Chaînes de Markov et génome

Etienne Pardoux

Leçon 1

Introduction

Le but de ces quelques leçons est de tenter d'expliquer le "pourquoi" et le "comment" des algorithmes basés sur les chaînes de Markov cachées pour l'annotation du génome. On parlera aussi des chaînes semi–markoviennes cachées, et on effleurera d'autres applications.

1 Comment lire l'ADN ?

On considère un fragment d'ADN, sous la forme d'un simple brin constitué d'une succession de 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*, et *régions non codantes* ; dans le cas des génomes eukaryotes il faut en outre découper les *régions codantes* en *introns* et *exons*.

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 $x_1, \ldots, x_n \in E$,*

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

$\Diamond$

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 Chaînes de Markov homogènes 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 $1 \leq k < n$

$\diamondsuit$

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. $\diamondsuit$

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*

$$\mu(k) = \mu P^k,$$

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

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

$\diamondsuit$