PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Structure-based identification of catalytic residues

Ran Yahalom,[1] Dan Reshef,[2] Ayana Wiener,[1] Sagiv Frankel,[1] Nir Kalisman,[1] Boaz Lerner,[3] and Chen Keasar[1,2]*

[1] Department of Computer Science, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[2] Department of Life Sciences, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel
[3] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ABSTRACT

The identification of catalytic residues is an essential step in functional characterization of enzymes. We present a purely structural approach to this problem, which is motivated by the difficulty of evolution-based methods to annotate structural genomics targets that have few or no homologs in the databases. Our approach combines a state-of-the-art support vector machine (SVM) classifier with novel structural features that augment structural clues by spatial averaging and *Z* scoring. Special attention is paid to the class imbalance problem that stems from the overwhelming number of non-catalytic residues in enzymes compared to catalytic residues. This problem is tackled by: (1) optimizing the classifier to maximize a performance criterion that considers both Type I and Type II errors in the classification of catalytic and non-catalytic residues; (2) under-sampling non-catalytic residues before SVM training; and (3) during SVM training, penalizing errors in learning catalytic residues more than errors in learning non-catalytic residues. Tested on four enzyme datasets, one specifically designed by us to mimic the structural genomics scenario and three previously evaluated datasets, our structure-based classifier is never inferior to similar structure-based classifiers and comparable to classifiers that use both structural and evolutionary features. In addition to the evaluation of the performance of catalytic residue identification, we also present detailed case studies on three proteins. This analysis suggests that many false positive predictions may correspond to binding sites and other functional residues. A web server that implements the method, our own-designed database, and the source code of the programs are publicly available at http://www.cs.bgu.ac.il/~meshi/functionPrediction.

Proteins 2011; 79:1952–1963.

Key words: catalytic residues; functional annotation; support vector machine (SVM); energy terms; spatial averaging; feature selection; class imbalance.

INTRODUCTION

Enzymes mediate virtually all the chemical reactions required for life. They do so by accurately positioning a few catalytic residues within a specific microenvironment known as the active site. Thus, the identification of these residues and ultimately understanding all the electrostatic, entropic, and allosteric aspects of the catalytic machinery are among the major goals of enzymology. Such a complete analysis, however, requires a considerable scientific effort that includes both experimental and computational studies. The current study focuses on the first step towards such studies, namely raising initial hypotheses regarding the identity of the catalytic residues.

Computational methods for the prediction of catalytic residues typically rely heavily on evolutionary inference, either direct annotation transfer from characterized proteins to their homologs, or interpretation of conservation patterns.[1,2] Evolutionary inference, however, has three inherent weaknesses. First and foremost, its basic assumptions do not always apply. Proteins may have different functions even if they are evolutionarily related[3] and residues may be conserved for reasons other than a catalytic role (e.g., structural stability). Second, evolution-based approaches for the prediction of catalytic residues fail when applied to orphan proteins (ORFans),[4] for which we cannot detect homologous proteins in the databases. Third, evolution-based methods are sensitive to errors due to misalignments.

Catalytic residues have characteristic structural features that may augment the evolutionary inference, as they are independ-

Author contributions: R.Y. developed the specialized SVM, did all the performance tests, and drafted the manuscript. D.R. characterized the new structural features, demonstrated their utility, and helped to draft the manuscript. A.W. and S.F. built the web server. N.K. developed the energy functions that underlie the structural features used in this project. B.L. helped in the design of this study, supervised its machine learning aspects, and took active part in the manuscript writing. C.K. conceived this study, coordinated it, and took an active part in the manuscript writing. All authors read and approved the final manuscript.
Additional Supporting Information may be found in the online version of this article.
*Correspondence to: Chen Keasar, Departments of Life Sciences and Computer Science, Ben-Gurion University of the Negev, P.O. Box 653, Beer-Sheva 84105, Israel.
E-mail: chen.keasar@gmail.com
Received 5 October 2010; Revised 14 January 2011; Accepted 28 January 2011
Published online 1 March 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23020

ent of homologous proteins and alignments.[5–7] However, structures are typically determined only after much experimental work has already characterized the protein function. The targets of the structural genomic initiative (SGI)[8] constitute a special case. Their structures are known but their function is often uncharacterized. Thus, their computational annotation is essential to gain full dividends from the SGI endeavor. The general trend in SGI annotation is that evolution is the major source of knowledge and the structural clues complement it. The current study, following Tong *et al.*,[9] goes one step further and uses only structural clues to predict catalytic residues. The major motivation behind this somewhat puritan approach is the need for methods that can handle ORFans, which are not negligible in number among the structural genomics targets.[10] Further, by providing a major challenge to our conceptions, the development of structure-based tools for catalytic residue identification may lead to new insights into the fundamental question of structure-function relationships.

We present here a two-fold strategy for the development of a structure-based predictor of catalytic residues. The first effort focuses on the development and evaluation of a set of structural features that represent the differences between catalytic and non-catalytic residues. The second effort concentrates on a classifier that uses these features to discriminate between the residues.

Our structural features are based on the observations that catalytic residues are typically spatially proximate, deeply occluded,[11,12] and they often destabilize the protein structure.[13–24] Further, there is a wide divergence of catalytic propensities among the different residue types.[25] These observations have already motivated several purely structural prediction methods for the identification of catalytic residues,[12,26–29] catalytic sites,[11,26] and functional residues.[30,31]

Two major problems complicate the use of structural features for catalytic residue prediction. The first is noise. Non-catalytic residues may occasionally have catalytic-like characteristics. Such outliers are relatively rare, but since non-catalytic residues are numerous, compared with the catalytic ones, even a small fraction of them may still result in many false predictions. Previous studies[9,32] coped with this problem by a post-processing step of spatial clustering. Since catalytic residues are clustered within the catalytic site, spatially isolated predictions are most likely false and therefore may be ignored. However, the strength of this solution is limited. Numerous false predictions, due to non-discriminative features, may still be clustered and falsely identified as catalytic. In this study, we present a novel solution that uses the same observation regarding the tendency of catalytic residues to be spatially proximate. However, our approach uses this observation in a pre-processing step. By using spatial averaging over proximate atoms, we smooth out many of these outliers to establish more discriminative features that produce fewer false predictions (Fig. 1).

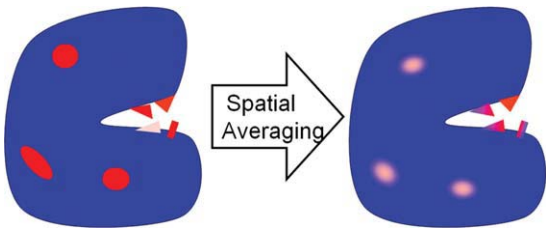

Figure 1

The reasoning behind the use of spatial averaging. In this schematic illustration of an enzyme (left), blue represents regions with no signal for catalytic residues; pink illustrates regions with a weak signal; red illustrates regions with a strong signal. Random non-catalytic outliers (ellipses) are likely to be distributed all over the structure and their noise is likely to be averaged out (right). Catalytic residues (triangles) are typically clustered and thus their signal is more likely to survive the averaging. Further, the averaging may even propagate some signal from nearby binding residues (rectangle). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The second major problem with structural features is that they are often systematically biased by irrelevant factors such as the protein size or even refinement procedure. We are not aware of any previous explicit treatment for this problem. It turns out, however, that feature normalization, that is, replacing the feature with its per-protein Z score (the difference between the feature and its average divided by the feature standard deviation), often results in much less biased features (see Results section).

The second major effort in this study focuses on optimizing a support vector machine (SVM) classifier[33,34] to discriminate residues based on the features selected as part of the first effort. We develop a methodology that tackles the severely low (~1%) ratio of catalytic to non-catalytic residues in enzymes. This so-called "class imbalance problem" has hampered previous studies that used a neural network[32] or SVM[5,9,35,36] for catalytic residue classification. In a nutshell, machine-learning classifiers that aim to maximize the accuracy of the assignment of residues to their correct classes, tend to over-predict the dominant class under class imbalance. If this problem is not considered, the classifier predicts all residues as non-catalytic and thereby achieves a (misleading) ~99% accuracy. To tackle this problem, previous studies used *ad hoc* manipulation of the classifier training phase. They either under-sampled (discarded) a random selection of the non-catalytic residues to reach better class ratio[32,35] or used different penalties for false predictions of catalytic and non-catalytic residues.[9] *Ad hoc* solutions, however, are prone to less-than-optimal performance and, worse, to overfitting. We present here a new, methodologically solid, implementation of these techniques, as well as a new one described below.

A key element in our methodology is the Matthews correlation coefficient (MCC)[37] (see Methods section for definition). MCC was originally used for the assessment

Table I
Summary of Datasets Used in This Study

	Dataset 1(benchmark)	Dataset 2	Dataset 3	Dataset 4
Reference	Designed especially for this research	Petrova and Wu[35]	Gutteridge *et al.*[32]	Tong *et al.*[9]
Annotations	CSA	CATRES	CATRES	CSA
# Enzymes	34	79	159	64
# Catalytic residues	102	252	546	386
# Non-catalytic residues	9546	23,548	57,614	32,977

There are minor discrepancies for the non-benchmark datasets between the counts of residues in this table and those reported in the original studies: Petrova and Wu reported 254 catalytic residues and 23,410 non-catalytic residues, Gutteridge *et al.* reported ~55,000 non-catalytic residues and 550 catalytic residues, and Tong *et al.* reported only 366 catalytic residues. We believe that these discrepancies are due to different preprocessing of the raw PDB files.

of early secondary structure prediction methods, yet another problem that class imbalance complicates.[37] Since it is symmetrical with respect to all possible outcomes of a binary classifier, it is more robust with respect to class imbalance than other common measures of classifier performance.[38] Thus, MCC is often used as a performance measure in catalytic residue prediction.[9,35] Our new methodology takes a further step and also uses MCC for classifier optimization.

Classifier optimization is performed as part of the SVM training phase by thoroughly sampling the parameter space and selecting the set of parameters that maximize MCC. This selection uses a validation set, which is a subset of the training set not used for the actual training; both are independent of a test set used for performance evaluation.[39] In addition to employing MCC for classifier optimization, we systematically under-sample non-catalytic residues and use different penalties for different false predictions. In addition, our new methodology copes with class imbalance through a "probability threshold method."[40] Usually, the probability estimates (i.e., SVM continuous outputs) are thresholded halfway, at a threshold of 0.5, which is a perfect threshold for maximizing the classification accuracy for a balanced problem. Here, for the imbalanced data, we seek for a probability threshold that maximizes the expected MCC.

For performance evaluation, previous studies used MCC and reported either the results of a single test[9] or those achieved using a cross validation experiment,[5,9,32,35,36] which is a more robust approach.[41] However, we noticed that even for cross-validation, the average MCC values are somewhat sensitive to the arbitrary partition of the dataset to training and test sets. Thus, we repeat our cross-validation experiment ten times with different partitions of the dataset and report averages and standard deviations. This protocol guarantees a high degree of replicability (i.e., consistency of results derived for random partitions of the data, as, for example, performed by different researchers).[42]

We evaluate the new prediction method using a benchmark dataset (dataset1 in Table I) that was specifically designed to imitate the most difficult structural genomics scenario. The dataset is non-redundant by both sequence and structure, thus SVM cannot learn any particular active site, only generic characteristics. Furthermore, the dataset does not include ligands bound to the active site, as such ligands might have sharpened structural clues.[16] Thus, this dataset is, by design, harder than the ones used in previous studies. Nevertheless, our prediction results are comparable to, or better than, previously published ones, demonstrating the strength of our structural features and SVM methodology. To further investigate the strengths and weaknesses of our approach, we also provide three detailed case studies taken from the benchmark dataset.

A fair and reliable comparison of our approach with other approaches to pure structure-based prediction of catalytic residues requires that we apply our approach to the same datasets on which the other approaches were applied. Thus, in addition to our benchmark results we also report here our prediction performance using three datasets taken from previous studies.[9,32,35] This direct comparison indicates that for purely structure-based prediction our approach is superior to those of Gutteridge *et al.*,[32] and Petrova and Wu,[35] and comparable to that of Tong *et al.*[9] The relative strengths and weaknesses of our approach and that of Tong *et al.*,[9] as well as prospective directions for further advances are discussed below.

MATERIALS AND METHODS

Datasets

We tested our new methodology on four datasets (Table I). The enzyme structures of all four were extracted from the Protein Data Bank (PDB).[43] Annotations of the benchmark dataset (dataset1) and dataset4 were taken from the Catalytic Site Atlas (CSA)[44] (version 2.2.2). For the other two datasets, we followed the original studies and used the annotations from the catalytic residue dataset (CATRES).[25]

Our benchmark dataset was specifically designed to mimic the structural genomics scenario by excluding hetero-oligomers, catalytic site mutants, and structures with hetero-atoms <5Å from a catalytic residue. The remaining structures were further filtered for high resolution

(up to 2.5Å) and non-redundancy (<30% pairwise-sequence identity and no common CATH fold). The final dataset includes 34 enzyme structures with 102 catalytic residues and 9546 non-catalytic residues. Of these 34 structures, eight were annotated according to literature references and 26 according to PSI-BLAST alignment with homologous literature entries. Two of these 26 structures (1agi and 1o0x) had E.C. (Enzyme Commission) numbers[45] that differed slightly from those of the literature entries they were aligned with, suggesting that the transfer of annotation may be incorrect. However, these structures were included due to their unique CATH topology. Detailed residue information about this dataset (including PDB id, chain, number, type and raw residue feature values, and annotation) is available upon request.

The enzymes of Dataset 2, Dataset 3, and Dataset 4 were taken from the articles of Petrova and Wu,[35] Gutteridge *et al.*,[32] and Tong *et al.*,[9] respectively. The datasets are available upon request.

Performance evaluation

Regarding the class of catalytic residues as positive and that of non-catalytic residues as negative, there are four possible outcomes to a binary (two-class) classifier: true positive (TP), false positive (FP) (Type I error), true negative (TN), and false negative (FN) (Type II error). For example, the outcome is FP when the classifier wrongly predicts a non-catalytic residue as a catalytic residue.

Using these four outcomes, the performance measures considered in this work are:[38]

1. Accuracy (ACC), $\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$;
2. True positive rate (TPR) (sensitivity, recall), $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$;
3. False positive rate (FPR) (1-specificity), $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$;
4. Precision (PCR), $\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$;
5. Filtration ratio (FR), $\mathrm{FR} = \frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$;
6. Matthews correlation coefficient (MCC);$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN}-\mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$; and
7. Area under the curve (AUC).

The range of the first five measures is between 0 and 1. The range of MCC[37] values is between −1 and 1, where a value of −1 indicates total disagreement between the predicted and true classes, a value of 1 indicates a perfect prediction, and a value of 0 indicates a completely random prediction. This measure was also used in the studies of Gutteridge *et al.*,[32] Petrova and Wu,[35] Pugalenthi *et al.*,[36] and Tong *et al.*[9] for performance evaluation but not for classifier optimization. By employing the same measure for both classifier optimization and performance evaluation—as we do in this study—we guarantee that the classifier will be optimized to the task for which it is being evaluated, that is, residue classification. AUC is the integral (area) under the receiver operator characteristic (ROC) curve[46] that plots TPR as a function of FPR. The highest possible AUC value of 1 is achieved when the classifier obtains a TPR of 1 with FPR of 0, that is, a perfect classifier. All the performance measures reported in this study are based on a ten-fold cross-validation (CV10) test.[41] To increase the test replicability, we repeated it for 10 permutations of the dataset and report the average results over this 10 × CV10 procedure.[42]

The SVM classifier

We developed a novel methodology (Fig. 2) for training and optimizing an SVM classifier that aims to alleviate the class-imbalance problem that is detailed in the Introduction. To correctly balance the classifier decisions between the majority non-catalytic class and the minority catalytic class, our methodology is based on: (1) optimization of the SVM classifier and its parameters to maximize MCC, and thereby to balance all types of classification errors. This maximization relies on a validation set that is independent of the training set to avoid over-fitting; (2) under-sampling non-catalytic residues before SVM training by discarding the non-catalytic residues that are least likely to be misinterpreted as catalytic (those that Kubat and Matwin[47] called "redundant"). Under-sampling is performed according to the amount ratio (AR) (Table II) between non-catalytic and catalytic residues providing the greatest value of MCC on a validation set. Values for this parameter are typically between 15 and 93.59 [e.g., see Fig. 3(A)]; 3) setting different penalties to erroneous identifications of catalytic residues and non-catalytic residues according to a penalty ratio (PR) (Table II) between these penalties. That is, we balance the enhanced contribution of the majority non-catalytic residue class to learning the SVM classifier by penalizing PR times more false negatives than false positives. Penalizing different SVM misclassifications differently was already applied to residue classification;[9] however there, the ratio between the penalties was arbitrarily set. We selected the PR value that maximized SVM MCC measured on a validation set. Values for this parameter are typically between 15 and 20 [e.g., see Fig. 3(A)]; and 4) conversion of the probabilistic output of SVM to a binary decision using a probability-threshold (PT) parameter (Table II). SVM outputs that were higher or lower than PT indicated catalytic or non-catalytic residues, respectively. Therefore, PT has implications on the numbers and rates of prediction errors; a too high PT value will increase the rate of catalytic residues that will be missed (i.e., FN error) and a too low value will increase the rate of non-catalytic residues that will wrongly be identified as catalytic (i.e., FP error). Here again, the optimal value of PT was determined as the one maximiz-

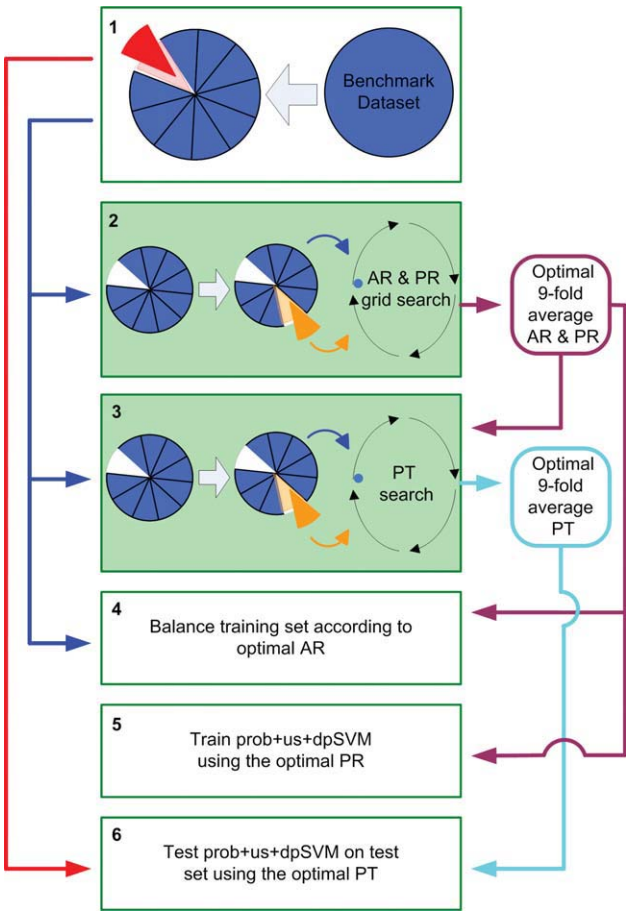

Figure 2

Nested CV10 experiment. For concurrent parameter optimization and performance evaluation, a nested CV10 experiment is applied to the benchmark dataset. The five training steps (rectangles numbered 1–5) and the final test step (numbered 6) are repeated 10 times with alternating test sets. Steps involving inner CV9 experiments are indicated by filled rectangles. Flow of information is indicated by arrows, the colors of which indicate the information type. A summary of the optimized parameters and techniques can be found in Table II. Step 1: The benchmark dataset is divided into 10 equal-sized and disjoint protein subsets, one of which is set aside as a test set (red). The remaining nine are used for training (blue). Step 2: An exhaustive grid search for the optimal amount and penalty ratios (AR and PR, respectively; see text) is conducted on the training set using a CV9 experiment for every AR-PR grid point. This CV9 experiment uses one subset (orange) for validation and the remaining eight for training (blue). The optimal AR and PR are used in the subsequent steps (plum arrows). Step 3: A CV9 search for the optimal probability threshold (PT) to be used in the test step is performed (cyan arrow). Step 4: The optimal AR is used to balance the training set. Step 5: The resulting balanced training set is used to train SVM with the optimal PR. Step 6: The learned SVM model is evaluated on the test set (red arrow) using the optimal PT (cyan arrow). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ing MCC on a validation set. Values for this parameter are typically around 0.05 [e.g., see Fig. 3(B)].

Our methodology uses an SVM having a polynomial kernel of degree two and an error penalty[33] of 55.71 that maximized MCC in a preliminary study. Moreover, before training the SVM classifier, we linearly scaled feature values to the [0, 1] range to prevent the dominance of residue features with the largest numeric ranges.

Manipulations of atom properties

The classification problem that we target is defined at the residue level (i.e., catalytic vs. non-catalytic) and thus our classifier uses residue-level features. However, some of these features are based on atomic properties. Below we present our approach to the conversion of eleven atom properties (see Results section) to residue-level features. Two types of manipulations—spatial averaging and Z score transformation—aid this conversion and considerably increase the power of the atom properties to distinguish between catalytic and non-catalytic residues.

To smooth out noise from sparse non-catalytic residues with catalytic-like properties, we apply spatial averaging at the atom level. For an atom property $f(a)$, we generate a family of averaged properties $f_\alpha(a) = \frac{\sum_{b\in P} f(b)\, e^{-\alpha\, d_{ab}}}{\sum_{b\in P} e^{-\alpha\, d_{ab}}}$, where a and b are atoms in protein P, and d_{ab} is the Euclidian distance between atoms a and b. The parameter $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$ determines the width of the effective neighborhood of atom a, over which the property is averaged (i.e., the neighborhood expands as α decreases). Each value of α produces a different averaged atomic property, and consequently, different structural features of residues.

To minimize systematic biases due to protein size and crystallographic refinement, we linearly transform each atomic property to its Z score, $f_Z(a) = \frac{f(a)-\mu}{\sigma}$, where μ is the protein average value of the property and σ is its standard deviation.

Using these manipulations (applying, or not, Z scoring with either of five neighborhood widths), we convert the basic properties to 110 atomic properties. Finally, to transform an atomic property to a residue feature, we assign each residue the greatest value among its atoms.

Residue features

Reasoning about structure-function relationships and the above manipulations suggest a set of 110 residue features that are based on atom properties. In addition, we consider two additional features that are inherently defined at the residue level, namely catalytic propensity[25] and hydrophobicity score.[25] We estimate the individual predictive power of each feature in two ways. First, we compare the distributions and median values of the feature for catalytic and non-catalytic residues. The distributions of structural features in the residue populations are far from normal. Thus, we use the one-sided Kolmogorov-Smirnov test and two-sided Wilcoxon rank-sum test to validate differences of distributions and medians, respectively. About half of the features indeed show stat-

Table II
Techniques Employed to Cope with the Class Imbalance Problem and the Parameters that Modulate them

Techniques	Abbreviation	Description	Associated parameter
Support vector machine	SVM	The standard binary classifier as described in Burges.[33] It uses an unaltered training set, and errors on both classes are penalized the same	–
Under-sampling	us	Reducing class imbalance in the training set by the removal of non-catalytic residues, such that the ratio between non-catalytic and catalytic residues in the training set is AR	Amount ratio (AR)
Differential-penalty SVM	dpSVM	The ratio between the penalty on catalytic residue misclassification and the penalty on non-catalytic residue misclassification, which are imposed during SVM-classifier training	Penalty ratio (PR)
Probability SVM	probSVM	An SVM classifier that compares the probability of a residue to be catalytic to a probability threshold (PT)	Probability threshold (PT)
Under-sampling and differential-penalty SVM	us+dpSVM	Consecutive application of us and dpSVM	AR and PR
Under-sampling and differential-penalty and probability SVM	prob+us+dpSVM	A us+dpSVM classifier that compares the probability of a residue to be catalytic to PT	AR and PR and PT

istically significant differences between their distributions and medians in catalytic and non-catalytic residues ($P < 0.05$). Second, we generate a ROC curve[46] for a linear classifier based on each feature and measure its AUC. The AUC values allow us to rank the significant features. The higher the value, the more discriminative is the feature.

The subset of features to be used by SVM is selected by a heuristic two-stage scheme. First, we manually cluster the residue features into 13 groups, where catalytic propensity and hydrophobicity constitute singleton groups and each of the other eleven groups includes features that result from the same atomic property by different manipulations. Then, the most predictive residue feature (by AUC) from each group is chosen to represent the group. Second, a wrapper approach[48] is applied to rank the SVM performance using each of the $2^{13} - 1 = 8191$ (Within a subset each feature may or may not appear. The empty subset was not tested) feature subsets. For this analysis, we use the benchmark dataset with PR = 20, AR = 35, and PT = 0.05 (Table II), which resulted in good SVM performance in our preliminary studies. A subset of six features (Table III) is selected for catalytic residue prediction although many other subsets of the original 112 features are almost as successful. The selected set is presented in the beginning of the Results section. For the complete list of features, the reasoning behind them, statistical evaluation of their predictive power, and details of the selection process see Table S1 and Table S2 in the Supporting Information.

Implementation

Protein structure analysis is performed using the MESHI software package.[49] The SVM classifier is trained

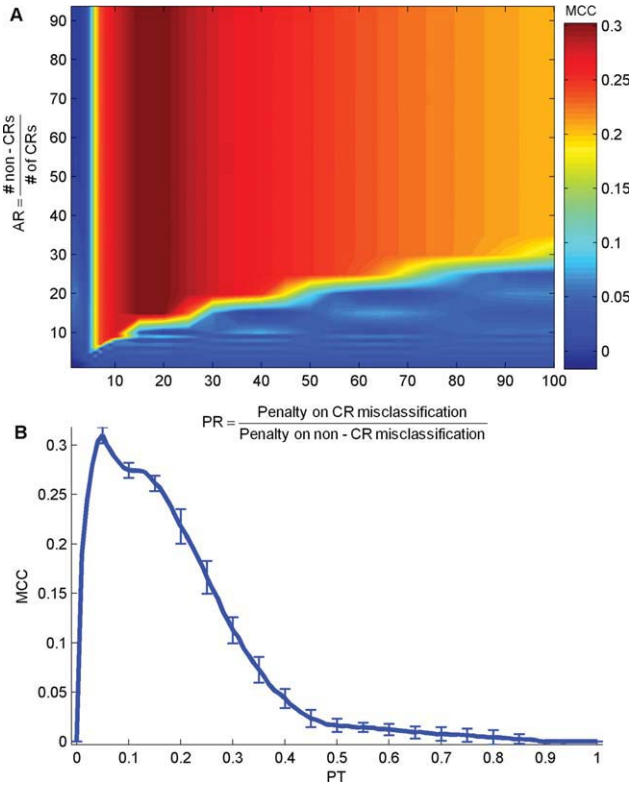

Figure 3
Optimization of amount ratio (AR), penalty ratio (PR), and probability threshold (PT) towards high MCC values. **A**: MCC values that correspond to a wide range of (AR and PR) pairs (Step 2 in Figure 2). A confined area, $15 \leq PR \leq 20$ and $15 \leq AR \leq 93.59$ (dark brown), stands out as optimal with respect to MCC. **B**: MCC as a function of PT given optimal AR and PR values (Step 3 in Figure 2); the highest MCC value is 0.32 for a PT of 0.05. Error bars represent the standard deviations from the average values on the curve (shown intermittently to maintain clarity). All MCC values were obtained by averaging over the CV experiment (Fig. 2). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table III
Performances of the Six Residue Features Used by the SVM Classifier

Feature	*P* Value KS	*P* Value WR	Linear classifier AUC
$brl0.0010_{10Z}$	3.97E-32	0.00E+00	0.859
CatalyticP	3.03E-21	0.00E+00	0.789
Hydrophobicity	1.17E-14	3.65E-12	0.699
$Temperature_{0.01Z}$	3.38E-08	5.65E-09	0.668
$Angle_{0.1Z}$	4.66E-07	7.89E-08	0.654
$Solvation_{0.01Z}$	1.83E-05	3.33E-06	0.633

These features were selected, first as having the highest linear classifier AUC values within their groups and then as the optimal subset of the latter. The complete set of the features appears in the Supporting Information (Table S4).
P values of the Kolmogorov-Smirnov (KS) and Wilcoxon rank-sum (WR) statistical comparison tests, along with values of AUC of linear classifiers discriminating residues, are reported for each of 13 feature group representatives. The rows are sorted in descending order according to AUC. The KS test statistic can be found in Table S4 of the Supporting Information.

and tested using the MATLAB®[50] interface of the LIBSVM package.[51]

All the software developed specifically for this study is publicly available at http://www.cs.bgu.ac.il/~meshi/functionPrediction. This includes a new catalytic residue prediction web server called MESHIfun that is available at http://www.cs.bgu.ac.il/~meshisa/meshifun. The server is trained using our benchmark dataset and it implements the new method. We also provide the raw data that we use for training and testing the SVM classifier to encourage and support testing of other machine learning approaches to the problem.

RESULTS

The structural features used by the SVM classifier

Our SVM uses six residue features that were selected from an initial set of 112 features (see Residue features section). Two of these six features, Hydrophobicity[25] and CatalyticP,[25] are veteran concepts in the field of catalytic residue prediction and represent some *a priori*, structure-independent expectation regarding the chemical nature of catalytic residues. The other four features are novel:

$Angle_{0.1Z}$ is a quadratic measure of angle distortion. The subscript indicates that the distortion values are averaged with an α value of 0.1 and *Z* scored (see

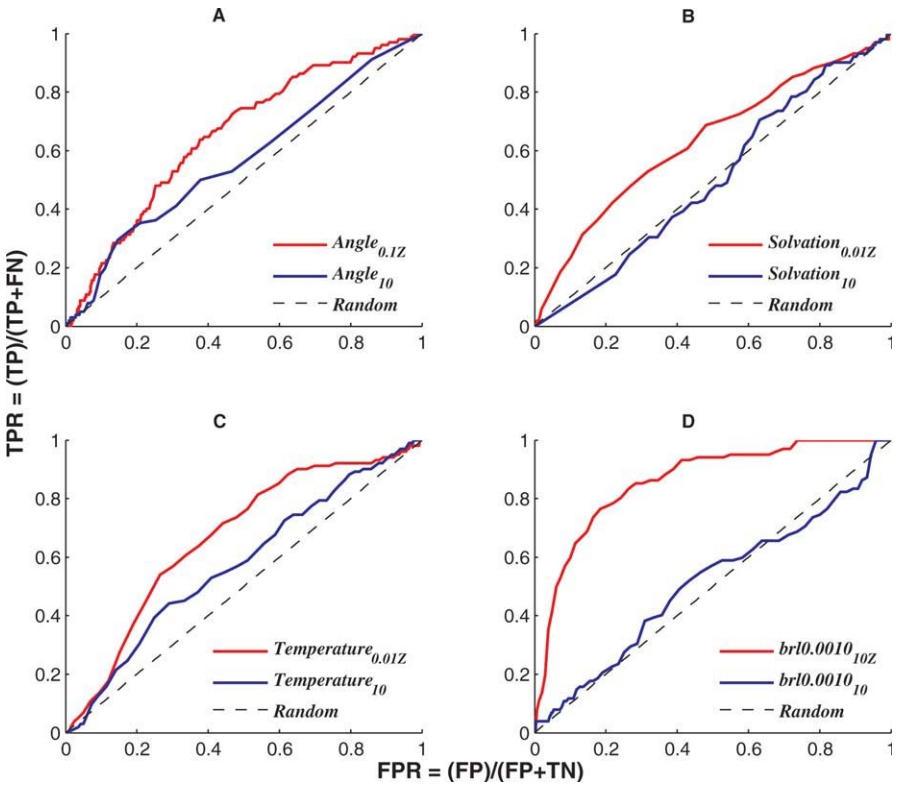

Figure 4
Effect of spatial averaging and *Z* scoring on the predictive power of four of the studied structural features. Linear classifier ROC curves derived for four pairs of residue features: **A**: $Angle_{0.1Z}$ vs. $Angle_{10}$; **B**: $Solvation_{0.01Z}$ vs. $Solvation_{10}$; **C**: $Temperature_{0.01Z}$ vs. $Temperature_{10}$; **D**: $brl0.0010_{10Z}$ vs. $brl0.0010_{10}$. For all four pairs of features, the ROC curves of the manipulated versions (red) are considerably higher than the curves of their non-manipulated counterparts (blue). The diagonal (dashed line) relates to a random classification (a pure guess). Consequently, manipulation improves AUC values (see text). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Manipulations of atom properties section). For catalytic residues, distortions from the ideal values are significantly ($P < 0.02$) more common than for non-catalytic residues. However, due to the far larger number of non-catalytic residues, sparse non-catalytic residues with distorted angles dominate. Further, systematic biases occur due to different crystallographic refinement methods. Thus, the feature without averaging or Z scoring has a low predictive power (AUC = 0.57 of the linear classifier). Spatial averaging and Z score transformation reduce these problems [Fig. 4(A)], resulting in an AUC value of 0.65. Interestingly, the contribution of the two manipulations is not additive; the combined AUC improvement is greater than the sum of the two individual ones (Table S1). The cumulative distributions of the feature after averaging and Z scoring for catalytic (blue) and non-catalytic (red) residues is presented in Figure 5(A). The difference between the distributions is statistically significant ($P < 0.001$, lower than before the manipulations). We are not aware of any previous work that used a similar feature in the context of structure-based functional prediction.

Having a predictive power for this feature is somewhat surprising. The geometry of molecules is determined by the strong interactions between covalently bonded atoms and thus one may expect that all angles would be at their ideal values, up to an experimental error. We see no chemical reason for distortions for either catalytic or non-catalytic residues. Further, the distortions that we observe are smaller than the crystallographic experimental error. We present a speculative explanation of this paradox in the Discussion.

Solvation$_{0.01Z}$ is based on an implicit solvent energy term, which is derived from a non-linear function of the number of carbons in the atom neighborhood (an indicator of burial) and the number of hydrogen bonds that the atom makes. High energies are more common in catalytic residues than in non-catalytic ones, as many of the catalytic residues are polar and pay an energetic price for being occluded in the active site. This difference, however, is not statistically significant. Here, again spatial averaging ($\alpha = 0.01$) and Z scoring considerably improved the predictive power of this term, resulting in a significant ($P < 0.02$) difference between the ROC curves [Fig. 4(B)] and the distributions [Fig. 5(B)] of catalytic and non-catalytic residues, and an AUC value of 0.63. We are not aware of previous studies that used such a term for functional residue prediction, but it is clearly related to the solvent accessibility feature that is often used.[5,32,35]

Temperature$_{0.01Z}$ is based on the inverse of the atom temperature factor. The significant ($P < 0.01$) difference between its distributions for catalytic and non-catalytic residues [Fig. 5(C)] is consistent with a previous study[25] that reported low temperature factors for catalytic residues. The authors of that study suggested that low values of the temperature factor of catalytic residues represent

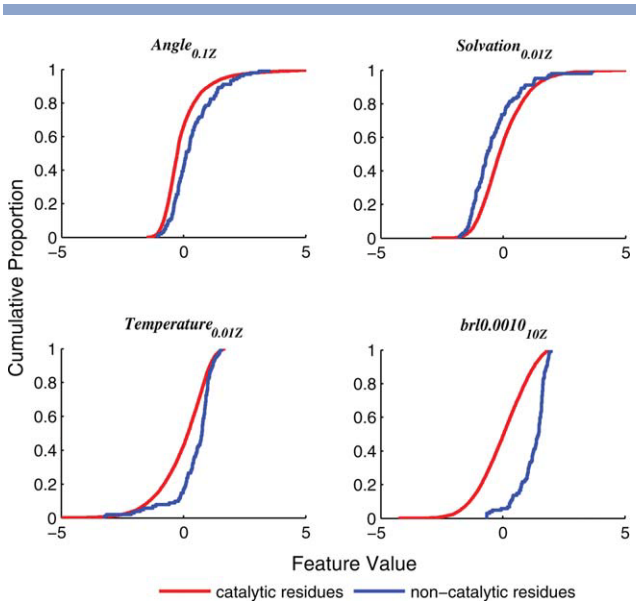

Figure 5

Cumulative distributions of four of the studied structural features for catalytic and non-catalytic residues. Each subplot depicts the difference between the empirical cumulative distribution function (ECDF) of the catalytic residues (blue) and the ECDF of the non-catalytic residues (red) for four studied residue features. The maximal distance between each pair of curves corresponds to the Kolmogorov-Smirnov statistic (see Table III for P values). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

low internal entropy, which is important for the catalytic role. In our study, spatial averaging and Z score transformation improve performance [Fig. 4(C)], increasing the AUC value from 0.58 to 0.67.

brl0.0010$_{10Z}$ is based on atom burial as measured by the number of the atom neighbors, where neighbor contribution is weighted by a Gaussian function of the distance. Catalytic residues are typically buried, but the protein size biases this term; in large proteins the average number of neighbors of all residues is greater than in smaller proteins. This bias renders the number of neighbors non-predictive. Z score transformation solves this problem [Fig. 4(D)], resulting in a significant ($P < 10^{-32}$) difference in distributions [Fig. 5(D)] and a high AUC value of 0.86. In this case, spatial averaging does not provide any improvement to predictive power, as neighboring atoms always have very similar values. For consistency, though, we do perform spatial averaging but with an α value of 10, which practically amounts to no averaging. This term is apparently related to observations that catalytic residues tend to be close to the protein center of mass. The current formalism however is novel.

SVM performance

We use an SVM classifier to discriminate catalytic from non-catalytic residues using the above six features.

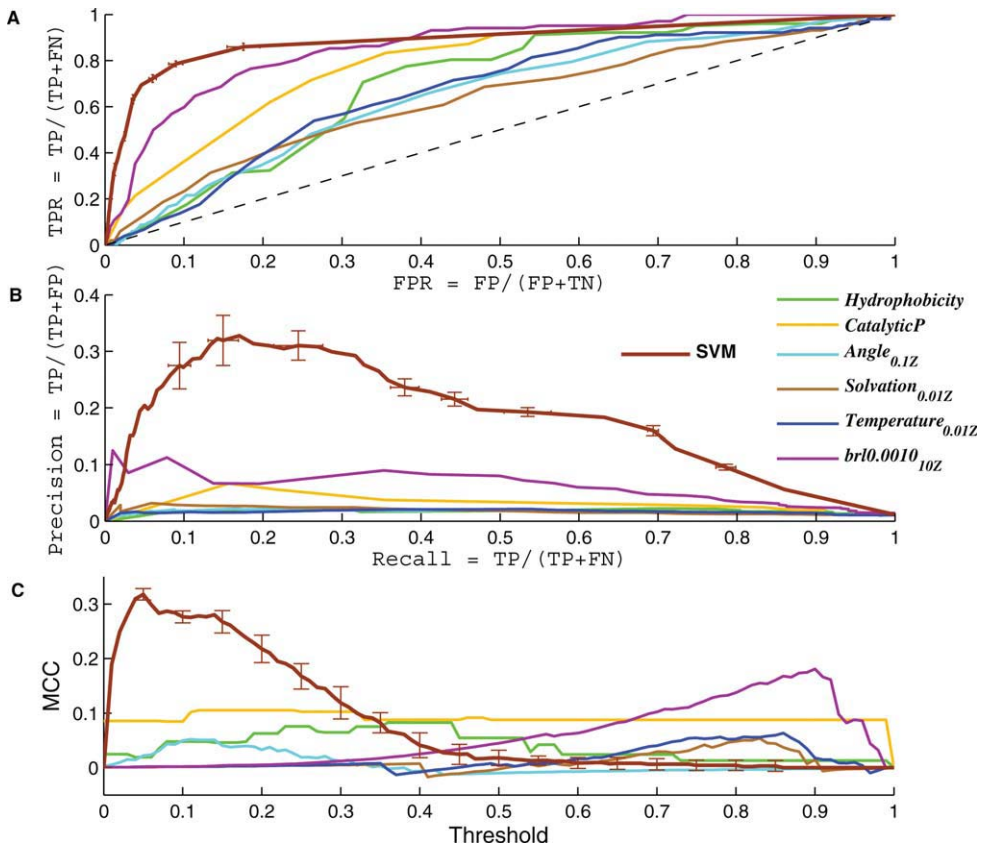

Figure 6

Performance of the SVM classifier on the benchmark dataset. **A**: ROC curve for the SVM classifier (brown) compared to those of linear classifiers for the six individual features. Each point on the SVM curve is an average of 10 CV10 experiments and corresponds to a specific probability threshold (PT). Error bars represent the standard deviations from the averages (not shown for all points to maintain clarity). **B**: Corresponding precision-recall curves are generated in the same way as the ROC curves for the above classifiers. **C**: MCC for SVM as a function of the probability threshold and for the linear classifiers as a function of feature thresholds. The highest MCC for the SVM is 0.32 for a threshold of 0.05. For both B and C, error bars are as in A.

Figure 6 assesses the added value of SVM to our methodology by comparing its performance on the benchmark dataset to those of linear classifiers that are based on the individual features. The three plots - ROC [Fig. 6(A)], precision-recall [Fig. 6(B)], and MCC versus probability threshold [Fig. 6(C)] - indicate an advantage of SVM over the linear classifiers with respect to all performance measures. The ROC curve of SVM is higher than those of the linear classifiers [Fig. 6(A)], resulting in an AUC value (0.892) that is above those of the linear classifiers (0.64–0.86). The precision-recall plots [Fig. 6(B)] show a far larger difference in favor of SVM. The precision values of the linear classifiers are relatively low (at best around 0.1) and almost indifferent to the recall value, whereas that of SVM peaks to more than 0.3. Finally and most importantly, since we focus on optimizing MCC, Figure 6(C) shows that SVM reaches an MCC value above 0.3, where the linear classifier peaks are below 0.2.

Table IV presents performance measure values of SVM using the classifier configuration and parameter settings that maximize MCC on the benchmark dataset. The performances of our new catalytic residue predictor, com-

Table IV

Average Performance Measures (with Standard Deviations in Parentheses; see Methods for Measure Definitions) of the SVM Classifier on the Benchmark Dataset

Performance measure	Performance value
MCC	0.305 (±0.013)
TPR	0.619 (±0.035)
PRC	0.172 (±0.008)
ACC (%)	98.77 (±0.049)
AUC	0.892 (±0.007)
FPR	0.036 (±0.002)
FR (%)	4.3 (±0.2)

The reported measure values are averages over 10 experiments when using the optimal AR, PR, and PT values in each CV fold. The relatively small standard deviations demonstrate the robustness of our method.

Table V
Comparison of our Results with those of Petrova and Wu (P&W)[35] Using Dataset 2 (Table I)

	Our method	P&W–S	Our method	P&W–S+E
The ratio between catalytic and non-catalytic residues in the test set	1:1	1:1	1:92	1:92
MCC	0.686 (±0.02)	0.515 (±0.028)	0.27 (±0.012)	0.23
TPR	0.842 (±0.018)	–	0.43 (±0.053)	0.9
PRC	0.845 (±0.02)	–	0.198 (±0.026)	–
AUC	0.922 (±0.007)	–	0.897 (±0.003)	–
ACC (%)	84.43 (±0.96)	75.55 (±1.382)	98.9 (±0.018)	86.96
FPR	0.164 (±0.027)	–	0.022 (±0.006)	0.13
FR (%)	50.3 (±2)	–	2.6 (±0.6)	–

Similar to Petrova and Wu, we present our performances on both balanced (1:1 ratio of catalytic to non-catalytic residues) and imbalanced (1:92 ratio) test sets. For our method, we report the average performance (with standard deviations in parentheses) of the SVM classifier using a 10 × CV10 experiment. The results of Petrova and Wu were taken from their article. Performance measures that are not explicitly reported by the authors are represented by "–." S represents the use of only structural features. S+E represents the use of both structural and evolutionary data.

pared to those reported by Petrova and Wu,[35] Gutteridge *et al.*,[32] and Tong *et al.*,[9] are shown in Tables V–VII, respectively. In terms of MCC, our predictor is superior to the pure-structure predictors presented by Petrova and Wu and Gutteridge *et al.*, and on a par with the THEMATICS method presented by Tong *et al.* It should be noted, though, that THEMATICS is inherently restricted to ionizable residues. Thus, while the overall performances of the two methods are similar, they differ remarkably in their actual predictions.

DISCUSSION

The wealth of sequence and structural data poured by the recent genomics and structural genomics initiatives impose an unprecedented challenge to the biological community, to characterize as many newly discovered proteins as possible. High throughput computational methods can provide initial working hypotheses regarding the identity of functional residues. Although these hypotheses may include many flaws they can still serve as starting points for systematic studies that combine experiments with time consuming accurate computational methods.[53]

Notwithstanding the strength of other current approaches to the problem of catalytic residue identification, purely structural methods are essential for ORFan structural genomics targets,[52] for which even conservation patterns cannot be derived. The number of such proteins is not negligible. Within the first 687 *Mycoplasma pneumoniae* proteins studied by the Berkley Structural Genomics Center, 54 (7.9%) are ORFans.[10] It is hard at this stage to estimate how representative this fraction is for other centers, and if and how it will change over time. Still, considering the substantial efforts involved in the structural genomics initiative, we believe that even lower numbers justify the development of methods that would maximize the functional insight gained from this endeavor.

This work presents a purely structural approach to the prediction of catalytic residues. Further, as our approach does not rely on similarity to annotate enzyme structures, it may be able to identify novel types of active sites. The structural features that we use encapsulate much of the common knowledge that has accumulated over the years about catalytic residues. Specifically, we use the observations that catalytic residues tend to have unusual structural characteristics,[16,27–29,31] to be spatially clustered,[32] and to be centrally positioned within the protein structure.[11,12]

The current study extends previous ones in several directions. First, it introduces structural clues that were not previously used in this context (e.g., solvation and angle propensity). Second, the study enhances predictive power by spatial averaging and/or *Z* score transformation. Third, to the best of our knowledge, our work is the first that attempts to mimic the structural genomics scenario by excluding from the test set any structure in which the catalytic residues are close to a ligand or mutated. Finally, the current study identifies class imbalance as crucial to accurate and practical catalytic-residue prediction. Thus, it proposes a novel methodology based on the MCC performance measure to integratively (1) optimize the SVM classifier, (2) under-sample non-cata-

Table VI
Comparison of our Results with those of Gutteridge *et al.* (G)[32] Using Dataset 3 (Table I)

	Our method	G_B – S	G_A – S	G_B – S+E	G_A – S+E
MCC	0.272 (±0.004)	0.19	0.23	0.28	0.32
TPR	0.565 (±0.038)	0.41	0.57	0.56	0.68
PRC	0.146 (±0.01)	0.1	0.1	0.14	0.16
AUC	0.901 (±0.003)	–	–	–	–
ACC (%)	99.04 (±0.008)	–	–	–	–
FPR	0.032 (±0.003)	–	–	–	–
FR (%)	3.7 (±0.375)	–	–	–	–

For our method, we report the average performance (with standard deviation in parentheses) of the SVM classifier using a 10 × CV10 experiment. Performance measures that are not explicitly reported by Gutteridge *et al.* are represented by "–." G_B and G_A represent the results before and after spatial clustering, respectively. We did not perform spatial averaging. S represents the use of only structural features (excluding conservation and DOPS scores). S+E represents the use of both structural and evolutionary data.

Table VII
Comparison of our Results with those of Tong *et al.* (Table 2 in Ref. 9) using Dataset 4 (Table I)

	Our method	Tong *et al.*
MCC	0.31 (±0.18)	0.31
TPR	0.57 (±0.35)	0.61
PRC	0.22 (±0.21)	0.2
ACC (%)	96 (±2)	–
FPR	0.029 (±0.18)	0.031
FR (%)	3.4 (±1.8)	3.8

Both methods are purely structural. Our method was evaluated by training and testing on the same protein sets used by Tong *et al.* We report averages and standard deviations over the test set. Performance measures that are not explicitly reported by Tong *et al.* are represented by "–." Detailed predictions of Dataset 4 can be found in Table S10 in the Supporting Information.

lytic residues, (3) set different penalties for erroneous identifications of catalytic and non-catalytic residues, and (4) control the threshold used in classification.

By using this methodology and selecting all parameters to maximize MCC instead of classification accuracy, the SVM classifier proves to be highly successful on four distinct datasets of residues. A direct comparison of the performance of our method to three state-of-the-art structure-based prediction methods shows that our method performs, in terms of MCC, better than two of the methods (Table V and Table VI) and on a par with the third one (Table VII). Further, our results are at least comparable to these methods when they are augmented by evolutionary data.

The study of Tong *et al.*[9] is purely structural and based on perturbed theoretical titration curves using a state-of-the-art electrostatic model. This method is very accurate in identifying ionizable catalytic residues. Although the performance of this method (in terms of MCC) is the same as ours, when averaged over all residue types, it clearly highlights a possible route of improvement to our current group of features. In the initial phase of our study, we tested features that are based on a simple Coulomb electrostatic model. These features had low predictive power and were not selected by the SVM wrapper. One may expect that adopting the electrostatic approach of Tong *et al.* is a promising way to improve our results.

This study, as its predecessors, focuses on the identification of catalytic residues given that the studied protein is an enzyme. This is a somewhat artificial question as often even this bit of information is unknown for structural genomics targets. In a preliminary study, we have tried to perform the same analysis with proteins, such as immunoglobulin, that are not expected to have any enzymatic activity. Qualitatively these proteins were undistinguishable from enzymes in terms of the number of residues predicted as catalytic. Thus, apparently the discrimination between enzymes and non-enzymes cannot be a simple outcome of the current study and requires different classifiers.

CONCLUSION

The new approach of catalytic residue identification presented here combines novel residue features with a novel machine-learning methodology. The features manifest new structural clues for a catalytic role of residues. The machine-learning methodology copes with the notorious class imbalance problem. We demonstrated the utility of this approach using a benchmark that was designed specifically to mimic the most difficult structural genomics scenario. In addition, we tested our approach using benchmarks from the literature. The results obtained using these benchmarks suggest that our approach is a promising alternative to the current methods. Detailed case studies we explored suggest that this approach may also prove handy for binding site prediction.

ACKNOWLEDGMENTS

The authors thank El-ad David Amir for the helpful discussion and insightful comments. The authors are grateful to Ms. Rise Silverman and Ms. Edna Oxman for carefully editing the manuscript. *Funding:* This study was partially supported by the Israeli Science Foundation (grant no. 289/06), and the National Institutes of Health (award no. 1R01 GM081712-01).

REFERENCES

1. Jones S, Thornton JM. Searching for functional sites in protein structures. Curr Opin Chem Biol 2004;8:3–7.
2. Kinoshita K, Nakamura H. Protein informatics towards function identification. Curr Opin Struct Biol 2003;13:396–400.
3. Todd AE, Orengo CA, Thornton JM. Evolution of protein function, from a structural perspective. Curr Opin Chem Biol 1999;3:548–556.
4. Yin Y, Fischer D. Identification and investigation of ORFans in the viral world. BMC Genomics 2008;9:24.
5. Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. Protein Sci 2007;16: 216–226.
6. Cheng G, Qian B, Samudrala R, Baker D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 2005;33:5861–5867.
7. Ota M, Kinoshita K, Nishikawa K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. J Mol Biol 2003;327:1053–1064.
8. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, Gaasterland T, Lin D, Sali A, Studier WF, Swaminathan S. Structural genomics: beyond the human genome project. Nat Genet 1999;23: 151–157.
9. Tong W, Williams RJ, Wei Y, Murga LF, Ko J, Ondrechen MJ. Enhanced performance in prediction of protein active sites with THEMATICS and support vector machines. Protein Sci 2008;17: 333–341.
10. Chandonia JM, Kim SH, Brenner SE. Target selection and deselection at the Berkeley Structural Genomics Center. Proteins 2006;62: 356–370.
11. Ben-Shimon A, Eisenstein M. Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J Mol Biol 2005;351:309–326.
12. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.

13. Warshel A. Energetics of enzyme catalysis. Proc Natl Acad Sci USA 1978;75:5250–5254.
14. Warshel A, Sussman F, Hwang JK. Evaluation of catalytic free energies in genetically modified proteins. J Mol Biol 1988;201:139–159.
15. Herzberg O, Moult J. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins 1991;11:223–229.
16. Heringa J, Argos P. Strain in protein structures as viewed through nonrotameric side chains: II. effects upon ligand binding. Proteins 1999;37:44–55.
17. Beadle BM, Shoichet BK. Structural bases of stability-function tradeoffs in enzymes. J Mol Biol 2002;321:285–296.
18. Gleason FK. Mutation of conserved residues in Escherichia coli thioredoxin: effects on stability and function. Protein Sci 1992;1:609–616.
19. Hibler DW, Stolowich NJ, Reynolds MA, Gerlt JA, Wilde JA, Bolton PH. Site-directed mutants of staphylococcal nuclease. Detection and localization by 1H NMR spectroscopy of conformational changes accompanying substitutions for glutamic acid-43. Biochemistry 1987;26:6278–6286.
20. Meiering EM, Serrano L, Fersht AR. Effect of active site residues in barnase on activity and stability. J Mol Biol 1992;225:585–589.
21. Pakula AA, Sauer RT. Amino acid substitutions that increase the thermal stability of the lambda Cro protein. Proteins 1989;5:202–210.
22. Schreiber G, Buckle AM, Fersht AR. Stability and function: two constraints in the evolution of barstar and other proteins. Structure 1994;2:945–951.
23. Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein stability and protein function. Proc Natl Acad Sci USA 1995;92:452–456.
24. Ferreiro DU, Hegler JA, Komives EA, Wolynes PG. Localizing frustration in native proteins and protein assemblies. Proc Natl Acad Sci USA 2007;104:19819–19824.
25. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002;324:105–121.
26. Wei Y, Ko J, Murga LF, Ondrechen MJ. Selective prediction of interaction sites in protein structures with THEMATICS. BMC Bioinformatics 2007;8:119.
27. Bate P, Warwicker J. Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. J Mol Biol 2004;340:263–276.
28. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
29. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
30. Dessailly BH, Lensink MF, Wodak SJ. Relating destabilizing regions to known functional sites in proteins. BMC Bioinformatics 2007;8:141.
31. Petock JM, Torshin IY, Weber IT, Harrison RW. Analysis of protein structures reveals regions of rare backbone conformation at functional sites. Proteins 2003;53:872–879.
32. Gutteridge A, Bartlett GJ, Thornton JM. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J Mol Biol 2003;330:719–734.
33. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min Knowledge Discov 1998;2:121–167.
34. Vapnik VN. The nature of statistical learning theory. New York, NY: Springer-Verlag New York, Inc.; 1995.188p.
35. Petrova NV, Wu CH. Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. BMC Bioinformatics 2006;7:312.
36. Pugalenthi G, Kumar KK, Suganthan PN, Gangal R. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochem Biophys Res Commun 2008;367:630–634.
37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
38. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412.
39. Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
40. David A, Lerner B. Support vector machine-based image classification for genetic syndrome diagnosis. Pattern Recog Lett 2005;26:1029–1038.
41. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence 1995;14:1137–1145.
42. Bouckaert RR, Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. Adv Knowledge Discovery Data Mining 2004;3–12.
43. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
44. Porter CT, Bartlett GJ, Thornton JM. The catalytic site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32:D129–D133.
45. Webb EC. Enzyme nomenclature 1992. Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology. Toronto: Academic Press Inc.; 1992;339:351.
46. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–874.
47. Kubat M, Matwin S. Addressing the curse of imbalanced training sets: one-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning 1997;186.
48. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.
49. Kalisman N, Levi A, Maximova T, Reshef D, Zafriri-Lynn S, Gleyzer Y, Keasar C. MESHI: a new library of Java classes for molecular modeling. Bioinformatics 2005;21:3931–3932.
50. The MathWorks, Inc., Natick, MA. Matlab. 1984–2007;7.0.0.19901 R14.
51. Chang CC, Lin CJ. LIBSVM: a library for support vector machines, 2001. Available at:http://www.csie.ntu.edu.tw/cjlin/libsvm 2001.
52. Fischer D, Eisenberg D. Finding families for genomic ORFans. Bioinformatics 1999;15:759–762.
53. Olsson MHM, Parson Warshel A. Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. Chem Rev 2006;106:1737–1756.