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 ?

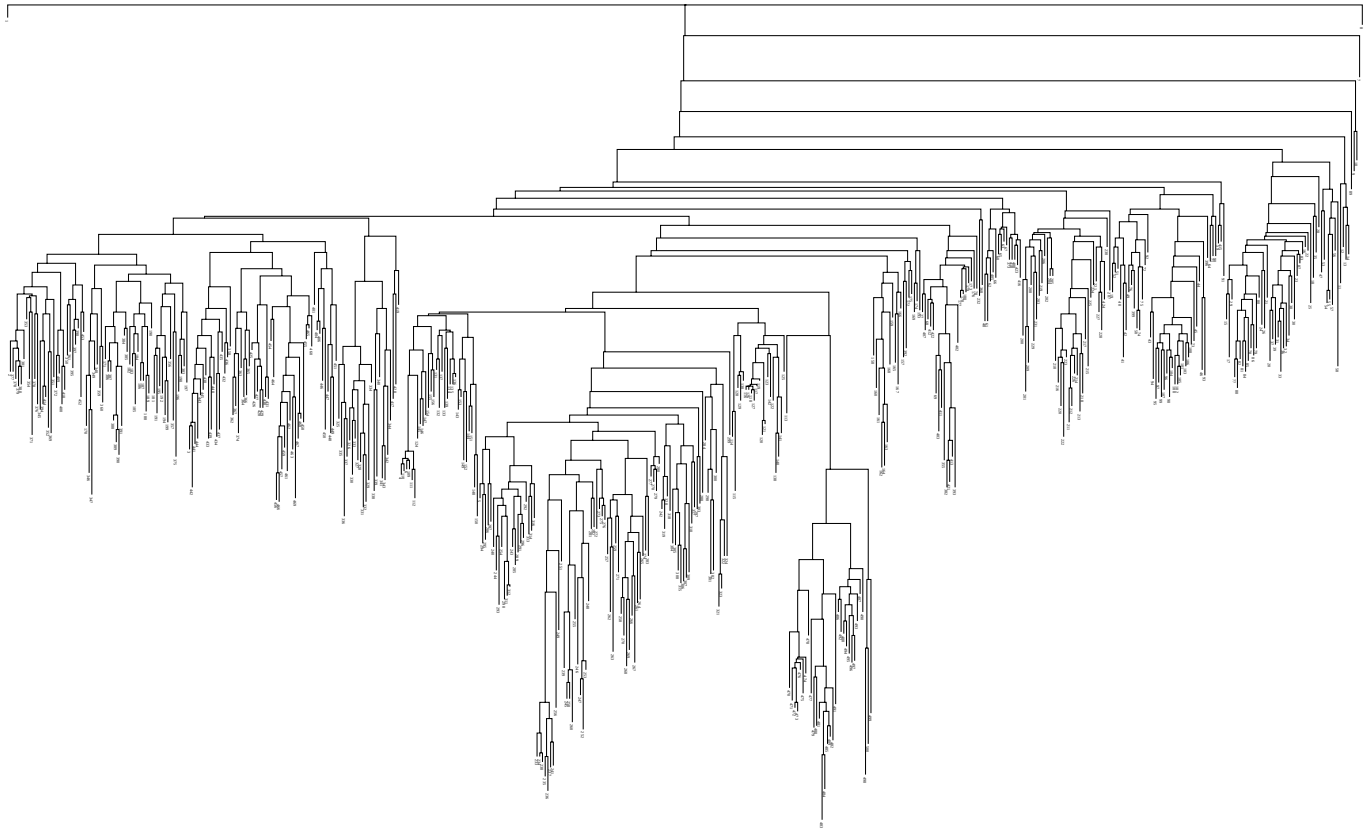
k séquences (reliées) : quels sont les traits communs ?

Applications

– représentation de familles de protéines
– identification et représentation de caractéristiques préservés et leur corrélation 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 ...

Approche 1 : Spécifier a valeur de toutes les colonnes possibles : $\left(\Sigma \cup \{-\}\right)^k$ si on a k séquences.

On peut simplifier un peu : match/mismatch ou indel seulement : $2^k - 1$ possibilités.

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

Scores 2

Approche 2 : fonction SP «sum of pairs»

trouver la solution exacte : NP-difficile

Approche 3 : séquence de consensus

Approche 4 : alignement à un arbre phylogénétique (viendra plus tard)

Représentation de familles

1. signatures («patterns»)

2. profile

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 altérnatifs 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`
déscription de cca. 10^{27} séquences possibles

Automate fini

automate fini déterministe $\mathcal{M} = \langle \Sigma, \mathcal{Q}, t \rangle$

– modèle de séquences sur un alphabet Σ
– ensemble d'états $\mathcal{Q}$; état initial $q_{\text{init}} \in \mathcal{Q}$ et état accepteur $q_{\text{acc}} \in \mathcal{Q}$.
– transitions t : $\mathcal{Q} \times \Sigma \mapsto \mathcal{Q}$

AF non-déterministe
transitions t : $\mathcal{Q} \times (\Sigma \cup \{\varepsilon\}) \mapsto 2^{\mathcal{Q}}$

génération d'une séquence $s = s_1 s_2 \cdots s_\ell$: (version AFD)
série d'états $q_{\text{init}} = q_0, q_1, \ldots, q_\ell = q_{\text{acc}}$ avec $t(q_{i-1}, s_i) = q_i$
pour chaque $i = 1, \ldots, \ell$

acceptance d'une séquence s : s'il existe une série d'états qui génère s

Automate fini 2

Problème : on a machine $\mathcal{M}$ et séquence $s = s_1 \cdots s_\ell$, décider si $\mathcal{M}$ accepte s.

Solution : facile si AF déterministe — suivre les transitions et vérifier si on arrive à l'état q_{acc}

```
1  soit q ← q_init
2  pour i ← 1 . . . ℓ
3        soit q ← t(q, s_i)
4  fin
5  si q = q_acc alors accepter else rejeter.
```

Temps de calcul : $O(\ell)$
Espace : entrée+ constante (stocker i et q seulement)

Expression régulière

E est une **expression régulière (ER)**
sur alphabet Σ ssi

1. $E = \emptyset$,
2. $E = a$ avec $a \in \Sigma_\varepsilon$,
3. $E = (E_1) \cup (E_2)$ avec E_1 et E_2 des ER,
4. $E = (E_1) \cdot (E_2)$ avec E_1 et E_2 des ER,
5. $E = (E_1)^*$ avec E_1 une ER.

(définition inductive)

Syntaxes altérnatifs : $|$ au lieu de $\cup$, omission de $\cdot$ et les parenthèses
règles de précédence : $* \cdot |$,
donc $b|aa^* = (b) \cup ((a) \cdot ((a)^*))$

Expression régulière 2

$\rightarrow$ les signatures de PROSITE sont des ER (avec un syntaxe différent, et sans l'étoile)

Exemple : $Cxx(x|\epsilon)(x|\epsilon)Cxxx(L|I|\cdots|C)xxxxxxxxHxxx(x|\epsilon)(x|\epsilon)H$

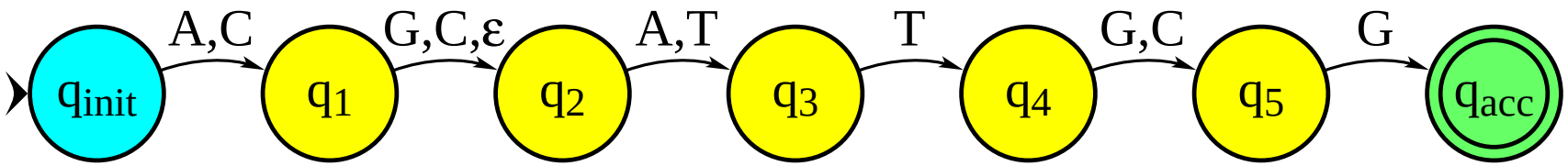
$\rightarrow$ étoile pour mutations avec triplets répétés
FMR1 (gène de syndrome de l'X fragile) `GCG(AGG|CGG)*CTG` ($>$ 200 : malade)
huntingtin (gène de maladie de Huntington `TTC(CAG)*CCG` ($>$ 45 : malade)

Thm. *Chaque ER corresponde à un AF et vice versa.* (v. IFT2102)

Automates finis et alignements

Un alignement multiple peut être representé par un AF :

```
CGATCG
C-ATGG
ACTTCG
```



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

Profile

$\longrightarrow$ 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 $j = 1, \ldots, n$

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 $j, \ldots, j+i-1$: $-\gamma_j - \delta_{j+1} - \delta_{j+2} - \cdots - \delta_{j+i-1}$

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.

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

$$E(i,j) = \max\Big\{G(i,j-1) - \gamma_j, E(i,j-1) - \delta_j\Big\}$$
$$F(i,j) = \max\Big\{G(i-1,j) - \alpha_j, F(i-1,j) - \beta_j\Big\}$$
$$G(i,j) = \max\Big\{E(i,j), F(i,j), G(i-1,j-1) + \mathsf{score}_j(S[i])\Big\}$$

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 : $\max\left\{G(i,n) : i = 1,\ldots,|S|\right\}$

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   HOMEOBOX_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;
```

Interprétation probabiliste d'un alignement

Alignement comme une série de mutations

- probabilité d'identité p
- probabilité de substitution q
- probabilité de supression/insertion : r

```
ACT-AG
AG-GAG
```

prob $= p^3 q r^2$,

on veut trouver l'alignement le plus probable — maximisation de prob

Interprétation probabiliste 2

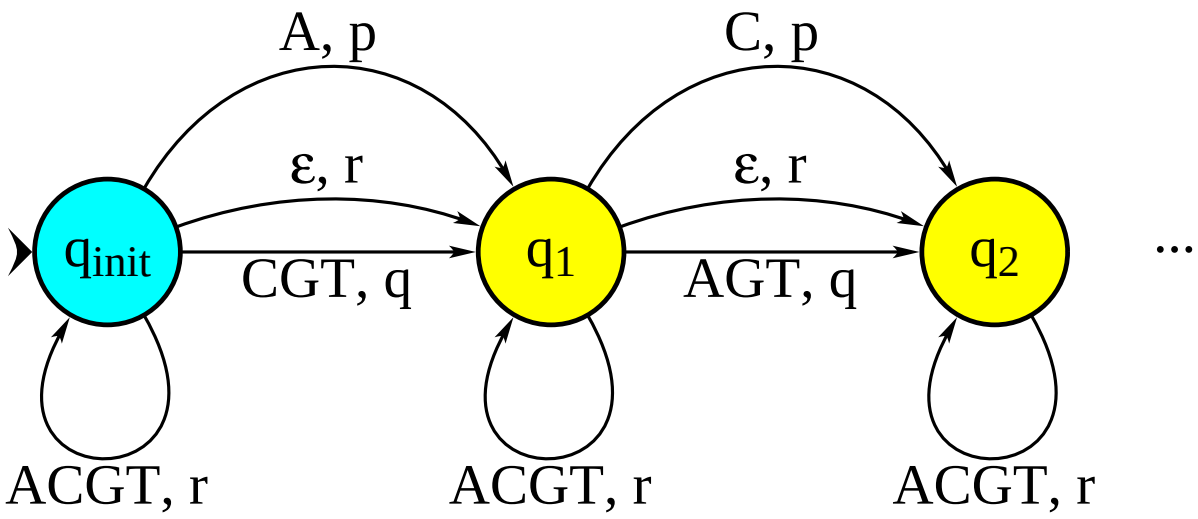
vraisemblance $\log_2(\text{prob}) = 3\log_2 p + \log_2 q + 2\log_2 r$
soit $s = p/2$, on a
$\log_2(\text{prob}) = 3 - \log_2 \frac{s}{q} - 2\log_2 \frac{\sqrt{s}}{r} + 5\log_2 s$

Avec $\mu = \log_2 \frac{s}{q}$ et $\delta = -\log_2 \frac{\sqrt{s}}{r}$, ça correspond à un système de scores pour un alignement optimal

Alignement généré par un automate probabiliste

ACT-AG

AG-GAG



Automate probabiliste

probabilité de l'alignement : produit des probabilités sur le chemin (ou la somme de ces produits sur tous les chemins)

Problème : calculer cette probabilité

Modèle de Markov caché

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

probabilités de transition τ : $\mathcal{Q} \times \mathcal{Q} \mapsto [0, 1]$ 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 π : $\mathcal{Q} \mapsto [0, 1]$ 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 inital q_1 au hasard par les probabilités π.
2 **pour** $i \leftarrow 1, \ldots, \ell$
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 $\mathbb{P}\left\{ s \;\middle|\; \mathcal{M} \right\}$, 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 $s = s_1 \cdots s_\ell$

$$L_{\mathcal{M}}(s) = \mathbb{P}\Big\{ s \;\Big|\; \mathcal{M} \Big\} = \sum_{q_1,\ldots,q_\ell} \pi(q_1) p(q_1, s_1) \tau(q_1, q_2) \cdots \tau(q_{\ell-1}, q_\ell) p(q_\ell, s_\ell).$$

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

Programmation Dynamique !

Soit $\alpha_i(q) = \mathbb{P}\Big\{ s_1 \cdots s_i, q_i = q \Big\}$ (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 : $\alpha_{i+1}(q) = \Big(\sum_{q'} \alpha_i(q') \tau(q', q) \Big) p(q, s_{i+1})$.

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 $\beta_i(q) = \mathbb{P}\Big\{s_{i+1} \cdots s_\ell, q_i = q\Big\}$ (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).

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

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

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

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

PD de nouveau

Problème 3 — apprentissage

C'est le plus 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.

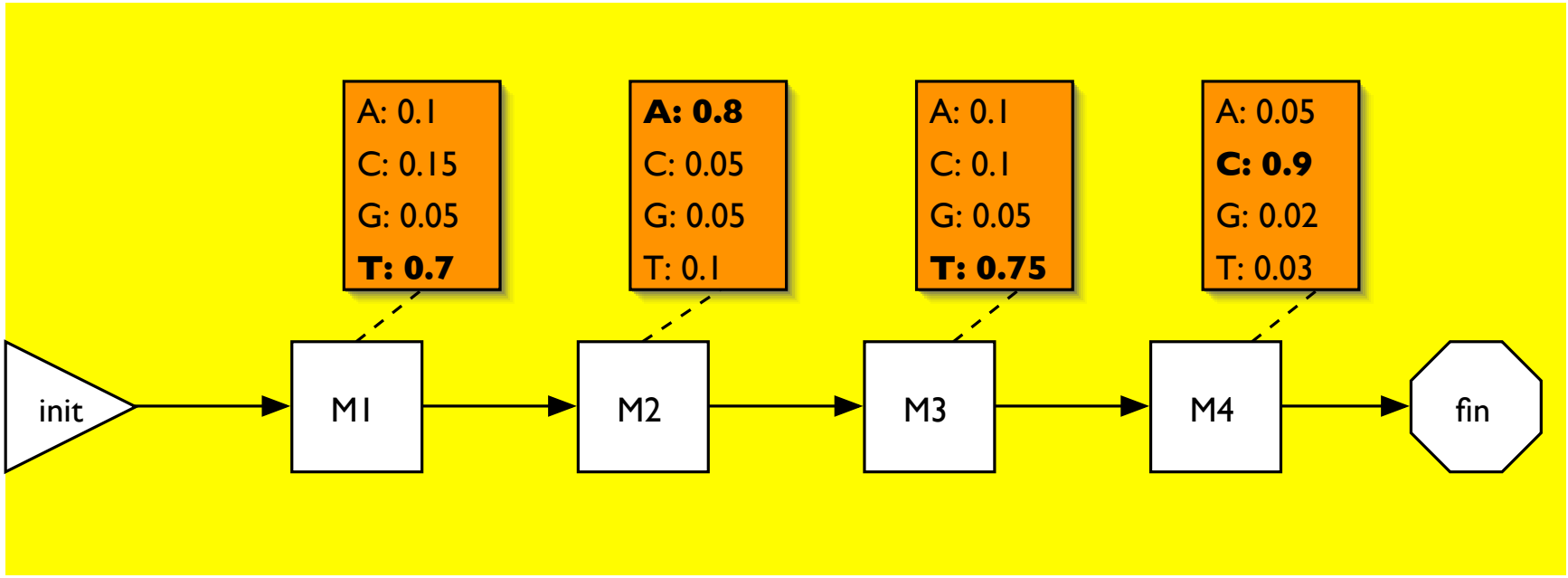
Baum-Welch : calculer tous les chemins pour toutes les séquences, les pondérer par leurs probabilités pour recalculation de τ, p, and π.

HMM pour génération d'une séquence

Séquence d'ADN ou séquence protéique

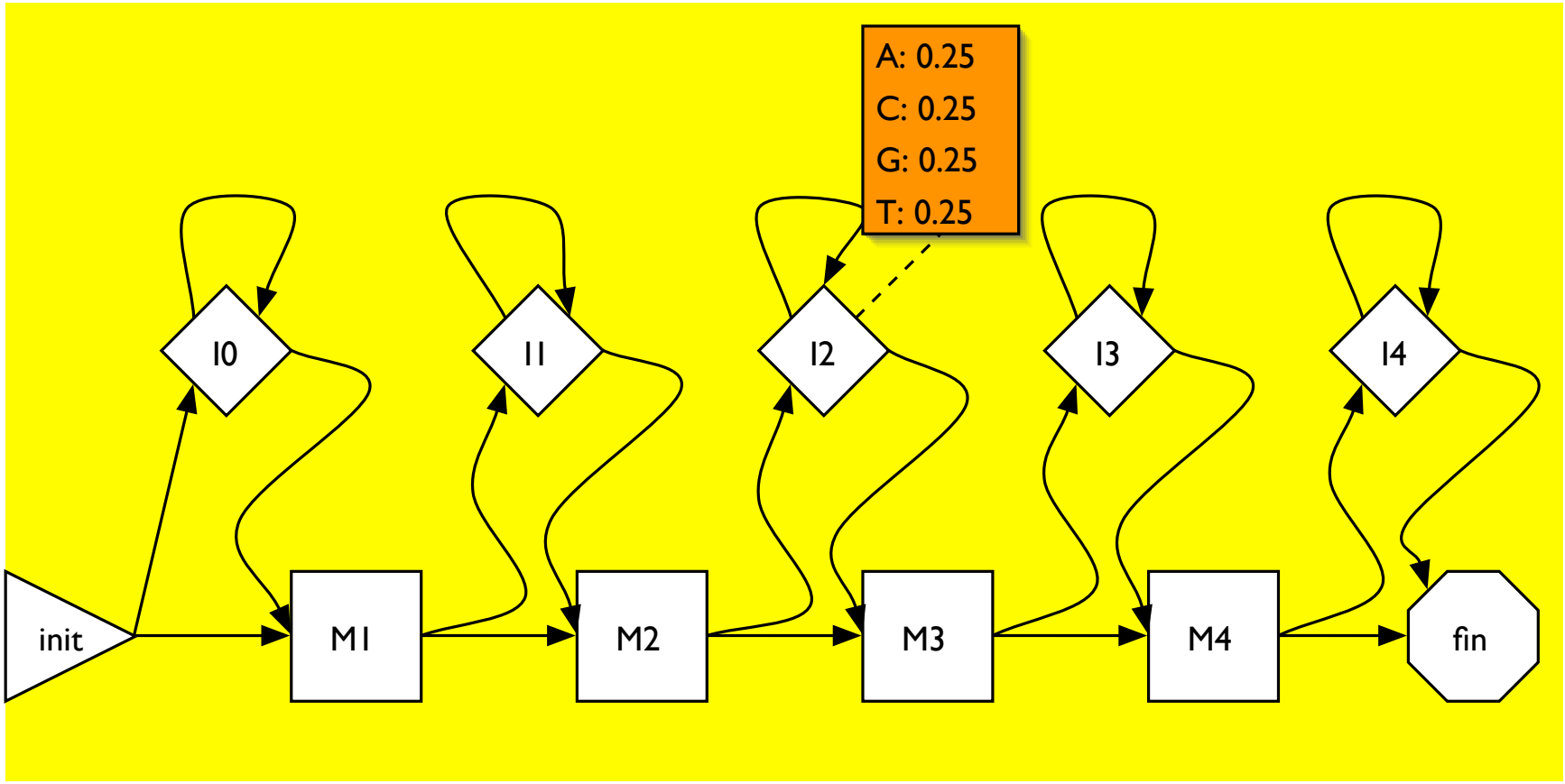
Construisons un modèle pour alignement [interprétation probabiliste 2, évolution].

1. Substitutions



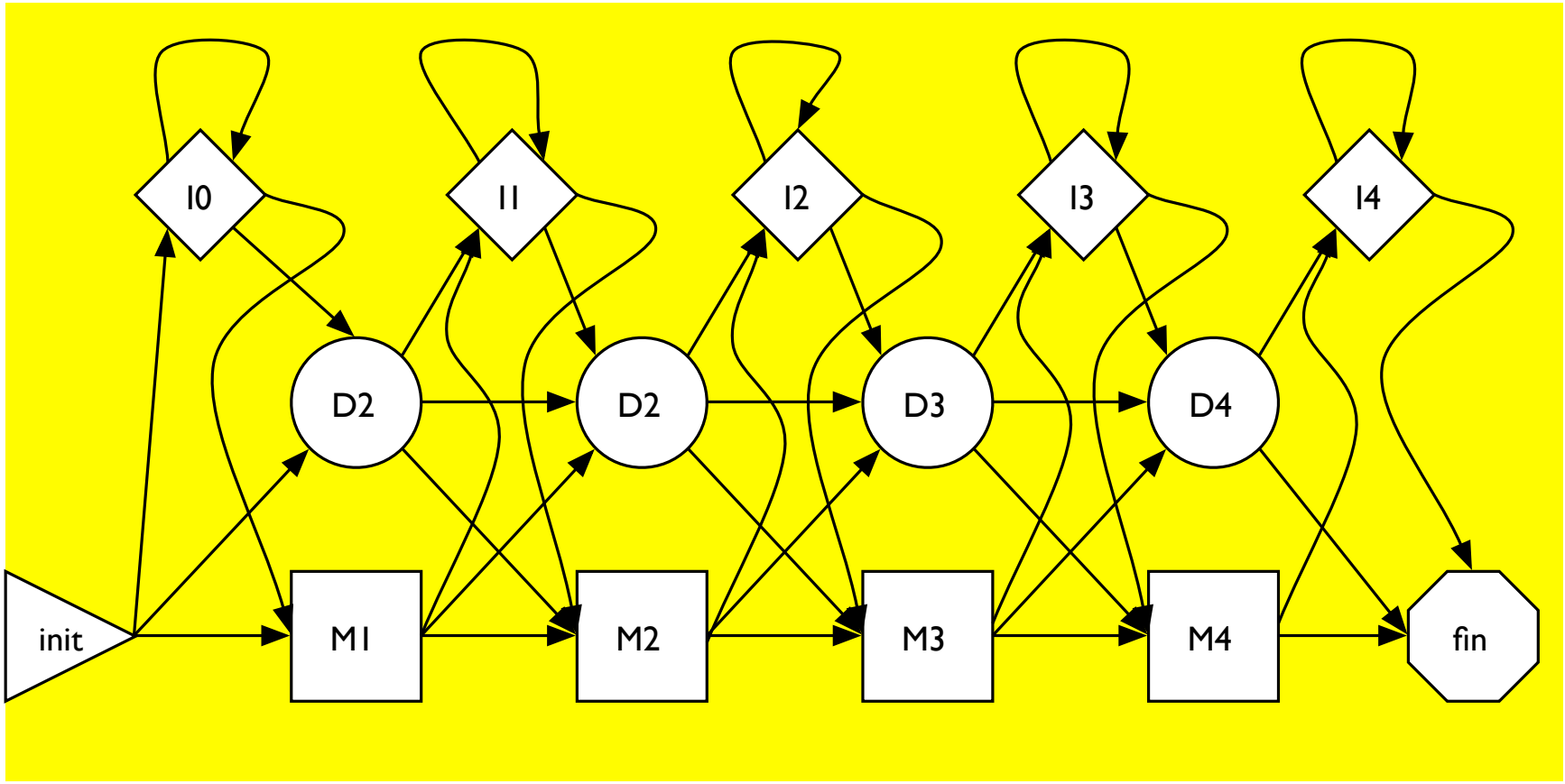
HMM de profile 2

2. Insertion de caractères



HMM de profile 3

3. Suppression de caractères



HMM de profile

Alignement à un HMM de profile : chemins Viterbi

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

Récurrences — attention aux états de suppression : pas d'émission

Construction d’un HMM de profile

- à partir de séquences : apprentissage [Baum-Welch ou d’autres méthodes]

- 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

HMM - philosphie

- log-vraisemblance pour un chemin : pondération d'un alignement, avec trous
- caractères émis dans un état d'insertion ne sont pas alignés
- pondération d'un alignement n'est pas arbitraire

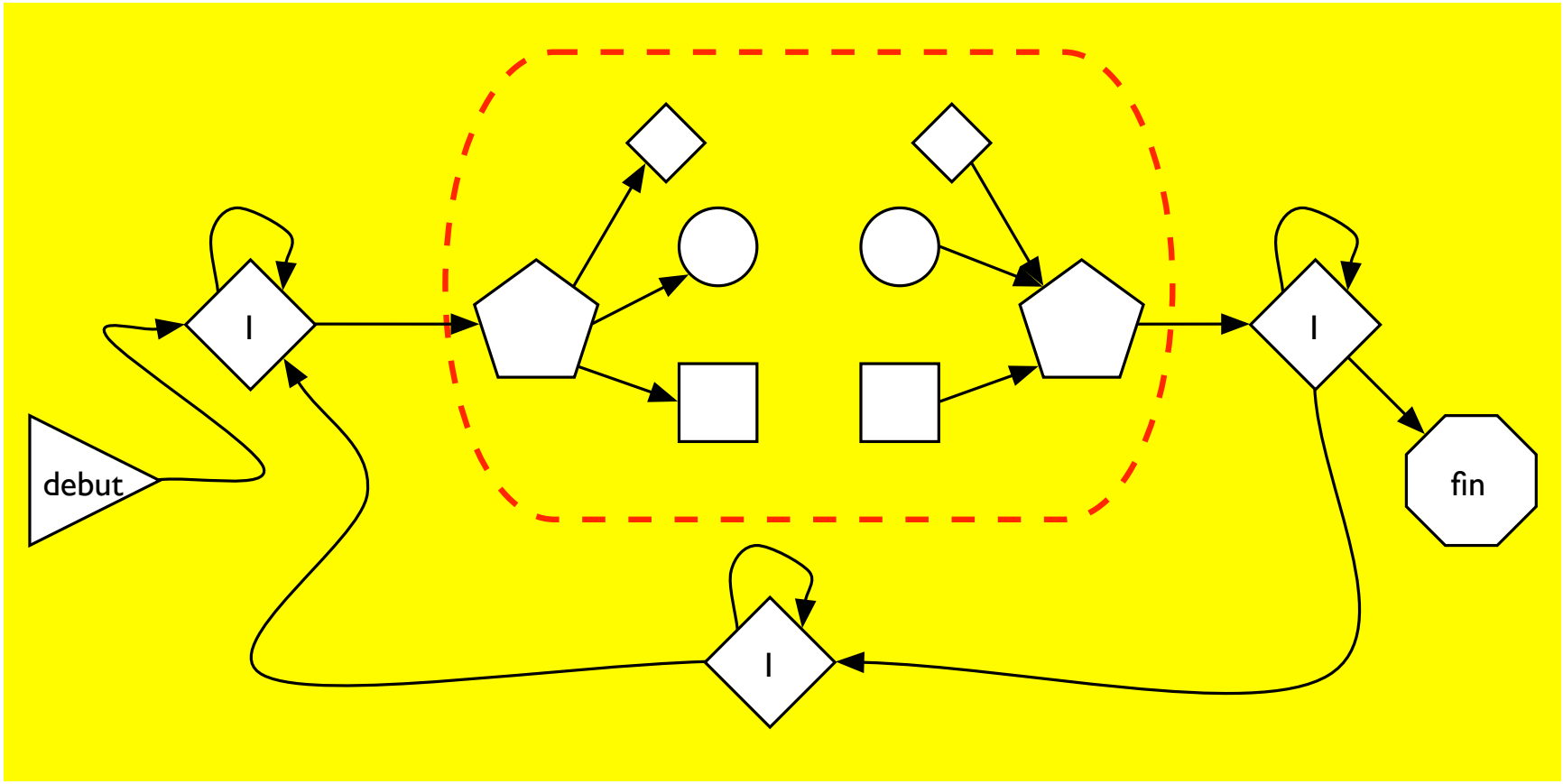
Classification

Pour une séquence s :
prendre profiles $\mathcal{M}_1, \ldots, \mathcal{M}_k$ et calculer les
vraisemblances $L_{\mathcal{M}_1}(s), \ldots, L_{\mathcal{M}_k}(s)$:
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 seiul L_{famille} : les membres de la famille ont $L > L_{\text{famille}}$,
autres ont $L < L_{\text{famille}}$.

HMMs - extensions

1. alignement local + plusieurs occurrences dans une séquence



HMMs - extensions 2

2. reconnaissance de gènes – architecture compliquée + combinaison de modèles

- «modules» pour exons et introns, régions entre gènes, promoteurs, etc.

- une difficulté : modéliser la distribution de longueurs des modules (duration de rester dans le même état)

HMMs - extensions 3

3. émission de «structure» (Bystroff)

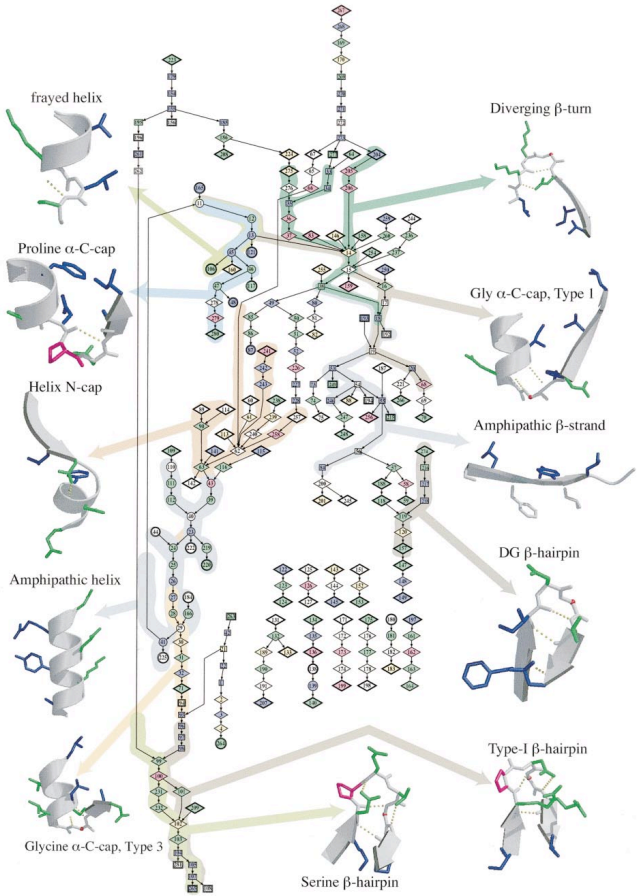


Figure 4 *(legend opposite)*