Published in final edited form as:

Proc Int Conf Mach Learn Appl. 2010 December 12; : 546–551. doi:10.1109/ICMLA.2010.167.

Non-Alignment Features Based Enzyme/Non-Enzyme Classification Using an Ensemble Method

Nicholas J Davidson and

Department of Mathematics, Boise State University, Boise, ID USA

Xueyi Wang

Department of Mathematics and Computer Science, Northwest Nazarene University, Nampa, ID USA

Nicholas J Davidson: njdavidson03@gmail.com; Xueyi Wang: xwang@nnu.edu

Abstract

As a growing number of protein structures are resolved without known functions, using computational methods to help predict protein functions from the structures becomes more and more important. Some computational methods predict protein functions by aligning to homologous proteins with known functions, but they fail to work if such homology cannot be identified. In this paper we classify enzymes/non-enzymes using non-alignment features. We propose a new ensemble method that includes three support vector machines (SVM) and two k -nearest neighbor algorithms (k -NN) and uses a simple majority voting rule. The test on a data set of 697 enzymes and 480 non-enzymes adapted from Dobson and Doig shows 85.59% accuracy in a 10-fold cross validation and 86.49% accuracy in a leave-one-out validation. The prediction accuracy is much better than other non-alignment features based methods and even slightly better than alignment features based methods. To our knowledge, our method is the first time to use ensemble methods to classify enzymes/non-enzymes and is superior over a single classifier.

Keywords

enzyme/non-enzyme classification; ensemble methods; support vector machine; k -nearest neighbour algorithm

I. INTRODUCTION

Understanding protein functions is essential for decoding molecular mechanisms in biological systems, but determining protein functions by experimental methods is difficult and time consuming. In recent years, with the efforts of various structural genomics projects [6, 11], a growing number of protein structures are determined without known functions and developing computational methods to predict protein functions becomes more and more important.

Many computational methods align an unknown protein to homologous proteins with known functions. These methods can be classified into three categories: sequence, motif, and structure. Sequence alignment methods, such as FASTA [23], PSI-BLAST [1], and CS-BLAST [3], check the similarity of two or more proteins' sequences to identify homologues. Motif search methods, such as ScanProsite [10] and Minimotif Miner [24], search for short sequential motifs in sequences and infer protein function by identifying functional motifs. Structure alignment methods, such as DaliLite [16], CE [26], and MAMMOTH [22], check the similarity of two or more protein structures to identify homologues. The alignment methods work well if homologous proteins with known functions can be found, but when

the homology is low or no homology can be found, these methods will have difficulty identifying the unknown protein function.

Some other methods use non-alignment features, such as inherent characteristics of proteins and genes, to predict protein functions. The common characteristics could be sequential, structural, or various other characteristics such as percentage of each amino acid, active site profiles, radius of gyration, subcellular localization, gene neighborhood, phylogenetic profiles, gene expression, protein interactions, sequential features, structural features, or chemical features. Fingerprint Clustering [15] and FFF [5] generate profiles (sequence, structure, and other characteristics) of structurally conserved active sites and classify/predict proteins by comparing and clustering the profiles. PA-SUB [19] and PSORTb [14] predict the location of a protein within a cell from its sequence and can help indicate the protein function. Salgado et al. [25] groups functional related genes by intergenic distances in *E. Coli* genome, based on the observation that genes operated by the same operon have short intergenic distances and tend to belong to the same functional class. Marcotte et al. [20] identifies functional links of proteins in *S. cerevisiae* and assigns a general function to an unidentified protein by analyzing phylogenetic profile, messenger RNA expression, and domain fusion data. Vazquez et al. [27] analyzes protein interaction data and classifies proteins by grouping them into functional categories and minimizing the interactions among them. The non-alignment based methods usually have lower prediction accuracy than alignment based methods.

In the previous work of classifying enzymes/non-enzymes that uses only non-alignment features, Dobson and Doig [12] used a data set of 1178 non-redundant high-resolution protein structures (691 enzymes and 487 non-enzymes) and 52 non-alignment features. They achieved 77% accuracy with all 52 features and 80% accuracy with an optimized subset of 36 features by a single support vector machine (SVM) and a leave- x %-out validation ($x = 1, 5, 10$, or 20). Concu et al. [7] used a similar data set to Dobson and Doig and two electrostatic potentials features and achieved 74.18% accuracy by Markov models.

In other work of classifying enzymes/non-enzymes that uses both alignment and non-alignment features, Borgwardt et al. [4] used the same data set of Dobson and Doig and the same features plus a graph model of sequential and structural edges and achieved 84.04% accuracy by an SVM and a 10-fold cross validation. Lu et al. [18] used a data set of 2443 enzymes and 4886 non-enzymes and functional domain features (multiple sequence alignments and hidden Markov model profiles) and achieved 86.03% accuracy by an SVM and a leave-one-out validation.

In this paper, we propose a new ensemble method to classify enzymes/non-enzymes using only non-alignment features. We use a data set of 1177 non-redundant high-resolution proteins (697 enzymes and 480 non-enzymes) adapted from Dobson and Doig [12] and 56 features. The ensemble method includes three support vector machines (SVM) and two k -nearest neighbor algorithms (k -NN). Two of the three SVMs use the same Gaussian radial basis kernel function but different feature subsets. The other SVM uses a polynomial kernel function and a feature subset. Both k -NNs use the same k value and the city-block metric but different feature subsets. Each SVM or k -NN classifier makes a prediction independently and the ensemble method outputs the majority vote. We test the ensemble method on our data set and achieve 85.59% accuracy by a 10-fold cross validation and 86.49% accuracy by a leave-one-out validation, while the best single SVM and k -NN achieve 84.24% and 82.46% accuracy by a 10-fold cross validation and 84.88% and 83.35% accuracy by a leave-one-out validation.

Our method is the first time to use an ensemble method to classify enzymes/non-enzymes. The prediction accuracy is much better than other non-alignment features based methods [7, 12] and is even slightly better than alignment features based methods [4, 18]. It shows that the ensemble method is superior over a single classifier when different classifiers are combined properly.

It should be noted that although we use a relatively simple feature set and a relatively simple k -NN method, more features or more complex methods may not result in a better accuracy, as Munteanu et al. [21] did a study that used 498 enzymes and 498 non-enzymes with 834 features and methods including general discriminant analysis models, neural networks, and machine learning but achieved worse prediction accuracies than before.

II. METHODS

A. Protein data set and non-alignment features

We use a data set adapted from Dobson and Doig [12] with a few corrections. The data set consists of 1177 non-redundant high-resolution X-ray crystallography structures with 697 enzymes and 480 non-enzymes from Protein Data Bank [2]. Each structure has a resolution $\leq 2.5\text{\AA}$ and an R-factor ≤ 0.25 . No protein chain has a similar fold to another protein chain in the data set (all Z-scores are less than 3.5 [26]).

From the original Dobson and Doig's data set, we moved 7 misclassified enzymes from non-enzyme category in the original Dobson and Doig data set, as shown in Table 1, replaced 9 obsolete PDB IDs, as shown in Table 2, and deleted one enzyme (1BGL) that we cannot calculate the surface pockets using the CASTp program [17].

We use a set of 56 non-alignment features in our study. It includes 52 features from Dobson and Doig [12] with modifications and four new features (number of magnesium ions, number of residues, total surface area, and number of surface pockets). The 56 features are classified into five categories, as shown in Table 3:

- Percentage of residues (20 features; one for each residue type)
- Percentage of surface area (20 features; one for each residue type). We first calculate the protein surface area using STRIDE program [13] and then calculate the percentage.
- Number of heterogens (8 features). We record a binary value (i.e. 1-present, 0-absent) for each heterogen. We add a new heterogen, a magnesium ion, which has more appearance in enzymes (69 or 8.9% of the total) than in non-enzymes (19 or 4.0% of the total).
- Percentage of secondary structure (3 features). We first obtain secondary structure assignment using the STRIDE program [13] and then calculate the percentage of each secondary structure as the percentage of the number of residues in each secondary structure category to the total number of residues. We take α , 3.10 , and π as the helix secondary structure, as did in Dobson and Doig [12].
- Others (5 features). We obtain the number of disulphide bonds and the number of residues from PDB file, the size of largest surface pocket and the number of surface pockets using the CASTp program [17], and the total surface area using the STRIDE program [13].

B. The ensemble method

We use an ensemble method including three support vector machines (SVMs) and two k -nearest neighbor algorithms (k -NNs). Each SVM or k -NN makes prediction independently and the ensemble method outputs the majority vote. Two SVMs use the same Gaussian radial basis kernel function with $\sigma = 3.5$, the other SVM uses a polynomial kernel function with $\text{degree} = 3$, and the two k -NNs use $k = 6$ and a city-block metric, each using a different feature subset.

SVM is a machine learning method widely used for classification and regression [9]. Given two sets of n -dimensional input feature vectors from two classes, an SVM transforms the vectors by a kernel function and finds a hyperplane that separates the two sets of vectors and maximizes margins to the vectors. The SVM is an eager learning method where the training step of finding the best hyperplane is computationally slow but the testing step of making classifications takes very little calculations. We use an SVM function provided in the MATLAB Bioinformatics Toolbox and use a soft margin method [8], which allows small misclassifications to improve the performance.

k -NN is a simple machine learning method used for classification. Given two sets of normalized n -dimensional input feature vectors from two classes, a value k , and a test vector v , a k -NN finds the best k feature vectors having the smallest distances to v and classifies v to the class owning a majority among the k vectors. If k is even, we can break a tie by classifying v to the same class of the feature vector with the smallest distance to v . The k -NN is a lazy learning method, where all the computation is done during classification, so it can be slow if having a large number of feature vectors. In our tests, the k -NN runs very fast because of a relatively small number of 1177 feature vectors.

We use a k -NN function from the MATLAB Bioinformatics Toolbox. Before using the k -NN function, we normalize all features into $[0, 1]$ except the number of surface pockets, which is normalized to $[0, 1.2]$ to emphasize the importance of the feature.

For finding a best feature subset for each individual SVM or k -NN classifier, first we use the individual classifier with all 56 features to classify enzymes/non-enzymes and calculate the prediction accuracy. Next we randomly remove one feature at a time until the accuracy begins to decrease and then randomly add back one feature at a time until the accuracy begins to decrease again, and repeat these two steps until reaching a feature subset with the best prediction accuracy. We run this algorithm many times and record best feature subsets. For the two SVMs and the two k -NNs with the same settings, we use the best and the second best feature subsets found by the above method. The details of the feature subsets are shown in Table 3.

Comparing to some previous work, although we use a relatively simple k -NN method, more complex methods may not result in a better accuracy, as Munteanu et al. [21] did a study that used 498 enzymes and 498 non-enzymes with methods including general discriminant analysis models, neural networks, and machine learning but achieved worse prediction accuracies than before.

III. RESULTS AND DISCUSSION

A. Prediction accuracy by the ensemble method

We test the ensemble method using both n -fold cross validation ($n = 5, 10, 20, 50, 100$) and a leave-one-out cross validation on the data set of 1177 proteins (697 enzymes and 480 non-enzymes). We repeat each n -fold cross validation 100 times and calculate the mean correct rate and standard mean error, as shown in Table 4.

The results show that our method achieves accuracy of at least 84.63% (85.59% in the 10-fold cross validation) and achieves the highest accuracy of 86.49% in the leave-one-out validation. The accuracy is much better than the previous best accuracy of 80% by Dobson and Doig [12] using non-alignment features. The standard mean errors of all validations are less than 0.63%, which shows that the ensemble method is stable with random division of training and testing subsets. With the increment of n in the cross validation, the prediction accuracy increases and the standard mean errors decrease, since a larger n value puts more data into the training set and makes a better prediction.

The results are also slightly better than the methods using both alignment and non-alignment features: 84% achieved by Borgwardt et al. [4], which uses an extra feature of a graph model of sequential and structural edges over the Dobson and Doig's data set and features, and 86% achieved by Lu et al. [18], which uses a data set of 2443 enzymes and 4886 non-enzymes with functional domain features. It shows that the performance of the 4 new non-alignment features is comparable to or better than that of the graph model feature in Borgwardt et al. [4]. The high percentage of prediction accuracy comparable to methods using alignment features shows that our method using the non-alignment features is effective in predicting protein functions. Since methods using alignment features may cause bias to homologous proteins but methods using non-alignment features do not, we suggest combining both methods together to better predict protein functions.

We tested ensemble methods with other combinations and this combination is the best we could find. The ensemble methods with 3 classifiers show little improvement of the prediction accuracy over the best individual classifier. An ensemble method with 7 or more classifiers can achieve better accuracy, but the improvement margin is small and the computation time can be much longer. Adding other prediction methods besides the SVM and the k -NN will be interesting to try in the future.

The best results we achieved using all 56 features are 81.38% for 10-fold cross validation and 81.90% for leave-one-out validation, where we use $k = 3$ and 16 for the two k -NNs in the ensemble method and leave the parameters of three SVMs untouched. We also test our ensemble method using all 52 features from Dobson and Doig and achieve 80.67% for 10-fold cross validation and 81.05% for leave-one-out validation. The results are much better than the previous best accuracies of 77.16% by Dobson and Doig [12] using all their features and show that the ensemble methods are better in classification than individual classifiers in classifying enzymes/non-enzymes.

B. Comparison of each individual classifier to the ensemble method

Figure 1 shows the prediction accuracy of each of the three SVMs and two k -NNs and the ensemble method. For individual classifiers, the two SVMs of the Gaussian radial basis kernel with $\sigma = 3.5$ achieve the two top prediction accuracies, where each uses a different subset of 30 and 35 features. The two k -NNs with $k = 6$ and a city-block metric achieve the next two top prediction accuracies, where each uses a different subset of 31 and 40 features. The SVM of the polynomial kernel with $\text{degree} = 3$ achieves the worst prediction accuracy, which uses a subset of 29 features. For each individual classifier, the prediction accuracy increases with the increment of n in the cross validation.

Except the SVM with polynomial kernel, the prediction accuracy of each of the other classifiers is better than the accuracy by Dobson and Doig [12], which shows that the addition of 4 extra features greatly improves the prediction performance. For the SVMs, the Gaussian radial basis kernel is much better than the polynomial and other kernels in prediction accuracy, which is also shown in other work [4, 12]. For the k -NNs, our tests showed that the city-block metric is much better than the Euclidian distance metric and the

next best k values after $k = 6$ will be 1 and 3. Before using the k -NNs, we first normalize all the features to $[0, 1]$. Since the feature of the number of surface pockets shows a slightly better distinction for enzymes and non-enzymes than other features (see the next section), we emphasize the importance of this feature by assigning a weight of 1.2 to it. It should be noted that the prediction accuracy of the best single k -NN is about 1.7% less than that of the best single SVM, while the k -NN is much simpler method compared to the SVM.

The prediction accuracy of the best individual classifier is about 1.3% less than that of the ensemble method. It shows that each classifier has its own distinct prediction advantage and covers the deficiency of other classifiers. Worse classifiers complement to the best classifier and improve the overall prediction accuracy of the ensemble method.

C. Performance of individual non-alignment feature

We extract 56 non-alignment features from the data set of 1177 proteins, where 4 of them are newly added: number of magnesium ions, number of residues, total surface area, and number of surface pockets. The analysis of 52 features can be found at Dobson and Doig [12]. For the feature of the number of magnesium ions, we believe that the Dobson and Doig [12] tried manganese ions instead of magnesium ions by mistake and discarded this feature, since none of proteins in the data set contains manganese ions. Our analysis shows that totally there are 62 magnesium ions in all enzymes in our data set (0.089 magnesium ion per enzyme) and there are 19 magnesium ions in all non-enzymes (0.0396 magnesium ion per non-enzyme), so an enzyme is more likely to contain magnesium ions than a non-enzyme.

For the other three new features, Figure 2 shows the histograms of the distribution of enzymes and non-enzymes. The histograms are generated by dividing the feature values into 100 bins and counting the number of values in each bin. All three features show larger values for enzymes and smaller values for non-enzymes. The features of the total surface area and the number of surface pockets divide the enzymes and non-enzymes very well. A simple classifier using only the feature of the total surface area achieves 71.20% accuracy (the total surface area of an enzyme ≥ 838 and that of a non-enzyme < 838) and a simple classifier using only the feature of the number surface pockets achieves 72.73% accuracy (the number of pockets of an enzyme ≥ 28 and that of a non-enzyme < 28). These two simple classifiers are actually comparable to the accuracy of 74.18% achieved by Concu et al. [7] that uses two electrostatic potentials features and the same data set. It will be interesting to see if we can improve the accuracy by including these two features once their program is available.

Table 5 shows the frequency of each feature used in the ensemble method. For the four new features, three of them (number of residues, total surface area, and number of surface pockets) are among the most used features. The percentages of the three secondary structures, number of the disulphide bonds, and the size of largest surface pocket are also among the most used features, the heterogens are among the least used features, and some of the percentages of residues and residue surface areas are used more often than others.

IV. CONCLUSIONS

We propose a new ensemble method for classifying enzymes/non-enzymes using non-alignment features. The results are much better than similar methods and are comparable to those methods using alignment features. It shows that the ensemble method is effective and promising in classifying enzymes/non-enzymes. The results also show that the ensemble method is better than an individual classifier, such as an SVM or a k -NN, and should be considered in the future for solving other classification problems.

Classifying enzyme classes is another important classification issue. We plan to work on classifying enzyme classes using the ensemble method by finding good individual classifiers and good feature sets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank for the helpful discussion with Dr. James Foster at University of Idaho. This research is supported by NIH Grant # P20 RR016454 from the INBRE Program of the National Center for Research Resources.

REFERENCES

1. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25(17):3389–3402. [PubMed: 9254694]
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research.* 2000; 28(1):235–242. [PubMed: 10592235]
3. Biegert A, Söding J. Sequence context-specific profiles for homology searching. *Proc Natl Acad Sci USA.* 2009; 106(10):3770–3775. [PubMed: 19234132]
4. Borgwardt KM, Ong CS, Schönaauer S, Vishwanathan SV, Smola AJ, Kriegel HP. Protein function prediction via graph kernels. *Bioinformatics.* 2005; 21(S1):i47–i56. [PubMed: 15961493]
5. Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS. Structure-based active site profiles for genome analysis and functional family subclassification. *J Mol Biol.* 2003; 334(3):387–401. [PubMed: 14623182]
6. Chandonia JM, Brenner SE. The impact of structural genomics: expectations and outcomes. *Science.* 2006; 311(5759):347–351. [PubMed: 16424331]
7. Concu R, Podda G, Uriarte E, González-Díaz H. Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials. *J Comput Chem.* 2009; 30(9):1510–1520. [PubMed: 19086060]
8. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning.* 1995; 20:273–297.
9. Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press; 2000.
10. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 2006; 34:W362–W365. [PubMed: 16845026]
11. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, Lee D, Fiser A, Godzik A, Rost B, Orengo C. PSI-2: structural genomics to cover protein domain family space. *Structure.* 2009; 17(6):869–881. [PubMed: 19523904]
12. Dobson PD, Doig AJ. Distinguishing enzyme structures from non-enzymes without alignments. *J Mol Biol.* 2003; 330(4):771–783. [PubMed: 12850146]
13. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. *Proteins.* 1995; 23(4):566–579. [PubMed: 8749853]
14. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics.* 2005; 21(5):617–623. [PubMed: 15501914]
15. Greenbaum DC, Arnold WD, Lu F, Hayrapetian L, Baruch A, Krumrine J, Toba S, Chehade K, Brömme D, Kuntz ID, Bogyo M. Small molecule affinity fingerprinting. A tool for enzyme family subclassification, target identification, and inhibitor design. *Chem Biol.* 2002; 9(10):1085–1094. [PubMed: 12401493]

16. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. *Bioinformatics*. 2008; 24:2780–2781. [PubMed: 18818215]
17. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci*. 1998; 7(9):1884–1897. [PubMed: 9761470]
18. Lu L, Qian Z, Cai YD, Li Y. ECS: an automatic enzyme classifier based on functional domain composition. *Comput Biol Chem*. 2007; 31(3):226–232. [PubMed: 17500036]
19. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Anvik J, Macdonell C, Eisner R. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*. 2004; 20(4):547–556. [PubMed: 14990451]
20. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature*. 1999; 402(6757):83–86. [PubMed: 10573421]
21. Munteanu CR, González-Díaz H, Magalhães AL. Enzymes/non-enzymes classification model complexity based on composition, sequence, 3D and topological indices. *J Theor Biol*. 2008; 254(2):476–482. [PubMed: 18606172]
22. Olmea O, Straus CE, Ortiz AR. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*. 2002; 11:2606–2621. [PubMed: 12381844]
23. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*. 1988; 85(8):2444–2448. [PubMed: 3162770]
24. Rajasekaran S, Balla S, Gradie P, Gryk MR, Kadaveru K, Kundeti V, Maciejewski MW, Mi T, Rubino N, Vyas J, Schiller MR. Minmotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res*. 2009; 37:D185–D190. [PubMed: 18978024]
25. Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc Natl Acad Sci USA*. 2000; 97(12):6652–6657. [PubMed: 10823905]
26. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*. 1998; 11(9):739–747. [PubMed: 9796821]
27. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*. 2003; 21(6):697–700. [PubMed: 12740586]

Figure 1. Prediction accuracy of the committee method and its 5 individual classifiers in 5-, 10-, 20-, 50-, and 100-fold and leave-one-out cross validation.

Figure 2. Histograms of the distribution of number of residues, total surface area, and number of surface pockets in enzymes and non-enzymes.

TABLE I

7 enzymes misclassified as non-enzymes in the Dobson and Doig data set [12]^{*}

PDB ID	Enzyme category
1A1R	Protease
1A2B, 1AS0, 1G7S, 1NG1	GTPase
1B63	ATPase
1B71	Ferroxidase

^{*} Enzyme categories are obtained from protein annotations (PDB and PubMed abstracts).

TABLE II

9 obsolete enzymes and their replacements

Obsolete PDB ID	Replacement
1AA8, 1AZ9, 1GMS, 1E7T, 1IB3, 1RMI, 1RUX, 2WRP, 4BCL	1VE9, 1XFG, 1WLN, 1GK6, 1M8Z, 1WU3, 1P30, 2OZ9, 3EOJ

TABLE III

The 56 non-alignment features for classifying enzymes/non-enzymes. Features marked as 1 and 2 belong to the subset used by the two SVMs with the Gaussian radial basis kernel in the ensemble method. Features marked as 3 belong to the subset used by the third SVM. Features marked as 4 and 5 belong to the subset used by the two k-NNs.

% of residues	% of surface area	Number of heterogens	% of secondary structure	Others
ALA ^{1,2,3,4,5}	ALA ^{1,2,3,4}	ATP ³	Helix ^{1,2,3,4,5}	Size of the largest surface pocket ^{2,3,4,5}
ARG ^{1,2,3,5}	ARG ^{4,5}	FAD ^{1,2,5}	Sheet ^{1,2,3,4,5}	# of disulphide bonds ^{1,2,3,4,5}
ASN ^{3,5}	ASN ^{1,2,3,4,5}	NAD ^{1,2}	Loop ^{1,2,4,5}	# of residues ^{1,2,3,4,5}
ASP ^{4,5}	ASP ^{1,2,3}	Magnesium ^{2,3}		Total surface area ^{1,2,4,5}
CYS ^{1,2,4,5}	CYS ^{3,4,5}	Copper ^{1,2}		# of surface pockets ^{1,2,3,4,5}
GLN ^{1,2,3,5}	GLN ^{4,5}	Calcium		
GLU ^{1,2,4}	GLU ^{1,2,5}	Heme ^{1,2}		
GLY ^{1,2,3,4,5}	GLY ^{2,3,4,5}	Iron (non-heme)		
HIS ⁴	HIS ^{4,5}			
ILE ^{1,2}	ILE ^{4,5}			
LEU ^{1,2,4,5}	LEU			
LYS ³	LYS ^{1,2,3,5}			
MET ^{1,2,3,5}	MET ^{4,5}			
PHE ^{3,4,5}	PHE ^{1,2,3,4,5}			
PRO ⁵	PRO			
SER ^{1,2,3,4,5}	SER ^{4,5}			
THR ⁵	THR ^{4,5}			
TRP ^{1,2,3}	TRP ^{2,3,5}			
TYR ^{1,2,3,4,5}	TYR ^{2,4,5}			
VAL ^{3,5}	VAL ^{1,2,3,5}			

TABLE IV

Prediction accuracy of the ensemble method

	5-fold	10-fold	20-fold	50-fold	100-fold	leave-one-out
correct rate	84.63%	85.59%	86.06%	86.26%	86.35%	86.49%
standard mean error	0.63%	0.52%	0.45%	0.26%	0.24%	--

TABLE V

Frequency of each feature used in the ensemble method

freq.	features				
	% of residues	% of surface area	# of heterogens	% of secondary structure	Others
5	ALA, GLY, SER, TYR	ASN, PHE		helix, sheet	# of residues, # of disulphide bonds, # of surface pockets
4	ARG, CYS, GLN, LEU, MET	ALA, GLY, LYS, VAL		turn	total surface area, maximum pocket size
3	GLU, PHE, TRP	ASP, CYS, GLU, TRP, TYR	FAD		
2	ASN, ASP, ILE, VAL	ARG, GLN, HIS, ILE, MET, SER, THR	NAD, Mg, Cu, heme		
1	HIS, LYS, PRO, THR		ATP		
0		LEU, PRO	Ca, Fe		