JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Number 6, 2000
Mary Ann Liebert, Inc.
Pp. 761–776

Modeling Amino Acid Replacement

TOBIAS MÜLLER[1] and MARTIN VINGRON[1,2]

ABSTRACT

The estimation of amino acid replacement frequencies during molecular evolution is crucial for many applications in sequence analysis. Score matrices for database search programs or phylogenetic analysis rely on such models of protein evolution. Pioneering work was done by Dayhoff *et al.* (1978) who formulated a Markov model of evolution and derived the famous PAM score matrices. Her estimation procedure for amino acid exchange frequencies is restricted to pairs of proteins that have a constant and small degree of divergence. Here we present an improved estimator, called the resolvent method, that is not subject to these limitations. This extension of Dayhoff's approach enables us to estimate an amino acid substitution model from alignments of varying degree of divergence. Extensive simulations show the capability of the new estimator to recover accurately the exchange frequencies among amino acids. Based on the SYSTERS database of aligned protein families (Krause and Vingron, 1998) we recompute a series of score matrices.

1. INTRODUCTION

MOST PROTEIN SEQUENCE ANALYSIS PROCEDURES rely on amino acid pair score matrices or exchange frequencies among amino acids. This is true in particular for sequence alignment methods (Waterman, 1995), database searching algorithms like BLAST (Altschul ***et al.***, 1990) or FASTA (Pearson and Lipman, 1988), and methods of phylogenetic reconstruction (Swofford and Olsen, 1990). In sequence alignment and database searching, score matrices attribute a similarity score to the matching of two amino acids. Thus, they are crucial for the correctness of the alignment or the ability of the search program to detect homologous sequences. In many cases, such a score matrix is derived from an estimator for amino acid exchange frequencies. Phylogenetic reconstruction is based on such frequencies, e.g., in the context of maximum likelihood estimation of phylogenies PROTML (Felsenstein, 1989).

Score matrices have been derived in different ways. Initially, the number of DNA mutations necessary to convert one amino acid into another (Fitch and Margoliash, 1967) was used as an indicator of distance or relatedness. Later attempts looked at physical parameters of amino acids (Rao, 1987) or studied associations in structural superpositions (Risler ***et al.***, 1988). The first statistical approach is described in McLachlan (1971). The widely used BLOSUM matrix series was derived from observed exchanges in block alignments of sequences of certain degree of divergence (Henikoff and Henikoff, 1992). In an alternative approach, one can interpret amino acid differences between homologous proteins as the result of a substitution process starting from a common though unknown ancestor. The formalization of this view was pioneered

[1]Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, 69120 Heidelberg, Germany.
[2]***Present address:*** Max-Planck-Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany.

by Dayhoff *et al.* (1978) who modeled this process as a Markov chain acting independently on each site of a protein.

Dayhoff *et al.* (1978) estimate the transition probabilities of this Markov chain. For each amino acid, they specify the probability that it is replaced by a particular other amino acid in some evolutionary period. A straightforward approach to this problem requires comparison of descendant sequences with their ancestors, which we do not know. The estimation thus needs to be based on indirect observations. For two homologous sequences, Dayhoff *et al.* (1978) estimate a common ancestor, restricting themselves to closely related pairs of sequences. Hence, they obtain a good estimation of replacement frequencies for short evolutionary periods. Modeling molecular evolution by a Markov chain, however, aims to interpret longer evolutionary periods by iterated application of this replacement dynamics. Dayhoff estimates exchange frequencies for short periods of time and extrapolates them to higher degree of divergence. Direct observation from alignments of remotely related sequences cannot be included in her method.

Today, large families of homologous proteins are available for comparison. Consequently, score matrices are being recomputed on these larger data sets, frequently based on Dayhoff's formalism (Jones *et al.*, 1992; Gonnet *et al.*, 1992). Furthermore, the advent of large numbers of structurally derived alignments has raised interest in using information also from very distant alignments for the purpose of estimating exchange frequencies and score matrices (Risler *et al.*, 1988; Overington *et al.*, 1990). The challenge arises to generalize Dayhoff's estimation procedure to handle input alignment of varying degree of divergence. First, Benner *et al.* (1994) suggest considering the evolutionary distances when deriving scoring matrices. In this paper, we present a different estimation procedure that fits a Markov model using alignment data of arbitrary degree of divergence. Our approach is more general and leads to different results than the one of Benner *et al.* (1994).

Section 2 of this paper will review some notions from Markov chain theory and link them to the biological phenomenon we want to describe. We will define a special class of Markov chains that are appropriate for modeling the evolution of proteins. Then we will introduce the resulting estimation problem. In Section 3 we will focus on the estimation of evolutionary time using a maximum likelihood method. The results will be validated by simulations in order to guarantee proper behavior of the method. Our novel method for estimating exchange frequencies will be introduced in Section 4. It relies on the concept of a resolvent associated to a Markov chain. Using this result, we will formulate an estimation procedure for the evolutionary process and evaluate its performance by simulations. In Section 5 we apply these developments to the calculation of a new "variable time" score matrix VT and compare it to the established BLOSUM and PAM series.

2. A STOCHASTIC MODEL FOR PROTEIN EVOLUTION

2.1. Definition of an evolutionary Markov process

A basic process in the evolution of proteins is the change of amino acids in time. Dayhoff *et al.* (1978) introduced the description of this process as a Markov chain. We start by summarizing the relevant notions in order to define "evolutionary Markov processes," abbreviated as "EMP." We assume that amino acids change independently at each site of the protein as depicted in Figure 1. Let us fix one site. Following Dayhoff *et al.* (1978), we model the evolution of this site as a continuous-time Markov chain X_t.

Let $\mathcal{A} = (a_1, \cdots, a_{20})$ denote the alphabet of the twenty amino acids. *Transition probabilities* are given by

$$p_{a_i a_j}(t) = Prob[X(s+t) = a_j | X(s) = a_i]$$

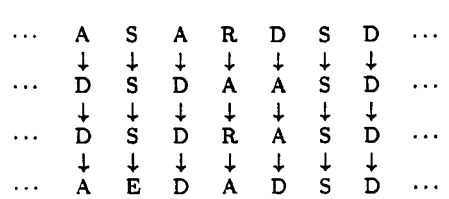

FIG. 1. The evolutionary process operates independently on each site of the proteins.

for times t, s and amino acids a_i, a_j. For two different amino acids a_i and a_j, $p_{a_i a_j}(t)$ denotes the probability that amino acid a_i has been replaced by amino acid a_j after a period t and $p_{a_i a_i}(t)$ is the probability that the amino acid a_i remains unchanged. We assume that these probabilities depend only on time increments t and not on absolute time points s. Mathematically, such a process is called *time homogeneous*.

Following Dayhoff, we calibrate the Markov chain, such that on average 1% of the amino acid are changed after one unit of time:

$$Prob[X(t) \neq X(t+1)] = 0.01.$$

Once calibrated, the time t in the Markov chain can be used as a measure of evolutionary time. Dayhoff introduced the acronym "PAM" (point accepted mutations) for these time units, where 1 PAM corresponds to 1% of amino acids changed.

We denote by $P(t)$ the transition probability matrix consisting of the entries $p_{a_i a_j}(t)$. For simplicity we write $p_{ij}(t)$ instead of $p_{a_i a_j}(t)$.

Note that:

- $P(0) = I$,
- $p_{ij}(t) \geq 0$ and $\sum_j p_{ij}(t) = 1$,
- $P_{s+t} = P_s P_t$ for $s, t \geq 0$,

where I denotes the identity matrix. The last equation is known as the Chapman–Kolmogorov equation.

We assume further that the Markov chain is continuous in the origin:

$$\lim_{t \searrow 0} p_{ij}(t) = \begin{cases} 1, & i = j \\ 0, & i \neq j, \end{cases} . \tag{1}$$

which is equivalent to the continuity of the functions $p_{ij}(\cdot)$. These functions are also differentiable in that the limit

$$\lim_{t \searrow 0} \frac{P(t) - I}{t} = Q \tag{2}$$

exists. $Q = (q_{ij})$ is called *rate matrix*. The diagonal entries of Q are negative; off-diagonal entries are positive.

Due to the time homogeneity assumption, the rate matrix Q provides an infinitesimal description of the process:

$$Prob[X(t+h) = i | X(t) = i] = 1 + q_{ii} h + o(h)$$
$$Prob[X(t+h) = j | X(t) = i] = q_{ij} h + o(h), \quad \text{for } i \neq j$$

or more concisely in matrix notation

$$P(h) \approx I + hQ$$

for small $h > 0$ and times $t \geq 0$.

From the Chapman–Kolmogorov equation, we get the forward and backward equation

$$\frac{d}{dt} P(t) = P(t)Q = QP(t). \tag{3}$$

This differential equation can be solved under the initial condition $P(0) = I$ and yields

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}.$$

This formula allows transition probabilities after any time of divergence t to be computed from the rate matrix. Vice versa, one can show that a matrix Q is a rate matrix of a family of transition probability matrices if and only if

$$q_{ij} \geq 0 \text{ for } i \neq j \quad \text{and} \quad \sum_j q_{ij} = 0 \text{ for all } i$$

(see, e.g., Grimmet and Stirzaker 1992).

A transition probability matrix determines the distribution of amino acids. We need to assume that any amino acid can mutate into any other during any period $t > 0$. In this case, there exists a unique limiting amino acid distribution $\pi = (\pi_1, \ldots, \pi_{20})$ where

$$\pi_j = \lim_{t \to \infty} p_{ij}(t) > 0$$

independently of the initial residue a_i. It fulfills the equations $\pi Q = 0$ and $\pi P(t) = \pi$ for all $t \geq 0$. Practically, this means that, if the overall amino acid usage in a set of proteins is equal to π, this distribution remains unaffected by evolutionary change. We only consider Markov processes which are in equilibrium and therefore all X_t are distributed according to π. Such a Markov chain is called *stationary* and π is called *stationary distribution.*

With the transition probabilities $(P_t)_{t \geq 0}$ and the overall amino acid distribution, we calculate the joint distribution $m_{ij}(t) = \pi_i p_{ij}(t)$ of (X_s, X_{s+t}). $M(t) = m_{i,j}(t)$ denotes the probability of finding amino acid a_i and amino acid a_j aligned with each other in two sequences that are t time units apart.

The Markov chain X_t describes the evolution of a single site in a protein from ancestors to descendants. While data from ancestor sequences are not available, we do observe pairs of proteins that have evolved from a common though unknown ancestor. To account for this problem, we assume that the evolution from ancestors to descendants can be modeled by the same process as its reverse. This enables us to explain the difference in both sequences by the same process. In Markov chain theory, this property is called *reversibility.* Formally, the probability of being and going from amino acid a_i to a_j in time t is the same as that of being and going from amino acid a_j to a_i in this time. This yields the *detailed balance* equation $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ for all $t > 0$ or equivalently, in terms of the entries of the rate matrix, $\pi_i q_{ij} = \pi_j q_{ji}$, and results in a symmetric matrix $M(t)$.

The above conditions on a Markov process allow a meaningful description of the amino acid replacement process and lead us to the following definition.

Definition 1 (EMP). *We call a continuous-time Markov chain X_t on the alphabet of amino acids an evolutionary Markov process (EMP), if*

- *X_t is time homogeneous,*
- *X_t is calibrated: $Prob[X(t) \neq X(t+1)] = 0.01$,*
- *X_t is stationary: $\exists \pi$, such that $\pi Q = 0$ and X_t is distributed according to π,*
- *X_t is reversible: $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$ for all i, j, and t,*
- *$P(t)$ is positive: $p_{ij}(t) > 0$ for all i, j, and t.*

Note that an EMP is uniquely specified by Q or $(Pt)_{t \geq 0}$.

2.2. Parameter estimation for an EMP

In modeling the evolutionary process, we are faced with the problem of estimating the rate matrix Q of an EMP. Typically, for this purpose we have pairwise sequence alignments at hand that are assumed to be correct. Dayhoff *et al.* (1978) restrict their estimation procedure to pairs of sequences that are very close to each other. With the enormous number of divergent alignments available today, this would imply a large loss of information.

Thus, it is desirable to exploit sequence alignments of widely different evolutionary distances for purposes of estimating the EMP.

An appropriate estimator must account for the evolutionary divergence of each alignment in the data set. We simplify the problem by proposing two distinct estimation problems to be solved:

- For each pairwise sequence alignment an evolutionary time t has to be estimated.
- A rate matrix has to be estimated from the alignment data.

The two problems are circular; the estimation of evolutionary time requires a rate matrix and vice versa. Our approach is iterative and cycles between estimating the evolutionary distances between the sequences in an alignment and updating the current rate matrix.

3. ESTIMATING THE TIME OF DIVERGENCE

Estimation of evolutionary distances is a major issue in molecular evolution. Several approaches are described in the literature, including stationary and nonstationary models, the log-det formula, and models with rate variation among sites. For a comprehensive review including a list of references, see Gu and Li (1998). We apply a maximum-likelihood approach similar to the one described in Adachi and Hasegawa (1996a); see also Felsenstein (1993).

3.1. Maximum likelihood estimator

Assume that a rate matrix is given. For a pairwise sequence alignment $\mathbb{A}$, a single parameter t, i.e., the evolutionary time of divergence, needs to be estimated. For this one-dimensional problem, a maximum likelihood approach is feasible. By definition, the maximum likelihood estimator $\hat{t}$ for the time of divergence is the time t that maximizes the log likelihood $\mathcal{L}(t|\mathbb{A})$ given the alignment $\mathbb{A}$. Due to the reversibility, we count unordered pairs of aligned amino acids. This yields a symmetric data matrix N, where N_{ij} denotes the number of amino acid pairs $\{a_i, a_j\}$ in the alignment. For reasons of consistency, we set N_{ii} equal to twice the count of $\{a_i, a_i\}$.

Since $P(t)$ is differentiable in t, a necessary condition that $\hat{t}$ is a maximum of the likelihood is that it is a solution of

$$0 = \frac{d}{dt}\mathcal{L}(t|\mathbb{A}) = \sum_{i=1}^{20}\sum_{j=1}^{20} N_{ij}\frac{d}{dt}\log M(t)_{ij} \tag{4}$$

$$= \sum_{i=1}^{20}\sum_{j=1}^{20} N_{ij}\frac{d}{dt}\log(Fe^{tQ})_{ij},$$

where F is a diagonal matrix with the amino acid frequencies as entries.

Using Equation (3), we get

$$0 = \frac{d}{dt}\mathcal{L}(t|\mathbb{A}) = \sum_{i=1}^{20}\sum_{j=1}^{20} N_{ij}\frac{(P(t)Q)_{ij}}{P(t)_{ij}}. \tag{5}$$

A maximum likelihood estimator for the time parameter t is a solution of Equation (5), which can be found, e.g., by Newton's method.

3.2. Simulations

In order to validate the performance of the estimation, we did simulation. We studied the behavior of the estimator, in particular for different sample sizes varying between 100 and 1,000 sites. Let an arbitrary EMP specified by a rate matrix Q be given. Artificial alignments of known distances t were generated according to this process and their divergence times were reestimated.

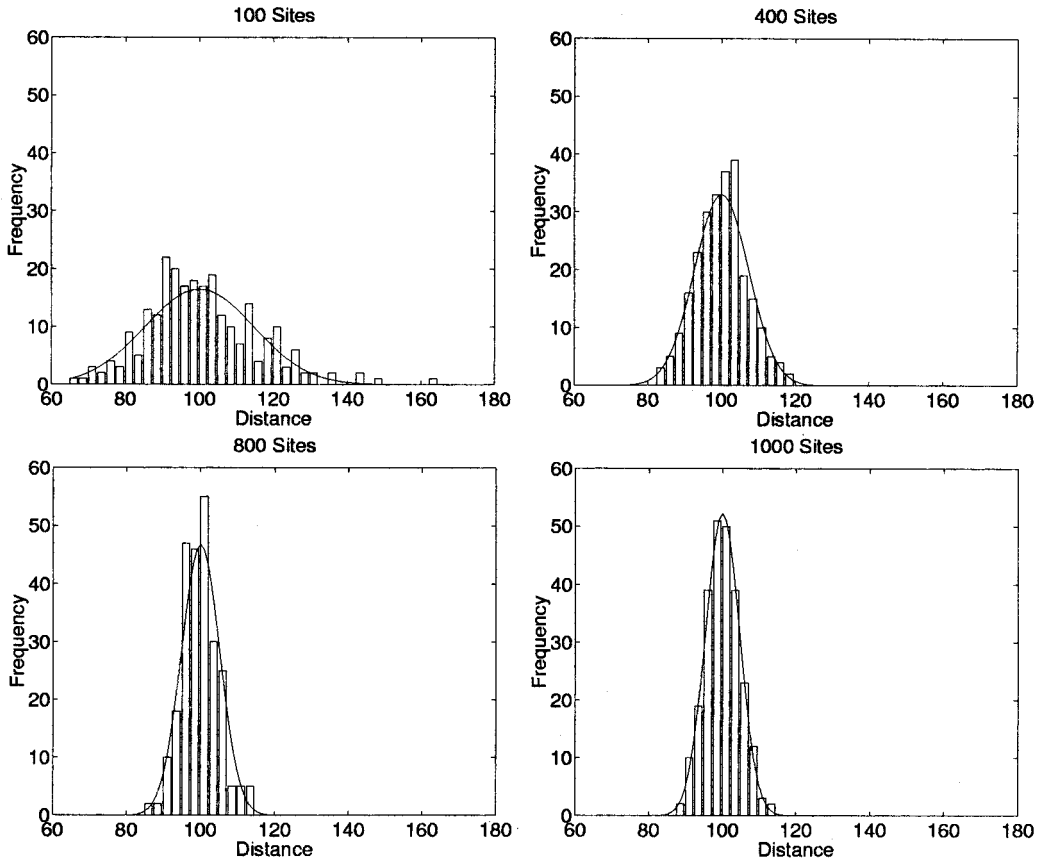

FIG. 2. Histograms of 250 time estimations of sequence alignments that are 100 PAM related.

We generated 1,000 pairwise sequence alignments of length 100, 400, 800, and 1,000, the sequences of each being 100 PAM apart. The resulting 250 estimates for each alignment length are plotted in the histograms in Figure 2. From the asymptotic theory, one would expect a normal distribution centered on the true distance of 100 PAM with variance equal to the inverse Fisher information of this alignment (Lindgren, 1993). The predicted distributions are also shown in this figure. This simulation shows that the normal approximation is also valid for practical alignment lengths. Theory and simulation agree on the wider spread of the distribution for shorter alignments. This coincides with the intuition that for small amounts of data the estimation is less reliable. Once the estimation is based on 800 or 1,000 sites, though, the distribution is sharply peaked around the true distance of 100 PAM, and most estimates remain within a window closely centered on this number.

4. ESTIMATION OF AN EMP

After having estimated the evolutionary distances for the given alignments one is confronted with the problem of estimating a rate matrix Q (the EMP) that explain in the best possible manner the observed exchange frequencies at their respective evolutionary distances. This question can, in principle, be solved by formulating a maximum-likelihood estimator for the EMP (Tavaré, 1986). In a phylogenetic framework this was suggested by Adachi and Hasegawa (1996b).

In order to specify this EMP, 210 parameters need to be estimated. Thus, optimizing this likelihood function is extremely computationally demanding, rendering this approach unpractical for large data sets. Here, we suggest an estimation procedure that is based on a representation of the rate matrix by the so-called resolvent of the transition probability matrices.

4.1. Resolvent method

Generally, the resolvent of a rate matrix Q is defined as the matrix $R_\alpha = (\alpha I - Q)^{-1}$, or, equivalently,

$$Q = \alpha I - R_\alpha^{-1} \tag{6}$$

independent of α.

One can show (see appendix) that the resolvent associated to an EMP equals the Laplace transform of transition probabilities of the process:

$$R_\alpha = \int_0^\infty e^{-\alpha t} P(t)\, dt. \tag{7}$$

The theory of resolvents is comprehensively described in Fukushima (1980) and in a more general framework in Ma and Röckner (1992).

Once the resolvent is computed, one can derive the rate matrix by applying Equation (6). The problem is to put this formalism to use in the estimation problem where we do not have perfect knowledge of all transition matrices but instead are given discrete sets of counts drawn at arbitrary distances.

Let a set of alignments $(\mathbb{A}_1, \cdots, \mathbb{A}_n)$ and the corresponding evolutionary distances $(t_1 \leq \cdots \leq t_n)$ be given. Our goal is to estimate an EMP from this data. We first estimate $P(t)$ by the empirical transition frequencies in the respective alignments. The estimator for $p_{ij}(t)$ is the count of all occurrences of (a_i, a_j) and (a_j, a_i) or twice (a_i, a_i), where each row is normalized by the corresponding amino acid frequency. This yields one estimated transition matrix $\hat{P}(t_i)$ for each time t_i.

The estimated transition matrices $\hat{P}(t_i)$ are used for estimating the resolvent R_α by approximating Equation (7). To this end, we split the integral (7) into n summands

$$(R_\alpha)_{ij} = \int_0^\infty e^{-\alpha t} p_{ij}(t)\, dt = \int_0^{t_1} e^{-\alpha t} p_{ij}(t)\, dt + \ldots + \int_{t_n}^\infty e^{-\alpha t} p_{ij}(t)\, dt,$$

and substitute each summand by the linear approximation

$$\int_{t_k}^{t_{k+1}} e^{-\alpha t} p_{ij}(t)\, dt \approx \int_{t_k}^{t_{k+1}} e^{-\alpha t} \left(y_k + \frac{y_{k+1} - y_k}{t_{k+1} - t_k}(t - t_k) \right) dt, \tag{8}$$

where $y_k = \hat{p}_{ij}(t_k)$. For a given parameter α, this sum is easily calculated.

Theoretically, the rate matrix is independent of α, but in the context of our approximation it is not. Therefore we require a rationale to select α. Again, maximum likelihood can yield a method of selecting α by maximizing the likelihood of the observed amino acid exchange counts. From Equations (6) and (4), we get the following expression for the log likelihood of n independent alignments

$$\mathcal{L}(\alpha | \mathbb{A}_1, \cdots, \mathbb{A}_n) \approx \sum_{k=1}^{n} \sum_{i=1}^{20} \sum_{j=1}^{20} N_{ij}^{(k)} \log(\pi_i (e^{t_k (\alpha I - \hat{R}_\alpha^{-1})})_{ij}), \tag{9}$$

where $\hat{R}_\alpha$ denotes the resolvent depending on the parameter α. This is not exact, because it relies on the estimate $\hat{R}_\alpha$ of the resolvent. Maximization of this expression for α yields the value that we use to estimate the rate matrix.

We integrate the above methods into the following overall procedure for the estimation of an EMP:

1. Choose a starting rate matrix **Q**, e.g., Dayhoff's model.
2. Estimate the pairwise evolutionary distances $\hat{t}_k$ of the given alignments $\mathbb{A}_k$ using **Q** according to Section 3.1.
3. Calculate empirical transition matrices $\hat{P}(\hat{t}_k)$.

4. Estimate the resolvent by a piecewise linear approximation between the estimated transition matrices according to (8).
5. Estimate the unknown parameter α by maximizing the likelihood (9).
6. Calculate the rate matrix of the EMP by Equation (6).
7. Iterate steps 2–6 until convergence is achieved.

Possibly the most surprising feature of the resolvent-based estimation procedure is that we have uncoupled the entries (i, j) of the transition matrices. An entry (i, j) of the rate matrix is computed using only (i, j) entries from transition matrices at other time points. This is the key to the computational simplicity of the method. In a maximum likelihood approach, on the other hand, one cannot achieve such an uncoupling but has to solve the high-dimensional optimization problem that reflects the interaction between all possible transitions.

4.2. Simulations

To evaluate the performance of the resolvent-based estimator, we generated alignments according to some fixed EMP but of various degree of divergence and reestimated the EMP parameters from the sample. We generated 5,000 pairs of letters for each PAM distance $10, 20, \cdots, 300$.

The first experiment shows the power of the resolvent in recovering a rate matrix. We demonstrate this by focusing on particular entries of a rate matrix. Figure 3 illustrates that the entries $(R_\alpha)_{ij}$ of the resolvent are well fitted by the exponentially-weighted empirical transition probabilities for an optimal choice of α. The thick line shows the theoretical values of $e^{-\alpha t} p_{ij}(t)$ based on the parameters of the simulation. This is approximated by a dashed line representing the connected, estimated values $e^{-\alpha t} \hat{p}_{ij}(\hat{t})$. The top figure

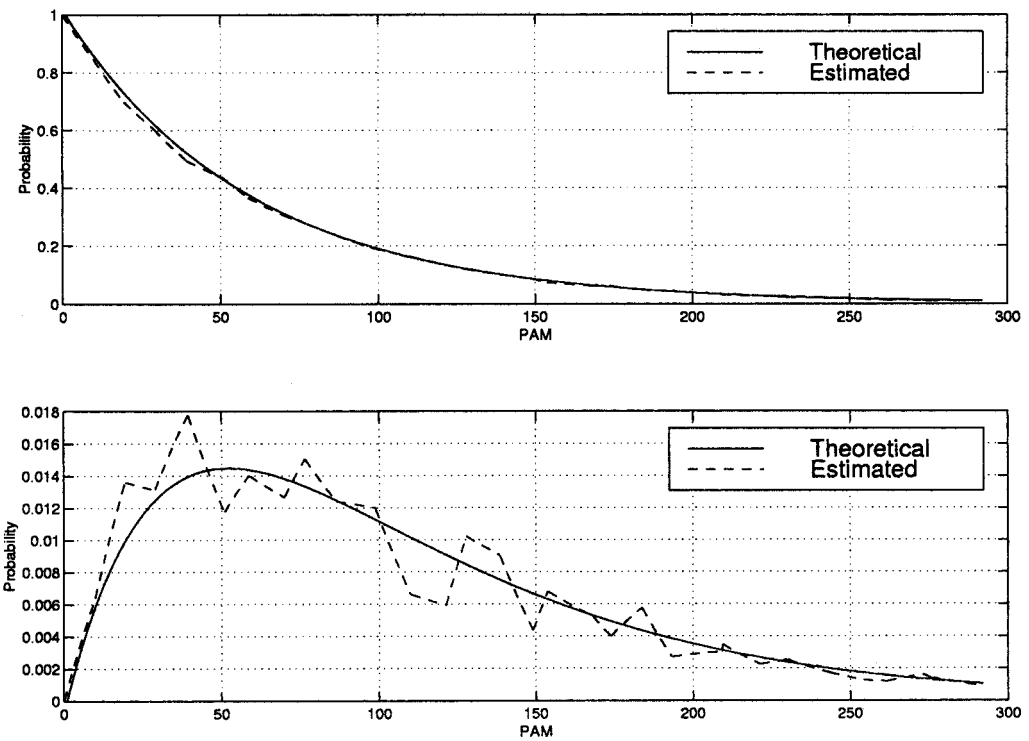

FIG. 3. Approximation of the weighted substitution probabilities $e^{-\alpha t} P(t)$ for different PAM distances as they appear in the integral that defines the resolvent (7). The thick line shows the true function generated with the parameters of the simulation. The dotted line depicts the corresponding observed counts weighted by the exponential function. The **top** figure is for a leucine-leucine match ($e^{-\alpha t} p_{LL}(t)$) while the **bottom** figure shows data for a leucine-methionine exchange ($e^{-\alpha t} p_{LM}(t)$).

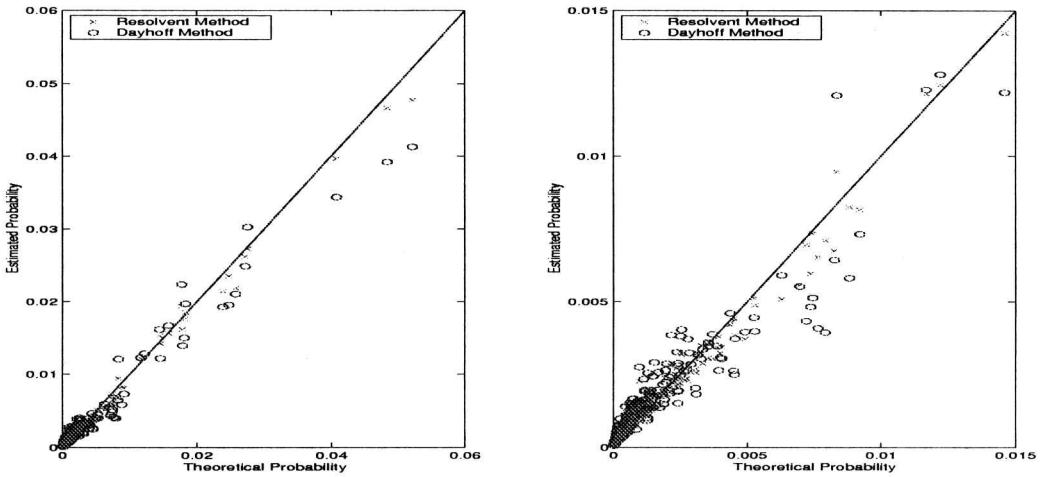

FIG. 4. Amino acid pair frequencies. The estimations are based on a sample of 150,000 letter pairs varying from 10–300 PAM. The pair frequencies are displayed at 100 PAM ($M(100)$). The true values (horizontal axis) are plotted versus the computed probabilities (vertical axis). The **right** plot is a zoom into the lower left corner of the **left** plot. Crosses denote results of our method while empty circles stand for Dayhoff's method.

shows these data for two identical amino acids (leucine vs. leucine), while on the bottom methionine-leucine exchanges are shown. For the identical pair there is hardly any fluctuation of the estimator around the true curve due to the large amount of data available. For the exchange, there is fluctuation but the areas under the two curves agree well.

It is difficult to compare our method to others, since we are not aware of another method that assigns the data points to their own evolutionary distance. In extensive comparisons to a more rigorous yet slow maximum likelihood method (data not shown), we have found the two approaches to be almost equally good. Here, we need to restrict ourselves to the comparison of the resolvent-based estimator to Dayhoff's method although the latter is theoretically less well suited to solve the problem than ours.

Figure 4 relates the entries of the matrix $M(100)$ according to the true model parameters with the estimated values. As introduced in Section 2.1, the matrix $M(100)$ describes the probability of observing a particular pair of amino acids after 100 time units. The full circles describe the resolvent-based estimator. The asterisks refer to an implementation of Dayhoff's method, where all available data were used. In reality, this would correspond to using Dayhoff's estimator on a large set of alignments, ignoring their individual evolutionary distances. The left plot shows the original data while the right one focuses on small values. It is obvious that the resolvent-based method recovers the original parameters with high accuracy.

5. RESULTS

5.1. The variable time score matrix family VT

A score matrix assigns a real number, a similarity score, to each pair of amino acids. Specifying an appropriate amino acid score matrix is central to protein comparison methods, and much effort has been devoted to defining, analyzing, and refining such matrices. We have applied the resolvent-based estimation procedure to a large set of protein sequence alignments in order to derive a matrix that is as accurate as possible. The result of this procedure is primarily a new rate matrix. Transition matrices and score matrices for the various PAM distances can be derived from this rate matrix.

Usually, score matrices are log-odds ratios of the form

$$S_{ij} = \log\left(\frac{m_{ij}}{\pi_i \pi_j}\right) \tag{10}$$

(Dayhoff *et al.*, 1978), where m_{ij} is the relative frequency of the amino acid pair (a_i, a_j) in a representative set of alignments and π_i denotes the overall frequency of amino acid a_i. By convention, score matrices are calculated by $S_{ij} = 10\log(m_{ij}/(\pi_i\pi_j))$. Multiplying the score matrix (and the gap cost) by some constant does not affect the optimal alignment. This ratio compares the probability of an event occurring under two alternative hypothesis; the numerator constitutes a model for related proteins, whereas the denominator is the probability that two amino acids are selected by chance. Obviously, the major issue is the estimation of the relative frequencies m_{ij}.

Our alignment data were extracted from the SYSTERS database (Krause and Vingron, 1998). SYSTERS provides a large-scale clustering of more than 100,000 protein sequences into families, each of which have automatically been multiply aligned. In order to avoid biases due to large families, we randomly selected one pair of aligned sequences from each cluster. We obtained on the order of 2.7 million pairs of aligned amino acids of which approximately 1 million are mismatch positions. The alignment data were binned into classes of highly similar estimated evolutionary distances ensuring that in each class the off-diagonal matrix entries are well populated. This has resulted in 80 such bins covering PAM distances from 1 to approximately 300. We used such distant alignments because they constitute distant pairs from within large multiple alignments and as such are still credible. We needed to apply the cycling over the resolvent-based estimation and the time estimation only three times for the rate matrix estimate to settle down. In this process, also the weight parameter α was estimated. Interpretation of the resulting exponential weight function shows that approximately 90% of the weight is put on alignments of a distance up to approximately 140 PAM while the alignments below 30 PAM account for only 22%. The result of our estimation procedure is a new rate matrix Q (see Table 1).

From Q, the stationary amino acid distribution π (see Table 3) and the relative pair frequencies

$$m_{ij}(t) = \pi_i \exp(tQ)_{ij}$$

can be derived. Score matrices can be derived using Equation (10). We call these score matrices $VT(t)$ (variable time), where t is the degree of evolutionary divergence that we are focusing on. The lower triangle of the matrix depicted in Table 2 shows the entries of the score matrix VT160.

The alignments used for the derivation of the VT series are the product of an alignment procedure which itself has used a score matrix. Therefore we wanted to check for the possible influence of this procedure on our result. We therefore also applied our estimation procedure to pairwise alignments extracted from the 3D-ALI database of structural alignments (see Pascarella and Argos, 1992). This data set should be independent of an a priori choice of a score matrix since the alignments have been derived from (or at least checked against) structural superpositions of proteins. The rate matrix derived from this data set agrees very well with the rate matrix derived from the SYSTERS alignments. We thus exclude the possibility of having reproduced the score matrix that was used in those alignments.

5.2. Comparison with other score matrices

The widely accepted BLOSUM series of score matrices has been derived by Henikoff and Henikoff (1992). They obtained amino acid substitution frequencies directly from multiple sequence alignments from the blocks database (Henikoff and Henikoff, 1991). Although not based on an evolutionary model, their alignments are selected from certain intervals of evolutionary distance. This is in contrast to the approaches of Dayhoff *et al.* (1978), Jones *et al.* (1992), and Gonnet *et al.* (1992) that rely on Markov models for protein evolution.

Here we focus on the comparison of our matrix with Dayhoff's PAM series and with the BLOSUM series of matrices. Such a comparison relies on identifying the comparable representatives from the respective series. According to Altschul (1991) and Henikoff and Henikoff (1992), the PAM160 and BLOSUM62 score matrices correspond to the same degree of divergence. Thus we compared our VT160 matrix (Table 2, lower triangle) to these two matrices. Figure 5 shows the correlation between the entries of VT160 with BLOSUM 62 and PAM160, respectively. It is apparent that VT160 is well correlated with BLOSUM62 while there is considerable difference to PAM160. Inspecting the differences between our matrix and each of the two others as given in the top triangle of the matrix in Table 2, this impression is reinforced. Especially for rare amino acids like tryptophan the PAM matrix seems to be the outlier. Due to the large data set and the improved derivation technique, we assume that the new data are more reliable.

Table 1. The Estimated Rate Matrix Q Multiplied by 10^{6a}

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	•	317	290	376	182	462	753	1115	107	277	540	450	213	146	666	2301	1520	36	91	1569
R	485	•	372	209	107	1146	435	438	546	134	417	3005	139	89	318	692	386	97	148	268
N	471	398	•	2148	71	624	643	803	695	276	202	1332	103	105	168	2217	1123	12	180	155
D	540	190	1848	•	16	341	3128	641	240	66	177	436	44	34	257	796	435	16	122	157
C	969	363	233	64	•	97	153	367	168	287	708	104	137	353	105	1291	509	104	347	766
Q	866	1404	712	457	33	•	2180	293	886	153	660	1822	213	75	560	835	624	58	153	218
E	917	346	475	2644	31	1426	•	462	182	123	278	1303	83	47	363	658	459	27	95	303
G	1306	334	568	521	81	182	447	•	95	90	180	271	66	77	187	1100	208	60	56	210
H	375	1258	1446	585	113	1669	527	285	•	178	535	543	153	288	448	708	468	41	1272	183
I	337	107	216	58	64	97	128	93	63	•	2975	186	767	479	113	216	768	35	140	5092
L	418	213	97	98	102	277	181	118	120	1810	•	168	903	918	249	326	293	98	204	1202
K	584	2559	1030	399	24	1258	1395	297	202	190	284	•	192	82	334	711	733	38	130	264
M	747	321	225	115	90	399	244	201	155	2012	3979	516	•	496	120	388	969	97	182	1373
F	267	107	120	47	126	74	77	124	151	691	2161	118	256	•	90	403	199	215	1708	580
P	1067	332	171	297	32	477	493	260	203	150	492	423	55	79	•	1266	562	28	80	205
S	2525	488	1430	607	271	481	588	1023	219	183	440	596	118	230	866	•	2554	43	157	316
T	2082	340	912	412	128	447	509	237	179	785	508	772	378	144	473	3182	•	29	142	1164
W	213	389	42	68	121	185	135	312	69	172	753	175	162	705	103	246	129	•	748	191
Y	208	225	255	205	162	188	190	112	867	264	620	234	121	2185	115	345	246	292	•	339
V	1837	197	109	125	167	132	289	207	59	4449	1767	236	458	358	144	336	990	37	165	•

[a]The diagonal equals minus the sum of the row.

TABLE 2. DATA FOR VT160

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3,1,−1	1,0	−1,1	−1,1	2,0	0,0	−1,0	−1,0	1,0	0,0	1,−1	1,0	0,0	1,0	−1,1	0,0	0,1	2,0	1,−1	0,0
R	−1	6,−1,1	1,0	1,1	2,1	0,0	2,−1	2,1	0,1	−1,0	1,0	−1,1	−1,−1	1,0	0,1	0,0	0,0	−3,2	2,0	1,1
N	−1	0	5,1,−1	0,1	3,1	0,0	−1,0	0,0	−1,0	0,0	0,0	−1,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1
D	−1	−1	2	5,1,−1	2,0	−1,0	−1,0	0,1	0,1	0,−1	1,0	0,1	0,0	2,−1	1,0	0,0	0,0	2,0	2,0	0,0
C	0	−2	−2	−3	10,0,1	3,1	3,1	2,1	2,2	1,0	5,0	3,0	4,0	5,1	1,0	0,1	2,0	7,1	0,2	2,0
Q	−1	1	0	0	−2	5,−1,0	−1,−1	1,0	−1,2	0,0	0,0	1,0	0,−1	2,0	0,1	1,0	1,0	3,0	2,−1	0,0
E	−1	−1	0	2	−3	1	5,0,0	−1,1	−1,−1	−1,0	0,0	2,0	0,0	1,−1	0,0	−1,−1	0,0	4,−1	2,−1	0,0
G	0	−1	0	0	−2	−2	−1	6,2,0	1,0	0,0	1,0	1,0	0,0	1,−1	−1,0	−1,0	0,1	5,0	2,−1	0,0
H	−2	1	1	0	−1	2	−1	−2	7,1,−1	1,0	0,1	1,1	2,0	1,0	0,1	0,0	1,1	1,0	2,0	0,0
I	−1	−3	−3	−4	−1	−3	−3	−4	−3	4,−1,0	0,0	−1,0	0,1	0,0	−1,0	0,0	0,1	3,1	1,−1	0,0
L	−2	−2	−3	−4	−1	−2	−3	−4	−2	2	4,−1,0	1,−1	−1,0	0,1	1,1	1,0	1,0	1,1	1,0	0,0
K	−1	3	1	0	−3	1	1	−2	0	−3	−3	5,0,0	−2,−1	2,0	1,0	1,−1	0,1	2,0	2,0	1,0
M	−1	−2	−2	−3	−1	−1	−2	−3	−2	2	2	−2	6,−1,1	0,0	0,−1	1,−1	1,1	3,0	2,0	0,0
F	−2	−3	−3	−4	−1	−3	−4	−4	−1	0	1	−3	0	7,−1,1	1,1	1,0	1,0	2,0	−2,1	2,1
P	0	−1	−2	−1	−3	0	−1	−2	−1	−3	−2	−1	−3	−3	7,1,0	−1,1	0,1	2,0	2,0	0,0
S	1	−1	1	0	0	0	−1	0	−1	−2	−2	−1	−2	−2	0	3,1,−1	0,0	−1,0	1,0	0,1
T	1	−1	0	−1	−1	−1	−1	−1	−1	0	−1	0	0	−2	0	1	4,0,−1	2,−1	1,0	0,0
W	−3	−1	−4	−4	−1	−2	−4	−2	−2	−2	−1	−3	−1	1	−4	−3	−3	12,−1,1	3,0	4,0
Y	−3	−2	−2	−3	0	−2	−3	−4	2	−2	−1	−2	−1	4	−3	−2	−2	2	8,−1,1	2,−1
V	0	−2	−2	−3	−1	−2	−2	−3	−3	3	1	−2	1	0	−2	−1	0	−3	−2	4,0,0

The lower triangle denotes the score matrix VT160. Log-odds are calculated by $S_{ij} = 10 \log(m_{ij}/(\pi_i \pi_j))$ and are rounded to the nearest integer. The upper triangle contains the differences VT160 − PAM160 and VT160 − BLOSUM62.

TABLE 3. THE STATIONARY AMINO ACID DISTRIBUTION

A	0.0771	L	0.0976
R	0.0501	K	0.0592
N	0.0462	M	0.0221
D	0.0538	F	0.0414
C	0.0146	P	0.0477
Q	0.0409	S	0.0707
E	0.0634	T	0.0568
G	0.0656	W	0.0127
H	0.0219	Y	0.0324
I	0.0592	V	0.0669

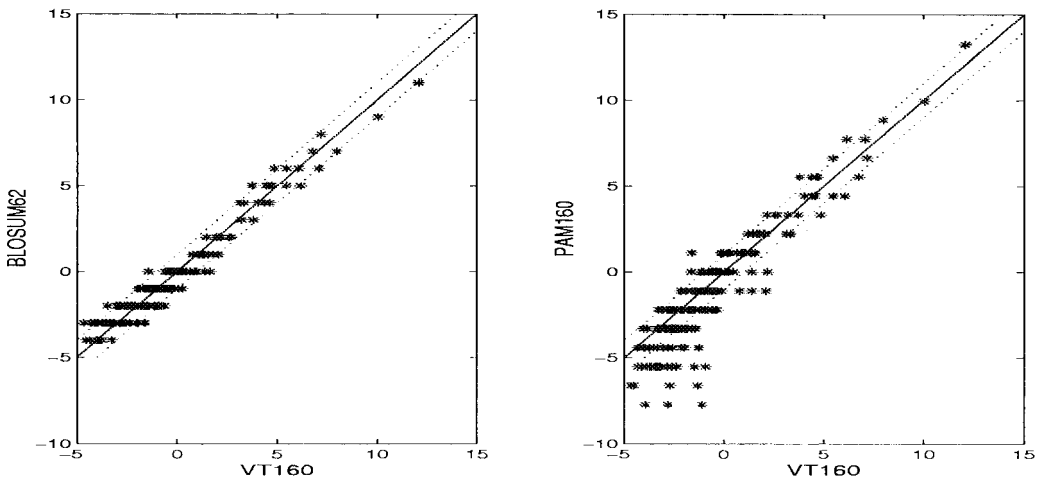

FIG. 5. The scores of VT160 against those of BLOSUM62 (**left**) and PAM160 (**right**). The dotted lines indicate a difference of one unit of score.

6. DISCUSSION

We have presented a new method to estimate the parameters of an amino acid substitution model. Following Dayhoff *et al.* (1978), we model protein evolution as a Markov process. In contrast to Dayhoff's estimation method, our estimator can handle alignment data of various degrees of divergence. This is of particular importance for exploiting information from the available distant sequence alignments. Our approach is iterative and cycles between estimating the evolutionary distances between the sequences in an alignment and updating the current estimator for the rate matrix.

The key to our estimation procedure is that we have avoided a full-fledged maximum likelihood computation by approximating the resolvent of the transition matrices. This technique has, to our knowledge, not been applied in this context before. It results in a tremendous simplification of the computation with a subsequent acceleration by several orders of magnitude. From a theoretical standpoint it is worth noting that the resolvent allows one to estimate the entries of the inverse of the rate matrix in a component-wise fashion. Interaction between the different transitions is only reflected in the subsequent inversion.

Adachi and Hasegawa (1996b) have approached similar estimation problems for a particular family given the phylogenetic tree of this family. Since we are studying many families simultaneously, we cannot assume that there exists one phylogenetic tree for our sequences. On the other hand, our method can also be applied to single families, particularly in the absence of a tree. However, when drawing many alignments from one family, these alignment will not be independent of each other. Avoiding this situation was also behind the decision to draw only one sequence pair from each SYSTERS family in our estimation procedure. Methods to weight sequence pairs in order to correct for the bias from sampling the same family repeatedly are under development.

The overall matrix that we derived from the SYSTERS database of protein families has been compared to other available matrices. It appears that at the respective time points where they are comparable, BLOSUM and VT are fairly similar. Our matrix, however, stems from a rate matrix. Thus, one can compute VT score matrices for every evolutionary distance. Furthermore, the new rate matrix can be applied in the context of maximum likelihood tree construction (e.g., PROTML (Adachi and Hasegawa, 1996a) or PUZZLE (Strimmer *et al.*, 1997)), where a good estimator of transition probabilities is essential.

The rate matrix, the overall amino acid frequencies, and the VT series of score matrices can be downloaded from *http://www.dkfz-heidelberg.de/tbi/people/tmueller/*. Matlab routines for the estimation procedure can be obtained from the first author.

APPENDIX: ASSOCIATION OF THE RESOLVENT TO THE RATE MATRIX

Fukushima (1980) and Ma and Röckner (1992) prove that the rate matrix Q of a Markov chain can be recovered from its resolvent in a very general setup of processes on infinite dimensional spaces. We give a simplified proof in the case of a finite state space Markov process.

Theorem 1. *The resolvent of transition probabilities of a Markov chain*

$$R_\alpha = \int_0^\infty e^{-\alpha t} P(t)\, dt$$

is related to its rate matrix by

$$Q = \alpha I - R_\alpha^{-1} \tag{11}$$

independent of the choice of $\alpha > 0$. In particular, R_α and R_β commute for $\alpha, \beta > 0$.

Proof. We need to show that Equation (11) is well defined. This means, R_α is invertible for all $\alpha > 0$. To do this, we need some more properties of the resolvent.

We start with a given family of transition matrices $(P(t))_{t\geq 0}$ fulfilling property (1). Using this and Lebesgue's theorem, we get

$$\lim_{\beta\to\infty} \beta R_\beta \lim_{\beta\to\infty} \int_0^\infty e^{-s} P(s/\beta)\, ds = I. \tag{12}$$

A crucial property of $(R_\alpha)_{\alpha>0}$ is the first resolvent equation

$$R_\alpha - R_\beta + (\alpha - \beta) R_\alpha R_\beta = 0 \text{ for all } \alpha, \beta > 0. \tag{13}$$

By definition of the resolvent and the Chapman–Kolmogorov equation, we get

$$\begin{aligned}(\beta - \alpha) R_\alpha R_\beta &= (\beta - \alpha) \int_0^\infty e^{-\alpha t} P(t) \int_0^\infty e^{-\beta s} P(s)\, ds\, dt \\ &= (\beta - \alpha) \int_0^\infty e^{(\beta-\alpha)t} \int_t^\infty e^{-\beta s} P(s)\, ds\, dt.\end{aligned}$$

But the right hand side is equal to

$$\begin{aligned}(\beta - \alpha) R_\alpha R_\beta &= (\beta - \alpha) \int_0^\infty e^{-\beta s} P(s) \int_0^s e^{(\beta-\alpha)t}\, dt\, ds \\ &= \int_0^\infty e^{-\beta s} P(s)(e^{(\beta-\alpha)s} - 1)\, ds \\ &= R_\alpha - R_\beta\end{aligned}$$

for all $\alpha, \beta > 0$. Interchanging the role of α and β in the proof shows that R_α and R_β commute.

In the sequel, we show that the resolvent is invertible as well as that the generator of the resolvent (11) coincides with the rate matrix of the process (2).

Let $R_\alpha u = 0$, then $R_\beta u = 0$ for all $\beta > 0$ by Equation (13). Using property (12) yields

$$u = Iu = \lim_{\beta \to \infty} \beta R_\beta u = 0$$

showing that R_α is injective, therefore surjective and thus bijective and invertible.

So far we have shown that the expression (11) is well defined for all $\alpha > 0$. The corresponding matrix Q_α is also independent of the choice of $\alpha > 0$, because for $u = R_\alpha v$ we get

$$\begin{aligned} R_\beta((\alpha - R_\alpha^{-1})u - (\beta - R_\beta^{-1})u) &= \alpha R_\beta R_\alpha v - R_\beta v - \beta R_\beta R_\alpha v + R_\alpha v \\ &= R_\alpha v - R_\beta v + (\alpha - \beta) R_\beta R_\alpha v \\ &= 0. \end{aligned}$$

By the injectivity of R_α, we see that $Q_\alpha = Q_\beta$ for all $\alpha, \beta > 0$.

It remains to identify the rate matrix Q_{R_α} associated to the resolvent by Equation (11) with the rate matrix of the Markov process $Q_{P(t)}$.

Due to the surjectivity of R_α, it suffices to show the assertion for $u = R_\alpha v$. By the definition of the rate matrix of a Markov process (2) and by the definition of u, we get

$$\begin{aligned} Q_{P(t)} u &= \lim_{t \searrow 0} \frac{1}{t}(P(t)u - u) \\ &= \lim_{t \searrow 0} \frac{1}{t} e^{\alpha t} \int_t^\infty e^{-\alpha s} P(s) v \, ds - \frac{1}{t} \int_0^\infty e^{-\alpha s} P(s) v \, ds \\ &= \frac{d}{dt} \left(e^{\alpha t} \int_t^\infty e^{-\alpha s} P(s) v \, ds \right) \Bigg|_{t=0} \\ &= \alpha \int_0^\infty e^{-\alpha s} P(s) v \, ds - v. \end{aligned}$$

By the definition of the resolvent and by using the invertibility of R_α, we get:

$$\begin{aligned} \alpha \int_0^\infty e^{-\alpha s} P(s) v \, ds - v &= \alpha R_\alpha v - v \\ &= \alpha u - R_\alpha^{-1} u \\ &= Q_{R_\alpha} u \end{aligned}$$

and therefore $Q_{P(t)} = Q_{R_\alpha}$. So we identify the rate matrix associated to the process with the rate matrix associated to the resolvent:

$$\alpha I - \left(\int_0^\infty e^{-\alpha t} P(t) \, dt \right)^{-1} = Q = \lim_{t \searrow 0} \frac{1}{t}(P(t) - I)$$

for all $\alpha > 0$. ■

ACKNOWLEDGMENTS

We would like to thank Rainer Spang and Marc Rehmsmeier for many helpful discussions.

REFERENCES

Adachi, J., and Hasegawa, M. 1996a. *Molphy: Programs for molecular phylogenetics*, ver. 2.3, Institute of Statistical Mathematics, Tokyo.

Adachi, J., and Hasegawa, M. 1996b. Model of amino acid substitution in protein encoded by mitochondrial DNA. *J. Mol. Evol.* 42, 459–468.

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Benner, S., Cohen, M., and Gonnet, G. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* 7, 1323–1332.

Dayhoff, M., Schwartz, R., and Orcutt, B. 1978. A model of evolutionary change in protein. *Atlas Protein Sequences Structure* 5, 345–352.

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* 5, 164–166.

Felsenstein, J. 1993. *PHYLIP manual, ver. 3.5c.* Department of Genetics, University of Washington Seattle.

Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* 155, 279–284.

Fukushima, M. 1980. *Dirichlet Forms and Markov Processes*, North Holland, New York.

Gonnet, G., Cohen, M., and Benner, S. 1992. Exhaustive matching of the entire protein sequence database. *Science* 256, 1443–1445.

Grimmet, G.R., and Stirzaker, D.R. 1992. *Probability and Random Processes.* Oxford Science Publications, New York.

Gu, X., and Li, W.-H. 1998. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Nat. Acad. Sci. USA* 95, 5899–5905.

Henikoff, S., and Henikoff, J.G. 1991. Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* 19(3), 6565–6572.

Henikoff, S., and Henikoff, J. 1992. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA* 89, 10915–10919.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8, 275–282.

Krause, A., and Vingron, M. 1998. A set-theoretic approach to database searching and clustering. *Bioinformatics* 14(5), 430–438.

Lindgren, W. 1993. *Statistical Theory*, Chapman and Hall, New York.

Ma, Z.M., and Röckner, M. 1992. *Dirichlet Forms*, Springer Verlag, New York.

McLachlan, A. 1971. Tests for comparing related amino acid sequences. cytochrome *c* and cytochrome c_{551}. *J. Mol. Biol.* 61, 409–424.

Overington, J., Johnson, M., Sali, A., and Blundell, T. 1990. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc. R. Soc. Lond.* 241, 132–145.

Pascarella, S., and Argos, P. 1992. A data bank merging related protein structures and sequences. *Prot. Eng.* 5, 121–137.

Pearson, W.R., and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Nat. Acad. Sci. USA* 85, 2444–2448.

Rao, M. 1987. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. Peptide Protein Res.* 29, 276–281.

Risler, J., Delorme, M., Delacroix, H., and Henaut, A. 1988. Amino acid substitutions in structurally related proteins. *J. Mol. Biol.* 204, 1019–1029.

Strimmer, K., Goldman, N., and von Haeseler, A. 1997. Bayesian probabilities and quartet puzzling. *Mol. Biol. Evol.* 14, 210–213.

Swofford, D.L., and Olsen, G.J. 1990. Phylogeny reconstruction, in *Molecular Systematics*, Sinauer Associates, Sunderland, MA.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures Mathematics Life Sciences* 17, 57–86.

Waterman, M.S. 1995. *Introduction to Computational Biology*, Chapman and Hall, London.

Address correspondence to:
Tobias Müller
Deutsches Krebsforschungszentrum
Theoretische Bioinformatik
IM Neuenheimer Feld 280
69120 Heidelberg, Germany

E-mail: {t.mueller/m.vingron}@dkfz-heidelberg.de