

Empirical profile mixture models for phylogenetic reconstruction

Le Si Quang¹, Olivier Gascuel¹ and Nicolas Lartillot^{1*}

¹ Méthodes et Algorithmes pour la Bioinformatique. LIRMM, CNRS-UM2, 141 rue Ada, 34392 Montpellier Cedex 5, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Previous studies have shown that accounting for site-specific amino acid replacement patterns using mixtures of stationary probability profiles offers a promising approach for improving the robustness of phylogenetic reconstructions in the presence of saturation. However, such profile mixture models were introduced only in a Bayesian context, and are not yet available in a Maximum Likelihood framework. In addition, these mixture models only perform well on large alignments, from which they can reliably learn the shapes of profiles, and their associated weights.

Results: In this work, we introduce an expectation-maximization algorithm for estimating amino-acid profile mixtures from alignment databases. We apply it, learning on the HSSP database, and observe that a set of 20 profiles is enough to provide a better statistical fit than currently available empirical matrices (WAG, JTT), in particular on saturated data.

Availability: We have implemented these models into two currently available Bayesian and Maximum Likelihood phylogenetic reconstruction programs. The two implementations, PhyloBayes, and PhyML, are freely available on our web site (<http://atgc.lirmm.fr/cat>). They run under Linux and MacOSX operating systems.

Contact: nicolas.lartillot@lirmm.fr

1 INTRODUCTION

Capturing the evolutionary properties of the amino-acid replacement process in protein sequences has traditionally been done using empirical matrices. Such matrices are meant to account for the biochemically conservative pattern of amino-acid replacement in natural protein sequences. In practice, they are empirically derived from databases of pairwise or multiple alignments, either using counting methods (Dayhoff *et al.*, 1978; Jones *et al.*, 1992), or by direct likelihood maximization (Adachi and Hasegawa, 1996; Adachi *et al.*, 2000; Whelan and Goldman, 2001; Le and Gascuel, 2008).

Empirical matrices appear to lead to simple and fairly accurate phylogenetic models (Whelan *et al.*, 2001). On the other hand, they do not explicitly encode the fact that biochemical constraints are essentially site-specific features, related to context-dependent purifying selection. More precisely, depending on their position and role in the protein's overall shape and structure, sites will in general

accept only a very specific subset of the 20 amino-acids, all other possibilities being strongly selected against. For instance, buried sites will preferentially accept hydrophobic amino-acids, whereas residues at an active site are more likely to be electrostatically charged. This suggests, as an alternative way of describing protein evolution, probabilistic models explicitly formulated in terms of variations of the amino acid propensities across sites (Bruno, 1996; Halpern and Bruno, 1998; Crooks and Brenner, 2005). Such models need to be devised with caution, however, so as to control the overall parameterization. In this respect, two essential simplifications can be proposed.

First, assuming that biochemical constraints can be explained through equilibrium frequencies, one can rely on simple processes, such as F81 (Felsenstein, 1981), that are entirely characterized by their profile of stationary probabilities over the 20 amino acids; the replacement rate between two amino acids does not depend on the initial state, and is simply proportional to the stationary probability (or equilibrium frequency) of the target state. Second, to avoid relying on site-specific stationary probability profiles, which may create an over-parameterization problem, we can assume that the variation of amino acid constraints is explainable as a finite mixture.

Mixture models have been used several times in phylogenetics, to model variations across sites of various aspects of the substitution or replacement process, such as the dN/dS ratio (Yang *et al.*, 2000), or even the complete substitution matrix (Koshi and Goldstein, 1998; Pagel and Meade, 2004). More sophisticated mathematical devices, such as Hidden Markov Models, of which mixture models can be seen as a degenerate case, have also been proposed (Felsenstein and Churchill, 1996; Goldman *et al.*, 1996; Thorne *et al.*, 1996; Goldman *et al.*, 1998). In another direction, Markov Modulated Models have been introduced, as a way of allowing sites to switch their substitution behavior at any time along the lineages (Holmes and Rubin, 2002; Gascuel and Guindon, 2007). Mixture models can be seen as a degenerate case of Markov Modulated Models, in which the switching rate is equal to 0.

A mixture of profiles has been introduced in a previous work (Lartillot and Philippe, 2004). It is based on a Dirichlet process mixture (Ferguson, 1973; Antoniak, 1974), which is a flexible non-parametric device often used in Bayesian studies (Neal, 2000; Huelsenbeck *et al.*, 2006; Huelsenbeck and Suchard, 2007). The resulting model was shown to provide a better fit than empirical matrices in several cases (Lartillot and Philippe, 2006). It also

*to whom correspondence should be addressed

appears to better accommodate saturation (i.e. multiple substitutions), making it more robust to phylogenetic artifacts (Lartillot *et al.*, 2007).

Non-parametric modelling based on Dirichlet process mixtures is a powerful method. One of its essential advantages is its flexibility, adapting to each particular dataset. However, a large amount of data is necessary for this non-parametric device to converge to a reasonably faithful description of the true distribution of profiles across sites. For this reason, it performs well on large alignments (more than 1,000 aligned positions), such as multiple gene concatenations. In contrast, it appears to be less efficient on smaller datasets, in particular, on single gene alignments. Another limiting feature is that Dirichlet process mixtures can only be practically implemented in a Monte Carlo framework, which makes them a method of choice for Bayesian studies, but a much less attractive tool in a Maximum Likelihood (ML) paradigm. Yet, the availability of a profile mixture model in a ML context could be useful in many situations.

Thus, we still lack an efficient, versatile (ML and Bayes) version of the profile mixture model, whose parameters would be pre-determined on empirical grounds, so that it could be used on small single-gene alignments as well as on large phylogenomic datasets. In the present paper, we develop such a model. Starting from a variant of the Expectation-Maximization method (Dempster *et al.*, 1997) especially devised for phylogenetic models (Holmes and Rubin, 2002), we specialize it to profile mixture models, and apply it on a subset of HSSP database (Sander and Schneider, 1991), to estimate a series of profile mixture models with an increasing number of components. We find that a good compromise is obtained in the form of a mixture of 20 profiles. When tested on phylogenetic alignments, this model gives, on average, a better fit than classical replacement matrices, such as WAG and JTT. Interestingly, profile mixtures are particularly more fit on saturated data, confirming that matrices specifically fail at correctly describing multiple conservative substitutions (Lartillot *et al.*, 2007). On the other hand, matrices are often better than profile mixtures when a low level of saturation is detected in the data. Finally, we show that 20 profiles are enough to reproduce some of the essential phylogenetic robustness properties previously observed using the non-parametric Bayesian version.

2 METHODS

Notations

The training database consists of a series of alignments $(D_r)_{r=1..R}$. For simplicity, we explain the overall method for a single alignment D . The generalization to several alignments is straightforward. The alignment D is made of P sequences, of length N , which are assumed to be related by an unknown, rooted phylogenetic tree τ . In all cases investigated in this work, the topology is considered fixed. The sequences are taken from an alphabet of size S . In practice, we will work on amino-acid alignments ($S = 20$), although the method introduced here is also valid for nucleic acid data.

Let i index the columns C_i , or sites, and j the nodes ($0 \leq j \leq 2P - 3$), with root node having index 0 and leaf node indexed in $[1, P]$. Branches are also indexed by j , $1 \leq j \leq 2P - 3$, with the convention that a branch has the same index as the node which is at its tip. Branch lengths are denoted by $l = (l_j)_{1 \leq j \leq 2P - 3}$, the relative rate of substitution at each site by $\mathbf{r} = (r_i)_{1 \leq i \leq N}$.

Each cell of the data matrix is referred to as x_{ij} for i and j running over sites (columns) and taxa (rows), respectively. Thus, x_{ij} is simply the state

of the process operating at site i , at the leaf indexed by j ($1 \leq j \leq P$). In addition, one will need to specify the state of the substitution process at all interior nodes as well. Let us simply expand the range of j so that now, x_{ij} is defined for $0 \leq j \leq 2P - 3$.

Profile mixture models

Substitutions occur independently at each site, according to Poisson (or F81, Felsenstein, 1981) Markov processes running along the branches of the tree. One such Markov process is entirely characterized by a vector of stationary probabilities, or equilibrium frequencies $\pi = (\pi(a))_{a=1..S}$, such that $\sum_a \pi(a) = 1$. In the following, we will call π the *profile* of the process.

The finite-time probability transition for a Poisson process takes a simple form: the probability of observing state b , after time $t = l$, and given that the process was in state a at $t = 0$ is

$$p(b | a, l) = e^{-l} \delta_{ab} + (1 - e^{-l}) \pi_b, \quad (1)$$

where δ_{ab} is the Kronecker symbol (i.e. is equal to 1 iff $a = b$).

We will consider a mixture of K Poisson processes. $\pi = (\pi_k)_{k=1..K}$, where each π_k is a vector of frequencies over the S states of the alphabet. To each component of the mixture is associated a weight w_k , such that $\sum_k w_k = 1$. We denote by $\mathbf{w} = (w_k)_{k=1..K}$ the weight vector.

In addition, we assume variations of the overall rate of substitutions across sites. As in most standard phylogenetic models, we do this by introducing a discretized gamma distribution (Yang, 1994), of mean 1 and (inverse variance) parameter α . In effect, this amounts to introducing a series of Q rates $\mathbf{r}(\alpha) = (r_q(\alpha))_{q=1..Q}$, each of which is the mean rate within the q th. bin of the discretization. In general, bins are chosen so that all the weights are equal to $1/Q$. The values of the discretized rates depend on the α parameter. We use $Q = 4$ throughout the present article.

The full parameter vector is thus $\Theta = (1, \pi, \mathbf{w}, \alpha)$. The likelihood at site i is a weighted average over all $K \times Q$ combinations of rates and profiles:

$$p(C_i | \Theta) = p(C_i | 1, \pi, \mathbf{w}, \alpha) = \frac{1}{Q} \sum_k w_k \sum_q p(C_i | 1, \pi_k, r_q(\alpha)), \quad (2)$$

and the total likelihood is the product over all sites:

$$p(D | \Theta) = \prod_i p(C_i | \Theta). \quad (3)$$

Expectation-Maximization algorithm

An Expectation-Maximization (EM) method (Dempster *et al.*, 1997) for estimating amino-acid replacement models has been introduced previously (Holmes and Rubin, 2002). It assumes a general (possibly Markov-modulated) time-reversible process, applied uniformly across sites. In the present context, we adapt the method so as to deal with mixture models, and to account for rate variations across sites. Concomitantly, we specialize it to profile models.

The EM algorithm is particularly convenient for Markov processes, compared to other numerical likelihood maximization methods, essentially because the conditional expectations involved can often be analytically maximized with respect to the parameters of the model (Holmes and Rubin, 2002). In the present case, the use of profiles leads to further simplifications in the analytical part of the computations.

Specifically, our EM algorithm relies on a combination of data augmentation and parameter expansion. The parameter expansion part consists in specifying the allocation of each site to one of the K components of the mixture, and to one of the Q rates of the discretized gamma distribution. To do this, we introduce two allocation vectors $\mathbf{y} = (y_i)_{i=1..N}$, and $\mathbf{z} = (z_i)_{i=1..N}$, such that $y_i \in [1..K]$ and $z_i \in [1..Q]$ are two integer indicators specifying the allocation status of site i . The likelihood can be rewritten as a sum over all possible joint realizations of these two allocation vectors:

$$p(D | \Theta) = \sum_{\mathbf{y}, \mathbf{z}} p(D | \mathbf{y}, \mathbf{z}, \Theta) p(\mathbf{y}, \mathbf{z} | \Theta) \quad (4)$$

where

$$p(D | \mathbf{y}, \mathbf{z}, \Theta) = \prod_i p(C_i | \mathbf{1}, \pi_{z_i}, r_{y_i}(\alpha)) \quad (5)$$

and

$$p(\mathbf{y}, \mathbf{z} | \Theta) = p(\mathbf{y})p(\mathbf{z} | \mathbf{w}) = \frac{1}{Q^N} \prod_i w_{z_i}. \quad (6)$$

The data augmentation part consists in specifying the state of the process at all internal nodes of the tree, and for each site $(\mathbf{x} = (x_{ij}))$, as well as the number of substitutions along each branch and for each site, which we denote by $\mathbf{n} = (n_{ij})$, n_{ij} being the number of substitutions on branch j and for site i . We call $\Xi = (\mathbf{x}, \mathbf{n})$ a *substitution mapping*. Note that a more complete account of the substitution history would be possible (Nielsen, 2002; Holmes and Rubin, 2002), by including the states of the process after each of the n_{ij} substitutions, and the exact position of each successive substitution event along the branch. However, we do not need it in the present case. Conditional on a particular allocation (\mathbf{y}, \mathbf{z}) , the likelihood is a sum over all possible mappings compatible with the data at the leaves D :

$$p(D | \mathbf{y}, \mathbf{z}, \Theta) = \sum_{\Xi | D} p(\Xi | \mathbf{y}, \mathbf{z}, \Theta). \quad (7)$$

The function to be maximized in our EM is then the expectation of $\ln p(\Xi | \mathbf{y}, \mathbf{z}, \Theta)$ over all possible realizations of mappings Ξ and allocations (\mathbf{y}, \mathbf{z}) , conditional on the current value Θ^* of the parameter:

$$Q(\Theta, \Theta^*) = E_{\Xi, \mathbf{y}, \mathbf{z}}[\ln p(\Xi | \mathbf{y}, \mathbf{z}, \Theta) | \Theta^*, D] \quad (8)$$

This expectation takes on a very simple form (see Appendix for details), that makes maximization with respect to Θ straightforward.

In practice, the optimization proceeds by alternating between four modules, performing an expectation followed by a maximization with respect to either $\mathbf{1}$, α , π or \mathbf{w} . In addition, as the alignments of the training database are independent given the parameters of the mixture (\mathbf{w} and π), we have implemented two separate programs: one performing the maximizations with respect to $\mathbf{1}$ and α conditional on the current value of \mathbf{w} and π for each alignment taken in turn, and returning the relevant expectations to another program, whose task is then to sum up the expectations obtained from all alignments, and perform the maximization with respect to \mathbf{w} and π .

3 DATA, LEARNING AND TESTING PROCEDURE

Models were estimated on a subset of the HSSP database (Sander and Schneider, 1991). This database contains more than 32,012 alignments, corresponding to globular proteins for which the three dimensional structure is known. Many alignments contain a large number of sequences, which we assume might help inferring more accurate amino-acid replacement models. On the other hand, most alignments contain a high proportion of missing data. Furthermore, there is a high level of redundancy, many sequences being represented more than one time in the database. We therefore proceeded to a cleaning process (Supplementary Information), so as to retrieve a subset of 1,030 ungapped and non-redundant alignments (with an average of 40 sequences and 253 aligned positions per alignment).

We also tried an alternative database, HOGENOM release 3 (Dufayard *et al.*, 2005). Each alignment of HOGENOM was scanned with G-blocks (Castresana, 2000), using the default options, and 1,200 alignments, with number of taxa ranging from 15 to 50, were selected at random.

For each alignment, the topology of the phylogenetic tree was inferred under the WAG model (Whelan and Goldman, 2001), using PhyML (Guindon and Gascuel, 2003). We used the EM algorithm mentioned above to estimate profile models, with K , the number of components of the mixture, ranging from 10 to 60, every 10. The corresponding model configurations will be referred to as C10, C20... C60. We used Xrate (Holmes and Rubin, 2002) and the same procedure as in Whelan and Goldman (2001) to estimate a single matrix model from the 1,030 HSSP alignments, which we named WAG_HSSP.

The fit of the models was evaluated by estimating the maximum likelihood under 57 protein alignments of TreeBase (Sanderson *et al.*, 1993). For both matrix and profile models, the likelihood has to be maximized with respect to the topology, the branch lengths and the α parameter of the gamma distribution of rates across sites. The topology was optimized using the SPR search developed in PhyML (Hordijk and Gascuel, 2005), starting from the tree provided by TreeBase. In the case of profile mixture models, we can set the weights (\mathbf{w}) associated to each component of the mixture equal to those inferred on the training database. In that case, all model configurations have the same number of parameters, and the likelihood score can then be directly compared between all models, including the WAG empirical matrix. Alternatively, we also tested the possibility of reoptimizing the weights on the test alignment, in which case the number of additional parameters, compared to WAG or to mixture models with fixed weights, is then equal to $K - 1$. To account for such variations in model dimensionality, we used the Bayesian Information Criterion (Schwartz, 1978):

$$BIC = -2 \ln \hat{L} + m \ln N,$$

as well as the second order Akaike information criterion (Akaike, 1974):

$$AIC = -2 \ln \hat{L} + 2m + \frac{2m(m+1)}{N-m-1},$$

where $\ln \hat{L}$ is the maximum likelihood estimate under the model of interest, N is the number of aligned positions of the test dataset, and m is the number of parameters specific to the model under investigation (here, $m = K - 1$). The second order correction is necessary for small alignments. In any case, it automatically reduces to the standard first order AIC when $N \gg m$. All AIC and BIC scores are displayed using WAG_HSSP as the reference. Thus, for a given model M , the score is defined as:

$$\Delta AIC_M = AIC_{WAG_HSSP} - AIC_M,$$

so that better models have higher scores.

Significance of the difference in log likelihood was assessed, between alternative models with the same number of parameters, or between a priori specified alternative topologies under the same model, using the Kishino Hasegawa test (Kishino and Hasegawa, 1989). When two alternative models, M_1 and M_2 , return distinct ML trees T_1 and T_2 when applied on a given dataset, we may be interested in knowing whether the two trees are contained in the confidence sets of each model, which we assessed using the Shimodaira Hasegawa test (Shimodaira and Hasegawa, 1999). Specifically, if we call L_{i1} and L_{i2} the log-likelihoods of the two trees under model M_i , $i = 1, 2$, we assess the null-hypothesis that $E[L_{i2}] = E[L_{i1}]$. Note that in the present case, where we test only two topologies, the Shimodaira Hasegawa test reduces to the Kishino Hasegawa test.

4 RESULTS

We used an Expectation-Maximization algorithm (see methods) to estimate by Maximum Likelihood a series of mixture models, with the number of components ranging from 10 to 60. We tried two alternative training databases: a subsample of HSSP (Sander and Schneider, 1991), and a subset of HOGENOM (Dufayard *et al.*, 2005). The resulting mixture models were then assessed by measuring their AIC and BIC score on a series of 57 alignments taken from TreeBase. We reasoned that testing on alignments that do not come from the same database as those that were used for estimating the parameters of the models avoids possible biases of the database. In addition, TreeBase contains alignments that have been produced especially for phylogenetic analyses, and thus, provide a priori a good benchmark for comparing models meant for phylogenetic reconstruction.

Reliability of the learning method

Focusing on the C20 model, we first evaluated the reliability and the reproducibility of the learning process. Five independent runs were performed, starting from random initial parameter configurations, and using the full training set (1,030 non-redundant alignments from HSSP). In addition, as a way of testing the effect of the finite size of the training set, we randomly split the training database into two subsets of 515 alignments, and estimated the parameters of the C20 model on each of them.

The independent runs performed on the full training database each time led to a distinct point of the parameter space, indicating that the likelihood surface has many local maxima. Such a complicated behavior was also observed in previous EM applications for Markov modulated models (Holmes and Rubin, 2002). However, the profiles obtained from different runs are very similar. Specifically, we could unequivocally identify 18 of the 20 profiles across two runs taken at random (Tables 2 and 3 of Supplementary Information). In addition, the differences in the AIC scores per site obtained when testing two independent runs against TreeBase are small (0.008), compared to the differences observed between the two independent training sets of 515 alignments (0.04), suggesting that effects due to the finite size of the database are dominant over problems related to the presence of local maxima. And more fundamentally, both sources of variability are smaller than the differences observed between different model configurations (see below).

Similar results were obtained under the C50 model: 41 of the 50 profiles could be unequivocally identified between two independent runs (Tables 4-7 of Supplementary Information), and a difference of 0.03 logarithmic units was observed between the resulting scores when testing on TreeBase. The difference was of 0.01 units when comparing the two independent training sets. Thus, for C50, the variations across independent runs and across independent subsets of the database are comparable, but still small compared to the differences between models.

The profiles obtained here confirm previous observations (Lartillot and Philippe, 2004). Importantly, they are sparse, in that most of them give a high probability for only 2 or 3 amino-acids, generally similar in their biochemical or geometrical properties. In higher dimensional models, such as C50, they also display some overlap and redundancy, and some profiles found for smaller models tend to split into more specialized forms in higher dimensional models. Some correlation can be found between the profile preferred at each site and other known structural attributes, such as exposure to solvent or, more weakly, secondary structure (Figures 1 and 2 of Supplementary Information). However, these correlations are weak, suggesting that the constraints experienced by each site in a protein are both more complex and more local than what can be predicted based on such simple structural and functional correlates.

Model comparison

We compared the profile mixture models C10-C60 with the best current time-reversible amino-acid replacement matrix, WAG (Whelan and Goldman, 2001). In addition, to compare models trained on the same database, we estimated a single matrix model from the 1,030 HSSP alignments (see Methods). This matrix, which we named WAG_HSSP, was used as our reference in the following model comparisons.

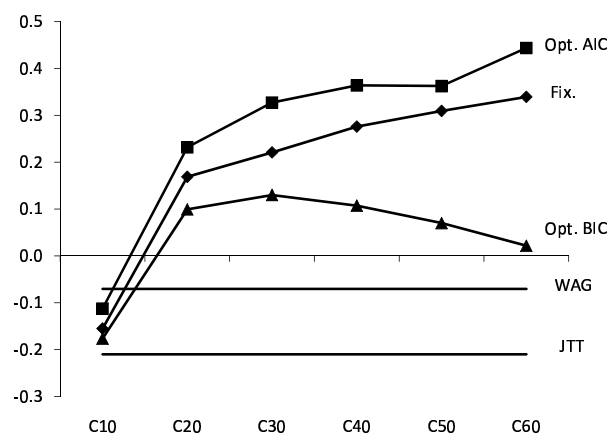


Fig. 1. AIC and BIC scores per site, using WAG_HSSP as the reference, as a function of the number of profiles in the mixture. Weights were either fixed (Fix.) or optimized (Opt.).

The WAG_HSSP matrix is clearly superior to JTT and WAG (Figure 1). The mean gain is 0.210 and 0.082 AIC/site, and a higher likelihood is obtained for 47 and 48 alignments out of 57, respectively. There are many reasons for the difference between these three matrix models among which the size and the quality of the training database, and the learning method (Le and Gascuel, 2008).

We first consider the mixture models with the weights of the components fixed to the values estimated along with the other parameters (Figure 1, Fix.). In this situation, all models have the same number of parameters, so that AIC and BIC scores are equivalent to twice the logarithm of the likelihood.

We see a clear improvement of the fit of the model as the number of categories increases (Figure 1). In fact, even with 60 categories, we did not reach the point where the fit would start to decrease. Among all mixture configurations, only C10 is worse than both WAG and WAG_HSSP. Thus, C10 may be too simple a model compared to single matrix models such as WAG or WAG_HSSP. In contrast, all other mixture models, from C20 to C60 clearly outperform both matrices.

Interestingly, the highest improvement in fit is accomplished when going from C10 to C20 (a gain of more than 0.3 AIC per site), whereas the difference between C20 and the best model tested here, C60, is of less than 0.2 AIC per site (Figure 1). This suggests that 20 profiles (but not 10) may be sufficient to capture the essential aspects of across site heterogeneities in amino-acid propensities.

If, instead of keeping the weights of the mixture fixed, we try to reoptimize them on the test alignments (Figure 1, Opt.) the conclusion are ambiguous: whereas AIC tends to favor weight reoptimization, BIC draws the opposite conclusion. In fact, according to BIC, not only do mixture models with reoptimized weights have a poorer fit, compared to their counterparts with fixed weights, but in addition, they are increasingly less fit as the number of components increases between 20 and 60. How to interpret this disagreement between AIC and BIC is not clear. A likelihood ratio test (LRT) can be performed, for assessing whether weights should be fixed or reoptimized for a given number of components. By this test, the fixed-weight null hypothesis is rejected in favor of weight reoptimization in about two thirds of the alignments at the 0.05 level (Table

Table 1. Number of alignments with significantly better (first number) or significantly worse (second number) likelihood scores under each model, compared to WAG_HSSP, as a function of saturation index. Significance is assessed by the Kishino-Hasegawa test, with a threshold p -value of 0.01.

	#Aln	C10	C20	C30	C40	C50	C60
all	57	5-13	17-6	18-6	19-6	24-5	27-5
$SAT < 1$	35	1-10	6-4	6-4	7-4	8-3	11-3
$SAT > 1$	22	4-3	11-2	12-2	12-2	16-2	16-2
$SAT > 2$	8	3-1	3-1	4-1	4-1	5-1	5-1

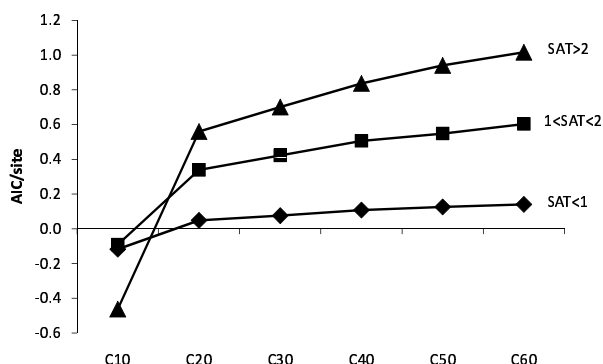


Fig. 2. Correlation between the AIC gain per site and the saturation level. We used as reference the mean AIC score per site obtained under WAG_HSSP, and averaged over the subset of alignments with given saturation index

8, Supplementary Information), thus suggesting that BIC puts too strong a penalty on the additional parameters represented by the weights. On the other hand, according to LRT, one third of the alignments still prefer the weights to be fixed to their predefined values. For simplicity, in the following, we focus exclusively on mixture models with fixed weights.

Finally, as an alternative to HSSP, we tried another database, HOGENOM (Dufayard *et al.*, 2005), for estimating a complete series of profile mixtures. These mixtures were then tested against TreeBase, and were found to lead to a poorer fit in all cases (Figure 3 of Supplementary Information). In the following, we only consider the profiles obtained on HSSP.

Mixture models are better on saturated data

Not all alignments prefer mixture models with more than 20 profiles over matrices. For instance, C20 was found significantly better than WAG_HSSP on 17 alignments out of 57 (Table 1), that is, 30% of the alignments. This proportion increases along with the number of components. Yet, even under the richest model proposed in the present work, at least 21% of the alignments seem to favor single matrix models over mixture models, among which 5 (9%) do so significantly. This raises the question of which aspects of the data are important in determining this preference.

Previous analyses have suggested that mixture models are especially better in the presence of saturation, i.e., when many sites of the protein have undergone a high level of multiple substitutions over the phylogenetic history. To investigate a potential relationship between fit and saturation in the present context, we computed a

Table 2. Number of sites, saturation index (SAT) and AIC score per site on genomic data sets

	# sites	SAT	AIC / site	
			C20	C50
Metazoans (first half)	17,136	1.96	0.61	0.91
Metazoans (second half)	17,135	2.14	0.60	0.91
Nematodes	35,366	1.45	0.31	0.52
Microsporidia	24,294	2.55	0.62	0.95
Nucleomorph	24,294	2.51	0.63	0.89

mean saturation index (SAT) for each of the 57 alignments of the test database. This saturation index is essentially the average minimum number of convergence or reversions per site, as inferred by a Maximum Parsimony reconstruction (see Lartillot *et al.*, 2007).

We observe that, as more and more saturated data are analyzed, the gain in fit obtained by profile mixtures is indeed higher (Figure 2, see also Figure 6 of Supplementary Information), and similarly, that the proportion of datasets for which profile mixtures are significantly better than WAG_HSSP is larger (Table 1). In particular, mixture models with 20 profiles or more were found significantly worse on only one alignment with a saturation index greater than 2. Altogether, this experiment confirms that profile models are better at describing saturated data.

Inferring the topology of the phylogenetic tree

We implemented WAG_HSSP and the profile mixture models in the program PhyML (Guindon and Gascuel, 2003). For each alignment of our test set, we searched for the maximum likelihood topology under WAG, WAG_HSSP, C20 and C50.

A different topology is found by WAG and by C20 in 51 out of 57 cases, with significance achieved in 4 cases. Similar observations were made when comparing WAG and C50, WAG and WAG_HSSP, or C20 and C50 (Table 9 of Supplementary Information). Note that the present statistical test is different from the previous one (Table 1). In Table 1, we were comparing the maximum likelihood scores obtained by the two models, each under its preferred topology. In the present case, we assess the two alternative trees under the same model, which is a more stringent assessment.

Previous analyses, based on phylogenomic datasets, have suggested that profile mixtures may be more robust to long branch attractions. To check that the profile mixtures developed here have similar properties, we analyzed a series of phylogenomic datasets taken from previously published phylogenomic studies (Brinkmann *et al.*, 2005; Philippe *et al.*, 2005). These datasets have been built by concatenating single gene alignments, and display well documented phylogenetic artifacts. Based on the literature, we can thus define for each dataset two alternative topologies: one, obtained using most currently available phylogenetic models, including WAG, and likely to be artifactual; and another one, thought to be closer to the true topology. In the following, we will refer to them directly as the 'artifactual' and the 'true' topology (see Supplementary Information).

We first compared the models. As can be seen from Table 2, all phylogenomic datasets display a strong preference for profile mixtures (C20 and C50), over WAG_HSSP ($p < 10^{-4}$ for the Kishino Hasegawa test in all cases). They also have a high saturation index

Table 3. Topology comparison on genomic data sets. $\Delta \ln L$ is between true and artifactual topologies, under the specified model

	WAG_HSSP		C20		C50	
	$\Delta \ln L$	p	$\Delta \ln L$	p	$\Delta \ln L$	p
Metazoans (1)	-27	0.26	14	0.29	16	0.26
Metazoans (2)	-67	0.07	1	0.49	13	0.29
Nematodes	-106	0.07	-2	0.48	7	0.43
Microsporidia	-180	0.00	144	0.00	147	0.00
Nucleomorph	-9	0.43	62	0.00	63	0.00

(Table 2), further indicating the correlation between high saturation and good fit for profile models.

Next, using the two alternative 'true' and 'artifactual' trees, we performed a Kishino Hasegawa test, successively under WAG_HSSP, C20 and C50. In all cases (Table 3), the WAG_HSSP matrix prefers the artifactual trees, and the true tree is significantly rejected at the 0.05 level in one case (microsporidians). In contrast, C50 always prefers the true topology, with significant rejection of the alternative tree achieved in two cases (microsporidians and Guillardia's nucleomorph). In one case, namely the position of microsporidians, there is a complete reversion from a significant rejection of the 'true' tree by WAG, to a significant rejection of the 'artifactual' tree under C50. A basal position of microsporidia, as favored by WAG, is probably the result of a long branch attraction artifact (LBA) (Keeling and Fast, 2002; Brinkmann *et al.*, 2005). Consequently, the reversion observed under C50 can be interpreted as a case of suppression of the LBA by the use of a more adequate model. Concerning C20, the results are similar: for all datasets but one (nematodes), the true topology is preferred, which suggests that C20 is a sufficiently rich model for phylogenetic reconstruction.

5 DISCUSSION AND CONCLUSION

In this work, we have developed and tested an empirical profile mixture model. Our aim was two-fold. First, we wanted to transpose some aspects of earlier work initially conducted in a Bayesian framework within a Maximum Likelihood paradigm. But also, we wanted to dispense with the need of inferring all the parameters of the profile mixture directly on the dataset under investigation, as it precludes the analysis of small alignments. Accordingly, we tried to provide phylogeneticists with a model whose parameters are pre-defined, and should be suitable for a large spectrum of alignments of phylogenetic interest.

Our analysis indicates that a moderate number of profiles, around 20, is enough to accomplish essential improvements in two directions, compared to standard amino-acid replacement matrices. First, it provides a better fit in many cases, and second, it reconstructs more reliable phylogenetic trees, with a lesser sensitivity to long-branch attraction artifacts. In particular, the striking reversion observed in the case of microsporidians illustrates how relaxing the assumption of substitutional homogeneity along the sequence can dramatically alleviate systematic errors in phylogenetic reconstruction. Similar results have been obtained previously in a Bayesian context, based on a more complex non-parametric profile mixture model. In this respect, the present work shows that those results are robust, with respect to the details of the model, the implementation, and the statistical framework.

We also observe that, in a non negligible fraction of test alignments, the empirical matrix remains the best alternative. In a theoretical perspective, this could mean that, if profile mixtures account for some fundamental aspects of the amino-acid replacement patterns that empirical matrices fail at capturing, on the other hand, they seem to miss something that these matrices correctly describe. As was previously suggested (Lartillot *et al.*, 2007), and further confirmed in the present analysis, what profile mixtures are good at modelling is the fact that some sites undergo repeated replacements among very small subsets of the amino-acid alphabet, something which is not easily encoded into one single global matrix. Conversely, we observe here that empirical matrices still better describe the amino-acid replacement process at shorter evolutionary distance. This suggests that mixtures of matrices could be even better than mixtures of profiles, provided that the equilibrium frequency vectors of some of those matrices are concentrated on small subsets of amino-acids.

Note that progress is still possible in the way empirical matrices are estimated. In this direction, a better empirical matrix was recently obtained by Le and Gascuel (2008). In the present context, we observe that it provides a significantly better fit than WAG_HSSP in most cases (averaged AIC score difference of 0.10 over Tree-Base). However, the essential result obtained here, namely, that empirical profile models are better on saturated data, is not fundamentally changed if this new matrix is used instead of WAG_HSSP.

In a more practical perspective, all these observations raise the question of how to decide for the best model in individual cases. A straightforward method would consist in systematically running phylogenetic analyses under the two models, WAG_HSSP and C20, in parallel, and then choosing the best model on a case by case basis. As we have shown above, a simple alternative to this systematic duplication would consist in computing the saturation index of the dataset based on a maximum parsimony reconstruction, and choose C20 whenever this saturation index is too high. Otherwise, WAG_HSSP, or another empirical matrix, should be used.

An important feature of the profile mixtures is their overall complexity. We did not investigate mixtures with more than 60 profiles, yet the likelihood scores indicate that higher dimensional mixtures would still bring further improvement. On the other hand, the resulting profiles would probably display some redundancy, as is already observed under C60. A possible explanation of this phenomenon is that the variations across sites form a continuum, imperfectly described by a finite mixture. In such a situation, arbitrarily rich mixtures can be favored given a sufficient amount of data, and the decision to stop has to be made based on a tradeoff between tractability and usefulness (Steel, 2005). In the present case, the C20 model represents such a pragmatic tradeoff, representing about as many parameters as 2 time-reversible empirical matrices, while achieving a substantial improvement over one matrix.

The profile mixture models developed in this paper are suitable for both ML and Bayesian analyses. Accordingly, we have developed them in the two frameworks in parallel, by simultaneously implementing them in two phylogenetic reconstruction programs, PhyML (Guindon and Gascuel, 2003) and PhyloBayes (<http://atgc.lirmm.fr/cat>). In the ML framework, C20 probably provides the best practical compromise between fit and computational efficiency. Thanks to a recoding trick (possible only for F81 processes), the C20 model represents an overall 4 to 5 fold increase,

both in computational time and memory requirements, instead of the 20-fold increase that would be obtained without the recoding.

In the Bayesian context, any model from C20 to C60 can be used, with only marginal differences in the resulting computational load, thanks to Monte Carlo parameter expansion methods. Empirical mixtures are especially suited to phylogenetic analyses using small alignments (e.g. single gene analyses), whereas large analyses should preferably be conducted under the more flexible non-parametric CAT model developed previously.

Altogether, our two implementations make it possible to perform extensive comparisons between matrix and vector models, under two distinct statistical frameworks, and using both single-gene alignments or large concatenations. Such a generalized cross-talk between alternative models and methods will greatly enlarge the experimental perspective, and will hopefully provide important clues as to what further improvements are still needed for converging towards more reliable phylogenetic trees.

ACKNOWLEDGEMENT AND FUNDING

We wish to thank Nicolas Rodrigue and the anonymous referees for their useful comments on the manuscript. This work was financially supported by the ANR-BIOSYS MITOSYS project and by the ACI-IMPBIO ModelPhylo project.

REFERENCES

- Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, **42**, 459–468.
- Adachi, J., Waddell, P. J., Martin, W., and Hasegawa, M. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J. Mol. Evol.*, **50**, 348–358.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **AC-19**(6), 716–723.
- Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.*, **2**, 1152–1174.
- Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., and Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.*, **54**, 743–757.
- Bruno, W. J. (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol. Biol. Evol.*, **13**, 1368–74.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignment for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
- Crooks, G. E. and Brenner, S. E. (2005). An alternative model of amino-acid replacement. *Bioinformatics*, **21**, 975–980.
- Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In M. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, pages 345–352. National Biomedical Research Foundation, Washington, DC.
- Dempster, A., Laird, N., and Rubin, D. (1997). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.
- Dufayard, J. F., Duret, L., Penel, S., Gouy, M., F. R., and G., P. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**, 2596–2603.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
- Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, **13**, 93–104.
- Ferguson, T. (1973). A bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, 209–230.
- Gascuel, O. and Guindon, S. (2007). Modelling the variability of evolutionary processes. In O. Gascuel and M. Steels, editors, *Reconstructing Evolution: new mathematical and computational advances*, pages 65–99. Oxford University Press.
- Goldman, N., Thorne, J. L., and Jones, D. T. (1996). Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, **263**(2), 196–208.
- Goldman, N., Thorne, J., and Jones, D. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.
- Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
- Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol.*, **15**, 910–917.
- Holmes, I. and Rubin, G. M. (2002). An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.*, **317**, 753–764.
- Hordijk, W. and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics*, **21**, 4338–4347.
- Huelsenbeck, J. P. and Suchard, M. A. (2007). A nonparametric method for accommodating and testing across-site rate variation. *Syst. Biol.*, **56**, 975–987.
- Huelsenbeck, J. P., Jain, S., Frost, S. W., and Pond, S. L. (2006). A dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 6263–6268.
- Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.
- Keeling, P. J. and Fast, N. M. (2002). Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu. Rev. Microbiol.*, **59**, 93–116.
- Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topology from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.*, **29**, 170–179.
- Koshi, J. M. and Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins*, **32**, 289–295.
- Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.
- Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, **55**, 195–207.
- Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppressing long branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.*, **7**, S4.
- Le, S. Q. and Gascuel, O. (2008). An improved general amino-acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, **9**, 249–265.
- Nielsen, R. (2002). Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.
- Pagel, M. and Meade, A. (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, **53**, 561–581.
- Philippe, H., Lartillot, N., and Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecysozoa, Lophotrochozoa and Protostomia. *Mol. Biol. Evol.*, **22**, 1246–1253.
- Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
- Sanderson, M. J., Baldwin, B. G., Bharathan, G., Campbell, C. S., Ferguson, D., Porter, J. M., Von Dohlen, C., Wojciechowski, M. F., and Donoghue, M. J. (1993). The growth of phylogenetic information and the need for a phylogenetic database. *Syst. Biol.*, **42**, 562–568.
- Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**(2), 461–464.
- Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, **16**, 1114–1116.
- Steel, M. (2005). Should phylogenetic models be trying to 'fit an elephant'? *Trends in Genetics*, **21**, 310–311.
- Thorne, J. L., Goldman, N., and Jones, D. T. (1996). Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, **13**, 666–673.
- Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
- Whelan, S., Lio, P., and Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet.*, **17**, 262–272.
- Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, **39**, 306–314.
- Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino-acid sites. *Genetics*, **155**, 431–449.