Les méthodes probabilistes en phylogénie moléculaire: (1) Les modèles d'évolution des séquences et le maximum de vraisemblance

Frédéric Delsuc, Emmanuel J.P. Douzery

LES MÉTHODES PROBABILISTES EN PHYLOGÉNIE MOLÉCULAIRE

(1) Les modèles d'évolution des séquences et le maximum de vraisemblance

Frédéric DELSUC et Emmanuel J. P. DOUZERY

Laboratoire de Paléontologie, Phylogénie et Paléobiologie,
Institut des Sciences de l'Évolution de Montpellier (ISEM), UMR 5554-CNRS,
Université Montpellier II, Montpellier, France
delsuc@isem.univ-montp2.fr

Abstract. — The success of the phylogenetic approach within its broad range of potential applications requires the development of powerful inference methods and a clear definition of their conditions for application. The theoretical efforts thus attempted to consider the problem of phylogenetic reconstruction within a statistical framework with the development of methods incorporating sequence evolution on their basic assumptions. In this context, probabilistic methods based on the likelihood function appeared particularly adapted to the analysis of molecular data. These methods based on explicit models of sequence evolution have the advantage to allow the application of a broad range of statistical tests to evaluate various evolutionary hypotheses. The maximum likelihood method profited from many successive improvements aiming at describing as well as possible the processes of biological sequence evolution. This method appears particularly powerful and its desirable statistical properties make it the currently most used method in the fields of molecular evolution and phylogenetic reconstruction.

Résumé. — Le succès de l'approche phylogénétique à l'intérieur d'un large champ d'applications potentielles nécessite le développement de méthodes d'inférence performantes et une définition claire de leurs conditions d'application. Des efforts théoriques se sont ainsi attachés à considérer le problème de la reconstruction phylogénétique dans un cadre statistique avec le développement de méthodes incorporant l'évolution des séquences dans leurs hypothèses de base. Dans ce contexte, les méthodes probabilistes fondées sur la fonction de vraisemblance apparaissent particulièrement adaptées au traitement des données moléculaires. Ces méthodes basées sur des modèles explicites de l'évolution des séquences possèdent l'avantage de permettre l'application d'une large gamme de tests statistiques pour évaluer différentes hypothèses évolutives. La méthode du maximum de vraisemblance a ainsi bénéficié de nombreuses améliorations successives visant à décrire au mieux les processus d'évolution des séquences biologiques. Cette méthode est particulièrement performante et ses propriétés statistiques séduisantes en font la méthode la plus utilisée à l'heure actuelle dans les domaines de l'évolution moléculaire et de la reconstruction phylogénétique.

INTRODUCTION AUX MÉTHODES PROBABILISTES

Les méthodes probabilistes sont fondées sur le concept de vraisemblance. Dans sa forme générale, la vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier. Étant donné un modèle qui spécifie les probabilités d'observer différents évènements, la vraisemblance L d'obtenir les données observées peut être calculée : $L_X = \Pr(X \mid H)$, où $\Pr(X \mid H)$ est la probabilité conditionnelle d'observer les données X sous l'hypothèse H. Dans le contexte phylogénétique, la fonction de vraisemblance peut-être vue comme le véhicule de l'information phylogénétique contenue dans les données (HUELSENBECK & BOLLBACK, 2001). Deux types de méthodes probabilistes basées sur le concept de vraisemblance ont été développées

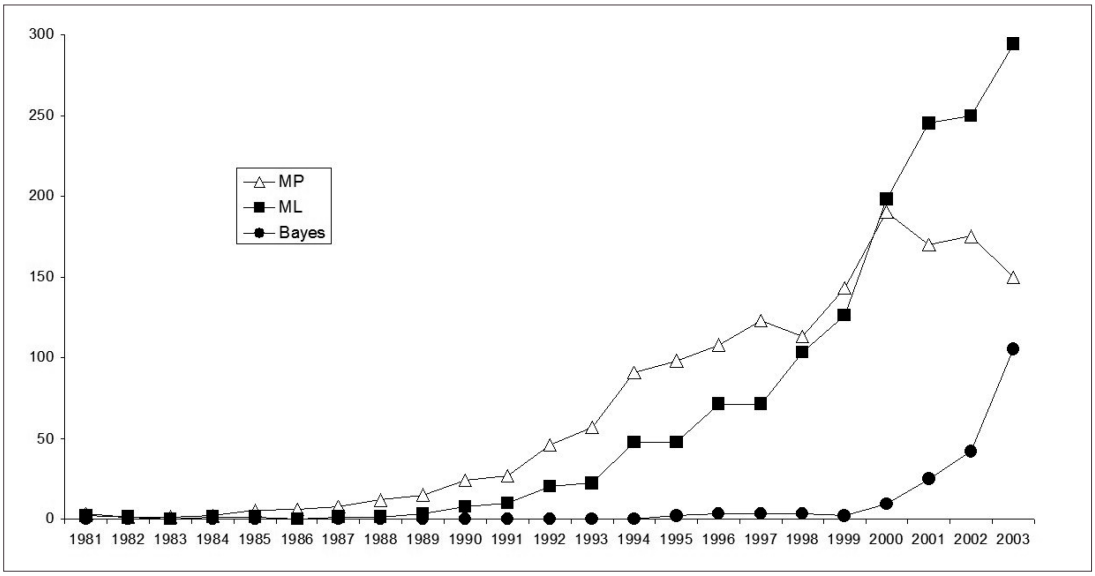

Figure 1. Evolution de l'utilisation de trois méthodes de reconstruction phylogénétique (MP : Maximum de parcimonie, ML : Maximum de vraisemblance, Bayes : Approche Bayésienne) depuis l'article de référence sur le maximum de vraisemblance de Felsenstein (1981). Le nombre de citations par année correspond au nombre de publications référencées dans la base de données PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) contenant les mots : « Parsimony » ET « Phylogeny », « Likelihood » ET « Phylogeny », et « Bayes » ET « Phylogeny » dans le titre ou le résumé.

et appliquées au problème statistique de l'estimation des phylogénies : la méthode du maximum de vraisemblance et plus récemment l'approche bayésienne. Bien qu'étant fondées toutes les deux sur la fonction de vraisemblance utilisant les mêmes modèles d'évolution des séquences, ces deux méthodes diffèrent par le concept de probabilité qu'elles emploient. Ainsi, la méthode du maximum de vraisemblance sélectionne l'arbre qui maximise la vraisemblance (l'arbre ayant la plus forte probabilité d'avoir conduit aux données), alors que l'approche bayésienne assigne une distribution de probabilités *a priori* aux différents arbres et réalise des inférences à partir de la distribution des probabilités *a posteriori* de ces arbres.

Introduite par EDWARDS et CAVALLI-SFORZA dans les années 1960 (EDWARDS & CAVALLI-SFORZA, 1967), puis développée par FELSENSTEIN (1973), la méthode du maximum de vraisemblance a été initialement limitée par la puissance de calcul des ordinateurs de l'époque. Ce n'est qu'au début des années 1980 que ce problème a été contourné grâce à la description de l'algorithme de « pruning » par FELSENSTEIN (1981) permettant le calcul de la vraisemblance à partir de séquences d'ADN pour un nombre élevé de taxons. À partir de ce moment là, l'utilisation de la méthode du maximum de vraisemblance va connaître une croissance quasi exponentielle, parallèlement à l'accumulation des données moléculaires dans les banques de séquences (fig. 1). La popularité grandissante des méthodes probabilistes réside sans doute dans leurs fondements statistiques bien définis (WHELAN *et al*., 2001). L'incorporation de modèles explicites d'évolution des séquences, dont les paramètres peuvent être estimés au cours de l'analyse, confère en effet à ces méthodes la capacité d'estimer de façon simultanée la phylogénie et le mode d'évolution des séquences. Une telle propriété rend ces méthodes particulièrement prometteuses dans la mesure où elles sont capables d'incorporer, dans des modèles de plus en plus complexes, les processus d'évolution des séquences au fur et à mesure qu'ils sont découverts.

LA FONCTION DE VRAISEMBLANCE

Le concept de vraisemblance appliqué à la phylogénie postule que la vraisemblance L_X est la probabilité conditionnelle d'observer les données X (un alignement de séquences) étant donné un arbre T (avec une topologie et des longueurs de branches données) :

$$L_X = \Pr(X \mid T)$$

L'alignement de séquences peut être représenté par une matrice X avec s taxons (lignes) et c sites (colonnes) dont les cases x_{ij} dénotent l'état nucléotidique observé pour le taxon i au site j :

$$X = [x_{ij}] = \begin{bmatrix} A & C & T & G & \cdots\cdots & C & G & T & G & C \\ A & C & G & T & \cdots\cdots & A & G & C & A & C \\ C & C & T & G & \cdots\cdots & T & G & T & A & C \\ A & G & T & G & \cdots\cdots & T & G & T & A & C \end{bmatrix}$$

Les sites x sont considérés comme étant indépendants et sont notés x_1 = (AACA), x_2 = (CCCG), …, x_c = (CCCC). Il faut ensuite calculer la probabilité d'obtenir chaque site de l'alignement. La probabilité d'observer les états de caractères pour un site donné dépend de la topologie τ de l'arbre, de ses longueurs de branches $v = (\upsilon_1, \upsilon_2, \ldots, \upsilon_b)$ et du modèle décrivant l'évolution des séquences le long de ses branches. Le modèle d'évolution définit la probabilité de changement $p_{ij}(\upsilon, \theta)$ d'un état i à un état j le long d'une branche de longueur υ selon les paramètres θ du modèle. Ces probabilités sont obtenues en prenant l'exponentielle de la matrice Q de taux instantanés de changement entre les différents nucléotides, laquelle est spécifique de chaque modèle d'évolution des séquences (voir § 3).

Le calcul de la probabilité d'observer les données au site considéré est la somme des probabilités d'observer les différents états nucléotidiques possibles à ce site et à chaque nœud interne de l'arbre. Les taxons étant numérotés 1, 2,…, s, les nœuds internes de l'arbre seront notés $s + 1$, $s + 2$,…, $2s - 2$ pour un arbre raciné dont la racine est numérotée $2s - 1$. De plus, si l'ancêtre du nœud k est désigné par $\sigma(k)$ et la longueur de la $k^{\text{ième}}$ branche par υ_k, alors la probabilité d'observer les données au site i sera le résultat de l'expression (1) où y_{ij} est le nucléotide (non observé) au $j^{\text{ième}}$ nœud pour le site i et π_{y2s-1} est la fréquence du nucléotide y (non observé) à la racine de l'arbre. La somme se fait sur les 4^{s-1} possibilités d'assignation des quatre nucléotides aux nœuds internes de l'arbre. Cependant, le nombre de combinaisons possibles devenant très vite immense pour un nombre de taxons supérieur à 10, la sommation se fait grâce à l'algorithme de « pruning » de Felsenstein (1981) qui tire avantage de l'information topologique dans son calcul.

Ainsi, en admettant l'indépendance des différents sites du jeu de données, la probabilité d'observer l'alignement de séquences est simplement le produit des probabilités des différents sites (2).

Cette quantité étant en général un nombre très petit (multiplication de nombres compris entre 0 et 1), le logarithme de vraisemblance est utilisé pour obtenir des valeurs plus facilement interprétables, et surtout manipulables d'un point de vue informatique. La vraisemblance de l'alignement de séquences (X) devient alors la somme des logarithmes de vraisemblance aux différents sites (3).

$$f(x_i \mid \tau, v, \theta) = \sum_{y}^{4^{s-1}} \left[\pi_{y2s-1} \left(\prod_{k=1}^{s} p_{y\sigma(k),\, x_{ik}}(\upsilon_k, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{y\sigma(k),\, y_{ik}}(\upsilon_k, \theta) \right) \right] \qquad (1)$$

$$f(X \mid \tau, v, \theta) = \prod_{i=1}^{c} f(x_i \mid \tau, v, \theta) \qquad (2)$$

$$\ln L_X = \sum_{i=1}^{c} \ln L_i \qquad (3)$$

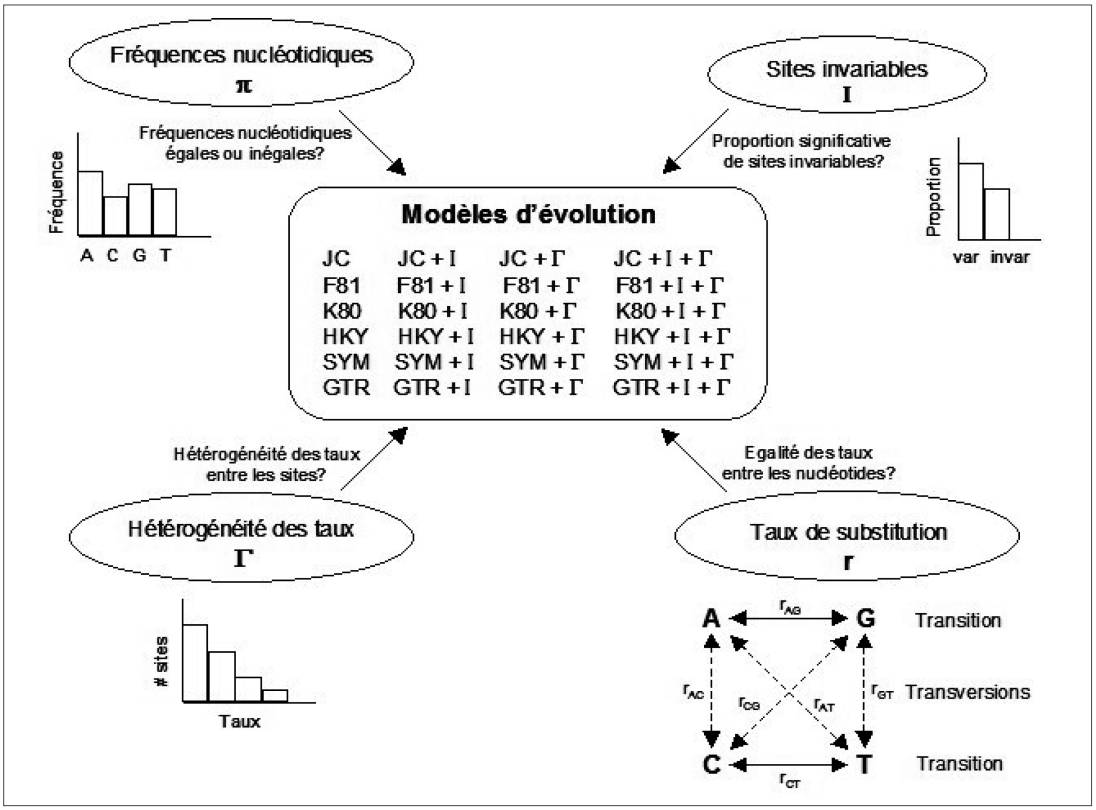

Figure 2. Constitution des principaux modèles d'évolution des séquences nucléotiques. Adapté d'après Posada & Crandall (2001).

LES MODÈLES D'ÉVOLUTION DES SÉQUENCES

Comme nous l'avons vu, le calcul de la fonction de vraisemblance dans sa forme la plus simple nécessite un modèle décrivant les probabilités de substitution d'un nucléotide par un autre. Cependant, pour être statistiquement robustes et performantes, les méthodes probabilistes nécessitent l'incorporation de modèles qui décrivent de la façon la plus réaliste possible les processus biologiques d'évolution des séquences. Ces modèles incorporent ainsi généralement des paramètres prenant en compte les fréquences nucléotidiques, les taux de substitutions, la présence d'une fraction de sites invariables et l'hétérogénéité des taux de substitution entre sites (fig. 2).

Les modèles de substitution

Les modèles de substitution permettent de calculer les probabilités des différents changements observés entre les séquences sous différentes hypothèses concernant leur processus d'évolution. Une hypothèse de base de ces modèles réside dans le fait que les processus de substitution (remplacement d'un nucléotide par un autre à un site donné) correspondent à un processus de Markov où la probabilité de passer de l'état *i* à l'état *j* dépend uniquement de l'état *i* et non pas des événements précédant cet état. De plus, il est en général supposé que ce processus de Markov est homogène c'est-à-dire que les probabilités de substitutions ne changent pas le long de chacune des branches de l'arbre. Sous ces conditions, un modèle d'évolution des séquences d'ADN

peut être mathématiquement représenté par une matrice 4×4 de taux instantanés décrivant la vitesse à laquelle chaque nucléotide est remplacé par chaque nucléotide alternatif. Cette matrice notée Q possède des composantes Q_{ij} qui correspondent aux taux r_{ij} de changement du nucléotide *i* en nucléotide *j* dans un intervalle de temps infinitésimal (Swofford *et al*., 1996).

De nombreuses versions successives de la matrice générale Q incorporant un nombre sans cesse croissant de paramètres ont été proposées depuis le modèle pionnier de Jukes & Cantor (1969). Ces modèles sont considérés comme étant réversibles dans le temps, c'est-à-dire que le taux de changement global d'un nucléotide *i* vers un nucléotide *j* est égal au taux de changement global du nucléotide *j* vers le nucléotide *i* ($r_{ij} = r_{ji}$). Ainsi, ces matrices sont généralement symétriques et le modèle le plus général (Yang, 1994a), autorisant des taux différents pour les six types de substitutions est appelé modèle GTR (pour « General Time Reversible ») ou REV (pour « Reversible »). Notons que ce dernier modèle s'accompagne aussi de fréquences en A, C, G, et T propres à chaque nucléotide. La figure 3 présente les matrices de taux instantanés correspondant aux principaux modèles utilisés pour décrire l'évolution des séquences d'ADN, ainsi que leurs relations hiérarchiques. Les valeurs numériques des paramètres de ces modèles n'étant pas connues *a priori*, elles sont généralement estimées par maximum de vraisemblance au cours de l'analyse. Chaque jeu de données possède donc virtuellement son propre ensemble de valeurs optimales (c'est-à-dire maximisant la vraisemblance des données observées) pour les différents paramètres du modèle.

Aussi généraux soient-ils, les modèles de substitution présentés précédemment sont basés sur l'hypothèse de l'indépendance des taux de substitution entre les sites. Ils ne tiennent donc pas compte de l'organisation structurale des molécules. Ainsi certains auteurs ont développé des modèles structuraux prenant en compte la structure en « tiges » et « boucles » des ARN (Schöniger & Von Haeseler, 1994 ; Muse, 1995 ; Tillier & Collins, 1998). Savill *et al*. (2001) présentent une comparaison statistique et une discussion de ces différents modèles. De la même manière, des modèles tenant compte de la structure secondaire et des propriétés fonctionnelles des protéines ont été proposés (Thorne *et al*., 1996 ; Goldman *et al*., 1998 ; Lio & Goldman, 1999). Une revue des améliorations successives apportées à ces différents modèles et de leur intérêt pour la compréhension des modalités d'évolution des protéines a été réalisée par Thorne (2000). Enfin, des modèles prenant en compte la structure en codons des gènes protéiques dans les analyses menées au niveau nucléotidique ont également été développés (Goldman & Yang, 1994 ; Muse & Gaut, 1994). Ces modèles codon-spécifiques, qui incorporent explicitement le code génétique par le biais des taux de substitutions synonymes et non-synonymes, ont été à l'origine de nouvelles méthodes permettant de détecter l'adaptation au niveau moléculaire en identifiant les sites évoluant sous un régime de sélection positive (Yang & Bielawski, 2000).

La modélisation de l'hétérogénéité des taux de substitution entre sites

Les modèles de substitution présentés font généralement l'hypothèse d'un taux de substitution unique et uniforme pour tous les sites. Cependant, l'existence de variations dans le taux de substitution entre les sites a été très tôt démontrée dès lors que les premières tentatives d'estimation du nombre de substitutions par site ont été conduites (Uzzell & Corbin, 1971). En effet, si le taux de substitution est uniforme au sein d'une séquence, le nombre de substitutions par site doit suivre une loi de Poisson. Testant cette hypothèse sur le cytochrome c, Fitch & Margoliash (1967) se sont aperçus que pour ajuster cette distribution au nombre minimum de substitutions inférées par site, il fallait exclure de l'analyse les sites qui apparaissent invariables. Ainsi, un modèle simple autorisant une fraction des sites à être invariables et les autres à évoluer à un taux uniforme s'ajuste souvent mieux aux données que le modèle uniforme (Adachi & Hasegawa, 1995). Cette fraction (I) de sites invariables peut-être esti-

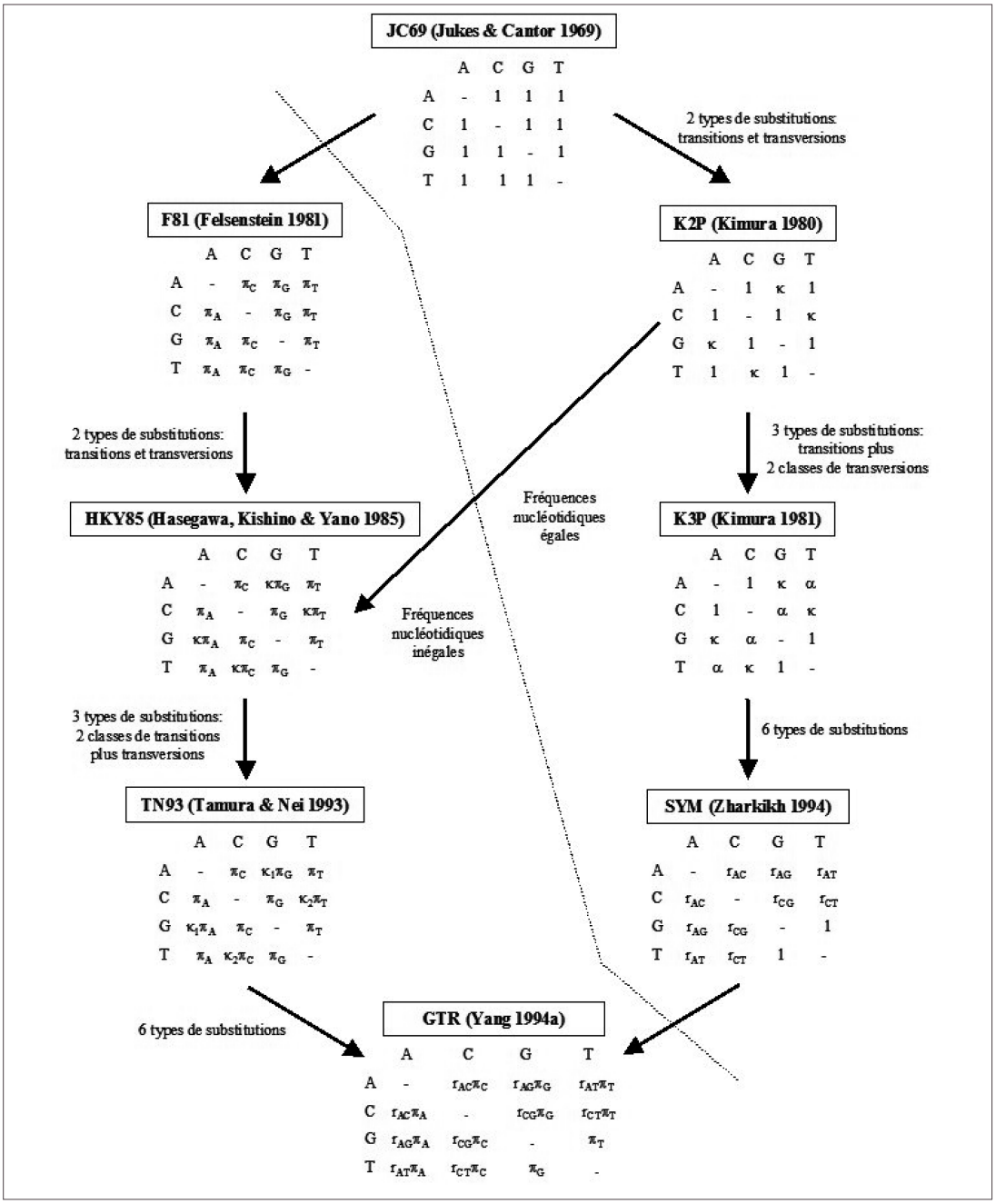

Figure 3. Présentation des principaux modèles d'évolution des séquences d'ADN réversibles en temps et de leurs filiations. Les flèches indiquent le passage d'un modèle simple vers un modèle plus complexe par la prise en compte des paramètres indiqués à chaque transition. La ligne en pointillés sépare les modèles à fréquences nucléotidiques égales et inégales. Modifié d'après Swofford *et al.* (1996) et Huelsenbeck & Bollback (2001).

mée par maximum de vraisemblance au cours des analyses phylogénétiques.

Des distributions continues ont également été utilisées pour modéliser le continuum de la variabilité des taux entre sites. La distribution gamma (Γ) a ainsi été proposée sur la base de ses propriétés mathématiques intéressantes (YANG, 1993). Cette loi statistique permet en effet de décrire une large gamme de distributions en changeant la valeur d'un paramètre (α) de forme (fig. 4). Ainsi, de faibles valeurs ($\alpha < 1$) définissent des distributions en forme de L proches de l'exponentielle négative et représentant une forte hétérogénéité d'un site à l'autre, alors que de fortes valeurs ($\alpha > 1$) décrivent des distributions en forme de cloche proches de la loi normale et caractérisant une faible hétérogénéité. Lorsque α tend vers l'infini, nous revenons à un modèle à taux uniforme. L'utilisation d'une distribution continue dans le modèle nécessite d'intégrer la fonction de vraisemblance sur l'ensemble des valeurs de la distribution, et l'équation (1) devient alors l'équation (4) où $f(r \mid \alpha)$ est la fonction de densité du taux r sous le modèle gamma et α le paramètre de forme de la distribution. Cependant, l'intégration de cette fonction continue requiert un énorme temps de calcul qui devient impraticable pour des jeux de données incluant plus de quelques espèces (YANG, 1993). Il a ainsi été proposé une approximation (Γ_K) de cette loi gamma continue en la discrétisant en K catégories de poids égal, et dont le taux moyen est utilisé pour représenter le taux de chaque catégorie (YANG, 1994b). L'équation précédente (4) devient alors une somme sur les K catégories discrètes : (5).

Cette discrétisation permet de réduire de manière drastique le temps de calcul associé à l'estimation du paramètre α, et est adéquate même en considérant seulement quatre catégories de taux (YANG, 1994b). L'incorporation de ce paramètre modélisant l'hétérogénéité des taux de substitutions entre sites dans les analyses phylogénétiques apporte un gain particulièrement significatif de vraisemblance et améliore très sensiblement les reconstructions (YANG, 1996a). Il est également possible de combiner les modèles gamma (Γ) et invariable (I) en autorisant une fraction des sites à être invariables et les taux des autres à être distribués selon une loi gamma (GU *et al.*, 1995).

La considération de partitions du jeu de données original a également été considérée comme un moyen de tenir compte de l'hétérogénéité du taux de substitution entre sites. Les modèles site-spécifiques (SSR pour « Site Specific Rate ») postulent que le taux de substitution est homogène au sein des différentes partitions mais autorisent des taux différents d'une partition à l'autre (SWOFFORD *et al.*, 1996). Ces modèles sont souvent utilisés pour décrire l'hétérogénéité au sein des séquences codantes dans lesquelles les partitions naturelles correspondent aux trois positions du codon. Cependant, il semble que ce modèle s'ajuste moins bien aux données que les modèles en loi gamma, vraisemblablement à cause du fait qu'il ignore l'hétérogénéité qui peut être importante au sein même de chacune des différentes partitions. Une approche intéressante de par son réalisme biologique est celle proposée par YANG (1996b) pour l'analyse combinée de gènes multiples. En effet, cette approche permet de gérer les variations de taux de substitution existant à la fois au sein des gènes analysés mais aussi entre ces différents gènes. Ceci est réalisé par l'attribution d'un modèle

$$f(x_i \mid \tau, \mathrm{v}, \theta, \alpha) = \int_0^{\infty} \left\{ \sum_{y}^{4^{s-1}} \left[\pi_{y_{2s-1}} \left(\prod_{k=1}^{s} p_{y_{i\sigma(k)}, x_{ik}}(\upsilon_k r, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{y_{i\sigma(k)}, y_{ik}}(\upsilon_k r, \theta) \right) \right] \right\} f(r \mid \alpha)\, dr \quad (4)$$

$$f(x_i \mid \tau, \mathrm{v}, \theta, \alpha) = \sum_{n=1}^{K} \left\{ \sum_{y}^{4^{s-1}} \left[\pi_{y_{2s-1}} \left(\prod_{k=1}^{s} p_{y_{i\sigma(k)}, x_{ik}}(\upsilon_k r_n, \theta) \right) \left(\prod_{k=s+1}^{2s-2} p_{y_{i\sigma(k)}, y_{ik}}(\upsilon_k r_n, \theta) \right) \right] \right\} \frac{1}{K} \quad (5)$$

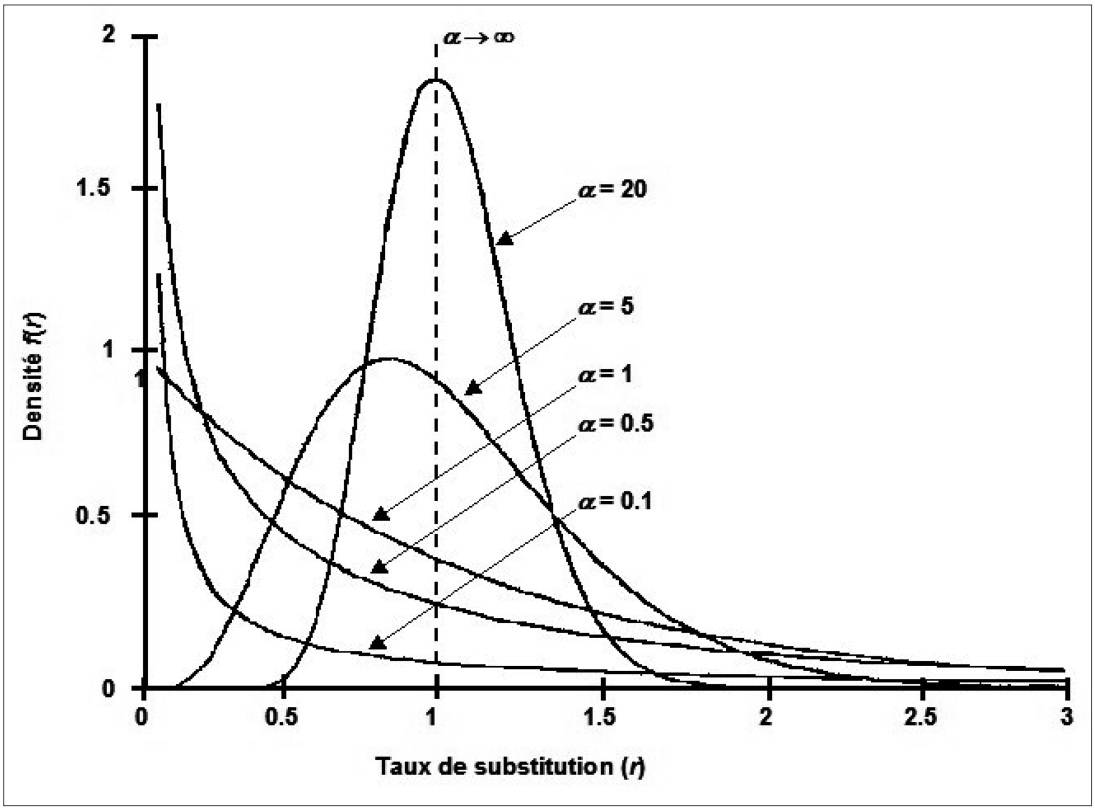

Figure 4. Fonction de densité f (r) de la distribution gamma (G) des taux de substitution par site (r) pour différentes valeurs du paramètre de forme (a). Adapté d'après Yang (1996a).

de substitution propre à chacune des partitions définies *a priori* lors de l'analyse phylogénétique d'une combinaison de gènes. Malgré le nombre de paramètres supplémentaires qu'il introduit, l'utilisation de ce modèle à partitions multiples conduit généralement à un gain significatif de vraisemblance par rapport à un modèle unique, montrant ainsi qu'il s'ajuste mieux aux données combinées (DELSUC *et al*., 2002 ; PUPKO *et al*., 2002).

Les modèles non-homogènes

L'hypothèse d'homogénéité des modèles de substitution peut-être relâchée de façon à ce que différents paramètres puissent varier au cours du temps, c'est-à-dire le long des branches de l'arbre. GALTIER & GOUY (1998) ont ainsi développé et implémenté en maximum de vraisemblance un modèle non-homogène et non-stationnaire permettant des variations de composition en bases le long des branches de l'arbre. Cette sophistication supplémentaire des modèles permet d'améliorer la fiabilité des reconstructions phylogénétiques lors de l'analyse de jeux de données présentant de fortes déviations par rapport à l'hypothèse de stationnarité. Cette méthode permet également d'inférer par maximum de vraisemblance les compositions en bases ancestrales à tous les nœuds de l'arbre. Son application à l'évolution des ARN ribosomiques a par exemple permis d'estimer la composition en bases ancestrale probable de ces gènes chez l'ancêtre de toutes les formes de vie actuelles (GALTIER *et al*., 1999).

Par ailleurs, l'hypothèse de la variabilité des taux de substitution au cours du temps a été très tôt évoquée pour expliquer la présence dans les séquences d'un nombre réduit de positions susceptibles de varier à chaque instant (Fitch & Markovitz, 1970). Ces observations ont donné naissance au modèle dit COVARIOTIDE/COVARION (pour « COncomitantly VARIable nucleOTIDEs/codONs ») de l'évolution moléculaire (Fitch, 1971) qui postule que bien que quelques sites essentiels au maintien de la fonction d'une molécule ne peuvent pas accepter de substitutions, la plupart des sites peuvent en fait passer d'un état libre de varier à un état fixé au cours du temps. Le corollaire de ce type de modèle est qu'il doit être possible d'observer au sein d'un alignement de séquences homologues, des sites qui sont fixés chez certains groupes taxonomiques distants alors qu'ils sont variables chez d'autres. La différence de localisation des sites invariables entre groupes taxonomiques différents et son importance dans les reconstructions phylogénétiques ont ainsi été soulignées (Lockhart *et al*., 1996), et des tests ont été développés pour identifier la présence d'une structure en covarions (Lockhart *et al*., 1998 ; Lopez *et al*., 1999). Ces derniers ont permis de montrer l'importance du phénomène en identifiant plusieurs cas où le modèle covarion explique mieux l'évolution des séquences qu'un modèle où le taux de substitution des sites est constant au cours de l'histoire évolutive des molécules (Lockhart *et al*., 2000).

La première formalisation du modèle de type covarion originel proposé par Fitch (1971) est due à Tuffley & Steel (1998). Leur modèle introduit un paramètre unique représentant le taux de passage des sites d'un état variable (« on ») à un état invariable (« off ») au cours du temps. Récemment, Galtier (2001) a proposé une généralisation de ce modèle en autorisant les sites à passer avec un certain taux d'une catégorie à l'autre d'une loi gamma décrivant les taux des différentes catégories (modèle SSRV pour « Site Specific Rate Variation »). Un second modèle (modèle USSRV pour « Unequal Site Specific Rate Variation »), généralisant le modèle précédent, autorise une certaine proportion des sites à évoluer selon le modèle SSRV alors que les sites restants évoluent selon un modèle classique en loi gamma (ASRV pour « Among Site Rate Variation »). L'implémentation en maximum de vraisemblance et l'application de ces modèles à des jeux de données empiriques montrent que la considération d'une structure en covarions augmente de manière significative la vraisemblance des données par rapport aux modèles homogènes classiques (Galtier, 2001). Ainsi, la variation au cours du temps du taux de substitution d'un site (hétérotachie) apparaît comme un phénomène majeur de l'évolution moléculaire (Lopez *et al*., 2002). Sa prise en compte explicite dans les modèles d'évolution des séquences apparaît comme un objectif majeur des développements méthodologiques futurs (Penny *et al*., 2001).

La sélection du modèle adéquat

Une critique souvent adressée à l'encontre des méthodes probabilistes en phylogénie moléculaire est leur dépendance vis-à-vis du modèle d'évolution sous-jacent. Ainsi, bien que ces méthodes soient considérées comme robustes vis-à-vis du non respect de leurs hypothèses de base (Gaut & Lewis, 1995 ; Huelsenbeck, 1995 ; Swofford *et al*., 2001), il est important d'employer des modèles qui s'ajustent le mieux possible aux données analysées. Comme nous l'avons vu précédemment, un nombre de modèles de complexité croissante et incorporant des paramètres de plus en plus nombreux sont à disposition du systématicien moléculaire. Un choix intuitif serait de considérer le modèle le plus riche en paramètres, ceci afin de disposer d'une estimation de la valeur individuelle de chacun d'entre eux. Cependant, cette stratégie a deux inconvénients : un nombre élevé de paramètres à estimer nécessite un temps de calcul conséquent, et la considération simultanée d'un nombre élevé de paramètres augmente la variance associée à l'estimation de chaque paramètre. Il est donc indispensable de disposer d'un critère objectif de choix entre les différents modèles existants (Posada & Crandall, 2001). Des tests statistiques probabilistes comme celui du rapport des vraisemblances (LRT pour « Likelihood Ratio Test » ; Huelsenbeck & Crandall, 1997)

et le critère d'information d'Akaike (AIC pour « Akaike Information Criterion » ; Akaike, 1974) ont ainsi été proposés pour comparer les différents modèles. POSADA & CRANDALL (1998) ont développé le programme MODELTEST qui permet de choisir le modèle d'évolution qui s'ajuste le mieux aux données parmi toute une gamme de modèles sur la base de LRT hiérarchiques et de comparaisons d'AIC. MININ *et al*. (2003) ont récemment proposé une procédure de sélection du modèle basée sur la performance qui est implémentée dans le programme DT-ModSel.

LA MÉTHODE DU MAXIMUM DE VRAISEMBLANCE

Comme nous l'avons vu précédemment, la fonction de vraisemblance dépend de nombreux paramètres : la topologie de l'arbre reliant les taxons, les longueurs de branches de cet arbre, et les paramètres du modèle d'évolution des séquences. Ces paramètres doivent être estimés car ils sont propres au jeu de données considéré et par conséquent inconnus avant l'analyse. La méthode du maximum de vraisemblance consiste à estimer les valeurs des paramètres qui maximisent la vraisemblance, c'est-à-dire celles qui confèrent la plus forte probabilité d'avoir conduit aux données observées. Cependant, la surface de la fonction de vraisemblance pour le problème phylogénétique est si complexe – c'est un espace de dimensions correspondant aux nombre de paramètres incorporés – qu'il est impossible de localiser le(s) pic(s) de maximum de vraisemblance de manière analytique. Ce problème est ainsi traité numériquement en le décomposant en deux étapes : (1) estimer les paramètres qui maximisent la vraisemblance d'une topologie donnée et (2) maximiser ensuite cette vraisemblance sur l'ensemble des topologies. L'arbre qui a la vraisemblance maximale est l'arbre de maximum de vraisemblance. Une limitation pratique de l'approche par maximum de vraisemblance est qu'il est très difficile de maximiser la vraisemblance de manière simultanée pour de nombreux paramètres. Cela se traduit par un temps de calcul rédhibitoire lorsque des modèles d'évolution complexes sont utilisés pour des jeux de données conséquents (> 20 taxons). Une possibilité est alors d'utiliser une stratégie d'estimation en boucle qui est répétée jusqu'à stabilisation de la topologie et des paramètres du modèle (SWOFFORD *et al*., 1996). Cependant, de très récents progrès en algorithmique ont abouti à l'obtention d'une méthode de maximum de vraisemblance extrêmement rapide qui permet l'analyse de jeux de données beaucoup plus conséquents, et qui est implémentée dans le logiciel PHYML (GUINDON & GASCUEL, 2003). L'application à la reconstruction phylogénétique d'algorithmes empruntés à la génétique des populations permet aussi de réduire remarquablement la durée de ces analyses (LEWIS 1998 ; BRAUER *et al*., 2002 ; LEMMON & MILINKOVITCH 2002).

LES TESTS STATISTIQUES D'HYPOTHÈSES ALTERNATIVES

L'une des forces des approches probabilistes en phylogénie repose sur le fait qu'elles permettent, sur la base de leur vraisemblance, la comparaison d'hypothèses évolutives alternatives dans un cadre statistique clairement défini. Les tests statistiques développés à cet effet sont de deux types : paramétriques et non paramétriques. KISHINO & HASEGAWA (1989) ont été les premiers à proposer un test basé sur la différence de vraisemblance δ entre deux topologies en compétition pour un même jeu de données. Leur célèbre test (test KH) peut-être vu comme un LRT non-paramétrique pour comparer deux topologies définies *a priori* (c'est-à-dire sur des critères extérieurs au jeu de données considéré). Ce test calcule de façon analytique la variance (σ^2) de δ et compare le rapport δ/σ à une loi normale afin de déterminer si cette différence de vraisemblance est statistiquement significative (KISHINO & HASEGAWA, 1989). L'utilité de ce test en a fait l'un des tests statistiques les plus utilisés en phylogénie au cours des quinze dernières années. Cependant, malgré le fait que ce test ait été initialement développé pour comparer deux topologies définies *a priori*, il a souvent été utilisé de façon inadéquate dans le cas de comparaisons incorporant des topologies définies *a posteriori*

(c'est-à-dire après analyse du jeu de données ; GOLDMAN *et al*., 2000). Pour pallier ce manque de robustesse du test KH par rapport au non respect de ses hypothèses de base, SHIMODAIRA & HASEGAWA (1999) en ont proposé une version modifiée (test SH) permettant également la comparaison de multiples topologies de façon simultanée. En corrigeant pour les comparaisons multiples, le test SH est plus conservatif que le test KH.

GOLDMAN *et al*. (2000) ont proposé une revue des différents tests de topologies disponibles qui en précise clairement leurs hypothèses de base et leurs conditions d'application. Ces auteurs ont également décrit un test paramétrique, brièvement évoqué par SWOFFORD *et al*. (1996), et applicable à la comparaison de topologies définies *a posteriori* (test SOWH par référence aux initiales des auteurs de l'idée originale). En pratique, le test SOWH est d'utilisation et d'interprétation faciles mais nécessite un temps de calcul important dans sa version exacte (sans approximation). D'autre part, de très larges différences ont été observées entre les résultats des tests SOWH et SH, le test SOWH apparaissant beaucoup plus libéral que le test SH. L'origine des très larges différences observées entre les résultats de ces deux tests n'ont pas encore été clairement identifiées (GOLDMAN *et al*., 2000). Cependant, un début de réponse a été apportée par STRIMMER & RAMBAUT (2002) qui suggèrent que le test paramétrique SOWH serait très sensible aux hypothèses du modèle d'évolution utilisé. Ces auteurs proposent également un test simple basé sur les vraisemblances pondérées qui semble être moins conservatif que le test SH. Ces résultats ont également été corroborés par une étude empirique menée par BUCKLEY (2002) qui montre une tendance du test SOWH à être sujet aux erreurs de type I (rejet de l'hypothèse nulle alors qu'elle est vraie), vraisemblablement due à sa sensibilité par rapport à l'adéquation au modèle d'évolution utilisé. Enfin, SHIMODAIRA (2002) a proposé un nouveau test (AU pour « Approximately Unbiased ») qui apparaît également moins conservatif que le test SH. Le programme CONSEL (SHIMODAIRA & HASEGAWA, 2001) implémente ces différents tests non paramétriques.

UN EXEMPLE DE RECONSTRUCTION PHYLOGÉNÉTIQUE PAR MAXIMUM DE VRAISEMBLANCE

Afin d'illustrer la méthode du maximum de vraisemblance, un jeu de données d'ARNr 12S mitochondrial pour 13 espèces de mammifères xénarthres (tatous, fourmiliers et paresseux) a été analysé (DELSUC *et al*., 2003). La sensibilité de la reconstruction phylogénétique à l'inadéquation potentielle du modèle d'évolution sous-jacent a été évaluée (fig. 5A). Une recherche heuristique utilisant l'algorithme TBR (pour « Tree Bisection and Reconnection ») de réarrangement des branches avec estimation simultanée des paramètres du modèle a été menée par le logiciel PAUP* (Swofford, 2002) sous les 56 modèles d'évolution différents définis par POSADA & CRANDALL (2001). Le modèle sélectionné sur la base du critère d'information d'Akaike (AIC) par le programme MODELTEST est le modèle GTR + Γ_8 qui identifie la topologie T1 comme la topologie de maximum de vraisemblance. Cependant, deux topologies alternatives (T2 et T3) apparaissent également selon le modèle utilisé (fig. 5B), mais les variations de logarithme de vraisemblance qu'elles impliquent ne sont pas statistiquement significatives. La topologie T1 est sélectionnée à la fois par les modèles les plus simples et les plus complexes alors que la topologie T2 n'apparaît que pour des modèles n'incorporant pas d'hétérogénéité des taux de substitution entre sites. La considération d'une fraction de sites invariables dans le modèle n'améliore que très peu la vraisemblance et ne change pas l'identité de la topologie sélectionnée. Ce genre d'analyse de robustesse de la méthode face aux hypothèses du modèle permet d'explorer le comportement des données afin d'identifier les biais potentiellement présents. La relative instabilité topologique révélée ici par cette analyse reflète ainsi le faible signal phylogénétique présent dans ce gène pour reconstruire la phylogénie des tatous (fig. 5C).

CONCLUSION

Grâce à l'utilisation de la fonction de vraisemblance, les méthodes probabilistes sont les seules méthodes

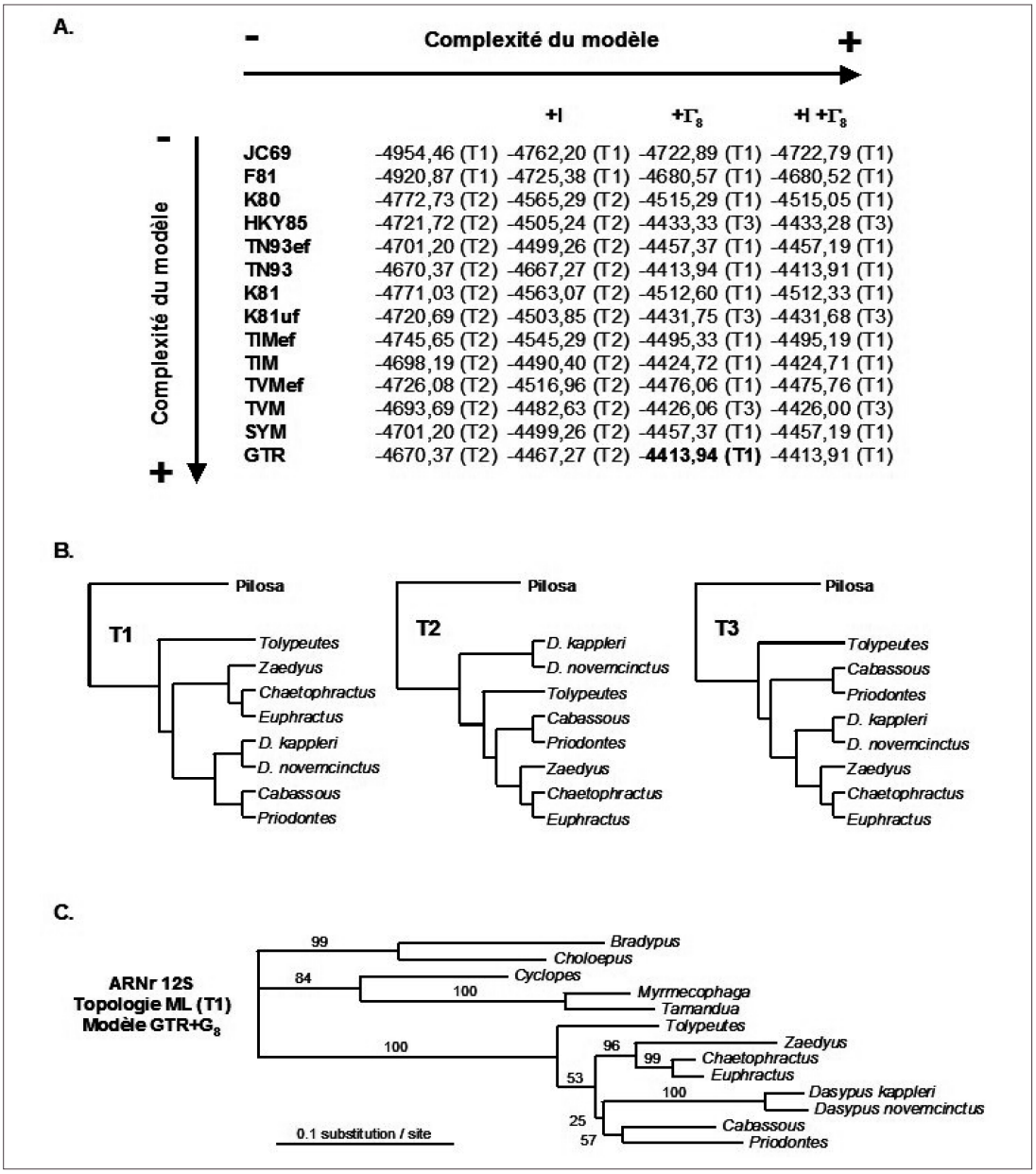

Figure 5. Exemple d'analyse phylogénétique par maximum de vraisemblance de l'ARNr 12S mitochondrial pour 13 espèces de Xénarthres. A. Analyse de sensibilité au modèle d'évolution des séquences, montrant les valeurs de vraisemblance obtenues sous les 56 modèles définis par Posada & Crandall (2001). B. Trois topologies alternatives apparues au cours de l'analyse de sensibilité et qui sont optimales pour des modèles différents. C. Phylogramme de maximum de vraisemblance obtenu sous le modèle optimal GTR + Γ_8 sélectionné pour ce jeu de données par le programme MODELTEST (Posada & Crandall, 1998). Les valeurs aux nœuds indiquent les pourcentages de bootstrap obtenus après 100 réplications. Données originales d'après Delsuc *et al.* (2003).

capables de coupler l'inférence des modalités évolutives des séquences moléculaires avec la reconstruction de la phylogénie. La capacité de ces méthodes à incorporer les progrès réalisés au cours des 30 dernières années dans la modélisation de l'évolution des séquences permet d'obtenir des inférences statistiques de plus en plus fiables (Whelan *et al*., 2001). Ainsi la méthode du maximum de vraisemblance est considérée comme une méthode de reconstruction phylogénétique particulièrement robuste par rapport au non respect de ses hypothèses de base (Swofford *et al*., 2001). Cette approche présente également l'avantage de se placer dans un cadre statistique bien défini qui permet – par une batterie de tests développés à cet égard – l'évaluation d'hypothèses alternatives quant aux modalités d'évolution des séquences et aux phylogénies sous-jacentes (Huelsenbeck & Crandall, 1997). Il n'en demeure pas moins que les modèles d'évolution implémentés actuellement sont loin de décrire toute la complexité du processus d'évolution des séquences, ce qui peut rendre ces méthodes inconsistantes dans les cas où leurs hypothèses de base sont fortement transgressées. Ainsi, si des avancées majeures restent à effectuer sur ce plan là, l'approche probabiliste fournit un cadre idéal pour leur développement.

Remerciements. — Cet article est fondé sur une partie de la thèse de l'Université Montpellier II dirigée par EJPD, soutenue par FD en décembre 2002, et intitulée : « Phylogénie moléculaire des Xénarthres (tatous, fourmiliers et paresseux) : Application des méthodes probabilistes à la reconstruction de leur histoire évolutive au sein des Mammifères placentaires ». Les auteurs tiennent à remercier Nicolas Galtier et Hervé Philippe pour leurs commentaires éclairés en tant que membres du jury de cette thèse, ainsi qu'Alice Cibois et Jean-François Silvain pour leurs remarques pertinentes sur le manuscrit. Cet article représente la contribution ISEM 2004-028 de l'Institut des Sciences de l'Évolution de Montpellier (UMR 5554/CNRS).

RÉFÉRENCES BIBLIOGRAPHIQUES

Adachi J. & Hasegawa M., 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree : heterogeneity among amino acid sites. *J. Mol. Evol*., 40 : 622-628.

Akaike H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.*, AC-19 : 716-723.

Brauer M.J., Holder M.T., Dries L.A., Zwickl D.J., Lewis, P.O. & Hillis D.M., 2002. Genetic algorithms and parallel processing in maximum-likelihood phylogeny inference. *Mol. Biol. Evol*., 19 : 1717-1726.

Buckley T.R., 2002. Model misspecification and probabilistic tests of topology : evidence from empirical data sets. *Syst. Biol*., 51 : 509-523.

Delsuc F., Scally M., Madsen O., Stanhope M.J., De Jong W.W., Catzeflis F.M., Springer M.S. & Douzery E.J.P., 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. *Mol. Biol. Evol*., 19 : 1656-1671.

Delsuc F., Stanhope M.J. & Douzery E.J.P., 2003. Molecular systematics of armadillos (Xenarthra ; Dasypodidae) : contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes. *Mol. Phylogenet. Evol*., 28 : 261-275.

Edwards A.W.F. & Cavalli-Sforza L.L., 1967. Phylogenetic analysis : models and estimation procedures. *Evolution*, 21 : 550-570.

Felsenstein J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool*., 22 : 240-249.

Felsenstein J., 1981. Evolutionary tree from DNA sequences : a maximum likelihood approach. *J. Mol. Evol*., 17 : 368-376.

Fitch W.M., 1971. Rate of change of concomitantly variable codons. *J. Mol. Evol*., 1 : 84-96.

Fitch W.M. & Margoliash E., 1967. A method for estimating the number of invariant amino acid

coding positions in a gene using cytochrome c as a model case. *Biochem. Genet*., 1 : 65-71.

FITCH W.M. & MARKOWITZ E., 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet*., 4 : 579-593.

GALTIER N. & GOUY M., 1998. Inferring pattern and process : maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol*., 15 : 871-879.

GALTIER N., TOURASSE N. & GOUY M., 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science*, 283 : 220-221.

GALTIER N., 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol*., 18 : 866-873.

GAUT B.S. & LEWIS P.O., 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol*., 12 : 152-162.

GOLDMAN N. & YANG Z., 1994. A codon based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol*., 11 : 725-736.

GOLDMAN N., THORNE J.L. & JONES D.T., 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149 : 445-458.

GOLDMAN N., ANDERSON J.P. & RODRIGO A.G., 2000. Statistical tests of topologies in phylogenetics. *Syst. Biol*. 49 : 652-670.

GU X., FU Y.-X. & LI W.-H., 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol*., 12 : 546-557.

GUINDON S. & GASCUEL O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*., 52 : 696-704.

HUELSENBECK J.P., 1995. The robustness of two phylogenetic methods : four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.*, 12 : 843-849.

HUELSENBECK J.P. & CRANDALL K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst*., 28 : 437-466.

HUELSENBECK J.P. & BOLLBACK J.P., 2001. Application of the likelihood function in phylogenetic analysis. In : D.J. Balding, M. Bishop & C. Cannings (eds.), *Handbook of Statistical Genetics*, pp. 415-439, John Wiley and Sons, Inc., New York.

JUKES T.H. & CANTOR C.R., 1969. The evolution of protein molecules. In : H.M. Munro (ed.), *Mammalian protein metabolism*, pp. 21-132, Academic Press, New York.

KISHINO H. & HASEGAWA M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol*., 29 : 170-179.

LEMMON A.R. & MILINKOVITCH M.C., 2002. The metapopulation genetic algorithm : An efficient solution for the problem of large phylogeny estimation. *Proc. Natl. Acad. Sci. USA*, 99 : 10516-10521.

LEWIS P.O., 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol*., 15 : 277-283.

LIO P. & GOLDMAN N., 1999. Using protein structural information in evolutionary inference : transmembrane proteins. *Mol. Biol. Evol*., 16 : 1696-1710.

LOCKHART P.J., LARKUM A.W., STEEL M., WADDELL P.J. & PENNY D., 1996. Evolution of chlorophyll and bacteriochlorophyll : the problem of invariant sites in sequence analysis. *Proc. Natl. Acad. Sci. USA*, 93 : 1930-1934.

LOCKHART P.J., STEEL M.A., BARBROOK A.C., HUSON D.H., CHARLESTON M.A. and HOWE C.J., 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol*., 15 : 1183-1188.

LOCKHART P.J., HUSON D., MAIER U., FRAUNHOLZ M.J., VAN DE PEER Y., BARBROOK A.C., HOWE C.J. & Steel M.A., 2000. How molecules evolve in eubacteria. *Mol. Biol. Evol*., 17 : 835-838.

LOPEZ P., FORTERRE P. & PHILIPPE H., 1999. The root of the tree of life in the light of the covarion model. *J. Mol. Evol.*, 49 : 496-508.

LOPEZ P., CASANE D. & PHILIPPE H., 2002. Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.*, 19 : 1-7.

MININ V., ABDO Z., JOYCE P. & SULLIVAN J., 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.*, 52 : 674-683.

MUSE S.V. & GAUT B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates with application to the chloroplast genome. *Mol. Biol. Evol.*, 11 : 715-724.

MUSE S.V., 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139 : 1429-1439.

PENNY D., MCCOMISH B.J., CHARLESTON M.A. & HENDY M.D., 2001. Mathematical elegance with biochemical realism : the covarion model of molecular evolution. *J. Mol. Evol.*, 53 : 711-723.

POSADA D. & CRANDALL K.A., 1998. MODELTEST : testing the model of DNA substitution. *Bioinformatics*, 14 : 817-818.

POSADA D. & CRANDALL K.A., 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50 : 580-601.

PUPKO T., HUCHON D., CAO Y., OKADA N. & HASEGAWA M., 2002. Combining multiple data sets in a likelihood analysis : which models are the best ? *Mol. Biol. Evol.* 19 : 2294-2307.

SAVILL N.J., HOYLE D.C. & HIGGS P.G., 2001. RNA sequence evolution with secondary structure constraints : comparison of substitution rate models using maximum-likelihood methods. *Genetics*, 157 : 399-411.

SCHÖNIGER M. & VON HAESELER A., 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.*, 3 : 240-247.

SHIMODAIRA H. & HASEGAWA M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, 16 : 1114-1116.

SHIMODAIRA H. & HASEGAWA M., 2001. CONSEL : for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17 : 1246-1247.

SHIMODAIRA H., 2002. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.*, 51 : 492-508.

STRIMMER K. & RAMBAUT A., 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. Lond. BS*, 269 : 137-142.

SWOFFORD D.L., OLSEN G.J., WADDELL P.J. & HILLIS D.M., 1996. Phylogenetic inference. In : D.M. HILLIS, C. MORITZ & B.K. MABLE (eds.), *Molecular systematics*, p. 407-514, Sinauer Associates, Sunderland, Massachusetts.

SWOFFORD D.L., WADDELL P.J., HUELSENBECK J.P., FOSTER P.G., LEWIS P.O. & ROGERS J.S., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, 50 : 525-539.

SWOFFORD D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Sinauer Associates, Sunderland, Massachusetts.

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

THORNE J.L., 2000. Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.*, 10 : 602-605.

TILLIER E.R.M. & COLLINS R.A., 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, 148 : 1993-2002.

TUFFLEY C. & STEEL M., 1998. Modelling the covarion hypothesis of nucleotide substitution. *Math. Biosci.*, 147 : 63-91.

UZZELL T. & CORBIN K.W., 1971. Fitting discrete probability distributions to evolutionary events. *Science*, 172 : 1089-1096.

WHELAN S., LIO P. & GOLDMAN N., 2001. Molecular phylogenetics : state-of-the-art methods for looking into the past. *Trends Genet.*, 17 : 262-272.

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

YANG Z., 1994a. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39 : 105-111.

YANG Z., 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites : approximate methods. *J. Mol. Evol.*, 39 : 306-314.

YANG Z., 1996a. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.*, 11 : 367-372.

YANG Z., 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.*, 42 : 587-596.

YANG Z. & BIELAWSKI J.P., 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.*, 15 : 496-503.