Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Automated method for predicting enzyme functional surfaces and locating key residues with accuracy and specificity

Yan Yuan Tseng and Jie Liang

Abstract— **Locating functionally important protein surfaces and identifying the catalytic site residues are critical for studying enzyme functions. Here, we present methods for predicting and characterizing catalytic sites of enzymes at atomic level that is fold-independent. By extract atomic patterns of catalytic residues in surface pockets computed geometrically, we develop a library of atomic patterns on protein functional surfaces of *ca* 700 structures. Together with propensities of secondary structures and residue occurrence in active sites, we develop methods to identify functionally important surfaces on protein structures and to locate key residues. We discuss application of our methods to amylase, dioxygenase, deaminase, dehalogenase, and hydratase. A large scale cross-validated prediction study shows that our method is sensitive and specific.**

I. INTRODUCTION

Identifying protein residues that play functional roles is an important task. Proteins have a large number (100–1,000) of residues, but only a small fraction of them are directly involved in biochemical functions. These residues often are dispersed in primary sequence, but fold spatially together to form a binding or catalytic surface. A subset of them are key residues because they either directly participate in catalysis, or are important for substrate binding [8, 17].

Although a large number of protein structures in the Protein Data Bank (PDB) are annotated, *e.g.*, with an enzyme commission (E.C) number representing a specific chemical reaction, often such functional information is incomplete: the location of the binding surface is unknown, the identities of the key residues are incomplete, and there are well-known examples where the E.C. labels are misleading. As more protein structures are solved in the structural genomics project [6], a large number of structures have unknown functions. Identifying functionally important surfaces and locating key residues would provide important information for further characterizations [5, 10, 11].

In this study, we develop methods for identifying functional surface from a large set of precomputed surfaces. Our method is based on analysis of bias of functionally important key residues in composition, in secondary structure, and in atomic patterns. We formulate a probabilistic model for predicting whether a residue located in a surface pocket is functionally important. This model is further used to identify the surface precomputed as a pocket that is most likely to be important for biological functions. This paper is organized as follows: we first describe our methods and the data set , we then report results of functional site prediction using several enzymes as example, followed by a large scale cross-validation study.

This work is supported by grants from NSF (CAREER DBI0133856), NIH (GM68958), and ONR (N000140310329).

Department of Bioengineering, SEO, MC-063 University of Illinois at Chicago 851 S. Morgan Street, Room 218 Chicago, IL 60607–7052, U.S.A. jliang@uic.edu

II. METHODS

A. Data set from PDB database

We found there are 13,877 protein structures among >30,000 structures in the PDB databank that are annotated as enzymes and have enzyme commission (E.C.) numbers. However, in many cases there is no information about where the active site is located on the structure and what residues are involved. We use geometric algorithm to compute surface pockets (including buried voids), which are stored in the CASTP database [4]. We are able to identify a set $\mathcal{A}$ of 3,275 proteins whose surface pockets contain one or more annotated residues. From these, we select a subset $\mathcal{B}$ of $\approx$ 700 structure after further cleaning up by verifying the annotations for each of the key residues, as well as requiring that experimentally measured B-factor exist. Altogether, we obtain a final set of $\approx$ 700 proteins structures, containing 3,007 annotated resides. We define a functional surface as a surface pocket containing one or more of these annotated key residue(s).

Fig. 1a shows the size distribution of functional surface pockets in set $\mathcal{A}$. The mean size is 35 residues. Fig. 1b shows that the amino acid residue composition of these functional surfaces is very different from the one of full backbone protein sequences [13].

B. Characteristics of enzyme binding surfaces

An important property of the functional surface is its size, *e.g.*, measured in the number of residues it contains (Fig. 1a). We also calculated the ratio of size of functional surface over the total length of the full protein. We found that in general, about 10%-30% of all residues on a protein are involved in enzyme function (Fig. 1c), namely, proteins use 10-30% of their residues to form local binding surfaces for catalysis. Another informative attribute of enzyme functional site is the molecular volumes (Fig. 1d). Based on these observations, we select those precomputed surface pockets containing 10-30% of the residues as candidates for prediction of functional surface.

Enzyme functional surfaces have characteristic usage of amino acid residues. Fig. 2 shows the distribution of the 20 amino acids in annotated residues on the 3,275 surface pockets from set $\mathcal{A}$. Similar to previous studies [2, 14, 15], we found that His, Asp, Glu, Ser and Cys account for more

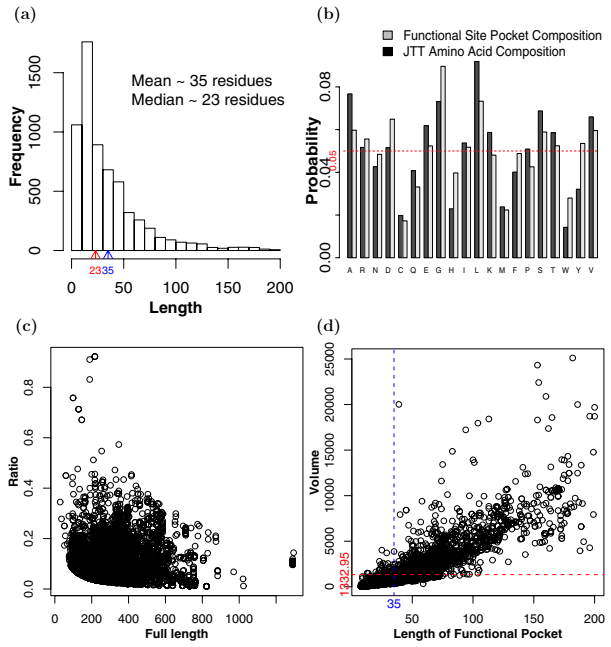


Fig. 1: The length distribution and unique residue composition of functional surfaces for 3,275 proteins with known key residues. (a) Functional surfaces usually consist of 8–200 residues, with the mean value of 35 residues. (b) The amino acid residue composition of functional surfaces on these proteins is different from the composition of sequences used to construct the JTT model [13]. (c) The ratio is defined as $\frac{length(pocket)}{length(backbone)}$. The ratio ranges from 0.1 to 0.3. Proteins commonly have size from 100 to 450 residues. They are most likely to have functional pockets of which length is from 10 to 80 residues. (d) The mean molecular volume of functional pockets is 1,332.95 Å^3. In general, the molecular volume of a functional pocket is less than 5,000 Å^3 and it's length is less than 80 residues.

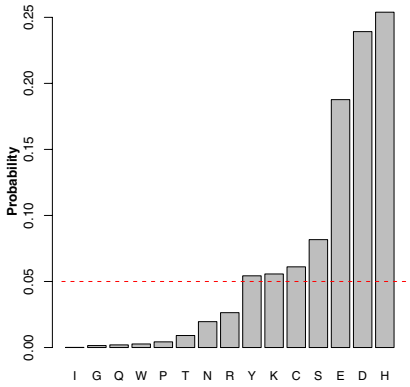


Fig. 2: Active site residues are mapped to functional pockets and based on annotation in SWISSPROT and PDB (17,930 pdb entries). His, Asp, Glu, Ser and Cys account for more than 80% of active site residues of functional pockets. In the contrast, Ala, Pro, Val, Leu and Met are completely missed because they are hydrophobic attracted in the core of proteins.

than 80% of active site residues in functional pockets. On the other hand, nonpolar residues (*e.g.*, Val, Leu, Pro) are absent. These hydrophobic resides are enriched in protein core for maintaining protein stability, but play little roles in enzyme activities.

For each annotated residue, we obtain the *atomic pattern* by listing the atoms in a consistent order from one specific residue that are exposed on the surface wall of the pocket. The secondary structure environment (β-sheet s, α-helix h, and coil c) of a residue provides useful information, as backbone N and O atoms form H-bonds in α-helix and β-strand and therefore are expected to be less likely to form H-bond involved in the interaction with substrates. We augment the atomic pattern by including information of the secondary structure environment of this residue: h for helix, s for β-sheet, and c for coil. For example, the Gln208 residue in the alpha-amylase structure `1bag` (see Fig. 3b) has the following atomic pattern:

```
GLN208 CD:NE2:O:OE1:c.
```

From annotated 3,007 key residues on proteins of set $\mathcal{B}$, we obtain 1,031 atomic patterns. This is used to construct the probability $\pi(a|\mathcal{K})$ of the occurrence of an atomic pattern a in key residues from protein set $\mathcal{B}$.

Integrated predictor of functionally important residues. For a residue i located in a surface pocket, because the identity r_i of this residue, its secondary structure environment s_i, and its atomic pattern a_i all provide useful discriminating information for identifying key residues important for enzyme functions, we use the following method to integrate these parameters and calculate the *key residue probability* $\mathbb{P}(i \in \mathcal{K})$ for the i-the residue to be from the set $\mathcal{K}$ of key residues:

$$\begin{aligned}\mathbb{P}(s_i, r_i, a_i, i \in \mathcal{K}) &= \pi(s_i, r_i, a_i | i \in \mathcal{K}) \cdot \pi(i \in \mathcal{K}) \\ &\approx \pi(s_i | i \in \mathcal{K}) \cdot \pi(r_i | i \in \mathcal{K}) \cdot \\ &\quad \pi(a_i | i \in \mathcal{K}) \cdot \pi(i \in \mathcal{K})\end{aligned} \quad (1)$$

where $\pi(r_i|i \in \mathcal{K})$ is the probability of a key residue to be amino acids type r_i, $\pi(s_i|i \in \mathcal{K})$ the probability of a key residue to be of the secondary structure type s_i, $\pi(a_i|i \in \mathcal{K})$ the probability of a key residue to be of the atomic pattern a_i, respectively.

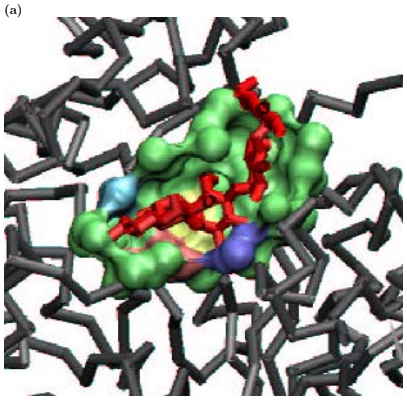
Identifying functional surface. A functional surface is where protein performs its biological roles. To identify key resides involved in biochemical reactions, a prerequisite is that the functional surface is identified correctly.

We identify the functional surface pocket p from a set of computed pockets $\mathcal{P}$ on a protein structure. We compute the summed probability

$$SP = \sum_{i \in P} \mathbb{P}(s_i, r_i, a_i, i \in \mathcal{K}).$$

If $SP \geq 10^{-3}$, we declare that pocket p is a functional surface.

(a)



(b)

```
ASP176 CG:OD1:OD2:c
HIS180 CD2:NE2:c
GLN208 CD:NE2:O:OE1:c
ASP269 CG:OD1:OD2:h
```

Fig. 3: Prediction of binding surface and predicted key residues of alpha-amylase. (a) The pocket (green, CASTP ID=60) predicted to be the functional surface interact with the substrate glucose (red). This functional surface contains 18 residues. Four of them are predicted to be functionally important: ASP176 (yellow), HIS180 (cyan), GLN208 (pink) and ASP269 (blue). (b) The four predicted key residues contains several high propensity atomic patterns from our library of 1,031 functional atom patterns. Secondary structure (β sheet s, helix h , and coil c) information used to compute $\pi_s(i)$ is also listed.

III. RESULTS

A. An example

We use alpha-amylases `1bag` as an example. Alpha-amylase ($\approx$ 420 residues) acts on starch, glycogen and related polysaccharides and oligosaccharides. Our task is to locate which pocket is the functional surface among the 60 pockets and further identify the key residues involved in the enzymatic reaction. Our only input is a structure of of the protein.

We first exhaustively compute all of the pockets (including voids) on this protein structure [4, 16]. We then compute the key residue probability $\mathbb{P}(i \in \mathcal{K})$ for each residues i in a pocket.

We first predict the functional surface. We rank the 60 surface pockets by summed probability SP. The largest pocket (CastP ID=60) contains the largest number (7) of predicted key residues, and has the largest $SP = 1.31 \times 10^{-3}$ value. It is predicted to be the functional surface pocket involved in enzyme reaction. This prediction is correct based on annotation and biochemical literature.

We then predict likely key residues important for enzymatic function. After we collect pocket surfaces with SP greater than a threshold $\theta = 10^{-3}$. For this protein structure, pocket 60 is the only one satisfying this condition. It contains 18 residues (Fig. 3). We found that there are 4 residues whose $\mathbb{P}(i \in \mathcal{K})$ values are significantly higher than the rest of 14 residues, and are predicted as key residues. These residues are identical from those reported in literature [7, 9].

TABLE I: Detecting functional surfaces and locating key residues. Predicted results and the true answers as recorded in the human curated SFLD [18] database are listed. Notice that some residues annotated with iron binding are not considered as catalytic ones and, therefore removed.

PDB structure	**Predicted surface ID/length**	**Predicted key residues**	SFLD (experimental data)
1bag[a]	**60/18**	**D176,H180**	D176,H180
EC 3.2.1.1		**Q208,D269**	Q208,D269
1qq5A	**71/11**	**D8,R39**	D8,R39
EC 3.8.1.2		**K147,N173**	S114,K147
		N175	S171,N173
			D176
1add–	**43/29**	**E217,H238**	E217,H238
EC 3.5.4.4		**D296,D66**	D296,D295
1ebgA	**134/18**	**S39,E211**	S39,E211
EC 4.2.1.11		**D246,D296**	D246,E295
		D320,K345	D320,K345
		H159	
1kmyA	**29/34**	**H195**	H195
EC 1.13.11.39			

[a]The active sites residues are verified by results reported in the literature [9].

B. Large-scale prediction of functional surfaces

Locating the functional surface is the most important task in studying enzyme mechanism, as the correct surface will guide further analysis of binding and catalysis mechanism, and will facilitate the correct prediction of the key residues on protein functional surfaces [12]. To evaluate the performance of our method in identifying functional surfaces, we use 10-fold cross validation on the $\mathcal{B}$ dataset. We remove 10% of the structures to test the performance of the prediction method, which is derived from analysis of the rest 90% of the data. The average of performace of predicting functional surfaces of proteins is 91.2% accuracy.

C. Prediction of key residues on protein functional surfaces

We compare our prediction results with enzymes contained in the Structure-Function Linkage Database (SFLD) [18], which links related sequences and structures of enzymes to their chemical reactions, with detailed annotation of enzyme active site residues. We select the four enzyme families that each has 8 or more structures. These are: 2,3-dihydroxybiphenyl dioxygenase (E.C. 1.13.11.39), adenosine deaminase (E.C. 3.5.4.4), 2-haloacid dehalogenase (E.C. 3.8.1.2), and phosphopyruvate hydratase (E.C. 4.2.1.11). We take a random template structure from each protein family, and apply our method to identify functional surfaces and then locating functionally important residues. As shown in Table I, we are able to accurately locate many functionally important residues.

Since our method successfully located functionally important residues, we believe the functional surfaces we predicted are also correct.

IV. CONCLUSIONS AND FUTURE WORKS

A. Conclusions

In this work, we have developed a method for identifying functional surfaces and for locating key resides. Our method is sequence and fold independent. We are able to identify systematically functional surfaces with $\geq 91.2\%$ accuracy. In the example of alpha-amylase, the key residues and their atomic patterns are identified, which fully agree with experimental data. Our work provides a fully automated method for locating functionally important surface and for identifying key residues. It can be used to study the mechanism of enzyme reaction, including interactions between residues and substrates. Its applications include drug design and engineered biochemical reactions.

B. Future Works

We plan to dynamically increase the size of the library of annotated functional surfaces, as more structures are deposited in the PDB databank. Additional annotations will be obtained by homology transfer when a surface is matched with another annotated surface satisfying stringent criterion (p-value $\leq 10^{-5}$ for cRMSD [3] distance of matched surfaces).

An important factor we plan to incorporate in our model is evolutionary conservation. Because residues in protein functional surface experience strong selection pressure [19], we expect this would further improve our method. Further consideration we plan to study is protein dynamics. Protein function often involves dynamic processes [1], and a crystal structure is only a snapshot conformation of a protein. The shape of the functional surface will change locally and may affect the shape of geometrically computed pockets. We expect that this problem will be alleviated as more structures are deposited and different functional conformations will be increasingly represented in the database. We will critically examine this issue and assess the robustness of current approach.

REFERENCES

[1] I. Bahar, A. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential." *Fold. Des.*, vol. 2, pp. 173–81, 1997.

[2] G. Bartlett, C. Porter, N. Borkakoti, and J. Thornton, "Analysis of catalytic residues in enzyme active sites." *J. Mol. Biol.*, vol. 324, pp. 105–21, 2002.

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

[4] T. A. Binkowski, S. Naghibzadeh, and J. Liang, "CASTp: Computed atlas of surface topography of proteins." *Nucleic Acids Res.*, vol. 31, pp. 3352–3355, 2003.

[5] T. Binkowski, A. Joachimiak, and J. Liang, "Protein surface analysis for function annotation in high-throughput structural genomics pipeline." *Protein. Sci.*, vol. 14, pp. 2972–81, 2005.

[6] J. Chandonia and S. Brenner, "The impact of structural genomics: expectations and outcomes." *Science*, vol. 311(5759), pp. 347–51, 2006.

[7] T. Collins, V. De, A. Hoyoux, S. Savvides, C. Gerday, B. Van, and G. Feller, "Study of the active site residues of a glycoside hydrolase family 8 xylanase." *J. Mol. Biol.*, vol. 354(2), pp. 425–35, 2005.

[8] S. Copley, W. Novak, and P. Babbitt, "Divergence of function in the thioredoxin fold suprafamily: evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor." *Biochemistry*, vol. 43, pp. 13 981–95, 2004.

[9] Z. Fujimoto, K. Takase, N. Doui, M. Momma, T. Matsumoto, and H. Mizuno, "Crystal structure of a catalytic-site mutant alpha-amylase from Bacillus subtilis complexed with maltopentaose." *J. Mol. Biol.*, vol. 277, pp. 393–407, 1998.

[10] R. George, R. Spriggs, G. Bartlett, A. Gutteridge, M. MacArthur, C. Porter, B. Lazikani, J. Thornton, and M. Swindells, "Effective function annotation through catalytic residue conservation." *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, pp. 12 299–304, 2005.

[11] F. Glaser, R. Morris, R. Najmanovich, R. Laskowski, and J. Thornton, "A method for localizing ligand binding pockets in protein structures." *Proteins*, vol. 62, pp. 479–88, 2006.

[12] N. Gold and R. Jackson, "Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships." *J. Mol. Biol.*, vol. 355, pp. 1112–24, 2006.

[13] D. T. Jones, W. R. Taylor, and J. M. Thornton, "The rapid generation of mutation data matrices from protein sequences." *CABIOS*, vol. 8, pp. 275–282, 1992.

[14] J. Kim, J. Mao, and M. Gunner, "Are acidic and basic groups in buried proteins predicted to be ionized?" *J. Mol. Biol.*, vol. 348, pp. 1283–98, 2005.

[15] R. Laskowski, J. Watson, and J. Thornton, "Protein function prediction using local 3D templates." *J. Mol. Biol.*, vol. 351, pp. 614–26, 2005.

[16] 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.*, vol. 7, pp. 1884–1897, 1998.

[17] E. Meng, B. Polacco, and P. Babbitt, "Superfamily active site templates." *Proteins*, vol. 55, pp. 962–76, 2004.

[18] S. Pegg, S. Brown, S. Ojha, J. Seffernick, E. Meng, J. Morris, P. Chang, C. Huang, T. Ferrin, and P. Babbitt, "Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database." *Biochemistry*, vol. 45, pp. 2545–55, 2006.

[19] Y. Tseng and J. Liang, "Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach." *Mol. Biol. Evol.*, vol. 23, pp. 421–36, 2006.