Statistique pour la bio-informatique

Séance 2 - 8 Octobre 2003

Mutations et horloge moléculaire

1 Évolution moléculaire

1.1 Un peu d'histoire

Darwin définissait l'évolution comme un processus de *descendance avec modifications*, un concept qui n'a pas été véritablement remis en cause depuis son époque (1859). En terme de génétique, la définition de Darwin peut être comprise de la manière suivante. Les individus transmettent leurs gènes à leurs descendants, les gènes sont modifiés par des mutations en de nouveaux allèles et ainsi de suite. À l'échelle de la population, ce sont ainsi les fréquences des allèles qui se trouvent modifiées.

Bien entendu, la notion de gène était inconnue de Darwin. Les fondements théoriques de la génétique des populations n'ont été posés que bien plus tard par Wright, Fisher et Haldane (années 30). Ces derniers ne connaissaient rien de la structure moléculaire de l'ADN. Ainsi le séquençage de brins d'ADN, notamment par la méthode de l'électrophorèse, a constitué une véritable révolution permettant l'accès à la variabilité de l'information génétique au niveau moléculaire (années 60). Dans un premier temps, l'accès à cette information conduisit à une utilisation quantitative des concepts développés en génétique des populations. Elle a aussi suscité de nouveaux développements théoriques destinés à mieux expliquer la variabilité observée au niveau moléculaire. Un exemple d'une telle théorie est le point de vue neutraliste de Kimura.

La théorie de Kimura repose essentiellement sur l'observation qu'une grande fraction de la variabilité génétique observée à l'échelle moléculaire est sélectivement neutre, c'est-à-dire que les differences entre allèles ne procurent pas d'avantage ou de désavantage sélectif. Dans ce cas, c'est le hasard des tirages aléatoires des différents allèles qui domine la dynamique d'évolution. Ce hasard est appelé **dérive génétique**. Cela ne signifie pas forcement que les mutations n'ont pas d'effet sur les gènes. Au contraire, cette hypothèse repose sur le fait que nombreuses mutations sont délétaires (létates). Ces mutations conduisent à des organismes non viables et les mutations en question ne sont donc pas observées. Seules les mutations n'ayant pas d'effet délétaires ont une chance d'être observées.

La théorie neutraliste émet l'hypothèse que la classe des mutations autorisées est composée intégralement de variants neutres. Elle s'oppose à l'alternative dans laquelle les mutations avantageuses, même si elles sont extrêmement rares, sont essentiellement responsables du polymorphisme des populations. Dans cette séance, nous nous intéressons aux aspects quantitatifs de l'hypothèse neutraliste et cherchons en particulier à décrire la destinée d'une mutation neutre apparue dans une population à une génération donnée.

1.2 Des poules et des hommes

Considérons par exemple le problème de la reconstruction des relations généalogiques liant différentes espèces. La suite d'acides aminés codant pour la β-globine a été séquencée dans le cas des espèces suivantes : *homme, chimpanzé, vache, ornithorynque, poule* et *requin*. Examinons le nombre de différences entre les séquences prises deux à deux. Les résultats sont résumés dans le tableau suivant

humain	8				
vache	27	25			
ornit.	35	34	41		
poule	44	45	55	42	
requin	93	96	97	95	91
	chimp.	humain	vache	ornit.	poule

Sous l'hypothèse neutre, le nombre de différences témoigne des relations évolutives. Si l'on observe 93 différences entre la séquence du chimpanzé et celle du requin, cela signifie que les deux espèces ont divergé il y a très longtemps, laissant le temps à 93 mutations neutres de se fixer dans l'une ou l'autre des deux branches séparées. Dans toutes les espèces, la fonction de la protéine est exactement identique. Grâce à ce tableau, une topologie d'arbre peut être définie sans trop d'ambiguité. *Ce ne sera pas toujours le cas.*

Le choix ou la pertinence d'un gène dans l'étude des relations phylogénétiques entre espèces dépend en partie de la vitesse d'évolution du gène. Par exemple, si l'on s'intéresse à des événements extrêmement anciens comme la diversification des plantes terrestres, il s'agit de choisir des gènes à évolution très lente, et donc très fortement sélectionnés. Pour des spéciations plus récentes, des fractions d'ADN à évolution plus rapides sont utiles (par exemple l'ADN mitochondrial).

1.3 Calcul de la probabilité de fixation

Afin de justifier la pertinence d'études phylogénétiques fondées sur le nombre de substitutions dans l'ADN, nous allons introduire un modèle simplifié décrivant la variation des fréquences alléliques au sein de populations. Les ingrédients de ce modèle sont : la reproduction, la mutation et la sélection. En d'autres termes, afin de définir la dynamique des fréquences alléliques, nous devons décrire le mode par lequel le matériel génétique est transmis, et comment ce matériel influence la survie de son porteur.

Le modèle adopté pour la *sélection* est la *neutralité* : c'est-à-dire donc l'absence de sélection. Dans une population de taille finie, la fixation au hasard des allèles (la dérive génétique) est responsable pour la majeure partie de la variabilité observée au niveau moléculaire. Cette variabilité est due aux rééchantillonnages successifs des gènes de génération en génération. Ainsi, certaines lignées peuvent s'éteindre par hasard car elles ne sont pas assez représentés.

Le modèle ci-dessous prend en compte cette notion. Il est appelé **modèle de Wright-Fisher**. Les hypothèses cruciales sont :

- la taille de la population est finie et constante ;
- les générations ne se chevauchent pas ;
- les gènes des descendants sont tirés avec remise dans une urne constituée des gènes des parents.

Les individus sont souvent diploïdes, c'est-à-dire qu'ils possèdent deux copies d'un gène. Nous supposerons qu'il y a $2N$ gènes dans une population de taille N. Rappelons que l'on appelle *allèle* une *variante* d'un gène ou d'une protéine, voire la séquence d'ADN en un locus précis pas forcément codant, observé dans une population. Nous supposons que le gène possède deux allèles A et a, et que l'allèle a n'est représenté qu'une unique fois (et A est présent $2N-1$ fois). Cette hypothèse correspond à l'apparition d'un nouvel individu (mutant) dans une population homogène au locus considéré. Nous cherchons la **probabilité de fixation** de l'allèle a dans cette population. En d'autres mots, nous cherchons la probabilité pour que la *dérive génétique* élimine tous les A.

Soit X_t le nombre de gènes a au temps t ($X_0 = 1$). Sachant $X_t = i$, X_{t+1} est une variable aléatoire de loi binomiale $\mathcal{B}(2N, i/2N)$

$$\forall j = 0, \ldots, 2N,\ \mathrm{P}(X_{t+1} = j \ / \ X_t = i) = C_{2N}^{j}(1 - \frac{i}{2N})^{2N-j}(\frac{i}{2N})^{j}$$

Remarquons que X_t est une chaîne de Markov. Nous avons de plus

$$E[\frac{X_{t+1}}{2N} \ / \ X_0, \ldots, X_t] = \frac{X_t}{2N}$$

Ainsi $X_t/2N$ est aussi une *martingale*. Elle converge presque-sûrement vers une variable dont la loi est de Bernoulli. En effet, lors de la fixation nous avons $X_t = 2N$. Dans le cas où la mutation disparaît au temps t nous avons $X_t = 0$. En notant T le temps d'atteinte de 0 ou $2N$, nous avons

$$p_{\text{fix}} = \mathrm{P}(X_T = 2N) = E[\frac{X_T}{2N}] = E[\frac{X_0}{2N}] = \frac{1}{2N}$$

Cette dernière probabilité correspond exactement à la probabilité de fixation de l'allèle a.

Imaginons désormais que les mutations se produisent à un taux par génération égal à μ très petit devant $2N$. *Nous voulons savoir à quel taux ces mutations, qui se produisent dans une population homogène, peuvent se fixer* et correspondre à une substitution au niveau du gène d'une espèce. On parlera de **taux de substitution** pour décrire la vitesse avec laquelle les gènes évoluent à l'échelle de la population.

Le résultat précédent nous éclaire

$$\text{taux de substitution} = 2N\mu\, p_{\text{fix}} = \mu.$$

Ce résultat provient de la superposition de $2N$ processus de Poisson homogènes indépendants. Ce résultat élémentaire montre que les données que l'on observe (les substitutions) ne sont en fait dépendantes que d'un seul facteur : le taux de mutation. Elles sont **indépendantes de la taille de la population**. Cette remarque est à la base d'une théorie très controversée : l'*horloge moléculaire*. Si le taux de mutation est constant, alors le taux de substitution l'est aussi et les substitutions se produisent à intervalles réguliers comme programmées par une horloge. Zuckerland et Pauling ont observé le phénomène pour les taux d'évolution de protéines dans le cytochrome c et l'hémoglobine (1965) mais il est clair cela n'est pas vrai en général et les méthodes de reconstruction phylogénétique récentes tentent parfois de rendre compte des variations des vitesses d'évolution.

Kimura (1962) a calculé la probabilité de fixation de l'allèle a dans un modèle où la probabilité de sélection est proportionnelle à une valeur appelée *fitness* égale respectivement à 1, $1+s$, $1+2s$ pour AA, Aa, et aa où s désigne la valeur sélective. Il obtient

$$p_{\text{fix}} \approx \frac{1 - e^{-2s}}{1 - e^{4Ns}} \approx 2s$$

pour s petit et N grand. Le taux de substitution est alors

$$\text{taux de substitution} = 4N\mu s$$

et dépend de la taille, du taux de mutation et de la valeur sélective de l'allèle. En général la valeur sélective est une grandeur difficile à estimer, de plus les tailles des populations doivent être elles aussi estimées. Cette remarque présage de manière évidente des difficultés pour reconstruire les relations phylogénétiques lorsque les gènes sont sélectionnés.

1.4 Horloge moléculaire et temps de divergence

Étant donné n espèces, nous supposons qu'il y a un unique arbre (phylogénétique) les reliant à un ancêtre commun. Un **taxon** est une entité dont la distance ou la similarité avec les autres entités peut faire l'objet d'une mesure. Un taxon peut correspondre à une espèce, une séquence, un langage, ou des propriétés morphologiques, culturelles, etc. Idéalement, une distance entre taxons doit refléter le temps écoulé depuis la divergence d'évolution. Nous allons examiner dans cette section le problème de calcul de distance et d'estimation de l'horloge moléculaire. Comme nous le verrons, ces deux questions sont interdépendantes.

Concernant les acides aminés, il est largement admis que le taux de substitution varie selon les organismes, la famille de protéines, etc du fait notamment de la pression de sélection qui s'exerce différemment. Pour l'insuline, par exemple, il est connu que le taux de substitution par site (et par an) est égal à $\lambda = 5.3\ 10^{-9}$ chez l'homme alors qu'il peut être de $0.3\ 10^{-9}$ pour d'autres espèces modèles. Les taux dans cet exemple peuvent être différent d'un facteur 10. Concernant les familles de protéines, le taux d'évolution de la fibropeptine (intervenant dans la structure de la fibrine et jouant un rôle dans la coagulation du sang) est de $9\ 10^{-9}$, le taux de l'histone H_4 (squelette de l'ADN) est de 10^{-11} ce qui signifie que les mutations sont très souvent létales (délétaires), le taux d'autres protéines telles que le cytochrome c ou l'hémoglobine prend une valeur intermédiaire autour de 10^{-10}. Par ailleurs, considérer qu'un taux est invariant d'un site à un autre est aussi une approximation. En effet, la structure tertiaire d'une protéine peut changer en modifiant sa fonction lorsqu'une mutation touche certaines zones de l'ADN (comme le noyau hydrophobe).

En gardant à l'esprit les limites mentionnées au paragraphe précédent, fixons comme unité de temps de l'évolution le temps τ nécessaire en moyenne pour obtenir une substitution pour 100 sites. Ce temps est relié à l'horloge de la manière suivante

$$\tau = \frac{1}{100\lambda}$$

où λ est le taux par site (penser au min de variables de loi exponentielle indépendantes).

Afin d'estimer le taux λ, un moyen consiste à observer le le temps de divergence de deux espèces reliées dans l'évolution. Pour l'histone par exemple, on observe deux différences seulement entre la séquence du pois (pea) et celle du veau (calf). Par ailleurs, les données fossiles (paléontologiques) démontrent que les plantes et les animaux ont divergé il y a de cela un milliard d'années. Puis que la séquence comporte 105 acides aminés, on en déduit que le temps de divergence est de l'ordre de τ, et

$$2T = 2 \times 10^{-9} = 2\tau \quad \Longrightarrow \quad \lambda = \frac{10^{-9}}{100} = 10^{-11}\,.$$

Toutefois, ce type d'argument est approximatif, car il ne tient pas compte du fait que plusieurs substitutions peuvent survenir en un même site (probabilité faible pour l'histone tout de même) et provoquer des coïncidences fortuites. Nous montrons comment corriger l'estimation du nombre de mutations. Supposons que les substitutions surviennent selon un processus de Poisson de paramètre λ. En un site précis, calculons la probabilité pour que l'on observe l'acide ancestral pour les sites des deux séquences

$$\mathrm{P}(\text{Observer l'acide ancestral}) = \prod_{i=1,2} \mathrm{P}(\text{pas de substitution sur la lignee } i)$$

La probabilité pour qu'il n'y a pas de mutation sur la lignée d'une espèce depuis l'ancêtre commun est

$$\mathrm{P}(\text{pas de substitution sur la lignee } i) = e^{-\lambda T}.$$

Ainsi, nous obtenons

$$\mathrm{P}(\text{Observer l'acide ancestral}) = e^{-2\lambda T}.$$

Concrètement, nous estimons cette probabilité en comptant le nombre de différences N_d entre les deux séquences divisé par la longueur commune L des deux séquences

$$\mathrm{P}(\text{Observer l'acide ancestral}) \approx 1 - \frac{N_d}{L}.$$

Puisque le nombre moyen de mutations est $2\lambda T$ (espérance de la loi de Poisson), nous obtenons une estimation du nombre de mutations

$$\delta \approx 2\lambda T = -\log(1 - \frac{N_d}{L})$$

De cette expression, on tire un estimateur de λ dès que T est connu. Par ailleurs, si l'on connait l'horloge moléculaire λ on tire une expression du temps de divergence, utile pour une reconstruction phylogénétique. Nous voyons qu'il n'est pas possible d'estimer simultanément T et l'horloge, ces deux paramètres sont liés.