

# Noyaux marginalisés

*Cours Master 2005/06*

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

# Plan

- Noyaux marginalisés
- Noyaux pour séquences par chaînes de Markov cachées

# Noyaux marginalisés

# Motivations

- On se place dans le cas où les données  $x \in \mathcal{X}$  ne sont que la *partie "visible" d'objets plus complexes*  $z = (x, y) \in \mathcal{Z} := \mathcal{X} \cup \mathcal{Y}$ .
- $y \in \mathcal{Y}$  n'est *pas observé*, mais contient beaucoup d'*information pertinente* sur  $x$  (ex:  $x$  est une phrase,  $y$  sa structure grammaticale)
- On a un noyau naturel sur  $\mathcal{Z}$  (*les données complètes*).
- Comment en déduire un noyau sur  $\mathcal{X}$  (*les données observées*)?

# Cadre probabiliste

- Etant donné une observation  $\mathbf{x} \in \mathcal{X}$ , la donnée cachée correspondante  $y \in \mathcal{Y}$  ne peut pas être déduite de manière certaine.
- On se place donc dans un *cadre probabiliste*:  $(\mathcal{Y}, \mathcal{B})$  est un espace mesurable, et on suppose que pour chaque  $\mathbf{x} \in \mathcal{X}$ , on a une probabilité "conditionnelle"  
 $P_{\mathbf{x}} \in \mathcal{M}_1^+(\mathcal{Y}, \mathcal{B})$  (une mesure positive de masse 1).

# Noyau marginalisé

**Définition 1** Si  $K_{\mathcal{Z}}$  est une fonction sur  $\mathcal{Z}^2 = (\mathcal{X} \times \mathcal{Y})^2$ , mesurable en tant que fonction de  $y$  pour tout  $x$ , l'opération de **marginalisation** consiste à définir la fonction  $K_{\mathcal{X}}$  sur  $\mathcal{X}^2$  par:

$$\begin{aligned} K_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') &:= E_{P_{\mathbf{x}}(d\mathbf{y}) \times P_{\mathbf{x}'}(d\mathbf{y}')} K_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') \\ &= \int \int K_{\mathcal{Z}}((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) P_{\mathbf{x}}(d\mathbf{y}) P_{\mathbf{x}'}(d\mathbf{y}') \end{aligned}$$

**Théorème 2** Si  $K_{\mathcal{Z}}$  est un noyau défini positif sur  $\mathcal{Z}$ , alors  $K_{\mathcal{X}}$  est un noyau défini positif sur  $\mathcal{X}$  appelé **noyau marginalisé**.

# Preuve du théorème 1

Si  $K_{\mathcal{Z}}$  est un n.d.p. sur  $\mathcal{Z}$ , alors il existe un espace de Hilbert  $\mathcal{H}$  et  $\Phi_{\mathcal{Z}} : \mathcal{Z} \rightarrow \mathcal{H}$  tel que

$$K_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') = \langle \Phi_{\mathcal{Z}}(\mathbf{z}), \Phi_{\mathcal{Z}}(\mathbf{z}') \rangle_{\mathcal{H}}.$$

La marginalisation donne donc:

$$\begin{aligned} K_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') &= E_{P_{\mathbf{x}}(d\mathbf{y}) \times P_{\mathbf{x}'}(d\mathbf{y}')} K_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') \\ &= E_{P_{\mathbf{x}}(d\mathbf{y}) \times P_{\mathbf{x}'}(d\mathbf{y}')} \langle \Phi_{\mathcal{Z}}(\mathbf{z}), \Phi_{\mathcal{Z}}(\mathbf{z}') \rangle_{\mathcal{H}} \\ &= \langle E_{P_{\mathbf{x}}(d\mathbf{y})} \Phi_{\mathcal{Z}}(\mathbf{z}), E_{P_{\mathbf{x}'}(d\mathbf{y}')} \Phi_{\mathcal{Z}}(\mathbf{z}') \rangle_{\mathcal{H}} \end{aligned}$$

donc  $K_{\mathcal{X}}$  est un n.d.p. sur  $\mathcal{X}$ .  $\square$

# Noyaux pour séquences par chaînes de Markov cachées



# Motivations

- Les chaînes de Markov cachées (*HMM*: hidden Markov models) sont *extrêmement utilisées* pour modéliser des séquences biologiques
- Les états cachés (voir les slides suivants) ont en général un *sens biologique*
- La marginalisation va permettre de déduire un n.d.p. basé sur cette information biologique codée dans le modèle probabiliste
- Cette applications n'est pas restreinte aux séquences biologiques: les HMM sont très utilisées dans de nombreux autres domaines (reconnaissance vocale, traitements du signal,...).

# Chaîne de Markov

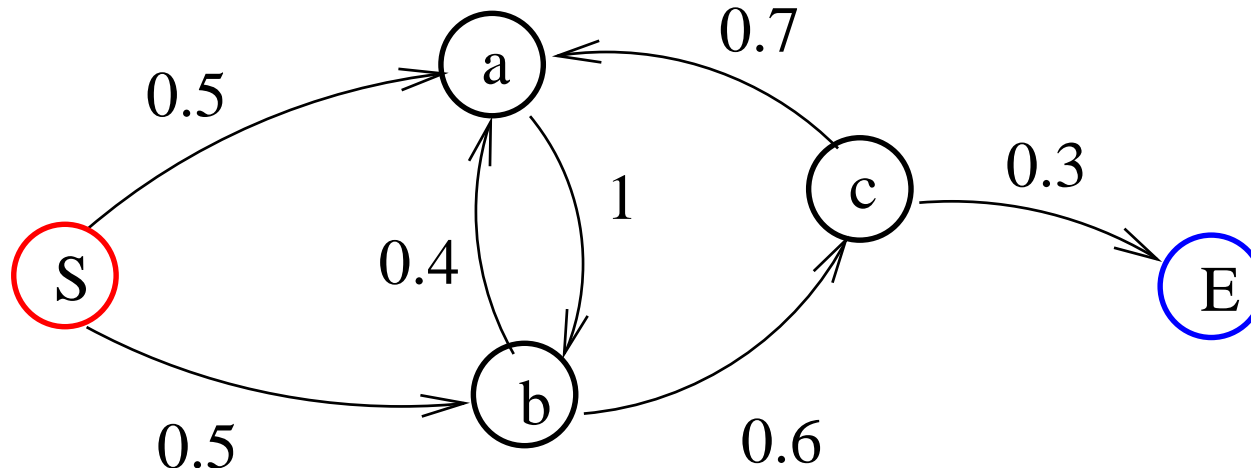
Soit  $\mathcal{S}$  un ensemble fini, comprenant un élément  $S$  (Start) et un élément  $E$  (End). Une chaîne de Markov est une distribution de probabilité  $P$  sur  $\mathcal{Y} = \mathcal{S}^*$  définie par:

$$P(y_0 y_1 \dots y_n) = \begin{cases} 0 & \text{si } y_0 \neq S \text{ ou } y_n \neq E \text{ ou } y_i = E \text{ pour } i < n, \\ \prod_{i=1}^n p(y_i | y_{i-1}) & \text{sinon,} \end{cases}$$

avec pour tout  $a, b \in \mathcal{Y}$

$$\begin{cases} p(a|b) \geq 0 \\ \sum_{c \in \mathcal{S}} p(c|a) = 1. \end{cases}$$

# Exemple



Calcul de la probabilité d'une séquence sous ce modèle:

$$P(SababcE) = 0.5 \times 1 \times 0.4 \times 1 \times 0.6 \times 0.3 = 0.0036$$

# Chaîne de Markov cachée (HMM)

On n'observe pas la suite des états, qui forment un chaîne de Markov. Par contre, on suppose que chaque état  $y \in \mathcal{S}$  (sauf  $S$  et  $E$ ) émet, indépendamment des autres, une variable  $x \in \mathcal{A}$  (fini) que l'on observe.

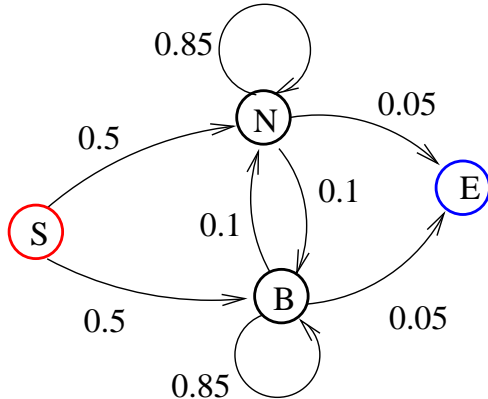
On obtient donc une probabilité sur les paires  $(\mathbf{x}, \mathbf{y}) \in \mathcal{A}^* \times \mathcal{S}^*$  définie par:

$$P(x_1 \dots x_n, y_0 \dots y_{n+1}) = p(y_0 \dots y_{n+1}) \times \prod_{i=1}^n \pi(x_i | y_i),$$

ou  $\pi$  est la probabilité conditionnelle d'émission

$(\sum_{x \in \mathcal{A}} \pi(x|y) = 1 \text{ pour } y \in \mathcal{S}).$

# Exemple: pile ou face biaisé



- Une pièce normale  $N$ , une biaisée  $B$  (non observé)

- On observe pile (0) ou face (1) avec probabilités:

$$\begin{cases} \pi(0|N) = 1 - \pi(1|N) = 0.5, \\ \pi(0|B) = 1 - \pi(1|B) = 0.8. \end{cases}$$

- Exemple de réalisation:

NNNNNBBBBBBBBBNNNNNNNNNNNBBBBBB  
1001011101111010010111001111011

# Noyau pour données observées

- Si on n'observe que  $\mathbf{x} \in \mathcal{A}^*$ , le n.d.p. le plus simple est le 1-spectral:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{a \in \mathcal{A}} n_a(\mathbf{x}) n_a(\mathbf{x}'),$$

ou  $n_a(\mathbf{x})$  est le nombre d'occurrences de  $a$  dans  $\mathbf{x}$ .

- Exemple:

$$\begin{aligned} \mathbf{x} &= 1001011101111010010111001111011, \\ \mathbf{x}' &= 0011010110011111011010111101100101, \end{aligned}$$

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= n_0(\mathbf{x}) n_0(\mathbf{x}') + n_1(\mathbf{x}) n_1(\mathbf{x}') \\ &= 11 \times 13 + 20 \times 21 = 563. \end{aligned}$$

# Noyaux pour données totales

- Si on observait les données totales  $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{A}^* \times \mathcal{S}^*$ , on pourrait utiliser le noyau:

$$K_{\mathbf{z}}(\mathbf{z}, \mathbf{z}') = \sum_{(a,s) \in \mathcal{A} \times \mathcal{S}} n_{a,s}(\mathbf{z}) n_{a,s}(\mathbf{z}'),$$

ou  $n_{a,s}(\mathbf{x}, \mathbf{y})$  est le nombre d'occurrences de  $s$  dans  $\mathbf{y}$  qui émettent  $a$  dans  $\mathbf{x}$ .

- Exemple:

$$\begin{aligned} \mathbf{z} &= 10010111101111010010111001111011, \\ \mathbf{z}' &= 0011010110011111011010111101100101, \end{aligned}$$

$$\begin{aligned} k(\mathbf{z}, \mathbf{z}') &= n_0(\mathbf{z}) n_0(\mathbf{z}') + n_1(\mathbf{z}) n_1(\mathbf{z}') + n_2(\mathbf{z}) n_2(\mathbf{z}') + n_3(\mathbf{z}) n_3(\mathbf{z}') \\ &= 7 \times 15 + 9 \times 12 + 13 \times 6 + 2 \times 1 = 293. \end{aligned}$$

# Noyau marginalisé

Le noyau marginalisé pour données observées est défini par:

$$\begin{aligned} K_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') &= \sum_{\mathbf{y}, \mathbf{y}' \in \mathcal{S}^*} K_{\mathcal{Z}}((\mathbf{x}, \mathbf{y}), (\mathbf{x}, \mathbf{y}')) P(\mathbf{y}|\mathbf{x}) P(\mathbf{y}'|\mathbf{x}') \\ &= \sum_{(a,s) \in \mathcal{A} \times \mathcal{S}} \Phi_{a,s}(\mathbf{x}) \Phi_{a,s}(\mathbf{x}'), \end{aligned}$$

avec

$$\Phi_{a,s}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{S}^*} P(\mathbf{y}|\mathbf{x}) n_{a,s}(\mathbf{x}, \mathbf{y})$$



# Calcul du noyau marginalisé

Nous allons calculer explicitement  $\Phi_{a,s}(\mathbf{x})$ , pour tout  $(a, s) \in \mathcal{A} \times \mathcal{S}$ . Supposons

$$\mathbf{x} = x_1 \dots x_n \in \mathcal{A}^n.$$

Alors  $P(\mathbf{y}|\mathbf{x}) > 0$  implique  $\mathbf{y}$  soit de la forme:

$$\mathbf{y} = Sy_1 \dots y_n E \in \mathcal{S}^{n+2}.$$

On a alors, en notant  $\delta$  est le symbole de Kronecker ( $\delta(u, v) = 1$  si  $u = v$ , 0 sinon):

$$n_{a,s}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \delta(x_i, a) \delta(y_i, b),$$

# Calcul du noyau marginalisé

On en déduit:

$$\begin{aligned}\Phi_{a,s}(\mathbf{x}) &= \sum_{\mathbf{y} \in \mathcal{S}^*} P(\mathbf{y}|\mathbf{x}) n_{a,s}(\mathbf{x}, \mathbf{y}) \\ &= \sum_{\mathbf{y} \in \mathcal{S}^*} P(\mathbf{y}|\mathbf{x}) \left\{ \sum_{i=1}^n \delta(x_i, a) \delta(y_i, s) \right\} \\ &= \sum_{i=1}^n \delta(x_i, a) \left\{ \sum_{\mathbf{y} \in \mathcal{S}^*} P(\mathbf{y}|\mathbf{x}) \delta(y_i, s) \right\} \\ &= \sum_{i=1}^n \delta(x_i, a) P(y_i = s|\mathbf{x}).\end{aligned}$$

# Calcul de $P(y_i = s | \mathbf{x})$

Il existe une méthode classique pour calculer  $P(y_i = s | \mathbf{x})$  de manière efficace: l'algorithme *Forward-Backward*. Soient:

$$\forall (s, i) \in \mathcal{S} \times [1, n], \quad \begin{cases} f_s(i) := P(x_1 \dots x_i, y_i = s), \\ b_s(i) := P(x_{i+1} \dots x_n | y_i = s) \end{cases}$$

Le calcul de  $P(y_i = s | \mathbf{x})$  se ramène à celui de  $f_s(i)$  et  $b_s(i)$ , car:

$$\begin{aligned} P(y_i = s | \mathbf{x}) &= \frac{P(x_1 \dots x_n, y_i = s)}{P(\mathbf{x})} \\ &= \frac{f_s(i)b_s(i)}{\sum_{s' \in \mathcal{S}} f_{s'}(i)b_{s'}(i)} \end{aligned}$$

# Calcul de $f_s(i)$

$f_s(i)$  (forward) se calcule *récurivement* pour par:

$$\forall t \in \mathcal{S}, \quad f_t(1) = P(x_1, y_1 = t) = p(t|S)\pi(x_1|t),$$

et pour  $t \in \mathcal{S}$  et  $j = 1, \dots, i$ :

$$f_t(j) = P(x_1 \dots x_j, y_j = t)$$

$$= \sum_{u \in \mathcal{S}} P(x_1 \dots x_j, y_{j-1} = u, y_j = t)$$

$$= \sum_{u \in \mathcal{S}} P(x_1 \dots x_{j-1}, y_{j-1} = u) P(x_j, y_j = t | y_{j-1} = u)$$

$$= \sum_{u \in \mathcal{S}} f_u(j-1) \times p(t|u)\pi(x_j|u).$$

# Calcul de $b_s(i)$

De même  $b_s(i)$  (backward) se calcule par:

$$\forall t \in \mathcal{S}, \quad b_t(n) = 1,$$

et pour  $t \in \mathcal{S}$  et  $j = n, n - 1, \dots, i$ :

$$\begin{aligned} b_t(j) &= P(x_{j+1} \dots x_n | y_j = t) \\ &= \sum_{u \in \mathcal{S}} P(x_{j+1} \dots x_n, y_{j+1} = u | y_j = t) \\ &= \sum_{u \in \mathcal{S}} P(x_{j+1}, y_{j+1} = u | y_j = t) P(x_{j+2} \dots x_n | y_{j+1} = u) \\ &= \sum_{u \in \mathcal{S}} b_u(j+1) \times p(u|t) \pi(x_{j+1}|u). \end{aligned}$$

# Remarque

Les algorithmes forward et backward sont les algorithmes de bases des HMM. On peut en particulier calculer la probabilité d'une séquence  $\mathbf{x} = x_1 \dots x_n$  par:

$$P(\mathbf{x}) = \sum_{s \in \mathcal{S}} P(x_1 \dots x_n, y_n = s) = \sum_{s \in \mathcal{S}} f_s(n).$$

# Extension

- Plutôt que le noyau 1-spectral, peut-on marginaliser des noyaux spectraux d'ordre supérieur?
- Par exemple:

$$K_{\mathcal{Z}}(\mathbf{z}, \mathbf{z}') = \sum_{(a,b,s,t) \in \mathcal{A}^2 \times \mathcal{S}^2} n_{a,b,s,t}(\mathbf{z}) n_{a,b,s,t}(\mathbf{z}'),$$

ou  $n_{a,b,s,t}(\mathbf{z})$  est le nombre d'occurrences du 2-mer  $st$  dans  $\mathbf{y}$  qui émettent  $ab$ .

- Intérêt: encoder plus d'information dans le noyau

# Extension (exercice)

Le noyau 2-spectral peut se marginaliser en:

$$K_2(\mathbf{x}, \mathbf{x}') = \sum_{(a,b,s,t) \in \mathcal{A}^2 \times \mathcal{S}^2} \Phi_{a,b,s,t}(\mathbf{x}) \Phi_{a,b,s,t}(\mathbf{x}'),$$

avec:

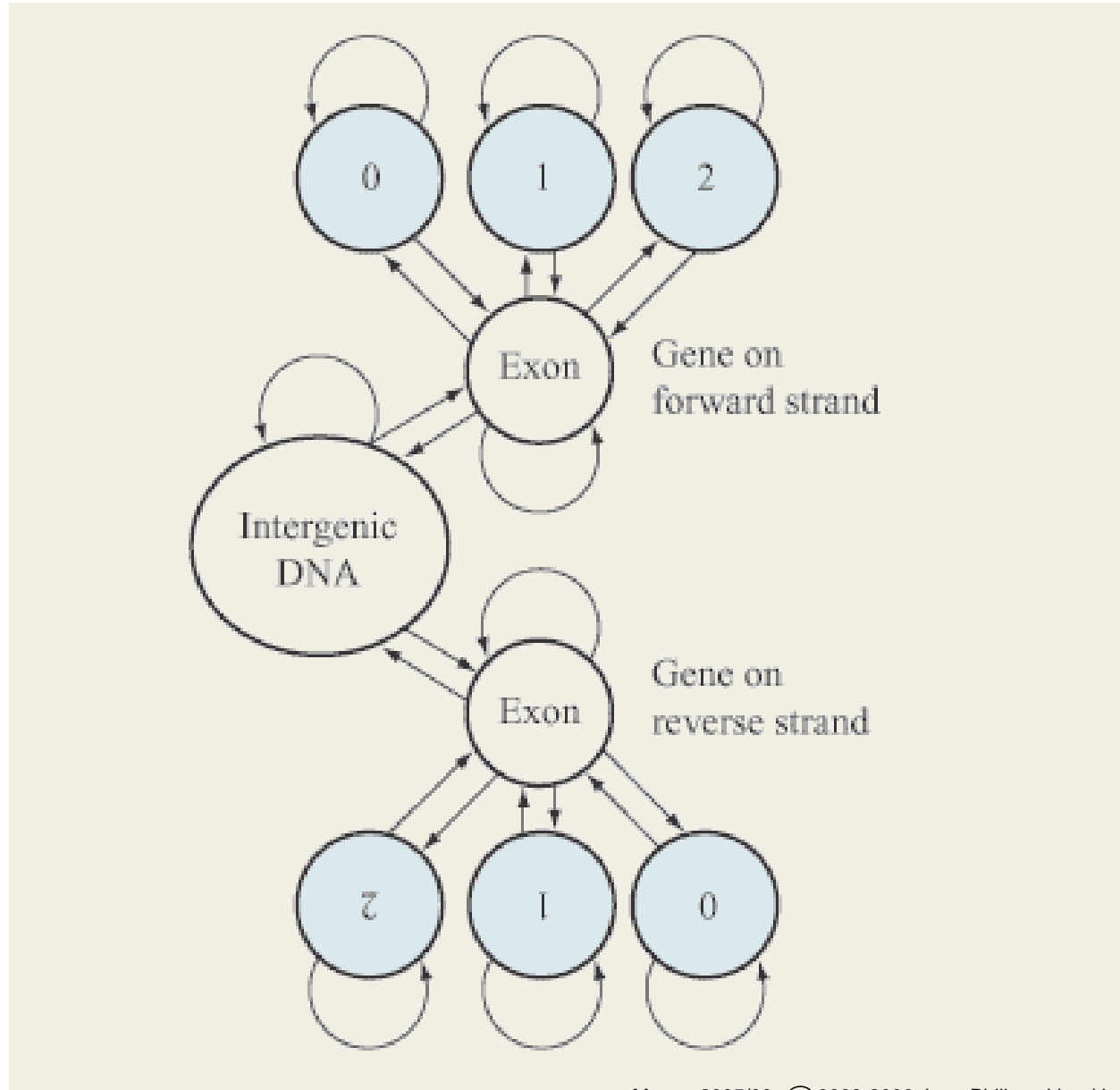
$$\Phi_{a,b,s,t}(\mathbf{x}) = \sum_{i=1}^{n-1} \delta(x_i, a) \delta(x_{i+1}, b) P(y_i = s, y_{i+1} = t | \mathbf{x}),$$

calculable par:

$$P(y_i = s, y_{i+1} = t | \mathbf{x}) = \frac{p(t|s)\pi(x_{i+1}|t)f_s(i)b_t(i+1)}{P(\mathbf{x})}.$$



# Exemple de HMM (DNA)



# Exemple de HMM (proteine)

