

Alignement multiple

Alignement multiple

alignement de deux séquences $\Rightarrow$ alignement de plusieurs séquences
: (extension naturelle pour l'informaticien)
: une philosophie différente pour le biologiste

2 séquences : est-ce qu'elles ont reliées ?

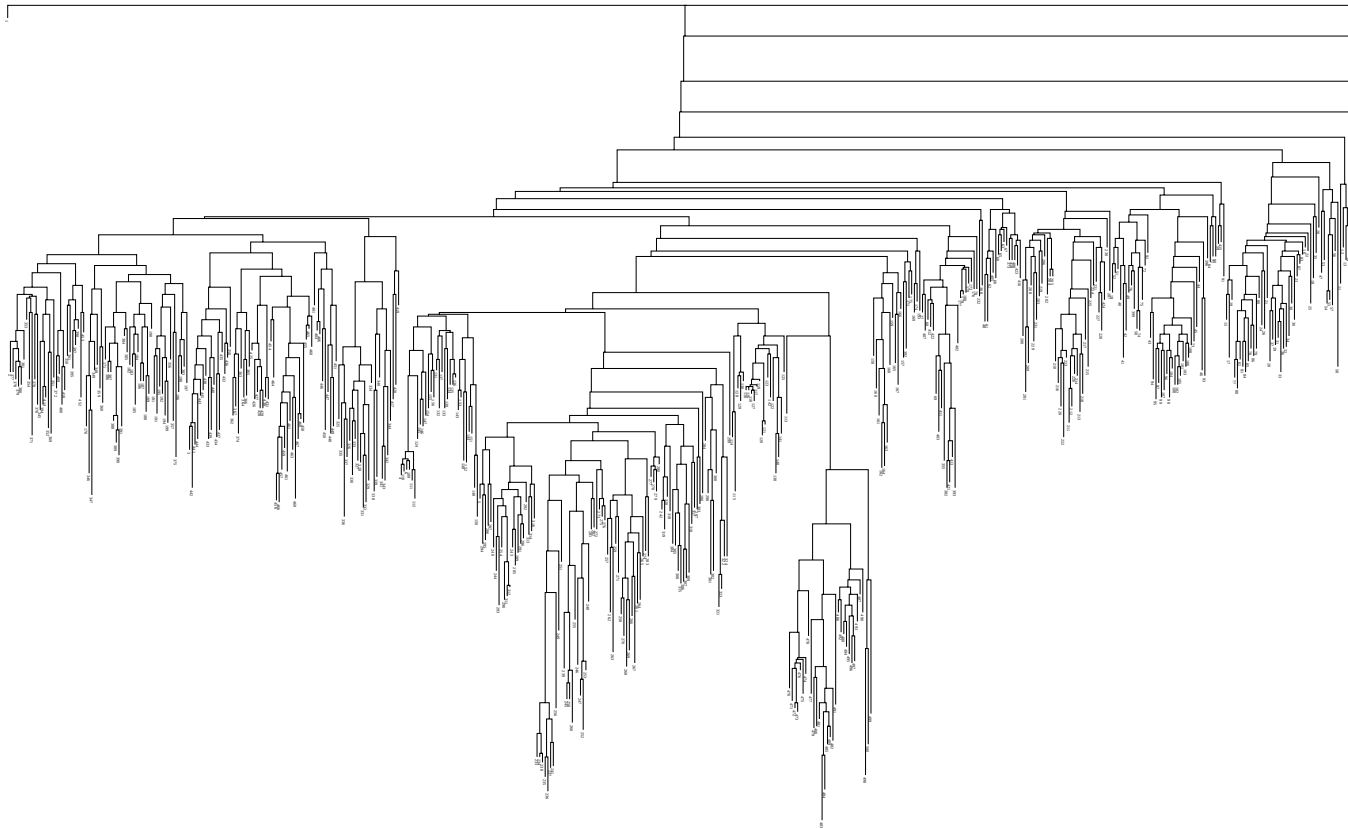
k séquences (reliées) : quels sont les traits communs ?

Applications

- représentation de familles de protéines
- identification et représentation de caractéristiques préservés et leur cor-
relation à la structure/fonction
- inférence de l'Évolution

un problème difficile . . .

Séquences distantes



Un alignement multiple

Lycopersi	CGAGCGCGTT	GTTGGAGAAA	AAGATCAATA	TATTGCTTAT
Convolvul	- - -GCGCGTT	ATTGGAGAAA	AAGATCAATT	TATTGCTTAT
Ipomoea	CATGCGCGTT	GTTGGAGAAA	AAGATCAATA	TATTGCTTAT
Borago	CGATGCCGTT	CCGGGAGAAG	AAAATCAATA	TATATGTTAT
Heliotrop	CGATCCCGTT	CCTGGAGACG	AAGATCAATA	TATTGCTTAT
Hydrophyl	CGATCCCGTT	CTTGGAGAAG	AAGATCAATA	TATTGCTTAT
Eriodicty	CGATCCCGTT	CCTGGAGAAG	AAGATCAATA	TATTTGTTAT
Digitalis	TGAGCCCGTT	CCTGGAGAAG	CAGATCAATA	TATCTGTTAT
Buddleja	CGAGCCCGTT	CCTGGAGAAA	CAGATCAATA	TATCTGTTAT

Valeur de l'alignement ?

Scores

On a k séquences de longueur n ...

Problème : comment spécifier la valeur d'une colonne dans l'alignement ?

Approche 1 : Spécifier la valeur de toutes les colonnes possibles :

$\left(\Sigma \cup \{-\}\right)^k$ si on a k séquences.

Approche 2 : fonction SP «sum of pairs» — considérer l'alignement de chaque paire de séquences :

$$\text{score} = \sum_{i < j} \text{score}(\text{ali de } S_i, S_j)$$

Calcul

Solution par programmation dynamique :

matrice de taille n^k , récurrence avec $2^k - 1$ termes ...

$$A(i_1, i_2, \dots, i_k) = \max \left\{ \begin{array}{l} A(i_1 - 1, i_2, \dots, i_k) + \text{score}(S[i_1], -, \dots, -), \\ \dots, \\ A(i_1 - 1, i_2 - 1, \dots, i_k) + \text{score}(S[i_1], S[i_2], \dots, S[i_k]) \end{array} \right\}$$

PD prend temps exponentiel ... est-ce qu'il y a une meilleure méthode ?

Thm. Il est NP-difficile de trouver l'alignement multiple qui est optimale selon SP.

D'autres définitions d'alignement

Séquence de consensus S^* , $\text{score} = \sum_i \text{score}(\text{ali de } S_i, S^*)$

Alignement à un arbre phylogénétique (viendra plus tard)

Représentation de l'information dans l'alignement

0. séquence de consensus

1. signature («patterns»)

2. profile [trou OK] ou PSSM [sans trou] «position-specific scoring matrix»

3. modèle de Markov caché : HMM «Hidden Markov Model»

problème : alignement à une famille

Signatures — PROSITE

entrée PS00028 : class I zinc-finger pattern

(un motif important dans facteurs de transcription)

cca. 600 séquences

```
ID    ZINC_FINGER_C2H2_1; PATTERN.  
AC    PS00028;  
DT    APR-1990 (CREATED);  
DT    JUN-1994 (DATA UPDATE);  
DT    JUL-1998 (INFO UPDATE).  
DE    Zinc finger, C2H2 type, domain signature.  
PA    C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.  
NR    /RELEASE=38,80000;  
NR    /TOTAL=2189(453); /POSITIVE=2147(412);  
NR    /UNKNOWN=6(6); /FALSE_POS=36(35);  
NR    /FALSE_NEG=3; /PARTIAL=2;
```

Syntaxe de PROSITE

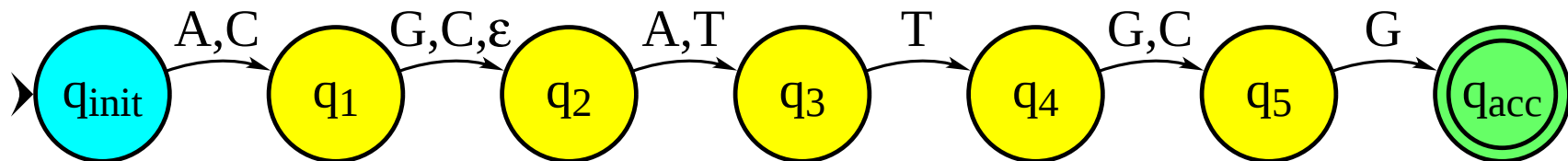
- lettres pour les acides aminés ; X acide arbitraire
- [...] : choix alternatifs dans une position
- {...} : choix exclus dans une position
- (i, j) : répété $i-j$ fois
- – séparateur entre les positions

zinc-finger motif : C-x(2, 4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3, 5)-H
description de cca. 10^{27} séquences possibles

Automates finis et alignements

Un alignement multiple peut être représenté par un automate fini (ou expression régulière) :

CGATCG
C-ATGG
ACTTCG



on a perdu l'info sur les fréquences des lettres ...

Profile

→ enregistrer la fréquence de symbols dans l'alignement multiple

Problème : trouver le profile dans une séquence

alignement de longueur n sur k séquences

calculer $p_j(a)$: fréquence/probabilité de caractère a dans colonne j ,

où $a \in \Sigma$ et $j = 1, \dots, n$

Alignement à un profile

Valeur d'un alignement d'une séquence S à un profile P

1. $S[i]$ aligné à colonne j de P

$$\text{score}(S[i], P[j]) = \sum_{a \in \Sigma} s(S[i], a) p_j(a) = \text{score}_j(S[i])$$

2. trou de longueur i dans P — insertion en colonne j : $-\alpha_j - \beta_j(i - 1)$

3. trou de longueur i dans S — suppression de colonnes $j, \dots, j + i - 1$:
 $-\gamma_j - \delta_{j+1} - \delta_{j+2} - \dots - \delta_{j+i-1}$

Alignement à un profile 2

PD : $G(i, j)$ score de l'alignement optimal entre $S[1..i]$ et les premières j colonnes de P .

Réurrences avec E (trou en S), F (trou en P), et G (match) :

$$E(i, j) = \max \left\{ G(i, j-1) - \gamma_j, E(i, j-1) - \delta_j \right\}$$

$$F(i, j) = \max \left\{ G(i-1, j) - \alpha_j, F(i-1, j) - \beta_j \right\}$$

$$G(i, j) = \max \left\{ E(i, j), F(i, j), G(i-1, j-1) + \text{score}_j(S[i]) \right\}$$

Alignement à un profile 3

Initialisation :

$G(i, 0) = E(i, 0) = 0$ (alignement peut commencer dans le milieu de S),
 $G(0, j) = F(0, j) = -\gamma_1 - \sum_{i=2}^j \delta_i$ (il commence au début du profile).

Score de l'alignement optimal : $\max \left\{ G(i, n) : i = 1, \dots, |S| \right\}$

Temps de calcul : $O(|S|n)$ (après le calcul de score_j)

Profile de PROSITE

exemple : Profile de homeobox

spécifie entre autres la pondération score_j , α_j , β_j , γ_j , et δ_j

```
ID    HOMEBOX_2; MATRIX.  
...  
          A    B    C    D    E    ...  
...  
/M: SY='E'; M= -5,  2, -25,  3, 11, ...  
/M: SY='Q'; M= -3, -4, -25, -4, 12, ...  
...  
/I:          I=-8; MI=-8; IM=-8; DM=-15; MD=-15;
```

Modèle de Markov caché

ensemble d'états $\mathcal{Q}$ de taille N

probabilités de transition $\tau: \mathcal{Q} \times \mathcal{Q} \mapsto [0, 1]$ où $\sum_{q' \in \mathcal{Q}} \tau(q, q') = 1$ pour tout $q \in \mathcal{Q}$

probabilités d'émission $p: \mathcal{Q} \times \Sigma \mapsto [0, 1]$ où $\sum_{c \in \Sigma} p(q, c) = 1$ pour tout $q \in \mathcal{Q}$.

probabilités d'état initial $\pi: \mathcal{Q} \mapsto [0, 1]$ où $\sum_{q \in \mathcal{Q}} \pi(q) = 1$.

Génération d'une séquence $s = s_1 \cdots s_\ell$ au hasard :

- 1 Choisir l'état initial q_1 au hasard par les probabilités π .
- 2 **pour** $i \leftarrow 1, \dots, \ell$
- 3 Choisir s_i au hasard par les probabilités $p(q_i, \cdot)$: émettre s_i .
- 4 **si** $i < \ell$ **alors** Choisir q_{i+1} au hasard par $\tau(q, \cdot)$.
- 5 **fin**

Trois problèmes

Problème 1. Si on observe la séquence s , comment est-ce qu'on peut calculer $\mathbb{P}\left\{s \mid \mathcal{M}\right\}$, la probabilité que $\mathcal{M}$ engendre s ?

Problème 2. Si on observe la séquence s , comment est-ce qu'on peut trouver la séquence d'états $q_1 \cdots q_\ell$ qui y correspond le mieux ?

Problème 3. Comment est-ce qu'on choisit les paramètres τ , p , π (apprentissage)

Problème 1

vraisemblance : probabilité que $\mathcal{M}$ émet $s = s_1 \cdots s_\ell$

$$L_{\mathcal{M}}(s) = \mathbb{P}\left\{s \mid \mathcal{M}\right\} = \sum_{q_1, \dots, q_\ell} \pi(q_1) p(q_1, s_1) \tau(q_1, q_2) \cdots \tau(q_{\ell-1}, q_\ell) p(q_\ell, s_\ell).$$

→ sommation sur tous les chemins : $O(N^\ell)$ termes dans la formule — trop

Programmation Dynamique !

Soit $\alpha_i(q) = \mathbb{P}\{s_1 \cdots s_i, q_i = q\}$ (production du préfixe en arrivant à l'état q).

Alors $L(s) = \sum_{q \in \mathcal{Q}} \alpha_\ell(q)$.

Calcul de $\alpha_i(q)$:

Initialisation : $\alpha_1(q) = \pi(q) p(q, s_1)$.

Récurrence : $\alpha_{i+1}(q) = \left(\sum_{q'} \alpha_i(q') \tau(q', q) \right) p(q, s_{i+1})$.

Problème 1 — cont.

On peut calculer les α_i d'une façon très efficace :
en temps $O(N^2\ell)$, avec $O(N)$ espace

Une autre possibilité

Soit $\beta_i(q) = \mathbb{P}\{s_{i+1} \cdots s_\ell, q_i = q\}$ (production du suffixe à partir de l'état q).

Alors $L(s) = \sum_{q \in \mathcal{Q}} \beta_1(q) \pi(q)$.

Initialisation : $\beta_\ell(q) = 1$.

Récurrence : $\beta_{i-1}(q) = \sum_{q'} \tau(q, q') p(q', s_i) \beta_i(q')$.

De nouveau : temps $O(N^2\ell)$, espace $O(N)$.

Problème 2

Séquence d'états la plus probable — qu'est-ce que ça veut dire ?

L'état le plus probable pour le i -ème caractère :

soit $\gamma_i(q) = \mathbb{P}\{q_i = q\}$ (que le i -ème état est q).

$$\gamma_i(q) = \frac{\alpha_i(q)\beta_i(q)}{\sum_{q' \in \mathcal{Q}} \alpha_i(q')\beta_i(q')}.$$

Alors $q_i^* = \arg \max \gamma_i(q)$ est l'état le plus probable pour le i -ème caractère.

Le chemin le plus probable — algorithme de Viterbi

Algo de Viterbi

Soit $\delta_i(q) = \max_{q_1, \dots, q_{i-1}} \mathbb{P}\{q_1 \cdots q_{i-1} q_i = q, s_1 \cdots s_i\}$ (meilleur chemin pour le préfixe).

Initialisation : $\delta_1(q) = \pi(q)p(q, s_1)$.

Récurrence : $\delta_{i+1}(q) = \left(\max_{q'} \delta_i(q') \tau(q', q) \right) p(q, s_{i+1})$

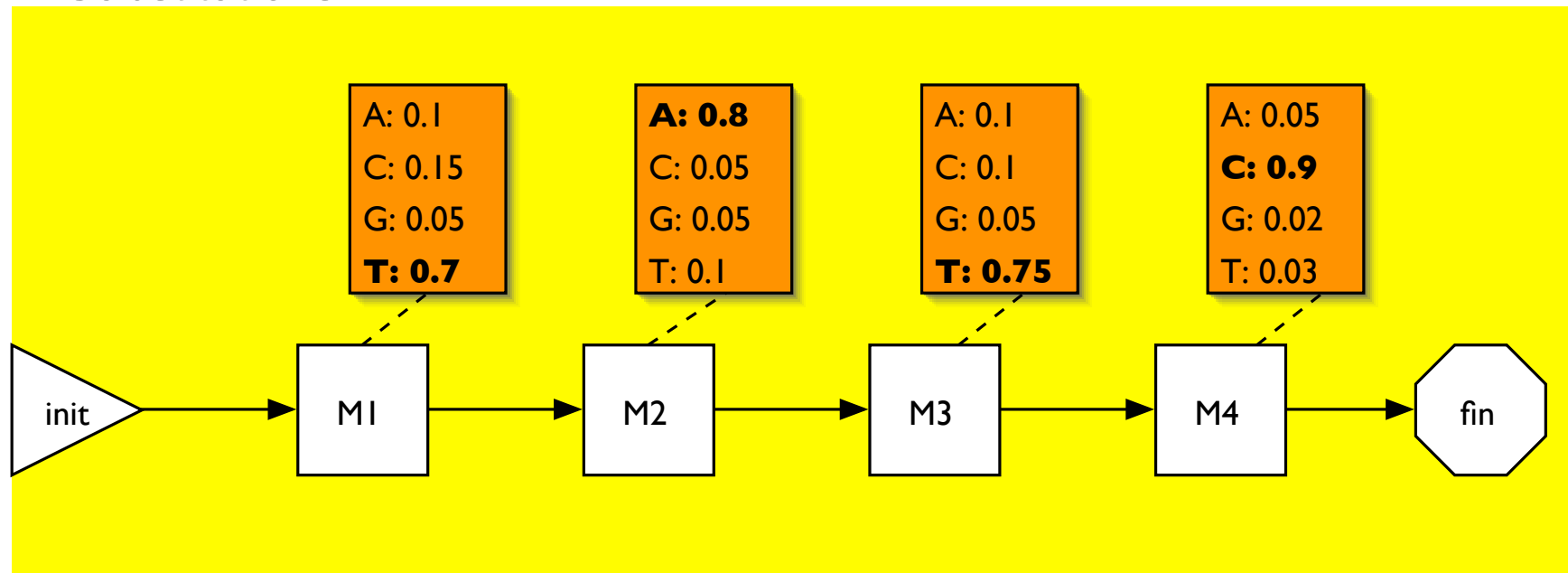
PD de nouveau

HMM pour génération d'une séquence

Séquence d'ADN ou séquence protéique

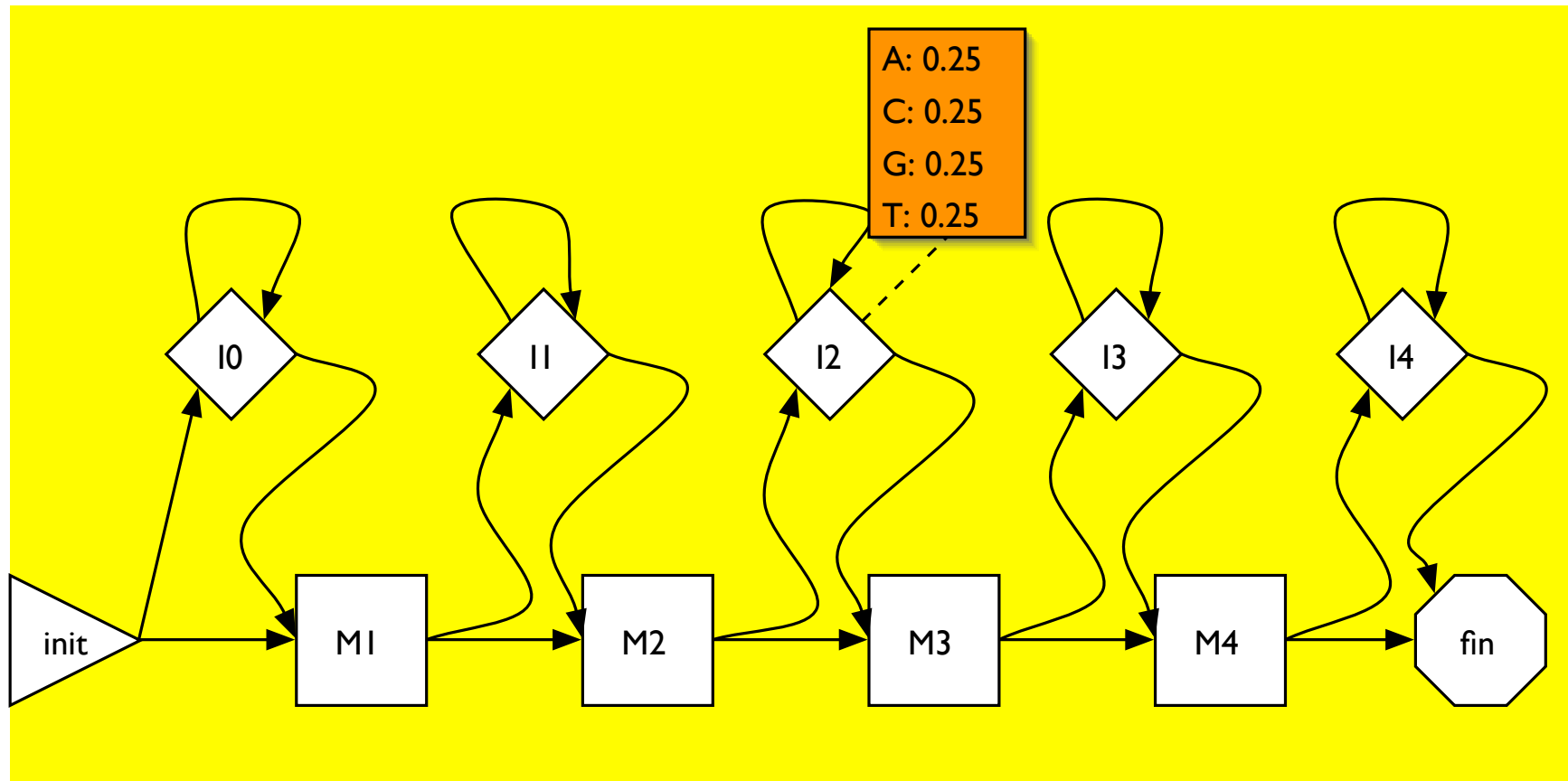
Construisons un modèle pour alignement

1. Substitutions



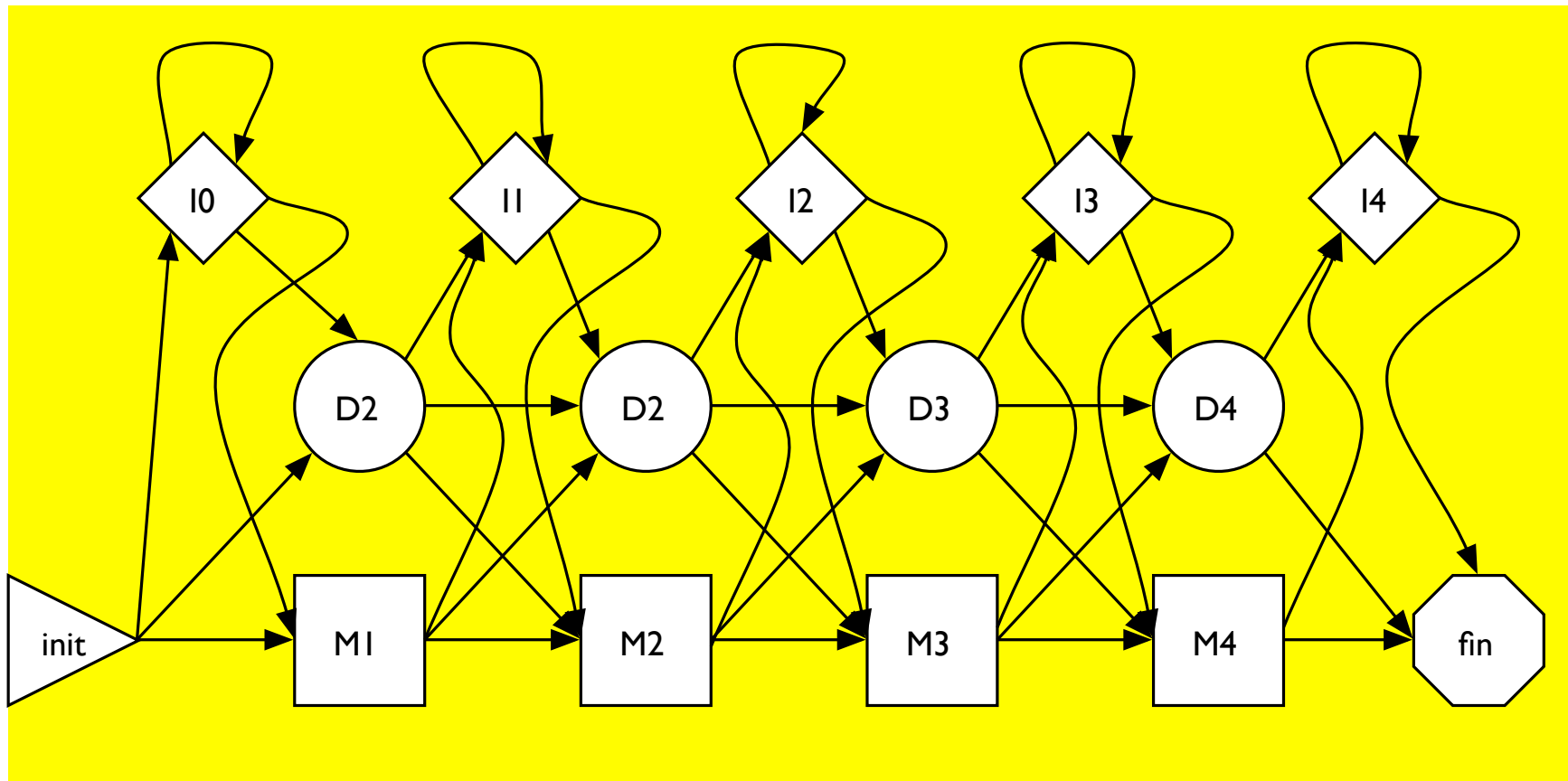
HMM de profile 2

2. Insertion de caractères



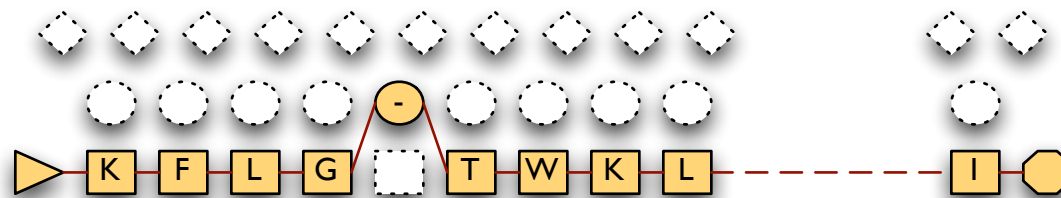
HMM de profile 3

3. Suppression de caractères



HMM de profile

Alignement à un HMM de profile : calculer le chemin Viterbi



Q9DAK4 /4-132	QLQG.TWKSVS	CDNF.ENYMKELGVGRASRK.LGCLAK.....PTVT
Q8QHA8 /4-132	HFVG.TWKLLS	SENF.EDYMKELGVGFATRК.MAGVAK.....PNLT
FABA_RAT /3-131	AFVG.TWKLVSS	ENF.DDYMKEVGVGFATRК.VAGMAK.....PNLI
MYP2_HUMAN /3-131	KFLG.TWKLVSS	ENF.DDYMKAIGVGLATRК.LGNLAK.....PTVI
TLBP_MOUSE /4-132	PFLG.TWKLVSS	ENF.ENYVRELGVCEPRK.VACLIK.....PSVS
O57663 /4-133	KFVG.TWKMISS	DNF.DDYMKAIGVGFATRQ.VGNRTK.....PNLV
FABE_RAT /6-134	DLEG.KWRLVES	HGF.EDYMKELGVGLALRK.MGAMAK.....PDCI
FABE_BOVIN /6-134	QLVG.RWRLVES	KGF.DEYMKEVGVGMALRK.VGAMAK.....PDCI

Viterbi : HMM de profile

Quel bonheur : un petit nombre de transitions, et pas de boucles !

Difficulté : états sans émission

États de match : $M_j : j = 1, \dots, n$

États d'insertion : $I_j : j = 0, \dots, n$

États de suppression : $D_j : j = 1, \dots, n$

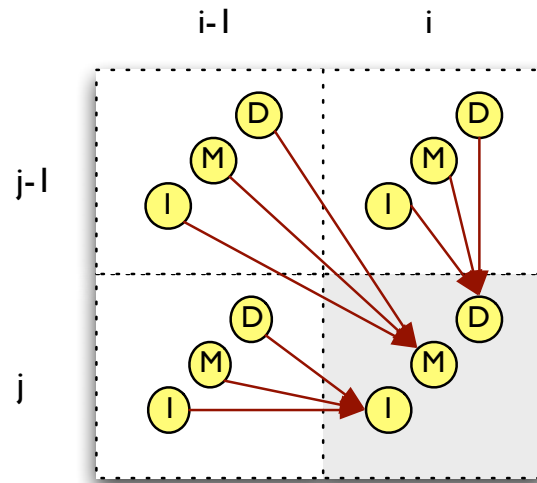
Sous-problèmes pour PD : $\delta_j^{(M)}(i), \delta_j^{(I)}(i), \delta_j^{(D)}(i)$ pour émission de $s_1 \cdots s_i$ en finissant dans un état M_j, D_j, I_j .

$$\delta_j^{(M)}(i) = p(M_j, s_i) \max \left\{ \delta_{j-1}^{(M)}(i-1) \tau(M_{j-1}, M_j), \right. \\ \delta_{j-1}^{(I)}(i-1) \tau(I_{j-1}, M_j), \\ \left. \delta_{j-1}^{(D)}(i-1) \tau(D_{j-1}, M_j) \right\}$$

$$\delta_j^{(I)}(i) = p(I_j, s_i) \max \left\{ \delta_j^{(M)}(i-1) \tau(M_j, I_j), \right. \\ \delta_j^{(I)}(i-1) \tau(I_j, I_j), \\ \left. \delta_j^{(D)}(i-1) \tau(D_j, I_j) \right\}$$

$$\delta_j^{(D)}(i) = \max \left\{ \delta_{j-1}^{(M)}(i) \tau(M_{j-1}, M_j), \right. \\ \delta_{j-1}^{(I)}(i) \tau(I_{j-1}, M_j), \\ \left. \delta_{j-1}^{(D)}(i) \tau(D_{j-1}, M_j) \right\}$$

Viterbi : HMM de profile 3



Cas de base (ali. global) : $\delta_0^{(I)}(0) = \tau(\text{start}, I_0)$; $\delta_1^{(M)}(0) = \tau(\text{start}, M_1)$;
 $\delta_1^{(D)}(0) = \tau(\text{start}, D_1)$.

Fin : $\delta^{(\text{fin})}(\ell) = \max \left\{ \delta_n^{(M)}(\ell) \tau(M_n, \text{fin}), \delta_n^{(I)}(\ell) \tau(I_n, \text{fin}), \delta_n^{(D)}(\ell) \tau(d_n, \text{fin}) \right\}$.

Score de l'alignement à HMM

Logarithme de la probabilité pour générer une séquence $s_1 \cdots s_\ell$ sur un chemin $q_1 \cdots q_m$: c'est la somme des termes comme $\log(\tau(q_j, q_{j+1}))$ et $\log(p(q_j, s_i))$.

Contributions à $\log \delta_j(i)$:

$M_{j-1} \rightarrow M_j : \log \tau(M_{j-1}, M_j) + \log p(M_j, s_i)$ (substitution)

$M_j \rightarrow I_j : \log \tau(M_j, I_j) + \log p(I_j, s_i)$ (début d'un trou)

$I_j \rightarrow I_j : \log \tau(I_j, I_j) + \log p(I_j, s_i)$ (extension d'un trou)

etc.

$\log \tau$ et $\log p$ donnent les pondérations pour substitutions ainsi que trou début, extension, et fin : tous les scores dépendent de la colonne j (comparez aux récurrences d'un alignement à un profile)

Score de l'alignement

Les probabilités définissent une pondération de l'alignement : alignement de score maximal (par log des probs) correspond au chemin Viterbi.

Avantage additionnel : évite l'underflow — probabilité pour un chemin long ne peut pas être représentée par une variable flottante p.e., $\epsilon_M = 10^{-324}$.

LODS

Hypothèse null : $s_1 \dots s_\ell$ est une séquence de caractères iid $\mathbb{P}\{s_i = \sigma\} = r_\sigma$ (souvent les mêmes probs d'émission aux états d'insertion).

Hypothèse alternatif : $s_1 \dots s_\ell$ est générée par notre HMM de profile $\mathcal{M}$.

LODS (logarithme des chances) : comparer $p_0 = \mathbb{P}\{s_1 \dots s_\ell \mid H_0\}$ et $p_{\mathcal{M}} = \mathbb{P}\{s_1 \dots s_\ell \mid H_1\}$:

$$\text{LODS} = \log \frac{p_{\mathcal{M}}}{p_0}$$

Calcul de LODS : par Viterbi, remplacer $p(q_j, s_i)$ par $\frac{p(q_j, s_i)}{r_{s_i}}$.

Pour des valeurs entières : $v = \left\lfloor \alpha \log_2 \frac{p}{r} \right\rfloor$ (r est la probabilité par H_0)

Construction d'un HMM de profile

- à partir de séquences : apprentissage
- à partir d'alignements (structuraux - très fiables) : utilisez les chemins [«seed alignments»] pour les paramètres

BD de HMMs de profile pour familles de protéines : Pfam

Apprentissage de paramètres

[Une approche de pseudo-compteurs.]

Si on a un alignement multiple, on identifie un chemin pour chaque séquence alignée.

- N nombre de séquences
- N_q nombre de chemins qui passent par q
- $N_{q \rightarrow q'}$ nombre de chemins avec une transition $q \rightarrow q'$
- $N_{q,\sigma}$ nombre de fois caractère σ est vu en état de match q
- x, y «pseudo-compteurs» (Laplace : $x = y = 1$)
- a_q nombre d'états q' avec $\tau(q, q') > 0$

$$\tau(q, q') = \frac{N_{q \rightarrow q'} + x}{N + a_q x} \quad p(q, \sigma) = \frac{N_{q,\sigma} + y}{N_q + y|\Sigma|}$$

Émission dans un état d'insertion : selon les fréquences de «background»

Problème 3 — apprentissage

Sans alignements initiaux — c'est difficile

Échantillon de séquences : séquences à aligner

- 1 Initialiser τ , p and π .
- 2 **répéter**
- 3 calculer le chemin Viterbi pour les séquences de l'échantillon ;
- 4 recalculer τ , p , et π par les chemins Viterbi ;
- 5 **jusqu'à** l'optimum local est achevé.

+ détails de l'initialisation, méthodes numériques, etc.

Baum-Welch : calculer tous les chemins pour toutes les séquences, les pondérer par leurs probabilités pour recalculation de τ , p , and π .

Classification

Pour une séquence s :

prendre profiles $\mathcal{M}_1, \dots, \mathcal{M}_k$ et calculer les

vraisemblances $L_{\mathcal{M}_1}(s), \dots, L_{\mathcal{M}_k}(s)$:

la meilleure* vraisemblance donne la classification de s .

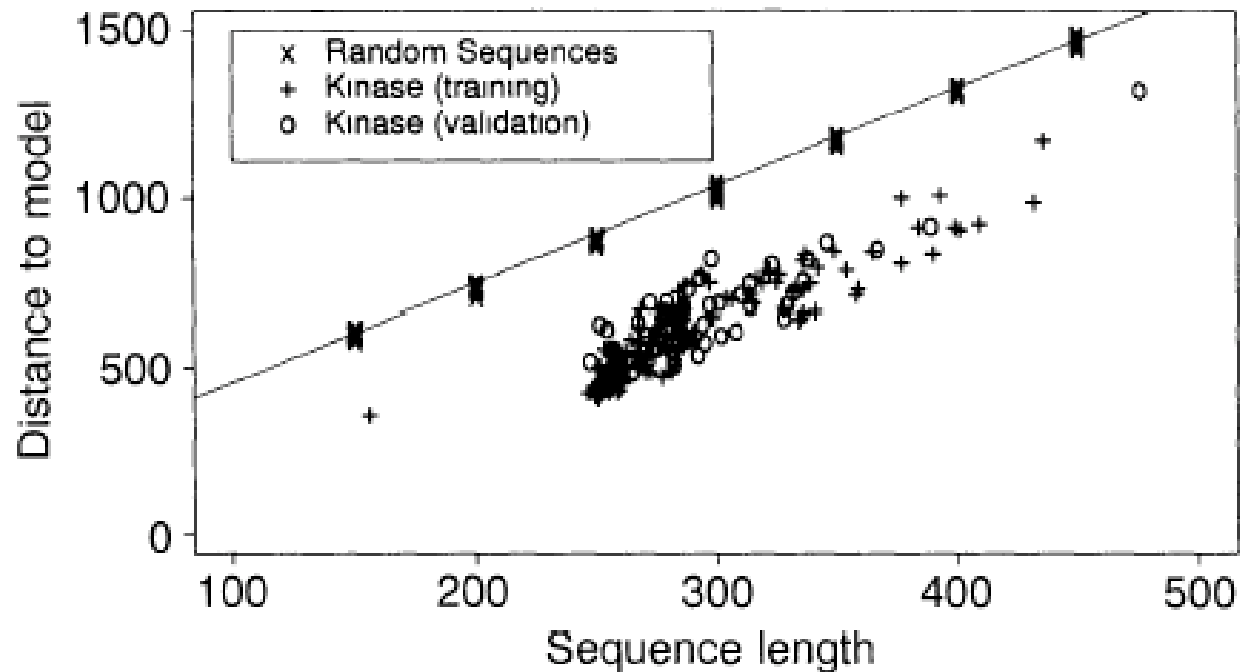
* n'est pas aussi simple que ça

1. la vraisemblance dépend de la longueur - il faut normaliser

2. comparer les L des membres de la famille à celle d'autres protéines

pour avoir un seuil L_{famille} : les membres de la famille ont $L > L_{\text{famille}}$,
autres ont $L < L_{\text{famille}}$.

Classification : exemple

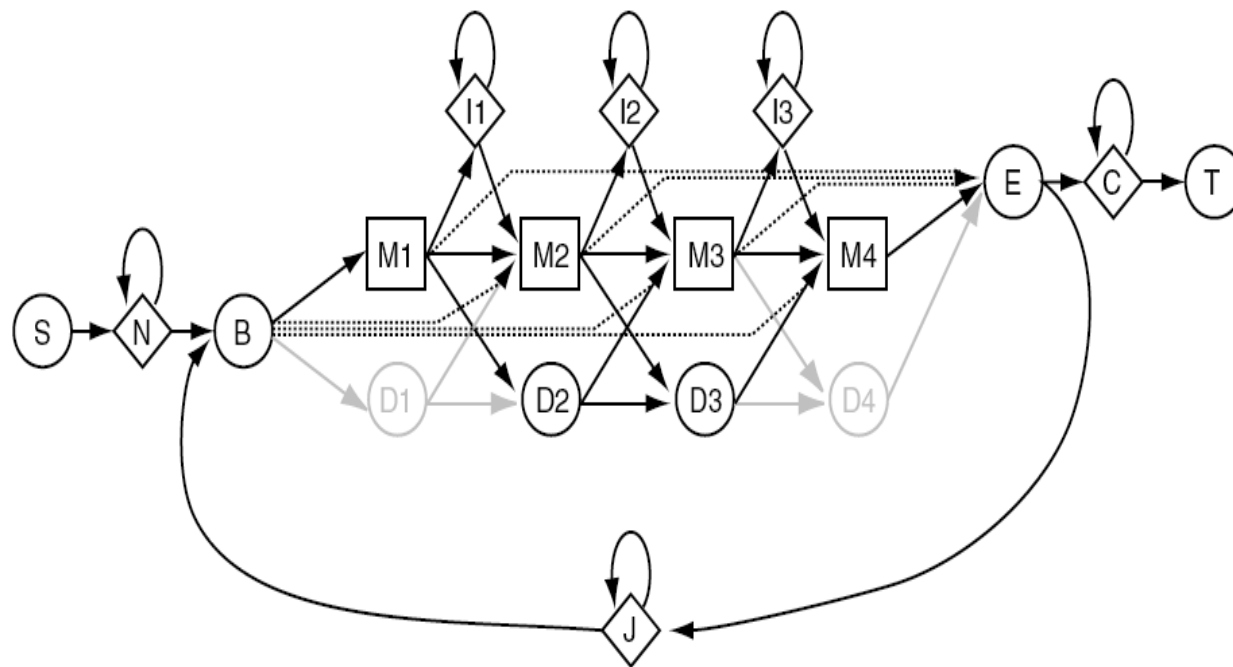


(Baldi et al. PNAS 91:1059)

(Y : $-\log$ vraisemblance, X : longueur de la séquence)

HMMs - extensions

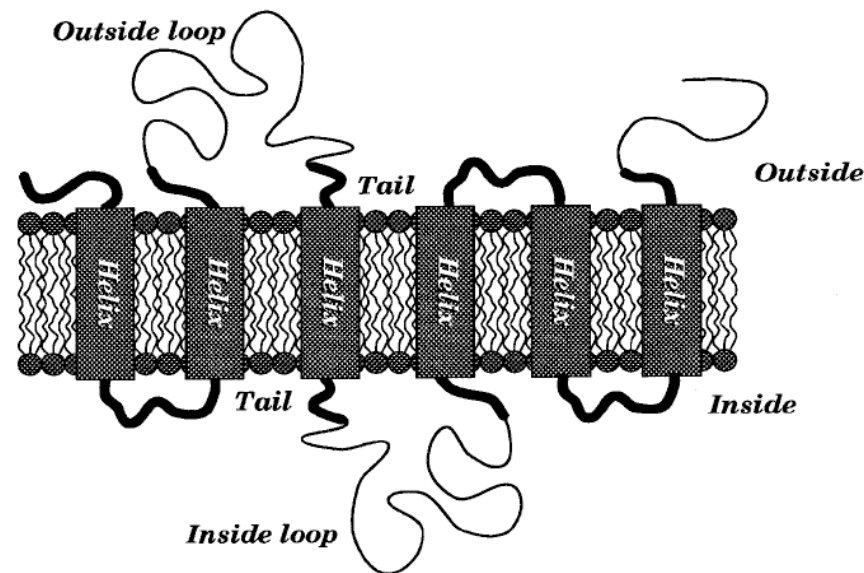
alignement local + plusieurs occurrences dans une séquence



(Eddy: HMMer manual v2.3.1)

HMMs - architecture spécialisé

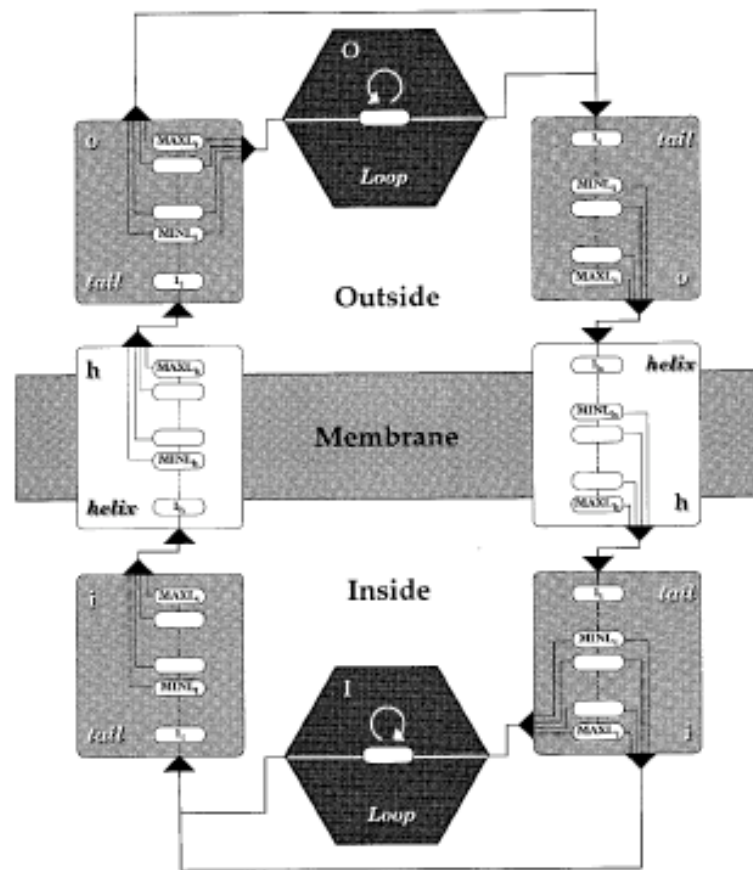
Exemple : protéines membranaires



Amino acid seq: MGDVCDTEFGILVA...SVALRPRKHGRWIV...FWVDNGTEQ...PEHMTKLHMM...

State seq: oooooooooohhhhh...hhhhiiiiiiihhh...hhhooooOO...OOoooooooohh...

(Tusnády & Simon, J.Mol.Biol. 283:489)



(Tusndy & Simon, J.Mol.Biol. 283:489)