Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior

Itay Mayrose, Dan Graur,† Nir Ben-Tal,‡ and Tal Pupko**

*Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel; †Department of Biology and Biochemistry, University of Houston; and ‡Department of Biochemistry, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv, Israel

The degree to which an amino acid site is free to vary is strongly dependent on its structural and functional importance. An amino acid that plays an essential role is unlikely to change over evolutionary time. Hence, the evolutionary rate at an amino acid site is indicative of how conserved this site is and, in turn, allows evaluation of its importance in maintaining the structure/function of the protein. When using probabilistic methods for site-specific rate inference, few alternatives are possible. In this study we use simulations to compare the maximum-likelihood and Bayesian paradigms. We study the dependence of inference accuracy on such parameters as number of sequences, branch lengths, the shape of the rate distribution, and sequence length. We also study the possibility of simultaneously estimating branch lengths and site-specific rates. Our results show that a Bayesian approach is superior to maximum-likelihood under a wide range of conditions, indicating that the prior that is incorporated into the Bayesian computation significantly improves performance. We show that when branch lengths are unknown, it is better first to estimate branch lengths and then to estimate site-specific rates. This procedure was found to be superior to estimating both the branch lengths and site-specific rates simultaneously. Finally, we illustrate the difference between maximum-likelihood and Bayesian methods when analyzing site-conservation for the apoptosis regulator protein Bcl-x_L.

Introduction

Rates of evolution in proteins are expected to vary among sites due to different selective constraints. Under the neutral theory of molecular evolution, amino acid positions that are under stringent selective constraints are expected to be highly conserved; positions that are more tolerant to replacement are most often variable (Kimura 1983). Conserved sites may point to functionally and structurally important regions involved in such activities as ligand binding, enzymatic activity, protein-protein interactions, or folding (Lichtarge and Sowa 2002).

Numerous site-specific conservation scores have been proposed over the years (reviewed in Valdar 2002; see also del Sol Mesa, Pazos, and Valencia 2003, Yao et al. 2003). Though evolution is the driving force that determines site conservation, none of these methods make full use of either the information contained in the phylogenetic tree or the stochastic nature of amino acid replacements. This deficit may lead to erroneous predictions. For example, when branch lengths are ignored, a replacement on a short branch will be given the same weight as one occurring on a long branch. However, an amino acid replacement between two divergent sequences is less surprising than one occurring between two closely related sequences. The incorporation of advanced evolutionary models was proved to greatly increase the accuracy of site-specific rate inference (Pupko et al. 2002). Evolutionary rates are commonly measured as number of replacements per amino acid site per year. The term site-specific evolutionary rate in the context of our conservation scores is different. Here, the rate is relative to the average evolutionary rate across all sites and hence is unitless. In addition, for each site we assume that the rate is constant across all lineages. Finally, in this paper we limit our discussion to site-specific rate inference that is based on probabilistic evolutionary models.

Currently, likelihood methods are considered state-of-the-art phylogenetic techniques, allowing robust statistical testing of evolutionary hypotheses (Whelan, Lio, and Goldman 2001). Several alternatives within the likelihood framework are currently being used for inferring evolutionary rates. These can be divided into two types: (1) Bayesian methods that presuppose a prior distribution of evolutionary rates, and (2) maximum-likelihood (ML) methods that do not. Both approaches have solid statistical foundations and are closely related, as they use the same models of evolution and operate within the same statistical framework.

The ML approach for estimating site-specific conservation scores chooses the rate that yields the highest probability to the observed data. The first site-specific rate estimation using ML was the DNArates program developed in the early 1990s by Gary Olsen. A paper describing DNArates was never published, but documentation can be found at http://geta.life.uiuc.edu/~gary/programs/DNArates.html (see also Felsenstein 2001). Nielsen (1997) also studied ML based estimation for DNA sequences and suggested incorporating a Gamma prior to avoid cases where the ML estimate is infinite. Using the same ML methodology, Pupko et al. (2002) developed the Rate4Site tool for the identification of functional regions in proteins. Rate4Site was embedded in the ConSurf server (Glaser et al. 2003; http://consurf.tau.ac.il) and successfully identified functional residues at the contact interface of several proteins (Donaudy et al. 2003; Mella et al. 2003; Ramelot et al. 2003; RamShankar et al. 2003).

Bayesian inference is based on the posterior probability distribution, which is directly proportional to the product of the prior distribution and the likelihood. A Bayesian approach, assuming a Gamma prior for DNA sequences, was suggested by Yang and collaborators (Yang and Wang 1995; Excoffier and Yang 1999). Computing a Bayesian estimate based on a continuous Gamma

Key words: rate variation among sites, evolutionary conservation, empirical Bayesian methods, bioinformatics, Bcl-x_L.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 21(9):1781–1791. 2004
doi:10.1093/molbev/msh194
Advance Access publication June 16, 2004

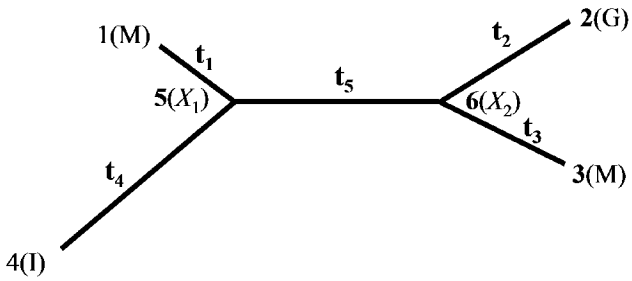

FIG. 1.—A four-taxon unrooted tree used to illustrate the likelihood calculations. The external nodes (leaves) are labelled 1 to 4; internal nodes are labelled 5 to 6. Branch lengths are marked by t_i, where i is the branch identifier. Capital letters in parentheses are one-letter abbreviations for amino acids.

distribution is computationally impracticable for even a modest number of sequences (Yang 1996). Yang (1994) suggested the discrete Gamma model as an approximate method and found that four categories are sufficient to provide a decent approximation to the continuous Gamma distribution.

Site-specific evolutionary rates are directly connected to the branch lengths of the phylogenetic tree (see *Materials and Methods*). The problems of estimating branch lengths and site-specific rates are thus inseparable. Two possible solutions exist: (1) estimate branch lengths first and then estimate site-specific rates, assuming that the branch lengths are known (e.g., Nielsen 1997, Pupko et al. 2002); or (2) estimate branch lengths and site-specific rate simultaneously, using an iterative procedure (e.g., Meyer and von Haeseler 2003).

The purpose of this study is to compare the performances of ML and Bayesian estimates through simulation. First, we shall study the effect of the number of discrete Gamma categories on the performance of the Bayesian method for the task of evaluating site-specific rates. Then, we shall study the effect of various evolutionary parameters, such as number of taxonomic units, branch lengths, sequence length, and the shape of the rate distribution, on the quality of predictions. These comparisons assume that the tree topology and branch lengths are known prior to rate inference. We then explore the accuracy of rate estimation in the more realistic scenario where branch lengths are not known in advance. We conclude with an illustrative biological example.

Materials and Methods

Maximum-Likelihood Estimation of Evolutionary Rates

The branch lengths of the phylogenetic tree represent the average evolutionary rate across all sites. A site-specific rate, r, indicates how fast this site evolves relative to the average. A rate of 2.0 indicates a site that evolves two times faster than the average. Thus, site-specific rates inferred here are not absolute evolutionary rates that require knowledge of divergence times, but rather they represent a comparative quantity. We follow Yang (1993) and present the likelihood computation using an example tree shown in figure 1. We assume here that the tree $T = (\tau, t)$, defined by its tree topology τ and associated branch lengths t, is known in advance. Nodes are labeled as in figure 1. The probability of the data given the rate parameter r is

$$P(data \mid r, T) = \sum_{X_1, X_2 \in \{Amino-acids\}} \pi_{X_1} \times P_{X_1,M}(rt_1) \times P_{X_2,G}(rt_2) \times P_{X_2,M}(rt_3) \times P_{X_1,I}(rt_4) \times P_{X_1,X_2}(rt_5), \quad (1)$$

where π_{X_1} is the frequency of amino acid X_1, and $P_{X_1,X_2}(rt)$ is the probability that amino acid X_1 will be replaced by amino acid X_2 along a branch of length t, given that the evolutionary rate at the site in question is r. Internal node 5 was arbitrarily chosen as the root of the tree. Because the model we have used is time-reversible, the tree could have its root anywhere with no effect on the calculations (Felsenstein 1981). Given r, the likelihood $P(data \mid r, T)$ can be calculated using Felsenstein's (1981) postorder tree traversal algorithm. The ML rate estimate is the rate that maximizes the likelihood function $P(data \mid r, T)$. In the rare case where all the characters at the leaves are different, the ML value of r is infinite (see also Nielsen 1997). To avoid this, we set an upper bound on r ($r_{\max} = 20.0$).

Empirical Bayesian Estimation of Evolutionary Rates

In the Bayesian case, a prior Gamma distribution over the rates is assumed (Jin and Nei 1990; Swofford et al. 1996; Yang 1996). The Gamma distribution with parameters α and β has a mean α/β and variance α/β^2. We set $\alpha = \beta$ so that the mean rate over all sites is 1.0 and the variance is $1/\alpha$. The shape of the Gamma distribution is then determined by α. When $\alpha > 1$, the distribution is bell-shaped, suggesting little rate heterogeneity. When $\alpha \to \infty$, there is a single rate for all sites. In the case of $\alpha < 1$, the distribution is highly skewed and is L-shaped. This situation indicates high levels of rate variation.

Within the Bayesian framework, the posterior probability is obtained from the likelihood function and the prior probability. Assuming that the topology, the branch lengths and α are known a priori, the probability of any given rate, r, is

$$P(r \mid data, T) \cong \frac{P(data \mid r, T)P(r)}{\sum_{i=1}^{k} P(data \mid r_i, T)P(r_i)}, \quad (2)$$

where $P(data \mid r, T)$ is obtained from equation (1), and $P(r)$ is the prior distribution on the rates. Here k discrete rate categories are used to approximate the continuous Gamma function (hence the approximation sign in eqs. 2 and 3), such that all categories have equal prior probabilities ($1/k$). The mean of each category, r_i, is used to represent all the rates within that category. The estimated rate is the expected value of the parameter:

$$E(r \mid data, T) \cong \sum_{i=1}^{k} P(r_i \mid data, T)r_i = \frac{\sum_{i=1}^{k} P(data \mid r_i, T)P(r_i)r_i}{\sum_{i=1}^{k} P(data \mid r_i, T)P(r_i)}. \quad (3)$$

Above it is assumed that the α parameter and the branch lengths are known a priori. This is rarely the case

when analyzing real data sets. If α is unknown and only the branch lengths are known a priori, one may estimate α by maximizing $P(data \mid \alpha, T)$ using a discrete distribution to approximate the Gamma distribution (Yang 1994; Yang and Wang 1995). The estimated α can then be used in the prior Gamma distribution for the Bayesian method. The replacement of α by its estimate has an empirical Bayesian justification (Yang and Wang 1995). Empirical Bayesian approaches differ from other Bayesian methods in that the prior is determined, in part, by the data (Leonard and Hsu 1999). Computing the rate estimate using equation (3) with an empirical Bayesian estimate of α is referred here as EB-EXP.

Estimating Branch Lengths

When the branch lengths are unknown one may estimate the branch lengths using the classical ML approach and then treat these branch lengths as known for the task of rate estimation, using either the ML or the Bayesian method. When inferring the branch length in this case we assumed a Gamma distribution and found the ML estimates of the α parameter and the branch lengths simultaneously. Alternatively, in the maximum-likelihood framework one can consider a rich model in which each site has its own rate. The tree and branch lengths can be estimated using this model (see, e.g., Meyer and von Haeseler 2003). In this case, assuming that the tree topology is known a priori, the parameters of the model (i.e., the site-specific rates) and the branch lengths are estimated simultaneously by an iterative procedure. In each iteration we first estimate site-specific rates, given the branch lengths. We then find the ML estimate of the branches given the rates. We continue until convergence of the likelihood function. We call this rich-model method ML-RICH.

Branch lengths and site-specific rate estimates are not independent. One can always multiply all branch lengths by a constant factor c and divide all rates by this factor, resulting in no change in the likelihood score. To avoid this circularity, in all methods, site-specific rates were scaled so that the average is 1. Our simulations indicate that scaling has a negligible effect on EB-EXP, whereas it increases the accuracy of the ML methods (data not shown).

Simulation

A simulated site-specific rate parameter was assigned to each site. Given a model tree and simulated rates, protein sequences were generated by simulating evolutionary changes along the branches. The simulation used the JTT model of amino acid replacement (Jones, Taylor, and Thornton 1992), in which each site evolves independently. For each run a total of 500 sites were generated in this manner.

For the simulation, one must determine the "true" rate in each site. If the true rates are sampled from a Gamma distribution, this could bias the results toward the Bayesian method, which assumes a Gamma prior. To avoid this bias, the rates used in our simulations were drawn from an empirical rate distribution inferred from a biological multiple sequence alignment (MSA) with many homologs.

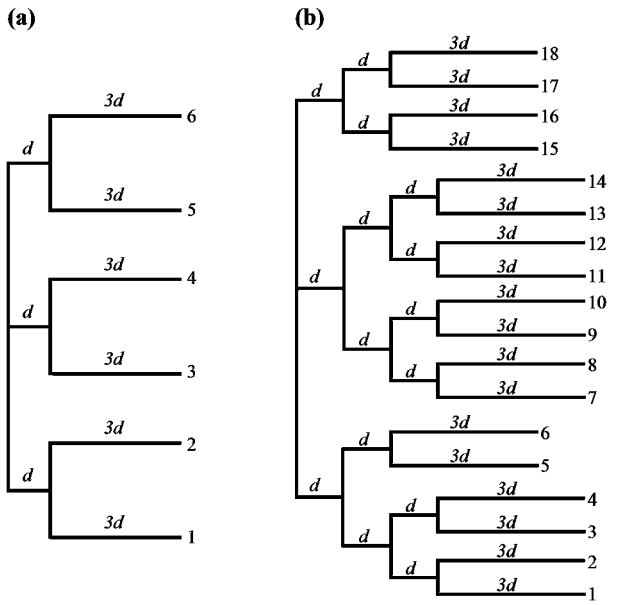

FIG. 2.—Illustrative unrooted model tree with (*a*) six sequences and (*b*) 18 sequences. The lengths of the internal and external branches are d and $3d$, respectively.

We used a distribution inferred by EB-EXP from an MSA of 34 homologous *Src*-homology-2 (SH2) domains (Pupko et al. 2002). The simulated rates were scaled so that the average was set to 1. To avoid a possible bias because the rate distribution was inferred by a specific method, the same MSA was used to infer a second empirical distribution using ML. The simulation results obtained with this distribution were similar with regard to the relative accuracy of the methods (data not shown).

Due to the complexity of the parameter space, we studied only several special cases. In all trees used, the length of interior branches was d and that of the exterior branches was $3d$. Our simulation runs varied in their value of d, number of sequences, sequence lengths, and rate distributions used for generating the data. Illustrative model tree with six and 18 sequences are shown in figure 2. The generated sequences, along with the model tree, were given as input to the EB-EXP and ML methods and rates were inferred for each position. We note that in these simulations the tree topology and branch lengths were assumed to be known a priori. The α parameter of EB-EXP was inferred from the data for each run. A different set of simulations were performed to study the more complicated case in which the branch lengths are not given a priori (see below). In each simulation run, the accuracy of inference was analyzed by computing the mean square error (MSE) between the simulated rates and the respective inferred estimates. MSE was calculated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}((\text{true } r^i) - (\text{estimated } r^i))^2, \qquad (4)$$

where n is the sequence length. Low MSE values indicate high accuracy. Ten simulation runs were performed for each simulated condition. As an accuracy measure we used mean MSE over these 10 runs.

Table 1
Simulated Data Sets Used for Testing the Dependency Between Number of Categories and Accuracy

Data set	Number of Sequences	Tree	Rate Distribution
1	6	Predetermined as in figure 2*a*, $d = 0.1$	Gamma, $\alpha = 0.3$
2	6	Predetermined as in figure 2*a*, $d = 0.1$	Gamma, $\alpha = 1.0$
3	18	Predetermined as in figure 2*b*, $d = 0.1$	Gamma, $\alpha = 0.3$
4	18	Predetermined as in figure 2*b*, $d = 0.1$	Gamma, $\alpha = 1.0$
5	6	NJ tree inferred from lysozyme *c* data set	Inferred from the data using ML
6	34	NJ tree inferred from SH2 data set	Inferred from the data using ML
7	24	NJ tree inferred from SH3 data set	Inferred from the data using ML

Choosing the Number of Discrete Gamma Categories

When using the discrete approximation to the continuous Gamma distribution, the more categories that are used the better the approximation will be. However, the computation time increases linearly with the number of categories. Thus, we evaluated the minimum number of categories needed to provide an acceptable approximation to the continuous Gamma distribution. To make sure that our results are not specific to a single tree, or to a specific MSA, seven different data sets were tested (table 1). In four data sets (1–4 in table 1) the dependency between the number of categories and the accuracy of the inferred rates was tested on the phylogenetic trees as in figure 2. The rate at each position was drawn from a Gamma distribution with a given value of α. Two values of α were considered: $\alpha = 0.3$ represents a severe among-site rate variation while $\alpha = 1.0$ is an example of little among-site rate variation. In data sets 5–7 (table 1), the trees used for the simulations were based on neighbor-joining (NJ) trees (Saitou and Nei 1987) inferred from real data sets. In this case, the rate at each position was drawn from a rate distribution that was obtained by analyzing the three real data sets using ML. In all cases 500 positions were simulated.

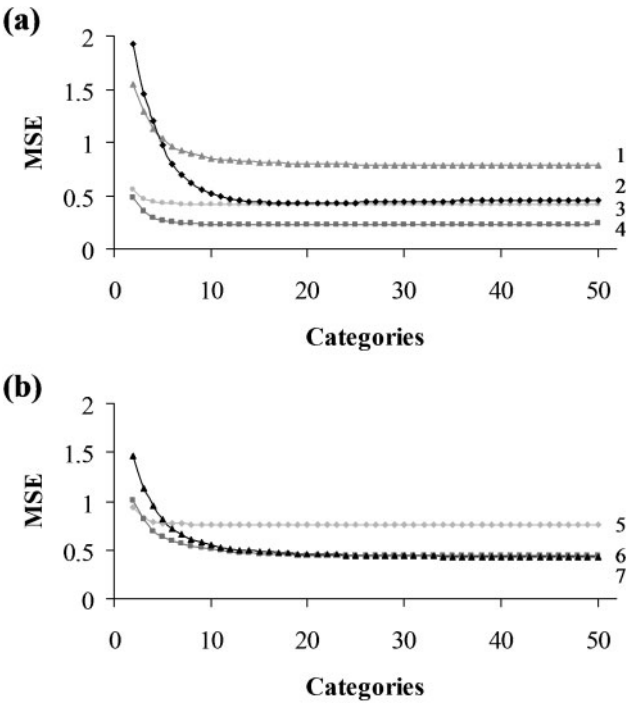

FIG. 3.—Accuracy of EB-EXP estimations as a function of number of categories. Accuracy was measured by using MSE. The data set number (table 1) is listed on the right-hand side of each curve. (*a*) Results obtained with a fixed phylogenetic tree. (*b*) Results obtained with trees and distributions inferred from real data sets.

Accuracy as a Function of Rate Variation

To study the effect of different levels of rate variation, the simulated rates were drawn from a 24-category discrete Gamma distribution with a specified α parameter. Fifteen different values of α were checked, ranging from 0.1 to 1.5 at equal intervals. This range appears to cover most of the values estimated from real data sets (Sullivan, Holsinger, and Simon 1996; Yang 1996). Three different sets of branch lengths were tested: $d = 0.1$, 0.25, and 0.5. In each case, trees with either six or 18 sequences were examined. The α used in the prior for EB-EXP was estimated from the simulated sequences.

Program Availability

The ML and EB-EXP rate-inference methods were implemented in computer programs written in C++ and are available at http://www.tau.ac.il/~talp/rate4site.html. A server for automatic inference of conserved regions in proteins and for projecting them onto the three dimensional structure is available through the ConSurf server (http://consurf.tau.ac.il/).

Results

Choosing the Number of Discrete Gamma Categories

When the number of categories was increased, the accuracy of inference increased until a plateau was reached (fig. 3). In all seven data sets, increasing the number of categories above 16 appeared to contribute little additional accuracy (fig. 3). We chose 16 categories for EB-EXP in all further analyses.

Accuracy as a Function of the Number of Sequences

Trees with six, 12, 18, 24, and 30 sequences were examined. Figure 4*a* shows the simulation results when *d* was fixed at 0.1. The accuracy of the estimates increased as the number of sequences increased. This is expected since more data are available at each position for rate inference. Noticeably, the Bayesian estimate is highly accurate for even a small number of sequences. The MSE between the inferred rates and their simulated values was

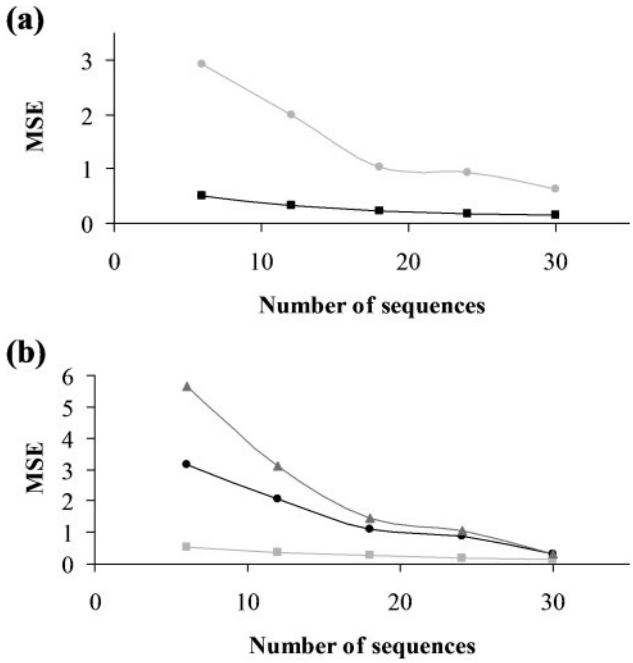

FIG. 4.—Accuracy of predictions as a function of the number of sequences. Accuracy was measured by using MSE. In all trees $d = 0.1$. (*a*) When the branch lengths are known prior to rate inference. (*b*) When the branch lengths are unknown. In this case the branch lengths of the model tree were optimized before the tree was given as input to the EB-EXP and ML inference methods. The results obtained using EB-EXP, ML, and ML-RICH are marked with squares, circles, and triangles, respectively.

below 0.52 for as few as six sequences. In contrast, a comparable level of accuracy is not achievable with ML even when the number of sequences is as large as 30. The considerable increase in accuracy for EB-EXP compared to ML is especially evident when data are scarce. EB-EXP superiority was reinforced when an extremely divergent tree was used to simulate the rate (i.e., $d = 1.0$). Though the prediction power decreased for both methods, the quality of the ML estimates dropped substantially, with MSE in the range of 2.14 (30 sequences) to 2.29 (six sequences), whereas the MSE for EB-EXP ranged between 0.58 (30 sequences) and 0.78 (six sequences).

Accuracy as a Function of Sequence Divergence

We investigated the accuracy of site-specific rate inference as a function of the degree of sequence divergence. For this purpose, we tested different model trees with branch lengths ranging from $d = 0.0125$ to $d = 1.0$. The result for a tree with six sequences is shown in figure 5. Again, EB-EXP appears superior to ML. For $d = 0.0125$, very low accuracies were observed for both methods. This can be explained by the insufficient evolutionary time needed to observe sufficient differences in the number of amino acid replacements among different positions. In this case, even positions with high rates of evolution are likely to exhibit no more than a handful of replacements. Thus, there is insufficient signal to infer accurate rates. For highly diverged sequences, the number of observed amino acid replacements may be saturated, which results in difficulties to distinguish between slow and fast evolving sites. Thus, for EB-EXP, a decreased accuracy was observed for high values of d (fig. 5). A similar behavior was reported by Yang and Wang (1995) on a tree with four sequences. For ML the accuracy as a function of d showed a peculiar pattern (fig. 5). This peculiarity was caused by the tendency of ML to infer extreme rates as well as by the dependency of MSE on the maximum rate chosen for ML (r_{max}) and the scaling procedure.

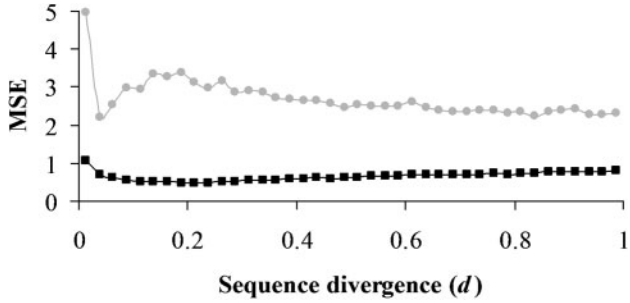

FIG. 5.—Accuracy of predictions as a function of the branch lengths d. The model tree used was as in figure 2*a*. The results obtained using EB-EXP and ML are marked with squares and circles, respectively.

Accuracy as a Function of α

Here we examined the effect of the amount of rate variation on the accuracy. For this purpose, different values of α were simulated. When a tree with six sequences was considered, the accuracy increased with α (fig. 6*a*). This is true for both methods. However, the increased accuracy with α is much more noticeable for

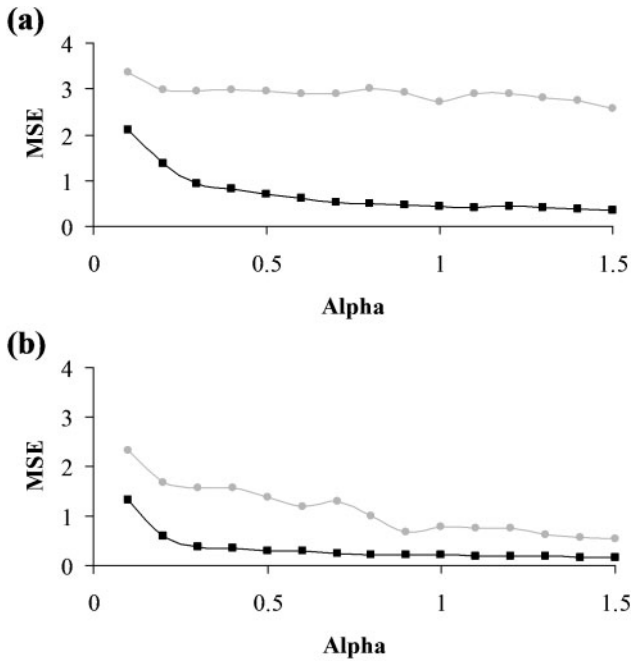

FIG. 6.—Accuracy of predictions as functions of α. The two inference methods are labeled as in figure 5. (*a*) The results obtained with a model tree as in figure 2*a* and $d = 0.1$. (*b*) The results obtained with a model tree as in figure 2*b* and $d = 0.1$.

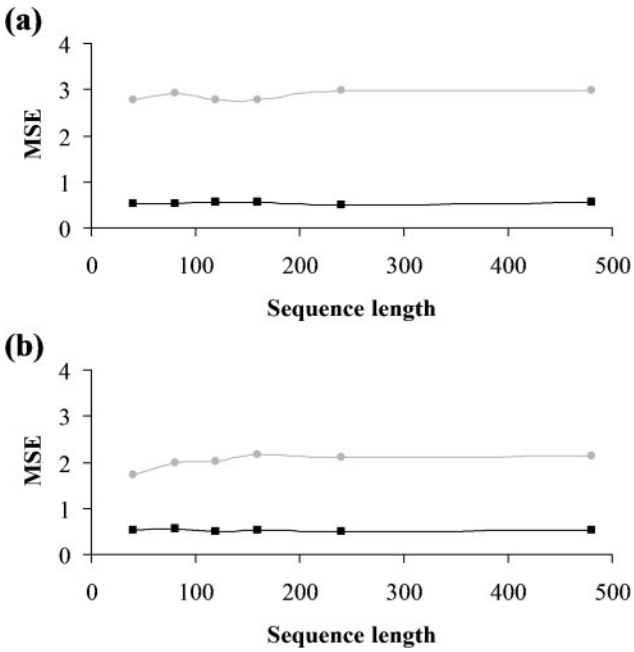

FIG. 7.—Accuracy of predictions as a function of sequence length obtained with a model tree as in figure 2*a* and $d = 0.1$. The two inference methods are labeled as in figure 5. (*a*) When the branch lengths of the model tree were given to the two inference methods. (*b*) When the branch lengths of the model tree were optimized before the tree was given as input to the inference methods.

EB-EXP compared to ML. The difference in accuracy between the two methods was less noticeable when trees with 18 sequences were used (fig. 6*b*).

Accuracy as a Function of Sequence Length

Sequence length might influence the quality of the inferred rates because it influences the accuracy of branch lengths and α parameter estimates. We first studied the effect of sequence length on the rate estimation when only the α parameter was estimated. Thus, the branch lengths were assumed to be known. As sites are considered independently, this need only influence EB-EXP; a ML does not require the estimation of α. As seen from figure 7*a*, the influence of the sequence length is not noticeable, as the accuracy is quite constant across different sequence lengths.

A second series of simulations were carried out to check the influence of sequence length on the rate inference when the branch lengths of the model tree were not given a priori but rather were optimized using ML (Felsenstein 1981) before estimating the site-specific rates. The results in figure 7*b* show that the effect of sequence length on accuracy is negligible in this case also.

Rate Estimation When the Branch Lengths Are Unknown a Priori

When the branch lengths are unknown two alternatives exists. Either the branch lengths are first estimated and then site-specific rates are inferred (using either ML or EB-EXP) or the branch lengths and site-specific rates are estimated simultaneously using the ML-RICH method (see *Materials and Methods*). We tested the accuracy of site-specific rate prediction using these three alternatives (ML, EB-EXP, and ML-RICH). Figure 4*b* presents our results for trees with different number of sequences. As expected, EB-EXP was superior to both ML methods, with ML better than ML-RICH in all cases studied. However, the differences between the methods diminished as the number of sequences increased, with the three methods reaching almost the same level of accuracy for 30 sequences.

Case Study

Will the differences between the various inference methods be noticeable when analyzing real data sets? To address this question we examined the evolutionary conservation pattern of the Bcl-2 protein family. This protein family plays a central role in the regulation of apoptotic cell death (Adams and Cory 1998). The family is divided into two subfamilies: anti-apoptotic and pro-apoptotic. All family members possess at least one of four conserved sequence motifs, known as Bcl-2 homology (BH) domains (BH1-BH4). Here we focus on the Bcl-x_L protein, for which the structure is known. Bcl-x_L contains all four BH domains, whereas distantly related proteins that promote apoptosis posses only BH3. The BH1, BH2, and BH3 domains strongly influence homo- and hetero-dimerization of Bcl-x_L. BH4 has been shown to be essential for Bcl-x_L to prevent apoptotic mitochondrial changes (Shimizu et al. 2000).

Homologous sequences were obtained from the SwissProt database (www.expasy.org/sprot/). Since only five orthologous sequences were obtained, we supplemented the alignment with 26 paralogous sequences. An MSA of these homologs was built using ClustalW (Thompson, Higgins, and Gibson 1994). We call this data set BCL-BIG. A smaller MSA consisting of the 14 closest homologs of Bcl-x_L was also constructed. This set only includes representatives from the anti-apoptotic family. We call this data set BCL-SMALL. For both data sets, an NJ tree was inferred using pairwise distances estimated by ML. Branch lengths in the resulting tree were then optimized using ML. The trees and the MSAs were given as input to the EB-EXP and the ML rate-inference methods. The inferred rates were then projected onto the three dimensional structure of a complex between Bcl-x_L and a Bak BH3 fragment (PDB ID: 1bxl; Sattler et al. 1997). In this step, the continuous evolutionary rates were partitioned into a discrete scale of 9 bins. The range of each bin varied such that each one contained 1/9 of the positions. Bin 9 contained the most conserved positions and bin 1 contained the most variable positions.

The conservation pattern obtained by both EB-EXP and ML using the BCL-BIG set of homologs yielded two main surface patches of conserved residues (fig. 8*a* and *b*). The first patch corresponds to a hydrophobic groove, formed by residues from the BH1, BH2, and BH3 regions. This patch is the binding site for the Bak peptide. The conservation pattern obtained by EB-EXP is slightly more pronounced than the patch obtained by ML. The second conserved patch corresponds to the BH4 domain. Empirical

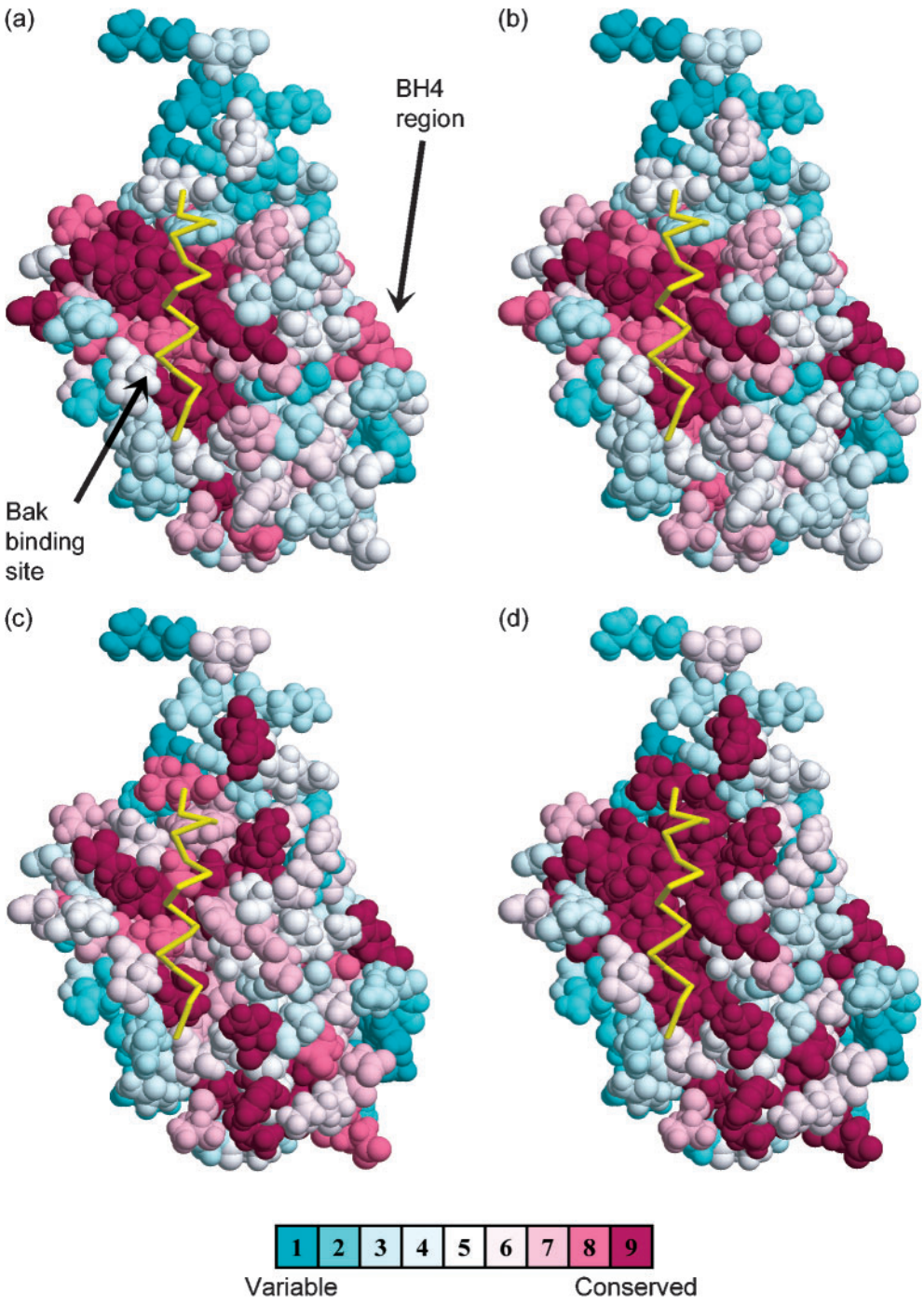

FIG. 8.—The conservation pattern obtained for the Bcl-x_L/Bak complex (PDB ID:1bxl) using (*a*) EB-EXP with BCL-BIG, (*b*) ML with BCL-BIG, (*c*) EB-EXP with BCL-SMALL, and (*d*) ML with BCL-SMALL. The Bcl-x_L protein is presented as a spacefill model. Conservation scores are color-coded onto the van der Waals surface of the protein. The Bak peptide is shown as a yellow backbone model. The color-coding bar shows the coloring scheme: burgundy corresponds to maximal conservation, white corresponds to average conservation, and turquoise to maximal variability.

evidence suggests that BH4 prevents apoptosis (Huang, Adams, and Cory 1998). However, this region is missing in more distant family members that also promote cell survival. In addition, no single residue in BH4 appeared to be essential for its function (Huang, Adams, and Cory 1998). EB-EXP graded the whole BH4 region as less conserved compared to ML.

A second analysis was carried out using only BCL-SMALL. This data set represents a difficult case for rate-inference since limited data are available. As the sequences used are quite similar to one another, many positions are uninformative (i.e., all the proteins exhibit the same amino acid at homologous positions). With EB-EXP the main conserved patch is still noticeable. However, the patch is

a bit scattered and extends beyond the interface boundaries (fig. 8*c*). Interestingly, the conservation pattern obtained with ML is expanded throughout the protein (fig. 8*d*). This is due to the fact that 51 out of 168 residues received the highest conservation score. Hence, much more than 1/9 of the positions are grouped in the most conserved bin. In this case ML cannot differentiate well between positions that are conserved because of their functionality and those that are conserved because of insufficient evolutionary signal.

Discussion

In this study we used simulations to compare the empirical Bayesian and ML site-specific rate-inference techniques. We also studied the effect of various parameters on the accuracy of each method.

One basic assumption in this study was that the rate at each site is constant during evolution. However, one might also try to find sites that are conserved in one part of the tree but are variable in the other. Such rate shifts may indicate change in the selection intensity at specific sites during evolution (reviewed in Gaucher et al. 2002). Rate shifts can also be inferred using an empirical Bayesian approach (Susko et al. 2002; Blouin, Boucher, and Roger 2003) or by using ML (Knudsen and Miyamoto 2001; Pupko and Galtier 2002). In our simulations we assumed that the tree topology is known a priori. In cases where this is not the case, one might use the Markov chain Monte Carlo technique to take the uncertainty of the tree topology into account (Huelsenbeck et al. 2001). Bayesian methods in phylogeny were recently criticized by Suzuki, Glazko, and Nei (2002) in the context of overestimation of Bayesian support for internal nodes. In our case, however, we limited the Bayesian part to a Gamma prior over the evolutionary rates, which is not the case with Bayesian methods that aim at inferring phylogenies.

When using a discrete approximation to the Gamma distribution, as in EB-EXP, the number of discrete categories must be specified. Yang (1994, 1995) suggested that four rate categories are sufficient to provide an optimum or near-optimum fit by the model to the data and to provide a good approximation to the continuous Gamma distribution. Our results showed that four categories are insufficient. For example, when four rate categories are used, 12.5% at each end of the distribution is not taken into account, i.e., 25% of the area below the rate distribution curve is ignored. Consequently, very high or very low substitution rates cannot be observed. This is unfortunate, since these are exactly the rates we seek to identify when predicting functionally important sites. We note, however, that Yang's (1994) emphasis was either phylogenetic tree reconstruction or estimating the shape of the Gamma distribution, which may not change dramatically with the number of categories. In contrast, here we were interested in the rates themselves. The discrete Gamma method with eight categories was recently used by Susko et al. (2002) to infer rate shifts in different subtrees and by Excoffier and Yang (1999), Meyer, Weiss, and von Haeseler (1999), and Misof et al. (2002) to infer substitution rates per site. In light of our findings, choosing 16 categories instead of eight may improve the results.

The simulation results showed that EB-EXP performs better than ML. Since both methods use the same likelihood function in their computations, the differences between EB-EXP and ML must be due to the incorporation of the prior distribution, which reduces the posterior probability of extreme unfavorable observed rates in EB-EXP. It can be claimed that the superiority of the Bayesian approach depends on how well the prior function fits the data. An empirical Bayesian approach is used here, in which the parameter of the prior Gamma function is inferred from the data. This gives more flexibility for the prior to fit the data.

There is another difference between the EB-EXP and ML methods. In EB-EXP, the inferred rate is the expectation over the posterior rate distribution (Yang and Wang 1995; Excoffier and Yang 1999; Susko et al. 2002), whereas the ML estimate is the rate that maximizes the likelihood function. A second Bayesian method, EB-MAP, is possible, in which the rate yielding the maximum a posteriori probability is chosen (i.e., $r_{map} = \text{argmax}_r P(r \mid data, T)$). One advantage of EB-MAP over EB-EXP is that there is no need to use a discrete approximation to the continuous Gamma distribution. This can be done by a maximization procedure directly on the continuous posterior distribution. However, taking the expectation of the continuous Gamma distribution is known to be asymptotically more accurate than EB-MAP when the accuracy is measured by a sum-of-square error function. Thus, the advantage of using EB-MAP is that there is no need to approximate the Gamma distribution, while the advantage of EB-EXP is that without the approximation it should be asymptotically more accurate. Simulation results obtained with EB-MAP were very similar, though slightly inferior, to those obtained with EB-EXP (data not shown). We chose EB-EXP for this study because in this method it is easier to obtain not only a point estimate but also its credibility interval (Susko et al. 2002). We note that a common way to infer site-specific rates (e.g., Meyer, Weiss, and von Haeseler 1999) is to choose the discrete rate category that contributes the most to the posterior distribution. This is not a real "Map" estimate: because the prior probability of each category is identical, this would in fact be a discrete version of the ML approach.

We note that in our simulation the accuracy of inference is overestimated, since we rarely know the true tree as was set up in the simulation. In addition, the substitution model used for the simulation is the same as the one used for inference, which is most certainly not the case for real data sets. Nonetheless, this discrepancy is the same for all inference methods, so our conclusions regarding the relative efficiency of the two inference methods should still hold. This uncertainty in the estimation of tree topology, branch lengths, and evolutionary model also results in underestimated credibility intervals obtained for EB-EXP.

We demonstrated that regardless of the inference method employed, accuracy of prediction depends strongly on the amount of data, i.e., the number of sequences in the MSA. We further showed that the degree of similarity in these sequences, represented by branch lengths in the phylogenetic tree, also affects results. A decrease in

prediction success was observed when the branch lengths were extremely short. In these cases the number of amino acid replacements was too small to allow reliable rate inferences. For EB-EXP, when branch lengths are very large, multiple replacements at a site might obscure the history of a character, resulting in a reduced accuracy. As ML tends to adopt extreme rates and MSE scores are highly sensitive to extreme rate values, a peculiar behavior for highly diverged sequences was observed in ML.

The shape of the rate distribution influences rate inference accuracy. Meyer and von Haeseler (2003) recently presented an ML variant that identifies site-specific substitution rates. In their simulation study that included different model trees, a decrease in accuracy was observed with increasing α values, which is in disagreement with our results. The discrepancy can be explained by the different approaches used to infer accuracy. While MSE was used in our study, Meyer and von Haeseler (2003) used the correlation coefficient between the inferred and simulated rates. To illustrate why these two criteria for accuracy may yield different results, consider two sites evolving at relative rates of 1.02 and 0.98, respectively. If the inferred rates are 1.0 and 1.01, respectively, the inferred rates are very close to the true values but they are in the wrong order. MSE measures the deviation of the inferred rate from its true value for each site independently from the other sites. The correlation coefficient, however, measures to what extent the inferred and simulated rates vary together. Thus, when the rates are nearly homogenous (i.e., high α values), rates with extreme values are rare and the inference is more accurate (low MSE). Correlation coefficients, however, are expected to be relatively low.

Another shortcoming of the ML method is that its point estimates tend to adopt extreme values when the amount of data drops below a critical threshold (Lewis 2001). Thus, when the data are scarce, as was the case when rates were inferred from less than 12 sequences, ML resulted in very rough estimates (MSE = 2.92 and 2.0 for six and 12 sequences, respectively, compared with 0.51 and 0.32, respectively, for EB-EXP). Figure 9*a* and *b* show scatter plots of inferred rates obtained using the ML and EB-EXP methods versus the simulated values. Whereas several extreme values were observed using the ML method (fig. 9*a*), the inferred rates of the EB-EXP method were clustered close to the $y = x$ line (fig. 9*b*).

When a large amount of sequences is available, one could be tempted to use the ML-RICH model, assuming a specific rate for each site. This model can be used to estimate both the phylogenetic tree and the site-specific rates simultaneously. Our results showed that the huge increase in the number of free parameters in the ML-RICH method results in decreased accuracy of site-specific rate estimates compared to ML. However, the difference in accuracy diminished as the number of sequences was increased, reaching almost the same accuracy for 30 sequences. It is expected that as the number of sequences increases, using the ML-RICH model would be more acceptable because more data would be available at each position. We note that Meyer and von Haeseler (2003) suggested a variant of the ML-RICH model for any number of sequences. Clearly, the reduced accuracy of rate

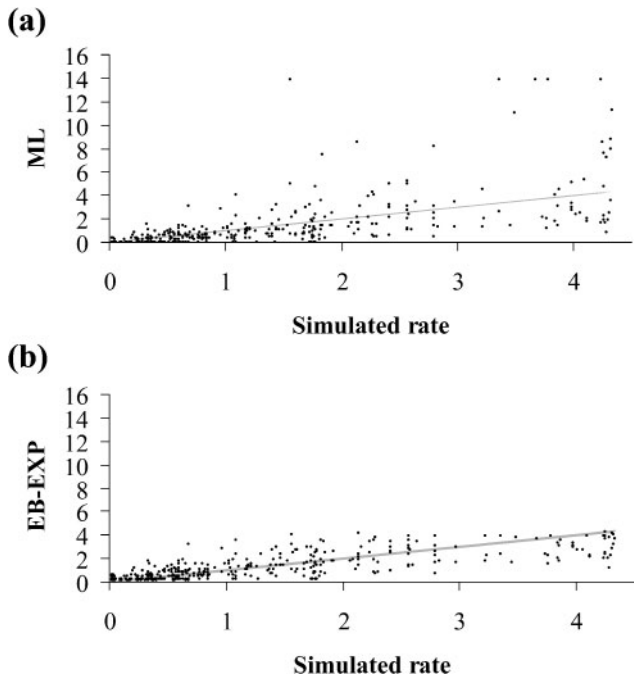

FIG. 9.—Scatter plots of 500 inferred rates versus their simulated values with a model tree with six sequences and $d = 0.1$ for (*a*) ML and (*b*) EB-EXP. The grey line marks the $y = x$ line.

inference show that this is not justified in the general case and could lead to a reduced accuracy of site-specific rate estimates.

One of the main difficulties in calculating site-specific conservation scores is to distinguish between amino acid sites that are conserved due to their functionality and those that appear to be conserved due to insufficient time since divergence. EB-EXP appears to differentiate better between these two cases. ML calculates only the most probable rate, which may be misleading when little data are available. Looking at the Bcl-x_L example, the arginines in positions 6 and 139 provide an illustration. Both positions are fully conserved, yet while Arg139 is present in all 32 homologs, Arg6 appears in only 11 of them. ML could not discriminate between these two positions and assigned both the highest conservation score. In contrast, EB-EXP rated Arg139 as the most conserved position. Indeed, mutating Arg139 to glutamine in Bcl-x_L has been shown to inhibit its anti-apoptotic function (Sattler et al. 1997). Arg6, on the other hand, was only the 29th conserved position (out of 169) when graded by EB-EXP, as it was missing in 21 homologs. This result is congruent with experiments: mutating Arg6 to alanine, in Bcl-x_L's close homolog Bcl-2, did not diminish the protein activity (Huang, Adams, and Cory 1998).

The distinction between the Bayesian and ML analyses was reinforced when using limited data, as was the case with BCL-SMALL. Whereas the conservation pattern using EB-EXP was a bit more scattered than in the complete analysis (fig. 8*c* as compared to 8*a*), ML graded a vast number of positions as extremely conserved (fig. 8*d*). As a consequence, the conserved patch expands far beyond the Bak binding groove.

A robust evolutionary analysis can provide means for the identification of patches of conserved residues on the protein surface, which are essential for the protein's function. The bottleneck for the *in silico* identification of these functional patches appears to be the availability of sequence data (Bell and Ben-Tal 2003). Too little variation in the MSA caused by too few sequences or too little diversity among them can render evolutionary analysis meaningless (Thornton et al. 2000). Ten available homologous proteins appear to be the sensitivity threshold when using ML (Bell and Ben-Tal 2003). Our study implies that these are exactly the conditions where EB-EXP is distinctly better than ML.

Acknowledgments

We thank Karen B. Avraham for introducing us to genes involved in hearing loss and Yossi Rosenberg for his help incorporating EB-EXP into the ConSurf server. N.B.-T. was supported by a Research Career Development Award from the Israel Cancer Research Fund. T.P. was supported by a grant in Complexity Science from the Yeshaia Horvitz Association. We thank three anonymous referees for insightful comments and suggestions.

Literature Cited

Adams, J. M., and S. Cory. 1998. The Bcl-2 protein family: arbiters of cell survival. Science **281**:1322–1326.

Bell, R. E., and N. Ben-Tal. 2003. *In silico* identification of functional protein interfaces. Comp. Funct. Genom. **4**:420–423.

Blouin, C., Y. Boucher, and A. J. Roger. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res. **31**: 790–797.

del Sol Mesa, A., F. Pazos, and A. Valencia. 2003. Automatic methods for predicting functionally important residues. J. Mol. Biol. **326**:1289–1302.

Donaudy, F., A. Ferrara, L. Esposito, R. Hertzano, O. Ben-David, R. E. Bell, S. Melchionda, L. Zelante, K. B. Avraham, and P. Gasparini. 2003. Multiple mutations of *MYO1A*, a cochlear-expressed gene, in sensorineural hearing loss. Am. J. Hum. Genet. **72**:1571–1577.

Excoffier, L., and Z. Yang. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Mol. Biol. Evol. **16**:1357–1368.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. **53**: 447–455.

Gaucher, E. A., X. Gu, M. M. Miyamoto, and S. A. Benner. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. **27**:315–321.

Glaser, F., T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. 2003. ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics **19**:163–164.

Huang, D. C., J. M. Adams, and S. Cory. 1998. The conserved N-terminal BH4 domain of Bcl-2 homologues is essential for inhibition of apoptosis and interaction with CED-4. Embo. J. **17**:1029–1039.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310–2314.

Jin, L., and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102

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

Kimura, M. 1983. The neutral theory of molecular evolution. Pp. 208–233 *in* M. Nei and R. Koehn, eds. *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland, Mass.

Knudsen, B., and M. M. Miyamoto. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA **98**:14512–14517.

Leonard, T., and J. S. J. Hsu. 1999. Bayesian methods: an analysis for statisticians and interdisciplinary researchers. Cambridge University Press, Cambridge.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Syst. Biol. **50**:913–925.

Lichtarge, O., and M. E. Sowa. 2002. Evolutionary predictions of binding surfaces and interactions. Curr. Opin. Struct. Biol. **12**:21–27.

Mella, M., G. Colotti, C. Zamparelli, D. Verzili, A. Ilari, and E. Chiancone. 2003. Information transfer in the penta-EF-hand protein sorcin does not operate via the canonical structural/functional pairing. A study with site-specific mutants. J. Biol. Chem. **278**:24921–24928.

Meyer, S., and A. von Haeseler. 2003. Identifying site-specific substitution rates. Mol. Biol. Evol. **20**:182–189.

Meyer, S., G. Weiss, and A. von Haeseler. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152**: 1103–1110.

Misof, B., C. L. Anderson, T. R. Buckley, D. Erpenbeck, A. Rickert, and K. Misof. 2002. An empirical analysis of mt 16S rRNA covarion-like evolution in insects: Site-specific rate variation is clustered and frequently detected. J. Mol. Evol. **55**:460–469.

Nielsen R. 1997. Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. Syst. Biol. **46**:346–353.

Pupko, T., R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics **18**:S71–S77.

Pupko, T., and N. Galtier. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. R. Soc. Lond. B Biol. Sci. **269**:1313–1316.

Ramelot, T. A., S. Ni, S. Goldsmith-Fischman, J. R. Cort, B. Honig, and M. A. Kennedy. 2003. Solution structure of Vibrio cholerae protein VC0424: A variation of the ferredoxin-like fold. Protein Sci. **12**:1556–1561.

RamShankar, M., S. Girirajan, O. Dagan, H. M. Ravi Shankar, R. Jalvi, R. Rangasayee, K. B. Avraham, and A. Anand. 2003. Contribution of connexin26 (*GJB2*) mutations and founder effect to non-syndromic hearing loss in India. J. Med. Genet. **40**:E68.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sattler, M., H. Liang, D. Nettesheim, et al. (12 co-authors). 1997. Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis. Science **275**:983–986.

Shimizu, S., A. Konishi, T. Kodama, and Y. Tsujimoto. 2000. BH4 domain of antiapoptotic Bcl-2 family members closes voltage-dependent anion channel and inhibits apoptotic mito-

chondrial changes and cell death. Proc. Natl. Acad. Sci. USA **97**:3100–3105.

Sullivan, J., K. E. Holsinger, and C. Simon. 1996. The effect of topology on estimates of among-site rate variation. J. Mol. Evol. **42**:308–312.

Susko, E., Y. Inagaki, C. Field, M. E. Holder, and A. J. Roger. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. Mol. Biol. Evol. **19**:1514–1523.

Suzuki, Y, G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA **99**:16138–16143.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp. 407–514 *in* C. M. D. M. Hillis and B. K. Mable, eds. Molecular systematics. Sinauer Associates, Sunderland, Mass.

Thompson, J. D., D. G. Higgins. and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Thornton, J. M., A. E. Todd, D. Milburn, N. Borkakoti, and C. A. Orengo. 2000. From structure to function: approaches and limitations. Nat. Struct. Biol. **7**:991–994.

Valdar, W. S. 2002. Scoring residue conservation. Proteins **48**: 227–241.

Whelan, S., P. Lio, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. **17**:262–272.

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

———. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1995. A space-time process model for the evolution of DNA sequences. Genetics **139**:993–1005.

———. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–372.

Yang, Z., and T. Wang. 1995. Mixed model analysis of DNA sequence evolution. Biometrics **51**:552–561.

Yao, H., D. M. Kristensen, I. Mihalek, M. E. Sowa, C. Shaw, M. Kimmel, L. Kawraki, and O. Lichtarge. 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J. Mol. Biol. **326**:255–261.

Pekka Pamilo, Associate Editor

Accepted June 3, 2004