Génomique comparative et Alignement de séquences biologiques

Catherine Matias

CNRS - Laboratoire Statistique & Génome, Évry

États de la Recherche de la SMF-oct 2009

Sommaire :

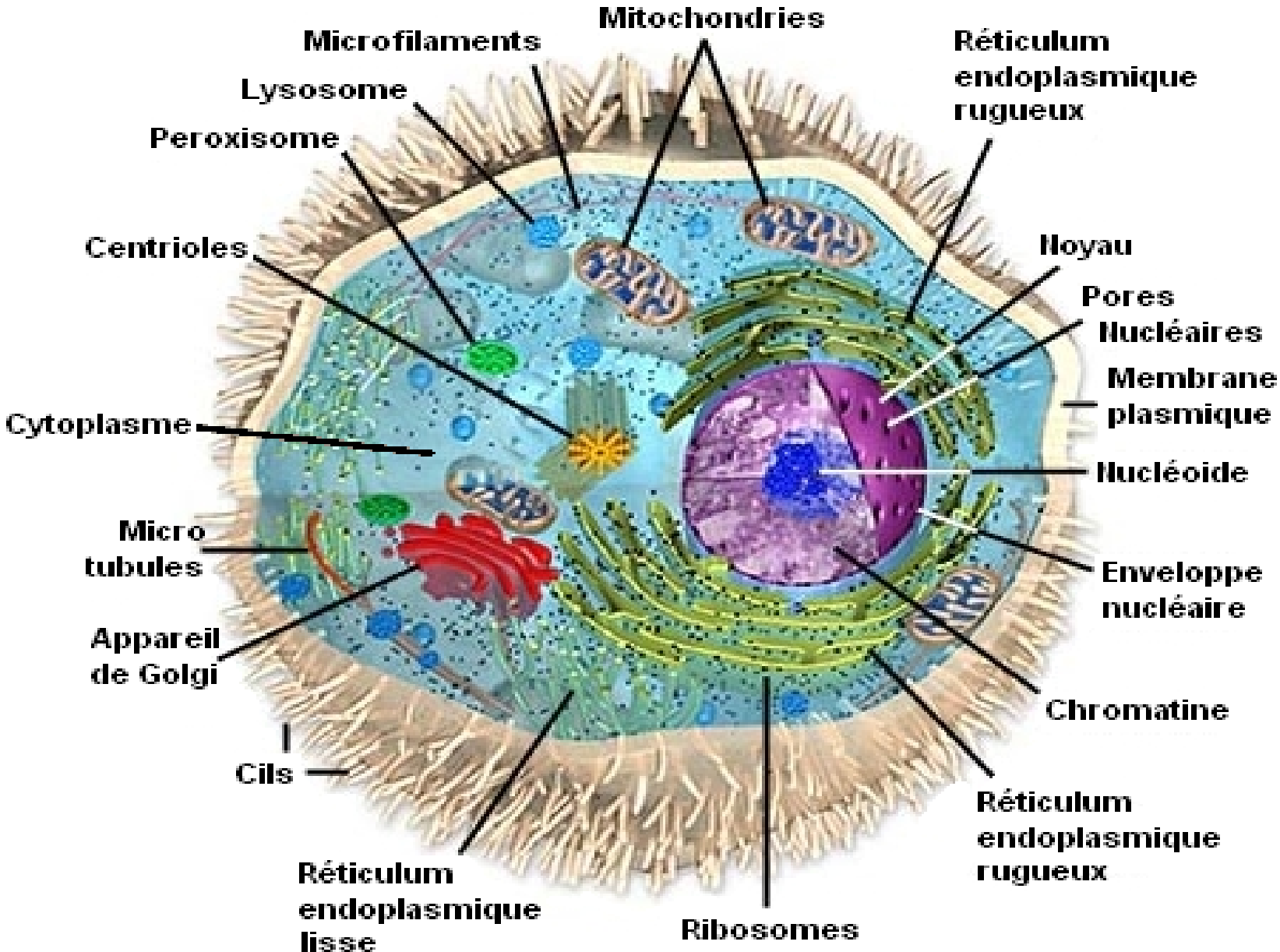
Première partie I

Introduction à la biologie moléculaire

Sommaire : Introduction à la biologie moléculaire

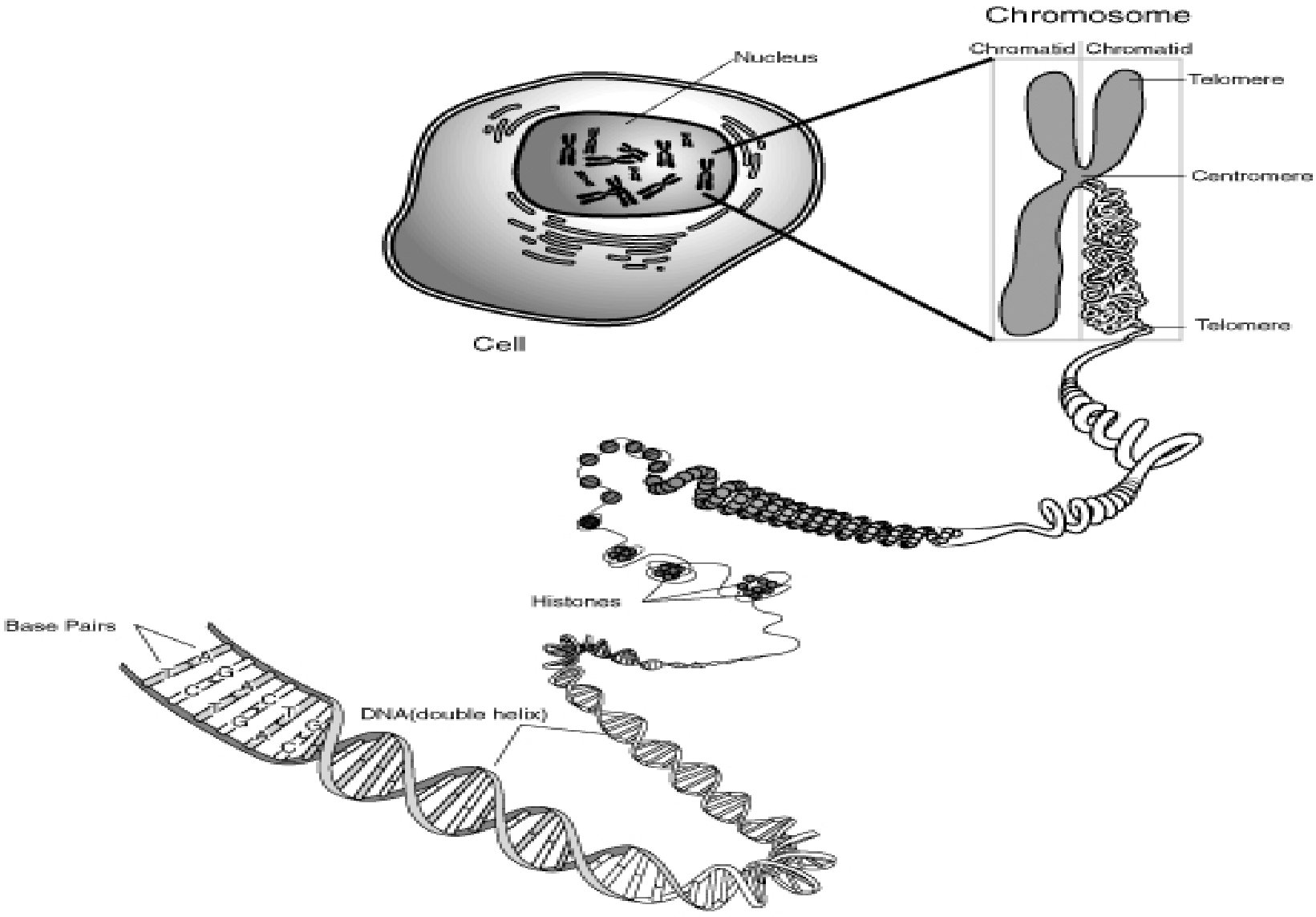
Sommaire

La cellule (avec ou sans noyau)



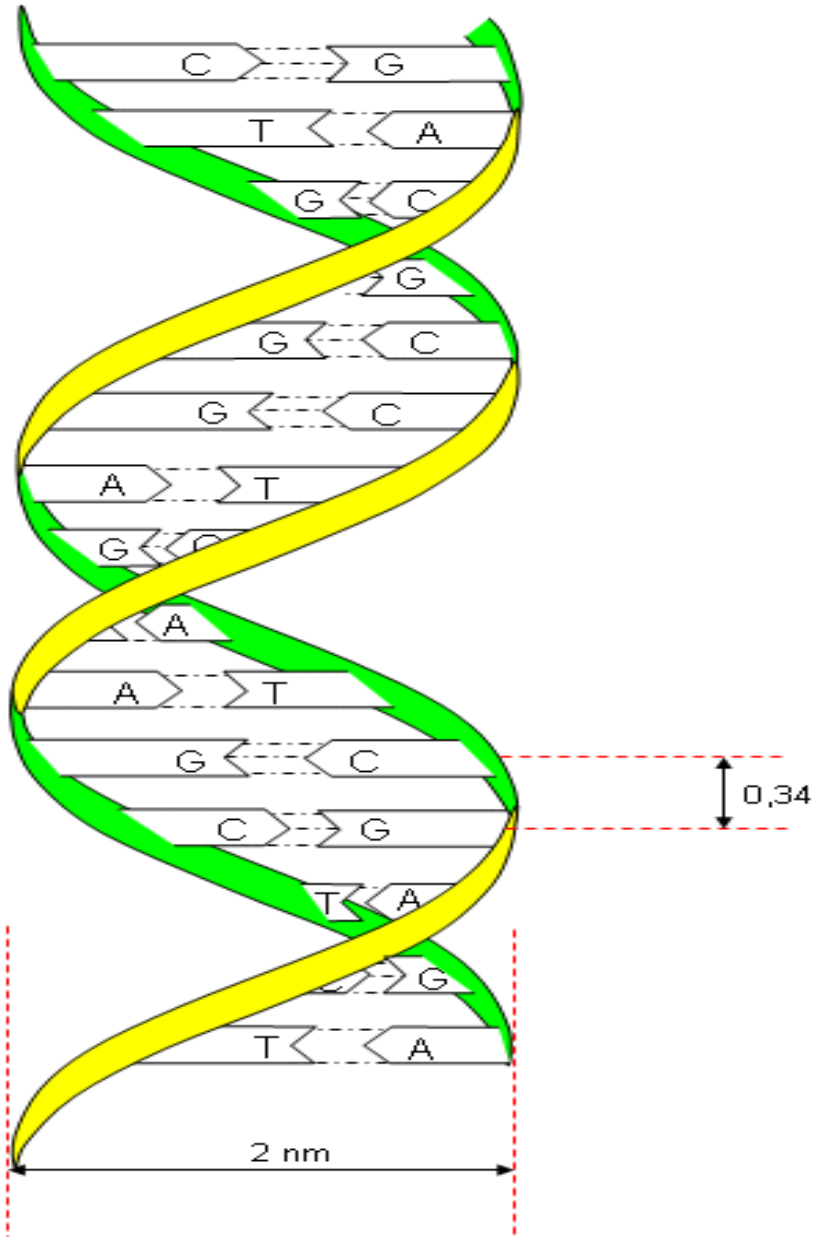
L'information génétique : les chromosomes

Chez l'Homme, 22 paires de chromosomes plus une paire de chromosomes sexuels



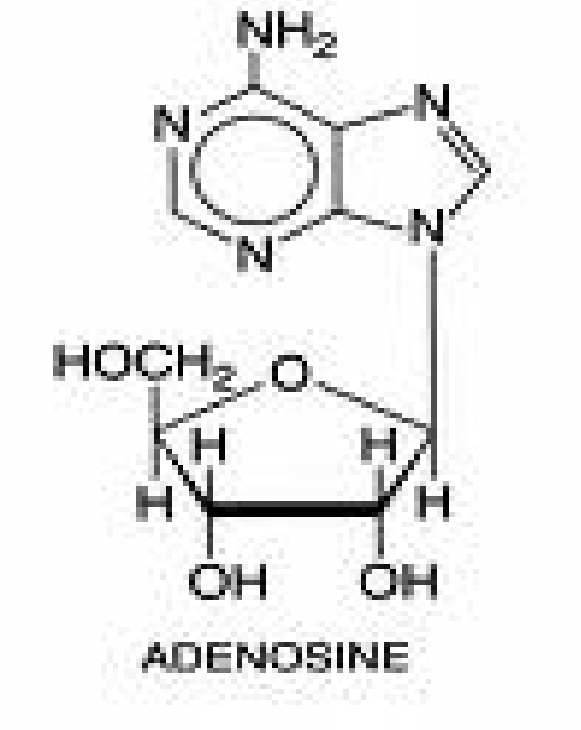
L'information génétique : l'ADN

Chez l'Homme, 3,4 milliards de *paires de bases*



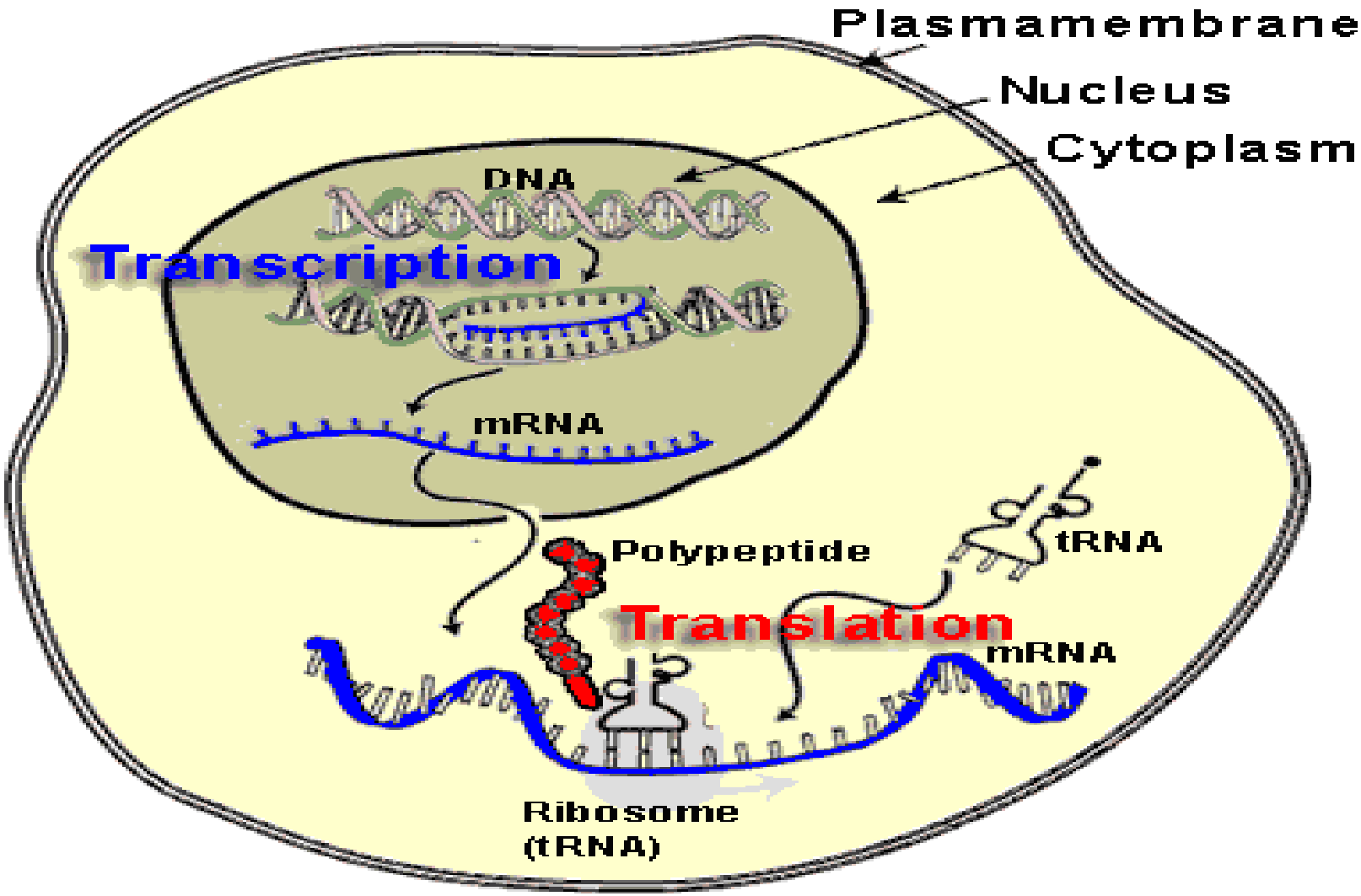
4 bases différentes : A,C, G et T

Appariement C/G et A/T



L'expression des gènes

Un gène = une petite unité d'ADN, qui code une protéine



Au final

- De nos jours, on est capables
  - de « séquencer les génomes des organismes »
  - de connaître la séquence d'une protéine
- Les séquences biologiques = soit des séquences d'ADN, soit des séquences d'acides aminés
- On ne sait rien sur la plupart des séquences (fonction, structure...)
- La génomique comparative consiste à comparer une séquence inconnue avec une base de données de séquences connues pour tenter d'en tirer de l'information

Sommaire

Génomique comparative

Définition et procédures

- Il s'agit de quantifier la similitude entre des séquences (d'ADN, de protéines).
- Les comparaisons peuvent se faire de multiple façon :
  - alignement (de portions de génomes, de génomes complets),
  - comparaison de l'ordre de certains gènes (ou de domaines),
  - comparaison de la composition des séquences en mots,
  - ...

Utilisations

- identification de sites fonctionnels,
- prédiction de fonctions,
- prédiction de structures secondaires de protéines,
- inférence de phylogénies,
- assemblages de séquences en contigs,
- ...

Génomique comparative

Définition et procédures

- Il s'agit de quantifier la similitude entre des séquences (d'ADN, de protéines).
- Les comparaisons peuvent se faire de multiple façon :
  - alignement (de portions de génomes, de génomes complets),
  - comparaison de l'ordre de certains gènes (ou de domaines),
  - comparaison de la composition des séquences en mots,
  - ...

Utilisations

- identification de sites fonctionnels,
- prédiction de fonctions,
- prédiction de structures secondaires de protéines,
- inférence de phylogénies,
- assemblages de séquences en contigs,
- ...

Sommaire

Qu'est-ce qu'un alignement ? (1/2)

- On a 2 (ou plus) séquences $X_{1:n}$ et $Y_{1:m}$ à valeurs dans le même alphabet fini $\mathcal{A}$.
- Est-ce qu'elles se « ressemblent » ?
- Un alignement c'est une correspondance entre les lettres de la première séquence et celles de la deuxième, sans en changer l'ordre, et en autorisant éventuellement des « trous ».

Exemple

$\mathcal{A} = \{A, C, G, T\}$ (les nucléotides de l'ADN),
$X_{1:9} = GAATCTGAC$, $Y_{1:6} = CACGTA$, et un alignement (global) des deux séquences est

$$\begin{array}{cccccccccc} G & A & A & T & C & - & T & G & A & C \\ C & A & - & - & C & G & T & - & A & - \end{array}$$

Qu'est-ce qu'un alignement ? (1/2)

- On a 2 (ou plus) séquences $X_{1:n}$ et $Y_{1:m}$ à valeurs dans le même alphabet fini $\mathcal{A}$.
- Est-ce qu'elles se « ressemblent » ?
- Un alignement c'est une correspondance entre les lettres de la première séquence et celles de la deuxième, sans en changer l'ordre, et en autorisant éventuellement des « trous ».

Exemple

$\mathcal{A} = \{A, C, G, T\}$ *(les nucléotides de l'ADN),*
$X_{1:9} = GAATCTGAC$, $Y_{1:6} = CACGTA$, *et un alignement (global) des deux séquences est*

$$\begin{array}{cccccccccc} G & A & A & T & C & - & T & G & A & C \\ C & A & - & - & C & G & T & - & A & - \end{array}$$

Qu'est-ce qu'un alignement ? (2/2)

Vocabulaire

- Deux lettres face à face = *match* (si ce sont les mêmes), ou *mismatch* (si les lettres sont différentes),
- une lettre en face d'un trou = indel (insertion-délétion) ou « gap ».

Premières remarques

- on peut faire de l'alignement sans autoriser les indels (lorsque les séquences sont très proches).

De plus, il existe 2 types d'alignement :

- **alignement global** : les séquences sont alignées en intégralité,
- **alignement local** : on cherche des portions des séquences qui s'alignent « bien ».

Qu'est-ce qu'un alignement ? (2/2)

Vocabulaire

- Deux lettres face à face = *match* (si ce sont les mêmes), ou *mismatch* (si les lettres sont différentes),
- une lettre en face d'un trou = indel (insertion-délétion) ou « gap ».

Premières remarques

- on peut faire de l'alignement sans autoriser les indels (lorsque les séquences sont très proches).

De plus, il existe 2 types d'alignement :

- **alignement global** : les séquences sont alignées en intégralité,
- **alignement local** : on cherche des portions des séquences qui s'alignent « bien ».

Alignement de portions de *A. tumephaciens* et *M. loti.*

Source : Hobolth, Jensen, JCB, 2005

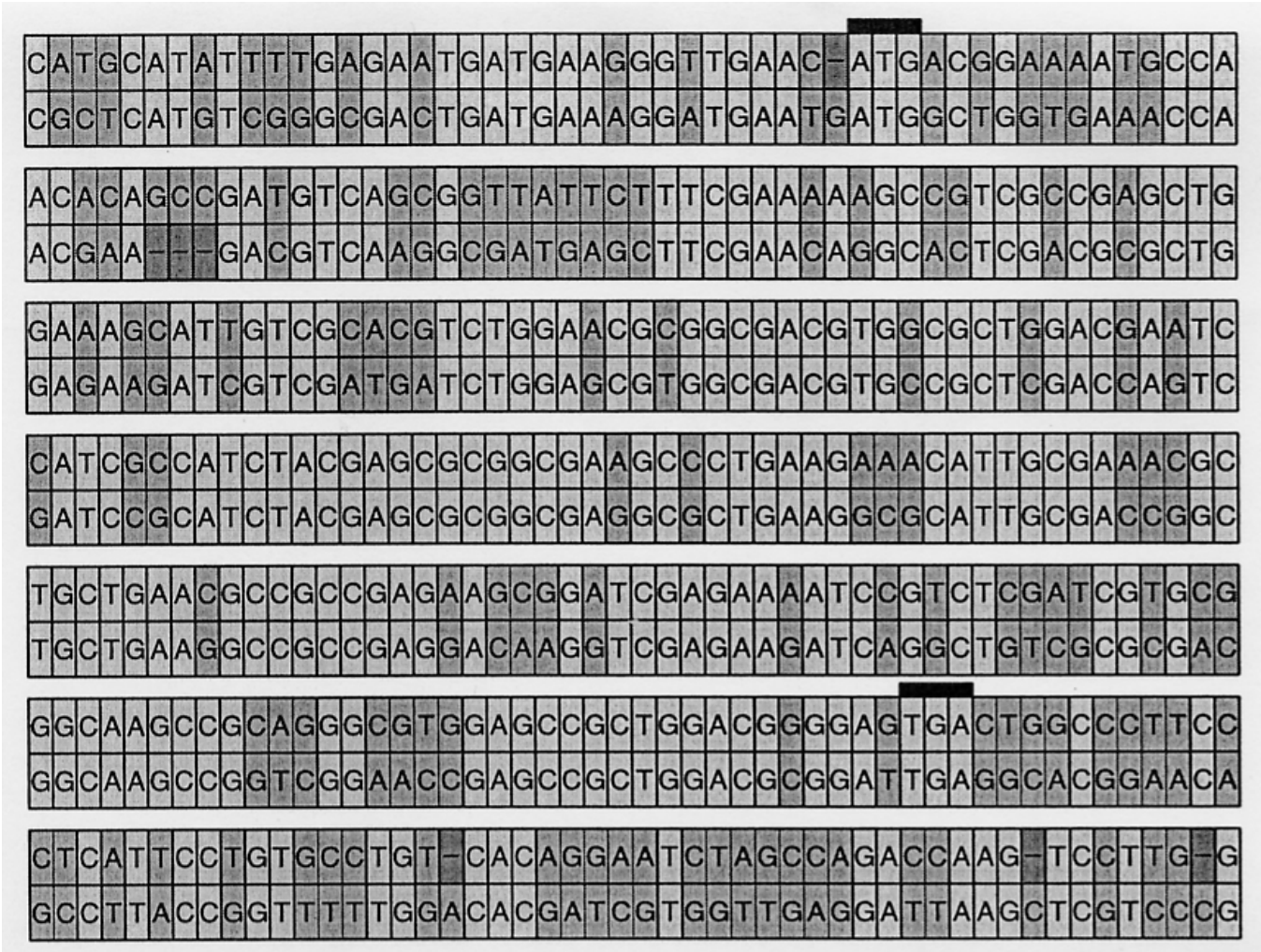


FIG. 3. Part of the pairwise alignment of *A.tumefaciens* and *M.loti.* Light gray color corresponds to conserved positions, and nonconserved positions and gaps are shown in dark gray. The two black bars on top of the alignment

Que représente un alignement ?

- Les séquences observées sont en fait issues d'un même ancêtre commun, par un processus d'évolution.
- Un processus d'évolution est constitué de modifications élémentaires (sûrement pas toutes connues à ce jour) qui sont des erreurs qui se produisent lors des réplications de l'ADN au cours du temps. Parmi les plus classiques
  - les mutations : un nucléotide (ie une lettre) est remplacé par un autre (éventuellement le même !),
  - les insertions et les délétions : un ou des nucléotides sont ajoutés ou supprimés de la séquence.
- Il y a bien sûr plein d'autres phénomènes (duplications, inversions, transferts horizontaux, ré-arrangements...) dont on ne tiendra pas compte ici.

Significativité d'un alignement

Contexte statistique

- On cherche à tester H_0 : « les deux séquences ont des distributions de lettres indépendantes » contre l'alternative H_1 : « les distributions des deux séquences sont liées ».
- Si les deux séquences dérivent du même ancêtre commun (et si cette divergence est suffisamment récente), alors cela sera détectable sur la distribution des lettres dans les séquences.

Sommaire

Représentation graphique

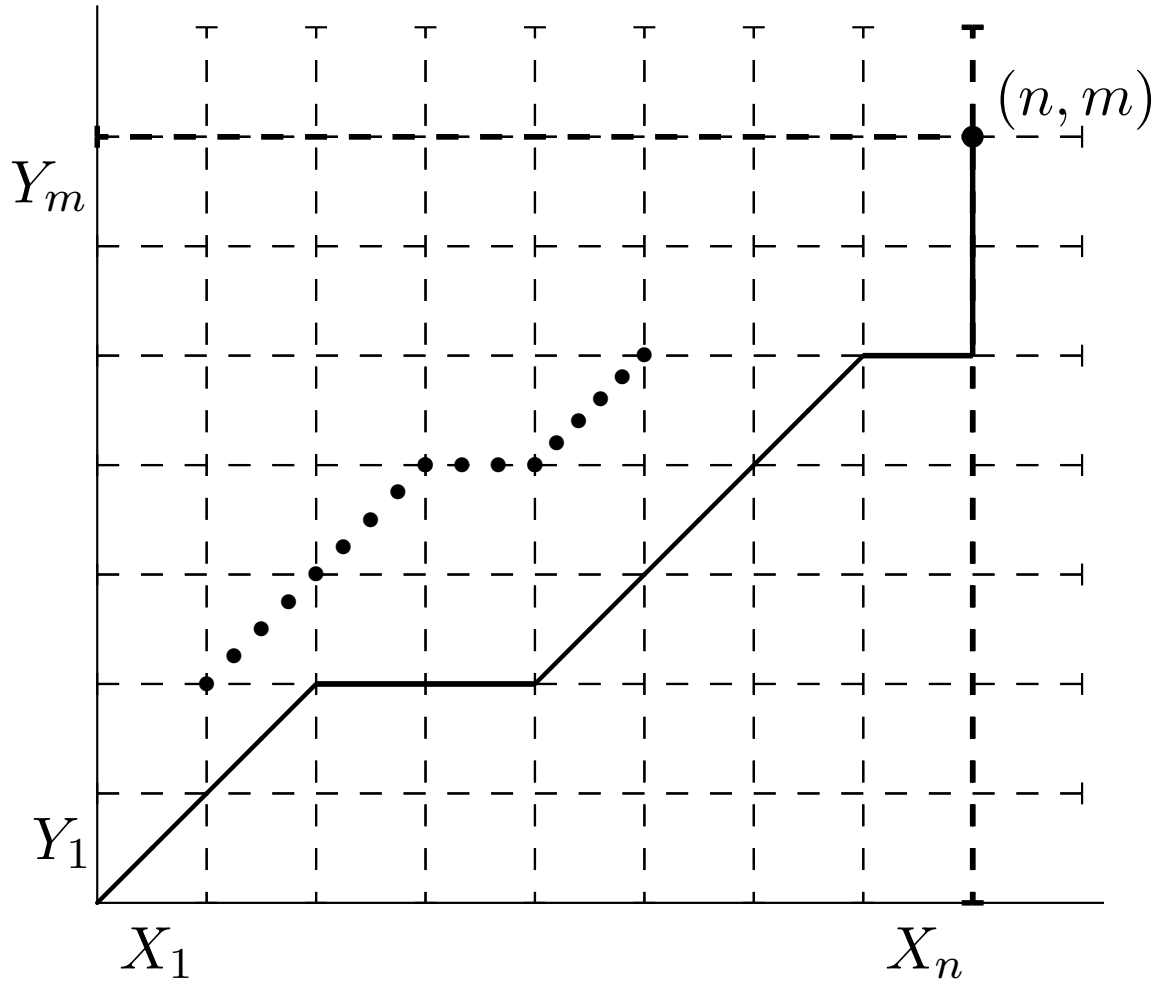


FIG.: Représentation graphique du meilleur alignement global (traits pleins) et local (traits pointillés) des séquences $X_{1:n}$ et $Y_{1:m}$.

Deuxième partie II

Alignement de deux séquences par fonction de score

Sommaire Partie II : Alignement par fonction de score

Sommaire

Alignement par fonction de score

Principe

La méthode d'alignement la plus classique consiste à utiliser une fonction de score : on attribue un certain nombre de points à chaque alignement et on sélectionne l'alignement (ou les alignements) de score le plus élevé. Ceci sous-entend qu'on est capable de calculer le score de **tous** les alignements possibles. C'est le cas pour certaines formes de score.

Quels scores ?

- Attribution des points « site par site »,
- Par exemple $+1$ pour un match, $-\mu$ pour un mismatch et $-\delta$ pour un indel ($\mu, \delta > 0$), puis on somme sur toutes les positions de l'alignement.
- Plus généralement, on considère une matrice de scores sur $\mathcal{A} \times \mathcal{A}$ qui attribue le score $s(a, b)$ à l'alignement de la lettre a en face de la lettre b.
- Pénalisation affine ou linéaire sur les indels : $-\Delta - \delta k$ où k est la longueur de l'indel et $\Delta \geq 0$ représente le coût de l'ouverture du « gap », alors que $\delta > 0$ représente le coût de l'agrandissement du « gap ».

Quels scores ?

- Attribution des points « site par site »,
- Par exemple $+1$ pour un match, $-\mu$ pour un mismatch et $-\delta$ pour un indel ($\mu, \delta > 0$), puis on somme sur toutes les positions de l'alignement.
- Plus généralement, on considère une matrice de scores sur $\mathcal{A} \times \mathcal{A}$ qui attribue le score $s(a, b)$ à l'alignement de la lettre a en face de la lettre b.
- Pénalisation affine ou linéaire sur les indels : $-\Delta - \delta k$ où k est la longueur de l'indel et $\Delta \geq 0$ représente le coût de l'ouverture du « gap », alors que $\delta > 0$ représente le coût de l'agrandissement du « gap ».

Formalisation mathématique (1/4)

- Le score d'alignement est une généralisation (non triviale) du score sur une seule séquence.
- Le score sur une séquence est un objet très général (utilisé par exemple pour la détection de zones d'intérêt).

Score sur une séquence

- On observe $X_1, \ldots, X_n$ i.i.d. (éventuellement Markov) de loi $\mathbb{P}^n$ à valeurs dans l'alphabet fini $\mathcal{A}$,
- On a une fonction de score $s : \mathcal{A} \to \mathbb{R}$ qui attribue des points à chaque lettre.
- Le score local $H_{i,j}$ de la portion de séquence entre les positions i et j est la somme des scores de chacune des lettres et le score local optimal M_n est le plus grand score local.

$$H_{i,j} = \sum_{k=i}^{j} s(X_k) \quad , \quad M_n = \max_{1 \leq i < j \leq n} H_{i,j}.$$

Formalisation mathématique (1/4)

- Le score d'alignement est une généralisation (non triviale) du score sur une seule séquence.
- Le score sur une séquence est un objet très général (utilisé par exemple pour la détection de zones d'intérêt).

Score sur une séquence

- On observe $X_1, \ldots, X_n$ i.i.d. (éventuellement Markov) de loi $\mathbb{P}^n$ à valeurs dans l'alphabet fini $\mathcal{A}$,
- On a une fonction de score $s : \mathcal{A} \to \mathbb{R}$ qui attribue des points à chaque lettre.
- Le score local $H_{i,j}$ de la portion de séquence entre les positions i et j est la somme des scores de chacune des lettres et le score local optimal M_n est le plus grand score local.

$$H_{i,j} = \sum_{k=i}^{j} s(X_k) \quad , \quad M_n = \max_{1 \leq i < j \leq n} H_{i,j}.$$

Formalisation mathématique (2/4)

Score d'alignement de 2 séquences. Cas sans indels

Soient $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ deux séquences i.i.d. à valeurs dans $\mathcal{A}$ et $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ une fonction de score. Le score local et le score local optimal se définissent alors respectivement comme

$$\begin{aligned} H_{(i,j),\ell} &= \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k}), \\ M_{n,m} &= \max_{\ell \geq 1} \max_{1 \leq i \leq n-\ell} \max_{1 \leq j \leq m-\ell} H_{(i,j),\ell}. \end{aligned}$$

Le score global d'alignement (pour deux séquences de même longueur n) est simplement la quantité $H_{(0,0),n}$, c'est-à-dire une simple somme de variables supposées i.i.d.

Formalisation mathématique (3/4)

Score d'alignement de 2 séquences. Cas avec indels

- Un alignement peut être décrit par un ensemble de positions alignées : $\{(i_k, j_k), 1 \leq k \leq \ell\}$, avec les contraintes

$$\begin{aligned}\text{al.} = \{&1 \leq i_1 < \cdots < i_\ell \leq n, \quad 1 \leq j_1 < \cdots < j_\ell \leq m \\ &\forall 1 \leq k \leq \ell, \quad i_{k+1} = i_k + 1 \text{ ou } j_{k+1} = j_k + 1\}.\end{aligned}$$

- Soient $I = \{i+1, \ldots, i+s\}$ et $J = \{j+1, \ldots, j+t\}$ deux ensembles d'indices consécutifs. On définit

$$H(I,J) = \max_{\text{al.}} \left\{ \sum_{k=1}^{\ell} s(X_{i_k}, Y_{j_k}) - \delta(s - \ell + t - \ell) \right\},$$

pour une pénalité linéaire.

- Le score local optimal entre les deux séquences est

$$M_{n,m} = \max_{I,J} H(I,J)$$

Formalisation mathématique (3/4)

Score d'alignement de 2 séquences. Cas avec indels

- Un alignement peut être décrit par un ensemble de positions alignées : $\{(i_k, j_k), 1 \leq k \leq \ell\}$, avec les contraintes

$$\begin{aligned}\text{al.} = \{&1 \leq i_1 < \cdots < i_\ell \leq n, \quad 1 \leq j_1 < \cdots < j_\ell \leq m \\ &\forall 1 \leq k \leq \ell, \quad i_{k+1} = i_k + 1 \text{ ou } j_{k+1} = j_k + 1\}.\end{aligned}$$

- Soient $I = \{i+1, \ldots, i+s\}$ et $J = \{j+1, \ldots, j+t\}$ deux ensembles d'indices consécutifs. On définit

$$H(I, J) = \max_{\text{al.}} \left\{ \sum_{k=1}^{\ell} s(X_{i_k}, Y_{j_k}) - \delta(s - \ell + t - \ell) \right\},$$

pour une pénalité linéaire.

- Le score local optimal entre les deux séquences est

$$M_{n,m} = \max_{I,J} H(I, J)$$

Formalisation mathématique (3/4)

Score d'alignement de 2 séquences. Cas avec indels

- Un alignement peut être décrit par un ensemble de positions alignées : $\{(i_k, j_k), 1 \leq k \leq \ell\}$, avec les contraintes

$$\begin{aligned}\text{al.} = \{&1 \leq i_1 < \cdots < i_\ell \leq n, \quad 1 \leq j_1 < \cdots < j_\ell \leq m \\ &\forall 1 \leq k \leq \ell, \quad i_{k+1} = i_k + 1 \text{ ou } j_{k+1} = j_k + 1\}.\end{aligned}$$

- Soient $I = \{i+1, \ldots, i+s\}$ et $J = \{j+1, \ldots, j+t\}$ deux ensembles d'indices consécutifs. On définit

$$H(I, J) = \max_{\text{al.}} \left\{ \sum_{k=1}^{\ell} s(X_{i_k}, Y_{j_k}) - \delta(s - \ell + t - \ell) \right\},$$

pour une pénalité linéaire.

- Le score local optimal entre les deux séquences est

$$M_{n,m} = \max_{I,J} H(I, J)$$

Sommaire

Algorithmes exacts

- Needleman et Wunsch pour l'alignement global [NW70], amélioré plus tard par Gotoh [Got82].
- Smith et Waterman [SW81] pour l'alignement local.
- Tous deux basés sur de la programmation dynamique (et donc utilisant la forme additive du score).

Algorithmes approchés

- L'algorithme de Smith et Waterman est trop lent si on veut comparer une séquence à toute une base de données.
- Des heuristiques existent pour accélérer ces procédures, par exemple en utilisant une première recherche rapide de segments identiques (points d'ancrage) à partir desquels on cherche à étendre l'alignement.
- voir BLAST, FASTA...

Sommaire

Matrices de comparaison (1/2)

- Le choix de la fonction $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ pose problème. [C'est aussi le cas de la pénalité pour les indels, mais les algorithmes existants limitent ce choix à des fonctions affines en la longueur de l'indel.]
- Pour $\mathcal{A} = \{A, T, G, C\}$, on utilise souvent soit une matrice identité, soit deux valeurs de score différentes : $s(X, X) = s(Y, Y) \neq s(X, Y)$ en fonction des groupes purines $X = \{A, G\}$ / pyrimidines $Y = \{C, T\}$.
- Pour $\mathcal{A} = \{\text{acides aminés}\}$ (taille 20), il existe deux grandes familles de matrices de comparaison de protéines
  - PAM ("Percent Accepted Mutations"), voir [DSO78].
  - BLOSUM ("Blocks Substitution Matrix"), voir [HH92].
  - Se distinguent par les méthodes par lesquelles elles ont été obtenues, mais basées toutes deux sur le principe des « log-odds ratios ».

Matrices de comparaison (1/2)

- Le choix de la fonction $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ pose problème. [C'est aussi le cas de la pénalité pour les indels, mais les algorithmes existants limitent ce choix à des fonctions affines en la longueur de l'indel.]
- Pour $\mathcal{A} = \{A, T, G, C\}$, on utilise souvent soit une matrice identité, soit deux valeurs de score différentes : $s(X, X) = s(Y, Y) \neq s(X, Y)$ en fonction des groupes purines $X = \{A, G\}$ / pyrimidines $Y = \{C, T\}$.
- Pour $\mathcal{A} = \{\text{acides aminés}\}$ (taille 20), il existe deux grandes familles de matrices de comparaison de protéines
  - PAM ("Percent Accepted Mutations"), voir [DSO78].
  - BLOSUM ("Blocks Substitution Matrix"), voir [HH92].
  - Se distinguent par les méthodes par lesquelles elles ont été obtenues, mais basées toutes deux sur le principe des « log-odds ratios ».

Matrices de comparaison (1/2)

- Le choix de la fonction $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ pose problème. [C'est aussi le cas de la pénalité pour les indels, mais les algorithmes existants limitent ce choix à des fonctions affines en la longueur de l'indel.]
- Pour $\mathcal{A} = \{A, T, G, C\}$, on utilise souvent soit une matrice identité, soit deux valeurs de score différentes : $s(X, X) = s(Y, Y) \neq s(X, Y)$ en fonction des groupes purines $X = \{A, G\}$ / pyrimidines $Y = \{C, T\}$.
- Pour $\mathcal{A} = \{\text{acides aminés}\}$ (taille 20), il existe deux grandes familles de matrices de comparaison de protéines
  - PAM ("Percent Accepted Mutations"), voir [DSO78].
  - BLOSUM ("Blocks Substitution Matrix"), voir [HH92].
  - Se distinguent par les méthodes par lesquelles elles ont été obtenues, mais basées toutes deux sur le principe des « log-odds ratios ».

Matrices de comparaison (2/2)

Alternative

Une solution à ce problème c'est de ne pas faire de l'alignement par score, mais par maximum de vraisemblance (voir Alignement statistique).

Sommaire

Introduction aux propriétés du score local

Dans la suite, $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes. On suppose $m = n$ par commodité uniquement.

Croissance linéaire du score global

Soit S_n le score maximal d'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Théorème (de sous-additivité de Kingman, 1968)

On a $S_{n+m} \geq S_n + S_m$ *donc* $\lim_{n\to\infty} n^{-1} S_n$ *existe p.s. et dans* $\mathbb{L}_1$ *vers a (score moyen par position).*

Introduction aux propriétés du score local

Dans la suite, $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes. On suppose $m = n$ par commodité uniquement.

Croissance linéaire du score global

Soit S_n le score maximal d'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Théorème (de sous-additivité de Kingman, 1968)

On a $S_{n+m} \geq S_n + S_m$ *donc* $\lim_{n\to\infty} n^{-1} S_n$ *existe p.s. et dans* $\mathbb{L}_1$ *vers a (score moyen par position).*

Introduction aux propriétés du score local

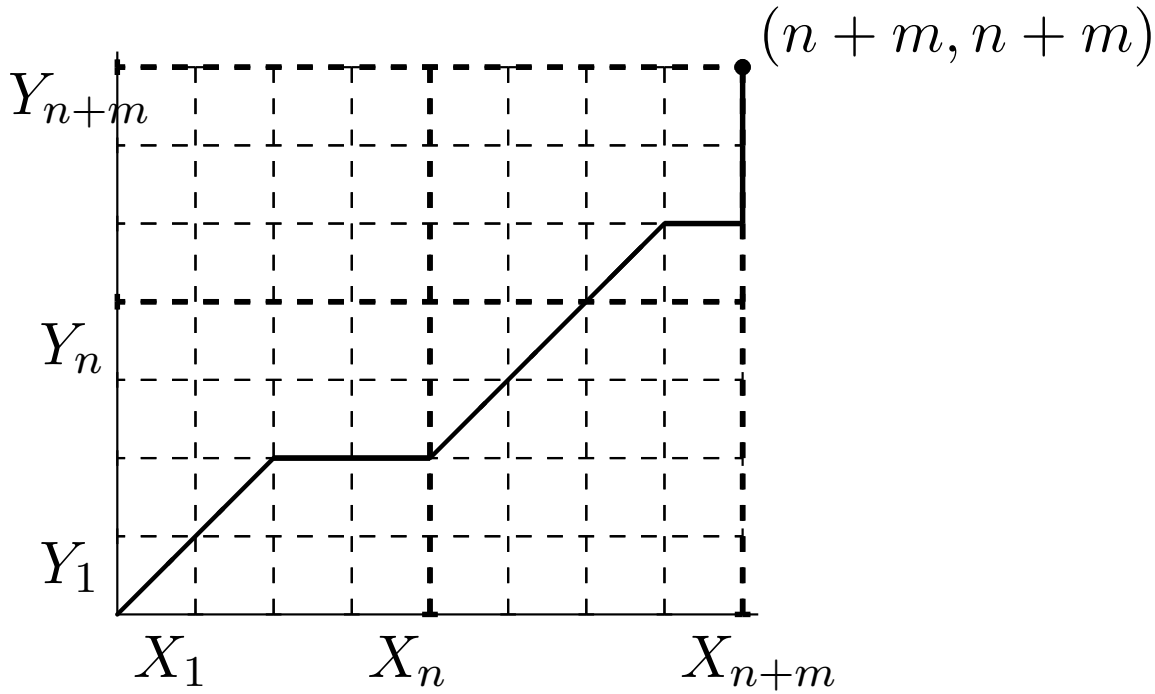
Dans la suite, $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes. On suppose $m = n$ par commodité uniquement.

Croissance linéaire du score global

Soit S_n le score maximal d'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Théorème (de sous-additivité de Kingman, 1968)

On a $S_{n+m} \geq S_n + S_m$ *donc* $\lim_{n\to\infty} n^{-1} S_n$ *existe p.s. et dans* $\mathbb{L}_1$ *vers a (score moyen par position).*



Introduction aux propriétés du score local

Dans la suite, $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes. On suppose $m = n$ par commodité uniquement.

Croissance linéaire du score global

Soit S_n le score maximal d'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Théorème (de sous-additivité de Kingman, 1968)

On a $S_{n+m} \geq S_n + S_m$ *donc* $\lim_{n\to\infty} n^{-1} S_n$ *existe p.s. et dans* $\mathbb{L}_1$ *vers a (score moyen par position).*

Heuristique pour le score local

- Si $a > 0$, le score local optimal M_n croît linéairement avec n ; les régions de score optimal ont une longueur de l'ordre de n.
- Si $a < 0$, les régions de score optimal sont sous un régime de grande déviation. Leur longueur est de l'ordre de $\log n$ et M_n croît comme $\log n$.

Changement de régime - premiers résultats

Arratia & Waterman [AW94]

- Considèrent le score suivant : $+1$ si match, $-\mu$ si mismatch, $-\delta$ pour chaque indel.
- Prouvent l'existence d'un changement de régime du score local optimal : suivant les valeurs des paramètres, M_n croît linéairement ou logarithmiquement en n.
- Pas de caractérisation explicite du point de changement de phase (comment choisir δ et μ pour être sûr d'être dans le régime logarithmique ?).

Comportement dans le régime logarithmique - premiers résultats

Zhang [Zha95]

Dans le même cadre qu'Arratia et Waterman, sous le régime logarithmique, on a

$$\frac{M_n}{\log n} \to 2b \text{ p.s}$$

et la limite b vérifie :

$$b = \max_{q \geq 0} \frac{q}{r(q)}; \quad r(q) = \lim_n \frac{-\log \mathbb{P}(S_n \geq qn)}{n}$$

où S_n = score de l'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Vers la construction d'une p-value

- Comportement asymptotique de $\mathbb{P}_{H_0}(M_n \geq t \log n)$?

Comportement dans le régime logarithmique - premiers résultats

Zhang [Zha95]

Dans le même cadre qu'Arratia et Waterman, sous le régime logarithmique, on a

$$\frac{M_n}{\log n} \to 2b \text{ p.s}$$

et la limite b vérifie :

$$b = \max_{q \geq 0} \frac{q}{r(q)}; \quad r(q) = \lim_n \frac{-\log \mathbb{P}(S_n \geq qn)}{n}$$

où S_n = score de l'alignement global de $X_{1:n}$ et $Y_{1:n}$.

Vers la construction d'une p-value

- Comportement asymptotique de $\mathbb{P}_{H_0}(M_n \geq t \log n)$?

Propriétés du score local sans indels (1/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

- Soient $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes, de distributions respectives μ_X et μ_Y.
- Soit $s = (s(x,y))_{x,y\in\mathcal{A}}$ matrice de score. On suppose

$$\mathbb{E}(s(X,Y)) = \sum_{(x,y)\in\mathcal{A}^2} s(x,y)\mu_X(x)\mu_Y(y) < 0$$

$$\mathbb{P}(s(X,Y) > 0) = \sum_{(x,y)\in\mathcal{A}^2} 1_{\{s(x,y)>0\}}\mu_X(x)\mu_Y(y) > 0.$$

- Ces hypothèses assurent le régime logarithmique.
- On suppose de plus une condition d'entropie.
- Rappel, le score local optimal (sans indels) d'alignement de $X_{1:n}$ et $Y_{1:n}$ est

$$M_n = \max_{\substack{1\le i,j\le n-\ell \\ \ell\ge 1}} \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k})$$

Propriétés du score local sans indels (1/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

- Soient $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes, de distributions respectives μ_X et μ_Y.
- Soit $s = (s(x,y))_{x,y\in\mathcal{A}}$ matrice de score. On suppose

$$\mathbb{E}(s(X,Y)) = \sum_{(x,y)\in\mathcal{A}^2} s(x,y)\mu_X(x)\mu_Y(y) < 0$$

$$\mathbb{P}(s(X,Y) > 0) = \sum_{(x,y)\in\mathcal{A}^2} 1_{\{s(x,y)>0\}}\mu_X(x)\mu_Y(y) > 0.$$

- Ces hypothèses assurent le régime logarithmique.
- On suppose de plus une condition d'entropie.
- Rappel, le score local optimal (sans indels) d'alignement de $X_{1:n}$ et $Y_{1:n}$ est

$$M_n = \max_{\substack{1\le i,j\le n-\ell \\ \ell\ge 1}} \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k})$$

Propriétés du score local sans indels (1/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

- Soient $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes, de distributions respectives μ_X et μ_Y.
- Soit $s = (s(x,y))_{x,y\in\mathcal{A}}$ matrice de score. On suppose

$$\mathbb{E}(s(X,Y)) = \sum_{(x,y)\in\mathcal{A}^2} s(x,y)\mu_X(x)\mu_Y(y) < 0$$

$$\mathbb{P}(s(X,Y) > 0) = \sum_{(x,y)\in\mathcal{A}^2} 1_{\{s(x,y)>0\}}\mu_X(x)\mu_Y(y) > 0.$$

- Ces hypothèses assurent le régime logarithmique.
- On suppose de plus une condition d'entropie.
- Rappel, le score local optimal (sans indels) d'alignement de $X_{1:n}$ et $Y_{1:n}$ est

$$M_n = \max_{\substack{1\le i,j\le n-\ell \\ \ell\ge 1}} \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k})$$

Propriétés du score local sans indels (1/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

- Soient $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes, de distributions respectives μ_X et μ_Y.
- Soit $s = (s(x,y))_{x,y\in\mathcal{A}}$ matrice de score. On suppose

$$\mathbb{E}(s(X,Y)) = \sum_{(x,y)\in\mathcal{A}^2} s(x,y)\mu_X(x)\mu_Y(y) < 0$$

$$\mathbb{P}(s(X,Y) > 0) = \sum_{(x,y)\in\mathcal{A}^2} 1_{\{s(x,y)>0\}}\mu_X(x)\mu_Y(y) > 0.$$

- Ces hypothèses assurent le régime logarithmique.
- On suppose de plus une condition d'entropie.
- Rappel, le score local optimal (sans indels) d'alignement de $X_{1:n}$ et $Y_{1:n}$ est

$$M_n = \max_{\substack{1\le i,j\le n-\ell \\ \ell\ge 1}} \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k})$$

Propriétés du score local sans indels (1/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

- Soient $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes, de distributions respectives μ_X et μ_Y.
- Soit $s = (s(x,y))_{x,y\in\mathcal{A}}$ matrice de score. On suppose

$$\mathbb{E}(s(X,Y)) = \sum_{(x,y)\in\mathcal{A}^2} s(x,y)\mu_X(x)\mu_Y(y) < 0$$

$$\mathbb{P}(s(X,Y) > 0) = \sum_{(x,y)\in\mathcal{A}^2} 1_{\{s(x,y)>0\}}\mu_X(x)\mu_Y(y) > 0.$$

- Ces hypothèses assurent le régime logarithmique.
- On suppose de plus une condition d'entropie.
- Rappel, le score local optimal (sans indels) d'alignement de $X_{1:n}$ et $Y_{1:n}$ est

$$M_n = \max_{\substack{1\le i,j\le n-\ell \\ \ell\ge 1}} \sum_{k=1}^{\ell} s(X_{i+k}, Y_{j+k})$$

Propriétés du score local sans indels (2/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

Théorème

Il existe des constantes $\theta^\star, K^\star > 0$ telles que le score local d'alignement sans indels M_n vérifie

$$\lim_{n\to+\infty} \mathbb{P}\left(M_n - \frac{2\log n}{\theta^\star} \leq t\right) = \exp(-K^\star \exp(-\theta^\star t)).$$

Remarques

- Généralisation facile à $n \neq m$.
- Les résultats de Dembo et al. sont étendus au cas des chaînes de Markov par Hansen [Han06].
- La loi limite est une loi des valeurs extrêmes de type I (ou loi de Gumbel) et apparaît dans la théorie des valeurs extrêmes.

Propriétés du score local sans indels (2/2)

Dembo, Karlin et Zeitouni [DKZ94b, DKZ94a]

Théorème

Il existe des constantes $\theta^\star, K^\star > 0$ telles que le score local d'alignement sans indels M_n vérifie

$$\lim_{n\to+\infty} \mathbb{P}\left(M_n - \frac{2\log n}{\theta^\star} \leq t\right) = \exp(-K^\star \exp(-\theta^\star t)).$$

Remarques

- Généralisation facile à $n \neq m$.
- Les résultats de Dembo et al. sont étendus au cas des chaînes de Markov par Hansen [Han06].
- La loi limite est une loi des valeurs extrêmes de type I (ou loi de Gumbel) et apparaît dans la théorie des valeurs extrêmes.

Propriétes du score local avec indels

Questions ouvertes

- Comment caractériser la transition de phase ?
- Quelle est la limite de $M_n / \log n$ et ses déviations dans le régime logarithmique ?

En pratique

- On utilise la forme de la queue de distribution du score sans indels même dans le cas avec indels. Cependant, il n'existe que des résultats partiels ou des heuristiques.
- Comment calculer en pratique les constantes $K^\star$ et $\theta^\star$ qui caractérisent la queue de distribution du score local optimal sans indels ?

Propriétes du score local avec indels

Questions ouvertes

- Comment caractériser la transition de phase ?
- Quelle est la limite de $M_n / \log n$ et ses déviations dans le régime logarithmique ?

En pratique

- On utilise la forme de la queue de distribution du score sans indels même dans le cas avec indels. Cependant, il n'existe que des résultats partiels ou des heuristiques.
- Comment calculer en pratique les constantes $K^\star$ et $\theta^\star$ qui caractérisent la queue de distribution du score local optimal sans indels ?

Résultats approchés/heuristiques pour le score local avec indels

Siegmund et Yakir [SY00b, SY00a]

- Approximation de la p-valeur si
  - le nombre maximum de trous est fixé,
  - ou bien le coût d'ouverture Δ d'un gap croît comme $\log n$.

Quelques références supplémentaires

- Grossmann et Yakir [GY04, Gro03] : inég. gdes déviations sur les scores global et local optimaux (avec indel).
- Chan [Cha03, Cha05] : « importance sampling » sur les p-valeurs + condition suffisante explicite sur les paramètres de la fonction de score pour assurer le régime log.
- Mott et Tribe [MT99] fournissent une méthode heuristique pour approcher la valeur de $\theta^\star$ et prédire le point de changement de transition de phase.
- Bibliographie non exhaustive !

Résultats approchés/heuristiques pour le score local avec indels

Siegmund et Yakir [SY00b, SY00a]

- Approximation de la p-valeur si
  - le nombre maximum de trous est fixé,
  - ou bien le coût d'ouverture Δ d'un gap croît comme $\log n$.

Quelques références supplémentaires

- Grossmann et Yakir [GY04, Gro03] : inég. gdes déviations sur les scores global et local optimaux (avec indel).
- Chan [Cha03, Cha05] : « importance sampling » sur les p-valeurs + condition suffisante explicite sur les paramètres de la fonction de score pour assurer le régime log.
- Mott et Tribe [MT99] fournissent une méthode heuristique pour approcher la valeur de $\theta^\star$ et prédire le point de changement de transition de phase.
- Bibliographie non exhaustive !

Lien avec la percolation de premier passage

Pour plus de détails, voir [Gro03]

- L'alignement optimal peut être vu comme un problème de percolation de premier passage $\longrightarrow$ problème mathématiquement difficile.
- Cependant, les questions qui se posent en percolation et en alignement optimal diffèrent.

Troisième partie III

Alignement statistique

Sommaire Partie III : Alignement statistique

Sommaire

Contexte

Alignement classique vs Alignement statistique

- Les fonctions de score traduisent l'évolution sous-jacente des séquences, et leur choix a priori introduit un biais dans le résultat.
- L'alignement statistique pallie à ce problème, en réalisant à la fois l'alignement des séquences et l'estimation des paramètres du modèle d'évolution sous-jacent.
- En pratique, l'alignement de deux séquences est réalisé par maximisation d'un critère de vraisemblance, dans un contexte de paires de séquences Markov caché.

Introduction à l'alignement statistique

Principe

On considère un modèle d'évolution (particulier) sur les séquences (avec des paramètres inconnus). On observe deux séquences, et on cherche à reconstruire leur « vrai alignement » (i.e. les positions homologues et les indels à partir desquels les séquences ont évolué) en maximisant leur vraisemblance sous ce modèle d'évolution.

Cadre

- Les modèles d'évolution qui permettent cette approche sont ceux introduits pas Thorne, Kishino et Felsenstein ([TKF91] et [TKF92]), ou encore des variantes [MLH04].
- Pour ces modèles d'évolution, le problème s'exprime dans le cadre des « pair-HMM ».
- L'avantage d'avoir un modèle probabiliste c'est qu'on peut non seulement faire de l'inférence, mais aussi des tests d'hypothèses...

Introduction à l'alignement statistique

Principe

On considère un modèle d'évolution (particulier) sur les séquences (avec des paramètres inconnus). On observe deux séquences, et on cherche à reconstruire leur « vrai alignement » (i.e. les positions homologues et les indels à partir desquels les séquences ont évolué) en maximisant leur vraisemblance sous ce modèle d'évolution.

Cadre

- Les modèles d'évolution qui permettent cette approche sont ceux introduits pas Thorne, Kishino et Felsenstein ([TKF91] et [TKF92]), ou encore des variantes [MLH04].
- Pour ces modèles d'évolution, le problème s'exprime dans le cadre des « pair-HMM ».
- L'avantage d'avoir un modèle probabiliste c'est qu'on peut non seulement faire de l'inférence, mais aussi des tests d'hypothèses...

Sommaire

Le modèle d'évolution TKF (1/2)

Modèle d'évolution

- Chaque site évolue indépendamment et peut subir une substitution ou être effacé.
- Les insertions (de lettres pour TKF91, de fragments pour TKF92) se font entre deux sites déjà existants, ou aux extrémités de la séquence.
- Chacun de ces évènements (mutation, insertion, délétion) a lieu avec un taux propre.
- Lors d'une substitution, une nouvelle lettre est tirée avec une certaine probabilité sur l'alphabet.

Conséquences (1/2)

- Chaque alignement des deux séquences peut être codé par une suite à valeurs dans $\{H, D, I\}$ qui indique les positions *homologues* (H, i.e. matchs/mismatchs), effacées (D) dans la première séquence ou insérées (I) dans la première séquence.

Le modèle d'évolution TKF (1/2)

Modèle d'évolution

- Chaque site évolue indépendamment et peut subir une substitution ou être effacé.
- Les insertions (de lettres pour TKF91, de fragments pour TKF92) se font entre deux sites déjà existants, ou aux extrémités de la séquence.
- Chacun de ces évènements (mutation, insertion, délétion) a lieu avec un taux propre.
- Lors d'une substitution, une nouvelle lettre est tirée avec une certaine probabilité sur l'alphabet.

Conséquences (1/2)

- Chaque alignement des deux séquences peut être codé par une suite à valeurs dans $\{H, D, I\}$ qui indique les positions *homologues* (H, i.e. matchs/mismatchs), effacées (D) dans la première séquence ou insérées (I) dans la première séquence.

Le modèle d'évolution TKF (2/2)

Conséquences (2/2)

- La suite $W_{1:L}$ où $W_i \in \{H, D, I\}$ qui code pour l'évolution entre deux séquences sous le modèle d'évolution TKF est une chaîne de Markov. Ici, L est la longueur du « vrai alignement ».
- Conditionnellement à cette suite $W_{1:L}$, le modèle émet de façon indépendante les lettres de deux séquences $\longrightarrow$ PairHMM.

Sommaire

Le modèle pair-Markov caché (1/4)

Rappel : représentation graphique d'un alignement

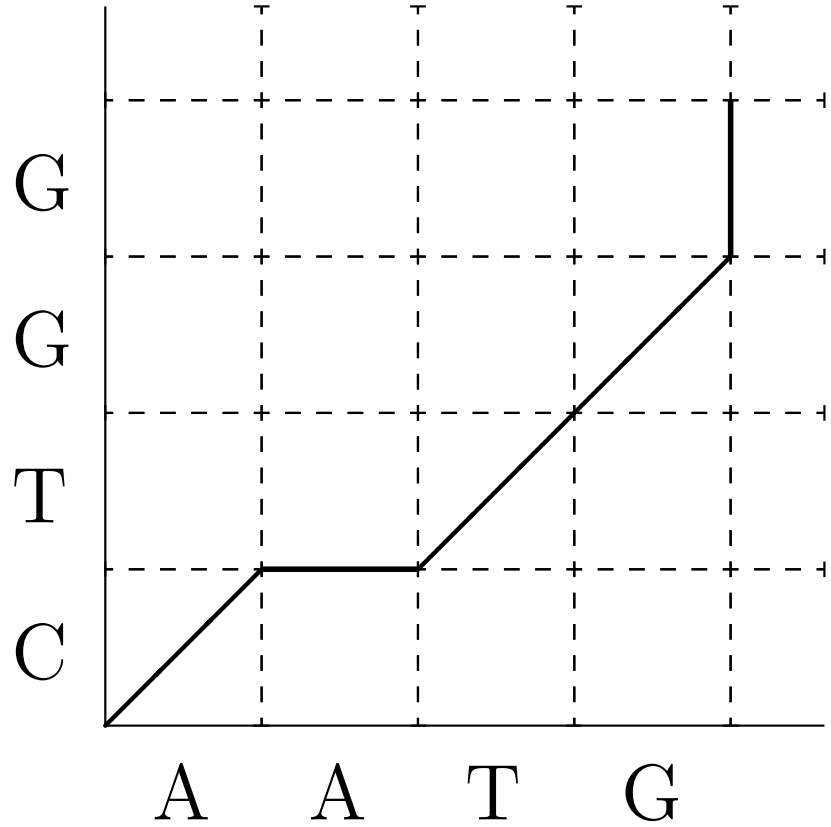


FIG.: Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{matrix} A & A & T & G & - \\ C & - & T & G & G \end{matrix}$.

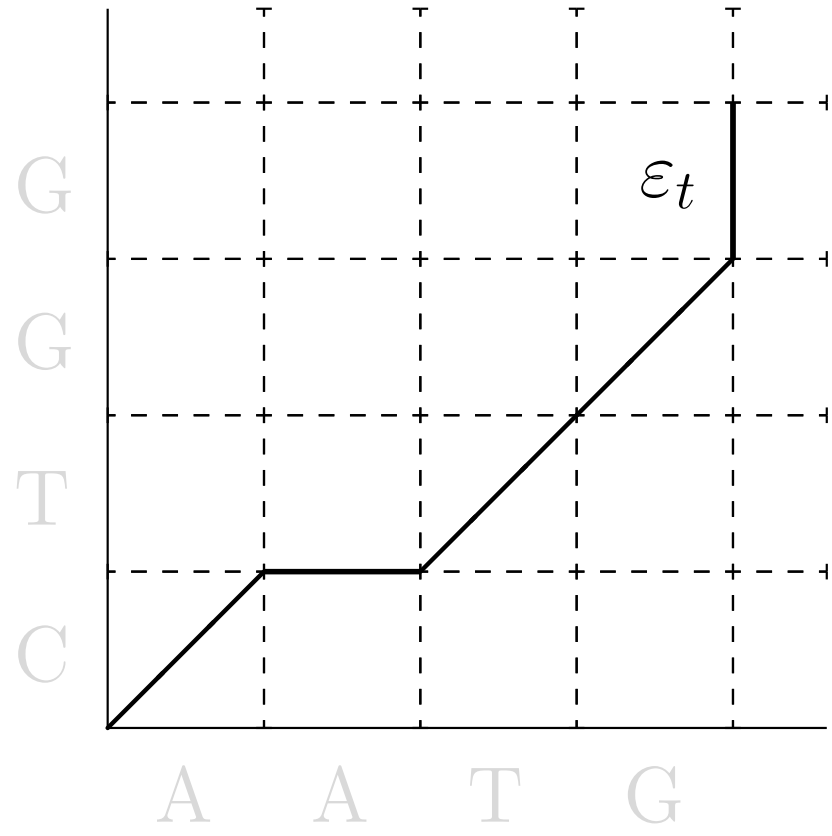
Le modèle pair-Markov caché (2/4)
Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.
- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.
- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

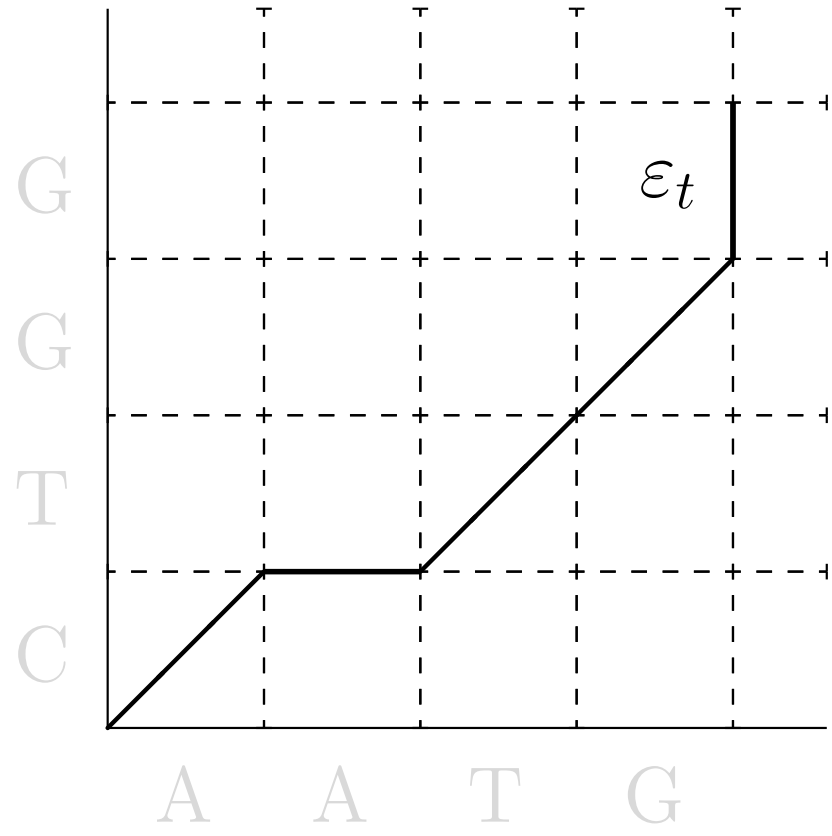


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

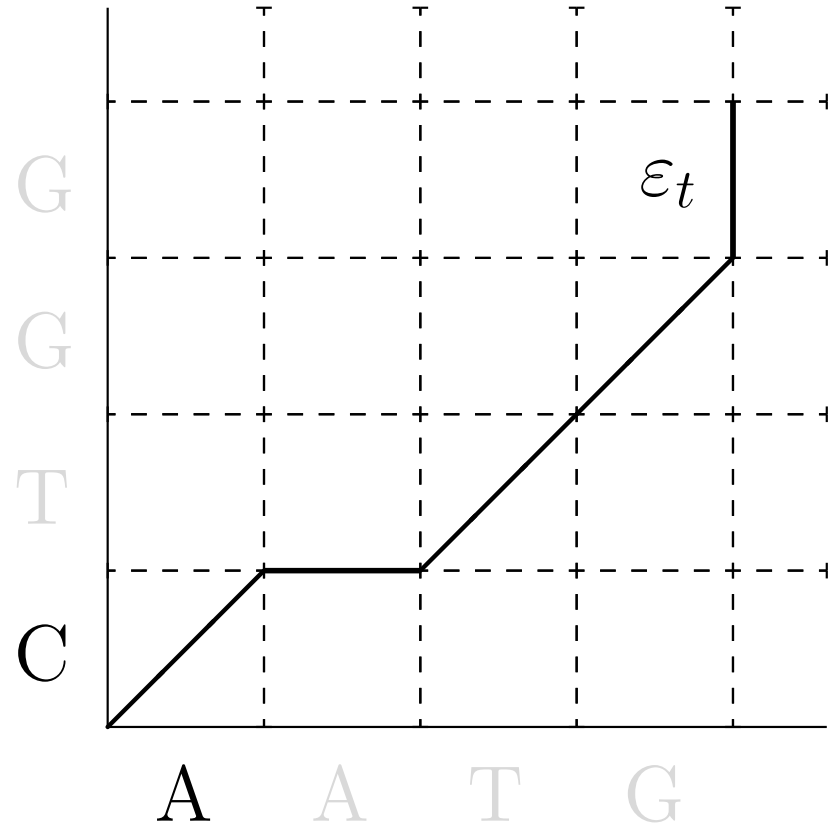


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

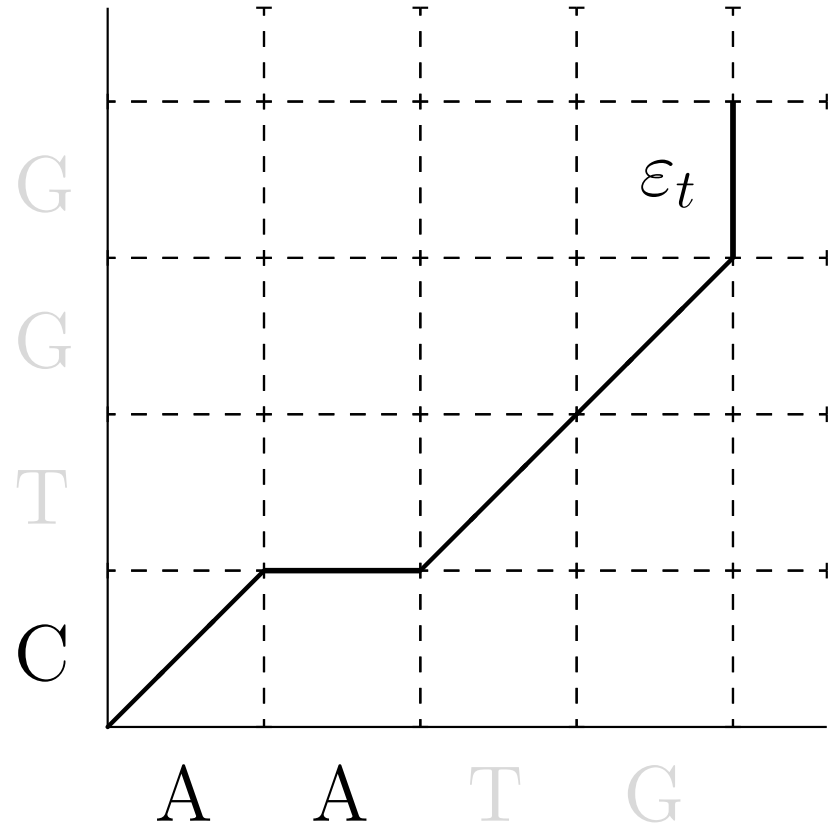


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

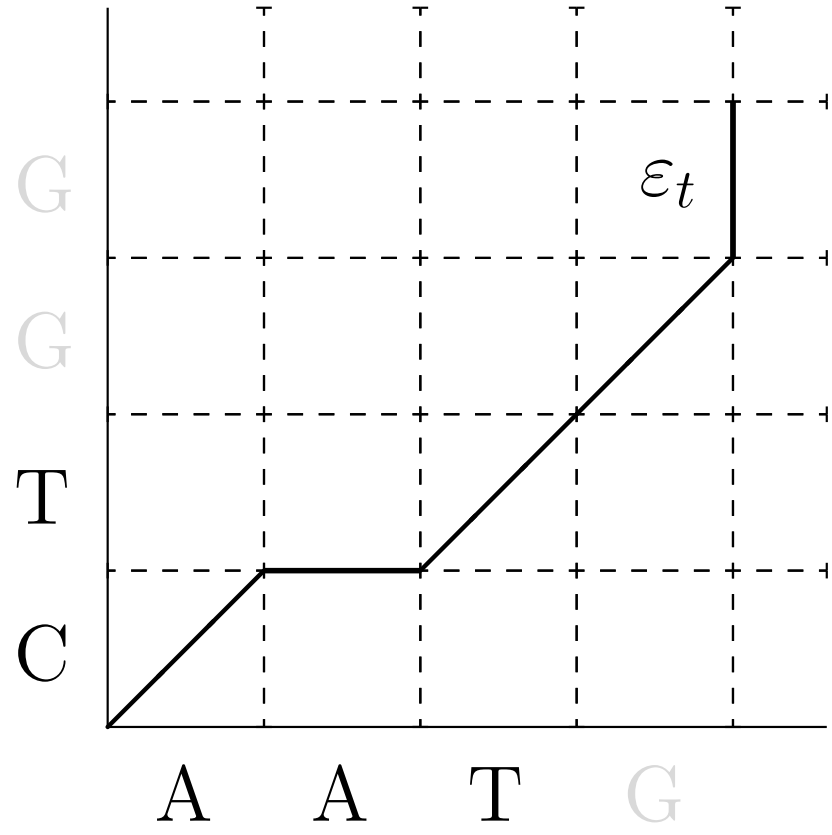


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

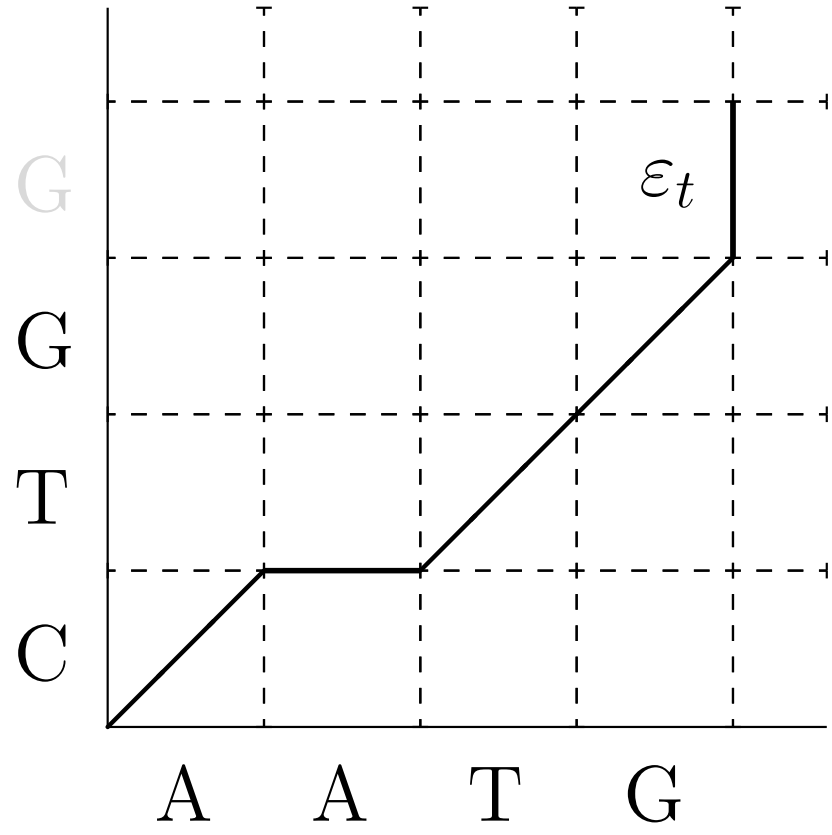


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.

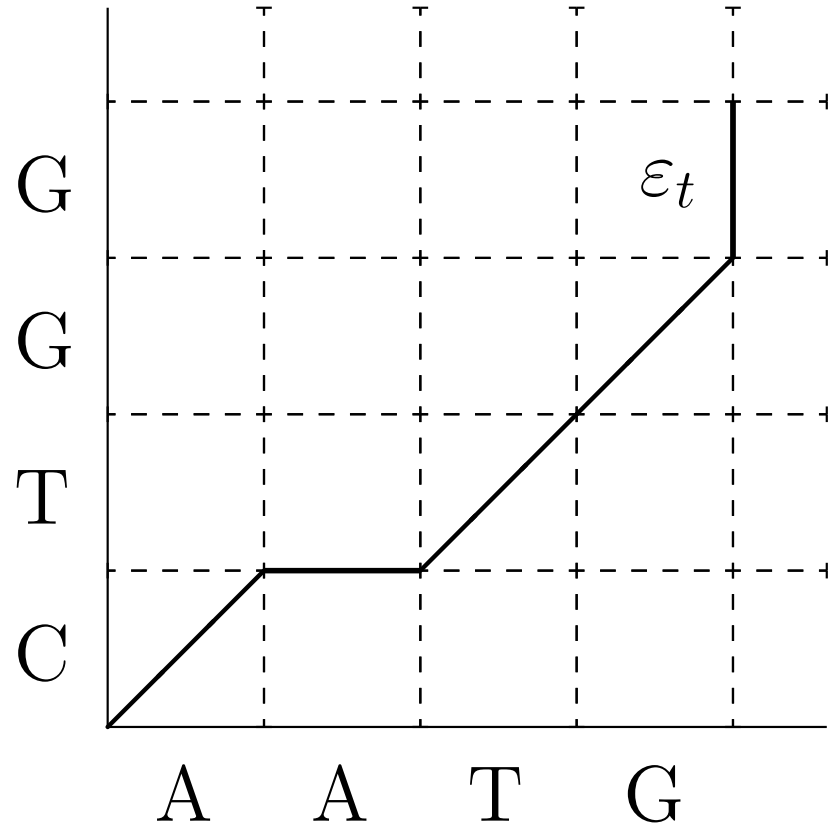


Le modèle pair-Markov caché (2/4)

Notations (1/2) [AGGM06]

- $\mathcal{A}$ alphabet fini (ex $\{A, C, G, T\}$).
- $\{\varepsilon_t\}_{t\geq 1}$ chaîne de Markov stationnaire ergodique sur $\mathcal{E} = \{(1,0);(0,1);(1,1)\}$. Matrice de transition π et loi stationnaire $\mu = (p, q, r)$.

- À l'instant t, conditionellement à $\{\varepsilon_s, s \leq t\}$ on tire indépendemment
  - Un couple de v.a. (X, Y) de loi h sur $\mathcal{A} \times \mathcal{A}$, si $\varepsilon_t = (1,1)$,
  - Une v.a. X de loi f sur $\mathcal{A}$ si $\varepsilon_t = (1,0)$,
  - Une v.a. Y de loi g sur $\mathcal{A}$ si $\varepsilon_t = (0,1)$.



Le modèle pair-Markov caché (3/4)
Notations (2/2) [AGGM06]

- $\theta = (\pi, f, g, h) \in \Theta$ sont les paramètres
- Soit $Z_0 = (0,0)$ et $Z_t = (N_t, M_t) = \sum_{s=1}^{t} \varepsilon_s,$, marche aléatoire sur $\mathbb{N} \times \mathbb{N}$.

On a

$$\mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t})$$

$$= \prod_{s=1}^{t} f(X_{N_s})^{1\{\varepsilon_s=(1,0)\}} g(Y_{M_s})^{1\{\varepsilon_s=(0,1)\}} h(X_{N_s}, Y_{M_s})^{1\{\varepsilon_s=(1,1)\}}$$

et $$\mathbb{P}(\varepsilon_{1:t} = e_{1:t}) = \mu_{e_1} \prod_{s=1}^{t-1} \pi(e_s, e_{s+1}).$$

Le modèle pair-Markov caché (3/4)
Notations (2/2) [AGGM06]

- $\theta = (\pi, f, g, h) \in \Theta$ sont les paramètres
- Soit $Z_0 = (0,0)$ et $Z_t = (N_t, M_t) = \sum_{s=1}^{t} \varepsilon_s,$, marche aléatoire sur $\mathbb{N} \times \mathbb{N}$.

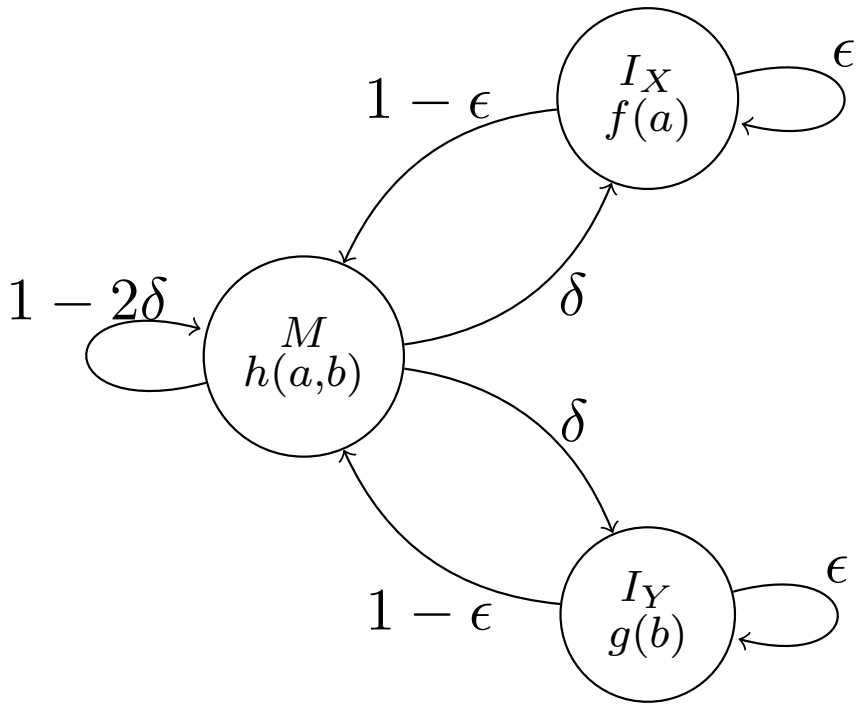
On a

$$\begin{aligned} & \mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}) \\ = & \prod_{s=1}^{t} f(X_{N_s})^{1\{\varepsilon_s=(1,0)\}} g(Y_{M_s})^{1\{\varepsilon_s=(0,1)\}} h(X_{N_s}, Y_{M_s})^{1\{\varepsilon_s=(1,1)\}} \end{aligned}$$

et $$\mathbb{P}(\varepsilon_{1:t} = e_{1:t}) = \mu_{e_1} \prod_{s=1}^{t-1} \pi(e_s, e_{s+1}).$$

Le modèle pair-Markov caché (4/4)

Représentation sous forme d'automate



$$\pi = \begin{pmatrix} \epsilon & 0 & 1-\epsilon \\ 0 & \epsilon & 1-\epsilon \\ \delta & \delta & 1-2\delta \end{pmatrix}$$

Sommaire

Vraisemblance

On observe $X_{1:n}$ et $Y_{1:m}$.

- L'algorithme forward-backward généralisé aux pair-HMM permet de calculer

$$\mathbb{P}_\theta(X_{1:n}, Y_{1:m}) = \sum_{e \in \mathcal{E}_{n,m}} \mathbb{P}(\varepsilon_{1:|e|} = e_{1:|e|}, X_{1:n}, Y_{1:m})$$

  où $\mathcal{E}_{n,m}$ est l'ensemble des chemins qui vont de $(0,0)$ à (n,m).
- L'algorithme `EM` appliqué aux pair-HMMs permet d'optimiser cette quantité par rapport à θ.
- On récupère une distribution a posteriori sur les alignements.
- (On peut également utiliser l'algorithme de Viterbi pour chercher l'alignement optimal).

Sommaire

Avantages du pairHMM sur les méthodes par score

- Les paramètres sont estimés. Ceci correspond à sélectionner la fonction de score optimale (au sens évolutif) pour l'alignement.
- Les pairHMM permettent d'obtenir une loi a posteriori sur les alignements.

Probabilités a posteriori d'alignements

(Source Metzler *et al.*, J. Mol. Evol. 2001)

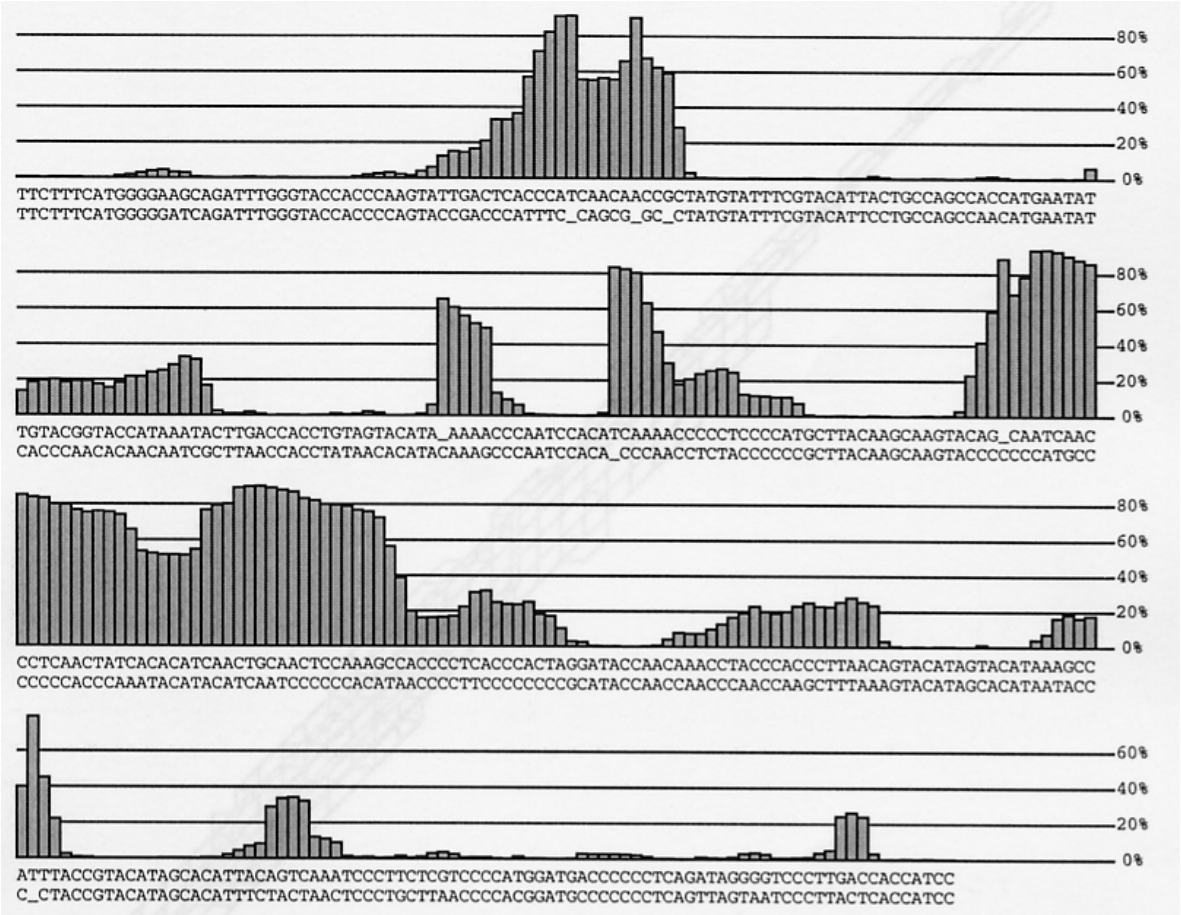


Figure 3: The most probable of the sampled alignments of a human and an orangutan HVR-1 sequence and the percentages of sampled alignments that differ from it in each position.

Quatrième partie IV

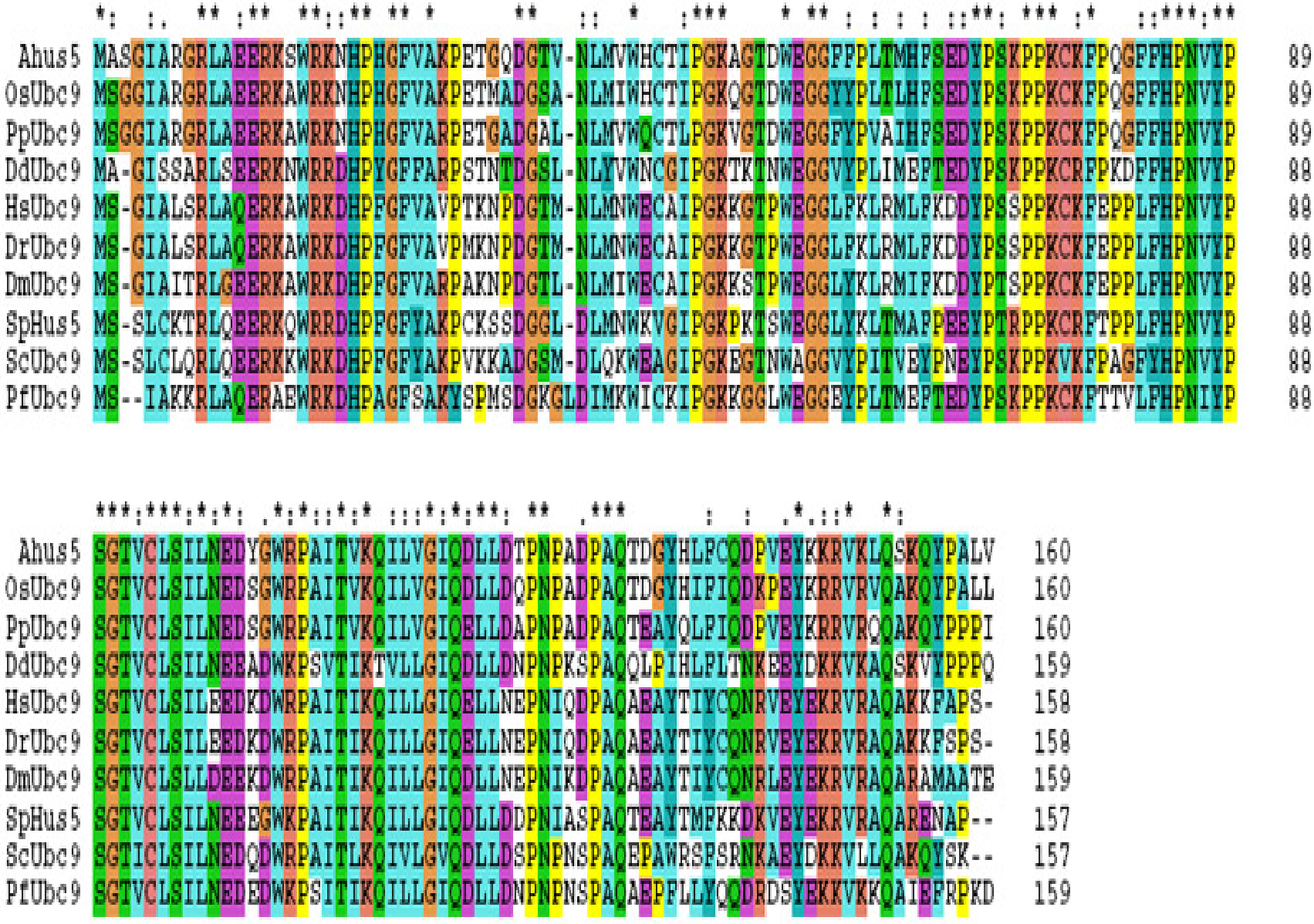
Alignement multiple

Sommaire Partie IV : Alignement multiple

Sommaire

Alignement multiple de séquences

Alignement de protéine Hus5/Ubc9 dans divers organismes



Introduction à l'alignement multiple (1/2)

Vocabulaire

- Pour les alignements de plus de 3 séquences, chaque site est soit un site *homologue* (i.e. présent dans la séquence ancestrale), soit *délété* (par rapport à la séquence ancestrale), soit *inséré* (par rapport à la séquence ancestrale).

Algorithmes d'alignement par score

- Au-delà de deux séquences, le problème devient rapidement très complexe car l'espace des alignements possibles explose. (Rappel, pour la programmation dynamique, la complexité croît comme le produit des longueurs des séquences.)
- Dans la pratique, il existe deux grands types de stratégies
  - progressives, basées sur de l'alignement par paires (Clustal W). Forte dépendance dans l'ordre des séquences.
  - par points d'ancrages multiples (DIALIGN2, MUSCLE).

Introduction à l'alignement multiple (1/2)

Vocabulaire

- Pour les alignements de plus de 3 séquences, chaque site est soit un site *homologue* (i.e. présent dans la séquence ancestrale), soit *délété* (par rapport à la séquence ancestrale), soit *inséré* (par rapport à la séquence ancestrale).

Algorithmes d'alignement par score

- Au-delà de deux séquences, le problème devient rapidement très complexe car l'espace des alignements possibles explose. (Rappel, pour la programmation dynamique, la complexité croît comme le produit des longueurs des séquences.)
- Dans la pratique, il existe deux grands types de stratégies
  - progressives, basées sur de l'alignement par paires (Clustal W). Forte dépendance dans l'ordre des séquences.
  - par points d'ancrages multiples (DIALIGN2, MUSCLE).

Introduction à l'alignement multiple (2/2)

Quelles séquences aligner ?

- En pratique, il faut faire attention à l'hétérogénéité dans les distances entre les séquences à aligner.
- Si un sous-ensemble de séquences est trop proche par rapport au reste des séquences, cela introduit un biais dans l'alignement.
- Certains logiciels pondèrent les (paires de) séquences en fonction de leur similitude (Note : la similitude est elle-même basée sur la matrice de score, avec un seuil qui n'est pas toujours explicite).

Sommaire

Alignement statistique multiple (1/2)

Principe (1/2)

- La généralisation du modèle pairHMM présenté ci-dessus à plusieurs séquences est non triviale.
- Il faut se donner une phylogénie des séquences (un arbre) pour avoir une probabilité d'apparition des séquences sous un modèle d'évolution. Dans la suite, on considère une phylogénie en étoile.
- Les états d'intérêt sont ici : les positions dans la séquence ancestrale, et pour chacune des séquences, les sites homologues (qui sont dérivés d'une position ancestrale), les sites délétés (par rapport à une position ancestrale), et les sites insérés (par rapport à une position ancestrale).

Alignement statistique multiple (2/2)

Principe (2/2)

- Pour k séquences, la chaîne cachée ε_t a pour longueur le nombre de positions dans la séquence ancestrale. À chaque temps t, l'état caché ε_t a pour valeur un vecteur de longueur k, dont chaque coordonnée i indique si la $i$ème séquence possède le site ancestral en position t (site homologue ou site délété) et le nombre d'insertions éventuelles après la position ancestrale (voir [AG07]).

Algorithme

- Les algorithmes d'alignement soufrent des mêmes problèmes d'efficacité que ceux qui utilisent l'alignement par score.

Alignement statistique multiple et phylogénie

- À noter : Dans Fleissner *et al.* [FMvH05] reconstruction de la phylogénie et alignement statistique multiple simultanés.

Sommaire

Chaînes de Markov cachées profils (1/3)

Références [Edd98, KBM+94]

Principe

- Le nombre de positions homologues L est fixé. Il existe une chaîne de Markov cachée (le profil) qui décrit la succession des états *homologue*, *inséré* et *délété*.
- Conditionnellement au profil, les séquences sont supposées indépendantes.
- Les paramètres de ce modèle profileHMM et l'alignement sous-jacent des séquences sont estimés à partir des séquences observées, par algorithme `EM`.

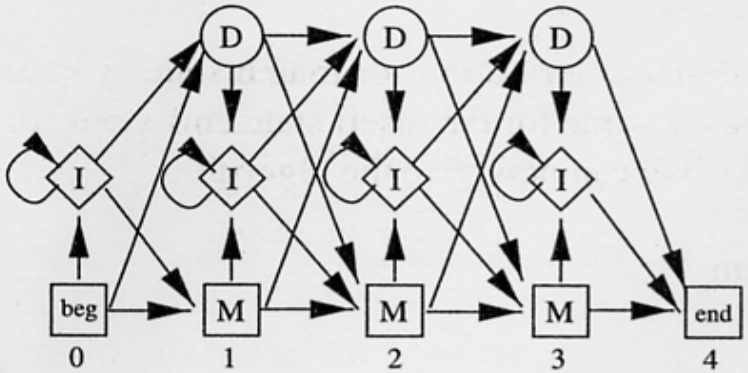
Chaînes de Markov cachées profils (2/3)

Chaîne profil (source Durbin *et al.* [DEKM98])

(a) Multiple alignment:

	x	x	.	.	.	x
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C
	1	2	.	.	.	3

(b) Profile-HMM architecture:



(c) Observed emission/transition counts

		model position 0	1	2	3
match emissions	A	-	4	0	0
	C	-	0	0	4
	G	-	0	3	0
	T	-	0	0	0
insert emissions	A	0	0	6	0
	C	0	0	0	0
	G	0	0	1	0
	T	0	0	0	0
state transitions	M-M	4	3	2	4
	M-D	1	1	0	0
	M-I	0	0	1	0
	I-M	0	0	2	0
	I-D	0	0	1	0
	I-I	0	0	4	0
	D-M	-	0	0	1
	D-D	-	1	0	0
	D-I	-	0	2	0

Figure 5.7 *As an example of model construction from an alignment, a small DNA multiple alignment is given (a), with three columns marked above with x's. These three columns are assigned to positions 1–3 in the model architecture (b). The assignment of columns to model positions determines the symbol emission and state transition counts (c) from which probability parameters would be estimated.*

Chaînes de Markov cachées profils (3/3)

En pratique

- L est souvent choisi comme la longueur moyenne des séquences à aligner.
- Présenté comme un alignement par « score spécifique à chaque position ». En effet, les paramètres d'émission des observations, conditionnellement à la chaîne cachée profil, sont différents suivants les positions dans l'alignement.

ProfileHMM vs Alignement statistique multiple

- La généralisation du modèle pairHMM à plus de deux séquences n'est pas le profileHMM.
- Différence = en profilHMM, conditionnellement à la chaîne profil, les séquences sont indépendantes.
- Dans un cadre d'align. stat. multiple, les lettres d'une colonne d'un site homologue sont émises selon une loi jointe, et les lettres correspondant à des sites insérés sont émises de façon indépendante (voir [AG07]).

Chaînes de Markov cachées profils (3/3)

En pratique

- L est souvent choisi comme la longueur moyenne des séquences à aligner.
- Présenté comme un alignement par « score spécifique à chaque position ». En effet, les paramètres d'émission des observations, conditionnellement à la chaîne cachée profil, sont différents suivants les positions dans l'alignement.

ProfileHMM vs Alignement statistique multiple

- La généralisation du modèle pairHMM à plus de deux séquences n'est pas le profileHMM.
- Différence = en profilHMM, conditionnellement à la chaîne profil, les séquences sont indépendantes.
- Dans un cadre d'align. stat. multiple, les lettres d'une colonne d'un site homologue sont émises selon une loi jointe, et les lettres correspondant à des sites insérés sont émises de façon indépendante (voir [AG07]).

Revues sur l'alignement

Bio-informatique

- Sur l'alignement statistique : [LDMH05].
- Sur la significativité d'un alignement par score : [PW04].
- Sur l'alignement de génomes complets [DP06].

Mathématique

- Sur l'alignement par score, le chapitre d'introduction de la thèse [Gro03].
- Sur l'alignement statistique, le chapitre d'introduction de la thèse [AG07].

[AG07] A. Arribas-Gil.
Estimation dans des modèles à variables cachées : alignement de séquences biologiques et modèles d'évolution.
PhD thesis, Université Paris-Sud, France, 2007.

[AGGM06] A.Arribas-Gil, E. Gassiat, and C. Matias.
Parameter estimation in pair-hidden Markov models.
Scand. J. Statist., 33(4) :651–671, 2006.

[AW94] R. Arratia and M. S. Waterman.
A phase transition for the score in matching random sequences allowing deletions.
Ann. Appl. Probab., 4(1) :200–225, 1994.

[Cha03] H. P. Chan.
Upper bounds and importance sampling of p-values of DNA and protein sequence alignments.
Bernoulli, 9(2) :183–199, 2003.

[Cha05] H. P. Chan.

Summation test for gap penalties and strong law of the local alignment score.
Ann. Appl. Probab., 15(2) :1492–1505, 2005.


[DEKM98] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison.
Biological sequence analysis. Probabilistic models of proteins and nucleic acids.
Cambridge : Cambridge University Press, 1998.


[DKZ94a] A. Dembo, S. Karlin, and O. Zeitouni.
Critical phenomena for sequence matching with scoring.
Ann. Probab., 22(4) :1993–2021, 1994.


[DKZ94b] A. Dembo, S. Karlin, and O. Zeitouni.
Limit distribution of maximal non-aligned two-sequence segmental score.
Ann. Probab., 22(4) :2022–2039, 1994.


[DP06] C. N. Dewey and L. Pachter.
Evolution at the nucleotide level : the problem of multiple whole-genome alignment.

Hum. Mol. Genet., 15(suppl 1) :R51–56, 2006.


[DSO78] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt.
A model of evolutionary change in proteins.
In *Atlas of Protein sequence and structure*, volume 5, Supplement 3, pages 345–352, Washington DC, 1978. National Biomedical Research Foundation.


[Edd98] Sean R. Eddy.
Profile hidden Markov models.
Bioinformatics Review, 14(9) :755–763, 1998.


[FMvH05] R. Fleissner, D. Metzler, and A. von Haeseler.
Simultaneous statistical multiple alignment and phylogeny reconstruction.
Systematic Biology, 54(4) :548–561, 2005.


[Got82] O. Gotoh.
An improved algorithm for matching biological sequences.
J. Mol. Biol., 162(3) :705–8, 1982.


[Gro03] S. Grossmann.

Statistics of optimal sequence alignments.
PhD thesis, Johann Wolfgang Goethe-Universität,
Frankfurt am Main, 2003. Available at
`http ://www.molgen.mpg.de/~grossman/dissertation.pdf`.


[GY04] S. Grossmann and B. Yakir.
Large Deviations for global maxima of independent superadditive processes with negative drift and an application to optimal sequence alignments.
Bernoulli, 10(5) :829–845, 2004.


[Han06] Niels Richard Hansen.
Local alignment of Markov chains.
Ann. Appl. Probab., 16(3) :1262–1296, 2006.


[HH92] S. Henikoff and J.G. Henikoff.
Amino acid substitution matrices from protein blocks.
Proc Natl Acad Sci U S A., 89(22) :10915–9, 1992.


[KBM+94] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler.

Hidden Markov models in computational biology : Applications to protein modelling.
J. Mol. Biol., 235 :1501–1531, 1994.


[LDMH05] Gerton Lunter, Alexei J. Drummond, István Miklós, and Jotun Hein.
Statistical alignment : recent progress, new applications, and challenges.
In *Statistical methods in molecular evolution*, Stat. Biol. Health, pages 375–405. Springer, New York, 2005.


[MLH04] I. Miklos, G. A. Lunter, and I. Holmes.
A "Long Indel" Model For Evolutionary Sequence Alignment.
Molecular Biology and Evolution, 21(3) :529–540, 2004.


[MT99] R. Mott and R. Tribe.
Approximate statistics of gapped alignments.
Journal of Comput. Biol., 6(1) :91–112, 1999.


[NW70] S.B. Needleman and C.D. Wunsch.

A general method applicable to the search for similarities in the amino acid sequence of two proteins.
J. Mol. Biol., 48(3) :443–53, 1970.


[PW04] W.R. Pearson and T.C. Wood.
Handbook of Statistical Genetics, chapter "Statistical Significance in Biological Sequence Comparison". Eds. Balding, D.J. and Bishop, M. and Cannings, C.
John Wiley & Sons, second edition, 2004.


[SW81] T.F. Smith and M.S. Waterman.
Identification of common molecular subsequences.
J. Mol. Biol., 147(1) :195–7, 1981.


[SY00a] David Siegmund and Benjamin Yakir.
Approximate p-values for local sequence alignments.
Ann. Stat., 28(3) :657–680, 2000.


[SY00b] David Siegmund and Benjamin Yakir.
Tail probabilities for the null distribution of scanning statistics.
Bernoulli, 6(2) :191–213, 2000.

[TKF91] J.L. Thorne, H. Kishino, and J. Felsenstein.
An evolutionary model for maximum likelihood alignment of DNA sequences.
J. Mol. Evol., 33 :114–124, 1991.

[TKF92] J.L. Thorne, H. Kishino, and J. Felsenstein.
Inching toward reality : an improved likelihood model of sequence evolution.
Journal of Molecular Evolution, 34 :3–16, 1992.

[YBH02] Yi-Kuo Yu, R. Bundschuh, and Terence Hwa.
Statistical significance and extremal ensemble of gapped local hybrid alignment.
In *Biological Evolution and Statistical Physics*, volume 585 of *Lecture Notes in Physics*, pages 3–21, Berlin/Heidelberg, 2002. Springer.

[YH01] Yi-Kuo Yu and Terence Hwa.
Statistical significance of probabilistic sequence alignment and related local hidden Markov models.
J. Comput. Biol., 8(3) :249–282, 2001.

[Zha95] Yu Zhang.
A limit theorem for matching random sequences allowing deletions.
Ann. Appl. Probab., 5(4) :1236–1240, 1995.