

Evolutionary model for predicting protein function by matching local surfaces: a Bayesian Monte Carlo approach

Yan Yuan Tseng and ¹ Jie Liang ²

Keywords: protein function, Markov chain Monte Carlo, substitution matrix, folding and evolution.

1 Introduction.

Inferring protein function is a challenging task, as global protein sequence and structure similarities are often unreliable for function prediction. Protein plays its role by interacting with other molecules, and local binding surfaces and their evolution histories contain direct functional information. To match local surfaces and to assess their similarity, scoring matrix such as PAM and BLOSUM are not suitable, because residues on protein functional surfaces experience very different selection pressure than residues in folding core. 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 functionally similar binding surfaces, and how such information can be integrated into a probabilistic model for predicting biological roles of enzyme structures. 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. We first validate our method by simulations and show that parameters of a known evolutionary model can be reliably recovered for surface patterns. We then test our method using alpha amylases, where family members often have $\leq 25\%$ sequence identities. We show our method can discover significant remote amylase proteins. Finally, we give an example of how to predict function of a protein structure with unknown function solved by a structural genomics project.

2 Results.

2.1 Evolutionary rates of binding sites are different

Residues on protein functional surface experience selection pressure that is different from functional surface. We compare the estimated substitution rate matrix of functional surface residues on alpha amylase with that of the remaining residues of the protein, the rest of surface residues, and the interior residues. In term of relative error by Fröbenius norm, we find that the rates differ by 28.95%, 25.25%, and 41.22%, respectively for the rest of protein, rest of surface and the interior of protein. In addition, it differs from the JTT rate matrix by 34.41%. It is clear that the selection pressures between functional site residues and other regions of the protein are different, and all are very different from the rates given by the JTT model [1]. This indicates that database search of functional surfaces will be far more effectively if we employ scoring matrix derived from estimated values of the binding surfaces.

¹Department of Bioengineering, University of Illinois at Chicago. E-mail: ytseng3@uic.edu

²Corresponding author. E-mail: jliang@uic.edu

2.2 Detecting and predicting functional pockets from a template binding surface.

We have developed a Bayesian Markov Chain Monte Carlo method for estimating residue substitution rates for functionally important binding pocket [2]. This approach allows effective modeling of evolution of protein function based on their binding surfaces. The 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. Surface similarity search based on scoring matrix constructed can lead to more sensitive and specific method for predicting protein function. Two examples are shown in Figure 1.

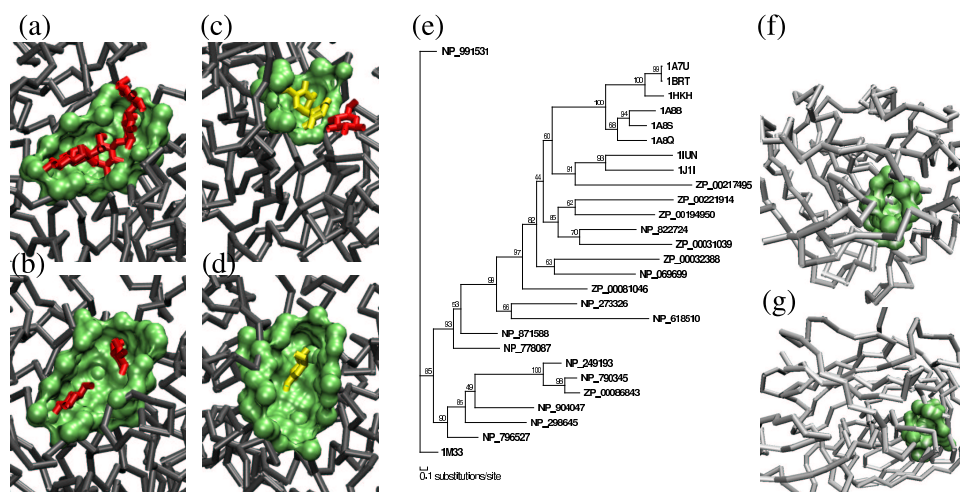


Figure 1: Validation of function prediction of alpha amylases, and predicting function of protein structure from structural genomics. (a) The binding pocket of alpha amylase on 1bag from *B. subtilis*. (b) A matched binding surface on a different protein structure (1b2y from human, full sequence identity 22%) obtained by querying with 1bag. (c) The binding pocket on alpha amylase 1bg9 from *H. vulgare*. (d) A matched binding surface on a different protein structure (1u2y from human, full sequence identity 23%) obtained by querying with 1bg9. (e) The phylogenetic tree of 28 sequences related to BioH, a protein from structural genomics with unknown function. Many of the homologous sequences are hypothetical genes. (f) The candidate binding pocket of BioH (1m33 [3] from structural genomics) and (g) a similar functional surface detected from proteinase A (2jxr, yeast, < 5% sequence identity) using scoring matrix derived based on estimated rates by Bayesian Monte Carlo with phylogenetic tree in (e).

References

- [1] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282, 1992.
- [2] 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.
- [3] Sanishvili, R., Yahunin, A. F., Laskowski, R. A., Evdokimova, E., Skarina, E., Doherty-Kirby, A., Lajoie, G. A., Thornton, J. M., Arrowsmith, C. H., Savchenko, A., Joachimiak, A. and Edwards, A. M. *J. Biol. Chem.* **278**(28), 26039 (2003).