CAT : Un Modèle Phylogénétique Bayésien permettant de prendre en compte l'Hétérogénéité des Processus de Substitution entre Sites dans les Alignements Protéiques

Nicolas Lartillot, Hervé Philippe

CAT : UN MODÈLE PHYLOGÉNÉTIQUE BAYÉSIEN PERMETTANT DE PRENDRE EN COMPTE L'HÉTÉROGÉNÉITÉ DES PROCESSUS DE SUBSTITUTION ENTRE SITES DANS LES ALIGNEMENTS PROTÉIQUES

Nicolas LARTILLOT [1] et Hervé PHILIPPE [2]

1. Laboratoire d'Informatique, de Robotique et de Mathématiques de Montpellier
161, rue Ada, 34392 Montpellier Cedex 5, France
nicolas.lartillot@lirmm.fr

2. Département de Biochimie
Université de Montréal, Succursale centre-ville, Montréal H3C3J7, Québec, Canada
herve.philippe@umontreal.ca

Abstract. — We propose a Bayesian mixture model, accounting for across-site heterogeneities of the substitutional processes in protein sequences. Our model, CAT, is based on the formalism of the Dirichlet processes, in which the total number of classes of the underlying mixture is not specified *a priori*, but rather, is considered an unknown of the problem, and is directly inferred from the available data. In this paper, we describe the model, and show its connections with the Bayesian non-parametric approach for modeling heterogeneity. We apply it to a series of alignments of real proteins, and uncover a significant level of heterogeneity across sites. Finally, by the evaluation of the Bayes factor, we show that the CAT model yields a significant improvement of the statistical fit over the standard models, based on one single substitution process describing all the sites of the alignment.

INTRODUCTION

Les méthodes probabilistes sont maintenant couramment utilisées en phylogénie moléculaire. Leur avantage essentiel réside dans le fait de rendre explicite l'ensemble des présupposés sous-jacents à la reconstruction, sous la forme d'un modèle d'évolution des séquences. De plus, elles garantissent la consistance asymptotique (Wald, 1949) de la reconstruction, à savoir, la convergence vers l'arbre vrai, au fur et à mesure que croît le nombre de caractères utilisés.

Toutefois, la garantie de consistance n'est bien-sûr effective que si le modèle utilisé décrit correctement l'évolution des séquences réelles. Or, les modèles actuellement utilisés en phylogénie sont loin de prendre en compte toute la complexité du comportement des séquences moléculaires au cours de l'évolution. En particulier, ils sont le plus souvent homogènes (tous les sites d'une protéine évoluent suivant les mêmes modalités) et stationnaires (ces modalités de plus sont constantes au cours du temps). De telles hypothèses ont toutes les chances d'être violées par les données réelles, et lors, les résultats obtenus au cours d'analyses conduites sous des modèles aussi peu réalistes ne sont plus du tout garantis (Sullivan & Swofford, 2001). De fait, de nombreux artefacts sont couramment observés en reconstruction phylogénétique, et il est probable qu'au moins une partie d'entre eux soient justement dus à la violation des hypothèses des modèles par les données.

En particulier, l'hétérogénéité du processus d'évolution entre les différentes positions de l'alignement semble être un aspect essentiel de l'évolution des séquences protéiques. Cette hétérogénéité est tout

Biosystema 22, *Avenir et pertinence des méthodes d'analyse en phylogénie moléculaire*, 2004 : pages 97 à 104.

d'abord quantitative : certaines positions sont virtuellement invariables sur de grandes échelles évolutives, tandis que d'autres évoluent très rapidement. La prise en compte de cette disparité du taux d'évolution relatif des différentes positions a abouti à la création de modèles dits à taux variables (YANG, 1993), dont les performances s'avèrent largement supérieures à celles des modèles à taux uniformes. Cependant, à cette disparité quantitative du taux de substitution se superpose également une hétérogénéité qualitative : la simple observation d'un alignement multiple de protéines suggère immédiatement que la plupart des positions n'acceptent qu'un nombre très restreint d'acides aminés au cours de leur évolution. Quelques tentatives ont déjà été proposées pour prendre en compte cet aspect essentiel de l'évolution moléculaire (BRUNO, 1996 ; DIMMIC *et al*., 2000 ; GOLDMAN *et al*., 1998 ; HALPERN & BRUNO, 1998 ; KOSHI & GOLDSTEIN, 1998, 2001 ; THORNE *et al*. 1996), mais cependant tous ces modèles présupposent un degré d'hétérogénéité fixé *a priori* (et généralement peu élevé), et de ce fait, aucun n'aborde vraiment la question de mesurer l'étendue réelle du phénomène.

Dans cette perspective, nous proposons un modèle probabiliste permettant de prendre en compte l'hétérogénéité substitutionnelle, en s'affranchissant de cette limitation. Ce modèle, appelé CAT (pour CATégories de substitution), fonctionne comme un modèle de mélange, dont le nombre de classes, plutôt que d'être fixé *a priori*, est un des paramètres du problème. Par un traitement bayésien, le nombre de classes, ainsi que tous les autres paramètres du modèle, sont estimés par échantillonnage, en utilisant le principe des Chaînes de Markov Monte Carlo (MCMC) (les principes de base de l'analyse bayésienne sont développés dans l'article de DELSUC & DOUZERY, dans ce volume de *Biosystema*).

Dans cet article, nous présentons brièvement le principe du modèle CAT. En particulier nous expliquons comment l'interpréter sous un angle non-paramétrique, c'est-à-dire, équivalent à un modèle où chaque site aurait son propre profil de substitution, tiré d'une distribution statistique directement inférée à partir des données. Enfin, nous appliquons le modèle CAT à des données réelles, et par évaluation du facteur de Bayes, nous montrons que ses performances sont supérieures à celles des modèles standards, présupposant un seul processus de substitution pour décrire l'évolution de l'ensemble des sites de l'alignement.

DONNÉES, MATÉRIELS ET MÉTHODES

Données et arbres

Trois alignements de séquences en acides aminés ont été obtenus ou constitués. Dans la suite, ils sont désignés selon une nomenclature indiquant successivement le nombre de taxons et le nombre de positions qu'ils contiennent. (1) EF30-627 : séquences du facteur d'élongation 2, chez un échantillon de 30 espèces eukaryotes. (2) Ek55-1525 : concaténation des séquences de tubulines alpha et béta, ainsi que de l'actine et du facteur d'élongation 1 alpha, chez 55 eukaryotes. L'alignement a été obtenu auprès de SANDRA BALDAUF (BALDAUF *et al*., 2000). Les séquences des diplomonadines et des trichomonadines, jugées trop divergentes, ont été supprimées de l'alignement initial. (3) Mt45-3596 : concaténation de l'ensemble des séquences codantes du génome mitochondrial de 45 mammifères. Dans les cas (1) et (3), les alignements ont été construits à l'aide de ClustalW (HIGGINS *et al*., 1994), puis repris manuellement, de manière à ne garder que les régions de l'alignement les moins ambiguës.

Pour des raisons de temps de calcul, les analyses présentées dans ce travail ont été conduites sous la contrainte d'une topologie fixe. Dans les cas (1) et (3), l'arbre utilisé est le consensus majoritaire *a posteriori*, tel que calculé par MrBayes (3.0) (HUELSENBECK & RONQUIST, 2001), sous le modèle JTT pour (1), et mtREV pour (3), avec des taux de substitution entre sites distribués selon une loi gamma discrétisée en 16 catégories. Pour (2), l'arbre est celui proposé par les auteurs dans BALDAUF *et al*. (2000), après avoir éliminé les branches correspondant aux diplomonadines et aux trichomonadines.

MODÈLES

L'ensemble des modèles et des méthodes employés dans cet article sont décrits en détail dans LARTILLOT & PHILIPPE (2004). Brièvement, nous reprenons la formulation classique des modèles phylogénétiques bayésiens (HUELSENBECK *et al*., 2002, LARGET & SIMON, 1999), les modèles plus spécifiquement présentés dans cet article ne différant fondamentalement des autres qu'au niveau des processus de substitution sous-jacents, qui font l'objet de l'essentiel des développements qui suivent.

Modèles Markoviens de substitution

Pour chaque site, la série des substitutions se produisant au cours du temps est modélisée par un processus de Markov réversible, opérant le long des branches de l'arbre. Un tel processus est entièrement caractérisé par une matrice 20×20, appelée matrice infinitésimale, ou matrice de substitution (Q). Cette matrice peut être décomposée en deux jeux de paramètres : d'une part, des taux relatifs d'échange (190 paramètres), et d'autre part, des probabilités stationnaires (ou fréquences d'équilibre) pour les 20 acides aminés $f = (f_l)_{l = 1..20}$ tels que $\sum_{l = 1..20} f_l = 1$. Un cas particulier de processus Markovien réversible est obtenu en fixant l'ensemble des taux relatifs à 1 : on obtient alors un processus de Poisson, qui est entièrement défini par son jeu de probabilités stationnaires *f*, que nous appelerons profil dans la suite de cet article.

Modélisation de l'hétérogénéité entre sites

Dans les modèles standards, l'ensemble des sites de l'alignement suivent tous le même processus de substitution, décrit par une seule matrice infinitésimale commune à toutes les positions. Dans cet article, nous modélisons l'hétérogénéité qualitative entre sites en introduisant un mélange de *K* classes. Chaque classe est caractérisée par sa propre matrice de substitution Q^k, pour k = 1..K. D'autre part, l'affiliation de chaque site (*i*, pour *i* = 1..N) à l'une de ces classes est spécifiée par une variable entière z_i, prenant une valeur comprise entre 1 et *K*. Le vecteur $\mathbf{z} = (z_i)_{i = 1..N}$, où *N* est le nombre de sites de l'alignement, est appelé vecteur d'allocation. De plus, pour ne pas rendre le modèle trop complexe, on ne considère ici que des mélanges de processus de Poisson. De la sorte, chaque classe est entièrement définie par son profil $f^k = (f^k_l)_{l = 1..20}$.

Le nombre de classes du mélange, *K*, est considéré comme une inconnue du problème, au même titre que les affiliations des sites (**z**) et les profils de chaque classe (f^k). Il nous faut alors définir une loi *a priori* sur l'ensemble de ces paramètres, y compris K. Pour ce faire, nous utilisons un processus de Dirichlet (FERGUSON, 1973). Un tel processus est entièrement caractérisé par son paramètre de concentration, appelé *a*. La distribution *a priori* résultant de ce processus peut être décomposée en deux facteurs : tout d'abord, sachant *a*, la probabilité *a priori* de la configuration (*K*, *z*) vaut :

$$(1) \quad p(K, z \mid a) \quad = \quad a^N \frac{\prod_{k=1..K}(n_k - 1)!}{\prod_{i=1..N}(a + i - 1)}$$

où n_k représente le nombre de sites affiliés à la classe *K*. Ensuite, sachant *K*, les vecteurs de probabilités stationnaires définissant chaque composante du mélange sont *a priori* i.i.d. selon une distribution uniforme sur l'ensemble des profils possibles (Dirichlet plate, appelée dans la suite G_0). L'hyperparamètre *a* permet de régler la taille moyenne *a priori* du mélange, avec de grandes valeurs de *a* favorisant un grand nombre de classes. Il est également inconnu, et nous considérons une distribution *a priori* uniforme sur sa valeur.

Alternativement, *K* peut être spécifié *a priori*. Sous ce cas de figure, nous ne considérerons en fait qu'une situation, *K* = N (modèle maximalement hétérogène, ou MAX). Sous le modèle MAX, chaque site possède son propre propre processus de substitution, différant des autres par son jeu de probabilités stationnaires. Les profils spécifiques de chaque site sont alors supposés i.i.d., tirés *a priori* suivant G_0.

Implémentation par la méthode des Chaînes de Markov Monte Carlo bayésiennes

Pour échantillonner la distribution de probabilité *a posteriori*, nous employons la méthode des Chaînes

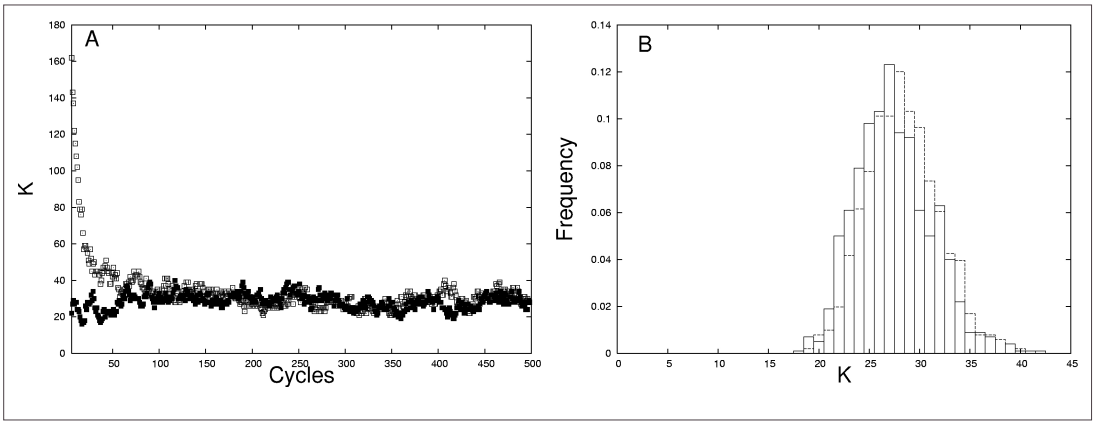

Figure 1.
a. Évolution du nombre de classes du mélange (K), au cours de deux chaînes MCMC (500 premiers cycles), avec comme conditions initiales $K = 1$ (cercles) et $K = N$ (carrés), N = 627 étant le nombre de sites de l'alignement (EF30-627).
b. Histogramme montrant les fréquences observées pour les différentes valeurs de K, au cours des deux chaînes figurées en A.

de Markov Monte Carlo. L'application de la technique du MCMC à la reconstruction phylogénétique est décrite dans HUELSENBECK *et al.* (2002), HUELSENBECK & RONQUIST (2001), et LARGET & SIMON (1999). Quant aux méthodes spécifiques employées dans le cadre des modèles qui font l'objet de cet article, elles sont détaillées dans LARTILLOT & PHILIPPE (2004).

ÉVALUATION DE LA PERFORMANCE STATISTIQUE DES MODÈLES

À des fins de comparaison, outre les modèles, MAX et CAT, nous considérons également une série de modèles standards, basés sur des matrices empiriques (Poisson + F, WAG + F, JTT + F ou mtREV + F, où le suffixe « + F » désigne le fait que les probabilités stationnaires du processus de substitution unique défini par le modèle sont identifiées aux fréquences empiriques observées dans le jeu de données). Enfin, nous envisageons également le modèle le plus général, sous l'hypothèse d'homogénéité substitutionnelle entre sites : le modèle GTR, correspondant à une matrice dont l'ensemble des paramètres, taux relatifs d'échange et probabilités stationnaires, sont considérés comme des inconnues, et sont donc inférés à partir des données.

En inférence bayésienne, la performance statistique d'un modèle M est mesurée à l'aune de sa vraisemblance marginale p (D | M), définie comme l'intégrale de la vraisemblance sur l'ensemble des paramètres, l'intégrale étant prise par rapport aux distribution de probabilité *a priori*. Une manière pratique d'exprimer ceci, lorsque l'on compare deux modèles particuliers, est de calculer le facteur de Bayes entre ces deux modèles, M_0 et M_1, défini comme le rapport de leurs vraisemblances marginales (JEFFREYS, 1935) :

(4) $BF = p(D \mid M_1) / p(D \mid M_0)$

On peut également définir le support en faveur de M_1, relativement à M_0, comme le logarithme naturel du facteur de Bayes. En pratique, lorsque plus de deux modèles sont considérés simultanément, on calculera le support relatif de chacun de ces modèles à un modèle de référence (présentement, le modèle Poisson + F).

Conceptuellement, la vraisemblance marginale est une manière simple et élégante d'évaluer la pertinence empirique d'un modèle. Par contre, l'évalua-

tion numérique du facteur de Bayes n'est pas du tout évidente. Dans le cadre de ce travail, nous avons utilisé la méthode dite d'intégration thermodynamique (Ogata, 1989). Les détails sont expliqués dans Lartillot & Philippe (2004).

RÉSULTATS

Analyse du nombre de classes moyen *a posteriori*

Dans un premier temps, nous avons appliqué le modèle CAT au jeu de données EF30-627. La figure 1.a montre l'évolution du nombre de classes (K), au cours de deux chaînes de Markov indépendantes, initialisées à K = 1 et K = N. Dans les deux cas, le nombre de classes se stabilise autour d'une valeur moyenne de K = 28,4 environ. Le modèle CAT infère donc une hétérogénéité substitutionnelle significative entre les positions de l'alignement étudié. La figure 1.b montre la distribution des différentes valeurs de *K* prises, à l'équilibre, pour chacune des deux chaînes. Conformément au principe de l'échantillonnage par MCMC, la fréquence à laquelle une valeur donnée de *K* est observée au cours d'une chaîne tend asymptotiquement vers la probabilité *a posteriori* de cette valeur de *K*. Ainsi, par exemple, peut-on estimer que $p\,(K = 27 \mid \mathrm{D}, \mathrm{CAT}) = 0{,}12$. En particulier, on constate que $p\,(K = 1 \mid \mathrm{D}, \mathrm{CAT})$ est virtuellement égale à 0 : autrement dit, le modèle homogène Poisson + F est fortement rejeté. D'un autre côté, la configuration *K* = N est également très peu probable *a posteriori*, ce qui suggère que la diversité substitutionnelle observée à travers l'alignement se ramène malgré tout à un nombre relativement restreint de catégories.

Modèles de mélange et inférence non-paramétrique

Un aspect fondamental des modèles de mélange basés sur des processus de Dirichlet (plus généralement, des modèles de mélange à nombre libre de composantes), est la double interprétation que l'on peut en faire : en tant que modèles de mélange *stricto sensu*, et en tant que modèles non-paramétriques (Green & Richardson, 1998). En effet, d'un côté, on peut

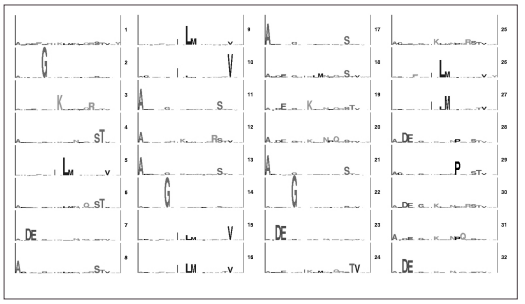

Figure 2. Profil moyen observé pour les 32 premiers sites de l'alignement EF30-627. Les acides aminés sont désignés selon le code usuel à une lettre. La hauteur de chaque lettre est proportionnelle à la probabilité stationnaire de l'acide aminé correspondant.

prendre le modèle au pied de la lettre, et considérer que les données sont effectivement un mélange composé d'un nombre fini de catégories bien définies d'observations, chaque catégorie présentant ses caractéristiques statistiques propres. Sous cet angle, le modèle CAT n'est finalement qu'un modèle de mélange classique, si ce n'est que l'on a essayé d'inférer également le nombre de composantes effectivement présentes dans les données réelles. Alternativement, et cette perspective est beaucoup moins naïve, le modèle CAT peut être vu comme un modèle non-paramétrique, c'est-à-dire équivalent à un modèle où chaque site aurait son propre profil de substitution, tiré d'une distribution statistique directement inférée à partir des données.

Une manière d'illustrer cet aspect non-paramétrique du modèle CAT consiste à calculer le profil moyen *a posteriori* en chaque site. En effet, si l'ensemble des paramètres définissant le mélange (nombre de classes, profils de chaque classe, affiliation des sites aux classes) fluctuent au cours d'une chaîne MCMC, il n'en reste pas moins que, à tout moment, chaque site est affilié à une classe bien définie, et l'on peut donc calculer, pour un site donné, la moyenne des profils des classes auxquelles ce site a été affilié au cours de la chaîne. Ces profils moyens peuvent ensuite être visualisés de manière compacte (fig. 2). On remarque que, bien que les profils semblent se ramener à un ensemble relativement

réduit de profils-types, ils sont néanmoins distincts en chaque site.

D'un point de vue biologique, l'ensemble de ces profils moyens semblent raisonnables, dans la mesure où ils regroupent en général des acides aminés biochimiquement équivalents. Il y a deux autres aspects marquants, concernant les profils : ils sont très disparates, même entre sites voisins, et de plus ils se concentrent presque tous sur un, deux ou trois acides aminés seulement. Ceci semble suggérer l'existence d'une pression de sélection à la fois pointue, ciblée, et hautement contextuelle, ce qui est tout à l'opposé de ce que présupposent les modèles couramment employés en reconstruction phylogénétique.

Application à d'autres jeux de données.

L'application du modèle CAT à d'autres jeux de protéines donne des résultats similaires (tableau 1) : dans tous les cas, le nombre de classes inféré est relativement élevé, les profils-types sont globalement de même nature que ceux observés lors de l'analyse effectuée sur les facteurs d'élongation (résultats non montrés). À noter que le nombre de classes tend à croître avec l'alignement, et également, que la longueur totale de l'arbre inférée sous CAT est systématiquement, et significativement, plus grande que sous les autres modèles.

Évaluation de la performance statistique

Enfin, nous avons mesuré la performance statistique des différents modèles utilisés au cours de cette analyse, sur les trois jeux de données EF, Ek et Mt, par évaluation du facteur de Bayes (tableau 2). On retrouve tout d'abord un certain nombre de résultats déjà connus : par exemple, les matrices empiriques sont nettement meilleures que le modèle Poisson ; parmi elles, Wag est meilleure que JTT, (sauf sur les données mitochondriales, pour lesquelles c'est sans surprise que l'on observe que mtREV est la matrice la plus performante). D'autre part, le modèle GTR n'est pas nécessairement meilleur que les matrices empiriques. Dans le cas présent, ceci est probablement dû au fait qu'il n'y a pas assez de signal dans les alignements étudiés,

	<K>	<L>		
	CAT	CAT	MAX	JTT + F
EF30-627	28,4 ± 3,5	8,13 ± 0,21	7,06 ± 0,14	7,46 ± 0,14
Ek55-1525	26,4 ± 3,7	5,36 ± 0,10	4,82 ± 0,07	4,78 ± 0,08
Mt45-3596	35,3 ± 3,2	7,54 ± 0,10	5,90 ± 0,05	5,91 ± 0,05

Tableau 1. Espérance *a posteriori* du nombre de classes sous Cat (<K>), et longueur totale de l'arbre (<L>) sous Cat, Max et JTT + F, estimée sur trois jeux de données distincts.

permettant de bien inférer l'ensemble des paramètres d'une matrice de substitution. Mais surtout, CAT est, dans tous les cas, le plus performant parmi tous les modèles. À noter enfin que le modèle MAX (K = N) est un très mauvais modèle, en réalité, le plus mauvais après Poisson.

Ce dernier point, la mauvaise performance de MAX, est à relever dans le cadre de l'interprétation non-paramétrique : en effet, ainsi que nous l'avons indiqué plus haut, CAT est en réalité équivalent à un modèle qui, à l'instar de MAX, permet à chaque site d'avoir son profil propre, la seule différence entre MAX et CAT résidant finalement dans la loi statistique suivie par ces profils site-spécifiques : dans le cadre de MAX il s'agit d'une loi uniforme sur l'ensemble de profils, tandis que CAT considère cette loi comme inconnue, et l'infère à partir des données. On voit donc ici que le facteur de Bayes, et donc la performance du modèle, est extrêmement sensible à la loi statistique sous-jacente.

	JTT + F	*WAG + F*	*mtREV + F*	*GTR*	*CAT*	*MAX*
EF30-627	1631	1760	1374	1550	1932	807
Ek55-1525	3324	3628	2848	–	3808	1112
Mt45-3596	10089	8970	10943	–	12389	855

Tableau 2. Valeurs du support (logarithme naturel du facteur de Bayes) en faveur de chacun des modèles, le modèle Poisson + F étant pris comme référence. Les estimations ont été obtenues par intégration thermodynamique.

Prises ensemble, ces comparaisons de modèles montrent que, s'il est important de prendre en compte l'hétérogénéité du comportement de chaque position des protéines au cours de l'évolution, il ne suffit pas pour autant d'invoquer une série de variables de sites permettant formellement de rendre le modèle hétérogène, mais il faut en outre bien cerner le problème des lois statistiques sous-jacentes. Le mauvais score du modèle MAX montre à quel point ce dernier point est crucial. Et inversement, la performance du modèle CAT illustre la manière dont les méthodes non-paramétriques permettent d'apporter à ce problème sa réponse la plus simple : estimer directement ces lois à partir des données disponibles. En outre, la démarche adoptée au cours de ce travail semble pouvoir facilement se généraliser aux autres composantes définissant les processus de substitutions (en particulier, les taux relatifs d'échange entre acides aminés), permettant d'envisager à terme, la construction d'un modèle général site-hétérogène.

Vers une application à la reconstruction phylogénétique

L'ensemble des résultats montrés dans cet article ont certes tous été obtenus en contraignant la topologie, suivant des critères extérieurs, et il reste encore à appliquer le modèle CAT plus directement au problème de la reconstruction phylogénétique, afin de déterminer si la prise en compte de l'hétérogénéité substitutionnelle entre sites apporte réellement une amélioration. Mais d'ores et déjà, deux observations suggèrent une réponse affirmative à cette question. D'une part, comme nous l'avons déjà observé précédemment, les profils moyens observés en la plupart des sites sont très pointus, au sens où ils se concentrent sur 2 ou 3 acides aminés. Ce point est crucial, dans une perspective de reconstruction phylogénétique, où un des enjeux majeurs, pour obtenir des résultats fiables, est de bien détecter les homoplasies (Felsenstein, 1978). En effet, prenons l'exemple de deux espèces présentant le même résidu à une position donnée. En caricaturant un peu, une méthode présupposant que les vingt acides aminés sont *a priori* également acceptables à cette position évaluera la probabilité que cette identité de séquence entre les deux espèces soit due à une homoplasie à environ 1/20. Si pourtant, du fait des contraintes biochimiques, et comme le suggèrent les profils observés sous CAT, cette position n'accepte en réalité que deux acides aminés (par exemple, une isoleucine ou une valine), cette probabilité devrait être plutôt estimée aux alentours de 1/2. Ce petit argument qualitatif suggère très fortement que l'absence de prise en compte de la spécificité des contraintes biochimiques en à chaque position, et de leur disparité à travers les séquences, est une cause potentiellement très importante d'artefacts de reconstruction. Une seconde observation vient conforter cette interprétation : pour tous les jeux de données, la longueur totale des arbres inférés sous le modèle CAT est significativement plus élevée que sous les autres modèles, ce qui semble suggérer que le modèle CAT propose effectivement une meilleure prise en compte de la saturation des données, et par là, pourrait aboutir à une reconstruction phylogénétique plus fiable.

BIBLIOGRAPHIE

Baldauf S.L., Roger A.J., Wenk-Siefert I., & Doolittle W.F., 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290 : 972-977.

Bruno W.J., 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, 13 : 1368-1374.

Dimmic M.W., Mindell D.P. & Goldstein R.A., 2000. Modeling evolution at the protein level using an adjustable amino acid fitness model. *Pac. Symp. Biocomput.*, 5 : 18-29.

Felsenstein J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27 : 401-410.

Ferguson T., 1973. A Bayesian analysis of some nonparametric problems. *Ann. Statistics*, 1 : 209-230.

Goldman N., Thorne J.L. & Jones D., 1998. Assessing the impact of secondary structure and solvent accessiblity on protein evolution. *Genetics*, 149 : 445-458.

GREEN P.J. & RICHARDSON S., 1998. *Modelling heterogeneity with and without the Dirichlet process*. Technical report, University of Bristol, Bristol, U.K.

HALPERN A.L. & BRUNO W.J., 1998. Evolutionary distances for protein-coding sequences : modeling site-specific residue frequencies. *Mol. Biol. Evol*., 15 : 910-917.

HIGGINS D., THOMPSON J., GIBSON T., THOMPSON J.D., Higgins D.G., Gibson T.J. (1994). Clustal W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22 : 4673-4680.

HUELSENBECK J.P., LARGET B., MILLER R.E. & RONQUIST F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol*., 51 : 673-688.

HUELSENBECK J.P. & RONQUIST F., 2001. MrBayes : Bayesian inference of phylogenetic trees. *Bioinformatics*, 17 : 754-755.

JEFFREYS H., 1935. Some tests of significance, treated by the theory of probability. *Proc. Camb. Phil. Soc*., 31 : 203-222.

KOSHI J.M. & GOLDSTEIN R.A., 1998. Models of natural mutations including site heterogeneity. *Proteins*, 32 : 289-295.

KOSHI J.M. & GOLDSTEIN R.A., 2001. Analysing site heterogeneity during protein evolution. *Pac. Symp. Biocomput*., 6 : 191-202.

LARGET B. & SIMON D., 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol*., 16 : 750-759.

LARTILLOT N. & PHILIPPE H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol*., 21 : 1096-1109.

NEAL R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graphical. Stat*., 9 : 249-265.

OGATA Y., 1989. A Monte Carlo method for high dimensional integration. *Numerische Mathematik*, 55 : 137-157.

SULLIVAN J. & Swofford D.L., 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site variations and nucleotide substitution pattern are violated ? *Syst. Biol*., 50 : 723-729.

THORNE J.L., GOLDMAN N. & JONES D.T., 1996. Combining protein evolution and secondary structure. *Mol. Biol. Evol*., 13 : 666-673.

WALD A., 1949. Note on the consistency of maximum likelihood. *Ann. Math. Stat*., 20 : 595-601.

YANG Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol*., 10 : 1396-1401.