BU-1316-M

Biases in Amino Acid Replacement Matrices and Alignment Scores Due to Rate Heterogeneity

Colleen Kelly*

Department of Computer Science and Statistics

University of Rhode Island, Kingston, RI 02881

phone: (401) 792-2701   email: kelly@stat.uri.edu

Gary A. Churchill

Department of Biometry and Plant Breeding
Cornell University, Ithaca, NY 14853

Running Head: Bias in Replacement Matrices

Keywords: rate variation, protein alignment, PAM matrix, evolutionary distance.

[0] To whom correspondence should be addressed.

Abstract

Empirically derived amino–acid replacement matrices are widely used in sequence comparison and database searches. We consider an extension of the usual Markov process model of protein evolution that admits site to site rate heterogeneity and demonstrate that rate heterogeneity can introduce a bias in estimated replacement probabilities and the corresponding alignment scores derived from these matrices. We suggest an approach to obtain unbiased estimates of replacement probabilities and alignment scores and derive the details for the case where rates are assumed to vary according to a Gamma distribution.

1 Introduction

Empirically derived amino acid replacement matrices (Dayhoff et al. 1978, Jones at al. 1992, Gonnet et al. 1992) are widely used in problems of sequence comparison and alignment (Altschul 1991, 1993) and in database searches (Altschul et al. 1990). Replacement matrices reflect the average (over many sites in many protein families) probabilities with which one amino acid may be substituted by another over evolutionary time. Because they are empirically derived, they should reflect exchangeability due to physical and chemical similarities of amino acids as well as effects due to properties of the underlying genetic code and the mutation processes acting at the DNA level. Empirically derived matrices are generally considered to reflect the true relationships among amino acids better than matrices derived from considerations of chemical or physical properties (Taylor 1986) or the genetic code (Feng and Doolittle 1985).

The pioneering work of Dayhoff and her collaborators (Dayhoff et al. 1978) has recently been updated to include the large amounts of protein sequence information that have accumulated since 1978 (Jones et al. 1992, Gonnet et al. 1992). The general pattern of the replacement probabilities appears to be remarkably stable in spite of the rather limited set of protein families that were available then as compared to the present (Jones et al. 1992). This suggests that it may be reasonable to use average replacement matrices although any particular family (or site within a family) may have its own characteristic pattern of replacements. It may be possible to develop a small set of distinct replacement matrices (Sander and Schneider 1991, Brown et al. 1993) that reflect different local environments within proteins or different protein families. If this is the case, the methods described here will continue to be useful in deriving appropriate unbiased estimates of the replacement probabilities.

The model of sequence change which is usually assumed in the construction of amino acid replacement matrices is a Markov process model that describes the pattern of replacements over time and acts independently and identically at each residue of a protein sequence. A critical discussion of this model is given by Wilbur (1985) who raises a number interesting points.

It is commonly accepted and evident from observation that rates of change are not identical at each site in a sequence (Uzzell and Corbin 1971, Holmquist 1983, Reeves 1992, Wakely 1993, Yang et al. 1994). Rate variation is probably present to a greater or lesser extent in every protein family. We are of the opinion that site to site rate variation presents one the most important and challenging problems now faced by methodological researchers in molecular evolution and that new methods will be needed by empirical scientists to properly analyze and interpret sequence data. It has already been observed that rate variation can introduce biases into estimates of sequence divergence (Kelly 1991, Kelly and Rice 1994, Yang et al. 1994, Ota and Nei 1994). We take the point of view that the Markov model is an acceptable approximation to the actual process of sequence evolution at any one site and address the issue of estimating an average replacement matrix in the presence of rate heterogeneity.

Clearly, any one pair of protein sequences will not provide sufficient information to estimate the large number of parameters required to specify a complete replacement matrix. Thus, many such pairs or families of sequences must be considered. A problem arises because not all sequence families are evolving at the same rates and not all pairs of sequences are separated by similar amounts of geological time. Thus to obtain estimates of replacement probabilities over a given interval of evolutionary time t (geological time × rate of evolution), the observed patterns of replacements must be adjusted to reflect a common amount of divergence.

Dayhoff et al. (1978) proposed measuring time in PAM units. One PAM unit corresponds to an average of one replacement per 100 sequence residues for a protein of average composition. In the past, it has been common to use the so called "PAM250" log–odds scoring matrix for sequence alignment and database searches. However, this matrix was developed for detecting very distant relationships (Schwartz and Dayhoff, 1978). It has been pointed out (Karlin and Altschul 1991) that the choice of a scoring matrix implies a particular target distribution of aligned amino acid pairs and that optimal results are obtained when the scoring matrix corresponds to the evolutionary distance between the particular sequences being compared. Altschul (1993) recommends the use of a small set of scoring matrices tuned to distinct evolutionary distances for database searching applications.

Replacement matrices at different PAM distances are typically computed by repeated mulitplication of a PAM1 matrix. We have noticed that there is a bias, due to rate heterogeneity, introduced into rescaled replacement matrices computed by this method. We describe the nature of this bias and suggest an adjustment in the case where rates are assumed to follow a Gamma distribution with known shape parameter α. In section 2 of this paper, we review the basic concepts of Markov models of sequence evolution for the case of homogenous rates and describe an extension to the heterogeneous rates case. We then briefly review the calculation of alignment scores from replacement matrices. In section 3 we demonstrate that the standard method of adjustment introduces a bias in both the replacement probabilities and their corresponding alignment scores. In section 4 we demonstrate these biases using data from the blocks data base (Henikoff and Henikoff 1991). In the final section, we discuss some of the consequences of rate heterogeneity on the estimated replacement rates.

2 Methods

2.1 A Homogeneous Rates Model

We assume that the process of amino acid replacement acting at one site can be described as a continuous time, time homogeneous Markov process. We further assume that the process is reversible in time and that it is has reached its equilibrium state. The reader is referred to Tavaré (1986) for a detailed description and criticisms of this model for the case of nucleotide sequences.

Under the Markov model, a site originally occupied by amino acid i will, after (evolutionary) time t, be occupied by amino acid j with probability $p_{ij}(t)$, for $i = 1, \ldots, 20$ and $j = 1, \ldots, 20$ where $1, \ldots, 20$ is an arbitrary numbering of the amino acids. The matrix of transition probabilities may be written as $\mathbf{P}(t) = [p_{ij}(t)]$. Such transition matrices are commonly expressed as the matrix exponential of a rates matrix $\mathbf{Q}$. Let $\mathbf{Q} = \mathbf{U}^{-1}\Sigma\mathbf{U}$ be the spectral decomposition of $\mathbf{Q}$. We can express $\mathbf{P}(t)$ as

$$\begin{aligned} \mathbf{P}(t) &= e^{\mathbf{Q}t} \\ &= \mathbf{U}^{-1}e^{\Sigma t}\mathbf{U} \\ &= \mathbf{U}^{-1}\Theta(t)\mathbf{U} \end{aligned}$$

where $\mathbf{U}$ is a matrix whose columns are the eigenvectors of $\mathbf{Q}$, Σ is a diagonal matrix of eigenvalues, $\Sigma = \text{diag}\{\sigma_1, \ldots, \sigma_{20}\}$ and $\Theta(t) = \text{diag}\{e^{\sigma_1 t}, \ldots, e^{\sigma_{20} t}\}$. This form is especially convenient for computing transition matrices at various times t.

If the spectral decompostion of $\mathbf{Q}$ can be obtained, the transition probability matrix can be adjusted to any other time τ by a mapping of the eigenvalues

$$\sigma_i \to \sigma_i^{\tau/t}.$$

The estimation and normalization procedures proposed by Dayhoff et al. (1978) and extended by Jones et al. (1992) are essentially discrete time approximations of this process.

2.2 A Heterogeneous Rates Model

To model the evolution of a sequence of sites where the rate of evolution may vary from site to site, we propose the following generalization of the Markov model (also see Kelly 1991, Kelly and Rice 1994, Yang 1993). Let λ_h be a multiplicative rate factor associated with site h in the sequence $h = 1, \ldots, n$. We assume that λ_h are independent and identically distributed realizations of a random variable Λ. For identifiability, we assume the distribution of Λ has mean one. The evolution at site h may now be described as a Markov process with transition matrix

$$\begin{aligned} \mathbf{P}^{(h)}(t) &= e^{\mathbf{Q}t\lambda_h} \\ &= \mathbf{U}^{-1} e^{\Sigma t \lambda_h} \mathbf{U}. \end{aligned} \tag{1}$$

An "average" replacement matrix that describes the mean behavior across many sites can be obtained by taking the expectation of $\mathbf{P}$ with respect to Λ,

$$\begin{aligned} E_\Lambda \mathbf{P}(t) &= \mathbf{U}^{-1} E_\Lambda \left(e^{\Sigma t \lambda_h} \right) \mathbf{U} \\ &= \mathbf{U}^{-1} \Phi(t) \mathbf{U} \end{aligned} \tag{2}$$

where

$$\Phi(t) = \operatorname{diag}\{\phi(\sigma_1 t), \ldots, \phi(\sigma_{20} t)\}$$

and $\phi(x)$ is the moment generating function of Λ evaluated at x (Kelly and Rice 1994).

Under this model, a replacement matrix can be adjusted to a time τ by the mapping

$$\sigma_i \to \phi\left(\frac{\tau}{t}\phi^{-1}(\sigma_i)\right).$$

For example, if the rates are gamma distributed with both scale and shape factors equal to α (this gives us a mean rate of one), then the mapping is

$$\sigma_i \to \left(1 - \frac{\tau}{t}(1 - \sigma_i^{\alpha})\right)^{-\alpha}.$$

2.3 Alignment Scores

Under the Markov model of evolution described above, the divergence time t measures the similarity of two protein sequences. An alternative measure used in sequence comparison is the alignment score, S. The calculation of an alignment score depends on assigning similarity scores, s_{ij}, to each pair of amino acids (i, j). The total score for two protein sequences is then just the sum of scores for each pair of amino acids in the sequence. Karlin and Altchul (1990) suggest scores based on the following log-odds ratio:

$$s_{ij} = \log_{\lambda}(\frac{q_{ij}}{\pi_i \pi_j})$$

which is the log (base λ) odds of the pair occuring by evolution as opposed to the pair occuring by chance. Notice that the probability of the pair occuring by evolution is

$$q_{ij} = \pi_i p_{ij}(t)$$

where $\{\pi_i, i = 1, 20\}$ are the amino acid frequencies. This method of calculating alignment scores thus requires a replacement matrix $\mathbf{P}(t)$.

3 Results

3.1 Bias in Replacement Matrices

If there is rate heterogeneity across the sites in a sequence, then adjusting a PAM matrix to time t by multiplying the PAM1 matrix t times will introduce a bias in the resulting replacement matrix. From Section 2.2, the average PAM1 matrix, assuming heterogeneous rates, can be written

$$E_\Lambda \mathbf{P}(1) = \mathbf{U}^{-1}\mathbf{\Phi}(1)\mathbf{U}$$

Assuming homogeneous rates, a replacement matrix $\mathbf{P}(t)$ is obtained by multipling out the PAM1 matrix t times. When rates vary across the positions in a sequence, this method of adjustment introduces the following bias.

$$\begin{aligned} \text{bias} &= \mathbf{P}(1)^t - \mathbf{P}(t) \\ &= \mathbf{U}^{-1}(\mathbf{\Phi}(1)^t - \mathbf{\Phi}(t))\mathbf{U} \end{aligned}$$

The diagonal matrix $\mathbf{\Phi}(1)^t - \mathbf{\Phi}(t)$ has entries $\varphi(\sigma_i)^t - \varphi(\sigma_i t)$. Jensen's inequality implies these eigenvalues will be underestimated when $t > 1$ and overestimated when $t < 1$.

$$\begin{aligned} \varphi(\sigma_i)^t - \varphi(\sigma_i t) &= E_\Lambda(\exp(\Lambda\sigma_i))^t - E_\Lambda(\exp(\Lambda\sigma_i t)) \\ &\begin{cases} \leq & E_\Lambda(\exp(\Lambda\sigma_i t)) - E_\Lambda(\exp(\Lambda\sigma_i t)) = 0 \quad \text{if } t > 1 \\ \geq & E_\Lambda(\exp(\Lambda\sigma_i t)) - E_\Lambda(\exp(\Lambda\sigma_i t)) = 0 \quad \text{if } t < 1 \end{cases} \end{aligned}$$

Equality holds above only when the rates are homogeneous.

An overall measure of the difference between the two replacement matrices is Barry and Hartigan's measure of distance $d = -1/20\log(\det(\mathbf{P}(t)))$. The bias in this distance is

$$\text{bias}(d) = -1/20(\log(\det(\mathbf{P}(1)^t)) - \log(\det(\mathbf{P}(t)))) \qquad (3)$$

$$= \; -1/20\sum_{i=1}^{20}(\log(\varphi(\sigma_i)^t) - \log(\varphi(\sigma_i t)) \qquad (4)$$

$$= \; -1/20\sum_{i=1}^{20}(t\log(\varphi(\sigma_i)) - \log(\varphi(\sigma_i t)). \qquad (5)$$

As in Kelly and Rice (1994), a Taylor's series expansion gives an expression for this bias:

$$\text{bias}(d) \; = \; -1/20\sum_{i=1}^{20}\sum_{n=1}^{\infty}(\frac{K_n t(\sigma_i)^n}{n!} - \frac{K_n(\sigma_i t)^n}{n!}) \qquad (6)$$

$$= \; -1/20\sum_{i=1}^{20}\sum_{n=2}^{\infty}\frac{K_n(t-t^n)(\sigma_i)^n}{n!} \qquad (7)$$

$$= \; 1/40\text{Var}(\Lambda)(t^2-t)\text{tr}(\mathbf{Q}^2) + 1/120K_3(t^3-t)\text{tr}(\mathbf{Q}^3) + \ldots \qquad (8)$$

Thus when $t > 1$ the bias in distance is positive and increases with increasing time and increasing rate variability. When $t < 1$, the bias is negative and increases with decreasing time and increasing rate variability.

If the rates have a Gamma distribution with parameter α, the cumulants are $K_n = (n-1)!\alpha^{1-n}$ (cf. Kendall p.91), and the bias simplifies to

$$\text{bias}(d) \; = \; -1/20\sum_{i=1}^{20}\sum_{n=1}^{\infty}\alpha\frac{(t-t^n)(\sigma_i/\alpha)^n}{n}. \qquad (9)$$

Figures 1 and 2 plot this bias (as a percentage of the Barry and Hartigan distance) against t for various values of α using eigenvalues estimated from Henikoff and Henikoff's (1991) data (see Section 4.1).

3.2 Bias in Alignment Score

Because the alignment scores are functions of the replacement probabilities, they will also be biased when heterogeneous rates are not accounted for. In this section we illustrate the bias in the expected total alignment score for sequences of length n. Suppose that two sequences are produced by evolution according to the heterogeneous rate model with gamma(α) distributed rates and divergence time t, then the expected data matrix is

$$E(\mathbf{N}) = n\mathbf{D}\mathbf{P}(t, \alpha)$$

where $\mathbf{D}$ is a diagonal matrix with entries π_i and $\mathbf{P}(t, \alpha)$ is the average replacement matrix calculated from Equation (2). Suppose also that a divergence time $\hat{t}_{homo}$ is estimated assuming the homogeneous rate model using the method of maximum likelihood, and that a total alignment score, $\hat{S}$, is calculated for the sequences using $\mathbf{P}(\hat{t}_{homo})$. Then the total score will have the following bias on average:

$$\begin{aligned} \text{bias}(\hat{S}) &= \sum_{i,j=1}^{20} E(N_{ij}) \log(\frac{p_{ij}(\hat{t}_{homo})}{\pi_j}) - \sum_{i,j=1}^{20} E(N_{ij}) \log(\frac{p_{ij}(t, \alpha)}{\pi_j}) \\ &= \sum_{i,j=1}^{20} E(N_{ij}) \log(\frac{p_{ij}(\hat{t}_{homo})}{p_{ij}(t, \alpha)}). \end{aligned}$$

The bias in the average alignment scores for a sequence of length $n = 1000$ with $\alpha = 1$ is plotted in Figure 3.

4 Example

In this section we illustrate the bias problems described above using Henikoff's blocks database Version 8.0 (Henikoff and Henikoff (1991)). This database contains 2880 blocks of aligned protein sequences from a number of species.

4.1 Bias in the Replacement Matrix

Using pairs of sequences with more than 85% similarity, we estimated a PAM1 matrix using the method described in Jones et al. (1992). This matrix is shown in Table 1. Tables 2 and 3 compare the PAM250 matrix calculated assuming a homogeneous rate model to the PAM250 matrix calculated assuming a heterogeneous rate model with a gamma rate distribution ($\alpha = 1$). The most noticable difference between these matrices is the overall rate of replacement. The probability of replacing any amino acid for another is the probability of observing amino acid i multiplied by the probability that amino acid i does not change:

$$\sum_i \pi_i(1 - p_{ii}(t))$$

For the homogeneous rate model replacement matrix this probability is .7819; for the heterogeneous rate model replacement matrix the probability is .6281. In order to compare replacement matrices with equal rates of change, we determined the divergence time t in the heterogeneous rate model (with $\alpha = 1$ that yielded a probability of change = .7819; this time is $t = 658$ PAMs. In comparing $\mathbf{P}(658, 1)$ to the PAM250 matrix, we note that rare replacements (those with zero entries in the PAM1 matrix) are much more likely under the heterogeneous rate model. Results obtained with larger values of α are similar in nature but more extreme.

4.2 Testing for Heterogeneous Rates

As discussed in Section 3, the severity of the biases encountered depends not only on whether rates are heterogeneous, but also on the extent of the variability. To illustrate that rates are typically heterogenous with a variability that will cause significant biases, we considered a subset of the blocks database. We fit both models to concatenated block sequences from E. coli and Homo sapiens from the Henikoff and Henikoff data set. The total length of the concatenated sequence was 958 amino acids, and the observed proportion of changes was .60. We estimated the divergence times for the two models, t_{homo} and t_{heter} and the heterogeneous rate model parameter α using the method of maximum likelihood. Maximizing the likelihood is equivalent to minimizing the goodness of fit statistic

$$\chi^2_{homo} = 2\sum_{i=1}^{20}\sum_{j=1}^{20} N_{ij} * \log(\frac{N_{ij}}{n\pi_i P_{ij}(t)}). \tag{10}$$

in the homogeneous rate model, and

$$\chi^2_{heter} = 2\sum_{i=1}^{20}\sum_{j=1}^{20} N_{ij} * \log(\frac{N_{ij}}{n\pi_i P_{ij}(t,\alpha)}). \tag{11}$$

in the heterogeneous rate model. Here n is the length of the sequence, $\{\pi_i, i = 1, \ldots, 20\}$ are the amino acid frequencies, and $\mathbf{P}(t)$ and $\mathbf{P}(t,\alpha)$ are the replacement matrices in the homogeneous rate and heterogeneous rate models, respectively.

For the E.coli - Homo sapiens data set:

$$t_{homo} = 137.66 \quad \chi^2_{homo} = 486.8$$

$$t_{heter} = 276.61 \quad \chi^2_{heter} = 412.9$$

$$\alpha = 1.432$$

The difference in the chi-square statistics is

$$\chi^2_{homo} - \chi^2_{heter} = 73.9 \tag{12}$$

which is highly significant if the statistic is chi-square distributed with one degree of freedom.

The sparseness of the raw data matrix (**N**), may make the chi-square approximation to statistic (12) questionable (see for example Goldman (1993)). To assess the distribution of this statistic, a parametric bootstrap confidence interval was calculated in the following way (also described in Goldman (1993)). To test

H_0 : rates are homogeneous

versus

H_A : rates are heterogeneous

we simulated one thousand data matrices, or one thousand multinomial random variables with parameters $n = 958$ and $\mathbf{p} = \mathbf{DP}(t_{homo})$). For each bootstrap matrix, the difference in chi-square statistics (12) is calculated. A histogram of this statistic is shown in Figure 4. An upper 95^{th} percentile of 2.80 and an upper 99^{th} percentile of 5.68 was obtained from the bootstrapped statistics. This provides significant evidence that the heterogeneous rate model with $\alpha = 1.432$ fits significantly better than the homogeneous rate model.

4.3 Bias in Alignment Scores

In this section we illustrate the biases in alignment scores that occur when heterogeneous rates are ignored using the concatenated Homo sapiens and E.coli sequences described above. We calculated the total alignment score for this pair of sequences for various times first using the homogeneous rate replacement matrix and then using a heterogenous rate replacement matrix assuming gamma distributed rates with $\alpha = 1.432$. These scores are displayed in Figure 5. As in the previous analysis, the homogeneous rate model suggests a divergence time of approximately 138 PAMs (which gives a maximum score of 536.6); whereas the heterogenous rate model suggests a divergence time of approximately 277 PAMs (which gives a maximum score of 573.5). Certainly, estimating divergence time with alignment scores will introduce a significant bias if heterogeneous rates are not considered.

5 Discussion

In an exhaustive matching of the entire protein sequence database, Gonnet et al. made the intriguing observation that "mutation matrices (normalized to a distance of 250 PAM ...) were found to differ, depending on whether they were derived from protein pairs that are distantly homologous or from protein pairs that are closely homologous." This observation may be interpreted as giving evidence that the PAM matrix is inadequate for aligning distantly related proteins. Results presented here, however, provide an alternate explanation for these findings. We demonstrate that the bias in normalizing a mutation matrix to 1 PAM may be positive or negative depending on whether the distance between the proteins is larger or smaller than 1. Gonnet et al. normalized to 250 PAMs, but the results are qualitatively equivalent: distantly homologous protein pairs ($t > 250$ PAMs) will have a positive bias and closely homogous pairs ($t < 250$ PAMs) will have a negative bias.

We note that Yang (1993) has found a heterogeneous rate model with $\alpha = 4$ provides a good fit to nucleotide sequences. Our results suggest that there is less rate variability in amino acid sequences ($\alpha < 2$) than in nucleotide sequences. This result might be expected since the amino acid model does not have to account for the rate differences in different codon positions. Estimates of rate variation in amino acid repalcements obtained by other methods (Uzzell and Corbin 1971, Ota and Nei 1994) suggest a value of approximately $\alpha = 2$. The true value of α will depend on the particular set of sequences under study. Thus it may not be possible to obtain a definitive value of α that is applicable to all amino acid sequences.

The results presented here demonstrate that there is a bias introduced into estimated replacement matrices due to rate heterogeneity. A method for correctly extrapolating replacement matrices estimated from closely related sequences to matrices appropriate for longer divergence times is provided. The evolution of protein sequences is a complex process and the methods described here are based on a number of simplifying assumptions. For example, the mutational spectrum as defined by the rate matrix $\mathbf{Q}$ in Equation (1) is likely to vary from site to site. This leads us to ask, does it make sense to estimate an average replacement matrix? For many practical purposes, such as database searching and sequence alignment, scoring methods based on average replacement matrices have proven to be very effective. We are hopeful that the bias correction proposed here will serve to improve their utility.

References

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

Altschul, S.F. 1993. A protein alignmnet scoring system sensitive at all evolutionary distances. *J. Mol. Evol.* 36, 290-300.

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

Brown M., Hughey R., Krogh A., Mian I.S., Sjölander K., Haussler D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. In *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology.*

Dayhoff M.O., Schwartz R.M., Orcutt B.C. 1978. A model of evolutionary change in proteins, pp. 345-352. In *Atlas of protein sequence structure, vol. 5, suppl. 3* M.O. Dayhoff (ed.., National Biomedical Research Foundation, Washington D.C.

Feng D.A., Doolittle R. 1985. Aligning amino acid sequences, comparison of commonly used methods. *J. Mol. Evol.* 21, 112-125.

Goldman N. 1993. Statistical tests of models of DNA substitution *J. Mol. Evol.* 36, 182-198.

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

Henikoff, S., Henikoff, J.G. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 19, 6565-6572.

Holmquist, R., Goodman, M., Conroy, T., Czelusniak, J. 1983. The spatial distribution of fixed mutations within genes coding for proteins. *J. Mol. Evol.* 19, 437-448.

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

Karlin, S., Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264-2268.

Kelly, C. 1991. PhD dissertation, UC San Diego

Kelly, C., Rice, J. 1994. Modeling molecular evolution: a heterogeneous rate analysis. *Mathematical Biosciences* (to appear)

Ota, T., Nei, M. 1994. Estimation of the number of amino acid substitutions per site when the substitution rate varies across site. *J. Mol. Evol.* 38, 642-643.

Reeves, J.A. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA *J. Mol. Evol.* 35, 17-31.

Sander, C., Schnieder, R. 1991. Database homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9, 56-68.

Schwartz, R.M., Dayhoff, M.O. 1978. Matrices for detecting distant relationships, pp/353-358 in *Atlas of protein sequence structure, vol. 5, suppl. 3* M.O. Dayhoff (ed.., National Biomedical Research Foundation, Washington D.C.

Stuart, A., Ord, J.K. 1987. *Kendall's Advanced Theory of Statistics.* Fifth Edition, Vol 1, Oxford Univeristy Press, New York

Taylor, W.R. 1986. The classification of amino acid conservation *J. Theor. Biol.* 119, 205-218.

Uzzell, T., Corbin, K.W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172, 1089-1096.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 if human mitchondrial DNA *J. Mol Evol.* 37, 613-623.

Wilbur, J.W. 1985. On the PAM matrix model of protein evolution. *Mol. Biol. Evol.* 2, 434-447.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10, 1396-1401.

Yang, Z., Goldman, N., Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic inference. *Mol. Biol. Evol.* 11, 316-324.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
9872	5	3	4	0	20	1	2	3	3	1	4	6	4	2	41	19	10	0	1
19	9907	0	0	2	3	0	7	0	2	1	2	0	0	2	15	12	22	0	2
4	0	9901	56	0	4	3	0	1	0	0	18	1	2	1	5	2	1	0	1
5	0	47	9909	0	3	1	1	11	0	0	2	1	12	1	3	2	2	0	0
0	1	0	0	9935	0	0	3	0	17	3	0	0	0	0	2	0	1	1	34
18	1	3	2	0	9951	1	0	1	0	0	5	0	1	2	13	2	1	0	0
1	0	4	2	1	2	9925	0	3	1	0	15	1	15	7	9	2	0	1	10
2	3	0	1	3	0	0	9833	1	33	13	1	0	0	0	1	9	98	0	0
5	0	1	13	0	2	2	1	9914	3	1	6	1	11	30	3	5	1	0	0
2	0	0	0	8	0	0	18	2	9929	18	1	1	2	3	2	2	12	0	0
3	1	0	0	6	1	0	34	3	86	9822	1	0	4	1	2	11	23	0	0
9	1	26	3	0	11	13	2	9	2	0	9862	1	7	2	29	17	3	0	2
10	0	1	1	0	1	1	1	1	2	0	1	9963	2	2	9	5	1	0	0
9	0	3	24	0	3	16	1	19	6	3	8	3	9879	8	6	7	2	0	4
3	1	0	1	0	3	5	0	31	5	1	2	2	5	9934	4	2	0	2	0
60	6	5	3	2	21	5	1	4	3	1	19	7	3	3	9805	48	2	0	2
24	4	1	2	0	2	1	8	5	3	4	10	3	3	2	42	9874	11	0	0
10	6	1	2	1	1	0	75	1	17	7	1	1	1	0	2	9	9865	0	0
0	0	0	0	2	1	1	0	0	2	0	0	0	0	7	2	0	1	9980	3
1	1	1	0	49	1	10	1	0	1	0	2	0	3	0	3	0	0	1	9923

Table 1: PAM1 Matrix estimated using Henikoff and Henikoff's data.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1374	288	433	491	171	1293	245	432	424	605	150	418	519	271	351	831	899	633	37	137
1039	1194	240	268	311	733	183	744	290	862	211	312	260	182	298	683	835	1092	52	212
658	101	2011	2032	101	719	342	184	562	273	70	606	208	416	315	496	479	280	21	123
623	94	1699	2297	91	626	305	198	749	305	78	481	209	513	418	444	441	298	23	107
280	141	109	117	3139	190	288	470	136	1411	247	144	95	150	143	241	249	519	128	1804
1159	182	424	442	104	3515	214	212	334	310	81	409	271	216	306	751	607	335	32	97
582	121	535	571	417	566	1875	199	572	416	94	574	280	583	644	526	487	274	97	588
594	284	166	215	395	325	115	1444	229	2026	445	195	188	152	208	376	605	1832	38	169
608	115	530	846	119	533	345	239	1928	524	119	388	265	506	1463	446	509	322	77	118
445	176	132	176	633	254	129	1082	269	3048	528	159	182	170	296	295	443	1259	61	261
529	206	163	216	530	318	140	1139	292	2530	557	190	181	189	281	353	547	1368	49	224
902	187	861	818	189	985	522	307	584	467	117	767	309	375	438	712	772	439	40	210
893	124	236	284	99	520	203	235	318	427	88	246	4100	234	311	618	615	344	21	80
687	128	695	1025	232	612	623	280	896	589	136	441	345	849	713	503	549	382	52	261
522	123	309	490	130	509	404	225	1521	600	119	302	269	418	2587	395	418	294	237	126
1198	273	471	504	212	1207	319	394	449	580	144	475	517	286	382	839	950	567	52	180
1130	291	397	437	191	851	258	553	446	758	195	450	449	272	353	829	1199	766	35	141
665	319	194	247	333	393	121	1401	236	1801	409	214	210	158	208	413	641	1856	36	145
176	69	68	87	373	174	195	131	258	400	67	89	60	99	764	173	132	164	6163	359
320	137	189	197	2568	253	577	287	191	829	149	226	109	240	197	291	261	323	175	2481

Table 2: PAM250 Matrix - Estimated Assuming Homogeneous Rates.

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
3033	233	327	381	145	999	180	323	329	475	116	320	396	216	271	837	755	513	34	115
841	3150	189	216	247	535	142	553	224	622	150	234	196	138	244	566	650	885	46	173
496	80	3606	1703	100	545	253	160	393	258	61	536	172	289	242	385	355	237	24	104
485	76	1424	3845	93	480	216	173	586	274	67	335	173	429	304	333	331	260	26	92
238	112	108	120	4544	187	201	360	128	1078	189	117	94	117	131	207	206	392	99	1372
896	133	321	339	102	4919	163	177	259	283	70	311	209	165	245	590	426	272	31	87
427	93	396	404	291	432	3703	172	405	353	79	491	219	496	486	427	357	233	77	459
445	211	145	187	302	271	100	2827	190	1554	380	156	156	120	172	281	481	1851	36	137
472	89	371	661	112	414	245	198	3665	426	99	312	209	416	1144	343	401	263	61	100
349	127	125	159	483	232	109	831	219	4504	472	131	156	142	242	236	332	909	53	191
409	147	142	185	406	273	117	973	244	2262	2127	151	148	168	224	277	464	1072	43	168
690	140	762	570	155	750	446	245	470	384	93	2548	232	301	325	677	661	346	37	170
682	94	195	234	99	401	159	195	251	365	73	185	5315	185	248	483	458	281	24	76
549	97	482	857	181	468	530	221	737	490	121	354	272	2717	532	393	435	305	45	212
404	101	237	356	119	408	305	187	1189	490	95	224	214	312	4204	310	318	244	176	106
1206	226	365	378	182	949	259	295	345	463	113	452	404	223	300	2253	965	424	47	150
949	227	294	328	158	598	189	439	352	569	166	385	334	215	268	842	2934	605	33	115
540	258	164	215	252	319	103	1415	193	1300	320	168	172	126	172	309	506	3310	35	121
165	61	77	97	289	168	155	124	202	343	59	81	66	85	568	156	125	158	6748	274
269	112	160	169	1953	226	451	233	163	607	111	183	103	194	166	243	214	269	133	4043

Table 3: PAM250 Matrix - Estimated Assuming Heterogeneous Rates with $\alpha = 1$

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
1818	233	425	499	264	1007	250	433	433	718	161	352	447	262	386	712	717	617	72	195
839	1592	322	378	347	706	221	594	358	859	192	295	311	211	358	569	662	863	82	240
645	136	2077	1332	219	728	305	305	501	532	118	482	291	325	381	471	485	421	63	182
634	133	1113	2271	214	682	284	313	611	545	122	378	291	398	426	441	469	434	65	174
433	158	236	276	2718	405	270	479	268	1181	221	206	207	190	267	328	363	571	139	1084
902	176	429	482	222	3026	240	325	392	560	126	354	326	234	368	604	543	456	68	169
594	145	476	531	392	635	1979	322	499	623	134	448	321	430	534	487	481	427	113	430
596	227	276	339	402	499	186	1673	324	1478	334	237	270	194	305	395	554	1411	74	227
620	143	472	690	235	627	301	337	2084	663	145	348	317	389	951	441	506	447	102	183
528	175	257	316	530	459	193	790	340	2874	386	219	264	204	347	360	464	929	90	274
568	188	274	337	474	494	198	855	357	1849	1080	234	264	220	339	388	537	1009	82	254
760	177	684	643	271	851	407	373	525	643	143	1344	337	317	429	625	657	506	76	231
769	148	329	394	218	626	233	338	381	619	129	269	3161	242	367	536	557	462	61	159
665	149	543	796	294	663	459	357	688	706	158	373	357	1404	565	471	527	478	85	260
574	148	374	499	242	612	335	330	988	704	144	296	318	332	2411	418	455	434	197	189
1026	227	447	501	288	971	295	414	444	707	159	417	449	268	404	1299	815	557	82	216
901	231	402	465	278	761	255	506	444	794	192	382	407	261	384	711	1676	682	72	198
649	252	292	359	367	535	189	1079	328	1330	301	247	282	198	306	414	571	2016	73	213
346	109	198	244	408	365	228	259	341	588	112	168	169	161	633	273	273	333	4447	344
454	155	280	320	1542	440	423	384	298	870	168	249	216	239	297	352	367	473	167	2306

Table 4: PAM658 Matrix - Estimated Assuming Heterogeneous Rates with $\alpha = 1$

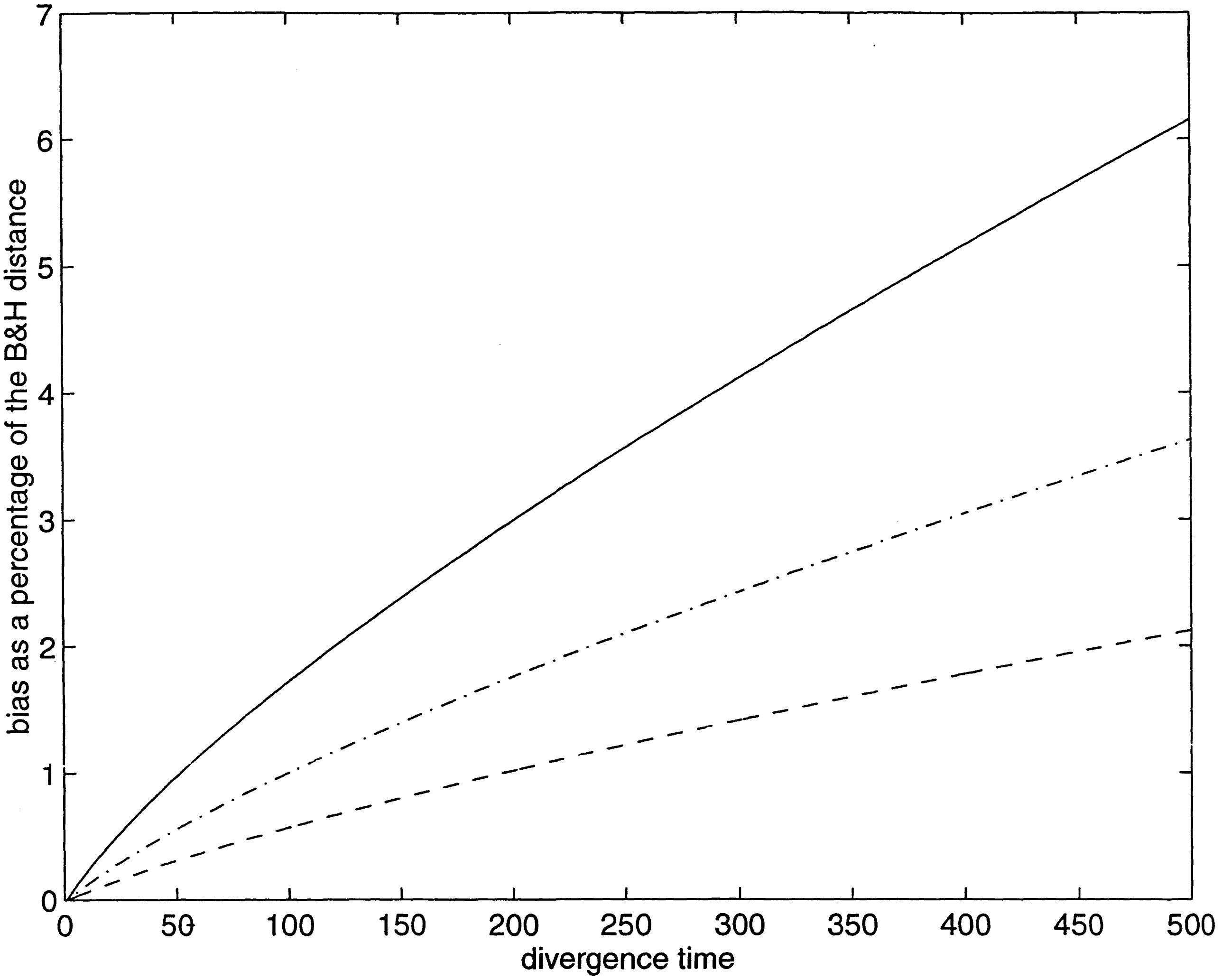

Figure 1: Bias in Barry and Hartigan's distance measure as a percentage of the distance for divergence times > 1. A heterogeneous rate model with gamma distributed rates was assumed ($\alpha = 1$: - - , $\alpha = 2$: - . , $\alpha = 4$: -).

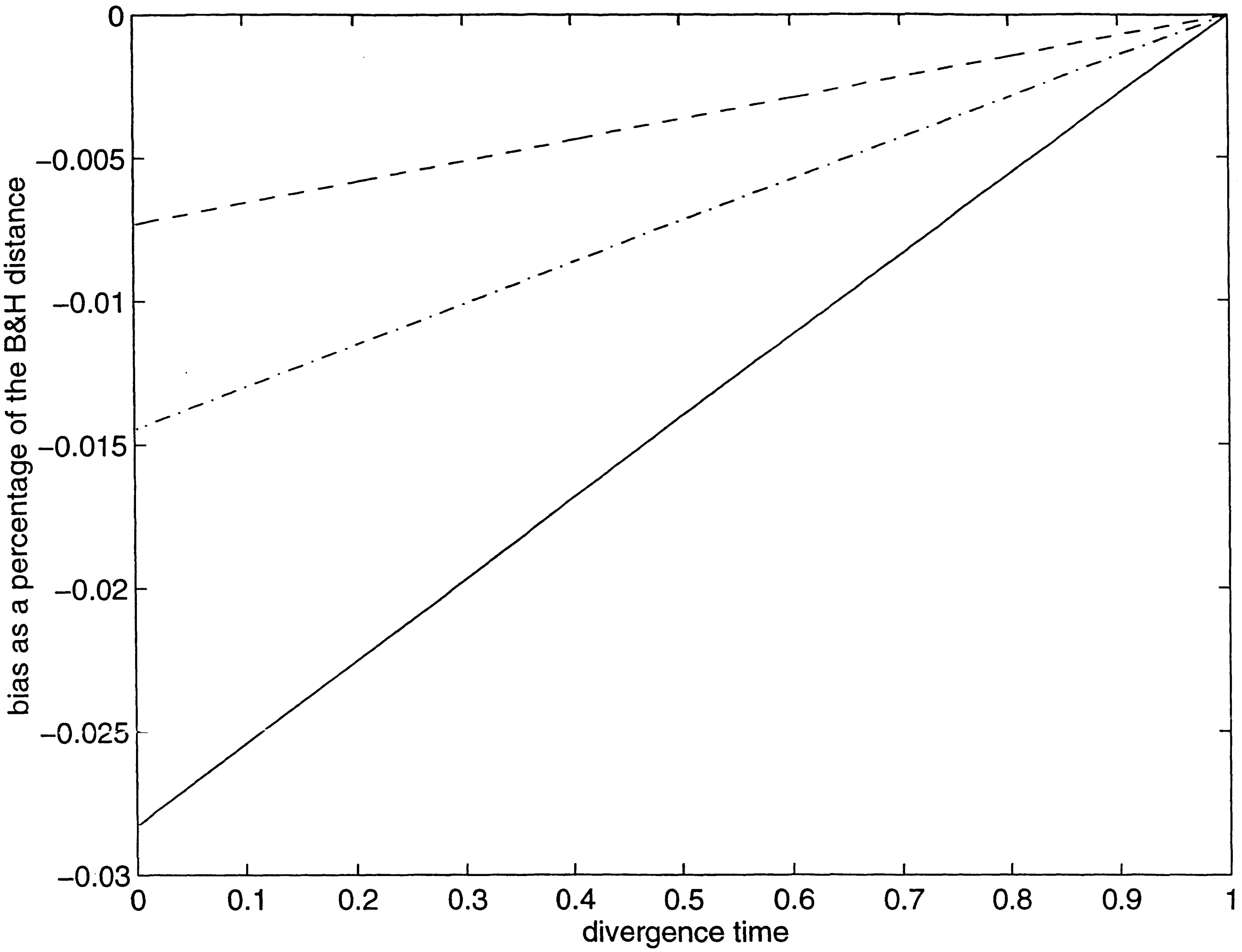

Figure 2: Bias in Barry and Hartigan's distance measure as a percentage of the distance for divergence times < 1. A heterogeneous rate model with gamma distributed rates was assumed ($\alpha = 1$: - - , $\alpha = 2$: - . , $\alpha = 4$: -).

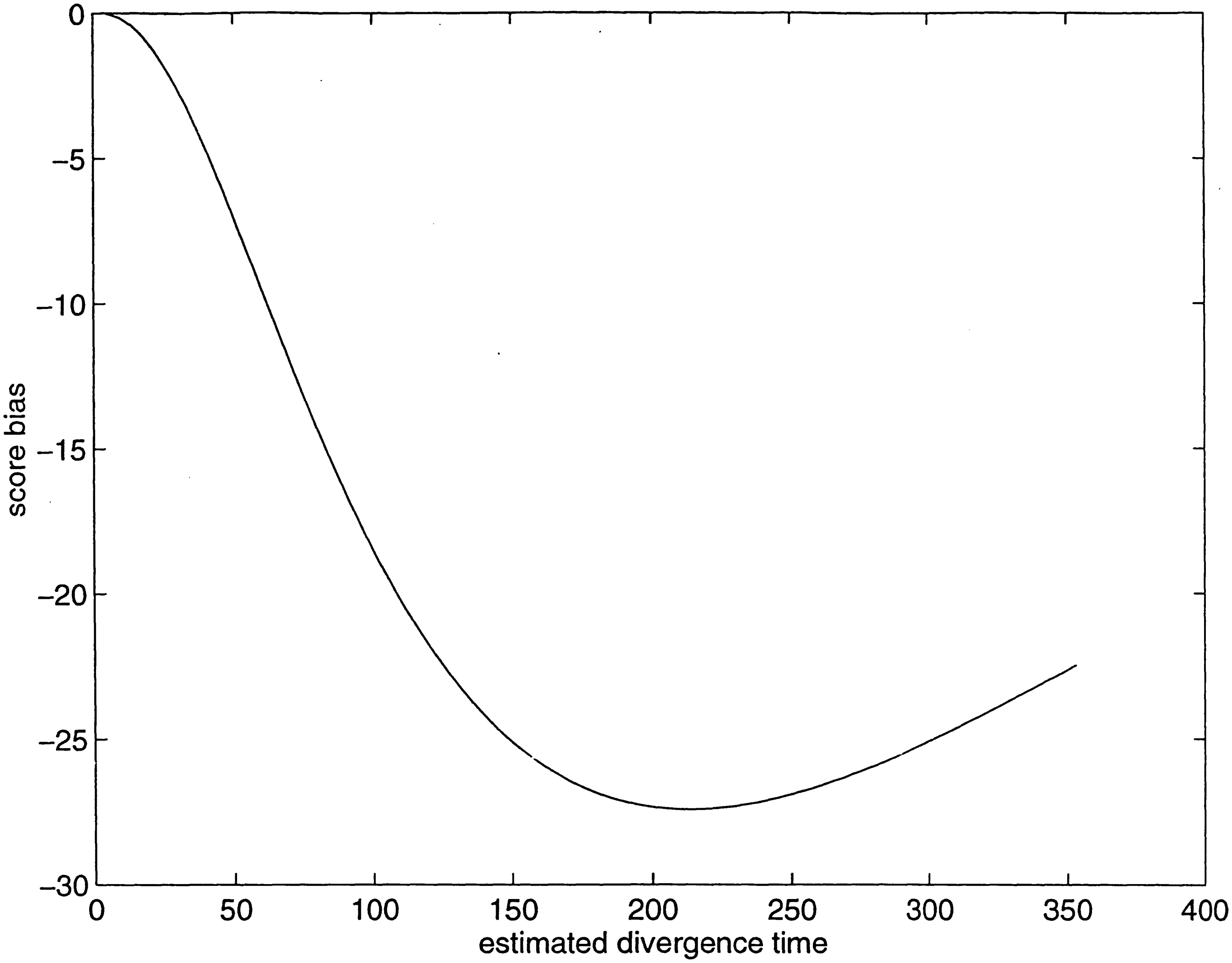

Figure 3: Bias in the expected alignment score for sequences of length $n = 1000$ as a function of divergence time when rates have a gamma($\alpha = 1$) distribution.

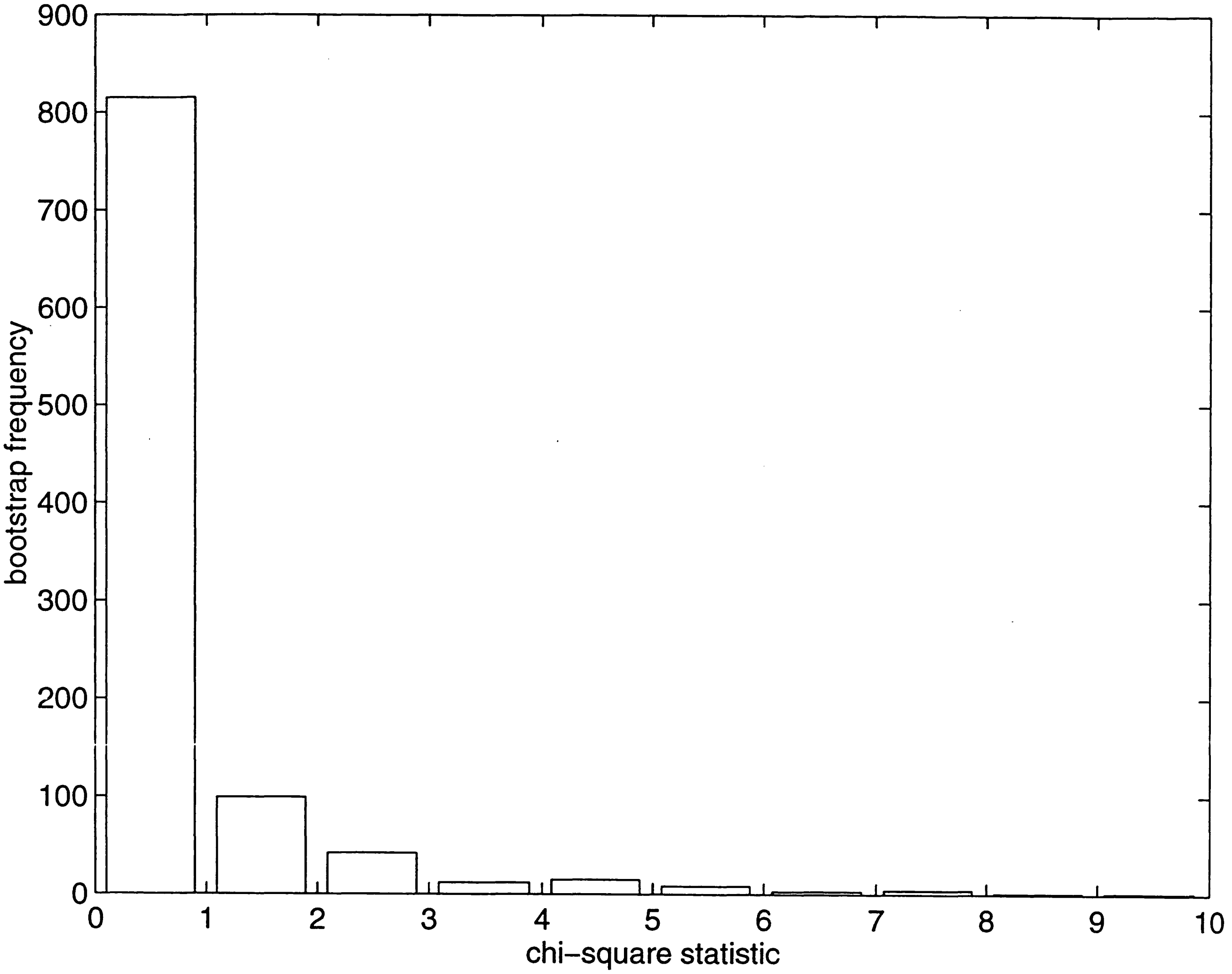

Figure 4: Histogram of bootstrapped chi-square statistics.

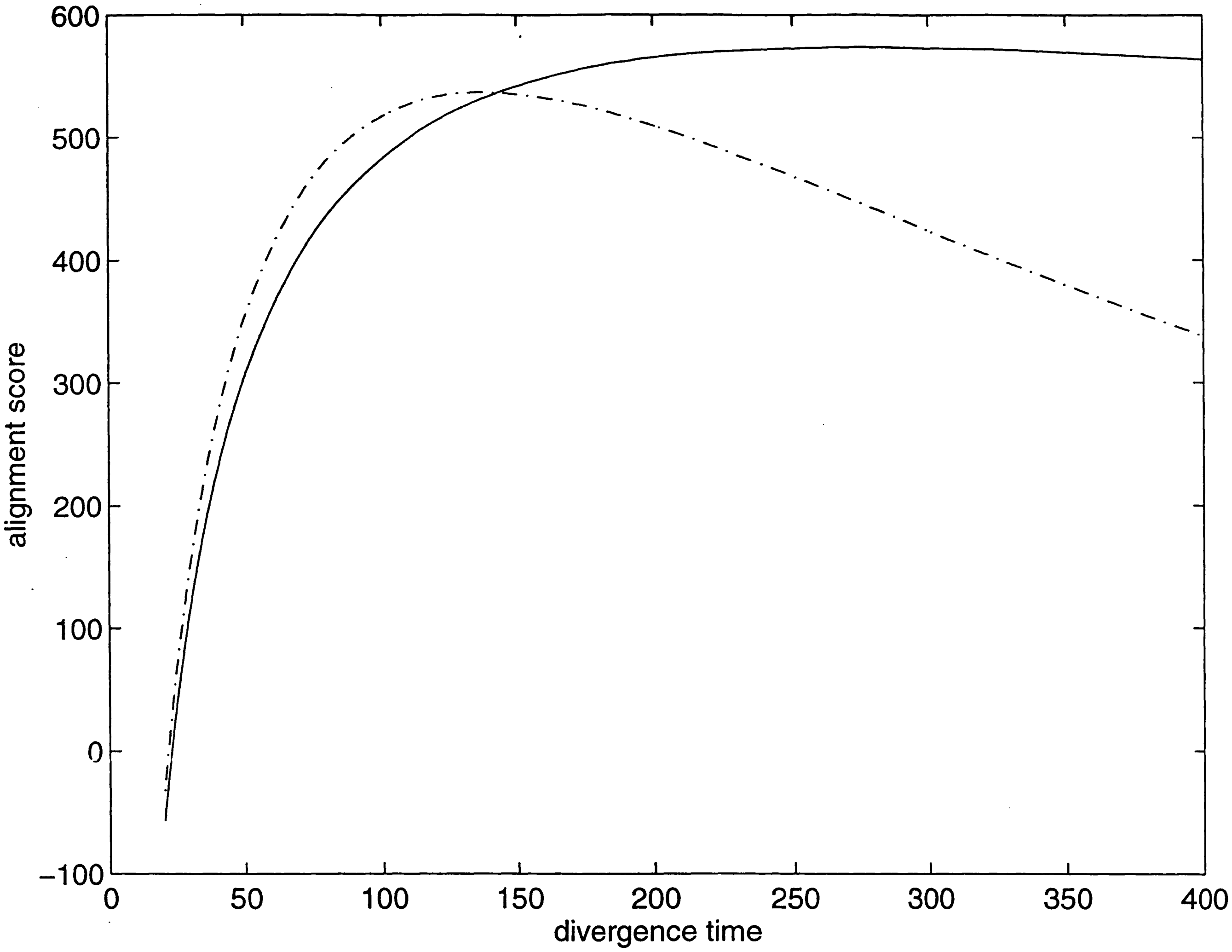

Figure 5: Alignment scores versus divergence time calculated for the Homo sapien - E.coli dataset assuming a homogeneous rate model (- .) and a heterogeneous rate model with $\alpha = 1.432$ (-).