Statistique pour la bio-informatique

Séance 3 - 15 Octobre 2003

Modèles de distances évolutives

1 introduction

Les données biologiques de séquences d'ADN à partir desquelles le biologiste tente de valider un grand nombre d'hypothèses résultent d'un processus stochastique indescriptiblement complexe : l'évolution. Des modèles extrêmement simplifiés de ce processus sont souvent utilisés, en particulier pour la reconstruction des phylogénies.

Bien qu'il y ait de grandes variations d'un individu à l'autre dans une population, nous supposerons qu'au niveau moléculaire (à l'echelle du nucléotide) les différences entre deux individus sont très faibles. Sans cette hypothèse, le concept même de *génome humain* n'aurait pas de sens. En fait, deux êtres humains diffèrent d'un nucléotide pour un nombre de sites de l'ordre de 1000. L'expression *nucléotide en un site de l'ADN* signifie donc **nucléotide le plus fréquent en un site de l'ADN** dans une population où une espèce.

Au cours d'une longue période de temps, un nucléotide en un site de l'ADN peut changer. Nous négligeons la courte période pendant laquelle la fixation a lieu (voir la séance précédente) et supposons désormais que le changement est instantané. Le nucléotide prédominant en un site verra son état spécifié par un processus de Markov à temps continu et nous considèrons par exemple la probabilité $p_{ag}(t)$ pour observer le nucléotide g partant de l'état "ancestral" a après une période t.

Lorsque l'on compare deux espèces à travers l'état d'un site, on est souvent à la recherche d'un "ancêtre commun" le plus récent. Dans cette démarche, on remonte le temps en partant de la première espèce jusqu'à l'ancêtre, puis on descend le temps depuis l'ancêtre jusqu'à la seconde espèce. Les modèles de l'évolution ne priviligient en général pas un sens du temps.

Dans cette séance, nous présentons les modèles les plus connus permettant de calculer une distance entre les séquences d'ADN d'un même gène pour deux espèces différentes. La distance en question est reliée au temps qu'il faut remonter pour trouver un ancêtre commun. Nous chercherons à estimer ce temps à partir du nombre de substitutions survenues chez l'une ou l'autre des deux espèces depuis l'ancêtre. Pour cela, nous nous fonderons sur la seule observation du nombre de différences entre les deux séquences étudiées et l'hypothèse d'une horloge moléculaire sera prépondérante.

2 Rappels sur les processus de Markov

Un processus aléatoire $\{X_t\ ;\ t \geq 0\}$ à valeurs entières est un processus de Markov si, pour tout $s, t \geq 0$, $i, j \in \mathbb{N}$ et $x_u \in \mathbb{N}$, $0 \leq u < s$, nous avons

$$\mathrm{P}(X_{t+s} = j \mid X_s = i,\ X_u = x_u,\ \forall 0 \leq u < s) = \mathrm{P}(X_{t+s} = j \mid X_s = i)\,.$$

En d'autres termes, un processus de Markov est un processus ayant la propriété d'indépendance conditionnelle du passé. La loi conditionnelle de la variable future X_{t+s} sachant l'état présent X_s et toute l'histoire du processus jusqu'au temps s ne dépend que du présent et est indépendante du passé. Si, de plus,

$$p_{ij}(s, t) = \mathrm{P}(X_{t+s} = j \mid X_s = i)$$

ne dépend pas de s, le processus $\{X_t\ ;\ t \geq 0\}$ est dit *homogène*.

Il existe une représentation intuitive d'un processsus de Markov homogène quelconque.

1) Pour tout i dans $\mathbb{N}$, le temps de séjour dans l'état i avant d'effectuer une transition vers un autre état est une variable aléatoire de loi exponentielle de paramètre ν_i.

2) Lorsque le processus "quitte" l'état i, il choisit d'aller en $j \neq i$ avec la probabilité p_{ij}. les probabilités de transition vérifient donc

$$p_{ij} \geq 0 \quad ; \quad \sum_{j \neq i} p_{ij} = 1\,.$$

3) Toutes les durées de séjour sont indépendantes.

On pose

$$\forall i \neq j\ , \quad \lambda_{ij} = \nu_i p_{ij}\,.$$

La vitesse avec laquelle le processus sort de l'état i est ν_i et la probabilité avec laquelle il entre dans j est p_{ij}. Ainsi, on peut interpréter λ_{ij} comme le *taux* avec lequel le processus partant de i entre en j. Bien entendu

$$\nu_i = \sum_{j \neq i} \nu_i p_{ij} = \sum_{j \neq i} \lambda_{ij}$$

et

$$\forall i \neq j \;, \quad p_{ij} = \lambda_{ij}/\nu_i \,.$$

Ainsi, la donnée de la famille $\{\lambda_{ij}\,;\, i \neq j\}$ détermine complètement le processus de Markov . Posons, pour tout $i, j \in \mathbb{N}$ et $s, t \geq 0$,

$$\mathrm{P}_{ij}(t) = \; \mathrm{P}(X_{t+s} = j \mid X_s = i)\,.$$

Il s'agit de la probabilité conditionnelle pour que le processus en i au temps s se trouve en j au temps $t+s$. Par homogénéité, cette grandeur ne dépend pas de s. La proposition suivante justifie la dénomination de taux de transition appliquée aux λ_{ij}.

Proposition 2.1 *Pour tout $i \in \mathbb{N}$, $h > 0$*

$$\begin{array}{rcl} 1 - p_{ii}(h) & = & \nu_i h + o(h)\;; \\ p_{ij}(h) & = & \lambda_{ij} h + o(h) \quad \textit{lorsque } j \neq i\,. \end{array}$$

On appelle **générateur infinitésimal** la matrice dont le terme général est λ_{ij} pour $i \neq j$ et $-\nu_i$ pour le terme diagonal d'ordre i. Cette matrice est notée Q dans la suite

$$Q = \left(\begin{array}{cccc} -\nu_0 & \lambda_{01} & \lambda_{02} & \dots \\ \lambda_{10} & -\nu_1 & \lambda_{12} & \dots \\ . & . & & \\ . & . & & \end{array} \right) .$$

Notons $p(t)$ la loi de la variable aléatoire X_t

$$\forall i \in \mathbb{N}, \quad p_i(t) = \; \mathrm{P}(X_t = i)\,.$$

La loi $p(t)$ peut se calculer comme la solution d'un système différentiel linéaire. Les équations de ce système sont appelées **équations de Kolmogorov**.

Théoreme 2.1 Equations de Kolmogorov.

$$\frac{d}{dt} p = pQ\,,$$

Soit

$$\forall j \in \mathbb{N}, \quad \frac{d}{dt} p_j = \sum_{i \neq j} p_i \, \lambda_{ij} \; - \; \nu_j \, p_j\,.$$

On dit qu'une loi de probabilité π est **invariante** ou stationnaire si elle est solution du système d'équations

$$\pi Q = 0 \,.$$

Dans le cas où le nombre d'états du processus est fini, l'existence de chemins reliant les états avec une probabilité non nulle garantit l'unicité de la solution à ce système. Cette hypothèse s'appelle l'irréductibilité. On dit de plus qu'un générateur infinitésimal est π-**réversible** si

$$\forall i, j \in \mathbb{N}, \quad \pi_i \lambda_{ij} = \pi_j \lambda_{ji}.$$

On vérifie facilement que la réversiblité implique que la loi de probabilité π est invariante. Dans le processus associé, il n'y a pas de sens privilégié au temps lorsque la dynamique démarre en régime stationnaire. On peut aussi facilement dans ce cas montrer que

$$\forall t \geq 0, \quad \pi_i p_{ij}(t) = \pi_j p_{ji}(t),$$

ou de manière équivalente en régime stationnaire

$$\forall t \geq 0, \quad \mathrm{P}(X_s = i \cap X_{s+t} = j) = \mathrm{P}(X_s = j \cap X_{s+t} = i).$$

Dans la suite, nous utiliserons des processus réversibles pour décrire les mutations le long des différentes branches d'un arbre d'évolution. Cette hypothèse évite d'affirmer qu'une espèce descend d'une autre. Elle permettra surtout de considérer des arbres non enracinés pour la reconstruction de phylogénies. Cette propriété sera pertinente pour la reconstruction d'arbre par la méthode de maximum de vraisemblance car elle laisse la liberté de choisir la racine sans modifier des propriétés statistiques importantes de l'arbre.

2.1 Le modèle de Jukes-Cantor

Les modèles d'évolution au niveau des nucléotides sont décrits par des processus de Markov à quatre états $\{a, c, t, g\}$. Le modèle de Jukes-Cantor est le plus simple imaginable. Il s'agit d'une chaîne Les taux de substitutions entre nucléotides sont tous égaux à α. La convention est d'écrire les nucléotides dans l'ordre a, g, c, t, regroupant purines et pyrimidines. L'évolution est décrite par un processus de Markov dont le générateur infinitésimal s'écrit

$$Q_{\mathrm{JC}} = \left(\begin{array}{cccc} a & g & c & t \\ -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{array} \right).$$

Les équations différentielles de Kolmogorov

$$\frac{d}{dt}p_{i.} = p_{i.}Q_{\mathrm{JC}}$$

se résument ici à

$$\frac{d}{dt}p_{ij} = \alpha - 4\alpha p_{ij}$$

Étant donné les conditions aux limites, $p_{ii} = 1$, $i = a, g, c, t$, nous obtenons

$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

et

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}, \quad \forall i \neq j.$$

Supposons maintenant que l'on dispose de séquences de longueurs égales à N pour deux espèces différentes supposées descendre d'un ancêtre commun éloigné dans le passé. Supposons que ces séquences résultent d'une évolution selon le modèle de Jukes-Cantor. Il est naturel de définir la distance entre ces deux séquences en relation avec le temps qui les séparent de l'ancêtre commun.

Dans le modèle de Jukes-Cantor, le nombre de substitutions par site suit un processus de Poisson homogène de paramètre $\lambda = 3\alpha$. Puisqu'il y a eu deux lignées indépendantes depuis l'ancêtre commun le plus récent et que nous considérons N sites, le nombre total de substitutions suit un processus de Poisson de paramètre

$$\nu = 2N\lambda = 6N\alpha.$$

Prenons la bifurcation comme instant initial. Les séquences n'ont pas alors divergé et en chaque site, nous avons

$$\sum_i p_{ii}(0) = 1.$$

D'après la remarque précédente, le nombre moyen de substitutions ayant eu lieu au temps t est égal à $6N\alpha t$. Nous allons chercher à relier cette grandeur au nombre de différences observées au temps t entre les deux séquences.

Par symétrie du modèle de Jukes-Cantor, la probabilité $\varphi(t)$ pour que *deux espèces descendant du même ancêtre aient le même nucléotide en un site donné au temps* t ne dépend pas du nucléotide de l'ancêtre et nous avons

$$\varphi(t) = \sum_j p_{ij}(t)^2.$$

Nous avons donc

$$\varphi(t) = \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}.$$

Afin de comparer deux séquences, nous estimons le nombre de nucléotides différents, d'où nous tirons une estimation $\hat{p}$ de la probabilité p que les deux nucléotides diffèrent en divisant par N. Au temps t, nous avons théoriquement

$$p(t) = 1 - \varphi(t) = \frac{3}{4}(1 - e^{-8\alpha t}).$$

En inversant, nous avons

$$6N\alpha t = -\frac{3N}{4}\log(1 - \frac{4}{3}p(t)).$$

Ainsi, un estimateur des moments du nombre de substitutions est

$$\hat{d} = -\frac{3N}{4}\log(1 - \frac{4}{3}\hat{p}).$$

Cet estimateur est souvent appelé *distance de Jukes-Cantor*.

Exercices supplémentaires.

1. On suppose que $\hat{p}$ est petit. Montrer que la variance de $\hat{d}$ est approximativement

$$Var[\hat{d}] \approx Np(1-p)e^{8\nu/3N}$$

2. Montrer que $\hat{d}$ est l'estimateur de maximum de vraisemblance de ν lorsque $\hat{p}$ est la fréquence empirique.

2.2 Le modèle de Kimura

Dans le modèle de Jukes-Cantor, la fréquence d'apparition de chaque nucléotide converge vers 0.25

$$p_{i.}(t) \to (1/4, 1/4, 1/4, 1/4).$$

De plus, il n'y a pas de distinction entre les transversions et les transitions. Ces hypothèses ne sont pas réalistes pour tous les organismes. Par exemple, sur un brin de l'ADN mitochodrial humain, la fréquence de chaque nucléotide est

$$a: \ 0.247 \quad g: \ 0.302 \quad c: \ 0.139 \quad t: \ 0.302.$$

Par ailleurs les transitions sont beaucoup plus fréquentes que les transversions dans les génomes eukaryotes.

Pour résoudre ce type de problème, Kimura introduisit une famille de modèles possédant jusqu'à 6 paramètres. Le modèle de Kimura à deux paramètres est l'un des plus simples. Il se décrit de la manière suivante

$$Q_{\mathrm{K}} = \begin{pmatrix} a & g & c & t \\ 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{pmatrix}.$$

Dans ce cas, la loi stationnaire est encore uniforme (par symétrie). Mais il y a une différence avec le modèle de Jukes-Cantor. Pour un nucléotide ancestral d'un type donné, la probabilité de retrouver le même type (purine ou pyrimidine) est approximativement 1/4, tandis que la probabilité de trouver le type opposé est approximativement 1/2. Le calcul de la distance de Kimura est un peu plus complexe que dans le cas de Jukes-Cantor. On obtient

$$\hat{d}/N = -\frac{1}{2}\log(1-2\hat{p}-\hat{q}) - \frac{1}{4}\log(1-2\hat{q})$$

où $\hat{p}$ est la proportion de transitions observées et $\hat{q}$ est la proportion de transversions observées.

Exercice. Résoudre les équations différentielles de Kolmogorov et vérifier les affirmations précédentes.