Chaînes de Markov et génome

Etienne Pardoux

Introduction

Le but de ces quelques leçons est double :

- introduire les chaînes de Markov et les chaînes de Markov cachées, et expliquer le principe des algorithmes basés sur les chaînes de Markov cachées pour l'annotation du génome ;

- introduire les chaînes de Markov sur les arbres, et leur utilisation en phylogénie, en particulier pour le calcul de vraisemblances d'arbres.

1 Comment lire l'ADN ?

On considère un fragment d'ADN, sous la forme d'un simple brin constitué d'une succession de nucléotides, que nous considèrerons comme des lettres dans l'alphabet `a, c, g, t`, par exemple

`a c c g t a a t t c g g a` $\cdots$ `t t g c`

"Lire" ou "annoter" cette séquence consiste essentiellement à la décomposer en *régions codantes à l'endroit* ou *à l'envers* (sachant que l'ADN est constitué en réalité de deux brins complémentaires appariés, qui ne sont pas lus dans le même sens), et *régions non codantes* ; dans le cas des génomes eukaryotes il faut en outre découper les *régions codantes* en *introns* et *exons*. Notons que les régions codantes sont lues par *codons*, i.e. triplets de nucléotides, chaque codons étant ensuite traduit en un *acide aminé*. L succession des acides aminés cosntitue un *gène*. Il est donc essentiel de lire chaque région codante dans la bonne phase de lecture. Oublier un codon n'est pas forcément très grave, mais se tromper en décalant la lecture d'un ou de deux nucléotides est catastrophique !

On est aidé dans cette démarche par la présence d'un codon START (resp. STOP) au début (resp. à la fin) de chaque région codante. Mais tout START ou STOP potentiel n'en est pas forcément un effectivement. Et il n'y a pas de signaux aussi nets marquant la transition entre *intron* et *exon*.

Une première possibilité est que les proportions respectives de `a`, de `c`, de `g` et de `t` soient nettement différentes entre plage codante et non codante. Une seconde possibilité est que ces proportions ne sont pas vraiment nettement différentes, et qu'il faut compter les *di* ou *trinucléotides*.

Dans le premier cas, on va distinguer entre région codante et non codante en comparant les proportions de `a`, de `c`, de `g` et de `t`. Dans le second cas, il faudra compter les paires ou les triplets. Et quelque soit le critère adopté, le plus difficile est de localiser correctement les ruptures (ou changements de plage).

Les méthodes que nous venons d'évoquer pour décomposer une séquence d'ADN en ses différentes plages – dans le but de détecter les gènes – peuvent être vues comme des procédures statistiques associées à une modélisation probabiliste. Cette modélisation n'est pas la même suivant que l'on regarde des fréquences de *nucléotides*, de *bi–* ou de *tri-nucléotides*.

Nous allons maintenant faire un détour par les modèles probabilistes possibles pour une séquence d'ADN.

2 Le modèle i.i.d

i.i.d. veut dire "indépendants et identiquement distribués". Ici on suppose que les nucléotides d'une sous–séquence donnée sont tirés indépendemment les uns des autres, tous avec la même loi de probabilité. La sous–séquence en question est une "plage homogène" (région codante, région intergénique, ...).

Définissons tout d'abord la notion d'*espace de probabilité* $(\Omega, \mathcal{F}, \mathbb{P})$.

• Ω est l'ensemble de toutes les réalisations possibles de l'expérience aléatoire, ou ensemble de tous les états du monde possibles, ou ensemble de toutes les suites possibles de nucléotides.

• $\mathcal{F}$ est la tribu des événements. En première approximation, on peut ici choisir $\mathcal{F}$ = ensemble des toutes les parties de Ω.

Exemples d'événement :

"Le 1er nucléotide est une purine"

"le triplet en position 7–8–9 est un codon START".

• $\mathbb{P}$ est la *probabilité*, qui à chaque événement $F \in \mathcal{F}$ associe un nombre réel $\mathbb{P}(F)$ de l'intervalle [0,1], et qui vérifie les deux axiomes :

i) $\mathbb{P}(\Omega) = 1$

ii) Si $\{F_n, n \geq 1\} \subset \mathcal{F}, F_n \cap F_m = \emptyset$ dès que $n \neq m$, $\mathbb{P}(\bigcup_n^\infty F_n) = \sum_1^\infty \mathbb{P}(F_n)$

Une *variable aléatoire* à valeurs dans E (par exemple $E = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, ou $E = \mathbb{N}, \ldots$) est une application

$$X = \Omega \to E$$

qui à $\omega \in \Omega$ associe $X(\omega)$ (qui doit vérifier la condition $\{\omega; X(\omega) = x\} \in \mathcal{F}$ pour tout $x \in E$).

Exemples de variable aléatoire

"le 5ème nucléotide de la séquence"

"le rang du 1er codon START dans la séquence"

"le nombre le `t a t a` dans la séquence"

Définition 2.1 *Une suite $(X_1, \ldots, X_n)$ de v.a. est dite indépendante si pour tout $x_1, \ldots, x_n \in E$,*

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$$

$\diamondsuit$

Exemple particulier avec $k = 5$:

$$\begin{aligned}
&\mathbb{P}(X_1 = \mathtt{a}, X_2 = \mathtt{c}, X_3 = \mathtt{a}, X_4 = \mathtt{t}, X_5 = \mathtt{g}) \\
&= \mathbb{P}(X_1 = \mathtt{a})\mathbb{P}(X_2 = \mathtt{c})\mathbb{P}(X_3 = \mathtt{a})\mathbb{P}(X_4 = \mathtt{t})\mathbb{P}(X_5 = \mathtt{g}) \\
&= \mathbb{P}(X_1 = \mathtt{a})\mathbb{P}(X_1 = \mathtt{c})\mathbb{P}(X_1 = \mathtt{a})\mathbb{P}(X_1 = \mathtt{t})\mathbb{P}(X_1 = \mathtt{g}).
\end{aligned}$$

Soit X_1 le premier nucléotide de notre séquence. Sa *loi de probabilité* est définie par le vecteur $p = (p_\mathtt{a}, p_\mathtt{c}, p_\mathtt{g}, p_\mathtt{t})$ donné par

$$p_\mathtt{a} = \mathbb{P}(X_1 = \mathtt{a}),\ p_\mathtt{c} = \mathbb{P}(X_1 = \mathtt{b}),\ p_\mathtt{g} = \mathbb{P}(X_1 = \mathtt{g}),\ p_\mathtt{t} = \mathbb{P}(X_1 = \mathtt{t})$$

Notons que $p_\mathtt{a}, p_\mathtt{c}, p_\mathtt{g}, p_\mathtt{t} \geq 0$ et $p_\mathtt{a} + p_\mathtt{c} + p_\mathtt{g} + p_\mathtt{t} = 1$.

On dit que les v.a. $(X_1, \ldots, X_n)$ sont i.i.d. (indépendantes et identiquement distribuées) si elles sont indépendantes et toutes de même loi. On dit aussi (dans le langage des statisticiens) que la suite $(X_1, \ldots, X_n)$ est un échantillon de taille n de la loi commune des X_i. A cet échantillon, on associe la loi de probabilité empirique

$$p^n_\mathtt{a} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathtt{a}\}},\ p^n_\mathtt{c} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathtt{c}\}},\ p^n_\mathtt{g} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathtt{g}\}},\ p^n_\mathtt{t} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathtt{t}\}}.$$

$p^n = (p^n_\mathtt{a}, p^n_\mathtt{c}, p^n_\mathtt{g}, p^n_\mathtt{t})$ est une probabilité sur E.

En pratique, la loi commune $p = (p_\mathtt{a}, p_\mathtt{c}, p_\mathtt{g}, p_\mathtt{t})$ des X_i est inconnue. Du moins si n est suffisamment grand, p^n est une bonne approximation de p. En effet, il résulte de la loi des grands nombres que

$$p^n_\mathtt{a} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i = \mathtt{a}\}} \to \mathbb{E}(\mathbf{1}_{\{X_1 = \mathtt{a}\}}) = \mathbb{P}(X_1 = \mathtt{a})$$

quand $n \to \infty$ (même résultat pour $\mathtt{c}, \mathtt{g}, \mathtt{t}$), et en outre d'après le théorème de la limite centrale,

$$\sqrt{n}(p_\mathtt{a} - p^n_\mathtt{a}) \xrightarrow{\mathcal{L}} N(0, p_\mathtt{a}(1 - p_\mathtt{a})),$$

c'est à dire pour tout $\delta > 0$,

$$\mathbb{P}\left(-\delta\sqrt{\frac{p_\mathtt{a}(1-p_\mathtt{a})}{n}} \leq p_\mathtt{a} - p^n_\mathtt{a} \leq \delta\sqrt{\frac{p_\mathtt{a}(1-p_\mathtt{a})}{n}}\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\delta}^{\delta} e^{-\frac{x^2}{2}} dx,$$

et donc puisque $\sqrt{p_{\mathtt{a}}(1-p_{\mathtt{a}})} \leq \frac{1}{2}$,

$$\mathbb{P}\left(|p_{\mathtt{a}} - p_{\mathtt{a}}^n| > \frac{\delta}{2\sqrt{n}}\right) \leq \sqrt{\frac{2}{\pi}} \int_{\delta}^{\infty} e^{-\frac{x^2}{2}} dx$$

On peut donc estimer la loi inconnue p, sous l'hypothèse que les nucléotides sont i.i.d., donc en particulier que la plage considérée est *homogène*.

L'hypothèse d'indépendance n'est pas forcément vérifiée, mais en réalité elle n'est pas absolument nécessaire pour que la démarche ci–dessus puisse être justifiée.

3 Le modèle de Markov

Supposer que les nucléotides sont indépendants les uns des autres n'est pas très raisonnable. On peut penser par exemple que, dans une région codante, la loi du 2ème nucléotide d'un codon dépend de quel en est le premier nucléotide.

D'où l'idée de supposer que la suite $(X_1, \ldots, X_n)$ forme une chaîne de Markov.

Rappelons la notion de probabilité conditionnelle $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Définition 3.1 *Une suite de v.a. $(X_1, \ldots, X_n)$ à valeurs dans l'ensemble E est une chaîne de Markov d'ordre 1 (modèle $M1$) si pour tout $1 < k \leq n$, $x_1, x_2, \ldots, x_k \in E$,*

$$\mathbb{P}(X_k = x_k | X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) = \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1}).$$

Plus généralement, la suite $(X_1, \ldots, X_n)$ est une chaîne de Markov d'ordre ℓ (≥ 1) (Modèle $M\ell$) si $\forall k > \ell$,

$$\mathbb{P}(X_k = x_k | X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) = \mathbb{P}(X_k = x_k | X_{k-\ell} = x_{k-\ell}, \ldots, X_{k-1} = x_{k-1})$$

$\diamondsuit$

Notons qu'une suite indépendante constitue un modèle $M0$. On va maintenant étudier le modèle $M1$, qui constitue le modèle de référence.

4 Chaîne de Markov homogène d'ordre 1

Même si notre but ultime est précisément d'étudier des situations non homogènes, il est essentiel de comprendre d'abord le cas homogène.

Définition 4.1 *Une chaîne de Markov $(X_1, \ldots, X_n)$ à valeurs dans l'ensemble fini E est dite homogène si pour tous $x, y \in E$, la quantité*

$$\mathbb{P}(X_{k+1} = y | X_k = x)$$

ne dépend pas de $1 \leq k < n$

$\Diamond$

Notons $P = (P_{xy})_{x,y\in E}$ la *matrice de transition* de la chaîne définie par

$$P_{xy} = \mathbb{P}(X_{k+1} = y | X_k = x), \;\; x, y \in E,$$

et $\mu_x = \mathbb{P}(X_1 = x), \;\; x \in E$ la *loi initiale.*

Lemme 4.2 *Soit F un autre ensemble fini, $f = E \times F \to E$, $\{Y_2, Y_3, \ldots Y_n\}$ une suite de v.a. i.i.d. à valeurs dans F, indépendante de X_1, avec X_1 = v.a. à valeurs dans E, de loi μ. Alors la suite $\{X_1, \ldots, X_n\}$ définie par la formule de récurrence*

$$X_k = f(X_{k-1}, Y_k), \;\; 2 \leq k \leq n$$

définit une chaîne de Markov de loi initiale μ et de matrice de transition

$$P_{xy} = \mathbb{P}(f(x, Y_2) = y).$$

Proposition 4.3 *La suite $(X_1, \ldots, X_n)$ est une chaîne de Markov de loi initiale μ et de matrice de transition P ssi pour tout $1 < k \leq n$, la loi de $(X_1, \ldots, X_k)$ est donnée par*

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \mu_{x_1} P_{x_1 x_2} \times \cdots \times P_{x_{k-1} x_k}.$$

Preuve : La CN s'établit en utilisant $k - 1$ fois la formule

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B),$$

et $k - 2$ fois la propriété de Markov.

Corollaire 4.4 *Si $(X_1, \ldots, X_n)$ est une chaîne de Markov de loi initiale μ, et de matrice de transition P, alors pour $k \geq 1$, $\mathbb{P}(X_{1+k} = y | X_1 = x) = (P^k)_{xy}$ et la loi de X_k $(1 < k \leq n)$ est la probabilité*

$$\mu(k) = \mu P^{k-1},$$

i.e. $\forall x \in E$,

$$\mu(k)_x = \sum_y \mu_y (P^{k-1})_{yx}$$

5 Chaîne de Markov homogène irréductible

On veut maintenant énoncer l'équivalent de la loi des grands nombres, pour les chaînes de Markov. Il nous faut d'abord ajouter une condition.

Définition 5.1 *On dit que la chaîne de Markov $(X_1, \ldots, X_n)$ de matrice de transition P est irréductible si $\forall x, y \in E$, $\exists k \geq 1$ tel que*

$$(P^k)_{xy} > 0,$$

i.e. si $\exists k$ tel que la chaîne passe avec probabilité non nulle de x à y en k itérations.

Théorème 5.2 *Soit $(X_k,\ k \geq 1)$ une chaîne de Markov à valeurs dans un ensemble fini E, de matrice de transition P irréductible. Alors il existe une unique probabilité π invariante par P, i.e. telle que $\pi = \pi P$, qui vérifie $\pi_x > 0$, $\forall x \in E$. Si $(X_k,\ k \geq 1)$ est une chaîne de Markov de loi initiale π et de matrice de transition P, alors π est la loi de X_k pour tout $k \geq 1$. Enfin on a le théorème ergodique (généralisation de la loi forte des grands nombres) : pour tout $f : E \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \sum_{x\in E} f(x)\pi_x \quad p.s.,$$

quelle que soit la loi de X_1.

Si on note à nouveau $\mu(k)$ la loi de X_k, il résulte du théorème ergodique, en prenant l'espérance, que pour toute probabilité $\mu(1)$ sur E,

$$\frac{1}{n}\sum_{k=1}^{n} \mu(k) \to \pi,$$

quand $n \to \infty$. Une question naturelle à se poser est de savoir si l'on a ou non $\mu(n) \to \pi$, quand $n \to \infty$. Ce n'est pas toujours vrai sous les hypothèses ci–dessus. Il faut en outre supposer que

Définition 5.3 *On dit qu'une chaîne de Markov de matrice de transition P irréductible est apériodique ssi pour un couple (x, y) dans $E \times E$ (et alors pour tous les couples, par l'irréductibilité), il existe N tel que $(P^n)_{xy} > 0$, pour tout $n \geq N$.*

Remarque 5.4 *On peut admettre que la chaîne des nucléotides successifs est irréductible et apériodique. Mais les chaînes périodiques ne sont pas que des curiosités mathématiques que l'on ne rencontrerait pas dans des modèles simples. Considérons la chaîne cachée (i.e. dont les valeurs ne sont pas données par la lecture des nucléotides), qui indique dans quelle plage (non codante, codante à l'endroit, codante à l'envers) se trouve le nucléotide que l'on lit. Codons par 0 l'état "non codant", 1 l'état "codant à l'endroit", 2 l'état "codant à l'envers". Quand on est à l'état 0, soit on y reste, soit on passe dans l'état 1, soit on passe dans l'état 2. Il est assez raisonnable de supposer que l'on ne passe pas directement de l'état 1 dans l'état 2 (et vice versa), sans repasser dans l'état 0. Considérons maintenant la chaîne qui décrit la suite des plages visitées, en "gommant" les longueurs de ces plages (donc en particulier la matrice de transition correspondante P a tous ses termes diagonaux nuls). Alors*

$$(P^n)_{00} = \begin{cases} 0, & \text{si } n \text{ est impair} \\ 1, & \text{si } n \text{ est pair,} \end{cases}$$

et la chaîne que nous venons de décrire est périodique.

Théorème 5.5 *Soit $(X_k, k \geq 1)$ une chaîne de Markov de matrice de transition P irréductible et apériodique. On désigne pour tout $n \geq 1$ par $\mu(n)$ la loi de probabilité de X_n. Alors $\mu(n) \to \pi$ quand $n \to \infty$.*

Donc dans le cas irréductible et apériodique on peut admettre que, en tout cas à partir d'un certain rang, la suite des X_k est identiquement distribuée, de loi commune l'unique probabilité invariante π de la chaîne.

6 Chaîne de Markov réversible

Soit $(X_1, \ldots, X_n)$ une chaîne de Markov de matrice de transition P. Posons $\hat{X}_k = X_{n+1-k}$. C'est un exercice facile sur les probabilités conditionnelles de montrer que la *suite retournée* $(\hat{X}_1, \ldots, \hat{X}_n) = (X_n, \ldots, X_1)$ est une chaîne de Markov. En général cette nouvelle chaîne de Markov n'est pas homogène. $(\hat{X}_1, \ldots, \hat{X}_n)$ est une chaîne homogène si $(X_1, \ldots, X_n)$ est initialisée avec sa probabilité invariante π. Dans ce cas, d'après la formule de Bayes, la matrice $\hat{P}$ de la chaîne retournée est donnée par

$$\hat{P}_{xy} = \frac{\pi_y P_{yx}}{\pi_x}.$$

On dit que la chaîne $(X_1, \ldots, X_n)$ est *réversible* si $\hat{P} = P$, ce qui est équivalent à ce que la *relation d'équilibre ponctuel* (en Anglais *detailed balance equation*) suivante soit satisfaite

$$\pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x \neq y,$$

autrement dit si la quantité $\pi_x P_{xy}$ est symmétrique en x, y.

Remarque 6.1 *Etant donnée une chaîne irréductible de matrice de transition P, cette chaîne possède une unique probabilité invariante, qui forcément satisfait la relation $\pi P = \pi$. Le couple (π, P) satisfait ou non la relation d'équilibre ponctuel (i.e. toutes les chaînes irréductibles ne sont pas réversibles). Pour construire une chaîne irréductible et non réversible, il suffit de choisir une matrice P irréductible, telle que pour un certain couple $x \neq y$, $P_{xy} = 0 < P_{yx}$.*

D'un autre coté, si P est une matrice de transition et π une probabilité sur E, telles que la relation d'équilibre ponctuel soit satisfaite, alors π est une probabilité invariante, comme on le vérifie en sommant la relation d'équilibre ponctuel par rapport à x (ou à y).

7 Statistique des chaînes de Markov homogènes $M1$

7.1 Estimation de la mesure invariante

On sait que $\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}} \to \pi_x$ p.s. quand $n \to \infty$. Donc

$$\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}}$$

est un estimateur de π_x, $x \in E$, au vu des observations $X_1, X_2, \ldots, X_n$.

7.2 Estimation de la matrice de transition

On va montrer que

$$\frac{\sum_1^n \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}}{\sum_1^n \mathbf{1}_{\{X_k=x\}}} \to P_{x,y}, n \to \infty.$$

Notons tout d'abord que la fraction ci–dessus vaut

$$\frac{\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}}{\frac{1}{n}\sum_1^n \mathbf{1}_{\{X_k=x\}}}.$$

Il suffit de considérer le numérateur. Plus précisément, il vaut

$$\frac{1}{n}\sum_{k \text{ pair}} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}} + \frac{1}{n}\sum_{k \text{ impair}} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}$$

Considérons par exemple la quantité

$$\frac{2}{n}\sum_{k \text{ pair, } 1\le k\le n} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}$$

On remarque que la chaîne $\{(X_1,X_2),(X_3,X_4),(X_5,X_6),\ldots\}$ est une chaîne de Markov à valeurs dans $E\times E$, de matrice de transition de (x,y) à (x',y') donnée par $P_{y,x'}P_{x'y}$, et de probabilité invariante $\tilde{\pi}_{xy}=\pi_x P_{xy}$.

Donc

$$\frac{2}{n}\sum_{k \text{ pair, } 1\le k\le n} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}} \to \pi_x P_{xy}.$$

Finalement

$$\frac{\sum_1^n \mathbf{1}_{\{X_k=x,X_{h+1}=y\}}}{\sum_1^n \mathbf{1}_{\{X_k=x\}}} \to P_{xy}$$

p.s., quand $n\to\infty$.

8 Statistique des chaînes de Markov Mk

Pour simplifier, on va se contenter de décrire le modèle $M2$. Dans ce cas, ce qui remplace la matrice P est une matrice de transition de $E\times E$ dans E, qui donne la loi de probabilité de X_{k+1}, sachant le couple (X_{k-1},X_k). Dans le cas $E=\{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$, on a donc une matrice transition à 16 lignes (indexées par les dinucléotides $\{\mathtt{aa},\mathtt{ac},\ldots,\mathtt{gt},\mathtt{tt}\}$) et 4 colonnes (indexées par $\{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$).

Remarque 8.1 *On peut aussi se ramener à un modèle $M1$ sur l'espace d'état $E\times E$, puisque si $(X_1,X_2,\ldots,X_n)$ est une chaîne de Markov d'ordre 2 à valeurs dans E, $((X_1,X_2),(X_2,X_3),\ldots,(X_{n-1},X_n))$ est une chaîne de Markov d'ordre 1 à valeurs dans $E\times E$. On se ramène à une matrice de transition carrée, et on peut introduire la notion de probabilité invariante...*

On estime la probabilité de transition $P_{xy,z}$ à l'aide de la quantité

$$\frac{\sum_{k=1}^{n} \mathbf{1}_{\{X_k=x,X_{k+1}=y,X_{k+2}=z\}}}{\sum_{k=1}^{n} \mathbf{1}_{\{X_k=x,X_{k+1}=y\}}},$$

qui converge p.s. vers $P_{xy,z}$ quand $n \to \infty$. Remarquons que cette statistique inclut le décompte des trinucléotides, donc en particulier des codons, ce qui fait que les chaînes d'ordre 2 sont très utilisées pour modéliser les régions codantes de l'ADN.

9 Chaîne de Markov non homogène

Plusieurs types de non homogénéités sont pertinents dans la modélisation de l'ADN.

9.1 Chaîne de Markov phasée

Dans une "plage codante", on peut penser que la probabilité de transition n'est pas indépendante du site, mais périodique de période 3. Comme la notion de "chaîne de Markov périodique" désigne tout autre chose (à savoir une chaîne qui n'est pas "apériodique" au sens de la définition 5.3), nous utiliserons, à la suite du livre de Robin, Rodolphe, Schbath, la terminologie "chaîne Markov phasée" pour désigner une chaîne de Markov $(X_n,\ 1 \leq n \leq N)$ telle que pour tous $x, y \in E$, l'application $n \to P(X_{n+1} = y | X_n = x)$ est périodique. Dans le cas qui nous occupe, on peut y compris songer à une chaîne de Markov d'ordre 2, telle que pour tout $y \in E$, la quantité $P(X_{n+1} = y | X_n = x, X_{n-1} = x')$ ne dépende pas de x, x' pour $n = 3k$, que de x pour $n = 3k+1$ et dépende de x, x' pour $n = 3k+2$, k entier. Cela veut dire en particulier que les codons successifs sont i.i.d. On pourraît aussi supposer que les codons successifs forment une chaîne de Markov d'ordre 1.

9.2 Chaîne de Markov localement homogène

Si l'on regarde plus globalement la séquence génomique, on s'attend à ce que la chaîne de Markov qui décrit celle–ci soit homogène dans la réunion des régions non codantes, dans celle des régions codantes à l'endroit, celle des régions non codantes, la réunion des introns et celle des exons, mais pas globalement homogène, et c'est d'ailleurs cette inhomogénéité qui doit nous permettre de réaliser l'annotation. Le principal problème est bien de détecter ce que l'on appelle les "ruptures de modèle".

Il existe une importante littérature statistique sur ces problèmes de rupture de modèle, mais il n'est pas clair que les algorithmes correspondant sont adaptables à la situation qui est la nôtre ici, où il est essentiel d'exploiter l'homogénéité de la chaîne sur la réunion des plages de même type (non codant, codant à l'endroit,...), et pas seulement sur chacune de ces plages prise isolément.

Cependant, Audic et Claverie ont mis au point un algorithme pour l'annotation des génomes prokaryotes, que nous allons maintenant décrire. On suppose que notre modèle

(qui peut être $M0$, $M1$, $M2$,...) est décrit par un paramètre θ (qui est une probabilité sur E dans le cas $M0$, une probabilité de transition dans le cas $M1$, ...), lequel prend trois valeurs distinctes (toutes trois inconnues !) $(\theta_0, \theta_1, \theta_2)$, suivant que l'on est dans une plage non codante, codante à l'endroit ou codante à l'envers.

- •*Étape d'initialisation* On découpe la séquence en plages de longueur 100 (éventuellement, la dernière plage est de longueur > 100). On décide au hasard de placer chaque plage dans l'une des trois "boîtes" 0, 1, et 2. Sur la base de tous les X_n se trouvant dans la boîte 0, on estime une valeur du paramètre θ, soit $\theta_0^{(1)}$. On estime de même les valeurs $\theta_1^{(1)}$ et $\theta_2^{(1)}$.
- •*Étape de mise à jour* Supposons que nos trois "boîtes" 0, 1, et 2 contiennent chacune des plages distinctes de longueur ≥ 100, sur la base desquelles on a estimé les valeurs $\theta_0^{(n)}$, $\theta_1^{(n)}$ et $\theta_2^{(n)}$. On commence par vider ces boîtes, et on reprend la séquence complète $\{X_n,\ 1 \leq n \leq N\}$. On extrait la sous–suite $\{X_n,\ 1 \leq n \leq 100\}$. On estime le paramètre θ sur la base de cette sous–suite, et on choisit laquelle des trois valeurs $\theta_0^{(n)}$, $\theta_1^{(n)}$ et $\theta_2^{(n)}$ est la plus proche de cette nouvelle valeur estimée. Puis on se pose le même problème avec la suite $\{X_n,\ 10 \leq n \leq 110\}$, avec la suite $\{X_n,\ 20 \leq n \leq 120\}$,.. jusqu'à ce que la valeur estimée devienne plus proche d'une autre des trois valeurs de l'étape précédente. Alors on revient en arrière de 50 nucléotides, et on place l'intervalle ainsi sélectionné depuis le début de la séquence dans la boîte 0, 1, ou 2, suivant le cas. On recommence, en prenant une plage de longueur 100, adjacente à l'intervalle que l'on vient de placer dans une des boîtes, et on répète les opérations précédentes. Lorsque l'on a épuisé la séquence, on se retrouve avec trois boîtes contenant chacune (du moins il faut l'espérer) des plages de longueur ≥ 100. On estime alors les trois nouvelles valeurs $\theta_0^{(n+1)}$, $\theta_1^{(n+1)}$ et $\theta_2^{(n+1)}$, sur la base du contenu des boîtes 0, 1, et 2 respectivement.

Si la séquence initiale est effectivement constituée de plages dont les compositions statistiques sont de trois types différents, l'algorithme converge rapidement, et quand on s'arrête, on a un découpage de la séquence initiale en sous–séquences de trois types différents. Il ne reste plus qu'à décider "qui est qui", ce qui requiert des connaissances a priori, acquises en observant des séquences qui ont déjà été annotées.

10 Chaîne de Markov cachée

Le point de vue Bayésien consiste à se donner une loi de probabilité a priori sur les paramètres inconnu θ_i, et leur évolution. Plus précisément on va maintenant se donner une nouvelle chaîne de Markov $(Y_1, \ldots, Y_N)$, dite "cachée" parce que non observée. Dans le cas des génomes prokaryotes, la chaîne (Y_n) prend par exemple ses valeurs dans l'ensemble à trois éléments $F = \{0, 1, 2\}$, et dans le cas eukaryote il faut différencier les états 1 et 2 entre les parties *intron* et *exon*. En réalité c'est encore un peu plus compliqué, car il faudrait prendre en compte les codons START et STOP, mais on verra cela un peu plus loin. L'avantage de cette approche est que l'on dispose d'algorithmes pour répondre aux questions que nous nous posons. On note F l'espace dans lequel la chaîne cachée prend ses valeurs, et $d = \text{card}(F)$.

Rappelons que $d \geq 3$.

Pour simplifier la présentation succincte de ces algorithmes, on va supposer que $(Y_1, \ldots, Y_N)$ est une chaîne de Markov (μ, P) à valeurs dans F, et que, connaissant les (Y_n), la suite des nucléotides $(X_1, \ldots, X_N)$ est indépendante, la loi de chaque X_n dépendant uniquement du Y_n correspondant, i.e. pour tout $1 \leq n \leq N$,

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{k=1}^{n} \mathbb{P}(X_k = x_k | Y_k = y_k) \\ &= \prod_{k=1}^{n} Q_{y_k x_k}.\end{aligned}$$

Le problème que l'on cherche à résoudre est le suivant : ayant observé la suite des nucléotides $(x_1, \ldots, x_N)$, quelle est la suite des états cachés $(y_1^*, \ldots, y_N^*)$ qui "explique le mieux" ces observations ? Autrement dit, il s'agit de calculer la suite qui maximise la vraisemblance de la loi a posteriori sachant les observations, i.e.

$$(y_1^*, \ldots, y_n^*) = \text{argmax}_{y_1,\ldots,y_n} \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y | X_1 = x_1, \ldots, X_n = x_n).$$

Notons que, dans ce modèle, on a comme paramètres inconnus le triplet (μ, P, Q). Pour résoudre le problème ci–dessus, on est obligé d'estimer d'abord les paramètres (mais nous discuterons ce problème à la fin). Si l'on admet que l'on connaît les paramètres, notre problème est résolu par :

10.1 L'algorithme de Viterbi

Définissons la suite de vecteurs ligne $\delta(n)$ par :

$$\delta_y(n) = \max_{y_1, y_2, \ldots, y_{n-1}} \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y, X_1 = x_1, \ldots, X_n = x_n)$$

$\delta_y(n)$ est en quelque sorte la plus forte probabilité d'une trajectoire des $\{Y_k, 1 \leq k \leq n-1\}$, qui se termine par $Y_n = y$, et correspondant à la suite des nucléotides observés $x_1, \ldots, x_n$. On a la formule de récurrence suivante entre les vecteurs $\delta(n)$:

$$\delta_y(n+1) = (\delta(n) * P)_y Q_{y x_{n+1}}$$

où l'opération $*$ qui à un vecteur ligne de dimension d et une matrice $d \times d$ associe un vecteur ligne de dimension d est définie comme suit :

$$(\delta * P)_y = \sup_{z \in F} \delta_z P_{zy}.$$

L'algorithme de Viterbi consiste à calculer les $\delta(n)$ de $n = 1$ à $n = N$, puis à retrouver la trajectoire optimale en cheminant pas à pas dans le sens "rétrograde" : connaissant y_n^*, on en déduit y_{n-1}^* par la formule :

$$y_{n-1}^* = \psi_{y_n^*}(n),$$

avec

$$\psi_y(n) = \text{argmax}_{z \in F} \delta_z(n-1) P_{zy}.$$

L'algorithme de Viterbi est décrit comme suit :

1. *Initialisation :*

$$\delta_y(1) = \mu_y Q_{yx_1}, \ \ y \in F;$$
$$\psi(1) = 0.$$

2. *Récurrence :* pour $1 < n \leq N$,

$$\delta_y(n) = (\delta(n-1) * P)_y Q_{yx_n},$$
$$\psi_y(n) = \text{argmax}_{z \in F} \delta_z(n-1) P_{zy}, \ \ y \in F.$$

3. *Etape finale :*

$$\delta^* = \max_{y \in F} \delta_y(N)$$
$$y_N^* = \arg\max_{y \in F} \delta_y(N).$$

4. *Récurrence rétrograde*

$$y_n^* = \psi_{y_{n+1}^*}(n+1), \ \ 0 \leq n < N.$$

10.2 Estimation des paramètres

Il y a deux stratégies possibles. L'une consiste à estimer les paramètres sur une séquence d'apprentissage déjà annotée. Dans ce cas, on estime les paramètres d'un modèle où toute la suite $\{(X_n, Y_n),\ 1 \leq n \leq N\}$ est observée. On utilise les algorithmes d'estimation bien connus que nous avons présentés dans les sections précédentes.

L'autre stratégie consiste à estimer les paramètres sur la base des seules observations de la suite des nucléotides. L'avantage est de faire l'estimation à partir du génome étudié, et non pas à partir d'un génome différent. L'inconvénient est bien sûr que l'on estime un modèle avec des observations très partielles. Il existe cependant des algorithmes maintenant classiques (l'algorithme EM, et sa variante SEM), qui permettent de résoudre ce problème.

Nous allons présenter de façon très sommaire l'algorithme SEM, qui est le plus utile dans les situations que nous décrirons plus loin. Pour chaque valeur du paramètre inconnu θ, on considère la loi conditonnelle des états cachés, sachant la suite des nucléotides, notée

$$\mathbb{P}_\theta \left(Y_1 = y_1, \ldots, Y_N = y_N | X_1 = x_1, \ldots, X_N = x_N \right),$$

ou plutôt

$$\mathbb{P}_\theta \left(Y_1^N = y_1^N | X_1^N = x_1^N \right).$$

L'algorithme SEM est un algorithme itératif, que l'on initie avec une valeur θ_0. L'itération qui remplace θ_n par θ_{n+1} se décompose en deux étapes comme suit :

– *Simulation* On tire au hasard une réalisation de la suite aléatoire Y_1^N, suivant la loi $\mathbb{P}_{\theta_n}(Y_1^N = \cdot | X_1^N = x_1^N)$. Notons $y_1^N(n)$ la suite obtenue ainsi.
– *Ré–estimation* On choisit

$$\theta_{n+1} = \text{argmax}_\theta \mathbb{P}_\theta \left(Y_1^N = y_1^N(n), X_1^N = x_1^N \right).$$

Dans l'algorithme EM, l'étape de simulation est remplacée par le calcul de $\mathbb{E}_{\theta_n}(Y_1^N | X_1^N = x_1^N)$.

11 Modèle semi–markovien caché

11.1 Les limites du modèle de Markov caché

On a vu ci–dessus à la section 7.1 que les temps de séjour d'une chaîne de Markov dans chacun des états visité suivent des lois géométriques. Le modèle de la section 11 implique donc que les longueurs des plages codantes et non codantes d'un génome prokaryote suivent des lois géométriques. Or cette hypothèse ne cadre pas avec les données dont on dispose. Il y a là un premier argument pour envisager un modèle plus général, mais on va voir maintenant un argument encore plus convainquant pour abandonner le modèle de Markov caché.

Examinons plus précisément notre problème, en nous limitant à nouveau pour simplifier au génome prokaryote. Il est bien sûr essentiel de prendre en compte l'information contenue dans les codons START et STOP. Si l'on renonce à un modèle phasé, on est obligé d'introduire 3 états START, 3 états codants et 3 états STOP, chacun correspondant à une des trois phases de lecture, le tout doit être multiplié par deux pour tenir compte du brin complémentaire. On ajoute un état non codant. Cela fait en tout 19 états. Certes, la plupart des termes de la matrice de transition sont nuls, mais cela fait quand même beaucoup d'états, et dans le cas eukaryote la situation est bien pire. On peut réduire ce nombre avec un modèle phasé, mais on récupère la même complexité en multipliant par trois le nombre de matrices de transition à estimer. Enfin on pourrait penser travailler sur la suite des codons plutôt que sur celle des nucléotides, mais ceci ne serait pas valable pour les parties non codantes.

On va voir ci–dessous que le modèle semi–markovien permet de réduire le nombre d'états à trois dans le cas prokaryote, en outre qu'il permet de choisir une loi plus réaliste que la loi géométrique pour la longueur des plages codantes.

11.2 Qu'est–ce qu'une chaîne semi–markovienne ?

La réponse dépend des auteurs. Je vais donner ma définition. Comme son nom l'indique, une chaîne semi–markovienne est "un peu moins markovienne" (i.e. oublie un peu moins son passé) qu'une chaîne de Markov. Étant donnée une suite aléatoire $(X_1, \ldots, X_N)$, on définit pour chaque $1 < n < N$ la v. a. η_n de la façon suivante

$$\eta_n = \sup\{k \geq 0,\ X_{n-k} = X_{n-k+1} = \cdots = X_n\}.$$

Dans l'application qui nous intéresse, c'est le nombre de sites à gauche du site n, qui sont dans la même plage que celui–ci. On notera

$$\varphi_n(x_1, \ldots, x_n) = \sup\{k,\ x_{n-k} = \cdots = x_n\}.$$

Donc $\eta_n = \varphi_n(X_1, \ldots, X_n)$.

Définition 11.1 *Une suite aléatoire $(X_1, \ldots, X_N)$ à valeurs dans E est une chaîne semi–markovienne ssi pour tout $1 < n \leq N$, pour tout $(x_1, \ldots, x_{n-1}, x, y) \in E^{n+1}$,*

$$\begin{aligned}&\mathbb{P}(X_{n+1} = y | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x)\\ &\quad = \mathbb{P}(X_{n+1} = y | X_n = x, \eta_n = \varphi_n(x_1, \ldots, x_{n-1}, x)).\end{aligned}$$

Le fait que l'état suivant d'une chaîne semi–markovienne dépende non seulement de l'état courant, mais de la longueur de la "visite" dans cet état jusqu'au site considéré fait que les lois des temps de séjour dans les divers états sont complètement arbitraires.

Plus précisément, une façon "générique" de préciser la loi d'une chaîne semi–markovienne (et aussi d'indiquer comment la simuler) est la suivante.

- D'une part on associe à chaque point $x \in E$ une loi de probabilité $(d_x(n),\ n \in \mathbb{N}\backslash\{0\})$ sur les entiers privés de 0, qui précise les lois des longueurs des "plages" sur lesquelles la chaîne est constante.
- D'autre part on se donne une matrice de transition P sur $E \times E$ d'une chaîne de Markov, dont tous les termes diagonaux sont nuls. Cette matrice décrit suivant quelle loi la chaîne change d'état, quand elle en change.

Voyons comment simuler une chaîne semi–markovienne dont la loi est caractérisée par les données : pour tout $x \in E$, d_x désigne la loi du temps de séjour à l'état x, et P_x. la loi du prochain état visité après l'état x. Si x est le point de départ $(X_1 = x)$, on tire une variable alétoire T_1 à valeurs dans $\mathbb{N}^*$ de loi d_x. Notons n la valeur simulée. Alors $X_1 = X_2 = X_3 = \cdot = X_n = x$. On tire une v.a. Z_1 de loi P_x. sur $E\backslash\{x\}$. Supposons que le résultat du tirage soit $Z_1 = y$. Alors $X_{n+1} = y$, et on recommence en tirant une v. a. T_2 de loi d_y, et une v.a. Z_2 de loi P_y., et ainsi de suite. Tous les tirages successifs sont bien entendu indépendants les uns des autres.

11.3 Le modèle semi–markovien caché

Encore une fois, limitons–nous pour simplifier au cas prokaryote. On considère 3 états cachés, l'état 0 pour *non codant*, l'état 1 pour *codant à l'endroit*, l'état 2 pour *codant à l'envers*. L'état 0 est un état *markovien* (on verra ci–dessous la raison de cette restriction), ce qui veut dire que la loi des longueurs des plages non codantes est une loi géométrique de paramètre q (à estimer). Les états 1 et 2 sont dits *semi–markoviens*. On choisira comme loi des longueurs des plages codantes à l'endroit et à l'envers l'image par l'application $x \to 3x$ d'une loi binomiale négative de paramètres $m \in \mathbb{N}^*$ et $0 < p < 1$ (i. e. cette loi décrit le nombre de codons plutôt que le nombre de nucléotides).

Définition 11.2 *On dit que la v.a. T suit la loi binomiale négative de paramètres m et p si T est le nombre de jets de pile ou face nécessaires pour obtenir exactement m piles, p désignant la probabilité d'obtenir pile à chaque coup. Soit*

$$\mathbb{P}(T = k) = C_{m-1}^{k-1}(1-p)^{k-m}p^k,$$

qui vaut la probabilité d'obtenir $m-1$ piles au cours des $k-1$ premiers coups, multipliée par la probabilité d'obtenir pile au k–ième coup.

La logique voudrait que l'on choisisse comme valeur du paramètre m le plus petit nombre d'acides aminés que contient un gène, plus deux (pour les codons START + STOP). Malheureusement, ce nombre minimal peut être extrêmement réduit dans des cas tout à fait exceptionnels, alors qu'il est de l'ordre de la dizaine hormis ces cas tout à fait exceptionnels. Un choix raisonnable semble être $m = 10$, mais ce choix doit être critiqué–validé par le Biologiste (et/ou confronté à la séquence étudiée).

Quant au paramètre p, il doit être estimé.

Quant à la loi de probabilité qui régit les changements d'état, elle est définie comme suit. On admet que tout plage codante (à l'endroit comme à l'envers) est suivie d'une plage non codante. Donc $P_{10} = P_{20} = 1$. En outre, $P_{01} + P_{02} = 1$, et on peut soit supposer que $P_{01} = 1/2$, soit chercher à estimer cette quantité.

Discutons maintenant de la loi des nucléotides, sachant l'état caché. On peut admettre que dans les plages non codantes, les nucléotides sont i. i. d., la loi commune étant à estimer. Dans une plage codante, on suppose par exemple que les codons sont mutuellement indépendants, le premier prenant ses valeurs dans l'ensemble des codons START possibles, le dernier dans l'ensemble des codons STOP possibles, les autres codons étant i. i. d., à valeurs dans les codons possibles qui codent pour un acide aminé (en particulier ces codons ne peuvent pas prendre comme valeur un des codons STOP). La description de la loi des codons d'une plage codante à l'envers se déduit aisément de celle que nous venons de donner pour les plages codantes à l'endroit.

11.4 Algorithme de Viterbi dans le cas semi–markovien caché

Nous allons maintenant décrire comment l'algorithme de Viterbi s'écrit dans notre nouvelle situation (qui est en fait une situation mixte markov caché – semi–markov caché).

Comme pour les chaînes de Markov cachées, l'idée est de calculer des quantités $\delta_y(n)$, pour tous les états cachés y, et $1 \leq n \leq N$. Pour $y = 0$, on pose comme précédemment

$$\delta_y(n) = \max_{y_1, y_2, \ldots, y_{n-1}} \mathbb{P}_\theta(Y_1 = y_1, \ldots, Y_{n-1} = y_{n-1}, Y_n = y, X_1 = x_1, \ldots, X_n = x_n).$$

Pour $y = 1$ (et de même pour $y = 2$), on pose

$$\delta_y(n) = \max_{y_1, \ldots, y_{n-1}} P_\theta(Y_1 = y_1, ..., Y_{n-1} = y_{n-1}, Y_n = y, Y_{n+1} \neq y, X_1 = x_1, ..., X_n = x_n)$$

La formule de récurrence est la suivante :

$$\delta_y(n) = \max \Big\{ P_\theta(X_1^n = x_1^n | Y_1^n = y, Y_{n+1} \neq y) d_y(n) \mu_y,$$
$$\max_{1 \leq k \leq n-1} \left[P_\theta(X_{n-k+1}^n = x_{n-k+1}^n | Y_{n-k} \neq y, Y_{n-k+1}^n = y, Y_{n+1} \neq y) \max_{z \neq y} [\delta_z(n-k) P_{zy}] d_y(k) \right] \Big\}$$

Nous n'allons pas détailler plus avant les calculs. Remarquons que le fait qu'à chaque étape on ait à calculer un max sur $1 \leq k \leq n$ rend l'algorithme a priori quadratique en N, ce qui est une mauvaise nouvelle. Cependant on peut limiter la portée de ce max, en arguant du fait qu'une région codante se termine au premier codon STOP rencontré. Cette remarque est encore vraie pour un exon dans le cas eukaryote : celui–ci se termine au plus loin lors de la rencontre du premier codon STOP (soit que ce codon marque effectivement la fin du gène, soit qu'il soit situé dans un intron). La même remarque ne s'applique pas aux régions non codantes, d'où le choix de prendre l'état 0 markovien.

12 Chaînes de Markov en temps continu et chaîne de Markov sur les arbres

On présenter les chaînes de Markov sur les arbres, qui sont couramment utilisées comme modèle en phylogénie. On verra au passage l'utilité de la notion de chaîne réversible. Dans ce cadre, la notion naturelle est celle de chaîne de Markov en temps continu, que nous allons introduire.

12.1 Chaîne de Markov en temps continu

Lorsque l'on modélise une séquence de nucléotides, on fait bien sûr appel à une collection indexée par les entiers de variables aléatoires. Lorsque l'on étudie l'évolution au cours du temps de cette même séquence, ou pour simplifier d'un site de cette séquence, il est naturel d'indexer les variables aléatoires par un paramètre t qui varie dans $[0, T]$.

On appelle *chaîne de Markov en temps continu* ou *processus markovien de sauts* une collection $(X_t,\ 0 \leq t \leq T)$ de v.a. à valeurs dans un ensemble fini E (E pourraît être dénombrable, mais le cas fini nous suffira), telle que pour tout n, $0 \leq t_1 < t_2 < \cdots < t_n \leq T$, $(X_{t_1}, \ldots, X_{t_n})$ est une chaîne de Markov. On ne considèrera que le cas homogène, où la transition de X_s à X_t $(s < t)$ ne dépend que de $t - s$.

Pour motiver la description que nous allons donner des chaînes de Markov en temps continu, revenons au cas d'une chaîne en temps discret de matrice de transition P. Considérons un état $x \in E$, tel que $p = P_{xx} > 0$. Alors la loi du temps de séjour de la chaîne à l'état x est la loi géométrique de paramètre p. En effet, si $X_1 = x$, et si $S = \inf\{k,\ X_{k+1} \neq x\}$,

$$\begin{aligned} \mathbb{P}(S = \ell | X_1 = x) &= \mathbb{P}(X_2 = x, \ldots, X_\ell = x, X_{\ell+1} \neq x | X_1 = x) \\ &= p^{\ell-1}(1-p). \end{aligned}$$

Cette propriété est intimement liée au fait que la propriété de Markov impose que la loi de S soit "sans mémoire", au sens où

$$\mathbb{P}(S > k + \ell | S > k) = \mathbb{P}(S > \ell),$$

propriété qui est caractéristique de la loi géométrique. Les lois sur $\mathbb{R}_+$ qui vérifient une propriété analogue sont les lois exponentielles, et il n'est pas difficile de se convaincre que les temps de séjour dans les différents états d'une chaîne de Markov en temps continu suivent des lois exponentielles. On dit que la v. a. S suit la loi exponentielle de paramètre $\lambda > 0$ si

$$\mathbb{P}(S > t) = \exp(-\lambda t), \quad \forall t > 0.$$

Une chaîne de Markov en temps continu est entièrement décrite par son *générateur infinitésimal* Q, qui est une matrice $d \times d$ (avec $d = |E|$), dont tous les termes hors diagonaux sont positifs ou nuls, et les termes diagonaux sont strictement négatifs (ou nul si l'état correspondant est absorbant). Supposons que la chaîne parte de x (i.e. $X_0 = x$). La chaîne reste à l'état x pendant un temps aléatoire T_1, de loi exponentielle de paramètre $q_x = -Q_{xx}$. L'état dans lequel la chaîne aboutit à l'issue du saut qui se produit à l'instant T_1 est choisi indépendemment de la valeur de T_1, suivant la loi de probabilité $(q_x^{-1} Q_{xy},\ y \in E \backslash \{x\})$, et ainsi de suite...

On peut donner une description alternative équivalente à la précédente, comme suit (l'équivalence entre les deux descriptions est un exercice de probabilités que le lecteur est invité à faire). Supposons qu'à l'instant 0 (ou à tout autre instant), la chaîne soit à l'état x. A chaque état $y \neq x$, on associe une v.a. S_y, de loi exponentielle de paramètre Q_{xy}, de telle sorte que les différentes v.a. S_y soient mutuellement indépendantes. La chaîne change d'état à l'instant $T_1 = \inf_{y \neq x} S_y$, et la position de la chaîne juste après ce saut est $\text{argmin}_{y \neq x} S_y$.

On a alors

$$\mathbb{P}(X_t = y | X_0 = x) = P(t)_{xy},$$

où la matrice $P(t)$ est donnée par

$$P(t) = \exp(tQ) = \sum_{k=0}^{\infty} \frac{t^k}{k!} Q^k.$$

La suite des valeurs à l'issue des sauts successifs est une chaîne de Markov appelée la *chaîne incluse*, de matrice de transition P dont les termes diagonaux sont nuls et les termes hors diagonaux sont donnés par $P_{xy} = q_x^{-1} Q_{xy}$. La chaîne en temps continu est dite irréductible ssi sa chaîne incluse l'est, et alors la chaîne en temps continu possède une unique probabilité invariante π, qui est la solution de l'équation $\pi Q = 0$. On a encore le théorème ergodique :

$$\frac{1}{t} \int_0^t f(X_s) ds \to \sum_x f(x) \pi_x \text{ p. s., quand } t \to \infty.$$

Enfin une chaîne en temps continu irréductible est *réversible* ssi la relation d'équilibre ponctuel

$$\pi_x Q_{xy} = \pi_y Q_{yx}, \quad \forall x \neq y$$

est satisfaite. Dire qu'une chaîne en temps continu est réversible, c'est dire que $(X_t,\, 0 \leq t \leq T)$ et $(\hat{X}_t = X_{T-t},\, 0 \leq t \leq T)$ ont même loi, donc évoluent suivant le même mécanisme.

12.2 Chaîne de Markov sur un arbre avec racine

La chaîne part de la racine (qui joue le rôle de l'instant initial 0) dans un certain état, disons x. Elle évolue jusqu'au premier noeud qui se trouve à la distance r de la racine, comme une chaîne en temps continu pendant l'intervalle de temps r. Notons y l'état de la chaîne en ce noeud. Sur chaque branche qui part de ce noeud court une chaîne en temps continu, partant de y, de telle sorte que les différentes chaînes sur les différentes branches sont mutuellement indépendantes, jusqu'au prochain noeud, et ainsi de suite...

12.3 Chaîne de Markov sur un arbre sans racine

Dans le cas d'un arbre sans racine, la même construction peut être faite en partant de n'importe quel point de l'arbre (soit d'un noeud, soit d'un point arbitraire d'une branche arbitraire), à condition d'utiliser une chaîne réversible (i.e. une chaîne irréductible dont le couple générateur infinitésimal – probabilité invariante satisfait la relation d'équilibre ponctuel).

13 Méthodes de vraisemblance en phylogénie

La comparaison des génomes de diverses espèces est maintenant le principal outil pour tenter de reconstruire des arbres phylogénétiques. Il existe plusieurs algorithmes qui construisent de tels arbres. Nous allons donner des indications sur la méthode du maximum de vraisemblance.

Notons que l'on peut comparer des gènes (i.e. des collections d'acides aminés), ou bien des séquences d'ADN. Nous nous limiterons pour fixer les idées aux séquences d'ADN.

13.1 Modèles d'évolution

Pour préciser la vraisemblance d'un arbre au vu des données, il nous faut préciser un modèle d'évolution, qui indique comment ces données ont été "fabriquées" par l'évolution le long des branches de l'arbre, pour chaque site sur l'ADN. Etant donné une loi de probabilité π sur l'espace $E = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, et un nombre positif ν, on définit le générateur infinitésimal Q d'une chaîne de Markov en temps continu à valeurs dans E par

$$Q_{xy} = \begin{cases} \nu(\pi_x - 1), & \text{if } x = y \\ \nu\pi_y, & \text{sinon.} \end{cases}$$

Notons que clairement pour $x \neq y$,

$$\pi_x Q_{xy} = \pi_y Q_{yx},$$

donc π est la probabilité invariante, et la chaîne est réversible. La matrice Q possède deux valeurs propres : $-\nu$, dont l'espace propre associé est constitué des vecteurs orthogonaux à π dans $\mathbb{R}^4$, et 0, dont l'espace propre associé est constitué des vecteurs colinéaires au vecteur $(1,1,1,1)$. Passant à l'exponentielle, on montre aisément que

$$P_{xy}(t) = \left(e^{tQ}\right)_{xy} = e^{-\nu t}\delta_{xy} + (1 - e^{-\nu t})\pi_y.$$

Dans le cas particulier où $\pi = (1/4, 1/4, 1/4, 1/4)$, ce modèle d'évolution est appelé le modèle de Jukes–Cantor.

Le temps t correspond ici à une distance sur l'arbre. Notons que le seul paramètre d'intérêt est le produit νt. Quitte à modifier en conséquence les longueurs des branches de l'arbre, on peut toujours se ramener à $\nu = 1$, ce que nous ferons dans la suite.

13.2 Calcul de la vraisemblance d'un arbre

On va supposer dans cette section que les différents sites évoluent indépendemment les uns des autres, et tous au même taux, ce taux étant également constant dans tout l'arbre. Cette hypothèse n'est pas très réaliste, et beaucoup de travaux récents se concentrent sur la détection des sites qui évoluent plus vite que les autres, éventuellement dans une partie seulement de l'arbre, mais pour démarrer l'étude et construire un premier arbre, il est naturel de faire l'hypothèse simplificatrice que nous venons d'énoncer. Une autre hypothèse assez utilisée est que les taux d'évolution des différents sites sont des v.a. i. i. d., de loi commune une loi Gamma.

L'information à notre disposition, les *données*, est constituée d'un jeu de k séquences alignées, de longueur m, i.e. pour chaque site s, $1 \leq s \leq m$, on a k lettres dans l'alphabet `a, c, g, t`, une pour chaque feuille de l'arbre. A chaque arbre enraciné T possédant k feuilles, on va associer la vraisemblance $L(T)$, fonction des données. La vraisemblance $L(T)$ est un produit de $s = 1$ à m des vraisemblances associées à chaque site s :

$$L(T) = \prod_{s=1}^{m} L_s(T).$$

Chaque $L_s(T)$ se calcule en utilisant la propriété de Markov, comme nous allons maintenant le voir.

Soit T un arbre enraciné. L'arbre est considéré la racine "en bas", les feuilles "en haut". On peut par exemple coder les noeuds d'un tel arbre comme suit, en remontant de la racine vers les feuilles :

- 0 désigne la racine ;
- 1, 2, 3, .. les "fils" de la racine, i.e. les noeuds qui sont directement reliés à la racine par une branche, numérotés dans un ordre arbitraire ;
- 1.1, 1.2, 1.3 ... désignent les fils de 1 ; 2.1, 2.2 .. les fils de 2, .. ;
- et ainsi de suite jusqu'aux feuilles.

Pour tout noeud $\alpha \in T\backslash\{0\}$, on note ℓ_α la longueur de la branche qui joint le "père" de α à α, et on associe à α l'ensemble Λ_α des feuilles du sous–arbre dont α est la racine. En particulier, Λ_0 désigne l'ensemble des feuilles de l'arbre. Si $\alpha \in \Lambda_0$, $\Lambda_\alpha = \{\alpha\}$. Si $\alpha \in T\backslash\Lambda_0$, on définit Γ_α l'ensemble des "fils" de α, notés $\alpha.1, \ldots, \alpha.i(\alpha)$ [i.e. $i(\alpha) = |\Gamma_\alpha|$].

On note $\{X_\alpha,\ \alpha \in T\}$ les nucléotides aux noeuds de l'arbre. On suppose qu'ils constituent les valeurs aux noeuds de l'arbre d'un processus de Markov sur l'arbre de générateur infinitésimal Q. Seules les valeurs des $\{X_\alpha,\ \alpha \in \Lambda_0\}$ sont observées. On note x_α la valeur observée de X_α, pour $\alpha \in \Lambda_0$. La vraisemblance de l'arbre, au vu des nucléotides au site s, est

$$L_s(T) = \mathbb{P}_T\left(\cap_{\alpha\in\Lambda_0}\{X_\alpha = x_\alpha\}\right).$$

On va expliciter le calcul de cette quantité, ce qui mettra en évidence sa dépendance par rapport à l'arbre T.

Pour tout $\alpha \in T$, $x \in E$, on définit $L^{(\alpha)}_{s,x}$, la vraisemblance conditionnelle du sous–arbre dont α est la racine, conditionnée par $X_\alpha = x$, par la récurrence montante suivante.

– Si $\alpha \in \Lambda_0$,

$$L^{(\alpha)}_{s,x} = \begin{cases} 1, & \text{si } x = x_\alpha; \\ 0, & \text{sinon.} \end{cases}$$

– Si $\alpha \in T\backslash\Lambda_0$ est tel que $\Gamma_\alpha \subset \Lambda_0$,

$$L^{(\alpha)}_{s,x} = \prod_{\beta\in\Gamma_\alpha} P_{xx_\beta}(\ell_\beta).$$

– Dans les autres cas,

$$L^{(\alpha)}_{s,x} = \sum_{x_{\alpha.1},\ldots,x_{\alpha.i(\alpha)}\in E} P_{xx_{\alpha.1}}(\ell_{\alpha.1})L^{(\alpha.1)}_{s,x_{\alpha.1}} \times \cdots \times P_{xx_{\alpha.i(\alpha)}}(\ell_{\alpha.i(\alpha)})L^{(\alpha.i(\alpha))}_{s,x_{\alpha.i(\alpha)}}.$$

Ce calcul conduit finalement à préciser les quantités $L^{(0)}_{s,x}$, $x \in E$. Enfin

$$L_s(T) = \sum_{x\in E} \pi_x L^{(0)}_{s,x},$$

et

$$L(T) = \prod_{s=1}^{m} L_s(T).$$

On aurait pu tout aussi bien écrire chaque $L_s(T)$ comme une somme de $4^{|T\backslash\Lambda_0|}$ termes. Mais les formules ci–dessus constituent l'algorithme qu'il faut utiliser en pratique.

13.3 Maximum de vraisemblance

Le calcul du maximum de vraisemblance sur tous les arbres possibles est complexe. La partie la moins difficile consiste à maximiser par rapport aux longueurs des branches. Encore

utilise-t-on un algorithme dont il n'est pas clair qu'il conduit à un maximum global. Celui–ci consiste à maximiser successivement par rapport à chaque longueur de branche, et à itérer la succession des maximisations tant que la vraisemblance augmente. On va voir maintenant que chaque maximisation par rapport à une longueur de branche se fait assez aisément.

Dans la mesure où les $\{X_\alpha,\ \alpha \in T\}$ sont issus d'un processus de Markov *réversible* sur l'arbre, la loi des $\{X_\alpha\}$ ne dépend pas du choix de la racine en n'importe quel noeud de l'arbre (ou plus généralement n'importe où sur une branche arbitraire).

Considérons deux noeuds voisins α et β de l'arbre. Désignons par $\ell_{\alpha\beta}$ la longueur de la branche qui les relie. Si l'on place la racine n'importe où sur cette branche, on définit comme ci–dessus des quantités $L^{(\alpha)}_{s,x}$ et $L^{(\beta)}_{s,y}$, $x, y \in E$. Alors

$$\begin{aligned} L_s(T) &= \sum_{x,y\in E} \pi_x P_{xy}(\ell_{\alpha\beta}) L^{(\alpha)}_{s,x} L^{(\beta)}_{s,y} \\ &= \sum_{x,y\in E} \pi_y P_{yx}(\ell_{\alpha\beta}) L^{(\alpha)}_{s,x} L^{(\beta)}_{s,y}. \end{aligned}$$

Cette procédure permet d'expliciter la dépendance de $L_s(T)$ et de $L(T)$ par rapport à la longueur d'une branche donnée, et de calculer le maximum par rapport à cette longueur. La recherche de ce maximum est en tout cas assez simple dans le cas du modèle d'évolution que nous avons décrit ci–dessus (on maximise le logarithme de $L(T)$, ce qui remplace le produit des $L_s(T)$ par une somme, et simplifie la maximisation).

14 Bibliographie

- J. Felsenstein, *Infering phylogenies*, Sinauer 2004.
- F. Muri–Majoube, B. Prum, Une approche statistique de l'analyse des génomes, *Gazette des Mathématiciens* **89**, 2001, 63–98.
- S. Robin, F. Rodolphe, S. Schbath, *ADN, mots et modèles*, Belin 2003.
- C. Semple, M. Steel, *Phylogenetics*, Oxford Univ. Press, 2003.