Annals of Biomedical Engineering, Vol. 35, No. 6, June 2007 (© *2007*) pp. 1037–1042
DOI: 10.1007/s10439-006-9241-2

Predicting Enzyme Functional Surfaces and Locating Key Residues Automatically from Structures

Yan Yuan Tseng and Jie Liang

Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, Chicago, IL 60607-7052, USA

(Received 20 September 2006; accepted 27 November 2006; published online 9 February 2007)

Abstract—Locating functionally important protein surfaces and identifying the catalytic site residues are critical for studying enzyme functions. Here, we present a method for predicting and characterizing catalytic sites of enzymes 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 a method 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. Our method can used to study enzyme function, drug design, and engineering novel biochemical function.

Keywords—Protein function, Enzyme function, Functional site, Key residues.

INTRODUCTION

Identifying protein residues that play functional roles is an important task. Proteins have a large number (100–1000) 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 unclear, 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 the 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 whether a precomputed surface is likely to be important for biological functions. Our 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. This is followed by a large scale cross-validation study.

METHODS

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 $\mathscr{A}$ of 3,275 proteins whose surface pockets contain one or more annotated residues as recorded either in the pdb or in the SwissProt database. From these, we select a subset $\mathscr{B}$ of ≈ 700 structure after further cleaning up by verifying the annotations for each of the key residues, as well as

Address correspondence to Jie Liang, Department of Bioengineering, SEO, MC-063, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, Chicago, IL 60607-7052, USA. Electronic mail: jliang@uic.edu

requiring that experimentally measured B-factor exist. Altogether, this final set of ≈ 700 protein structures contain 3,007 annotated residues. We define a functional surface as a surface pocket containing one or more of annotated key residue(s). Fig. 1a shows the size distribution of functional surface pockets in set $\mathscr{A}$. The mean size is 35 residues. Fig. 1b shows that the amino acid residue composition of these functional surfaces is very different from that of full backbone protein sequences.[13]

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 the size of functional surface over the total size 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 $\mathscr{A}$. Similar to previous studies,[2,14,15] we found that His, Asp, Glu, Ser and Cys account for more 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 its *atomic pattern* by listing the atoms that are exposed on the surface wall of the pocket in a consistent order. In addition, the secondary structural environment (e.g., β-sheet, denoted as *s*, α-helix *h*, and coil *c*) of a residue also provides useful information. For example, 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

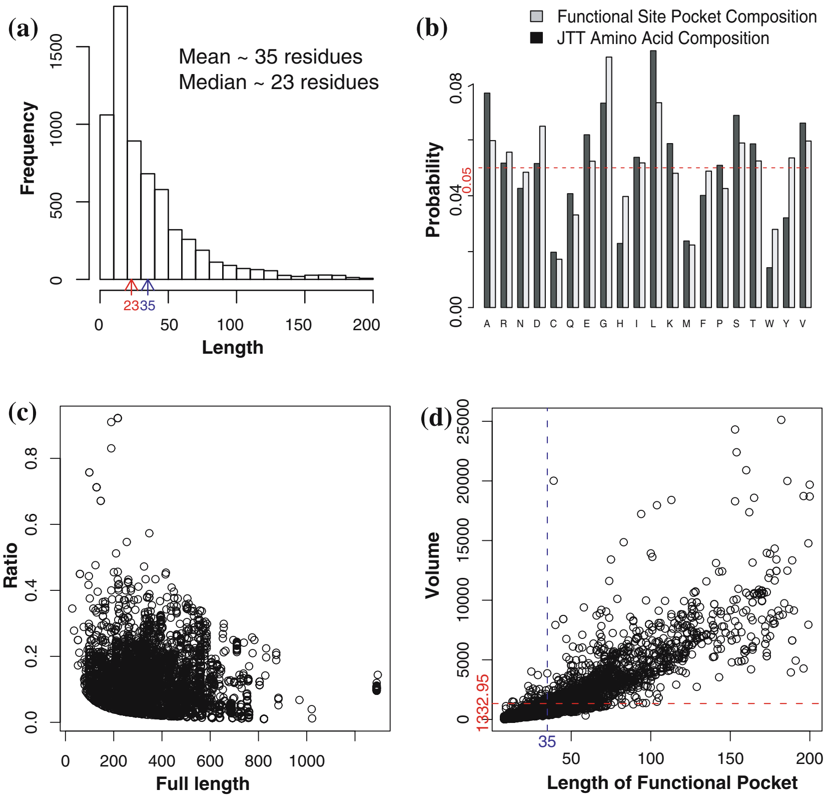


FIGURE 1. The length distribution and unique residue composition of functional surfaces for 3275 proteins with known key residues. (a) Functional surfaces usually consist of 8–200 residues, with the mean value of 35 residues. (b) The amino acid residue composition of functional surfaces on these proteins is different from the composition of sequences used to construct the JTT model.[13] (c) The distribution of the size ratio (defined as $\frac{\text{length(pocket)}}{\text{length(backbone)}}$). 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 length 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.

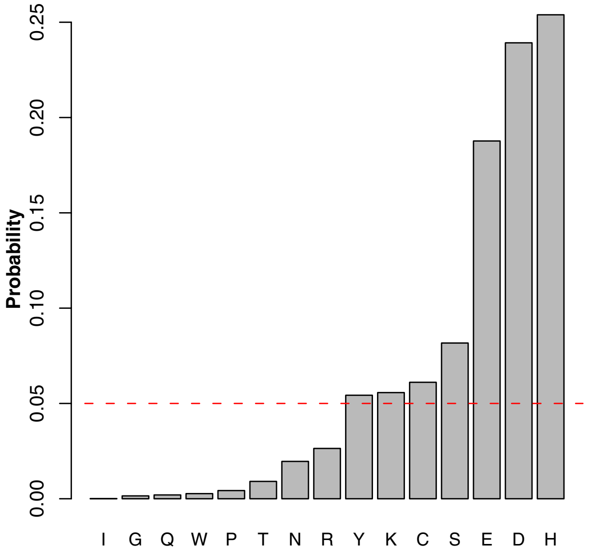


FIGURE 2. Active site residues are mapped to functional pockets and based on annotation in SwissProt and Pdb (17930 pdb entries). His, Asp, Glu, Ser and Cys account for more than 80% of active site residues of functional pockets. In the contrast, Ala, Pro, Val, Leu and Met are completely missed because they are hydrophobic attracted in the core of proteins.

substrates. Therefore, we also record the secondary structural 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. 4b) has the following atomic pattern:

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

From the 3007 annotated key residues on proteins of set $\mathscr{B}$, we obtain 1031 atomic patterns.

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* $P(i \in \mathscr{K})$ for the *i*-the residue to be from the set $\mathscr{K}$ of key residues:

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

where $\pi(s_i|i \in \mathscr{K})$ the probability of a key residue to be of the secondary structure type s_i, $\pi(r_i|i \in \mathscr{K})$ is the probability of a key residue to be amino acids type r_i, $\pi(a_i|i \in \mathscr{K})$ the probability of a key residue to be of the atomic pattern a_i, respectively. These are estimated from the $\mathscr{B}$ dataset of annontated key residues. For example, the probability $\pi(a_i|i \in \mathscr{K})$ is estimated from the occurrence of a specific atomic pattern a_i taken by residue *i* in all annotated key residues from set $\mathscr{B}$.

B-factors as a Filter for Atoms

Temperature B-factors or Debye-Waller factors are experimentally measured for atoms in X-ray crystallography and have been used to represent the atomic mobility. Residues exhibiting relatively low B-factors are generally those participating in forming secondary structures, neighboring disulfide bridges, or are involved in ligands binding. Atoms largely exposed to solvent generally experience more fluctuation and exhibit larger B-factors.

To test the hypothesis whether key resides potentially involved in ligands binding have lower B values, we use *ca* 500 structures without ligands or substrates from protein set $\mathscr{B}$ and compare B-factors of key residues and of non-key resides. Fig. 3 shows that in general key residues have smaller B-factors, and most are polar residues (e.g., His, Asp, Glu, Asn, Gln, Lys, Arg, and Ser).

Based on this observation, we use B-factors as a filter in our predicton. For a surface pocket, we first select only atomic patterns with high probabilities, namely, those appear with high frequencies among all patterns. For atomic patterns with single occurrence that is recorded in the database of known key residues, we compare their B-factors to that of key residues with

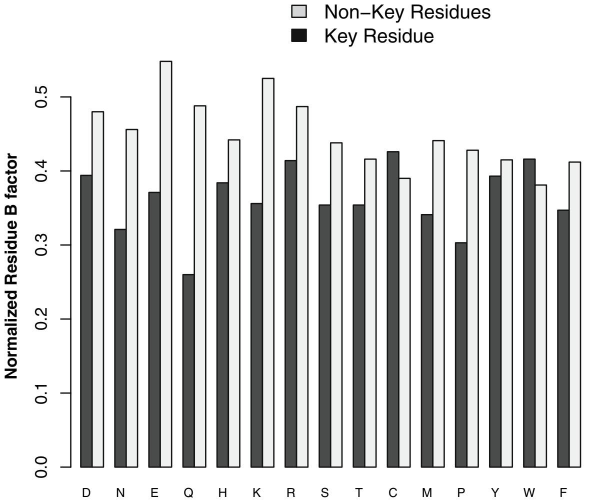


FIGURE 3. Functionally important key residues have overall smaller B-factors. B-factors of key residues from structures without bound-ligand are compared to B-factors of non-key resides. For residues from a protein structure, B-factors are normalized by the difference of the maximal and minimal values.

TABLE 1. B-factor can be used as a filter to improve the accuracy of predicting key residues. The $\mathscr{B}$ dataset is equally divided. One half is used as training set to predict key resides in the other half (containing 342 structures with a total 52,228 resideus). Here TP is the number of true positives, FP false positives, TN true negatives, and FN false negatives. TP/(TP+FP) is the positvie predicted value representing prediction accuracy. The accuracy of prediciton is improved if B-factor is used as a filter for predicting key residues in a protein surface.

Filter	TP	TN	FP	FN	TP/(TP+TP)
B-factor	1445	49770	305	708	82.6%
No filter	1349	49675	496	708	73.1%

the same atomic patterns from our database. If the B-factor is less than the highest one from the database, we accept this residue for further analysis, otherwise this residue is removed from further considertion. For multiply occurring patterns, we only accept the ones if their B-factors are less than the average B-factors of the same pattern in the database, or we choose the lowest one. With this implementation of B-factor as a filter, we can improve the accuracy of predicting key residues (Table 1).

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 $\mathscr{P}$ on a protein structure. We compute the summed probability SP(p) for a pocket surface p:

$$\mathrm{SP}(p) = \sum_{i \in p} P(s_i, r_i, a_i, i \in \mathscr{K}).$$

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

RESULTS

An Example

We use alpha-amylases1bag 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 the structure of the protein.

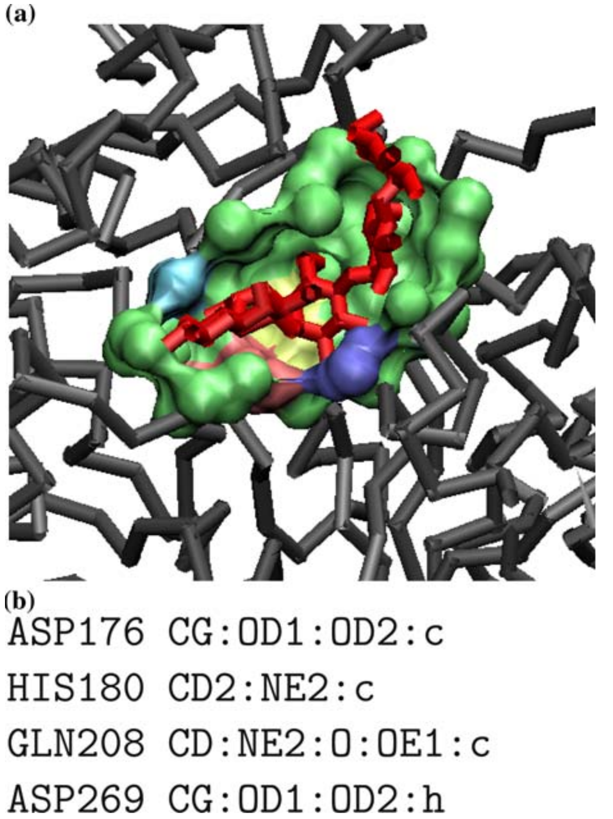
We first exhaustively compute all of the pockets (including voids) on this protein structure.[4,16] We then compute the key residue probability $P(i \in \mathscr{K})$ for each residues i in a pocket.

We first predict the functional surface. We rank the 60 surface pockets by summed probability SP(p). The largest pocket (CastP ID = 60) contains the largest number (7) of predicted key residues, and has the largest SP(p) = 1.31×10^{-3} value. It is therefore 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. 4). We found that there are four residues whose $P(i \in \mathscr{K})$ values are significantly higher than the rest of 14 residues, and are predicted as key residues. These residues are identical from the annotated residues reported in the literature.[7,9]

Large-scale Prediction of Functional Surfaces

Locating the functional surface is an important task in studying enzyme mechanism, as the correct surface



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

FIGURE 4. Predicting binding surface and key residues of alpha-amylase. (a) The pocket (green) with Castp ID = 60 is predicted to be the functional surface interacting 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 1031 functional atomic patterns. The class of secondary structural environment (β sheet *s*, helix *h* , and coil *c*) is also listed.

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 tests on the $\mathscr{B}$ dataset. We remove 10% of the structures to test the performance of the prediction method, which is derived from the analysis of the rest 90% of the data.

Our results are summarized in a Receiver Operating Characteristics (ROC) curve, where the sensitivity of our method is plotted against its specificity at various significance levels of summed probability values. Here the x-axis represents the false positive rate, namely, 1−specificity, or, 1- TN/(TN + FP), where TN is the number of true negatives, FP the number of false positives. The y-axis represents the true positive rate or sensitivity, defined as TP/(TP + FN), where FN is the number of false negatives.

An overall performance measure is the area under the ROC curve, which is 98.3%, indicating our method performs very well. At the confidence level of summed probability SP = 10^{-3}, the average specificity of our predictions of the functional surfaces of all 3,503 protein surfaces in these 70 proteins in 10-fold cross-validation tests is 99.88%, and the average sensitivity is 92.9% (Fig. 5). Table 2 further provides details of the performace assessed in accuracy (measured as TP/(TP + FP)), with an average value of 91.2%.

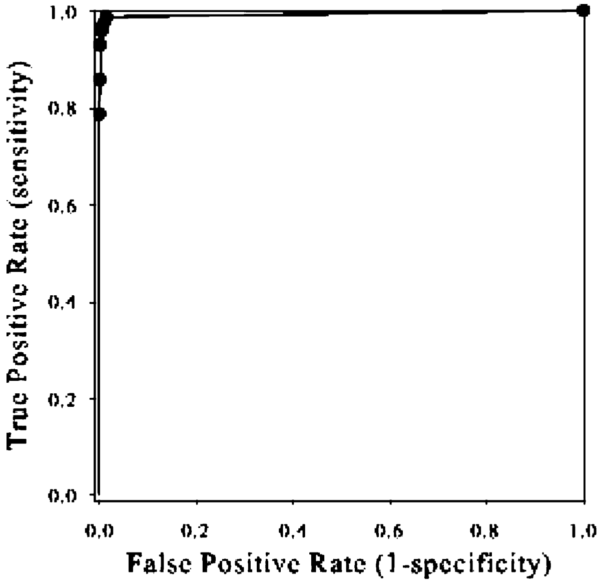


FIGURE 5. Performance in predicting functional surfaces of enzymes in 10-fold cross-validation tests summarized in a Receiver Operating Characteristics (ROC) curve. The overall area under the ROC curve is 98.3%, indicating our method has excellent performance.

TABLE 2. Results of functional surface prediction using 10-fold cross validations. The average accuracy is 91.2%.

Runs	TP	FP	TP/(TP+FP)
1	65	5	0.929
2	66	4	0.943
3	64	6	0.914
4	63	7	0.900
5	63	7	0.900
6	64	6	0.914
7	62	8	0.886
8	65	5	0.929
9	62	8	0.886
10	65	5	0.929

Prediction of Key Residues on Protein Functional Surfaces

We compare the predicted key residues 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 3, we are able to accurately locate many functionally important residues.

CONCLUSIONS AND FUTURE WORKS

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 ≥ 91.2% accuracy. In the example of alpha-amylase, functional surface and the key residues identified 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.

Future Works

We plan to increase the size of the library of annotated functional surfaces, as more structures are

TABLE 3. Detecting functional surfaces and locating key residues. Predicted results and the true answers as recorded in the human curated SFLD[18] database are listed. Several residues are annotated as iron binding are not considered to be catalytic and are therefore removed.

PDB structure	Predicted surface ID/length	Predicted key residues	SFLD (experimental data)
1bag	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			

being deposited in the PDB databank. Additional annotations can be incorporated 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).

We also plan to incorporate evolutionary information in our model. Because residues in protein functional surface experience strong selection pressure,[19] we expect this would further improve our method. We plan to further study 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 examine this issue and assess the robustness of current approach.

ACKNOWLEDGMENTS

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

REFERENCES

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

[2] Bartlett, G. J., C. T. Porter, N. Borkakoti, and J. M. Thornton. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* 324:105–121, 2002.

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

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

[5] Binkowski, T. A., A. Joachimiak, and J. Liang. Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci.* 14:2972–2981, 2005.

[6] Chandonia, J. M. and S. E. Brenner. The impact of structural genomics: Expectations and outcomes. *Science* 311(5759):347–351, 2006.

[7] Collins, T., D. De Vos, A. Hoyoux, S. N. Savvides, C. Gerday, J. Van Beeumen, and G. Feller. Study of the active site residues of a glycoside hydrolase family 8 xylanase. *J. Mol. Biol.* 354(2):425–435, 2005.

[8] Copley, S. D., W. R. Novak, and P. C. Babbitt. Divergence of function in the thioredoxin fold suprafamily: Evidence for evolution of peroxiredoxins from a thioredoxin-like ancestor. *Biochemistry* 43:13981–13995, 2004.

[9] Fujimoto, Z., 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.* 277:393–407, 1998.

[10] George, R. A., R. V. Spriggs, G. J. Bartlett, A. Gutteridge, M. W. MacArthur, C. T. Porter, B. Lazikani, J. M. Thornton, and M. B. Swindells. Effective function annotation through catalytic residue conservation. *Proc. Natl. Acad. Sci. USA* 102:12299–12304, 2005.

[11] Glaser, F., R. J. Morris, R. J. Najmanovich, R. A. Laskowski, and J. M. Thornton. A method for localizing ligand binding pockets in protein structures. *Proteins* 62:479–488, 2006.

[12] Gold, N. D. and R. M. Jackson. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J. Mol. Biol.* 355:1112–1124, 2006.

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

[14] Kim, J., J. Mao, and M. R. Gunner. Are acidic and basic groups in buried proteins predicted to be ionized?. *J. Mol. Biol.* 348:1283–1298, 2005.

[15] Laskowski, R. A., J. D. Watson, and J. M. Thornton. Protein function prediction using local 3D templates. *J. Mol. Biol.* 351:614–626, 2005.

[16] Liang, J., 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.

[17] Meng, E. C., B. J. Polacco, and P. C. Babbitt. Superfamily active site templates. *Proteins* 55:962–976, 2004.

[18] Pegg, S. C., S. D. Brown, S. Ojha, J. Seffernick, E. C. Meng, J. H. Morris, P. J. Chang, C. C. Huang, T. E. Ferrin, and P. C. Babbitt. Leveraging enzyme structure-function relationships for functional inference and experimental design: The structure-function linkage database. *Biochemistry* 45:2545–2555, 2006.

[19] Tseng, Y. Y. 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.* 23:421–436, 2006.