

Concepts for the Estimation of Amino Acid Substitution Models

Tobias Müller, Rainer Spang and Martin Vingron
Deutsches Krebsforschungszentrum
Theoretische Bioinformatik
Im Neuenheimer Feld 280
69120 Heidelberg, Germany
{t.mueller|r.spang|m.vingron}@dkfz-heidelberg.de

July 10, 2000

Abstract

In a simple model, the evolution of proteins can be viewed as the accumulation of amino acid substitutions. Amino acids with similar chemical and physical properties are replaced by each other more often than different types. M. Dayhoff *et al.* (Atlas of Protein Sequences and Structure, 1978, 5, 345–352) suggested to revert this relation and describe the similarity of amino acids by their replacement frequencies. Replacement frequencies depend on the degree of divergence between sequences and inhomogeneity with respect to divergence inside the alignment data is the major obstacle in a statistical approach.

We compare three approaches: First, the original method by M. Dayhoff (DH), secondly, the resolvent method (RV) (Müller & Vingron, Journal of Computational Biology, 2000), and finally a Maximum Likelihood approach (ML) first described here. We evaluate the methods using a highly divergence inhomogeneous set of sequence alignments. ML is the method of choice in the case of small sets of input data. The RV method is computationally much less demanding while it performs only slightly worse than the ML. Therefore it is perfectly appropriate for large scale applications.

1 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. If this replacement improves the fitness of the organism the new amino acid will be accepted by natural selection. Replacements of very unsimilar amino acids often drastically change the fold of the protein, often 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. This asks for a dynamical model which describes protein evolution on a time scale.

Modeling amino acid replacements by a Markoff chain has been introduced by (Dayhoff *et al.*, 1978). In the most simple version of this approach a set of identical Markoff chains acting independently on each site of the protein can be 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 divergence inhomogeneous sequence data.

In the original approach of (Dayhoff *et al.*, 1978) the actual estimation is restricted to only very closely related pairs of sequences. However, once a Markoff model is fitted by this data, replacement frequencies characteristic for distantly related sequences can be simply extrapolated from the model. Dayhoff's approach is the subject of Subsection 3.1.

Dayhoff's approach has been generalized and applied to larger data sets (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 related alignments (Risler *et al.*, 1988; Overington *et al.*, 1990). However, these authors do not provide a general statistical model which fully describes the time dependence of the input data.

First, Benner *et al.* (1994) pointed out the problem of estimating one consistent model from an inhomogeneous pool of alignment data. They sketch a normalization algorithm that is based on computing logarithms of transition matrices, which they approximate by power series. The approach is heuristic since the convergence of the power series can not be guaranteed for empirically derived matrices.

Müller & 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 Section 3.3. Alternatively, we describe a novel Maximum Likelihood based approach. Section 3.2 provides the details of the mathematical formalisms and computations.

In principle one has two important, but ambivalent criteria for evaluating the quality of the methods. For large scale applications time performance of the algorithms is crucial, whereas statistical efficiency of the estimator can be compensated 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 Section 4 we discuss the individual merits of the resolvent and of the Maximum Likelihood estimator.

2 Model

We start by reviewing some basic results on Markoff chains. Let $P(t)$ be the transition probability matrix of a Markoff chain with entries $p_{ij}(t) = \text{Prob}[X(s+t) = j | X(s) = i]$. We consider only Markoff chains for which

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

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

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

exists. $Q = (q_{ij})$ is called *rate matrix*. The diagonal entries of Q are negative, off diagonal entries are positive. Equation (2) provides an approximation for the rate matrix:

$$P(t) - I = tQ + o(t), \quad (3)$$

hence we have

$$p_{ij}(t) \approx tq_{ij} \quad (4)$$

for $i \neq j$ and small t . Alternatively, one can characterize the rate matrix of a Markoff chain by the resolvent:

$$R_\alpha := \int_0^\infty e^{-\alpha t} P(t) dt. \quad (5)$$

R_α is invertible for all $\alpha > 0$ (Müller & Vingron, 2000). From the Chapman–Kolmogorov equation we get the forward and backward equations

$$\frac{d}{dt} P(t) = P(t)Q = QP(t). \quad (6)$$

Differentiating $e^{-\alpha t}P(t)$ using the product rule for matrix differentiation yields

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

for $\alpha > 0$, $t \geq 0$. Using this and equation (6) we obtain:

$$\begin{aligned} \alpha \int_0^\infty e^{-\alpha t} P(t) dt + \int_0^\infty \frac{d}{dt} [e^{-\alpha t} P(t)] dt &= \int_0^\infty e^{-\alpha t} \frac{d}{dt} P(t) dt \\ &= \int_0^\infty e^{-\alpha t} P(t) dt Q. \end{aligned}$$

Using the fundamental theorem of calculus and multiplying by R_α^{-1} reduces the above equation to

$$\alpha I - R_\alpha^{-1} = Q, \quad (7)$$

where I denotes the identity matrix.

The differential equation (6) can be solved under the initial condition $P(0) = I$ and yields

$$P(t) = \exp(tQ) = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}. \quad (8)$$

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

$$q_{ij} \geq 0 \text{ for } i \neq j \quad \text{and} \quad \sum_j q_{ij} = 0 \text{ for all } j, \quad (9)$$

see Grimmet & Stirzaker (1992).

The concrete problem of modelling amino acid replacement frequencies requires additional assumptions on the Markoff chain. We discuss these requirements which leads to the notion of “evolutionary Markoff processes” (EMPs) as introduced in Müller & Vingron (2000).

Following Dayhoff *et al.* (1978) we model the evolution of each site of the proteins by a single time homogeneous Markoff chain $X(t)$, calibrated, such that on average 1% of the amino acid are changed after one unit of time:

$$\text{Prob}[X(t) \neq X(t+1)] = 0.01. \quad (10)$$

Once calibrated, the time t in the Markoff chain can be used as a measure of evolutionary divergence (Müller & Vingron, 2000). The acronym ‘‘PAM’’ (Accepted Point Mutations) is commonly used for this unit of divergence.

All possible mutation events of two arbitrary amino acids have been observed. Hence we do not loose generality, if we assume that any amino acid can mutate in any other during any period of time. This can be ensured by the assumption $q_{ij} > 0$ for all i, j . In this case there exists a unique limiting amino acid distribution $\pi = (\pi_1, \dots, \pi_{20})$ where $\pi_j = \lim_{t \rightarrow \infty} p_{ij}(t) > 0$ independently of the initial residue a_i . It fulfills the equations $\pi Q = 0$ and $\pi P(t) = \pi$ for all $t \geq 0$. We only considers Markoff processes which are in equilibrium and therefore all X_t are distributed according to π . Note that a transition matrix contains implicit assumptions on the distribution of amino acids.

With the transition probabilities $(P_t)_{t \geq 0}$ and the overall amino acid distribution we calculate the joint distribution

$$m_{ij}(t) = \pi_i p_{ij}(t) \quad (11)$$

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 Markoff 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. We can not decide the direction of the mutation, we only observe pairs of corresponding amino acids at a certain position in a protein. It would not be appropriate to model such a direction. We can account for the symmetry in the data by restricting our model to the class of time reversible Markoff chains. Formally, 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 . 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,

$$\pi_i q_{ij} = \pi_j q_{ji}. \quad (12)$$

In particular, $M(t)$ is a symmetric matrix. Following Müller & Vingron (2000) we call a process satisfying the above condition an evolutionary Markoff process (EMP).

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. Due to the reversibility the rate matrix Q and $P(t)$ are selfadjoint relative to $\langle \cdot, \cdot \rangle_F$, i.e. $\langle Px, y \rangle_F = \langle x, Py \rangle_F$, and can therefore be transformed into diagonal form by change of coordinates. The eigenvalues of Q are real by selfadjointness, and negative due to Gershgorin's theorem. Using definition (8) we can rewrite Q and $P(t)$ as

$$Q = S \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_{20} \end{pmatrix} S^{-1} \quad P(t) = S \begin{pmatrix} e^{t\lambda_1} & & 0 \\ & \ddots & \\ 0 & & e^{t\lambda_{20}} \end{pmatrix} S^{-1} \quad (13)$$

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 (13).

However, if one wants to estimate the rate matrix of an EMP from observed transition frequencies, one faces the problem that in general a family of empirical transition matrices $P(t_1), \dots, P(t_n)$ do not correspond to a single consistent EMP. Nevertheless, the estimation of one consistent EMP from time inhomogeneous data is the subject of this paper.

3 Algorithms

In this section we describe three approaches to the EMP estimation problem. We start with summarizing the original work of Dayhoff *et al.* (1978), develop the formalism of a Maximum Likelihood estimator, and finally explain the resolvent method.

3.1 Dayhoff's Method

Dayhoff's strategy is first to estimate $P(1)$ for then extrapolating to higher PAM distances. She pools input alignments of only closely related sequences. Nevertheless, this data can be still time inhomogeneous. From this data she derives a calibrated transition matrix using the fact that on small evolutionary distances the calibration can be carried out linearly. From equation (8) or equation (3) ,(3) it becomes clear that linear calibration fails for more divergent data.

For two distinct amino acids i and j , let A_{ij} be the number of times that the amino acid pair (a_i, a_j) is observed. Let A be the matrix that consists

of the A_{ij} for $i \neq j$ and is zero on the main diagonal. Let π be the overall amino acid frequencies. The estimator is based on the equation

$$P_{ij} = \text{Prob}[i \text{ mutates}] \text{Prob}[i \rightarrow j | i \text{ mutates}] \quad (14)$$

where $i \rightarrow j$ means amino acid i mutates into amino acid j . One wants to estimate the term on the left hand side, and has data to estimate both terms on the right hand side. Let us start with the case $i \neq j$. One needs to give the probability that amino acid i mutates. This can be done by introducing the relative amino acid mutability

$$m_i = \frac{\sum_j A_{i,j}}{\sum_{k,j} A_{k,j}}.$$

The second term on the right hand side of equation (14) is estimated by $A_{i,j}/\sum_j A_{i,j}$. The diagonal entries of the transition matrix are set to

$$P_{ii} = 1 - m_i.$$

Calibration is done by the transformation $m_i \rightarrow m_i/100\pi_i$. This definition ensures that P_{ij} is calibrated to 1 PAM. Transition frequencies for x PAM can be extrapolated by P^x . It is important to note that Dayhoff only 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.

3.2 Maximum Likelihood

With the enormous number of divergent alignments available today, Dayhoff's approach implies a large loss of information. 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. We split the general estimation problem into two parts: (1) Estimation of the evolutionary degree of divergence of all pairwise sequence alignments. (2) Estimation of a rate matrix from the alignment data given the distances. The iterative nature of this approach is discussed in Müller & Vingron (2000). Time estimations is discussed in Barry & Hartigan (1987a,b); Adachi & Hasegawa (1996); Baake & von Haeseler (1999); Müller & Vingron (2000). In this paper we focus on the estimation of the rate matrix, therefore we assume that the evolutionary distances of all alignments are known.

The main problem in formulating a Maximum Likelihood estimator is to develop a parameterization for the rate matrix, which reflects all requirements for an EMP. In order to specify the EMP, we estimate 210 parameters, the stationary amino acid distribution π and the rate matrix Q . For simplicity we first formulate the estimator only for a single given alignment \mathbb{A} .

For an alignment \mathbb{A} with evolutionary distance t the Maximum Likelihood method yields the following estimates for π and Q :

$$\begin{aligned} (\hat{\pi}, \hat{Q}) &= \operatorname{argmax}_{\pi, Q} \mathcal{L}(\pi, Q | t, \mathbb{A}) \\ &= \operatorname{argmax}_{\pi, Q} \sum_{i,j} N_{ij} \log((F e^{tQ})_{ij}), \end{aligned} \tag{15}$$

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 parametrization

$$Q = c\tilde{Q}, \tag{16}$$

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

$$\tilde{Q} = \begin{pmatrix} \bullet_1 & \frac{r_{1,2}\pi_2}{\pi_1} & \cdots & \cdots & \frac{r_{1,20}\pi_{20}}{\pi_1} \\ r_{2,1} & \bullet_2 & \cdots & \cdots & \frac{r_{2,20}\pi_{20}}{\pi_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & \frac{r_{19,20}\pi_{20}}{\pi_{19}} \\ r_{20,1} & \cdots & \cdots & r_{20,19} & \bullet_{20} \end{pmatrix}, \quad (17)$$

where $0 < \pi_i < 1, \sum_{i=1}^{20} \pi_i = 1$ and $\bullet_i = -\sum_{i \neq j} q_{ij}$. Note that any process (Q, π) of the form (16) are EMPs, in particular satisfying detailed balance and vice versa all EMPs can be parametrized as in (16). The factor $c = (200 \sum_{j>i} \pi_j r_{ij})^{-1}$ calibrates the process to approximately 1 PAM.

Estimating Q without the normalization constant c would yield a rate matrix calibrated to PAM distances close to 1 PAM. The factor c is used to make it exactly 1 PAM. This constant is, however, close to one. It needs to be chosen, such that $\text{tr}(Fe^Q) = 0.99$.

We get by equation (3)

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

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

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

due to the symmetry of $(r_{i,j})$ and hence we choose

$$c = (200 \sum_{j>i} \pi_j r_{ij})^{-1}. \quad (18)$$

We generalize formula (15) for the likelihood

$$[\hat{\pi}, \hat{Q}] = \underset{\pi, Q}{\text{argmax}} \sum_{k=1}^n \sum_{j,i} N_{ij}^{(k)} \log((Fe^{t_k Q})_{ij}), \quad (19)$$

of n independent alignments with corresponding times t_k and count matrices $N^{(k)}$. The parameters which maximizes (19) yields the Maximum Likelihood estimator among all Q which describe calibrated, stationary, time reversible Markoff chains.

Note, that there are some constraints operating. The entries of the stationary distribution π must be in the interval $(0, 1)$, and $r_{ij} > 0$. Therefore we reparameterize this likelihood maximization problem to obtain a new objective function in which the parameter are unconstrained. For example, if the constraint $x > 0$ is required, we can set $x = \exp(y)$. This new unconstrained problem can be solved by ordinary optimization algorithms. We use Powell's method (Brent, 1973) described in the numerical recipes (Press *et al.*, 1990).

3.3 Resolvent Method

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

$$Q = \alpha I - R_\alpha^{-1}, \quad (20)$$

where R_α denotes the resolvent of the Markoff chain (5). Once the resolvent is computed one can derive the rate matrix by applying this formula. 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 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. Therefore we require a rationale to select α . Maximum Likelihood can be applied. With (7) we obtain

$$\mathcal{L}(\alpha | N_1, \dots, N_n) \approx \sum_{k=1}^n \sum_{i,j} N_{ij}^k \log(\pi_i e^{t_k(\alpha I - \hat{R}_\alpha^{-1})_{ij}}). \quad (21)$$

The optimal α and formula (20) give us a rate matrix. In practice, this method is used iteratively with time estimation updates as in the case of the Maximum Likelihood method.

4 Results

Clearly, if applied to real sequence data, the different estimation procedures result in different substitution models. We do not see any obvious biological criterium that can be applied to decide whether one model is superior to another. If in contrast we start from a given model (Q, π) and a set of evolutionary degrees of divergence t_1, \dots, t_n and sample artificial pairwise alignment data according to the associated distribution $M(t_i) = F \exp(t_i Q)$. We pool alignments of various degrees of divergence and run the estimators on this data. This data is used to reestimate the parameters (Q, π) that are used for alignment generation. In this setup estimator evaluation is straight forward.

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

We show three simulation results. First, we test the estimators in the case of a small input data set. We reestimate model parameter from 10 alignments of 300 sites each, where the degree of divergence varies from 10 to 300 PAMs with one alignment for each distance. The result are shown in Figure (1). In this situation the Maximum Likelihood method is much more accurate than both, the resolvent method and Dayhoff's method. This is not surprising, because Maximum Likelihood approaches are known to yield highly efficient estimators in many situations. For small sets of input data the Maximum Likelihood estimator is the method of choice. On the other hand, one can clearly improve the accuracy of estimations by using more input data. In the case of protein evolution tens of thousands alignments are easily accessibly. Theoretically, one would assume that Maximum Likelihood estimator is also superior in this setup. But in practice, it is not applicable, because it is computationally to demanding. For real data sets the evolutionary degree of divergence of all alignments are very likely to be different. This is especially bad since it makes likelihood evaluations slow, see equation (19). In order to simulate the performance of the Maximum Likelihood estimator on a data set of moderate size we use a set containing 10 alignments of length 5000 sites, for the distances 10 to 300 PAM. This yields a large set of observed amino acid pairs but only at 10 different PAM distances. The results are shown in Figure 2. One can clearly observe that the resolvent method catches up ground when compared to the Maximum Likelihood method, while Dayhoff's

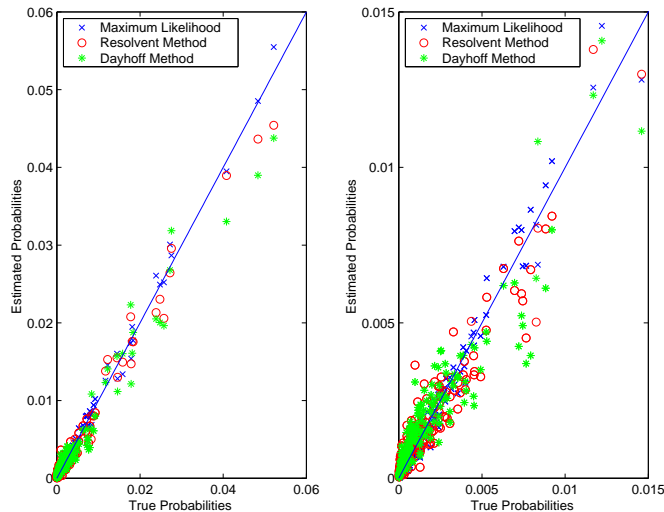


Figure 1: Comparison of all three methods on small data set. 10 artificial alignments of length 300 sites are used. Estimated values are plotted versus the simulation parameters.

method shows the expected bias resulting from ignoring the evolutionary distances.

For huge amounts of input data the Maximum Likelihood estimator can not be evaluated. Figure 3 compares Dayhoff's method with the resolvent estimator for 10,000 alignments of length 300 with PAM distances distributed uniformly on the interval $[0, 300]$. The resolvent method shows very satisfying accuracy, and as one would expect it clearly outperforms Dayhoff's method. Since Maximum Likelihood is not practical we recommend the use of the resolvent method for large scale applications.

5 Discussion

We discuss the problem of estimating amino acid replacement frequencies from large scale, inhomogeneous divergent alignment data. We tested the

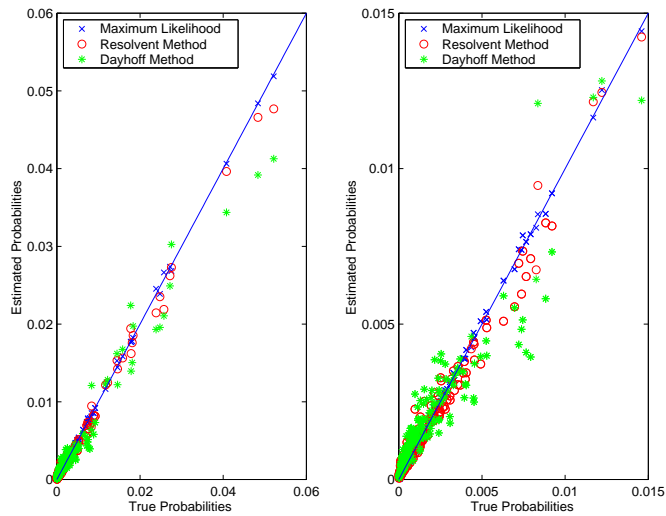


Figure 2: Comparison of all three methods on moderate sized data set. 10 artificial alignments of length 5000 sites are used. Again, estimated values are plotted versus the simulation parameters.

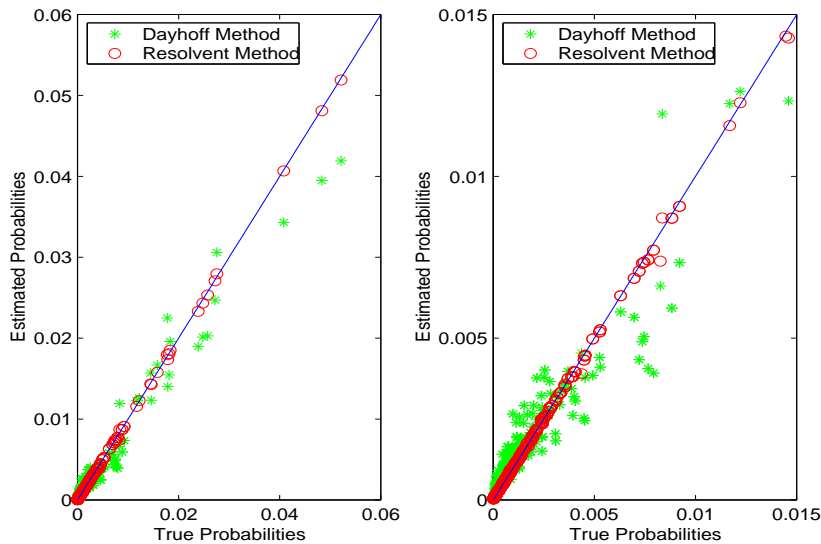


Figure 3: Evaluation of large scale estimations. Only results from Dayhoff's method and the resolvent method are shown. 10,000 artificial alignments of length 300 sites are used. Estimated values are plotted versus the simulation parameters.

applicability of Dayhoff's method to this kind of data, we reviewed and evaluated the Resolvent method and developed a novel Maximum Likelihood estimator. All have in common that they model protein evolution by a Markoff process acting independently on each site of the proteins. While methods (2) and (3) take evolutionary distances into account the first one does not. In simulation using time heterogeneous alignment data we prove the importance to include distances into the model. In particular, Maximum Likelihood proved to be optimal for small data sets whereas for larger data sets where Maximum Likelihood becomes computationally unfeasible the Resolvent method is a good alternative.

The EMP model reduces the phenomenon of protein evolution to 210 parameter. It is obvious, that this can not cover the entire complexity of evolution. One makes the assumptions that the positions in a protein evolve independently from each other with the same dynamics for each site, which can be modeled by a Markoff chain. Of course, it is well known that different sites in a protein may evolve at different speeds and possibly different replacement mechanisms are operating. All 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, such that model fitting becomes 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 few sequences were available and homology detection was refined to relatively closely related pairs of sequences. Clearly, their method is 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 these authors and has no the-

oretical justification. We use it to demonstrate the practical importance of including the divergence parameter into a model.

A natural shortcoming of using only closely related sequence alignments is that the estimator is biased towards the evolution of fast evolving positions in proteins. Hence basing the estimation on the large and divergent data set that we used does not only improve the model due to 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 as the Resolvent method. On the other hand it is restricted to input data 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 principal we have a trade of between the statistical and the computational efficiency of the estimators. For small data sets we recommend a Maximum Likelihood approach, while the resolvent method is a practical alternative tailored for huge data sets.

Acknowledgments

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

References

- Adachi, J. & Hasegawa, M. (1996). Molphy: Programs for molecular phylogenetics, ver. 2.3. *Tokyo: Institute of Statistical Mathematics*.
- Baake, E. & von Haeseler, A. (1999). Distance measures in terms of substitution processes. *Theor. Popul. Bio.* **5**, 166–175.

- Barry, D. & Hartigan, J. (1987a). Asynchronous distance between homologous dna sequences. *Biometrics*, **43**(2), 261–276.
- Barry, D. & Hartigan, J. (1987b). Statistical analysis of hominoid molecular evolution. *Stat. Sci.* **2**, 191–210.
- Benner, S., Cohen, M., & Gonnet, G. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Engineering*, **7**, 1323–1332.
- Brent, R. P. (1973). *Algorithms for Minimization without Derivatives*. Englewood Cliffs, N.J.: Prentice Hall.
- Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). A model of evolutionary change in protein. *Atlas of Protein Sequences and Structure*, **5**, 345–352.
- Gonnet, G., Cohen, M., & Benner, S. (1992). Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.
- Grimmet, G. R. & Stirzaker, D. R. (1992). *Probability and Random Processes*. New York: Oxford Science Publications.
- Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.
- Müller, T. & Vingron, M. (2000). Modeling amino acid replacement. *Journal of Computational Biology*, **1**. to appear.
- Overington, J., Johnson, M., Sali, A., & Blundell, T. (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., Flannery, B. P., Teukolsky, S., & Vetterling, W. T. (1990). *Numerical Recipes in C*. Cambridge: Press Syndicate of the University of Cambridge.

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