Introduction à la génomique comparative

Catherine Matias

CNRS - Laboratoire Statistique & Génome, Évry

Atelier de la SFDS - 23, 24 octobre 2008.

Sommaire

Première partie I

Introduction

Sommaire Introduction

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,
- ...

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).

Ainsi, 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

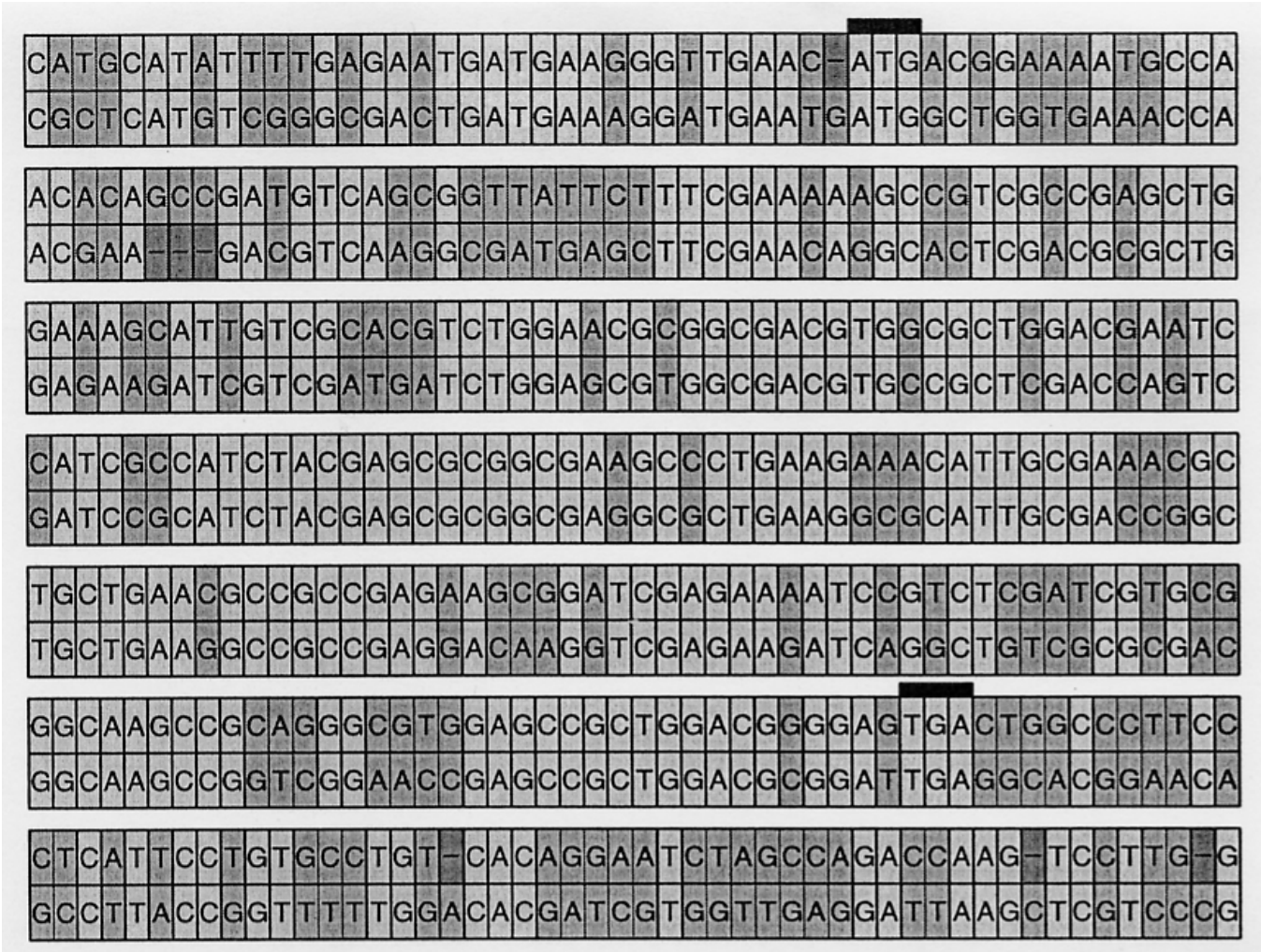


FIG. 3. Part of the pairwise alignment of *A.tumefaciens* and *M.loti.* Light gray color corresponds to conserved positions, and nonconserved positions and gaps are shown in dark gray. The two black bars on top of the alignment

Que représente un alignement ?(1/2)

- On présume que 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.

Que représente un alignement ?(2/2)

L'alignement est donc censé traduire la phylogénie sous-jacente aux séquences. [La phylogénie d'un groupe d'espèces, c'est l'arbre qui représente l'évolution de ces espèces à partir d'un ancêtre commun.] Il faut retenir que phylogénie et alignement sont très dépendants.

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.

Représentation graphique (1/3)

- alignement entre deux séquences de longueur n et m = un chemin (contraint à des pas élémentaires du type $(1,1), (1,0)$ et $(0,1)$ uniquement) sur la grille $[0,n] \times [0,m]$.
- les pas $(1,1)$ = *matchs* ou *mismatchs*
- les pas $(1,0)$ et $(0,1)$ = *indels*

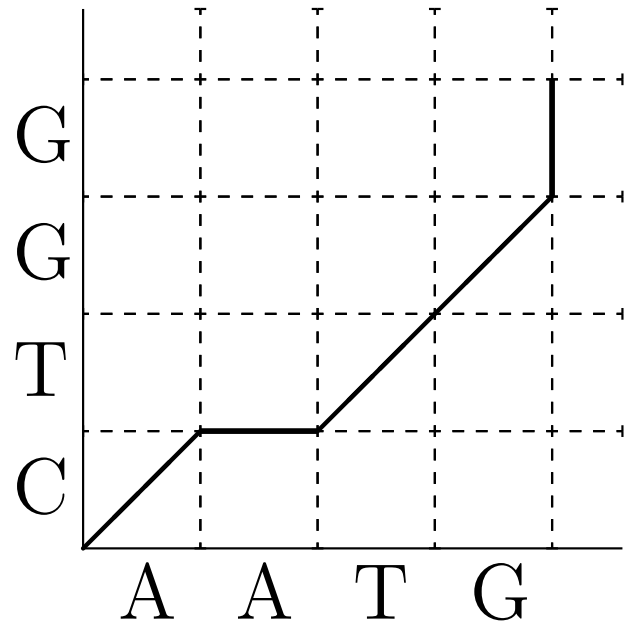


FIG.: Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{matrix} A & A & T & G & - \\ C & - & T & G & G \end{matrix}$.

Représentation graphique (2/3)

- alignement global = chemin qui commence en $(0, 0)$ et termine en (n, m),
- alignement local = chemin contraint à rester dans la grille $[0, n] \times [0, m]$.
- NB : le meilleur alignement global ne contient pas nécessairement le meilleur alignement local.

Représentation graphique (3/3)

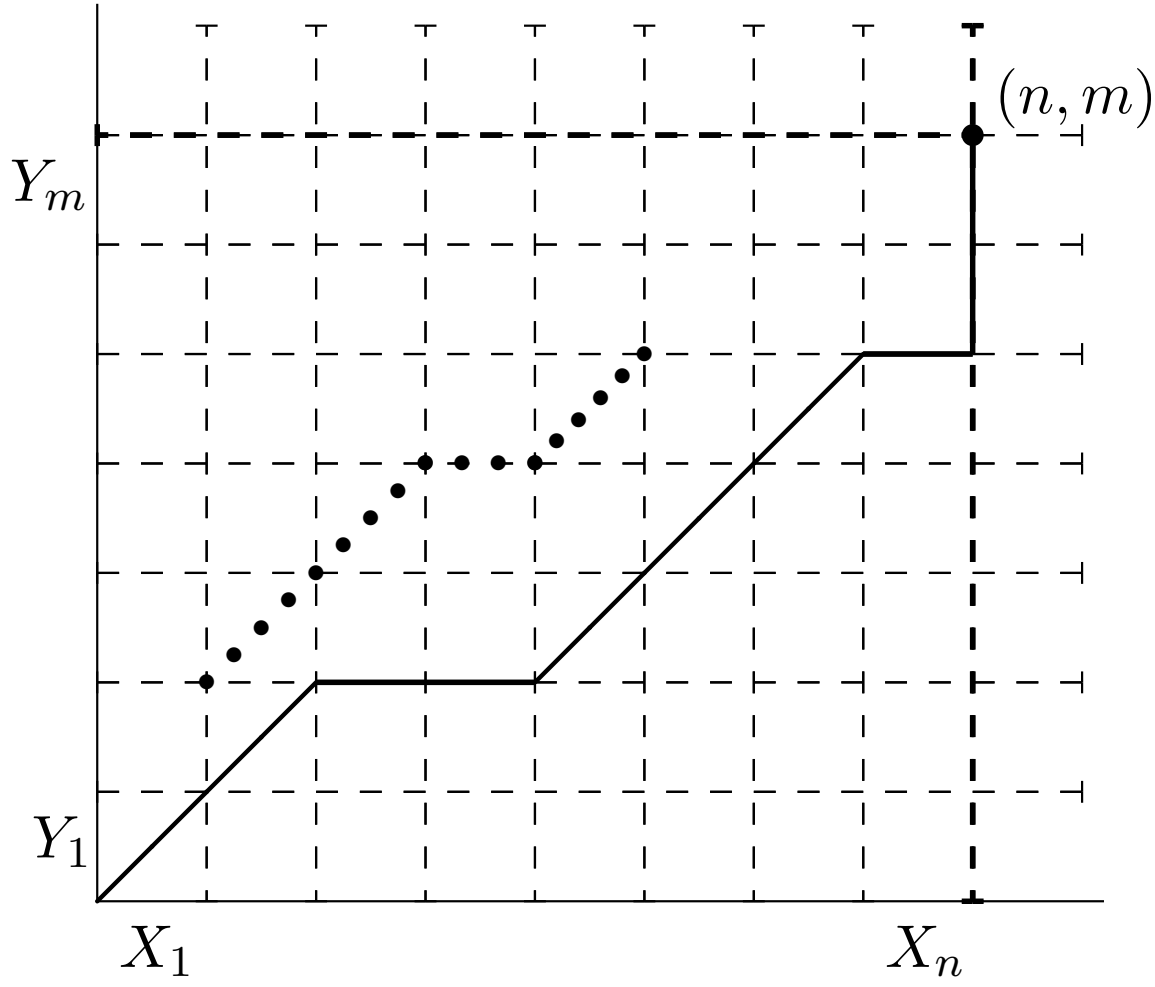


FIG.: Représentation graphique du meilleur alignement global (traits pleins) et local (traits pointillés) des séquences $X_{1:n}$ et $Y_{1:m}$.

Deuxième partie II

Alignement de deux séquences par score

Sommaire Partie II : Alignement par score

Alignement par 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 ».

Il faut noter que l'utilisation d'un score additif correspond à l'hypothèse que l'évolution sous-jacente des séquences se fait de façon indépendante sur les sites. C'est une hypothèse simplificatrice, délicate pour l'alignement de protéines, et très fausse pour les ARNs de structure.

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 (1/2)

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

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

- Soit s une fonction de score $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. Soient $I = \{i+1, \ldots, i+s\}$ et $J = \{j+1, \ldots, j+t\}$ deux ensembles d'indices consécutifs. On définit

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

pour une pénalité linéaire.

Formalisation mathématique (4/4)

Score d'alignement de 2 séquences. Cas avec indels (2/2)

- Le maximum est pris sur tous les alignements $\{(i_k, j_k), 1 \leq k \leq \ell, \ell \geq 0\}$ tels que

$$i + 1 \leq i_1 < \cdots < i_\ell \leq i + s$$
$$j + 1 \leq j_1 < \cdots < j_\ell \leq j + t.$$

- Le score local optimal entre les deux séquences est

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

  Quand $n = m$, on note simplement M_n.
- Le score global d'alignement est simplement la quantité $H(\{1, \ldots, n\}, \{1, \ldots, m\})$.

Remarques (1/2)

- En pratique, pour l'alignement qui réalise le score local optimal, il existe $i^\star, j^\star, \ell^\star, s^\star, t^\star$ tels que les positions de match/mismatch $\{(i_k, j_k), 1 \leq k \leq \ell^\star\}$ vérifient

$$i^\star + 1 = i_1 < \cdots < i_{\ell^\star} = i^\star + s^\star$$
$$j^\star + 1 = j_1 < \cdots < j_{\ell^\star} = j^\star + t^\star.$$

  En effet, on choisit un score négatif pour les indels et donc les premières et dernières positions de l'alignement local ne sont pas des indels, mais des matchs/mismatchs.

- La distribution asymptotique du score local optimal (ou des k plus grands scores locaux) sous l'hypothèse H_0 : « les deux séquences sont indépendantes » permet de dire si un alignement est significatif ou pas.

Remarques (2/2)

- Équilibre à faire entre les scores des lettres et la pénalité des insertions-délétions. Influence sur le type d'alignements obtenus.
- Le score augmente naturellement avec la longueur de la séquence. On distingue deux « phases » : la croissance linéaire ou la croissance logarithmique avec la longueur de la séquence.
- Le régime d'intérêt est le régime logarithmique.
- Il faut être capable de calculer le score de tous les alignements possibles entre deux séquences. La complexité du problème est très élevée, mais la solution est permise par l'existence d'algorithmes basés sur la programmation dynamique.

Algorithmes exacts (1/4)

- 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).

Principe

L'idée est assez simple : à chaque étape de la construction de l'alignement, on a trois possibilités : soit la prochaine lettre de la séquence X est alignée en face d'un blanc, soit la prochaine lettre de la séquence Y est alignée en face d'un blanc, soit on aligne la prochaine lettre de la séquence X avec la prochaine lettre de la séquence Y. Parmi ces trois possibilités, on garde celle qui maximise le score total (i.e le score de l'étape précédente + le coût de cette étape) et on continue.

Algorithmes exacts (2/4)

Programmation dynamique - al. global - pénalité linéaire

- Soit $F(i,j)$, le meilleur score d'alignement (global) entre $X_{1:i}$ et $Y_{1:j}$,
- On construit la matrice F de façon récursive :
- $F(0,0) = 0$, $F(i,0) = -\delta i$ et $F(0,j) = -\delta j$,
-

$$\text{Puis} \quad F(i,j) = \max \begin{cases} F(i-1,j-1) + s(X_i, Y_j) \\ F(i-1,j) - \delta \\ F(i,j-1) - \delta \end{cases}$$

Complexité : $O(nm)$ en temps et en mémoire.

Algorithmes exacts (3/4)

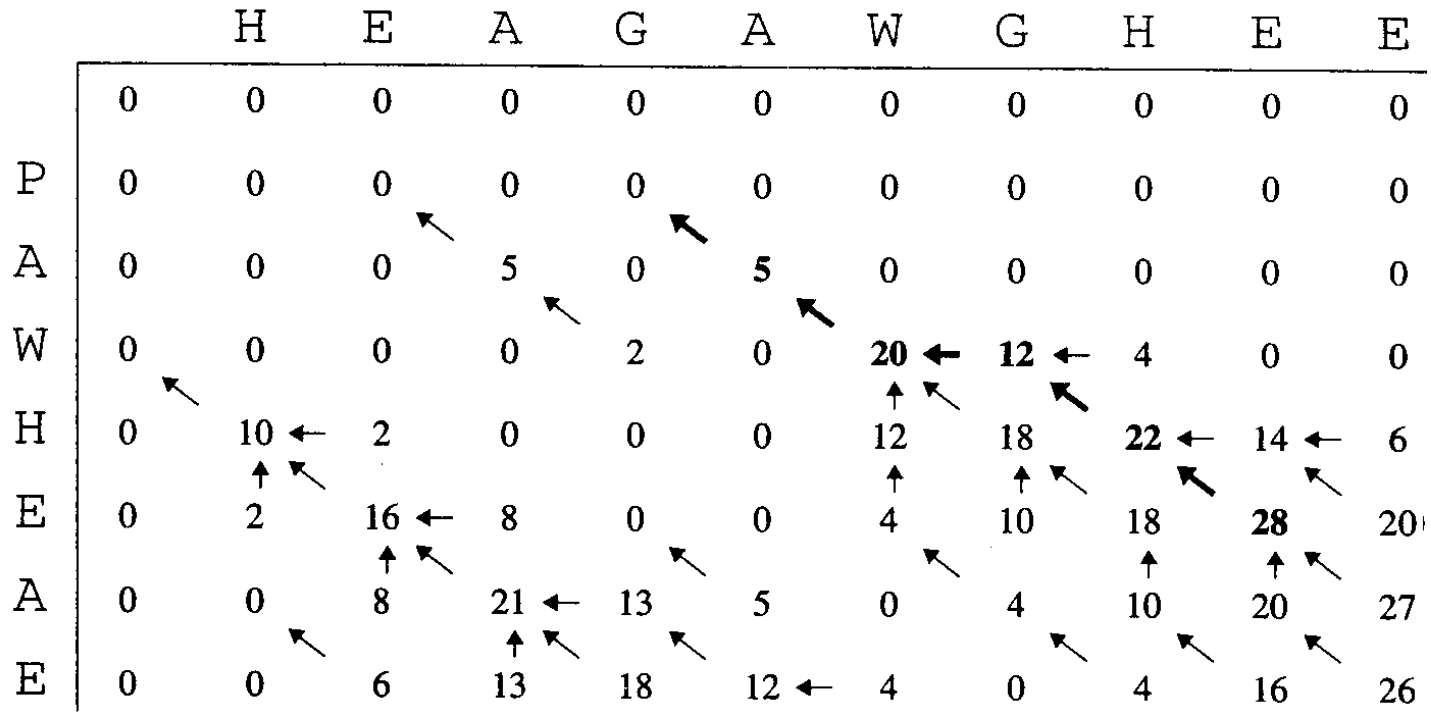
Programmation dynamique - al. local - pénalité linéaire

- Soit $F(i,j)$, le meilleur score d'alignement (local) entre $X_{1:i}$ et $Y_{1:j}$,
- On construit la matrice F de façon récursive :
- $F(0,0) = F(i,0) = F(0,j) = 0$,
-

$$\text{Puis} \quad F(i,j) = \max \begin{cases} 0 \\ F(i-1,j-1) + s(X_i, Y_j) \\ F(i-1,j) - \delta \\ F(i,j-1) - \delta \end{cases}$$

Complexité : $O(nm)$ en temps et en mémoire.
Pour plus de détails, voir [DEKM98].

Algorithmes exacts (4/4)
(Source Durbin *et al.* [DEKM98])



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE
AW-HE

Figure 2.6 *Above, the local dynamic programming matrix for the example sequences. Below, the optimal local alignment, with score* 28.

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...

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)

Log-odds ratios

Il s'agit de prendre $s(a,b) = \log \frac{p_{ab}}{q_a q_b}$ où q_a est la probabilité d'apparition de la lettre a dans les séquences, et $p_{a,b}$ probabilité d'apparition du match/mismatch (a,b) dans l'alignement.

En pratique

- Sur des séquences bien connues des biologistes, et alignées « à la main », on peut estimer q_a, p_{ab} par leurs fréquences d'apparition,
- Pour des séquences « proches », on utilise alors les valeurs ci-dessus de la matrice de score pour aligner les nouvelles séquences.

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).

Introduction aux propriétés du score local

- Pas de résultats sur le comportement de la queue de distribution du score global avec indels.
- Dans la suite, $X_{1:n}$ et $Y_{1:n}$ i.i.d. et indépendantes. On suppose $m = n$ par commodité uniquement.
- Existence d'un changement de régime : suivant la valeur des paramètres, le score local optimal croît linéairement ou logarithmiquement en n.
- Le régime d'intérêt est le régime logarithmique.
- Comportement asymptotique de $M_n / \log n$?

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 $Y_{1:n}$.

Propriétés du score local sans indels (1/5)

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 (2/5)

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 (3/5)

Loi de Gumbel

- Si Z_i suite de variables i.i.d. de loi exponentielle $\mathcal{E}(1)$, alors

$$\lim_{n \to +\infty} \mathbb{P}(\max_i Z_i - \log n \leq t) = \exp(\exp(-t)).$$

- Le même comportement apparaît pour le maximum de variables indépendantes dont la queue de distribution a une décroissance exponentielle.

Intuition de la preuve (voir [Gro03]) (1/3)

- On peut aisément prouver que le maximum de n^2 variables i.i.d. dont la queue de distribution décroît comme $K^\star \exp(-\theta^\star t)$ se comporte exactement comme M_n ci-dessus.
- Pourquoi M_n peut être interprété comme un tel maximum ?

Propriétés du score local sans indels (4/5)

Intuition de la preuve (voir [Gro03]) (2/3)

- Soit $\mathbf{i}$ un point de la grille $[0, n]^2$ et $T_{\mathbf{i}}$ le score maximal obtenu sur tous les chemins (alignements) qui commencent au point $\mathbf{i}$.
- M_n et le maximum des $T_{\mathbf{i}}$ ont à peu près le même comportement

$$M_n \simeq \max_{\mathbf{i} \in [0,n]^2} T_{\mathbf{i}}.$$

Mais, il n'y a pas égalité à cause des effets de bord.

Propriétés du score local sans indels (5/5)

Intuition de la preuve (voir [Gro03]) (3/3)

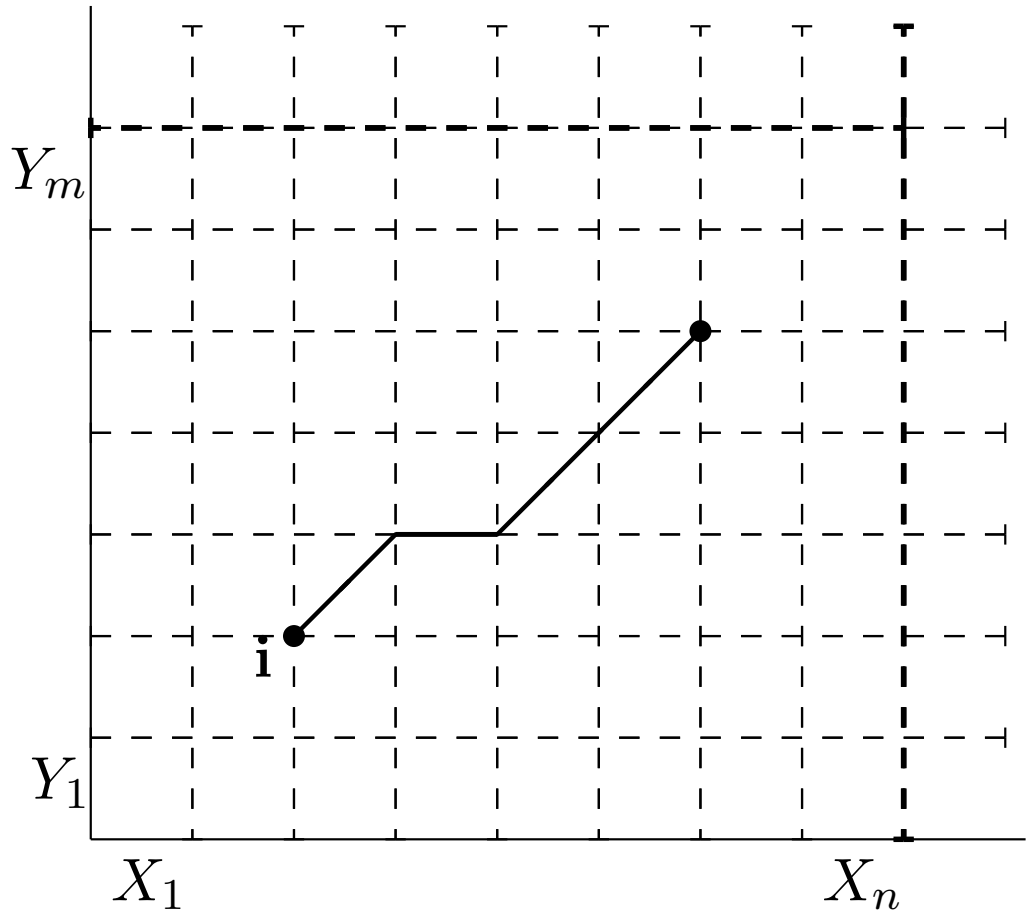


FIG.: Représentation graphique de $T_{\mathbf{i}}$.

Propriétes du score avec indels

Questions

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

- Les praticiens sont convaincus que la forme de la queue de distribution du score reste (approximativement) la 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 dans les cas particuliers suivants
  - Soit le nombre maximum de trous est fixé,
  - Soit le coût d'ouverture Δ d'un gap croît comme $\log n$.

Quelques références supplémentaires

- Grossmann et Yakir [GY04, Gro03] : inégalités de grandes 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 logarithmique.
- Mott et Tribe [MT99] fournissent une méthode heuristique pour approcher la valeur de $\theta^\star$ et prédire le point de

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

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...

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.

Le modèle pair-Markov caché (1/3)

Rappel : représentation graphique d'un alignement

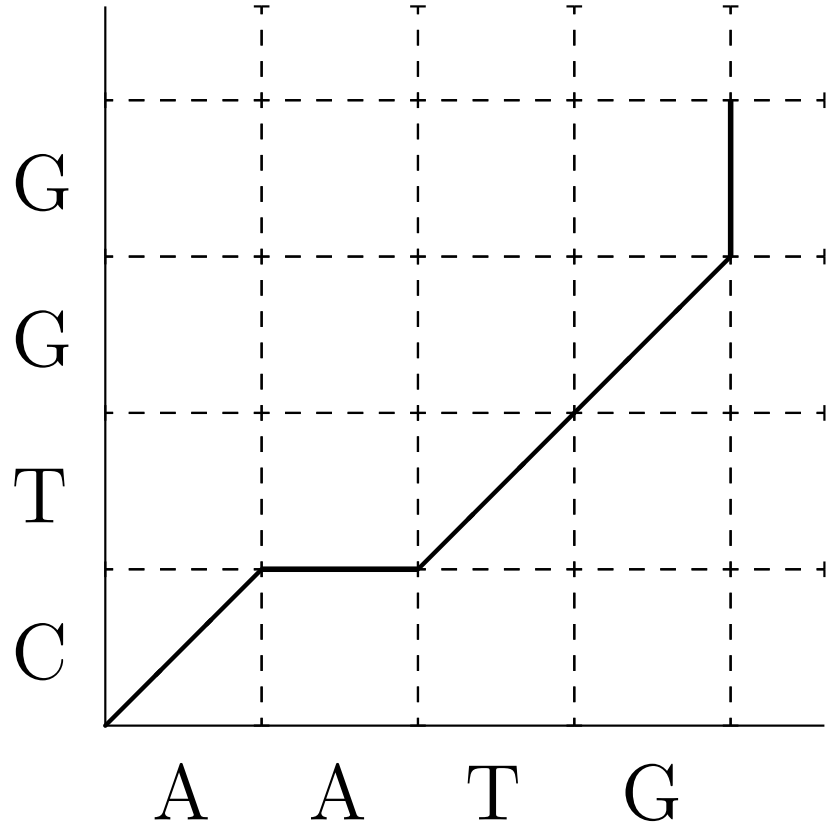
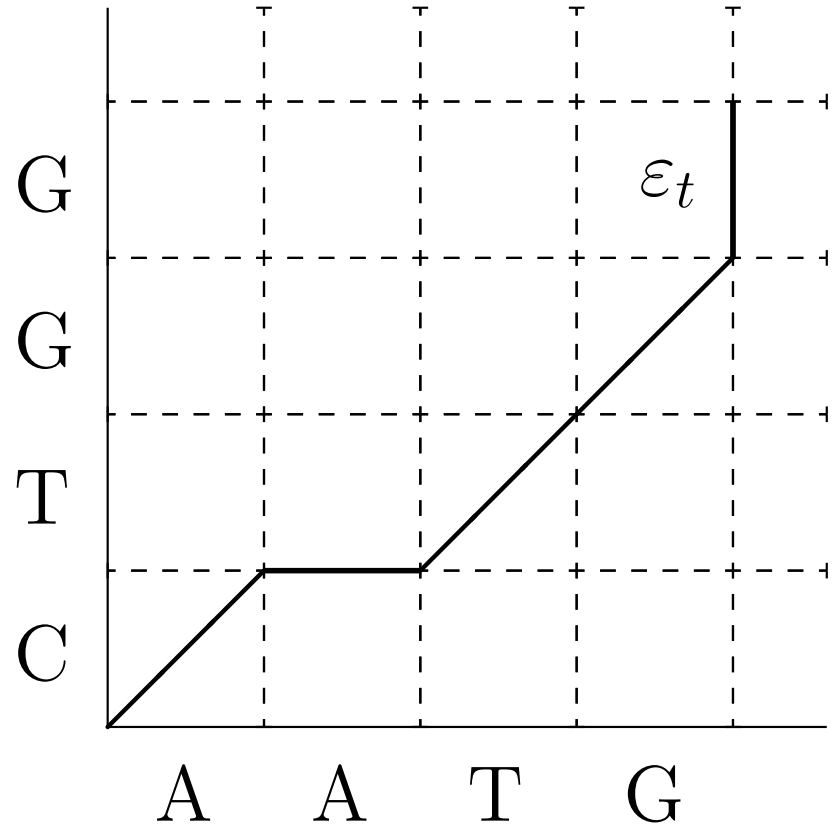


FIG.: Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{matrix} A & A & T & G & - \\ C & - & T & G & G \end{matrix}$.

Le modèle pair-Markov caché (2/3)
Notations (1/2)

- $\mathcal{A}$ 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 (3/3)
Notations (2/2)

- $\theta = (\pi, f, g, h) \in \Theta$ 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

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

- $\mathcal{E}_{n,m} = \{$ chemins de $(0,0)$ à $(n,m)\}$ et $\mathcal{E}_\infty = \{$ chemins dans $\mathbb{N} \times \mathbb{N}\}$.
- Si $e \in \mathcal{E}_{n,m}$, alors $|e|$ est la longueur du chemin e et on a $n \vee m \leq |e| \leq n + m$.

Modèle d'évolution / PairHMM

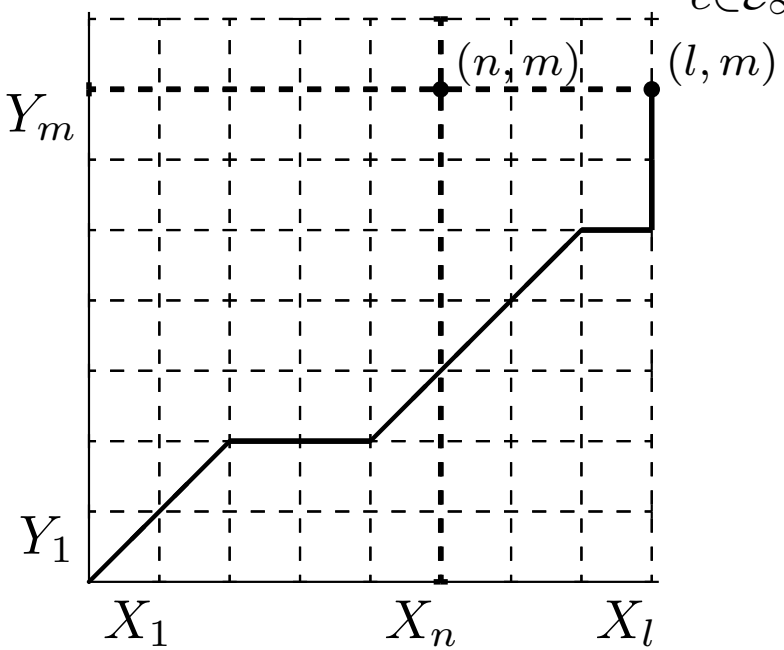
- Il faut garder en tête que le modèle pairHMM présenté ici est a priori plus général que les modèles d'évolution (comme TKF) présentés ci-dessus.
- En particulier, dans le cadre du modèle d'évolution TKF, les paramètres du modèle pairHMM induit sont contraints.

Observations et vraisemblances (1/3)

On observe $X_{1:n}$ et $Y_{1:m}$. Au moins deux interprétations

a) il s'agit de $X_{1:N_s}$ et $Y_{1:M_t}$ pour des valeurs inconnues de (s,t) (éventuellement $s \neq t$). Correspond au cas où le vrai alignement ne passe pas par le point (n,m). La vraisemblance est alors $\mathbb{P}(X_{1:n}, Y_{1:m})$ (i.e. distribution marginale)

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

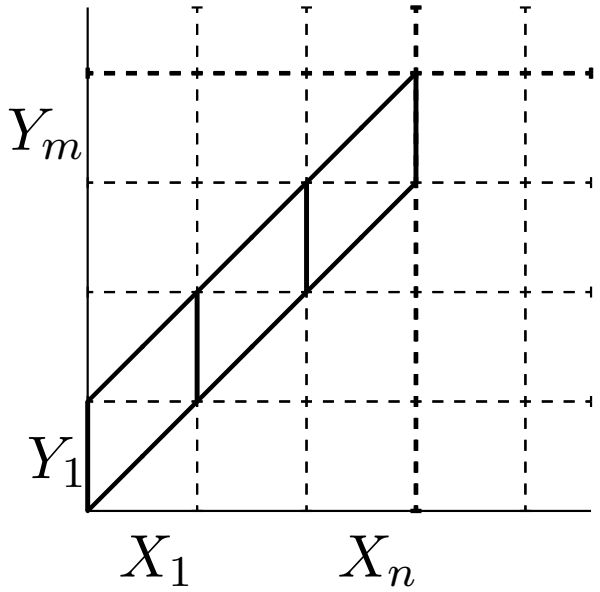


Trop compliqué pour que l'on puisse effectivement calculer la vraisemblance.

Observations et vraisemblances (2/3)

b) Il s'agit des séquences émises jusqu'à un temps t (inconnu). Alors, la vraisemblance est
$\mathbb{P}(X_{1:N_t}, Y_{1:M_t}) = \mathbb{P}(Z_t, X_{1:N_t}, Y_{1:M_t})$,

$$\mathbb{P}(X_{1:N_t}, Y_{1:M_t}) = \sum_{e \in \mathcal{E}_{N_t,M_t}, |e|=t} \mathbb{P}(\varepsilon_{1:t} = e_{1:t}, X_{1:N_t}, Y_{1:M_t})$$



$n = 3$, $m = 4$, $t = 4$

$\longrightarrow$ cependant la longueur t n'est pas observée !

- En fait, aucune de ces quantités n'est celle qui est calculée par l'algorithme `EM` dans le cadre des pair-HMMs.

Observations et vraisemblances (3/3)

Soit $\ell_t(\theta) = \log \mathbb{P}(X_{1:N_t}, Y_{1:M_t})$. Puisque t n'est pas observé, on considère $w_t(\theta) = \log Q_\theta(X_{1:N_t}, Y_{1:M_t})$ où

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

i.e., le point observé (n, m) appartient à la trajectoire, pour une longueur de trajectoire inconnue. Alors

$$Q_\theta(X_{1:n}, Y_{1:m}) = \mathbb{P}(\exists s \geq 1, Z_s = (n, m), X_{1:n}, Y_{1:m}).$$

- w_t est la quantité calculée par l'algorithme `EM` appliqué aux pair-HMMs.
- On a $n \vee m \leq t \leq n + m$ donc $t \to \infty \Rightarrow n, m \to \infty$

Hypothèses

- Condition nécessaire d'identifiabilité : $\exists x, y \in \mathcal{A}$ tels que $h(x,y) \neq f(x)g(y)$.
  En effet, sinon, $\mathbb{P}(X_{1:N_t}, Y_{1:M_t} | \varepsilon_{1:t}) = \prod_{s=1}^{t} f(X_{N_s}) g(Y_{M_S})$.
- Hypothèses biologiques possibles :
  H1 : $p = q$
  H2 : $h_X = f, h_Y = g$.

Résultats : existence d'un contraste limite

Arribas-Gil, Gassiat, M. [AGGM06]

Soit $\Theta_0 = \{\theta \in \Theta; \pi_{i,j} > 0, f(x) > 0, g(y) > 0, h(x,y) > 0\}$

Théorème

Pour tout $\theta \in \Theta_0$ *:*

i) $t^{-1}\ell_t(\theta)$ *converge* $\mathbb{P}_0$*-ps et dans* $\mathbb{L}_1$*, quand* $t \to \infty$*, vers*

$$\ell(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}_0\left(\log \mathbb{P}(X_{1:N_t}, Y_{1:M_t})\right)$$

$$= \sup_t \frac{1}{t}\mathbb{E}_0\left(\log \mathbb{P}(X_{1:N_t}, Y_{1:M_t})\right).$$

ii) $t^{-1}w_t(\theta)$ *converge* $\mathbb{P}_0$*-ps et dans* $\mathbb{L}_1$*, quand* $t \to \infty$*, vers*

$$w(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}_0\left(\log Q_\theta(X_{1:N_t}, Y_{1:M_t})\right)$$

$$= \sup_t \frac{1}{t}\mathbb{E}_0\left(\log Q_\theta(X_{1:N_t}, Y_{1:M_t})\right).$$

On défini alors les contrastes limites : pour tout $\theta \in \Theta_0$,

$$D(\theta|\theta_0) = w(\theta_0) - w(\theta) \quad \text{and} \quad D^*(\theta|\theta_0) = \ell(\theta_0) - \ell(\theta).$$

Résultats : propriétés des contrastes limites

Arribas-Gil, Gassiat, M. [AGGM06]

Soit $\Theta_{exp} = \{\theta \in \Theta_0 \ : \ \forall \lambda > 0, \mathbb{E}(\varepsilon_1) \neq \lambda \mathbb{E}_0(\varepsilon_1)\}$ et $\Theta_{marg} = \{\theta \in \Theta_0 \ : \ h_X = f, \ h_Y = g\}$.

Théorème

- *Pour tous* $\theta \in \Theta_0$, $D(\theta|\theta_0) \geq 0$ *et* $D^*(\theta|\theta_0) \geq 0$.
- *Pour tous* $\theta \in \Theta_{exp}$, $\theta \neq \theta_0$, *on a* $D(\theta|\theta_0) > 0$ *et* $D^*(\theta|\theta_0) > 0$.
- *Si* θ_0, $\theta \in \Theta_{marg}$, *alors* $D(\theta|\theta_0) > 0$ *et* $D^*(\theta|\theta_0) > 0$ *dès que* $f \neq f_0$ *ou* $g \neq g_0$.

Remarque

Sous l'hypothèse **H1**, les moyennes de ε_1 sous $\mathbb{P}$ et $\mathbb{P}_0$ sont alignées avec $(0, 0)$. Dans ce cas, nous ne prouvons pas que si $h \neq h_0$ alors $D(\theta|\theta_0) > 0$ or $D^*(\theta|\theta_0) > 0$.

Résultats : propriétés de continuité

Arribas-Gil, Gassiat, M. [AGGM06]

Soit $\Theta_\delta = \{\theta \in \Theta, \pi_{ij} \geq \delta, f(x) \geq \delta, g(y) \geq \delta, h(x,y) \geq \delta\}$.

Lemme

Les suites $\{t^{-1}w_t(\theta)\}_{t\geq 1}$ *et* $\{t^{-1}\ell_t(\theta)\}_{t\geq 1}$ *sont uniformément équicontinues sur* Θ_δ.

En conséquence,

Corollaire

i) $\{t^{-1}w_t(\theta)\}_t$ *(resp.* $\{t^{-1}\ell_t(\theta)\}_t$*) converge* $\mathbb{P}_0$*-ps vers* $w(\theta)$ *(resp. vers* $\ell(\theta)$*) uniformément sur* Θ_δ *;*

ii) $\ell(\theta)$ *et* $w(\theta)$ *sont uniformément continues sur* Θ_δ.

Re-paramétrisation

Arribas-Gil, Gassiat, M. [AGGM06]

- Rappel : les pairHMMs sont utilisés pour modéliser des processus d'évolution. Le paramètre d'intérêt n'est pas θ, mais les paramètres (β) du modèle d'évolution originel.
- On s'intéresse donc à une re-paramétrisation $\beta \mapsto \theta(\beta)$.
- ex :

$$h(x,y) = \begin{cases} f(x)(1-e^{-\alpha})f(y) & \text{si } x \neq y \\ f(x)\{(1-e^{-\alpha})f(x) + e^{-\alpha}\} & \text{sinon,} \end{cases}$$

où $\alpha > 0$ est le taux de substitution et $f(x)$ est la loi stationnaire de la lettre x dans le modèle d'évolution, que l'on supposera connue.
Le paramètre d'intérêt est alors $\beta = (\pi, \alpha)$.
(NB : ici $f = g = h_X = h_Y$.)

Résultats : propriétés des estimateurs

Arribas-Gil, Gassiat, M. [AGGM06]

Supposons que $\beta \in B \mapsto \theta(\beta) \in \Theta$ est continue et pour tout $\delta > 0$ on note $B_\delta = \theta^{-1}(\Theta_\delta)$. Supposons également $\beta_0 = \theta^{-1}(\theta_0) \in B_\delta$ for some $\delta > 0$. L'algorithme pairHMM calcule

$$\widehat{\beta}_t = \operatorname*{Argmax}_{\beta \in B_\delta} \; w_t(\theta(\beta)).$$

Théorème [version fréquentiste]

Si l'ensemble des maxima de $w(\theta(\beta))$ sur B_δ est réduit à $\{\beta_0\}$, alors $\widehat{\beta}_t$ converge $\mathbb{P}_0$-ps vers β_0.

Théorème [version bayésienne]

Si l'ensemble des maxima de $w(\theta(\beta))$ sur B_δ est réduit à $\{\beta_0\}$, et si β_0 appartient au support de l'a priori ν, alors la suite des lois a posteriori $\nu_{|X_{1:N_t}, Y_{1:M_t}}$ converge en loi $\mathbb{P}_0$-ps vers une masse de Dirac en β_0.

Simulations

Arribas-Gil, Gassiat, M. [AGGM06]

Modèle de substitution

$$h(x,y) = \begin{cases} f(x)(1-e^{-\alpha})f(y) & \text{si } x \neq y \\ f(x)\{(1-e^{-\alpha})f(x) + e^{-\alpha}\} & \text{sinon,} \end{cases}$$

où $\alpha > 0$ est le taux de substitution et $f(x)$ la distribution stationnaire de la lettre x, supposée connue. Les paramètres sont $\beta = (\pi, \alpha)$.

Nous considérons le cas $p = q$ et soit

- simulations de $(\varepsilon_s)_s$ i.i.d. avec p_0 la probabilité d'un pas horizontal ou vertical et $r_0 = 1 - 2p_0$ la probabilité d'un pas diagonal. Les paramètres sont $\beta = (p, \alpha)$.
- simulations de $(\varepsilon_s)_s$ chaîne de Markov stationnaire et $p_0 = q_0$. Il y a 6 paramètres.

Simulations, cas iid (1/3)

Arribas-Gil, Gassiat, M. [AGGM06]

$\alpha_0 = 0.05$ et $p_0 = q_0 = 0.25$

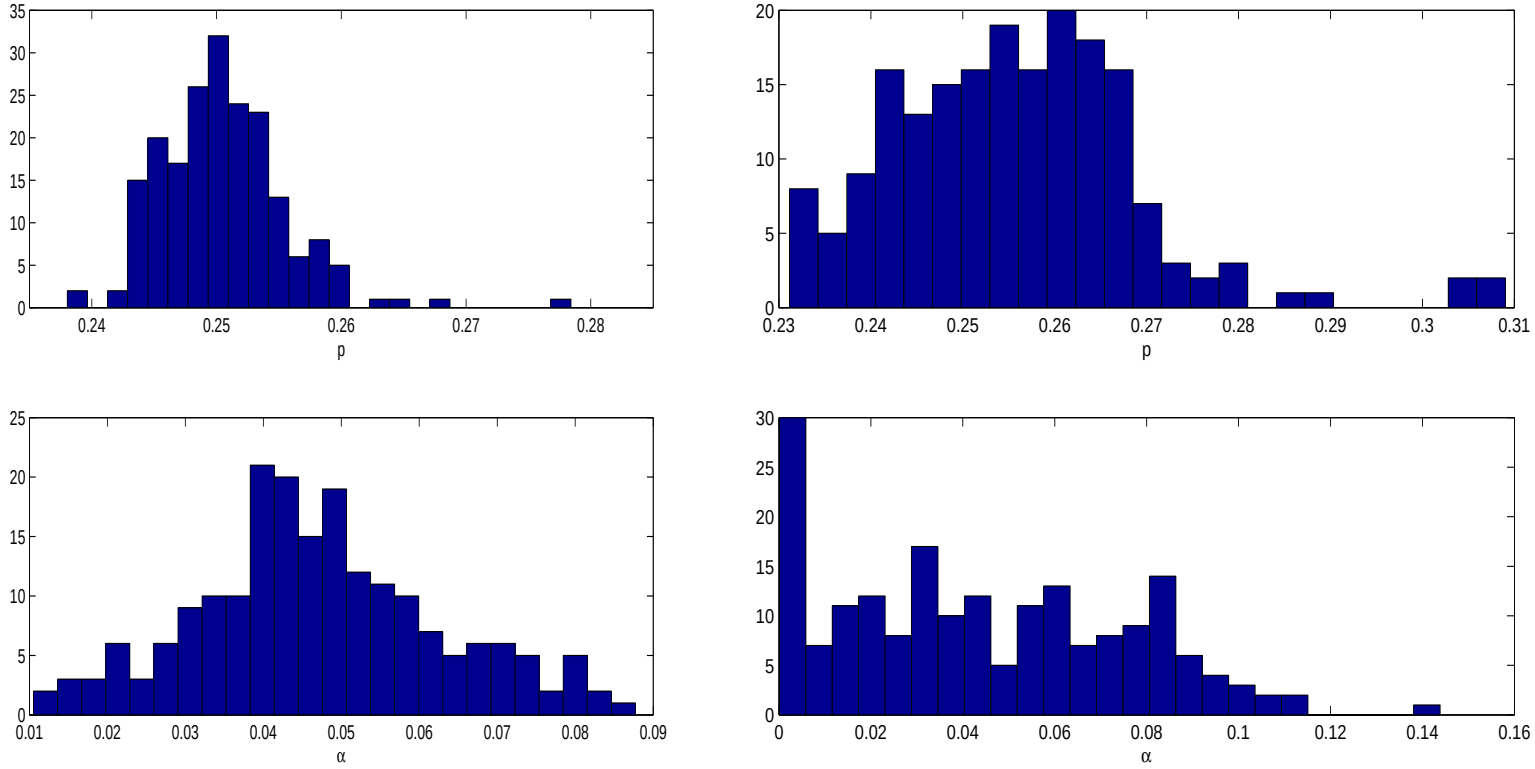


FIG.: Histogrammes de l'EMV obtenu avec 200 simulations d'alignements de longueur 15000 pour le modèle i.i.d. À gauche : estimation de p lorsque $\alpha = \alpha_0$ et estimation d'α lorsque $p = p_0$. À droite : estimation simultanée de p et α.

Simulations, cas iid (2/3)

Arribas-Gil, Gassiat, M. [AGGM06]

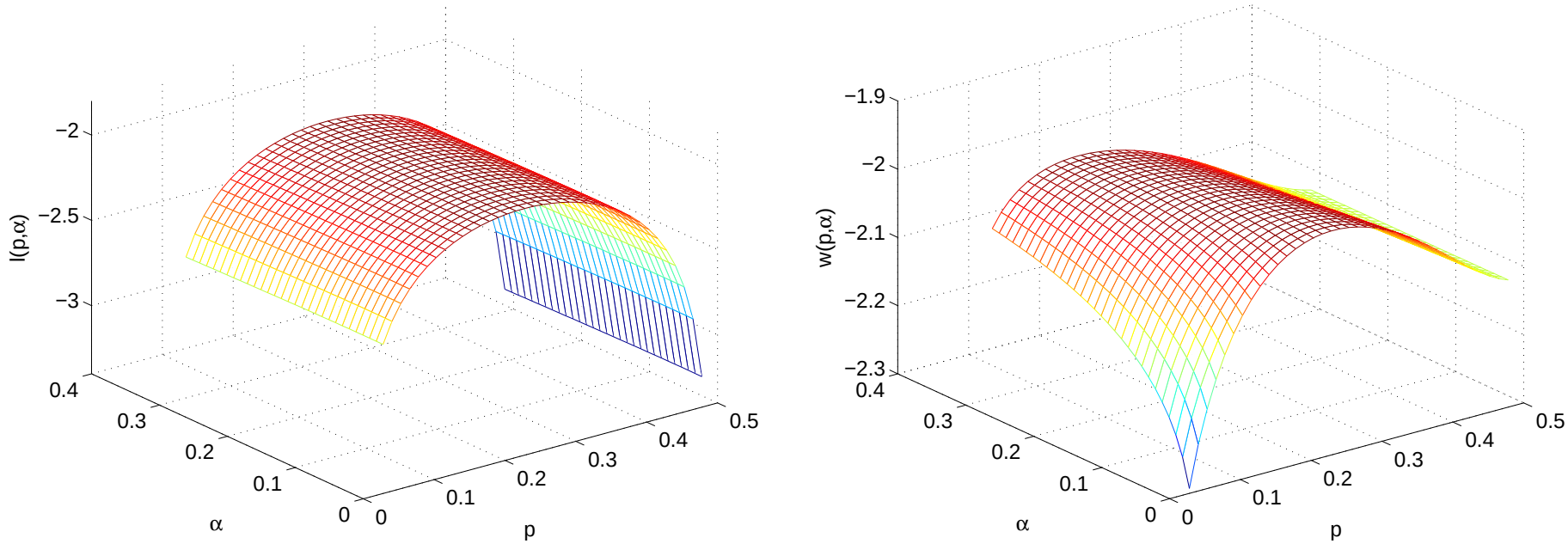


FIG.: ℓ et w pour le modèle i.i.d. $(p_0 = 0.25, \alpha_0 = 0.05)$.

Simulations, cas iid (3/3)

Arribas-Gil, Gassiat, M. [AGGM06]

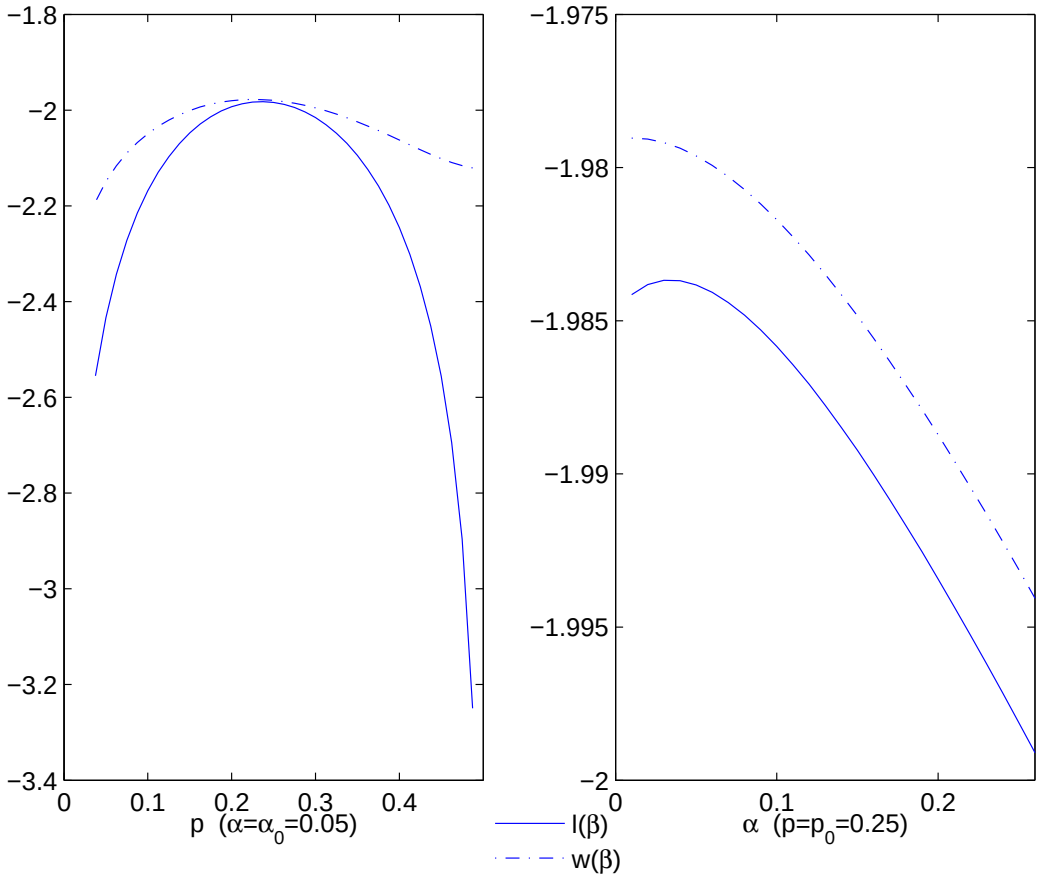


FIG.: Coupes de ℓ et w pour $\alpha = \alpha_0$ fixé (gauche) et pour $p = p_0$ fixé (droite).

Simulations, cas Markov (1/3)

Arribas-Gil, Gassiat, M. [AGGM06]

Nous simulons 200 alignements de longueur 15000, avec un taux de substitution $\alpha_0 = 0.05$ et matrice de transition

$$\begin{array}{c} \\ D \\ H \\ V \end{array} \begin{array}{c} \begin{array}{ccc} D & H & V \end{array} \\ \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.3 & 0.5 & 0.2 \\ 0.3 & 0.1 & 0.6 \end{pmatrix} \end{array}$$

de loi stationnaire associée $p_0 = q_0 = 0.25$.
Les paramètres sont π_{HH}, π_{HV}, π_{DV}, π_{VV} et π_{DH}. Nous fixons $f(x) = 0.25$ pour tout x.

Simulations, cas Markov (2/3)

Arribas-Gil, Gassiat, M. [AGGM06]

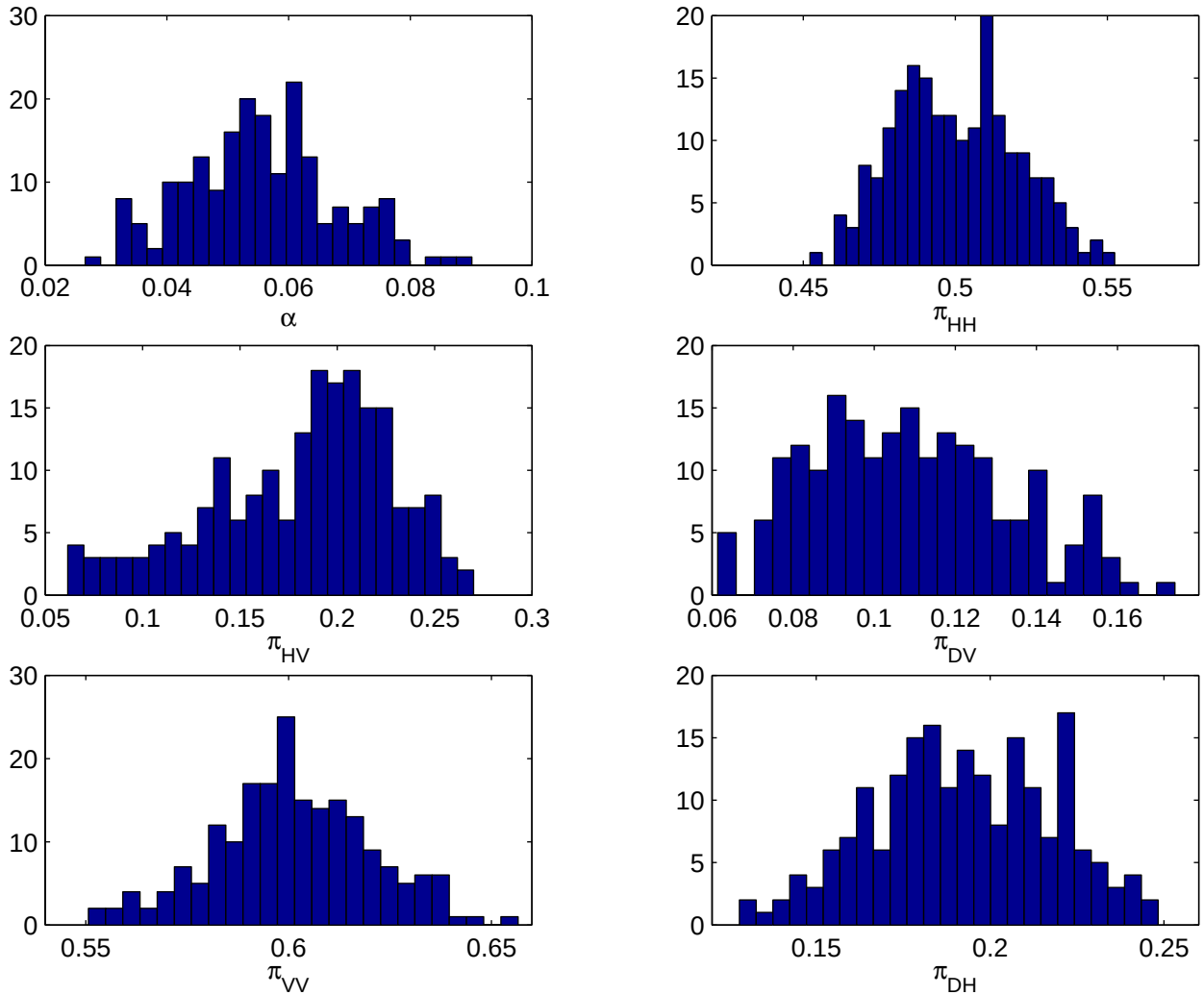


FIG.: Estimation des probas de transition sachant $\alpha = \alpha_0$ (gauche) et estimation de α sachant les vraies valeurs des transitions (droite).

Simulations, cas Markov (3/3)

Arribas-Gil, Gassiat, M. [AGGM06]

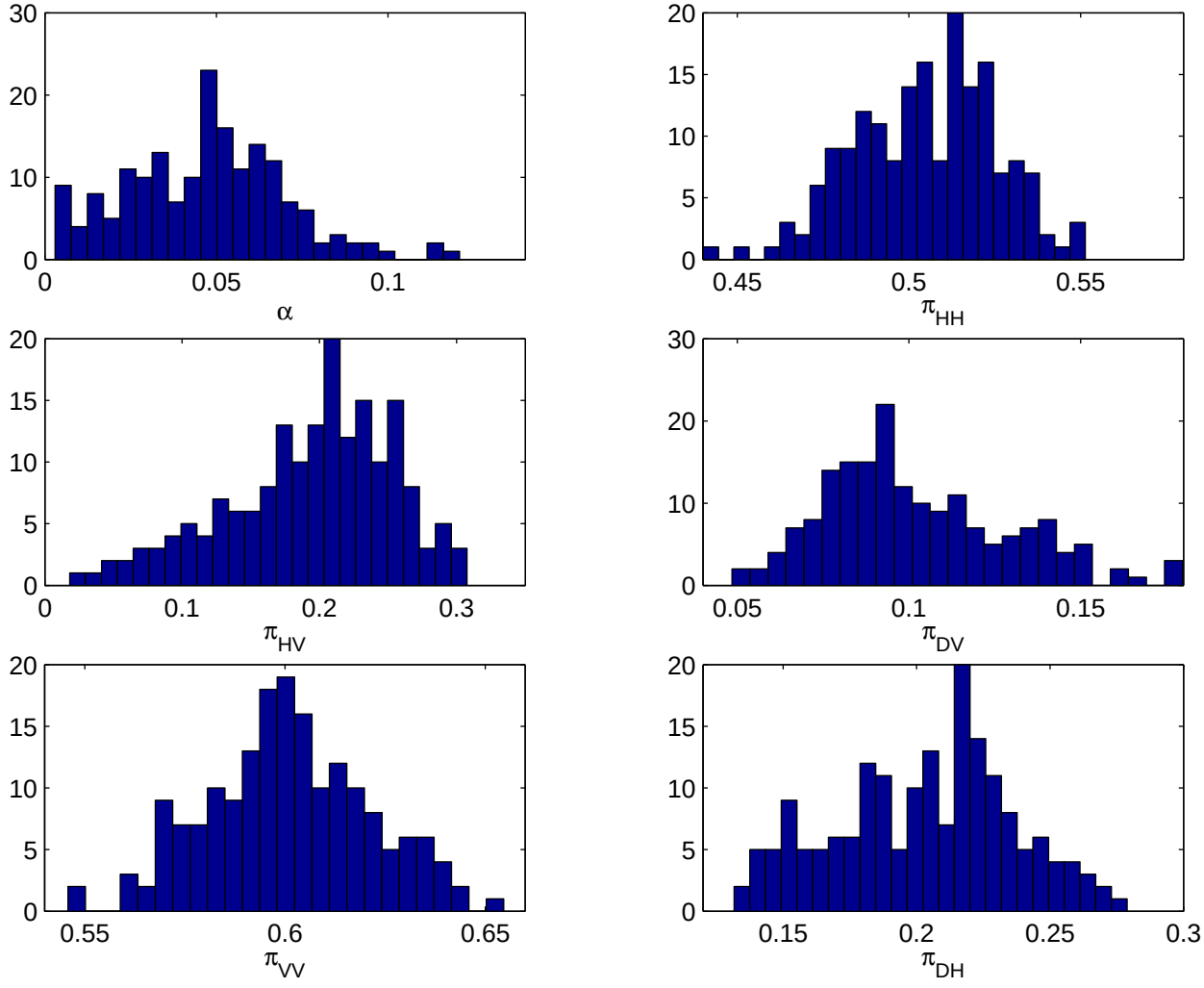


FIG.: Estimation simultanée des probabilités de transition et de α.

Conclusions

- L'estimation des paramètres du modèle d'évolution dans le contexte pairHMM est consistante dans de nombreux cas (et sûrement tout le temps).
- Un des avantages des modèles pairHMM par rapport aux méthodes de score est qu'ils permettent d'obtenir une loi a posteriori sur les alignements.

Probabilités a posteriori d'alignements

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

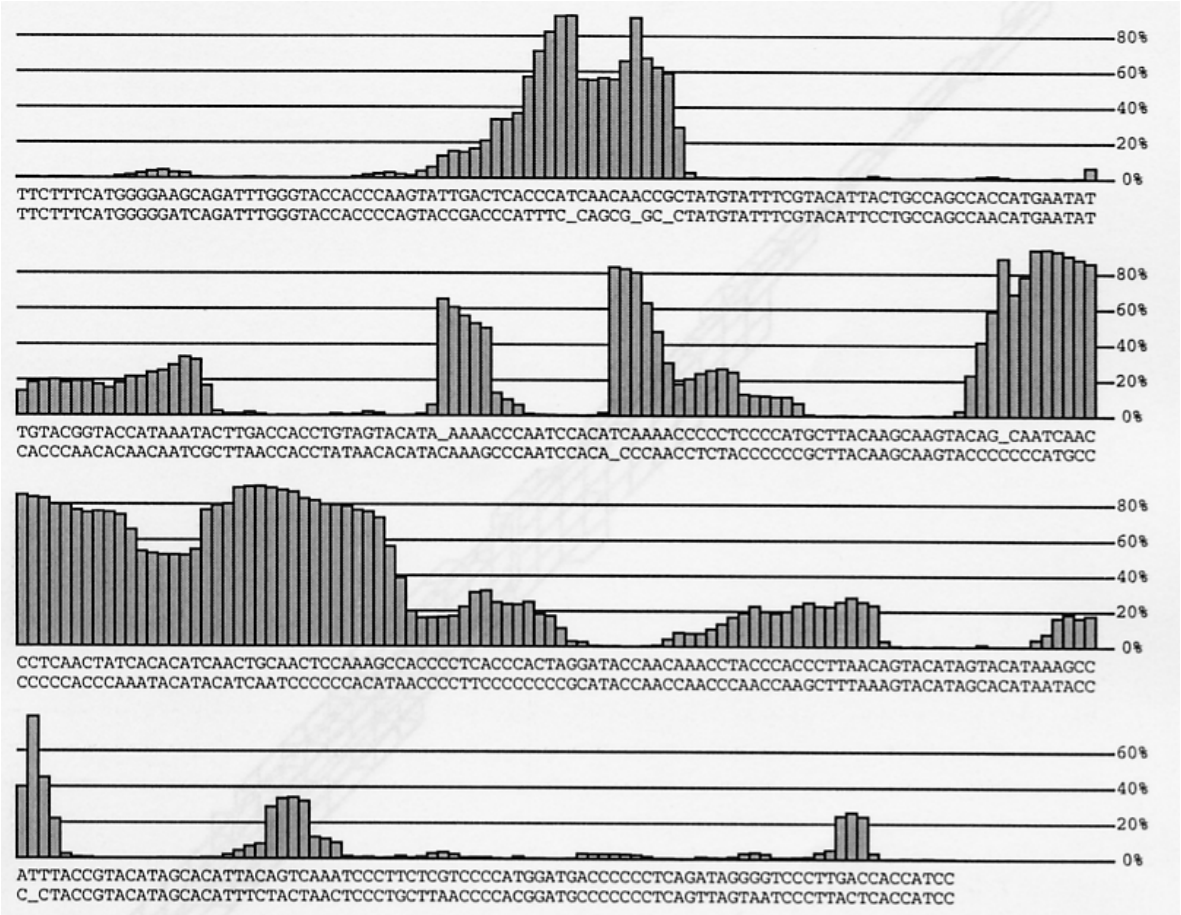


Figure 3: The most probable of the sampled alignments of a human and an orangutan HVR-1 sequence and the percentages of sampled alignments that differ from it in each position.

Perspectives

- Propriétés de cette loi a posteriori sur les alignements ?
- Le modèle pairHMM permet de faire un alignement global des séquences. Il existe des extensions pour permettre l'alignement local (« alignement hybride », [YBH02, YH01]). Propriétés de ces modèles ?
- Un des problèmes majeurs des modèles d'évolution à l'heure actuelle, c'est qu'ils font souvent l'hypothèse que les sites sont indépendants. Il existe des travaux pour sortir de ce cadre : [LH04, SH04, JP00]... Peut-on faire de l'alignement statistique avec dépendances au voisin du modèle d'évolution ?
- À noter : il existe des procédures d'alignement par score pour lesquelles le score dépend du contexte [GTT06, Hua94, WL84].
- L'article [LDMH05] contient une revue intéressante sur les problématiques de l'alignement par max. de vraisemblance.

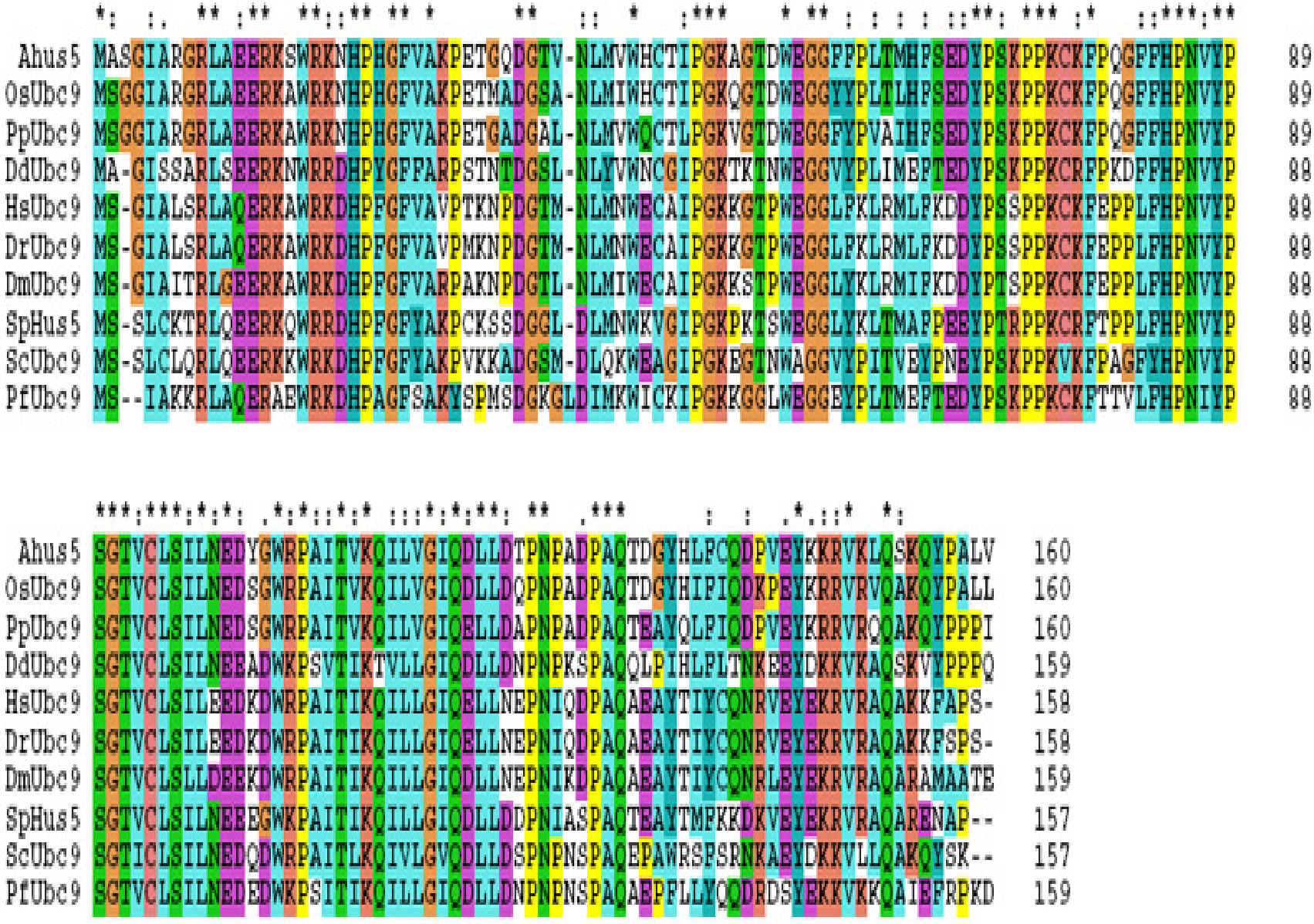
Quatrième partie IV

Alignement multiple

Sommaire Partie IV : Alignement multiple

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 (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).

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.

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`.

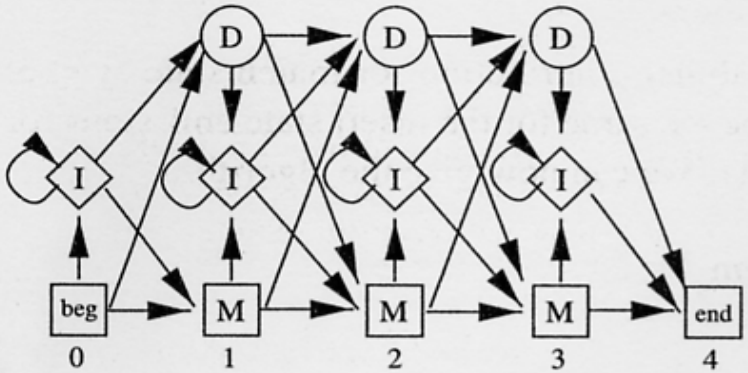
Chaînes de Markov cachées profils (2/3)

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

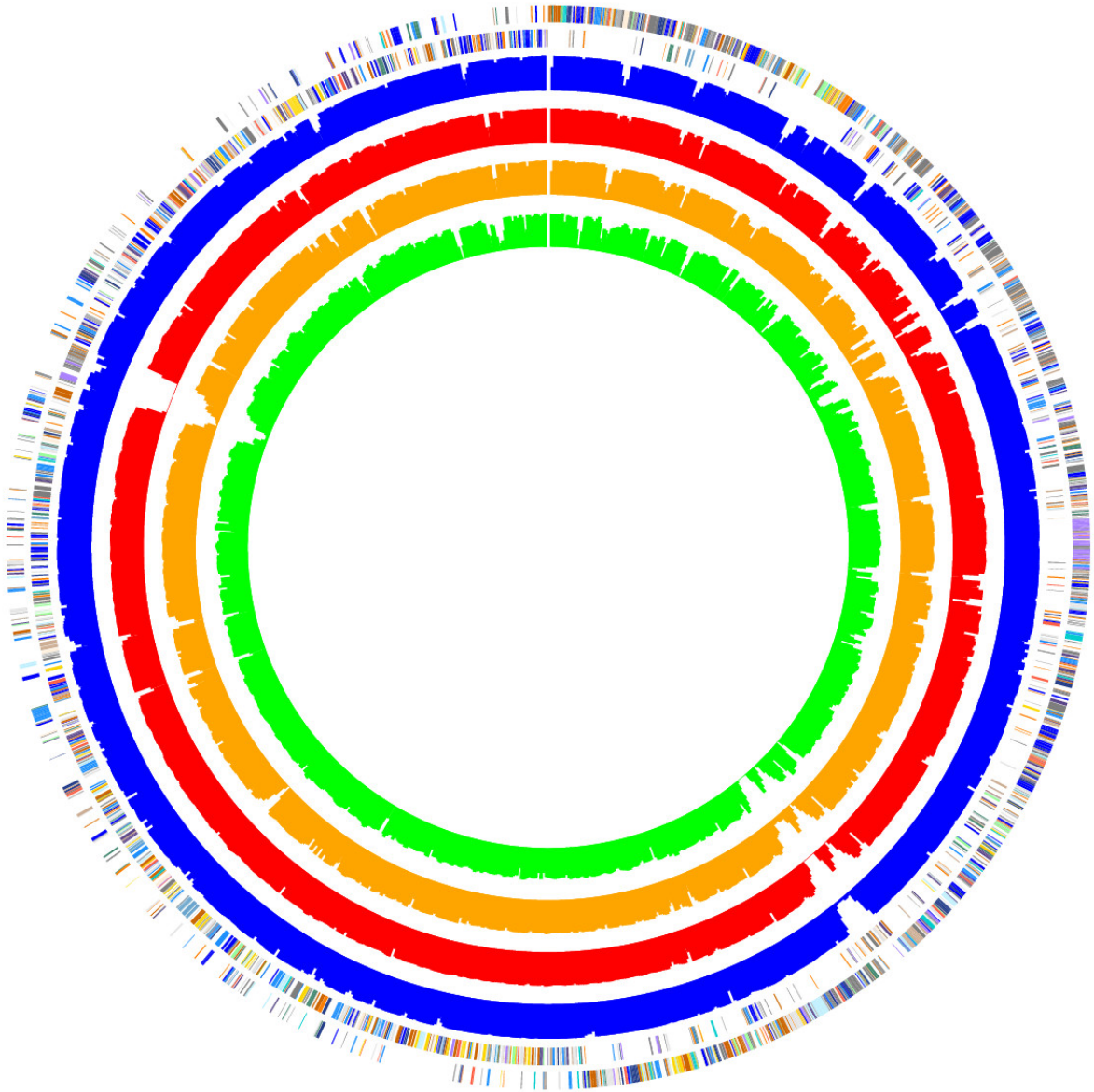
- La généralisation du modèle pairHMM à plus de deux séquences n'est pas le profileHMM.
- Différence = en profilHMM, conditionnellement à la chaîne profil, les séquences sont indépendantes.
- Dans un cadre d'align. stat. multiple, les lettres d'une colonne d'un site homologue sont émises selon une loi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

Cinquième partie V

Alignement de génomes complets

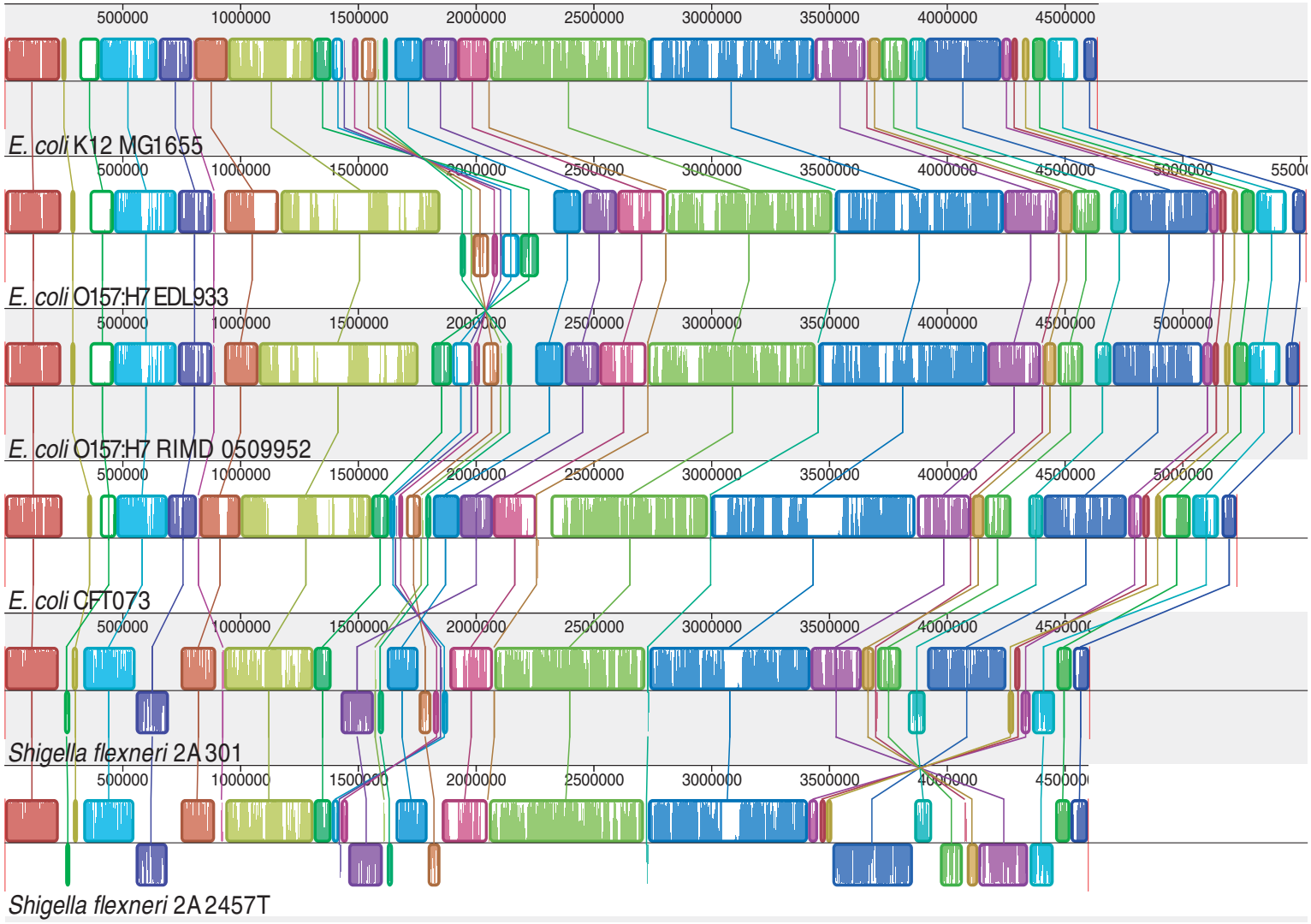
Exemple : alignement des génomes complets de 5 souches de *Listeria*

(Source Ghai *et al.*, BMC Bioinformatics 2004, **5** :198)



Exemple : alignement des génomes complets de 6 souches d'*E. Coli*

(Source Mau *et al.*, Genome Biology 2006, **7** :R44)



Introduction à l'alignement de génomes complets

Méthodes

- Combinaison d'alignements locaux de paires de séquences (MULTIZ).
- Méthodes hiérarchiques : combinent de l'alignement global avec des cartes d'homologie (MAVID, MLAGAN).

Commentaires

- Bcp de développements algorithmiques, peu de développements statistiques.

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] Ana Arribas-Gil, Elisabeth Gassiat, and Catherine Matias.
Parameter estimation in pair-hidden Markov models.
Scand. J. Statist., 33(4) :651–671, 2006.

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

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

[Cha05] Hock Peng Chan.

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


[DEKM98] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison.
Biological sequence analysis. Probabilistic models of proteins and nucleic acids.
Cambridge : Cambridge University Press., 1998.


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


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


[DP06] Colin N. Dewey and Lior 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] Roland Fleissner, Dirk Metzler, and Arndt 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`.

[GTT06] Anna Gambin, Jerzy Tiuryn, and Jerzy Tyszkiewicz.
Alignment with context dependent scoring function.
J. Comput. Biol., 13(1) :81–101 (electronic), 2006.

[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.

[Hua94] Xiaoqiu Huang.
A context dependent method for comparing sequences.
In *Combinatorial pattern matching (Asilomar, CA, 1994)*, volume 807 of *Lecture Notes in Comput. Sci.*, pages 54–63. Springer, Berlin, 1994.

[JP00] Jens Ledet Jensen and Anne-Mette Krabbe Pedersen.
Probabilistic models of DNA sequence evolution with context dependent rates of substitution.
Adv. in Appl. Probab., 32(2) :499–517, 2000.

[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.

[LH04] Gerton Lunter and Jotun Hein.
A nucleotide substitution model with nearest-neighbour interactions.
Bioinformatics, 20(1) :216–223, 2004.

[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.

[SH04] Adam Siepel and David Haussler.
Phylogenetic Estimation of Context-Dependent Substitution Rates by Maximum Likelihood.
Mol. Biol. Evol., 21(3) :468–488, 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.


[WL84] W. John Wilbur and David J. Lipman.
The context dependent comparison of biological sequences.
SIAM J. Appl. Math., 44(3) :557–567, 1984.

[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.