Cours 4

Modèles de Markov

Hélène Touzet

Le casino malhonnête

- ▷ *1 dé normal* : probabilité de 1/6 par chiffre
- ▷ *1 dé pipé* : le 6 a une probabilité de 0.5 et les autres faces de 0.1
- ▷ *passage du dé normal au dé pipé* : probabilité 0.05
- ▷ *passage du dé pipé au dé normal* : probabilité 0.1

Problème

À partir de l'observation d'une partie, trouver les endroits où le jeu se fait avec le dé pipé.

Régions hydrophobes dans les protéines

▷ zones caractérisées par un fort biais de composition :

hydrophobe $\rightarrow$ Ile, Leu, Val, Phe

▷ p_i, fréquence d'apparition de chaque acide aminé dans une zone hydrophobe

▷ q_i, fréquence d'apparition de chaque acide aminé dans une protéine

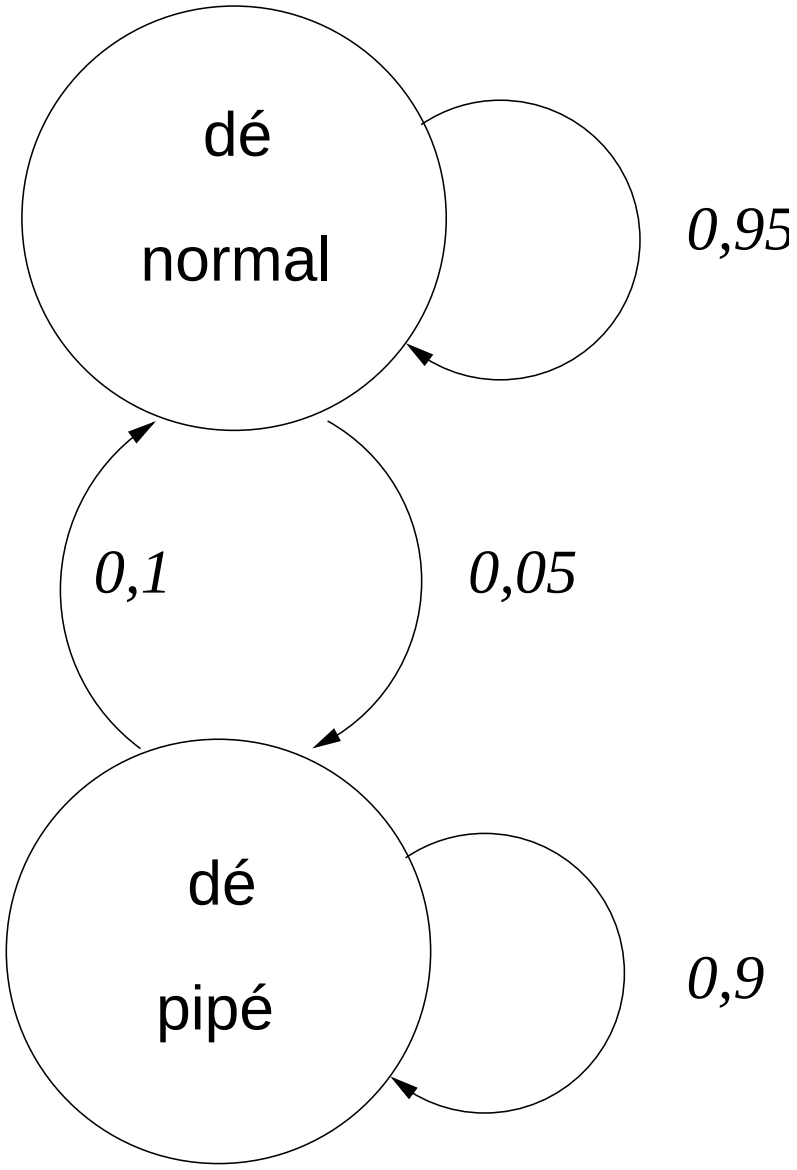
▷ m, longueur moyenne d'une région non hydrophobe

▷ n, longueur moyenne d'une région hydrophobe

Problème

Trouver les régions hydrophobes dans une séquence protéique

1 *1/6*
2 *1/6*
3 *1/6*
4 *1/6*
5 *1/6*
6 *1/6*



1 *0,1*
2 *0,1*
3 *0,1*
4 *0,1*
5 *0,1*
6 *0,5*

Processus markovien

(ou chaîne markovienne)

▷ Les événements ne sont pas indépendants.

▷ L'événement en $i+1$ dépend de celui en i, et uniquement de celui-ci.

mémoire limitée

Un **modèle de Markov** est donc déterminé par

▷ un ensemble fini d'états $\pi_0, \ldots, \pi_l$

▷ un ensemble de probabilités de transition

$$a_{kl} = P(\pi_{i+1} = l / \pi_i = k)$$

Probabilité d'accèder à l'état l alors que l'on est dans l'état k

Modèle de Markov caché

Perte d'information entre le modèle et l'observation

▷ $(\pi_i)_{i\geq 0}$, suite d'états qui suit un modèle de Markov

$$a_{kl} = P(\pi_{i+1} = l / \pi_i = k)$$

▷ $(x_j)_{j\geq 0}$, une suite d'observations

▷ e_k, des probabilités d'émissions

$$e_k(b) = P(x_i = b / \pi_i = k)$$

Probabilité d'observer b alors que l'on est dans l'état k

Exemple 1: *Protéines transmembranaires*

- ▷ protéines fichées dans la membrane d'une cellule

- ▷ permettent à la cellule de recevoir des informations extérieures

- ▷ le domaine transmembranaire est structuré en hélice α, avec un fort biais de composition en acides aminés hydrophobes

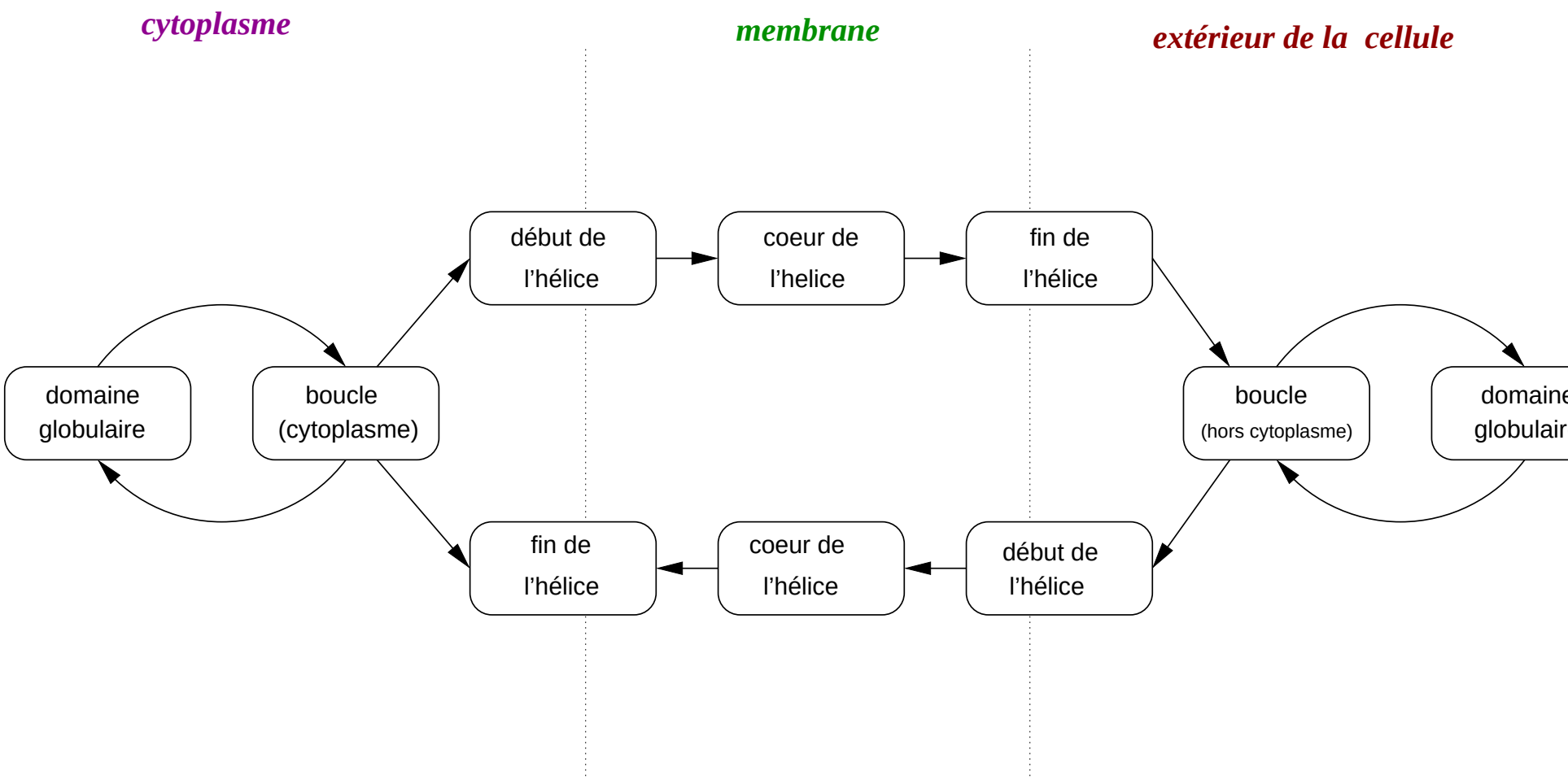
- ▷ la protéine contient souvent une succession de domaines transmembranaires

Hélice α

Motif continu :

chaque résidu en position n est lié au résidu en position $n + 4$

- ▷ en moyenne une dizaine de résidus (de 4 jusqu'à 40 dans des cas extrêmes),
- ▷ l'orientation de l'hélice est vers la droite,
- ▷ les angles entre résidus sont fixes,
- ▷ $3,6$ résidus par tour d'hélice.


cytoplasme
membrane
extérieur de la cellule
domaine globulaire
boucle (cytoplasme)
début de l'hélice
coeur de l'helice
fin de l'hélice
boucle (hors cytoplasme)
domaine globulair
fin de l'hélice
coeur de l'hélice
début de l'hélice

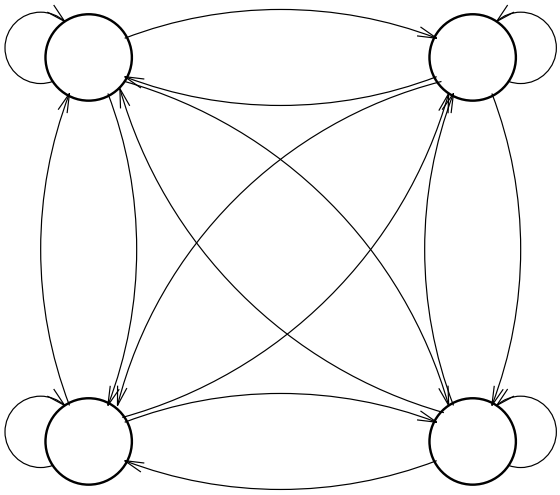
Exemple 2 : *les ilôts* CpG

- ▷ Méthylation : dans le dinucléotide ...CG..., le C mute souvent en T
- ▷ La distribution des nucléotides n'est donc pas indépendante

Probabilité de transition d'un nucléotide à l'autre

↗	A	C	G	T
A	0.30	0.21	0.28	0.21
C	0.32	0.30	0.08	0.30
G	0.25	0.25	0.30	0.20
T	0.17	0.23	0.30	0.30

Le modèle de Markov peut être vu comme un automate, où les transitions sont étiquetées par des probabilités.



automate probabiliste

. . . ou comme une grammaire régulière, où les règles de production sont étiquetées par des probabilités.

$$
\begin{array}{lcll}
A & \rightarrow & aA & [0.30] \\
A & \rightarrow & cC & [0.20] \\
A & \rightarrow & gG & [0.28] \\
A & \rightarrow & tT & [0.21] \\
C & \rightarrow & aA & [0.32] \\
 & etc. & &
\end{array}
$$

grammaire régulière stochastique

Exemple 2 : *suite*

▷ **ilôt** `CpG`**:** dans les zones du génome précédant un gène, le phénomène de méthylation disparait, et la proportion en `CpG` est donc plus importante.

Probabilités de transition dans un ilôt `CpG`

	`A`	`C`	`G`	`T`
`A`	0.18	0.27	0.43	0.12
`C`	0.17	0.37	0.27	0.19
`G`	0.16	0.34	0.37	0.13
`T`	0.08	0.36	0.38	0.18

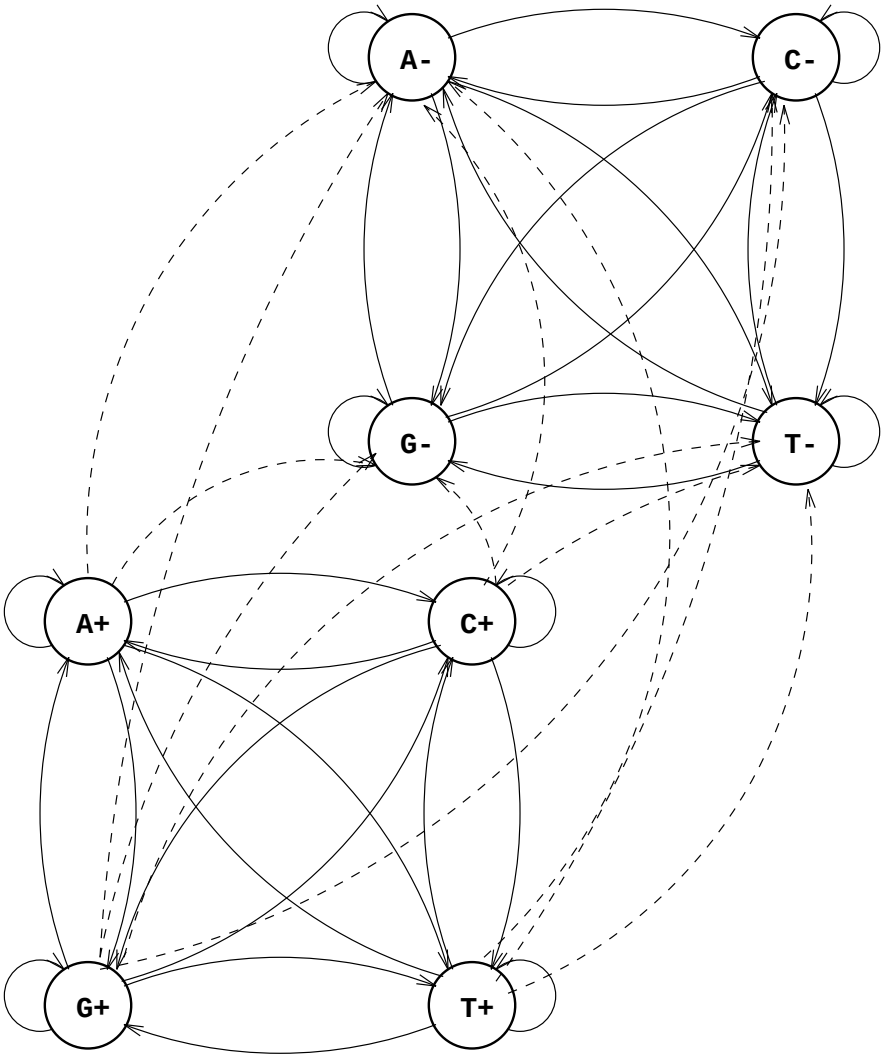
▷ Deux sous-modèles :

modèle **plus** (ilôts)
modèle **moins** (hors ilôts)

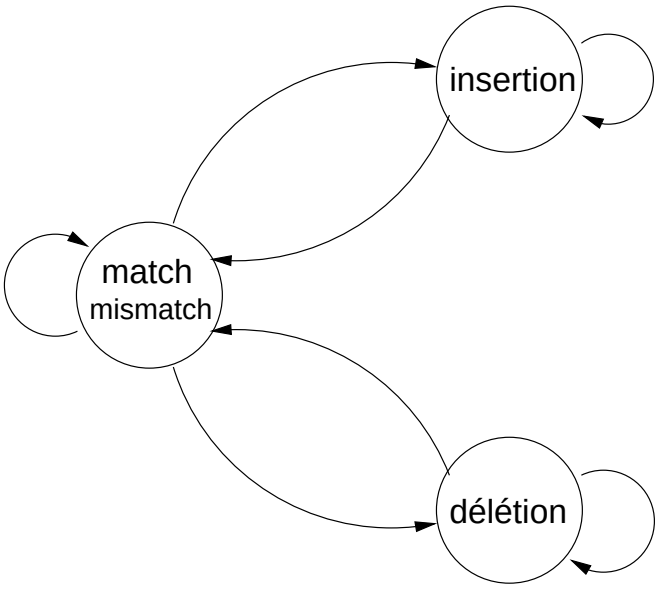
▷ Articulation des deux modèles ?

8 états `A -`, `C -`, `G -`, `T -`
`A +`, `C +`, `G +`, `T +`

4 observations `A`, `C`, `G`, `T`



Exemple 3: *alignement 2 à 2*



- ▷ Probabilités de transition :
- ▷ Emission : couple de symbôles (A, A), (A, C), $(-, T)$, etc.
- ▷ Probabilités d'émission :

- ▷ Étant donnée une protéine, où sont les domaines
  transmembranaires ?

- ▷ Étant donnée une séquence ADN, où sont les ilôts `CpG` ?

- ▷ Étant donné deux séquences, quel est le meilleur alignement ?

- ▷ Étant donnée une suite d'observations, quelle est la suite d'états la plus probable ?

Algorithme de Viterbi

- ▷ Étant donnée une séquence, quelle est son adéquation au modèle, sa vraisemblance?

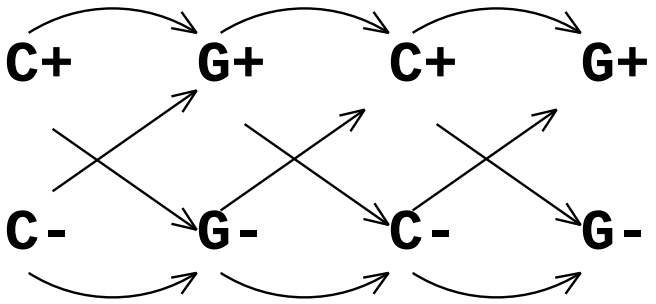
- ▷ Étant donnée une observation, quelle est sa probabilité ?

Algorithme Forward (ou Backward)

Algorithme de Viterbi

Trouver la suite d'états la plus probable pour une suite d'observations

Observation: CGCG



- ▷ $x_0, \ldots x_n$, suite d'observations
- ▷ $v_l(i)$: probabilité du chemin le plus probable entre x_0 et x_i, terminant sur l'état l
- ▷ $v_l(0) = 1$
- ▷ $v_l(i+1) = e_l(x_{i+1})\ \max_k\{v_k(i)\ a_{kl}\}$
- ▷ le chemin cherché est reconstruit à partir de $\max_l\{v_l(n)\}$.

Implémentation: programmation dynamique

▷ Table de $n \times k$

▷ **Complexité** en temps : $O(n)$ (modèle constant)

Variante: calculer $V_l(i) = \log v_l(i)$

$$V_l(i+1) = \log e_l(x_{i+1}) + \max_k \{V_k(i) + \log a_{kl}\}$$

Interêt numérique

Algorithme Forward

Trouver la probabilité d'une suite d'observations

- $f_l(i)$: probabilité de l'observation entre x_0 et x_i, le dernier état étant $\pi_i = l$.

$$f_l(i+1) \;=\; e_l(x_{i+1}) \sum_k f_k(i) a_{kl}$$

- **Implémentation :** programmation dynamique

- **Complexité :** $O(n)$ (modèle constant)

- Version symétrique : l'algorithme *Backward*

  $b_l(i)$: probabilité de l'observation entre x_{i+1} et x_n en partant de l'état $\pi_i = l$

Exemple 4: *modélisation d'un motif*

▷ Échantillon de motifs = une fonction

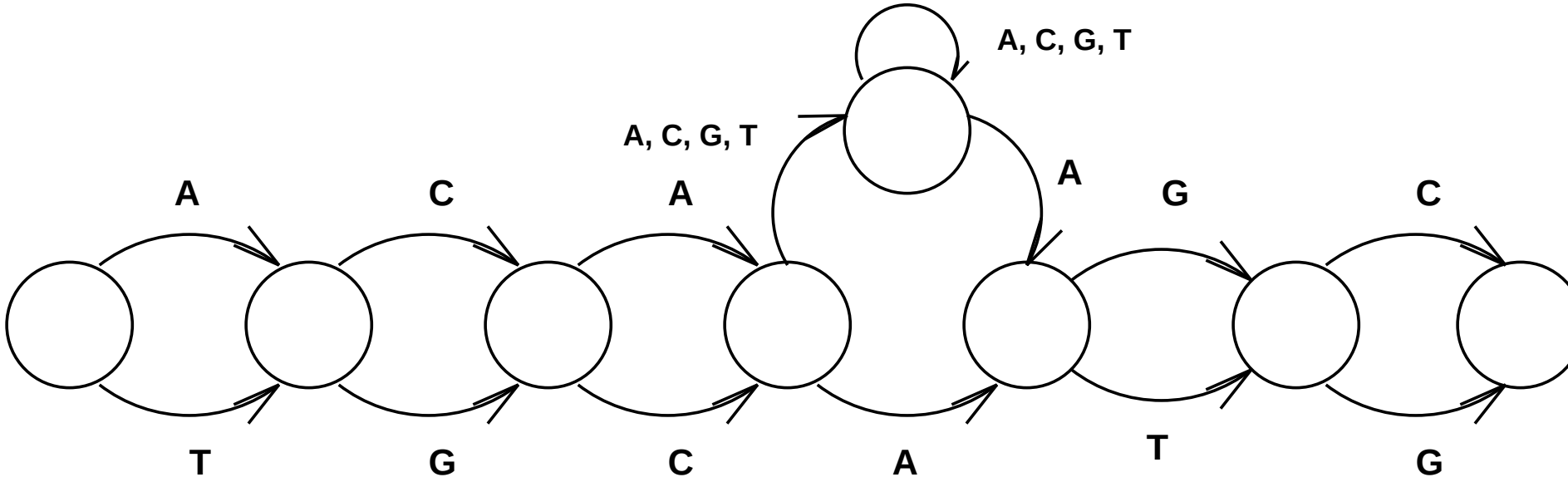
```
A C A A T G
T C A A C T A T C
A C A C A G C
A G A A T C
A C C G A T C
```

▷ Alignement multiple

```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

$$(\mathtt{A}+\mathtt{T})(\mathtt{C}+\mathtt{G})(\mathtt{A}+\mathtt{C})\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^{*}\mathtt{A}(\mathtt{G}+\mathtt{T})(\mathtt{C}+\mathtt{G})$$

$$(\mathtt{A}+\mathtt{T})(\mathtt{C}+\mathtt{G})(\mathtt{A}+\mathtt{C})((\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}^{+}\mathtt{A})+\mathtt{A})(\mathtt{G}+\mathtt{T})(\mathtt{C}+\mathtt{G})$$

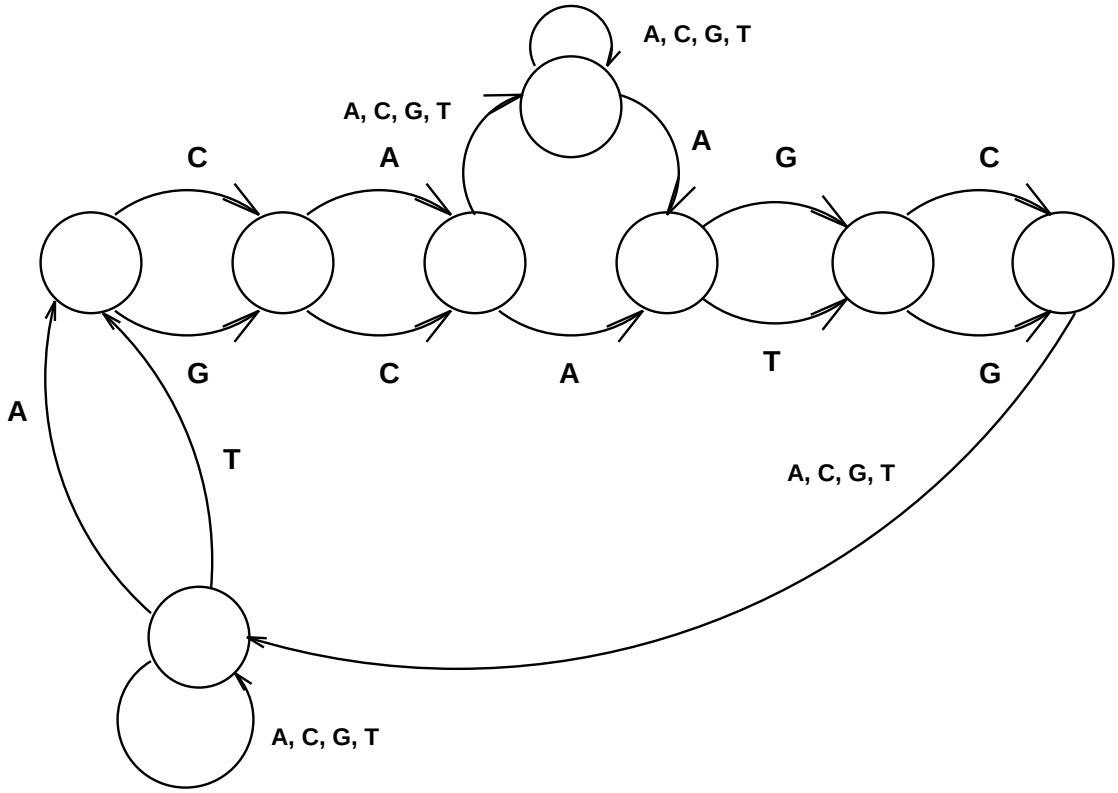

A, C, G, T
A, C, G, T
A
C
A
A
G
C
T
G
C
A
T
G

- ▷ États du modèle :
- ▷ Observations :
- ▷ Adéquation des motifs `ACACATC` et `TCCAGC`? algorithme **Forward**
- ▷ **Séquence consensus :** la séquence la plus probable

Localisation du motif dans des séquences arbitraires:

- ▷ la distribution des nucléotides hors du motif est indépendante et équiprobable

- ▷ les motifs ne sont pas chevauchants



Il ne reste plus qu'à appliquer l'algorithme de **Viterbi**.

Comment déterminer les paramètres d'un modèle de Markov ?

On dispose d'un échantillon, ensemble d'apprentissage.

Cas 1: observations = états

Modèle de Markov simple, sans probabilités d'émission

- On compte les transitions

  A_{kl} : nombre de transitions de l'état k vers l

- Problème de *sur-adaptation* : introduction de *pseudo-comptes*

$$A_{kl} \leftarrow A_{kl} + r_{kl}$$

- Normalisation pour avoir une probabilité

$$a_{kl} : \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

Cas 2: observations $\neq$ états, mais les états sont connus

Pour l'exemple 1, l'échantillon est constitué de séquences pour lesquelles on sait où sont les ilôts `CpG`*.*
Pour l'exemple 2, l'échantillon est constitué de séquences pour lesquelles on connait les occurrences du motif.

▷ Le calcul des probabilités de transition ne change pas

$$a_{kl} = \frac{A_{kl}}{\sum A_{kl'}}$$

▷ Pour les probabilités d'émission

- Recensement des observations
  $E_k(b)$: nombre de fois où l'état k donne l'observation b
- On corrige éventuellement avec des pseudo-comptes

  $E_k(b) \leftarrow E_k(b) + r_k(b)$
- Puis on normalise, pour avoir une probabilité

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Cas général : observations $\neq$ états

Modèle de Markov caché sans informations sur l'échantillon

Algorithme de Baum-Welch (1972):

- ▷ approximations successives des paramètres
- ▷ appréciation de la pertinence des paramètres: vraisemblance

  Probabilité globale de l'échantillon avec l'algorithme Forward

▷ Probabilité que a_{kl} soit utilisé à la position i dans la séquence $x = x_0 \dots x_n$

$$\frac{f_k(i)\ a_{kl}\ e_l(x_{i+1})\ b_l(i+1)}{P(x)}$$

f_k : algorithme Forward
b_k : algorithme Backward

▷ Espérance de a_{kl}

$$A_{kl} = \sum_x \sum_i \frac{f_k(i)\ a_{kl}\ e_l(x_{i+1})\ b_l(i+1)}{P(x)} \quad (1)$$

On somme sur toutes les positions de toutes les séquences de l'échantillon

▷ Pour les paramètres d'émission

$$E_k(b) = \sum_x \sum_{x_i=b} \frac{f_k(i)\ b_k(i)}{P(x)} \quad (2)$$

1. *Initialisation:* choix de valeurs initiales arbitraires pour les paramètres a et e

2. *Itération:*

   - Calculer les valeurs f_k et b_k pour toutes les séquences de l'échantillon
   - Détermination des valeurs pour E et A avec les équations 1 et 2
   - Nouvelles valeurs pour e et a (*cf* cas 2)

3. *Critère d'arrêt:* recommencer *2.* jusqu'à avoir convergence de la probabilité de l'échantillon.

Convergence vers un maximum **local** de la probabilité de l'échantillon.