Construction de noyaux autour de pair-HMMs

Frédéric Sur

Loria & CNRS

En collaboration avec Yann Guermeur

Problème :

Protéine représentée par une séquence d'acides aminés (résidus)
$\longrightarrow$ prédiction de la structure secondaire.

Modélisation :

– Ensemble des descriptions $\mathcal{X}$: segments de $2n+1$ acides aminés.
– Ensemble de Q catégories $\mathcal{Y}$: états conformationnels (hélices, brins ...).

$\longrightarrow$ prise en compte des résidus adjacents pour la prédiction de l'état conformationnel d'un résidu donné.

Classification d'une nouvelle fenêtre x à l'aide d'une Machine à Vecteurs Support Multi-classe (M-SVM, *e.g.* [Guermeur 04]) :

$\longrightarrow$ description x associée à la catégorie $C(x) = C_{j^*}$ telle que

$$j^* = \underset{j \in \{1,\ldots,Q\}}{\arg\max} \sum_{i=1}^{m} \beta_{i,j} k(x_i, x) + b_j$$

où :

– $b_j \in \mathbb{R}$ et $\beta_{i,j} \in \mathbb{R}$ estimés à partir d'un ensemble d'apprentissage $\big(x_1, C(x_1)\big), \big(x_2, C(x_2)\big), \ldots, \big(x_m, C(x_m)\big)$,

– le noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est à choisir.

Classification d'une nouvelle fenêtre x à l'aide d'une Machine à Vecteurs Support Multi-classe (M-SVM, *e.g.* [Guermeur 04]) :

$\longrightarrow$ description x associée à la catégorie $C(x) = C_{j^*}$ telle que

$$j^* = \underset{j \in \{1,\ldots,Q\}}{\arg\max} \sum_{i=1}^{m} \beta_{i,j} k(x_i, x) + b_j$$

où :

– $b_j \in \mathbb{R}$ et $\beta_{i,j} \in \mathbb{R}$ estimés à partir d'un ensemble d'apprentissage $\big(x_1, C(x_1)\big), \big(x_2, C(x_2)\big), \ldots, \big(x_m, C(x_m)\big)$,

– le noyau $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est à choisir.

Propriétés recherchées pour le noyau :

– bonne capacité de discrimination
idéalement : $k(x, y) = 0$ si $C(x) \neq C(y)$ et $k(x, y) = 1$ si $C(x) = C(y)$.

– prise en compte des phénomènes de l'évolution biologique (substitutions et insertions / délétions).

Noyau défini positif

Définition d'un noyau (semi-)défini positif $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:
pour tout $n > 0$ et tout $x_1, x_2, \ldots, x_n \in \mathcal{X}$, la matrice $M_k = \big(k(x_i, x_j)\big)_{1 \leqslant i,j \leqslant n}$ est symétrique, semi-définie positive.
(*i.e.* $\forall\, C \in \mathbb{R}^n,\ {}^tC \cdot M_k \cdot C \geqslant 0$)

$\longrightarrow$ Comment construire un noyau à partir de noyaux ? [Shawe-Taylor - Christianini 04]

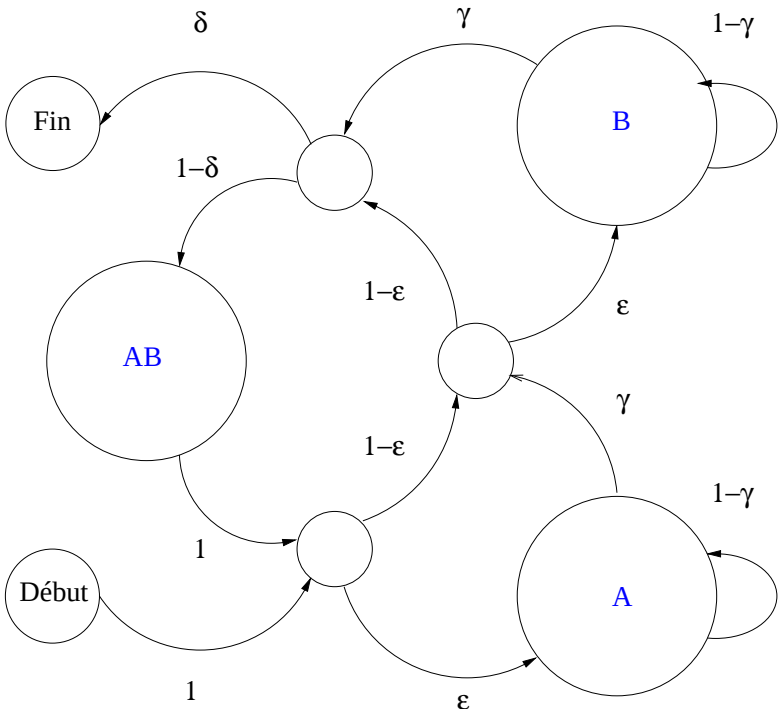
Proposition : si k et k' sont deux noyaux et $\alpha \in \mathbb{R}^+$, alors

- $(x, y) \mapsto k(x, y) + \alpha k'(x, y)$ est un noyau,
- $(x, y) \mapsto k(x, y) \cdot k'(x, y)$ est un noyau,
- $(x, y) \mapsto \exp\big(k(x, y)\big)$ est un noyau.

Proposition : si $f : \mathcal{X} \to \mathbb{R}^+$, alors $(x, y) \mapsto f(x) f(y)$ est un noyau.

Rappel 1 : le noyau $k(x, y)$ doit prendre en compte les phénomènes d'insertion / délétion.

$\longrightarrow$ Utilisation de Q pair-HMMs, un pour chaque catégorie. [Durbin *et al.* 98]



$\longrightarrow$ calcule $P(x, y|\Lambda_i)$: probabilité d'observer (x, y) sachant que $C(x) = C(y) = C_i$.

Proposition : $(x, y) \mapsto P(x, y|\Lambda_i)$ est un noyau sur $\mathcal{X} \times \mathcal{X}$. [Haussler 99] [Watkins 00]

Apprentissage d'un HMM (1)

Rappel 2 : le noyau $k(x, y)$ doit posséder une bonne capacité de discrimination.

Base de données $\mathcal{B}$: m échantillons $\big(x_1, C(x_1)\big), \big(x_2, C(x_2)\big), \ldots, \big(x_m, C(x_m)\big)$.

$\longrightarrow$ Ensemble d'apprentissage pour le pair-HMM Λ_i (associé à la catégorie C_i) :

$$\big\{(x_k, x_l) \in \mathcal{B}^2, C(x_k) = C(x_l) = C_i\big\}.$$

Rappel 2 : le noyau $k(x, y)$ doit posséder une bonne capacité de discrimination.

Base de données $\mathcal{B}$: m échantillons $\big(x_1, C(x_1)\big), \big(x_2, C(x_2)\big), \ldots, \big(x_m, C(x_m)\big)$.

$\longrightarrow$ Ensemble d'apprentissage pour le pair-HMM Λ_i (associé à la catégorie C_i) :

$$\big\{(x_k, x_l) \in \mathcal{B}^2, C(x_k) = C(x_l) = C_i\big\}.$$

Apprentissage standard d'un pair-HMM par maximum de vraisemblance :

$$\mathcal{L}(x_1, \ldots, x_n | \Lambda_i) = \prod_{C(x_i)=C(x_j)=C_i} P(x_i, x_j | \Lambda_i).$$

Algorithme : Baum-Welsh. [Rabiner 89]

$\longrightarrow$ utilise uniquement les exemples « positifs ».

$\Longrightarrow$ besoin d'un apprentissage discriminant.

Apprentissage d'un HMM (2)

Apprentissages discriminants, exemples : [Schlütter 00]

– critère *Maximum Mutual Information*

$$E_{\text{MMI}} = \sum_{k=1}^{Q} \left(\log(p(G_k|\lambda_k)p(\lambda_k)) - \log\left(\sum_{l=1}^{Q} P(G_k|\lambda_l)p(\lambda_l)\right)\right)$$

(algorithme : Baum-Welsh modifié)

– critère *Minimum Classification Error* ($\rho > 0$)

$$E_{\text{MCE}} = \sum_{k=1}^{Q} \frac{1}{1 + \left(\frac{p(G_k|\lambda_k)p(\lambda_k)}{\sum_{l\neq k} P(G_k|\lambda_l)p(\lambda_l)}\right)^{2\rho}}$$

(algorithmes : méthodes de descente de gradient)

où : $p(G_k|\lambda_l) = \prod_{C(x_i)=C(x_j)=C_k} P(x_i, x_j|\lambda_l).$

Comment classifier ?

$P(x, y|\Lambda_i)$: probabilité d'observer (x, y) sachant que $C(x) = C(y) = C_i$.

Soit $P(x, y|\Lambda_0)$ la probabilité d'observer (x, y) sachant que $C(x) \neq C(y)$.

Règle de classification : Maximum *a posteriori*.

$$P(\Lambda_0|x, y) < P\big(\Lambda_1 \text{ ou } \ldots \text{ ou } \Lambda_Q|x, y\big) \Longleftrightarrow C(x) = C(y)$$

d'où l'étude de : $R(x, y) := P\big(\Lambda_1 \text{ ou } \ldots \text{ ou } \Lambda_Q|x, y\big) = \sum_{i=1}^{Q} \frac{p(\Lambda_i)P(x, y|\Lambda_i)}{P(x, y)}$.

$$R(x,y) = \sum_{i=1}^{Q} \frac{p(\Lambda_i)P(x,y|\Lambda_i)}{P(x,y)}$$

$$R(x,y) > 1/2 \Longleftrightarrow C(x) = C(y).$$

Or on cherche k tel que (idéalement) : $\begin{cases} k(x,y) = 0 \text{ si } C(x) \neq C(y) \\ k(x,y) = 1 \text{ si } C(x) = C(y). \end{cases}$

Possibilité : $k(x,y) = \kappa' \exp\big(\sigma(R(x,y) - \rho)\big) = \kappa \exp\big(\sigma R(x,y)\big)$.

avec $\kappa, \kappa', \sigma, \rho > 0$ choisis avec soin.

Mais k doit être un noyau !

Proposition d'un noyau

$$k(x, y) = \kappa \exp\big(\sigma R(x, y)\big)$$

Proposition : si R est un noyau, alors k est un noyau.

Proposition d'un noyau

$k(x,y) = \kappa \exp(\sigma R(x,y))$

Proposition : si R est un noyau, alors k est un noyau.

$$R(x,y) = \sum_{i=1}^{Q} \frac{p(\Lambda_i) P(x,y|\Lambda_i)}{P(x,y)}$$

Modélisation de $P(x,y)$?

$\longrightarrow$ Sous l'hypothèse d'indépendance des observations :

$$P(x,y) = p(x)p(y)$$

avec $p(x) = \prod_{i=-n}^{n} \widehat{p}(\overline{x}_i)$ si $x = (\overline{x}_{-n}, \ldots, \overline{x}_n)$.
($\widehat{p}(\overline{x}_i)$ fréquence empirique du résidu $\overline{x}_i$)

Proposition : $(x,y) \mapsto 1/\big(p(x)p(y)\big)$ est un noyau sur $\mathcal{X} \times \mathcal{X}$.

Donc R est un noyau sur $\mathcal{X} \times \mathcal{X}$.

Comment fixer les paramètres ?

$\longrightarrow$ alignement de noyaux pour fixer κ et σ [Guermeur - Lifschitz - Vert 04].

Chercher κ et σ maximisant :

$$\widehat{A}(\kappa,\sigma) = \frac{\sum_{i,j} \delta_{C(x_i),C(x_j)} \cdot k_{\kappa,\sigma}(x_i,x_j)}{\left(\sum_{k=1}^{Q}(\#\{x_i, C(x_i) = C_k\})^2\right) \cdot \left(\sum_{i,j} k_{\kappa,\sigma}(x_i,x_j)^2\right)}.$$

où $\mathcal{B} = \left\{(x_i, C(x_i))\right\}$ est une base de données.

(corrélation avec un noyau « idéal »)

But : proposition de noyaux basés sur des pair-HMMs pour des M-SVM.

Avantage attendu : prise en compte des insertions / délétions dans les séquences de résidus.

Problèmes potentiels :
– modélisation de $P(x, y)$ (mais fortes contraintes pour obtenir un noyau. . .) ;
– complexité algorithmique (prédiction de ponts disulfures plus adaptée ?).

Autre approche envisageable : sorties de SVMs bi-classe (une pour chaque catégorie) « mixées ».

$\longrightarrow$ Reste à implémenter !