Méthode hiérarchique hybride de prédiction des angles de torsion ω des protéines

Yann Guermeur*, Thérèse E. Malliavin **

*LORIA Equipe ABC, Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy cedex
yann.guermeur@loria.fr
https://members.loria.fr/YGuermeur/
**Laboratoire de Physique et Chimie Théoriques, UMR CNRS 7019
therese.malliavin@univ-lorraine.fr

Résumé. Nous abordons un problème ouvert de biologie structurale, la prédiction des angles de torsion ω des protéines à partir de l'information locale de séquence, comme un problème de discrimination à catégories multiples. L'approche hiérarchique mise en œuvre pour le traiter combine des machines à vecteurs support multi-classes et un perceptron multi-couche au moyen d'une méthode d'ensemble linéaire. L'étude comparée du comportement de l'architecture globale et de ses composantes permet deux conclusions. D'une part, la séquence locale contient de l'information exploitable par notre méthode de prédiction. D'autre part, le taux de reconnaissance apparaît fortement corrélé à la qualité de l'estimation des probabilités a posteriori des catégories.

1 Introduction

La détermination des structures tridimensionnelles des protéines, c'est-à-dire la position relative de leurs atomes, est essentielle pour comprendre leur fonction. Cela fait environ 30 ans que la communauté de biologie structurale développe des méthodes prédisant ces structures à partir de la séquence d'acides aminés. Un cadre général pour ces prédictions se fonde sur l'exploitation de deux types de descripteurs : la structure locale et les contacts à longue distance, ces derniers étant inférés à partir d'alignements multiples de séquences. De telles approches sont confrontées à des limitations statistiques lorsque les alignements sont constitués d'un trop petit nombre de séquences (Radjasandirane et de Brevern, 2023). De plus, les régions désordonnées des protéines ne relèvent pas de ce cadre car les contacts à longue distance n'y sont pas définis. L'efficacité de la reconstruction de la structure d'une protéine à partir de la seule connaissance de la structure tridimensinnelle locale a récemment été étudiée par Hengeveld et al. (2023). L'approche retenue était la géométrie des distances (Lavor et al., 2012; Worley et al., 2018). Ces travaux ont établi que les variations de stéréochimie avaient un impact important sur les résultats. C'est particulièrement le cas des variations des mesures des angles de torsion ω.

Cette communication évalue la possibilité de prédire les valeurs des angles ω à partir de l'information locale uniquement. Elle traite le problème comme un problème de discrimination

à catégories multiples dont les descriptions correspondent au contenu d'une fenêtre d'analyse glissant sur la séquence. Deux types de systèmes discriminants sont mis en œuvre pour la prédiction initiale : des machines à vecteurs support multi-classes (M-SVM) et un perceptron multi-couche (PMC). Les sorties des M-SVM sont post-traitées afin de fournir, à l'image de celles du PMC, des estimations des probabilités a posteriori (p.a.p.) des catégories. Toutes les estimations sont ensuite présentées en entrée d'une méthode d'ensemble linéaire (LEM) qui fournit la prédiction finale.

L'organisation de l'article est la suivante. La section 2 formule le problème de biologie structurale comme un problème de discrimination. La présentation et l'évaluation des classifieurs de base (experts) fait l'objet de la section 3. La combinaison de modèles est évaluée dans la section 4. Enfin, la section 5 est consacrée aux conclusions et au travail en cours.

2 Caractérisation du problème de classification

Cette section reformule un problème de biologie structurale ouvert comme un problème de classification supervisée.

2.1 Problème de biologie structurale

Les protéines sont des polymères d'acides aminés, et chaque acide aminé contient les atomes lourds N, C_α et C définissant le squelette de la protéine, ainsi que, pour chaque résidu (d'acide aminé) sauf la Proline, l'atome d'hydrogène amide HN lié de manière covalente à l'azote N. Pour le résidu d'indice i dans la séquence, l'angle de torsion ω est l'angle dièdre défini par les atomes C_α et C du résidu d'indice $i-1$ et N et C_α du résidu d'indice i. Il se caractérise aussi comme l'angle de torsion définissant la planarité des atomes C, O, N et HN. La prédiction de sa mesure à partir de l'information locale (dans la séquence) est un problème ouvert pour lequel il n'existe pas d'état de l'art.

2.2 Problème d'inférence empirique

Les quatre catégories du problème de discrimination sont déterminées en fonction de la valeur du paramètre $\delta\omega$ défini par :

$$\delta\omega = \text{sign}(\omega)180^\circ - \omega.$$

Elles sont associées à des seuils comme suit :

$$\begin{cases} \delta\omega \leqslant -3^\circ & \text{: classe 1} \\ -3^\circ < \delta\omega \leqslant 0^\circ & \text{: classe 2} \\ 0^\circ < \delta\omega \leqslant 4^\circ & \text{: classe 3} \\ 4^\circ < \delta\omega & \text{: classe 4} \end{cases}.$$

Compte tenu du contexte biologique dans lequel s'inscrit cette étude, les descriptions sont issues de la séquence seule et non d'un alignement multiple. La description associée à l'angle ω d'un résidu est l'heptapeptide centré sur ce résidu.

Les données utilisées correspondent à 1470 structures de protéines issues de la Protein Data Bank (Berman et al., 2000). Ces structures, déterminées par cristallographie aux rayons X, contiennent une seule chaîne. Dans chaque structure, seule la chaîne A a été retenue, afin d'éviter les biais induits par des homodimères. La résolution des structures est de 0.48 à 1.5 Å et les facteurs R sont tous meilleurs que 0.26. Le nombre d'acides aminés (descriptions) correspondant est de 89033. La classe la plus représentée, la troisième, rassemble approximativement un tiers des exemples.

3 Systèmes discriminants de base

Les experts utilisés dans cette étude sont des classifieurs à marge. Les M-SVM s'appuient sur le concept de marge géométrique tandis que le PMC, estimant les p.a.p. des catégories, est un classifieur à marge analytique.

3.1 M-SVM et PMC

Trois M-SVM sont mises en œuvre. En suivant l'ordre chronologique, il s'agit des modèles de Weston et Watkins (1998), Lee et al. (2004) et Guermeur et Monfrini (2011). Dans la suite, elles sont respectivement nommées WW-M-SVM, LLW-M-SVM et M-SVM^2. Elles utilisent le même noyau gaussien, incorporant de l'information sur le problème d'intérêt à travers le produit scalaire qu'il applique aux acides aminés (résidus). La matrice de ce produit scalaire est obtenue suivant la méthode décrite par Guermeur et al. (2004). Précisément, elle résulte de la projection de la matrice de substitution BLOSUM62 (Henikoff et Henikoff, 1989) sur le cône convexe des matrices symétriques semi-définies positives. Les sorties des machines sont post-traitées au moyen d'un modèle de régression logistique polytomique (PLR) (Hosmer et Lemeshow, 1989) de manière à produire des estimations des p.a.p. des catégories.

Le PMC (Anthony et Bartlett, 1999) utilisé se caractérise par l'emploi de l'exponentielle normalisée comme fonction d'activation des unités de sortie et de l'entropie croisée comme fonction de perte. Cette combinaison est ordinairement reconnue comme fournissant de meilleures estimations des p.a.p. des catégories que la combinaison sigmoïde et coût quadratique.

3.2 Evaluation des modèles

Les modèles décrits ci-dessus sont évalués dans le cadre d'une validation croisée à sept pas. Pour un ensemble de test fixé, le PMC utilise l'ensemble des données d'apprentissage pour son entraînement. Les M-SVM n'utilisent que les cinq premiers sous-ensembles de ce jeu pour leur apprentissage, le dernier servant à l'entraînement de la PLR. Les performances obtenues sont regroupées dans le tableau 1. Dans ce tableau, $\%$ rec. désigne le taux de reconnaissance, les C_k représentent les coefficients de corrélation de Pearson-Matthews (Matthews, 1975) et CE désigne l'entropie croisée moyenne (rapportée au nombre d'exemples).

Tous les experts sont nettement meilleurs que le classifieur naïf assignant l'ensemble des exemples à la catégorie la plus représentée (la catégorie 3). Cette supériorité est soulignée par la positivité de l'ensemble des coefficients de Pearson. Le test de comparaison de deux pourcentages établit que la supériorité de chacune des M-SVM sur le PMC est statistiquement

Classifieur	% rec.	C_1	C_2	C_3	C_4	CE
WW-M-SVM + PLR	49.3	0.41	0.19	0.24	0.39	1.14
M-SVM^2 + PLR	48.9	0.40	0.19	0.23	0.38	1.15
LLW-M-SVM + PLR	48.7	0.39	0.19	0.22	0.37	1.15
PMC	44.8	0.33	0.14	0.17	0.32	1.27

TAB. 1: Performances comparées des classifieurs de base.

significative avec une confiance supérieure à 0.95. Le gain s'explique entièrement par l'utilisation du noyau dédié. Les performances des trois M-SVM sont très similaires. La combinaison d'experts doit fournir une indication sur leur complémentarité.

4 Méthode d'ensemble

Dans cette section, nous considérons un problème de discrimination à C catégories pour lequel on dispose de N classifieurs de base (experts). Le classifieur d'indice j calcule la fonction (à valeurs vectorielles) $g^{(j)} = \left(g_k^{(j)}\right)_{1 \leqslant k \leqslant C}$. Pour tout $n \in \mathbb{N}^*$, soit U_n le $(n-1)$-simplexe unité défini par $U_n = \left\{u = (u_p)_{1 \leqslant p \leqslant n} \in \mathbb{R}_+^n : \sum_{p=1}^n u_p = 1\right\}$. Chaque modèle est supposé prendre ses valeurs dans U_C (une hypothèse évidemment vérifiée par les classifieurs de la section 3).

4.1 LEM

La méthode d'ensemble linéaire (Guermeur, 2013) considérée s'appuie sur le modèle de régression linéaire multivariée donné par :

$$\forall k \in [\![1; C]\!],\ \forall x \in \mathcal{X},\ \ g_{\boldsymbol{\theta},k}(x) = \sum_{j=1}^{N} \theta_j \sum_{l=1}^{C} \theta_{kjl} g_l^{(j)}(x)$$

avec

$$\begin{cases} (\theta_j)_{1 \leqslant j \leqslant N} \in U_N \\ \forall\, (j,l) \in [\![1; N]\!] \times [\![1; C]\!],\ \ (\theta_{kjl})_{1 \leqslant k \leqslant C} \in U_C \end{cases}.$$

Ce modèle étend la combinaison convexe, obtenue en posant tous les coefficients θ_{kjl} égaux à $\delta_{k,l}$ où δ est le symbole de Kronecker. Notons $V_{C,N}$ le polytope convexe auquel appartient le vecteur $\boldsymbol{\theta} = (\theta_j \theta_{kjl})$ des paramètres. L'entraînement de la LEM sur le jeu de données $((x_i, y_i))_{1 \leqslant i \leqslant m}$ correspond à la résolution du problème de programmation convexe suivant :

Problème 1

$$\min_{\boldsymbol{\theta} \in V_{C,N}} \left\{ -\sum_{i=1}^{m} \ln\left(g_{\boldsymbol{\theta},y_i}\left(x_i\right)\right) \right\}.$$

La consistance de ce principe inférentiel est caractérisée par la proposition 3.4 de (Guermeur, 2013).

4.2 Evaluation de la combinaison

L'évaluation de la LEM s'effectue de nouveau au moyen d'une validation croisée à sept pas. La décomposition de la base d'apprentissage est modifiée de manière à fournir des données pour l'apprentissage supplémentaire. Ainsi, pour un jeu de test fixé, les quatres premiers sous-ensembles du jeu d'apprentissage sont utilisés pour entraîner les M-SVM, le cinquième est utilisé pour entraîner la PLR et le sixième pour entraîner la combinaison. Le PMC est entraîné sur les cinq premiers sous-ensembles du jeu d'apprentissage. Les performances des experts de base et de leur combinaison sont regroupées dans le tableau 2.

Classifieur	$\%$ rec.	C_1	C_2	C_3	C_4	CE
LEM	49.8	0.41	0.20	0.24	0.39	1.13
M-SVM^2 + PLR	48.7	0.39	0.19	0.22	0.37	1.16
WW-M-SVM + PLR	48.5	0.40	0.17	0.23	0.39	1.15
LLW-M-SVM + PLR	48.4	0.39	0.18	0.22	0.37	1.16
PMC	44.7	0.32	0.13	0.17	0.31	1.28

TAB. 2: Performances comparées de la méthode d'ensemble et des classifieurs de base.

La combinaison de modèles apporte un accroissement du taux de reconnaissance de 0.5% par rapport au meilleur des experts considéré individuellement (voir le tableau 1). Ce faible gain s'explique par l'observation suivante : la taille du jeu de données utilisé ne permet pas de multiplier les apprentissages. Précisément, diminuer la taille de l'ensemble d'apprentissage des experts pour permettre l'apprentissage de la LEM a un coût. La combinaison obtenue est en fait une combinaison convexe dans laquelle seules les trois M-SVM ont un poids non nul (le PMC ne contribue pas).

5 Conclusions et travail en cours

Cette étude a principalement établi un résultat qualitatif : l'existence au sein de la séquence protéique d'une information locale sur la valeur des angles de torsion ω. L'exploitation de cette information au moyen d'une méthode de prédiction combinant des SVM multi-classes et un PMC fournit déjà des performances prometteuses. Elle s'appuie sur une forte dépendance entre le taux de reconnaissance et la valeur de l'entropie croisée.

Le travail en cours porte sur une sélection des descripteurs exploitant plus efficacement l'information locale et une extension de la méthode de prédiction tirant un meilleur parti de la nature séquentielle des données. Il profite de la disponibilité d'un jeu de données nettement plus grand.

Références

Anthony, M. et P. Bartlett (1999). *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, Cambridge.

Berman, H., J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, et P. Bourne (2000). The protein data bank. *Nucleic Acids Research 28*, 235–242.

Guermeur, Y. (2013). Combining multi-class SVMs with linear ensemble methods that estimate the class posterior probabilities. *Communications in Statistics - Theory and Methods 42*(16), 3011–3030.

Guermeur, Y., A. Lifchitz, et R. Vert (2004). A kernel for protein secondary structure prediction. In B. Schölkopf, K. Tsuda, et J.-P. Vert (Eds.), *Kernel Methods in Computational Biology*, Chapter 9, pp. 193–206. The MIT Press, Cambridge, MA.

Guermeur, Y. et E. Monfrini (2011). A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica 22*(1), 73–96.

Hengeveld, S., M. Merabti, F. Pascale, et T. Malliavin (2023). A study on the covalent geometry of proteins and its impact on distance geometry. In *6th International Conference on Geometric Science of Information (GSI'23)*, pp. 520–530.

Henikoff, S. et J. Henikoff (1989). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A 89*(22), 10915–10919.

Hosmer, D. et S. Lemeshow (1989). *Applied Logistic Regression*. Wiley, London.

Lavor, C., L. Liberti, N. Maculan, et A. Mucherino (2012). The discretizable molecular distance geometry problem. *Computational Optimization and Applications 52*, 115–146.

Lee, Y., Y. Lin, et G. Wahba (2004). Multicategory support vector machines : Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association 99*(465), 67–81.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta 405*, 442–451.

Radjasandirane, R. et A. de Brevern (2023). Structural and Dynamic Differences between Calreticulin Mutants Associated with Essential Thrombocythemia. *Biomolecules 13*(3), 509–533.

Weston, J. et C. Watkins (1998). Multi-class support vector machines. Technical Report CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science.

Worley, B., F. Delhommel, F. Cordier, T. Malliavin, B. Bardiaux, N. Wolff, M. Nilges, C. Lavor, et L. Liberti (2018). Tuning interval Branch-and-Prune for protein structure determination. *Journal of Global Optimization 72*, 109–127.

Summary

We address an open problem in structural biology, the prediction of the torsion angles ω of the proteins from local information coming from sequence, as a multi-category pattern classification problem. The hierarchical approach implemented to deal with it combines multi-class support vector machines and a multi-layer perceptron by means of a linear ensemble method. The comparative study of the behaviour of the global architecture and its components makes it possible to draw two conclusions. On the one hand, our prediction method can make use of the knowledge of the local sequence. On the other hand, the recognition rate appears highly correlated to the quality of the class posterior probability estimates.