Prédiction de la structure 3D des boucles protéiques

Christelle Reynes, Leslie Regad, Robert Sabatier, Anne-Claude Camproux

Prédiction de la structure 3D des boucles protéiques

Christelle Reynès[1], Leslie Regad[3], Robert Sabatier[2] & Anne-Claude Camproux[1]

[1] *MTi, Unite Inserm, Paris 7 Diderot U973, Bât Lamarck, 35, rue Hélène Brion 75205 Paris Cedex 13*

[2] *Laboratoire de Physique Industrielle et Traitement de l'Information, EA 2415, Faculte de Pharmacie, 15 Avenue Charles Flahault, BP 14491, 34093 MONTPELLIER Cedex 5*

[3] *Faculty of Pharmacy, Centre for Drug Research, Computational Drug Discovery group P.O. Box 56, Viikinkaari 5 E, FI-00014 University of Helsinki*

Résumé

La connaissance de la structure tridimensionnelle des protéines est essentielle pour comprendre leur fonction. Devant la complexité de l'obtention expérimentale de ces informations, la mise au point de méthodes permettant leur prédiction à partir des séquences en acides aminés est un défi majeur. Par ailleurs, dans ce contexte, l'étude des boucles est particulièrement importante car leur rôle fonctionnel est prépondérant et leur structure tridimensionnelle particulièrement variable. Nous proposons une méthode de prédiction reposant sur l'encodage des structures 3D par alphabet structural. Cette méthode comporte une première étape de prédiction locale d'une lettre structurale à partir de quatre acides aminés par une méthode d'arbre. La prédiction est ensuite affinée en tenant compte des informations telles que les ressemblances entre lettres structurales et la succession non aléatoire des lettres dans les séquences. Un modèle de Markov caché est utilisé dans ce but. La méthode sera testée sur une large base de données de boucles contenant un jeu d'apprentissage ainsi qu'un jeu de validation.

Mots clés : boucles protéiques, prédiction de structure 3D, alphabet structural, algorithme génétique, modèle de Markov caché.

Abstract

Knowledge about tridimensional structure of proteins is of prime importance in understanding their fonction. However, in view of the complexity of their experimental obtaining, the design of prediction methods from amino acid sequences is really challenging. Moreover, in such a context, studying loops is particularly interesting because they play an important part in function and their structure is very variable. A method relying on 3D structure encoded by a structural alphabet is proposed. This method consists of a first tree-based prediction step of local structural letters from a sequence of four amino acid

sequences. The prediction is then refined by taking into account through a hidden Markov model, similarities between structural letters and the non random succession of letters in sequences. The method will be applied on a large loop dataset divided into learning and validation sets.
Keywords : protein loops, 3D structure prediction, structural alphabet, genetic algorithm, hidden Markov model.

1 Introduction

La forme tridimensionnelle (3D) des protéines est la forme active qui permet leur caractérisation fonctionnelle. Cette forme peut être déterminée expérimentalement (par cristallographie par exemple). Actuellement, les bases de données fournissent les structures 3D de très nombreuses protéines. Cependant, la possibilité de prédire la structure 3D à partir de la séquence en acides aminés (AA) d'une protéine pourrait permettre de la comparer à des protéines connues sans avoir à passer par la phase expérimentale.
Actuellement, il est déjà courant de prédire les structures bidimensionnelles (2D), c'est-à-dire identifier les zones de la protéine organisées en feuillets β, hélices α ou boucles. Alors que les autres structures secondaires ont une organisation dans l'espace relativement stable, les boucles sont particulièrement variables. De plus, elles sont très souvent impliquées dans la fonction de la protéine. La connaissance de leur structure 3D est donc particulièrement importante. Par exemple, dans le domaine du *drug design* (conception de médicaments) et spécialement du criblage de molécules, la connaissance de la conformation tridimensionnelle d'une zone d'intérêt (une poche susceptible de recevoir un ligand par exemple) peut permettre de sélectionner ou d'écarter des molécules candidates.
Afin de faciliter l'apprentissage, un alphabet structural a été construit. Il permet d'encoder les structures 3D sous une forme unidimensionnelle (1D). Cet alphabet repose sur l'utilisation de 27 lettres structurales, chacune étant représentative de la structure 3D de quatre AA successifs.
La méthode proposée ici se décomposera en deux parties. Tout d'abord, une étape de construction d'une méthode de discrimination permet de prédire, à partir de la séquence de quatre AA consécutifs, la lettre structurale correspondante. Cependant, il est très difficile d'obtenir un taux de bien classés important en utilisant cette seule information. C'est pourquoi, nous utilisons une matrice décrivant les risques de confusion entre lettres structurales ainsi que des matrices décrivant les probabilités de transition entre lettres successives. Ces matrices sont utilisées dans le cadre d'un modèle de Markov caché pour améliorer la qualité de la prédiction.

2 Encodage de la structure 3D par alphabet structural

Afin de décrire la structure 3D d'une chaîne protéique et plus précisément de son squelette carboné, on peut utiliser un alphabet structural (Camproux *et al.*, 2004, Camproux et Tufféry, 2005, Etchbest *et al.*, 2005). Pour cela, on peut décrire l'agencement des atomes dans l'espace à partir des distances entre carbones α. Dans cette étude, nous nous basons sur l'alphabet construit par Camproux *et al.* (2004). Pour chaque groupe de quatre atomes de carbone α consécutifs, on mesure quatre valeurs, (d_1, d_2, d_3, p_4). Les trois premières valeurs correspondent aux distances entre carbones non consécutifs et la quatrième à la valeur algébrique de la projection du dernier carbone α sur le plan formé par les trois premiers (cette valeur est positive si le dernier carbone est au-dessus du plan formé par les trois premiers et négative dans le cas contraire). Ces valeurs sont représentées dans la Fig. 1.
Pour identifier l'alphabet structural, un modèle de Markov caché a été construit dans lequel les états cachés correspondent aux lettres structurales et les états observés sont les vecteurs (d_1, d_2, d_3, p_4). Différents paramètres du modèle ont dû être estimés : le nombre d'états cachés, les fonctions reliant les états cachés aux états observés, les probabilités initiales des états cachés et les probabilités de transition entre états cachés. Pour un nombre d'états cachés donné (c'est-à-dire pour un modèle de dimension donnée), les autres paramètres ont été optimisés par maximum de vraisemblance. Pour choisir entre les modèles de différentes dimensions, le critère BIC (Bayesian Information Criterion) a été employé (Hastie *et al.*, 2003).
L'alphabet structural optimal (obtenu sur deux jeux de données indépendants), basé sur ce BIC, est un modèle à 27 lettres structurales : $(a, A, B, C, \ldots, X, Y, Z)$. Dans une première approche, ces lettres ont été reliées aux structures 2D. Il en résulte que les lettres (a, A, V, W) sont plutôt caractéristiques des hélices α, les lettres (L, N, M, T, X) sont plus souvent observées dans des feuillets β, les lettres (Z, B, C, J, T) marquent plus spécifiquement des flancs entre boucles et hélices α ou boucles et feuillets β. Enfin, les autres lettres sont plus spécifiques aux boucles.
Des études préalables (Regad *et al.*, 2006) tendent à montrer que des mots de quatre lettres structurales consécutives (obtenues donc sur sept carbones α consécutifs) sont bien adaptés pour décrire avec précision les chaînes peptidiques, et particulièrement les boucles. En effet, si on mesure l'écart moyen entre carbones α pour des fragments structuraux ayant le même encodage dans cet alphabet, on trouve un écart moyen inférieur à 1 Å. C'est pourquoi nous nous basons, dans la suite de ce travail, sur l'extraction de mots de quatre lettres structurales dans les zones identifiées comme boucles. Ces études ont également montré que les matrices de transition entre lettres structurales sont creuses, c'est-à-dire qu'on trouve uniquement un nombre limité de trajectoires entre lettres. Cette observation sera particulièrement importante dans la suite de ce travail.

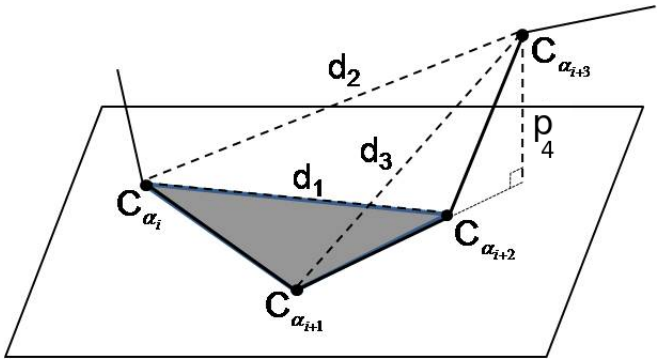

FIG. 1 – Représentation graphique des valeurs mesurées pour décrire la structure 3D de quatre carbones consécutifs.

3 Prédiction des lettres structurales individuelles

Nous travaillons sur une banque de données de 3186 protéines, appelée PDB50 (2008) et correspondant aux chaînes protéiques présentant moins de 50% d'identités. Ceci assure une hétérogénéité et donc, une certaine représentativité des protéines présentes. La PDB50 est composée de chaînes protéiques d'au moins 30 résidus, obtenues par diffraction aux rayons X (résolution supérieure à 2.5 Å). Les protéines présentant des ambiguïtés sur les coordonnées (résidus manquants, conformations alternatives,...) ont été écartées. Ces données ont été utilisées pour réaliser l'encodage en lettres structurales à l'aide du modèle de Markov caché décrit précédemment. Parallèlement, on dispose, pour les mêmes protéines, de leur séquence en acides aminés. Nous sommes donc en mesure, pour ces données, de mettre en parallèle, les informations de séquence en acides aminés et de structure 3D par l'intermédiaire du codage par alphabet structural.
Dans ce travail, afin de travailler sur données fiables, nous considérons uniquement les mots de quatre lettres structurales vus plus de 20 fois dans la base de données. On obtient alors 5103 mots différents, présents, au total, en 292767 occurrences.
Dans une première étape, les données ont été condensées en une matrice de dimension 5103×140 pour les acides aminés. En fait, on a 7×20 colonnes. C'est une table de contingence comptant, pour chaque mot, le nombre de chacun des 20 acides aminés à chacune des 7 positions. Afin d'éviter un poids trop important des mots présents en de nombreuses occurrences, les effectifs sont divisés par le nombre d'occurrences du mot. On obtient donc, pour chacun des 5103 mots, une sorte de *profil* en terme d'acides aminés.
Une première approche peut consister à appliquer l'Analyse Factorielle Discriminante au sens de Fisher (AFD, Hastie *et al.*, 2003) directement sur ces données. Pour s'affranchir du problème des données de contingence, une transformation $log(1+x)$ a été appliquée. Un modèle d'AFD a été construit pour chacune des quatre lettres structurales. Les premiers résultats sont très encourageants puisqu'ils permettent, en description, d'obtenir environ

55 % de bonnes classifications. Ces résultats sont évidemment provisoires et partiels mais ils seront rapidement précisés à l'aide de validation croisée d'une part et d'un échantillon de validation d'autre part. Toutefois, ces premiers résultats nous donnent déjà des indications sur les lettres qu'il est plus ou moins facile de prédire ainsi que sur les confusions les plus probables entre lettres.
L'inconvénient de cette approche est la généralisation destinée à prédire les lettres pour une séquence donnée. Dans un tel cas, au lieu des pourcentages que l'on avait obtenu, après normalisation dans les données de l'AFD, on se trouve en présence uniquement de 0 et de 1, ce qui a tendance à projeter les nouveaux points en dehors de l'espace défini par les données d'apprentissage. Il est donc nécessaire de proposer une méthode de discrimination travaillant directement sur les données binaires.
Une nouvelle méthode sera donc présentée. Elle repose sur un principe de classification par arbre de décision dans lequel chaque nœud repose uniquement sur la valeur d'un des quatre AA. Chaque AA ayant vingt modalités possibles et afin d'éviter une trop grande complexité de l'arbre (et donc des problèmes de généralisation), les différents acides aminés trouvés dans une position sont regroupés. On prend donc des décisions en fonction de l'appartenance ou non de l'acide aminé considéré à l'un des groupes construits. Les différents paramètres de cette méthode sont estimés à l'aide d'un algorithme génétique. Les détails de la méthode ainsi que ses résultats seront donnés lors de la présentation orale.

4 Amélioration de la prédiction par utilisation d'un modèle de Markov caché

La méthode de prédiction se heurte à plusieurs problèmes. Tout d'abord, il existe certaines ressemblances entre lettres qui peut amener le modèle à faire certaines confusions de manière récurrente. De plus, il arrive qu'une même séquence de quatre AA conduise à deux lettres structurales différentes. Cependant, ce dernier problème peut être en partie réglé par la prise en compte de l'environnement de la lettre considérée. En effet, il a été observé (Camproux *et al.*, 2004, Regad *et al.*, 2006) que les lettres successives ne sont pas indépendantes. Le fait d'obtenir une lettre donnée pour une position est fortement lié aux lettres observées dans les positions adjacentes.
Toutes ces observations nous ont conduits à chercher à construire un modèle de Markov caché (Cappé *et al.*, 2005) dont les états cachés seraient les *vraies* lettres structurales et les états observés, les lettres prédites par la méthode de discrimination. Cette façon de procéder est assez proche de l'utilisation qui est faite des modèles de Markov cachés dans la reconnaissance de caractères (Premaratne *et al.*, 2006). Pour estimer les probabilités de transition entre états cachés, l'ensemble de la base de données (mots présents au moins une fois) a été utilisée. Par ailleurs, le lien entre états cachés (vraies lettres) et états observés (lettres prédites) est modélisé par la matrice de confusion entre lettres. Cette

matrice provient d'études antérieures (Guyon *et al.*, 2004).
La correction des prédictions par le modèle de Markov caché obtenu est alors effectuée en utilisant l'algorithme de Viterbi (Cappé *et al.*, 2005). Les résultats sur la base de donnée considérée seront montrés lors de la présentation orale.

5 Conclusion

La prédiction de la structure 3D des protéines à partir de la séquence en AA est un enjeu majeur dans la compréhension de la fonction des nouvelles séquences obtenues chaque jour dans le cadre des programmes de séquençage massif. L'utilisation d'un alphabet structural a déjà permis de simplifier significativement l'information relative à cette structure 3D tout en conservant une précision satisfaisante. La prédiction des lettres structurales à partir des séquences d'AA est donc un bon moyen de faire le lien entre séquences 1D et 3D. Cependant, ces données révèlent de nombreuses difficultés et imposent l'utilisation de méthodes de prédiction spécifiques et de méthodes de correction a posteriori qui tiennent compte de l'ensemble des connaissances acquises sur les lettres structurales.

Bibliographie

[1] Camproux, A.-C., Gautier, R. et Tufféry, P. (2004) A hidden Markov model derived structural alphabet for proteins. *Journal of Molecular Biology*, **339**, 591-605.
[2] Camproux, A.-C. et Tufféry (2005) Hidden Markov Model-derived structural alphabet proteins : the learning of protein local shapes captures sequence specificity. *Biochimica et Biophysica Acta*, **1724**, 394-403.
[3] Cappé, O., Moulines, E., Rydén, T. (2005), *Inference in Hidden Markov Models*, Springer.
[4] Etchebest, C., Benros, C., Hazout, S. et de Brevern A.G. (2005) A structural alphabet for local protein structures : improved prediction methods. *Proteins*, **58**, 810-827.
[5] Guyon, F., Camproux, A.-C., Hochez, J. et Tufféry, P. (2004) SA-Search : a web tool for protein structure mining based on a Structural Alphabet. *Nucleic Acids Research*, **32**, 545-548.
[6] Hastie, T., Tibshirani, R. & Friedman, J. (2003) *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, Springer, New York.
[7] Premaratne, H.L., Järpe, E. et Bigun, J. (2006) Lexicon and hidden Markov model-based optimisation of the recognised Sinhala script. *Pattern Recognition Letters*, **27**, 696-705.
[8] Regad, L., Martin, J. et Camproux, A.-C. (2006) Identification of non random motifs in loops using a structural alphabet. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 1-9.