Apprentissage semi-supervisé
et
prédiction de ponts disulfures dans les protéines

autour du mémoire de stage de DEA de

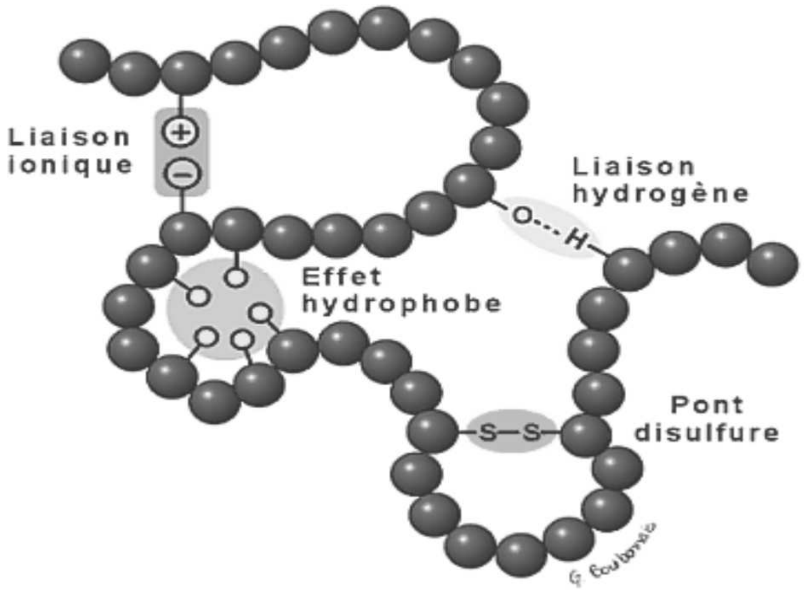
Christophe Magnan

encadré par : *François Denis* et *Cécile Capponi*,

ACI GENOTO3D

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

Prédiction de ponts disulfures



- Nombreux travaux sur la prédiction de l'état d'oxydation des cystéines (biblio du GT sequences et de l'ACI).
- Nous cherchons à prédire les appariements dans une protéine dont les cystéines sont oxydées.

Données disponibles

```
>/disk2/Pdb/Entries/pdb1e0f.ent  Chaine=K
IRFGMGKVPCPDGEVGYTCDCGEKICLYGQSCNDGQCSGDPKPSSEFEEFEIDE

TRUE–INTRA : CYS  K    10 – CYS  K    19 Dist =    2.026
TRUE–INTRA : CYS  K    21 – CYS  K    32 Dist =    2.031
TRUE–INTRA : CYS  K    26 – CYS  K    37 Dist =    2.026
```

Figure 1: Données extraites de la PDB par C. Geourjeon

L'information utile à la prédiction est vraisemblablement répartie

+ localement autour de chaque cystéine (contexte local)

+ dans des propriétés globales de la protéine (descripteurs globaux)

Une approche possible : extraire au mieux l'information disponible localement et la réinjecter dans des systèmes combinant les deux niveaux.

Données utilisées

>/disk2/Pdb/Entries/pdb1e0f.ent Chaine=K

IRFGMGKVPCPDGEVGYTCDCGEKICLYGQSCNDGQCSGDPKPSSEFEEFEIDE

TRUE–INTRA : CYS K 10 – CYS K 19 Dist = 2.026
TRUE–INTRA : CYS K 21 – CYS K 32 Dist = 2.031
TRUE–INTRA : CYS K 26 – CYS K 37 Dist = 2.026

1	GMGKVPPDGEVG	GEVGYTDCGEKI	0	GMGKVPPDGEVG	VGYTCDGEKICL	0	GEVGYCTDCGEKI	CLYGQSNDGQCS
1	VGYTCDGEKICL	CLYGQSNDGQCS	0	GMGKVPPDGEVG	DCGEKILYGQSC	0	GEVGYTDCGEKI	SCNDGQSGDPKP
1	DCGEKILYGQSC	SCNDGQSGDPKP	0	VGYTCDGEKICL	GEVGYTDCGEKI	0	GMGKVPPDGEVG	CLYGQSNDGQCS
0	VGYTCDGEKICL	DCGEKILYGQSC	0	CLYGQSNDGQCS	SCNDGQSGDPKP	0	VGYTCDGEKICL	SCNDGQSGDPKP
0	DCGEKILYGQSC	GEVGYTDCGEKI	0	DCGEKILYGQSC	CLYGQSNDGQCS	0	GMGKVPPDGEVG	SCNDGQSGDPKP

Figure 2: Les 15 paires de fenêtres de diamètre 12 centrées sur les 6 cystéines de pdb1e0f

Espace de représentation : paires de fenêtres de taille t, centrées sur les cystéines, appariées ou non.

Objectif : prédire les paires appariées.

Classification supervisée

X : espace des descriptions, $P(x)$.

Y : espace des classes $\{+,-\}$, $P(y|x)$.

Objectif : construire un classifieur $f : X \to Y$ minimisant $P\{(x,y) \in X \times Y | y \neq f(x)\}$ où $P(x,y) = P(x)P(y|x)$.

Classification supervisée. $S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, échantillon d'exemples étiquetés (x, y) i.i.d. selon $P(x, y)$.

Données biologiques : deux problèmes

1	GMGKVPPDGEVG	GEVGYTDCGEKI	0	GMGKVPPDGEVG	VGYTCDGEKICL	0	GEVGYCTDCGEKI	CLYGQSNDGQCS
1	VGYTCDGEKICL	CLYGQSNDGQCS	0	GMGKVPPDGEVG	DCGEKILYGQSC	0	GEVGYTDCGEKI	SCNDGQSGDPKP
1	DCGEKILYGQSC	SCNDGQSGDPKP	0	VGYTCDGEKICL	GEVGYTDCGEKI	0	GMGKVPPDGEVG	CLYGQSNDGQCS
0	VGYTCDGEKICL	DCGEKILYGQSC	0	CLYGQSNDGQCS	SCNDGQSGDPKP	0	VGYTCDGEKICL	SCNDGQSGDPKP
0	DCGEKILYGQSC	GEVGYTDCGEKI	0	DCGEKILYGQSC	CLYGQSNDGQCS	0	GMGKVPPDGEVG	SCNDGQSGDPKP

- L'hypothèse d'indépendance est malmenée : les paires non appariées peuvent être déduites des paires appariées. Pour $2n$ cystéines, n ponts et $2n(n-1)$ paires non appariées.
- Que signifie le fait qu'une paire n'est pas appariée ?
  - qu'elle ne peut pas l'être ?
  - qu'elle ne l'est pas mais qu'elle pourrait l'être dans d'autres circonstances ?

Un exemple artificiel

- L'environnement local d'une cystéine est de type A ou B.
- Règle d'appariement : deux cystéines peuvent former un pont si
  - leur environnement est de même type (règle locale) **et**
  - elles sont séparées par au plus une autre cystéine.

On observe la chaîne `-C-C-C-C-C-C-`, les types des environnements locaux correspondant étant `-A-B-A-B-A-A-`.

Une seule manière d'apparier les cystéines : (1,3), (2,4) et (5,6).

Les observations de paires sont :

1 AA	1 BB	1 AA	0 AB	0 AB	0 AA	0 AA	0 AB
0 AB	0 AB	0 AB	0 AA	0 AA	0 AB	0 AB	

Quel statut pour les paires non appariées ?

1	GMGKVPPDGEVG	GEVGYTDCGEKI	0	GMGKVPPDGEVG	VGYTCDGEKICL	0	GEVGYCTDCGEKI	CLYGQSNDGQCS
1	VGYTCDGEKICL	CLYGQSNDGQCS	0	GMGKVPPDGEVG	DCGEKILYGQSC	0	GEVGYTDCGEKI	SCNDGQSGDPKP
1	DCGEKILYGQSC	SCNDGQSGDPKP	0	VGYTCDGEKICL	GEVGYTDCGEKI	0	GMGKVPPDGEVG	CLYGQSNDGQCS
0	VGYTCDGEKICL	DCGEKILYGQSC	0	CLYGQSNDGQCS	SCNDGQSGDPKP	0	VGYTCDGEKICL	SCNDGQSGDPKP
0	DCGEKILYGQSC	GEVGYTDCGEKI	0	DCGEKILYGQSC	CLYGQSNDGQCS	0	GMGKVPPDGEVG	SCNDGQSGDPKP

Que signifie le fait qu'une paire n'est pas appariée ?

- qu'elle ne peut pas l'être ?
- qu'elle ne l'est pas mais qu'elle pourrait l'être dans d'autres circonstances ?

Si l'on veut extraire au mieux l'information locale, vaut-il mieux considérer une paire non appariée

- comme un exemple négatif (tiré selon $P(\cdot|-)$) **ou**
- comme un exemple non étiqueté (tiré selon $P(\cdot)$) ?

Classification supervisée et semi-supervisée

X : espace des descriptions, $P(x)$.

Y : espace des classes $\{+,-\}$, $P(y|x)$.

Objectif : construire un classifieur $f : X \rightarrow Y$ minimisant $P\{(x,y) \in X \times Y | y \neq f(x)\}$ où $P(x,y) = P(x)P(y|x)$.

Classification supervisée.

$S_{lab} = \{(x_1, y_1), \ldots, (x_l, y_l)\}$, échantillon i.i.d. selon $P(x,y)$.

Classification semi-supervisée (usuelle).

$S_{lab} = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ i.i.d. selon $P(x,y)$ et

$S_{unl} = \{x'_1, \ldots, x'_u\}$ i.i.d. selon $P(x)$.

Classification semi-supervisée asymétrique

$S_{pos} = \{x_1, \ldots, x_p\}$ i.i.d. selon $P(x|+)$ et

$S_{unl} = \{x'_1, \ldots, x'_u\}$ i.i.d. selon $P(x)$.

Consistance de la classification semi-supervisée asymétrique

- En général, $P(\cdot)$ et $P(\cdot|+)$ ne déterminent pas la distribution sous-jacente $P(\cdot,\cdot)$. En effet,

$$P(x,+) = P(x|+)\underline{P(+)} \text{ et } P(x,-) = P(x) - P(x,+)$$

  et $P(+) \in [0,\lambda]$ où $\lambda = Min\{\frac{P(x)}{P(x|+)}|P(x|+) > 0\}$.

- Pour certaines classes de distributions, $P(\cdot)$ et $P(\cdot|+)$ déterminent $P(+)$ (et donc $P(\cdot,\cdot)$)
  - Cas déterministe : $P(+|x)P(-|x) = 0$.
  - Lorsque $P(\cdot,\cdot)$ suit l'hypothèse du classifieur naïf de Bayes (+ quelques conditions naturelles ...).

Classifieur naïf de Bayes

Classifieur de Bayes : attribuer à une description x la classe y qui maximise $P(y|x)$.

$$f_B(x) = ArgMax_y P(y|x) = ArgMax_y P(x|y)P(y)$$

Classifieur d'erreur minimale mais généralement inatteignable.

Hypothèse du classifieur naïf de Bayes :

- $X = (X^1, \ldots, X^d)$
- $P(x^1, \ldots, x^d|y) = \prod_{i=1}^{d} P(X^i = x^i|y) = \prod_{i=1}^{d} P(X = x^i|y)$

Propriétés : ces hypothèses

- sont souvent insoutenables,
- conduisent à de mauvaises estimations de la densité,
- **conduisent à des performances très correctes en classification.**

Classifieur naïf de Bayes : cas supervisé

Modèle : $\theta = (P(+), P(X = x|+), P(X = x|-))$ où x décrit le domaine de X.

Principe du maximum de vraisemblance (ML) : ayant observé S_{lab}, choisir θ maximisant la vraisemblance de S_{lab} relativement à θ.

$$L(S_{lab};\theta) = \prod_{(x,y)\in S_{lab}} P_\theta(x,y)$$

$$= \prod_{(x,+)\in S_{lab}} P_\theta(x|+)P(+) \prod_{(x,-)\in S_{lab}} P_\theta(x|-)(1-P(+)).$$

$$\hat{P}(+) = \frac{\#\{(x,+) \in S_{lab}\}}{\#S_{lab}} \text{ et } \hat{P}(X = x^i|y) \simeq \frac{TF(x^i, y)}{\sum_{x^j \in Dom(X)} TF(x^j, y)}.$$

Le principe ML est consistant.

Classifieur naïf de Bayes : cas semi-supervisé asymétrique

Principe ML : ayant observé S_{pos} et S_{unl}, choisir θ maximisant la vraisemblance des échantillons relativement à θ.

$$L(S_{pos}, S_{unl}; \theta) = \prod_{x \in S_{pos}} \alpha P_\theta(x|+) \prod_{x \in S_{unl}} (1-\alpha) P_\theta(x)$$

$$= \prod_{x \in S_{pos}} \alpha P_\theta(x|+) \prod_{x \in S_{unl}} (1-\alpha) \left[P_\theta(x|+) P_\theta(+) + P_\theta(x|-)(1 - P_\theta(+)) \right]$$

- Le principe ML est consistant mais
- à part α, les paramètres du modèle ne sont plus calculables analytiquement.

Méthode EM (Expectation - Maximisation)

- Construction d'une suite de modèles $\theta_0, \ldots, \theta_i, \theta_{i+1}, \ldots$ telle que $L(S, \theta_{i+1}) \geq L(S, \theta_i)$. [Dempster et al. 77]
- Itération autour de "valeurs manquantes" prises comme pivot, calculées à partir du modèle courant (Expectation) puis utilisées pour calculer le modèle suivant (Maximisation).
- EM est utilisé
  - pour l'inférence de modèles de mélange $\alpha P_1 + (1 - \alpha) P_2$,
  - pour l'inférence des paramètres d'un HMM (Baum-Welch),
  - en apprentissage semi-supervisé (McCallum 99).

Algorithme NB semi-supervisé asymétrique

`Entrée :` S_{pos}, S_{unl}

$M = \emptyset$

`Répéter`

`Initialisation aléatoire de` $P(+)$

`Construction d'un modèle initial` θ_0 `à l'aide de` S_{pos}, S_{unl} `et` $P(+)$

`Pour` $i = 0$ `à` N `faire`

`Calculer` $\hat{P}(x|+)$ `et` $\hat{P}(x|-)$ `à l'aide de` θ_i `pour tout` $x \in S_{unl}$ `(Expectation)`

`Déterminer` θ_{i+1} `à l'aide de` S_{pos}, S_{unl} `et de ces estimations (Maximisation)`

$M = M \cup \{\theta_N\}$

`Choisir` $\theta \in M$ `maximisant la vraisemblance de` S_{pos}, S_{unl}

Un exemple de déroulement de l'algorithme

Données artificielles : $|S_{pos}| = 19, |S_{unl}| = 1000, |S_{test}| = 1000, P(+) = 0,5798, logL(\theta) = -5651,84, TxErreur(\theta) = 0,045$

Etape	$P(+)$	logL	TxErreur
0	0.7868	-6813.95	0.248
1	0.8286	-5009.26	0.188
2	0.7762	-5804.40	0.124
3	0.7221	-5735.62	0.111
4	0.6744	-5694.87	0.046
5	0.6368	-5673.13	0.046
6	0.6104	-5663.93	0.045
7	0.5934	-5661.03	0.045
8	0.5830	-5660.45	0.044
9	0.5767	-5660.38	0.045

Les données

Fichier de base : 4667 (domaines de) protéines dont 1278 comportent des ponts disulfures et 223 dont toutes les cystéines sont impliquées dans des ponts "intra".

# cystéines	#protéines	# cystéines	# protéines
2	59	16	4
4	49	20	2
6	50	24	2
8	26	32	1
10	20	34	1
12	9		

Ces 223 protéines comportent 5794 paires de cystéines dont 674 ponts.

Représentation des données.

Trois représentations d'une paire $(x_1 \dots x_d; x'_1 \dots x'_d)$: un "texte" dont les "mots" sont

- $\{x_i x'_i | 1 \leq i \leq d\}$
- $\{x_i x'_i | 1 \leq i \leq d\} \cup \{x_i x'_{d-i} | 1 \leq i \leq d\}$
- $\{x_i x'_j | 1 \leq i, j \leq d\}$

Une méthode de test

Construire un système de ponts à partir d'un classifieur estimant la probabilité qu'une paire forme un pont.

Exemple avec 6 cystéines $c_1, c_2, \ldots c_6$: le classifieur estime

$$\mathbf{P}(\mathbf{c_1}, \mathbf{c_3}) > P(c_1, c_5) > P(c_2, c_3) > \mathbf{P}(\mathbf{c_2}, \mathbf{c_6}) > P(c_3, c_5) > \ldots$$

Un problème : nous souhaitons comparer les approches supervisée et semi-supervisée. Mais comment évaluer l'erreur d'un classifieur en contexte semi-supervisé ?

Nous comparons le nombre de ponts bien prédits (rappel sur les exemples positifs, contrôlé par le nombre de prédictions autorisées).

Protocole

Validations croisées 10-folds sur les 223 protéines retenues, mélangées, pour des “rayons” de fenêtres variant de 1 à 8 ; comparaisons du nombre de ponts bien prédits par NB supervisé et semi-supervisé asymétrique.

Le nombre de ponts aléatoirement bien prédits est de $n/(2n-1)$ pour une protéine comportant $2n$ cystéines. Sur la totalité des données, une méthode aléatoire prédirait environ 158 ponts sur 674.

Résultats

- un rayon inférieur ou égal à 5 ne donne rien de tangible,
- le codage simple (C1) ne donne rien ; le codage simple et inversé (C2) est un peu meilleur ; le codage croisé (C3) semble le meilleur,
- l'algorithme NB supervisé à des résultats très proches d'une classification aléatoire (158 ponts),
- l'algorithme NB semi-supervisé asymétrique “décolle” assez nettement.

Rayon	NBsupC1	NBsupC2	NBsupC3	NBsemiC1	NBsemiC2	NBsemiC3
6	165	176	188	192	191	220
7	175	177	204	187	194	222
8	174	176	184	186	207	217

Conclusion et perspectives

- Hypothèse en passe d'être validée : mieux vaut traiter les paires non appariées comme des exemples non étiquetés
- Améliorations à prévoir : méthode de test, information évolutionnaire, traiter la non indépendance des données, ...
- Algorithmes plus performants que NB : version semi-supervisée des SVM
- Intégrer des infos plus globales
- à suivre ...