Comparaison de séquences
Méthodes combinatoires
versus
Méthodes probabilistes
Pour une réconciliation

Dominique CELLIER
Laboratoire de Mathématiques Raphäel Salem
ABISS - Université de Rouen

Dominique.Cellier@univ-rouen.fr

Journées Statistiques et Probabilités en Génomique

Lyon - 1, 2 et 3 octobre 2002

Alignement global

On considère deux séquences

$$\mathbb{X} = x_1 x_2 \cdots x_n \qquad \text{et} \qquad \mathbb{Y} = y_1 y_2 \cdots y_m$$

$x_i,\ y_j \in \mathcal{A}$.

Alignement $(\mathbb{X}^*, \mathbb{Y}^*)$ des deux séquences :

$$\begin{array}{lcccc} (\mathbb{X}^* =) & x_1^* & x_2^* & \cdots & x_L^* \\ (\mathbb{Y}^* =) & y_1^* & y_2^* & \cdots & y_L^* \end{array}$$

$L \geq \max\{n, m\}$ et $x_i^*,\ y_j^* \in \mathcal{A}^* = \mathcal{A} \bigcup \{-\}$

exemple : $\mathbb{X} =$ VLSPADK et $\mathbb{Y} =$ HLAESK

V	L	S	P	A	D	-	K
H	L	-	-	A	E	S	K

Convention sur l'ordre des insertions/délétions

$$\begin{array}{ccccc} \alpha & \gamma & - & \delta & \varepsilon \\ \beta & - & \eta & - & \theta \end{array} \qquad \begin{array}{ccccc} \alpha & \gamma & \delta & - & \varepsilon \\ \beta & - & - & \eta & \theta \end{array}$$

Représentation d'un alignement

Ensemble des indices $\mathbb{A}^*$ des paires de lettres alignées

$$\mathbb{A}^* = \{(1,1), (2,2), (5,3), (6,4), (7,6)\}$$

Problème :

trouver le meilleur alignement possible !

Alignement global optimal

1. Un critère de qualité
2. Rechercher le meilleur alignement pour ce critère

Deux approches possibles

– Approche classique combinatoire
– Approche probabiliste par Chaîne de Markov Cachée (HMM)

Approche combinatoire

Score d'un alignement :

▷ *fonction de score s pour les appariements de lettres*

$$s \ : \ \mathcal{A}^2 \ \mapsto \mathbb{R}$$

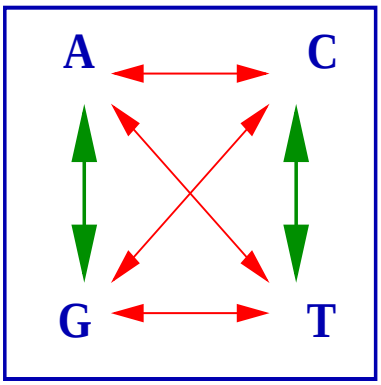
◇ **ADN :**

1. Fonction classique (cf BLAST, FASTA)

$$s(\alpha,\beta) \ = \ \begin{cases} \lambda & \text{si} \ \ \alpha = \beta \\ -\mu & \text{si} \ \ \alpha \neq \beta \end{cases}$$

2. Modèle d'évolution

	a	c	g	t
a	2			
c	-7	2		
g	-5	-7	2	
t	-7	-5	-7	2



◇ Protéines : matrices PAM, BLOSUM ...

Matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

▷ *fonction de score pour les gaps*

La plus fréquente est une fonction de gap affine

$$g(l_1, l_2) = \begin{cases} 0 & l_1 = 0 \ , \ l_2 = 0 \\ d + e \cdot (l_1 - 1) & l_1 \geq 1 \ , \ l_2 = 0 \\ d + e \cdot (l_2 - 1) & l_1 = 0 \ , \ l_2 \geq 1 \\ d' + 2 \cdot d + e \cdot (l_1 + l_2 - 2) & l_1 \geq 1 \ , \ l_2 \geq 1 \end{cases}$$

Cas particuliers : $d' = 0$, $e = 0$

▷ *score d'un alignement*

$$S(\mathbb{X}^*, \mathbb{Y}^*) = \sum_{k=1}^{l} s(x_{n_k}, y_{m_k}) - \sum_{k=0}^{l} g(n_{k+1} - n_k - 1, n_{k+1} - n_k - 1)$$

$\{(n_1, m_1), (n_2, m_2), \cdots, (n_l, m_l)\} = \mathbb{A}^*$,
$(n_0, m_0) = (0, 0)$ et $(n_{l+1}, m_{l+1}) = (n + 1, m + 1)$

exemple :

V	L	S	P	A	D	-	K
H	L	-	-	A	E	S	K

$$S(\mathbb{X}^*, \mathbb{Y}^*) = s(V, H) + s(L, L) + s(A, A) + s(D, E) + s(K, K) - 2 \cdot d - e$$

Alignement global optimal

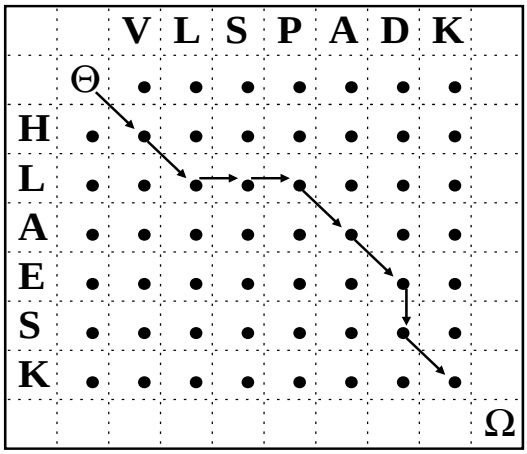
$$\mathbf{S}(\mathbb{X}, \mathbb{Y}) = \max_{(\mathbb{X}^*, \mathbb{Y}^*)} S(\mathbb{X}^*, \mathbb{Y}^*)$$

Algorithme de Needleman et Wunsch

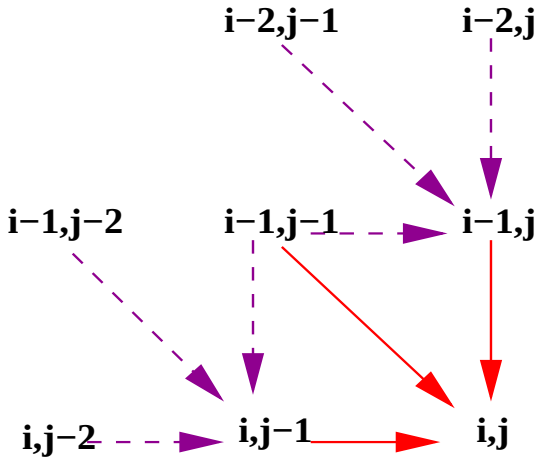
alignement $\Longleftrightarrow$ chemin dans un graphe orienté

exemple : $\mathbb{X}$ = VLSPADK et $\mathbb{Y}$ = HLAESK

V	L	S	P	A	D	-	K
H	L	-	-	A	E	S	K

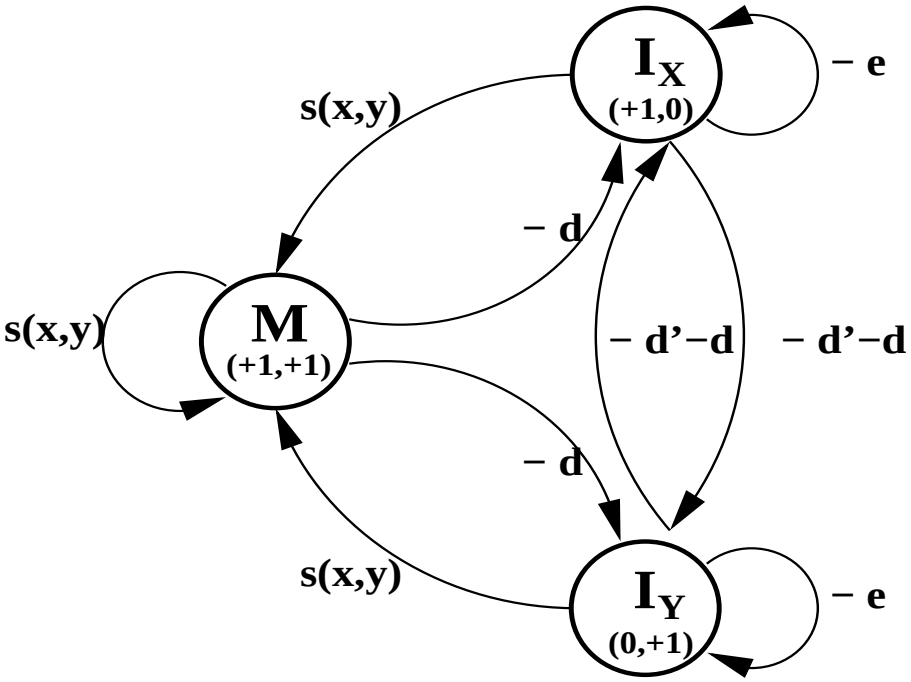


Formules de récurrence



$\Longrightarrow$ algorithme de programmation dynamique - complexité $O(nm)$

Représentation par un automate fini



M : état de **match**
$\mathrm{I}_{\mathbb{X}}$: état d'**insertion** sur la séquence $\mathbb{X}$
$\mathrm{I}_{\mathbb{Y}}$: état d'**insertion** sur la séquence $\mathbb{Y}$

A chaque état est attaché un indicateur de décallage d'indice.
A chaque transition est associé un accroissement du score.

Retour à l'exemple Représentation de la succession d'états associés à l'alignement :

V	L	S	P	A	D	-	K
H	L	-	-	A	E	S	K
M	M	$\mathrm{I}_{\mathbb{X}}$	$\mathrm{I}_{\mathbb{X}}$	M	M	$\mathrm{I}_{\mathbb{Y}}$	M

Les équations de récurrence associées à l'automate

$$\mathbf{S}_{i,j} = S(\mathbb{X}^{(i)}, \mathbb{Y}^{(j)})$$

$\mathbb{X}^{(i)} = x_1 x_2 \cdots x_i$ et $\mathbb{Y}^{(j)} = y_1 y_2 \cdots y_j$.

$$\begin{aligned}
V^{\mathrm{M}}(i,j) &= s(x_i, y_j) + \max \begin{cases} V^{\mathrm{M}}(i-1, j-1) \\ V^{\mathrm{I}_{\mathbb{X}}}(i-1, j-1) \\ V^{\mathrm{I}_{\mathbb{Y}}}(i-1, j-1) \end{cases} \\
V^{\mathrm{I}_{\mathbb{X}}}(i,j) &= \max \begin{cases} V^{\mathrm{M}}(i-1, j) - d \\ V^{\mathrm{I}_{\mathbb{X}}}(i-1, j) - e \\ V^{\mathrm{I}_{\mathbb{Y}}}(i-1, j) - d - d' \end{cases} \\
V^{\mathrm{I}_{\mathbb{Y}}}(i,j) &= \max \begin{cases} V^{\mathrm{M}}(i, j-1) - d \\ V^{\mathrm{I}_{\mathbb{Y}}}(i, j-1) - e \\ V^{\mathrm{I}_{\mathbb{X}}}(i, j-1) - d - d' \end{cases}
\end{aligned}$$

On note $V^{\clubsuit}(i,j)$ le score de l'alignement global optimal se terminant au couple de sites (i,j) dans l'état $\clubsuit$.

Similitude troublante avec un **HMM** et les équations de récurrence liées à l'**algorithme de Viterbi** associé !

Algorithme de Needleman et Wunsch :

1- Initialisation :
$V^{\mathbf{M}}(0,0) = 0$ et $V^{\mathbf{I}_{\mathbb{X}}}(0,0) = V^{\mathbf{I}_{\mathbb{Y}}}(0,0) = -\infty$
Pour $i = -1$ à n et $j = -1$ à m
$V^*(i,-1) = V^*(-1,j) = -\infty$

2- Pour $i = 1$ à n et $j = 1$ à m faire

$$V^{\mathbf{M}}(i,j) = s(x_i, y_j) + \max \begin{cases} V^{\mathbf{M}}(i-1,j-1) \\ V^{\mathbf{I}_{\mathbb{X}}}(i-1,j-1) \\ V^{\mathbf{I}_{\mathbb{Y}}}(i-1,j-1) \end{cases}$$

$$V^{\mathbf{I}_{\mathbb{X}}}(i,j) = \max \begin{cases} V^{\mathbf{M}}(i-1,j) - d \\ V^{\mathbf{I}_{\mathbb{X}}}(i-1,j) - e \\ V^{\mathbf{I}_{\mathbb{Y}}}(i-1,j) - d - d' \end{cases}$$

$$V^{\mathbf{I}_{\mathbb{Y}}}(i,j) = \max \begin{cases} V^{\mathbf{M}}(i,j-1) - d \\ V^{\mathbf{I}_{\mathbb{Y}}}(i,j-1) - e \\ V^{\mathbf{I}_{\mathbb{X}}}(i,j-1) - d - d' \end{cases}$$

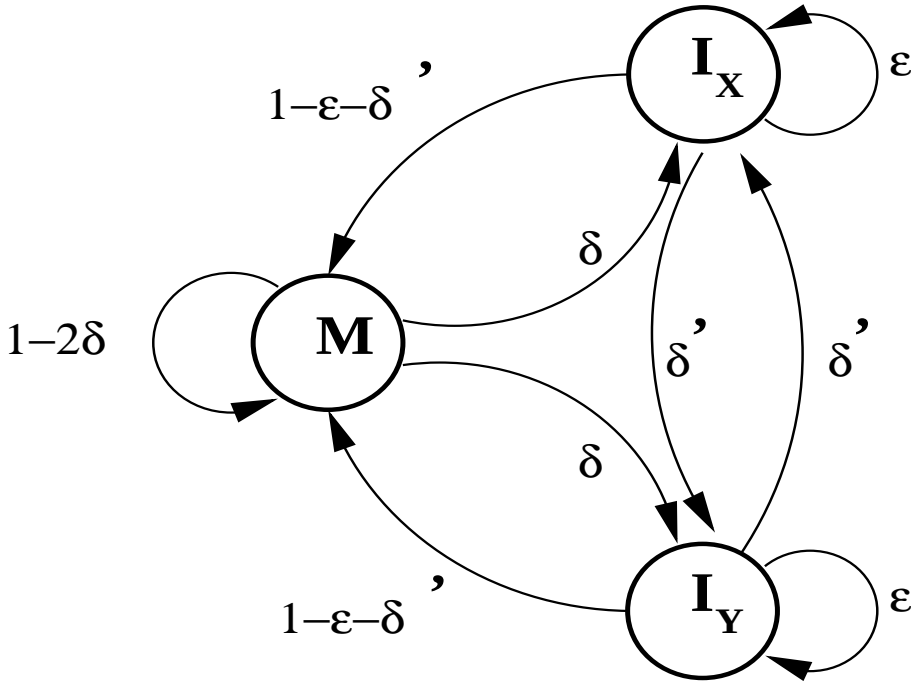
3- Fin :

$$\mathbf{S}(\mathbb{X}, \mathbb{Y}) = \mathbf{S}_{n,m} = \max \left\{ V^{\mathbf{M}}(n,m), V^{\mathbf{I}_{\mathbb{X}}}(n,m), V^{\mathbf{I}_{\mathbb{Y}}}(n,m) \right\}$$

et remontée pour obtenir l'alignement

Complexité : $O(nm)$

Approche probabiliste par HMM

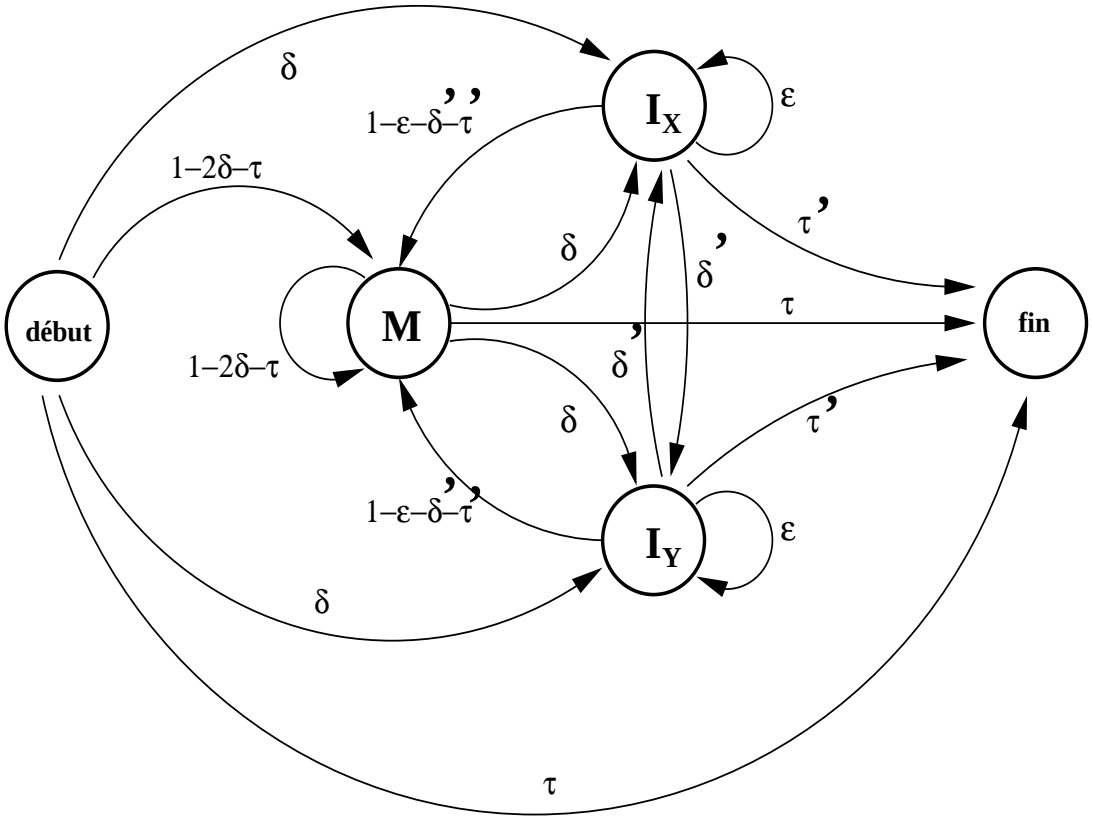


Les **états**, les **transitions** et les **probabilités d'émission** :

$\mathbf{M}$ - *état de match* : alignement (α, β) avec probabilité $p_{\alpha,\beta}$
$\mathbf{I}_{\mathbb{X}}$ - *état d'insertion* sur la séquence $\mathbb{X}$: alignement $(\alpha, -)$ avec probabilité q_α
$\mathbf{I}_{\mathbb{Y}}$ - *état d'insertion* sur la séquence $\mathbb{Y}$: alignement $(-, \beta)$ avec probabilité q_β

Les *transitions entre états* : probabilités associées aux arcs.

Alignement global par HMM : modèle $\mathcal{M}$



Retour à l'exemple

Représentation de la succession d'états cachés associés à l'alignement :

	V	L	S	P	A	D	-	K	
	H	L	-	-	A	E	S	K	
début	M	M	I_X	I_X	M	M	I_Y	M	fin

Vraisemblance d'un alignement de deux séquences

$$
\begin{aligned}
\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}) \quad = \quad & (1-2\delta-\tau)^{n_{\mathbf{dM}}} \cdot (\delta)^{n_{\mathbf{dI_X}}} \cdot (\delta)^{n_{\mathbf{dI_Y}}} \cdot (\tau)^{n_{\mathbf{df}}} \\
& \times (1-2\delta-\tau)^{n_{\mathbf{MM}}} \cdot (\delta)^{n_{\mathbf{MI_X}}} \cdot (\delta)^{n_{\mathbf{MI_Y}}} \cdot (\tau)^{n_{\mathbf{Mf}}} \\
& \times (\epsilon)^{n_{\mathbf{I_X I_X}}} \cdot (1-\epsilon-\delta'-\tau')^{n_{\mathbf{I_X M}}} \cdot (\delta')^{n_{\mathbf{I_X I_Y}}} \cdot (\tau')^{n_{\mathbf{I_X f}}} \\
& \times (\epsilon)^{n_{\mathbf{I_Y I_Y}}} \cdot (1-\epsilon-\delta'-\tau')^{n_{\mathbf{I_Y M}}} \cdot (\delta')^{n_{\mathbf{I_Y I_X}}} \cdot (\tau')^{n_{\mathbf{I_Y f}}} \\
& \times \prod_{\alpha,\beta \in \mathcal{A}} (p_{\alpha\beta})^{n_{(\alpha,\beta)}} \\
& \times \prod_{\alpha \in \mathcal{A}} (q_\alpha)^{n_{(\alpha,-)}} \cdot \prod_{\beta \in \mathcal{A}} (q_\beta)^{n_{(-,\beta)}}
\end{aligned}
$$

Retour à l'exemple La vraisemblance de l'alignement : l'alignement :

	V	L	S	P	A	D	-	K	
	H	L	-	-	A	E	S	K	
début	**M**	**M**	$\mathbf{I_X}$	$\mathbf{I_X}$	**M**	**M**	$\mathbf{I_Y}$	**M**	fin

$$
\begin{aligned}
\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}) \quad = \quad & (1-2\delta-\tau) \times p_{VH} \times (1-2\delta-\tau) \times p_{LL} \\
& \times \delta \times q_S \times \epsilon \times q_P \\
& \times (1-\epsilon-\delta'-\tau') \times p_{AA} \times (1-2\delta-\tau) \times p_{DE} \\
& \times \delta \times q_S \times (1-\epsilon-\delta'-\tau') \times p_{KK}
\end{aligned}
$$

Algorithme de Viterbi d'alignement global par HMM de deux séquences

Rechercher l'alignement $\widetilde{\mathbb{A}} = (\widetilde{\mathbb{X}}, \widetilde{\mathbb{Y}})$ de $\mathbb{X}$ et $\mathbb{Y}$ de vraisemblance maximum dans le modèle HMM, $\mathcal{M}$, choisi :

$$\widetilde{\nu} = \mathbb{P}\left(\mathbb{X}, \mathbb{Y}, \widetilde{\mathbb{A}} \mid \mathcal{M}\right) = \max_{\mathbb{X}^*, \mathbb{Y}^*} \mathbb{P}\left(\mathbb{X}, \mathbb{Y}, (\mathbb{X}^*, \mathbb{Y}^*) \mid \mathcal{M}\right)$$

On note $\nu^{\mathbf{S}}(i, j)$ la vraisemblance de l'alignement optimal se terminant au couple de sites (i, j) dans l'état $\mathbf{S}$.

Algorithme de Viterbi

1. **initialisation :**

   $\nu^{\mathrm{M}}(0,0) = 1$ et $\nu^{\mathrm{I}_{\mathbb{X}}}(i,0) = \nu^{\mathrm{I}_{\mathbb{Y}}}(0,j) = 0$

2. **pour** $i = 1, \cdots, n$ **et** $j = 1 \cdots, m$ **faire**

$$\nu^{\mathrm{M}}(i,j) = p_{x_i,y_j} \cdot \max \begin{cases} (1-2\delta-\tau) \cdot \nu^{\mathrm{M}}(i-1,j-1) \\ (1-\epsilon-\delta'-\tau') \cdot \nu^{\mathrm{I}_{\mathbb{X}}}(i-1,j-1) \\ (1-\epsilon-\delta'-\tau') \cdot \nu^{\mathrm{I}_{\mathbb{Y}}}(i-1,j-1) \end{cases}$$

$$\nu^{\mathrm{I}_{\mathbb{X}}}(i,j) = q_{x_i} \cdot \max \begin{cases} \delta \cdot \nu^{\mathrm{M}}(i-1,j) \\ \epsilon \cdot \nu^{\mathrm{I}_{\mathbb{X}}}(i-1,j) \\ \delta' \cdot \nu^{\mathrm{I}_{\mathbb{Y}}}(i-1,j) \end{cases}$$

$$\nu^{\mathrm{I}_{\mathbb{Y}}}(i,j) = q_{y_j} \cdot \max \begin{cases} \delta \cdot \nu^{\mathrm{M}}(i,j-1) \\ \epsilon \cdot \nu^{\mathrm{I}_{\mathbb{Y}}}(i,j-1) \\ \delta' \cdot \nu^{\mathrm{I}_{\mathbb{X}}}(i,j-1) \end{cases}$$

3. **fin :**

   $\widetilde{\nu} = \max\{\tau \cdot \nu^{\mathrm{M}}(n,m), \tau' \cdot \nu^{\mathrm{I}_{\mathbb{X}}}(n,m), \tau' \cdot \nu^{\mathrm{I}_{\mathbb{Y}}}(n,m)\}$

   et remontée pour obtenir l'alignement

Complexité : $O(nm)$

Approche combinatoire versus HMM

Automate d'alignement associé à un HMM d'alignement

Problème : *Passer du modèle probabiliste au système de score.*

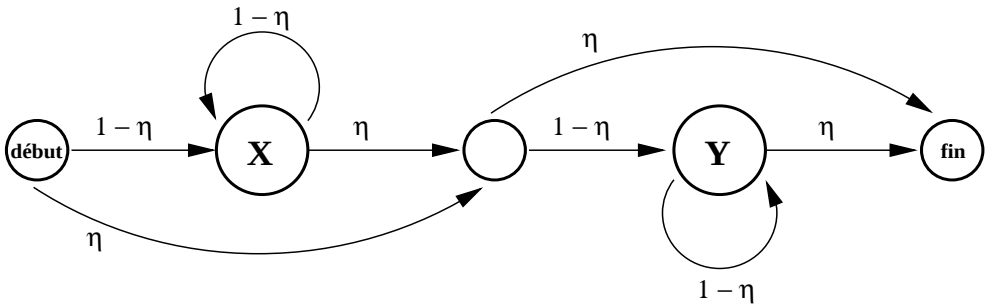
Vraisemblance d'un alignement dans le modèle $\mathcal{M}$

En choisissant $\tau' = \tau \cdot \frac{1-\epsilon-\delta'}{1-2\delta}$ la vraisemblance se réécrit

$$\begin{aligned}\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}) \quad = \quad & \left[(1-2\delta-\tau)^{n_{\mathrm{M}}} \cdot \prod_{\alpha,\beta \in \mathcal{A}} (p_{\alpha,\beta})^{n_{(\alpha,\beta)}}\right] \\ & \times \left[\left(\delta \cdot \frac{1-\epsilon-\delta'}{1-2\delta}\right)^{n_{\mathrm{MI}}} \right. \\ & \left. \cdot \prod_{\alpha \in \mathcal{A}} (q_\alpha)^{n_{(\alpha,-)}} \cdot \prod_{\beta \in \mathcal{A}} (q_\beta)^{n_{(-,\beta)}}\right] \\ & \times (\epsilon)^{n_{\mathrm{I}^2}} \times (\delta')^{n_{\mathrm{I}\to\mathrm{I}}} \times \tau\end{aligned}$$

Modèle d'alignement $\mathcal{M}_0$:

HMM sous l'hypothèse d'indépendance des séquences :



Vraisemblance dans le modèle $\mathcal{M}_0$

$$\begin{aligned}\mathbb{P}\left(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}_0\right) &= (1-\eta)^{n+m} \cdot \eta^2 \\ &\quad \times \prod_{\alpha \in \mathcal{A}} (q_\alpha)^{n_{(\alpha,-)}} \cdot \prod_{\beta \in \mathcal{A}} (q_\beta)^{n_{(-,\beta)}}\end{aligned}$$

Le rapport de vraisemblance du modèle $\mathcal{M}$ versus le modèle $\mathcal{M}_0$

$$\begin{aligned}\frac{\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M})}{\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}_0)} &= \left[\left(\frac{1-2\delta-\tau}{(1-\eta)^2}\right)^{n_{\mathbf{M}}} \cdot \prod_{\alpha,\beta\in\mathcal{A}} \left(\frac{p_{\alpha,\beta}}{q_\alpha q_\beta}\right)^{n_{(\alpha,\beta)}}\right] \\ &\quad \times \left[\left(\frac{\delta}{1-\eta} \cdot \frac{1-\epsilon-\delta'}{1-2\delta}\right)\right]^{n_{\mathbf{MI}}} \\ &\quad \times \left(\frac{\epsilon}{1-\eta}\right)^{n_{\mathbf{I}^2}} \\ &\quad \times \left(\frac{\delta'}{1-\eta}\right)^{n_{\mathbf{I}\to\mathbf{I}}} \\ &\quad \times \tau \times \frac{1}{\eta^2}\end{aligned}$$

Le logarithme du rapport de vraisemblance

$$\begin{aligned}\ln\left[\frac{\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M})}{\mathbb{P}(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}_0)}\right] &= \sum_{\alpha,\beta\in\mathcal{A}} n_{(\alpha,\beta)} \cdot \left[\ln\left(\frac{p_{\alpha,\beta}}{q_\alpha q_\beta}\right) + \ln\left(\frac{1-2\delta-\tau}{(1-\eta)^2}\right)\right] \\ &\quad + n_{\mathbf{MI}} \cdot \ln\left(\frac{\delta}{1-\eta} \cdot \frac{1-\epsilon-\delta'}{1-2\delta}\right) \\ &\quad + n_{\mathbf{I}^2} \cdot \ln\left(\frac{\epsilon}{1-\eta}\right) \\ &\quad + n_{\mathbf{I}\to\mathbf{I}} \cdot \ln\left(\frac{\delta'}{1-\eta}\right) \\ &\quad + \ln\left(\frac{\tau}{\eta^2}\right)\end{aligned}$$

Construction d'un système de score additif

$$\begin{cases} s(\alpha,\beta) &= \ln\left(\frac{p_{\alpha,\beta}}{q_\alpha q_\beta}\right) + \ln\left(\frac{1-2\delta-\tau}{(1-\eta)^2}\right) \\ d &= -\ln\left(\frac{\delta}{1-\eta}\cdot\frac{1-\epsilon-\delta'}{1-2\delta}\right) \\ e &= -\ln\left(\frac{\epsilon}{1-\eta}\right) \\ d'+d &= -\ln\left(\frac{\delta'}{1-\eta}\right) \end{cases}$$

En remarquant que

$$\arg\max_{\mathbb{X}^*,\mathbb{Y}^*} \mathbb{P}\left(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}\right) = \arg\max_{\mathbb{X}^*,\mathbb{Y}^*} \frac{\mathbb{P}\left(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}\right)}{\mathbb{P}\left(\mathbb{X}^*, \mathbb{Y}^* \mid \mathcal{M}_0\right)}$$

On peut déduire une ≪ version automate ≫ de l'algorithme de Viterbi :

1. **initialisation :**
   $V^{\mathbf{M}}(0,0) = \ln\tau - 2\ln\eta$, $V^{\mathbf{I}_{\mathbb{X}}}(0,0) = V^{\mathbf{I}_{\mathbb{Y}}}(0,0) = -\infty$
   $V^*(i,-1) = V^*(-1,j) = -\infty$.
2. **pour** $i = 0\cdots n$ **et** $j = 0\cdots m$ **faire**

$$V^{\mathbf{M}}(i,j) = s(x_i,y_j) + \max\left\{\begin{array}{l} V^{\mathbf{M}}(i-1,j-1) \\ V^{\mathbf{I}_{\mathbb{X}}}(i-1,j-1) \\ V^{\mathbf{I}_{\mathbb{Y}}}(i-1,j-1) \end{array}\right.$$

$$V^{\mathbf{I}_{\mathbb{X}}}(i,j) = \max\left\{\begin{array}{l} V^{\mathbf{M}}(i-1,j) - d \\ V^{\mathbf{I}_{\mathbb{X}}}(i-1,j) - e \\ V^{\mathbf{I}_{\mathbb{Y}}}(i-1,j) - d - d' \end{array}\right.$$

$$V^{\mathbf{I}_{\mathbb{Y}}}(i,j) = \max\left\{\begin{array}{l} V^{\mathbf{M}}(i,j-1) - d \\ V^{\mathbf{I}_{\mathbb{Y}}}(i,j-1) - e \\ V^{\mathbf{I}_{\mathbb{X}}}(i,j-1) - d - d' \end{array}\right.$$

3. **fin :**
   $V = \max\left\{V^{\mathbf{M}}(n,m), V^{\mathbf{I}_{\mathbb{X}}}(n,m), V^{\mathbf{I}_{\mathbb{Y}}}(n,m)\right\}$

▷ On reconnaît l'algorithme de programmation dynamique classique de Needleman et Wunsch

▷ Cet algorithme fournit un alignement optimal identique au *chemin de Viterbi* dans le modèle HMM

Première conclusion :

1. A tout modèle HMM d'alignement global on peut associer un automate fini d'alignement.
2. L'alignement optimal obtenu par cet automate (Needleman et Wunsch) est le même que l'alignement de Viterbi pour le modèle HMM.

Réciproquement :

HMM associé à un Automate d'alignement

▷ A tout modèle d'alignement global classique (automate et système de score) peut-on associer un modèle HMM d'alignement ?

▷ Si ce modèle existe, l'alignement de Viterbi est-il la solution optimale (Needleman et Wunsch) du problème d'alignement pour l'automate ?

▷ Est-il possible d'écrire le système de score comme logarithme d'un rapport de vraisemblance ?

▷ Pour un système de score donné, les paramètres du modèle HMM doivent vérifier les quatre équations

$$\begin{cases} s(\alpha,\beta) &= \ln\left(\frac{p_{\alpha,\beta}}{q_\alpha q_\beta}\right) + \ln\left(\frac{1-2\delta-\tau}{(1-\eta)^2}\right) \\ d &= -\ln\left(\frac{\delta}{1-\eta}\cdot\frac{1-\epsilon-\delta'}{1-2\delta}\right) \\ e &= -\ln\left(\frac{\epsilon}{1-\eta}\right) \\ d'+d &= -\ln\left(\frac{\delta'}{1-\eta}\right) \end{cases}$$

Nous devons avoir

$$\sum_{\alpha,\beta\in\mathcal{A}} p_{\alpha,\beta} = 1$$

La première équation induit la **contrainte**

$$\sum_{\alpha,\beta\in\mathcal{A}} e^{s(\alpha,\beta)} q_\alpha q_\beta = \frac{1-2\delta-\tau}{(1-\eta)^2}$$

Les fonctions de score classiques pour les protéines (Matrices PAM, BLOSUM) sont obtenues à partir d'un modèle d'évolution et s'écrivent comme un log de rapport de vraisemblance.

La contrainte sera vérifiée en imposant la condition

$$\frac{1-2\delta-\tau}{(1-\eta)^2} = 1$$

Exemple

▷ Alignement de protéines avec une matrice PAM et le score affine pour les gaps :

$$d = +12 \quad , \quad e = 2 \quad et \quad d' = 0$$

▷ On choisit $\eta = 2.10^{-3}$ (*longueur moyenne d'une protéine* $\simeq 500$).

▷ Les équations induisent les paramètres du HMM associé

η	δ	ϵ	δ'	τ
$0,002$	$7,1.10^{-6}$	$0,135$	$6,13.10^{-6}$	$0,004$

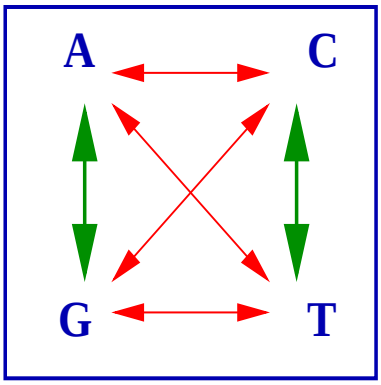
Système de Score et et modèle probabiliste d'évolution

La matrice PAM1 de similarité pour l'ADN.

Le point de départ est une matrice d'évolution 1PAM représentant une conservation de 99% des squences au cours d'une durée 1PAM d'évolution.

Modèle d'évolution 1PAM : matrice de Markov Π d'évolution

	a	g	t	c
a	0.990	0.006	0.002	0.002
g	0.006	0.990	0.002	0.002
t	0.002	0.002	0.990	0.006
c	0.002	0.002	0.006	0.990



Hypothèse : fréquences égales des nucléotides :

$$q_a = q_c = q_g = q_t = \frac{1}{4}$$

Les probabilités d'alignement :

$$p_{\alpha\beta} = \mathbb{P}(\alpha, \beta) = q_\alpha \cdot \pi_{\alpha\beta} = \frac{1}{4} \cdot \pi_{\alpha\beta}$$

Le système de score :

$$s(\alpha, \beta) = \log_2 \left(\frac{p_{\alpha\beta}}{q_\alpha \cdot q_\beta} \right) = \log_2 \left(\frac{\pi_{\alpha\beta}}{q_\beta} \right)$$

La matrice PAM1 :

	a	g	t	c
a	1.98	-5.38	-6.96	-6.96
g	-5.38	1.98	-6.96	-6.96
t	-6.96	-6.96	1.98	-5.38
c	-6.96	-6.96	-5.38	1.98

	a	g	t	c
a	2	-5	-7	-7
g	-5	2	-7	-7
t	-7	-7	2	-5
c	-7	-7	-5	2

Les matrices PAMk.

La matrice d'évolution (kPAM) :

$$(kPAM) = \Pi^k$$

La matrice PAMk

$$PAMk(\alpha, \beta) = \log_2 \left(\frac{(kPAM)_{\alpha\beta}}{q_\beta} \right)$$

Exemples :

– k=64

64PAM	a	g	t	c
a	0.5776	0.2214	0.1004	0.1004
g	0.2214	0.5776	0.1004	0.1004
t	0.1004	0.1004	0.5776	0.2214
c	0.1004	0.1004	0.2214	0.5776

PAM64	a	g	t	c
a	1.2081	-0.1751	-1.3149	-1.3149
g	-0.1751	1.2081	-1.3149	-1.3149
t	-1.3149	-1.3149	1.2081	-0.1751
c	-1.3149	-1.3149	-0.1751	1.2081

– k=128

128PAM	a	g	t	c
a	0.4028	0.2759	0.1605	0.1605
g	0.2759	0.4028	0.1605	0.1605
t	0.1605	0.1605	0.4028	0.2759
c	0.1605	0.1605	0.2759	0.4028

PAM128	a	g	t	c
a	0.6883	0.1426	-0.6386	-0.6386
g	0.1426	0.6883	-0.6386	-0.6386
t	-0.6386	-0.6386	0.6883	0.1426
c	-0.6386	-0.6386	0.1426	0.6883

– k=256

256PAM	a	g	t	c
a	0.2900	0.2739	0.2180	0.2180
g	0.2739	0.2900	0.2180	0.2180
t	0.2180	0.2180	0.2900	0.2739
c	0.2180	0.2180	0.2739	0.2900

PAM256	a	g	t	c
a	0.2142	0.1319	-0.1974	-0.1974
g	0.1319	0.2142	-0.1974	-0.1974
t	-0.1974	-0.1974	0.2142	0.1319
c	-0.1974	-0.1974	0.1319	0.2142

– k=512

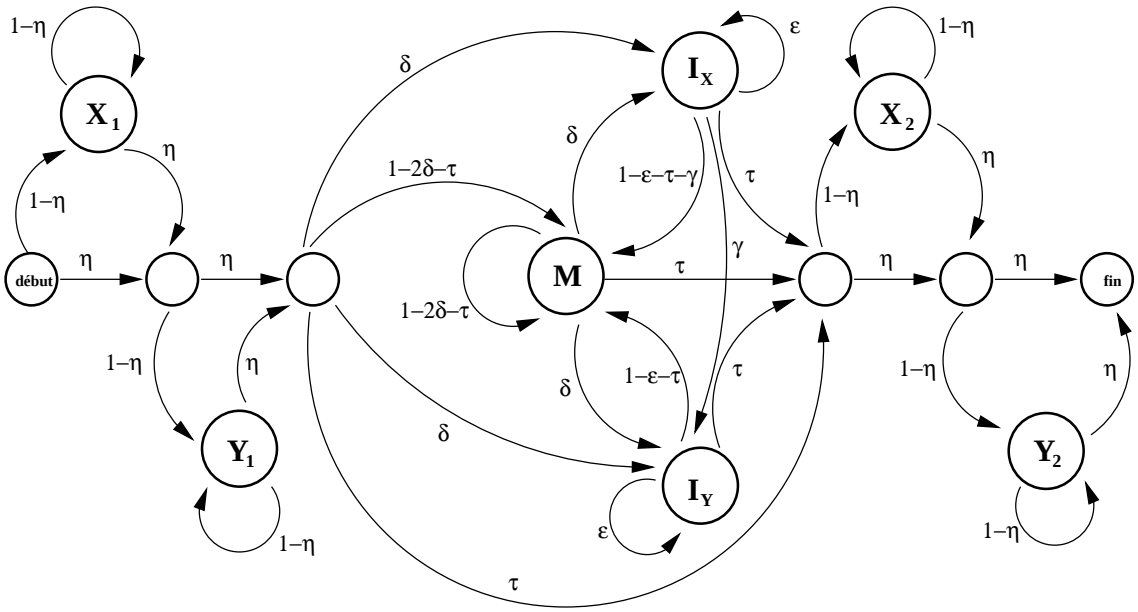
512PAM	a	g	t	c
a	0.2542	0.2539	0.2459	0.2459
g	0.2539	0.2542	0.2459	0.2459
t	0.2459	0.2459	0.2542	0.2539
c	0.2459	0.2459	0.2539	0.2542

PAM512	a	g	t	c
a	0.02415	0.02268	-0.02380	-0.02380
g	0.02268	0.02415	-0.02380	-0.02380
t	-0.02380	-0.02380	0.02415	0.02268
c	-0.02380	-0.02380	0.02268	0.02415

– k=1024

1024PAM	a	g	t	c
a	0.2500	0.2500	0.2499	0.2499
g	0.2500	0.2500	0.2499	0.2499
t	0.2499	0.2499	0.2500	0.2500
c	0.2499	0.2499	0.2500	0.2500

PAM1024	a	g	t	c
a	0.0003	0.0003	-0.0003	-0.0003
g	0.0003	0.0003	-0.0003	-0.0003
t	-0.0003	-0.0003	0.0003	0.0003
c	-0.0003	-0.0003	0.0003	0.0003


début
fin
M
X1
X2
Y1
Y2
IX
IY
1−η
η
δ
1−2δ−τ
1−ε−τ−γ
τ
γ
ε
1−ε−τ

Références :

1. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998).*Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge University Press.
2. Evens, W. J. and Grant, G. R. (2002). *Statistical methods in bioinformatics : an introduction*. Springer-Verlag.
3. Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of molecular biology*. 162, pp. 705-708.
4. Needleman, S.B. and Wunsch, C.D. (1970).A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*. 48, pp. 443-453.
5. Yu, Yi-Kuo and HWA, Terence. (2001).Statistical significance of probabilistic sequence alignment and related local hidden markov models. *Journal of Computational Biology*. Vol. 8, N. 3, pp. 249-282.