HAL open science

”Hybrid Protein Model (HPM)” : une nouvelle approche pour caractériser les relations séquence-structure dans les protéines

Alexandre de Brevern, Serge A. Hazout

To cite this version:

Alexandre de Brevern, Serge A. Hazout. ”Hybrid Protein Model (HPM)” : une nouvelle approche pour caractériser les relations séquence-structure dans les protéines. 2000, pp.105-112. inserm-00134567

HAL Id: inserm-00134567

https://inserm.hal.science/inserm-00134567

Submitted on 2 Mar 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL author manuscript

Recueil des actes des Premières Journées Ouvertes de Biologie, Informatique et Mathé

HAL author manuscript inserm-00134567, version 1

“ Hybrid Protein Model (HPM) ” : une nouvelle approche pour caractériser les relations séquence-structure dans les protéines.

Alexandre de Brevern & Serge Hazout

Equipe de Bioinformatique Génomique et Moléculaire (EBGM), Unité INSERM U436,

Université Denis Diderot-Paris7, case 7113,

2, place Jussieu, 75251 Paris Cedex 05

Résumé

Le passage de la séquence protéique codante à la structure tridimensionnelle est une des difficultés majeures de la biologie structurale. La recherche d'un alphabet structural permettant de coder la structure locale du squelette protéique a été précédemment réalisée au moyen d’un classifieur non supervisé (A. de Brevern *et al*.). L’alphabet obtenu en 16 Blocs Protéiques (BPs) assure une bonne approximation des structures protéiques. La prédiction locale de ces BPs montre que l'information de la séquence induit fortement le type de repliement local, mais reste imparfaite (taux de prédiction correcte = 40,7%). La méthode “ Hybrid Protein Model ” présentée dans cette étude 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. L’analyse de la répartition le long de la protéine hybride en relation avec la nature du bloc structural a permis d’affiner le rôle de certains acides aminés dans les structures secondaires et des régions flanquantes. L’étude aboutit à un concept de “ modèle flou ” entre la séquence et la structure.

Introduction

Les protéines se replient suivant un nombre limité de conformations (Govindarajan et al., 1999), toutefois la complexité et le nombre important de paramètres physico-chimiques, cinétiques, dynamiques et stériques rentrant en jeu rend la prédiction de la structure tridimensionnelle d'une protéine difficile, si cette dernière n'a pas de similitudes importantes avec des protéines dont la structure est déjà connue. L'augmentation des bases de données génomiques rend ce problème encore plus crucial aujourd'hui.

Une première analyse des structures protéiques fait apparaître le rôle essentiel des structures secondaires (correspondant à un alphabet structural à 3 niveaux), structures répétitives (hélices α et feuillets β), stables du fait de liaisons internes importantes et structures intermédiaires plus labiles, les boucles. Les premiers algorithmes simples de prédiction, tel GOR (Garnier *et al*., 1978) donnaient des taux de prédictions proches de 60%. Des approches plus sophistiquées, telles les réseaux neuronaux couplés à des alignements de séquences, aboutissent à des performances en constante progression (Rost et Sander, 1993; Salamov et Solovyev, 1997). Toutefois ce gain de prédiction induit une impossibilité d'analyse des facteurs rentrant en jeu (en particulier sur le rôle de certains acides aminés). L'augmentation du nombre de structures de protéines cristallisées a permis le développement de travaux de recherche visant à étudier la structure des boucles, celles-ci étant plus variables de par leur nature propre (Kwasikroch *et al.*, 1996, Wodjick *et al.*, 1999).

HAL author manuscript inserm-00134567, version 1

D'autres groupes de recherche se sont attachés à établir un alphabet structural plus large prenant en compte l'hétérogénéité du squelette protéique. On peut citer les travaux de Rooman et collaborateurs (1990) et de Fetrow et collaborateurs (1997) dans lesquels ils ont constitué un petit nombre (4 à 7) de structures de petites tailles fortement dépendantes des structures secondaires. Il en ressort que l'information séquentielle est relativement caractéristique, cependant leur approximation de la structure protéique réelle reste pauvre. Pour rechercher de façon systématique les formes les plus communes de repliements, Unger *et al.* (1989) et Schuchhardt *et al*. (1996) ont développé des algorithmes de regroupement de blocs protéiques. Ils en obtiennent 100, ce qui paraît relativement excessif pour une stratégie de prédiction de structure. Bystroff et Baker (1998) ont élaboré une méthode de construction et de prédiction de blocs de tailles variables allant de 3 à 17 acides aminés. Ils ont extrait des formes caractéristiques, cependant ils ne donnent aucune précision moyenne sur l'approximation d'une structure protéique réelle par leurs blocs. Récemment, nous avons mis au point une méthode d'obtention de Blocs Protéiques (PBs), il s'agissait d'un classifieur non supervisé tenant compte des transitions entre blocs protéiques (de Brevern *et al*., soumis) d'une manière comparable à l'approche " Hidden Markov Model (HMM) " développée par Camproux *et al*. (1999). Un alphabet structural en 16 blocs a été défini. Le taux de prédiction correcte de ces blocs par une approche bayésienne est de 34,2% en ne conservant que le plus probable en chaque site d'une protéine, et passe à 40,7% en subdivisant les blocs les plus fréquents suivant leurs spécificités de séquences. L'information locale de la séquence est beaucoup plus informative sur la structure locale que l'on peut croire au premier abord; on voit ainsi que 75,8% des vrais blocs se retrouvent parmi les 4 blocs les plus probables (au sens de la stratégie bayésienne) et ceci indépendamment du type de bloc.

De cette précédente remarque, on peut essayer caractériser la "dépendance floue" entre la séquence et la structure locale du squelette protéique si l'on admet l'hypothèse vraie (cela signifie que pour une chaîne d'acides aminés, il existe une loi de probabilité d'apparition des blocs structuraux, et réciproquement). Pour établir cette dépendance, nous avons mis au point une méthode d'apprentissage étiquetée "Hybrid Protein Model" (HPM) liant séquence et structure en une même observation et qui tend à conserver la séquentialité des observations. Dans ce papier, nous décriverons le principe général de la méthode, puis nous montrerons les résultats obtenus sur la dépendance séquence-structure en termes de choix d'acides aminés. Afin de mieux les interpréter, nous les avons mis en correspondance avec notre alphabet structural en 16 blocs. L'utilisation des Blocs Protéiques permet d'analyser la répartition des observations de façon plus détaillée qu'avec un simple alphabet à trois labels, puisque les BPs permettent une décomposition fine de la structure protéique tridimensionnelle.

Matériels et Méthodes

Les données

La base de données est composée de 342 protéines ayant moins de 25% d'identités (Hobohm et al., 1992; Hobohm et Sander, 1994). Les protéines ont été découpées en fragments de 5 acides aminés consécutifs, soit 86980 fragments. Chaque acide aminé a été recodé selon trois variables : l'hydrophobicité (Kyle et Doolitle, 1982), le volume de la chaîne latérale (Zamyatin,1972), et la charge

HAL author manuscript inserm-00134567, version 1

(-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 et +1. 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, après avoir été préalablement décalés pour les angles Ψ supérieurs à 120° de -360° et pour les angles Φ inférieurs à -120° de +360°. Ainsi chaque fragment de 5 résidus est défini par un vecteur $\boldsymbol{V}$ de 23 composantes (15 pour la séquence et 8 pour la structure).

Hybrid Protein Model (HPM)

Dans notre étude, la protéine hybride correspond à une succession de L fragments de 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 HPM est d'apprendre "au mieux" l'ensemble de la base de vecteurs (au nombre de 86980) par cet hybride de L vecteurs. L'apprentissage est similaire à celui d'une carte auto-organisée de Kohonen ("Self-Organizing Map" ou SOM; Kohonen 1989, 1997). Cependant, dans notre cas, on se limite à un apprentissage monodimensionnel et la diffusion de l'information le long de l'hybride n'est pas réalisée artificiellement, elle est implicite quand on prend plusieurs vecteurs successifs dans la protéine étudiée. 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 1 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) On calcule pour chacune position p de l'hybride un score $S(p)$ (distance euclidienne) de dissemblance entre la sous-matrice $\boldsymbol{V}$ et celle $\boldsymbol{W}(p)$ de même taille prise dans l'hybride. Ainsi on établit un profil de scores le long de l'hybride. (3) On recherche le score minimum S_{min} et donc la position p^* = argmin{$S(p)$} associée 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 $\boldsymbol{V}$. La transformation est définie par l'équation :

$$\boldsymbol{W}(p) \rightarrow \boldsymbol{W}(p) + (\boldsymbol{V} - \boldsymbol{W}(p)).a(n) \quad \text{et} \quad a(n) = a_0/(1 + n/N)$$

n désignant le nombre de sous-matrices présentées à l'hybride, N le nombre total de sous-matrices de la base de données et a_0 le coefficient d'apprentissage initial. Le coefficient d'apprentissage $a(n)$ décroît au cours de l'apprentissage. Ayant modifié l'hybride, on passe au fragment suivant jusqu'à traiter complètement la base.

HAL author manuscript inserm-00134567, version 1

(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.

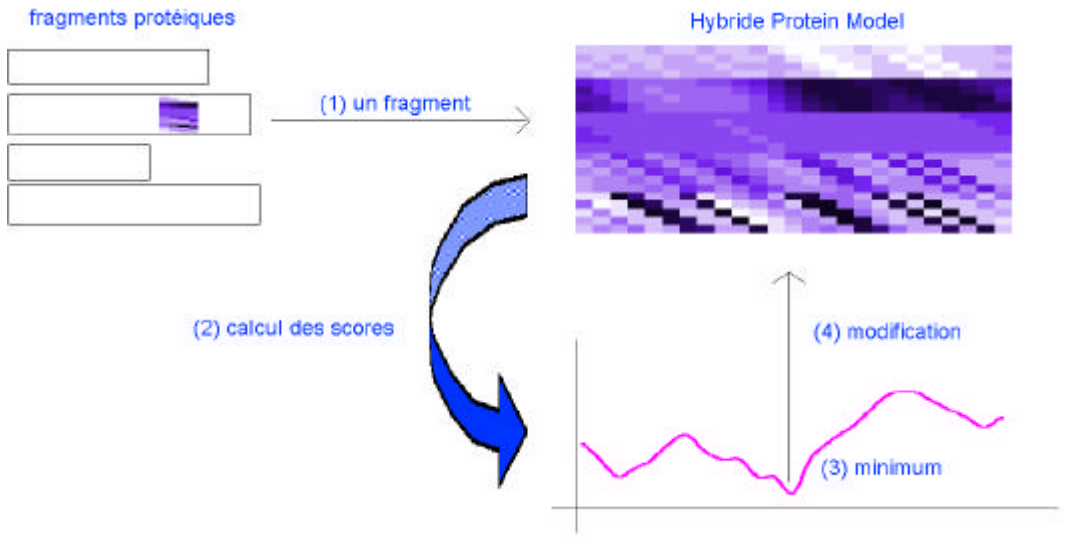

Figure 1 : Schéma représentant le principe d'apprentissage des informations séquence-structure par la "protéine hybride" (voir les explications dans le texte). (1) Sélection d'un fragment défini par une sous-matrice d'observations. (2) Calcul d'un score local entre ce fragment et une région de l'hybride 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.

Blocs Protéiques

Les 16 blocs protéiques (labellés de BP*a* à BP*p*) ont été définis sur la même base de données par un classifieur non supervisé tenant en compte des transitions entre blocs structuraux. Ils permettent une bonne approximation des structures tridimensionnelles des protéines (de Brevern *et al.*, soumis). Les blocs BP*a* à BP*f* représentent les blocs associés aux feuillets β, la forme régulière étant BP*d*, les blocs le précédant étant ses extrémités N-caps, les suivant étant ses extrémités C-caps. De même, pour les blocs liés aux hélices α, le bloc BP*m* correspond à la forme régulière (partie centrale d'une hélice droite) entouré par les blocs BP*k* et BP*l* (N-cap), et, BP*n* à BP*p* (C-cap). Les derniers blocs BP*g* à BP*j* sont principalement présents dans les structures en boucle.

Résultats

Nous ne présenterons que les principaux résultats de la phase d'apprentissage des structures protéiques par HPM : (i) Description de la " protéine hybride " en termes de séquence - structure et (ii) Correspondance entre la protéine hybride et les blocs structuraux.

HAL author manuscript inserm-00134567, version 1

Description de la " protéine hybride " en termes de séquence - structure.

La figure 2 donne le résultat de l'apprentissage effectué avec un coefficient d'apprentissage α_0 = 0,03 sur C = 20 cycles pour une protéine hybride de longueur L = 25. Une première constatation évidente apparaît : la séquentialité des fragments est visible, cependant on note que 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 Ψ 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. On note que les 25 patterns (ou fragments) présentent globalement des caractéristiques différentes. Cependant un regroupement des patterns par une classification hiérarchique montre la constitution de deux groupes homogènes distincts ayant comme frontières les 2éme et 13éme positions. Après apprentissage, on a noté le nombre de fois où une position de l'hybride est sélectionnée. La répartition des observations est relativement homogène le long de l'hybride, les nombres variant entre 2950 et 4500 observations.

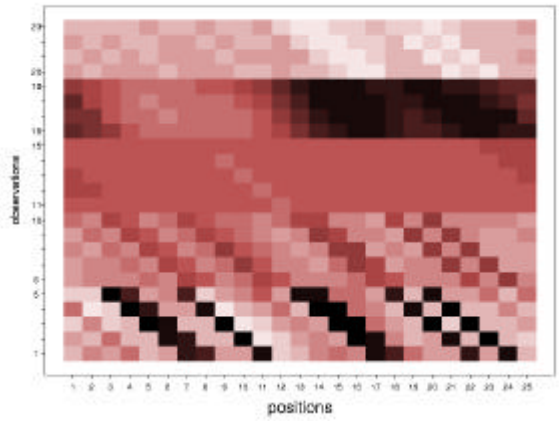

Figure 2 : La " Protéine Hybride " après apprentissage.

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, on a par tranches de 5 lignes : l'hydrophobicité, le volume, la charge et les angles dièdres Ψ et Φ.

Correspondance entre la protéine hybride et les blocs structuraux.

La figure 3 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, on retrouve les blocs qui concernent les hélices α et leurs régions flanquantes, et de 13 à 25, les

HAL author manuscript inserm-00134567, version 1

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 (G) et des prolines (P) dans ces zones en association avec les blocs BP*h*, BP*i* et BP*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 bloc BP*d* 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 côtés est enfoui et l'autre exposé.

(iii) certains se retrouvent majoritairement en certaines positions de l'hybride. par exemple, BP*l* (N-cap d'une hélice α) est localisé en positions 1-4 où apparaissent des résidus chargés, BP*m* (région centrale d'une hélice α) en positions 4-11 dans lesquelles des résidus hydrophobes (I, V, L, F, M et partiellement A) et où l'on retrouve le motif hydrophobe (i, i+1, i+4), c'est à dire l'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.

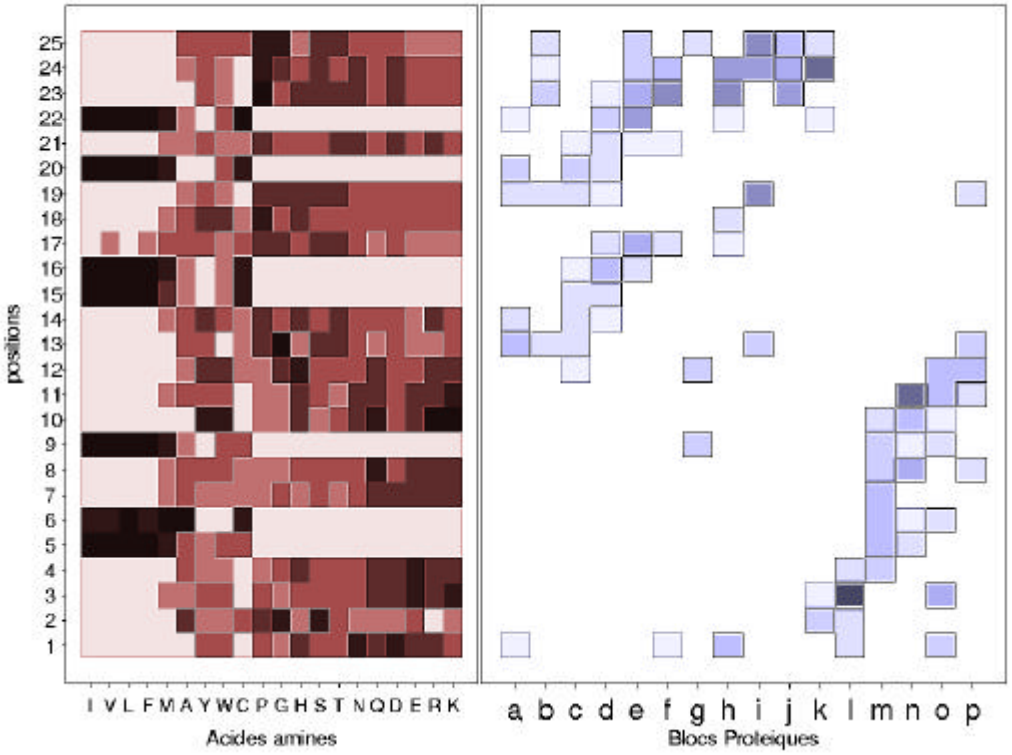

Figure 3 : Correspondances entre la " protéine hybride et les blocs structuraux. (a) Figure de gauche : Répartition des acides aminés le long de la protéine hybride uniquement pour l'élément central des fragments. (b) Figure de droite : Répartition des fragments dans les blocs structuraux.

HAL author manuscript inserm-00134567, version 1

Discussion et conclusion

La répartition des Blocs Protéiques montrent 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, difficilement appréciable avec une séquentialité à 3 niveaux. De plus l'utilisation des structures secondaires se heurte à l'établissement toujours arbitraire de règle d'attribution à l'une des 3 classes (Colloc'h *et al.*, 1993). 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 façon 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 modélisation *ab initio* . Différents développements complémentaires sont en cours tel le tracé du squelette protéique dans la table séquence-structure de la figure 3. Cette méthode peut aussi de servir de validation lors de l'élaboration de modèles structuraux complets.

Références (Times 12 points, gras)

DE BREVERN A.G., ETCHEBEST C. and S. HAZOUT , Bayesian probabilistic approach for prediction backbone structures in terms of protein blocks, *submitted*.

BYSTROFF C., and D. BAKER (1998), Prediction of local structure in proteins using a library of sequence-structure motif, *Journal of Molecular Biology*, **281**, 565-77.

HAL author manuscript inserm-00134567, version 1

CAMPROUX A.C., TUFFERY P., CHEVROLAT J.-P., BOISVIEUX J.-F. and S. HAZOUT (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone, *Protein Engineering* , **12**,.

COLLOC'H N., ETCHEBEST C., THOREAU E., HENRISSAT B., AND J.-P. MORNON (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment, *Protein Engineering* ;**6**:377-382.

FETROW J.S., PALUMBO M.J., and R. BERG (1997), Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme, *Proteins*, **27**, 249-71.

GARNIER J., OSGUTHORPE D.J., and R. ROBSON (1978), Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular protein, *Journal of Molecular Biology*, **120**, 97-120.

GOVINDARAJAN S., RECABARREN R., and R.A. GOLDSTEIN (1999), Estimating the total number of protein folds, *Proteins*, **35**, 408-14.

HOBOHM U, SCHARF F., SCHNEIDER R., and C. SANDER (1992), Selection of a representative set of structures from the Brookhaven Protein Databank ,*Protein Science*, **1**, 409-417.

HOBOHM U. , and C. SANDER (1994), Enlarged representative set of protein structures, *Protein Science*, **3**, 522-524.

KOHONEN T. (1989), An introduction to neural computing, *Neural Networks*, **1**, 3-16.

KOHONEN T. (1997), Self-Organizing Maps. (2nd edition) In T.S. Huang, T. Kohonen, R.M. Schroeder and H.K. Loetsch (eds.): *Springer Series in Information Sciences 30*., Springer-Verlag Berlin.

KWASIGROCH J.-M., J. CHOMILIER, and J.-P. MORNON (1996), A global taxonomy of loops in globular proteins, *Journal of Molecular Biology*, **259**, 855-872.

KYLE J., and R.F. DOOLILTTLE (1986), A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology*, **157**, 105-132.

ROOMAN M.J., RODRIGUEZ J., and S.J. WODAK (1990), Automatic definition of reccurent local structure motifs in proteins, *Journal of Molecular Biology*, **213**, 327-336.

ROST B. , and C. SANDER (1993), Prediction of protein secondary structure at better than 70\% accuraccy, *Journal of Molecular Biology* , **232**, 584-599.

SALAMOV A.A., and V.V. SOLOVYEV (1997), Protein secondary structure prediction using local alignements, *Journal of Molecular Biology*, **268**, 31-36.

SCHUCHHARDT J., SCHNEIDER G., REICHELT J., SCHOMBURG D., and P. WREDE (1996), Local structural motifs of protein backbones are classified by self-organizing neural networks, *Protein Engineering*., **9**, 833-842.

UNGER R., D. HAREL, S. WHERLAND, and J.L. SUSSMAN (1989), A 3D building blocks approach to analyzing and predicting structure of proteins, *Proteins*, **5**, 355-373.

WODJICK J., MORNON J.-P., and J. CHOMILIER (1999), New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification, *Journal of Molecular Biology*, **289**, 1469-1490.

ZAMYATIN A.A. (1972), Protein volume in solution, *Prog. Biophys. Mol. Biol*., **24**, 107-123.