

**Modélisation probabiliste en biologie**  
**Chaînes de Markov et**  
**chaînes de Markov cachées**

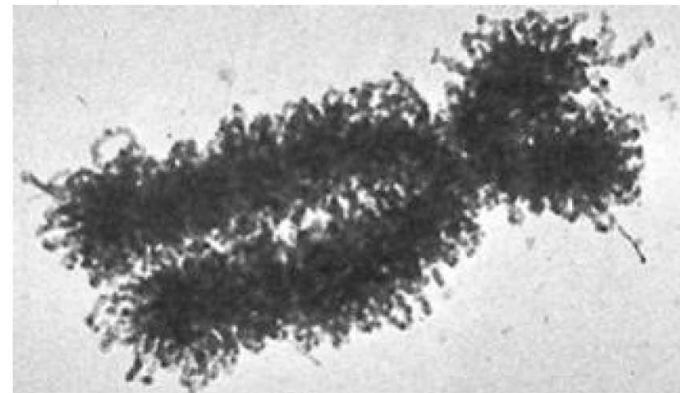
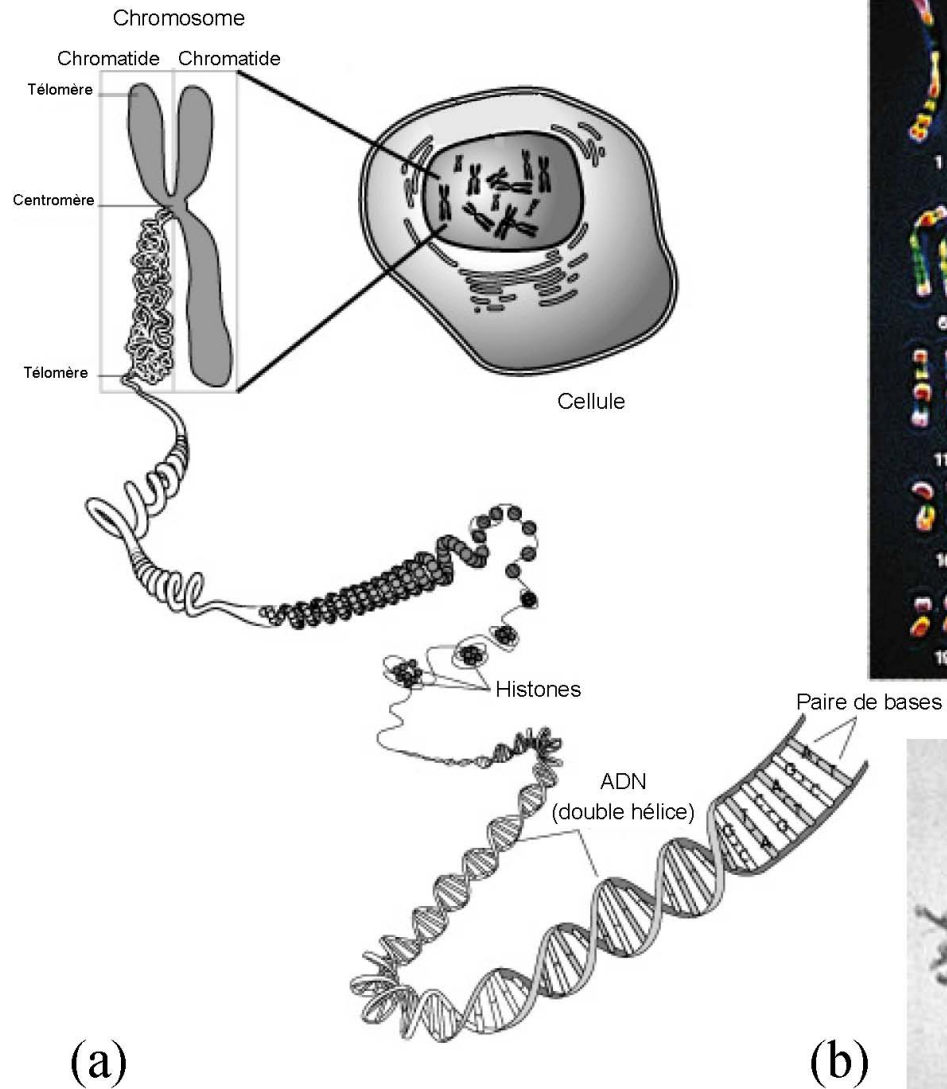
**Christelle Gonindard**

[Christelle.gonindard@univ-grenoble-alpes.fr](mailto:Christelle.gonindard@univ-grenoble-alpes.fr)

# Plan du cours

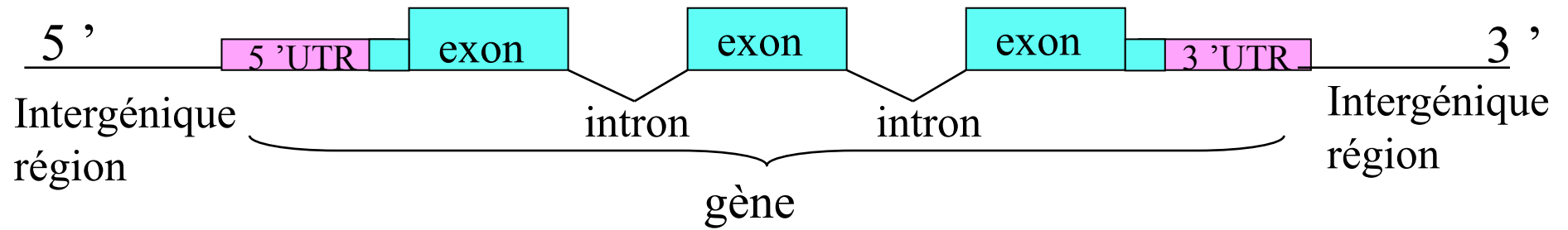
- 1- Quelques rappels biologiques
- 2- Quelques rappels probabilistes
- 3- Modélisation d'une séquence - Motivation biologique
- 4- Le modèle de chaîne de Markov cachées (CMC ou HMM)
- 5- Quelques applications Biologiques

# 1- Quelques rappels biologiques



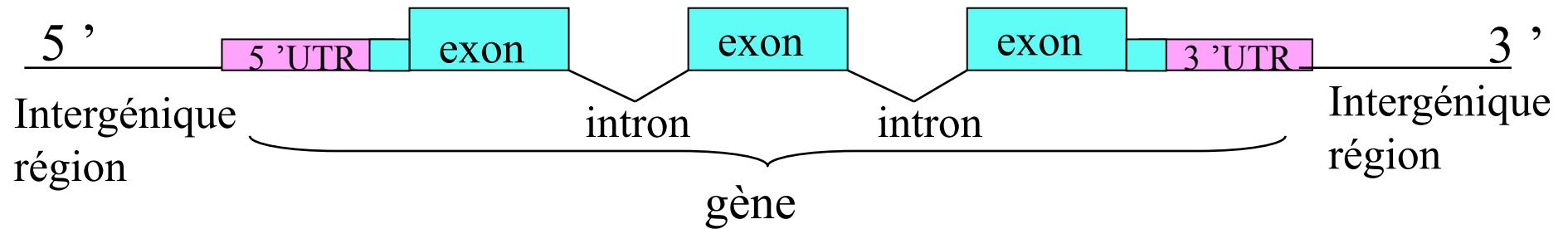
# 1- Quelques rappels biologiques

ADN

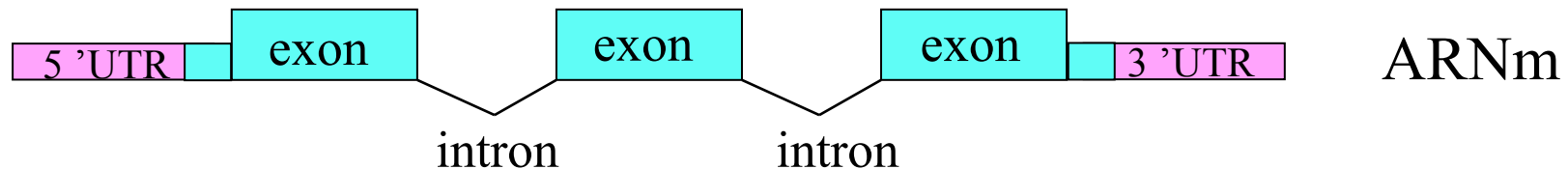


# 1- Quelques rappels biologiques

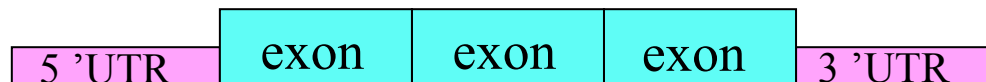
ADN



⇒ Transcription

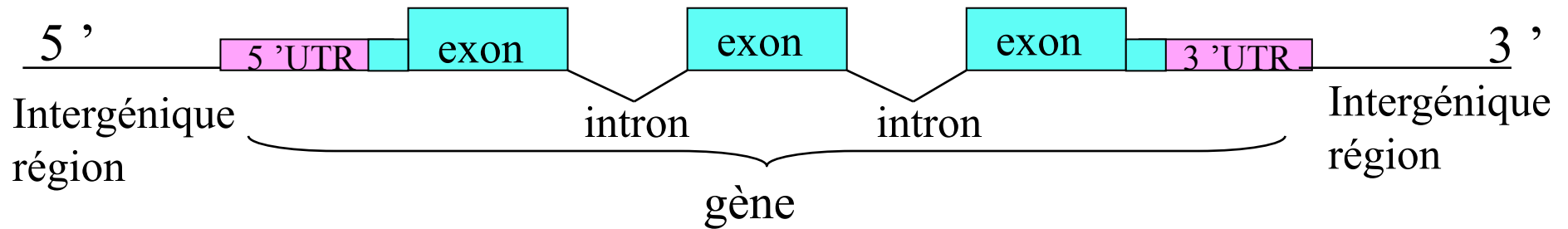


↓ épissage

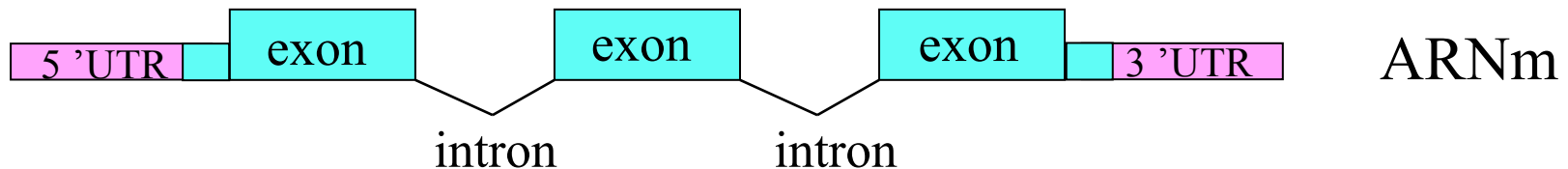


# 1- Quelques rappels biologiques

ADN



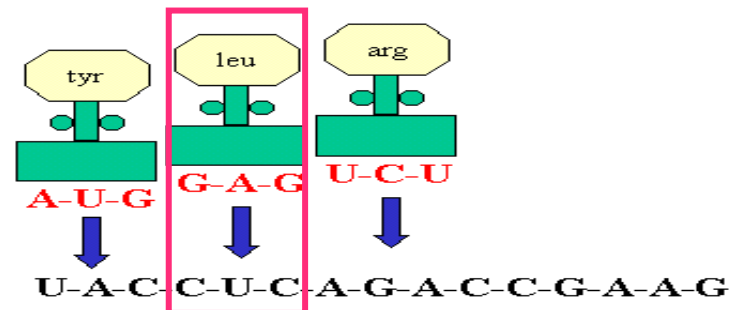
⇒ Transcription



↓ épissage

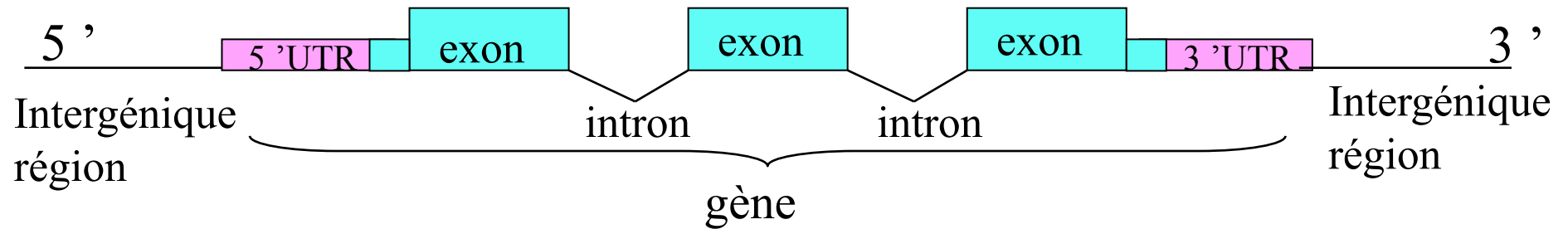


⇒ Traduction

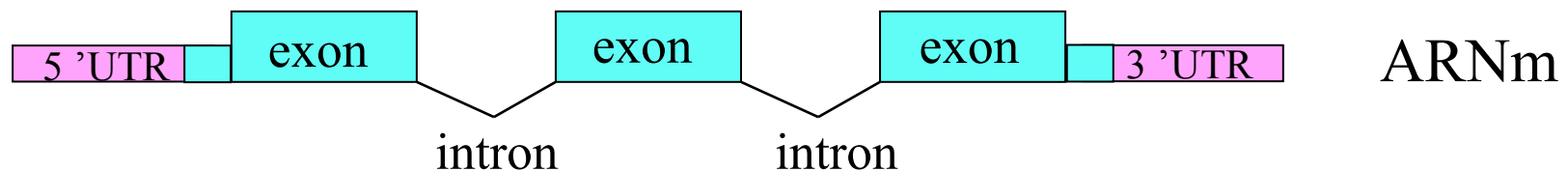


# 1- Quelques rappels biologiques

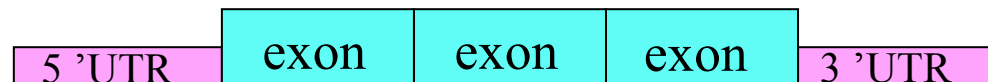
ADN



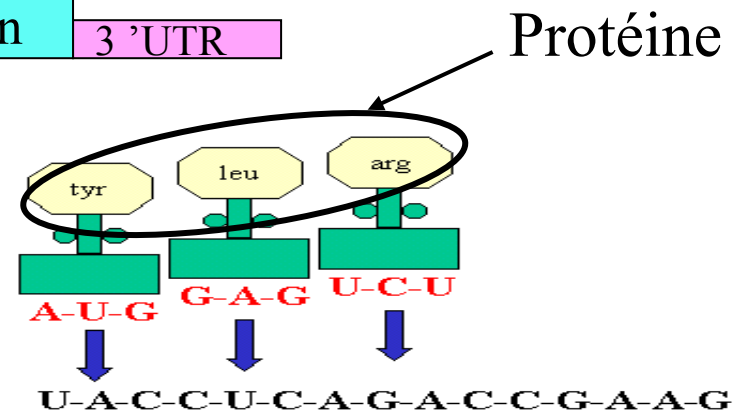
⇒ Transcription



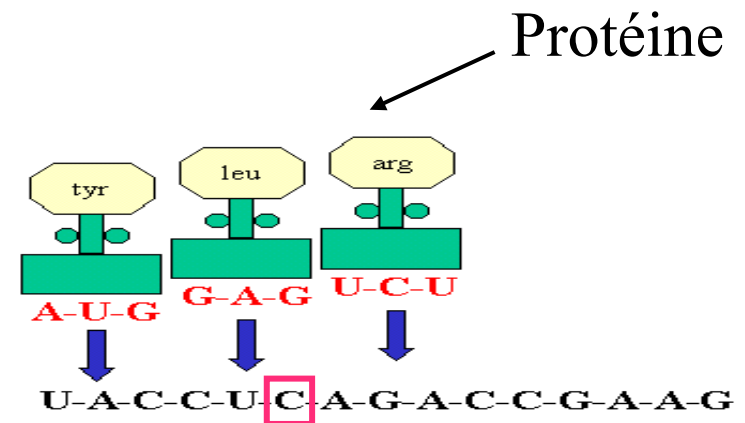
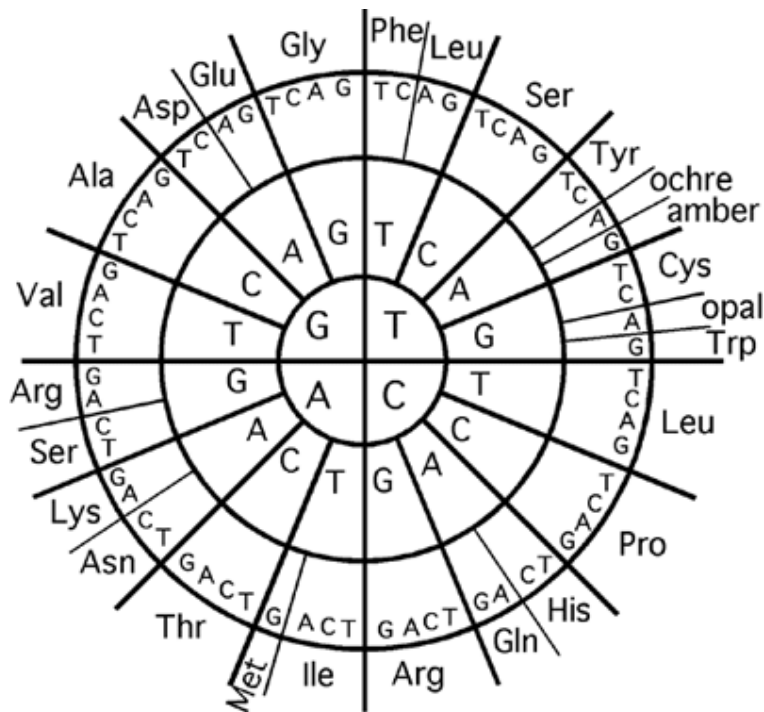
↓ épissage



⇒ Traduction

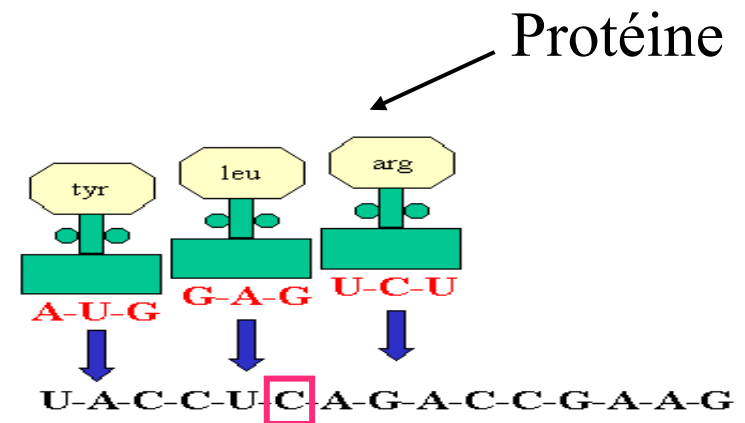
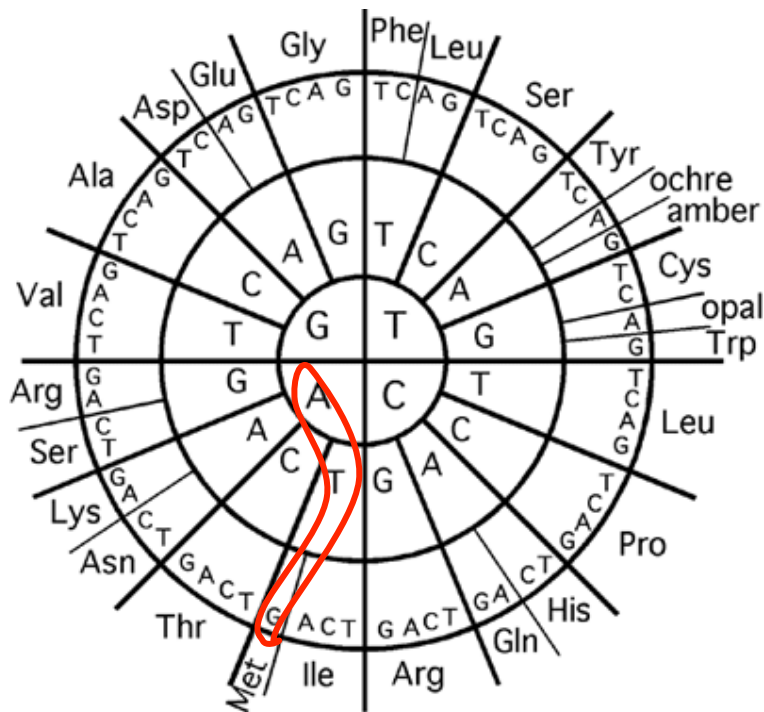


# 1- Quelques rappels biologiques

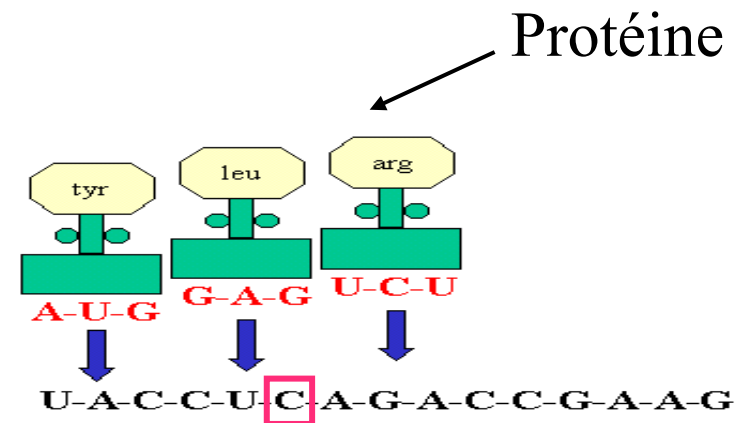
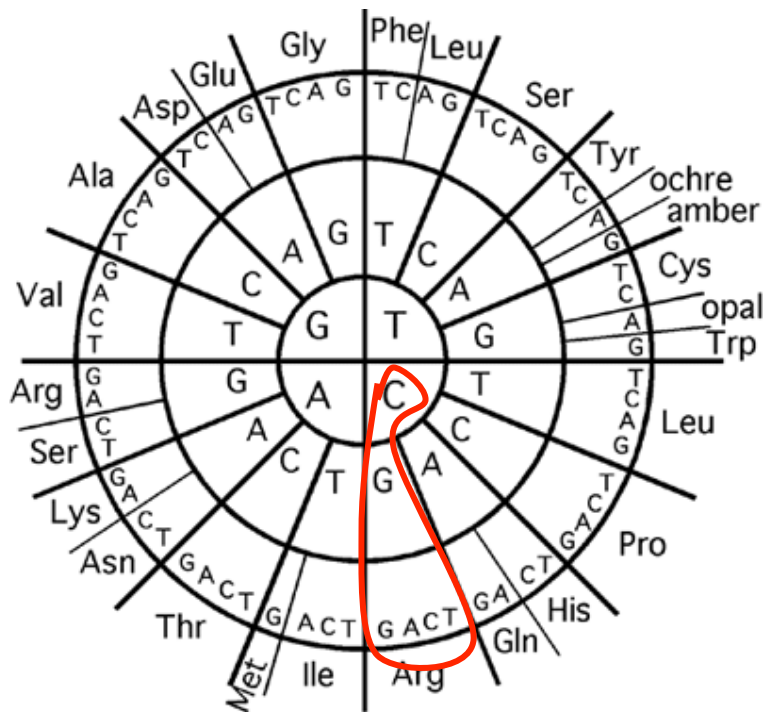




# 1- Quelques rappels biologiques



# 1- Quelques rappels biologiques



## 2- Quelques rappels probabilistes

### Modélisation :

Séquence génomique de longueur  $n$  modélisée par une suite de variables aléatoire  $X_1, X_2, \dots, X_n$  avec

$$X_i \in A$$

$$A = \{a, c, g, t\}$$

ou bien

$$A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$$

## 2- Quelques rappels probabilistes

### Qu'est ce qu'une variable aléatoire ?

On se donne un espace de probabilité  $(\Omega, P)$  assez gros pour faire toutes les mesures/expériences qui nous intéressent.

Une variable aléatoire est une fonction  $X : \Omega \rightarrow A$ .

Elle est décrite par les nombres  $p(x) = P(X = x)$  pour tout  $x \in A$ .

Donc

$$p(x) \geq 0 \text{ et } \sum_{x \in A} p(x) = 1$$

L'ensemble des  $p(x)$  pour  $x$  appartenant à  $A$  s'appelle la **loi** de  $X$  ou la **distribution** de  $X$ .

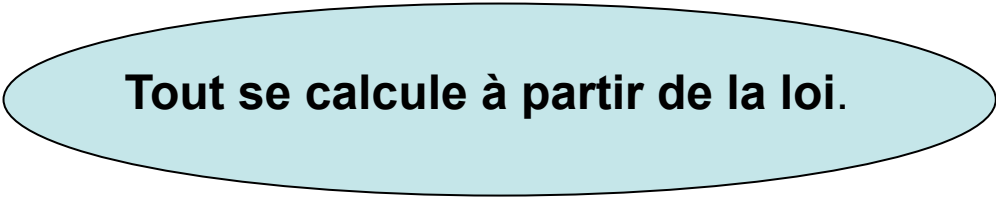
## 2- Quelques rappels probabilistes

*En pratique* : la loi de  $X$  donne  $P(x \in B)$  pour tout  $B \subset A$

et permet de calculer les moyennes.

*Exemple* : Pour calculer un taux de  $gc$ ,  $B=\{g,c\}$  et

$$P(x \in B) = p(g) + p(c)$$



**Tout se calcule à partir de la loi.**

## 2- Quelques rappels probabilistes

*La loi conjointe :*

Si  $X_1, \dots, X_n : \Omega \rightarrow A$

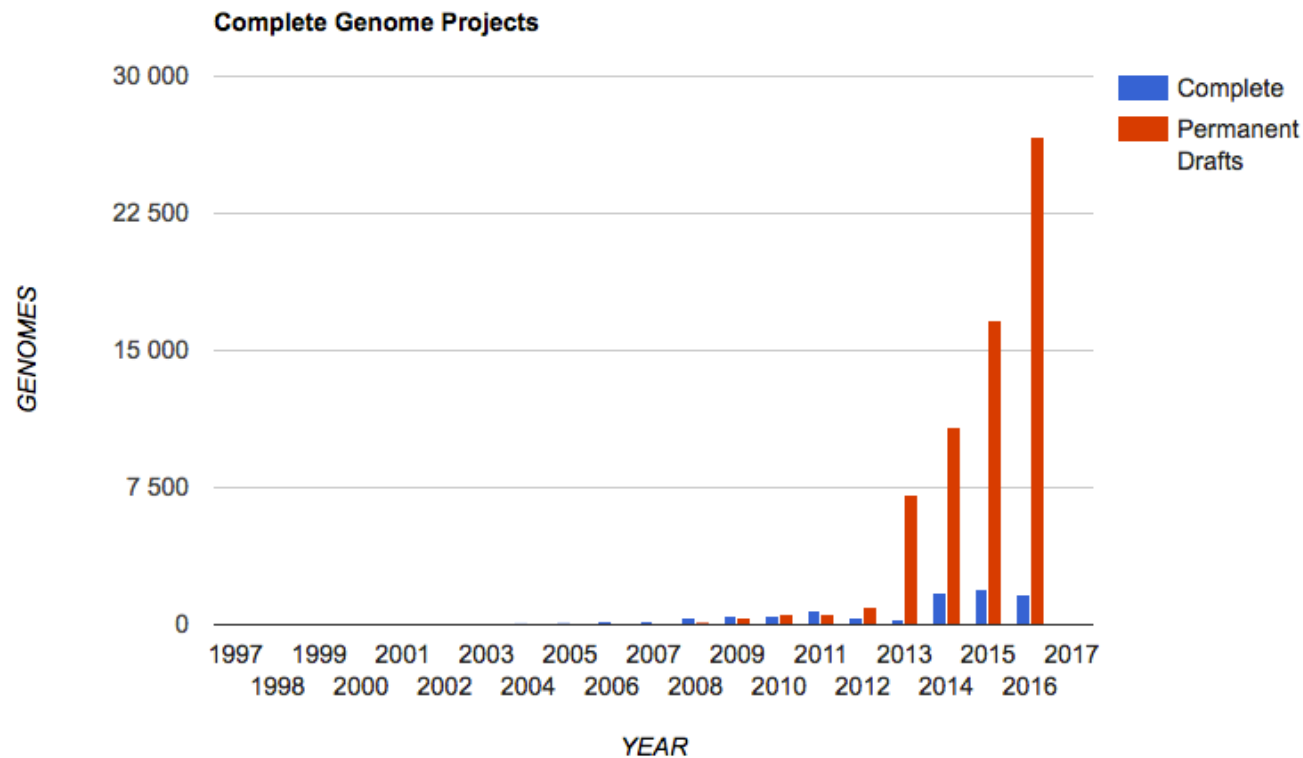
On se donne  $P(X_{1:n} = x_{1:n})$  pour tout  $x_{1:n} \in A^n$

Notation :  $X_{1:n} = (X_1, \dots, X_n)$

$x_{1:n} = (x_1, \dots, x_n)$

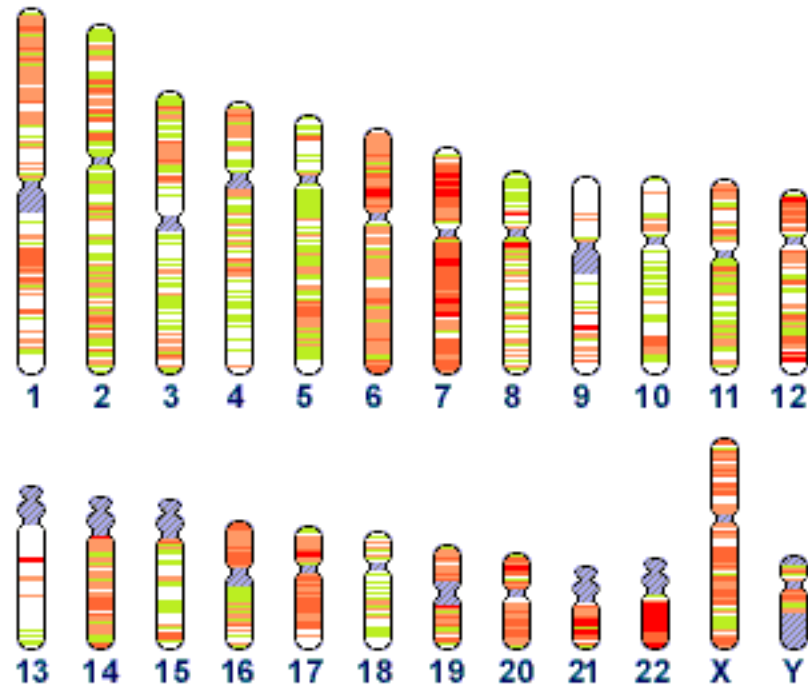
Donc  $X_{1:n} = x_{1:n}$  signifie que  $X_k = x_k$  pour tout  $1 \leq k \leq n$

### 3- Modélisation d' une séquence



La quantité d'information disponible est gigantesque nécessite de traitements automatiques et le stockage.

### 3- Modélisation d' une séquence



- Le génome humain :  $3 \cdot 10^9$  bp
- La partie codante 1 à 3% seulement du génome

La quantité d'information disponible est gigantesque nécessité de traitements automatiques et le stockage.



### 3- Modélisation d' une séquence

**Modèle** : Outil pour extraire de l' information

Un bon modèle devrait permettre de **révéler des caractéristiques relatives à la fonction ou à la structure de la séquence.**

On **ne prétend pas donner une description exacte** de la séquence avec un modèle, mais nécessite une adéquation correcte.

### 3- Modélisation d' une séquence

L' utilisation de modèles probabilistes pour l' analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu' un événement observé est significatif ou simplement le fruit du hasard ?

- fréquence ou présence d' un motif,
- score d' alignement de séquences,
- nombre de répétitions, etc.

### 3- Modélisation d' une séquence

L' utilisation de modèles probabilistes pour l' analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu' un événement observé est significatif ou simplement le fruit du hasard ?
  - fréquence ou présence d' un motif,
  - score d' alignement de séquences,
  - nombre de répétitions, etc.
- modéliser l' alternance d' états dans une séquence et caractériser cette structure le mieux possible sur une séquence observée :
  - codant/non codant (introns/exons/intergénique),
  - transferts horizontaux chez les bactéries,
  - régions variables/constantes des virus, etc.

### 3- Modélisation d' une séquence

L' utilisation de modèles probabilistes pour l' analyse de séquences biologiques intervient dans de nombreux problèmes :

- est-ce qu' un événement observé est significatif ou simplement le fruit du hasard ?
  - fréquence ou présence d' un motif,
  - score d' alignement de séquences,
  - nombre de répétitions, etc.
- modéliser l' alternance d' états dans une séquence et caractériser cette structure le mieux possible sur une séquence observée :
  - codant/non codant (introns/exons/intergénique),
  - transferts horizontaux chez les bactéries,
  - régions variables/constantes des virus, etc.
- L' analyse de l' évolution des séquences au cours du temps,etc.

### 3- Modélisation d' une séquence

Une séquence d' ADN de longueur  $n$  est modélisée par une suite de variable aléatoire (v.a.) :

$$X_1, X_2, \dots, X_n \in A = \{a, c, g, t\}$$

#### **Le Modèle M00**

- Chaque  $X_n$  vaut  $x$  **avec la même probabilité** pour chaque valeur de  $x$  dans  $A$
- Chaque  $X_n$  est **indépendant des autres  $X_k$**  pour tout  $k$  différent de  $n$

*Donc pour tout  $n \geq 1$  et tout  $x_{1:n}$*

$$P(X_{1:n} = x_{1:n}) = \frac{1}{|A|^n}$$

La propriété d'indépendance signifie que :

$$P(X_{1:n} = x_{1:n}) = P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n)$$

Avantages : calculs faciles et beaux théorèmes

### 3- Modélisation d' une séquence

Une question récurrente :

« Dans une longue séquence  $X_{1:n}$  décrite par le modèle M00, que peut-on dire de la proportion de A? »

Notation : fonction indicatrice  $1(B)$

$1(B) = 1$  si  $B$  est vrai

$1(B) = 0$  si  $B$  est faux

Comptage des proportions : 
$$N_n(x) = \sum_{k=1}^n 1(X_k = x_k)$$

$$R_n(x) = \frac{N_n(x)}{n}$$

Loi exacte (pas intéressante) :  $P(N_n(x) = k) = C_n^k \frac{3^{n-k}}{4^n}, \quad 0 \leq k \leq n$

Approximation (plus intéressante) :  $R_n(x) \rightarrow \frac{1}{4}$ , quand  $n$  devient grand

### 3- Modélisation d' une séquence

Donc : Si les proportions observées sur une longue séquence d'ADN s'éloignent nettement de 25%, 25%, 25% et 25%, cela signifie que le modèle n'est pas adapté aux données.

Exemple : le génome d'*Escherichia coli* comporte 4.6 – 5.4 Mb et :

$$\% a = 23.66$$

$$\% g = 27.89$$

$$\% c = 25.3$$

$$\% t = 23.15$$

On peut montrer que se sont des écarts trop grands sous *M00*

### 3- Modélisation d' une séquence

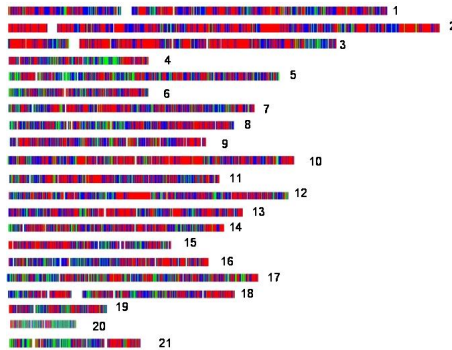
Exemple :

*Tetraodon* | *Zebrafish*

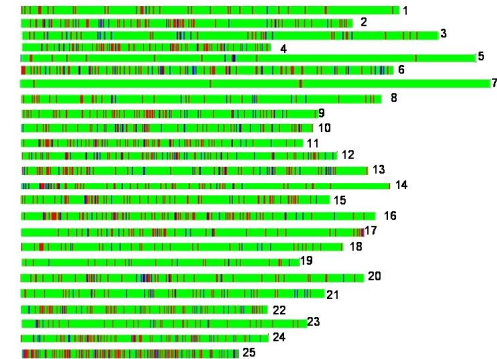


450 millions d' années divergence  
mammifère | poisson

GC riche



GC faible



Pourquoi cette différence ?  
Comment est-elle survenue ? Origine ?



### 3- Modélisation d' une séquence

#### *Le Modèle M0*

On **garde l'indépendance** mais à présent :

$$P(X_k = x) = p(x)$$

pour  $p(x) \geq 0$  avec  $\sum_{x \in A} p(x) = 1$ .

Vocabulaire :  $(p(x))_{x \in A}$  s'appelle la loi ou la distribution des  $X_n$

Formule :  $P(X_{1:n} = x_{1:n}) = p(x_1)p(x_2)...p(x_n) = \prod_{x \in A} p(x)^{N_n(x)}$

Modélise la composition en nucléotides

### 3- Modélisation d' une séquence

#### *Le Modèle M0*

On **garde l'indépendance** mais à présent :

$$P(X_k = x) = p(x)$$

pour  $p(x) \geq 0$  avec  $\sum_{x \in A} p(x) = 1$ .

Vocabulaire :  $(p(x))_{x \in A}$  s'appelle la loi ou la distribution des  $X_n$

Formule :  $P(X_{1:n} = x_{1:n}) = p(x_1)p(x_2)...p(x_n) = \prod_{x \in A} p(x)^{N_n(x)}$

Modélise la composition en nucléotides

Conséquence :

Estimer les probabilités des lettres par la fréquence des lettres d' une séquence (maximum de vraisemblance)

### 3- Modélisation d' une séquence

#### *Le Modèle M1*

A présent, les positions successives  $X_n$  ne sont plus **indépendantes**.

On commence par le cas le plus simple : la distribution des  $X_n$  est influencée par la valeur  $X_{n-1}$ . C'est ce que l'on appelle **un modèle de Markov**.

Définition : Si  $P(B_2) \neq 0$ , la probabilité conditionnelle de  $B_1$  sachant  $B_2$  est :

$$P(B_1 | B_2) = \frac{P(B_1 \cap B_2)}{P(B_2)}$$

Cela correspond à l'intuition suivante :

Si  $B_1$  = je suis en retard en cours et  $B_2$  = il neige

On peut penser que si  $B_2$  est vrai alors la circulation dans Grenoble est plus difficile et donc que  $B_1$  a plus de risque d'être vérifié.

Il vaudrait dans ce cas mieux évaluer  $B_1$  par une probabilité éventuellement différente de  $P(B_1)$ , qui rende compte du fait que  $B_2$  est réalisé (il neige) : cette nouvelle valeur est  $P(B_1 | B_2)$

### 3- Modélisation d' une séquence

#### *Le Modèle M1*

Définition :

La suite  $X_{1:n}$  est une chaîne de Markov si, pour tout  $1 \leq k \leq n-1$  et tout  $x_{1:k+1}$

$$P(X_{k+1} = x_{k+1} \mid X_{1:k} = x_{1:k}) = P(X_{k+1} = x_{k+1} \mid X_k = x_k)$$

On parle aussi de mémoire à distance 1 : si on s'intéresse à la position  $k+1$ , on peut oublier les valeurs aux positions  $1, \dots, k-1$  et ne garder que la position  $k$ .

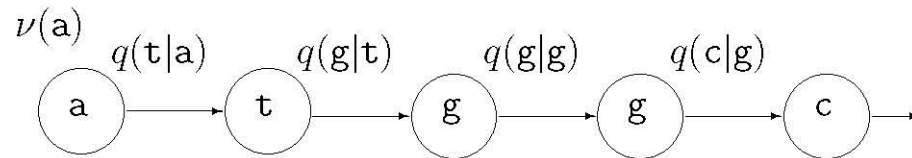
On voit que la loi de la chaîne de Markov est décrite complètement dès que l'on connaît  $P(X_1 = x)$  pour tout  $x$  dans  $A$  et  $P(X_k = x' \mid X_{k-1} = x)$  pour tout couple  $(x, x')$ .

$$v(x) = P(X_1 = x), \quad q(x' \mid x) = P(X_k = x' \mid X_{k-1} = x)$$

On note aussi  $q(x, x') = q(x' \mid x)$  (**attention : l'ordre de  $x$  et  $x'$  change**)

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*



Par exemple :

$$P(X_{1:5} = atggc) = v(a)q(t|a)q(g|t)q(g|g)q(c|g)$$

Paramètre de la chaîne de Markov

- **Loi initiale**  $v$  :  $v(x) \geq 0$  pour tout  $x \in A$  et  $\sum_{x \in A} v(x) = 1$
- **Matrice de transition**  $q$  :  $q(x, x') \geq 0$  pour tous  $x$  et  $x' \in A$

$$\text{et } \sum_{x' \in A} q(x, x') = 1$$

donc  $0 \leq v(x) \leq 1$  et  $0 \leq q(x, x') \leq 1$

$$v = (v(a), v(c), v(g), v(t))$$

Matrice stochastique

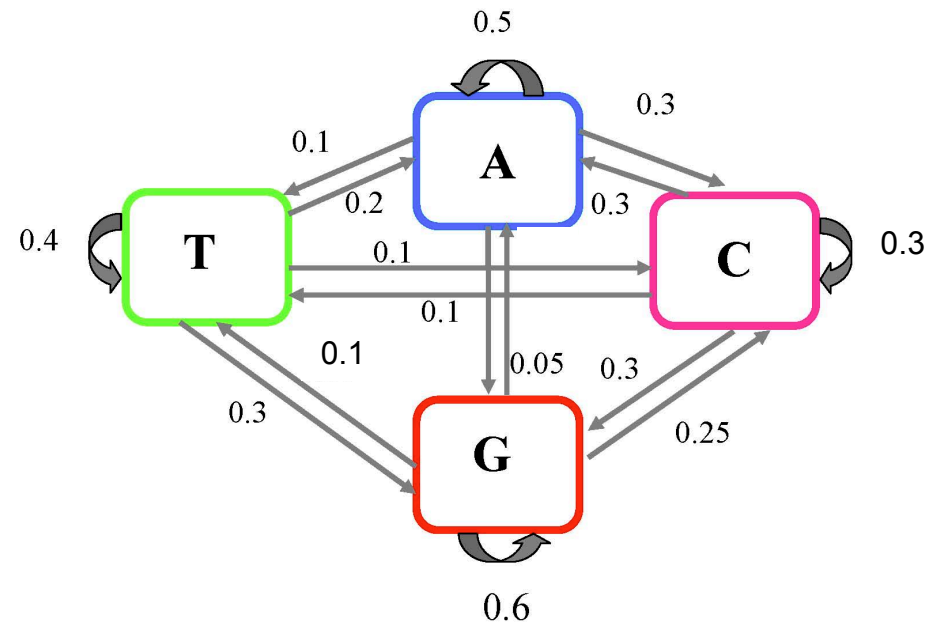
$$\left\{ \begin{array}{l} q = \begin{pmatrix} q(a, a) & q(a, c) & q(a, g) & q(a, t) \\ q(c, a) & q(c, c) & q(c, g) & q(c, t) \\ q(g, a) & q(g, c) & q(g, g) & q(g, t) \\ q(t, a) & q(t, c) & q(t, g) & q(t, t) \end{pmatrix} \end{array} \right.$$

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Exemple*

Séquence observée

GGAATTGTGCGTGCACCGGTGAACGTGCAACTGC...

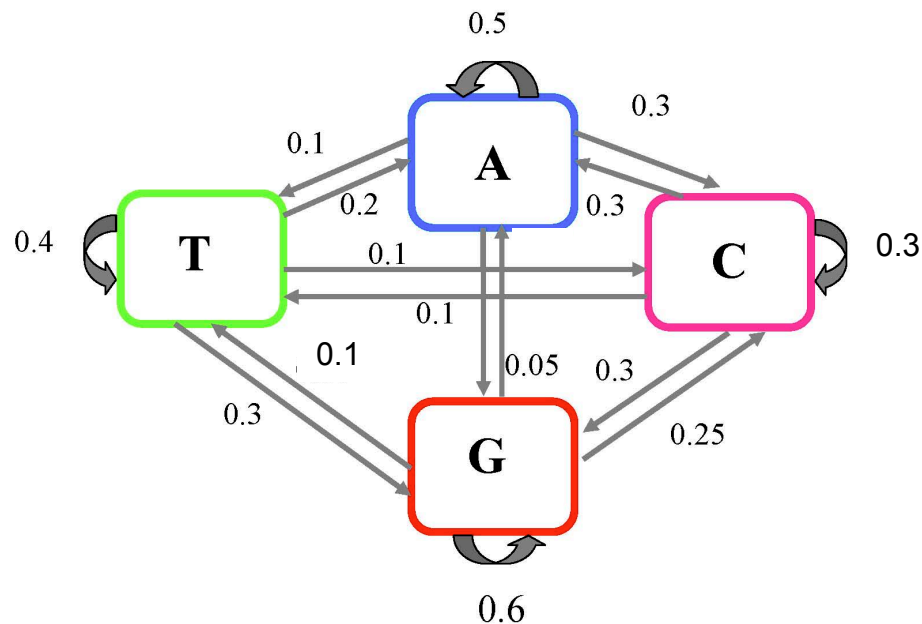


### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Exemple*

Séquence observée

GGAATTGTGCGTGCACCGGTGAACGTGCAACTGC...



↩

	A	C	G	T
A	0.5	0.3	0.1	0.1
C	0.3	0.3	0.3	0.1
G	0.05	0.25	0.6	0.1
T	0.2	0.1	0.3	0.4

Modèle M1

=> Somme des lignes =1

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*

Une chaîne de Markov (modèle M1) est un processus avec mémoire (autre que celle actuelle) :

$$P(X_{k+1} = x_{k+1} \mid X_{1:k} = x_{1:k}) = P(X_{k+1} = x_{k+1} \mid X_k = x_k)$$

Cas ADN : M1 donne une meilleur approximation de la réalité que le modèle indépendant (M0). Mais, en fait, les dépendances sont encore plus complexes.

On utilisera très vite des dépendances à  $m$  pas (modèle Mm). Le principe reste le même que pour M1



### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*

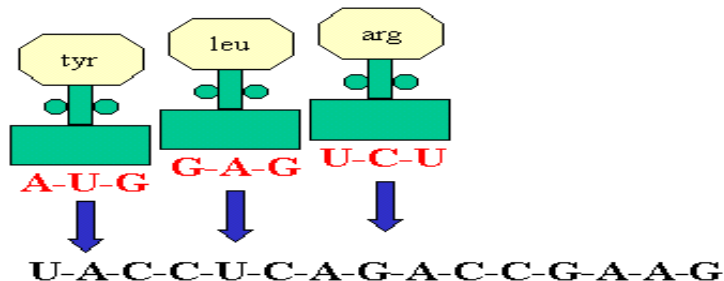
Une chaîne de Markov (modèle M1) est un processus dans mémoire (autre que celle actuelle) :

$$P(X_{k+1} = x_{k+1} \mid X_{1:k} = x_{1:k}) = P(X_{k+1} = x_{k+1} \mid X_k = x_k)$$

Cas ADN : M1 donne une meilleur approximation de la réalité que le modèle indépendant (M0). Mais, en fait, les dépendances sont encore plus complexes.

On utilisera très vite des dépendances à  $m$  pas (modèle Mm). Le principe reste le même que pour M1

Exemple : les codons



### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*

##### Quelques remarques :

- On peut utiliser les modèles pour une séquence génomique donnée.

Alors  $q(x, x')$  donne la probabilité que le site  $n+1$  soit occupé par un  $x'$  sachant que le site  $n$  est occupé par un  $x$ , dans ce cas  **$n$  est spatial.**

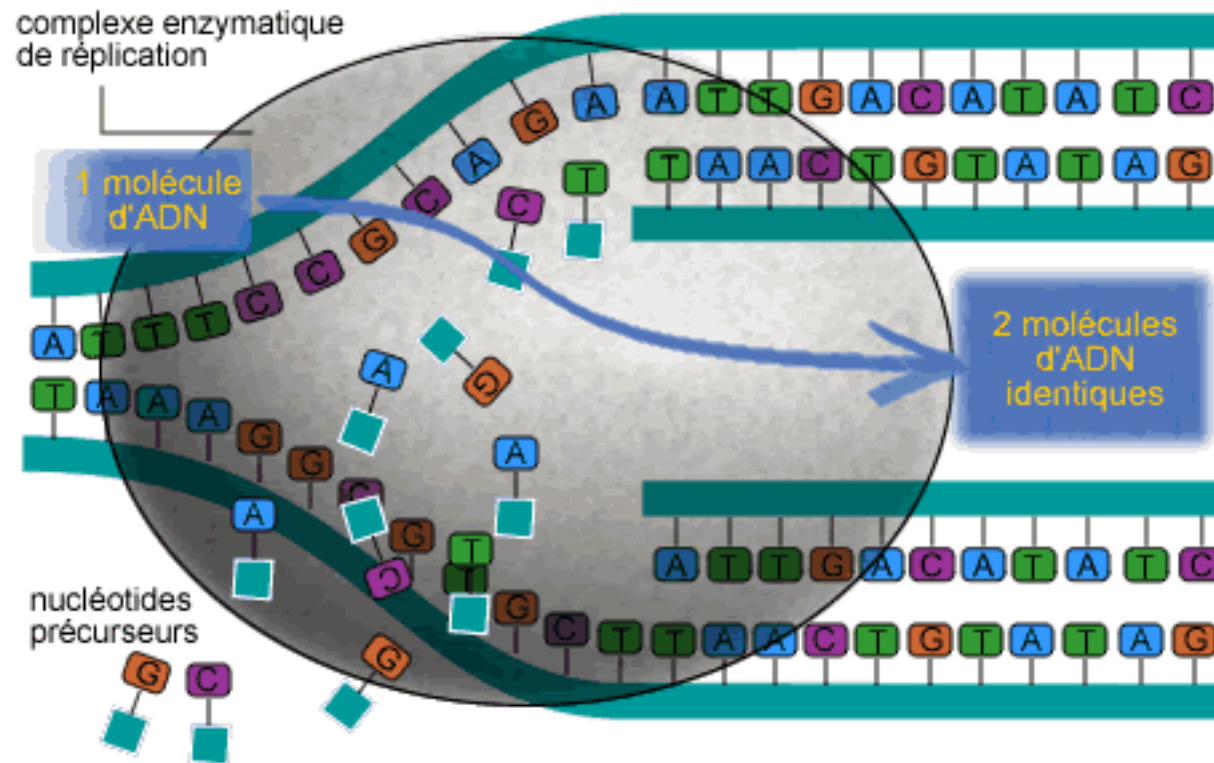
- On peut aussi utiliser  **$n$  comme un indice temporel.**

Donc  $X_n$  est le nucléotide en un site donné après  $n$  réplifications de la molécule d'ADN et par exemple, les sites évoluent indépendamment les uns des autres .

On peut penser qu'il y a eu beaucoup de réplifications donc on s'intéresse à la distribution quand  $n$  devient grand.

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*



Réplication

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*

**Deux exemples classiques de modèles M1 d'évolution :**

On s'intéresse à un site fixé et on suppose qu'il évolue indépendamment du reste de la séquence (ce qui est tout à fait faux biologiquement!!).

**Jukes-Cantor :** pour tous  $x \neq x'$ ,  $q_{JC}(x, x') = p$  avec  $0 \leq p \leq \frac{1}{3}$

$$q_{JC} = \begin{pmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{pmatrix}$$

Le paramètre  $p$  dépend de l'échelle de temps considérée.

### 3- Modélisation d' une séquence

#### *Le Modèle M1 : Modélisation d'une séquence d'adn*

**Deux exemples classiques de modèles M1 d'évolution :**

On s'intéresse à un site fixé et on suppose qu'il évolue indépendamment du reste de la séquence (ce qui est tout à fait faux biologiquement!!).

**Kimura :** Purine (a,g) versus pyrimidines (c,t)

Pour chaque transition, probabilité  $u$ .

Pour chaque transversion, probabilité  $v$ .

Donc  $0 \leq u + 2v \leq 1$ . dans l'ordre a,c,g,t:

$$q_{JC} = \begin{pmatrix} 1-u-2v & v & u & v \\ v & 1-u-2v & v & u \\ u & v & 1-u-2v & v \\ v & u & v & 1-u-2v \end{pmatrix}$$

Même remarque que pour Jukes et Cantor.

### 3- Modélisation d' une séquence

#### *Quelques définitions*

**Théorème** : Dans un modèle M1, la distribution après  $n$  pas vaut  $vq^n$

### 3- Modélisation d' une séquence

#### Quelques définitions

**Théorème** : Dans un modèle M1, la distribution après  $n$  pas vaut  $vq^n$

*Remarque* : Si  $Q$  est la matrice de transition alors  $P(x_{i+2} = e' \mid x_i = e) = Q_{ee'}^2$

**Convergence  
vers  
un équilibre**

- 1) Les  $vq^n$  varient avec  $n$ , on sent l'effet de l'âge
- 2) Chaque  $vq^n$  dépend de  $v$ , on se souvient de son état initial.
- 3) Mais tout ceci disparaît quand  $n$  devient grand, on finit par tout oublier.

Convergence vers l'équilibre : si  $n$  grand,

$$P_v(X_n = x) \approx \pi(x)$$

*Remarque* :  $\pi(x)$  est indépendant de  $n$  et de  $v$ .

Que vaut  $\pi(x)$ ? C'est un exemple de distribution stationnaire :  $\mu = q\mu$

Si on converge, c'est la distribution stationnaire.

### 3- Modélisation d' une séquence

#### Quelques définitions

**Théorème** : Dans un modèle M1, la distribution après  $n$  pas vaut  $vq^n$

*Remarque* : Si  $Q$  est la matrice de transition alors  $P(x_{i+2} = e' \mid x_i = e) = Q_{ee'}^2$

**Convergence  
vers  
un équilibre**

- 1) Les  $vq^n$  varient avec  $n$ , on sent l'effet de l'âge
- 2) Chaque  $vq^n$  dépend de  $v$ , on se souvient de son état initial.
- 3) Mais tout ceci disparaît quand  $n$  devient grand, on finit par tout oublier.

Convergence vers l'équilibre : si  $n$  grand,

$$P_v(X_n = x) \approx \pi(x)$$

*Remarque* :  $\pi(x)$  est indépendant de  $n$  et de  $v$ .

Que vaut  $\pi(x)$ ? C'est un exemple de distribution stationnaire :  $\mu = q\mu$

Si on converge, c'est la distribution stationnaire.

- Etat absorbant : un état est absorbant si  $P(x_{i+1} = e \mid x_i = e) = 1$



### 3- Modélisation d' une séquence

#### *Le Modèle Mm*

##### Chaine de Markov d'ordre $m$ : *Mm*

les  $X_i$  dépendent des  $m$  lettres précédentes et sont générées selon :

$$\mu(a_1 \dots a_m) = P(X_1 \dots X_m = a_1 \dots a_m) \quad \forall a_j \in A$$

$$\pi(a_1 \dots a_m, b) = P(X_i = b \mid X_1 \dots X_{i-1} = a_1 \dots a_m)$$

Peut s'ajuster sur la fréquence observée des  $(m+1)$  mots d'une séquence

### 3- Modélisation d' une séquence

#### *Le Modèle Mm*

##### Chaîne de Markov d'ordre $m$ : *Mm*

les  $X_i$  dépendent des  $m$  lettres précédentes et sont générées selon :

$$\mu(a_1 \dots a_m) = P(X_1 \dots X_m = a_1 \dots a_m) \quad \forall a_j \in A$$

$$\pi(a_1 \dots a_m, b) = P(X_i = b \mid X_1 \dots X_{i-1} = a_1 \dots a_m)$$

Peut s'ajuster sur la fréquence observée des  $(m+1)$  mots d'une séquence.

Ces modèles sont basés sur une **hypothèse d'homogénéité** de la séquence : les probabilités d'émission des lettres sont identiques tout au long de la séquence.

### 3- Modélisation d' une séquence

#### *Utilisation pratique des modèles*

##### Apprentissage dans un modèle $M_m$

Comptage des mots jusqu'à la longueur  $m+1$  incluse.

### 3- Modélisation d' une séquence

#### *Utilisation pratique des modèles*

##### Apprentissage dans un modèle $M_m$

Comptage des mots jusqu'à la longueur  $m+1$  incluse.

##### Vraisemblance dans un modèle $M_m$

Les Comptages des mots de longueur  $m+1$  (et la loi initiale) suffisent à calculer  $P(x)$

### 3- Modélisation d' une séquence

#### *Utilisation pratique des modèles*

##### Apprentissage dans un modèle $M_m$

Comptage des mots jusqu'à la longueur  $m+1$  incluse.

##### Vraisemblance dans un modèle $M_m$

Les Comptages des mots de longueur  $m+1$  (et la loi initiale) suffisent à calculer  $P(x)$

##### Discrimination entre modèles $M_m$

On utilise la vraisemblance pour déterminer si une nouvelle séquence  $x$  est plutôt décrite par un modèle  $+$  ou  $-$ , on calcule donc :

$$l(x) = \log \left( \frac{P_+(x)}{P_-(x)} \right) = \sum_{x,w} N(wx) \log \left( \frac{P_+(x|w)}{P_-(x|w)} \right)$$

### 3- Modélisation d' une séquence

#### *Utilisation pratique des modèles*

##### Discrimination entre modèles $M_m$

On utilise la vraisemblance pour déterminer si une nouvelle séquence  $x$  est plutôt décrite par un modèle + ou -, on calcule donc :

$$l(x) = \log \left( \frac{P_+(x)}{P_-(x)} \right) = \sum_{x,w} N(w,x) \log \left( \frac{P_+(x|w)}{P_-(x|w)} \right)$$

Première étape : estimation de  $q_+$  et  $q_-$ .

Deuxième étape : loi empirique de  $l(x)$  quand  $x$  suit un modèle + puis quand  $x$  suit un modèle -

Si les deux lois empiriques diffèrent nettement, on peut tester de nouvelles séquences.

### 3- Modélisation d' une séquence

#### *Exemple : les îlots CpG*

**Attention** : CpG désigne c puis g sur le même brin, et non pas une paire complémentaire c-g en un locus donné des deux brins.

**Principe biologique** : la cytosine c des CpG a tendance à être méthylée, souvent en thymine t. Donc les dinucléotides cg sont plus rares que le produit des fréquences de c et de g.

Sauf autour des promoteurs de certains gènes, où la méthylation est réprimée.

**Fait d'expérience** : plus de cg et de c et g autour des régions promotrices qu'ailleurs: on parle d'îlots CpG.

### 3- Modélisation d' une séquence

#### *Exemple : les îlots CpG*

**Objectifs** : trouver les îlots CpG.

**Remarque** : problème de dinucléotides donc M1 naturel

**Référence** : Durbin, Eddy, Krogh, Mitchison (1998)

- Ensemble d'entraînement de 60kb, 48 îlots GpG
- Deux modèles M1 (estimation par maximum de vraisemblance : comptage), notés + pour les ilots GpG et – pour le reste.

$$q_+ = \begin{pmatrix} .180 & .274 & .426 & .120 \\ .171 & .368 & .274 & .188 \\ .161 & .339 & .375 & .125 \\ .079 & .355 & .384 & .182 \end{pmatrix}.$$

$$q_- = \begin{pmatrix} .300 & .205 & .285 & .210 \\ .322 & .298 & .078 & .302 \\ .248 & .246 & .298 & .208 \\ .177 & .239 & .292 & .293 \end{pmatrix}.$$



### 3- Modélisation d' une séquence

#### *Exemple : les îlots CpG*

$$q_+ = \begin{pmatrix} .180 & .274 & .426 & .120 \\ .171 & .368 & .274 & .188 \\ .161 & .339 & .375 & .125 \\ .079 & .355 & .384 & .182 \end{pmatrix} .$$
$$q_- = \begin{pmatrix} .300 & .205 & .285 & .210 \\ .322 & .298 & .078 & .302 \\ .248 & .246 & .298 & .208 \\ .177 & .239 & .292 & .293 \end{pmatrix} .$$

#### **Premier problème :**

Identifier une séquence  $x$  comme étant un îlot CpG ou non.

Calculs de vraisemblance : le  $\log(\text{score})$  de  $x$  est :

$$l(x) = \log\left(\frac{P_+(x)}{P_-(x)}\right) = \sum_{x, x' \in A} N(x, x') \log\left(\frac{q_+(x, x')}{q_-(x, x')}\right)$$

### 3- Modélisation d' une séquence

#### *Problème lié à ces modèles*

Ces modèles sont basés sur une **hypothèse d'homogénéité** de la séquence : les probabilités d'émission des lettres sont identiques tout au long de la séquence.

Pour l'ADN : gènes/régions intégéniques, introns/exons, etc ...

Idée : décrire chaque type de région par un modèle  $M_m$  spécifiques puis recoller ces différents modèles.