Improving Phylogenetic Inference with a Semiempirical Amino Acid Substitution Model

Stefan Zoller*1,2 and Adrian Schneider3
1Computational Biochemistry Research Group, ETH Zürich, Zürich, Switzerland
2Swiss Institute of Bioinformatics, Switzerland
3Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, The Netherlands
Corresponding author: E-mail: stefan.zoller@inf.ethz.ch.
Associate editor: Jeffrey Thorne

Abstract

Amino acid substitution matrices describe the rates by which amino acids are replaced during evolution. In contrast to nucleotide or codon models, amino acid substitution matrices are in general parameterless and empirically estimated, probably because there is no obvious parametrization for amino acid substitutions. Principal component analysis has previously been used to improve codon substitution models by empirically finding the most relevant parameters. Here, we apply the same method to amino acid substitution matrices, leading to a semiempirical substitution model that can adjust the transition rates to the protein sequences under investigation. Our new model almost invariably achieves the best likelihood values in large-scale comparisons with established amino acid substitution models (JTT, WAG, and LG). In particular for longer alignments, these likelihood gains are considerably larger than what could be expected from simply having more parameters. The application of our model differs from that of mixture models (such as UL2 or UL3), as we optimize one rate matrix per alignment, whereas mixture models apply the variation per alignments site. This makes our model computationally more efficient, while the performance is comparable to that of UL3. Applied to the phylogenetic problem of the origin of placental mammals, our new model and the UL3 mixed model are the only ones of the tested models that cluster Afrotheria and Xenarthra into a clade called Atlantogenata, which would be in correspondence with recent findings using more sophisticated phylogenetic methods.

Key words: amino acid substitution model, Markov model, principal component analysis, semiempirical substitution model, origin of mammals, Atlantogenata.

Introduction

Probabilistic models of amino acid replacement are an essential component for a wide range of evolutionary protein analyses, such as sequence alignments, tree inferences, ancestral reconstructions, or structure predictions. Typically, they are Markov models, defined by a 20 × 20 rate matrix Q from which the substitution probabilities P for an evolutionary distance t can be derived by P(t) = exp(tQ). Amino acid substitution models are traditionally empirical models, where the rate matrix Q has been estimated once from a large data set and is then used for all subsequent analyses. Only the amino acid frequencies are normally estimated again from the specific data that is analyzed.

The first amino acid replacement model was presented as early as 1978 based on observed substitutions in 71 protein groups (Dayhoff et al. 1978). Since then, the models have continuously been improved, be it by using larger sequence databases (Gonnet et al. 1992; Jones et al. 1992) or by employing maximum-likelihood (ML) estimation and an expectation maximization algorithm to estimate the rate matrix directly (Whelan and Goldman 2001). More recently, a rate matrix has been estimated under consideration of rate variation among sites, by dividing the sites of the training alignments into categories according to their relative evolutionary rates. This was shown to improve the performance of the model, especially when applied using rate variation, which is a standard option for current phylogenetic analyses (Le and Gascuel 2008).

Interestingly, the rate matrix of amino acid models is almost never parametrized, whereas this is common for codon and nucleotide models where the substitution rates can often be adjusted to the sequences of the individual analyses. Matrix parametrization can range from estimating only a couple of parameters from which the complete rate matrix is derived (e.g., transition/transversion and nonsynonymous/synonymous rate ratios in the Goldman and Yang [1994] codon model) to all substitution rates being estimated from the data. The latter is called general time-reversible model (GTR), and, because of the high number of parameters, it is normally only used in models for nucleotide substitutions (Tavaré 1986). It is not completely clear why all widely used amino acid models are empirical. Certainly, there is a lack of obvious parametrization (as opposed to the rather evident parameters mentioned earlier for codon models). Also, empirical models have been used successfully for more than three decades.

Although the rate matrices of amino acid models are normally empirical, there are other free parameters used to
model the evolution of proteins. The most common examples are models that assume a distribution of rates among sites (typically a discrete gamma distribution with shape parameter $\alpha$, Yang 1994) or models that include an estimate of the proportion of invariant sites (Shoemaker and Fitch 1989). Furthermore, models have been proposed which assume different amino acid distributions or substitution matrices at different sites or classes of sites (Thorne et al. 1996; Halpern and Bruno 1998; Koshi and Goldstein 1998) to reflect structural and functional heterogeneity along proteins. Le et al. (2008) presented so-called mixture models, where, similar to rate variation among sites, each site is modeled as a mixture of several substitution processes. These processes had been estimated empirically by dividing the training alignments into different site categories (e.g., exposed vs. buried sites) or by letting an unsupervised learning algorithm find the optimal site partitions (two or three site categories for the UL2 or UL3 model, respectively).

Although many of the existing models somehow parametrize the amino acid substitution process, they do not explicitly parametrize the relative substitution rates, as it is common for nucleotide or codon substitution model. In this study, we present a method based on principal component analysis (PCA) to find the most relevant parameters for an amino acid substitution model. The obtained principal components (PCs) describe the substitution rates that covary the most among different protein families. Thus, the PCs are an empirically determined parametrization for an amino acid substitution model. We have previously applied a similar analysis to codon substitution models and thereby not only confirmed well-known codon substitution parameters but also found new parameters (Zoller and Schneider 2010). Using those results, we proposed a semiempirical codon substitution model, which we integrated into phylogeny software (Zoller and Schneider 2012).

Although seemingly similar, there is a fundamental difference between the type of model presented herein and mixture models. Although both model types are based on a set of pre-estimated empirical matrices, mixture models optimize mixture probabilities that are applied to each site individually (i.e., for each site, the likelihood under different models is computed and then averaged), and our model optimizes one substitution rate matrix as a combination of several matrices and applies this single matrix to all sites. In other words, mixture models tackle the variation in the substitution process across sites, whereas semiempirical models as this one are concerned with the variation of the process across genes. Thus, the two approaches are almost orthogonal and could in principle also be combined.

Recently, a non-negative matrix factorization (NNMF) algorithm was used to create a semiempirical amino acid substitution model (Murrell et al. 2011). Although the goal and the methodology is very similar to the PCA-based model proposed here, there are technical differences that also impact the performance of the model. Thus, we include a detailed discussion of the differences between the PCA and the NNMF methods for constructing a semiempirical substitution model, together with extensive performance analyses of those two and other amino acid substitution models.

In a Bayesian framework, much more parameter-rich models are thinkable. A prominent example is the CAT model family by Lartillot and Philippe (2004), where different site categories are modeled to have different equilibrium frequencies but the same relative substitution rates. The relative rates (exchangeabilities) are either taken from existing empirical models (e.g., CAT-JTT) or estimated from the data (CAT-GTR model). In this article, however, we will focus on the ML framework and models formulated therein.

Materials and Methods

Data Sets

Multiple sequence alignments (MSAs) were extracted from two different databases: from Pandit (Whelan et al. 2003) we created a training data set to estimate the model components and from the Orthologous Matrix (OMA) project (Dessimoz et al. 2005) we obtained testing data sets.

The Pandit database contains MSAs of homologous sequences covering a wide range of species. To construct the training set, we first discarded all MSAs with less than 4 sequences or less than 100 amino acid sites. The sites of the Pandit MSAs are annotated with a quality score; we used this information to exclude alignments if less than 60% of the sites are marked as reliable. Next, we applied “Gblocks” (Castresana 2000) with default options to filter sites that contain too many gaps. Finally, we discarded all MSAs that contained fewer than 50 sites, as they are unlikely to hold enough information to allow for a reliable estimation of substitution rates. This procedure left 2,752 MSAs (which we call the “Pandit set”) and was used to estimate the input matrices for our PCA model (see later). It includes approximately 71% of all training MSAs used for the LG model (Le and Gascuel 2008).

OMA is a database for orthologous genes from more than 1,200 species. The orthology property is not necessary for our purposes, but OMA is a convenient source of diverse clusters of homologous protein sequences to create MSAs. We extracted orthologous groups separately for each domain of life (Archaea, Bacteria, and Eukaryota), each group containing sequences from between 5 and 20 different species. The sequences of each group were aligned with “mafft” (Katoh et al. 2005) using the default options. As with the Pandit set, we applied Gblocks with default options and discarded MSAs with less than 50 sites left after this procedure. We sampled 1,500 MSAs from each domain, thus we obtained a set of 4,500 diverse MSAs, which we called the “OMA set.” It is used as testing data set for the models. The OMA set is an ideal testing data set, because none of the models was estimated from OMA alignments. Thus, although the different models were trained on different data set, the comparison presented here is performed on a neutral data set.

The alignments in the OMA set contain less distant species than the alignments in the Pandit set, but the sequences are longer in the average. In the Pandit set, the mean sequence
length is 156 with a standard deviation of 113, whereas in the OMA set, we find a mean sequence length of 345 with a standard deviation of 249. In the OMA set, we count on average 10 species per MSA with a variance of 19, whereas in the Pandit set, the average is 11 with a variance of 82. As for the sequence diversity, the OMA sequences within groups show an average pairwise distance of 61 PAM with a variance of 1,466, whereas in the Pandit set, the average pairwise distance is 89 PAM with a variance of 2,239.

To test the amino acid models on a well-known phylogenetic problem, the origin of placental mammals, we constructed a concatenated protein sequence alignment of the following nine species: opossum and the Tasmanian devil as outgroups; the African elephant and tenrec representing the superorder Afrotheria; armadillo and sloth from Xenarthra; and dog (Laurasiatheria) and human and mouse (Euarchontoglires) from the superorder Boreotheria. The orthologs were extracted from OMA (Dessimoz et al. 2005). All 1,426 orthologous groups containing proteins from all nine species have been selected. We used mafft (Katoh et al. 2005) with default options to create an MSA from each orthologous group separately. Unreliable regions were removed using Gblocks, also using the default options. Finally, we concatenated all alignments to one large MSA with a total of 1,044,566 amino acid sites.

The PCA Amino Acid Model

The general time-reversible PCA amino acid model (PCMA) presented here was constructed in a similar manner as the PCA codon model (PCM) we presented previously (Zoller and Schneider 2012). Besides the natural differences between codon and amino acid data, there are also changes in the data preparation step. We followed the approach presented by Le and Gascuel (2008) who introduced the idea of splitting the alignments at the estimation step of the rate matrices by substitution rate categories and accordingly scaled trees: for each MSA, we categorized each site into one of four rate classes using the output of “PhyML” (Guindon et al. 2009) run with default options and LG as substitution matrix (Le and Gascuel 2008). We then separated the sites according to the assigned classes, ending up with four shorter MSAs per original alignment, each containing the sites for one distinct rate category. The estimated tree for the original alignment was scaled by each of the four mean rates and stored together with the respective partial alignment. This lead to 11,008 new alignments, each with its scaled tree.

Next, “XRRate” (Klosterman et al. 2006) was used to estimate a single initial time-reversible (symmetric) exchangeability matrix from all 11,008 partial MSAs jointly. We used 0.0001 as the minimum fractional increase in log likelihood per round of expectation–maximization (EM) and set the number of consecutive nonincreasing rounds of EM to 3. The exchangeability matrix R together with the frequency π define the rate matrix Q by the equation Qij = πjRij, where the diagonal elements are determined by the basic property of a rate matrix that the row sum up to 0. The thus estimated preliminary rate matrix served as a starting point to estimate 2,752 new exchangeability matrices, one from each quadruple of partial alignments. Splitting up the original alignments into four rate-dependent alignments with correspondingly scaled trees is expected to result in better rate estimates than using the simpler approach of estimating the matrices directly from the original MSAs (Le and Gascuel 2008).

The final step was the PCA on these 2,752 exchangeability matrices, which were represented as vectors of length 190 (half the number of all off-diagonal elements, as the matrix is symmetric). The PCA resulted in 190 PCs P(k), k = 1...190, sorted by corresponding eigenvalues (largest first), a scale matrix S (to normalize the variances of the input values), and a mean matrix M (to center the input vectors). The PCs, S, and M obtained from the PCA were all vectors of 190 elements but were back transformed to symmetric 20 matrices with undefined diagonal elements (as they are not needed). All PCs are orthogonal to each other and describe the correlations of the variance in the input data. Thus, the first PC corresponds to the combination of exchangeability parameters that covaries the most among the input matrices, and the second PC corresponds to the second most covarying combination and so on.

Using the results of the PCA, the amino acid model PCMA is constructed analogous to the PCM codon model by reversing the PCA. The exchangeability matrix R is a linear combination of the mean matrix M and the PCs P(k), with the coefficients ck being free parameters to be estimated from the data. The matrix R together with the equilibrium frequencies (typically estimated from the data) define the rate matrix Q. The elements Rij are computed as follows:

\[ R_{ij} = \max\left(0, M_{ij} + S_{ij} \sum_k c_k p_{ij}^{(k)} \right) \quad \forall i \neq j \]  

(1)

From the resulting symmetric exchangeability matrix R, the rate matrix Q can be derived as described earlier. After normalization of Q, the substitution probability matrix P(t) for the underlying Markov process at distance t can be computed as \( P(t) = e^{tR} \). We denote an instance of PCMA with n PCs (and n free parameters) as PCMA+nC.

For every component added to the model, one additional parameter has to be optimized. PCs are orthogonal to each other by definition, hence we expect the optimization to be quite efficient, even with larger numbers of parameters. We implemented our model in PhyML and, therefore, use its built-in optimization code; the additional free parameters are treated the same way as existing model parameters such as the proportion of invariant sites or the alpha parameter of the gamma distribution that models rate variation across sites. The results from the test alignments indicate that optimization time increases approximately linearly with the number of PCs used. Given a parameter combination \( c_1, \ldots, c_n \), one rate matrix is computed (following eq. 1) and then used for the computation of the likelihood for all sites. Thus, the optimization of the branch lengths is not more expensive than under a fixed-matrix model such as LG; only the additional parameters from the model have to be optimized. This is in contrast to mixture models, where on
top of the mixture parameter optimization, the likelihood computation itself becomes more expensive, because for each site, the likelihood has to be computed using each matrix separately.

Comparison of Substitution Models

Comparing models with different numbers of parameters cannot be done directly on likelihoods, as parameter-rich models would have an unfair advantage by having more degrees of freedom. Especially for small sample sizes, there is a risk of overfitting the data, leading to skewed results. Thus, we measured the goodness of the fit of a substitution model to the testing data with the corrected Akaike Information Criterion (AICc; Hurvich and Tsai 1989). The original Akaike Information Criterion (AIC; Akaike 1974) describes how good a model fits the data, based on the likelihood $L$ of the fit and the number of free model parameters $k$:

$$\text{AIC} = 2k - 2\ln(L)$$

The model with the lowest AIC value describes the data the best. Models that use a higher number of free parameters receive a higher penalty on their score. Differences of up to 7 are considered as "weak," whereas differences larger than 7 show "strong support" for one model over the other (Burnham and Anderson 2002).

Hurvich and Tsai showed that the original AIC value contains a bias toward more complex models in situations where the sample size is relatively small. Thus, they proposed to extend the AIC with a second-order term to construct a AICc:

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1}$$

where $n$ denotes the sample size. In our analysis of trees over MSAs without gaps, this sample size corresponds to the length (number of sites) of the MSA.

Note that in the AICc, higher numbers of free model parameters lead to more severe penalties than in the original AIC. Furthermore, only the difference of free parameters has an influence on the original AIC, but because of the quadratic nature of the additional correction term in AICc, the absolute number of free parameters increases in importance. If the number of parameters exceeds the sample size, the AICc becomes ill defined. Therefore, we had to discard some combinations of models and MSAs from the comparison.

We used PhyML with default options to construct trees for each of the testing MSAs under all models in question. The topologies were fixed to the ML trees under the Whelan and Goldman (WAG) substitution model (Whelan and Goldman 2001), but we let the branch lengths of the ML trees be optimized under the respective models. The number of branches of an unrooted tree with $t$ taxa is $b = 2t - 3$. A further two free parameters were the proportion of invariable sites and the $\alpha$ parameter of a discrete gamma distribution with four rate categories (Yang 1994). The amino acid frequencies can either be estimated from the data (called +F variant of a model) or the original, fixed frequencies from the model can be used (the $-F$ variant). If the frequencies are estimated, 19 extra parameters have to be counted in the model comparison. Although we avoid direct comparisons of models with $-F$ against models with $+F$, the AICc calculations still depend on the frequency parameters due to their nonlinear influence. The total number of free parameters to calculate the AICc values for each tree under a $+F$ model is $b + 2 + 19 + s$, where $s$ denotes the number of free parameters of the substitution matrix. For example, the ML tree estimation for an MSA with 10 species under the LG+F model would lead to $(2 \times 10 - 3) + 2 + 19 + 0 = 38$ free parameters for the calculation of the AICc value.

We compared the PCMA model to the nonparametric JTT, WAG, and LG models and to several instances of the NNMF model (Murrell et al. 2011). The NNMF model (Murrell et al. 2011) was presented as a linear combination of a variable number of predefined factor matrices with the coefficients of those matrices being free parameters. A specific instance with $n$ factor matrices has $n - 1$ free parameters (because they can be normalized) and will here be denoted as NNMF+$n$. M. Murrell et al. propose to use a large range of $n$ and then select the best-fitting instance based on AICc. They suggest to count one additional free parameter for selecting the best instance, to obtain a fair AICc value for the NNMF model. However, we do not think that this method is correct, because one optimizes for all instance from 1 to $n$ not only $n$ parameters but also $1 + 2 + \cdots + n$ parameters. Therefore, we use only one instance of NNMF in the comparisons, namely NNMF+10M. Only for the direct comparison of PCMA and NNMF, we treat the models equally by choosing the best-fitting instances (in terms of AICc) from 40 instances of each model (NNMF+1M up to NNMF+40M and PCMA+0C up to PCMA+39C) and then comparing those two best instances, again in terms of AICc. The model frequencies for NNMF are not specified, thus we can only use the $+F$ variant for this model.

We also included the mixture models UL2 and UL3 in the comparisons by using a special version of PhyML, called "PhyML-mix" (Le et al. 2008). During the training step, these two models were allowed to find the best site partition of the sequence data and thus should be at least as good as comparable models with predefined site partitions (and equal numbers of parameters), such as EX2 or EX3. Obviously, using a priori information of the protein structure would not allow for a fair comparison, and thus, it is not done here. For UL2 and UL3, as for the other models, the estimation included the $\alpha$ parameter for a discrete gamma distribution with four rate categories and the proportion of invariant sites. Unfortunately, PhyML-mix does not allow for the estimation of frequencies from the data. Therefore, UL2 and UL3 had to be applied using the built-in frequencies of each model ($-F$) and thus cannot be compared directly to NNMF. For the mixture proportions, UL2 and UL3 include one and two free parameters, respectively.

Application on a Phylogenetic Problem

On the concatenated alignment of nine mammals, PhyML was run with the same options as described earlier using JTT,
WAG, LG, UL2, UL3, and several instances of PCMA and NNMF as substitution models. First, under each model, we globally optimized tree topology, branch lengths, and model parameters using the concatenated alignment. This resulted in only two different topologies for all models. Then, we only optimized the branch lengths and model parameter values for each model on the two topologies, to obtain the likelihood values for both topologies under each model.

**Results**

**Interpreting the PCs**

The obtained PCs are the result of unsupervised learning and thus could in principle come up with any possible combination of rates. However, it can be expected that at least part of the signal would be biologically meaningful and that it could be interpreted as a specific feature of amino acids. Therefore, we attempted to find an interpretation of the first few PCs by measuring their correlation with predefined feature vectors, which describe well-known or potential parameters. This has been done previously for codon and amino acid substitution models (Murrell et al. 2011; Zoller and Schneider 2012).

To this end, we used the 544 amino acid properties defined by the AAindex (Kawashima et al. 2008). From those 20-element vectors, we generated feature vectors with 190 elements that encode the upper triangular half of a $20 \times 20$ change matrix. Given two index elements $a, b$ for the amino acids A, B, the change from A to B has been calculated as $|a - b|$, to obtain the differences between the amino acid properties. Measuring the Spearman rank correlation between these vectors and the first few PCs allowed for some interpretation of the components with respect to these properties. Figure 1 shows some of the correlations: the size of the discs corresponds to the Spearman rank correlation coefficient, whereas the color informs about the significance of the correlation. For each feature vector, we used 50,000 random vectors to estimate the distribution of random correlation coefficients to assess the significance of the obtained correlations: discs in blue stand for coefficients that are not distinguishable from the random noise on the 99% level, whereas red discs show a significantly higher correlation. AAindex currently contains 544 different indices; a complete list of all significant correlations with the first 20 PCs can be found in supplementary material, Supplementary Material online. The features shown in figure 1 have been chosen because they show strong correlations with more than one of the first six PCs. “Hydration,” “Hydrophobic Parameter,” “Hydrophobicity,” “Polarity,” and “Sidechain Hydrophobicity” correspond to accession numbers HOPA770101, LEVM760101, PRAM900101, ZIMJ680103, and BLAS910101, respectively. “Minimal Nucleotide Change” encodes the minimal number of nucleotides needed to mutate from one amino acid to another.

The strongest correlation can be found between PC1 and the feature vector that encodes the minimum number of nucleotide changes to mutate from one amino acid to another; the Spearman coefficient for this feature vector

![Figure 1](https://academic.oup.com/mbe/article-abstract/30/2/469/1014942/1014942)
is 0.63, whereas the next strongest correlation for this PC amounts to only 0.24 (for the side chain hydropathy). This indicates that the largest variation in amino acid substitution models is driven by the genetic code. Interestingly, this signal also reappears a little bit weaker in PC3 and, still weaker, in PCs.

Another important feature is the notion of hydropathy. As can be seen in the complete table in the supplementary material, Supplementary Material online, all the first few PCs except for PC1 correlate mostly with features in connection with hydropathy or polarity. Although those indices are expected to be somewhat correlated among themselves, they still encode different features; it can be seen that all the listed features have significant matches with a different PC or combination of PCs.

Very likely, there are other signals encoded in the PCs that are not captured with the feature vectors used. The aim of this correlation analysis is only to give an interpretation to some of the signal found by the unsupervised PCA algorithm that empirically identified the most important parameters for an amino acid substitution model.

### Distribution of PCMA Coefficients

Another indication of biological relevance of the PCs could be found by studying their influence when applied to different alignments. As described in Materials and Methods, PCMA+\(n\)C uses a linear combination of the first \(n\) PCs, which are multiplied by coefficients \(c_i\) (the free parameters of the model). The contribution of the PCs to adapt to specific protein alignments would be indicated by varying coefficient estimates across different alignments. Table 1 lists mean and variances for the first 10 coefficients collected from the optimization on the OMA test set. The means are relatively close to zero and the variances considerably larger, indicating that all the coefficients vary substantially, both to the positive and to the negative. This indicates that the substitution process among the gene groups analyzed differs substantially and that the PCMA model is able to adjust to the particularities by choosing different coefficients for the different alignments.

### Comparison of Selected Amino Acid Models

We compared the performance of several amino acid substitution models on the OMA test data set of 4,500 MSAs.

In a first analysis, we compared the performance of a selection of six amino acid models (WAG and LG; UL2 and UL3; and PCMA+1C and PCMA+10C) against each other. All models were applied with the \(-F\) variant, thus NNMF could not be included in this comparison. Performance was measured by the AICc (Hurvich and Tsai 1989): for each model, we counted the number of times this model obtained the lowest AICc value for an MSA. In this first comparison, we counted only the best-fitting model per MSA. The results of these comparisons are shown in figure 2. The MSAs were grouped into bins depending on alignment length and the plots show the distribution of best models per length bin. The last bin contains all MSAs with sequence lengths 1,450 and larger.

For short alignments, models with no parameters, in particular LG, dominate. However, the longer the alignments, the more often parameter-rich models, especially PCMA+10C achieve the best AICc values. Among similar models, more parameters are almost always beneficial (UL3 favored over UL2, and PCMA+10C over PCMA+1C), only for short alignments can the model variants with less parameters sometimes achieve better AICc values. It also has to be noted that PCMA+10C almost always reaches the highest likelihood values among the tested models, but the AICc cost for 10 extra parameters is often too large on small alignments to be compensated by likelihood gains.

In figure 3, we show selected direct comparisons between two model instances at a time. The lower part of the bars denotes the number of MSAs with strong support for a model and the upper part in weaker color the additional number of MSAs with a lower AICc but only weak support (difference <7 AICc points, see Burnham and Anderson 2002). Figure 3a shows the comparison of PCMA+10C against LG, with a similar outcome as noted previously, that is, LG performing stronger on shorter alignments but PCMA+10C scoring better on longer ones (more than 350 sites). The direct comparison of PCMA and NNMF, the two models based on a similar principle, shows that PCMA is clearly superior, independent of alignment length (fig. 3b). For this comparison, we compared for each MSA the best instance of PCMA against the best instance of NNMF, similar to what was suggested in the NNMF article (Murrell et al. 2011). Finally, the last two panels of figure 3 show the performance of PCMA+10C against the two versions of the UL mixture models. Against UL2, PCMA performs clearly better except for the smallest alignments, where again the AICc cost for the extra parameters is too high. In addition, except on the shortest alignments, UL3 and PCMA+10C seem to be quite evenly matched. A full comparison of PCMA+1C and PCMA+10C against all other models and model variants can be found in the supplementary material, Supplementary Material online.

It should also be noted that all these models were estimated from different training data sets. This could potentially explain some of the performance differences, in addition to the inherently different methodologies. However, because the training data sets were always very large and often quite similar, the average substitution rates in the data are expected to be quite generic.

### Table 1. Mean and Variance of the First 10 Coefficients.

<table>
<thead>
<tr>
<th>Principal Component</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(c_1)</td>
<td>0.373</td>
<td>3.021</td>
</tr>
<tr>
<td>(c_2)</td>
<td>−0.293</td>
<td>3.809</td>
</tr>
<tr>
<td>(c_3)</td>
<td>0.052</td>
<td>2.242</td>
</tr>
<tr>
<td>(c_4)</td>
<td>0.658</td>
<td>3.684</td>
</tr>
<tr>
<td>(c_5)</td>
<td>0.049</td>
<td>1.808</td>
</tr>
<tr>
<td>(c_6)</td>
<td>1.216</td>
<td>3.884</td>
</tr>
<tr>
<td>(c_7)</td>
<td>−0.939</td>
<td>5.334</td>
</tr>
<tr>
<td>(c_8)</td>
<td>0.714</td>
<td>3.719</td>
</tr>
<tr>
<td>(c_9)</td>
<td>0.272</td>
<td>3.228</td>
</tr>
<tr>
<td>(c_{10})</td>
<td>0.572</td>
<td>2.991</td>
</tr>
</tbody>
</table>
FIG. 2. Model selection based on AICc values for selected models. For each bin of sequence lengths and for each model, the number of best AICc values per MSA and model are shown. The last bin contains all MSAs with 1,450 or more sites.

FIG. 3. Direct comparison of selected models. Each bin with mean value x summarizes alignments with x ± 50 sites. The first bin corresponds to the MSAs with 50–150 sites, and the last bin contains all MSAs with 1,450 or more sites. All comparisons were done on the –F variant, except for (b) where the +F variant was used. (a) LG vs. PCMA+10C; (b) PCMA vs. NNMF; (c) PCMA+10C vs. UL2; and (d) PCMA+10C vs. UL3.
Optimal Number of Parameters

Next, we analyzed the optimal number of matrices used (the \( n \) parameter) in the models PCMA and NNMF. Figure 4 gives an overview of how the optimal number of parameters (i.e., with the lowest AICc value) changes as a function of sequence length of the MSA. On longer alignments, where more information is available, more free parameters help to adapt better to the data. However, for shorter sequences, the extra parameters of instances such as PCMA +39C or NNMF +40M are not increasing the likelihood values enough to compensate for the penalty of the AICc calculation for using such a large number of free parameters. Interestingly, PCMA requires in general less parameters than NNMF on alignments of similar length. This is possibly because the PCs are sorted according to their relevance to the evolutionary signal, and thus, most variation in the data can be approximated with only a few PCs. In the NNMF model, on the other hand, the variation of the training data is evenly reflected in all matrix factors.

Analysis of Tree Lengths

As an additional test, we compared the length of trees (sum of all branch lengths) estimated using different model instances. Although this is not a real statistical criteria, it has been suggested before to assess the quality of substitution models (Le and Gascuel 2008). The main reasoning is that a better model should be superior in inferring ancestral states and multiple substitutions along the branches and thus obtain longer branch-length estimates (Pagel and Meade 2005).

Figure 5 shows the mean difference in tree length (in percent) for each instance of the PCMA and NNMF model family, compared with the tree lengths under LG. On the horizontal axis, the number of free model parameters is shown. The value for the parameterless model LG has been marked with a dot at 0, and a flat line is provided to alleviate the comparison to the different instances of PCMA and NNMF.

A first comparison of LG and the two semiempirical models shows that trees under NNMF are clearly smaller (11–17%) than under LG, but LG and PCMA produce trees of similar length; PCMA with fewer PCs (<8) causes smaller trees, and with more PCs, the trees become larger than under LG.

The values for NNMF show a clear correlation between the number of free model parameters and the estimated tree length: the more matrices are used, the longer the trees become. These results indicate that NNMF could even benefit from using more than 40 free parameters, but then, the risk of overfitting increases too. PCMA shows a different behavior. Up to PCMA +6C, the tree length is relatively constant. Next, we observe a strong increase in tree length between PCMA +6C and PCMA +10C from −4% to +3.25% compared with the tree length under LG. Finally, from PCMA +22C on, the tree length declines. These results suggest an optimal number of parameters of between 10 and 22, which is in accordance with the findings based on AICc (shown in fig. 4).

Phylogenetic Application: The Origin of Placental Mammals

We applied our new model to a phylogenetic problem that is still somewhat debated, namely the origin of placental mammals. The problem centers around the branching order of three major clades: Afrotheria (mammals living in Africa, in our study represented by elephant and tenrec), Xenarthra (mammals that today are only native to South and Central America, here represented by armadillo and sloth), and a clade called Boreotheria, which combines Laurasiatheria (here the dog) with Euarchontoglires (here human and mouse). The traditional view based on morphology suggests an early branching of Xenarthra, leaving Afrotheria and Boreotheria as sister clades (Novacek 1992; Shoshani and McKenna 1998). However, some studies of molecular data sets sometimes obtain a different topology, where Afrotheria acts as an outgroup to Xenarthra and Boreotheria (shown in fig. 6a, e.g., Murphy et al. 2001; Nikolaev et al. 2007). Most recent studies, using more sophisticated phylogenetic approaches, often find a topology where Afrotheria and Xenarthra form a separate clade (called Atlantogenata) to the exclusion of Boreotheria.

Fig. 4. Optimal number of free parameters for a given sequence length, using the +F model variants of PCMA and NNMF. Each bin with mean value \( x \) summarizes alignments with \( x \pm 50 \) sites. The last bin contains all MSAs with more than 1,450 sites per sequence.
The analyses came to this conclusion based on ML trees from genome scale data sets (Hallström et al. 2007; Prasad et al. 2008), from examination of insertions of long interspersed nuclear element sequences (Waters et al. 2007), from looking at insertion/deletion patterns (Murphy et al. 2007), or from extracting signals by using different outgroups (Schneider and Cannarozzi 2009).

The ML tree reconstruction of the concatenated alignment of 1,426 proteins from 9 mammals was performed using JTT, WAG, LG, UL2, UL3, and various instances of PCMA and NNMF as described in the Materials and Methods section. It resulted in only two different topologies, shown in figure 6: almost all models (JTT, WAG, LG, UL2, and all NNMF instances and PCMA using >10 PCs) resulted in a topology where Afrotheria are an outgroup to Xenarthra and Boreotheria (Tree 1, shown in fig. 6a). Only for PCMA with 10 and more PCs and UL3 did the ML reconstruction prefer Tree 2, the Atlantogenata hypothesis (shown in fig. 6b).

We then estimated the likelihood for both topologies under all models (+F and –F variants, where available) by optimizing only the branch lengths and the substitution parameters. These results are summarized in table 2: for both frequency variants, the first column lists for each model the obtained log-likelihood value (ln $L$) on Tree 1, whereas the second column lists the difference in ln $L$ to Tree 2. Positive differences indicate support for Tree 2, and negative differences imply support for Tree 1.

UL3 and the PCMA variants with 10 or more free parameters are the only models to favor Tree 2. The PCMA instances with 10 or more PCs are also the models with the highest ln $L$ values. This shows that PCMA with enough parameters is able to find what is currently viewed as the correct tree, in a scenario where most other models fail to do so. Whether the frequencies are estimated or not does not make a difference in terms of which tree is selected. However, as can be expected, the likelihood always improves using the +F variant. Although this is only one example, it is a difficult phylogenetic problem where generally many more taxa are needed to obtain a good phylogenetic tree. It seems that the very long, concatenated alignment provides an advantage for a parametric model such PCMA, which can apparently better fit the substitution patterns between the sequences.

**Discussion**

Although it may seem at a first glance that NNMF and PCMA are very similar methods as both try to decompose empirically derived substitution matrices to obtain a parametric model, their performance is very different. Although NNMF does not seem to improve much over current models such as
Positive differences (in bold) indicate support for Tree 2, whereas negative differences indicate support for Tree 1.

Another important difference lies in the computation time, just adjusted to the problem, depending on the alignment size. The advantage that the number of parameters can be estimated with six mixture matrices (Le and Gascuel 2010).

Table 2. Log-Likelihood Values for Selected Models on the Two Trees Shown in Figure 6.

<table>
<thead>
<tr>
<th>Model</th>
<th>Tree 1</th>
<th>Tree 2</th>
<th>Tree 1</th>
<th>Tree 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(ln L)</td>
<td>(ln L)</td>
<td>(ln L)</td>
<td>(ln L)</td>
</tr>
<tr>
<td>JTT</td>
<td>-6,198,176</td>
<td>-25</td>
<td>-6,190,508</td>
<td>-20</td>
</tr>
<tr>
<td>LG</td>
<td>-6,279,157</td>
<td>-69</td>
<td>-6,278,379</td>
<td>-54</td>
</tr>
<tr>
<td>WAG</td>
<td>-6,286,539</td>
<td>-83</td>
<td>-6,271,739</td>
<td>-80</td>
</tr>
<tr>
<td>NNMF+1M</td>
<td>Not available</td>
<td></td>
<td>-6,326,840</td>
<td>-89</td>
</tr>
<tr>
<td>NNMF+10M</td>
<td>-6,277,456</td>
<td></td>
<td>-6,240,801</td>
<td>-52</td>
</tr>
<tr>
<td>NNMF+20M</td>
<td>-6,240,801</td>
<td></td>
<td>-6,277,456</td>
<td>-52</td>
</tr>
<tr>
<td>NNMF+30M</td>
<td>-6,223,263</td>
<td></td>
<td>-6,240,801</td>
<td>-56</td>
</tr>
<tr>
<td>PCMA+0C</td>
<td>-6,306,223</td>
<td>-81</td>
<td>-6,281,177</td>
<td>-66</td>
</tr>
<tr>
<td>PCMA+10C</td>
<td>-6,204,356</td>
<td>1</td>
<td>-6,178,526</td>
<td>18</td>
</tr>
<tr>
<td>PCMA+20C</td>
<td>-6,201,916</td>
<td>5</td>
<td>-6,174,755</td>
<td>12</td>
</tr>
<tr>
<td>PCMA+30C</td>
<td>-6,201,164</td>
<td>7</td>
<td>-6,174,339</td>
<td>12</td>
</tr>
<tr>
<td>UL2</td>
<td>-6,340,460</td>
<td>27</td>
<td>-6,314,711</td>
<td>17</td>
</tr>
<tr>
<td>UL3</td>
<td>-6,314,711</td>
<td>17</td>
<td>Not available</td>
<td></td>
</tr>
</tbody>
</table>

Note.—The value for Tree 1 denotes the obtained log likelihood (ln L), whereas the value for Tree 2 is the difference to the ln L value of Tree 1 under the same model. Positive differences (in bold) indicate support for Tree 2, whereas negative differences indicate support for Tree 1.

WAG and LG, PCMA consistently produces higher likelihood values than those models, and for longer alignments, these gains are significantly higher than what could be expected from simply having more parameters. The analyses based on tree lengths and phylogenetic reconstruction indicate that the improvement is not only in terms of likelihood values but that the PCMA model better captures the substitution process than other models.

The results of the comparison with the mixture models UL2 and UL3 are quite informative. Although most PCMA instances achieve higher AICc values than UL2, PCMA and UL3 seem to be closely matched. Also applied to the phylogenetic problem of the origins of the placental mammals, UL3 and PCMA with 10 or more PCs agree on the currently favored topology. Although the different nature of the two model types, site-wise variation versus gene-wise variation in the substitution process, makes it difficult to compare in a completely fair way, it seems that UL3 and PCMA+10C perform about equally in the tests in this study. PCMA also has the advantage that the number of parameters can be adjusted to the problem, depending on the alignment size. Another important difference lies in the computation time, both for estimating the models and when applying them. Although it is no problem to use more PCs (the PCA yields many more PCs that make sense to use), the estimation of mixture models is computationally more demanding: for example, it already took approximately 12 days on a cluster to estimate UL3 (Le et al. 2008). A possible approach to more efficiently estimate models with more mixture categories is using a priori information. For example, using secondary structure information, the EX_EHO model could be estimated with six mixture matrices (Le and Gascuel 2010). And very recently, evolutionary rate categories were used to estimate individual matrices, which replace the traditional rate factors for modeling a rate distribution and thus allow for matrix mixing without additionally increasing the computation time (LG4M and LG4X models, Le et al. 2012).

The difference in computation time when applying the models can be demonstrated on the large phylogenetic data set of approximately 1 million amino acids that we used to analyze the origin of placental mammals: PCMA+10C used 4.5 h of CPU time, whereas the time for UL3 amounted to more than 7 days to optimize the branch lengths and the model parameters (although this difference is likely exaggerated due to a less efficient implementation of PhyML-mix, it would still be large using the same implementation). Under the nonparametric LG model, this optimization took a little more than 2 h.

There are several factors that could explain the strength of the PCMA model. First, the training matrices are estimated using state-of-the-art methodology, including EM of the exchangeability parameters and matrix estimation under consideration of rate variation among sites. Furthermore, PCA seems to be the ideal method to find the relevant factors, because the PCs are ranked such that the first PC explains the largest part of the variation and the lower ranked PCs encode less important factors of protein evolution, until we find only noise. This feature makes it possible to separate signal from noise, as opposed to the NNMF method, where the noise is also incorporated into the n factors that had been chosen for the matrix decomposition.

A possible difficulty when using PCMA is the optimal number of PCs to choose. The best option is to use PCMA by testing a range of instances and then selecting the best one (based on AICc value). However, this would often be impractical, as it requires several runs of the phylogeny software and the computation of AICc values. We have shown that a single instance (e.g., PCMA+10C; fig. 2) performs already reasonably well. Using the observations of how the optimal number of parameters depends on the length of the alignments (fig. 4), the following configurations should yield good results when using only one model instance: for MSAs of up to 350 sites, PCMA+1C or PCMA+2C seem optimal. For MSAs with between 350 and 1,000 sites, PCMA+10C can be used, and for longer alignments, PCMA+20C is appropriate. For large phylogenetic analyses where the best topology is the main focus, it would still be worthwhile to run different instances of PCMA to find the absolute best configuration of the model.

Considering the results presented in this study, which show that phylogenetic inference can be improved using a semiempirical amino acid substitution model, this seems a promising direction for future analyses. Most phylogeny software is already equipped with a mechanism to optimize parametric models, thus it should be feasible to integrate PCMA into current software packages. Our model implementation in PhyML is available at http://people.inf.ethz.ch/zollers/phymlpcma.html (last accessed 5 October 2012).

Supplementary Material

Supplementary material is available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).
Acknowledgments

The authors thank Olivier Gascuel and the two anonymous reviewers, as well as Jeffrey Thorne, the associate editor, for their input and help on improving the manuscript. They also thank Jaap Heringa for pointing them to the AAindex.

References


Shoemaker J, Fitch W. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol Biol Evol. 6:270–289.


