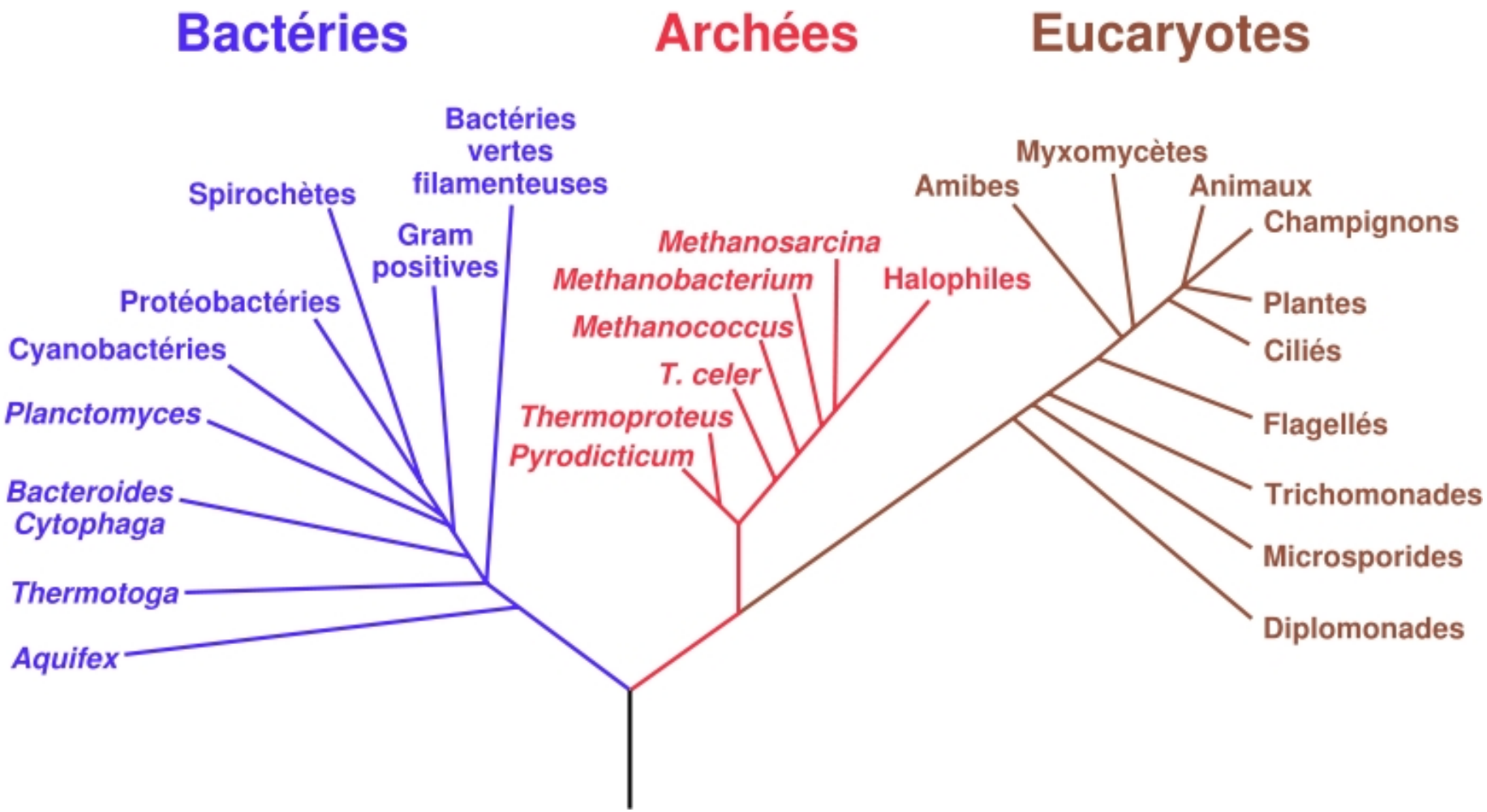
Un modèle d'évolution des séquences d'ADN avec dépendance au voisin de gauche

Audrey Finkler

Colloque Jeunes Probabilistes et Statisticiens, Mont-Dore

3 mai 2010

L'arbre phylogénétique des vivants

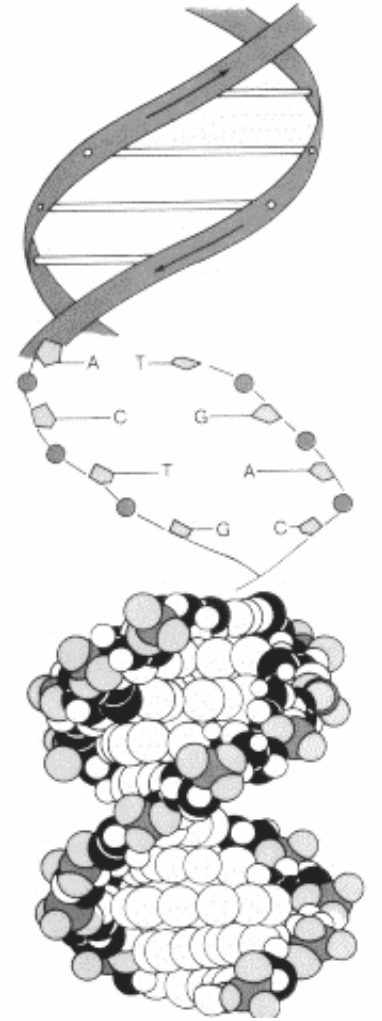


Distance évolutive : $K = kT$.

Plan

Plan

Quelques notions de biologie moléculaire



- Séquence d'ADN :

$$\mathbf{X} = (X_1, \ldots, X_n),$$

suite finie de variables aléatoires avec $X_i \in \mathcal{A} = \{A, C, G, T\}$.

- Complémentarité des bases :

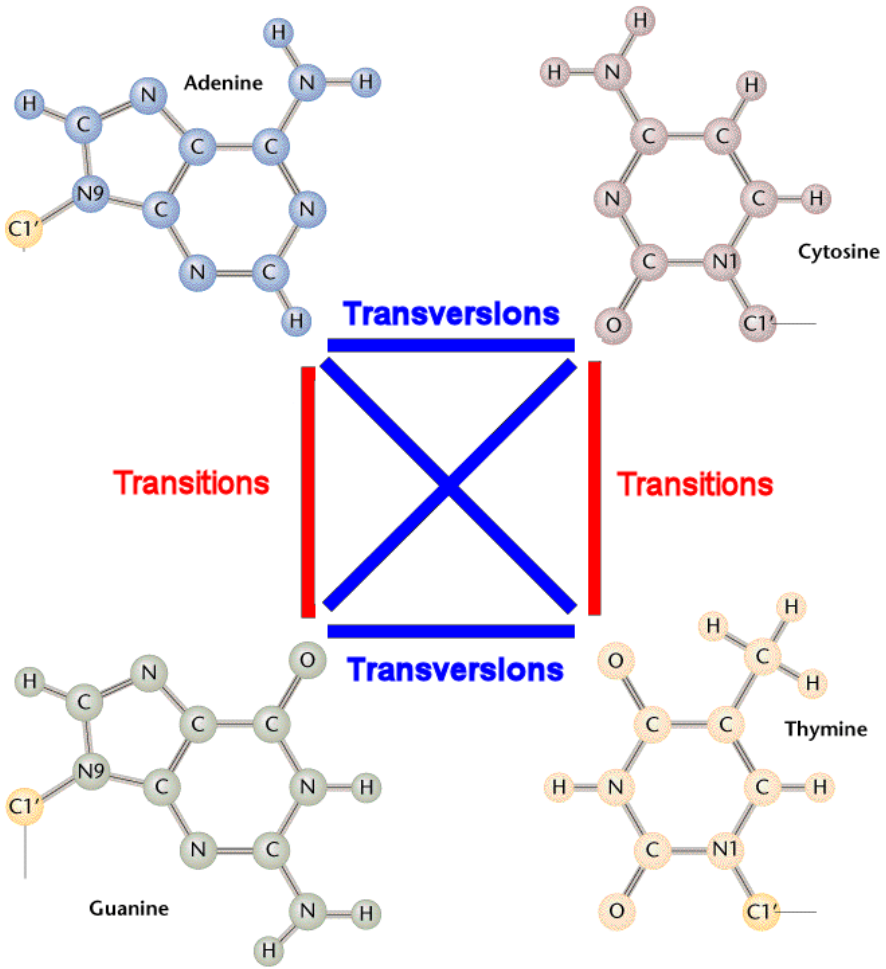
$$A - T \quad \text{et} \quad C - G$$

- Purines : $\mathcal{R} = \{A, G\}$.
Pyrimidines : $\mathcal{Y} = \{C, T\}$.

Processus de substitution markovien à temps continu :

- substitution : mutation ponctuelle sans insertion ni délétion,
- temps continu,
- markovien d'ordre 1,
- homogène, uniforme et irréductible,
- sites indépendants.

Transitions et transversions



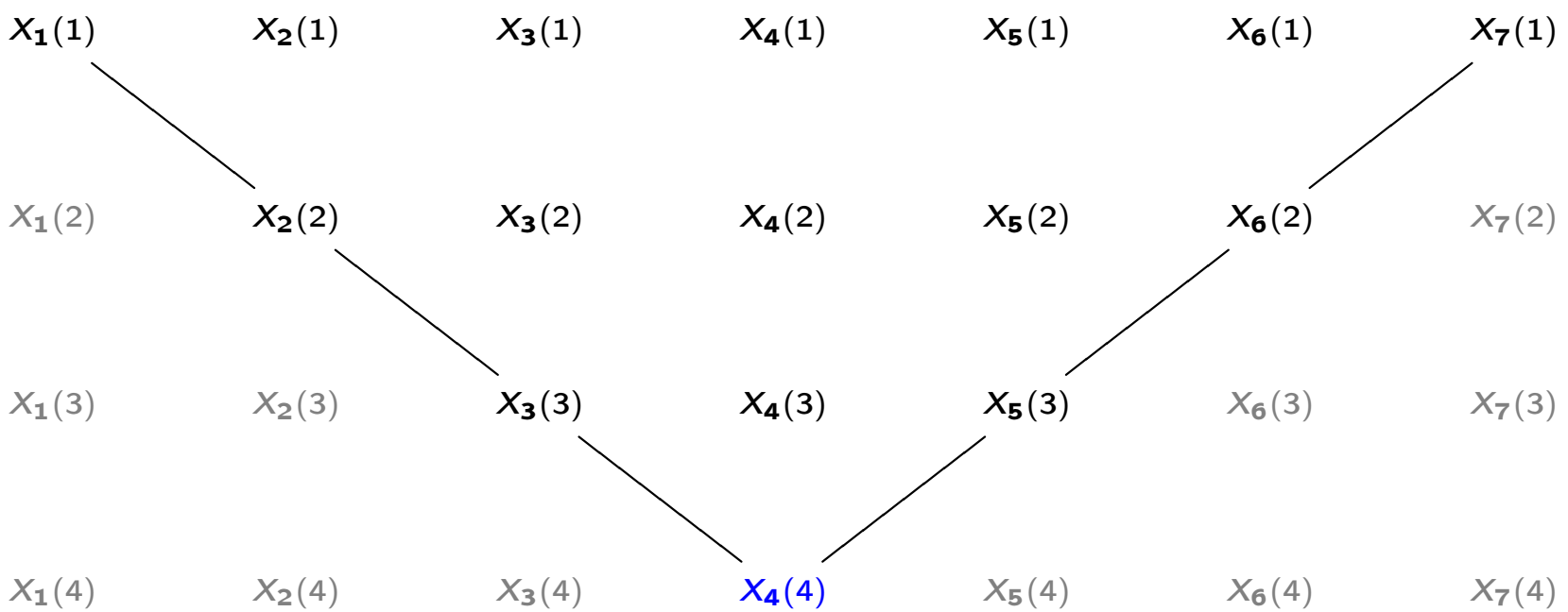
- Histoire du site i :
  $X_i := (X_i(t))_{t \in [0,T]}$.
- Transition (ts) : $\mathcal{R} \longrightarrow \mathcal{R}$ ou $\mathcal{Y} \longrightarrow \mathcal{Y}$.
  Transversion (tv) : $\mathcal{R} \longrightarrow \mathcal{Y}$ ou $\mathcal{Y} \longrightarrow \mathcal{R}$.
- $1 < \dfrac{ts}{tv} < 10$.

Effet *CpG*

$$\overrightarrow{CG} \;\xrightarrow{\text{mute en}}\; \overrightarrow{TG} \qquad \text{ou} \qquad \overrightarrow{CG} \;\xrightarrow{\text{mute en}}\; \overrightarrow{CA}$$
$$\underleftarrow{GC} \qquad\qquad \underleftarrow{AC} \qquad\qquad \underleftarrow{GC} \qquad\qquad \underleftarrow{GT}$$

Substitutions à partir d'un dinucléotide *CpG* sur un brin et son complémentaire.

Cône de dépendances



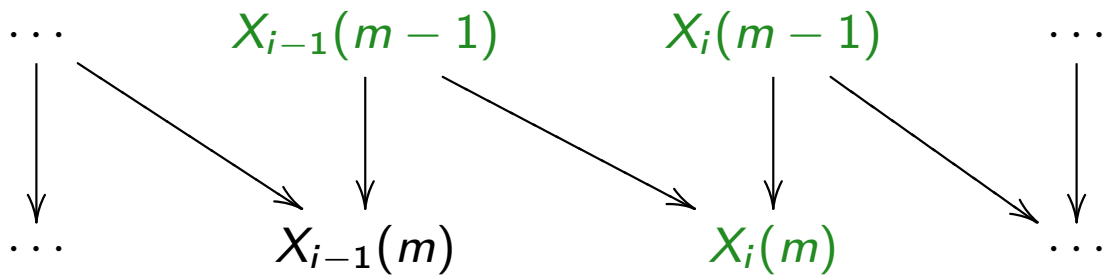
Plan

Modèle

Temps discrétisé :

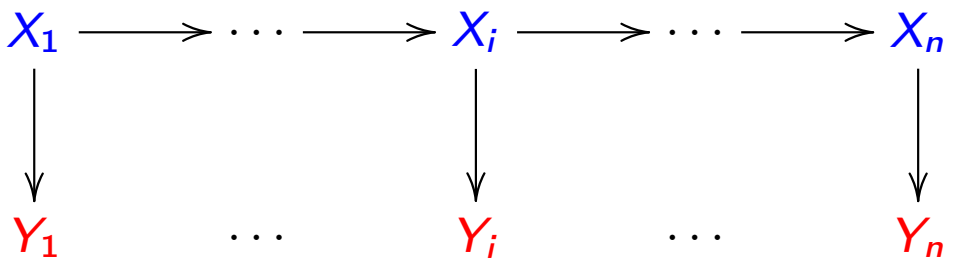
$$\begin{array}{cccc} X_1(1) & X_2(1) & \dots & X_n(1) \\ X_1(2) & X_2(2) & \dots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(M-1) & X_2(M-1) & \dots & X_n(M-1) \\ X_1(M) & X_2(M) & \dots & X_n(M) \end{array} \quad \downarrow \text{ temps}$$

Graphe des dépendances du modèle :



Modèle vu comme une chaîne de Markov cachée (HMM)

Deux processus emboîtés :



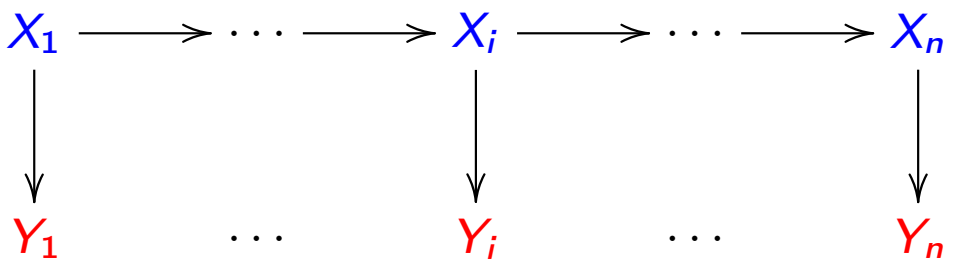
Pour $i \in \{1, \ldots, n\}$,

- états cachés $X_i = (X_i(m))_{1 \leqslant m \leqslant M}$,
- états observés $Y_i = \{X_i(1), X_i(M)\}$.

$$\begin{array}{cccc} X_1(1) & X_2(1) & \ldots & X_n(1) \\ X_1(2) & X_2(2) & \ldots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(M-1) & X_2(M-1) & \ldots & X_n(M-1) \\ X_1(M) & X_2(M) & \ldots & X_n(M) \end{array} \quad \Big\downarrow \text{ temps}$$

Modèle vu comme une chaîne de Markov cachée (HMM)

Deux processus emboîtés :



Pour $i \in \{1, \ldots, n\}$,

- états cachés $X_i = (X_i(m))_{1 \leqslant m \leqslant M}$,
- états observés $Y_i = \{X_i(1), X_i(M)\}$.

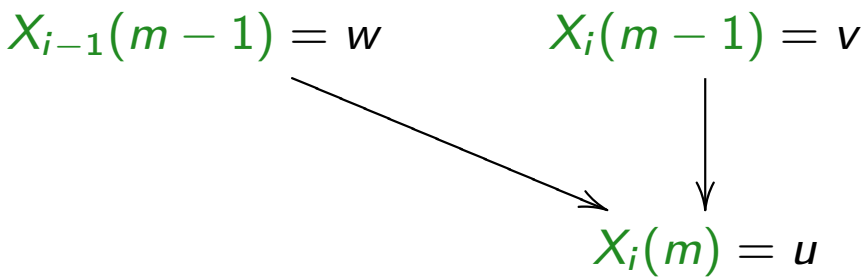
$$\begin{array}{cccc} X_1(1) & X_2(1) & \ldots & X_n(1) \\ X_1(2) & X_2(2) & \ldots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(M-1) & X_2(M-1) & \ldots & X_n(M-1) \\ X_1(M) & X_2(M) & \ldots & X_n(M) \end{array} \quad \Big\downarrow \text{ temps}$$

Paramètres

- processus observé :

$$\mathbb{P}(Y_i = y | X_i = U) = \mathbb{1}_{((u_1, u_M)=(y_1, y_2))},$$

- processus caché : chaîne de Markov de matrice $D(U, V) = \mathbb{P}(X_{i+1} = V | X_i = U)$.



$$c(u; v, w) = \mathbb{P}(X_i(m) = u | X_i(m-1) = v, X_{i-1}(m-1) = w),$$

avec $\sum_{u \in \mathcal{A}} c(u; v, w) = 1, \quad \forall (v, w) \in \mathcal{A}^2$ et

$$D(U, V) = \pi_{v_1} \prod_{m=2}^{M} c(v_m; v_{m-1}, u_{m-1}).$$

Paramètres à estimer

Posons :

$$c_\theta(u; v, w) = \begin{cases} \mu \cdot \tau^{\epsilon(u,v)} \cdot \left(\frac{1-\tau}{2}\right)^{1-\epsilon(u,v)} \cdot \gamma^{\kappa(w,v)} & \text{si } v \neq u, \\ 1 - \mu \cdot \gamma^{\kappa(w,v)} & \text{si } v = u, \end{cases}$$

où $\theta = (\mu, \tau, \gamma) \in [0,1]^3$ avec

- μ : probabilité pour un site de subir une substitution,
- τ : facteur multiplicatif Ts/Tv, avec $\epsilon(u, v)$ indicatrice d'une transition entre u et v,
- γ : facteur multiplicatif prenant en compte le contexte, $\kappa(w, v)$ indiquant un contexte différent de *CpG* ou *TpA*.

Estimation de θ par l'algorithme EM-HMM

$$Q(\theta, \theta_n^{(k)}) = q + \sum_{i=2}^{n} \sum_{U \in \mathcal{A}^M} \sum_{V \in \mathcal{A}^M} \mathbb{P}_k(X_{i-1} = U, X_i = V | Y) \ln D(U, V).$$

On approche itérativement l'estimateur du maximum de vraisemblance de θ par l'algorithme EM appliqué aux HMM. Le paramètre est initialisé à $\theta_n^{(0)}$ puis on alterne les étapes suivantes :

E : calculer $Q(\theta, \theta_n^{(k)})$,

M : poser $\theta_n^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta_n^{(k)})$.

Plan

Simulations

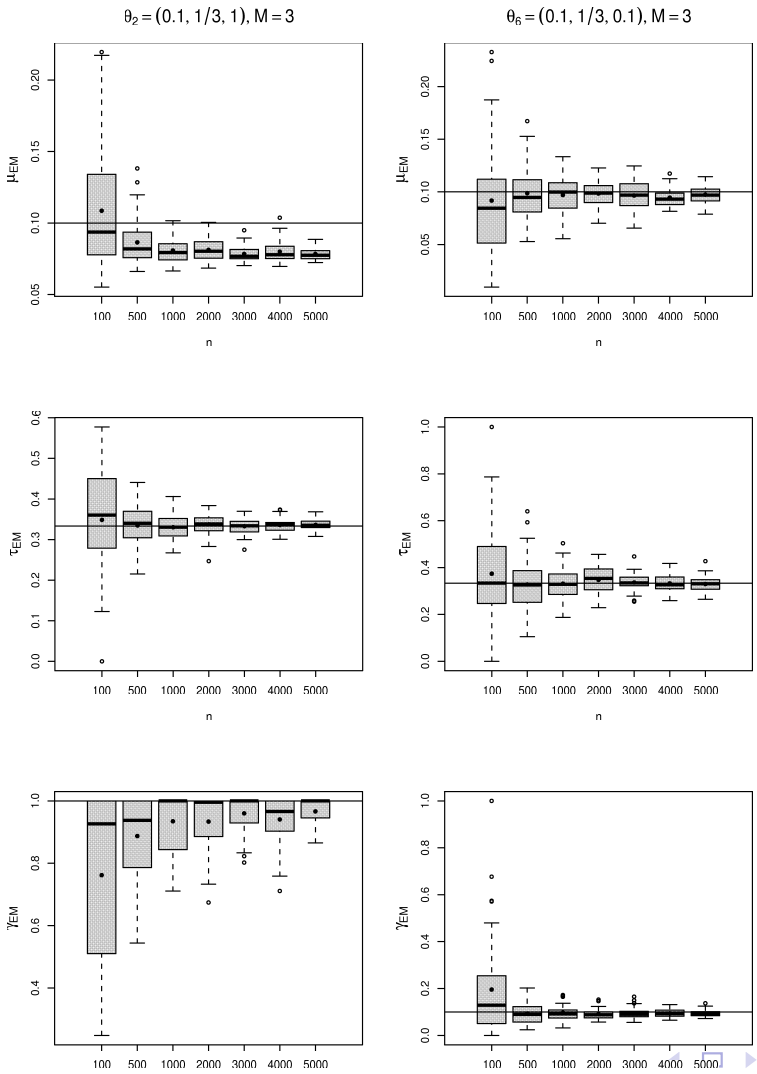
Pour $n \in \{100, 500, 1\,000, 2\,000, 3\,000, 4\,000, 5\,000\}$:

- 50 répétitions pour $M = 3, 4$,
- 25 répétitions pour $M = 5$,

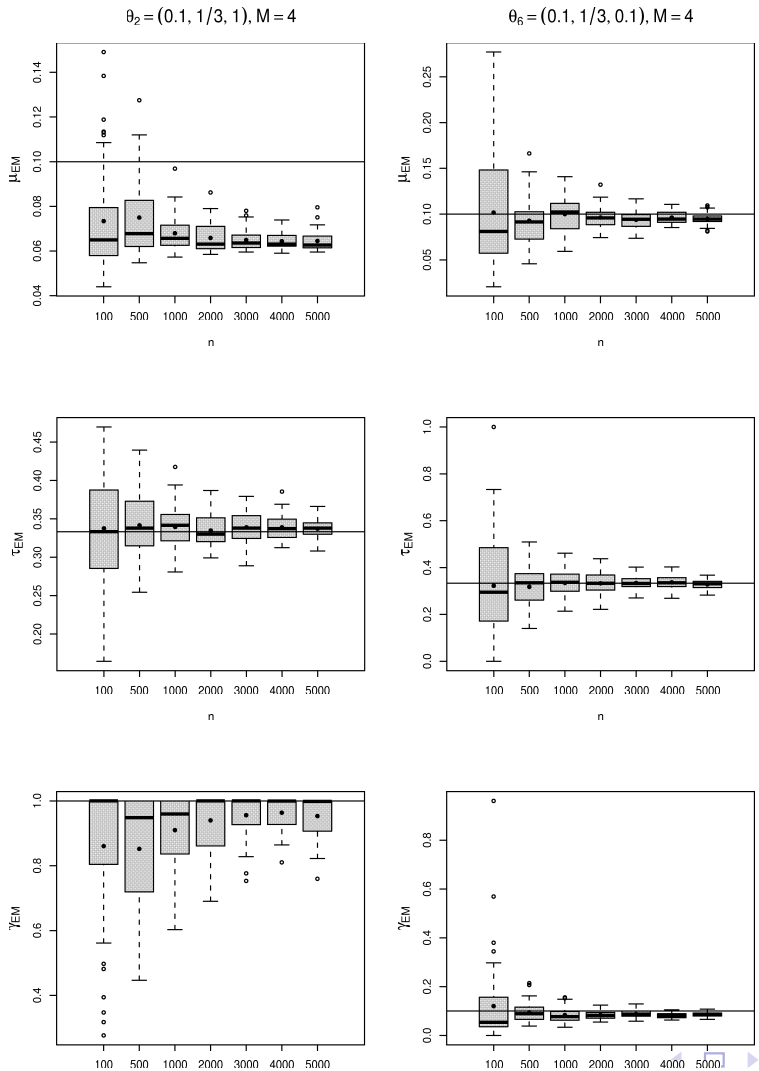
pour chacune des sept valeurs de θ :

	μ	τ	γ
θ_1	0.01	1/3	1
θ_2	0.1	1/3	1
θ_3	0.1	0.5	1
θ_4	0.1	0.8	1
θ_5	0.1	1/3	1/3
θ_6	0.1	1/3	0.1
θ_7	0.01	0.8	1/3

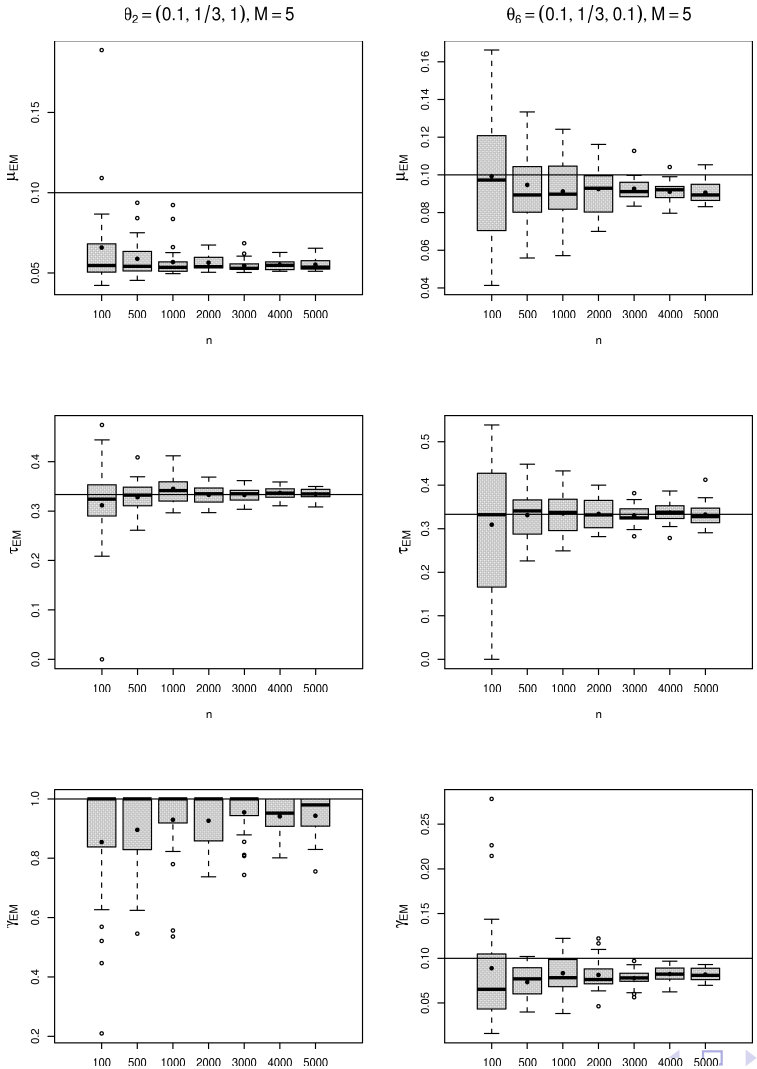
Résultats pour $M = 3$



Résultats pour $M = 4$



Résultats pour $M = 5$



Application à des données réelles

Le taux global de substitution k_{EM} est tel que $\mu_{\mathrm{EM}} = \dfrac{k_{\mathrm{EM}}\, T}{M-1}$.

- séquence codante de l'insuline : Souris et Rat pour $n = 327$ et $M = 5$:

$$\mu_{\mathrm{EM}} = 0.060, \quad \tau_{\mathrm{EM}} = 0.717, \quad \gamma_{\mathrm{EM}} = 0.241.$$

  Durée depuis divergence : 41 Ma, soit $k_{\mathrm{EM}} = 0.3\%$ par site et Ma.

- séquence non-codante de la bêta-globine : Homme et Chimpanzé pour $n = 2\ 147$ et $M = 5$:

$$\mu_{\mathrm{EM}} = 0.004, \quad \tau_{\mathrm{EM}} = 0.634, \quad \gamma_{\mathrm{EM}} = 0.839.$$

  Durée depuis divergence : 4.7 Ma, soit $k_{\mathrm{EM}} = 0.17\%$ par site et Ma.

Généralisations possibles

- Considérer plus de deux séquences : famille de séquences alignées.
- Remplacer les nucléotides par des codons.
- Étendre la dépendance au voisin de droite : champs markoviens.

Merci pour votre attention !


L. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains Ann. Math. Stat., 37 :1554-1563, 1966


L. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains Ann. Math. Stat., 41(1) :164-171, 1970


M. Bulmer, Neighboring base effects on substitution rates in pseudogenes Mol. Biol. and Evol., 3 :322-329, 1986

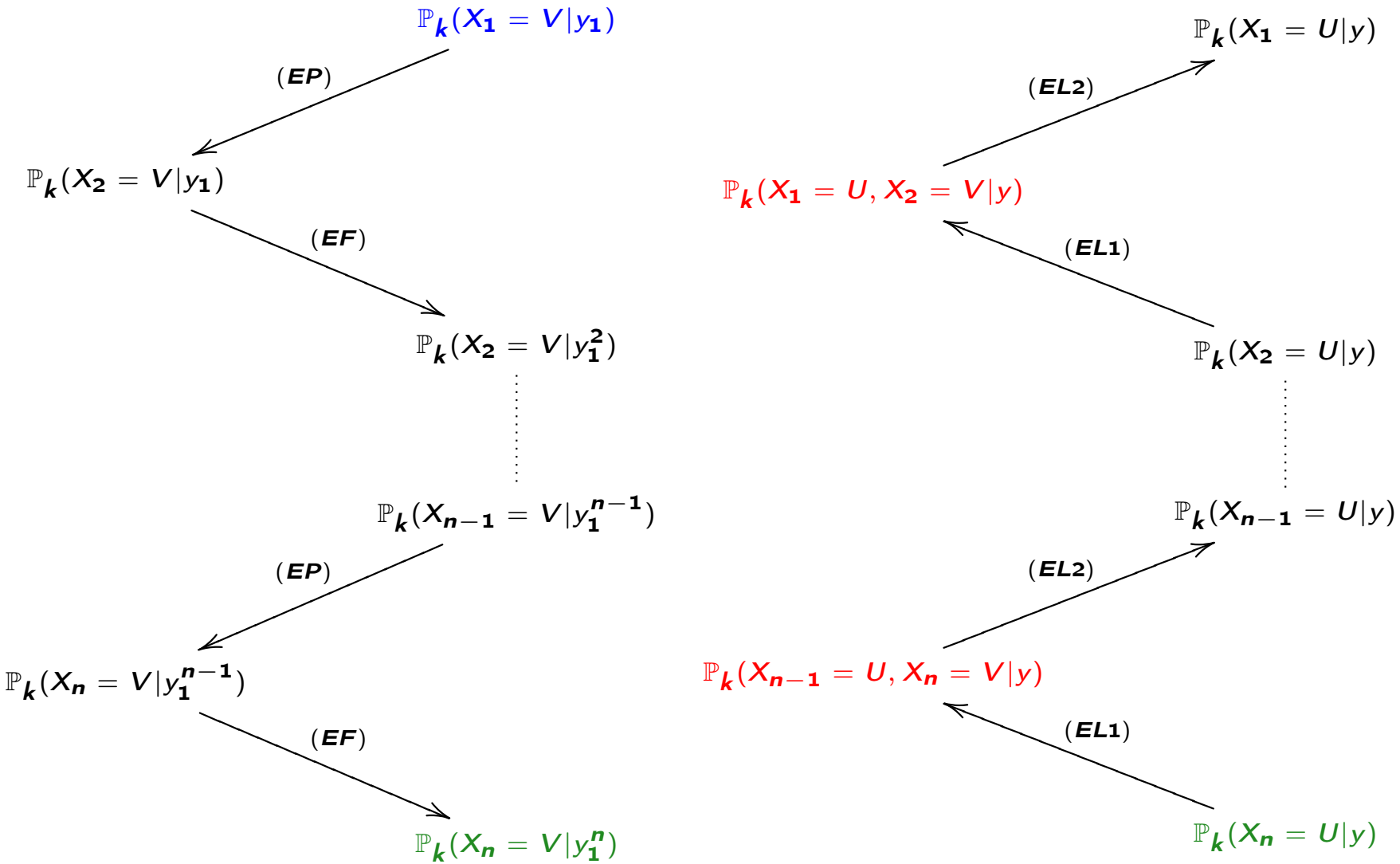

J.L. Jensen, Context dependent DNA evolutionary models Research Reports No.458, May 2005


J.L. Jensen, A.-M. Pedersen, Probabilistic models of DNA sequence evolution with context dependent rates of substitution Adv. Appl. Probab., 32 :499-517, 2000


A.-M. Pedersen, J.L. Jensen, A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames Mol. Biol. Evol., 18(5) :763-776, 2001


A.-M. Pedersen, C. Wiuf, F.B. Christiansen, A codon-based model designed to describe lentiviral evolution Mol. Biol. Evol., 15(8) :1069-1081, 1998


C. Wu, On the convergence properties of the EM algorithm Ann. Stat., 11(1) :95-103., 1983


$\mathbb{P}_k(X_1 = V|y_1)$
(EP)
$\mathbb{P}_k(X_2 = V|y_1)$
(EF)
$\mathbb{P}_k(X_2 = V|y_1^2)$
$\mathbb{P}_k(X_{n-1} = V|y_1^{n-1})$
(EP)
$\mathbb{P}_k(X_n = V|y_1^{n-1})$
(EF)
$\mathbb{P}_k(X_n = V|y_1^n)$
$\mathbb{P}_k(X_1 = U|y)$
(EL2)
$\mathbb{P}_k(X_1 = U, X_2 = V|y)$
(EL1)
$\mathbb{P}_k(X_2 = U|y)$
$\mathbb{P}_k(X_{n-1} = U|y)$
(EL2)
$\mathbb{P}_k(X_{n-1} = U, X_n = V|y)$
(EL1)
$\mathbb{P}_k(X_n = U|y)$

Equations prédictives (EP)

Notons $Y_i^j = (Y_i, \ldots, Y_j)$.

$$\begin{aligned}
\mathbb{P}_k(X_i = V | y_1^{i-1}) &= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_{i-1} = U, X_i = V | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_i = V | X_{i-1} = U, y_1^{i-1}) \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} \mathbb{P}_k(X_i = V | X_{i-1} = U) \mathbb{1}_{(u_1, u_M) = y_{i-1}} \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}), \\
&= \sum_{U \in \mathcal{A}^M} C^{(k)}(U, V) \mathbb{P}_k(X_{i-1} = U | y_1^{i-1}) \mathbb{1}_{(u_1, u_M) = y_{i-1}}.
\end{aligned}$$

Equations de filtrage (EF)

$$
\begin{aligned}
\mathbb{P}_k(X_i = U | y_1^i) &= \mathbb{P}_k(X_i = U | y_1^{i-1}, y_i), \\
&= \frac{\mathbb{P}_k(X_i = U, y_i | y_1^{i-1})}{\mathbb{P}_k(y_i | y_1^{i-1})}, \\
&= \frac{\mathbb{P}_k(X_i = U, y_i | y_1^{i-1})}{\sum_{V \in \mathcal{A}^M} \mathbb{P}_k(X_i = V, y_i | y_1^{i-1})}, \\
&= \frac{\mathbb{P}_k(y_i | X_i = U, y_1^{i-1}) \mathbb{P}_k(X_i = U | y_1^{i-1})}{\sum_{V \in \mathcal{A}^M} \mathbb{P}_k(y_i | X_i = V, y_1^{i-1}) \mathbb{P}_k(X_i = V | y_1^{i-1})}, \\
&= \frac{\mathbb{1}_{(u_1, u_M) = y_i} \mathbb{P}_k(X_i = U | y_1^{i-1})}{\sum_{V \in \mathcal{A}^M} \mathbb{1}_{(v_1, v_M) = y_i} \mathbb{P}_k(X_i = V | y_1^{i-1})}.
\end{aligned}
$$

Equations de lissage 1 (EL1)

$$
\begin{aligned}
\mathbb{P}_k(X_{i-1}=U, X_i=V|y) &= \mathbb{P}_k(X_{i-1}=U|X_i=V,y)\mathbb{P}_k(X_i=V|y),\\
&= \mathbb{P}_k(X_{i-1}=U|X_i=V,y_1^{i-1})\mathbb{P}_k(X_i=V|y)\mathbb{1}_{(v_0,v_M)=y_i},\\
&= \frac{\mathbb{P}_k(X_i=V,X_{i-1}=U|y_1^{i-1})\mathbb{P}_k(X_i=V|y)}{\mathbb{P}_k(X_i=V|y_1^{i-1})}\mathbb{1}_{(v_1,v_M)=y_i},\\
&= \frac{\mathbb{P}_k(X_i=V|X_{i-1}=U,y_1^{i-1})\mathbb{P}_k(X_{i-1}=U|y_1^{i-1})}{\mathbb{P}_k(X_i=V|y_1^{i-1})}\cdot\\
&\qquad \mathbb{P}_k(X_i=V|y)\mathbb{1}_{(v_1,v_M)=y_i},\\
&= \frac{\mathbb{P}_k(X_i=V|X_{i-1}=U)\mathbb{1}_{(u_1,u_M)=y_{i-1}}\mathbb{P}_k(X_{i-1}=U|y_1^{i-1})}{\mathbb{P}_k(X_i=V|y_1^{i-1})}\cdot\\
&\qquad \mathbb{P}_k(X_i=V|y)\mathbb{1}_{(v_1,v_M)=y_i},\\
&= \frac{\mathbb{P}_k(X_{i-1}=U|y_1^{i-1})\mathbb{P}_k(X_i=V|y)}{\mathbb{P}_k(X_i=V|y_1^{i-1})}\cdot\\
&\qquad C^{(k)}(U,V)\mathbb{1}_{(u_1,u_M)=y_{i-1}}\mathbb{1}_{(v_1,v_M)=y_i}.
\end{aligned}
$$

Equations de lissage 2 (EL2)

$$\begin{aligned}
\mathbb{P}_k(X_{i-1} = U|y) &= \sum_{V \in \mathcal{A}^M} \mathbb{P}_k(X_{i-1} = U, X_i = V|y), \\
&= \mathbb{P}_k(X_{i-1} = U|y_1^{i-1})\mathbb{1}_{(u_1,u_M)=y_{i-1}} \cdot \\
&\qquad \sum_{V \in \mathcal{A}^M} \frac{\mathbb{P}_k(X_i = V|y)}{\mathbb{P}_k(X_i = V|y_1^{i-1})} C^{(k)}(U,V)\mathbb{1}_{(v_1,v_M)=y_i}.
\end{aligned}$$