Noyau basé sur des pair-HMMs (pour la prédiction des ponts disulfures)

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, ou de ponts disulfures.

Modélisation :

– Ensemble des descriptions $\mathcal{X}$: segments de $2n+1$ acides aminés.
– Ensemble de Q catégories $\mathcal{Y}$: structure secondaire, ou présence/absence d'un pont disulfure.

$\longrightarrow$ prise en compte des résidus adjacents pour la prédiction.

Approche : classification d'une nouvelle fenêtre à l'aide d'une Machine à Vecteurs Support (éventuellement multi-classe).

But : construction d'un noyau robuste vis-à-vis des propriétés de l'évolution biologique (substitutions et insertions / délétions).

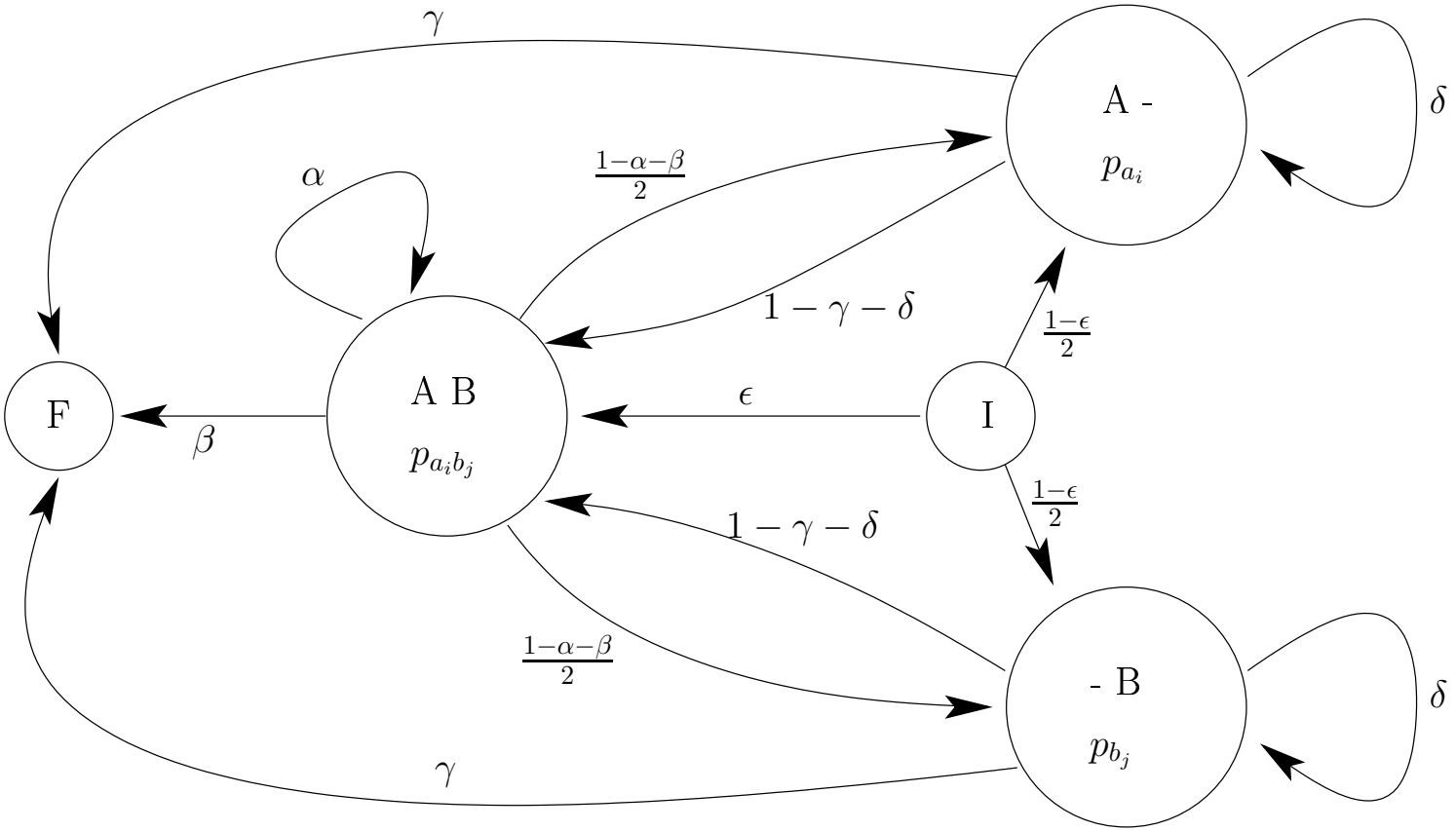
$\longrightarrow$ Utilisation d'un pair-HMM.

La loi jointe calculée par un pair-HMM est un noyau admissible sous certaines hypothèses.

D. Haussler. Convolution kernels on discrete structures. Rapport technique UCS-CRL-99-10, UC Santa Cruz, 1999.

C. Watkins. Dynamic alignment kernels. In A.J. Smola, P. Bartlett, B. Schölkopf et D. Schuurmans, éditeurs, *Advances in large margin classifiers*, pages 39-50. The MIT Press, 2000.

Pair-HMM :



Inspiré de :

R. Durbin, S. Eddy, A. Krogh, et G. Mitchison, *Biological sequence analysis*, chapitre 4, Cambridge University Press, 1998.

Algorithme d'apprentissage : Baum-Welch sur des paires de séquences (taille fixée) dont la distance d'édition est petite.

$\longrightarrow$ estimation des probabilités de transition et émission maximisant la vraisemblance des observations.

Distance d'édition pénalise les substitutions, insertions et délétions.

- permet d'obtenir un « meilleur alignement » par programmation dynamique.

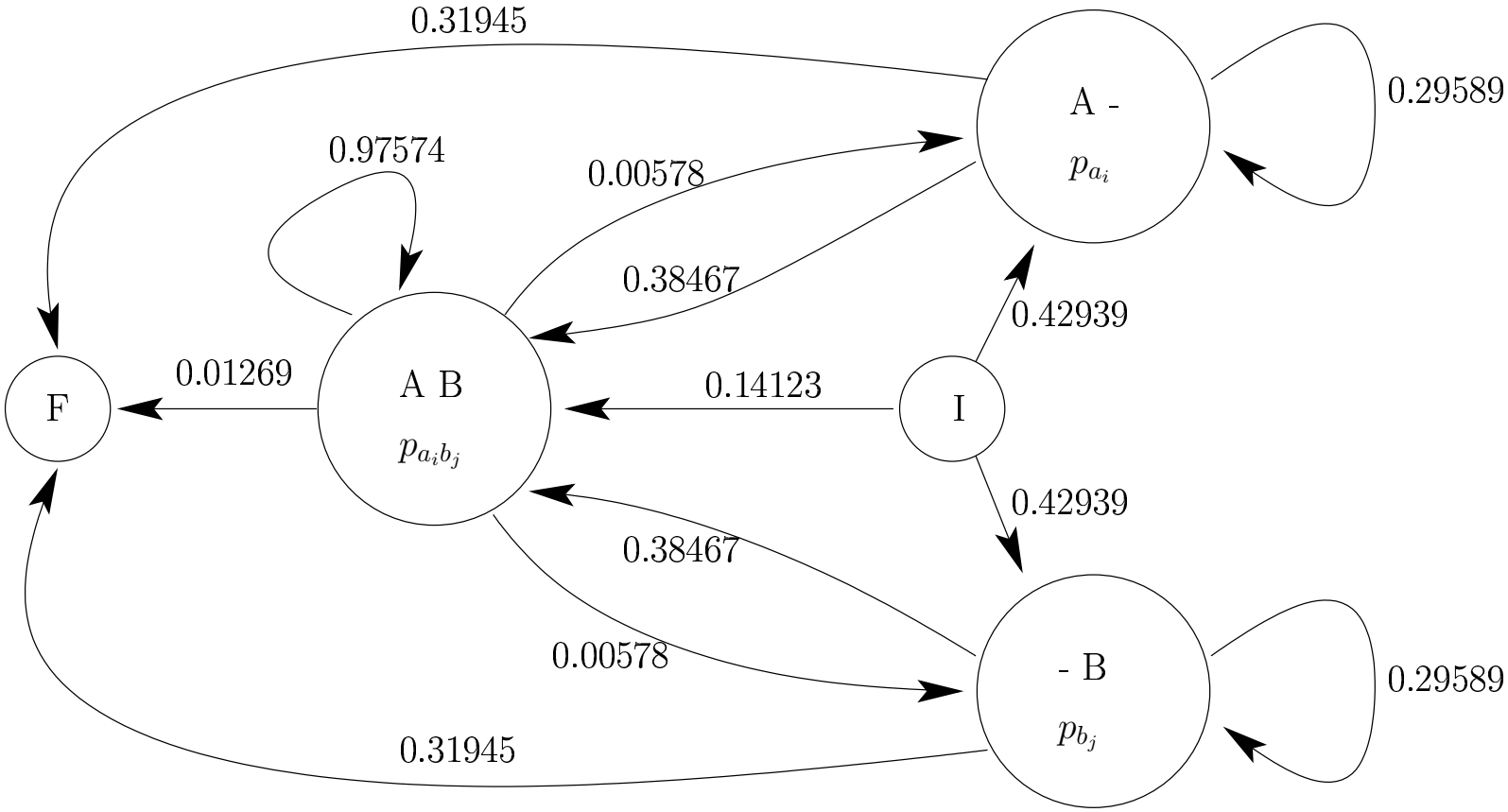
Mais :

- ne permet pas d'apprendre les coûts de substitution et insertion / délétion.

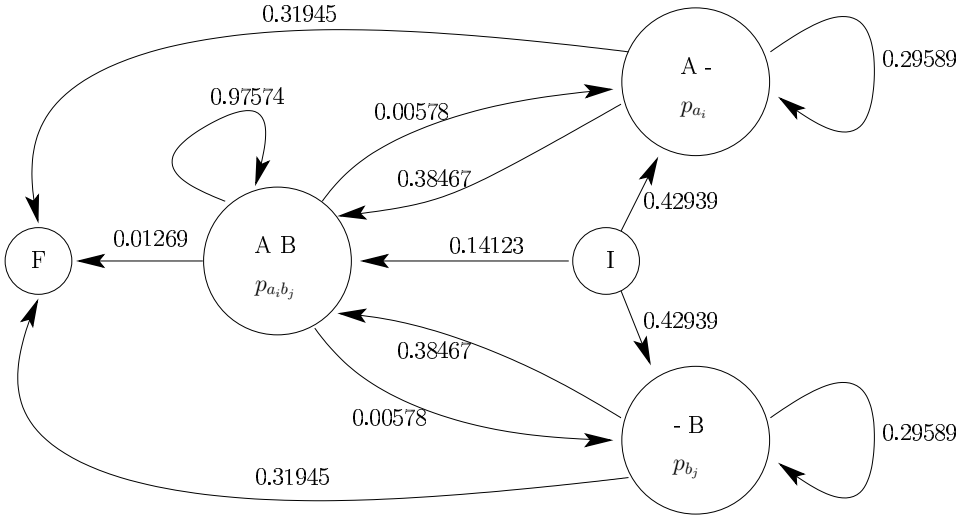
- n'est pas un noyau.

Estimation des probabilités de transition et émission

Estimation sur une base de 62359 paires de séquences de taille 13.



Remarque



Espérance du nombre d'émissions dans un état donné (proba. de boucler : p) :

$$\sum_{k=1}^{\infty} k p^{k-1}(1-p) = \frac{1}{1-p}\,.$$

Pour A B : 41.22

Pour A - ou - B : 1.42

Exemples d’alignements obtenus

Algorithme : Viterbi donne la séquence d’états la plus probable pour une paire de séquences.

Exemple d’alignements obtenus :

```
- - - R C V N F N C R A V H C P -
C A H D C - - - - C R C V H C P H

  H C - D L V A V K V - H C H C
  H C H D - - A A C V P H C H C

  - R C V N F N C R A V H C P -
  K H C V - H N C R A - H C C N
```

Étude de la loi d'émission

valeurs de $p(., L) = p(.|L)/p(L)$ (x10^4) :

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
7	0	3,9	4,8	10,6	3,6	5,6	0	15,1	0	875	4	1,9	4,2	4,5	0	0	1	2,6	6,2

score dans la matrice de PAM250 :

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2

Problème : calcul biaisé par la base d'apprentissage. . .

Autre approche : imposer les probabilités d'émission suivant les matrices de substitution PAM ou BLOSUM ?

À faire. . .

- Comparer les probabilités de transition obtenues à :

B. Knudsen et M. Miyamoto, Sequence alignments and pair hidden Markov models using evalutionary history, *Journal of Molecular Biology*, num. 333, pages 453-460, 2003.

(estimation basée sur la longueur moyenne des insertions / délétions et leur taux d'apparition)

- Comparer les probabilités d'émission et les matrices de substitutions classiques.

- Étudier la dépendance du PHMM estimé à la base d'apprentissage.

- Intégrer dans le noyau de l'information pertinente pour la prédiction des ponts disulfures : influence de la longueur de la séquence de prédiction, un PHMM modélisé par classe de prédiction. . .