Modèles markoviens d'évolution et Phylogénie

Etienne Pardoux

1 Introduction

On va 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.

On passera ensuite en revue les divers modèles couramment utilisés pour modéliser l'évolution des séquences d'ADN, et on examinera la calcul de la vraisemblance d'un arbre.

2 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.

3 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...

4 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).

5 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. On va décrire pusieurs modèles markoviens d'évolution de l'ADN, en précisant les taux dde transition d'un nucléotide à l'autre. C'est à dire que l'on précisera la matrice Q sous la forme

$$Q \quad = \quad \begin{array}{c|cccc} & \mathtt{a} & \mathtt{c} & \mathtt{g} & \mathtt{t} \\ \mathtt{a} & x & x & x & x \\ \mathtt{c} & x & x & x & x \\ \mathtt{g} & x & x & x & x \\ \mathtt{t} & x & x & x & x \end{array}$$

5.1 Le modèle de Jukes–Cantor (1969)

C'est le plus simple, qui suppose que toutes les mutations se font au même taux, i.e. pour un certain $\alpha > 0$,

$$Q \quad = \quad \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}.$$

La probabilité invariante associée est la probabilité uniforme sur les 4 nucléotides. Les probabilités de transition se calculent explicitement

$$P(t) = \begin{pmatrix} 0,25+0,75e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} \\ 0,25-0,25e^{-4\alpha t} & 0,25+0,75e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} \\ 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25+0,75e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} \\ 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25-0,25e^{-4\alpha t} & 0,25+0,75e^{-4\alpha t} \end{pmatrix}.$$

5.2 Les modèles de Kimura (1980 et 1981)

Il est raisonnable de supposer que les transitions (remplacement d'une purine par une autre, ou d'une pyrimidine par une autre) sont plus fréquentes que les transversions (remplacement d'une purine par une pyrimidine ou vice versa). Donc on est amené à supposer que les taux de sustitution entre `a` et `g` ou entre `c` et `t` sont plus élevés que tous les autres, d'où le modèle

$$Q = \begin{pmatrix} -2\alpha-\beta & \alpha & \beta & \alpha \\ \alpha & -2\alpha-\beta & \alpha & \beta \\ \beta & \alpha & -2\alpha-\beta & \alpha \\ \alpha & \beta & \alpha & -2\alpha-\beta \end{pmatrix}.$$

La probabilité invariante est encore la probabilité uniforme. Les probabilités de transition sont données par

$$P_{xx}(t) = 0,25+0,25e^{-4\beta t}+0,5e^{-2(\alpha+\beta)t},$$
$$P_{xy}(t) = 0,25+0,25e^{-4\beta t}-0,5e^{-2(\alpha+\beta)t},$$

si $x \neq y$ sont soit tous deux des purines, soit tous deux des pyrimidines,

$$P_{xy} = 0,5-0,5e^{-4\beta t}$$

dans le dernier cas.

Kimura a proposé un second modèle, de la forme

$$Q = \begin{pmatrix} -\alpha-\beta-\gamma & \alpha & \beta & \gamma \\ \alpha & -\alpha-\beta-\gamma & \gamma & \beta \\ \beta & \gamma & -\alpha-\beta-\gamma & \alpha \\ \gamma & \beta & \alpha & -\alpha-\beta-\gamma \end{pmatrix},$$

pour lequel la probabilité invariante est encore la probabilité uniforme.

5.3 Le modèle de Felsenstein

Etant donné une loi de probabilité π sur l'espace $E = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, et un nombre positif u, Felsenstein propose

$$Q = \begin{pmatrix} u(\pi_{\mathtt{a}}-1) & u\pi_{\mathtt{c}} & u\pi_{\mathtt{g}} & u\pi_{\mathtt{t}} \\ u\pi_{\mathtt{a}} & u(\pi_{\mathtt{c}}-1) & u\pi_{\mathtt{g}} & u\pi_{\mathtt{t}} \\ u\pi_{\mathtt{a}} & u\pi_{\mathtt{c}} & u(\pi_{\mathtt{g}}-1) & u\pi_{\mathtt{t}} \\ u\pi_{\mathtt{a}} & u\pi_{\mathtt{c}} & u\pi_{\mathtt{g}} & u(\pi_{\mathtt{t}}-1) \end{pmatrix}.$$

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 : $-u$, 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

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

Dans le cas particulier où $\pi = (1/4, 1/4, 1/4, 1/4)$, ce modèle se réduit à celui de Jukes–Cantor.

5.4 Le modèle d'Hasegawa, Kishino, Yano (1985)

Il s'agit d'une généralisation à la fois du premier modèle de Kimura, et de celui de Felsenstein. Soit à nouveau π une probabilité sur E, et u, v deux nombres positifs.

$$Q = \begin{pmatrix} 1 - u\pi_{\mathtt{g}} - v\pi_2 & v\pi_{\mathtt{c}} & u\pi_{\mathtt{g}} & v\pi_{\mathtt{t}} \\ v\pi_{\mathtt{a}} & 1 - u\pi_{\mathtt{t}} - v\pi_1 & v\pi_{\mathtt{g}} & u\pi_{\mathtt{t}} \\ u\pi_{\mathtt{a}} & v\pi_{\mathtt{c}} & 1 - u\pi_{\mathtt{a}} - v\pi_2 & v\pi_{\mathtt{t}} \\ v\pi_{\mathtt{a}} & u\pi_{\mathtt{c}} & v\pi_{\mathtt{g}} & 1 - u\pi_{\mathtt{c}} - v\pi_1 \end{pmatrix},$$

où $\pi_1 = \pi_{\mathtt{a}} + \pi_{\mathtt{g}}$, $\pi_2 = \pi_{\mathtt{c}} + \pi_{\mathtt{t}}$. A nouveau π est la probabilité invariante. On peut encore donner une expression explicite pour $P(t)$.

Il y a de bonnes raisons de supposer que $\pi_{\mathtt{c}} = \pi_{\mathtt{g}}$ et $\pi_{\mathtt{a}} = \pi_{\mathtt{t}}$, puisque l'ADN est une molécule à deux brins, constitué de paires `c` : `g` et `a` : `t`. L'égalité ci–dessus est une conséquence de l'hypothèse assez naturelle que l'évolution est la même sur les deux brins. Le modèle HKY, avec cette restriction, devient un modèle avec trois paramètres, à savoir u, v et $\theta = \pi_{\mathtt{c}} + \pi_{\mathtt{g}}$, qui a été proposé par Tamura en 1992.

5.5 Le modèle réversible général

Comme le cardinal de E est très petit, on peut chercher à utiliser le modèle le plus général. Tavaré a proposé une paramétrisation du modèle le plus général, qui prend la forme

$$Q = \begin{pmatrix} -uW & uA\pi_{\mathtt{c}} & uB\pi_{\mathtt{g}} & uC\pi_{\mathtt{t}} \\ uD\pi_{\mathtt{a}} & -uX & uE\pi_{\mathtt{g}} & uF\pi_{\mathtt{t}} \\ uG\pi_{\mathtt{a}} & uH\pi_{\mathtt{c}} & -uY & uI\pi_{\mathtt{t}} \\ uJ\pi_{\mathtt{a}} & uK\pi_{\mathtt{c}} & uL\pi_{\mathtt{g}} & -uZ \end{pmatrix},$$

où u est un paramètre positif, π la probabilité invariante,

$$\begin{aligned} W &= A\pi_{\mathtt{c}} + B\pi_{\mathtt{g}} + C\pi_{\mathtt{t}} \\ X &= D\pi_{\mathtt{a}} + E\pi_{\mathtt{g}} + F\pi_{\mathtt{t}} \\ Y &= G\pi_{\mathtt{a}} + H\pi_{\mathtt{g}} + I\pi_{\mathtt{t}} \\ Z &= J\pi_{\mathtt{a}} + K\pi_{\mathtt{c}} + L\pi_{\mathtt{g}}, \end{aligned}$$

et les paramètres $A, B, \ldots, L$ sont douze paramètres libres. Comme on le verra ci–dessous, il est important pour le calcul du maximum de vraisemblance que le modèle soit réversible. La contrainte de réversibilité impose six contraintes, à savoir

$$A = D,\ B = G,\ C = J,\ E = H,\ F = K,\ I = L.$$

Il reste donc six paramètres à choisir, par exemple A, B, C, E, F et I. Il y a en outre les 3 paramètres de la probabilité invariante, cela fait donc en tout 9 paramètres à choisir.

5.6 Modèles de codons

Un codon est un triplet de nucléotides qui code pour un acide aminé. Parmi les $4^3 = 64$ codons possible, 3 sont des codons STOP possibles, les 61 autres codent pour les 20 acides aminés. Donc parmi les mutations possibles de codons, il faut distinguer les mutations synonymes (qui transforment un codon en un autre qui code pour le même acide aminé) des mutations non synonymes. Les dernières sont soit freinées, soit favorisées par la sélection, alors que les changements synonymes s'accumulent au taux des mutations. En général, le rapport mutations synonymes / mutation non synonymes est plus grand que 1.

Goldman et Yang on proposé en 1994 un modèle comportant 63 paramètres, à savoir 60 paramètres pour les fréquences π_{xyz}, le taux de transition α, le taux de transversion β, et le rapport

$$\omega = \text{taux des mutations non synonymes / taux des mutations synonymes.}$$

Le modèle GY s'écrit

$$Q_{(x_1y_1z_1)(x_2y_2z_2)} = \begin{cases} 0 & \text{si les codons 1 et 2 diffèrent par plus d'une base,} \\ \alpha\pi_{x_2y_2z_2} & \text{pour une transition synonyme,} \\ \beta\pi_{x_2y_2z_2} & \text{pour une transversion synonyme,} \\ \omega\alpha\pi_{x_2y_2z_2} & \text{pour une transition non synonyme,} \\ \omega\beta\pi_{x_2y_2z_2} & \text{pour une transversion non synonyme.} \end{cases}$$

Notons que sur les 63 paramètres à estimer, les 60 paramètres de la probabilité invariante π ne sont en général pas estimés par le maximum de vraisemblance, mais par les fréquences empiriques des divers codons dans les données. Une autre possibilité est d'estimer π_{xyz} par le produit $\pi_x^1\pi_y^2\pi_z^3$ des fréquences des divers nucléotides aux positions 1, 2 et 3 du codon.

5.7 Modèles non homogènes

Une hypothèse implicite dans les modèles markoviens considérés jusqu'ici est la stationnarité. Le générateur infinitésimal est le même sur les diverses branches de l'arbre phylogénétique. Donc la probabilité invariante est la même sur les diverses branches, autrement dit les diverses séquences doivent avoir aproximativement la même composition en bases. Certaines données contredisent nettement cette situation. On peut alors relâcher l'hypothèse d'homogénéité du processus de Markov sur tout l'arbre. Galtier et Gouy adoptent le modèle de Tamura, avec des paramètres α et β homogènes sur tout l'arbre, et un paramètre θ (qui règle la proportion de $\mathsf{g} + \mathsf{c}$) qui peut varier d'une branche à l'autre de l'arbre.

5.8 Dépendance ou indépendance entre les sites

La plupart des modèles markoviens supposent que le comportement des divers sites au cours de l'évolution est un comportement i.i.d. Evidemment cette hypothèse n'est pas raisonnable, mais elle simplifie grandement les calculs.

Il y a à ce jour très peu de travaux qui proposent des modèles markoviens où les évolutions des différents sites sont corrélées. Citons le travail de G. Didier, qui propose un modèle où l'évolution de chaque site dépend des sites voisins. Indiquons une autre approche, utilisée par Pollock, Taylor et Goldman pour modéliser l'évolution de séquences de protéines.

Considérons un modèle du type

$$Q_{xy} = s_{xy}\pi_y,$$

qui est un modèle réversible, si $s_{xy} = s_{yx}$. On propose alors de modéliser l'évolution d'une paire de protéines en choisissant un générateur infinitésimal de la forme

$$\begin{aligned} Q_{xx',yx'} &= s_{xy}\overline{\pi}_{yx'}, \\ Q_{xx',xy'} &= s_{x'y'}\overline{\pi}_{xy'}, \\ Q_{xx',yy'} &= 0, \text{ si } x \neq y \text{ et } x' \neq y', \end{aligned}$$

où $\overline{\pi}$ est une probabilité invariante sur l'ensemble des paires de protéines.

5.9 Variation du taux d'évolution entre branches

Etant donné un générateur infinitésimal Q, pour tout $u > 0$, uQ est encore un générateur infinitésimal. Supposons que Q est constant sur tout l'arbre. Si u est lui aussi constant sur l'arbre, puisque les feuilles (les espèces d'aujourd'hui) sont équidistantes (les distances sont mesurées en temps) de l'ancêtre commun situé à la racine, alors on est dans la situation de l'hypothse d'une "horloge moléculaire". Certains jeux de données sont incompatibles avec une telle hypothèse. On doit alors, pour utiliser un modèle cohérent avec de telles données, permettre au paramètre u de prendre une valeur différente sur chaque branche de l'arbre. On a donc un nouveau paramètre par branche, ce qui fait au total beaucoup de paramètres.

Un autre point de vue est de supposer que u est la valeur prise par un processus stochastique, qui évolue sur l'arbre comme un processus de Markov, soit en temps continu, soit en temps discret (et alors la valeur du processus est constante sur chaque branche, les changements se produisant aux noeuds). Conditionnellement en les valeurs prises par ce processus, les divers nucléotides évoluent comme des processus de Markov non homogènes sur l'arbre. On est alors dans un cadre bayésien, qui se prête à des calculs grâce aux méthodes de simulation dites "Monte Carlo par Chaînes de Markov" (en Anglais MCMC, Markov Chains Monte Carlo).

5.10 Variation du taux d'évolution entre sites

Le modèle le plus fréquent de variation de taux entre sites est de supposer que les taux associés aux divers sites sont i.i.d., de loi commune une loi gamma (ou une discrétisation de cette loi).

Une autre approche, dûe à Felsenstein et Churchill consiste à supposer que les taux forment, le long de la séquence d'ADN considérée, une chaîne de Markov (qui est en fait "cachée"), à valeurs dans un ensemble de cardinal petit (pour des raisons pratiques).

5.11 Modèles dit "covarion"

Il s'agit de modèles où le taux d'évolution est non seulement différent d'un site à l'autre, mais aussi, pour un site donné, d'une branche à l'autre de l'arbre. Covarion est un acronyme pour "COncomitantly VARIable codON".

Posons $E = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, $G=$ l'ensemble des valeurs possibles pour le taux u. Galtier considère en chaque site un processus de Markov indépendant, à valeurs dans $E \times G$.

6 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.

6.1 Calcul de la vraisemblance d'un arbre

Pour fixer les idées, supposons que l'on utilise le modèle de Felsenstein. Le temps t correspond ici à une distance sur l'arbre. Notons que le seul paramètre d'intérêt est le produit $u \times t$. Quitte à modifier en conséquence les longueurs des branches de l'arbre, on peut toujours se ramener à $u = 1$, ce que nous supposons dorénavant.

Nous ne considérerons dans la suite que des it arbres binaires.

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 binaire 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 sont les "fils" de la racine, i.e. les noeuds qui sont directement reliés à la racine par une branche ;
- 1.1, 1.2 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 note $\Gamma_\alpha = \{\alpha.1, \alpha.2\}$ les "fils" de α.

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 $L_{s,x}^{(\alpha)}$, la vraisemblance conditionnelle du sous–arbre dont α est la racine, conditionnée par $X_\alpha = x$, par la récurrence montante suivante.

– Si $\alpha \in \Lambda_0$,

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

– Sinon,

$$L_{s,x}^{(\alpha)} = \sum_{x_{\alpha.1}, x_{\alpha.2} \in E} P_{x x_{\alpha.1}}(\ell_{\alpha.1}) L_{s,x_{\alpha.1}}^{(\alpha.1)} \times P_{x x_{\alpha.2}}(\ell_{\alpha.2}) L_{s,x_{\alpha.2}}^{(\alpha.2)}.$$

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

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

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 $4^{|T \setminus \Lambda_0|}$ termes. Mais les formules ci–dessus constituent l'algorithme qu'il faut utiliser en pratique.

6.2 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 $L_{s,x}^{(\alpha)}$ et $L_{s,y}^{(\beta)}$, $x, y \in E$. Alors

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

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).

7 L'approche bayésienne en Phylogénie

Reprenons l'écriture de la vraisemblance. Notons D le vecteur des variables aléatoires qui sont observées, et d le vecteur des valeurs observées (d comme "données"), i. e. d est constitué des diverses séquences génomiques alignées.

Précisons maintenant les paramètres dont dépend la vraisemblance. Dans les paramètres inconnus (que l'on cherche à préciser), il y

• d'une part la forme de l'arbre, que nous noterons τ, qui est une inconnue dans un ensemble fini $\mathcal{T}$ (de cardinal $(2n-3)!!$ dans le cas d'un arbre enraciné avec n feuilles, $(2n-5)!!$ dans le cas sans racine),

• d'autre part les longueurs des diverses branches, et la matrice de transition Q du modèle d'évolution (ou du moins les paramètres de cette matrice autres que la probabilité invariante). Les longueurs de branche et les paramètres inconnus de la matrice Q varient dans une partie d'un espace euclidien $V \subset \mathbb{R}^d$. On notera cet ensemble de paramètres λ.

Le paramètre inconnu est donc le couple $\theta = (\tau, \lambda)$, dont la valeur est arbitraire dans $\Theta = \mathcal{T} \times V$, et la vraisemblance est la fonction

$$L(\theta) = \mathbb{P}_\theta(D = d).$$

La vraisemblance de la valeur θ du paramètre inconnu est la probabilité d'observer les données que nous avons sous les yeux, si θ est la vraie valeur de ce paramètre.

Dans le point de vue bayésien, le paramètre inconnu θ est la réalisation d'une variable aléatoire, autrement dit (τ, λ) est la réalisation d'un vecteur aléatoire (T, Λ). Prendre ce point de vue impose de se donner une *loi de probabilité a priori*, dont le bayésien nous dit qu'elle permet d'intégrer des informations a priori sur le paramètre inconnu, ce que refusent les anti–bayésiens, pour qui la seule information est apportée par les données.

On va donc se donner une loi *a priori* pour le vecteur (T, Λ), que l'on prendra de la forme suivante :

• on se donne la loi de T, qui est une loi sur un ensemble fini $\mathcal{T}$, donc on se donne des $\alpha_\tau = \mathbb{P}(T = \tau)$, $\tau \in \mathcal{T}$;

• on se donne la loi conditionnelle de Λ, sachant la valeur de T, et on suppose que pour tout $\tau \in \mathcal{T}$, la loi conditionnelle de Λ, sachant que $T = \tau$, admet une densité $q_\tau(\lambda)$, autrement dit pour toute fonction mesurable

$$f : \mathcal{T} \times V \ \to \ \mathbb{R}_+,$$

$$\mathbb{E}(f(T, \Lambda) = \sum_{\tau \in \mathcal{T}} \int_V f(\tau, \lambda) p_\tau(\lambda) d\lambda,$$

à condition de noter $p_\tau(\lambda) = \alpha_\tau \times q_\tau(\lambda)$.

Dans ce contexte, on a un couple aléatoire, formé d'une part du "paramètre" (T, Λ), et d'autre part des données D. La loi de ce couple est précisé par

- d'une part la loi a priori de (T, Λ) ;
- d'autre part la loi conditionnelle des données, sachant le paramètre.

Plus précisément, la vraisemblance s'interprète dans le contexte bayésien comme la probabilité conditionnelle :

$$L(\tau, \lambda) = \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right).$$

On va alors chercher à calculer la loi *a posteriori*, qui est la loi conditionnelle du "paramètre" (T, Λ), sachant les données, i. e. sachant que $D = d$. Cette loi conditionnelle est donnée par la célèbre "formule de Bayes", qui dans notre situation précise la loi de jointe de (T, Λ) sachant que $D = d$ sous la forme

$$p_\tau(\lambda | D = d) = \frac{\mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda)}{\sum_{\tau \in \mathcal{T}} \int_V \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}.$$

Autrement dit, à nouveau si

$$f : \mathcal{T} \times V \ \rightarrow \ \mathbb{R}_+,$$

$$\mathbb{E}\left(f(T, \Lambda) | D = d\right) = \frac{\sum_{\tau \in \mathcal{T}} \int_V f(\tau, \lambda) \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}{\sum_{\tau \in \mathcal{T}} \int_V \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}.$$

Par exemple, on peut s'intéresser à la loi de probabilité *a posteriori* de la forme de l'arbre, i. e. de la v. a. T. Celle–ci est donnée par la formule suivante : pour tout $\tau \in \mathcal{T}$,

$$\mathbb{P}\left(T = \tau | D = d\right) = \frac{\int_V \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}{\sum_{\tau \in \mathcal{T}} \int_V \mathbb{P}\left(D = d | (T, \Lambda) = (\tau, \lambda)\right) p_\tau(\lambda) d\lambda}.$$

8 Méthode de calcul MCCM

Supposons que l'on veuille calculer cette dernière quantité, pour un petit nombre de valeurs de τ. Un calcul explicite est sans espoir, compte tenu de la taille des données (nombre d'espèces considérées) et de la complexité croissante des modèles utilisés. On est donc amené à utiliser une méthode de type Monte Carlo, utilisant des tirages aléatoires. Cependant , il se présente une sérieuse difficulté pour réaliser des tirages aléatoires suivant la loi a posteriori de (T, Λ), sachant les données, c'est que pour identifier cette loi, il est nécessaire de calculer le dénominateur dans les formules ci–dessus. Pour peu en particulier que le cardinal de $\mathcal{T}$ soit gigantesque, ce calcul peut se révéler totalement impossible.

On se trouve eaxctement dans la même situation que les physiciens Metropolis et al., qui au début des années 1950, voulaient réaliser, en mécanique statistique, des calculs d'espérance sur un esemble fini (mais de cardinal gigantesque), par rapport à une probabilité qui n'était connue qu'à une constante multiplicative près (i. e. la constante de normalisation n'était pas calculable en pratique).

8.1 Le principe de la méthode MCCM

L'idée géniale de Metropolis et al., qui depuis a été reprise dans de nombreux domaines d'application (en particulier le traitement d'images et la statistique bayésienne), est que

pour mettre en oeuvre une méthode de Monte Carlo, on peut s'appuyer sur le théorème ergodique des chaînes de Markov irréductibles et récurrentes positives, au lieu de la loi des grands nombres. Autrement dit, pour calculer une valeur approchée d'une intégrale de la forme (on se limite dorénavant à une intégrale sur un ensemble fini F, ce qui signifie que l'on a discrétisé V) :

$$\sum_{x\in F} f(x)\pi(x),$$

on a le choix entre une *méthode de Monte Carlo classique*, qui consiste à approcher cette quantité par

$$\frac{1}{N}\sum_{k=1}^{N} f(X_k),$$

où les $\{X_k\}$ sont i. i. d., de loi commune la probabilité π sur F, et une *méthode de Monte Carlo par Chaîne de Markov* (en Anglais MCMC), qui consiste à approcher la même quantité à nouveau par

$$\frac{1}{N}\sum_{k=1}^{N} f(X_k),$$

mais où cette fois les $\{X_k\}$ forment une chaîne de Markov irréductible à valeur dans F, admettant π comme probabilité invariante.

Or s'il n'est pas toujours facile d'identifier la probabilité invariante d'une chaîne de Markov de matrice de transition donnée, il est assez facile d'associer à une probabilité π sur un ensemble fini F une matrice de transition P d'une chaîne irréductible, telle que π soit P–invariante. La démarche la plus simple est de choisir P matrice de transition irréductible, telle que la chaîne soit réversible par rapport à π, autrement dit telle que

$$(*) \qquad \pi_x P_{xy} = \pi_y P_{yx}, \quad \forall x \neq y,\ x, y \in F.$$

La remarque essentielle est que la connaissance de la probabilité π à une constante multiplicative près suffit parfaitement à trouver une matrice P satisfaisant les conditions requises.

On va présenter l'algorithme de Metropolis–Hastings. L'idée est la suivante. Soit Q une matrice de transition irréductible sur F (qui n'a a priori rien à voir avec la probabilité π), dont on soit capable de simuler aisément les transitions. On choisit comme matrice P la matrice de transition dont les termes hors diagonaux sont donnés par

$$P_{xy} = \min\left(Q_{xy}, \frac{\pi_y}{\pi_x} Q_{yx}\right),$$

et les termes diagonaux sont donnés par

$$P_{xx} = 1 - \sum_{y\neq x} P_{xy},$$

sous réserve que P ainsi défini soit irréductible, ce qui est par exemple vrai si Q vérifie la propriété : pour tout x, y, $Q_{xy} > 0 \Leftrightarrow Q_{yx} > 0$. Il est clair que cette matrice P est bien une

matrice de transition ($P_{xy} \leq Q_{xy}$, $x \neq y$ implique que $P_{xx} \geq 0$), et qu'elle vérifie la relation $(*)$. Pour expliquer comment on simule les transitions de la chaîne de matrice de transition P, posons pour $x, y \in F$,

$$r_{xy} = \frac{P_{xy}}{Q_{xy}} = \min\left(1, \frac{\pi_y Q_{yx}}{\pi_x Q_{xy}}\right).$$

Une façon de simuler une transition de la chaîne $\{X_k\}$ de matrice de transition P est la suivante. Supposons que $X_k = x$, et on veut simuler X_{k+1}. On simule d'abord une transition de la chaîne $\{Y_k\}$ de matrice de transition Q, partant de $Y_k = x$. Supposons que le résultat de ce tirage soit $Y_{k+1} = y$. On *accepte cette transtion* (et dans ce cas $X_{k+1} = y$) avec la probabilité r_{xy} ; on *refuse cette transition* (et dans ce cas $X_{k+1} = x$) avec la probabilité $1 - r_{xy}$. Remarquons $r_{xx} = 1$, donc dans le cas $y = x$, la "non–transition" est toujours acceptée.

Autrement dit, pour passer de $X_k = x$ à X_{k+1}, on effectue

- le tirage de Y_{k+1}, suivant la probabilité $Q_{x\cdot}$;
- le tirage de U_{k+1}, de loi uniforme sur $[0, 1]$;

et on pose

$$X_{k+1} = Y_{k+1}\mathbf{1}_{\{U_{k+1} \leq r_{xY_{k+1}}\}} + X_k \mathbf{1}_{\{U_{k+1} > r_{xY_{k+1}}\}}.$$

8.2 Mise en oeuvre de la méthode MCCM

- Faut–il éliminer le début de la simulation ("burn–in") ?
- Faut–il échantillonner les tirages ?
- Faut–il simuler plusieurs chaînes en parallèle ?
- Comment faire avec les transitions rares ?
- Combien de temps faut–il simuler la chaîne ?

9 Bibliographie

- W. Ewens, G. Grant, *Statistical methods in Bioinformatics*, Springer 2001.
- J. Felsenstein, *Infering phylogenies*, Sinauer 2004.
- R. Nielsen ed., *Statistical methods in Molecular Evolution*, Springer 2005.
- C. Semple, M. Steel, *Phylogenetics*, Oxford Univ. Press, 2003.