II. Génomique comparative

Catherine Matias

CNRS - Laboratoire de Probabilités et Modèles Aléatoires, Paris
catherine.matias@math.cnrs.fr
http://cmatias.perso.math.cnrs.fr/

ENSAE - 2014/2015

Sommaire

Première partie I

Génomique comparative

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,
- …

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 ? (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

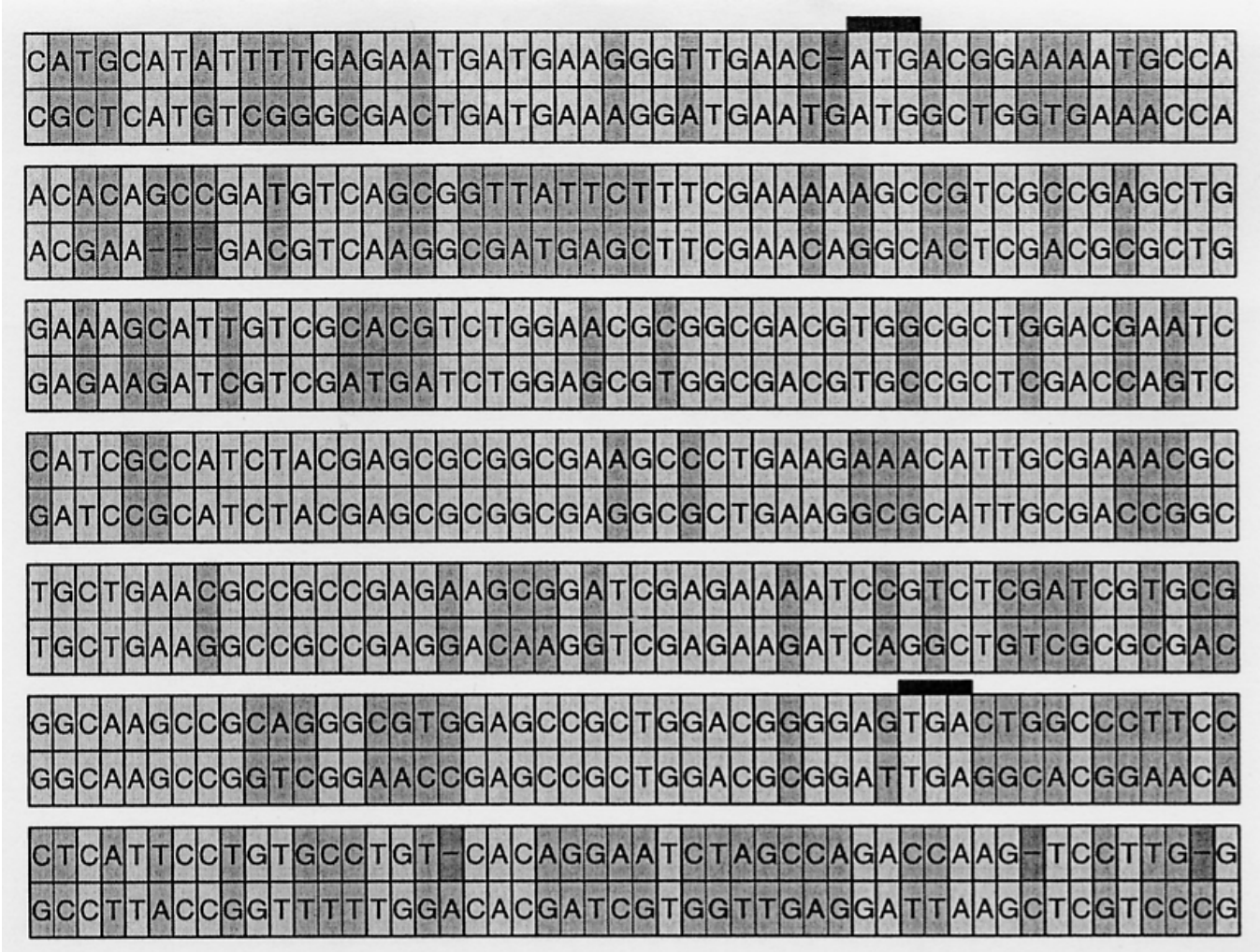


FIG. 3. Part of the pairwise alignment of *A.tumefaciens* and *M.loti.* Light gray color corresponds to conserved positions, and nonconserved positions and gaps are shown in dark gray. The two black bars on top of the alignment

Que représente un alignement ? I

- 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 ? II

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.

Sommaire

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*

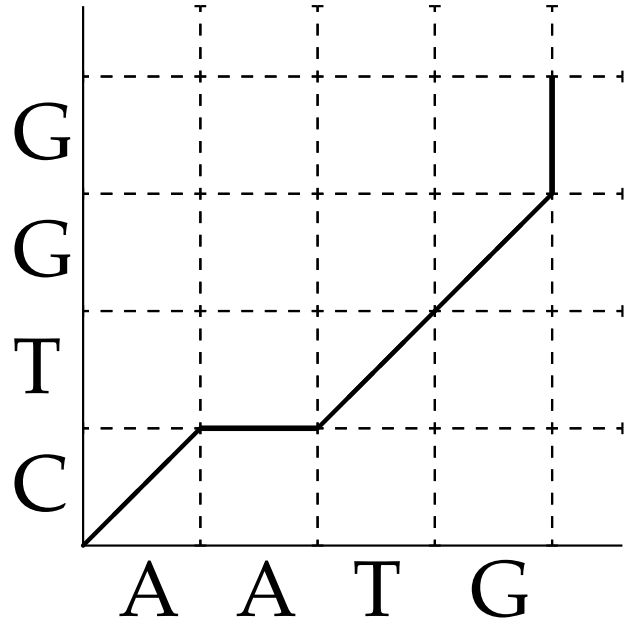


FIGURE : Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{matrix} A & A & T & G \\ C & - & T & G & G \end{matrix}$ -.

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)

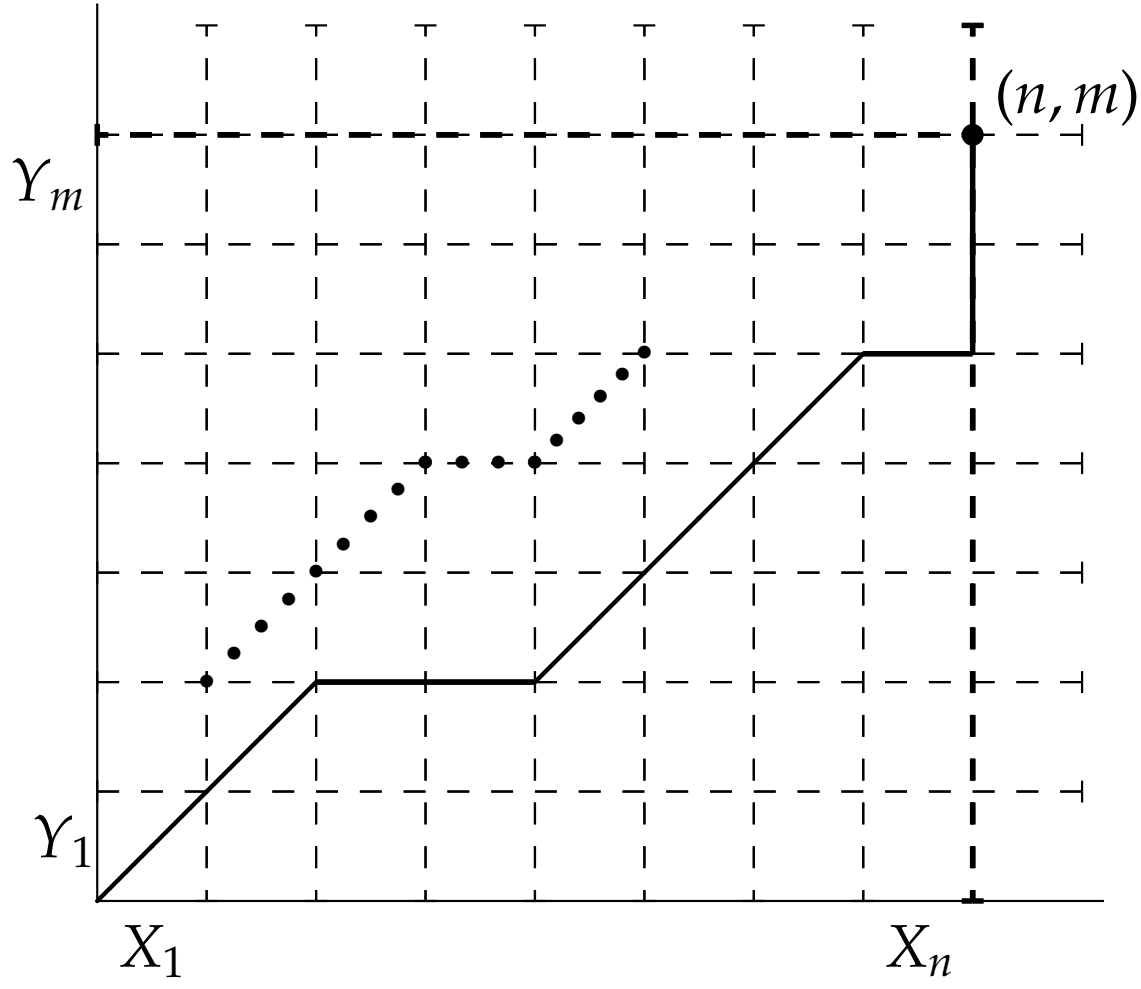
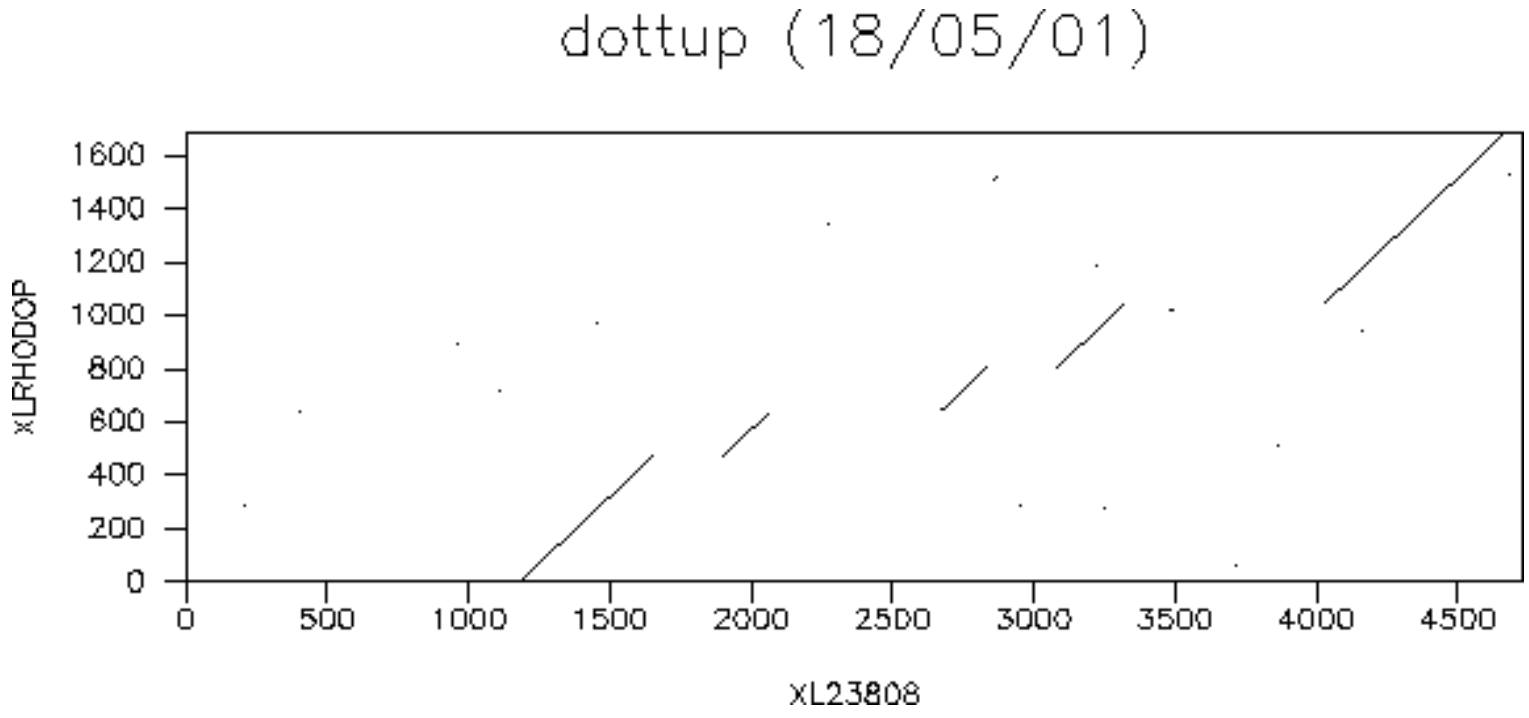


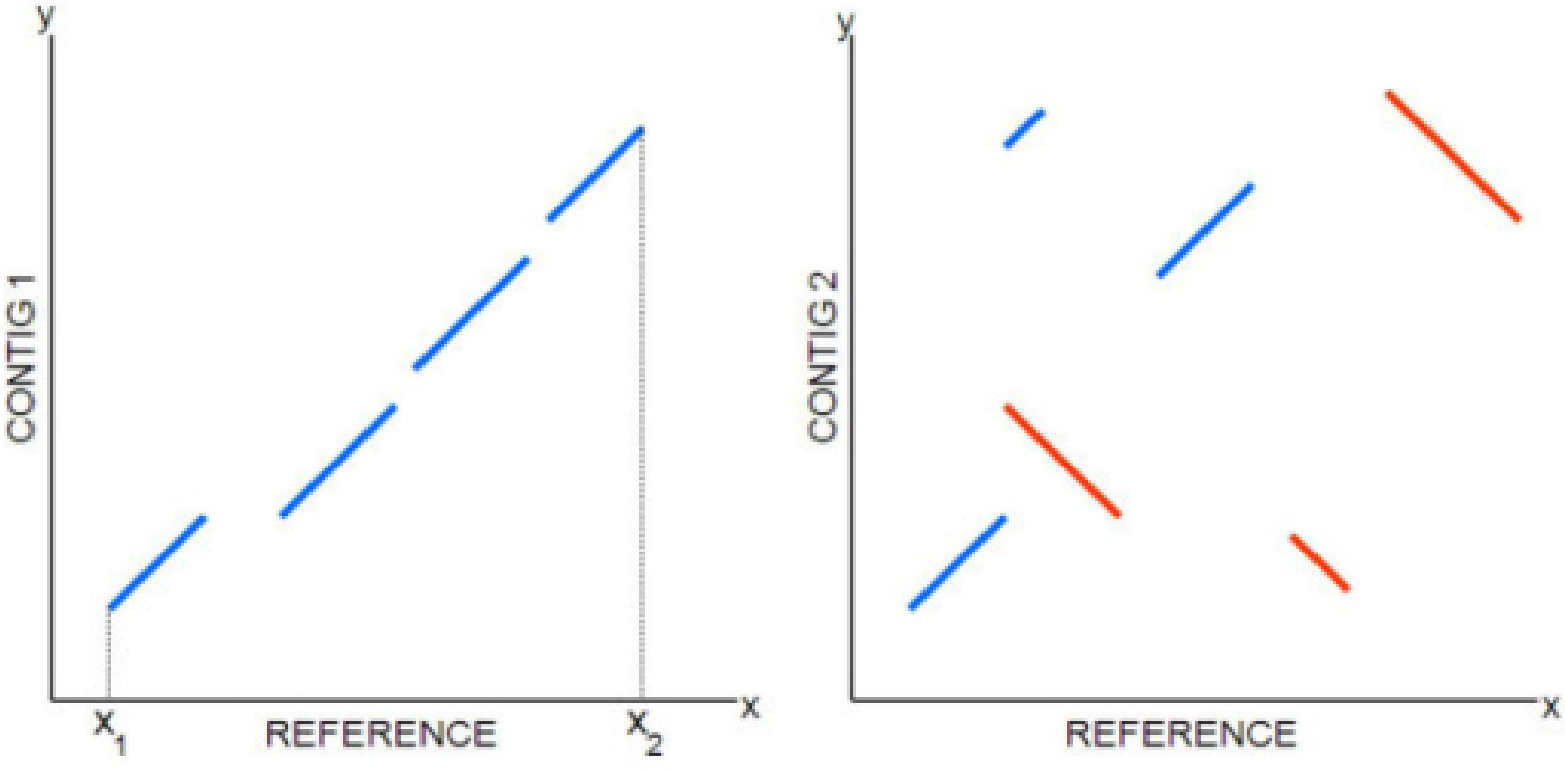
FIGURE : Représentation graphique du meilleur alignement global (traits pleins) et local (traits pointillés) des séquences $X_{1:n}$ et $Y_{1:m}$.

Dotplots : comparaison visuelle I

On dispose chaque séquence le long des axes et un point représente une identité entre deux positions.



Dotplots : comparaison visuelle II



Inconvénient : on ne voit que les identités, pas les similarités.

Deuxième partie II

Alignement de deux séquences 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 ».

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 I

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

Formalisation mathématique II

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.

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

- But : Détecter une zone de score maximal et sa valeur.
- NB : méthode distincte du principe de fenêtre glissante.

Remarques

- 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.
- Équilibre à faire entre les scores des lettres et la pénalité des insertions-délétions. Influence sur le type d'alignements obtenus.
- 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.

Sommaire

Algorithmes exacts I

- 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 II

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 initialise $F(0,0) = 0$, $F(i,0) = -\delta i$ et $F(0,j) = -\delta j$,
- Puis

$$\begin{array}{ccc} F(i-1,j-1) & & F(i-1,j) \\ & \searrow & \downarrow \\ F(i,j-1) & \rightarrow & F(i,j) \end{array}$$

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

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

Algorithmes exacts III

Programmation dynamique - al. local - pénalité linéaire

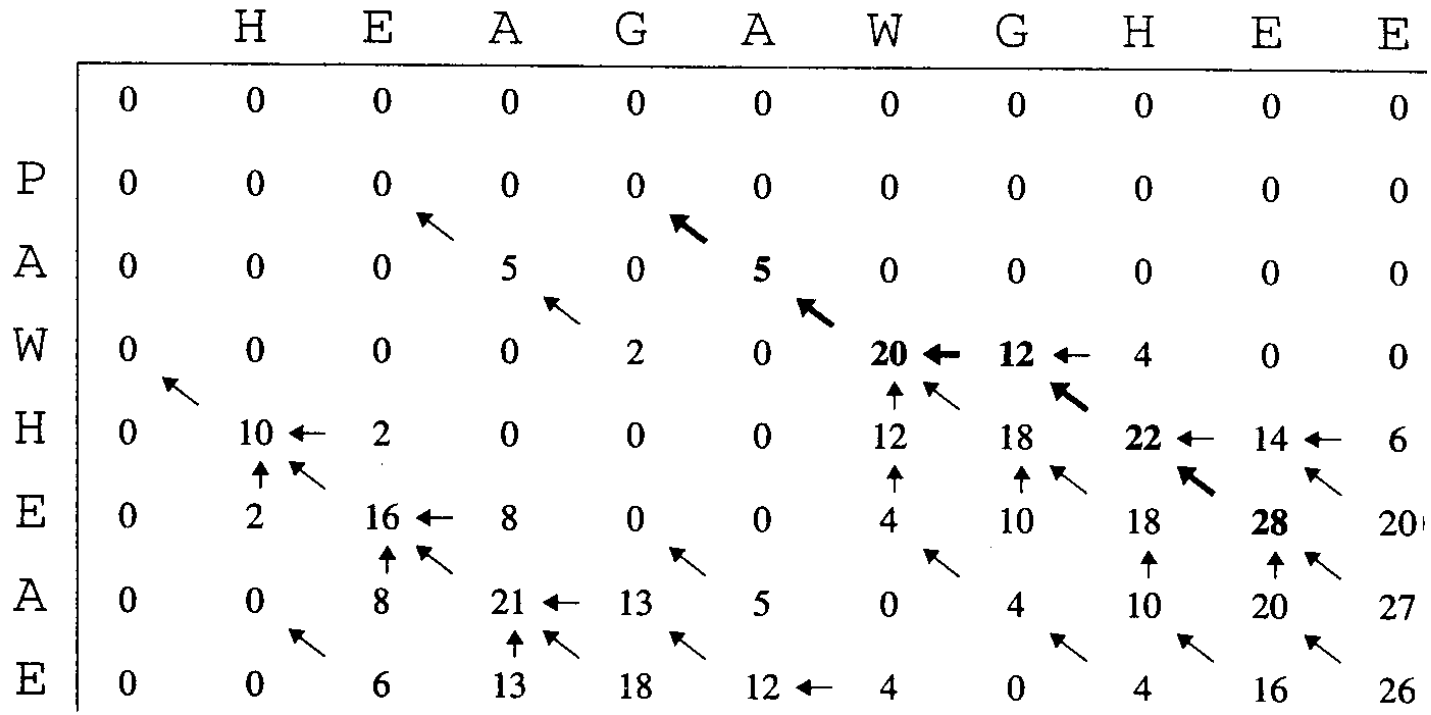
- Soit $F(i,j)$, le meilleur score d'alignement (local) entre $X_{1:i}$ et $Y_{1:j}$,
- On initialise $F(0,0) = F(i,0) = F(0,j) = 0$,
- Puis

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

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

Algorithmes exacts IV

(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...

Sommaire

Matrices de comparaison I

- 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 $\mathcal{A} =$ {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 II

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).

Troisième partie III

Alignement statistique

Sommaire

Contexte

Alignement par score 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...

Sommaire

Le modèle d'évolution TKF I

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.

Le modèle d'évolution TKF II

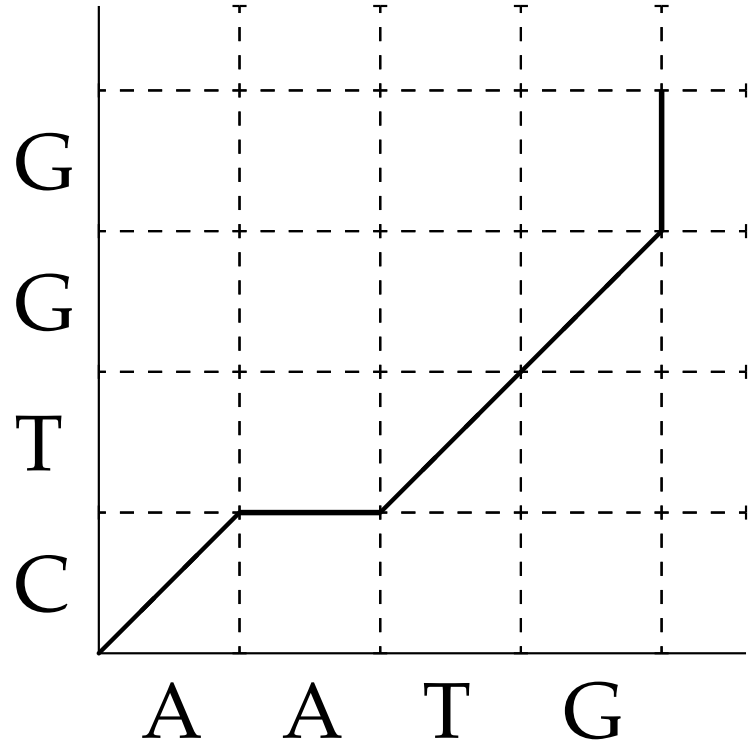
Conséquences

- 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.
- 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 I

Rappel : représentation graphique d'un alignement



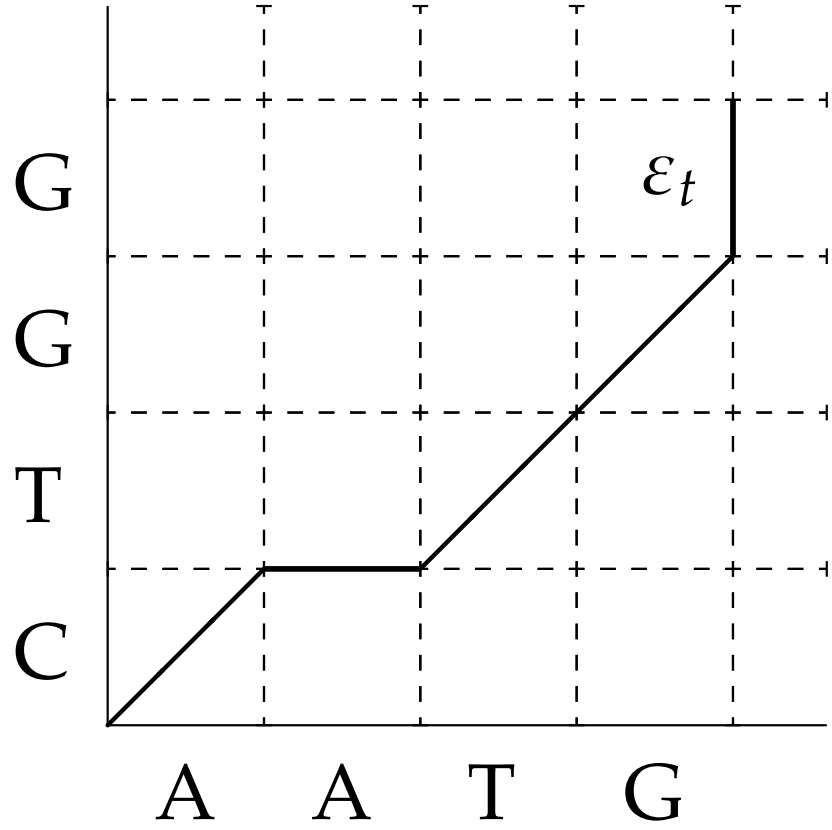
Représentation graphique d'un alignement entre les deux séquences $X = AATG$ et $Y = CTGG$. L'alignement représenté correspond à $\begin{smallmatrix} A & A & T & G & - \\ C & - & T & G & G \end{smallmatrix}$.

Le modèle pair-Markov caché II

Notations [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 III

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

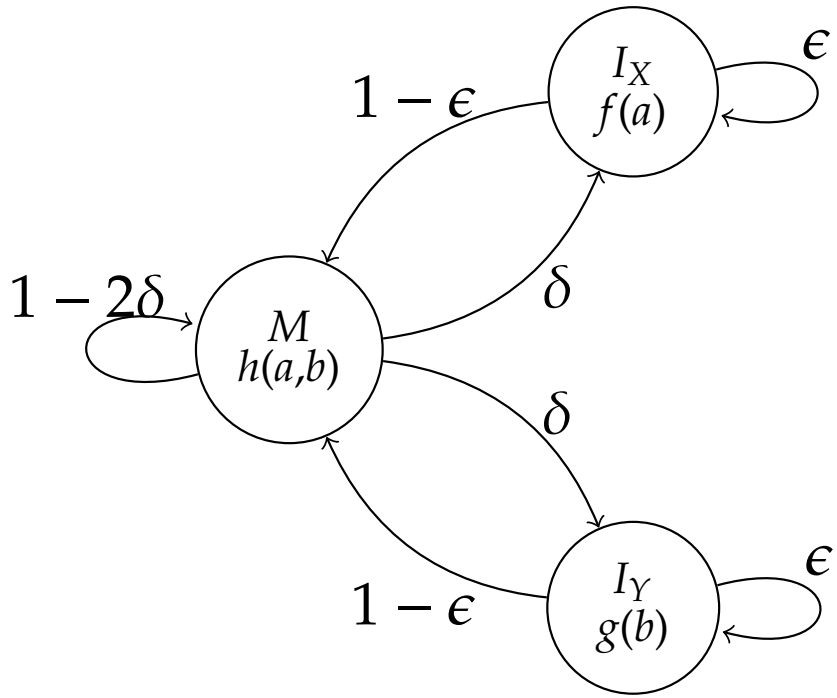
On a

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

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 IV

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.
- NB : L'article [LDMH05] contient une revue intéressante sur les problématiques de l'alignement par max. de vraisemblance.

Probabilités a posteriori d'alignements

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

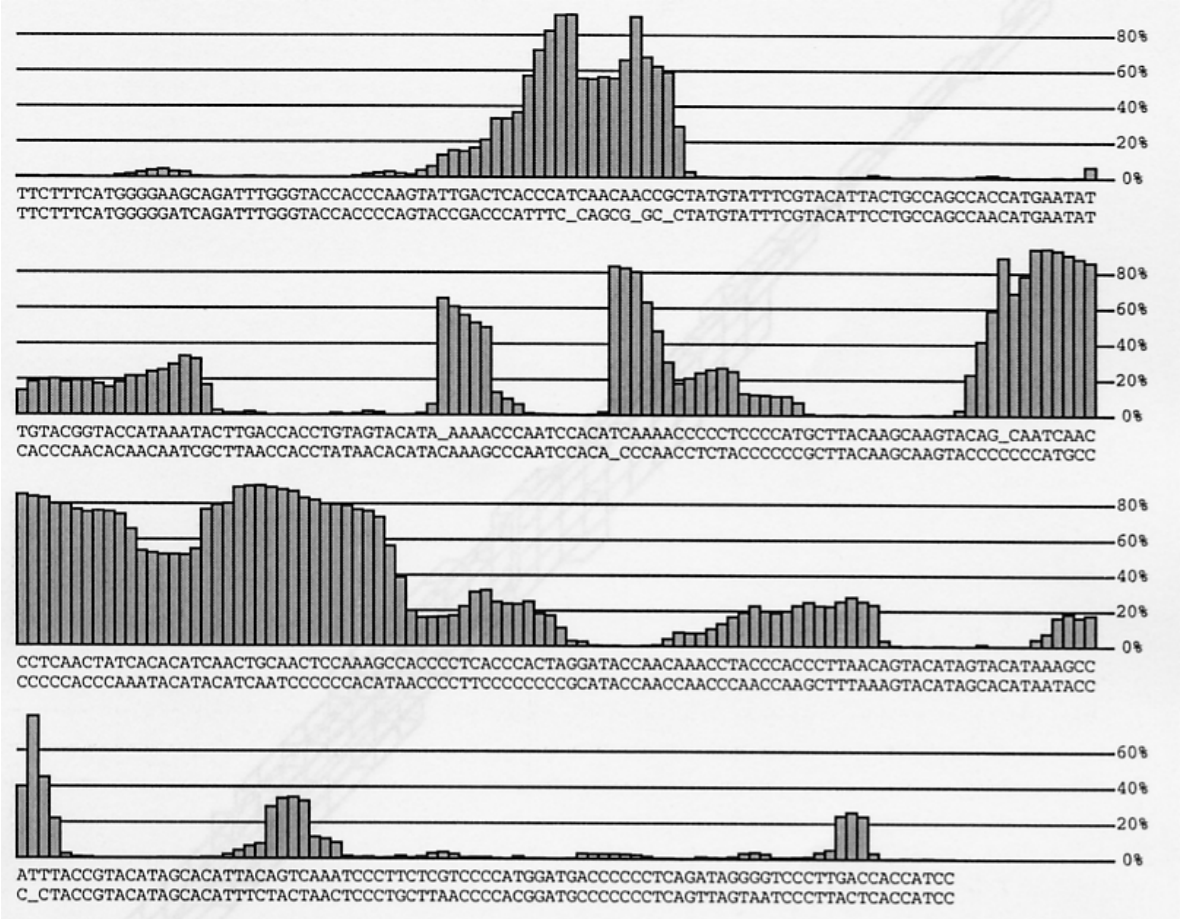


Figure 3: The most probable of the sampled alignments of a human and an orangutan HVR-1 sequence and the percentages of sampled alignments that differ from it in each position.

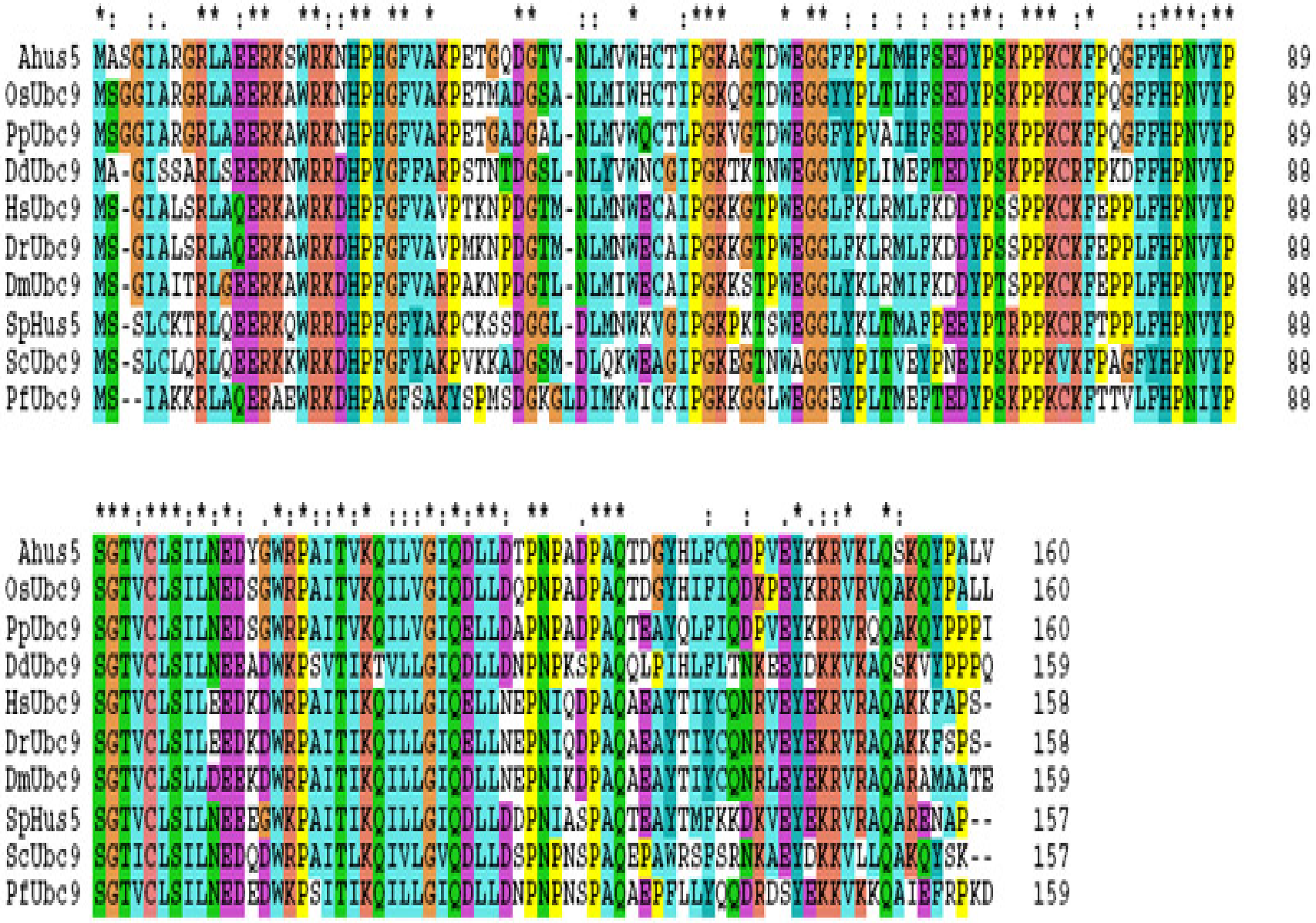
Quatrième partie IV

Alignement multiple

Sommaire

Alignement multiple de séquences

Alignement de protéine Hus5/Ubc9 dans divers organismes



Introduction à l'alignement multiple I

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.
- 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 II

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

Principe

- La généralisation du modèle pairHMM présenté ci-dessus à plusieurs séquences est non triviale (voir [AG07]).
- 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 I

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`.

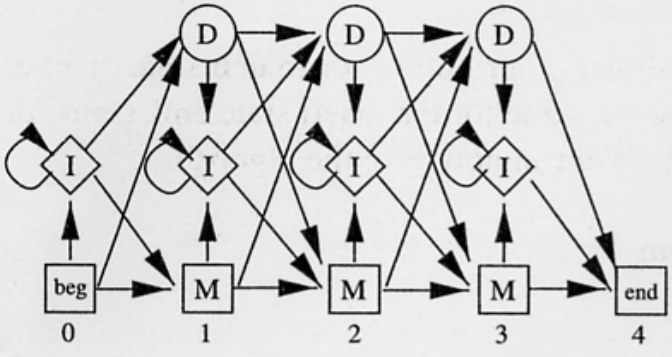
Chaînes de Markov cachées profils II

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 III

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.

Chaînes de Markov cachées profils IV

ProfileHMM vs Alignement statistique multiple

- La généralisation du modèle pairHMM à plus de deux séquences n'est pas le profileHMM.
- Différence = en profileHMM, 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].

Références I


[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.


[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.

Références II


[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.

Références III


[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 `www.math.uni-frankfurt.de/~stoch/Leute/grossmann/dissertation.pdf`

Références IV


[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.

Références V


[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.


[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.

Références VI


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


[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.