

Estimation de la distance génétique entre deux espèces

Mots-clés : chaîne de Markov, estimateur du maximum de vraisemblance, méthode delta.

Introduction

L'information génétique des organismes vivants est portée par les molécules d'ADN. Cette information se transmet de génération en génération, mais peut subir au cours de ces transmissions des variations dues à des erreurs de transcription, variations qui sont à l'origine de l'évolution des espèces.

L'étude de l'ADN permet donc de mesurer cette évolution, en particulier d'estimer la distance séparant des espèces actuelles distinctes dont on suppose qu'elles ont eu dans le passé un ancêtre commun. On s'intéresse plus précisément dans ce texte à deux espèces dérivant par mutations d'un ancêtre commun à partir duquel elles ont évolué indépendamment l'une de l'autre. Le but de l'étude est d'estimer le temps écoulé depuis cette divergence, sachant qu'on ne peut observer que les espèces présentes et non l'ancêtre commun dont on ne sait a priori rien.

1. Le modèle

Une molécule d'ADN est composée de séquences de nucléotides, caractérisés par un type de base azotée. Il existe quatre types de bases azotées distinctes, notées A (adénine), C (cytosine), G (guanine), T (thymine). Un brin d'ADN peut donc être vu comme une suite de sites en lesquels figurent des lettres, prises parmi A, C, G, T. Ces molécules se reproduisent de génération en génération, permettant ainsi la transmission de l'information génétique. Les mutations sont dues principalement à des erreurs lors de la réplication de ces séquences (ajout, délétion ou substitution de nucléotides). On ne s'intéressera dans ce texte qu'au cas des substitutions.

Le modèle le plus simple consiste à supposer qu'à chaque réplication la probabilité d'une substitution en un site donné est constante, égale à un réel $\alpha > 0$, la mutation s'effectuant avec équiprobabilité vers l'une des trois autres bases indépendamment de tout le passé du processus. On supposera de plus que les processus de substitution en des sites distincts sont indépendants.

Le but est d'étudier la distance génétique entre deux espèces relativement proches A et B dont on pense qu'elles ont eu dans le passé un ancêtre commun à partir duquel elles ont divergé à une certaine époque, évoluant ensuite indépendamment l'une de l'autre. On ne sait rien de cet ancêtre ni du temps de divergence (c'est précisément lui qu'on veut estimer).

On observe pour cela deux séquences d'ADN fonctionnellement homologues de l'espèce A et de l'espèce B, par exemple :

A T T C ... G A A

pour A et

G T T C ... G A T

pour B, et on se propose d'estimer à partir de cette comparaison le temps séparant chacune de ces espèces de leur ancêtre commun (si beaucoup de mutations se sont produites, on peut penser que cette distance est grande).

2. Une chaîne de Markov à temps discret

On s'intéresse dans cette partie à la variation de la base en un site donné. On discrétise le temps et on note X_n la base en ce site à l'instant n .

Les hypothèses introduites au début du texte amènent à considérer la suite $(X_n)_{n \geq 0}$ comme une chaîne de Markov homogène d'espace d'états $E = \{A, C, G, T\}$ (qu'on peut aussi prendre égal à $\{1, 2, 3, 4\}$ en numérotant ces états) et de matrice de transition

$$P = \begin{pmatrix} 1 - \alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1 - \alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1 - \alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1 - \alpha \end{pmatrix}.$$

Cette chaîne est irréductible apériodique. Elle admet une unique probabilité invariante π , qui est la loi uniforme sur E .

Elle est de plus réversible, i.e. vérifie $\pi_i P_{i,j} = \pi_j P_{j,i}$ pour tout couple (i, j) d'états. Cette réversibilité traduit la réversibilité temporelle du processus de mutation : elle équivaut à

$$P_\pi(X_0 = i_0, X_1 = i_1, \dots, X_n = i_n) = P_\pi(X_n = i_0, X_{n-1} = i_1, \dots, X_0 = i_n)$$

pour tout entier n et tout $(n + 1)$ -uplet (i_0, i_1, \dots, i_n) d'états, où l'on a noté P_π la probabilité pour la chaîne stationnaire, i.e. de loi initiale π .

La matrice P est diagonalisable, et on vérifie aisément que sa puissance n -ième P^n , qui représente la matrice de transition en n pas de la chaîne, est de la forme

$$P^n = \begin{pmatrix} r(n) & s(n) & s(n) & s(n) \\ s(n) & r(n) & s(n) & s(n) \\ s(n) & s(n) & r(n) & s(n) \\ s(n) & s(n) & s(n) & r(n) \end{pmatrix}$$

où $s(n) = \frac{1}{4} - \frac{1}{4} \left(1 - \frac{4}{3}\alpha\right)^n$.

La probabilité $P(X_n \neq X_0)$ que la base ait changé au site considéré entre les générations 0 et n est $p_n = 3s(n)$, d'où on tire

$$n = \frac{\ln\left(1 - \frac{4}{3}p_n\right)}{\ln\left(1 - \frac{4}{3}\alpha\right)}$$

d'où l'approximation, pour α petit :

$$n \sim -\frac{3}{4\alpha} \ln\left(1 - \frac{4}{3}p_n\right).$$

L'espérance du nombre de substitutions au site considéré au cours des n premiers pas de la chaîne est $K = n\alpha$, d'où l'approximation, pour α petit :

$$K \sim -\frac{3}{4} \ln \left(1 - \frac{4}{3} p_n \right).$$

Si on observe, non plus seulement un site, mais un grand nombre N de sites, en supposant que les processus de substitutions $X_n^{(k)}$ ($k = 1, \dots, N$) en ces différents sites sont indépendants et suivent tous la même loi, on peut estimer p_n par la fréquence $F_n = \frac{1}{N} \sum_{k=1}^N \mathbf{1}_{X_n^{(k)} \neq X_0^{(k)}}$ d'observation d'une substitution en ces N sites et ainsi obtenir un estimateur de K en remplaçant p_n par F_n dans la formule précédente.

On remarque en particulier que si les séquences aux instants 0 et n sont proches (F_n petit), K est proche de F_n comme on pouvait s'y attendre.

3. Du temps discret au temps continu

Si la probabilité α de substitution en un pas en un site donné est faible, mais si on observe la chaîne sur une grande échelle de temps, on peut approcher le processus en temps discret par une chaîne de Markov à temps continu. Plus précisément, si on suppose que $\alpha = \frac{\lambda}{n}$ pour un certain réel $\lambda > 0$, et si on s'intéresse aux transitions sur un temps nt , i.e. pour $\lfloor nt \rfloor$ pas, où $\lfloor x \rfloor$ désigne la partie entière de x , on voit que les transitions se font de i vers i avec une probabilité tendant vers $\frac{1}{4}(1 + 3e^{-4\lambda t/3})$ et de i vers un état distinct j avec une probabilité tendant vers $\frac{1}{4}(1 - e^{-4\lambda t/3})$ quand n tend vers l'infini.

On est ainsi amené à introduire la matrice

$$Q = \lambda \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}.$$

La matrice de transition $P^{\lfloor nt \rfloor}$ est alors proche, quand n est grand, de

$$P^{(t)} = e^{tQ} = \begin{pmatrix} \rho(t) & \sigma(t) & \sigma(t) & \sigma(t) \\ \sigma(t) & \rho(t) & \sigma(t) & \sigma(t) \\ \sigma(t) & \sigma(t) & \rho(t) & \sigma(t) \\ \sigma(t) & \sigma(t) & \sigma(t) & \rho(t) \end{pmatrix}$$

où $\rho(t) = \frac{1}{4}(1 + 3e^{-4\lambda t/3})$ et $\sigma(t) = \frac{1}{4}(1 - e^{-4\lambda t/3})$.

On peut voir ainsi l'évolution au cours du temps de la base en un site donné : les temps de substitution sont les instants de saut d'un processus de Poisson d'intensité λ et à chaque instant de saut de ce processus de Poisson, le saut s'effectue avec équiprobabilité vers un des trois types possibles, indépendamment de tout le passé du processus.

Ici encore la probabilité uniforme π sur E est l'unique probabilité stationnaire pour la chaîne à temps continu, i.e. l'unique probabilité vérifiant $\pi Q = 0$.

La chaîne à temps continu est elle aussi réversible, ce qui signifie que $\pi_i q_{i,j} = \pi_j q_{j,i}$ pour tout couple (i, j) d'états, et traduit la réversibilité temporelle du processus de substitution.

4. Estimation de la distance génétique

On revient à la situation décrite dans l'introduction. On dispose d'échantillons d'ADN de deux espèces A et B dont on pense qu'elles ont eu dans le passé un ancêtre commun. On fait l'hypothèse que ces deux espèces ont évolué indépendamment l'une de l'autre à partir de cet ancêtre commun et on voudrait estimer leur distance génétique en observant deux séquences d'ADN fonctionnellement homologues de longueur N de ces deux espèces.

En prenant comme origine du temps l'instant de divergence, on considère donc pour chaque site k deux processus $(X_t^{(k)})_t$ et $(Y_t^{(k)})_t$ à temps continu vérifiant $X_0^{(k)} = Y_0^{(k)}$ qui évoluent indépendamment suivant la loi décrite dans la partie 3. On a donc, pour tout couple (i, j) d'états et tout site k , $P(X_t^{(k)} = j \mid X_0^{(k)} = i) = P(Y_t^{(k)} = j \mid Y_0^{(k)} = i)$, et on suppose que $X_0^{(k)} = Y_0^{(k)}$ suit la loi stationnaire π . On suppose de plus que les processus correspondant à des sites différents sont indépendants.

On définit la distance génétique entre A et B comme

$$K = 2t \sum_{i \in E} \pi_i q_i = 2\lambda t$$

où $q_i = -q_{i,i}$ est le taux de substitution, i.e. le nombre moyen de substitutions par site et par unité de temps, et t l'instant présent. Le nombre K représente donc le nombre moyen de substitutions par site entre A et B quand on commence par remonter l'arbre généalogique de A vers l'ancêtre commun, puis qu'on le redescend de cet ancêtre commun vers B.

On peut observer pour chaque site son état actuel dans les espèces A et B et donc, pour tout couple (i, j) d'états, la variable aléatoire

$$N_{i,j} = \sum_{k=1}^N \mathbf{1}_{\{i\}}(X_t^{(k)}) \mathbf{1}_{\{j\}}(Y_t^{(k)})$$

représentant le nombre de sites dans l'état i pour A et dans l'état j pour B, et donc la variable aléatoire

$$D_N = \sum_{i \neq j} N_{i,j}$$

représentant le nombre de sites en lesquels les bases sont différentes pour les deux espèces.

La variable aléatoire D_N suit la loi binomiale de paramètres N et

$$p = \frac{3}{4}(1 - e^{-8\lambda t/3}) = \frac{3}{4}(1 - e^{-4K/3}).$$

On a donc

$$P(D_N = d) = \binom{N}{d} p^d (1-p)^{N-d}$$

pour tout entier $d \in \{0, \dots, N\}$.

L'estimateur \widehat{K}_N du maximum de vraisemblance pour K est alors

$$\widehat{K}_N = -\frac{3}{4} \ln \left(1 - \frac{4D_N}{3N} \right).$$

Cet estimateur est convergent : \widehat{K}_N converge presque sûrement vers K quand N tend vers l'infini. On peut montrer de plus qu'il est asymptotiquement normal : sa loi est approximativement normale quand N est grand. Cette normalité résultera de la proposition suivante, appelée en statistique méthode delta :

Proposition : Soit $(Y_n)_n$ une suite de variables aléatoires réelles, θ un réel et $(r_n)_n$ une suite de réels positifs tendant vers l'infini tels que $r_n(Y_n - \theta)$ converge en loi vers une variable aléatoire suivant la loi normale $N(0, \sigma^2)$. Alors, pour toute fonction g de \mathbb{R} dans \mathbb{R} dérivable en θ , $r_n(g(Y_n) - g(\theta))$ converge en loi vers une variable aléatoire suivant la loi normale $N(0, g'(\theta)^2 \sigma^2)$.

Démonstration : L'hypothèse implique que Y_n converge en probabilité vers θ . Soit alors h la fonction de \mathbb{R} dans \mathbb{R} définie par

$$g(x) = g(\theta) + (x - \theta) g'(\theta) + (x - \theta) h(x)$$

pour $x \neq \theta$, $h(\theta) = 0$. La fonction h est continue en θ ; il en résulte que $h(Y_n)$ tend en probabilité vers 0. Par ailleurs, $r_n(Y_n - \theta) g'(\theta)$ converge en loi vers une variable aléatoire suivant la loi normale $N(0, g'(\theta)^2 \sigma^2)$. En appliquant successivement la forme multiplicative, puis la forme additive du lemme de Slutsky, on en déduit que $r_n(Y_n - \theta) h(Y_n)$ converge en loi (ou en probabilité) vers 0, puis que $r_n(g(Y_n) - g(\theta))$ converge en loi vers une variable aléatoire suivant la loi normale $N(0, g'(\theta)^2 \sigma^2)$.

En appliquant la proposition avec $Y_N = D_N/N$, $r_N = \sqrt{N}$, $\sigma^2 = p(1-p)$ et $g(x) = -\frac{3}{4} \ln \left(1 - \frac{4x}{3}\right)$, on obtient que \widehat{K}_N suit approximativement la loi normale de moyenne K et de variance $\frac{9p(1-p)}{N(3-4p)^2}$.

Suggestions de développements

- On pourra détailler les propriétés de la chaîne de Markov de la partie 2, en particulier la réversibilité, justifier le calcul de P^n et préciser les propriétés asymptotiques de cette chaîne.
- On pourra simuler la chaîne de Markov de la partie 2 pour diverses valeurs de n et comparer la fréquence F_n de substitution observée et le nombre total de substitutions au cours des n premiers pas pour les N sites observés. Si N est grand (de l'ordre de 1000, par exemple) et si on tire à chaque pas un site au hasard sur lequel on effectue une substitution aléatoire, les processus de substitutions en les différents sites ne sont plus indépendants; la formule reste-t-elle approximativement valable dans cette situation?
- Détailler le passage du temps discret au temps continu esquissé dans la partie 3, en expliquant en particulier l'intervention du processus de Poisson dans ce modèle.
- Expliciter, pour le modèle à temps continu de la partie 3, la loi de X_t en fonction de la loi de X_0 et de t .
- Justifier l'interprétation de K donnée dans la partie 4.
- Expliciter le calcul de l'estimateur du maximum de vraisemblance \widehat{K}_N .
- Implémenter l'estimateur de K pour le modèle à temps discret ou pour le modèle à temps continu. Vérifier la normalité asymptotique de la loi de \widehat{K}_N en effectuant un grand nombre de simulations.

Références :

S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences, *Lectures on Mathematics in the Life Sciences*, 17, 1986.

M. Cristianini, M.W. Hahn, *Introduction to computational genomics : a case studies approach*, Cambridge University Press, 2007.