A New Semiempirical Codon Substitution Model Based on Principal Component Analysis of Mammalian Sequences

Stefan Zoller\textsuperscript{1,2} and Adrian Schneider*\textsuperscript{3}

\textsuperscript{1}Department of Computer Science, Eidgenössische Technische Hochschule Zurich, Zürich, Switzerland
\textsuperscript{2}Swiss Institute of Bioinformatics, Zürich, Switzerland
\textsuperscript{3}Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands

*Corresponding author: E-mail: schneadr@gmail.com.
Associate editor: Jeffrey Thorne

Abstract

Codon substitution models have traditionally been parametric Markov models, but recently, empirical and semiempirical models also have been proposed. Parametric codon models are typically based on $61 \times 61$ rate matrices that are derived from a small number of parameters. These parameters are rooted in experience and theoretical considerations and generally show good performance but are still relatively arbitrary. We have previously used principal component analysis (PCA) on data obtained from mammalian sequence alignments to empirically identify the most relevant parameters for codon substitution models, thereby confirming some commonly used parameters but also suggesting new ones. Here, we present a new semiempirical codon substitution model that is directly based on those PCA results. The substitution rate matrix is constructed from linear combinations of the first few (the most important) principal components with the coefficients being free model parameters. Thus, the model is not only based on empirical rates but also uses the empirically determined most relevant parameters for a codon model to adjust to the particularities of individual data sets. In comparisons against established parametric and semiempirical models, the new model consistently achieves the highest likelihood values when applied to sequences of vertebrates, which include the taxonomic class where the model was trained on.

Key words: Markov model, codon substitution model, principal component analysis.

Introduction

Markov models have been used to model sequence evolution for more than four decades (Jukes and Cantor 1969), and for almost two decades, they have also been employed to model codon substitutions (Goldman and Yang 1994; Muse and Gaut 1994). Current codon models are Markov models based on $61 \times 61$ rate matrices to generate substitution probabilities between codons. Substitutions to and from stop codons are typically not considered.

The differences among the models lie mostly in the generation of the rate matrix that defines the Markov chain. There are two main approaches for creating this matrix. Parametric models derive it from relatively few parameters that can be estimated directly from the sequences in question. For example, the model by Goldman and Yang (1994) incorporated parameters for the ratio of the rate of transitions to the rate of transversions (\(\kappa\)) and for the ratio of nonsynonymous to synonymous substitution rates (\(\omega\)). Muse and Gaut (1994) followed a similar approach but introduced explicit parameters for synonymous and nonsynonymous substitution rates and defined the rate matrix in terms of nucleotide frequencies instead of codon frequencies. Typically, the parameters of these models are estimated from the data to which the model is applied; the model can therefore adjust to the particularities of the analyzed sequences. Over the years, parametric models benefited from various improvements such as varying \(\omega\) among lineages (Yang 1998) or among sites (Nielsen and Yang 1998).

For empirical models, all elements of the rate matrix are estimated once from large data sets and then kept constant for all further applications. The first empirical codon model was presented by Schneider et al. (2005) who used alignments of vertebrate coding sequences to estimate a rate matrix. Unlike previous models, this model also allowed for instantaneous multinucleotide substitutions and realistic rates between different amino acids. It was shown to be better suited for certain evolutionary analyses such as sequence alignments, but for many other purposes, the use of a fixed set of substitution rates, independent from the analyzed sequences, is too strict. Therefore, a new generation of codon models, so-called semiempirical models, have been presented (Doron-Faigenboim and Pupko 2007; Kosiol et al. 2007). These are based on empirical rate matrices but combine them with parameters similar to those of the parametric models. Thus, they are able to adjust to the data sets in question.

The parameters of these models were basically the same as those used in the first parametric codon models. The \(\kappa\) parameter had to be refined to also model multiple transitions or transversions between codons and the meaning of \(\omega\) has changed: It no longer describes the absolute nonsynonymous to synonymous rate ratio but expresses the deviation from the average ratio of nonsynonymous and synonymous rates in the original data. However, the basic assumptions have not changed substantially. The choice for this set of parameters is reasonable and roots in years of
experience and theoretical considerations on the underlying biological principles. We previously identified the most relevant parameters for codon models with an empirical approach (Zoller and Schneider 2010) using principal component analysis (PCA). The PCA computation is based on the covariance matrix from many observed data sets and results in a set of uncorrelated vectors (called principal components or PCs), which point in the directions of the largest variances in the data. Applying this method to a large set of parameters for modeling codon substitutions. The most relevant parameters were found to be the well-established nonsynonymous–synonymous rate ratio ($\omega$), followed by the rate of multinucleotide substitutions (we call it $\nu$, see also below), and the transition–transversion rate ratio ($\kappa$).

A graphical representation of the findings by Zoller and Schneider (2010) is displayed in figure 1.

Here, we present a new codon model, which roots directly in those PCA results. The principal component analysis codon model (PCM) reverses the PCA by building a linear combination of the first few PCs to approximate the optimal substitution rates for the data sets in question. The coefficients for the PCs in the linear combination are free model parameters and can be fitted to the individual data sets. Therefore, the PCM model is not only based on an empirical rate matrix but also allows to adjust exactly those combinations of rates that have empirically been found to vary the most among data sets.

In our earlier study, one important parameter that has been identified was the relative rates of codon substitutions involving more than one nucleotide change and substitutions with only one change. This parameter seemed to be second most important after $\omega$, even more important than the commonly used $\kappa$. Similar observations have been made by Kosiol et al. (2007) as well as Doron-Faigenboim and Pupko (2007). To test the efficiency of this new parameter, we have created another new model, called ECM+$\omega$+$\nu$, that incorporates only the two parameters $\omega$ and $\nu$ (for the relative rate of multinucleotide substitutions) and compared it with other models that also include a variant of the $\kappa$ parameter.

Both new models, PCM and ECM+$\nu$, have been implemented in “BEAST/BEAGLE” (Drummond and Rambaut 2007), an application that uses Bayesian Markov chain Monte Carlo techniques to analyze molecular sequences, and in “CodonPhyML” (Zanetti 2010), which is an extension of “PhyML” (Guindon et al. 2010) towards codon models. All models described here will be available with the forthcoming update of BEAST and the release of CodonPhyML, respectively.

Materials and Methods

The PCM

The rate matrix for the “PCM” is based on a linear combination of PCs, which describe the most important features of codon evolution (Zoller and Schneider 2010). We used 3,666 symmetric substitution matrices as input data, each matrix encoded as a vector of length \(61 \times (61-1)/2 = 1,830\). Performing PCA on such a set of data vectors consists of three steps: First, the vectors are centered (by subtracting the means $M$ from all input vectors), then the individual parameters are normalized to a standard deviation (SD) of 1 (by dividing all elements of the centered input vectors by the SDs $S$), leading to centered and normalized vectors $A^{(i)}$. Computing the eigenvector/eigenvalue decomposition of the covariance matrix of the $A^{(i)}$ results in the PCs, a list of vectors $P^{(i)}$ which are an orthogonal reprojection of the input data. After sorting the PCs decreasingly according to their corresponding eigenvalues, the first PCs indicate the directions of the largest variances in the input data. Linear combinations of the first few PCs normally allow for reasonable approximations of all input vectors. Therefore, as the input for this work were codon substitution matrices estimated from single multiple sequence alignments (MSAs), the resulting PCs describe the most varying features of codon substitution matrices.

The PCM model presented here is created by reversing the PCA with fewer components but free coefficients, using $M$, $S$, and the $P^{(i)}$ represented as symmetric $61 \times 61$ matrices. The parameters of the models are the coefficients $c_i$ of the linear combination of the PCs. We use PCM+nC to denote a PCM with n PCs. The symmetric relative rate matrix (also called exchangeability matrix) is a linear combination of $M$ and the scaled PCs $P^{(i)}$. If an element becomes less than 0, thus violating properties of Markov models, it is set to 0.
\[ R_{uv} = \max \left( 0, M_{uv} + S_{uv} \sum_{i} c_i \rho^{(i)} \right). \] (1)

Throughout this article, we are working with matrices of dimension 61 \times 61 that model only the substitutions between sense codons and ignore the stop codons. Given the exchangeability matrix \( M \) with the diagonal elements chosen such that the row sums of the rate matrix will add up to 0, the rate matrix \( Q \) is computed as \( Q_{ij} = \pi_i R_{ij} \), where \( \pi_i \) denotes the frequency of codon \( j \). At last, the substitution matrix \( P(t) \) with the substitution probabilities for evolutionary distance \( t \) is computed as \( P(t) = e^{Qt} \).

Interpreting PCs

Figure 1 visualizes the contributions of some predefined parameters to the PCs found by Zoller and Schneider (2010). We chose eight possible codon substitution parameters and encoded them as vectors of length 1,830, corresponding to the 1,830 substitution rate parameters in the PCs. These so-called feature vectors were then compared with the obtained PCs. The absolute value of the Pearson correlation coefficient between a feature vector and a PC denotes the similarity between the two vectors, allowing to measure the contributions of the individual feature vectors to a particular PC.

The tested feature vectors encode \( \kappa, \omega, \) and \( \nu \) (for multinucleotide substitutions) as well as changes in five physicochemical properties of the encoded amino acids: weight, acidity, charge, hydropathy, and polarity.

For each PC and for each feature vector, a 95% confidence interval from correlations to 10,000 random vectors has been estimated. This showed that correlations lower than 0.045 can be considered as random (the inner disc in fig. 1).

Multinucleotide Substitutions as a Model Parameter

The second new model, denoted ECM+\( \omega+\nu \), was created after the interpretation of the first few PCs of the Mammalia data set. It is based on an initial empirical rate matrix—hence the Empirical Codon Model (ECM) as given by Kosiol et al. (2007)—and includes the two most important parameters according to the analysis: \( \omega \), the ratio of nonsynonymous to synonymous substitutions and \( \nu \), a factor for multinucleotide substitutions. Given an initial rate matrix \( l \) of empirical substitution rates, the relative rate matrix \( R \) is computed as (“nt” denoting “nucleotide”):

\[
R_{uv} = \begin{cases} 
1_{uv} & \text{if } u \rightarrow v \text{ is a syn. single-nt change} \\
1_{uv,\omega} & \text{if it is a nonsyn. single-nt change} \\
1_{uv,\nu} & \text{if it is a syn. multi-nt change} \\
1_{uv,\omega,\nu} & \text{if it is a nonsyn. multi-nt change}
\end{cases}
\] (2)

CodonPhyML

CodonPhyML (Zanetti 2010; Zanetti, Gil, Anisimova, in preparation) extends the phylogeny software PhyML (Guindon et al. 2010) to also implement codon models in addition to the nucleotide and amino acid models available in the standard version. It uses maximum likelihood estimation to optimize model and tree parameters. Previously, CodonPhyML included already variants of the models of Goldman and Yang (1994) as well as several models of the ECM family. For this work, Marcelo Zanetti also kindly included our new models PCM+\( \nu \)C and ECM+\( \omega+\nu \).

Codon frequencies in CodonPhyML can either be estimated from the MSA with different frequency models or they can be given as input parameters. Similarly, the initial rate matrix of the semiempirical model families ECM and PCM can be provided by the user.

Comparison of Different Codon Models

We compared our new models with the well-established M0 model (Goldman and Yang 1994) and the semiempirical ECM+F+\( \omega+2\kappa \), a member of the ECM family of codon models (Kosiol et al. 2007). Of the ECM variants, ECM+F+\( \omega+2\kappa \) was chosen because it performs very well compared with other ECM variants (Kosiol et al. 2007). This model uses two different parameters \( \kappa_1 \) and \( \kappa_2 \) for the rates of transitions and transversions, respectively. The rate of substitution between codons \( i \) and \( j \) is multiplied by \( \kappa(i,j) = \kappa_1^{n_{iv}} \kappa_2^{n_{iv}} \), where \( n_{iv} \) is the number of transitions and \( n_{iv} \) the number of transversions between \( i \) and \( j \) (see Kosiol et al. 2007). For all models and computations, we used the ‘+F’ variant where the codon frequencies are estimated from the alignment. Thus, for better readability, we will discard the ‘+F’ in future references to models in this paper.

Model comparison in a maximum likelihood (ML) framework was done by calculating and interpreting the second-order Akaike Information Criterion value (\( \text{AIC}_C \) Sugiura 1978; Hurvich and Tsai 1989). Given the log-likelihood (logL) \( l \) of a model and a tree, with \( p \) free parameters (in the tree and the model) and \( n \) columns in the codon-wise alignment, the \( \text{AIC}_C \) is computed as follows:

\[
\text{AIC}_C = (-2l + 2p) + \frac{2p(p+1)}{n-p-1}.
\] (3)

In some MSAs, the number of codons per sequence of a given MSA was too low, that is, the term \( (n-p-1) \) became very small or even negative, suggesting that the amount of information is too small for \( \text{AIC}_C \) to work. We therefore excluded MSAs where \( (n-p-1) > 10 \) from the comparison.

When the AIC is used to rank models, the best model is assumed to be the one with the smallest \( \text{AIC}_C \) value. For the other models, the \( \text{AIC}_C \) difference to the best model is computed, but unlike likelihood ratio tests, the AIC cannot be used to reject a model. However, the following thresholds have been proposed as rules of thumb: a difference of less than 2 is still interpreted as “considerable support,” a difference between 4 and 7 is called “considerably less support,” and differences larger than 10 indicate “no support” for that model (Burnham and Anderson 2002, chap. 2.6, p. 70).

In this study, we only compared two models at a time (e.g., an instance of the ECM family against an instance of PCM). Thus, we calculated for two models \( A \) and \( B \) the values \( a = \text{AIC}_C(A) \) and \( b = \text{AIC}_C(B) \), respectively. If \(|a-b| \)
was larger than 10, strong support in favor of the model with the lower AIC_C value was concluded.

Alignments Used in the Comparison
Kosiol et al. (2007) used 200 MSAs from the PANDIT database (Whelan et al. 2003) to evaluate their ECM model family. We extracted those 200 MSAs and the corresponding trees from the database; 17 MSAs with less than four species were discarded. Also, 199 new “Mammalia” MSAs were chosen from the Orthologus Matrix Project (OMA) project (version of 24 May 2010), but it was ensured that all the test MSAs were different from the ones used for performing the PCA. Furthermore, 200 MSAs each were taken from the OMA predictions for “Archaea” (version of 24 March 2011), “Cyanobacteria” (8 April 2011), “Other Eukaryota” (28 March 2011), and “Vertebrata” (17 June 2011). These groups have been chosen as test cases because their in-group distances as well as their codon usage patterns differ from the ones in the Mammalia data set. All of the OMA MSAs and the corresponding trees were constructed in the same way as described for the PCA analysis (Zoller and Schneider 2010). The MSAs used from Vertebrata and Mammalia are not overlapping.

Results and Discussion
Codon Model Comparison
We compared variants of four different codon models: M0 (Goldman and Yang 1994), ECM+ω+2κ (Kosiol et al. 2007), ECM+ω+ν, and PCM in an ML framework. Besides using the ECM models with the initial rate matrix of Kosiol et al. (2007), we also tested them with an initial matrix estimated from the Mammalia data set, denoting these models “ECM+ω+2κ (mam)” and “ECM+ω+ν (mam).” Of the PCM model family, we evaluated PCM+2C and PCM+8C, the variants using two and eight PCs, respectively.

The model comparison is based on the logL per alignment site, computed over all alignments in the respective data sets. Figure 2 gives an overview of these results, comparing all models on six different data sets. Because the differences in logL/site are much larger among the data sets than among the models, only the improvements over M0 on each data set are shown in figure 2. The absolute logL/site values for M0 are shown in figure 3 and allow to compare the performance of codon models on different taxonomic ranges. Higher bars correspond to better logL values; the error bars denote the 95% confidence interval based on 1,000 bootstrap samples for each data set.

A first observation is that all models that include empirical rates achieve higher likelihood values than M0 (with ECM+ω+ν on the Mammalia data set being the only exception). The most important difference between ECM and M0 is the use of empirical substitution rates, although the parameters remain similar. This clearly shows that including empirical rates improves the performance of a codon substitution model. The PCM models perform better than ECM on MSAs from vertebrate sequences (i.e., the data sets Mammalia and Vertebrata). Using its original initial rate matrix, ECM+ω+2κ outperforms PCM+νC in the data sets Archaea, Cyanobacteria, and Other Eukaryota. The ECM models achieve better results than the PCM models on the Pandit data set.

The performance of the codon models on different data sets (fig. 3) shows that codon models are not equally suited for all taxonomic ranges. Note that the logL differences among the data sets are an order of magnitude larger than...
among the models (fig. 2). Codon models perform well on mammalian and vertebrate sequences, whereas for other eukaryotes, bacteria, and archaeas, they appear to be less suited. This can probably be attributed to longer evolutionary distances in these data sets (and thus saturation of synonymous substitutions) and also to the stronger selection on codon usage in prokaryotic species (Ikemura 1985). The Pandit data set, which spans all three domains of life, seems to be particularly unsuited for codon models.

When comparing PCM and ECM models, not only the performance of the different parameters but also the influence of the underlying empirical rate matrices are taken into account. The initial ECM rate matrix was estimated from mammalian sequences only. By changing the initial rate matrix of the ECM model to the initial rate matrix estimated in the first steps of the PCA on the Mammalia data, the minimum AICc value of both ECM+ω and ECM+ν models was then compared with the minimum AICc value of both ECM+ω+2κ models. Differences smaller than 10 are not included.

 ECM+ω+ν performed clearly better than ECM+ω+2κ on Cyanobacteria and also often better on the Pandit and Other Eukaryota data sets. In Mammalia, Vertebrata, and Archaea, ECM+ω+2κ outperforms ECM+ω+ν in every MSA with strong enough AICc difference.

Interestingly, many MSAs result in no significant support for either model. In the Pandit data set, only 37% of the AICc value per model family has been considered; we are thus comparing the potential of a model (what the best variant could achieve), not the performance of a specific instance of the model. If the difference between the two lowest AICc values of each family was smaller than 10, the result was discarded for lack of sufficient support for either model (Burnham and Anderson 2002). In all data sets, there is always one family that clearly outperformed its competitor. In Mammalia and Vertebrata, PCM dominates the picture. In Pandit, Archaea, Cyanobacteria, and Other Eukaryota, the ECM models clearly outperform PCM.

**Comparison of Model Families**

Figure 4 gives another perspective on the results of the model comparison. For each data set, only the lowest AICc value per model family has been considered; we are thus comparing the potential of a model (what the best variant could achieve), not the performance of a specific instance of the model. If the difference between the two lowest AICc values of each family was smaller than 10, the result was discarded for lack of sufficient support for either model (Burnham and Anderson 2002). In all data sets, there is always one family that clearly outperformed its competitor. In Mammalia and Vertebrata, PCM dominates the picture. In Pandit, Archaea, Cyanobacteria, and Other Eukaryota, the ECM models clearly outperform PCM.

**Comparison of ECM Models**

By comparing only ECM+ω+2κ and ECM+ω+ν, we can test if parameters for transition and transversion rates are important or if a simple parameter for multinucleotide substitutions is just as effective. Table 1 compares and summarizes the lowest AICc values per ECM model of each data set. For both ECM+ω+2κ and ECM+ω+ν, AICc values have been calculated using Kosiol’s initial rate matrix and using the rate matrix estimated in the first steps of the PCA on the Mammalia data. The minimum AICc value of both ECM+ω+2κ models was then compared with the minimum AICc value of both ECM+ω+ν models. Differences smaller than 10 are not included.

<table>
<thead>
<tr>
<th>Family</th>
<th>Mean(κ)</th>
<th>Var(κ)</th>
<th>ECM+ω+2κ</th>
<th>ECM+ω+ν</th>
<th>ECM+ω+2κ values (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandit</td>
<td>1.487</td>
<td>0.287</td>
<td>33 (24)</td>
<td>17 (13)</td>
<td></td>
</tr>
<tr>
<td>Mammalia</td>
<td>2.722</td>
<td>0.289</td>
<td>0</td>
<td>95 (48)</td>
<td></td>
</tr>
<tr>
<td>Archaea</td>
<td>1.607</td>
<td>0.020</td>
<td>0</td>
<td>41 (21)</td>
<td></td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>1.255</td>
<td>0.013</td>
<td>100 (50)</td>
<td>9 (5)</td>
<td></td>
</tr>
<tr>
<td>Other Eukaryota</td>
<td>1.275</td>
<td>0.022</td>
<td>16 (8)</td>
<td>10 (5)</td>
<td></td>
</tr>
<tr>
<td>Vertebrata</td>
<td>2.443</td>
<td>0.252</td>
<td>0</td>
<td>99 (50)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1. Means and Variances of the Distributions of κ for Each Data Set as well as the Numbers of Best Supported AICc Values.**

**Fig. 4.** Comparison of the ECM and PCM model families. The bars display the percentages of MSAs, on which a model outperformed the other by a notable difference (> 10) in AICc (Burnham and Anderson 2002). The numbers in the data set labels inform about the total number of MSAs in a particular set.
Fig. 5. Estimations of $\kappa$ for each MSA (under M0) plotted against the difference of logL/site between ECM+$\omega+2\kappa$ and ECM+$\omega+\nu$ ($\Delta\text{logL/site}$). $\Delta\text{logL/site}>0$ implies that ECM+$\omega+2\kappa$ performed better than ECM+$\omega+\nu$.

$\kappa_1 = \kappa_2 = \kappa$, the substitution matrix of ECM+$\omega+2\kappa$ would be defined as

$$Q_{ls} = \begin{cases} 
\pi_1 l_{is} \omega^n & \text{if } l \rightarrow s \text{ is a synonymous change} \\
\pi_1 l_{is} \omega^m & \text{if } l \rightarrow s \text{ is a non-syn. change}
\end{cases}$$

with $n$ denoting the number of nucleotide changes. ECM+$\omega+\nu$ uses a very similar approach:

$$Q_{ls} = \begin{cases} 
\pi_1 l_{is} \nu^n & \text{if } l \rightarrow s \text{ is a synonymous change} \\
\pi_1 l_{is} \nu^m & \text{if } l \rightarrow s \text{ is a non-syn. change}
\end{cases}$$

but with $m = 0$ for a single nucleotide change and $m = 1$ for a multinucleotide substitution. If the $\kappa$ values in ECM+$\omega+2\kappa$ are equal, the only difference between these two models are the exponents $n$ and $m$. Although ECM+$\omega+\nu$ is not strictly nested within ECM+$\omega+2\kappa$, the models can—depending on the parameters of ECM+$\omega+2\kappa$—become so similar that it is very hard to make a distinction, resulting in very similar logL values.

The biggest difference between the two models is the introduction of separate rates for transitions and transversions in ECM+$\omega+2\kappa$. We would therefore expect ECM+$\omega+\nu$ to perform as well as ECM+$\omega+2\kappa$ on data sets where the $\kappa$ values are near 1 and the variance of its distribution is very small. This is indeed the case as can be seen in table 1, which lists for each data set the means and variances of the $\kappa$ values estimated from the corresponding MSAs as well as the results of the model comparisons. Low mean $\kappa$ values with small variances correlate with better performance of ECM+$\omega+\nu$; when the mean $\kappa$ in the data set is high, ECM+$\omega+2\kappa$ more often outperforms its competitor.

This correlation between estimated $\kappa$ values and the relative performance of the two ECM model variants is visualized in fig. 5. The graph shows estimations of $\kappa$ for each MSA (under M0) plotted against the difference of logL/site for ECM+$\omega+2\kappa$ and ECM+$\omega+\nu$ ($\Delta\text{logL/site}$). If $\Delta\text{logL/site}>0$ for a particular MSA, ECM+$\omega+2\kappa$ performed better than ECM+$\omega+\nu$. It can be seen that on MSAs with a higher $\kappa$ value, ECM+$\omega+2\kappa$ has a clear advantage, whereas for $\kappa$ values below 2, both models give very similar results. In the 290 (of 1,172) MSAs for which $\Delta\text{logL/site}<0$, the maximum value for $\kappa$ is 1.86. In 830 MSAs (70.8%), $\Delta\text{logL/site}$ does not differ by more than 0.03.

Conclusions

Using the results from our previous study on applying PCA on codon sequences, we built a new semiempirical PCM, which uses linear combinations of a given number of PCs to approximate the empirically found variation of codon substitution rate matrices. Our results show that this approach outperforms any other tested model when applying it to vertebrate sequences. This also holds on other data sets when comparing PCM to an ECM model with an initial rate matrix estimated from mammalian alignments. It is not unexpected that the PCM model performs best on vertebrate data, given that this class also contains Mammalia, where the empirical analysis to construct the model was based on. However, the model is not simply overfitting a given set of sequences; all tests were performed on a completely new set of MSAs from an extended taxonomic range. This indicates that the PCA approach is able to capture more information from the empirical analysis than the ECM method and thus is able to improve the modeling of codon evolution, as long as the model is applied to sequences close to the taxonomic range of the training set. The PCM model was trained on mammalian data, as this is where codon models probably make the most sense: the interspecies distances are small enough for the synonymous substitutions not to saturate, and selection on codon usage is weaker here than in single-celled organisms (Ikemura 1985).

Codon models are often formulated such that the $\omega$ parameter is allowed to vary across sites, reflecting the varying degrees of selection at different positions of a gene (Nielsen and Yang 1998). The PCM model presented here does currently not allow for such rate variation, and an analogous implementation of rate variation is not straightforward since there is no explicit selection parameter in our model. Possible solutions would be to either include an extra $\omega$ parameter in addition to the PCs or to again use PCA to empirically estimate a component (or several components) that vary the most across sites. Since both options are not without drawbacks, we decided to ignore this problem for now and try to tackle it in future work.

The most relevant component of the first PC in our analysis is the selection coefficient $\omega$, which is already widely used in parametric codon models. PC2 consists mostly of a linear combination of $\kappa$ and $\nu$, the latter being more relevant. Although the mechanisms of multinucleotide substitutions are not yet fully understood, they appear to be important, and one might want to include this parameter in new parametric models of codon substitution. This finding is consistent with the conclusion given by Kosiol et al. (2007). Furthermore, our own model comparisons showed that models including solely $\omega$ and $\nu$ perform similarly to models that include $\kappa$, but that it is often beneficial to model...
κ as well. On data sets with a high mean and variance in the distribution of κ, the lack of modeling transitions and transversions led to lower logL/site values for our test model ECM+ω+ν compared with ECM+ω+2κ.

The PCM presented here combines all these important parameters in the first couple of PCs. Therefore, it is able to model the most relevant aspects of evolution on codon level using only few parameters, the coefficients of the PCs.

Acknowledgments

We would like to thank Gaston Gonnet, Maria Anisimova, and Daniel Dalquen for helpful discussions. Many thanks also to Marcelo Zanetti who helped substantially with implementing the new codon models in CodonPhyML. Thanks also to the reviewers that contributed to improving the manuscript. A.S. was supported by a grant PBEZP2-129595 from the Swiss National Science Foundation (SNF).

References


