Evaluation of features for catalytic residue prediction in novel folds

EUNSEOG YOUN,[1] BRANDON PETERS,[1] PREDRAG RADIVOJAC,[2] AND SEAN D. MOONEY[1]

[1]Center for Computational Biology and Bioinformatics, Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA
[2]School of Informatics, Indiana University, Bloomington, Indiana 47408, USA

(RECEIVED August 25, 2006; FINAL REVISION November 8, 2006; ACCEPTED November 10, 2006)

Abstract

Structural genomics projects are determining the three-dimensional structure of proteins without full characterization of their function. A critical part of the annotation process involves appropriate knowledge representation and prediction of functionally important residue environments. We have developed a method to extract features from sequence, sequence alignments, three-dimensional structure, and structural environment conservation, and used support vector machines to annotate homologous and nonhomologous residue positions based on a specific training set of residue functions. In order to evaluate this pipeline for automated protein annotation, we applied it to the challenging problem of prediction of catalytic residues in enzymes. We also ranked the features based on their ability to discriminate catalytic from noncatalytic residues. When applying our method to a well-annotated set of protein structures, we found that top-ranked features were a measure of sequence conservation, a measure of structural conservation, a degree of uniqueness of a residue's structural environment, solvent accessibility, and residue hydrophobicity. We also found that features based on structural conservation were complementary to those based on sequence conservation and that they were capable of increasing predictor performance. Using a family nonredundant version of the ASTRAL 40 v1.65 data set, we estimated that the true catalytic residues were correctly predicted in 57.0% of the cases, with a precision of 18.5%. When testing on proteins containing novel folds not used in training, the best features were highly correlated with the training on families, thus validating the approach to nonhomologous catalytic residue prediction in general. We then applied the method to 2781 coordinate files from the structural genomics target pipeline and identified both highly ranked and highly clustered groups of predicted catalytic residues.

Keywords: catalytic residue prediction; structural environment conservation; feature evaluation

Identifying residues of importance in the protein products of genes is a challenging and important problem for informatics, genomics, biochemistry, and drug discovery. A particular challenge for the computational biologist is identifying features that are correlated with or, preferably, govern biochemical/cellular processes and are useful for prediction. In contrast to previous efforts that define functional sites broadly (Lichtarge et al. 1996; Elock 2001; Porter et al. 2002), we apply supervised machine learning methods with the goal of predicting the specific functional roles of amino acids. To address this, we are investigating how similarity among amino acid sequence, evolutionary, and structural descriptors can be used to quantify specific amino acid functional roles in proteins. We have developed a pipeline for knowledge-based annotation of residue function using support vector

Reprint requests to: Sean D. Mooney, 714 N. Senate Avenue, EF 250, Indianapolis, IN 46202, USA; e-mail: sdmooney@iupui.edu; fax: (317) 278-9217.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.062523907.

machines (SVMs) and evaluated it on the problem of automated annotation of catalytic residues in enzymes. We chose catalytic residues because it is only a partially understood problem and very challenging, given that only ~1% of residues in a single enzyme are catalytic and only a subset of proteins are enzymes. Throughout this study, we adopt the definition of a catalytic residue provided by Gutteridge et al. (2003) and use the Catalytic Site Atlas (CSA) as a gold standard (Porter et al. 2002).

Many approaches for identification of functional residues rely on broadly identifying residues of importance based on evolutionary sequence conservation, phylogenetic trees, and/or protein structure (Casari et al. 1995; Lichtarge et al. 1996; Aloy et al. 2001). Few methods are able to specifically annotate the role certain residues play in protein function (Ofran and Rost 2003; Ota et al. 2003; Iakoucheva et al. 2004; Yan et al. 2004), and even less have incorporated protein structure into their predictions. These methods have many applications, including annotation of structural genomics targets and predicting whether mutation will have an effect on that specific functional role. Annotating a specific functional role is a challenging endeavor, usually only a subset of functional residues actually participates in that role (such as catalysis, phosphorylation, ligand stabilization, etc.), and building good training sets is often difficult.

Some approaches (Gutteridge et al. 2003; Petrova and Wu 2006) have addressed these questions. Previously, the Thornton group used a neural network approach to predict catalytic residues using sequence conservation, residue type, and structural features (Gutteridge et al. 2003). They found that features including sequence conservation, secondary structure, residue type, and solvent accessibility were important (Bartlett et al. 2002; Jones and Thornton 2004; Torrance et al. 2005). In our study, we assembled a diverse set of features based on local sequence neighborhood, two-dimensional and three-dimensional structure, and evolutionary conservation. Then, with the goal of predicting catalytic residues in unannotated proteins, we developed a prediction model based on SVMs. SVMs have been extensively used in various machine learning problems, especially prediction, as an alternative to standard neural network approaches (Haykin 1999). Their previously successful applications include microarray analysis (Brown et al. 2000), disorder prediction in proteins (Ward et al. 2004), protein secondary structure prediction (Hua and Sun 2001; Zhang et al. 2005), and protein solvent accessibility prediction (Nguyen and Rajapakse 2005), to name a few. The popularity of an SVM is due to its high generalization performance, its intuitive idea, its sound mathematical foundation, and its few numbers of free parameters to adjust.

We describe a method for automated catalytic site annotation and evaluate features in our predictor for their ability to discriminate these residues. Of these features, we find that structural features of residue environments such as solvent accessibility, in addition to sequence conservation, are important to prediction of catalytic residues. Encouragingly, we find that higher SVM scores are closer to the catalytic machinery of enzymes, and that the method is able to make predictions on novel folds at a small cost to accuracy.

The utility of this method can be illustrated along with other structure and functional inference tools such as ProFunc (Laskowski et al. 2005), DALI (Holm and Sander 1993), Structure-Based Local Environment Search Tool (S-BLEST) (Mooney et al. 2005), and PSI-BLAST (Altschul et al. 1997) by application to known structures of partially characterized function. To do this, we applied the method to 2781 structures determined as part of the structural genomics projects (Chandonia and Brenner 2006) and scored all residues. We identified highly ranked residues and found that they tend to be clustered with other highly ranked residues, and we further characterized several of the top hits in concert with the previously mentioned function analysis tools.

Results

Method training and evaluation

Our approach was to first identify features we believe might be important for prediction and then use those features to compare the experiments of the Thornton group (Gutteridge et al. 2003) using an SVM. After validation of the method on the previously mentioned data set, prediction was performed upon a 40% nonredundant set of proteins (ASTRAL 40 v1.65 data set) (Chandonia et al. 2004) with an annotation in CSA. We chose to use ASTRAL 40 because its sequences are well annotated and classified into the SCOP hierarchy (Murzin et al. 1995). This allowed us to evaluate how well our predictors perform on proteins from new structural families, new structural superfamilies, and new folds and to understand the relationship between the features that are useful in each of these situations.

We developed three data sets for evaluation based on fold, superfamily, and family. In order to avoid a bias from imbalanced protein domain distribution between classes (families, superfamilies, and folds), we randomly selected one protein domain from each class. In the first case, the data set was determined by selecting a random protein member from each family. In the second case, the data set was constructed by selecting a random member from each superfamily. Similarly, the final case was constructed by selecting a random member from each

fold. We then performed 10-fold cross-validation on each set, where no protein had residues in both testing and training.

As expected, we find that our results are similar to those of the Thornton group with a slight increase in both sensitivity and precision (sensitivity, 65.3%; precision, 14.4% vs. 56.0% and 14.0% without clustering). However, because we are applying our method to the ASTRAL 40 v1.65 data set, we are able to evaluate how well we can predict catalytic sites when varying degrees of structural similarity are present between the query sequence and the training set. This performance can likely be further improved if structural clustering methods are employed (Gutteridge et al. 2003), but for our purposes, the raw SVM scores are sufficient. When training by SCOP structural family, we find that we achieve 57.0% sensitivity and 18.5% precision for 10-fold cross-validation on the ASTRAL 40 nonredundant database (Table 1). This sensitivity reduces to 51.1% when trained on the fold level, suggesting, not surprisingly, that family neighbors are important for improved prediction. On the other hand, the performance drop is relatively small, which is encouraging for the annotation of the new folds. We confirmed this also by plotting the receiver operating characteristics (ROC) (Fawcett 2003) curves of family, superfamily, and fold level experiments (Fig. 1). To calculate the contribution of structural features, we estimate the performance using only sequence related features. We used the same experiment protocol as before on family level experiments but used only sequence features. Sequence only features have 16.6% sensitivity, 15.1% precision, and 86.6% area under the ROC curve (AUC).

Feature ranking was performed on each data set using the AUC (Fawcett 2003). To do this, we constructed an ROC for each feature independently and then ranked them according to the decreasing AUC values. Overall, in all sets we find that the best features are conservation in a sequence alignment, structural conservation score (SCS) using S-BLEST (Mooney et al. 2005), solvent accessibility, and residue class. Not surprisingly, the most important structural features are those calculated for the catalytic residue itself, while those related to its sequence and spatial neighbors were less discriminatory. Top-ranked features are shown in Table 2. Next, we evaluated whether the best features when training between families and training between folds are highly correlated (Fig. 2). The *X*-axis in Figure 2 is a feature ranking based on the family level data set, and *Y*-axis is the corresponding feature's ranking based on the fold level data set. We find that the features are highly correlated, with the best features having the least difference.

We also wanted to compare usefulness of structural conservation versus sequence conservation to catalytic site prediction and then evaluated whether these features were complementary. First, we computed the AUC of SVM prediction using only the information per position score feature at the family level. We employed 10-fold cross-validation experiments on the data set and collected SVM prediction scores to compute the AUC. Next, the AUC was computed using the information per position score from the output of PSI-BLAST and SCS features from S-BLEST. We find that the SCS feature increases the AUC value (0.87) compared with the case when sequence conservation was used alone (0.84). This analysis suggests that the structural conservation-based features can improve prediction over using sequence conservation alone. Figure 3 shows the distributions of these two features for catalytic and noncatalytic residues. For the information per position score, 83% of catalytic residues have scores $\geq$0.2, while 87% of noncatalytic residues have scores $\leq$0.2. For the SCS, 76% of catalytic residues have scores $\leq$0.7, while 88% of noncatalytic residues have scores $\geq$0.7.

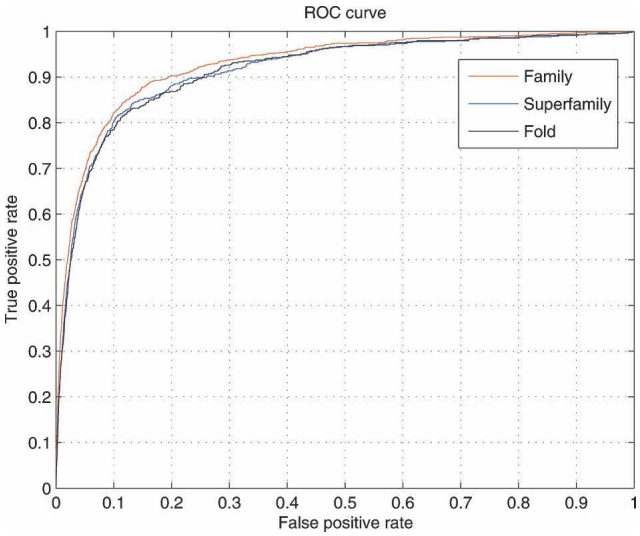

Figure 1. ROC curves for different data sets based on SCOP family, superfamily, and fold. This plot is analogous to the data in Table 1.

Table 1. *Tenfold cross-validation performance by data set*

Data set	Sensitivity (%)	Precision (%)	AUC (%)
Family	57.02	18.51	92.90
Superfamily	53.93	16.90	91.35
Fold	51.11	17.13	91.44

SCOP family based data collection is done by randomly picking one protein from each family. This approach is also applied for SCOP superfamily-based and SCOP fold-based data collection. The SVM is used to do 10-fold cross-validation evaluation on resultant data sets. As samples in test set are distantly related to those in training set, it is more difficult to predict, thus resulting in reduced performance.

Table 2. *The most highly ranked features used*

Rank	AUC	Source	Description
1	0.847	PSI-BLAST	PSSM
2	0.845	PSI-BLAST	Information per position (IPP)
3	0.837	PSI-BLAST	Weighted observed percentage
4	0.762	S-BLEST	SCS
5	0.688	Shell 1	Residue class2 is nonpolar
6	0.688	Shell 1	Residue class1 is hydrophobic
7	0.686	Shell 2	Residue class1 is hydrophobic
8	0.684	Shell 2	Mobility
9	0.681	Shell 2	Residue class2 is nonpolar
10	0.681	Shell 2	Charge with His
11	0.665	PSI-BLAST	IPP adjacent (N − 1)
12	0.665	Shell 4	Solvent accessibility
13	0.661	Shell 3	Solvent accessibility
14	0.659	Shell 4	Residue class1 is polar
15	0.657	PSI-BLAST	IPP adjacent (N + 1)
16	0.651	PSI-BLAST	IPP adjacent (N − 2)
17	0.650	Shell 3	Charge with His
18	0.641	PSI-BLAST	IPP adjacent (N − 3)
19	0.640	Shell 2	Atom name is any
20	0.640	Shell 3	Mobility
21	0.640	Shell 2	Hydrophobicity
22	0.639	Shell 1	Residue class1 is charged
23	0.637	Shell 4	Residue class2 is polar
24	0.632	PSI-BLAST	IPP adjacent (N + 2)
25	0.632	Shell 3	Residue class1 is polar
26	0.632	Shell 4	Atom name is O
27	0.629	Shell 4	Secondary structure2 is beta
28	0.626	Shell 3	Secondary structure2 is beta
29	0.622	PSI-BLAST	IPP adjacent (N − 4)
30	0.620	Shell 3	Amine

Features are ranked based on the area under the ROC curve using the family data set described in the text. Among 314 features, the top 30 features are listed. Score is the area under the ROC curve (AUC) value. Source describes the source of the feature, and Shell i is the *i*th shell of the S-BLEST vector.

For each residue, the SVM outputs a likelihood that a residue is catalytic. We computed the distance between the Cα atom of the residue being evaluated and the Cα atom of the closest catalytic residue in the protein. We find that high scoring noncatalytic residues are approximately 7 Å away from a catalytic residue, while residues near the decision threshold are nearly twice that far (Fig. 4). The tradeoff between decision threshold and precision is presented in Figure 5, and shows that precision can be increased significantly with higher score thresholds.

Using the fold data set, Figure 6 shows the performance of the different types of catalytic residues (left) and the distribution of all the residue type frequencies (right). Histidine is a rare residue type, but it has the highest catalytic residue frequency and sensitivity. That is, histidine constitutes only 2.6% of all residues but 18% of all catalytic residues in fold data set. On the contrary, lysine has the highest residue frequency (9%) but constitutes only 1% of catalytic residues. Figure on the left suggests that rare or hydrophobic catalytic residue types are very difficult to predict and consequently have very low sensitivity and precision.

Analysis of solved structural genomics targets

In an effort to illustrate the utility of this method, we applied it to all solved structural genomics targets listed on the structural genomics website. These included 5143 chains contained in 2781 unique Protein Data Bank (PDB) identification numbers. The highest scoring residues are dominated by divalent ion binding sites or similar, as can be seen from the examples below. The top scoring residues are shown in Table 3.

Overall, the highest scoring residue from all structural genomics targets was histidine 145 from pdb:1XM7 chain A, the crystal structure of the hypothetical protein aq_1665 from *Aquifex aeolicus* as determined by the Midwest Structural Genomics Center (R. Zhang, M. Zhou, F. Collart, and A. Joachimiak, in prep.). This protein was originally deposited in October 2004, and its function is listed with only "hydrolase activity." ProFunc (Laskowski et al. 2005) identifies this protein as a purple acid phosphotase. Similarity matches on ProFunc include hits to the calcineurin-like phosphoesterases (InterPro), metallo-dependent phosphatases (InterPro), and apo structure of methanococcus jannaschii phosphodiesterase mj0936 pdb:2AHD (SSM [Krissinel and Henrick 2004] and DALI). S-BLEST (Mooney et al. 2005) identifies a region that includes residues 49, 77, 111, and 145 as being closely associated with its likely SCOP superfamily (as determined

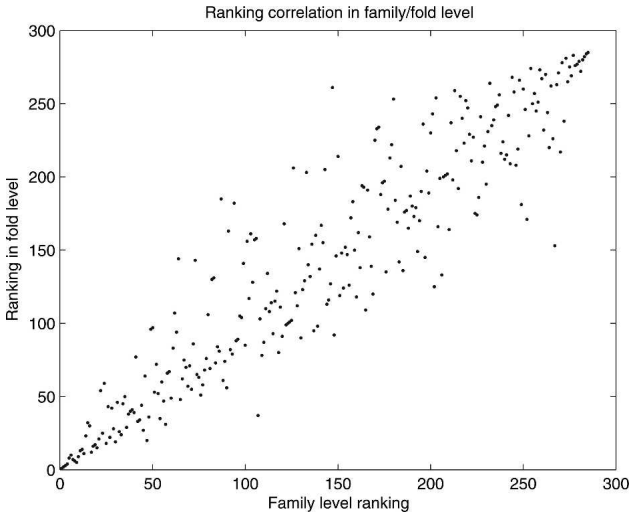

Figure 2. Feature ranking correlation between family-based and fold-based data sets. Each of 314 features was ranked based on their class discriminating value, by determining the AUC value. Several S-BLEST features are not informative (constant in both catalytic and noncatalytic residues), and these were removed from this plot.

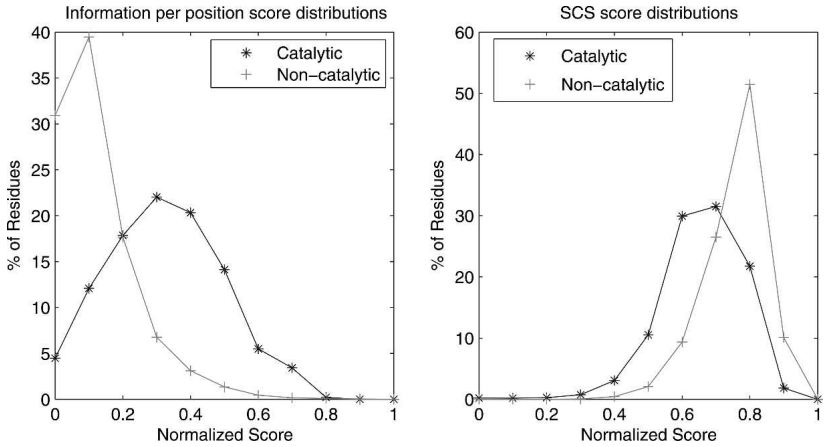

Figure 3. Information per position score from PSI-BLAST and SCS from S-BLEST distributions for catalytic and noncatalytic residues.

by S-BLEST), metallo-dependent phosphatases (d.159.1). This region appears to form a large cleft and is solvent accessible. The catalytic residue predictor described here identifies four residues with very high prediction scores. These residues are spatially tightly clustered and are histidine 145 (SVM score of 4.5), aspartate 50 (3.5), histidine 111 (3.4), and aspartate 7 (3.3). All of these residues are located in the region described by S-BLEST. It is interesting that the crystal structure does not contain a metal ion at this site. Figure 7a shows the region on a ribbon representation of chain A.

The second highest scoring residue from all structural genomics targets was aspartate 189 from pdb:2AZ4 chain A, crystal structure of a hypothetical protein from *Enterococcus faecalis* V583 as determined by the Midwest Structural Genomics Center (R. Zhang, N. Maltseva, S. Moy, F. Collart, M. Cymborowski, W. Minor, and A. Joachimiak, in prep.). This protein was originally deposited in September 2005 without any functional annotation. ProFunc again sheds light onto potential functions of this protein. The superfamily is predicted to be metallo-hydrolase/oxidoreductase (superfamily) and several hits to the Zn-dependent hydrolase of metallo-β-lactamase superfamily (SSM and DALI). S-BLEST identifies a region from the metallo-hydrolase/oxidoreductases superfamily (d.157.1). Using these proteins, a region roughly defined by the most associated residue environments centered at residues 92, 94, 167, 168, 169, and 189 is found to be closely associated to this superfamily, metallo-hydrolase/oxidoreductases (d.157.1). The catalytic residue predictor identifies several residues that are scored very highly and again are tightly clustered, aspartate 189 (4.2), histidine 94 (3.8), histidine 92 (3.5), histidine 167 (3.3), and histidine 97 (3.2). This region is adjacent to two Zn^{2+} atoms that directly interact with these residues. Figure 7b shows the region on a ribbon representation of chain A.

The third highest scoring residue from all structural genomics targets was histidine 9 from pdb:1ZZM chain A, the crystal structure of YJJV, TATD Homolog from *Escherichia coli* k12 as determined by the New York Structural Genomics Research Consortium (NYSGRC; www.nysgrc.org). This protein was deposited June 2005 with no annotated function. S-BLEST identifies two structurally similar structures from ASTRAL 40 v1.69, both from the TIM β/α-barrel (c.1.9) superfamily. Again, several residues surrounding a metal ion binding site are high scoring, histidine 9 (4.1), histidine 133 (3.5),

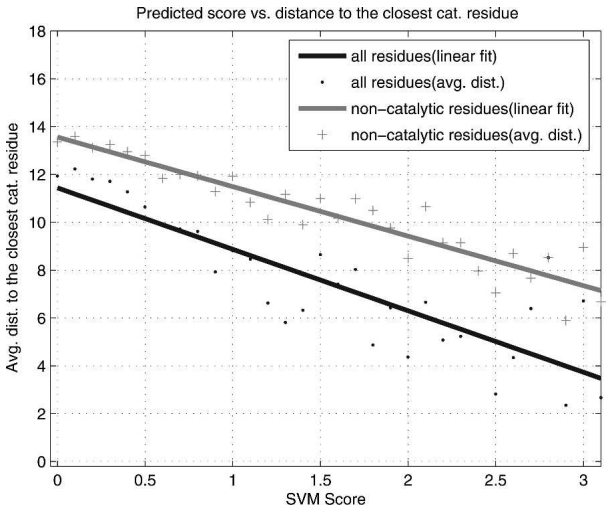

Figure 4. SVM predicted score vs. distance (Å) to the closest catalytic residue. The SVM outputs a predicted score for each residue. Distance (Å) between the residue and the closest catalytic residue is computed. The predicted score x is binned by rounding x to one decimal place (X-axis). The distances in each bin are averaged (Y-axis). Averaged distances are then fitted by a linear regression. Averaged distance decreases almost linearly as predicted score increases.

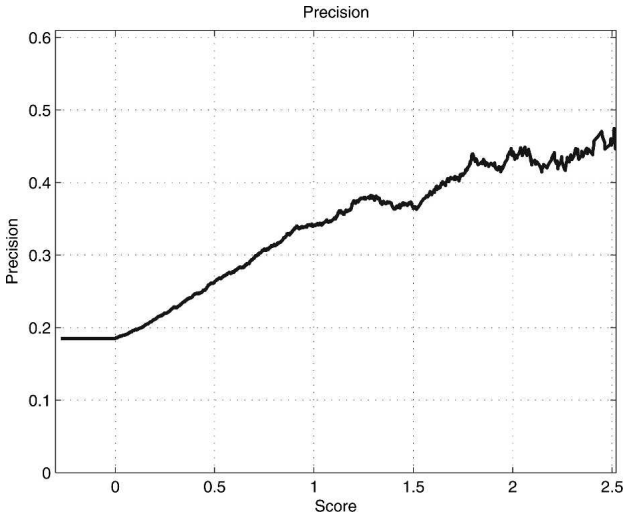

Figure 5. Precision of prediction as decision threshold is shifted. After the SVM predicted scores are sorted in descending order, the top k scores (X-axis) are used to compute the precision (Y-axis), where $k = 1,\ldots,5000$.

glutamate 97 (3.4), and histidine 11 (3.3). Figure 7c shows the region on a ribbon representation of chain A.

The final example is crystal structure of conserved hypothetical protein from *Pseudomonas aeruginosa* PAO1 from pdb:1Z7A chain E (although other chains in this file had similarly high scores) as determined by the Midwest Structural Genomics Center (C. Chang, T. Skarina, A. Savchenko, A. Edwards, and A. Joachimiak, in prep.). In this case, the score is high, but not among the top proteins. Again, a cluster of histidine and charged residues is identified, and here there is no bound ion. Similarity searches on ProFunc predict this to be a polysaccharide deacetylase (HMMPfam), a glycoside hydrolase/deacetylase (superfamily), and DALI associates this domain with several carbohydrate esterase structures. S-BLEST does not find any significant similarity to any known structural environments in ASTRAL 40 v1.69, although carbohydrate metabolism and esterase activity are represented in the top hits. The highly ranked residues are histidine 126 (3.1), histidine 259 (3.0), and glutamate 36 (2.3). Figure 7d shows the region on a ribbon representation of chain A.

Discussion

Structural genomics projects are experimentally determining structures of proteins of novel folds without full or even partial characterization of their function. While many methods predict the function and functionally important regions of these structures, few methods identify residues that are participating in specific functional roles. Nonhomologous site prediction methods, that is, methods that do not transfer function via sequence homology alone, are critical for this hypothesis generation in functionally uncharacterized proteins and identification of residues playing specific functional roles.

In this study we investigated influence of various sequence, structural, and evolutionary features to the problem of annotation of catalytic residues and then developed a model for automated identification of catalytic residues in unannotated protein structures. We find that catalytic residues can be predicted on enzymes that come from new folds with a small cost to both sensitivity and precision over proteins from previously characterized families. In an effort to enable researchers to access our method, we have developed a Web site and coordinate

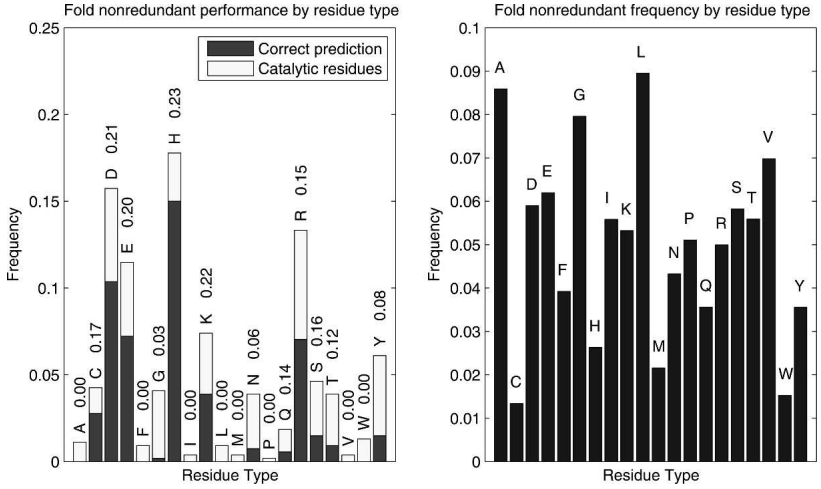

Figure 6. Performance by the different types of catalytic residues. Plot on the *left* considers only catalytic residues. Light bars show the distribution of different types of catalytic residues. Dark bars show sensitivity. Numbers *above* each residue type letter in the *left* panel represent precisions of corresponding residue type. Plot on the *right* shows frequencies of different residue types.

Table 3. *The most highly ranked residues from the data set of solved structural genomics targets*

PDB_ID	CHAIN	RES_NUM	RES_TYPE	SVM_SCORE	PDB_TITLE
1XM7	A	145	H	4.51	STRUCTURAL GENOMICS, THE CRYSTAL STRUCTURE OF THE HYPOTHETICAL PROTEIN AQ_1665 FROM *AQUIFEX AEOLICUS*
1XM7	B	145	H	4.32	STRUCTURAL GENOMICS, THE CRYSTAL STRUCTURE OF THE HYPOTHETICAL PROTEIN AQ_1665 FROM *AQUIFEX AEOLICUS*
2AZ4	A	189	D	4.18	CRYSTAL STRUCTURE OF A HYPOTHETICAL PROTEIN FROM *ENTEROCOCCUS FAECALIS* V583
1ZZM	A	9	H	4.13	CRYSTAL STRUCTURE OF YJJV, TATD HOMOLOG FROM *ESCHERICHIA COLI* K12, AT 1.8 Å RESOLUTION
2AZ4	B	189	D	4.08	CRYSTAL STRUCTURE OF A HYPOTHETICAL PROTEIN FROM *ENTEROCOCCUS FAECALIS* V583
1S3L	B	320	H	4.03	STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF A NOVEL ARCHAEAL PHOSPHODIESTERASE
1S3L	A	120	H	4.03	STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF A NOVEL ARCHAEAL PHOSPHODIESTERASE
1Z2I	A	125	H	4.02	CRYSTAL STRUCTURE OF AGROBACTERIUM TUMEFACIENS MALATEDEHYDROGENASE, NEW YORK STRUCTURAL GENOMICS CONSORTIUM
1ZKP	C	155	D	4.01	1.5 Å RESOLUTION CRYSTAL STRUCTURE OF A METALLO BETALACTAMASE FAMILY PROTEIN, THE ELAC HOMOLG OF *BACILLUS ANTHRACIS*, A PUTATIVE RIBONUCLEASE
2FFI	A	18	H	4.01	CRYSTAL STRUCTURE OF PUTATIVE 2-PYRONE-4,6-DICARBOXYLICACID HYDROLASE FROM *PSEUDOMONAS* PUTIDA, NORTHEAST STRUCTURAL GENOMICS TARGET PPR23.
1ZTC	C	71	H	3.98	CRYSTAL STRUCTURE OF HYPOTHETICAL PROTEIN (TM0894) FROM *THERMOTOGA MARITIMA* AT 2.10 Å RESOLUTION
1UUF	A	40	C	3.95	CRYSTAL STRUCTURE OF A ZINC-TYPE ALCOHOL DEHYDROGENASE-LIKEPROTEIN YAHK
1ZKP	B	155	D	3.92	1.5 Å RESOLUTION CRYSTAL STRUCTURE OF A METALLO BETALACTAMASE FAMILY PROTEIN, THE ELAC HOMOLG OF *BACILLUS ANTHRACIS*, A PUTATIVE RIBONUCLEASE
1YIX	A	7	H	3.91	CRYSTAL STRUCTURE OF YCFH, TATD HOMOLOG FROM *ESCHERICHIA COLI* K12, AT 1.9 Å RESOLUTION
1ZKP	A	155	D	3.91	1.5 Å RESOLUTION CRYSTAL STRUCTURE OF A METALLO BETALACTAMASE FAMILY PROTEIN, THE ELAC HOMOLOG OF *BACILLUS ANTHRACIS*, A PUTATIVE RIBONUCLEASE
1ZKP	D	155	D	3.91	1.5 Å RESOLUTION CRYSTAL STRUCTURE OF A METALLO BETALACTAMASE FAMILY PROTEIN, THE ELAC HOMOLG OF *BACILLUS ANTHRACIS*, A PUTATIVE RIBONUCLEASE
1M65	A	40	H	3.91	YCDX PROTEIN
1ZTC	D	71	H	3.88	CRYSTAL STRUCTURE OF HYPOTHETICAL PROTEIN (TM0894) FROM *THERMOTOGA MARITIMA* AT 2.10 Å RESOLUTION
2FFI	B	18	H	3.88	CRYSTAL STRUCTURE OF PUTATIVE 2-PYRONE-4,6-DICARBOXYLICACID HYDROLASE FROM *PSEUDOMONAS* PUTIDA, NORTHEAST STRUCTURAL GENOMICS TARGET PPR23.
1ZTC	A	71	H	3.87	CRYSTAL STRUCTURE OF HYPOTHETICAL PROTEIN (TM0894) FROM *THERMOTOGA MARITIMA* AT 2.10 Å RESOLUTION
1ZTC	B	71	H	3.87	CRYSTAL STRUCTURE OF HYPOTHETICAL PROTEIN (TM0894) FROM *THERMOTOGA MARITIMA* AT 2.10 Å RESOLUTION
1Z2I	C	125	H	3.86	CRYSTAL STRUCTURE OF AGROBACTERIUM TUMEFACIENS MALATEDEHYDROGENASE, NEW YORK STRUCTURAL GENOMICS CONSORTIUM
1XM8	A	131	D	3.85	X-RAY STRUCTURE OF GLYOXALASE II FROM *ARABIDOPSIS THALIANAGENE* AT2G31350
1XM8	B	131	D	3.85	X-RAY STRUCTURE OF GLYOXALASE II FROM *ARABIDOPSIS THALIANAGENE* AT2G31350
2A9V	D	80	C	3.85	CRYSTAL STRUCTURE OF (NP_394403.1) FROM *THERMOPLASMA ACIDOPHILUM* AT 2.45 Å RESOLUTION
2A9V	B	80	C	3.84	CRYSTAL STRUCTURE OF (NP_394403.1) FROM *THERMOPLASMA ACIDOPHILUM* AT 2.45 Å RESOLUTION

(continued)

Table 3. *Continued*

PDB_ID	CHAIN	RES_NUM	RES_TYPE	SVM_SCORE	PDB_TITLE
2GFQ	B	140	H	3.83	STRUCTURE OF HYPOTHETICAL PROTEIN PH0006 FROM *PYROCOCCUSHORI KOSHII*
1YIX	B	7	H	3.81	CRYSTAL STRUCTURE OF YCFH, TATD HOMOLOG FROM *ESCHERICHIA COLI* K12, AT 1.9 Å RESOLUTION
2A9V	A	80	C	3.80	CRYSTAL STRUCTURE OF (NP_394403.1) FROM *THERMOPLASMA ACIDOPHILUM* AT 2.45 Å RESOLUTION
1S3N	B	297	H	3.78	STRUCTURAL AND FUNCTIONAL CHARACTERIZATION OF A NOVEL ARCHAEAL PHOSPHODIESTERASE

The top residues are overwhelmingly associated enzymes with metal ion binding sites in the active site. Additionally, the residues are generally either charged or histidine.

submission method at http://sblest.org/crp/. Observing the usefulness of the features we use, we can apply the same attributes on other similar residue-based data sets.

Across all experiments, the best features are based on sequence conservation, structural conservation, or structural uniqueness. This finding is similar to those that have been seen in other methods that predict catalytic residues (Gutteridge et al. 2003) and predictions of other functional sites including deleterious mutation prediction (Saunders and Baker 2002). Although these are the best features, many noncatalytic residues are sequence and structurally conserved (Fig. 3). This must be taken into account when hypothesizing that a conserved residue is catalytic, because catalysis is an interplay of several features and is more complicated than conservation in sequence alignments.

Not all residue types are equally predictable. As observed in Figure 6, hydrophobic residues are very difficult to predict. This is likely due to the large fraction of hydrophobic residues in proteins, and the very small fraction of hydrophobic catalytic residues. When retraining with only hydrophobic catalytic residues, we got poor results (data not shown), and it may be because catalytic residues (1 in 600) were much rarer in the training set and they were not distinguishable to the vast majority of noncatalytic residues.

It is truly an important endeavor for computational biology to develop new features for the prediction of important sites in proteins. While sequence, sequence conservation, and structural information have been well characterized for these problems, structural conservation has been difficult to quantify. It seems logical that quantifying structural conservation around a site in a protein could improve the ability to predict that site, as structural conservation adds information not present in the other features. Indeed, when we test whether knowledge that another similar structural site exists in a nonredundant database, our prediction accuracy improves over conservation in a multiple sequence alignment. This approach to measuring conservation is very simple (matching feature vectors) and likely could be improved with more sophisticated approaches that take into account all significant environments simultaneously.

By applying this approach to newly determined structures of unknown function, we are able to identify likely active site regions of enzymes and hypothesize about residues involved in catalytic mechanism. We find that the highest scoring residues tend to be histidine or charged residues, near ion binding sites, and tightly clustered. Since ion binding sites are not necessarily catalytic sites, this may be a potential area of focus for removing false positives. Lesser high scoring residues in other proteins tend to cluster near the active site of the protein. Obviously, performance of the method must be taken into account, so specific residue predictions should be considered hypotheses and not definitive.

In conclusion, the future for prediction of sites in new folds is possible using a set of bioinformatic features based on structure and evolution. We find that sequence conservation, structure conservation, residue class, and solvent accessibility represent the top features for prediction. Because the training data are highly imbalanced, this is a very challenging and important problem. As more features are identified, we believe that sensitivity and precision will improve.

Materials and methods

Data set

The training data set is based on ASTRAL 40 v1.65 and contains 987 protein SCOP domains (Murzin et al. 1995), 396 SCOP families, 236 SCOP superfamilies, and 189 SCOP folds. The catalytic residues in this data set are the CSA, which is heavily imbalanced, containing 2897 catalytic and 267,608 noncatalytic residues. Although prediction with imbalanced data is an important problem, in this study we did not try to optimize performance by evaluating different approaches for imbalanced data. To overcome this, we undersampled noncatalytic residues for training data while keeping the original ratio on test data. In particular, during data set construction, we included all catalytic

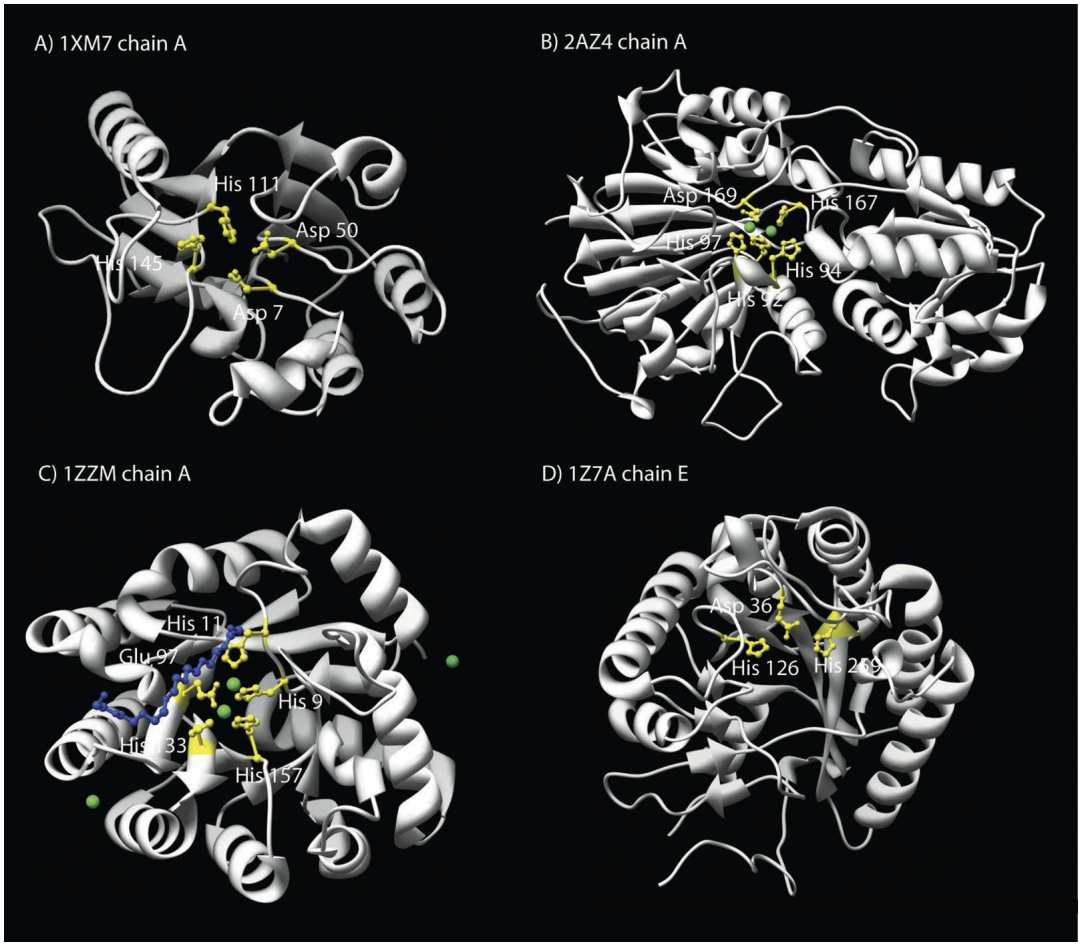

Figure 7. Analysis of the top-ranked residues from the solved structural genomics targets. (*A*) The crystal structure of the hypothetical protein aq_1665 from *Aquifex aeolicus* as determined by the Midwest Structural Genomics Center (pdb:1XM7 chain A). Here we highlight the four top-scored residues: histidine 145 (SVM score of 4.5), aspartate 50 (3.5), histidine 111 (3.4), and aspartate 7 (3.3). (*B*) The crystal structure of a hypothetical protein from *Enterococcus faecalis* V583 as determined by the Midwest Structural Genomics Center (pdb:2AZ4 chain A). Here we highlight the top-scoring residues: aspartate 189 (4.2), histidine 94 (3.8), histidine 92 (3.5), histidine 167 (3.3), and histidine 97 (3.2) along with the bound Zn^{2+} ions. (*C*) The crystal structure of YJJV, TATD Homolog from *Escherichia coli* k12 as determined by the New York Structural Genomics Research Consortium (NYSGRC) (pdb:1ZZM chain A). Again, the top residues are highlighted along with the bound ion. (*D*) The crystal structure of conserved hypothetical protein from *Pseudomonas aeruginosa* PAO1 as determined by the Midwest Structural Genomics Center (pdb:1Z7A chain E). Here we highlight the three top-scored residues: histidine 126 (SVM score of 3.1), histidine 259 (SVM score of 3.0), and glutamate 36 (SVM score of 2.3).

residues from each protein domain and at most 100 noncatalytic residues to prevent our system from being more influenced by the long proteins. After this step, the training data set was still imbalanced, and we overcame this by giving more weight to catalytic residues during the training process. The weighting on catalytic residues was done so that it has similar effect as having the ratio of 1:6 between the number of catalytic and noncatalytic residues in training data. This same ratio was used by the Thornton group (Gutteridge et al. 2003).

Feature encoding

Four classes of features were evaluated, based on sequence, sequence conservation, structure, and structure conservation. To construct sequence-based features, we used a protocol similar to those published for predicting protein phosphorylation sites (Iakoucheva et al. 2004). For the window of residue type, we considered window of size 21: 10 residues on the left, 10 residues on the right, and itself. After counting the frequency for each residue type, the raw frequency was divided by the window size. Beginning and ending residues do not have enough flanking residues, and thus, amino acid counts were divided by the appropriate window sizes to get relative frequencies.

Sequence conservation was evaluated using a position-specific scoring matrix (PSSM) determined from the output of PSI-BLAST (Altschul et al. 1997). To do this, the following automated protocol was employed. Each sequence in ASTRAL 40 was queried against the NCBI NR database using the NCBI blastpgp program. The output was parsed, and three scores were

extracted: a PSSM, a weighted observed percentage, and information per position score of the sequence conservation of the corresponding residue. For the sequence conservation of adjacent residues, we considered the window of size 20 (left: 10, right: 10) using the information per position score. Since N-terminal residues do not have left residues, we used an average conservation score of the protein to fill in the left residue scores. Similarly, we used the average for the right residues for the C-terminal residues.

Features based on protein structure utilized the representation described using S-BLEST (Mooney et al. 2005). A total of 264 atom-based structural properties are calculated in four radial shells, each extending 1.875 Å from the Cβ atom of the residue are considered. The properties include atom-based features (such as the number of carbon atoms in the shell), residue-based features (such as the number of atoms associated with alanine residues), secondary structure–based features, solvent accessibility, charge, and others. The full feature list is described previously (Bagley and Altman 1995) and is available upon request from the authors.

Another set of structural features we used are the B-factor, cysteine bridged pair information, solvent accessibility, and secondary structure of a residue. The B-factor is extracted from Cα atom of the residue from a PDB file. We used the definition of a cysteine bridged pair as if the distance between their sulfur atoms is within 2.5 Å. The solvent accessibility and secondary structure features were extracted from DSSP output (Kabsch and Sander 1983). Secondary structure classifications from DSSP program were simplified to helix, sheet, or coil/other.

We also included a feature that quantified conservation and uniqueness of the properties in the S-BLEST features. S-BLEST calculates a Manhattan distance between two vectors and determines a Z-score based on the distribution of distances relative to the query in the database being used. This feature, which we call the SCS, was the magnitude of the maximum absolute valued Z-score between the query residue and the top matched residue in ASTRAL 40 v1.65 using the previously described S-BLEST method. In order to compute the SCS, ideally, we need to construct a new database that does not contain any protein domains which belong to the same SCOP class (fold, superfamily, and family) for each protein in training, since these proteins will affect the value of the Z-score. This was not computationally tractable, so instead, we used the entire ASTRAL 40 v1.65 database and selected the top matched residue that does not belong to the same SCOP class (fold, superfamily, and family) in the training set. We found that this compromise does not make a significant performance difference compared with the ideal experimental design (data not shown).

Performance measures

The overall fraction of catalytic residues in our data set is 1.1%. Considering this highly imbalanced nature, a simple prediction accuracy, or equivalently error rate, is not a sufficient prediction evaluation measure. Using a similar approach that was performed previously (Gutteridge et al. 2003), more sensible measures are computed from the following quantities: TP, number of correctly classified catalytic residues; FP, number of noncatalytic residues incorrectly predicted to be catalytic; FN, number of catalytic residues incorrectly predicted to be noncatalytic; and TN, number of correctly classified noncatalytic residues. Our measure of performance was as follows: Sensitivity = TP/(TP + FN), and Precision = TP/(TP + FP). We also use a ROC curve to compare and visualize predictors' performances. The ROC curve is a two-dimensional plot of predictor performance (Fawcett 2003). A single number summarization of ROC curve is its AUC. The AUC of a predictor is a probability that the predictor ranks a randomly selected positive sample higher than a randomly picked negative sample. The area under the ROC curve is also used to rank the features and identify the best features (Theodoridis and Koutroumbas 1998).

SVM prediction

For SVM prediction (Vapnik 1995), we used a linear kernel (Cristianini and Shawe-Taylor 1999). Throughout the experiments, we used default regularization parameter (*C*). We employed an SVM for prediction using the SVM^{light} and its Matlab interface (Joachims 2002). Since each feature has a different scaling, examples were normalized to [0, 1] interval.

Feature selection

The ROC curve has various utilities. One of them is to compute a measure of class discrimination capability of a feature. That is, for each feature, we rank the feature in a database and compute the area under the ROC curve. This area is the score of the class discrimination capability of the feature. Although the AUC ranges in [0, 1], we adjusted the score by $1 - \text{AUC}$ if $\text{AUC} < 0.5$ since we were interested in relative ordering the samples of this feature and doing this is equivalent to flipping the sign of the samples in the corresponding feature. Therefore, the AUC score ranges in [0.5, 1], and the higher the score the more discriminating the feature is.

Structural genomics targets

Coordinate data for solved structural genomics targets were extracted from the PDB, using the list of solved structures from the TargetDB Web site (http://targetdb.pdb.org/). All residues in all chains were run against the method.

Acknowledgments

We thank J.W. Torrance for providing the training set used in the early study of catalytic sites (Torrance et al. 2005). This research is supported by K22LM009135 (S.D.M.), a grant from IU Biomedical Research Council, the Showalter Trust, and the Indiana Genomics Initiative. The Indiana Genomics Initiative (INGEN) is supported in part by the Lilly Endowment.

References

Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J. 2001. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311:** 395–408.

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

Bagley, S. and Altman, R.B. 1995. Characterizing the microenvironments surrounding protein sites. *Protein Sci.* **4:** 622–635.

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

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.J., and Haussler, D. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.* **97:** 262–267.

Casari, G., Sander, C., and Valencia, A. 1995. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **2:** 171–178.

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

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. 2004. The ASTRAL compendium in 2004. *Nucleic Acids Res.* **32:** D189–D192.

Cristianini, N. and Shawe-Taylor, J. 1999. *An introduction to support vector machines*. Cambridge University Press, Boston, MA.

Elock, A.H. 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.* **12:** 885–896.

Fawcett, T. 2003. ROC Graphs: Notes and practical considerations for data mining researchers. In *HP Labs Technical Report*. HP Laboratories, Palo Alto, CA.

Gutteridge, A., Bartlett, G.J., and Thornton, J.M. 2003. Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330:** 719–734.

Haykin, S. 1999. *Neural networks: A comprehensive foundation*, 2 ed. Prentice Hall, Upper Saddle River, NJ.

Holm, L. and Sander, C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233:** 123–138.

Hua, S. and Sun, Z. 2001. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector approach. *J. Mol. Biol.* **308:** 397–407.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32:** 1037–1049.

Joachims, T. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Jones, S. and Thornton, J.M. 2004. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* **8:** 3–7.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

Krissinel, E. and Henrick, K. 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D Biol. Crystallogr.* **60:** 2256–2268.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. 2005. ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Res.* **33:** W89–W93.

Lichtarge, O., Bourne, H.R., and Cohen, F.E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257:** 342–358.

Mooney, S.D., Liang, M.H., DeConde, R., and Altman, R.B. 2005. Structural characterization of proteins using residue environments. *Proteins* **61:** 741–747.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of protein database for the investigation of sequence and structures. *J. Mol. Biol.* **247:** 536–540.

Nguyen, M.N. and Rajapakse, J.C. 2005. Prediction of protein relative solvent accessibility with a two-stage SVM approach. *Proteins* **59:** 30–37.

Ofran, Y. and Rost, B. 2003. Predicted protein–protein interaction sites from local sequence information. *FEBS Lett.* **544:** 236–239.

Ota, M., Kinoshita, K., and Nishikawa, K. 2003. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.* **327:** 1053–1064.

Petrova, N.V. and Wu, C.H. 2006. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics* **7:** 312.

Porter, C.T., Bartlett, G.J., and Thornton, J.M. 2002. The catalytic site atlas: A resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **32:** D129–D133.

Saunders, C.T. and Baker, D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322:** 891–901.

Theodoridis, S. and Koutroumbas, K. 1998. *Pattern recognition*. Academic Press, San Diego, CA.

Torrance, J.W., Bartlett, G.J., Porter, C.T., and Thornton, J.M. 2005. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologues families. *J. Mol. Biol.* **347:** 565–581.

Vapnik, V.N. 1995. *The nature of statistical learning theory*. Springer Verlag, New York, NY.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337:** 635–645.

Yan, C., Honavar, V., and Dobbs, D. 2004. Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. *Neural Comput. Appl.* **13:** 123–129.

Zhang, Q., Yoon, S., and Welsh, W.J. 2005. Improved method for predicting β-turn using support vector machine. *Bioinformatics* **21:** 2370–2374.