Un protocole de détection d'affinités locales dans les protéines

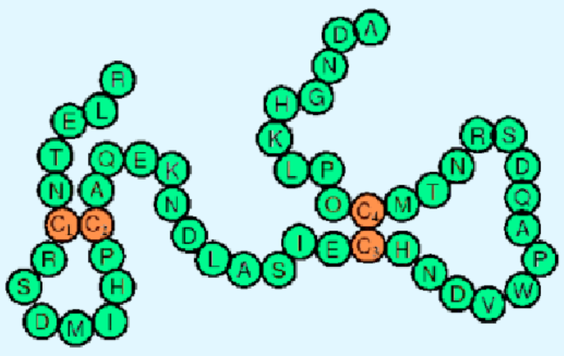
Christophe N. Magnan, Cécile Capponi, François Denis

Laboratoire d'Informatique Fondamentale, UMR CNRS 6166
CMI, 39, rue F. Joliot Curie 13453 Marseille Cedex 13
prénom.nom@lif.univ-mrs.fr

Le contexte

La prédiction de liaisons locales dans les protéines

Ex: ponts disulfures, ponts salins

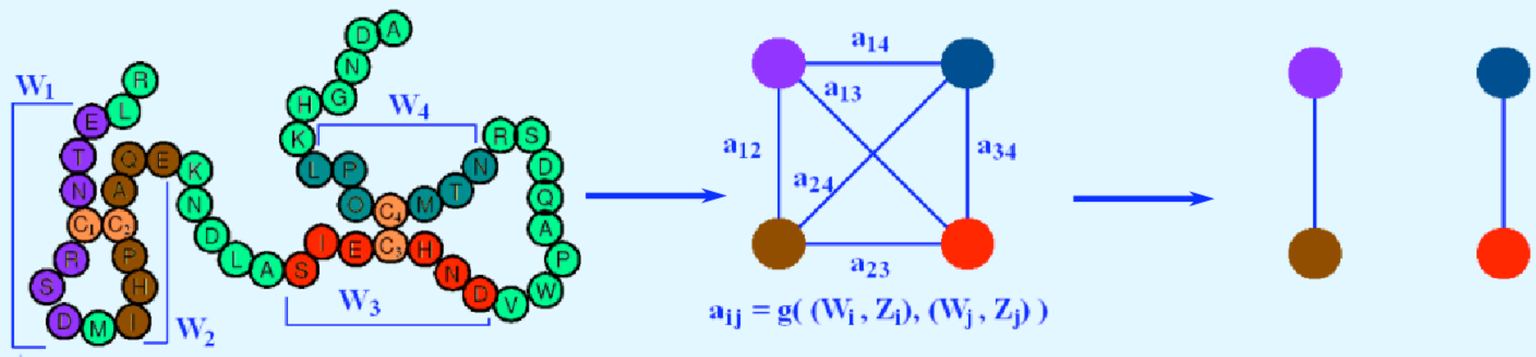


La méthodologie majoritairement utilisée pour résoudre le problème

Utilisation du contexte local comme principal attribut descriptif (Notons Zi les autres attributs)

Apprentissage d'une fonction g modélisant la propension pour deux résidus à former un pont

Sélection des paires appariées correspondant au couplage parfait de poids maximal (polynomial)



Motivations de notre travail

Qualité des prédictions obtenues grâce à cette méthodologie: peu encourageante et insuffisante pour que la méthode soit intégrée sur des serveurs de prédiction de la structure 3D des protéines.

Le point de vue des biologistes: il n'est pas clair que l'information située autour des résidus appariés joue un rôle dans ces appariements qui ne pourraient être qu'une conséquence de la structure.

Ce que nous apportons:

1) Un protocole générique (indépendant du contact) de détection d'une information locale impliquée dans l'appariement de résidus distants sur la séquence.

2) Une expérimentation de ce protocole sur des données de type ponts disulfures et ponts salins avec une adaptation de l'algorithme du perceptron répondant aux critères du protocole.

Un protocole de détection de l'affinité locale

Certains couples (w,w') d'environnements locaux ont-ils plus de chances que d'autres d'être liés par un pont?

Approche triviale: pour une protéine contenant k ponts:

Pas d'information locale <=> $P(B(w,w') \mid w,w',k) = 1/(2k-1)$

Problème: impossible d'estimer $P(B(w,w') \mid w,w',k)$ sans hypothèse supplémentaire.

La solution que nous proposons:

supposer l'existence d'une fonction d'affinité discrète g censée représenter la probabilité que deux environnements locaux w et w' soient liés par un pont et telle que:

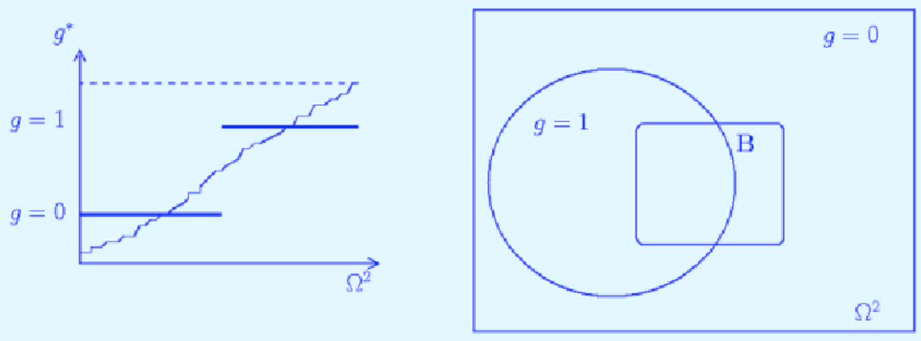
1) $g(w_1,w_2) = g(w_1',w_2') \Rightarrow P(B(w_1,w_2 \mid w_1,w_2,k) = P(B(w_1',w_2') \mid w_1',w_2',k)$

2) $y < y' \Rightarrow P(B(w1,w2) \mid g(w1,w2)=y) < P(B(w1',w2') \mid g(w1',w2')=y')$

Le cas le plus simple est Y = {0,1} et dans ce cas:

$$P(B(w_1,w_2) \mid w_1,w_2,k) = P(B(w_1,w_2) \mid g(w_1,w_2),k) = \begin{cases} \alpha_{1k} \text{ si } g(w_1,w_2) = 1 \\ \alpha_{0k} \text{ si } g(w_1,w_2) = 0 \end{cases}$$

Représentation schématique d'une telle fonction g:



L'observation d'un pont est donc liée de façon non déterministe à la valeur de g.

Les paires d'environnements (w,w') correspondent à des exemples d'étiquette g(w,w') qui auraient subi un **bruit de classification conditionnel à chaque classe (CCCN):**

$$\begin{cases} \eta^+ = 1 - \alpha_{1k} \\ \eta^- = \alpha_{0k} \end{cases}$$

Cette catégorie de bruit (RALAIVOLA et al., 2006) correspond à une généralisation du bruit uniforme CN et dans le cadre PAC, on a CN=CCCN.

Ce résultat permet de chercher g dans toutes les classes de concepts CN-apprenables.

Si on peut apprendre une fonction g telle que α1k > α0k, on aura prouvé l'existence d'une information locale impliquée dans l'appariement des résidus.

Algorithme du perceptron CCCN

Nous avons montré que l'algorithme du perceptron pouvait être adapté pour apprendre à partir de données corrompues par du bruit de classification conditionnel à chaque classe CCCN.

Deux étapes permettent de montrer ce résultat:

1) Bruits η^+ (bruit exemples positifs) et η^- (bruit exemples négatifs) connus

On peut établir un estimateur consistant de la somme des exemples mal classés par le perceptron courant de l'algorithme ne dépendant que des observations, de η^+ et de η^-.

Ce vecteur est un bon vecteur de mise à jour du perceptron et lui permet de converger en un nombre d'itérations de l'ordre du perceptron classique.

2) Bruits η^+ et η^- inconnus

- scanner les différentes valeurs de η^+ et η^- dans l'intervalle [0,1] ($\eta^+ + \eta^- \neq 1$),
- lancer l'algorithme du perceptron CCCN pour chacune des paires de valeurs de η^+ et η^-,
- sélectionner un modèle parmis toutes les hypothèses obtenues.

Le principe de minimisation du risque empirique n'est pas valide dans ce contexte d'apprentissage.

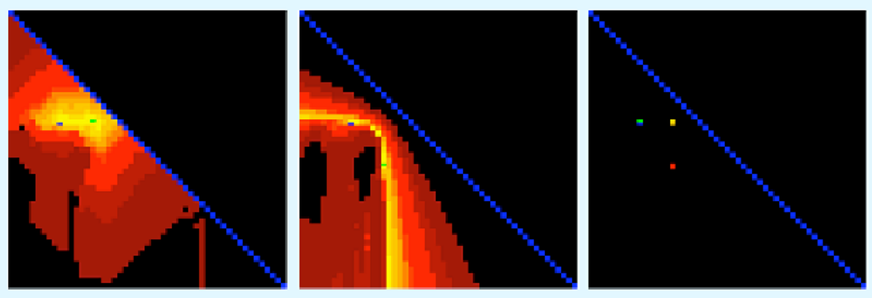
Nous avons montré que le choix du premier modèle qui minimise les deux critères suivants est consistant:

1) la différence absolue entre les valeurs de η^+ et η^- données à l'algorithme en entrée et les valeurs calculées avec l'hyperplan final obtenu sur les données d'apprentissage

2) la norme du dernier vecteur de mise à jour du perceptron (somme des exemples mals classés)

Exemple de sélection (η^+ en abscisse, η^- en ordonnée, cible en bleu foncé, plus le critère considéré est minimal, plus la couleur est claire, le minima est vert clair):

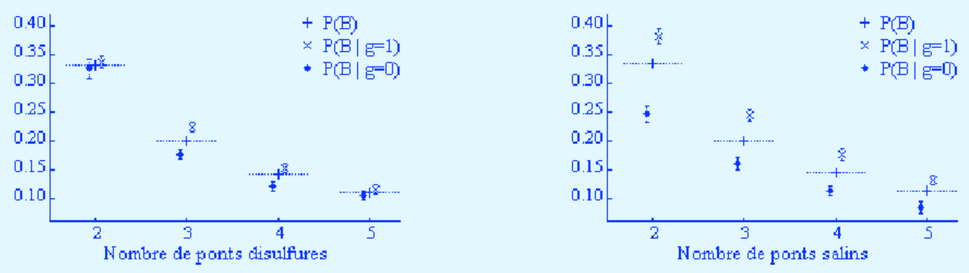
Critère 1 | Critère 2 | Combinaison des deux critères



Expérimentation du protocole sur des données réelles

Expériences sur des jeux de données de protéines contenant:

- des ponts disulfures (jeu SPX)
- des ponts salins (ACI GENOTO3D)



Ponts salins:

- une affinité locale est clairement détectée (forte stabilité)
- les ponts sont majoritairement classés comme des paires possédant un haut niveau d'affinité
- valide le protocole

Ponts disulfures:

- les séparateurs linéaires ne permettent pas de détecter la présence d'une information locale
- plusieurs explications:
  * pas d'information locale contribuant à l'appariement des cystéines
  * la classe des séparateurs linéaires ne permet pas de détecter une affinité
  * l'affinité locale est peut-être insuffisante pour contrarier les ponts disulfures (liaisons fortes)

Discussion:

- protocole validé expérimentalement: permet de détecter une information locale
- protocole générique (indépendant du contact): une aide pour les travaux du domaine
- nécessité d'explorer d'autres classes de concepts

Références

BLUM A., FRIEZE A. M., KANNAN R. & VEMPALA S. (1996). A polynomial-time algorithm for learning noisy linear threshold functions. In *IEEE Symposium on Foundations of Computer Science*, p. 330–338.

BYLANDER (1994). Learning linear threshold functions in the presence of classification noise. In *COLT : Proceedings of the Workshop on Computational Learning Theory*.

FARISELLI P. & CASADIO R. (2001). Prediction of disulfide connectivity in proteins. In *Bioinformatics*, number 17(10), p. 957–964.

FISER A. & SIMON I. (2000). Predicting the oxidation state of cysteines by multiple sequence alignment. In *Bioinformatics, Vol 16, No 3, p.251-256*.

RALAIVOLA L., DENIS F. & MAGNAN C. (2006). Cn = cpcn. In *ICML '06 : Proceedings of the 23rd international conference on Machine learning*, p. 721–728.

ROSENBLATT F. (1962). Principles of neurodynamics. In *Spartan Books*.

VULLO A. & FRASCONI P. (2004). Disulfide connectivity prediction using recursive neural networks and evolutionary information. In *Bioinformatics*, number 20.