PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 3903–3911
doi:10.1098/rstb.2008.0177
Published online 7 October 2008

A likelihood framework to analyse phyletic patterns

Ofir Cohen[1], Nimrod D. Rubinstein[1], Adi Stern[1], Uri Gophna[2] and Tal Pupko[1,*]

[1]*Department of Cell Research and Immunology, and* [2]*Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel*

Probabilistic evolutionary models revolutionized our capability to extract biological insights from sequence data. While these models accurately describe the stochastic processes of site-specific substitutions, single-base substitutions represent only a fraction of all the events that shape genomes. Specifically, in microbes, events in which entire genes are gained (e.g. via horizontal gene transfer) and lost play a pivotal evolutionary role. In this research, we present a novel likelihood-based evolutionary model for gene gains and losses, and use it to analyse genome-wide patterns of the presence and absence of gene families. The model assumes a Markovian stochastic process, where gains and losses are represented by the transition between presence and absence, respectively, given an underlying phylogenetic tree. To account for differences in the rates of gain and loss of different gene families, we assume among-gene family rate variability, thus allowing for more accurate description of the data. Using the Bayesian approach, we estimated an evolutionary rate for each gene family. Simulation studies demonstrated that our methodology accurately infers these rates. Our methodology was applied to analyse a large corpus of data, consisting of 4873 gene families spanning 63 species and revealed novel insights regarding the evolutionary nature of genome-wide gain and loss dynamics.

Keywords: phyletic pattern; probabilistic evolutionary models; genome evolution; gene gain and loss; horizontal gene transfer; gene content

1. INTRODUCTION

Sequence evolution at the nucleotide level is now commonly described using probabilistic models of sequence evolution, which rely on continuous-time Markov process theory (Felsenstein 1981; Yang 1995). These techniques have revolutionized our understanding of the evolutionary dynamics acting at each position of a single coding gene. Nevertheless, single base-pair substitutions are but a fraction of all events that constitute the evolution of a genome, since gene content itself varies substantially across species (Mira *et al.* 2002). Variation in gene content is the result of either gene loss or gene gain via, for example, horizontal gene transfer (HGT) events (Lawrence & Roth 1996; Ochman *et al.* 2000; Koonin *et al.* 2001). These macroevolutionary events of gene gains and losses are collectively termed gene gain and loss (GGL) events.

In the last decade, a large volume of data from the ever-increasing number of sequenced genomes was analysed. It was shown that HGT events are common in all three domains of life (Doolittle *et al.* 1990; Nelson *et al.* 1999; Koonin *et al.* 2001; Gogarten & Townsend 2005). Analysing these events contributed to more accurate reconstruction of evolutionary histories and provided new insights on evolutionary processes such as the speciation of lineages, the adaptation of organisms to new ecological niches and the evolution of novel functions and the pruning of existing ones (Syvanen 1994; Jain *et al.* 2003; Lake & Rivera 2004; Pennisi 2004; Gogarten & Townsend 2005; Wapinski *et al.* 2007).

For analysing GGL events, one must have a representation scheme of the presence and absence of genes across species. These data are organized in a phyletic pattern, in which the character '1' represents the presence of at least one homologue from a certain gene family in a given genome, and the character '0' represents its absence. In this representation, the aligned phyletic pattern from all gene families is analogous to a multiple sequence alignment. In the phyletic pattern alignment, each column represents a specific gene family, and thus this alignment encapsulates the fundamental information about the content of all gene families among all the analysed genomes. Notably, using this representation, only the evolutionary dynamics of entire gene families are accounted for. The $1 \rightarrow 0$ transition indicates a deletion event of all the members of a specific gene family, and the $0 \rightarrow 1$ transition covers both the gene family de novo appearance ('birth') and HGT endowing the acceptor with the first copy of the gene family. These definitions imply that events such as gene duplication, deletion of one paralogue or even HGTs to a genome in which a representative of a gene family already exists are ignored. This increases the reliability of the data, since modelling gene family content evolution (rather than gene-content evolution) alleviates the need to reliably detect orthologous sequences. Furthermore, the gain or loss of an entire gene family is likely to represent major functional evolutionary shifts.

* Author for correspondence (talp@post.tau.ac.il).

One contribution of 17 to a Discussion Meeting Issue 'Statistical and computational challenges in molecular phylogenetics and evolution'.

A reliable inference of gains and losses of gene families necessitates accurate modelling, which describes the dynamics of these events. Initial approaches towards this goal were based on the Dollo parsimony criterion (Mirkin *et al.* 2003). More advanced models were based on the maximum-likelihood (ML) paradigm (Hao & Golding 2006); however, they assumed that the gain rate is equal to the loss rate, thus ensuring that the genome sizes do not change. Nevertheless, as stated by these authors, this assumption is unlikely to be correct, given evidence that even within closely related species genome content may change considerably (Thompson *et al.* 2005). Moreover, these models assume that the same process of gene gains and losses is shared among all gene families, i.e. the gain and loss rate parameters are equal across gene families. However, a large body of evidence suggests that there is in fact a tremendous variance in the tendency of various gene families to undergo GGLs (Jain *et al.* 1999; Lerat *et al.* 2005).

In this paper, our goal was to progress towards more realistic models for analysing phyletic data, by developing a better methodology that captures the macro-evolutionary dynamics of GGLs. Our models rely on an evolutionary methodology developed for sequence data and adapt and extend it to the task of analysing phyletic patterns.

2. MATERIAL AND METHODS

(a) *Evolutionary models for GGLs*

The analysis of phyletic patterns using evolutionary models requires three components: the data; the phylogenetic tree; and the underlying stochastic model of transitions. (i) The data are represented as an $S\times F$ matrix D, in which $D_{sf}=1$ if the gene family $f\in\{1,\ldots,F\}$ is present in species $s\in\{1,\ldots,S\}$, and zero otherwise (S is the number of species analysed and F is the number of gene families in the data). (ii) The phylogenetic tree (topology and branch lengths) is assumed to be known (see below for details). (iii) In accordance with the sequence evolution paradigm, we assume that the GGL dynamics follow a continuous Markov process over a two-state alphabet $\{0,1\}$.

In our models, the transitions between zeros and ones and vice versa are defined by a 2×2 rate matrix Q and initial probabilities π_{root}. More specifically, Q is given in the following form:

$$Q=\left(\begin{array}{c|cc} & 0 & 1\\ \hline 0 & -g & g\\ 1 & l & -l\end{array}\right), \tag{2.1}$$

where $g\equiv Q_{0\to1}$ denotes the $0\to1$ instantaneous rate parameter and $l\equiv Q_{1\to0}$ denotes the $1\to0$ rate parameter. As in other evolutionary models, the transition probability $P_{ij}(t)$ denotes the probability that character i will be replaced by character j along a branch of length t and is computed by $P(t)=\mathrm{e}^{Qt}$. Notably, $P_{ij}(t)$ in this case can be calculated analytically (Ross 1996), which enables shorter running times since the costly numerical evaluation of matrix exponentials is saved. This calculation is given by

$$\left.\begin{aligned} P_{01}(t)&=g/(g+l)-\mathrm{e}^{-(g+l)\cdot t}\cdot(1-l/(g+l)),\\ P_{10}(t)&=l/(g+l)-\mathrm{e}^{-(g+l)\cdot t}\cdot(1-g/(g+l)),\\ P_{11}(t)&=g/(g+l)+\mathrm{e}^{-(g+l)\cdot t}\cdot(1-g/(g+l))\\ &\text{and}\\ P_{00}(t)&=l/(g+l)+\mathrm{e}^{-(g+l)\cdot t}\cdot(1-l/(g+l)).\end{aligned}\right\} \tag{2.2}$$

We have developed several models, which can be defined by two separate characteristics—the stochastic process and the rate distribution among gene families. We first consider a simple model in which the gain and loss rates are equal and the character frequencies at the root (π_{root}) are the same as the stationary frequencies (π_Q). In this model, $\pi_Q(0)=\pi_Q(1)=0.5$ and the only free parameter is the rate coefficient of the Q matrix (r_Q),

$$Q=r_Q\cdot\begin{pmatrix}-0.5 & 0.5\\ 0.5 & -0.5\end{pmatrix}. \tag{2.3}$$

The parameter r_Q is introduced to transform branch lengths from substitution per amino acid site to GGL events per gene family. This allows one to use branch lengths estimated from sequence data to analyse the GGL dynamics. It is thus implicitly assumed that the proportions among branches are the same for amino acid substitutions and GGL events (see §4).

The second model is similar, but allows separate rates for gain and loss, which necessitates an additional free parameter $\pi_Q(0)$. In this model, $\pi_Q(0)$ and $\pi_Q(1)$ are the stationary character frequencies and equal to the root character frequencies, π_{root},

$$Q=r_Q\cdot\begin{pmatrix}-\pi_Q(1) & \pi_Q(1)\\ \pi_Q(0) & -\pi_Q(0)\end{pmatrix}. \tag{2.4}$$

The third model further allows for the character frequencies at the root to differ from the stationary frequencies, thus adding an additional free parameter, $\pi_{\mathrm{root}}(0)$.

(b) *Introducing among-gene family rate variation*

We implemented two variations for the rate distribution among gene families. The first, termed the homogeneous rate model, assumes that the Q matrix is equal for all gene families. Since some gene families undergo virtually no gains and losses (e.g. ribosomal proteins), while others are characterized by rampant gains and losses, clearly this homogeneous assumption is not adequate (e.g. Jain *et al.* 1999). To capture this phenomenon, we introduced a rate parameter r_f for the gene family f. Under this assumption, the rate matrix of the gene family f equals

$$Q=r_f\cdot r_Q\cdot\begin{pmatrix}-\pi_Q(1) & \pi_Q(1)\\ \pi_Q(0) & -\pi_Q(0)\end{pmatrix}. \tag{2.5}$$

We further assume that for each gene family r_f is sampled from an *a priori* rate distribution, the gamma distribution. This is equivalent to the classical among-site rate variation model, first introduced by Yang (1993). We make the same restriction that r_f is sampled from a distribution with mean 1. Thus, only one additional free parameter for this model is added—the shape parameter of the gamma distribution (α). We term this model the among-gene family rate variation (AGRV) model for the phyletic pattern.

There are two widely used methods for estimating the evolutionary rate at each site f. The first, based on the ML paradigm, searches for the rate that maximizes the likelihood given that rate (Yang & Nielsen 2000; Pupko *et al.* 2002),

$$\hat{r}_f=\arg\max_{r_f}P(D_f|T,t,Q), \tag{2.6}$$

where T and t are the phylogenetic tree topology and the set of branch lengths, respectively. We note that Q is a function of r_f as defined in equation (2.5). The lower the rate, the more conserved the gene family is, i.e. it has experienced fewer gains and losses compared with the average over all gene families.

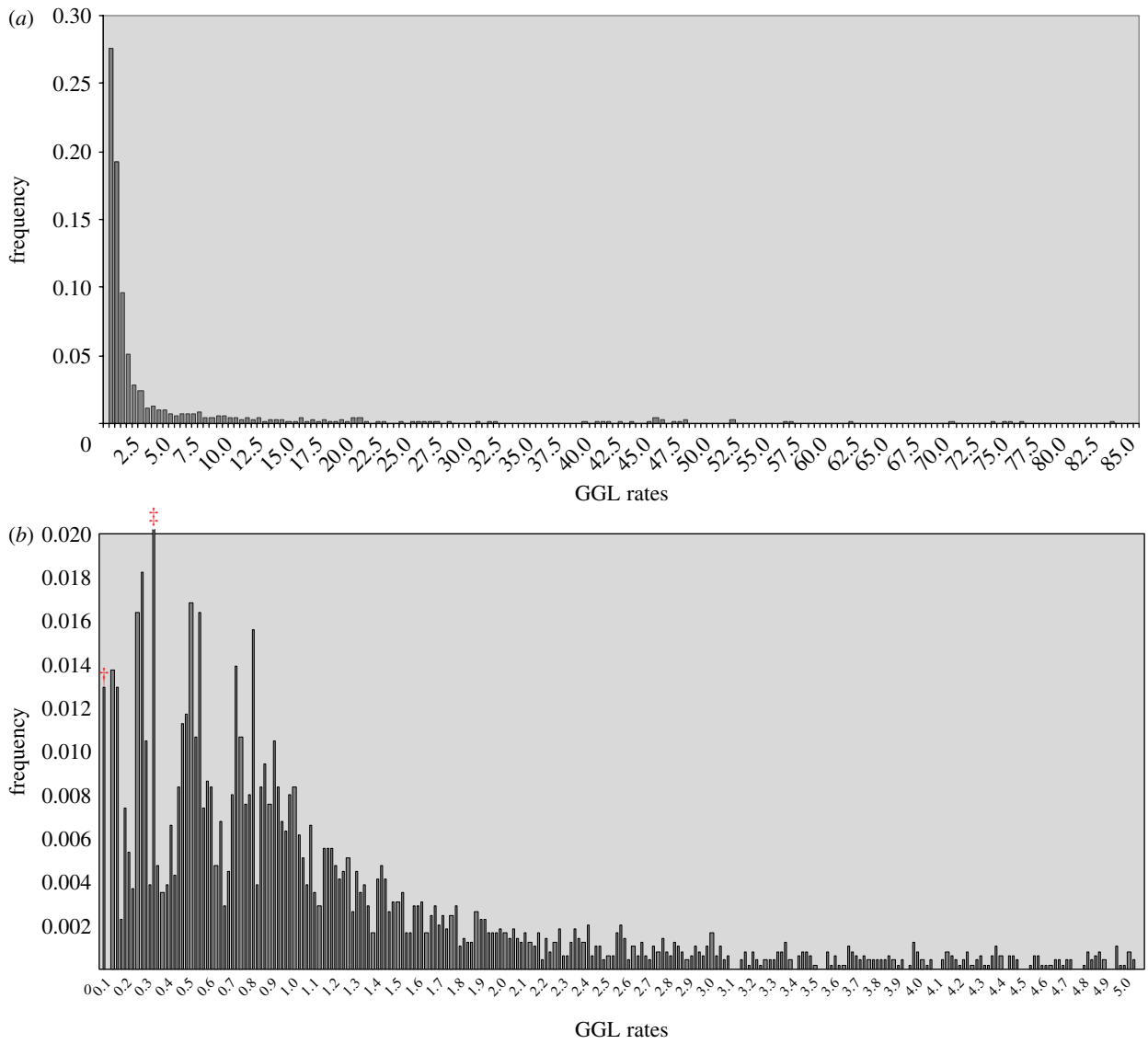

Figure 1. Empirical distribution of GGL rates inferred using ML from 4873 COG gene families. (*a*) The empirical distribution spanning GGL rates inferred from 0 to 87, capturing approximately 90% of all gene families, with bin size 0.5 and (*b*) from 0 to 5, capturing approximately 70% of all gene families, with bin size 0.02. The bin denoted by the single dagger represents the GGL rate of the 63 gene families that are present in all species. The bin denoted by the double dagger represents the GGL rate of the 286 gene families that are present only in the two eukaryotes.

Alternatively, once the gamma distribution is assumed and its parameters are optimized using ML, it is straightforward to compute the posterior expectation of the rate for each gene family using Bayes' theorem. This is equivalent to the Bayesian estimation of site-specific conservation scores, as used for sequence analysis (Yang & Wang 1995; Mayrose *et al.* 2004). The implementation of the rate distribution is achieved by partitioning it into K discrete rate categories (Yang 1994). The rate of gene family f is thus computed as follows:

$$\hat{r}_f = \frac{1}{P(D_f)} \sum_{k=1}^{K} r^{(k)} P(D_f | T, t, Q^{(k)}) P(r^{(k)}), \tag{2.7}$$

where $Q^{(k)}$ is the rate matrix Q, in which $r_f = r^{(k)}$.

We applied the categorization with several numbers of categories, ranging from 4 to 32. All results presented in this work are computed with 16 discrete categories. We note, however, that essentially the same results were obtained when either 32 or only 4 categories were used (not shown).

(c) *Likelihood correction accounting for unobservable data*

Gene families that are absent in all taxa are unobservable. To correct for unobservable data, we used the approach suggested for restriction site analysis (Felsenstein 1992). The likelihood computation of each gene family is conditioned on observing '1' in at least one of the species. In terms of probability, instead of simply computing $P(D)$, we compute $P(D|\text{'observable'})$, which is equal to $P(D$ & 'observable')$/(1-P(\text{'missing'}))$,

$$L_f^{(+)} = \frac{L_f}{1 - L^{(-)}}, \tag{2.8}$$

where $L_f^{(+)}$ is the required conditional likelihood for the gene family f; L_f is the likelihood of the observable data for the gene family f; and $L^{(-)}$ is the likelihood of a gene family to be absent in all taxa, i.e. the likelihood of a column of zeros.

Computing the likelihood of all gene families in the tree is straightforward, achieved by multiplying the likelihoods of all

Table 1. Maximum log-likelihood comparisons of different evolutionary models used for the analysis of phyletic patterns. (Models were tested using 63 species spanning 4873 gene families, using the COG phyletic pattern (Tatusov *et al.* 2003) and the tree of Ciccarelli *et al.* (2006). Each row in the table represents a single model and includes its underlying assumptions, the model parameters and the maximum log likelihood.)

model	assumptions	parameters	maximum log likelihood
M0	rate ~ homogeneous, $\pi_{root}=\pi_Q$ gain rate = loss rate	$r_Q=0.98$	−119 317
M1	rate ~ homogeneous $\pi_{root}=\pi_Q$	$r_Q=1.68$ $\pi_Q(0)=0.8$	−111 513
M2	rate ~ homogeneous	$r_Q=1.65$ $\pi_Q(0)=0.81$ $\pi_{root}(0)=0.31$	−111 354
M0 + Γ	rate ~ $\Gamma(\alpha)$ $\pi_{root}=\pi_Q$ gain rate = loss rate	$r_Q=1.55$ $\alpha=0.7$	−114 585
M1 + Γ	rate ~ $\Gamma(\alpha)$ $\pi_{root}=\pi_Q$	$r_Q=6.59$ $\pi_Q(0)=0.85$ $\alpha=0.31$	−102 844
M2 + Γ	rate ~ $\Gamma(\alpha)$	$r_Q=7.32$ $\pi_Q(0)=0.87$ $\pi_{root}(0)=0.15$ $\alpha=0.64$	−102 417

gene families,

$$L^{(+)}=\prod_{f=1}^{F}L_f^{(+)}=\prod_{f=1}^{F}\frac{L_f}{1-L^{(-)}} \tag{2.9}$$

and after taking logs,

$$\begin{aligned}\log L^{(+)} &= \sum_{f=1}^{F}\log L_f-\sum_{f=1}^{F}\log(1-L^{(-)}),\\ &= \sum_{f=1}^{F}\log L_f-F\cdot\log(1-L^{(-)}).\end{aligned} \tag{2.10}$$

We note that for data extracted from the COG database (Tatusov *et al.* 2003), observable phyletic patterns are only those that are present in at least three species. In this case, the presented computation of $L^{(-)}$ is only an approximation for the unobservable data.

(d) *Likelihood computation and parameter estimation*

Likelihood computation is achieved using Felsenstein's (1981) pruning algorithm. All free parameters of the model described above are estimated such that they maximize the likelihood of the data. This is achieved using standard numerical algorithms.

(e) *Model comparison*

The likelihood-ratio test (LRT) is used to compare couples of nested models. The LRT statistic is $2\cdot\Delta\ell$ (where $\Delta\ell$ is the difference between the maximum log likelihoods of the compared models). It is approximated by the χ^2 distribution, where the number of degrees of freedom is equal to the difference in the number of free parameters between the two models.

(f) *Tree topology, branch lengths and the position of the root*

The reference tree topology and branch lengths were taken from the 'tree of life' (Ciccarelli *et al.* 2006). The topology and branch lengths are fixed for all models. The branch lengths in the tree of life can be transformed to branch lengths of GGL events per gene family, by multiplying all branches by the rate factor r_Q (detailed above). In the models in which the character frequencies at the root are identical to the stationary distribution, the position of the root does not affect the likelihood computations and hence the unrooted tree of Ciccarelli *et al.* (2006) was used. For the models in which the character frequencies at the root are allowed to differ from the stationary ones, the position of the root must be determined. Thus, for these models, we numerically found the position of the root by maximizing the likelihood of the data, given the root position.

(g) *Phyletic pattern*

The presence and absence of gene families for each species were extracted from the COG database (Tatusov *et al.* 2003). The intersection between the 66 species of the COG database and the Ciccarelli *et al.* (2006) tree retained 63 species including bacteria, archaea and eukaryotes, spanning 4873 gene families.

3. RESULTS

(a) *Analysis of phyletic patterns requires elaborate evolutionary models*

We implemented six models to analyse phyletic patterns. The performance of all models was evaluated over a dataset spanning 63 species including bacteria, archaea and eukaryotes, and 4873 gene families obtained from the COG database (Tatusov *et al.* 2003), using a reference phylogeny (Ciccarelli *et al.* 2006). We performed stepwise comparisons of the maximum log likelihoods obtained by the six models, starting with the most naive one and gradually alleviating the unrealistic assumptions by introducing more complex models.

The simplest model (M0 in table 1) assumes the following: (i) the rate of gain is equal to the rate of loss (in this case, the stationary frequencies are $\pi_Q(0)=\pi_Q(1)=0.5$), (ii) the character frequencies at the root (π_{root}) are the same as the stationary frequencies (π_Q), and (iii) the rate is equal for all gene families (a homogeneous rate model).

We first tested the assumption that the gain and loss rates are equal. Model M1 is identical to M0, except that an additional parameter is allowed to account for unequal gain and loss rates. The maximum log likelihood under M1 was a few thousand points higher than M0 (table 1). Comparing these scores using the LRT suggests that M1 fits the data overwhelmingly better than M0. Furthermore, under M1 $\pi_Q(0)=0.8$, which is substantially different from 0.5.

We next tested the assumption that the character frequencies at the root equal the stationary frequencies (table 1, M1 versus M2). The maximum log-likelihood scores under M2 were approximately 160 points higher than M1. Furthermore, the character frequencies at the root were different from the stationary one ($\pi_Q(0)=0.81$ versus $\pi_{\text{root}}(0)=0.31$).

Finally, the assumption that the gene family evolutionary rate is identical for all gene families was tested. This assumption was significantly rejected regardless of the model compared (table 1, M0 versus M0$+\Gamma$, M1 versus M1$+\Gamma$ and M2 versus M2$+\Gamma$). The increase in the maximum log likelihood was strikingly significant. We note that when the assumptions regarding character frequencies are tested by comparing the models, which account for AGRV, the same conclusions as above are obtained, i.e. the gain and loss rates differ (M0$+\Gamma$ versus M1$+\Gamma$) and the root frequencies differ from the stationary ones (M1$+\Gamma$ versus M2$+\Gamma$).

To conclude, our data clearly support a stochastic process with different rates for gain and loss, separate character frequencies at the root and a non-homogeneous GGL rate across gene families.

(b) *Estimating a gain–loss rate for each gene family*

Our probabilistic models can be used to infer for each gene family its underlying rate of gene gains and losses. In order to estimate the rate distribution, we first computed the ML estimate of the evolutionary rate for each gene family. The empirical distribution of gene-family-specific rates (figure 1) appears to follow a non-uniform distribution. It is evident that the empirical GGL rate distribution is highly asymmetric. The vast majority of the gene families have a low GGL rate, whereas several gene families with higher GGL rates comprise the heavy outlier. In figure 1*a*, the distribution of 90 per cent of all gene families is shown, spanning GGL rates from 0 to 87. The 10 per cent outlying gene families were truncated since it was impractical to present them in the same graph. In figure 1*b*, the distribution of 70 per cent of all gene families is shown. From this visualization it can be seen there that there are considerable differences in rates within the gene families with relatively low rates. Indicated in this figure is a GGL rate of approximately 0 that is shared by all gene families present in all species. Also indicated is the GGL rate of 286 out of the 4873 gene families that are only present in both eukaryotic genomes.

Our likelihood analysis revealed that a gamma distribution captures the variability of GGL rates among gene families. We thus adopted the empirical Bayesian approach for inferring the GGL rate of each gene family, in which a gamma prior distribution over GGL rates is explicitly assumed. This method allows an informative inference of GGL rates for all gene families studied. To exemplify the usefulness of the gene-family-specific rate inference, we identified the gene family with the highest GGL rate. The 'adenine-specific DNA methylase' gene family (COG2189) is sparsely present in 17 out of the 63 species analysed and was inferred to have a GGL rate of 4.46. This high rate is also apparent when projecting the presence and absence states of this gene family onto the underlying phylogenetic tree (figure 2*a*). The projection on the tree reveals that often the phyletic pattern has changed in a relatively short evolutionary time. For example, this gene is absent in *Thermoplasma acidophilum* but present in *Thermoplasma volcanium*, two relatively closely related taxa.

To emphasize the conceptual difference between the absence/presence frequencies and GGL rate inference using the phylogeny and the evolutionary model, we present the phyletic pattern of the '20S proteasome, alpha and beta subunits' gene family (COG0638). This gene family has the same presence/absence frequencies as the gene family above, i.e. it is also present in 17 out of the 63 species. However, projecting the phyletic pattern on the phylogeny reveals that here, the gene family is highly clustered (it is present in all eukaryotes and archaea and in a single bacterial clade, the Mycobacteriaceae). Accordingly, the estimated GGL rate for this gene family is 0.0525, which is in the lowest fifth percentile (figure 2*b*).

(c) *Simulation studies*

We conducted a simulation study to estimate the accuracy of our method for inferring gene-family-specific rates. We first obtained a set of reference rates produced using the inferred rates of the analysed data of 4873 gene families with model M2$+\Gamma$. Using the reference rates, we simulated the stochastic process under the same evolutionary model and established 1000 sequences. For each of the simulated sequences the rates were inferred and the Spearman correlation was computed between the true rates and the inferred rates. The average correlation coefficient was 0.857 with a standard deviation of 0.0037. Since the same model was used for simulating the data and for the inference, the lack of perfect correlation is due to the stochastic nature of the evolutionary process. Our simulation results thus suggest that given an accurate model for the GGL dynamics, it is possible to infer evolutionary rates of gain and loss with relatively high reliability.

4. DISCUSSION

In this paper, we have devised a new methodology to analyse phyletic patterns. Our motivation was to develop biologically realistic evolutionary models that capture the GGL dynamics. Our phyletic models are adapted from sequence-based likelihood models, and as such they provide a firm probabilistic approach and are supported by a set of pre-developed techniques and algorithms, allowing hypothesis testing and comparisons of parameters across datasets. We have

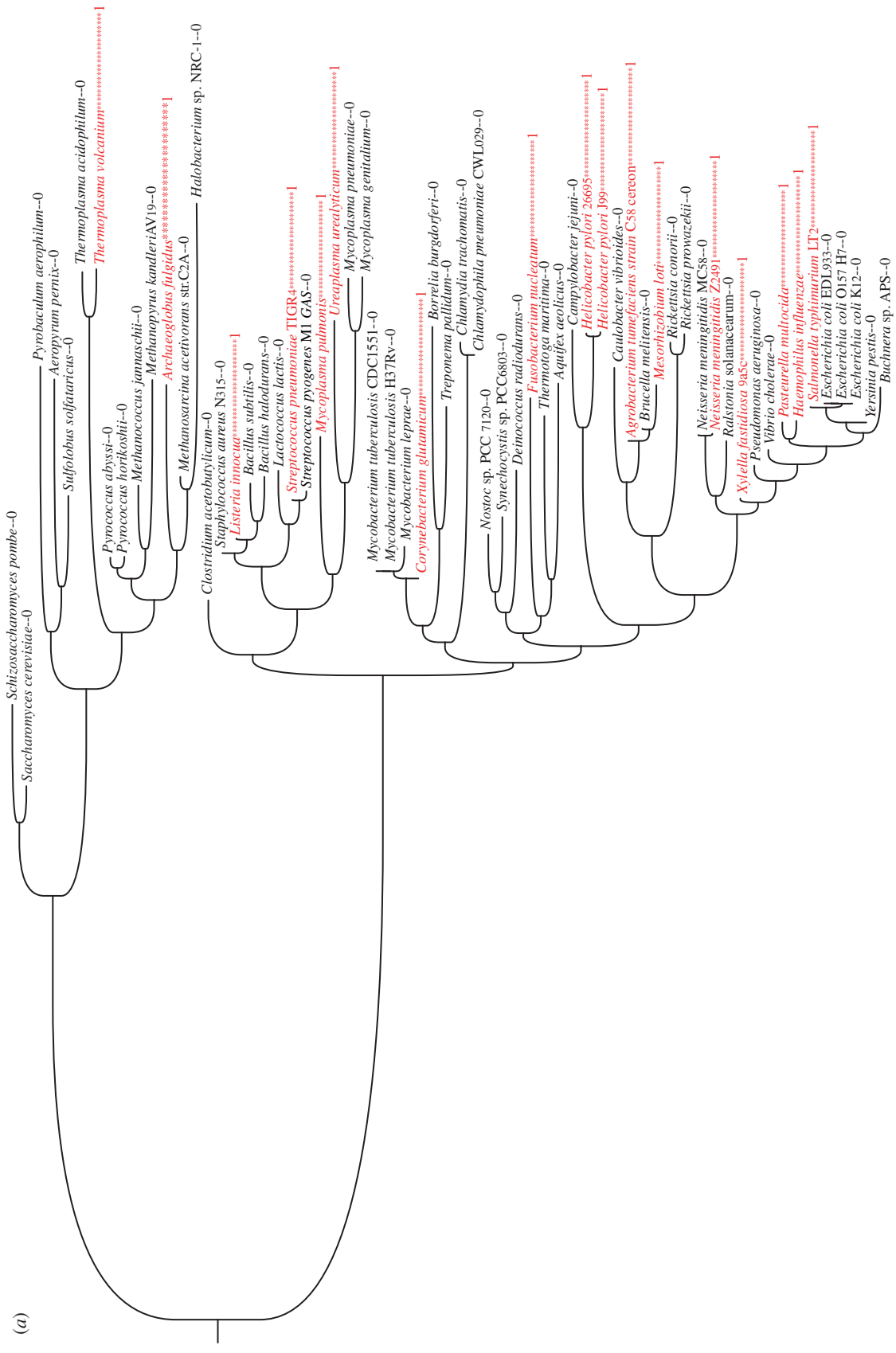

Figure 2. The projection of the phyletic pattern of two gene families on the species tree. The presence and absence are indicated by '1' and '0' extensions to the species name, respectively. (*a*) The adenine-specific DNA methylase gene family, inferred to have a very high GGL rate. (*b*) The 20S proteasome, alpha and beta subunits gene family, inferred to have a low rate. (*Continued opposite.*)

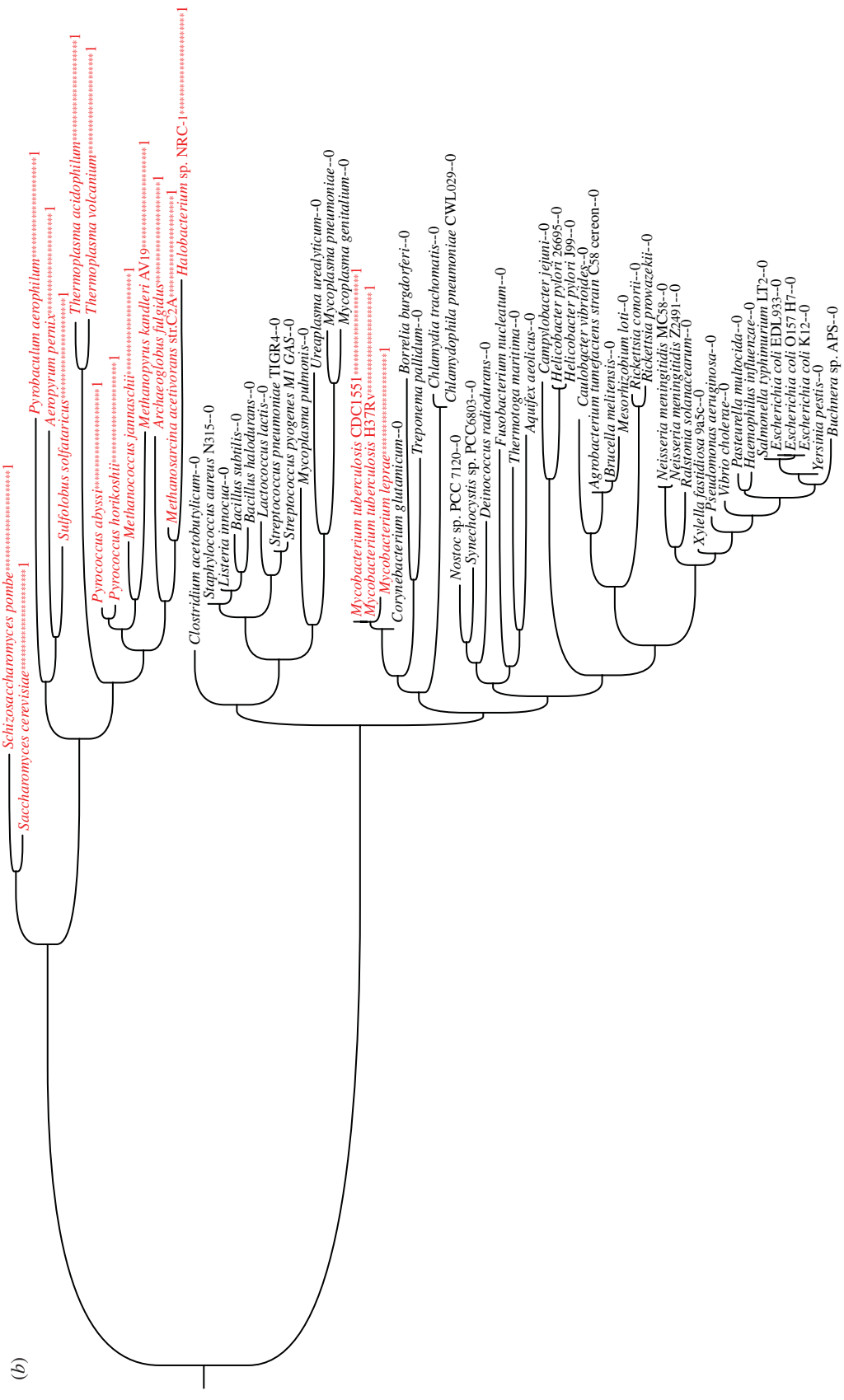

Figure 2. (*Continued.*)

implemented several models, increasingly realistic, while gradually relaxing oversimplified assumptions. The various models were subsequently used to analyse the phyletic pattern extracted from the COG database (Tatusov *et al.* 2003) using a reference phylogeny (Ciccarelli *et al.* 2006). The dataset consists of 63 species, spanning 4873 different gene families.

Our analysis indicates that it is appropriate to allow GGL rates to differ from each other and to allow the rate of gains and losses to vary among gene families. Since oversimplified models can often lead to erroneous conclusions, we suggest that a framework of probabilistic models, such as the one presented here, should be used for alleviating unrealistic assumptions concerning the underlying evolutionary process generating phyletic data.

In a model that allowed the root frequencies to differ from the stationary ones, a large difference was observed both in terms of maximum log-likelihood scores and between the estimates of the frequencies (root versus stationary). We speculate that this difference indicates that the stochastic process is not homogeneous (across the tree). In other words, all of the above models assumed a single set of stationary frequencies across the entire phylogeny. This assumption, however, fails in scenarios where, for example, some genomes underwent reductive evolution. The discrepancy in the estimated character frequencies at the root ($\pi_{\mathrm{root}}(1)=0.85$) and the stationary ones ($\pi_Q(1)=0.16$) suggests that such a scenario of reductive evolution of specialist microbes is rather common.

In this work, we assumed a pre-determined tree topology, and that branch lengths are fixed up to a proportional scale, which is ML estimated by the scaling factor r_Q. Since the reference tree is unrooted, the location of the root is also ML estimated. The root was estimated to reside in the edge connecting the bacteria with the archaea and eukaryota groups, as seen in figure 2. Notably, this ML estimate of the position of the root from gene content is consistent with the accepted evolutionary history.

The fixed parameters can also be estimated from the phyletic pattern data. Regarding the tree topology, it was not re-estimated here, since it has been suggested that the most commonly used methods (e.g. multiple sequence alignment based, gene order based and shared gene content based) result in a similar species phylogeny (Snel *et al.* 2005). It has also been suggested that the estimation of topology based on gene content is prone to systematic error (Spencer *et al.* 2007). Concerning branch lengths, multiplying them by our scaling factor transforms their measure to units of GGL events per gene family. One shortcoming of this approach is that we implicitly assume that higher substitution rates are perfectly correlated to higher GGL rates. It is interesting to test how good this correlation is, by comparing branch lengths estimated from sequence data versus branch lengths estimated based solely on the phyletic pattern data. Such comparisons may reveal lineages that underwent dramatic GGL events in a short evolutionary time, when the latter is measured in terms of base-pair substitutions.

To conclude, we have developed a novel framework for analysing the phyletic pattern, which is a step towards more realistic description of the GGL dynamics. Our main effort was to better capture variation of GGL rates among genes. In sequence-based analysis, substantial efforts are invested in developing evolutionary models that can capture variation of the stochastic process among tree lineages. As a case in point, recent models were developed to identify positive selection in specific branches and sites using codon models (e.g. the branch-site model of Zhang *et al.* (2005)). Initial work in this direction regarding the phyletic pattern analysis was done by Marri *et al.* (2007), in which each branch was allowed to have its own GGL rate. The next challenge with respect to the phyletic pattern models will thus be to integrate lineage-specific processes with AGRV, i.e. to allow for model variation both among gene families and among specific lineages.

We thank Matthew Spencer, Itay Mayrose, Eyal Privman and David Burstein for critical reading of the manuscript and providing helpful suggestions. A.S. is a fellow of the Complexity Science Scholarship programme. N.D.R. is a fellow of the Edmond J. Safra programme in bioinformatics. T.P. and U.G. were supported by the Research Networks Programme in Bioinformatics of the Ministry of Science and Technology of Israel, and the Ministries of Foreign Affairs and National Education and Research of France. This study was also supported by an Israeli Science Foundation grant to T.P.

REFERENCES

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B. & Bork, P. 2006 Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283–1287. (doi:10.1126/science.1123061)

Doolittle, R. F., Feng, D. F., Anderson, K. L. & Alberro, M. R. 1990 A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J. Mol. Evol.* **31**, 383–388. (doi:10.1007/BF02106053)

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)

Felsenstein, J. 1992 Phylogenies from restriction sites: a maximum-likelihood approach. *Evol. Int. J. Org. Evol.* **46**, 159–173.

Gogarten, J. P. & Townsend, J. P. 2005 Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3**, 679–687. (doi:10.1038/nrmicro1204)

Hao, W. & Golding, G. B. 2006 The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res.* **16**, 636–643. (doi:10.1101/gr.4746406)

Jain, R., Rivera, M. C. & Lake, J. A. 1999 Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806. (doi:10.1073/pnas.96.7.3801)

Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. 2003 Horizontal gene transfer accelerates genome innovation and evolution. *Mol. Biol. Evol.* **20**, 1598–1602. (doi:10.1093/molbev/msg154)

Koonin, E. V., Makarova, K. S. & Aravind, L. 2001 Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* **55**, 709–742. (doi:10.1146/annurev.micro.55.1.709)

Lake, J. A. & Rivera, M. C. 2004 Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* **21**, 681–690. (doi:10.1093/molbev/msh061)

Lawrence, J. G. & Roth, J. R. 1996 Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843–1860.

Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. 2005 Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3**, e130. (doi:10.1371/journal.pbio.0030130)

Marri, P. R., Hao, W. & Golding, G. B. 2007 The role of laterally transferred genes in adaptive evolution. *BMC Evol. Biol.* **7**(Suppl. 1), S8. (doi:10.1186/1471-2148-7-S1-S8)

Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. 2004 Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791. (doi:10.1093/molbev/msh194)

Mira, A., Klasson, L. & Andersson, S. G. 2002 Microbial genome evolution: sources of variability. *Curr. Opin. Microbiol.* **5**, 506–512. (doi:10.1016/S1369-5274(02)00358-2)

Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. 2003 Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2. (doi:10.1186/1471-2148-3-2)

Nelson, K. E. *et al.* 1999 Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329. (doi:10.1038/20601)

Ochman, H., Lawrence, J. G. & Groisman, E. A. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304. (doi:10.1038/35012500)

Pennisi, E. 2004 Microbiology. Researchers trade insights about gene swapping. *Science* **305**, 334–335. (doi:10.1126/science.305.5682.334)

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. 2002 Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**(Suppl. 1), S71–S77.

Ross, S. M. 1996 *Stochastic processes*. New York, NY: Wiley.

Snel, B., Huynen, M. A. & Dutilh, B. E. 2005 Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* **59**, 191–209. (doi:10.1146/annurev.micro.59.030804.121233)

Spencer, M., Bryant, D. & Susko, E. 2007 Conditioned genome reconstruction: how to avoid choosing the conditioning genome. *Syst. Biol.* **56**, 25–43. (doi:10.1080/10635150601156313)

Syvanen, M. 1994 Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.* **28**, 237–261.

Tatusov, R. L. *et al.* 2003 The COG database: an updated version includes eukaryotes. *BMC Bioinform.* **4**, 41. (doi:10.1186/1471-2105-4-41)

Thompson, J. R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D. E., Benoit, J., Sarma-Rupavtarm, R., Distel, D. L. & Polz, M. F. 2005 Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**, 1311–1313. (doi:10.1126/science.1106028)

Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. 2007 Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61. (doi:10.1038/nature06107)

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. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314. (doi:10.1007/BF00160154)

Yang, Z. 1995 A space-time process model for the evolution of DNA sequences. *Genetics* **139**, 993–1005.

Yang, Z. & Nielsen, R. 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43.

Yang, Z. & Wang, T. 1995 Mixed model analysis of DNA sequence evolution. *Biometrics* **51**, 552–561. (doi:10.2307/2532943)

Zhang, J., Nielsen, R. & Yang, Z. 2005 Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479. (doi:10.1093/molbev/msi237)