New cascade architecture for protein secondary structure prediction

Yann GUERMEUR and Fabienne THOMARAT

LORIA, UMR7503 CNRS, Campus Scientifique, BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
{Yann.Guermeur, Fabienne.Thomarat}@loria.fr

Keywords Protein secondary structure prediction, M-SVMs, ensemble methods.

Prédiction de la structure secondaire des protéines : mise en œuvre optimisée de l'architecture en cascade

Mots-clefs Prédiction de la structure secondaire des protéines, M-SVM, combinaison de modèles.

1 Introduction

La prédiction du repliement des protéines à partir de leur séquence (*ab initio*) est un problème majeur en biologie structurale. Du fait de sa difficulté, elle est ordinairement réalisée par étapes, l'une des principales étant la prédiction d'éléments structuraux périodiques locaux reliés entre eux par des éléments apériodiques. Ces éléments constituent la structure secondaire. La prédiction de cette structure consiste à affecter aux résidus d'une séquence protéique leur état conformationnel : hélice α, brin β ou apériodique (coil).

Les méthodes statistiques fournissent actuellement les meilleures prédictions. L'architecture de base sur laquelle elles s'appuient est pratiquement toujours la même : deux classifieurs sont utilisés en cascade. Le premier, dit "séquence-structure", effectue la prédiction de la structure à partir de la séquence (ou d'un alignement multiple), le second, dit "structure-structure", lissant la prédiction initiale en exploitant la corrélation des états conformationnels des résidus consécutifs [5,6,7,8]. Cette brique de base est incorporée dans des combinaisons prenant des formes variées, ce qui soulève des questions relevant de la sélection de modèles. Ainsi, la méthode décrite dans [7] combine plusieurs centaines de perceptrons multicouches (PMC), tandis que les auteurs de la méthode Jnet [2] ont récemment fait passer la taille de la couche cachée des PMC qu'ils utilisent de neuf à cent.

Nous proposons une version optimisée de l'architecture en cascade. Ses performances sont proches de l'état de l'art, pour une complexité en échantillon inférieure d'au moins un ordre de grandeur. Les sorties sont des estimations des probabilités a posteriori (p.a.p.) des catégories, ce qui autorise leur post-traitement par des modèles génératifs. Le premier niveau est constitué par les quatre principales machines à vecteurs support multi-classes (M-SVM) proposées dans la littérature, le second est une combinaison linéaire multivariée des sorties de ces machines.

2 Architecture en cascade

Nous donnons une définition générale des modèles impliqués dans l'architecture pour un problème de discrimination caractérisé par une mesure de probabilité P sur l'espace produit $\mathcal{Z}$ d'un espace de description $\mathcal{X}$ et d'un ensemble de catégories $\mathcal{Y} = [\![1, Q]\!]$.

2.1 Prédiction séquence-structure

Soient κ un noyau de Mercer sur $\mathcal{X}^2$ et H_κ l'espace de Hilbert à noyau reproduisant qu'il engendre. Soit $\mathcal{H} = (H_\kappa + \{1\})^Q$. Pour toute fonction $h = (h_k)_{1 \leqslant k \leqslant Q} \in \mathcal{H}$, $\bar{h} \in \bar{\mathcal{H}} = H_\kappa^Q$ est sa partie linéaire.

DÉFINITION 2.1. *Soient* $((x_i, y_i))_{1 \leqslant i \leqslant m} \in \mathcal{Z}^m$, $\lambda \in \mathbb{R}_+^*$ *et* $\xi \in \mathbb{R}^{Qm}$ *tel que pour* $(i,k) \in [\![1,m]\!] \times [\![1,Q]\!]$, ξ_{ik} *est sa composante d'indice* $(i-1)Q + k$, *avec* $(\xi_{iy_i})_{1 \leqslant i \leqslant m} = 0_m$. *Une M-SVM à* Q *catégories est un classifieur obtenu en résolvant un problème de programmation quadratique de la forme*

$$\min_{h,\xi} \left\{ \|M\xi\|_p^p + \lambda \left\|\bar{h}\right\|_{\bar{\mathcal{H}}}^2 \right\}$$

$$s.c. \begin{cases} \forall i, \forall k \neq y_i, K_1 h_{y_i}(x_i) - h_k(x_i) \geqslant K_2 - \xi_{ik} \\ \forall i, \forall (k,l) \in ([\![1,Q]\!] \setminus \{y_i\})^2, K_3 (\xi_{ik} - \xi_{il}) = 0 \\ \forall i, \forall k \neq y_i, K_4 \xi_{ik} \geqslant 0 \\ \sum_{k=1}^Q h_k = 0 \end{cases}$$

où $p \in [\![1,2]\!]$, $(K_1, K_3, K_4) \in [\![0,1]\!]^3$ *et* $K_2 \in \mathbb{R}_+^*$. M *est telle que* $\|M\xi\|_p$ *définisse une norme sur* ξ.

Nous considérons ici les M-SVM de Weston et Watkins (WW), Crammer et Singer (CS) et Lee,

Lin et Wahba (LLW), ainsi que la M-SVM^2 [4]. En notant δ le symbole de Kronecker et N la matrice de dimension Qm de terme général $n_{ik,jl} = (1-\delta_{y_i,k})\left(1-\delta_{y_j,l}\right)\delta_{i,j}\left(\delta_{k,l}+1\right)$, ces machines se caractérisent au moyen du Tab. 1.

M-SVM	M	p	K_1	K_2	K_3	K_4
WW	I_{Qm}	1	1	1	0	1
CS	I_{Qm}	1	1	1	1	1
LLW	I_{Qm}	1	0	$\frac{1}{Q-1}$	0	1
M-SVM^2	$N = M^T M$	2	0	$\frac{1}{Q-1}$	0	0

Tab. 1. Spécifications des quatre M-SVM.

2.2 Prédiction structure-structure

Soient $U_Q = \left\{u = (u_k)_{1\leqslant k\leqslant Q} \in \mathbb{R}_+^Q : \sum_{k=1}^Q u_k = 1\right\}$ et $\left\{g^{(j)} : 1 \leqslant j \leqslant N\right\}$ un ensemble de N classifieurs à valeurs dans U_Q. Le module de prédiction structure-structure s'appuie sur les régressions linéaires

$$\forall x \in \mathcal{X},\ \forall k \in [\![1, Q]\!],\quad g_{\beta_k}(x) = \sum_{j=1}^{N}\sum_{l=1}^{Q} \beta_{kjl} g_l^{(j)}(x)$$

où les vecteurs $\beta_k \in \mathbb{R}^{NQ}$ satisfont :

$$\forall v \in U_Q^N, \left(\beta_k^T\right)_{1\leqslant k\leqslant Q} v \in U_Q. \qquad (1)$$

Soient $\beta = (\beta_k)_{1\leqslant k\leqslant Q}$, $t_k = (\delta_{k,l})_{1\leqslant l\leqslant Q}$ et ℓ_{LEM} une fonction de perte convexe. L'apprentissage du modèle de régression est obtenu en résolvant :

$$\min_{\beta} \sum_{i=1}^{m} \ell_{\text{LEM}}\left(t_{y_i}, (g_{\beta_k}(x_i))_{1\leqslant k\leqslant Q}\right)$$

$$s.c.\ \ (1).$$

L'expression des contraintes (1) en fonction des β_{kjl} et la preuve que le choix pour ℓ_{LEM} du coût quadratique ou de l'entropie croisée produit des sorties tendant vers les p.a.p. des catégories sont données dans [4].

3 Evaluation des performances

Les prédicteurs du traitement séquence-structure (vecteur x) correspondent au contenu d'une fenêtre d'analyse de taille treize glissant sur les lignes de la matrice PSSM [1] associée à la séquence en cours de traitement. Le noyau des M-SVM est un noyau RBF elliptique, pondérant les positions de la fenêtre. Ces poids sont obtenus par alignement noyau-cible. Les sorties des machines sont ramenées dans U^3 ($Q = 3$) par application d'une exponentielle normalisée. La prédiction structure-structure utilise une extension du modèle de régression décrit dans la section 2.2. Celle-ci exploite les sorties post-traitées des quatre M-SVM pour une fenêtre de taille dix-sept. Nous fournissons des résultats expérimentaux obtenus sur la base CB513 [3]. Ils résultent d'une procédure de validation croisée à deux niveaux nommée *stacked generalization*.

Les taux de reconnaissance des M-SVM varient entre 76,0% (LLW) et 76,9% (CS et M-SVM^2). La différence avec le taux d'un PMC, 72,2%, est statistiquement significative avec une confiance >0,95. L'architecture globale atteint un taux de 78,3% (ℓ_{LEM} = coût quadratique), ce qui correspond aux dernières performances annoncées pour PSIPRED [5].

4 Conclusions et perspectives

Une version optimisée de l'architecture en cascade à la base des meilleures méthodes de prédiction de la structure secondaire a été introduite. Ses performances, encourageantes, peuvent facilement être améliorées, en incorporant la prédiction des différents types de ponts, ou celle des angles dièdriques de l'épine dorsale. Notre premier objectif est de dépasser les limitations induites par l'approche locale en post-traitant les sorties au moyen de HMM.

Références

[1] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17) :3389–3402, 1997.

[2] C. Cole, J.D. Barber and G.J. Barton, The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.*, 36 :W197–W201, 2008.

[3] J. Cuff and G. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34(4) :508–519, 1999.

[4] Y. Guermeur, Ensemble methods of appropriate capacity for multi-class support vector machines, *Proceedings of SMTDA'10*, Chania. (accepté).

[5] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2) :195–202, 1999.

[6] M.N. Nguyen and J.C. Rajapakse, Two-stage multi-class support vector machines to protein secondary structure prediction, *Pac Symp Biocomput. 10*, pp. 346–357, 2005.

[7] T.N. Petersen, C. Lundegaard, M. Nielsen, H. Bohr, J. Bohr, S. Brunak, G.P. Gippert and O. Lund, Prediction of protein secondary structure at 80% accuracy. *Proteins*, 41(1) :17–20, 2000.

[8] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol.*, 232(2) :584–599, 1993.