**	PAGIA	m	4,82	p(3,03)	d(1,12)	c(0,44)
DFKRL	**QQDPP**	AGIAG	c	2,55	$\underline{c}$(4,43)	d(0,19)	p(0,12)
FKRLQ	**QDPPA**	GIAGA	c	3,10	f(13,43)	$\underline{c}$(2,87)	k(0,23)
KRLQQ	**DPPAG**	IAGAG	e	5,45	b(7,14)	$\underline{e}$(1,94)	g(1,68)
RLQQD	**PPAGI**	AGAGI	h	5,34	b(12,39)	$\underline{h}$(6,72)	l(3,16)
LQQDP	**PAGIA**	GAGIS	i	4,97	$\underline{i}$(11,29)	p(5,29)	c(1,40)
QQDPP	**AGIAG**	AGISG	a	4,75	g(15,58)	$\underline{a}$(6,16)	e(3,24)
QDPPA	**GIAGA**	GISGA	c	6,75	b(7,15)	h(4,01)	$\underline{c}$(2,32)

TAB. 4.2 – *Exemple de prédiction de la partie N-terminale de la protéine de conjugaison à l'ubiquitine, avec les 15 acides aminés de chaque séquence prédite avec la partie centrale représentant le bloc structural et son environnement avec les 5 résidus de part et d'autre, le bloc réel, le* N_{eq} *et les trois blocs les plus probables classés par ordre décroissant. Le bloc souligné est le réel.*

Cette partie N-terminale est composée d'une hélice α formée par 10 blocs protéiques m suivi par une boucle de 7 blocs qui mène à un feuillet β. Cet exemple est basé sur l'utilisation des familles séquentielles précédemment décrite. Chaque ligne correspond à une séquence, par exemple, la cinquième fenêtre centrée sur MRDFK est assignée au bloc protéique m. Les trois premières solutions ont été ordonnées suivant leur score de prédiction R_k, pour BP m, BP f et BP b, leurs scores respectifs étant de 22,13, 1,25 et 0,40.

Ainsi, le premier score indique que la probabilité du bloc m est 22,13 fois plus élevée que celle d'avoir ce bloc de façon purement aléatoire. En cette position, la prédiction est correcte. Les scores élevés des premières positions sont justifiés par la présence de résidus Leucine, Méthionine, Arginine, Lysine, Aspartate et Leucine en position (-3), (-2), (-1), (+2), (+3) et (+4). De même, le BP f est classé en seconde position du fait de la prsence de l'Aspartate en position centrale de la fenêtre. En ne conservant que les premiers rangs, 10 blocs protéiques sont correctement prédits sur 18. Sur l'ensemble des protéines, le taux de prédiction Q(1)* est de 40,8%. Sans tenir compte des familles séquentielles, il était de 30,4%, soit un gain de plus de 10%. Classiquement, les taux de prédiction ne sont calculés que pour les solutions optimales. Mais, en observant, les solutions des trois premiers rangs, 17 des 18 blocs y sont. La position erronée correspond à une fin d'hélices α qui possède une composition inhabituelle en acides aminés, KRLQQDPPA en [-4;+4].

Aussi, au lieu de ne prendre en compte que les premiers rangs, une approche pertinente revient à examiner les taux de prédiction Q(r) pour un rang donné r. Le N_{eq} permet de quantifier cette dispersion parmi les scores. Ainsi dans la première partie de l'hélice α, le N_{eq} varie entre 2,06 et 3,78; il est ainsi corrélé avec une bonne prédiction. Inversement, à la fin de l'hélice α, la probabilité de trouver le bloc protéique réel décroît alors que le N_{eq} augmente au-delà de 4,82. Les sites sont de moins en moins informatifs. Des N_{eq} intermédiaires sont observés pour les 7 derniers résidus, le nombre de rang à conserver est alors de 2.

Cet exemple montre l'intérêt des stratégies de prédiction basées sur un nombre variable de blocs sélectionnés par site.

4.4.4 Stratégie globale

En utilisant la base de données, nous avons établi la relation qui existe entre la probabilité de trouver le bloc réel parmi les r blocs les plus probables pour un N_{eq} donné. Cette démarche permet d'obtenir la figure 4.11 qui met en relation le taux de prédiction en fonction du N_{eq} et du nombre de solutions conservées. Cet exemple utilise les familles séquentielles.

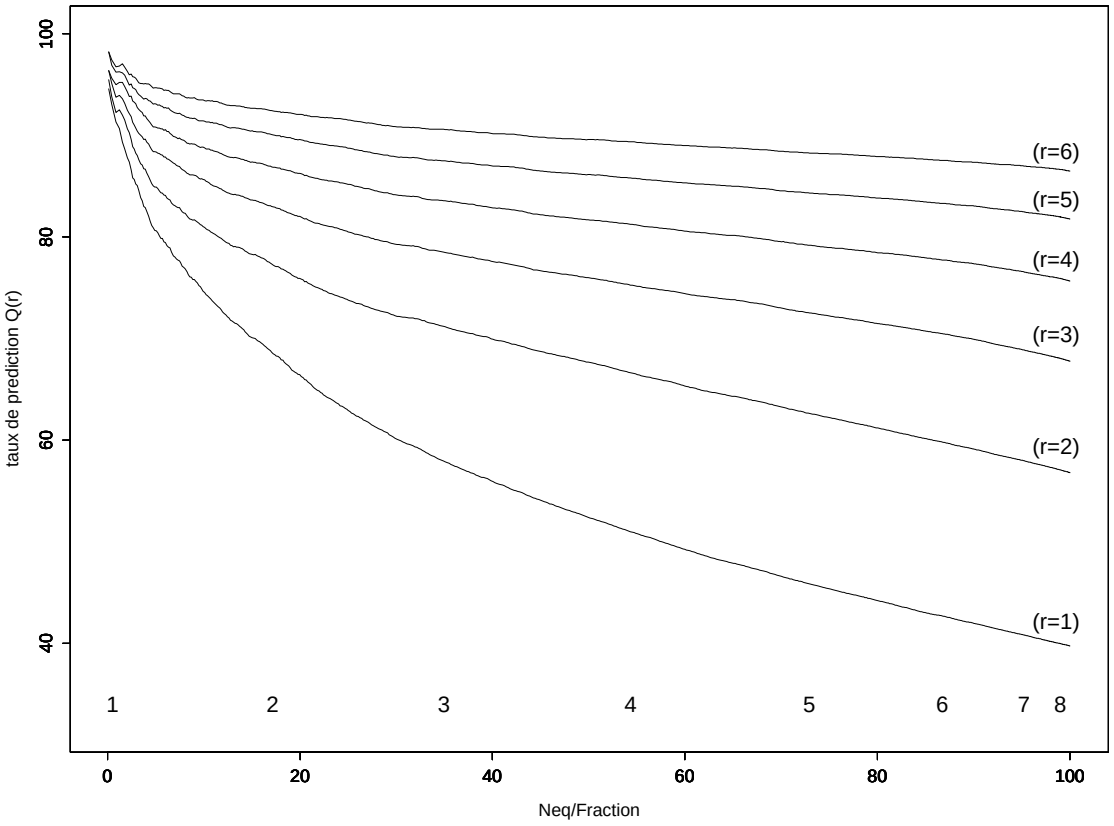

FIG. 4.11 – *Taux de prédiction* $\boldsymbol{Q}(r)$ *associé à chaque taux de Neq pour r variant entre 1 et 6.*

La distribution du taux de prédiction associé à chaque N_{eq} a été calculée. Pour chaque taux de prédiction Q_g, le nombre de rangs, (i.e. nombre de blocs) à conserver pour atteindre Q_g en fonction du N_{eq} a été déterminé. Par exemple, pour un N_{eq} inférieur à 6,32, il faut sélectionner les 3 PBs les plus probables pour avoir un taux de bonne prédiction de 70 %.

La figure 4.12 montre le résultat de cette stratégie pour la protéine *1aak* (cf. paragraphes 3.3.2.4 et 4.2). Le profil des N_{eq} (figure 4.12a) montre la variation de cet indice entre 1,06 et 9,79. La figure 4.12b donne en chaque site le rang véritable du bloc dans la prédiction. 77,8% des blocs réels sont parmi les 3 solutions les plus probables. Certaines zones de la protéine nécessite de conserver un grand nombre PBs, comme les deux boucles reliant les deux feuillets

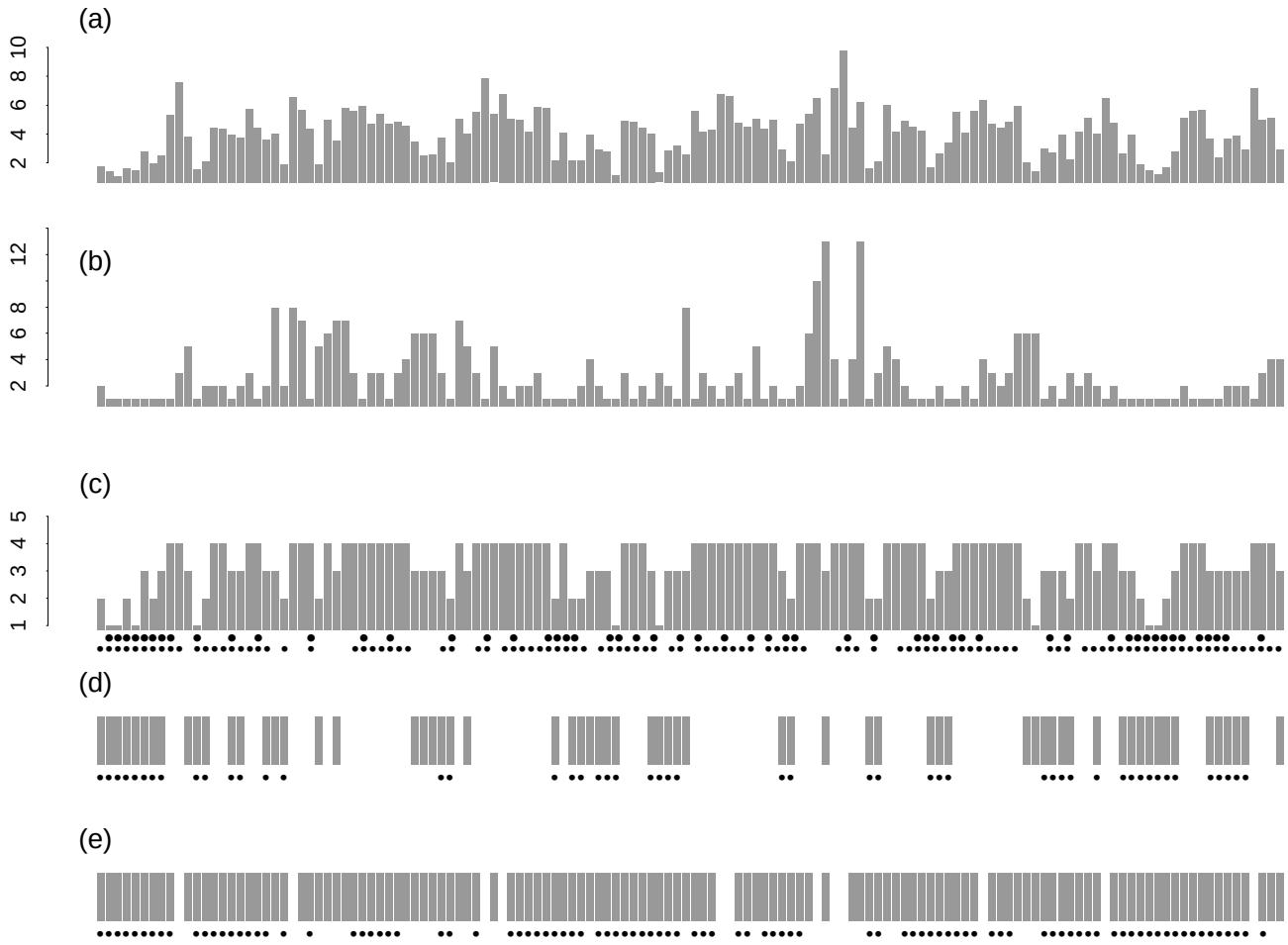

FIG. 4.12 – *Application des 2 stratégies à la protéine de liaison à l'ubiquitine avec (a) évolution du* N_{eq} *le long de la séquence, (b) position du bloc réel dans les solutions, (c) taux global de prédiction* Q_g *de 75 %, avec le nombre à conserver alors, les petits points donnent les positions où le bloc réel est au premier rang, les points plus importants sont les positions où le bloc réel est dans les rangs conservés, (d) zone conservée pour la stratégie locale avec un taux de prédiction* Q_l *de 75 %, en conservant* r = *3 rangs, et, (e) même stratégie pour* Q_l*=70 % et* r = *3 rangs.*

β (positions 22 à 46), et la large boucle (positions 82 à 90) contenant une petite hélice α.

La figure 4.12c montre le nombre de blocs protéiques devant être sélectionné pour avoir un taux global de prédiction Q_g de 75%. Les séries de points en-dessous définissent les sites où le bloc réel est trouvé au premier rang, et parmi les rangs sélectionnés. Le nombre maximal de blocs protéiques est de 4. Le taux de prédiction au premier rang est de 40,7%; pour Q_g=75%, de 1 à 4 blocs protéiques par site sont sélectionnés (8, 17, 37 et 72 sites respectivement). Les structures répétitives et les blocs proches de ces dernières (cf. figure 3.10) sont bien délimitées. Comme attendu, les boucles sont plus difficilement prédictibles. En observant les deux séries de points, il est net que les zones possédant des blocs bien prédits au premier rang sont les plus aisées à prédire avec un nombre restreint de blocs.

Cette stratégie amène à un excès de blocs en chaque position. Cependant, en contrepartie le taux de prédiction est toujours assuré.

4.4.5 Stratégie locale

La seconde stratégie diffère de la précédente qui prenait un nombre variable de blocs protéiques sur l'ensemble de la protéine, en utilisant un nombre constant de blocs sur une partie de la protéine. Ainsi, un taux de prédiction Q_l est garanti pour r blocs choisis. La figure 4.11 montre chaque valeur de N_{eq} et pour un nombre r variant de 1 à 5. Pour calculer ces courbes, les sites compris entre 1 et chaque valeur de N_{eq} ont été conservés et le taux de prédiction associé calculé. Cet étude a été répétée depuis r=1 (juste le premier rang, le plus probable des blocs) jusqu'au 6 premiers rangs inclus.

Pour donner un exemple, si l'on désire avoir 70 % des sites le N_{eq} doit être inférieur à 4,8. En sélectionnant 1 (de même 2, 3 et 4) rang(s), le taux de prédiction associé est de 46,8 % (de même 63,4 % 73,1 % et 79,6 %). De la même manière pour un taux de prédiction donné de 80 %, le N_{eq} est de 1,28 et 5,5 % des sites seront sélectionnés avec un seul rang; ils passent respectivement à 1,6 et 11,5 % pour les deux premiers rangs, 2,6 et 26,9 % pour les trois premiers, et, 4,6 et 66,4 % pour les quatres premiers.

La figure 4.12d est un exemple de cette stratégie appliquée à la protéine *1aak*, les zones prises en compte représentent un taux de prédiction Q_l de 75%, en conservant 3 rangs. Le N_{eq} correspondant est alors inférieur à 5,11. 62 ont été sélectionnés, les points en dessous montrent

les 49 positions où le bloc réel se trouve parmi les blocs sélectionnés. Le taux final de prédiction est de 79 % pour 46,3 % des sites de la protéine pris en compte. En comparant avec la précédente approche, il est clair que prendre 3 blocs de manière fixe est un excès. De la même manière avec $r = 4$ et $Q_l = 75$ %, 52 % des résidus de la protéine sont alors utilisés.

La figure 4.12e montre la même stratégie pour Q_l=70 % et $r = 3$ rangs. En utilisant un N_{eq} maximal de 6,32, 122 sites, soit 91 % des sites de la protéine ont été sélectionnés et 95 de ces sites possèdent le bloc réel parmi ceux choisis soit un taux de prédiction de 77,9 %.

Ainsi, cette stratégie permet de localiser les sites les plus prédicitibles, cependant une recherche préalable doit être menée quand au nombre r de rangs qui doit être sélectionné. Par exemple, pour un taux de prédiction de Q_l=70 %, la proportion des sites sélectionnés augmente fortement avec un passage de r =2 à 3 rangs. (une augmentation de 49 %). Pour de future application de ces stratégies, comme dans des méthodes *ab initio*, le choix du nombre de blocs sélectionnés par site pose un certain problème : augmenter le nombre de rangs conservé r permet une prise en compte d'une plus grande partie des sites, mais aussi induit une combinatoire plus complexe pour reconstruire un modèle moléculaire.

4.5 Conclusion

La figure 4.13 récapitule l'ensemble du processus bayésien mis en place. Dans un premier temps, les séquences ont été directement utilisées en donnant un taux de prédiction plus que convenable de 34,4 % pour 16 états possibles. La définition des familles séquentielles (*1 bloc* $\rightarrow$ *n séquences*), permet à la fois un gain global de prédiction et une homogénéisation des taux de prédiction des blocs en conservant une homogénéité structurale des blocs protéiques. L'augmentation du taux de prédiction est significatif avec un passage de 34,4 à 40,7 %.

La succession d'une certaine série d'acides aminés n'obligent pas un seul type de repliements [186], mais notre approche permet de voir que cette succession induit un certain type de repliement qui peut être particulièrement bien caractérisé. La disposition préférentielle des blocs réels parmi les blocs les plus probables a permis l'élaboration du concept (*1 séquence* $\rightarrow$ *n blocs*) avec l'utilisation d'un indice de confiance le N_{eq} qui permet de bien localiser les zones les plus probables.

Ces recherches ont permis la mise au point de deux stratégies distinctes pour rechercher les

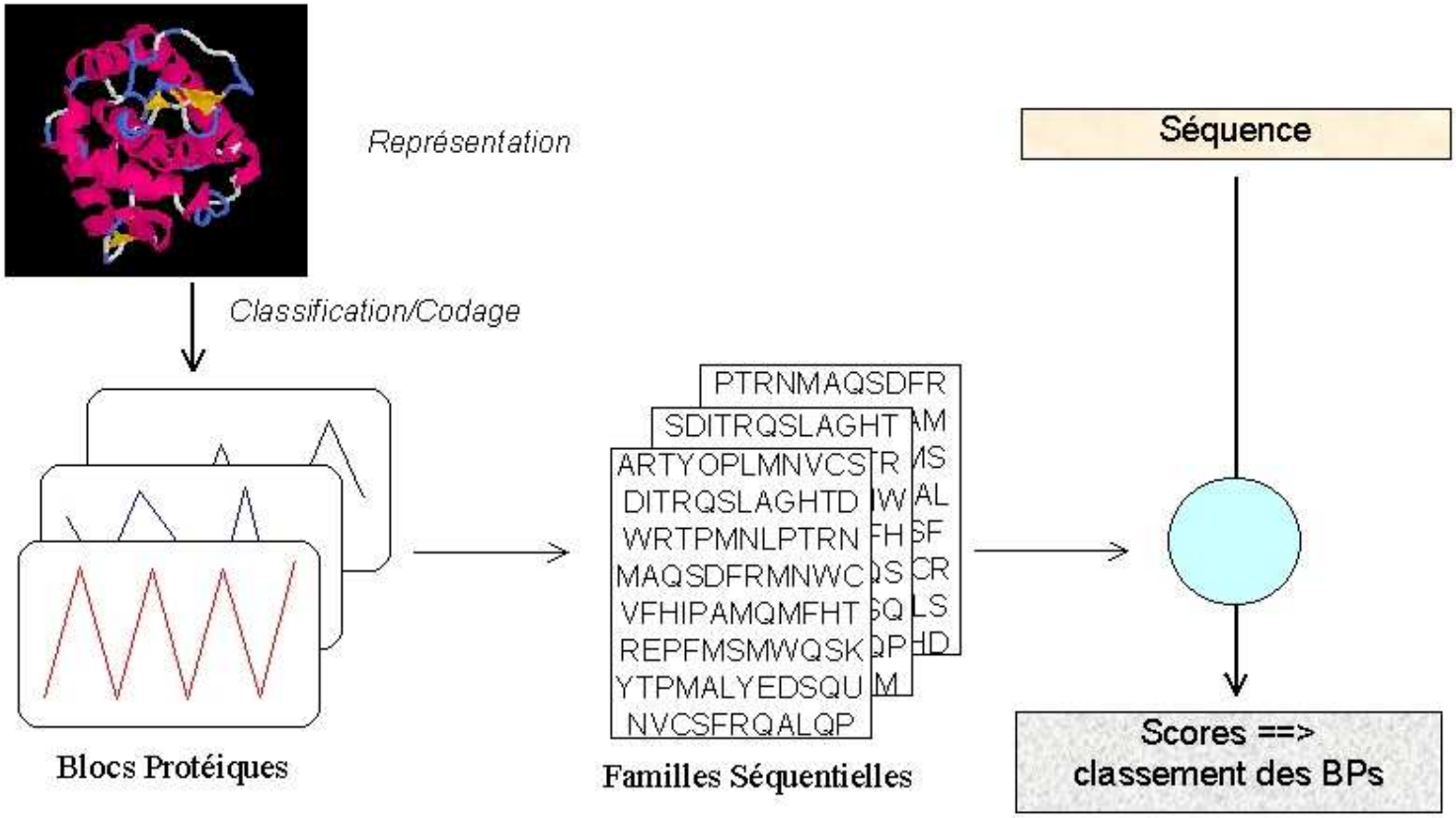

FIG. 4.13 – *Schéma récapitulatif de la prédiction bayésienne, aprés la réalisation des blocs protéiques, les fragments protéiques sont associés à leur bloc et cette information est directement utilisée par la prédiction.*

zones et/ou le nombre de blocs à conserver pour aboutir à un taux de prédiction donné. La première stratégie donne un nombre variable de blocs en chaque résidu suivant le N_{eq} local, la seconde avec un nombre fixe de blocs donne un taux de prédiction garanti pour un nombre de zones prédite limitée. Ces deux stratégies donnent toujours un nombre trop élevé de blocs à conserver, il faudra y remédier dans un proche futur. L'utilisation de méthode classique type réseau de neurone artificiel et/ou alignements de séquence aurait surement donné des résultats quantitativement supérieurs. Toutefois, ces méthodes sont des techniques qui ne permettent en aucun cas de comprendre les tenant et les aboutissant de manière globale. Développant un nouvel alphabet largement plus complexe que les structures secondaires classiques, il aurait été regrettable de ne pas comprendre les spécificités de chacun des blocs et de passer directement à un apprentissage type "boite noire". Maintenant que cette première étape a été effectuée, avec le Pr. Hazout, nous développons une approche de type réseau neuronal artificiel dont nous attendons beaucoup.