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

Résumé : Nous étudions la question de l'existence d'une affinité locale dans les protéines contribuant à l'appariement d'acides aminés distants. Nous proposons un protocole, indépendant du contact considéré, pour répondre à cette question. Nous avons pu expérimenter ce protocole sur des données artificielles grâce à une adaptation de l'algorithme du perceptron répondant aux besoins du protocole. Ces résultats montrent que lorsqu'une telle information locale existe et participe aux contacts, on doit pouvoir la détecter, en estimer la qualité et l'extraire.
Mots-clés : structure 3D des protéines, affinité locale, bruit de classification.

1 Contexte

La prédiction de la structure 3D des protéines est un défi en recherche. Des méthodes permettent déjà une prédiction de la structure secondaire (hélices α et brins β principalement) avec une précision supérieure à 90 %. Parallèlement, des méthodes de prédiction d'interactions entre deux résidus distants d'une protéine ont été proposées. Pour les biologistes et chimistes, il n'est pas clair que le voisinage local (acides aminés voisins sur la séquence) des résidus appariés joue un rôle dans ces appariements ni que ces voisinages contiennent une information utilisable pour les prédire. L'objet de cette étude est de proposer un protocole permettant de répondre à la question de l'existence d'une information locale impliquée dans l'appariement de certains résidus distants.

2 Prédiction d'affinités locales dans les protéines

Une protéine peut se définir par sa structure primaire (mot de Σ^* où Σ est l'ensemble des 20 acides aminés ou un autre alphabet similaire), à partir de laquelle la structure 3D est formée de façon quasi-déterministe (Anfinsen, 1973) dans un milieu donné. Un pont est une liaison (covalente, hydrogène ou électrostatique) entre deux résidus distants de cette séquence. Considérons les environnements locaux des résidus appariés, c'est-à-dire les acides aminés situés autour de ces résidus sur la séquence primaire. Ils sont représentés par des fragments de la protéine, appelés fenêtres, centrés sur les résidus impliqués dans les ponts et de taille $2r+1$ où r est appelé rayon de la fenêtre.

Pour une catégorie de pont donnée (ponts disulfures ou salins, liaisons hydrogène,...), on désignera par $\mathcal{P} \subset \Sigma^*$ (resp. $\mathcal{P}_l \subset \mathcal{P}$) l'ensemble des protéines avec des ponts (resp. l ponts). Soit P une distribution de probabilité sur $\mathcal{P}$ et soit $\Omega_r = \Sigma^{2r+1}$ l'ensemble des fragments potentiels de protéines de rayon r centrés sur les résidus appariés. Une première approche de l'existence d'une information locale impliquée dans le fait qu'il y ait ou non un pont entre deux résidus peut se décrire ainsi : soient e et $e' \in \Omega_r$ les contextes locaux de deux résidus r et r' pris parmi les $2l$ résidus d'une protéine de $\mathcal{P}_l$

susceptibles de former un pont et soit $P(B(e,e')|e,e',l)$ la probabilité d'observer un pont entre r et r'. Remarquons qu'il n'y a aucune raison que les paires (e,e') déterminent à elles seules la présence d'un pont. e et e' ne contiennent pas d'information sur le fait que r et r' forment un pont si et seulement si $P(B(e,e')|e,e',l) = \frac{1}{2l-1}$. Le problème est qu'il n'est pas possible d'estimer $P(B(e,e')|e,e',l)$ sans hypothèse supplémentaire. Pour $r=3$, $|\{(e,e')\in\Omega_r^2\}|=20^{12}\simeq 4.10^{14}$ et seules quelques centaines d'exemples sont disponibles.

La solution que nous proposons est de supposer l'existence d'une *fonction d'affinité discrète* g censée modéliser la propension de deux environnements locaux à former un pont : $g : \Omega_r \text{ x } \Omega_r \rightarrow Y$ (avec Y ordonné et $|Y|$ petit) et telle que :

- $g(e_1,e_2) = g(e_1',e_2') \Rightarrow P(B(e_1,e_2)|e_1,e_2,l) = P(B(e_1',e_2')|e_1',e_2',l)$
- $y < y' \Rightarrow P(B(e_1,e_2)|g(e_1,e_2)=y) < P(B(e_1',e_2')|g(e_1',e_2')=y')$ $(y,y'\in Y)$

Le cas le plus simple envisageable est $Y = \{0,1\}$: les paires de fenêtres sont partitionnées en deux classes correspondant à deux niveaux d'affinité et dans ce cas :

$$P(B(e,e')|e,e',l) = P(B(e,e')|g(e,e'),l) = \begin{cases} \alpha_1^l \text{ si } g(e,e')=1 \\ \alpha_0^l \text{ si } g(e,e')=0 \end{cases}$$

avec $\alpha_1^l > \alpha_0^l$. L'observation d'un pont est donc liée de façon non déterministe à la valeur de g. Dans ce modèle, des paires d'environnements (e,e') formant un pont (resp. n'en formant pas) correspondent exactement à des exemples (e,e') d'étiquette $g(e,e')$ qui auraient subi un bruit de classification conditionnel à chaque classe : $\eta^1 = 1-\alpha_1^l$, $\eta^0 = \alpha_0^l$ (Ralaivola et *al.*, "CN=CPCN", ICML 2006). Si l'on peut apprendre une telle fonction g, on aura prouvé l'existence d'une information locale impliquée dans l'appariement des résidus.

3 Perceptron CCCN, expérimentations et conclusion

Le protocole proposé nécessite une méthode pouvant apprendre à partir de données sujettes à du bruit CCCN. Ce bruit est une généralisation du bruit uniforme (noté CN) car il n'impose pas l'égalité des taux de bruit sur chacune des classes. Nous souhaitons utiliser des séparateurs linéaires, avec l'idée d'utiliser des méthodes à noyaux par la suite. Nous avons montré que l'algorithme du perceptron peut se généraliser pour apprendre des séparateurs linéaires dans le contexte CCCN. En effet, lorsque les taux de bruits sont connus, on peut obtenir une expression analytique du vecteur de mise à jour du perceptron qui permet à celui-ci de converger en un nombre d'itérations similaire au perceptron classique. Lorsque les taux de bruit sur chacune des classes ne sont pas connus, nous avons construit un critère de sélection de modèle consistant permettant ainsi d'obtenir un algorithme efficace pour apprendre des séparateurs linéaires en présence de bruit CCCN.

Cet algorithme nous a permis de tester le protocole proposé dans cette étude sur des données artificielles de type ponts disulfures. Les résultats confirment que si les ponts sont formés en favorisant les paires de contextes locaux telles que la fonction d'affinité est forte, alors il est possible de détecter cette information locale et de la retrouver. L'expérimentation de ce protocole sur des protéines réelles est en cours pour les catégories de ponts disulfures et salins.