Estimation of Amino Acid Residue Substitution Rates at Local Spatial Regions and Application in Protein Function Inference: A Bayesian Monte Carlo Approach

Yan Y. Tseng and Jie Liang

Department of Bioengineering, Science and Engineering Offices, MC-063, University of Illinois at Chicago

The amino acid sequences of proteins provide rich information for inferring distant phylogenetic relationships and for predicting protein functions. Estimating the rate matrix of residue substitutions from amino acid sequences is also important because the rate matrix can be used to develop scoring matrices for sequence alignment. Here we use a continuous time Markov process to model the substitution rates of residues and develop a Bayesian Markov chain Monte Carlo method for rate estimation. We validate our method using simulated artificial protein sequences. Because different local regions such as binding surfaces and the protein interior core experience different selection pressures due to functional or stability constraints, we use our method to estimate the substitution rates of local regions. Our results show that the substitution rates are very different for residues in the buried core and residues on the solvent-exposed surfaces. In addition, the rest of the proteins on the binding surfaces also have very different substitution rates from residues. Based on these findings, we further develop a method for protein function prediction by surface matching using scoring matrices derived from estimated substitution rates for residues located on the binding surfaces. We show with examples that our method is effective in identifying functionally related proteins that have overall low sequence identity, a task known to be very challenging.

Introduction

Amino acid sequences are an important source of information for inferring distant phylogenetic relationships and for predicting the biochemical functions of protein. Because the substitutions of nucleotides can become rapidly saturated, and the likelihood of unrelated identical substitutions is high for nucleotides, the information of evolutionary conservation of nucleotides is quickly obscured after a number of generations. The mapping of DNA sequences by the genetic code to amino acid sequences frequently can reveal more remote evolutionary relation with more interpretable sequence similarity (Liò and Goldman 1999). In addition, statistical analysis of protein sequence alignment is also more reliable as it is much more difficult to detect and correct for deviations from independent identical distributions in DNA sequences due to possible translation of normal complexity DNA sequences into low complexity protein sequences such as tandem repeats of simple patterns of a few residues (Pearson 1998).

The success in detecting evolutionarily related protein sequences through sequence alignment depends on the use of a scoring matrix, which determines the similarity between residues. Rate matrices of amino acid residue substitutions can be the basis for developing many scoring matrices for sequence alignment. Dayhoff, Schwartz, and Orcutt (1978) were the first to develop empirical models of amino acid residue substitutions. They used a counting method to obtain accepted point mutation matrices (called Pam matrices). The widely used Blosum matrices can be viewed as analogous to transition matrices of residues at different time intervals (S. Henikoff and J. G. Henikoff 1992; Liò and Goldman 1998). They were developed following a heuristic counting approach similar to that of Pam and were derived from structure-based alignments of blocks of sequences of related proteins (S. Henikoff and J. G. Henikoff 1992). Both Pam and Blosum matrices are widely used for sequence alignment (e.g., in software tools such as Fasta, Blast, and ClustalW) (Altschul et al. 1990; Pearson 1990; Thompson, Higgins, and Gibson 1994). An update of the Pam matrices based on the same counting approach using a much enlarged database is the Jones-Taylor-Thornton (JTT) amino acid substitution matrix, which is widely used for phylogenetic analysis (Jones, Taylor, and Thornton 1992; Adachi and Hasegawa 1996; Yang 1997).

Whelan and Goldman pointed out that these counting methods are effectively equivalent to the maximum parsimony method, and therefore suffer from several drawbacks: the systematic underestimation of substitutions in certain branches of a phylogeny and the inefficiency in using all information contained in the amino acid residue sequences (Whelan and Goldman 2001). This can be a serious problem for applications such as inferring protein functions from a protein sequence, as the number of sequence homologs available for multiple sequence alignment is often limited. In addition, matrices such as Pam and Blosum have implicit parameters whose values were determined from the precomputed analysis of large quantities of sequences, while the information of the particular protein of interest has limited or no influence. A more effective approach for studying amino acid residue substitutions is to employ an explicit continuous time Markov model based on a phylogenetic tree of the protein (Yang, Nielsen, and Hasegewa 1998; Whelan and Goldman 2001). Markovian evolutionary models are parametric models and do not have prespecified parameter values. These values are estimated from data specific to the protein of interest (Whelan, Liò, and Goldman 2001). Recent work using this approach has shown that more informative rate matrices can be derived, with significant advantages over matrices obtained from counting method (Whelan and Goldman 2001).

Despite these important results, current studies of the substitution rates of amino acid residues are based on the assumption that the whole protein sequence experience similar selection pressure and therefore have the same substitution rates. There is no distinction for different regions of

Key words: continuous time Markov process, Bayesian Markov chain Monte Carlo, amino acid substitution matrix, protein function prediction.

E-mail: jliang@uic.edu.

Mol. Biol. Evol. 23(2):421–436. 2006
doi:10.1093/molbev/msj048
Advance Access publication October 26, 2005

proteins, namely, all sites have the same evolutionary rates. This is an unrealistic assumption. For example, regions that directly participate in biochemical functions, such as binding surfaces, are likely to experience very different selection pressure from other regions. In the protein interior, hydrophobic amino acid residues may be conserved not due to their functional roles, but due to the constraints of maintaining protein stability, as hydrophobic interactions are the driving force of protein folding (Dill 1990; Govindarajan and Goldstein 1997; Parisi and Echave 2001; Li and Liang 2005). Similarly, residues in the transmembrane segments of membrane proteins experience different selection pressure from soluble parts of the proteins (Liò and Goldman 1999; Tourasse and Li 2000). It is therefore important to study region-specific residue replacement rates.

An important advance in the reconstruction of phylogeny is the consideration of heterogenous substitution rates among different sites (Yang et al. 2000; Mayrose et al. 2004). However, these are based on substitution models of either nucleotides or codons, with sometimes discretized categories of rates. Because of the large number of parameters due to an alphabet size of 20 for amino acid residues, it is impractical to estimate site-specific rates for amino acid residue sequences.

In this study, we use a continuous time Markov model to estimate residue substitution rates for spatially defined regions of proteins based on known three-dimensional structures of proteins (Liang, Edelsbrunner, and Woodward 1998; Binkowski, Adamian, and Liang 2003). Different from previous studies of rate estimation based on maximum likelihood methods (Felsenstein and Churchill 1996; Yang, Nielsen, and Hasegawa 1998; Whelan and Goldman 2001; Siepel and Haussler 2004), we develop a Bayesian method to estimate the posterior mean values of the instantaneous rates of residue substitution. Our approach is based on the technique of Markov chain Monte Carlo, a method that has been widely used in phylogenetic analysis (Yang and Rannala 1997; Mau, Newton, and Larget 1999; Huelsenbeck, Rannala, and Larget 2000). To derive well-defined spatial regions of proteins which are formed by residues well separated in primary sequences, we rely on computational analysis of protein structures (Liang, Edelsbrunner, and Woodward 1998). In our study, these distant residues in sequences are spatial neighbors that participate in direct molecular binding events and can be regarded as belonging to the same class of substitution rates. Our study is also motivated by the need to deduce related functions from protein structures, that is, to identify functionally related protein structures. As structural biology proceeds, there is an increasing number of proteins whose atomic structures are resolved, yet their biological functions are completely unknown (Sanishvili et al. 2003).

Our results show that residue substitution rates are significantly different for different regions of the proteins, for example, for the buried protein core, solvent-exposed surfaces, and specific binding surfaces on protein structures. We also develop a novel method for inferring protein functions. Using residue similarity scoring matrices derived from estimated substitution rates for protein surfaces, our method is far more effective than several other methods in detecting similar binding surface that are functionally related from different protein structures. This is a challenging task, as it is well known that function prediction becomes difficult when the sequence identity between two proteins is below 60%–70% (Rost 2002; Tian and Skolnick 2003).

This paper is organized as follows. We first describe the continuous time Markov model for residue substitution rates. We then discuss how to compute the likelihood function of substitution rate matrices given a specific phylogeny and a multiple sequence alignment. The Markov chain Monte Carlo method is then briefly described, including the design of move sets that helps to improve the rate of mixing. We then describe simulation results in estimating substitution rates. This is followed by discussion of the results of different substitution rates estimated for different regions of a set of proteins. We then give examples to show how residue scoring matrices derived from the estimated rate matrix can improve detection of functionally related proteins.

Model and Methods

Continuous Time Markov Process for Residue Substitution

For a given phylogenetic tree, we use a reversible continuous time Markov process as our evolutionary model (Felsenstein 1981; Yang 1994*a*). This model has several advantages over empirical methods. For example, Markovian evolutionary models are parametric models and do not have prespecified parameter values. These values are all estimated from data specific to the protein of interest (Whelan, Liò, and Goldman 2001). In addition, previous works showed that the effects of secondary structure and solvent accessibility are important for protein evolution, and such effects can be captured by a Markovian evolutionary model, while it is difficult for empirical methods to take these effects into account (Goldman, Thorne, and Jones 1996, 1998; Liò and Goldman 1999; Robinson et al. 2003).

Once the tree topology and the time intervals of sequence divergence $\{t\}$ (or the branch lengths) of the phylogenetic tree are known, the parameters of the model are the 20×20 rate matrix $\boldsymbol{Q}$ for the 20 amino acid residues. Because substitution rate and divergence time t are confounded, t cannot be expressed in absolute units. We follow the approach of Adachi and Hasegawa (1996) to represent the divergence time t as the expected number of residue changes per 100 sites between the sequences. The entries q_{ij} of matrix $\boldsymbol{Q}$ are substitution rates of amino acid residues for the set $\mathcal{A}$ of 20 amino acid residues at an infinitesimally small time interval. Specifically, we have: $\boldsymbol{Q} = \{q_{ij}\}$, where the diagonal element is $q_{i,i} = -\sum_{i,j\neq i} q_{i,j}$. The transition probability matrix of size 20×20 after time t is (Liò and Goldman 1998):

$$\boldsymbol{P}(t) = \{p_{ij}(t)\} = \boldsymbol{P}(0)\exp(\boldsymbol{Q} \cdot t),$$

where $\boldsymbol{P}(0)=\boldsymbol{I}$. Here $p_{ij}(t)$ represents the probability that a residue of type i will mutate into a residue of type j after time t. To ensure that the nonsymmetric rate matrix $\boldsymbol{Q}$ is diagonalizable for easy computation of $\boldsymbol{P}(t)$, we follow Whelan and Goldman (2001) and insist that $\boldsymbol{Q}$ takes the form of $\boldsymbol{Q}=\boldsymbol{S} \cdot \boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix who entries are the composition of residues from the region of interest on the protein structure, and $\boldsymbol{S}$ is a symmetric matrix whose

entries need to be estimated. Because symmetric $\boldsymbol{S}$ is diagonalizable as $\boldsymbol{S}=\boldsymbol{V}\Lambda\boldsymbol{V}^T$, the matrix $\boldsymbol{Q}=\boldsymbol{S}\cdot\boldsymbol{D}=\boldsymbol{D}^{1/2}\boldsymbol{V}\Lambda\boldsymbol{V}^T\boldsymbol{D}^{-1/2}$ is also diagonalizable, hence $\boldsymbol{P}(t)=\boldsymbol{P}(0)(\boldsymbol{D}^{1/2}\boldsymbol{V})\exp(\Lambda t)(\boldsymbol{V}^T\boldsymbol{D}^{-1/2})$.

Likelihood Function of a Fixed Phylogeny

For node k and node l separated by divergence time t_{kl}, the time reversible probability of observing residue x_k in a position h at node k and residue x_l of the same position at node l is:

$$\pi_{x_k}p_{x_kx_l}(t_{kl})=\pi_{x_l}p_{x_lx_k}(t_{kl}).$$

For a set $\mathcal{S}$ of s multiple-aligned sequences $(x_1,x_2,\ldots,x_s)$ of length n amino acid residues in a specific region, we assume that a reasonably accurate phylogenetic tree $\boldsymbol{T}=(\mathcal{V},\mathcal{E})$ of the proteins is given. Here $\mathcal{V}$ is the set of nodes, namely, the union of the set of observed s sequences $\mathcal{L}$ (leaf nodes), and the set of $s-1$ ancestral sequences $\mathcal{I}$ (internal nodes). $\mathcal{E}$ is the set of edges of the tree. Let the vector $\boldsymbol{x}_h=(x_1,\ldots,x_s)^T$ be the observed residues at position h for the s sequences, h ranges from 1 to n. Without loss of generality, we assume that the root of the phylogenetic tree is an internal node k. Given the specified topology of the phylogenetic tree $\boldsymbol{T}$ and the set of edges, the probability of observing s number of residues $\boldsymbol{x}_h$ at position h is:

$$p(\boldsymbol{x}_h|\boldsymbol{T},\boldsymbol{Q})=\pi_{x_k}\sum_{\substack{i\in\mathcal{I}\\ x_i\in\mathcal{A}}}\prod_{(i,j)\in\mathcal{E}}p_{x_ix_j}(t_{ij}).$$

After summing over the set $\mathcal{A}$ of all possible residue types for the internal nodes $\mathcal{I}$. The probability $P(\mathcal{S}|\boldsymbol{T},\boldsymbol{Q})$ of observing all residues in the functional region is:

$$P(\mathcal{S}|\boldsymbol{T},\boldsymbol{Q})=P(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_s|\boldsymbol{T},\boldsymbol{Q})=\prod_{h=1}^{n}p(\boldsymbol{x}_h|\boldsymbol{T},\boldsymbol{Q}).$$

This can be used to calculate the log-likelihood function $l=\log P(S|\boldsymbol{T},\boldsymbol{Q})$.

Bayesian Estimation of Instantaneous Rates

Our goal is to estimate the values of the $\boldsymbol{Q}$ matrix. The continuous time Markov model for residue substitutions has been implemented in several studies using maximum likelihood estimation (Yang 1994*a*; Whelan and Goldman 2001) and has also been applied in a protein folding study (Tseng and Liang 2004). Different from these prior studies, here we adopt a Bayesian approach. We use a prior distribution $\pi(\boldsymbol{Q})$ to encode our past knowledge of amino acid substitution rates for proteins. We describe the instantaneous substitution rate $\boldsymbol{Q}=\{q_{ij}\}$ by a posterior distribution $\pi(\boldsymbol{Q}|\mathcal{S},\boldsymbol{T})$, which summarizes prior information available on the rates $\boldsymbol{Q}=\{q_{ij}\}$ and the information contained in the observations $\mathcal{S}$ and $\boldsymbol{T}$. After integrating the prior information and the likelihood function, the posterior distribution $\pi(\boldsymbol{Q}|\mathcal{S},\boldsymbol{T})$ can be estimated up to a constant as:

$$\pi(\boldsymbol{Q}|\mathcal{S},\boldsymbol{T})\propto\int P(\mathcal{S}|\boldsymbol{T},\boldsymbol{Q})\cdot\pi(\boldsymbol{Q})\,\mathrm{d}\boldsymbol{Q}.$$

Our goal is to estimate the posterior means of rates in $\boldsymbol{Q}$ as summarizing indice:

$$\mathbb{E}_\pi(\boldsymbol{Q})=\int\boldsymbol{Q}\cdot\pi(\boldsymbol{Q}|\mathcal{S},\boldsymbol{T})\,\mathrm{d}\boldsymbol{Q}.$$

In this study, we use uniform uninformative priors. Others choices are also possible.

Markov Chain Monte Carlo

We run a Markov chain to generate samples drawn from the target distribution $\pi(\boldsymbol{Q}|\mathcal{S},\boldsymbol{T})$. Starting from a rate matrix $\boldsymbol{Q}_t$ at time t, we generate a new rate matrix $\boldsymbol{Q}_{t+1}$ using the proposal function: $T(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})$. The proposed new matrix $\boldsymbol{Q}_{t+1}$ will be either accepted or rejected, depending on the outcome of an acceptance rule $r(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})$. Equivalently, we have:

$$\boldsymbol{Q}_{t+1}=A(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})=T(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})\cdot r(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1}).$$

To ensure that the Markov chain will reach stationary state, we need to satisfy the requirement of detailed balance, that is,

$$\pi(\boldsymbol{Q}_t|\mathcal{S},\boldsymbol{T})\cdot A(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})=\pi(\boldsymbol{Q}_{t+1}|\mathcal{S},\boldsymbol{T})\cdot A(\boldsymbol{Q}_{t+1},\boldsymbol{Q}_t).$$

This is achieved by using the Metropolis-Hastings acceptance ratio $r(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})$ to either accept or reject $\boldsymbol{Q}_{t+1}$, depending on whether the following inequality holds:

$$u\le r(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})=\min\left\{1,\frac{\pi(\boldsymbol{Q}_{t+1}|\mathcal{S},\boldsymbol{T})\cdot T(\boldsymbol{Q}_{t+1},\boldsymbol{Q}_t)}{\pi(\boldsymbol{Q}_t|\mathcal{S},\boldsymbol{T})\cdot T(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})}\right\},$$

where u is a random number drawn from the uniform distribution $\mathcal{U}[0,1]$. With the assumption that the underlying Markov process is ergodic, irreducible, and aperiodic (Grimmett and Stizaker 2001), a Markov chain generated following these rules will reach the stationary state (Robert and Casella 2004).

We collect m correlated samples of the $\boldsymbol{Q}$ matrix after the Markov chain has reached its stationary state. The posterior means of the rate matrix are then estimated as:

$$\mathbb{E}_\pi(\boldsymbol{Q})\approx\sum_{i=1}^{m}\boldsymbol{Q}_i\cdot\pi(\boldsymbol{Q}_i|\mathcal{S},\boldsymbol{T}).$$

Move Set

A move set determines the proposal function $T(\boldsymbol{Q}_t,\boldsymbol{Q}_{t+1})$, which is critical for the rapid convergency of a Markov chain. To improve mixing, we design two types of moves for proposing a new rate matrix $\boldsymbol{Q}_{t+1}$ from a previous matrix $\boldsymbol{Q}_t$. When the state variable s for these two types of moves takes the value $s=1$, we take Type 1 move. When the state $s=2$, we take Type 2 move. For Type 1 moves, a single entry of the rate matrix with index ij is randomly chosen, and with equal probability we assign:

$$q_{ij,t+1}=\alpha_1q_{ij,t}\quad\text{or}\quad q_{ij,t+1}=\alpha_2q_{ij,t},$$

where $\alpha_1=0.9$ and $\alpha_2=1.1$. For Type 2 moves, we use a simplified residue alphabet of size 5 to represent the 20

amino acid residue types, based on the analysis described by Li, Hu, and Liang (2003). The five residue types are: $\{G, A, V, L, I, P\}$, $\{F, Y, W\}$, $\{S, T, C, M, N, Q\}$, $\{D, E\}$, and $\{K, R, H\}$. We select one of the five reduced residue types following $\mathcal{U}[1,2,\ldots,5]$, and scale with equal probability all entries in $\boldsymbol{Q}$ corresponding to the residues contained in one of the simplified residue type, with a constant of either $\alpha_1 = 0.9$ or $\alpha_2 = 1.1$ at equal probability. The transition between these two types of moves is determined by the transition matrix:

$$\begin{pmatrix} s_{1,1} & s_{1,2} \\ s_{2,1} & s_{2,2} \end{pmatrix} = \begin{pmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \end{pmatrix}.$$

Overall, the acceptance ratio of Type 1 moves is 50%–66%, and the acceptance ratio of Type 2 move is $<10\%$.

Rate Matrix $\boldsymbol{Q}$ and Residue Similarity Score

To derive residue similarity scoring matrices for sequence alignments and database searches from the evolutionary model, we calculate the residue similarity scores (Karlin and Altschul 1990) $b_{ij}(t)$ between residues i and j at different evolutionary time t from the rate matrix $\boldsymbol{Q}$:

$$b_{ij}(t) = \frac{1}{\lambda}\log\frac{p_{ij}(t)}{\pi_j} = \frac{1}{\lambda}\log\frac{m_{ij}(t)}{\pi_i\pi_j},$$

where $m_{ij}(t)$ is joint probability of observing both residue type i and j at the two nodes separated by time t, and λ is a scalar. Here $b_{ij}(t)$ satisfies the equality $\sum \pi_i\pi_j e^{\lambda b_{ij}} = 1$, because of the property of the joint probability $\sum_{ij} m_{ij}(t) = \sum_{ij} \pi_i p_{ij}(t) = \sum_i \pi_i = 1$ holds for Markov matrix which has the property $\sum_j p_{ij}(t)$ (Grimmett and Stizaker 2001). The overall expected score of this matrix is then $\sum_{ij} m_{ij}(t) b_{ij}(t)$, usually in bit units (Karlin and Altschul 1990).

Computation of Surface Pockets and Interior Voids

We use the Volbl method to compute the solvent accessible (SA) surface area of protein structures (Edelsbrunner et al. 1995; Liang et al. 1998). We use the CastP method (Liang, Edelsbrunner, and Woodward 1998; Binkowski, Naghibzadeh, and Liang 2003) to identify residues located on surface pockets. Both Volbl and CastP are based on precomputed alpha shapes (Edelsbrunner and Mücke 1994), where the dual simplicial complex is constructed from the Delaunay triangulation of the atomic coordinates of the protein. We use the pocket algorithm (Edelsbrunner, Facello, and Liang 1998; Liang, Edelsbrunner, and Woodward 1998) in CastP to identify residues located in surface pockets and interior voids. Details and other applications of these methods can be found in Edelsbrunner, Facello, and Liang (1998), Liang, Edelsbrunner, and Woodward (1998), Liang and Dill (2001), and Binkowski, Adamian, and Liang (2003).

Results

There are a large number of parameters (189) characterizing the substitutions of amino acid residues. We first need to understand at what accuracy these parameters can be estimated. Because we are studying regions (e.g., binding surfaces) on a protein structure, we often only have a few dozen instead of a few hundred residue positions available for parameter estimation. In addition, we are frequently limited by the available sequence data, and the size of the phylogenetic tree may be moderate. Even if the parameters of the substitution model can be estimated, it is not clear how effective they are for applications such as inferring protein functions from protein structures. We describe our results addressing each of these issues.

Rate Estimation: Simulation Studies

We first carry out a simulation study to test the accuracy of the estimated residue substitution rates. We generate a set of artificial sequences based on an evolutionary model with known substitution rates. We ask whether our method can recover the original substitution rates reasonably well and how many sequences and residues are necessary so an accurate estimation can be made. For this purpose, we first take the sequence of the alpha-catalytic subunit of cyclic adenosine monophosphate (cAMP)-dependent protein kinase (SwissProt P36887, pdb 1cdk, with length 343) and the sequence of carboxypeptidase A2 precursor (SwissProt P48052, pdb 1aye, length 417).

Statistics for Estimation Accuracy

We use the JTT evolutionary model (Jones, Taylor, and Thornton 1992), which is characterized by a frequency-independent amino acid interconversion rate matrix $\boldsymbol{S}_{JTT}$ and the diagonal matrix $\boldsymbol{D}$ of the composition of the 20 amino acid residues for the set of sequences that were used to derive the original JTT model (Yang 1997). The substitution rate matrix $\boldsymbol{Q}_{JTT}$ is then: $\boldsymbol{Q}_{JTT} = \boldsymbol{S}_{JTT}\cdot\boldsymbol{D}$. To avoid potential bias, we use the composition $\boldsymbol{D}$ of the protein kinase and the frequency-independent amino acid interconversion rate matrix of $\boldsymbol{S}_{JTT}$ to calculate the instantaneous rate matrix $\boldsymbol{Q}$ for the protein kinase, which is then used to generate 16 artificial kinase sequences at different time intervals t using the probability $\boldsymbol{P}(t) = \exp(\boldsymbol{Q}t)\boldsymbol{I}$. Here we use a simple balanced phylogenetic tree of 16 leaf nodes with equal branch lengths of $t = 0.1$ for all edges. We compare the estimated frequency-independent amino acid interconversion rate matrix $\tilde{\boldsymbol{S}}$ to the true matrix $\boldsymbol{S}_{JTT}$.

For comparison, we first normalized the estimated and true JTT frequency-independent interconversion rate matrices, such that:

$$\frac{1}{20}\sum_{ij,i\neq j} s_{ij} = 1 \quad \text{and} \quad \frac{1}{20}\sum_{ij,i\neq j} \tilde{s}_{ij} = 1,$$

where s_{ij} is the (i, j)-th entry of the matrix $\boldsymbol{S}$.

We are interested in the rates of substitution that occur in a specific spatial region of the protein. Because these regions contain only a subset of the residues and often are under different selection pressure, not all possible substitutions are observed with adequate frequency for estimation. In addition, the usually moderate size of the phylogenetic tree limits the observed frequency of some substitutions. Nevertheless, the frequently observed substitutions

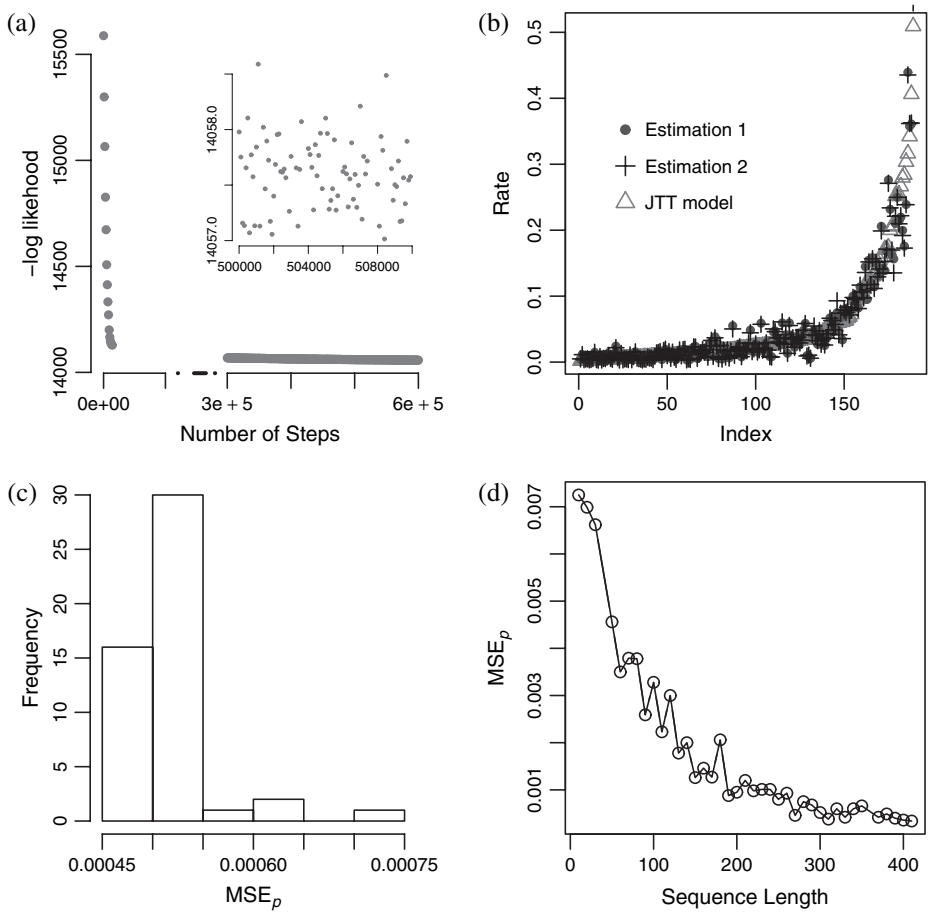

FIG. 1.—Estimating residue substitution rates using simulated carboxypeptidase sequences. (*a*) The Markov chain converges after 3×10^5. The insert shows negative log-likelihood ($-\ell$) values in stationary state after the burning-in period. (*b*) s_{ij} values estimated in two simulations are all similar to the true rates. In the first simulation, the 189 initial values are set such that $s_{ij} = 0.1$ for all entries. In the second simulation, the 189 initial s_{ij} values are sorted numerically by index i then by index j, and the values are assigned from 0.1 with an increment of 0.01 for the next entry. (*c*) The $MSE_{\mathcal{P}}$ values from 50 repeated estimations of substitution rates of carboxypeptidase with random initial values are all less than 8×10^{-4}. The mean value of $MSE_{\mathcal{P}}$ is 5.2×10^{-4}. (*d*) The value of $MSE_{\mathcal{P}}$ depends on the length of available subsequences. For subsequence of length $\geq$20, the $MSE_{\mathcal{P}}$ value is $<$0.008.

for a specific protein region are likely to be the most important ones, and the estimation of their rates should be better than the rates of infrequently observed substitutions.

We need to quantitatively assess our estimation error. Because it is very difficult to estimate accurately the absolute values of the individual rates, we assess instead the errors in estimated $\tilde{s}_{ij}$ in terms of their effects on the overall patterns of residue substitution on a specific protein region. This is more appropriate for many applications such as the analysis of the evolution of binding surfaces and the evolution of the folding core, as only a subset of substitutions occur at a functional surface or in the core. We develop some quantitative measures for this purpose.

We call a residue pair (i, j) an "occurring pair" if both residues i and j occur simultaneously in one column of the multiple-aligned sequences of a specific region. For the subset of rates $\mathcal{S}=\{s_{ij}\}$ for a residue pair (i, j) from the set of occurring pairs $\mathcal{P}$, we obtain the "relative contribution" of a specific frequency-independent interconversion rate between a pair of residues as:

$$s'_{ij} = s_{ij} / \sum_{ij \in \mathcal{P}} s_{ij}.$$

The Δe_{ij} "weighted error in contribution" is computed as:

$$\Delta e_{ij} \equiv \frac{f_{ij}}{\sum_{ij, i \neq j} f_{ij}} [\tilde{s}'_{ij} - s'_{ij}],$$

where $\tilde{s}_{ij}$ is the estimated value of s'_{ij} and f_{ij} is the number count of how often the (i, j) substitutions occur.

To measure the overall differences of the estimated $\tilde{\boldsymbol{S}}$ and the original $\boldsymbol{S}_{JTT}$ matrices for the occurring substitutions, we use the "weighted mean square error" ($MSE_{\mathcal{P}}$) Mayrose et al. (2004):

$$MSE_P \equiv \frac{1}{|P|} \sum_{ij \in P, i \neq j} \Delta e_{ij}^2.$$

Error Analysis in Estimated Rates

Using the 16 artificial sequences generated from the sequence of carboxypeptidase and a simple balanced phylogenetic tree with equal branch length $t = 0.1$ for all edges between nodes, the Markov chain converges quickly after

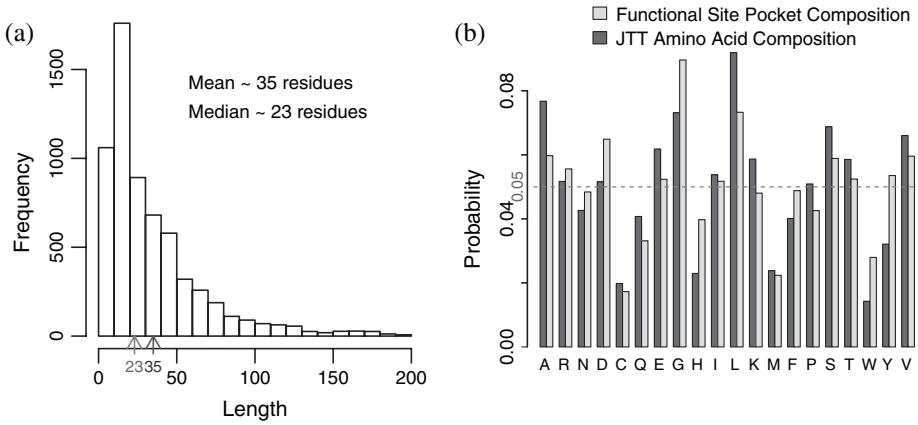

FIG. 2.—The length distribution and amino acid composition of functional pockets. (*a*) The length distribution of 6,273 functional pockets. The average length of functional pockets is 35 residues, and the median is 23 residues. (*b*) Comparison of amino acid compositions of residues in 6,273 functional pockets with the composition of 16,300 protein sequences used to derive the JTT substitution matrix. The dashed line is the expected probability of 0.05 if all substitution rates following the uniform distribution.

3×10^5 Monte Carlo steps (fig. 1*a*), as shown by the value of $-\ell$ for the negative likelihood function. After a burning-in period of 3×10^5 steps, we collect $m = 4 \times 10^5$ samples for estimating $\{s_{ij}\}$ values. Figure 1*b* shows the estimation results for two simulations started from two different sets of initial values of $\{s_{ij}\}$. It is clear that both sets of estimated rates $\{\tilde{s}_{ij}\}$ for the occurring pairs are in general agreement to the set of true values from the JTT model.

To further assess how robust the estimations are, we repeated the Markov chain Monte Carlo simulation 50 times using random initial values of $\{s_{ij}\}$ drawn from a uniform distribution of $\mathcal{U}(0,1)$. On an average, the estimation error is small. The mean of the overall weighted $MSE_{\mathcal{P}}$ from 50 simulations is 5.2×10^{-4} for occurring pairs (fig. 1*c*).

Length Dependency of Errors in Estimated Parameters

To estimate region-specific substitution rates, it is important to assess how the accuracy of the estimation depends on the size of the region. For example, the functional region of a protein contains only a small number of residues, which varies depending on the size of the binding site. We carry out another simulation study for this purpose. Starting from the N-termini of the 16 artificially generated carboxypeptidase sequences, we take a subsequence from each sequence, with the length increasing from 10 to 417, at an increment of ten residues. We then estimate the substitution rates at each length. Each simulation of a different length was started from a random set of initial values drawn from $\mathcal{U}(0,1)$, and the same burning-in period and sample size m are used as before. The $MSE_{\mathcal{P}}$ values obtained using sequences of different lengths are plotted in figure 1*d*. Our results show that for this set of sequences, as long as the number of residues is $\geq$20, the $MSE_{\mathcal{P}}$ of the estimated parameters will be less than 0.008.

Based on analysis of the protein structures in the Protein Data Bank, we found that among the surface pockets from 6,273 protein structures that all contain annotated functional residues (as recorded either in the Feature table of the SwissProt database or the Active Site field of the PDB file), the average size of a functional site pocket is 35 residues, and the median is 23 residues (fig. 2*a*). This suggests

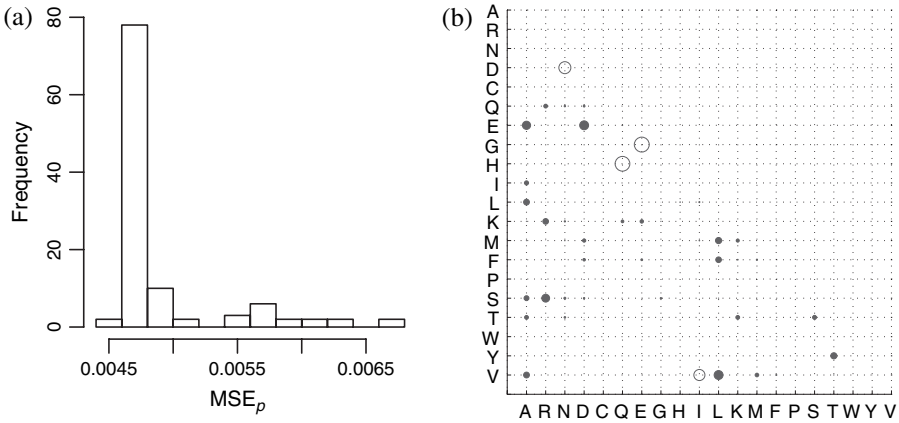

FIG. 3.—Estimating the substitution rates of residues on the binding surface of cAMP-dependent protein kinase from simulated sequences. (*a*) For 110 independent estimations of the substitution rates with random initial values, the $MSE_{\mathcal{P}}$ values are all $<8 \times 10^{-4}$. The mean $MSE_{\mathcal{P}}$ value of the 110 estimations is 0.0048. (*b*) There are only four substitutions (empty circles) whose error Δe_{ij} is greater than 3.0%, although all of the 90 occurring pairs have $\Delta e_{ij} < 4.5\%$.

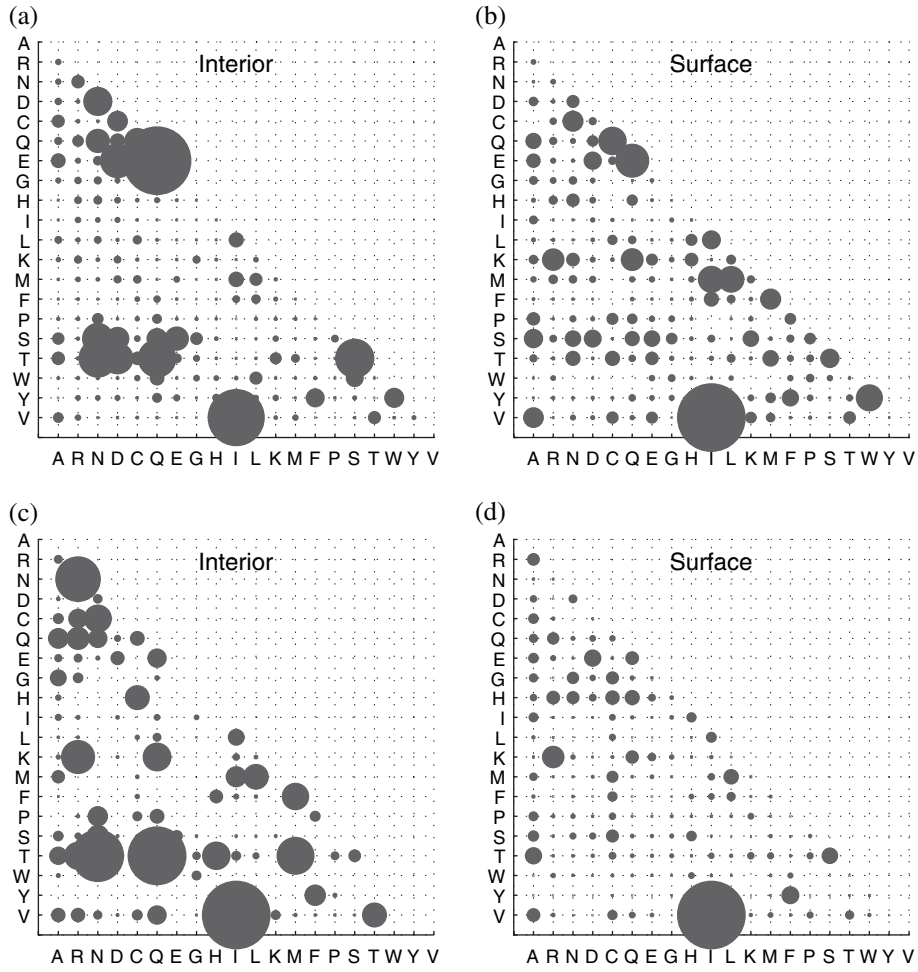

FIG. 4.—Substitution rates of residues on solvent-exposed surface and in buried interior. (*a*) Substitution rates of buried interior residue on 2-haloacid dehalogenase (pdb 1qh9). There are 100 occurring pairs. (*b*) Substitution rates of surface-exposed residues of 1qh9. There are 188 occurring pairs. (*c*) Substitution rates of buried interior residue of alpha amylase (pdb 1bag). There are 190 occurring pairs. (*d*) Substitution rates of surface-exposed residues of 1bag. There are 177 occurring pairs.

that our method will be applicable for the analysis of protein functional pockets.

We carried out another simulation study estimating substitution rates only for the binding surface of a protein. Using the same phylogenetic tree as that of the carboxypeptidase simulations and the same JTT model, we generate 16 artificial sequences of the alpha-catalytic subunit of cAMP-dependent protein kinase (SwissProt P36887, pdb 1cdk, length 343). Our goal is to estimate rates only for the subset of 38 residues located in the binding site. Figure 3*a* shows that the $MSE_{\mathcal{P}}$ values of the estimated rates from 110 independent simulations for the 90 occurring pairs of residues are all small. The estimated rates from all simulations have $MSE_{\mathcal{P}} < 8 \times 10^{-3}$, and the mean of the overall $MSE_{\mathcal{P}}$ from 110 simulations is 4.8×10^{-3} for the 90 occurring pairs. Clearly, the estimation errors measured in $MSE_{\mathcal{P}}$ are larger when only residues in the binding site are used compared to the estimation errors of carboxypeptidase where all 417 residues are used. Nevertheless, the estimations are still useful, as the mean $MSE_{\mathcal{P}}$ value remains small. Figure 3*b* plots the individual mean value of weighted errors Δe_{ij} for the 90 occurring pairs obtained from 110 simulations. There are only four substitutions whose weighted error in contribution Δe_{ij} is greater than 3%, although all occurring pairs have $\Delta e_{ij} < 4.5\%$.

Evolutionary Rates are Region Specific

Exposed Surface and Buried Interior Have Different Substitution Rates

Residues on protein surfaces that are exposed to solvent are under different physicochemical constraints from residues in the buried interior. We estimate the substitution rates for exposed and buried regions on a protein structure. We use a simple criterion to classify residues as either exposed or buried: based on the calculation of SA surface area using Volbl (Liang et al. 1998), we declare a residue to be buried if its SA area is 0 $Å^2$ and exposed if SA area > 0 $Å^2$.

For the protein 2-haloacid dehalogenase (pdb 1qh9), figure 4 shows that the residues on the exposed surfaces and in the buried interior have very different substitution patterns. For example, the substitution of threonine (T) with asparagine (N), aspartate (D), or glutamine (Q) occurs much more frequently in the buried interior than on the

Table 1
Substitutions Rate of Residues in the Interior and on the Exposed Surface Are Different

Protein Family	pdb	Interior Occurring Pairs	Surface Occurring Pairs	*P* Value of K-S Test
EC 3.4.11.18	1b6a	80	175	0.016
EC 3.2.1.1	1bag	190	177	0.015
EC 2.3.3.1	1csc	55	163	0.009
EC 3.8.1.2	1qh9	139	169	0.023
EC 3.2.1.21	1h49	60	169	0.024
EC 3.5.1.5	1udp	92	162	0.014
EC 1.1.1.37	1b8v	97	150	4.8×10^{-5}

NOTE.—K-S, Kolmogorov-Smirnov.

surface (fig. 4*a* and *b*). A similar pattern is also seen for alpha amylase (pdb 1bag, fig. 4*c* and *d*). In general, ionizable and polar residues in the protein interior have higher propensities to mutate to other ionizable and polar residues.

The frequent substitutions between T and {N, D, Q} observed in the protein interior of l-2-haloacid dehalogenase and amylase suggest that to maintain the H-bonding interactions in the protein interior, it is far more common to have substitutions among ionizable residues and polar residues. These substitution patterns point to the importance of preserving polar interactions, which provide important structural stability in the protein interior, as the high dielectric constants inside proteins makes the electrostatic contribution of salt bridges and H-bonds in the protein interior stronger than H-bonds on protein surfaces.

The conclusion that residues in the protein interior experience different selection pressure from residues on the protein surfaces are likely to be true for other proteins. We estimated the substitution rates of buried residues and exposed residues for six additional proteins with different biological functions as indicated by different enzyme classification numbers (table 1). In all cases, we find that surface residues have different evolutionary patterns overall. Although not all substitution rates are noticeably different, table 1 shows that for each of the eight proteins studied, we can reject the null hypothesis, based on the nonparametric Kolmogorov-Smirnov test, that the two distributions of substitution rates for the set of exposed residues and the set of buried residues are the same.

Residues in Functional Sites and on the Rest of the Surface Have Different Substitution Rates

Protein functional sites are the regions where a protein interacts with ligand, substrate, or other molecules. Because proteins fold into their three-dimensional native structures, functional sites often involve residues that are distant in sequence but are in spatial proximity. As can be seen in figure 5, two proteins with a low sequence identity (<16%) may be very different overall, but their functional binding pockets may be quite similar. In this study, we use the CastP database of precomputed surface pockets for our analysis of functional sites on protein structures. This approach has been applied in studies of protein function prediction (Binkowski, Adamian, and Liang 2003; Binkowski, Naghibzadeh, and Liang 2003) and in structural analysis of nonsynonymous single-nucleotide polymorphisms (Stitziel et al. 2003).

Residues that are located in functional pockets are under different selection pressures. This can be clearly seen in figure 2*b*, such that the composition of residues in functional pockets is very different from the composition of residues in the set of full protein sequences from which the JTT substitution matrix was derived. Here we examine only protein surface pockets that contain functionally important residues as annotated by either SwissProt or PDB. In functional pockets, Tyr, Trp, His, Asp, and Gly residues are far more enriched, but Leu, Ser, and Ala are less if compared to sequences used in the JTT rate matrix analysis. Tyr, Trp, His, and Asp are residues that play important roles in enzyme reactions through electrostatic interactions, change of protonation states, and aromatic interactions. Gly residues are important in the formation of turns and other geometric features for binding site formation. The enrichment of hydrophobic Leu and small residues Ser and Ala in the full sequence are probably important for structural stability.

We examine the patterns of residue substitutions on protein functional surfaces in some detail. Taking a structure of alpha amylase (pdb 1bag) as an example, we compare the estimated substitution rate matrix of functional surface residues with that of the remaining surface residues of the protein (fig. 6). It is clear that the selection pressures for residues located in functional site and for residues on the rest of the protein surface are different and they are also both different from the JTT matrix (data not shown). This suggests that identifying functionally related protein surfaces will be more effective if we employ scoring matrices specifically derived from residues located on functional surface instead of using a general precomputed substitution matrix.

Application: Detecting Functionally Similar Biochemical Binding Surfaces

For proteins carrying out similar functions such as binding similar substrates and catalyzing similar chemical reactions, the binding surfaces experience similar physical and chemical constraints. The sets of allowed and forbidden substitutions will therefore be similar because of these constraints. The continuous time Markov model can provide evolutionary information at different time intervals once the instantaneous substitution rates are estimated. This information is encoded in the time-dependent residue substitution probabilities. An objective test of the utility of the estimated evolutionary model is to examine if we can discover functionally related proteins, namely, whether we can identify protein structures that have similar binding surfaces and carry out similar biological functions.

Identification of Functionally Related Proteins from a Template Binding Surface

We use alpha amylases as our test system. Alpha amylase (Enzyme Classification [EC] number, EC 3.2.1.1) acts on starch, glycogen, and related polysaccharides and oligosaccharides. Detecting functionally related alpha amylase is a challenging task, as many of them have very low overall sequence identities (<25%) to the query protein template. If two proteins have a sequence identity below 60%–70%, it becomes difficult to make functional inferences based on sequence alignment (Rost 2002).

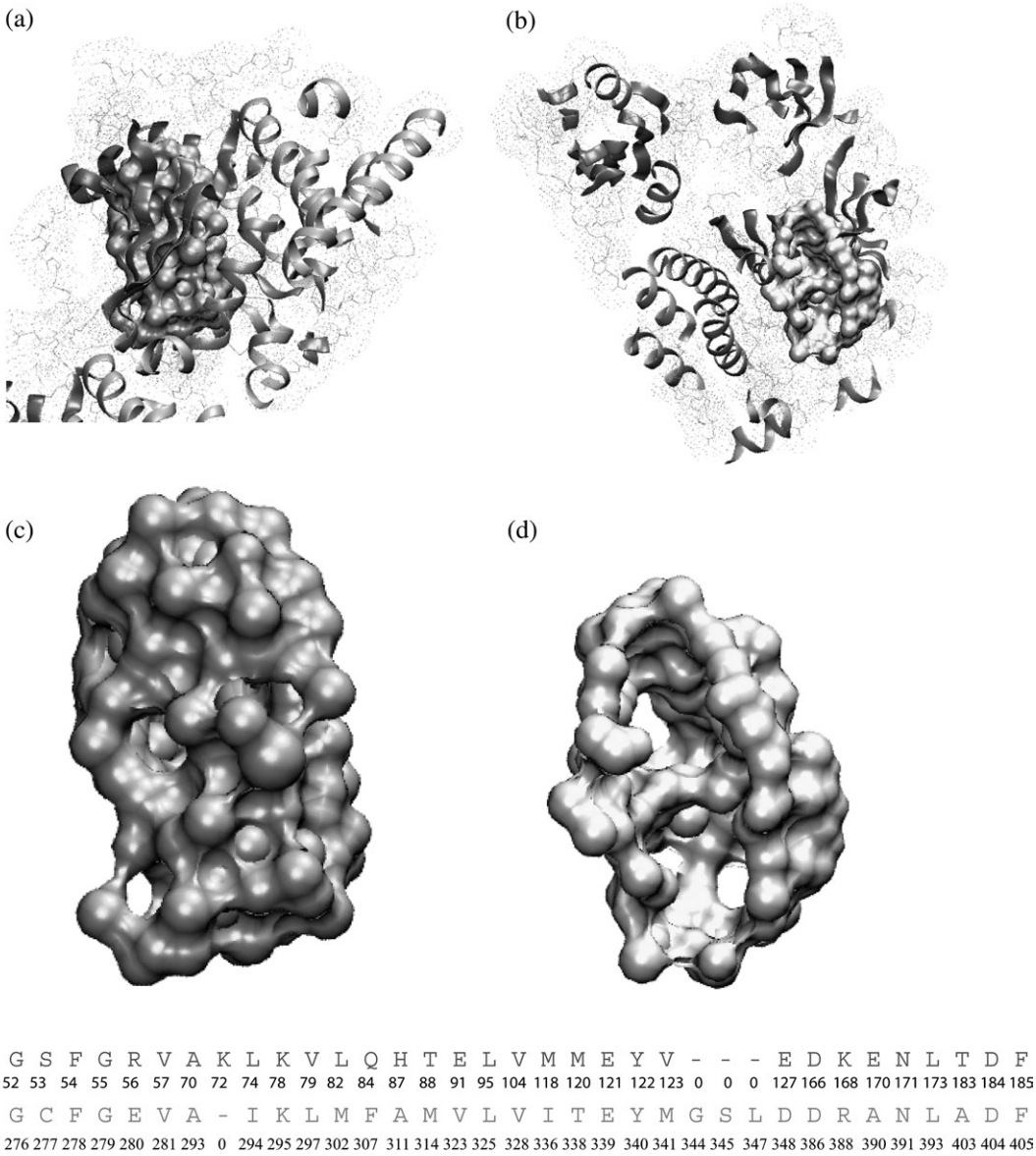

FIG. 5.—Protein functional pockets of kinases. Functional site of (*a*) the catalytic subunit of cAMP-dependent protein kinase (1cdk chain A), and (*b*) tyrosine protein kinase c-src (2src). Both kinases bind to AMP or AMP analogs. Their global primary sequence identity is as low as 16%. However, if we extract their binding surfaces (as shown in *c* and *d*) out, (*e*) the residues forming the binding pockets have much a higher sequence identity (51%).

Given a template binding surface from an alpha amylase (1bag, pdb), we wish to know how many protein structures can be identified that have the same EC number at an accuracy of all four EC digits. These protein structures all carry out the same or related reactions. By the convention of the EC system, the EC numbers represent a progressively finer classification of the enzyme with the first digit about the basic reaction and the last digit often about the specific functional group that is cleaved during reaction.

We first exhaustively compute all of the voids and pockets on this protein structure (Liang, Edelsbrunner, and Woodward 1998; Binkowski, Naghibzadeh, and Liang 2003). Based on biological annotation contained in the Protein Data Bank, the 60th pocket containing 18 residues is identified as the functional site (fig. 7*b*). To construct an evolutionary model, we use sequence alignment tools to gather sequences homologous to that of 1bag (Altschul et al. 1997). After removing redundant sequences and sequences with $>90\%$ identity to any other identified sequences or the query sequence of 1bag, we obtain a set of 14 sequences of amylases. These 14 sequences are used to construct a phylogenetic tree of alpha amylase (fig. 7*a*). We use the maximum likelihood method implemented in the Molphy package for tree construction (Adachi and Hasegawa 1996).

We then calculate the similarity scoring matrices from the estimated values of the rate matrix. Because a priori we do not know how far a particular candidate protein is separated in evolution from the query template protein, we calculate a series of 300 scoring matrices, each characterizing the residue substitution pattern at a different time separation, ranging from 1 time unit to 300 time unit. Here 1 time unit represents the time required for 1 substitution per 100

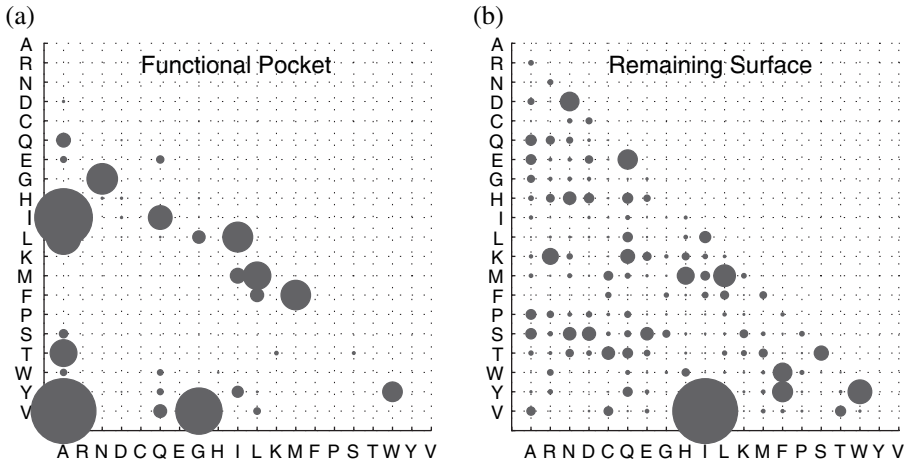

FIG. 6.—Substitution rates of residues in the functional binding surface and the remaining surface of alpha amylase (pdb 1bag). (*a*) Substitution rates of the functional binding surface. There are 39 occurring pairs. (*b*) Substitution rates of the remaining surface on 1bag. There are 177 occurring pairs.

residues (Dayhoff, Schwartz, and Orcutt 1978). We use the Smith-Waterman algorithm as implemented in the Ssearch method of Fasta (Pearson 1991) with each of the 300 scoring matrices in turn to align sequence patterns of candidate binding surfaces from a database of >2 million protein surface pockets contained in the pvSoar database (Binkowski, Freeman, and Liang 2004). We use an E value of 10^{-1} as the threshold to decide if a matched surface pocket is a hit. Surfaces similar to the query binding pocket identified (with E values $< 10^{-1}$) are then subjected to further shape analysis, where those that cannot be superimposed to the residues of the query surface pattern at a statistically significant level (P value < 0.01) by either the coordinate squared root of mean square deviations (RMSD) measure or the orientational RMSD (Binkowski, Adamian, and Liang 2003) measure are excluded. The P value is estimated using methods developed by Binkowski, Adamian, and Liang (2003).

A total of 58 PDB structures are found to have similar binding surfaces to that of 1bag, and hence are predicted as amylases. All of them turn out to have the same EC number of 3.2.1.1 as that of 1bag. We repeat this study but using a different amylase structure as the query protein. Using the functional pocket on 1bg9, we found 48 PDB structures with EC 3.2.1.1 labels. The union of the results from these two searches gives 69 PDB structures with EC 3.2.1.1 labels. Examples of matched protein surfaces are shown in figure 7.

Comparison with Others

We compare our results with other studies. The Enzyme Structure Database (ESD) (http://www.ebi.ac.uk/thornton-srv) collects protein structures for enzymes contained in the Enzyme data bank (Bairoch 1993) for study. Here we take the ESD database as the gold standard, and all true answers are contained in this human curated database. There are 75 PDB entries with enzyme class label EC 3.2.1.1 in ESD (version Oct, 2004). Out of the 75 structures, our method discovered 69 PDB structures (no redundancy) using 1bag and 1bg9 as queries.

We also compare our results with those obtained from a database search using sequence alignment methods. Using the Smith-Waterman algorithm as implemented in Ssearch of the Fasta package with the default Blosum50 matrix, only 32 structures are identified as alpha amylase (see table 2 in Binkowski, Adamian, and Liang [2003]). When using Psi-Blast and the NR database with default parameters, an E value threshold of 10^{-3}, and <10 iterations to generate position-specific weight matrices, 65 structures (no redundancy) among the 75 known structures of alpha amylase are found after combining results from queries with 1bag and 1bg9.

We next tested search results using the standard JTT matrix instead of the estimated protein-specific and surface-specific matrix. In this case, we find 52 hits instead of 58 using 1bag as the query protein and 8 hits instead of 48 using 1bg9 as the query protein.

Our method differs from Ssearch (Pearson 1998) in two aspects: first, we use short sequence patterns generated from the binding surface of the protein structure instead of the full protein sequences. Second, we use the customized scoring matrix derived from the estimated evolutionary model instead of the standard Blosum matrix. Psi-Blast differs from our method in that it also uses full-length primary sequences and it effectively uses an empirical model of position-specific weight matrices to extract evolutionary information from a set of multiple-aligned sequences, without the benefit of using a phylogeny and an explicit parametric model.

Compared to the Fasta sequence alignment and Psi-Blast search, our method can identify more alpha amylases. In addition, because we directly detect binding surface similarity instead of global sequence similarity, our prediction has stronger implications for inferring functional relationships. In contrast, Psi-Blast search does not provide information about which residues are important for function. We have also shown that our estimated rate matrix works much better than the generic precomputed JTT matrix, especially when the query template surface has a relatively small size.

To examine whether our method works for proteins of other functions, we repeated our test using four additional enzymes of different biochemical functions. These are: 2,

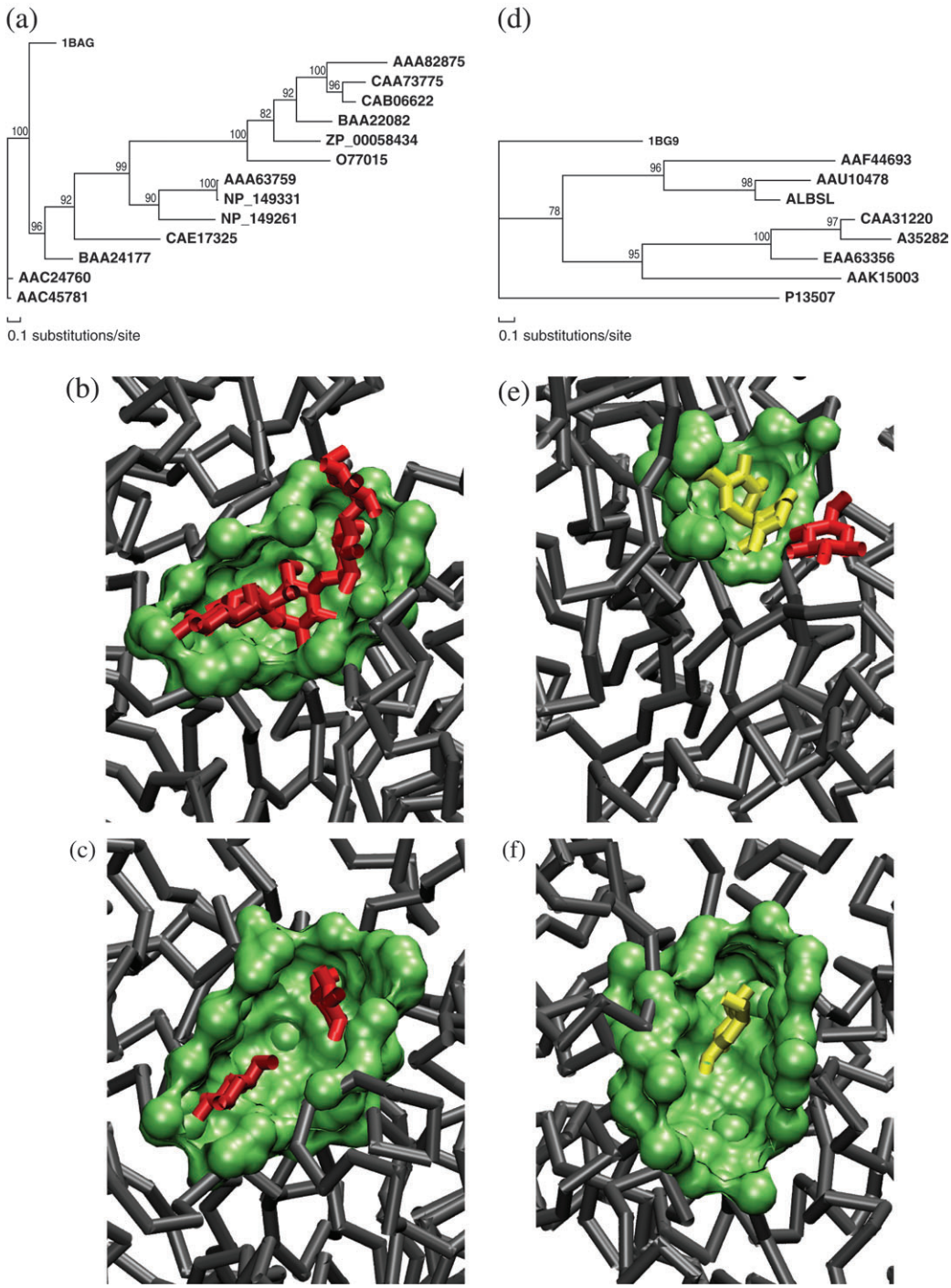

FIG. 7.—Function prediction of alpha amylases. (*a*) The phylogenetic tree for PDB structure 1bag from *Bacillus subtilis*. (*b*) The functional binding pocket of alpha amylase on 1bag. (*c*) A matched binding surface on a different protein structure (1b2y from human, full sequence identity 22%) obtained by querying with the binding surface of 1bag. (*d*) The phylogenetic tree for 1bg9 from *Hordeum vulgare*. (*e*) The binding pocket on 1bg9. (*f*) A matched binding surface on a different protein structure (1u2y from human, full sequence identity 23%) obtained by querying with 1bg9.

3-dihydroxybiphenyl dioxygenase (EC 1.13.11.39), adenosine deaminase (EC 3.5.4.4), 2-haloacid dehalogenase (EC 3.8.1.2), and phosphopyruvate hydratase (EC 4.2.1.11). As shown in table 2, we are able to find all other protein structures of the same EC numbers contained in the ESD in all four cases. Our results are better than using Psi-Blast or using the JTT matrix.

Discussion

We have developed a Bayesian method for estimating residue substitution rates. Bayesian inference of phylogeny was independently introduced by Yang and Rannala (1997), Mau, Newton, and Larget (1999), and Li, Pearl, and Doss (2000). Bayesian methods have found wide applications (Huelsenbeck et al. 2001, 2002), including host-parasite

Table 2
Detecting Functionally Related Proteins

Protein Family	Query Structure	Pocket[a] identification	Pocket Length	Our[b] Result	Results by Psi-Blast[c]	Results by JTT[d]	ESD[e] (true answers)
EC 3.2.1.1	1bag	60	18	58	45	52	75
EC 3.2.1.1	1bg9	61	12	48	21	8	75
EC 3.8.1.2	1qh9	23	16	8	8	3	8
EC 3.5.4.4	2ada	49	28	23	17	19	23
EC 4.2.1.11	1ebh	122	35	22	20	19	22
EC 1.13.11.39	1kw9	34	23	18	16	18	18

[a] Pocket id could be referenced through CastP database (http://cast.engr.uic.edu).
[b] Our results are obtained from querying with a template binding surface and customize scoring matrices.
[c] The true answers are taken as those recorded in the human curated ESD database.
[d] Results using Psi-Blast sequence alignment.
[e] Results using our method with a standard JTT matrix.

cospeciation (Huelsenbeck, Rannala, and Yang 1997), estimation of divergence times of species (Thorne, Kishino, and Painter 1998), simultaneous sequence alignment and phylogeny estimation (Mitchison 1999), inference of ancestral states (Huelsenbeck and Bollback 2001), and determination of the root position of a phylogenetic tree (Huelsenbeck, Bollback, and Levine 2002). Similar to others, our approach is based on the Markov chain Monte Carlo sampling technique. Although we are not aware of any other studies using Bayesian models for the direct estimation of substitution rates between amino acid residues, our approach is a natural extension of existing work on maximum likelihood estimation (Goldman and Yang 1994; Yang, Nielsen, and Hasegawa 1998) of codon substitution rates for amino acid residues and other studies based on Bayesian statistical analysis (Huelsenbeck, Rannala, and Yang 1997; Yang and Rannala 1997; Thorne, Kishino, and Painter 1998; Mau, Newton, and Larget 1999; Huelsenbeck et al. 2001).

In this work, we studied the substitution of residues using amino acid sequences instead of nucleotide sequences. In our model, the parameters of the continuous time Markov process are the rates of direct substitutions between residues. A more established model of residue substitution is that of the substitutions between codons. This model can provide rich information about detailed mechanisms of molecular evolution. For example, the differential effects of transition versus transversion and synonymous versus nonsynonymous substitutions all can be modeled (Goldman and Yang 1994; Yang, Nielsen, and Hasegawa 1998). Our choice of the current model of direct residue substitution is based on two practical considerations. First, for the application of predicting protein functions, we find it is far easier to gather amino acid residue sequences than nucleotide sequences when large scale database searches are carried out. Second, when using scoring matrices derived from substitution rates to detect remotely related proteins, amino acid sequences give far better results in sensitivity and specificity than nucleotide sequences (Pearson 1998; Liò and Goldman 1999). An interesting future study would be one that is based on codon substitution models, which will help to identify possible bias in the current approach, where the effects of transition/transversion and synonymous/nonsynonymous substitutions are not considered.

It has long been recognized that the evolutionary divergence of protein structures is far slower than that of sequences (Chothia and Lesk 1986). Because physical constraints on protein structure would give rise to associations between patterns of amino acid replacement and protein structure (Koshi and Goldstein 1996, 1997), the substitution rates of residues in different secondary structural environments and of different solvent accessibility have been well studied (Lesk and Chothia 1982; Goldman, Thorne, and Jones 1996, 1998; Thorne, Goldman, and Jones 1996; Bustamante, Townsend, and Hartl 2000). In a pioneering work, Thorne, Goldman, and Jones developed an evolutionary model that combines secondary structure with residue replacement, and showed that the incorporation of secondary structure significantly improves the evolutionary model for sucrose synthase (Thorne, Goldman, and Jones, 1996). The impact of secondary structure and solvent accessibility on protein evolution were further studied in detail using a hidden Markov model in (Goldman, Thorne, and Jones 1998). Additional work showed that an accurate evolution model can in turn lead to accurate prediction of protein secondary structure (Goldman, Thorne, and Jones 1996; Liò et al. 1998). Parisi and Echave have further developed a simulation model to study the effects of selection of structural perturbation on the site-dependent substitution rates of residues (Parisi and Echave 2001, 2005; Robinson et al. 2003). These studies highlighted the importance of physical constraints on protein evolution.

Our work is a continuation in the direction of assessing substitution rates of residues in different structural environments, but with an important novel development. Here we proposed to study substitution rates of residues in a new structural category, namely, residues from local binding surface regions that are directly implicated in biochemical functions. Because a fundamental goal of studying protein evolution is to understand how biological functions emerge, evolve, and disappear (J. Gu and X. Gu 2003; Vogel et al. 2004; Lecomte, Vuletich, and Lesk 2005), estimation of the substitution rates of residues on functional surfaces is critically important.

Proteins are selected to fold to carry out necessary cellular roles. In many cases, they are involved in binding interactions with other molecules. Surface binding pockets and voids are therefore the most relevant structural regions, which can be computed using exact algorithms (Liang, Edelsbrunner, and Woodward 1998). A unique advantage of this novel structural category is that it allows better

separation of residues experiencing selection pressure due to the constraints of biochemical functions from those due to the constraints for physical structural integrity. In contrast, the structural categories of residues in different secondary structural environments and solvent accessibility are more suited to study how substitutions are related to protein stability because they inevitably will include many conservation patterns due to the requirements of structural stability.

For example, solvent accessibility directly relates to the driving force of hydrophobic effects for protein folding, and secondary structures are essential for maintaining protein stability (Dill 1990; Dill et al. 1995). The structural categorizations developed in (Goldman, Thorne, and Jones 1996, 1998; Thorne, Goldman, and Jones 1996) are well suited for studying how protein evolution is constrained by physical interactions important for protein folding and stability. For example, the patterns of hydrophobic residues in the buried interior, polar residues on the surface, and small residues in β-turns are all due to structural constraints and do not have direct functional implications. Indeed, the study of Koshi and Goldstein found strong correlation between transfer free energy ΔG of amino acid residues, a physicochemical property of amino acid solvation energy, and residue substitution rates (Koshi and Goldstein 1996). The categorization of residues proposed here are designed for studying how protein evolution is constrained by function (i.e., protein-ligand/substrate binding and protein-protein interactions). To our best knowledge, this is the first study in which a structure-derived category amenable for computation is proposed that separates residues selected for function from residues selected for stability.

Our results showed that residues located in functional pockets have different substitution rates from residues in the remaining parts of the protein. The differences are mostly due to residues such as His and Asp that are known to be important for protein function. All of these region-specific substitution rate matrices are different from the precomputed Blosum matrix.

It is informative to examine the difference of the substitution rates in the JTT matrix and the binding site-specific rate matrices we estimated. The JTT matrix was developed using a very large database of sequences, and the overall composition D_{JTT} of amino acid residues is very different from the composition D of the binding surfaces. Hence, the conserved residues, or the values of the diagonal elements s_{ii} of the substitution matrix, are very different. This is reflected in the different residue composition for functional surfaces and the full protein sequence (fig. 2*b*). This would result in different overall patterns of substitutions. For substitution after a long time interval, it is necessary to estimate the off-diagonal elements s_{ij} with some accuracy as the substitutions would accumulate with time, and identifying remotely related binding surfaces becomes difficult.

It is challenging to estimate substitution rates of amino acid residues in a local region. The number of residue positions for a specific region may be small, and the available sequences in the phylogenetic tree may also be limited. It is unlikely that all 189 independent substitution rates of the 20 $\times$ 20 matrix can be estimated accurately when only limited data are available. In this study, we can only estimate substitution rates for occurring pairs, namely, substitutions between residues that occur in the same position in different sequences. However, for applications such as inferring protein functions by matching similar binding surfaces, our results show that the constructed scoring matrices are very effective. It is likely that the substitutions (or lack thereof) that occur in the sampled data for a specific region are the most important ones in overall patterns of evolution of residues in this specific region. For example, the most important features in a functional pocket on a protein structure are the conserved residues. Accurate estimation of the diagonal rates (s_{ii}) is therefore the most important task. Because conserved residues appear in relatively higher frequency, they often can be estimated well. If some substitutions never occur in the sampled data, they probably are not important and setting their values to a baseline offset value such as that from a uniform prior would be reasonable. We have carried out detailed studies on identifying functionally related alpha amylases and other enzymes by querying with one or more template binding surface and assessing similarity using scoring matrices derived from the estimated rates. As shown in table 2, our approach works very well in practice. In a control study, we assign random values to the matrix entries, which conform to the normalization condition. Scoring matrices derived from this randomized rate matrix are ineffective, and we were not able to find any functionally related proteins for any example listed in table 2.

One might wish to estimate a 20 $\times$ 20 substitution rate matrix that is specific to an individual site or position in the sequence. However, this would require a very large amount of data that are not available in practice. In addition, it is conceivable that estimating site-specific rate matrices may not be necessary or possible. For example, if a residue is critical for protein folding stability, it might be conserved through all stages of the evolution, and there is no variation at this particular position of the amino acid sequences. In such cases, it is difficult to estimate a full substitution matrix for this site. In our approach, we essentially pool residues that are located in the same region together and assume that they experience similar evolutionary pressure.

Ultimately, the effectiveness of incorporating structural information in phylogenetic analysis and evolutionary models can be tested on the criterion whether it in turn helps to understand the organization principles of protein structures and their biochemical functions. As indicated by successful applications in protein function prediction reported here, structure-based phylogenetic analysis provides a powerful framework for studying significant problems in structural biology.

Our method benefits from existing computational techniques. Without the mathematical theory that formalizes our intuitive notion of protein shapes such as pockets and voids (Edelsbrunner, Facello, and Liang 1998), efficient algorithms for their computation (Edelsbrunner, Facello, and Liang 1998; Liang, Edelsbrunner, and Woodward 1998), strategies for shape similarity assessment (Binkowski, Adamian, and Liang 2003), as well as demonstrated success of these computational techniques (Liang, Edelsbrunner, and Woodwart 1998; Binkowski, Adamian, and Liang 2003; Li, Hu, and Liang 2003;

Li and Liang 2005), the novel category of functionally important surface pockets would not be possible.

There are, however, some limitations in our method. If the number of homologous sequences is too few (<10) or the length of the functionally important binding pocket is too short (<8 residues), there will not be enough data for parameter estimation. Another limitation of our study is the assumption that all sites in a protein evolve according to the same rate matrix along all branches of the phylogenetic tree. Although simulation studies and applications indicate that the estimated rates are sufficiently accurate for the purpose of detecting functionally related protein surfaces, this assumption may not be realistic for studying detailed evolutionary history and mechanisms for a specific protein (Yang 1993, 1994*b*; Huelsenbeck and Nielsen 1999; Felsenstein 2001).

Our simulation study is simple and cannot provide a full picture of the estimation errors under different biological conditions. The focus of our simulation study is to assess how estimation error is affected by the length of a functional pocket. In our method, the proper and accurate construction of a high quality phylogenetic tree is essential. We find it important to carefully select amino acid sequences to ensure quality multiple sequence alignments, where few gaps are introduced and proteins of different divergence are well represented. In our practice, we find that the maximum likelihood estimator of Molphy works well with amino acid sequences for constructing phylogenetic trees. The effects of the assumption that the input phylogenetic tree is optimal, as well as the effects of different input branch lengths on the accuracy of estimation, needs further detailed studies. Our preliminary results suggest that the estimated scoring matrices for protein functional sites and database search results are insensitive to small perturbations in the phylogenetic tree and the branch lengths. For instance, in a database search of alpha amylase, we are able to use different surface templates, each from a different protein structure with its own slightly different phylogenetic tree and branch lengths. Our results show that the sets of functionally related proteins are nearly identical (data not shown).

Furthermore, the choice of a prior is an important and complex issue in Bayesian statistics. We assume that the likelihood function dominates and the information from the prior is limited. More detailed study is needed for a clear understanding of the influence of the choice of prior.

In summary, we have extended existing continuous time Markov models of residue substitution from that of codon-codon replacement to a model of residue-residue replacement. We have also developed a novel structural category of local surface regions that is well suited for studying the evolution of protein functions. We have implemented an effective Bayesian Monte Carlo method that can successfully estimate the substitution rates of residues in small local structural regions in proteins. In addition, we have developed a database search method using scoring matrices derived from estimated residue substitution rates. Our results in solving the fundamental problem of inferring protein functions from protein structures show very encouraging results. There are other novel technical developments. For example, we find it necessary to develop an efficient move set for rapid mixing in Monte Carlo estimation of substitution rates. We have also explored how reliability of estimated substitution rates depends on the size of the local region. As indicated by the successful applications reported here, we believe that phylogenetic analysis of protein evolution provides powerful tools for the important bioinformatic task of protein function prediction.

Acknowledgments

We thank Andrew Binkowski for help in pvSoar search, Ronald Jackups, Jr for proofreading the manuscript, Rong Chen, Susan Holmes, Art Owen, and Simon Whelan for rewarding discussions. We also thank Jeffrey Thorne and an anonymous referee for insightful suggestions. This work is supported by grants from the National Science Foundation (CAREER DBI0133856), the National Institutes of Health (GM68958), and the Office of Naval Research (N000140310329).

Literature Cited

Adachi, J., and M. Hasegawa. 1996. MOLPHY, version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. Inst. Stat. Math. Tokyo **28**:1–150.

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

Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Bairoch, A. 1993. The ENZYME data bank. Nucleic Acids Res. **21**:3155–3156.

Binkowski, T. A., L. Adamian, and J. Liang. 2003. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J. Mol. Biol. **332**:505–526.

Binkowski, T. A., P. Freeman, and J. Liang. 2004. pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res. **32**:W555–W558.

Binkowski, T. A., S. Naghibzadeh, and J. Liang. 2003. CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res. **31**:3352–3355.

Bustamante, C., J. Townsend, and D. Hartl. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. Mol. Biol. Evol. **17**:301–308.

Chothia, C., and A. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. EMBO. J. **5**: 823–826.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. Atlas of protein sequence and structure national biomedical research foundation. National Biomedical Research Foundation: Washington, D.C.

Dill, K. 1990. Dominant forces in protein folding. Biochemistry **29**:7133–7155.

Dill, K., S. Bromberg, K. Yue, K. Fiebig, D. Yee, P. Thomas, and H. Chan. 1995. Principles of protein folding—a perspective from simple exact models. Protein Sci. **4**:561–602.

Edelsbrunner, H., M. Facello, P. Fu, and J. Liang. 1995. Measuring proteins and voids in proteins. Pp. 256–264 *in* Proceedings of the 28th Annual Hawaii International Conference on System Sciences, Vol. 5. IEEE Computer Society Press, Los Alamitos, Calif.

Edelsbrunner, H., M. Facello, and J. Liang. 1998. On the definition and the construction of pockets in macromolecules. Discrete Appl. Math. **88**:83–102.

Edelsbrunner, H., and E. Mücke. 1994. Three-dimensional alpha shapes. ACM Trans. Graph. **13**:43–72.

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.

Felsenstein, J., and G. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

Goldman, N., J. Thorne, and D. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263**:196–208.

———. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**: 445–458.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

Govindarajan, S., and R. Goldstein. 1997. Evolution of model proteins on a foldability landscape. Proteins **29**:461–466.

Grimmett, G. R., and D. R. Stizaker. 2001. Probability and random processes. Oxford University Press, New York.

Gu, J., and X. Gu. 2003. Natural history and functional divergence of protein tyrosine kinases. Gene **317**:49–57.

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

Huelsenbeck, J., and J. Bollback. 2001. Empirical and hierarchical Bayesian estimation of ancestral states. Syst. Biol. **50**: 351–366.

Huelsenbeck, J., J. Bollback, and A. Levine. 2002. Inferring the root of a phylogenetic tree. Syst. Biol. **51**:32–43.

Huelsenbeck, J., B. Larget, R. Miller, and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. **51**:673–688.

Huelsenbeck, J., and R. Nielsen. 1999. Variation in the pattern of nucleotide substitution across sites. J. Mol. Evol. **48**:86–93.

Huelsenbeck, J., B. Rannala, and B. Larget. 2000. A Bayesian framework for the analysis of cospeciation. Evolution Int. J. Org. Evolution **54**:352–364.

Huelsenbeck, J., B. Rannala, and Z. Yang. 1997. Statistical tests of host-parasite cospeciation. Evolution **52**:410–419.

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

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.

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

Koshi, J., and R. Goldstein. 1996. Correlating structure-dependent mutation matrices with physical-chemical properties. Pac. Symp. Biocomput. 488–499.

———. 1997. Mutation matrices and physical-chemical properties: correlations and implications. Proteins **27**:336–344.

Lecomte, J., D. Vuletich, and A. Lesk. 2005. Structural divergence and distant relationships in proteins: evolution of the globins. Curr. Opin. Struct. Biol. **15**:290–301.

Lesk, A., and C. Chothia. 1982. Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. **160**:325–342.

Li, S., D. Pearl, and H. Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. J. Am. Stat. Assoc. **95**: 493–508.

Li, X., C. Hu, and J. Liang. 2003. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins **53**:792–805.

Li, X., and J. Liang. 2005. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins **60**:46–65.

Liang, J., and K. Dill. 2001. Are proteins well-packed? Biophys. J. **81**:751–766.

Liang, J., H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam. 1998. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. Proteins **33**:1–17.

Liang, J., H. Edelsbrunner, and C. Woodward. 1998. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. **7**:1884–1897.

Liò, P., and N. Goldman. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

———. 1999. Using protein structural information in evolutionary inference: transmembrane proteins. Mol. Biol. Evol. **16**:1696–1710.

Liò, P., N. Goldman, J. Thorne, and D. Jones. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. Bioinformatics **14**:726–733.

Mau, B., M. Newton, and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics **55**:1–12.

Mayrose, I., D. Graur, N. Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. **21**:1781–1791.

Mitchison, G. 1999. A probabilistic treatment of phylogeny and sequence alignment. J. Mol. Evol. **49**:11–22.

Parisi, G., and J. Echave. 2001. Structural constraints and emergence of sequence patterns in protein evolution. Mol. Biol. Evol. **18**:750–756.

———. 2005. Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. Gene **345**:45–53.

Pearson, W. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183**:63–98.

———. 1991. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics **11**:635–650.

———. 1998. Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. **276**:71–84.

Robert, C. P., and G. Casella. 2004. Monte Carlo statistical methods. Springer-Verlag Inc., New York.

Robinson, D., D. Jones, H. Kishino, N. Goldman, and J. Thorne. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. **20**:1692–1704.

Rost, B. 2002. Enzyme function less conserved than anticipated. J. Mol. Biol. **318**:595–608.

Sanishvili, R., A. F. Yahunin, R. A. Laskowski et al. (12 co-authors). 2003. Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from Escherichia coli. J. Biol. Chem. **278**:26039–26045.

Siepel, A., and D. Haussler. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. **21**:468–488.

Stitziel, N., Y. Tseng, D. Pervouchine, D. Goddeau, S. Kasif, and J. Liang. 2003. Structural location of disease-associated single-nucleotide polymorphisms. J. Mol. Biol. **327**:1021–1030.

Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: 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.

Thorne, J., N. Goldman, and D. Jones. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

Thorne, J., H. Kishino, and I. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. **15**:1647–1657.

Tian, W., and J. Skolnick. 2003. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. **333**:863–882.

Tourasse, N., and W. Li. 2000. Selective constraints, amino acid composition, and the rate of protein evolution. Mol. Biol. Evol. **17**:656–664.

Tseng, Y. Y., and J. Liang. 2004. Are residues in a protein folding nucleus evolutionarily conserved? J. Mol. Biol. **335**:869–880.

Vogel, C., M. Bashton, N. Kerrison, C. Chothia, and S. Teichmann. 2004. Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. **14**:208–216.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. **18**: 91–699.

Whelan, S., P. Liò, 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*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

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

———. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yang, Z., R. Nielsen, and M. Hasegawa. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15**:1600–1611.

Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol. Biol. Evol. **14**:717–724.

Jianzhi Zhang, Associate Editor

Accepted October 19, 2005