Un protocole pour détecter la présence d'information locale pour la prédiction des ponts disulfures

François Denis et **Christophe Magnan**

ACI GENOTO3D

Laboratoire d'Informatique Fondamentale de Marseille, UMR
CNRS 6166, Université de Provence

Motivations

- prédire les ponts disulfures : deux avis
  - une aide pour prédire la structure 3D ?
  - les ponts sont déterminés par le repliement ?
- information locale - information globale : les fenêtres centrées sur les cystéines oxydées contiennent-elles une information sur les ponts qu'elles forment ?

Problématique : poser un cadre formel pour étudier la présence d'information locale.

Intérêt ?

- permettra de répondre à la question sur les ponts disulfures ?
- pourra être réinvesti dans d'autres problèmes biologiques ?
  brins beta ?
- pas inintéressant du point de vue de l'apprentissage.

Modélisation (1)

Structure primaire d'une protéine : mot de Σ^* où Σ est l'ensemble des acides aminés ou tout autre alphabet dérivé.

$\mathcal{P} \subset \Sigma^*$: sous ensemble des protéines contenant un nombre pair de cystéines, toutes impliquées dans des ponts disulfures.

$\mathcal{P}_l \subset \mathcal{P}$: protéines de $\mathcal{P}$ contenant $2l$ cystéines.

$\mathcal{G}$: ensemble de graphes non orientés dont les sommets sont des entiers (position des cystéines) et ont une arité égale à 1.

$\Phi : \mathcal{P} \to \mathcal{G}$, associe de manière déterministe un graphe de connexion à une protéine de $\mathcal{P}$.

Modélisation (2)

Soit P une distribution de probabilités sur $\mathcal{P}$. Soit $r \in \mathbb{N}$ un rayon et soit $\Omega_r = \Sigma^{2r+1}$. Pour $w, w' \in \Omega_r$, on notera :

$P(w)$: probabilité que w soit un contexte local d'une cystéine dans une protéine $p \in \mathcal{P}$.

$P(w, w')$: probabilité que w et w' soient des contextes locaux distincts d'une cystéine dans une protéine $p \in \mathcal{P}$.

$P(w, w'|l)$: probabilité que w et w' soient des contextes locaux distincts d'une cystéine dans une protéine de $\mathcal{P}_l$.

$P(B(w, w')|w, w', l)$: probabilité que w et w' forment un pont sachant que ce sont des contextes locaux distincts d'une cystéine d'une protéine de $\mathcal{P}_l$.

Qu'est ce que l'information locale ?

Pas d'information locale

$$\Leftrightarrow$$

$P(B(w,w')|w,w',l) = 1/(2l-1)$ ne dépend que de l.

Pb : inenvisageable d'estimer $P(B(w,w')|w,w',l)$ sans hypothèses supplémentaires.

Si $r=3$, $|\Omega_r| = 20^{12} \simeq 4 \cdot 10^{15}$!

Seulement quelques centaines d'observations disponibles.

Fonction d'affinité

Idée : supposer l'existence d'une fonction $g : \Omega_r \to Y$ ($|Y|$ petit) :

$g(w_1, w_2) = g(w'_1, w'_2) \Rightarrow \forall l, P(B(w_1, w_2)|w_1, w_2, l) = P(B(w'_1, w'_2)|w'_1, w'_2, l)$.

Cas le plus simple : $|Y| = 2$. Les paires de fenêtres se répartissent en deux classes, correspondant à deux niveaux d'affinités.

Affinité, ponts et bruit de classification

Sous l'hypothèse de l'existence d'une fonction g avec $Y = \{0, 1\}$,

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

L'observation d'un pont (B) correspond à $g = 1$ avec un bruit de classification $\eta^+ = 1 - \alpha_1^l$ et $\eta^- = \alpha_0^l$.

Fonctions apprenables avec ce type de données

Soit $S \subset \mathcal{P}$ l'échantillon d'apprentissage, $S_l = S \cap \mathcal{P}_l$.

Chaque paire de contextes locaux de chaque protéine de S_l, formant un pont ou non, a été tirée selon un oracle distribuant les exemples de g avec bruit de classification : $EX(g, \eta_l^+, \eta_l^-)$.

Question : Quelles fonctions g peut on espérer apprendre à l'aide de sources de données distinctes de la forme

$$EX(g, \eta_2^+, \eta_2^-), \ldots, EX(g, \eta_l^+, \eta_l^-) \ ?$$

- les fonctions apprenables avec bruit de classification uniforme $EX(g, \eta)$ (Goldberg 2005) : les séparateurs linéaires en particulier,
- les fonctions apprenables par requêtes statistiques (Kearns 93?).

Fonctions apprenables par requêtes statistiques

Les fonctions apprenables à l'aide d'estimations de la forme :

$$\hat{P}(A), \hat{P}(A \cap \{g = 1\}), \hat{P}(A \cap \{g = 0\}).$$

Exemples :

- k-DNF : disjonction de monomes d'au plus k variables. Soit $g = m_1 \cup \ldots \cup m_s$: m figure dans g ssi $P(m \cap \{g = 0\}) = 0$.
- arbres de décisions : algorithmes basés sur des calculs d'entropie utilisant $\hat{P}(A), \hat{P}(A \cap \{g = 1\}), \hat{P}(A \cap \{g = 0\})$.

Requêtes statistiques et bruit de classification

Soit e l'étiquette observée sur les exemples distribués selon $EX(g, \eta^+, \eta^-)$. On a :

$$\begin{cases} P(A \cap \{e = 1\}) = P(A \cap \{g = 1\})(1 - \eta^+) + P(A \cap \{g = 0\})\eta^- \\ P(A \cap \{e = 0\}) = P(A \cap \{g = 1\})\eta^+ + P(A \cap \{g = 0\})(1 - \eta^-). \end{cases}$$

- Bruit connu : ok.
- Bruit inconnu : sélectionner les paramètres de bruit qui minimisent les désaccords sur les données observées.
- Passage à l sources distinctes : pas de pbs.

Premières conclusions

La présence d'information locale doit *théoriquement* pouvoir être détectée, sous réserve que la fonction d'affinité puisse être représentée par une fonction apprenable par requêtes statistiques.

Dans ce cas, la fonction d'affinité doit pouvoir être approchée de manière arbitrairement proche, en supposant qu'on dispose de suffisamment d'exemples, même en l'absence d'information sur la fonction Φ calculant le graphe de connexion d'une protéine.

Questions :

- combien d'exemples sont nécessaires ? bornes de convergence ?
- quelles classes de fonctions g choisir pour le pb des ponts disulfures ?

Simulations

$\Sigma = \{x_0, \ldots, x_{19}\}$; la cystéine est x_0.

Génération aléatoire de 150 “protéines” comportant 4, 6, 8 ou 10 cystéines selon un modèle multinomial généré aléatoirement, longueur des protéines choisie uniformément entre 50 et 150.

Fonction d’affinité : $g = x_3x_7 \vee x_0x_3 \vee x_2x_5 \vee x_1x_5$.

Les ponts sont formés en privilégiant les paires tq $g = 1$.

Fenêtres de rayon 4.

Codage “simple” des paires de fenêtres :
$(z_1, \ldots, z_9), (z'_1, \ldots, z'_9) \to (z_1z'_1, \ldots, z'_9)$.

Paramètres de bruit : 200 tirages aléatoires dans $[0, 1]^8$.

Fonction d'affinité : $x_3x_7 \vee x_0x_3 \vee x_2x_5 \vee x_1x_5$.

nbp/prot	nbprot	nb paires	$P(g=1)$	$P(g=1\|B,nbp)$	η^-	η^+
2	44	264	0.23	0.39	0.27	0.44
3	31	465	0.21	0.51	0.12	0.52
4	34	952	0.23	0.54	0.09	0.67
5	41	1845	0.24	0.62	0.06	0.72

La cible est retrouvée exactement!

	2	3	4	5
x1x5	18, 14	24, 22	85, 37	182, 74
x3x7	4, 3	6, 2	10, 5	31, 14
x0x3	5, 6	15, 16	36, 25	91, 35
x2x5	4, 11	8, 12	33, 13	51, 18

Conclusion

- Un modèle pour extraire une information locale portée par une fonction d'affinité apprenable par requêtes statistiques.
- Le modèle prédisait que g serait retrouvée. L'expérience montre que le nombre de données nécessaires est compatible avec le volume des données disponibles pour la prédiction des ponts disulfures.
- à suivre