I. Analyse de séquences biologiques

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

- Partie I : Introduction à la biologie moléculaire
- Partie II : Les modèles de Markov cachés (HMMs)
- Partie III : Exemples d'application

Première partie I

Introduction à la biologie moléculaire

Sommaire

Règnes du vivant

Les 3 règnes

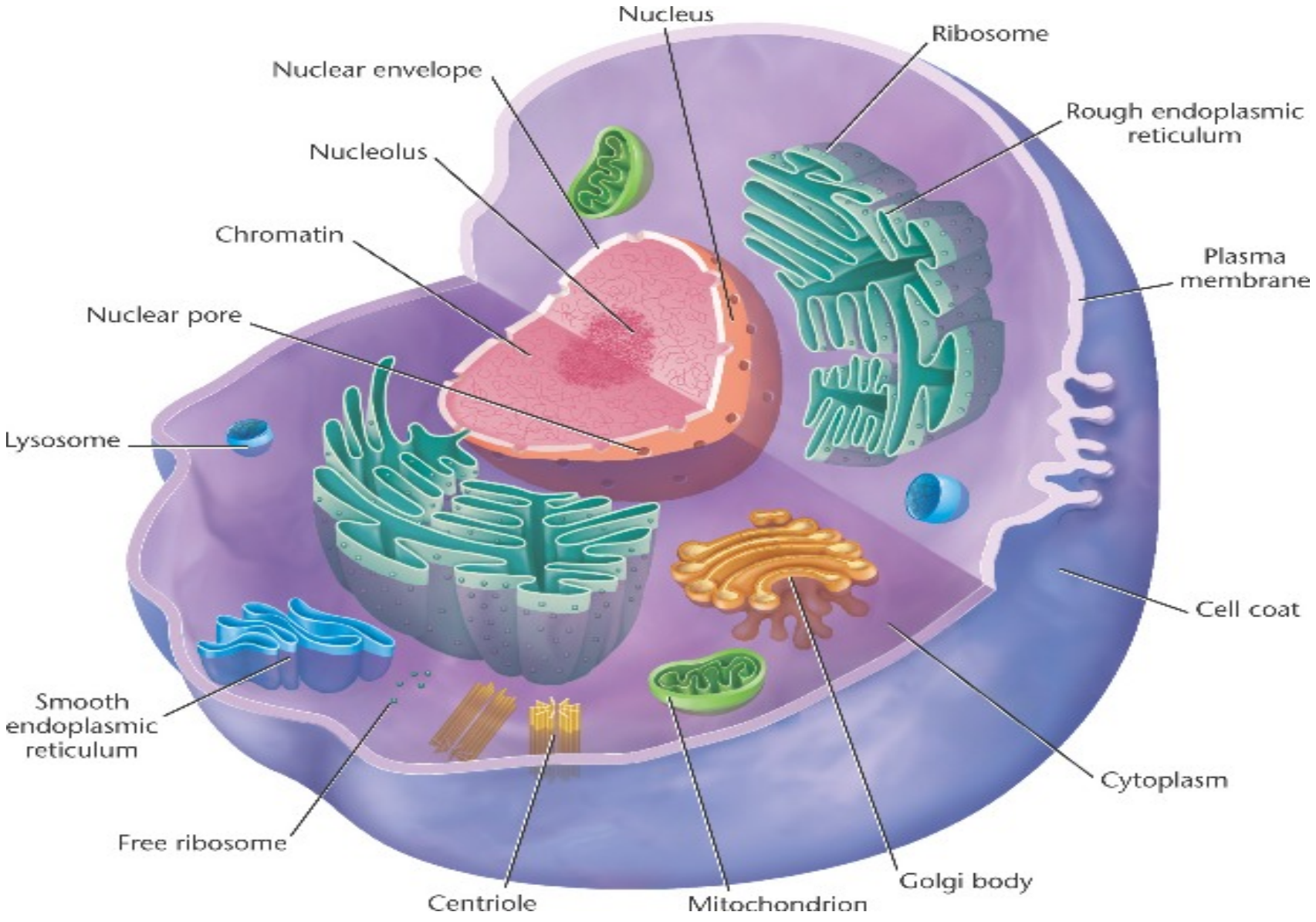
- Bactéries (ou eubactéries),
- Archées,
- Eucaryotes (cellules à noyau) : animaux, plantes, champignons, protistes.

Les bactéries et les archées sont procaryotes (cellules sans noyau).

Quelques organismes modèles

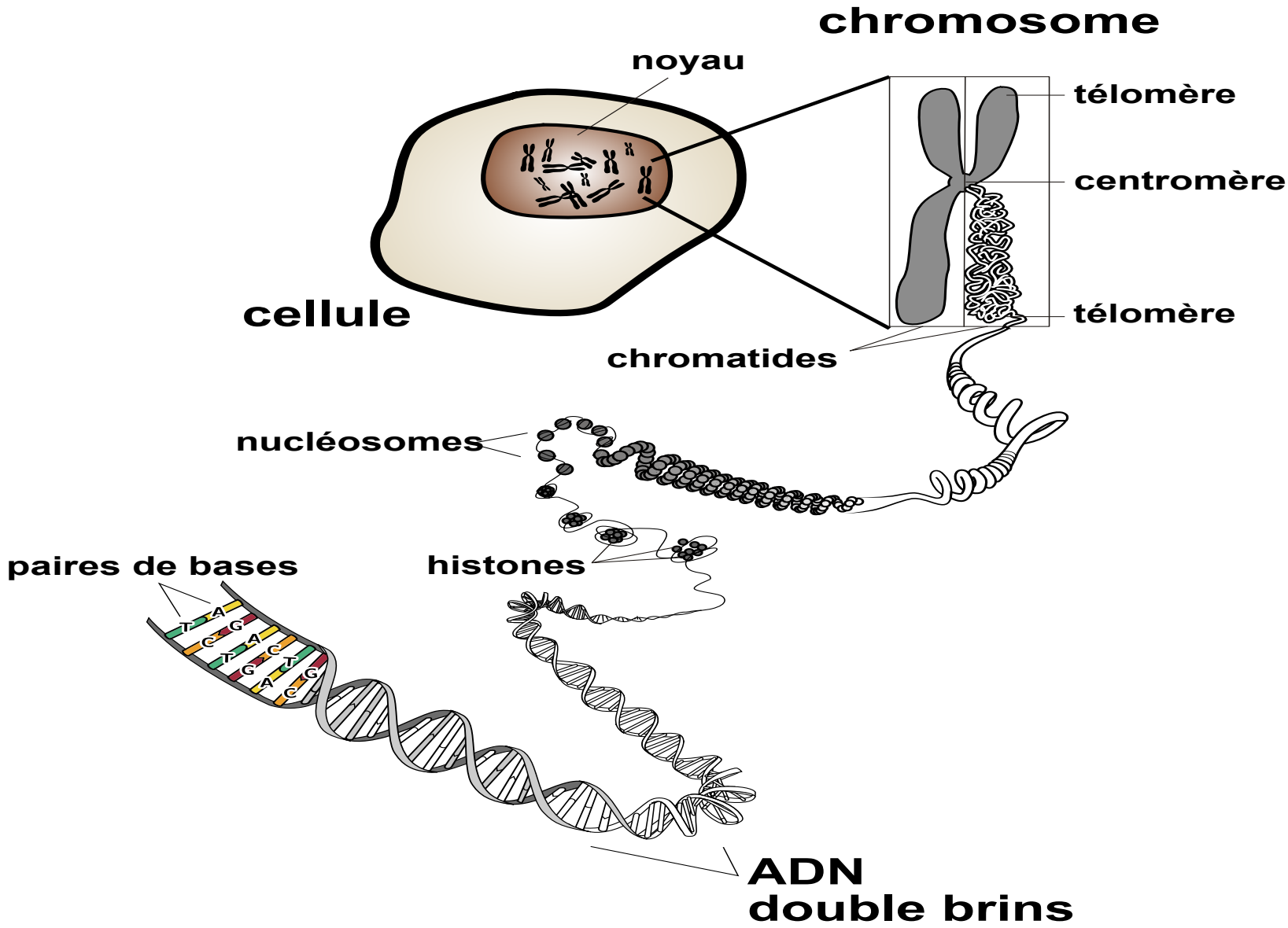
- *Escherichia coli*, *Bacillus subtilis* : bactéries,
- *Arabidopsis thaliana* (ou arabette) : plante,
- *Saccharomyces cerevisiae* : levure (champignons),
- *Homo sapiens*, *Rhesus macaque*, souris ...
- *Drosophila melanogaster* (ou mouche du vinaigre) ...

La cellule (ici avec noyau)



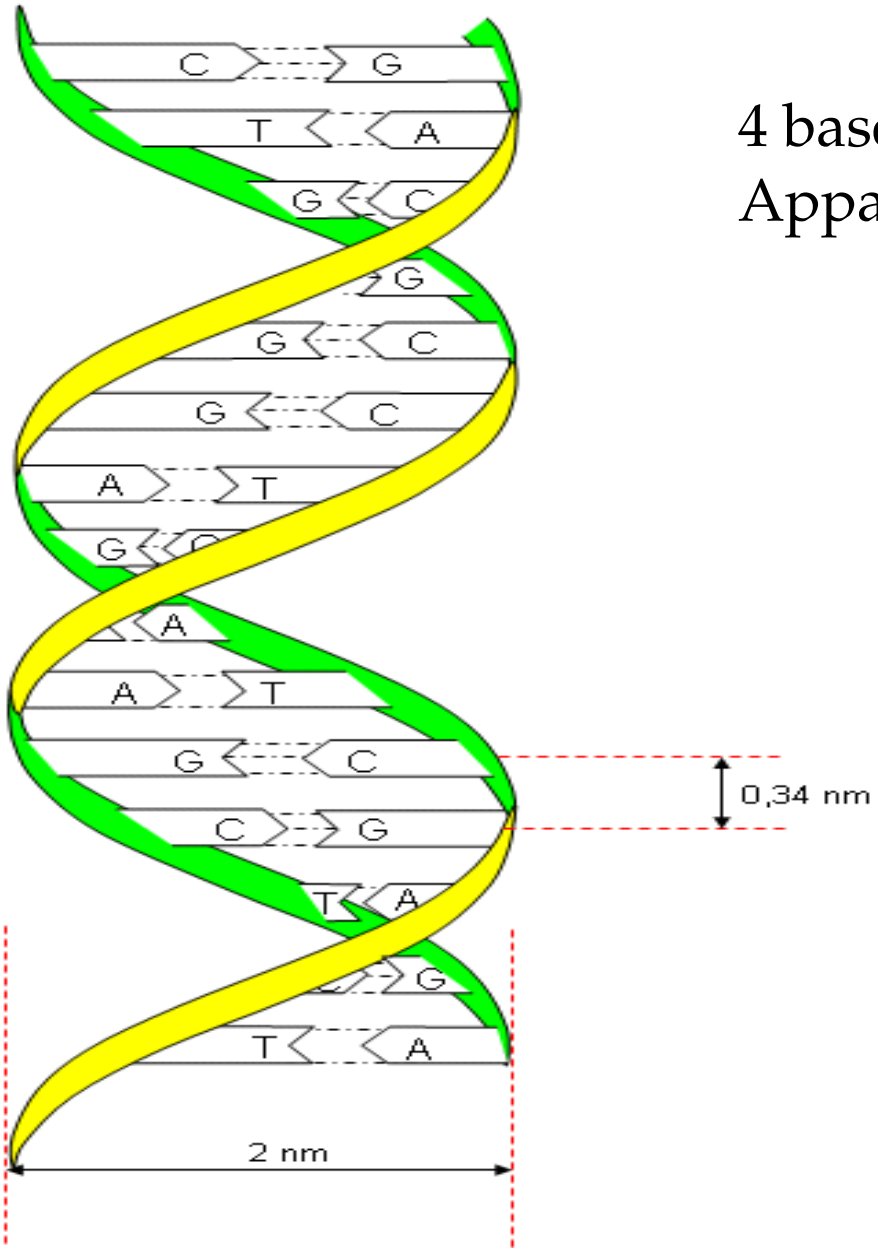
L'information génétique : les chromosomes

Chez l'Homme, 22 paires de chromosomes plus une paire de chromosomes sexuels.



L'information génétique : l'ADN

Chez l'Homme : 3.4 milliards de nucléotides (*paires de bases*).



4 bases différentes : A,C,G et T
Appariement C/G et A/T

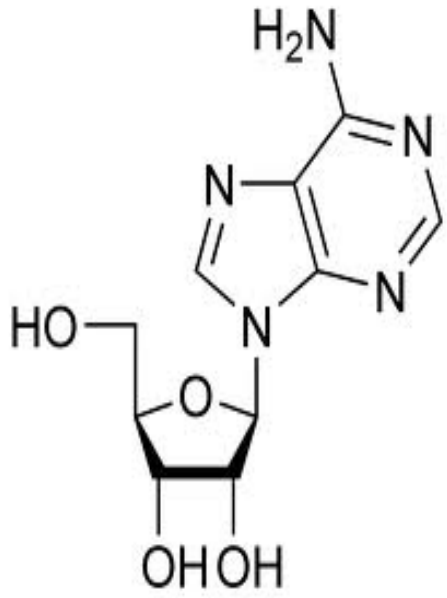
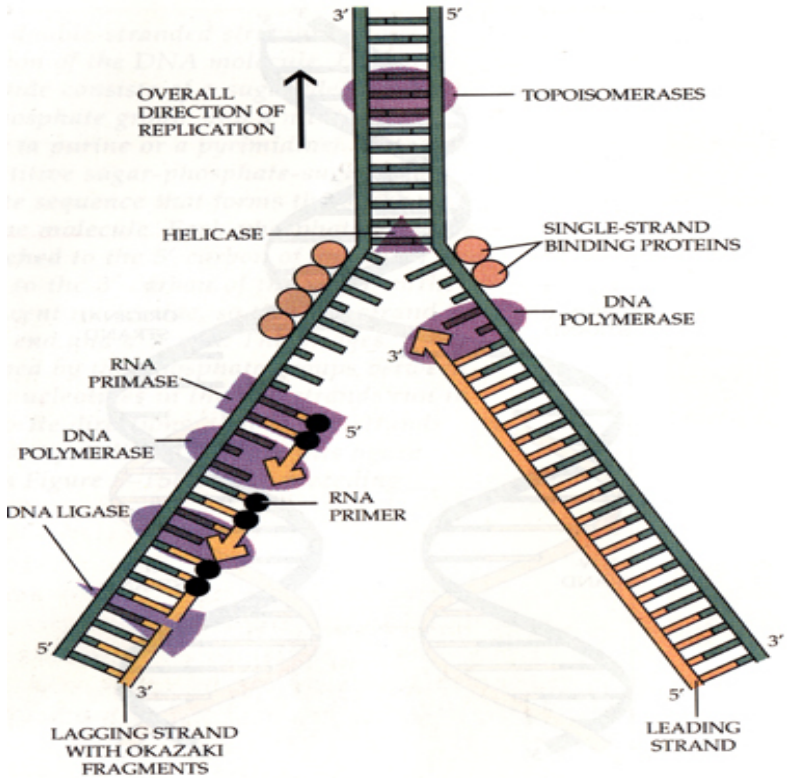


FIGURE : Adénosine.

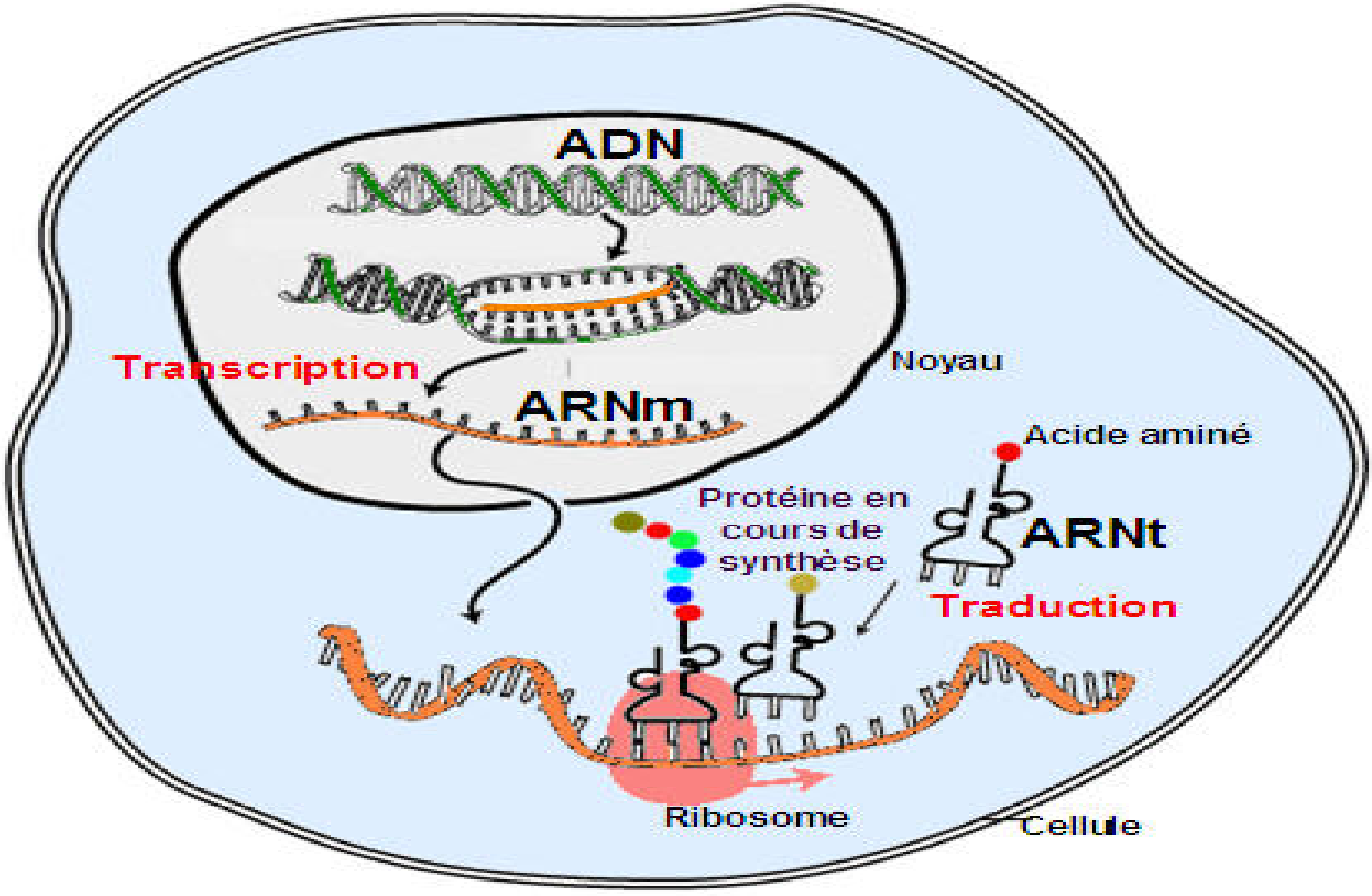
Réplication de l'ADN et variabilité de l'information génétique



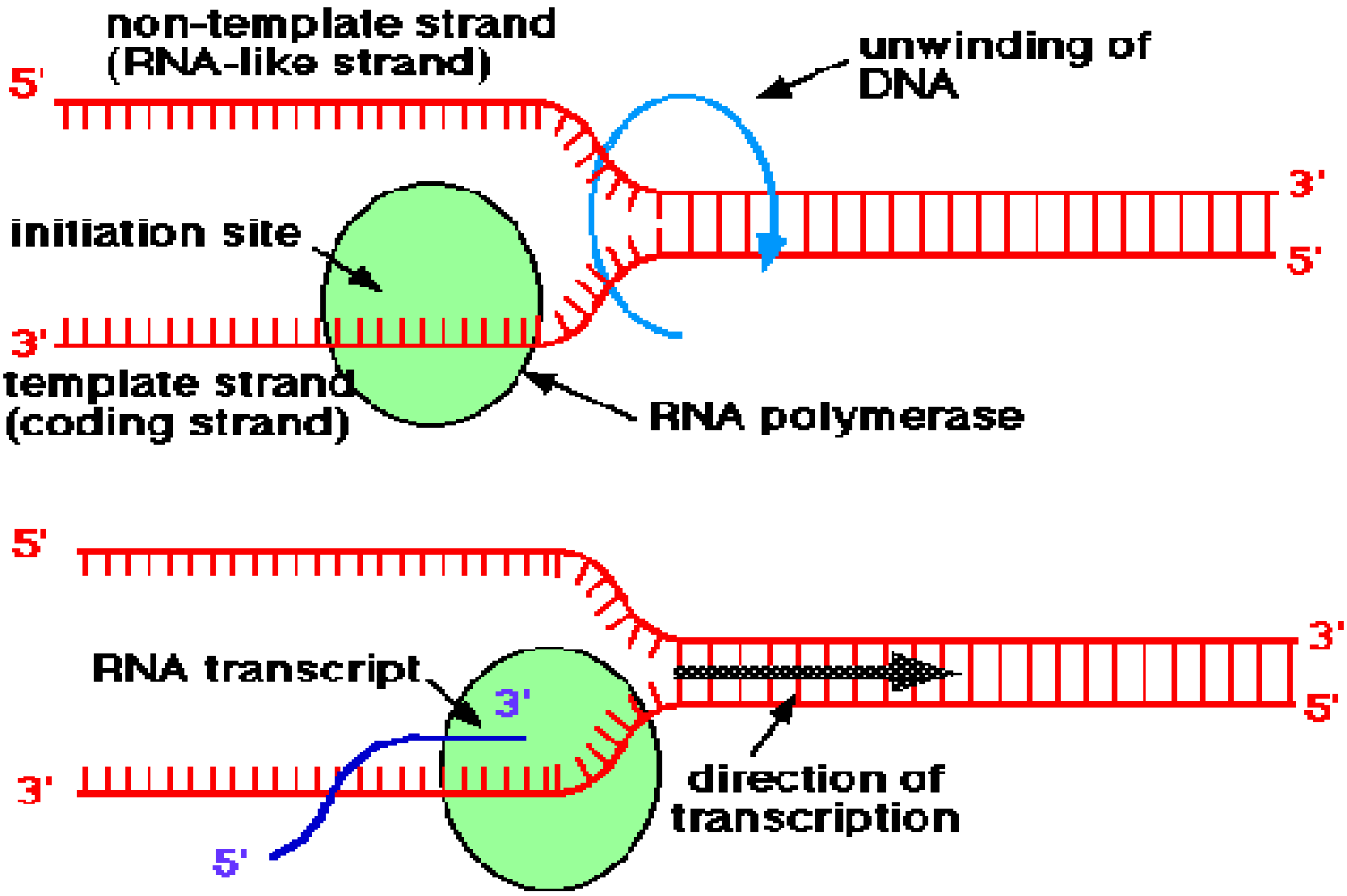
- réplication très fidèle, dite *semi-conservative*,
- cependant, quelques erreurs : mutations ponctuelles, insertions/délétions/déplacements/inversions de segments de segments (*cf.* Cours 2 : génomique comparative)

Expression de l'information génétique

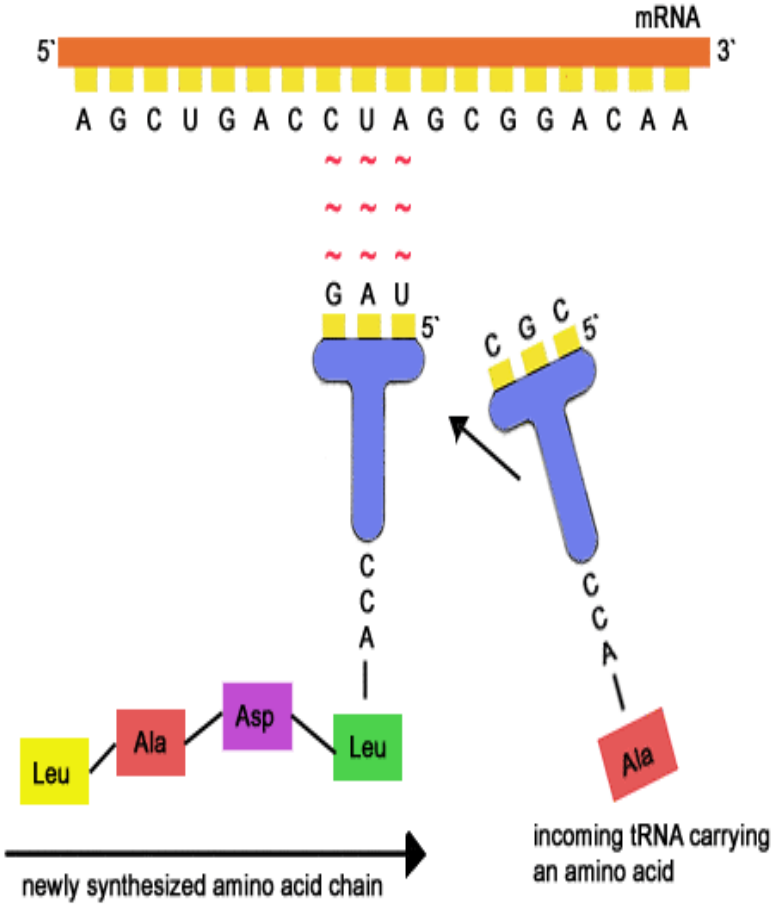
Un gène = une petite unité d'ADN, qui code une protéine.



L'expression des gènes : 1.transcription (de l'ADN en ARN)



L'expression des gènes : 2.traduction (des ARNs en protéines)



Universal Genetic Code Chart

Messenger RNA Codons and Amino Acids for Which They Code

First base	Second base U	Second base C	Second base A	Second base G	Third base
U	UUU, UUC: PHE; UUA, UUG: LEU	UCU, UCC, UCA, UCG: SER	UAU, UAC: TYR; UAA, UAG: STOP	UGU, UGC: CYS; UGA: STOP; UGG: TRP	U C A G
C	CUU, CUC, CUA, CUG: LEU	CCU, CCC, CCA, CCG: PRO	CAU, CAC: HIS; CAA, CAG: GLN	CGU, CGC, CGA, CGG: ARG	U C A G
A	AUU, AUC, AUA: ILE; AUG: MET or START	ACU, ACC, ACA, ACG: THR	AAU, AAC: ASN; AAA, AAG: LYS	AGU, AGC: SER; AGA, AGG: ARG	U C A G
G	GUU, GUC, GUA, GUG: VAL	GCU, GCC, GCA, GCG: ALA	GAU, GAC: ASP; GAA, GAG: GLU	GGU, GGC, GGA, GGG: GLY	U C A G

Les protéines sont des séquences d'acides aminés (21 lettres).

L'expression des gènes : la régulation

Quand réguler ?

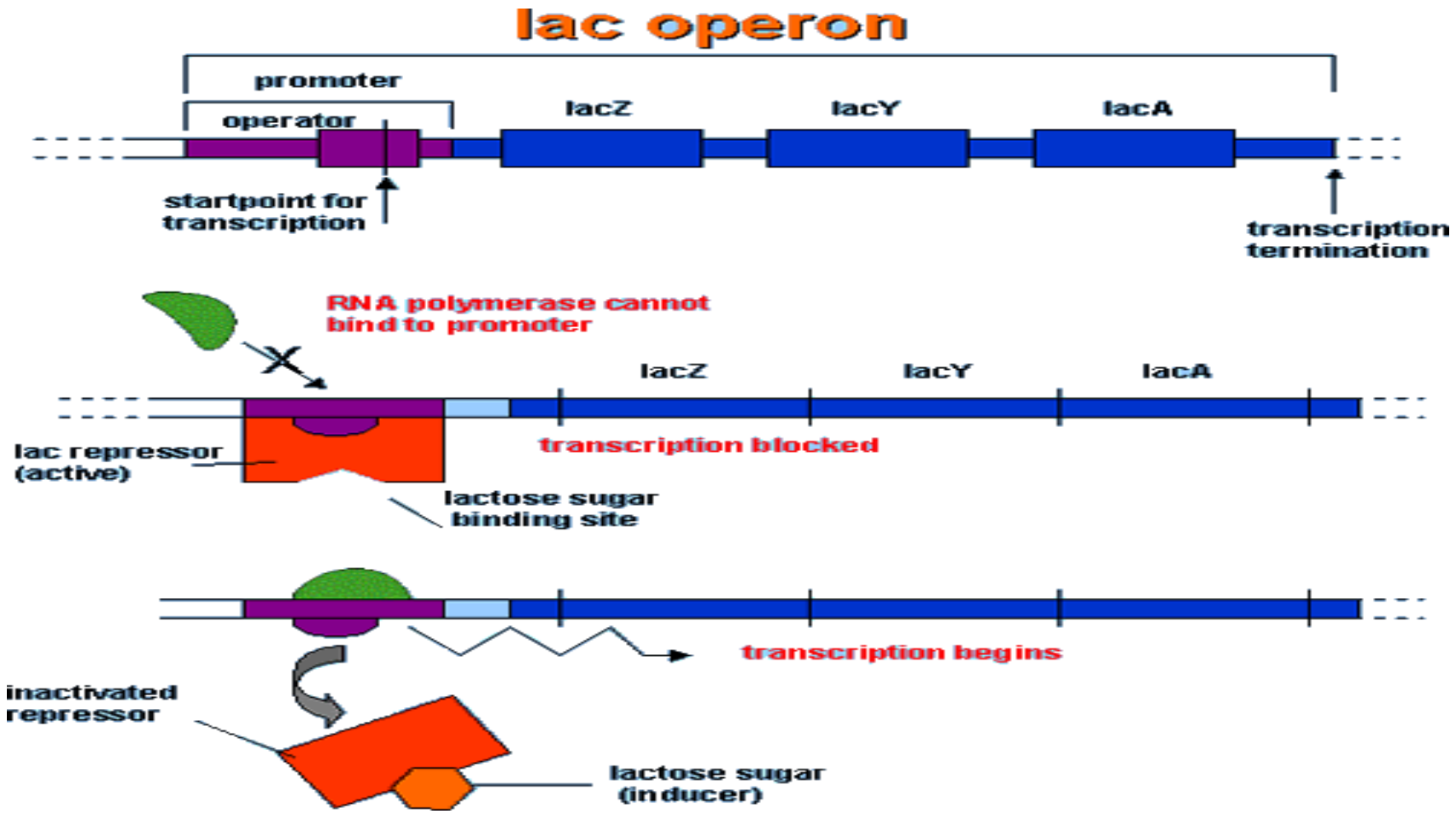
- au niveau de la transcription,
- ou post-transcriptionel (voir Épissage alternatif),
- ou modification post-translation de la protéine.

Comment réguler ?

- facteurs de transcription (initient la transcription),
- séquences promotrices en amont des gènes,
- régulation épigénétique (chez les eucaryotes) : modifications héritables de l'ADN (ex : nucléosomes, méthylation des cytosines ...)
- ...

Ex : l'operon lactose

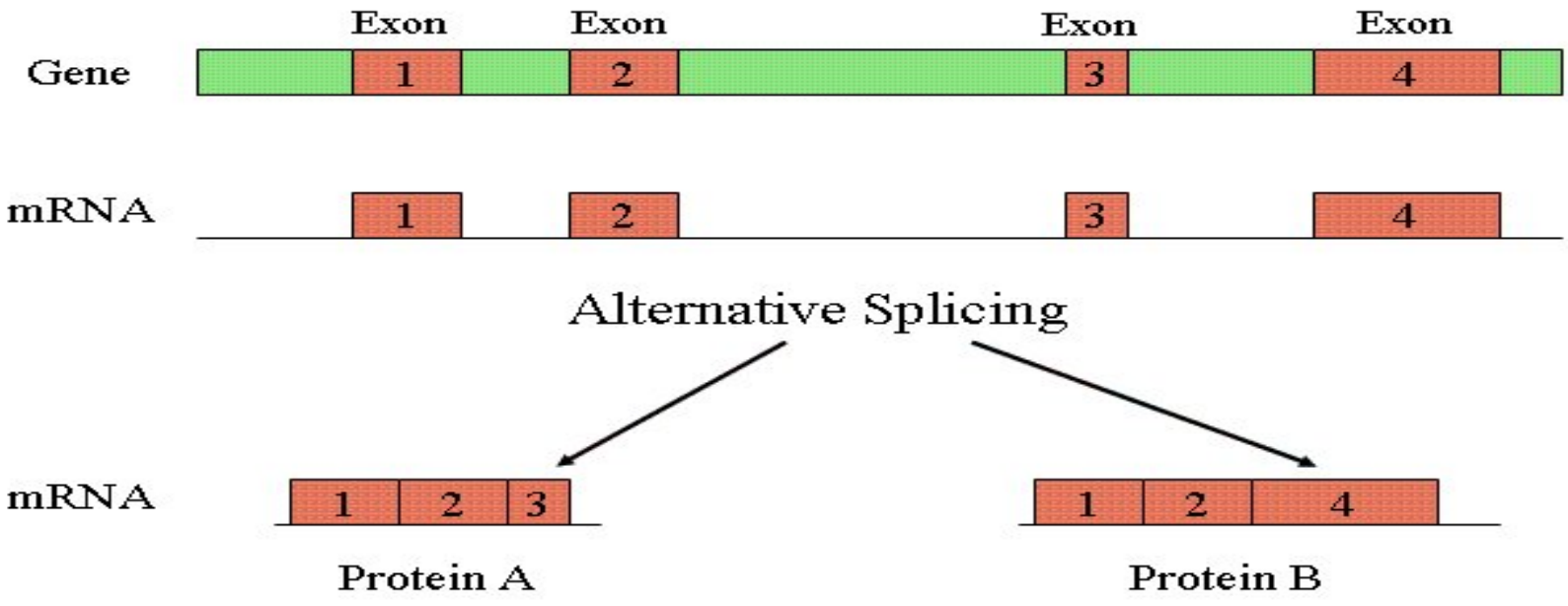
Régulation de la β-galactosidase, enzyme qui initie la chaîne du métabolisme du lactose.



L'expression des gènes : cas particulier de l'épissage alternatif

(modification post-transcriptionnelle chez les eucaryotes).

- Gènes organisés en exons (transcrits) et introns (supprimés).
- Plusieurs protéines peuvent ainsi être obtenues à partir du même gène.
- Existe aussi chez les archées, mais pas chez les bactéries.



Le génome

Le génome d'un organisme

C'est l'ensemble des molécules d'ADN et d'ARN présentes dans ses cellules :

- les chromosomes,
- les ARNs (ARNm, ARNt, . . .) de la cellule,
- plus, chez les eucaryotes, l'ADN mitochondrial.

La variation génétique

- Les individus d'une même espèce ont des génomes essentiellement identiques.
- Chaque individu possède cependant un génome unique et son génome diffère de celui d'un autre individu de la même espèce par des *variants génétiques* (*cf.* Cours sur Variants génétiques).
- Dans ce cours, nous nous plaçons au niveau des espèces. Les séquences considérées sont des séquences *consensus* de l'espèce étudiée.

Les données de séquences

- Les technologies de séquençage permettent d'obtenir la séquence *consensus* d'un organisme.
- Après séquençage, les séquences sont traitées par des outils bio-informatiques afin d'être annotées.
- Les séquences annotées sont stockées dans des bases de données (ex : EMBL Nucleotide Sequence Database, http ://www.ebi.ac.uk/embl/), régulièrement mises à jour.

Sommaire

Principes de l'annotation (de structure)

- Identifier et indiquer sur la séquence les informations des gènes : position (ORFs ou « open reading frames »), sens de lecture, structure (introns/exons), séquences promotrices, structures uniques ou en opéron, ...
- Pour cela, nécessité de mettre au point des outils d'annotation automatique, soit en utilisant uniquement l'information de la séquence, soit en la comparant à d'autres séquences déjà annotées (*cf.* Cours 2 : Génomique comparative).
- Les chaînes de Markov cachées (HMMs pour hidden Markov models) et leurs variantes, sont un des outils les plus répandus en bio-informatique pour l'analyse des séquences.

Quelques références

P. Nicolas.
Mise au point et utilisation de modèles de chaînes de Markov cachées pour l'étude des séquences d'ADN.
PhD Thesis, Université d'Évry, France, 2003.
stat.genopole.cnrs.fr/_media/publications/nicolas.pdf

T. Lionnet et V. Croquette.
Introduction à la Biologie Moléculaire.
Cours, 2005.
http://pimprenelle.lps.ens.fr/biolps/sites/default/files/teaching/4/my_biomol.pdf

Deuxième partie II

Les modèles de Markov cachés (HMMs)

Sommaire

Modèles de mélange finis

Définition

- Famille (finie) de densités $\{f_q; q \in \{1, \ldots, Q\}\}$ (par rapport à la mesure de comptage ou la mesure de Lebesgue),
- Proportions a priori $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$, telles que $\pi_q \geq 0$ et $\sum_{q=1}^{Q} \pi_q = 1$,

La loi de mélange est $\sum_{q=1}^{Q} \pi_q f_q$.

Intérêts

- Permet de modéliser l'hétérogénéité des observations : celles-ci sont issues de Q groupes homogènes et non observés,
- les paramètres π_q représentent les proportions a priori de chaque groupe,
- les paramètres f_q sont les lois a priori de chaque groupe homogène.

Modèles de mélange finis : illustration

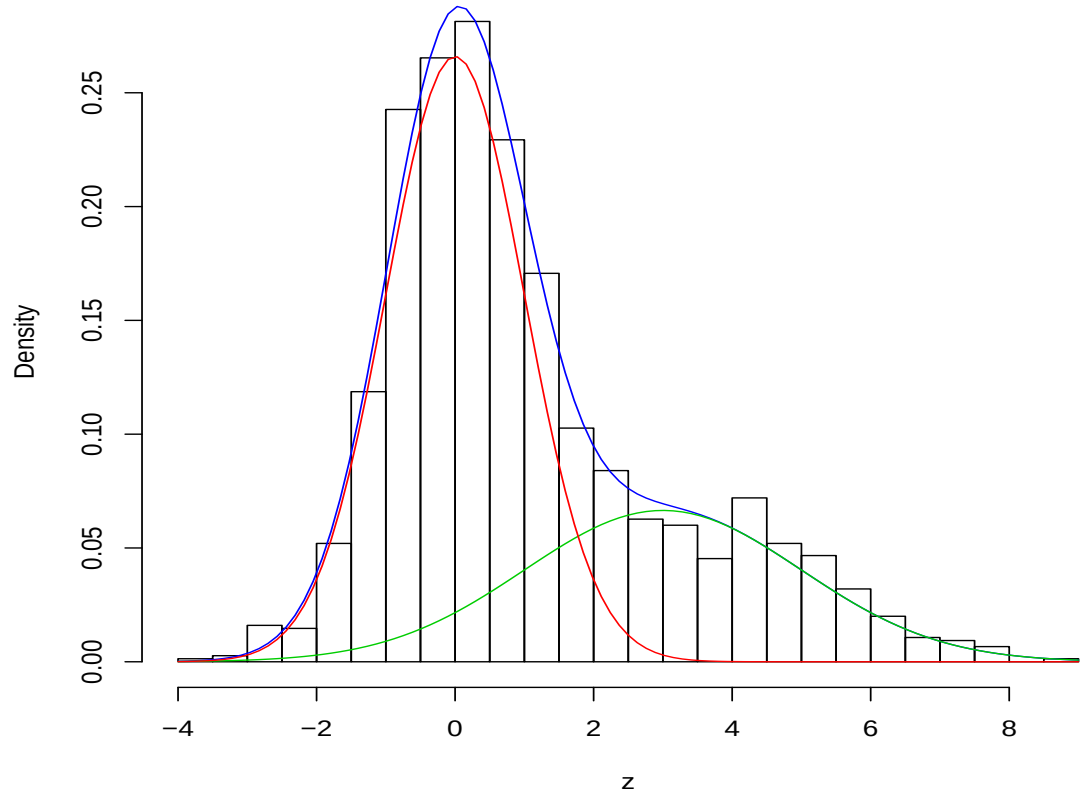


FIGURE : Histogramme d'un échantillon de taille 1500 du mélange $\frac{2}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(3,2)$. En bleu, la densité du mélange, en rouge et vert, les densités de chaque composante.

Modèles de mélange finis : vus comme modèles à variables cachées

Notations

Soit $\{S_k\}_{k\geq 1}$ une suite de v.a. i.i.d. à valeurs dans $\mathcal{S} = \{1, \ldots, Q\}$ avec $\mathbb{P}(S_k = q) = \pi_q$ et $\{X_k\}_{k\geq 1}$ t.q. cond. à $S_1, \ldots, S_n$, les observations $X_1, \ldots, X_n$ sont indépendantes, et la loi de chaque X_k ne dépend que de S_k :

$$\mathcal{L}(X_1^n | S_1^n) = \otimes_{k=1}^n \mathcal{L}(X_k | S_k), \text{ de densité } f_{S_k}.$$

Alors, les $\{X_k\}_{k\geq 1}$ sont i.i.d. de densité $\sum_{q=1}^Q \pi_q f_q$.

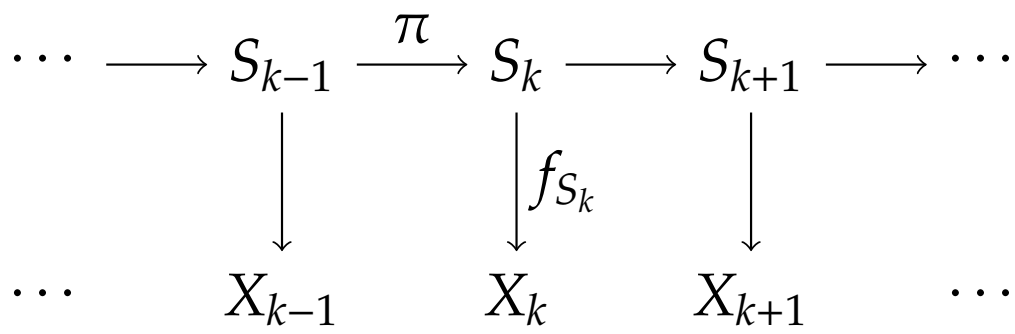
Représentation graphique

$$\begin{array}{ccccccc} \cdots & & S_{k-1} & & S_k & & S_{k+1} & & \cdots \\ & & \downarrow & & \downarrow f_{S_k} & & \downarrow & & \\ \cdots & & X_{k-1} & & X_k & & X_{k+1} & & \cdots \end{array}$$

Sommaire

Chaînes de Markov cachées (hidden Markov models, HMMs)

On introduit de la dépendance dans les états cachés



(i) $\{S_k\}$ est une chaîne de Markov, non observée, à valeurs dans un espace $\mathcal{S} = \{1, \ldots, Q\}$, de matrice de transition π et loi initiale μ. C'est la suite des régimes,

(ii) $\{X_k\}$ est la suite des observations, à valeurs dans $\mathcal{X}$,

(iii) Conditionnellement à la donnée des régimes $S_1, \ldots, S_n$, les observations $X_1, \ldots, X_n$ sont indépendantes, et la loi de chaque X_k ne dépend que de S_k :

$$\mathcal{L}(X_1^n | S_1^n) = \otimes_{k=1}^n \mathcal{L}(X_k | S_k), \text{ de densité } f_{S_k}.$$

Mélanges vs HMMs

Similarités/Différences

- Les $\{X_k\}_{k\geq 1}$ ne sont plus des v.a. indépendantes.
  En fait, les $\{X_k\}_{k\geq 1}$ ne sont même pas une chaîne de Markov ! On parle de dépendance *longue*.
- Les observations sont globalement hétérogènes, mais elles sont *ordonnées* et le modèle induit des plages de loi homogène.
- Retrouver les états cachés revient donc à segmenter la séquence observée en zones de distribution homogène.

HMMs pour analyser les séquences

Buts

- Retrouver la suite des régimes qui composent la séquence
- Pour cela, nécessité d'estimer les paramètres du modèle $\theta = (\mu, \pi, \{f_q\}_{1 \leq q \leq Q})$.

Moyens

- Estimation des paramètres : Algorithme EM
- Reconstruction des régimes : versions stochastiques de EM (SEM ou EM à la Gibbs)

Sommaire

La vraisemblance HMM

Vraisemblance des observations

Paramètre du modèle $\theta = (\mu, \pi, \{f_q\}_{1 \le q \le Q})$.

$$\ell_n(\theta) := \log \mathbb{P}_\theta(X_{1:n}) = \log \Big(\sum_{s_1, \dots, s_n} \mathbb{P}_\theta(X_{1:n}, S_{1:n} = s_{1:n}) \Big).$$

Calcul nécessite une somme sur Q^n termes, impossible numériquement si n n'est pas très petit !

Modèles à données manquantes

- Remarquer que $\{(S_k, X_k)\}_{k \ge 1}$ est une chaîne de Markov et la vraisemblance *complète* a une expression très simple
  $\log \mathbb{P}_\theta(S_{1:n}, X_{1:n}) =$
  $\log \mu(S_1) + \sum_{i=2}^n \log \pi(S_{i-1}, S_i) + \sum_{i=1}^n \log f_{S_i}(X_i)$.
- L'algorithme EM (expectation-maximization) est un algo itératif qui permet de maximiser (localement) la vraisemblance dans des modèles à données manquantes.

L'algorithme Expectation-Maximization

Modèle à données observées $X_{1:n}$, données manquantes $S_{1:n}$ et données complètes $(S_{1:n}, X_{1:n})$.

Principe

- On part d'une valeur initiale θ^0,
- À l'itération k, on effectue les deux étapes
  - Expectation on calcule $Q(\theta, \theta^k) := \mathbb{E}_{\theta^k}(\log \mathbb{P}_\theta(S_{1:n}, X_{1:n})|X_{1:n})$.
  - Maximization on maximise $\theta^{k+1} := \text{Argmax}_\theta Q(\theta, \theta^k)$.
- Arrêt lorsque $\delta := \|\theta^{k+1} - \theta^k\|/\|\theta^k\| \leq \epsilon$ ou un nb. max. d'itérations est atteint.

Conséquences

- À chaque itération, la vraisemblance observée augmente (preuve : Inégalité de Jensen).
- Avec plusieurs initialisations, on devrait atteindre le maximum global.

Algo EM : augmentation de la vraisemblance observée

Preuve

On écrit $Q(\theta^{k+1}, \theta^k) \geq Q(\theta^k, \theta^k)$, i.e :

$$
\begin{aligned}
0 &\leq \mathbb{E}_{\theta^k}\left[\log \frac{\mathbb{P}_{\theta^{k+1}}(S_{1:n}, X_{1:n})}{\mathbb{P}_{\theta^k}(S_{1:n}, X_{1:n})}\middle| X_{1:n}\right] \\
&\underset{\text{Jensen}}{\leq} \log \mathbb{E}_{\theta^k}\left[\frac{\mathbb{P}_{\theta^{k+1}}(S_{1:n}, X_{1:n})}{\mathbb{P}_{\theta^k}(S_{1:n}, X_{1:n})}\middle| X_{1:n}\right] \\
&= \log \int_{\mathcal{S}^n} \frac{\mathbb{P}_{\theta^{k+1}}(s_{1:n}, X_{1:n})}{\mathbb{P}_{\theta^k}(s_{1:n}, X_{1:n})} \mathbb{P}_{\theta^k}(s_{1:n}|X_{1:n}) ds_1 \dots ds_n \\
&= \log \int_{\mathcal{S}^n} \frac{\mathbb{P}_{\theta^{k+1}}(s_{1:n}, X_{1:n})}{\mathbb{P}_{\theta^k}(X_{1:n})} ds_1 \dots ds_n = \log \frac{\mathbb{P}_{\theta^{k+1}}(X_{1:n})}{\mathbb{P}_{\theta^k}(X_{1:n})}.
\end{aligned}
$$

Ainsi, $\mathbb{P}_{\theta^{k+1}}(X_{1:n}) \geq \mathbb{P}_{\theta^k}(X_{1:n})$.

Algo EM pour les HMMs I

Log-vraisemblance complète

$$\log \mathbb{P}_\theta(S_{1:n}, X_{1:n}) = \sum_{q=1}^{Q} 1_{S_1=q} \log \mu_q$$

$$+ \sum_{i=2}^{n} \sum_{1\le q,l\le Q} 1_{S_{i-1}=q, S_i=l} \log \pi(q,l) + \sum_{i=1}^{n} \sum_{q=1}^{Q} 1_{S_i=q} \log f_q(X_i).$$

Espérance conditionnelle aux observations sous θ^k

$$Q(\theta, \theta^k) = \sum_{q=1}^{Q} \mathbb{P}_{\theta^k}(S_1 = q | X_{1:n}) \log \mu_q$$

$$+ \sum_{i=2}^{n} \sum_{1\le q,l\le Q} \mathbb{P}_{\theta^k}(S_{i-1} = q, S_i = l | X_{1:n}) \log \pi(q,l) + \sum_{i=1}^{n} \sum_{q=1}^{Q} \mathbb{P}_{\theta^k}(S_i = q | X_{1:n}) \log f_q(X_i).$$

Algo EM pour les HMMs II

Principe de l'algo

- étape E : calculer $\mathbb{P}_{\theta^k}(S_i|X_{1:n})$ et $\mathbb{P}_{\theta^k}(S_{i-1}, S_i|X_{1:n})$, *cf.* équations forward-backward.
- étape M : immédiat (comme pour le max de vraisemblance d'une chaîne de Markov).

Étape E pour les HMMs : équations forward-backward

Équations forward : calcul de $\alpha_k(\cdot) := \mathbb{P}_\theta(S_k = \cdot, X_{1:k})$

- Initialisation $\forall q,\ \alpha_1(q) := \mathbb{P}_\theta(S_1 = q, X_1) = f_q(X_1)\mu(q)$,
- Pour $k = 2, \dots, n$, pour tout l, $\alpha_k(l) = [\sum_{q=1}^{Q} \alpha_{k-1}(q)\pi(q,l)]f_l(X_k)$.

Rem : On peut obtenir la vraisemblance des observations via $\mathbb{P}_\theta(X_{1:n}) = \sum_{q=1}^{Q} \alpha_n(q)$, mais la maximisation est alors non triviale !

Équations backward : calcul de $\beta_k(\cdot) := \mathbb{P}_\theta(X_{k+1:n}|S_k = \cdot)$

- Initialisation $\beta_n(\cdot) := 1$,
- Pour $k = n, \dots, 2$ on calcule pour tout q,
  $\beta_{k-1}(q) = \sum_{l=1}^{Q} f_l(X_k)\beta_k(l)\pi(q,l)$.

Probas a posteriori de l'étape E

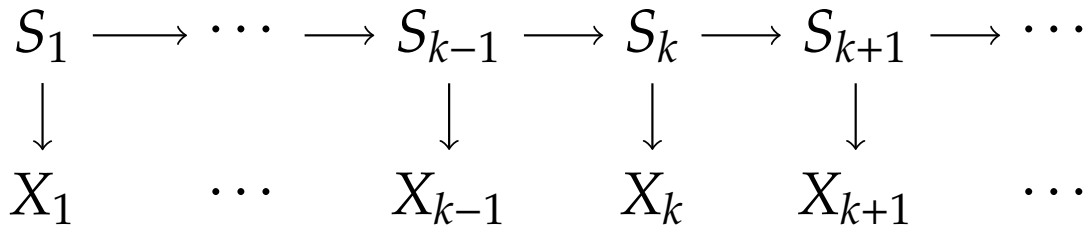
$\mathbb{P}(S_k = q|X_{1:n}) \propto \alpha_k(q)\beta_k(q)$
et $\mathbb{P}(S_{k-1} = q, S_k = l|X_{1:n}) \propto \alpha_{k-1}(q)\pi(q,l)f_l(X_k)\beta_k(l)$.

Outil : les Graphes Acycliques Dirigés (DAGs, [Lau96])

Distributions factorisées

On considère un ensemble de variables $\mathcal{V} = \{V_i\}_{1 \leq i \leq N}$ et un graphe acyclique dirigé $\mathcal{G} = (\mathcal{V}, E)$. La distribution $\mathbb{P}$ sur $\mathcal{V}$ est dite *factorisée* selon G si $\mathbb{P}(\mathcal{V}) = \mathbb{P}(V_{1:N}) = \prod_{i=1}^{N} \mathbb{P}(V_i | pa(V_i, \mathcal{G}))$, où $pa(V_i, \mathcal{G})$ désigne l'ensemble des parents de V_i dans $\mathcal{G}$.

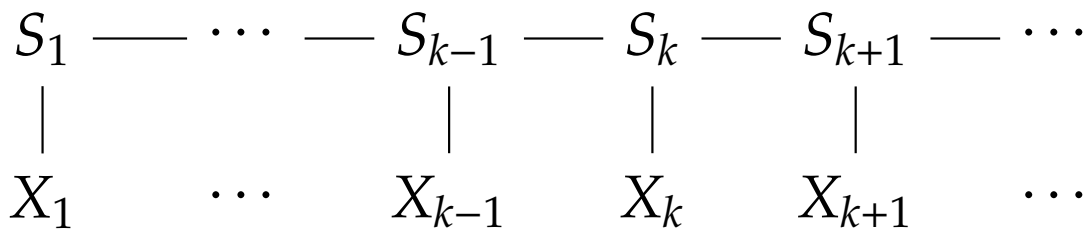
ex : HMM



Propriétés des distributions factorisées selon DAGs

Moral graph

Le *graphe moral* d'un DAG $\mathcal{G}$ est obtenu à partir de $\mathcal{G}$ en « mariant » les parents puis en retirant les directions des arêtes.
ex : Moral graph associé à un HMM



Propriétés d'indépendance

Soient I, J, K des sous ensembles de $\{1, \dots N\}$. Alors

- Dans un DAG $\mathcal{G}$, conditionnellement à ses parents, une variable est indépendante de ses non-descendants.
- Dans le moral graph associé à $\mathcal{G}$, si tous les chemins de I à J passent par K, alors $\{V_i\}_{i \in I} \perp\!\!\!\perp \{V_j\}_{j \in J} \mid \{V_k\}_{k \in K}$.

Preuves

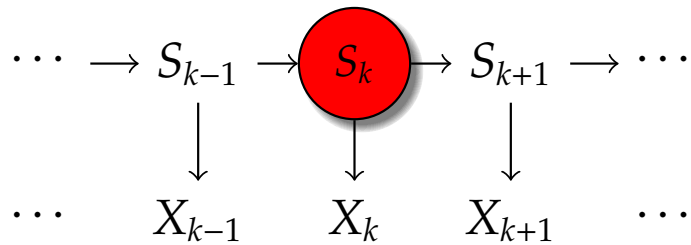
Forward equations

$$\alpha_k(l) = \mathbb{P}_\theta(S_k = l, X_{1:k}) = \sum_{q=1}^{Q} \mathbb{P}_\theta(S_{k-1} = q, S_k = l, X_{1:k})$$

$$= \sum_{q=1}^{Q} \mathbb{P}_\theta(X_k | S_{k-1} = q, S_k = l, X_{1:k-1}) \mathbb{P}_\theta(S_k = l | S_{k-1} = q, X_{1:k-1}) \mathbb{P}_\theta(S_{k-1} = q, X_{1:k-1})$$

$$= \sum_{q=1}^{Q} f_l(X_k) \pi(q, l) \alpha_{k-1}(q).$$

DAG

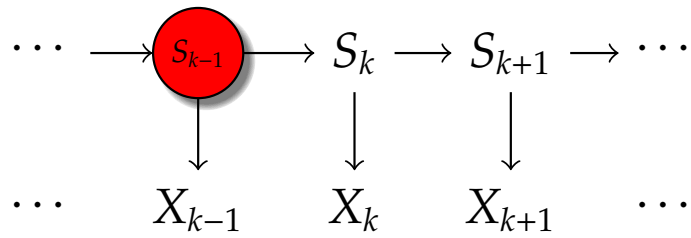


Preuves

Forward equations

$$\begin{aligned}
\alpha_k(l) &= \mathbb{P}_\theta(S_k = l, X_{1:k}) = \sum_{q=1}^{Q} \mathbb{P}_\theta(S_{k-1} = q, S_k = l, X_{1:k}) \\
&= \sum_{q=1}^{Q} \mathbb{P}_\theta(X_k | S_{k-1} = q, S_k = l, X_{1:k-1}) \mathbb{P}_\theta(S_k = l | S_{k-1} = q, X_{1:k-1}) \mathbb{P}_\theta(S_{k-1} = q, X_{1:k-1}) \\
&= \sum_{q=1}^{Q} f_l(X_k) \pi(q, l) \alpha_{k-1}(q).
\end{aligned}$$

DAG



Preuves

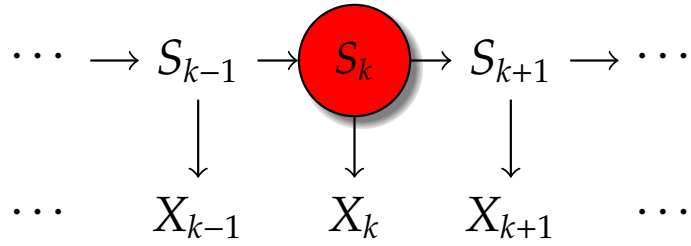
Backward equations

$$\beta_{k-1}(q) = \mathbb{P}_\theta(X_{k:n}|S_{k-1} = q) = \sum_{l=1}^{Q} \mathbb{P}_\theta(X_{k:n}, S_k = l|S_{k-1} = q)$$

$$= \sum_{l=1}^{Q} \mathbb{P}_\theta(X_k|S_k = l, S_{k-1} = q, X_{k+1:n})\mathbb{P}_\theta(X_{k+1:n}|S_{k-1} = q, S_k = l)\pi(q, l)$$

$$= \sum_{l=1}^{Q} f_l(X_k)\beta_k(l)\pi(q, l).$$

DAG



Preuves

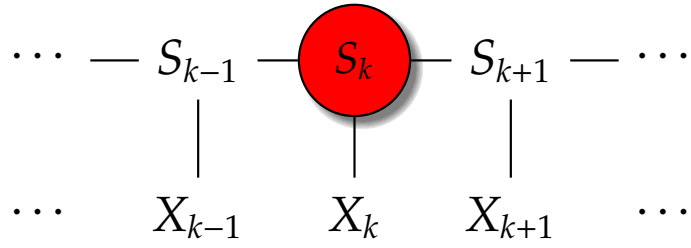
Backward equations

$$\beta_{k-1}(q) = \mathbb{P}_\theta(X_{k:n}|S_{k-1} = q) = \sum_{l=1}^{Q} \mathbb{P}_\theta(X_{k:n}, S_k = l|S_{k-1} = q)$$

$$= \sum_{l=1}^{Q} \mathbb{P}_\theta(X_k|S_k = l, S_{k-1} = q, X_{k+1:n})\mathbb{P}_\theta(X_{k+1:n}|S_{k-1} = q, S_k = l)\pi(q, l)$$

$$= \sum_{l=1}^{Q} f_l(X_k)\beta_k(l)\pi(q, l).$$

Moral graph

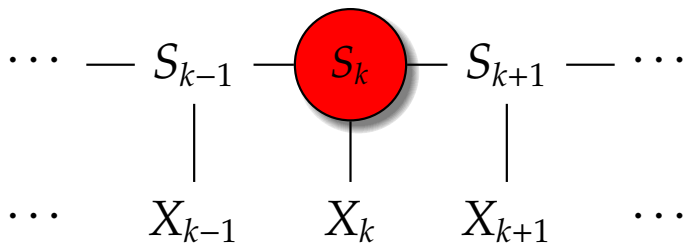


Preuves

Probas a posteriori de l'étape E (pour un seul état caché)

$$\mathbb{P}(S_k = q|X_{1:n}) \propto \mathbb{P}(S_k = q, X_{1:k}, X_{k+1:n}) = \mathbb{P}(X_{k+1:n}|S_k = q, X_{1:k})\mathbb{P}(S_k = q, X_{1:k})$$
$$\propto \beta_k(q)\alpha_k(q).$$

Moral graph

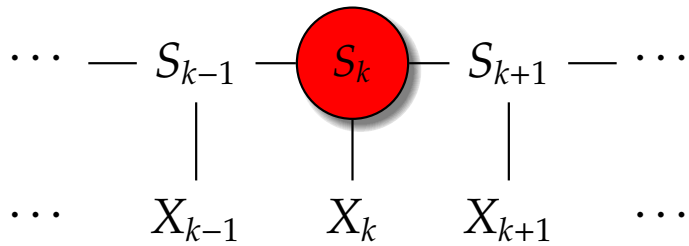


Preuves

Probas a posteriori de l'étape E (pour deux états cachés)

$$\begin{aligned}
\mathbb{P}(S_{k-1}=q, S_k=l|X_{1:n}) &\propto \mathbb{P}(X_{1:k-1}, X_k, X_{k+1:n}, S_{k-1}=q, S_k=l) \\
&\propto \mathbb{P}(X_{k+1:n}|S_{k-1}=q, S_k=l, X_{1:k-1}, X_k)\mathbb{P}(X_k|S_{k-1}=q, S_k=l, X_{1:k-1}) \\
&\quad \times \mathbb{P}(S_k=l|S_{k-1}=q, X_{1:k-1})\mathbb{P}(S_{k-1}=q, X_{1:k-1}) \\
&\propto \beta_k(l) f_l(X_k)\pi(q,l)\alpha_{k-1}(q).
\end{aligned}$$

Moral graph

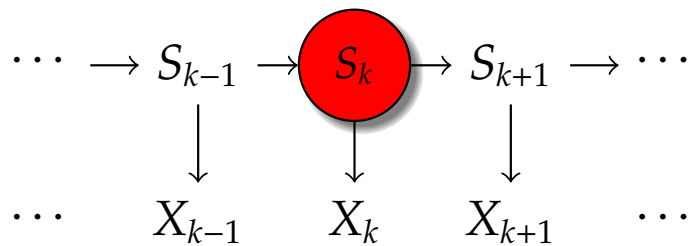


Preuves

Probas a posteriori de l'étape E (pour deux états cachés)

$$\begin{aligned}
\mathbb{P}(S_{k-1} = q, S_k = l | X_{1:n}) &\propto \mathbb{P}(X_{1:k-1}, X_k, X_{k+1:n}, S_{k-1} = q, S_k = l) \\
&\propto \mathbb{P}(X_{k+1:n} | S_{k-1} = q, S_k = l, X_{1:k-1}, X_k) \mathbb{P}(X_k | S_{k-1} = q, S_k = l, X_{1:k-1}) \\
&\quad \times \mathbb{P}(S_k = l | S_{k-1} = q, X_{1:k-1}) \mathbb{P}(S_{k-1} = q, X_{1:k-1}) \\
&\propto \beta_k(l) f_l(X_k) \pi(q, l) \alpha_{k-1}(q).
\end{aligned}$$

DAG

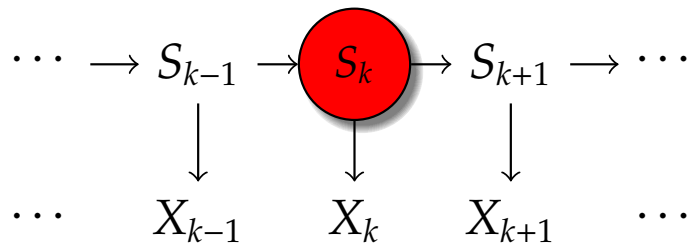


Preuves

Probas a posteriori de l'étape E (pour deux états cachés)

$$\begin{aligned}
\mathbb{P}(S_{k-1} = q, S_k = l | X_{1:n}) &\propto \mathbb{P}(X_{1:k-1}, X_k, X_{k+1:n}, S_{k-1} = q, S_k = l) \\
&\propto \mathbb{P}(X_{k+1:n} | S_{k-1} = q, S_k = l, X_{1:k-1}, X_k) \mathbb{P}(X_k | S_{k-1} = q, S_k = l, X_{1:k-1}) \\
&\quad \times \mathbb{P}(S_k = l | S_{k-1} = q, X_{1:k-1}) \mathbb{P}(S_{k-1} = q, X_{1:k-1}) \\
&\propto \beta_k(l) f_l(X_k) \pi(q, l) \alpha_{k-1}(q).
\end{aligned}$$

DAG



Étape M pour les HMMs

$$\begin{aligned}
\theta^{k+1} &= \text{Argmax}_\theta Q(\theta, \theta^k) \\
&= \text{Argmax}_\theta \Big\{ \sum_{q=1}^{Q} \mathbb{P}_{\theta^k}(S_1 = q | X_{1:n}) \log \mu_q \\
&\quad + \sum_{i=2}^{n} \sum_{1 \le q,l \le Q} \mathbb{P}_{\theta^k}(S_{i-1} = q, S_i = l | X_{1:n}) \log \pi(q,l) \\
&\quad + \sum_{i=1}^{n} \sum_{q=1}^{Q} \sum_{x \in \mathcal{X}} \mathbb{P}_{\theta^k}(S_i = q | X_{1:n}) 1_{X_i = x} \log f_q(x) \Big\}.
\end{aligned}$$

Une maximisation sous contraintes de $(\pi, \{f_q\}_{1 \le q \le Q})$ donne

$$\begin{aligned}
\pi(q,l)^{k+1} &\propto \sum_{i=2}^{n} \mathbb{P}_{\theta^k}(S_{i-1} = q, S_i = l | X_{1:n}) \\
f_q^{k+1}(x) &\propto \sum_{i=1}^{n} \mathbb{P}_{\theta^k}(S_i = q | X_{1:n}) 1_{X_i = x}.
\end{aligned}$$

Et en régime stationnaire, on peut estimer μ par

$$\mu^{k+1}(q) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{\theta^k}(S_i = q | X_{1:n}).$$

Algo EM et initialisation multiples

- En pratique, on lance plusieurs fois l'algorithme, avec des initialisations différentes,
- À la fin de chaque algo EM, on peut calculer la log-vraisemblance obtenue

$$\ell_n(\hat{\theta}) := \log \mathbb{P}_{\hat{\theta}}(X_{1:n}) = \sum_{l=1}^{Q} \hat{f}_l(X_1)\hat{\beta}_1(l)\mathbb{P}_{\hat{\theta}}(S_1 = l)$$

- On sélectionne le paramètre correspondant à la meilleure log-vraisemblance finale.

Sommaire

Estimation des régimes

Viterbi algorithm

- The most popular method. It consists in finding the maximum a posteriori path

$$\hat{S}_{1:n} = \text{Argmax}_{s_{1:n} \in \mathcal{S}^n} \mathbb{P}_{\hat{\theta}}(X_{1:n}, S_{1:n} = s_{1:n}), \tag{1}$$

  where $\hat{\theta}$ is the solution of EM-algorithm.
- Viterbi is an exact recursive algorithm for solving (1).
- Main drawback : unstable w.r.t. sequence length. E.g. remove the last observation, then $\hat{S}_{1:n}$ is completely changed.

Alternative solution

At the end of EM algorithm, one has access to $\hat{\mathbb{P}}(S_k = q|X_{1:n}) \propto \hat{\alpha}_k(q)\hat{\beta}_k(q)$. Thus, one may consider $\hat{S}_k = \text{Argmax}_{1 \leq q \leq Q} \hat{\mathbb{P}}(S_k = q|X_{1:n})$

Variante à EM

Algorithme SEM (stochastic EM)

Variante de EM, chaque itération contient 3 étapes

- Étape E : calcul de la loi jointe des $\{S_i\}_{i\geq 1}$ conditionnelle aux obs. $\{X_i\}_{i\geq 1}$, sous le param. courant θ^k, *cf.* équations forward-backward.
- Étape S : simule des variables cachées $s^k_{1:n}$ de façon indépendante, selon la loi $s_i \sim \mathbb{P}_{\theta^k}(S_i = \cdot | X_{1:n})$
- Étape M : $\theta^{k+1} = \text{Argmax}_\theta \log \mathbb{P}_\theta(S_{1:n} = s^k_{1:n}, X_{1:n})$

Conséquences

À la cv de l'algo, on a un estimateur de $\mathbb{P}(S_k = q | X_{1:n})$: on peut soit prendre le MAP (maximum a posteriori), soit simuler sous cette loi, ou garder la loi telle quelle.

Choix du nombre d'états cachés (sélection de modèles)

- Peut-être imposé par le problème. Ex : détection de gènes chez les bactéries, $Q = 2$ pour régimes codant / non codant.
- En général, sélectionner Q est un problème difficile. Pour les HMMs, on peut utiliser un critère BIC

$$\hat{Q} = \text{Argmin}_Q \Big\{ -\log \mathbb{P}_{\hat{\theta},Q}(X_{1:n}) + \frac{N_Q}{2} \log n \Big\},$$

où $N_Q = Q(Q-1) + Q(|\mathcal{X}| - 1)$ nombre de paramètres du modèle et $\mathbb{P}_{\hat{\theta},Q}(X_{1:n})$ est obtenu après cv de l'algo EM, via équations forward.

Troisième partie III

Exemples d'application

Détection de gènes

ex : Chez *B. subtilis*, [Nic03, NBM02].

Principe

Les séquences codantes suivent une distribution des lettres différente des séquences non codantes : une segmentation en deux régimes (codant / non codant) devrait permettre de détecter les gènes sur la séquence.

Raffinement du HMM

En pratique, on utilise des modèles de Markov à régimes markoviens, *i.e*, conditionnellement à $\{S_i\}_{i\geq 1}$, la suite des observations $\{X_i\}_{i\geq 1}$ est une chaîne de Markov d'ordre k, la loi de chaque X_i dépend de S_i et $X_{i-k:i-1}$.
Ex : $k = 1$

$$\begin{array}{ccccccccc} \cdots & \longrightarrow & S_{k-1} & \longrightarrow & S_k & \longrightarrow & S_{k+1} & \longrightarrow & \cdots \\ & & \downarrow & & \downarrow & & \downarrow & & \\ \cdots & \longrightarrow & X_{k-1} & \longrightarrow & X_k & \longrightarrow & X_{k+1} & \longrightarrow & \cdots \end{array}$$

Détection de gènes (chez *B. subtilis*, [Nic03])

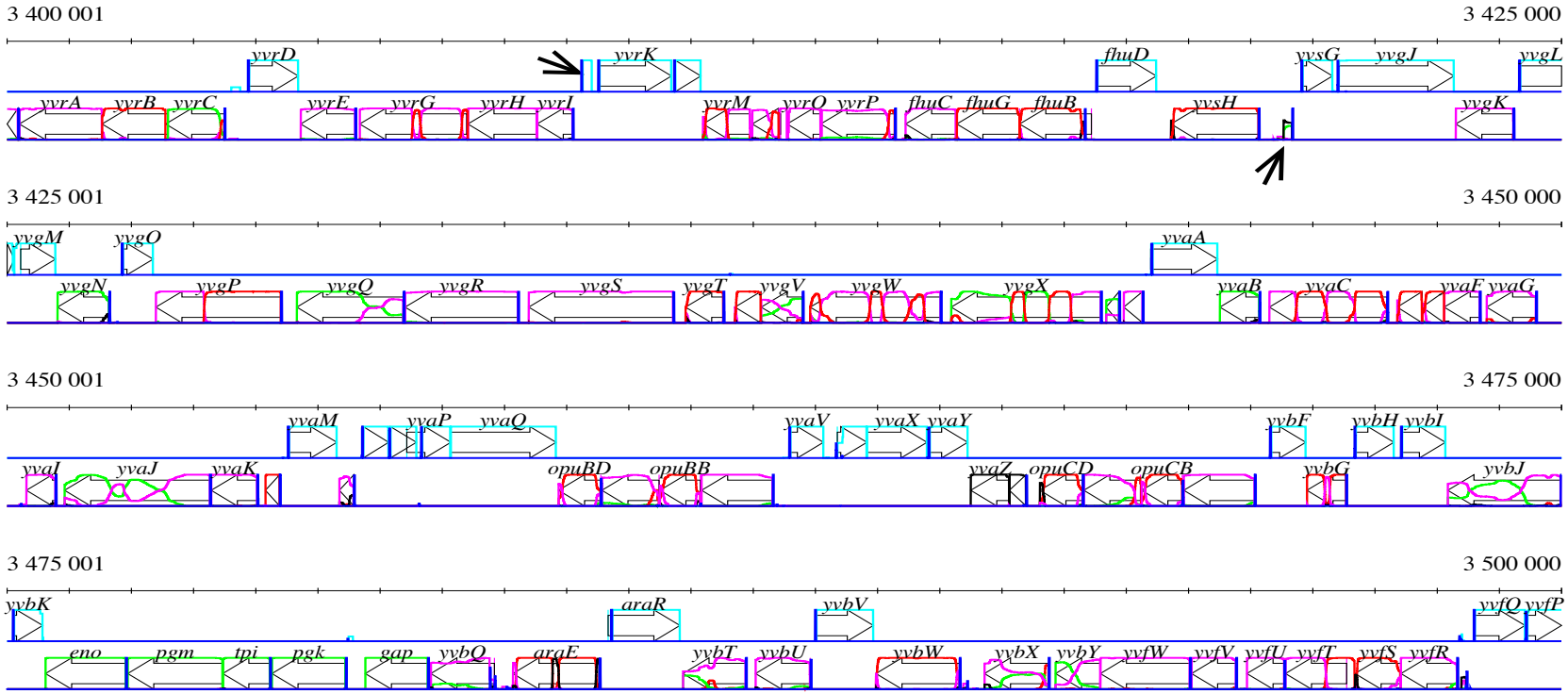
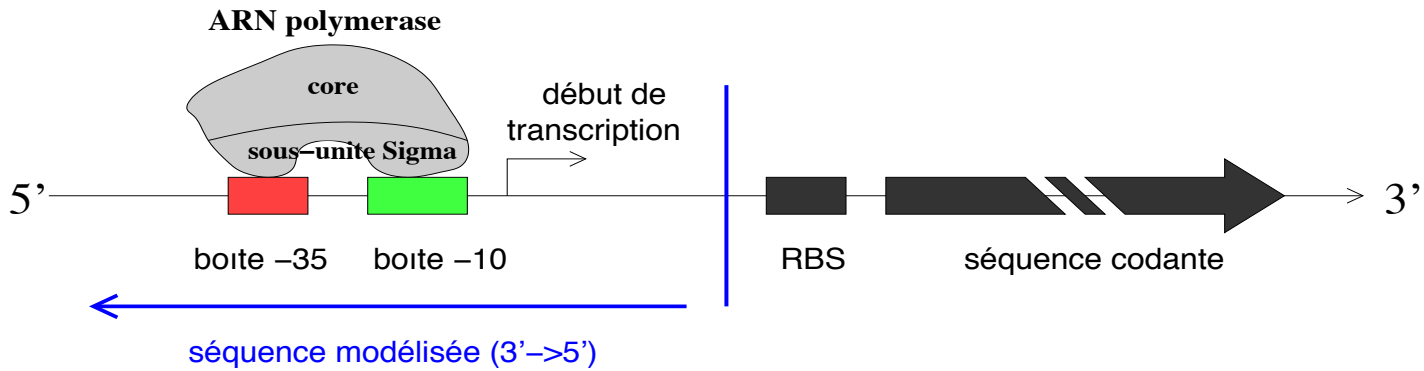


FIGURE : Segmentation d'une portion du génome de *B. subtilis* avec 5 états cachés [Nic03]. Les probas a posteriori de chaque état caché sont proches de 0 ou 1. Les annotations de GenBank sont superposées sur la séquence.

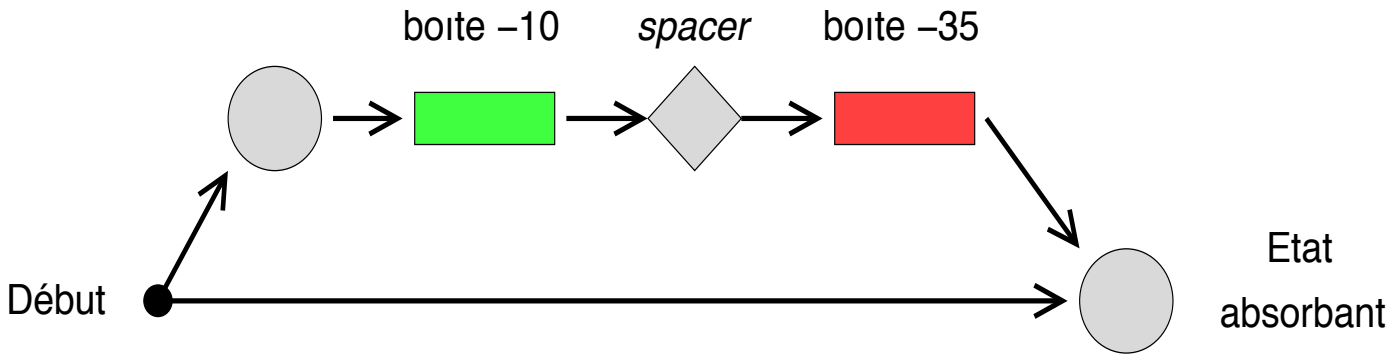
Détection de motifs ([Nic03])

Ex : sites promoteurs



Principe

- Contraindre le HMM pour forcer la détection de structure,
- Utiliser des semi-Markov cachés (HSMM) qui généralisent les HMMs au cas où les temps de séjour dans un état ne suivent pas une loi géométrique.



Détection de motifs ([Nic03])

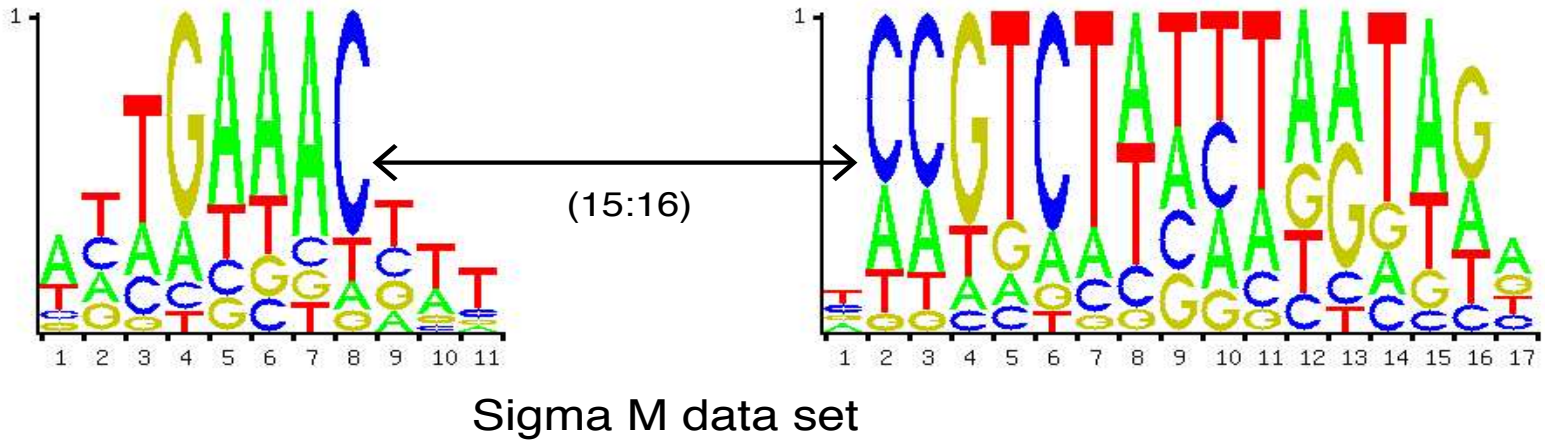


FIGURE : Exemple de motifs promoteurs estimés sur un jeu de données issu de *B. subtilis*. La taille de chaque lettre est proportionnelle à sa proba a posteriori. À gauche, la boîte « -35 » et à droite, la boîte « -10 ».

Références

S. L. Lauritzen.
Graphical models, volume 17 of *Oxford Statistical Science Series*.
Oxford University Press, New York, 1996.

P. Nicolas, L. Bize, F. Muri, M. Hoebeke, F. Rodolphe, S.D. Ehrlich, B. Prum, and P. Bessières.
Mining *Bacillus subtilis* chromosome heterogeneities using hidden Markov models.
Nucleic Acids Res., 30(6) :1418–1426, 2002.

P. Nicolas.
Mise au point et utilisation de modèles de chaînes de Markov cachées pour l'étude des séquences d'ADN.
PhD Thesis, Université d'Évry, France, 2003.
`stat.genopole.cnrs.fr/_media/publications/nicolas.pdf`