Estimating Amino Acid Substitution Models: A Comparison of Dayhoff's Estimator, the Resolvent Approach and a Maximum Likelihood Method

Tobias Müller, Rainer Spang,† and Martin Vingron**

*Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, Im Neuenheimer Feld 280, Heidelberg, Germany and †Duke University, Institute of Statistics and Decision Sciences, Durham, North Carolina

Evolution of proteins is generally modeled as a Markov process acting on each site of the sequence. Replacement frequencies need to be estimated based on sequence alignments. Here we compare three approaches: First, the original method by Dayhoff, Schwartz, and Orcutt (1978) Atlas Protein Seq. Struc. **5**:345–352, secondly, the resolvent method (RV) by Müller and Vingron (2000) J. Comput. Biol. **7**(6):761–776, and finally a maximum likelihood approach (ML) developed in this paper. We evaluate the methods using a highly divergent and inhomogeneous set of sequence alignments as an input to the estimation procedure. ML is the method of choice for small sets of input data. Although the RV method is computationally much less demanding it performs only slightly worse than ML. Therefore, it is perfectly appropriate for large-scale applications.

Introduction

Differences between homologous proteins are the result of a mutation process starting from a common, though unknown ancestor. In a mutation event, an amino acid at a certain position in a protein is replaced by another one. These exchanges are constrained by the requirement to maintain protein structure or function. Certain mutations tend to have little effect in this respect and are observed more frequently than exchanges that clearly influence the protein structure.

If this replacement improves the fitness of the organism the new amino acid will be accepted by natural selection. Replacements of very dissimilar amino acids often drastically change the fold of the protein, leading to a complete loss of function. Hence, such mutations are less often observed than those of similar amino acids, which have only slight effects with respect to fold and function. Therefore, amino acid similarity is reflected in replacement frequencies. It is important to note that actual mutation counts depend not only on these similarities but also on the degree of divergence of the sequences that one compares. We thus need a model which describes protein evolution as a function of time.

Modeling amino acid replacements by a Markov chain has been introduced by Dayhoff, Schwartz, and Orcutt (1978). In this approach a set of identical Markov chains acting independently on each site of the protein is used. The time index of the process is interpreted as a measure of evolutionary divergence. The challenge is to estimate the parameters of the process from divergent and time-inhomogeneous sequence data.

In the original approach of Dayhoff, Schwartz, and Orcutt (1978) the actual estimation is restricted to only very closely related pairs of sequences. However, once a Markov model is fitted to this data, replacement frequencies characteristic for distantly related sequences can be extrapolated from the model.

Dayhoff's approach has been generalized and applied to larger data sets (Gonnet, Cohen, and Benner 1992; Jones, Taylor, and Thornton 1992). Furthermore, the advent of large numbers of structurally derived alignments has raised interest in using information also from very distant related alignments (Overington et al 1990; Risler et al 1988). However, these authors do not provide an estimation procedure which would account for the time-inhomogeneity of the input data.

Benner, Cohen, and Gonnet (1994) were the first to point out the problem of estimating one consistent model from an inhomogeneous pool of alignment data. They sketch a normalization algorithm, which is based on computing logarithms of transition matrices which they approximate by power series. The approach is heuristic, as the convergence of the power series cannot be guaranteed for empirically derived matrices.

Müller and Vingron (2000) present a rigorous estimation procedure which is based on an entirely different mathematical formalism. We refer to this method as resolvent method and briefly review it in the *Resolvent Method* section. Alternatively, we describe a novel maximum likelihood based approach. The *Maximum Likelihood* section provides the details of the mathematical formalisms and computations.

In principle, one has two important, although conflicting, criteria for evaluating the quality of the methods. For large-scale applications, time performance of the algorithms is crucial, whereas low statistical efficiency of the estimator can be compensated for by the huge amount of data that is used. On the other hand, if one is restricted to only a small set of input data, the accuracy of the estimator is more important. In the *Results* section we discuss the individual merits of the resolvent and the maximum likelihood estimator.

Model

Let $P(t)$ be the transition probability matrix of a time continuous Markov chain with entries $p_{ij}(t) = \text{Prob}(X[s+t] = j \mid X[s] = i)$. We consider only Markov chains for which the rate matrix $Q = \lim_{t \searrow 0} (P[t] -$

Key words: amino acid replacement, amino acid score matrix, maximum-likelihood, protein evolution.

Address for correspondence and reprints: Tobias Müller, Deutsches Krebsforschungszentrum, Theoretische Bioinformatik, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. E-mail: t.mueller@dkfz.de.

Mol. Biol. Evol. 19(1):8–13. 2002

$I)/t$ exists, where I denotes the identity matrix. Hence, we get the following linear approximation

$$P(t) = I + tQ + o(t), \tag{1}$$

for small t. The resolvent

$$R_\alpha = \int_0^\infty e^{-\alpha t} P(t)\, dt \tag{2}$$

provides an alternative characterization of the rate matrix. From the Chapman-Kolmogorov equation, we get the forward and backward equations $(d/dt)P(t) = P(t)Q = QP(t)$. Differentiating $e^{-\alpha t}P(t)$ using the product rule for matrix differentiation we get

$$\alpha e^{-\alpha t}P(t) + \frac{d}{dt}(e^{-\alpha t}P[t]) = e^{-\alpha t}\frac{d}{dt}P(t)$$

for $\alpha > 0$, $t \geq 0$. Using this we obtain:

$$\alpha \int_0^\infty e^{-\alpha t}P(t)\, dt + \int_0^\infty \frac{d}{dt}(e^{-\alpha t}P[t])\, dt$$
$$= \int_0^\infty e^{-\alpha t}\frac{d}{dt}P(t)\, dt = \int_0^\infty e^{-\alpha t}P(t)\, dtQ.$$

Multiplication by R_α^{-1} reduces the above equation to

$$\alpha I - R_\alpha^{-1} = Q \qquad (\alpha > 0). \tag{3}$$

Note that R_α is in fact invertible for all $\alpha > 0$ (see Müller and Vingron 2000).

The forward and backward equations can be solved under the initial condition $P(0) = I$ and yield

$$P(t) = \exp(tQ). \tag{4}$$

This allows transition probabilities for any time of divergence t to be computed from the rate matrix.

The problem of modeling amino acid replacement frequencies requires additional assumptions on the Markov chain. Following Dayhoff, Schwartz, and Orcutt (1978) we model the evolution of each site of the proteins by a single time-homogeneous Markov chain $X(t)$, calibrated, such that on average 1% of the amino acids are changed after one unit of time: $\text{Prob}[X(t) \neq X(t + 1)] = 0.01$. Once calibrated, the time t in the Markov chain can be used as a measure of evolutionary divergence. The acronym PAM (Point Accepted Mutations) is commonly used for this unit of divergence.

We assume that any amino acid can change into any other one. This is ensured by the requirement that the exchange rates q_{ij} be strictly positive for all i, j. Then there exists a unique limiting amino acid distribution $\pi = (\pi_1, \ldots, \pi_{20})$ where $\pi_j = \lim_{t\to\infty} p_{ij}(t) > 0$ is independent of the initial residue i. The distribution π fulfills the equations $\pi Q = 0$ and $\pi P(t) = \pi$ for all $t \geq 0$ and is called the equilibrium distribution. We only consider Markov processes which are in equilibrium. 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}). $M(t) = m_{ij}(t)$ 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. Whereas data from ancestor sequences are not available, we do observe pairs of proteins that have evolved from a common, though unknown, ancestor. We cannot decide the direction of the mutation, we only observe pairs of corresponding amino acids at certain positions in a protein. It would not be appropriate to model such a direction. Such processes are described by the class of time-reversible Markov chains. This means that the probability of being in amino acid a_i and going from a_i to a_j in time t is equal to that of being in amino acid a_j and going from amino acid a_j to a_i. Consequently, we get the detailed balance equation $\pi_i q_{ij} = \pi_j q_{ji}$. In particular, $M(t)$ is a symmetric matrix. We use the shortcut Evolutionary Markov Process (EMP) for a process satisfying all the conditions discussed earlier.

The transition and the rate matrix of an EMP have the following mathematical properties. Denote by F the diagonal matrix with entries π_i. Then F constitutes a symmetric, positive definite matrix and $\langle x, y\rangle_F = \langle x, Fy\rangle$ defines an inner product. Because of the reversibility, Q and $P(t)$ are selfadjoint relative to $\langle\cdot, \cdot\rangle_F$, i.e., $\langle P(t)x, y\rangle_F = \langle x, P(t)y\rangle_F$, and can therefore be transformed into diagonal form by a change of coordinates. The eigenvalues of Q are real by selfadjointness, and negative owing to Gershgorin's theorem. Using definition (4) we can rewrite Q and $P(t)$ as

$$Q = S\begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_{20}\end{pmatrix}S^{-1} \qquad P(t) = S\begin{pmatrix}e^{t\lambda_1} & & 0\\ & \ddots & \\ 0 & & e^{t\lambda_{20}}\end{pmatrix}S^{-1}, \tag{5}$$

where λ_i are the eigenvalues of Q, and the matrix S consists of the joint orthonormal basis of eigenvectors of Q and $P(t)$. According to the Perron-Frobenius theorem the largest eigenvalue of $P(t)$ equals 1 and therefore the largest eigenvalue of Q equals 0. In general, given a rate matrix Q of an EMP, one can easily calculate $P(t)$ for all $t > 0$ by formula (5).

Estimation Algorithms

The problem we are dealing with is the estimation of the parameters of an EMP from observed replacement frequencies. However, practically these frequencies are derived from different sequence alignments and will generally not conform to one EMP. We describe three approaches to the EMP estimation problem. We start with summarizing the original work of Dayhoff, Schwartz, and Orcutt (1978), develop the formalism of a maximum likelihood estimator, and finally explain the resolvent method.

Dayhoff's Method

Dayhoff's method has a strategy consisting of estimating $P(1)$ and then extrapolating to higher PAM distances. She pools input alignments of only closely related sequences. Nevertheless, this data can still be time-inhomogeneous. From this data she derives a calibrated transition matrix using the fact that on small evolution-

ary distances the calibration can be carried out linearly, as in equation (1). It is important to note that Dayhoff intended to use only alignments of very closely related pairs of sequences. There is no theoretical justification for applying it to more divergent input alignments. In fact, from our discussion earlier it becomes clear that linear calibration fails for more divergent data. The computational details of Dayhoff's method are summarized in various textbooks, see e.g., Setubal and Meidanis (1997).

Maximum Likelihood

With the enormous number of divergent alignments available today, Dayhoff's approach implies a huge loss of information. Yet it is highly desirable to exploit sequence alignments of widely different evolutionary distances. However, this requires progress in the theory of EMP estimation. An appropriate estimator should account for the evolutionary divergence of each alignment in the data set.

Our procedure is based on a two-step algorithm to improve a given set of rates and an equilibrium distribution. In the first step, the degree of evolutionary distance for each given alignment is estimated under the input parameters. In the second step, new parameters are estimated assuming the divergence times just calculated. This two-step procedure was started using Dayhoff's parameters and then iterated until no change of the parameters could be detected. This iterative approach is discussed in Müller and Vingron (2000). The time estimation part is discussed in Barry and Hartigan (1987*a*, 1987*b*); Adachi and Hasegawa (1996); Baake and von Haeseler (1999); Müller and Vingron (2000). For the following exposition we thus focus on the estimation of the rate matrix assuming that the evolutionary distances of all alignments are given.

The main problem in formulating a maximum likelihood estimator is to develop a parameterization for the rate matrix that reflects all requirements for an EMP. In order to specify the EMP, we estimate 210 parameters, namely the equilibrium distribution π and the rate matrix Q.

For simplicity we start by formulating the estimator only for a single given alignment $\mathbb{A}$ with evolutionary distance t. The maximum likelihood method yields the following estimates for π and Q:

$$(\hat{\pi}, \hat{Q}) = \underset{\pi,Q}{\operatorname{argmax}}\, \mathcal{L}(\pi, Q \mid t, \mathbb{A}) = \underset{\pi,Q}{\operatorname{argmax}} \sum_{i,j} N_{ij} \log([Fe^{tQ}]_{ij}), \quad (6)$$

where N_{ij} counts aligned amino acid pairs, F is a diagonal matrix with entries π_i, and Q is a rate matrix. We use the parameterization

$$Q = c\tilde{Q}. \quad (7)$$

where $c = (200\ \Sigma_{j>i}\ \pi_j r_{ij})^{-1}$ and

$$\tilde{Q} = \begin{pmatrix} \bullet_1 & \frac{r_{1,2}\pi_2}{\pi_1} & \cdots & \cdots & \frac{r_{1,20}\pi_{20}}{\pi_1} \\ r_{2,1} & \bullet_2 & \cdots & \cdots & \frac{r_{2,20}\pi_{20}}{\pi_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \frac{r_{19,20}\pi_{20}}{\pi_{19}} \\ r_{20,1} & \cdots & \cdots & r_{20,19} & \bullet_{20} \end{pmatrix}, \quad (8)$$

where $0 < \pi_i < 1$, $\Sigma_{i=1}^{20}\ \pi_i = 1$, and $\bullet_i = -\Sigma_{i \neq j}\ q_{ij}$. Note that Q is a rate matrix and satisfies the detailed balance equation. Conversely, all rate matrices that fulfill detailed balance can be parameterized as in (8). We proceed to show that multiplication by the factor $c = (200\ \Sigma_{j>i}\ \pi_j r_{ij})^{-1}$ calibrates the process to 1 PAM. To this end, we search for a c such that tr $(Fe^{c\tilde{Q}}) = 0.99$. From equation (1) we obtain

$$0.99 = \operatorname{tr}(Fe^{c\tilde{Q}}) \approx \operatorname{tr}(F[I + c\tilde{Q}]) = 1 + c\ \operatorname{tr}(F\tilde{Q}),$$

or equivalently $c \approx -(100\ \operatorname{tr}[F\tilde{Q}])^{-1}$. We want to express c in terms of the parameterization variables:

$$\operatorname{tr}(F\tilde{Q}) = \sum_i \pi_i \tilde{q}_{ii} = -\sum_{j \neq i} \pi_i \tilde{q}_{ij} = -2 \sum_{j>i} \pi_j r_{ij},$$

because of the symmetry of (r_{ij}) and hence we choose

$$c = \left(200 \sum_{j>i} \pi_j r_{ij}\right)^{-1}. \quad (9)$$

Practically, we are not given one alignment but many, each of possibly different divergence time. This leads to the following expression for the log-likelihood

$$[\hat{\pi}, \hat{Q}] = \underset{\pi,Q}{\operatorname{argmax}} \sum_{k=1}^{n} \sum_{j,i} N_{ij}^{(k)} \log([Fe^{t_k Q}]_{ij}), \quad (10)$$

where we assume that the n alignments are drawn independently and have divergence times t_k, $k = 1, \ldots, n$. $N^{(k)}$ are the respective matrices of exchange counts. The parameters that maximize equation (10) yield the maximum likelihood estimator among all Q that describe an EMP.

The maximization problem is tackled using standard optimization algorithms (Brent 1973; Press et al 1990). To this end, the constrained optimization problem needs to be mapped to an unconstrained one. For example, we map the distribution values to the interval (0,1) using the function $x \mapsto \arctan(x)/\pi + 0.5$ and we map the positive relative rates to the positive real line by the function $x \mapsto \exp(-x)$.

Resolvent Method

An alternative method for estimating the rate matrix Q of an EMP is presented in Müller and Vingron (2000). It is based on the relation

$$Q = \alpha I - R_\alpha^{-1} \quad \text{for all } \alpha > 0, \quad (11)$$

where R_α denotes the resolvent of the Markov chain (2). Once the resolvent is computed, one can derive the rate

matrix by applying this formula. The problem is putting 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 n alignments be given and assume that t_k is the degree of divergence of the sequences in alignment k. The goal is to estimate an EMP from the alignment data using the distances t_k. We first estimate $P(t_k)$ by the empirical transition frequencies in the respective alignments. We estimate $p_{ij}(t_k)$ by counting all occurrences of (a_i, a_j) and (a_j, a_i), and then normalizing by the overall frequency of amino acid i. For each alignment, this yields one estimated transition matrix $\hat{P}(t_k)$ for each time t_k. We want to approximate the integral $(R_\alpha)_{ij} = \int_0^\infty e^{-\alpha t} p_{ij}(t)\, dt$. This is done using linear interpolation of the $p_{ij}(t_k)$ and then integrating the piecewise functions. Note that the 20×20 entries of the resolvent can be calculated separately and independently of each other. Theoretically, the rate matrix is independent of α, but for empirical integrals it is not. The approximation of the integral $(R_\alpha)_{ij} = \int_0^\infty e^{-\alpha t} p_{ij}(t)\, dt$ is most accurate if the lattice points $p_{ij}(t_k)$ lie in the high-density region of the integrand. Whether this is the case or not depends on the parameter α. Consequently, a sensitive choice for α is called for. Our approach is likelihood based. With equation (3) we obtain

$$\mathcal{L}(\alpha \,|\, N_1, \ldots, N_n) \approx \sum_{k=1}^{n} \sum_{i,j} N_{ij}^k \log(\pi_i e_{ij}^{t_k(\alpha I - \hat{R}_\alpha^{-1})}). \quad (12)$$

Choosing a parameter α with maximal likelihood gives us a rate matrix. In practice, the resolvent method is used iteratively with time estimation updates as in the case of the maximum likelihood method.

Results

Clearly, if applied to real sequence data, the different estimation procedures result in different substitution models. We do not see any obvious biological criterion that can be applied to decide whether one model is superior to another. In contrast, we start from a given model (Q, π) and a set of evolutionary degrees of divergence $t_1, \ldots, t_n$ and sample artificial pairwise alignment data according to the associated distribution $M(t_i) = F \exp(t_i Q)$. More concretely, we draw independent pairs of amino acids from this distribution, and the concatenation of these pairs gives us a simulated gap-free alignment at PAM distance t_k.

We pool alignments of various degrees of divergence and run the estimators on this data, which is then used to reestimate the parameters (Q, π) that are used for alignment generation. In this setup, estimator evaluation is straightforward.

The resulting estimated substitution models can be equivalently represented by the rate matrix Q, any transition matrix $P(t)$, or a matrix of pair frequencies $M(t)$. As the first two display strong diagonal dominance, graphical comparison of them is not appropriate. Instead we choose the $M(100)$ pair frequencies.

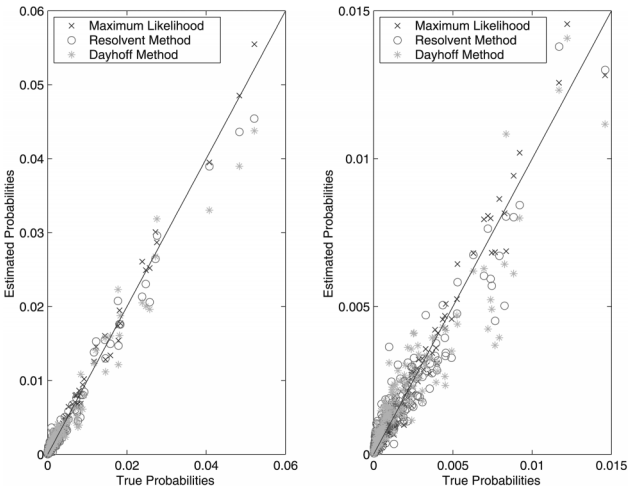

FIG. 1.—Comparison of all three methods on a small data set. Thirty artificial alignments of length 300 sites are used. Estimated values are plotted versus the simulation parameters. The right picture zooms into the lower left corner of the picture on the left side.

We show three simulation results. First, we test the estimators in the case of a small input data set. We reestimate the model parameters from 30 alignments of 300 sites each, where the degree of divergence varies from 10 to 300 PAMs with one alignment for each distance. The results are shown in figure 1. In this situation the maximum likelihood method is more accurate than either the resolvent method or Dayhoff's method. This is not surprising, because maximum likelihood approaches are known to yield highly efficient estimators. Yet one can clearly improve the accuracy of estimations by using more input data. In the case of protein evolution, tens of thousands of alignments are easily accessible. Theoretically, one would assume that the maximum likelihood estimator would be superior in this setup, but in practice, it is not appropriate because it is computationally too demanding. For real data sets the evolutionary degree of divergence of all alignments is very likely to be different. This is especially bad as it makes likelihood evaluations slow, see equation (10). In order to simulate the performance of the maximum likelihood estimator on a data set of medium size we use a set containing 30 alignments of 5,000 sites, for distances ranging from 10 to 300 PAM. This yields a large set of observed amino acid pairs but only at 30 different PAM distances. The results are shown in figure 2. One can clearly observe that the resolvent method catches up when compared to the maximum likelihood method, whereas Dayhoff's method shows the expected bias resulting from ignoring the evolutionary distances.

For huge amounts of input data, the maximum likelihood estimator cannot be evaluated. Figure 3 compares Dayhoff's method with the resolvent estimator for 10,000 alignments of length 300 with PAM distances distributed uniformly at the interval [0, 300]. The resolvent method shows very satisfying accuracy, and as one would expect, it clearly outperforms Dayhoff's method.

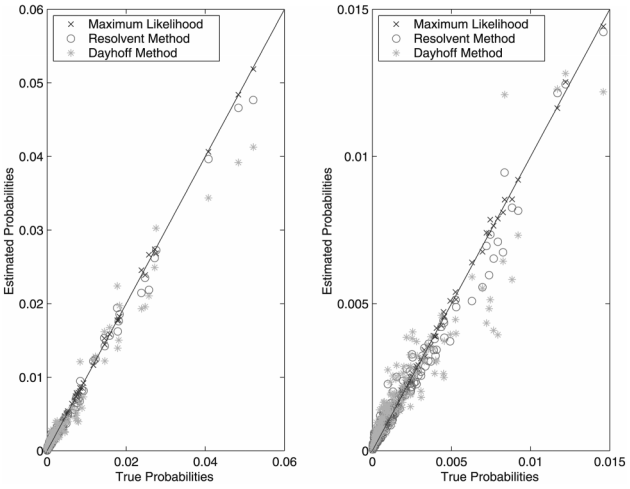

FIG. 2.—Comparison of all three methods on a medium sized data set. Thirty artificial alignments of 5,000 sites are used. Again, estimated values are plotted versus the simulation parameters. The right picture focuses on lower probabilities.

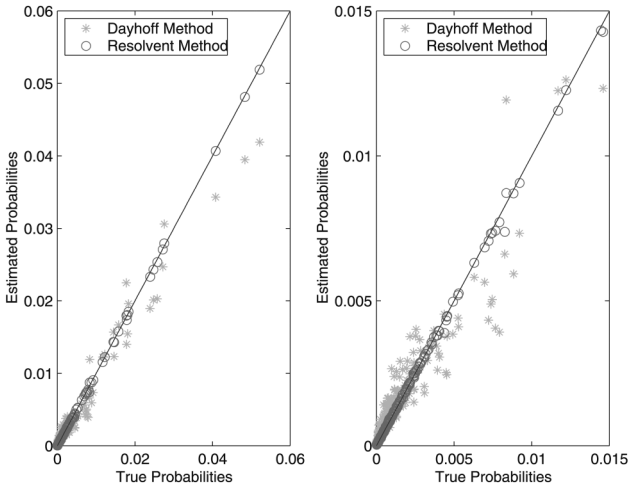

FIG. 3.—Evaluation of large-scale estimations. Only results from Dayhoff's method and the resolvent method are shown. Ten thousand artificial alignments of 300 sites are used. Estimated values are plotted versus the simulation parameters. The right picture zooms into the lower left corner of the picture on the left side.

Table 1 summarizes the three comparison experiments. Here, we assess the performance of the three methods in terms of the relative entropy $H(\hat{M}[100]|M[100]) = \Sigma_{ij}\ m_{ij}(100)\log([m_{ij}\{100\}]/[\hat{m}_{ij}\{100\}])$ of the joint distributions $\hat{M}(100)$ of the estimated models to the model $M(100)$ that is underlying the simulations.

As we pointed out, a sensitive choice of the resolvent parameter α is critical for the performance of the resolvent method. This is clearly supported by the simulation experiments. Figure 4 shows a plot of the average log-likelihood of the fitted model versus α. It becomes clear that although the rate matrix Q is theoretically independent of α, numerical problems can significantly weaken the resolvent method if α is inappropriately chosen.

Discussion

We discussed the problem of estimating amino acid replacement frequencies from inhomogeneous divergent alignment data. We tested the applicability of Dayhoff's method to this kind of data, reviewed and evaluated the resolvent method, and developed a novel maximum likelihood estimator. All of these methods model protein evolution by a Markov process acting independently on each site of the proteins. Whereas the RV and ML methods take evolutionary distances into account, Dayhoff's method does not. In simulations with time-heterogeneous alignment data we prove the importance of including distances into the model. In particular, maximum likelihood proved to perform best for small data sets, whereas for larger data sets where maximum likelihood becomes computationally infeasible, the resolvent method is a good alternative.

The EMP model reduces the phenomenon of protein evolution to 210 parameters. It is obvious that this cannot cover the entire complexity of evolution. One assumes that the positions in a protein evolve independently of each other with the same dynamics for each site, which can be modeled by a Markov chain. Of course, it is well known that different sites in a protein may evolve at different speeds and that possibly different replacement mechanisms are operating. All of our assumptions are questionable from a biological point of view. However, from the perspective of data analysis it is obvious that one needs to simplify to make model fitting practical. The challenge is to reflect as much of the reality as possible with 210 parameters.

In 1972 when Dayhoff et al. proposed the first solution to this problem, only a few sequences were available, and homology detection was restricted to relatively closely related pairs of sequences. Clearly, their method was intended for the use of this kind of data. Our use of Dayhoff's method in the context of divergent alignments is not in the sense of its authors and has no theoretical justification. We use it to demonstrate the practical importance of including the divergence parameter into a model.

Table 1
A Summary of the Evaluation Experiments. The Efficiency of the Estimators is Measured in Terms of the Relative Entropy of the Estimated Models to the Model that is Underlying the Simulation

	Dayhoff	Resolvent	Maximum Likelihood
Small data set	$7.1453e^{-4}$	$2.9946e^{-4}$	$8.6068e^{-5}$
Medium size data set	$6.6340e^{-4}$	$1.5976e^{-4}$	$2.7474e^{-6}$
Huge data set	$6.54e^{-2}$	$8.735e^{-3}$	Not computed

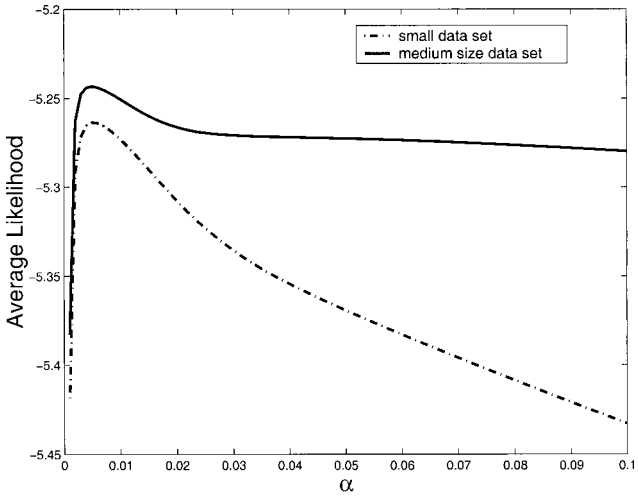

FIG. 4.—The resolvent method: a plot of the average log-likelihood per site of the fitted model versus the resolvent parameter α. The continuous line refers to the medium size data set, whereas the dotted line refers to the small data set.

A natural shortcoming of using only closely related sequence alignments is that the estimator is biased toward the evolution of fast-evolving positions in proteins. Thus, basing the estimation on the large and divergent data set that we used does not only improve the model parameters because of the much larger amount of input data, but might also reflect protein evolution on longer time scales more appropriately.

Our simulation results show that the maximum likelihood estimator is more efficient than the resolvent method. On the other hand, it is restricted to input data sets of moderate size. But more input data clearly gives more accurate estimates for the transition probabilities, which may well compensate for the theoretical suboptimality of the estimator. In principle, we have a trade off between the statistical and the computational efficiency of the estimators. For small data sets we recommend a maximum likelihood approach, whereas the resolvent method is a practical alternative tailored for huge data sets.

Acknowledgments

We would like to thank two anonymous referees for helpful comments on an earlier version of this paper. We are also grateful to Sven Rahmann and Marc Rehmsmeier for many stimulating discussions.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. Molphy: (programs for molecular phylogenetics) Version 2.3. Institute of Statistical Mathematics, Tokyo.

BAAKE, E., and A. VON HAESELER. 1999. Distance measures in terms of substitution processes. Theor. Popul. Biol. **5**: 166–175.

BARRY, D., and J. HARTIGAN. 1987*a*. Asynchronous distance between homologous DNA sequences. Biometrics **43**(2): 261–276.

———. 1987*b*. Statistical analysis of hominoid molecular evolution. Stat. Sci. **2**:191–210.

BENNER, S., M. COHEN, and G. GONNET. 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng. **7**:1323–1332.

BRENT, R. P. 1973. Algorithms for minimization without derivatives. Prentice Hall, Englewood Cliffs, NJ.

DAYHOFF, M., R. SCHWARTZ, and B. ORCUTT. 1978. A model of evolutionary change in protein. Atlas Protein Seq. Struct. **5**:345–352.

GONNET, G., M. COHEN, and S. BENNER. 1992. Exhaustive matching of the entire protein sequence database. Science **256**:1443–1445.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

MÜLLER, T., and M. VINGRON. 2000. Modeling amino acid replacement. J. Comput. Biol. **7**(6):761–776.

OVERINGTON, J., M. JOHNSON, A. SALI, and T. BLUNDELL. 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. R. Soc. Lond. B. **241**:132–145.

PRESS, W. H., B. P. FLANNERY, S. TEUKOLSKY, and W. T. VETTERLING. 1990. Numerical recipes in C. Press Syndicate of the University of Cambridge, Cambridge.

RISLER, J., M. DELORME, H. DELACROIX, and A. HENAUT. 1988. Amino acid substitutions in structurally related proteins. J. Mol. Biol. **204**:1019–1029.

SETUBAL, J., and J. MEIDANIS. 1997. Introduction to computational molecular biology. PWS Publishing Company, Boston.

WILLIAM TAYLOR, reviewing editor

Accepted July 23, 2001