Statistique pour la bio-informatique

Séance 7

Phylogénies

1 Reconstruction de phylogénies

L'étude de la phylogénie est un vaste domaine et quelle que soit la méthode utilisée, des hypothèses très simplificatrices sont faites sur l'évolution biologique des séquences. L'évolution est représentée par un arbre binaire. Nous confondons "espèces" et "séquences" étant entendu que les relations de descendance des espèces sont plus complexes que celles des séquences et ne coincident pas nécessairement. Il aussi est supposé que les espèces étudiées ont un ancêtre commun (évolution divergente).

Définition. *Soit S un ensemble de points. Une distance sur S est une application telle que i) $d(x, y) \geq 0$, ii) $d(x, y) = 0$ ssi $x = y$, iii) $d(x, y) \leq d(x, z) + d(y, z)$.*

On dit qu'une distance dérive d'un arbre s'il existe un arbre tel que les éléments de S puissent être considérés comme les feuilles et la distance entre deux feuilles est égale à la somme des distances sur les branches qui les séparent. Une distance dérivée d'un arbre est aussi appelée *distance additive*.

1.1 La méthode UPGMA

Définition. *Soit S un ensemble de points. Une distance ultramétrique sur S est une distance telle que pour tout triplet de points, deux distances sont égales et la troisième est strictement inférieure aux deux autres. Cette condition est parfois appelée condition des trois points.*

Théoreme 1.1 *Étant donné une distance ultramétrique, il existe un unique (aux permutations triviales près) arbre enraciné dont la distance dérive.*

Il existe une démonstration constructive (admise) de ce résultat. L'arbre obtenu est équivalent à celui obtenu par la méthode UPGMA dont voici la description. On cherche tout d'abord les deux espèces x et y de distance $d(x, y)$ minimale. Soit k le noeud qui les joint. Puisque d dérive d'un arbre, nous avons

$$d(k, z) = (d(x, z) + d(y, z) - d(x, y))/2$$

pour tout $z \neq x, y$. Par conséquence de la définition de distance ultramétrique, nous avons

$$d(x, k) = d(y, k) = d(x, y)/2.$$

On remplace x et y par le noeud n et on réitère le processus. Cela construit un arbre enraciné.

Lorsque la distance est ultramétrique, la méthode UPGMA reconstruit donc l'arbre dont la distance dérive. Dans les applications en biologie, cette méthode est parfois utilisée sans que cette hypothèse puisse être vérifiée. On parle alors de méthode heuristique de reconstruction.

1.2 La méthode Neighbor-Joigning

On dit qu'une distance vérifie *la condition des quatre points* si pour tout x, y, z, t, deux des trois grandeurs suivantes sont égales et supérieures à la troisième

$$d(x, y) + d(z, t)$$

$$d(x, t) + d(z, y)$$

$$d(x, z) + d(t, y).$$

Étant donné une métrique vérifiant la condition des quatre points, nous souhaitons construire un arbre dont la distance dérive. Nous énonçons le résultat suivant, qui établit la faisabilité de ce programme.

Théoreme 1.2 *Soit S un ensemble fini et d une distance vérifiant la condition des quatre points. Alors il existe un unique arbre non-enraciné dont la distance dérive. La condition des quatre points est une condition nécessaire et suffisante d'addivité. L'arbre est donné par l'algorithme appelé Neighbor-Joining.*

L'algorithme **Neighbor-Joining** procède globalement d'une manière analogue à UPGMA. On cherche tout d'abord une paire de taxons voisins x et y, puis on postule que ces taxons ont un parent commun k. On construit alors une nouvelle matrice de distance en remplaçant x et y par k.

Puisque k est un ancêtre commun de x et y, nous pouvons le considérer comme un noeud interne du sous-arbre dont les feuilles sont x et y. La distance de k aux autres taxons doit être calculée indirectement, car nous ne disposons pas de données pour ce noeud. Soit z un taxon tel que $z \neq x, y$ $d(k, z)$ est solution du système d'équations suivant

$$\begin{aligned} d(x,k) + d(j,k) &= d(x,y) \\ d(x,k) + d(k,z) &= d(x,z) \\ d(y,k) + d(k,z) &= d(y,z) \end{aligned}$$

Ainsi, nous avons de manière identique à UPGMA

$$d(k,z) = \frac{1}{2}(d(x,z) + d(y,z) - d(x,y)).$$

Ce calcul permet donc d'itérer le procédé de construction emboité des matrices de distances. Il reste à préciser la manière de déterminer les deux taxons voisins. Pour cela, definissons

$$D_{xy} = d(x,y) - (r_x + r_y)$$

où

$$r_x = \frac{1}{n-2} \sum_{z \neq x}^{n} d(x,z).$$

Ainsi, r_x représente la distance moyenne du taxon x aux autres taxons (pondérée par un facteur $(n-1)/(n-2)$). Dans l'algorithme NJ on dira que x et y sont voisins si D_{xy} est minimal (en général négatif). La paire x, y sera choisie pour itérer la construction de l'arbre.

Exemple Itérons le premier pas de construction de l'arbre associé à la matrice de distance suivante

1	$a+b$	$a+c+d$	$a+c+e$
2		$b+c+d$	$b+c+e$
3			$d+e$
	2	3	4

Les valeurs de r_i sont égales à

$$\begin{aligned} r_1 &= (3a + 2c + b + d + e)/2 \\ r_2 &= (3b + 2c + a + d + e)/2 \\ r_3 &= (3d + 2c + a + b + e)/2 \\ r_4 &= (3e + 2c + a + b + d)/2 \end{aligned}$$

Pour les D_{ij} nous obtenons

$$\begin{aligned} \delta_{12} &= -a - b - 2c - d - e = \delta_{34} \\ \delta_{13} = \delta_{23} = \delta_{14} = \delta_{24} &= -a - b - c - d - e \end{aligned}$$

Sans surprise, nous obtenons que les paires $1, 2$ et $3, 4$ sont à égalités pour le minimum. Cela signifie qu'il n'y pas une manière unique de démarrer la construction de l'arbre (lui, unique).

Théoreme 1.3 *(Studier et Keppler, 1988) Soit d une distance dérivant d'un arbre non enraciné et*

$$\delta(x, y) = (n - 4)d(x, y) - \sum_{z \neq x,y} d(x, z) + d(y, z), \quad n = \#S.$$

si $\delta(x, y)$ est minimal, alors x et y sont voisins (séparés par un noeud intermédiaire uniquement).

Démonstration du Théoreme. Le raisonnement est par l'absurde. Supposons que

$$\delta(i, j) = \min_{k,\ell} \delta(k, \ell)$$

et que i et j ne sont pas voisins. Dans ce cas, i et j sont séparés par r noeuds internes, tous racine d'un sous-arbre non trivial ($r \geq 1$). Il y a trois cas à considérer (les détails sont ommis).

Cas 1. Le premier arbre intermédiaire (de i vers j) ne comporte qu'une feuille ℓ. Dans ce cas (à vérifier)

$$\delta(i, \ell) = d(i, \ell) - r_i - r_\ell < \delta(i, j)$$

Cas 2. Symétrique en remplaçant i par j

Cas 3. L'arbre 1 et l'arbre r, peut-être identiques, contiennent chacun deux feuilles voisines que nous notons respectivement k, ℓ et m, n. Si l'arbre 1 contient plus feuilles que l'arbre 2 alors $\delta(i,j) > \delta(m,n)$, sinon $\delta(i,j) > \delta(k,\ell)$.

Théoreme 1.4 *Soit S un ensemble fini. Une distance ultramétrique vérifie la condition des quatre points.*

Pour une telle métrique les méthodes de reconstruction NJ et UPGMA sont équivalentes. En un sens, une distance ultramétrique représente la situation idéale où l'horloge moléculaire est constante le long de toutes les branches de l'évolution.

Exercice. Construire l'arbre associé à la matrice de distance suivante

x	4	7	11	9
y		9	13	11
z			8	6
t				8
	y	z	t	u

Exemple. Boucle D de controle dans l'ADN mitochondrial humain échantillonné pour 55 individus de la tribu Nuu Chah Nulth (Indiens d'Amérique). Les 55 séquences comportent 360 paires de bases et diffèrent en 18 sites appelés *sites de ségrégation*. Le tableau 1.2 présente les 14 séquences ou allèles en ce locus.

Exercice. On suppose $\#S = n \geq 3$. Dénombrer les arbres non-enracinés (solution : $(2n-5)!/(2^{n-3}(n-3)!)$) et enracinés. Les nœuds internes ne sont pas numérotés.

site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	effectif
allele																			55
a	*a*	*g*	*g*	*a*	*a*	*t*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*t*	*c*	*t*	*t*	*c*	2
b	*a*	*g*	*g*	*a*	*a*	*t*	*c*	*c*	*t*	*t*	*t*	*t*	*c*	*t*	*c*	*t*	*t*	*c*	2
c	*g*	*g*	*a*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*t*	*t*	*t*	1
d	*g*	*a*	*g*	*g*	*a*	*c*	*c*	*c*	*c*	*c*	*t*	*t*	*c*	*c*	*c*	*t*	*t*	*c*	3
e	*g*	*g*	*g*	*a*	*a*	*t*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*t*	*c*	*t*	*t*	*c*	19
f	*g*	*g*	*g*	*a*	*g*	*t*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*t*	*c*	*t*	*t*	*c*	1
g	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*c*	*c*	*c*	*c*	*c*	*t*	*t*	*t*	1
h	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*c*	*c*	*t*	*c*	*c*	*t*	*t*	*t*	1
i	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*c*	*c*	*t*	4
j	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*c*	*t*	*t*	8
k	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*t*	*t*	*c*	5
l	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*t*	*t*	*t*	4
m	*g*	*g*	*g*	*g*	*a*	*c*	*c*	*c*	*t*	*c*	*t*	*t*	*c*	*c*	*c*	*t*	*t*	*c*	3
n	*g*	*g*	*g*	*g*	*a*	*c*	*t*	*t*	*t*	*c*	*t*	*t*	*c*	*c*	*t*	*t*	*t*	*c*	1

TAB. 1 – Jeu de données de *Ward et al., 1991* : ADN mitochondrial humain échantillonné pour les individus de la tribu Nuu-Chah-Nulth (Indiens d'Amérique). En chaque site de ségrégation est présente soit une purine (*a*-*g*) soit une pyrimidine *c*-*t*. On n'observe pas de transversion.

1.3 Estimer les distances

Supposons que les taxons consistent en n séquences d'ADN et considérons pour ces n séquences les nombres de différences observées lorsque l'on compare les séquences deux à deux. Cette matrice constitue une notion naïve de similarité entre les séquences. En effet, la matrice sous-estime le nombre réel de mutations qui se sont produites, et qui ont peut-être touché plusieurs fois le même site.

Afin de produire une distance d'évolution plus réaliste, il est possible d'utiliser l'un des modèles markoviens vus précedemment, tels que les modèles de Jukes-Cantor ou de Kimura. Supposons par exemple que le taux de mutation d'une base i en une autre base j, ne dépend que de j

$$\lambda_{ij} = \lambda_j.$$

Bien entendu, cette hypothèse est simplificatrice (On peut effectuer les calculs en général au prix d'une discussion demandant des notations supplémentaires). Sous cette hypothèse, la probabilité $p_a(t)$ d'observer la base a au temps t obéit à l'équation suivante

$$p'_a = -(\lambda_c + \lambda_t + \lambda_g)p_a + \lambda_a(1 - p_a)$$

À l'équilibre, nous obtenons

$$\pi_a = \frac{\lambda_a}{\lambda}$$

où l'on a noté $\lambda = \lambda_a + \lambda_c + \lambda_t + \lambda_g$. Connaissant la base ancestrale, par exemple $i = c$, l'intégration de l'équation différentielle donne

$$p_{ij}(t) = \mathrm{P}(j \text{ au temps } t \mid i \text{ au temps } 0) = e^{-\lambda t}\delta_{ij} + (1 - e^{-\lambda t})\pi_j.$$

Supposons que la base ancestrale soit tirée selon la loi de probabilité π, c'est-à-dire que le système est en équilibre à l'instant $t = 0$. Nous obtenons alors

$$\mathrm{P}(j \text{ au temps } t \ \cap \ i \text{ au temps } 0) = \pi_c p_{ij}(t) = \alpha_{ij} + \beta_{ij}e^{-\lambda t}$$

où α_{ij} et β_{ij} sont des constantes.

On souhaite maintenant estimer le temps de divergence T de deux séquences. En considérant que le temps n'a pas de direction privilégié (évolution réversible), nous supposons que la séquence 1 correspond à l'instant 0 et la séquence 2 correspond à l'instant $2T$. Soit N_{ij} le nombre de substitutions de i vers j. Posons $\chi = \exp(-\lambda t)$. Nous estimons χ par la méthode de maximum de vraisemblance

$$\hat{\chi} = \mathrm{argmax} \sum_i \sum_j N_{ij} \ \log(\alpha_{ij} + \beta_{ij}\chi).$$

par une méthode numérique. Ensuite, nous avons

$$T = -\frac{1}{\lambda} \log \hat{\chi}.$$

Exercice. Décrire les équations correspondant à l'annulation de la dérivée de la vraisemblance.

2 Construction de phylogénie par maximum de vraisemblance

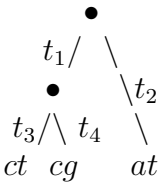
Dans cette section, nous souhaitons estimer simultanément la topologie d'un arbre phylogénétique et les longueurs des branches, c'est à dire les distances évolutives séparant les taxons. Dans un premier temps, nous montrons comment calculer la vraisemblance d'un arbre lorsque la topologie est donnée, ensuite nous décrivons un algorithme d'estimation des longueurs des branches (EM).

2.1 Vraisemblance d'un arbre phylogénétique

Supposons données n séquences de longueur L. Pour une topologie d'arbre donnée, nous considérons qu'une branche est une arête séparant un noeud interne de l'arbre (la racine étant considérée comme un noeud) d'un autre noeud ou d'une feuille de l'arbre. Nous prenons comme paramètre

$$T = t_1, \ldots, t_m, \quad m = 2(n-1)$$

l'ensemble des longueurs des branches de l'arbre. Considèrons par exemple, le modèle suivant



pour les 3 séquences $s_1 = ct$, $s_2 = cg$, $s_3 = at$. Nous supposons 1) que les délétions et les insertions sont de probabilité nulle, 2) qu'il n'y a pas de sens privilégié au temps sur une arête. La première hypothèse permet d'utiliser des modèles simplifiés de distance provenant d'un processus de Markov. Elle permet aussi de ne pas faire intervenir de dépendances entre les différents sites. La seconde hypothèse se traduit par la réversibité

du processus dans le temps et permet en fait d'enraciner l'arbre comme on le souhaite (justification dans le paragraphe suivant).

Pour l'arbre ci-dessus, nous souhaitons calculer la probabilité d'observer les $n = 3$ séquences s_1, s_2, s_3

$$L(T) = L(t_1, \ldots, t_m) = \mathrm{P}(s_1, \ldots, s_n; T).$$

Lorsque l'on cherche à calculer $L(T)$, les variables aux noeuds internes de l'arbre ne sont pas observées. Elles apparaissent donc comme des données manquantes. La vraisemblance $L(T)$ se calcule théoriquement en sommant sur toutes les séquences de longueurs L $(= 2)$. Il y a $16^2 = 256$ séquences possibles. En supposant l'indépendance des sites à l'intérieur de la séquence, le calcul se réduit au produit de L $(= 2)$ expressions comportant 4^{n-1} (16) termes. L'hypothèse d'indépendance se résume de la manière suivante

$$L(T) = \mathrm{P}(c, c, a\,;\, T)\,\mathrm{P}(t, g, t\,;\, T)$$

Cela revient à considérer les deux sous-arbres de la figure **??**. Chacune des deux probabilités intervenant dans cette expression est immédiatement calculable. Notons s_5 la base présente à la racine et s_4 la base présente au noeud interne parent de s_1 et s_2, la vraisemblance s'écrit

$$\mathrm{P}(c, c, a\,;\, T) = \sum_{s_5, s_4} \pi_{s_5} p_{s_5 s_4}(t_1) p_{s_5 s_3}(t_2) p_{s_4 s_1}(t_3) p_{s_4 s_2}(t_4)$$

où $p_{ij}(t)$ désigne la probabilité de passer de i à j durant un intervalle de temps de longueur t et π la loi invariante associée. La vraisemblance est dite incomplète, de la même manière que dans le chapitre concernant l'estimation dans les modèles à données manquantes.

Le calcul de la vraisemblance suivant la formule précédente suggère un algorithme dont la complexité est de l'ordre de $L(\#\Sigma)^{n-1}$ opérations, où Σ est l'ensemble des symboles pour un site donné. Ce dernier est de cardinal égal à 4 pour les nucléotides et de cardinal égal à 20 pour les acides aminés. Ainsi, on aboutit très rapidement à une explosion du nombre de calculs rendant la formule impossible à exploiter directement.

2.2 L'algorithme d'estimation de Felsenstein

Calcul récursif La complexité de calcul de la vraisemblance $L(T)$ peut être réduite de manière considérable par une méthode de programmation dynamique reposant sur un calcul récursif de la vraisemblance. L'algorithme est dû à Felsenstein (1981). Son principe de fond est analogue au schéma de Horner pour les polynômes. Considérons le

sous-arbre de T enraciné en k, d'état ou de symbole σ_k. De cette racine partent deux sous-arbres peut-être triviaux. Les fils de k sont notés ℓ et m et ont pour symbole σ_ℓ et σ_m

$$\begin{array}{cc} & \sigma_k \\ t_1/ & \backslash t_2 \\ \sigma_\ell & \sigma_m \end{array}$$

Notons donc L^k_σ la vraisemblance de ce sous-arbre enraciné en k, étant donné que $\sigma_k = \sigma$.

La vraisemblance peut être calculée récursivement en posant

$$L^k_{\sigma_k} = \left(\sum_{\sigma_\ell} p_{\sigma_k,\sigma_\ell}(t_{k\ell}) L^\ell_{\sigma_\ell} \right) \left(\sum_{\sigma_m} p_{\sigma_k,\sigma_m}(t_{km}) L^m_{\sigma_m} \right)$$

L'algorithme de calcul est initialisé aux feuilles en posant, pour toute une feuille i,

$$L^i_\sigma = \begin{cases} 1 & \text{si le symbole observe en } i \text{ est } \sigma \\ 0 & \text{sinon.} \end{cases}$$

Il se termine à la racine r en calculant

$$L(T) = \sum_{\sigma_r} \pi_{\sigma_r} L^r_{\sigma_r}$$

et la complexité algorithmique est de l'ordre de $L \times (\#\Sigma)^2$.

Exercice. Développer le calcul de la vraisemblance sur un petit exemple (4 séquences).

Longueurs de branches optimales Soit T une topologie d'arbre fixée, comportant n feuilles, pour laquelle les longueurs des branches $t_1, \ldots, t_m$ sont considérées comme un paramètre statistique à optimiser. Nous cherchons

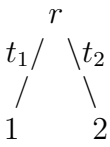
$$(\hat{t}_1, \ldots, \hat{t}_m) = \text{argmax } L(T) = \text{argmax } \mathrm{P}(s_1, \ldots, s_m; T)$$

L'algorithme de Felsenstein s'appuie sur l'heuristique suivante. Chaque composante t_i est optimisée séquentiellement, les autres composantes t_j, $j \neq i$, restant provisoirement constantes.

Un principe (dit de la poulie) permet de placer la racine de l'arbre de manière arbitraire.

Lemme 2.1 *Supposons que le modèle markovien d'évolution permettant de calculer les distances soit réversible. Alors, le calcul de la vraisemblance est insensible au choix d'une racine de l'arbre.*

Démonstration. Par indépendance des sites de la séquence, il suffit de prouver le résultat pour l'arbre correspondant à un unique site. Nous supposons construit un arbre non-enraciné, et nous plaçons la racine r de manière arbitraire entre deux noeuds internes 1 et 2 sont deux noeuds internes séparés par une distance égale à $t_1 + t_2$



D'après la définition récursive, la vraisemblance s'écrit

$$L(T) = \sum_{\sigma_r} \pi_{\sigma_r} L^r_{\sigma_r}$$

où

$$L^r_{\sigma_r} = \left(\sum_{\sigma_1} p_{\sigma_r,\sigma_1}(t_1) L^1_{\sigma_1} \right) \left(\sum_{\sigma_2} p_{\sigma_r,\sigma_2}(t_2) L^2_{\sigma_2} \right).$$

Nous souhaitons montrer que cette expression ne dépend que de $t_1 + t_2$ et non du choix arbitraire de t_1 et t_2 lié au fait que nous avons fixé r. Cet état de fait découle de la réversibilité

$$\pi_{\sigma_r} p_{\sigma_r,\sigma_1}(t_1) = \pi_{\sigma_1} p_{\sigma_1,\sigma_r}(t_1)$$

Nous avons donc en permutant les sommes

$$L(T) = \sum_{\sigma_1} \sum_{\sigma_2} L^1_{\sigma_1} L^2_{\sigma_2} \sum_{\sigma_r} \pi_{\sigma_1} p_{\sigma_1,\sigma_r}(t_1) p_{\sigma_r,\sigma_2}(t_2)$$

Par la propriété de Markov, nous avons

$$\sum_{\sigma_r} p_{\sigma_1,\sigma_r}(t_1) p_{\sigma_r,\sigma_2}(t_2) = p_{\sigma_1,\sigma_2}(t_1 + t_2)$$

et l'expression $L(T)$ ne dépend plus que de $t_1 + t_2$.

L'optimisation de la vraisemblance consiste donc tout d'abord à considérer les distances sur un arbre **non -enraciné**. Puis le principe de poulie permet de faire glisser une racine sur un nœud quelconque. Notons n_1 ce nœud (ou cette feuille). Il se trouve relié directement à un autre nœud (ou feuille) n_2 par une branche de longueur t. Par le principe de poulie, la vraisemblance vue comme une fonction de t s'écrit

$$L(t) = \sum_{\sigma_1,\sigma_2} \pi_{\sigma_1} L^1_{\sigma_1} L^2_{\sigma_2} p_{\sigma_1,\sigma_2}(t).$$

Afin de fixer les idées, considèrons à nouveau le modèle d'évolution simplifié suivant (JK avec $\alpha = 1/3$)

$$p_{\sigma_1,\sigma_2}(t) = e^{-t}\delta_{\sigma_1,\sigma_2} + (1 - e^{-t})\pi_{\sigma_2}$$

où π_{σ_2} est la loi uniforme. En arrageant les différents termes de la somme, nous obtenons

$$L(p) = pA + qB$$

où $p = e^{-t}$ et

$$A = \sum_{\sigma} \pi_\sigma L^1_\sigma L^2_\sigma, \qquad B = \sum_{\sigma_1,\sigma_2} \pi_{\sigma_1}\pi_{\sigma_2} L^1_{\sigma_1} L^2_{\sigma_2}$$

Puisque les séquences sont de longueur L, nous avons L sites indépendants et

$$\ln L(q) = \sum_{\ell=1}^{L} \log(pA_\ell + qB_\ell)$$

En annulant la dérivée

$$\frac{\partial}{\partial q} \log L(q) = 0,$$

nous obtenons l'équation de point fixe suivante

$$q = \frac{1}{L} \sum_{\ell=1}^{L} \frac{qB_\ell}{pA_\ell + qB_\ell}.$$

Il est naturel de résoudre cette équation par itérations

$$q_{n+1} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{q_n B_\ell}{p_n A_\ell + q_n B_\ell}.$$

La vraisemblance augmente alors jusqu'à atteindre un maximum local. Et l'algorithme constitue une version de l'algorithme EM.