Testing Substitution Models Within a Phylogenetic Tree

Gunter Weiss*† and Arndt von Haeseler†‡

*Max-Planck-Institut für evolutionäre Anthropologie, Leipzig, Germany; †Heinrich-Heine Universität Düsseldorf, Germany; and ‡Forschungszentrum Jülich, Germany

Phylogenetic tree reconstruction frequently assumes the homogeneity of the substitution process over the whole tree. To test this assumption statistically, we propose a test based on the sample covariance matrix of the set of substitution rate matrices estimated from pairwise sequence comparison. The sample covariance matrix is condensed into a one-dimensional test statistic $\Delta = \sum \ln(1 + \delta_i)$, where $\delta_i$ are the eigenvalues of the sample covariance matrix. The test does not assume a specific mutational model. It analyses the variation in the estimated rate matrices. The distribution of this test statistic is determined by simulations based on the phylogeny estimated from the data. We study the power of the test under various scenarios and apply the test to X chromosome and mtDNA primate sequence data. Finally, we demonstrate how to include rate variation in the test.

Introduction

The reconstruction of evolutionary trees from molecular sequence data is nowadays a common tool with which to infer the ancestral relationship of living organisms. This is in part due to the ease of getting sequence data from any species one might be interested in either by standard sequencing techniques or simply by database searches. In addition, interest in the theory of sequence evolution and tree reconstruction has led to a large number of studies on substitution models (Zharkikh 1994), methods of reconstructing trees, exact and heuristic tree-search algorithms, and statistical procedures testing model assumptions and/or tree topologies (cf. Swofford et al. 1996). Today several software packages are available which put many of the theoretical results into practice (Swofford 1993; Felsenstein 1995; Adachi and Hasegawa 1996; Strimmer and von Haeseler 1996; Yang 1999; Kumar et al., 2000).

Given a set of sequence data, the first step in the analysis is the alignment of the sequences. This is a topic of its own which we do not address here (cf Gotth 1999). From a sequence alignment a phylogenetic tree can be reconstructed by several means. Each method, however, uses different assumptions (stated either explicitly or implicitly), different optimality criteria, and different tree search algorithms. For example, most distance-based methods as well as maximum likelihood methods assume an explicit model of sequence evolution. With the exception of LogDet and paralinear distances, these models require homogeneity, reversibility, and stationarity. The model is usually defined by a matrix containing the instantaneous rates of change from one nucleotide to another. In addition it is assumed that the positions in a sequence evolve mutually independently (for review see Baake and von Haeseler 1999). However, the model might allow the overall rate of change to differ between positions, such that rate heterogeneity among sites is incorporated. Well known are attempts to model rate heterogeneity according to a gamma distribution (Yang 1993, 1994) or by a log-normal distribution (Olsen 1987).

Although methods for testing the model of nucleotide substitutions abound (i.e., Goldman 1993a,b), little work has been put into the assumption that the evolutionary model is identical on all branches of a tree (for an exception see Yang and Roberts 1995). Recently, Devauchelle et al. (2001) observed large variations across taxa in substitution matrices for mtDNA encoded proteins. These observations were obtained by singular value decomposition, with no obvious way to support the results statistically.

In this study we question the justifiability of the assumption of a single evolutionary model acting on all branches of a tree. To this end, we develop a statistical test based on the sample covariance matrix $S$ of the set of substitution rate matrices estimated from pairwise sequence comparison. The sample covariance matrix is condensed into an one-dimensional test statistic $\Delta = \sum \ln(1 + \delta_i)$, where the $\delta_i$ are the eigenvalues of $S$. $\Delta$ is used to test the assumption of a single substitution model acting on all branches of a tree—i.e., testing the homogeneity assumption of the substitution process in the tree.

The outline of the paper is as follows: First, we define models of sequence evolution and describe how to estimate substitution rate matrices from pairwise sequence comparisons. Next, we derive and motivate the test statistic $\Delta$ and show how to determine the null-distribution of $\Delta$ by simulations. The reliability and sensitivity of the test are demonstrated by simulation studies. Finally, we apply the test to primate sequence data from the X chromosome and mitochondrion.

Models and Matrices

Suppose that substitution in a DNA sequence follows the Markovian model of molecular evolution (Felsenstein 1983; Lanave et al. 1984; Tavaré 1986; Gu and Li 1996; Kelly and Rice 1996)—i.e., a continuous-time stationary Markov process defined by a rate matrix $Q$. This matrix has off-diagonal entries $q_{ij} > 0$ that give the instantaneous substitution rates at which nucleotide $i$ is replaced by nucleotide $j$ ($i \neq j \in \{A, C, G, T\}$). The diagonal elements of $Q$ are given by

$$q_{ii} = -\sum_{j \neq i} q_{ij},$$
such that rows sum up to zero. Let $\pi_i, i \in \{A, C, G, T\}$ be the equilibrium frequency of nucleotide $i$. We scale time such that the expected rate of substitutions

$$-\sum_i \pi_i q_{ij} = 1,$$

(1)

i.e., time is measured in expected number of substitutions. We require the substitution process to be reversible in time. Then,

$$\pi_i q_{ij} = \pi_j q_{ji}$$

must hold for all $i \neq j \in \{A, C, G, T\}$. This implies that the off-diagonal entries of $Q$ can be expressed as $q_{ij} = \pi_j b_{ij}$ with $b_{ij} = b_{ji}$, such that the model contains nine free parameters.

Assume that the rate matrix $Q$ is diagonalizable (which is the generic case). Then, $Q$ has a spectral decomposition

$$Q = U \text{diag}[\lambda_0, \lambda_1, \lambda_2, \lambda_3] U^{-1},$$

(2)

where $U$ is the matrix of (right) eigenvectors and $\text{diag}[\lambda_0, \lambda_1, \lambda_2, \lambda_3]$ a diagonal matrix containing the eigenvalues of matrix $Q$. The constraints on $Q$ imply that $\lambda_0 > \lambda_1 \geq \lambda_2 \geq \lambda_3$. (For details we refer to Chapter 4.8 and Appendix of Karlin and Taylor (1975)). The transition probability matrix for a branch of length $t$ (i.e., average number of substitutions) is given by

$$P(t) = \exp(Qt).$$

From equation (2) this can be expressed as

$$P(t) = U \text{diag}[\eta_0, \eta_1, \eta_2, \eta_3] U^{-1},$$

(3)

with $\eta_s = e^{\lambda_s t}$, $s = 0, 1, 2, 3$, and $\eta_0 = 1 > \eta_1 \geq \eta_2 \geq \eta_3 > 0$.

While the structure of the substitution process defined by $Q$ is assumed to be the same for all positions along the sequence, the rate of evolution might differ from position to position. We model this rate heterogeneity among sites by a nonnegative random variable $R$ with mean $\mathbb{E}(R) = 1$ and moment-generating function $\phi(x) = \mathbb{E}(\exp(Rx))$ (Gu and Li 1996; Kelly and Rice 1996). The relative rate $r$ at a position is then given by a realization of $R$. However, the requirement $\mathbb{E}(R) = 1$ ensures that the evolutionary rate averaged over the whole sequence remains the same independent of the specific distribution of $R$ (for an overview see Waddell and Steel [1997]).

Conditional on $R = r$, the transition probability matrix of a position for a branch of length $t$ is given by

$$P(t | R = r) = \exp(Qr).$$

Suppose $R$ has density $f(r)$. Then, the unconditional transition probability matrix is

$$P(t) = \int_0^\infty P(t | R = r) f(r) dr$$

(4)

$$= U \text{diag}[\eta_0, \eta_1, \eta_2, \eta_3] U^{-1},$$

(5)

with eigenvalues

$$\eta_r = \phi(\lambda_r t) = \int_0^\infty e^{\lambda_r r} f(r) dr.$$

(6)

In molecular evolution rate heterogeneity is frequently modelled by a $\Gamma$-distribution with density

$$f(r) = \frac{\alpha^\gamma}{\Gamma(\alpha)} r^{\gamma-1} e^{-\alpha r},$$

(7)

where the shape of the distribution depends on the parameter $\alpha > 0$ (Uzzel and Corbin 1971; Wakeley 1993). For this model equation (6) yields

$$\phi(\lambda_r t) = \left( \frac{\alpha - \lambda_r t}{\alpha} \right)^{\gamma}.$$

If substitution rates are uniform among sites (i.e., $\alpha \to \infty$) then $\phi(\lambda_r t) = e^{\lambda_r t}$ as above (eq. 3). Note that the eigenvalues of $P(t)$ are functions of the product $\lambda_r t$.

**Inference and Testing**

Transition probability matrices as defined above are not directly observable from sequence data. However, they might be inferred from pairwise sequence comparisons as follows. Let $h_{i,j}$ denote the probabilities of having nucleotides $i$ and $j$ in a pair of sequences ($i, j \in \{A, C, G, T\}$). These probabilities are collected in a so-called divergence matrix $H$ (for review see Baake and von Haeseler 1999). Assuming stationarity and reversibility of the substitution process, the divergence matrix can be expressed as

$$H = \Pi P(t),$$

(8)

where $\Pi = \text{diag}[\pi_A, \pi_C, \pi_G, \pi_T]$. Using equations (5, 6, and 8), we get the following functional relationship between the substitution rate matrix $Q$ and observables from data:

$$Q \cdot t = U \cdot \text{diag}[^{-1}(\eta_0), ^{-1}(\eta_1), ^{-1}(\eta_2), ^{-1}(\eta_3)] \cdot U^{-1},$$

(9)

where $\eta_s, s = 0, 1, 2, 3$, are the eigenvalues of $P(t) = \Pi^{-1}$ and we assume that $^{-1}(\cdot)$ exists, which is the case for the $\Gamma$-distribution. $H$ can easily be estimated by relative frequencies of letter pairs in a pairwise alignment, an estimate for $\Pi$ is given by the average base composition in the corresponding alignment. Because we measure time in expected substitutions (eq. 1) we can use equation (9) to estimate $Q$ directly from pairwise sequence comparisons. Note that the constraint of equation (1) yields the identifiability of $Q$ in equation (9).

Now, suppose we were given a set of $n$ aligned sequences that are related by a phylogenetic tree. Furthermore, each of the $\binom{n}{2}$ sequence pairs yields an estimate of the substitution process that acted on the path connecting the sequence pair in the tree. Under the assumption of homogeneity of the substitution process over the whole phylogeny—i.e., all branches in the tree evolve according to the same substitution model $Q$—we expect all $\binom{n}{2}$ estimates of $Q$ to be close to each other. In this case differences between estimates are due to stochastic effects of sampling finite sequences. Alternatively,
when there is heterogeneity of the substitution patterns among branches, we expect that there is variation between estimates that is not attributable to random effects but reflects differences of the substitution patterns among branches.

To quantify variation between estimates we utilize the sample covariance matrix of the estimates of $Q$. To this end, let $q_{k,l}$ denote the vector of the 12 off-diagonal entries of the estimate of $Q$ for sequence pair $(k,l)$. Note that the diagonal entries of $Q$ are given by the negative of the respective row sum. Let further

$$
\bar{q} = \frac{1}{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} q_{k,l}
$$

be the mean of the vectors $q_{k,l}$. The sample covariance matrix $S$ is then given by

$$
S = \frac{1}{n} \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} (q_{k,l} - \bar{q})(q_{k,l} - \bar{q})',
$$

where $A'$ denotes the transpose of a matrix $A$. The matrix $S$ measures the variability between the estimates $q_{k,l}$. However, for constructing a statistical test the $12 \times 12$ matrix $S$ is not manageable. As it is typical of multivariate tests in general we condense the information contained in $S$ to a scalar by introducing the following quantity:

$$
\Delta = \sum_{i=1}^{12} \ln(1 + \delta_i),
$$

where $\delta_i, i = 1, \ldots, 12$ are the eigenvalues of $S$. Note that the definition of $S$ differs slightly from the logarithm of the determinant of $S$ defined by $\sum_{i=1}^{12} \ln(\delta_i)$. The addition of one to each eigenvalue in equation (12) improves numerical stability and prevents unfavorable behavior of singular matrices.

We use $\Delta$ as a statistic to test the null-hypothesis of homogeneity of substitution patterns over the complete phylogeny. To do this, we need to determine the distribution of $\Delta$ given the null-hypothesis is true. Because we are using estimates from pairwise sequence comparisons, where sequences are related by a phylogeny, we expect those estimates to be highly correlated. Hence, we do not presume any standard distribution to fit our problem, but resort to a simulation procedure to determine the null-distribution of $\Delta$ described next.

From equations (1) and (9) we get not only an estimate for $Q$ but also estimates for the distance between each pair of sequences. The resulting distance matrix is used to reconstruct the phylogenetic tree and estimate its branch lengths by standard-distance-based tree reconstruction methods, like Neighbor-Joining (Saitou and Nei 1987). Along this estimated phylogeny (tree and its branch lengths) we apply Seq-Gen (Rambaut and Grassly 1997) to simulate the evolution of sequences according to $q$, our best estimate for $Q$ under the null-hypothesis. For the resulting simulated data set, we compute $\Delta$ according to equation (12). By repeating this simulation, say 1,000 times, we get an approximation to the null-distribution of $\Delta$. The $P$ value of the actually observed $\Delta_{\text{data}}$ is then estimated by the proportion of simulated $\Delta$ values equal to or larger than $\Delta_{\text{data}}$.

**Simulation Studies**

To check reliability and sensitivity of our test procedure we conducted simulation studies for a variety of trees and models. We used an eight-taxon tree, shown in figure 1A, to simulate data sets under the null-model of homogeneous substitution patterns. Thus, we can determine the size of the test, i.e., the maximum probability of type I error (Bickel and Doksum 1977).

To demonstrate the statistical power of the approach, we simulated data along trees where on specific subtrees (see fig. 1 B, C, D) the substitution model $Q^*$ differed from that on the rest of the tree, $Q$. The branches in the trees (respecting the root node) are all of the same length $t$ and are measured in expected substitutions per sequence. Simulations were carried out for branch lengths $t = 10, 20, 50, 100, 200, \text{ and } 500$ and a sequence length of 5,000 nucleotides.

The substitution models are defined as pairs $(B, \pi)$ of symmetric matrices $B$ and base frequency vectors $\pi$, such that entries of $Q$ are given by

$$
q_{ij} = \begin{cases} 
    b_{ij}\pi_i & \text{if } i \neq j, \\
    -\sum_{k \neq i} b_{ik}\pi_k & \text{if } i = j.
\end{cases}
$$

Matrices $B$ and vectors $\pi$ were chosen to show whether small changes in the substitution patterns can be detected by our procedure. The entries of the basic matrix $B_0$, which specifies $Q$ in figure 1, are all equal to one. Thus, $(B_0, \pi)$ constitutes a Felsenstein-type substitution model (Felsenstein 1981). Matrices $B_1$ and $B_2$ differ from $B_0$ by only one entry. More precisely, we set $b_{C,F} = b_{T,G}$ equal to 2.0 and 4.0, respectively. That is to say, $B_1$ and $B_2$ correspond to situations where the rate of one type of base exchange is increased. We used four different
Table 1
Simulated Size (Tree A) and Power (Tree B-D) of the Test of Homogeneity for Various Substitution Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Q^*</th>
<th>Tree</th>
<th>Branch Length t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>(B_0, \pi_0)</td>
<td>–</td>
<td>A</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.110</td>
</tr>
<tr>
<td>(B_0, \pi_1)</td>
<td>–</td>
<td>A</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.362</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.620</td>
</tr>
<tr>
<td>(B_0, \pi_2)</td>
<td>–</td>
<td>A</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.080</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.063</td>
</tr>
<tr>
<td>(B_0, \pi_3)</td>
<td>–</td>
<td>A</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.350</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.570</td>
</tr>
<tr>
<td>(B_0, \pi_0)</td>
<td>–</td>
<td>A</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.032</td>
</tr>
<tr>
<td>(B_0, \pi_1)</td>
<td>–</td>
<td>A</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D</td>
<td>0.055</td>
</tr>
</tbody>
</table>

Note.—See fig. 1 and text for further details.

The power of the test is generally low with short branches that evolve differently from the rest of the tree. The "balanced" tree D shows generally the highest power. Joining primate Xq13.3 tree; the estimated transition rate matrix $\pi$ becomes clear that differences between the two models $Q$ and $Q^*$ in terms of matrices B are detected more easily (upper four ABCD blocks of table 1) than the minor differences in the base frequency distributions $\pi$ (last two blocks of table 1).

Applications

With the encouraging results from simulations, we applied our test to two data sets of hominoid DNA. The first data set consists of 10 kb sequences from the Xq13.3 locus sampled from human, bonobo, chimpanzee, gorilla, orangutan, and gibbon (Kaessmann et al. 1999, 2001). The second data set comprises the complete mitochondrial DNA sequences from the same species except gibbon (Horai et al. 1995).

As a first step we performed the test assuming no rate heterogeneity among sites. Figure 2A shows the Neighbor-Joining primate Xq13.3 tree; the estimated transition rate matrix $\pi$ and the estimated nucleotide frequencies $\pi$ are displayed in figure 2B. Based on these estimates the null-distribution of $\Delta$ was approximated by the simulations in figure 2C. The observed value $\Delta_{Xq13} = 1.071$ falls right into the distribution of $\Delta$ corresponding to a $P$ value of 0.203. Thus, the test does not reject the homogeneous substitution patterns assumption. The power of the test is rather low, however, when sequences of this level of diversity are analyzed (see scale of phylogeny in figure 2A).

A different result emerges for the complete mitochondrial DNA. The estimated tree (fig. 3A) agrees with the tree obtained from the Xq13.3 sequence. However, based on the estimates in figure 3B our simulations show, that
$\Delta_{mtDNA} = 1.033$ is way out of the range of the simulated null-distribution yielding a p-value of less than 0.001 (see fig. 3C). The hypothesis of homogeneity of the substitution patterns is rejected.

We further investigated the impact of possible rate heterogeneity among positions on the outcome of the test. More specifically, we modeled rate heterogeneity by a $\Gamma$-distribution (eq. 7) and varied the shape parameter $\alpha$ in our analyses. Note that small values of $\alpha$ correspond to high levels of rate heterogeneity. Figures 2D and 3D show the relationship between the upper 95% quantile of the null-distribution (based on 1,000 simulations for each choice of $\alpha$) and the observed value $\Delta_{Xq13}$ as a function of $\alpha$. It becomes clear that for the mtDNA data the null-hypothesis is rejected independently of levels of rate heterogeneity (fig. 3D).

The effect of possible rate heterogeneity on the analysis of Xq13 data is more interesting. For increasing level of rate heterogeneity, i.e. a decreasing value of $\alpha$, $\Delta_{Xq13}$ approaches the 95% quantile curve. For $\alpha < 0.03$ the value of $\Delta_{Xq13}$ is even larger than the corresponding 95% quantile. Thus, under the assumption of extreme rate heterogeneity, which may be indicated by extra evidence, the test would reject the homogeneity assumption for the Xq13 data. These results suggest that not obeying possible levels of rate heterogeneity results in a more conservative test. However, we note that with $\alpha$ close to zero numerical problems during estimation and simulation of the null-distribution become apparent.

**Discussion**

The assumption of homogeneity of the substitution process on a phylogeny can be tested statistically using DNA sequence data. We showed how to construct such a test and how simulations yield an approximation of the null-distribution of the proposed test statistic $\Delta$. The test itself does not assume a specific mutational model. It does not estimate parameters of the substitution process from data, but rather evaluates differences in estimated rate matrices.

However, if rate heterogeneity among sites should be incorporated, a model and its parameters have to be specified in advance of the application of the test.

Simulation studies and the analyses of two real data sets demonstrated that the test is admissible and has a reasonable power of rejection. We also showed that the problem of unknown rate heterogeneity can be circumvented by analyzing the data under various levels of rate heterogeneity. Application to the two real data sets suggests that assuming wrongly homogeneous substitution rates leads to the test being conservative.

What are the effects of a finding of heterogeneous substitution patterns within a phylogenetic tree? On the
one hand the heterogeneity is interesting in itself, because there is need for a biological explanation. We could imagine that substitution patterns do not change purely by chance but could result from a change in the sequence environment—e.g., less selective constrains due to a gene duplication, different replication machinery for a mitochondrial insert in the nucleus, or need for adaptation because of changes in selective constraints either on the locus studied or on genes that affect the locus indirectly.

On the more technical side the test results can improve the way phylogenies are built. Varying substitution patterns affect the estimation of branch lengths and possibly also the branching pattern of the reconstructed tree. For example, it is well known that compositional changes in nucleotide frequencies can produce misleading trees (Lockhart et al. 1994; Galtier and Gouy 1995). With this in mind, results of a phylogenetic reconstruction should be interpreted with care, if our test suggests a rejection of the homogeneity assumption.

Because a fully parameterized probability model on a phylogeny that allows for branch-specific substitution models is not feasible, the test result can help in improving the process of model building for a specific data set by using appropriate methods that take into account different rate matrices over the tree (e.g., Steel 1993; Lake 1994).

The proposed test clearly has its limits in that divergence matrices that are sparse easily lead to complex eigenvalues and problems in assessing the null-distribution of $\Lambda$ through simulations.

Although it was not demonstrated here, an obvious extension of the test is to answer which type of base exchange contributes significantly to a significant result of the overall test. The more challenging problem is to provide a definite answer to which branches contribute most to a significant result. For this purpose estimates of branch-specific substitution models seem to be necessary. We will address this question in the near future (Weiss and von Haeseler, in preparation).

**Acknowledgments**

We thank the MPG for financial support. We also thank Roland Fleißner for critically reading the manuscript. The program is available free of charge and can be downloaded from www.tree-puzzle.de.

**Literature Cited**


Wolfgang Stephan, Associate Editor

Accepted December 2, 2002