Estimating evolutionary rate of local protein binding surfaces: a Bayesian Monte Carlo approach

Yan Yuan Tseng and Jie Liang*
Dept of Bioengineering
University of Illinois at Chicago
Chicago, IL 60607, USA
{ytseng3,jliang}@uic.edu

***Abstract*— To infer protein function by matching local surface patterns, an effective scoring matrix for evaluating surface similarity is critical. In this study, we develop an evolution model of binding surfaces using a continuous time Markov process. We develop a Bayesian Markov chain Monte Carlo method to estimate the substitution rates of amino acid residues with specialized move sets. We then develop scoring matrices of residue similarity specific to a functional site and show how they can be used to identify similar binding surfaces, and how such information can be used for predicting biological roles of proteins. Our method is especially effective in extracting evolutionary information from the phylogeny of sequences homologous to a protein structure, all of which may be of unknown functions.**

I. INTRODUCTION

Predicting protein function is a challenging task, as functional annotation cannot be transferred reliably based on global sequence or structure similarity [1–3]. Because protein carries out its biological roles by interacting with other molecules, binding surfaces on protein structures play important roles in determining protein functions. A promising approach therefore is to examine candidate functional surface region on protein structures and identify similar local spatial motifs on other protein structures that are functionally related [4–9]. This approach allows the detection of remote functional relationship, and often does not require global similarity between the protein backbones. An example of this approach is that of reference [7], which is based on matching computed surface Several novel functional relationship between proteins of different fold were uncovered using this method (*e.g.*, HIV-1 protease and HSP-70) [7].

The success of such methods hinges upon the use of a scoring matrix. The evolutionary history of protein provides essential information for understanding its biological functions and how these functions emerge and evolve. Several empirical residue scoring matrices have been developed based on evolutionary history of proteins [10, 11]. A deficiency of all these scoring matrices is the neglect of the distinction between regions on proteins that directly participate in biological function through binding, *v.s.* other regions of the protein which may be important for protein stability. Functional region of proteins are likely to experience very different selection pressure than the rest of the proteins.

*Corresponding author.

In this study, we develop a method for obtaining residue scoring matrix by estimating residue substitution rates at functional surface. Our evolution model is based on a continuous time Markov process and incorporates explicit information contained in a phylogenetic tree. We estimates Bayesian posterior mean values of the instantaneous rates of residue substitution using the technique of Markov chain Monte Carlo. These rates are then used to construct a series of specialized and surface specific scoring matrices for detecting similar surface patterns that are functionally relevant on other protein structures. We illustrates our method using simulated sequences, as well as proteins with known functions. We then show how this method can help to infer the biochemical roles of a protein structure of unknown function solved by structural genomics project.

II. MODEL AND METHODS

a) Continuous time Markov process for residue substitution: For a given phylogenetic tree, we use a reversible continuous time Markov process as our evolutionary model [12, 13]. This model has several advantages over empirical methods. For example, empirically constructed matrices such as PAM and BLOSUM have implicit parameters whose values were determined from pre-computed analysis of large quantities of data, while the information about the particular protein of interest has limited or no influence. In contrast, Markovian evolutionary models are parametric models and do not have pre-specified parameter values. These values are all estimated from data specific to the protein of interest [14]. Empirical position specific weight matrix such as the ones generated by PSI-BLAST takes no consideration of the phylogeny of the proteins, and can be very biased if the sequences are unevenly distributed along a subset of branches of the tree. In addition, previous works showed that the effects of secondary structure and solvent accessibility are important for local amino acid replacement on protein evolution, and such effects can be captured by a Markovian evolutionary, while it is difficult for empirical methods to take these effects into account [15–17].

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. The divergence time represents expected number of changes between

sequences which are nodes in a phylogenetic tree. 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:

$$\boldsymbol{Q} = \{q_{ij}\} = \begin{pmatrix} - & q_{1,2} & \cdots & q_{1,20} \\ q_{1,2} & - & \cdots & q_{2,20} \\ & & \ddots & \\ q_{1,20} & q_{2,20} & \cdots & - \end{pmatrix},$$

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 [18]:

$$\boldsymbol{P}(t) = \{p_{ij}(t)\} = \exp(\boldsymbol{Q} \cdot t) = \boldsymbol{U} \exp(\boldsymbol{\Lambda} t) \boldsymbol{U}^{-1},$$

where $p_{ij}(t)$ represents the probability that a residue of type i will mutate into a residue of type j after time t. $\boldsymbol{U}, \boldsymbol{U}^{-1}$, and $\boldsymbol{\Lambda}$ are right eigenvectors, left eigenvectors, and diagonal matrix formed by the eigenvalues of $\boldsymbol{Q}$ sorted in descending-order, respectively. In practice, to ensure that the nonsymmetric rate matrix Q is diagonalizable for easy computation of $\boldsymbol{P}(t)$, we follow reference [19] and insists 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 on the protein functional surface, and $\boldsymbol{S}$ is a symmetric matrix whose entries need to be estimated. Because symmetric $\boldsymbol{S}$ is diagonalizable as $\boldsymbol{S} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T$, the matrix $\boldsymbol{Q} = \boldsymbol{S} \cdot \boldsymbol{D} = \boldsymbol{D}^{1/2} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^T \boldsymbol{D}^{-1/2}$ is also diagonalizable. Since the $\boldsymbol{S}$ matrix is normalized, there are $210 - 20 - 1 = 189$ unknown parameters.

b) Bayesian estimation of instantaneous rates: Our goal is to estimate the values of the $\boldsymbol{Q}$ matrix. Continuous time Markov model for residue substitutions has been implemented in several studies using maximum likelihood estimator [12, 19] and has found applications in protein folding studies [20]. 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. For a multiple sequence alignment $\mathcal{S}$ and a given phylogenetic tree $\boldsymbol{T}$), we describes instantaneous substitution rate $\boldsymbol{Q} = \{q_{ij}\}$ by a posterior distribution $\pi(\boldsymbol{Q}|\mathcal{S}, \boldsymbol{T})$, which summarizes information available on the rates $\boldsymbol{Q} = \{q_{ij}\}$ and information brought by the observations $\mathcal{S}$ and $\boldsymbol{T}$. After integrating the prior information $\pi(\boldsymbol{Q})$ and the likelihood function $P(\mathcal{S}|\boldsymbol{T}, \boldsymbol{Q})$ (see Appendix) and assuming a given phylogenetic tree $\boldsymbol{T}$, the posterior distribution $\pi(\boldsymbol{Q}|\mathcal{S}, \boldsymbol{T})$ can be estimated up to a constant as:

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

Our goal is to estimate the posterior mean of rates in $\boldsymbol{Q}$ as a summarizing index:

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

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

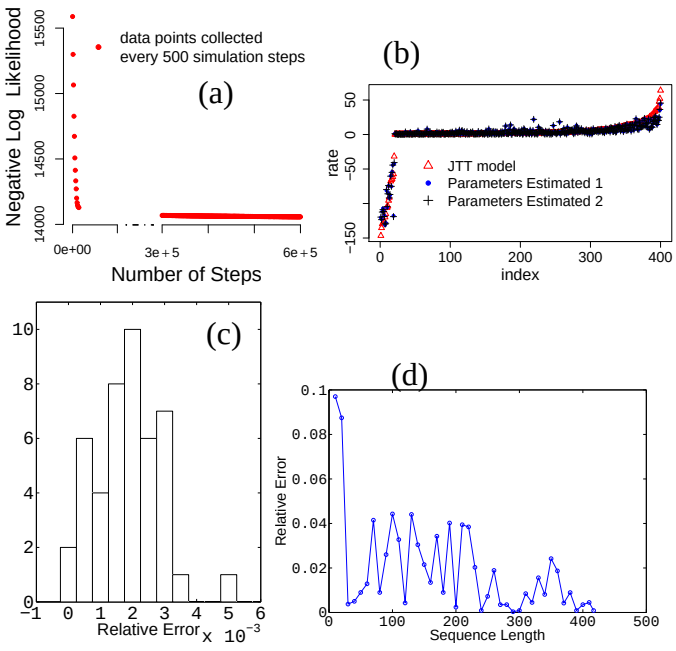

Fig. 1. Estimating substitution rates. (a) The Markov chain converges. (b) Rates estimated in two simulations are all similar to the true rates. In the first simulation, $q_{i,j} = 0.1$ for all entries. In the second, the 189 initial $q_{i,j}$ values are sorted numerically by index i then by index j, and the values are assigned from 0.1 with an increment of 0.01 for the next entry. (c) The relative errors in 50 estimations with random initial values are all $< 5\%$. (d) The relative error remains $< 5\%$ if the length of sequence is ≥ 20.

c) Rate matrix $\boldsymbol{Q}$ and residue similarity score: To derive residue scoring matrix from the evolutionary model for database search, we derive residue similarity scores [21] $b_{ij}(t)$ between residues i and j at different evolutionary time t from the rate matrix $\boldsymbol{Q}$:

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

where $m_{ij}(t)$ is the joint probability of observing both residue type i and j at the two nodes separated by time t, and λ is a scalar. $b_{ij}(t)$satisfies $\sum \pi_i \pi_j e^{\lambda b_{ij}} = 1$, because of the property of $\sum_{ij} \pi_i p_{ij}(t) = 1$ for Markov matrix.

III. Results

d) Rates estimation: simulation studies: To validate our method, we first carry out a simulation study to test how accurate estimated residue substitution rates are. We generate a set of artificial sequences based on a known evolutionary model, with known substitution rates. We ask the question whether our method can recover the original rates reasonably well, and how many sequences and residues are needed so a good estimation can be made. For this purpose, we take the sequence of carboxypeptidase A2 precursor (SwissProt P48052, length 417), and generate 16 artificial sequences using the JTT evolutionary model, where the residue substitution rates are taken from reference [22].

Figure 1b shows the estimation results for two simulations. We started from two different sets of initial values of $\{q_{i,j}\}$. It is clear that both sets of the estimated rates $\{\widetilde{q_{i,j}}\}$ are very similar to the set of true values. To further quantitatively

assess how similar the estimated rates and the true rates are, we calculate the relative error Δe, defined as:

$$\Delta e \equiv |(||\boldsymbol{Q}||_F - ||\widetilde{\boldsymbol{Q}}||_F)/||\boldsymbol{Q}||_F|, \quad (1)$$

where $||\boldsymbol{Q}||_F$ denotes the Fröbenius norm of matrix $\boldsymbol{Q}$: $||\boldsymbol{Q}||_F = (\sum_{i=1}^{189} |q_{i,j}|^2)^{1/2}$. Figure 1c shows that the relative errors from 50 simulations are all very small ($\Delta e \approx 10^2$). Each of these 50 simulations had a different set of random initial values of $\{q_{i,j}\}$ drawn from a uniform distribution of $\mathcal{U}(0,1)$.

The functional region of a protein contains only a small number of residues, which varies depending on the size of the binding site. It is important to assess how the accuracy of rate estimation is affected by the size of the binding site. Starting from the N-termini of these sequences, we take a substring from each sequence, with the length increasing from 10 to 417, at an increment of 10 residues. Each simulation of a different length was started from a random set of initial values drawn from $\mathcal{U}(0,1)$. The duration of these simulations is longer than the $70,000$ time steps required for a typical Markov chain to converge. Our results show that as long as the number of residues is ≥ 20, the relative error Δe of estimated parameters and true parameters will be less than 5% (Figure 1d).

e) Detecting functionally similar binding surfaces: An objective test of the effectiveness of our method is to see if we can discover related protein binding surfaces, namely, whether we can discover protein structures that have similar binding surfaces and carrying out similar biological functions.

We use alpha-amylases as our test system. Alpha-amylase (Enzyme Classification number $E.C.3.2.1.1$) acts on starch, glycogen and related polysaccharides and oligosaccharides. Detecting functionally related alpha amylase is a challenging task, as many of them have overall very low sequence identities ($< 25\%$) to the query protein template. Below a sequence identity of 70%, it becomes difficult to make functional inference for proteins based on sequence alignment [3].

Given a template structure of binding surface of an alpha amylase (`1bag`, pdb), we wish to find out how many structures of proteins that are of the same E.C. number of all four digits can be identified. These protein structures all carry out the same or related reactions.

To construct the evolutionary model, we use sequence alignment tools to find sequences homologous to that of `1bag` [23]. After removing redundant sequences, 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 (Figure 2a). We use the maximum-likelihood method as implemented in MORPHY for tree construction [24].

We then calculate a similarity scoring matrix 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 characterizes the residue

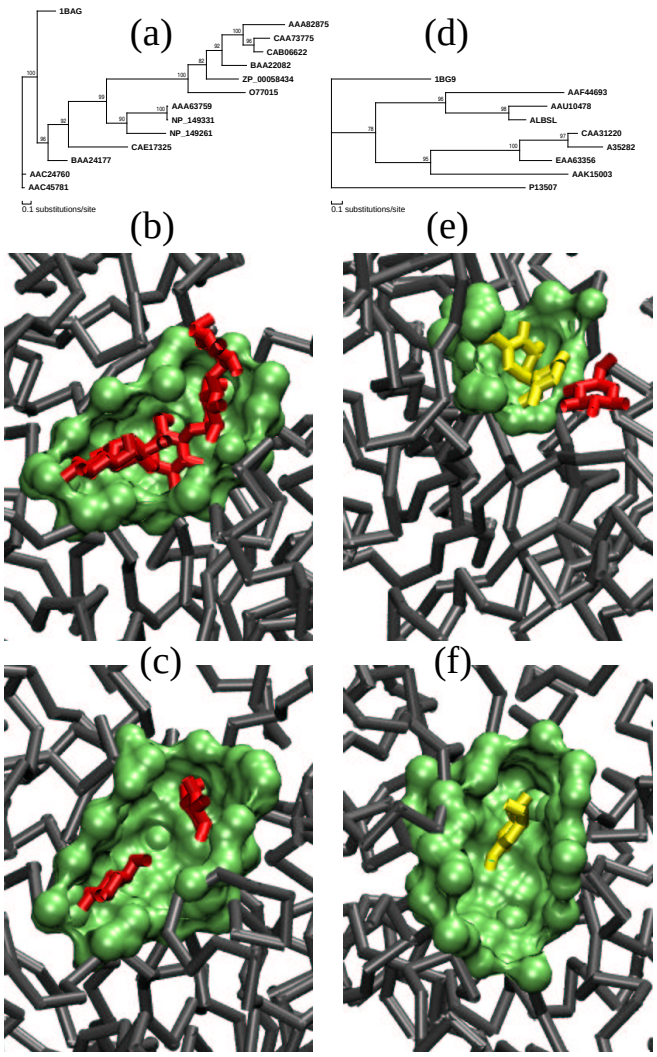

Fig. 2. Validation of function prediction of alpha amylases. (a) The phylogenetic tree for PDB structure `1bag` from *B. subtilis*. (b) The binding pocket of alpha amylase on `1bag`. (c) A matched binding surface on a different protein structure (`1b2y` from human, full sequence identity 22%) obtained by querying with `1bag` (d) The phylogenetic tree for `1bg9` from *H. 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`

substitution pattern at a different time separation, ranging from 1 time unit to 300 time unit. We use Smith-Waterman algorithm with each of the 300 scoring matrices to align sequence patterns of candidate binding surfaces from >2 million protein surface pockets contained in the PVSOAR database [25]. Surfaces similar to the query binding pocket identified are subject 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 coordinate RMSD measure or orientational RMSD measure are excluded [7].

A total of 58 PDB structures are identified to have similar binding surfaces as that of `1bag`, and hence are predicted as amylase. All of them have the same $E.C.3.2.1.1$ label as that of `1bag`. Similarly, we found 48 PDB structures of

$E.C.3.2.1.1$ label when using the functional site of `1bg9`. The union of the results from these two searches gives 69 PDB structures with $E.C.3.2.1.1$ labels. Examples of matched protein surfaces are shown in Figure 2

The Enzyme Structures Database (ESD) (`www.ebi.ac.uk/thornton-srv`) collects protein structures for enzymes contained in the ENZYME databank [26]. There are 75 PDB entries with enzyme class label $E.C.3.2.1.1$ in ESD (version Oct 2004). Out of the 75 structures, our method discovered 69 PDB structures using `1bag` and `1bg9` as queries. We also compare our results with those obtained from database search using sequence alignment tools. Using SSEARCH in FASTA with default set-up of BLOSUM50 matrix, only 32 structures are identified as alpha amylase (see Table 2 in [7]). When using PSI-BLAST with the E-value threshold of 10^{-3}, BLOSUM62, default parameters, the NR database for generating position-specific weight matrix, and < 10 iterations, only 41 structures among the 75 known structures of alpha-amylase are found.

IV. DISCUSSION

We have developed a method for estimating residue substitution rates for functionally important binding pocket. This approach allows effective modeling of evolution of protein function based on their binding surfaces. Our work is also the first study using Bayesian estimation and Markov chain Monte Carlo to infer amino acid residue substitution rates. We show that the estimated substitution rates can be used to construct scoring matrix for effective database search of functionally similar binding surfaces, with better performance than other methods such as PSI-BLAST.

Our results suggest that binding surfaces on proteins often contain distinctive evolutionary information, and such information can be effectively extracted using the continuous time Markov model proposed here. Surface similarity search based on scoring matrix constructed using our method can lead to more sensitive and specific method for predicting protein function.

An advantage of our method is that sequences with unknown functions, (*e.g.*, hypothetical proteins obtained from genome sequencing efforts) become an important source of information about the evolutionary history of protein functional site. Sequences that are used to construct the phylogenetic tree can be all of unknown structures, or of unknown function. The majority of the sequences in the phylogenetic tree can be of hypothetical proteins, and they may provide critical information for predicting protein function.

V. ACKNOWLEDGMENT

We thank Dr. Andrew Binkowski for generous help in PVSOAR search, Drs. Rong Chen, Susan Holmes, Art Owen, and Simon Whelan for rewarding discussions, and Dr. Andzrej Joachimiak for suggesting BioH for our study. This work is supported by grants from NSF (CAREER DBI0133856), NIH (GM68958), and ONR (N000140310329).

REFERENCES

[1] C.A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: quantifying the relations bet ween protein sequence, structure and function through traditional and probabilistic score s. *J Mol Biol*, 297(1):233–49, 2000.

[2] A.E. Todd, C.A. Orengo, and J.M. Thornton. Evolution of function in protein superfamilies, from a structural perspec tive. *J Mol Biol*, 307(4):1113–43, 2001.

[3] B. Rost. Enzyme function less conserved than anticipated. *J Mol Biol*, 318(2):595–608, 2002.

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

[5] R. A. Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics*, 13:323–330, 1995.

[6] R. A. Laskowski, N. M. Luscombe, M. B. Swindells, and J. M. Thornton. Protein clefts in molecular recognition and function. *Protein Sci.*, 5:2438–2452, 1996.

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

[8] R. B. Russell. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.*, 279:1211–27, 1998.

[9] F. Glaser, T. Pupko, I. Paz, R.E. Bell, D. Shental, E. M artz, and N. Tal. Consurf: identification of functional regions in proteins by surface-mapp ing of phylogenetic information. *Bioinformatics*, 19(1):163–4, 2003.

[10] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. Atlas of protein sequence and structure. 5 suppl. 3:345, 1978.

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

[12] Z. H. Yang. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39:105–111, 1994.

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

[14] S. Whelan, P. Liò, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genet.*, 17(5):262–272, 2001.

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

[16] N. Goldman, J. L. Thorne, and D. T. Jones. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, 149:445–458, 1998.

[17] P. Liò and N. Goldman. Using protein structural information in evolutionary inference: transmembrane proteins. *Mol. Biol. Evol.*, 16(12):1696–1710, 1999.

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

[19] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, 18(5):691–699, 2001.

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

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

[22] D. T. Jones, W. R. Taylar, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.

[23] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[24] J. Adachi and M. Hasegawa. A computer program package for molecular phylogenetics. ver 2.3, 1996.

[25] T.A. Binkowski, P. Freeman, and J. Liang. pvsoar: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nuc. Aci. Res.*, 32:W555–558, 2004.

[26] A. Bairoch. The enzyme data bank. *Nucleic Acids Res*, 21(13):3155–6, 1993.