Arch Pharm Res Vol 33, No 9, 1451-1459, 2010
DOI 10.1007/s12272-010-0920-z

Archives of Pharmacal Research
www.springer.com/12272

Support Vector Machine Based Classification of 3-Dimensional Protein Physicochemical Environments for Automated Function Annotation

Hyeyoung Min[1,*], Seunghak Yu[2,*], Taehoon Lee[2], and Sungroh Yoon[2]

[1]College of Pharmacy, Chung-Ang University, Seoul 156-756, Korea, [2]School of Electrical Engineering, Korea University, Seoul 136-713, Korea

(Received July 16, 2010/Revised August 10, 2010/Accepted August 15, 2010)

The knowledge of protein functions as well as structures is critical for drug discovery and development. The FEATURE system developed at Stanford is an effective tool for characterizing and classifying local environments in proteins. FEATURE utilizes vectors of a fixed dimension to represent the physicochemical properties around a residue. Functional sites and non-sites are identified by classifying such vectors using the Naïve Bayes classifier. In this paper, we improve the FEATURE framework in several ways so that it can be more flexible, robust and accurate. The new tool can handle vectors of a user-specified dimension and can suppress noise effectively, with little loss of important signals, by employing dimensionality reduction. Furthermore, our approach utilizes the support vector machine for a more accurate classification. According to the results of our thorough experiments, the proposed new approach outperformed the original tool by 20.13% and 13.42% with respect to true and false positive rates, respectively.

Key words: Protein function, 3-dimensional structure, FEATURE, Dimensionality reduction, Normalization, Support vector machine

INTRODUCTION

The knowledge of protein functions as well as structures is critical for drug discovery and development, because most drug targets are endogenous proteins including enzymes and receptors. In addition, information on the function of drug targets allows us to understand the mechanism of action (MOA) of the drugs and, sometimes, the pathogenesis of diseases.

The recent advances in structural genomics have produced an enormous amount of three-dimensional (3D) protein structural data (Friedberg, 2006), and this ample structural information enables us to understand protein functions which are hardly recognizable when protein structures are analyzed at two-dimensional level. However, although the 3D protein structure can provide a detailed knowledge of its functions, the exponential increase in structural data exceeds the capacity of the current experimental methods to analyze them. Accordingly, many computational methods have been proposed to fulfill such a demand for large-scale 3D structural data analysis.

The fuzzy functional form (FFF; Barker and Thornton, 2003) and JESS (Di Gennaro et al., 2001) methods were developed in order to build structural templates which specify amino acid residues and allowable geometric relationships to define sites of interest using examples of constructed models. The evolutionary trace (ET) method and the ConSurf tool (Landau et al., 2005) exploit evolutionary conservation given the idea that structurally and functionally important regions in the protein remain unaltered during evolution. They determine invariant residues or scores to identify evolutionarily critical residues, respectively, and use them to predict protein function. Query3d is a method that functions both as a structural database management system (DBMS) and as a local structural comparison algorithm (Ausiello et al., 2005). This method saves functional and structural information of all residues in the Protein Data Bank

*These authors contributed equally to this work.
Correspondence to: Sungroh Yoon, School of Electrical Engineering, Korea University, Seoul 136-713, Korea
Tel: 82-2-3290-4826, Fax: 82-2-3290-3844
E-mail: sryoon@korea.ac.kr

(PDB; Berman et al., 2002). New annotation information can be added for an update, and stored information can be queried based on the residues' function and can be compared to find similar residue environments. In contrast to the conventional algorithms which use amino acid sequences for represent-ation, the Surfing the Molecules tool (SuMo; Jambonet al., 2003) represents the protein structure by a set of stereochemical groups. This system can compare protein structures or substructures to find local spatial similarities. The FLORA method is based on structural pattern recognition (Redfern et al., 2009). FLORA creates finds structural motifs which are specific to each of different functional sub-groups within diverse domain superfamilies. An input protein structure is then compared against the motifs and its function is predicted. FLORA differs from other methods by focusing on the domain, rather than on the entire protein complex.

FEATURE (Liang et al., 2003) is a method of analyzing protein structure by finding 'sites' in a protein. The site of a protein refers to a location which has a specific function or structure. In order to characterize protein microenvironments, in FEATURE, a set of concentric shells (of radius 6-10 Angstroms) are placed around a central point and the number of physicochemical properties within the shells in sites of interest and controls (or 'non-sites') are counted. The collected physicochemical properties are then used to build a statistical site model of the 3D distribution of properties by comparing the numerical features of a site with those of a non-site. In contrast to other methods, in FEATURE, no prior information on the properties of the protein of interest is required and it is able to provide information on protein function even when its sequence and structure have no close homology to any other proteins.

However, the FEATURE method is certainly not free from its own limitations and leaves room for improvements. First, when selecting non-sites as controls, FEATURE randomly picks those non-sites whose atom density is similar to that of sites and does not consider the possible similarities in physicochemical properties between sites and non-sites. Thus, there is a chance that the selected non-site has similar physicochemical properties to those of a site, which may obscure the construction of an accurate site model and adversely affect prediction quality.

The second limitation of FEATURE is related to the problem of solving an underdetermined system. The implementation of FEATURE uses 264-dimensional data (44 features per shell and 6 shells in total) to hold structural information of a protein. However, the number of site samples available for analysis is less than 264, and the lack of enough data would cause failure in providing a unique solution. In order to overcome this problem, two approaches can be used. The first is to decrease the dimensionality of data, and the second is to increase the number of site samples. Since it costs more to increase the number of samples in various aspects, it would be more reasonable and effective to reduce data dimensionality.

Another weakness of FEATURE is the performance of the supervised machine learning algorithm employed in data analysis. FEATURE uses the Naïve Bayes classifier (Bishop, 2007) to discern sites from non-sites within protein structure. In general, the Naïve Bayes classifier works reasonably well so long as the conditional independence assumption holds (Domingos and Pazzani, 1997; Frank et al., 2000). The conditional independence assumption means that each feature is conditionally independent of every other feature under a given condition, and in most cases, it simplifies the learning process and makes the Naïve Bayes classifier powerful. However, in some biological conditions under which FEATURE is used, the physicochemical properties comprising features of a vector of a site are not always independent. For example, the physicochemical properties of FEATURE include 'solvent accessibility', 'hydrophobicity', 'hydroxyl oxygen', 'carboxyl oxygen', 'charge' and 'negative charge' (Yoon et al., 2007). Among these, the 'solvent accessibility' and 'hydrophobicity' properties are related, and 'hydroxyl oxygen' and 'carboxyl oxygen' tend to increase 'charge' and 'negative charge'. Accordingly, the conditional independence assumption does not hold in these situations and the classifier fails to work accurately.

As a solution to these problems which are intrinsically present in the FEATURE system and which diminish its performance, we propose in this paper, a new approach with fine tuning of the above mentioned four issues. Firstly, we lowered the possibility of randomly choosing non-sites which are quite similar to sites in physicochemical properties. Secondly, we normalized the 264-dimensional data to prevent errors caused by wide dynamic range of features. Each vector is composed of 264 features and the value of each feature is very complex, including categorical value and numerical value ranging from negatives to positives. As a result, some features with a wide range of numerical values may become decisive, when calculating distance between two sample vectors, and may lead to inaccurate results. The third issue is dimensionality reduction of protein data. We reduced dimensionality without loss of properties, which is critical for determining the characteristics of positive

samples. Finally, we used the support vector machine (SVM) which is one of the most powerful and popular classifiers (Bishop, 2007). Taken together, the results demonstrated that our proposed new approach showed better performance in comparison with the original tool by 20.13% and 13.42% in terms of true and false positive rates, respectively.

MATERIALS AND METHODS

System overview

The overall flow of the proposed method is depicted in Fig. 1. The input is a set of FEATURE vectors, each of which is a collection of local physicochemical properties around a protein residue. More details of the FEATURE vector used in this study will be given shortly. From the input FEATURE vectors, we selected only those that have PROSITE annotations. The PROSITE labels were used to assess the classification performance of the system, which was tested to determine if the PROSITE annotations could be rediscovered. It would be ideal to use experimentally verified annotations, but the number of verified annotations is rather small for a large-scale functional study, and we decided to use the PROSITE annotations instead. Out of the FEATURE vectors with PROSITE annotations, we further identified a set of functional sites and their corresponding non-sites. Both the size of a (non-)site and their statistical characteristics were considered in this process. Then, using the nonparametric Wilcoxon rank-sum test, we reduced the number of dimensions of each vector. This resulted in improved classification performance. Lastly, the classification was performed using the preprocessed site and non-site data. For training SVM classifier, we used 30% of positive and negative samples. The rest was used as validation data for performance evaluation. This training and validation process was repeated 10 times to ensure statistical reliability.

Data collection

By following the procedures described by Yoon et al. (2007), we generated approximately 2,000,000 FEATURE vectors, each of which represented the 44 physicochemical properties listed in the paper by Yoon et al. (2007). This set of vectors was based on approximately 10,000 non-redundant protein chains in the Protein Data Bank (PDB; Berman et al., 2002). Any pair of the structures used to derive the 2 million vectors share smaller than 50% sequence similarity. Since the number of physicochemical properties characterized per shell is fixed in the original implementation of FEATURE, we upgraded it first so that any user-specified number of physicochemical properties per shell can be used.

Out of these 2 million vectors, we selected 9,828 vectors that represent 203 sites. Among these 203 sites, 15 were manually chosen by a domain expert according to the biological importance, and the other 188 sites were selected by grouping the selected vectors according to their PROSITE annotations and then choosing the ones in which the member vector variances were small.

Site data were considered as positive samples, and non-site negative samples were chosen by calculating the Mahalanobis distance (De Maesschalck et al., 2000) between vectors and selecting data with relatively long distance (more details will follow). Each site sample consisted of several vectors, and each vector was made up of 264-dimensional properties (44 property features per shell and 6 shells in total) as explained above.

Site and non-site sample selection

For selecting these functional sites to be used for downstream classification analysis, we considered two aspects. Firstly, we computed the variance of each of the 264 features of the vectors that belong to a site and then created a list of the sites in which such variances are small. Secondly, we removed the sites that have too few vectors from the list. The remaining sites were used as the positive samples.

It is important to note that the positive and negative samples were defined for each functional site. In other words, for a functional site, any vectors that do not belong to the functional site can be included in the negative sample for that functional site. To form a non-site from a functional site, we exploited the Mahalanobis

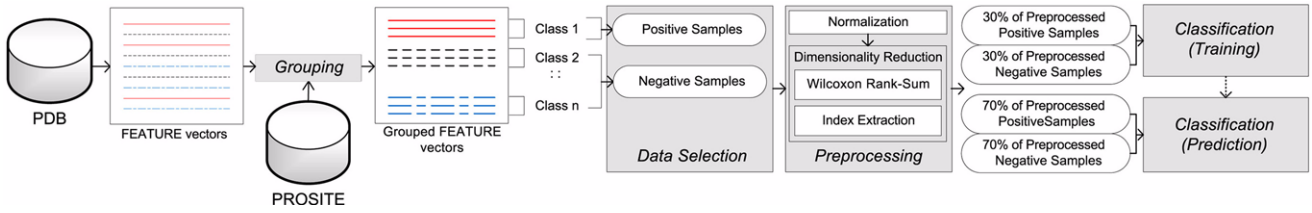

Fig. 1. The overview of the proposed method

distance. For each vector that belonged to the functional site, we calculated its Mahalanobis distance between functional site and all the vectors out of the site. After sorting the distance values in a descending order, we randomly chose the vectors that ranked in the upper 25% of the sorted list.

Data preprocessing

In this step, normalization and dimensionality reduction were performed. Firstly, to handle data with wide range of numerical values, we calculated the average of the values belonging to each of the 264 properties and divided the values in each property by its average. Secondly, we conducted the Wilcoxon rank-sum test and obtained P-values to discern over-represented properties of sites in comparison with those of non-sites. The Wilcoxon rank-sum test is a nonparametric significance test which is based solely on the order in which the observations from the two samples fall, and its statistic is the sum of ranks from observed positive samples. To reduce dimensionality, we accepted over-represented properties and eliminated under-represented properties as well as the properties which were not statistically different between sites and non-sites. The under-represented properties were eliminated since their effect on the classification result was negligible in most cases, in contrast to the over-represented properties.

Classification using support vector machine

We trained the support vector machine (SVM) with the preprocessed data. The SVM classifier works by constructing the optimal hyperplane which maximizes the margin between two input sets (Bishop, 2007). In this study, 30% of positive (sites) and negative (non-sites) sample data were used to train SVM, and the rest of them were used for validation. For fairness, we used the same training and test sets for the compared alternative techniques.

RESULTS AND DISCUSSION

Normalization

When calculating distance metrics between multi-dimensional vectors, wide-ranged numerical values of different features cause erroneous results, and hence, a normalization process is typically required to correct the differences in the range of values. For data normalization, we calculated the average of each feature from 9,828 vectors and divided the value of each feature by its average. As a result, although there were no noticeable differences between the positive and negative samples as raw data, the differences became distinct after normalization (Fig. 2). In addition, as shown in Fig. 3, the normalization of sample site 'PROTEIN_KINASE_STDR' increased the area under the curve (AUC) for the receiver operating characteristic (ROC) curve (Rosner, 2000) by 57.7%, indicating that classification performance was improved through

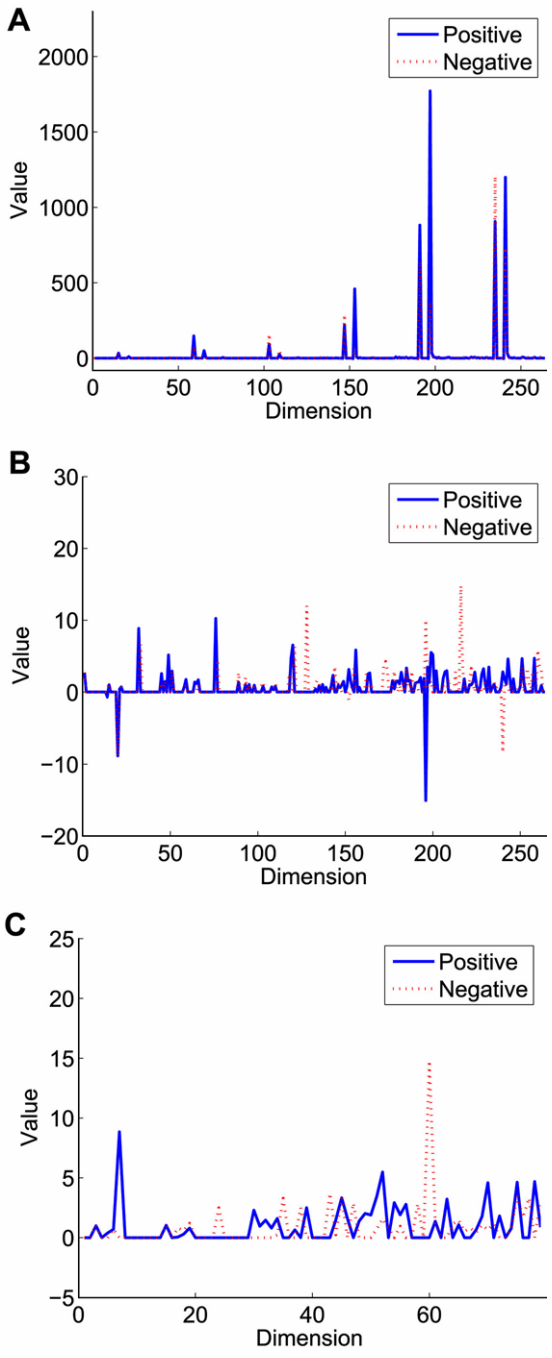

Fig. 2. Effects of preprocessing of vectors representing the PROTEIN_KINASE_STDR site and a randomly selected set of non-site vectors. (**A**) Performance without preprocessing, (**B**) after normalization, and (**C**) after normalization and dimensionality reduction.

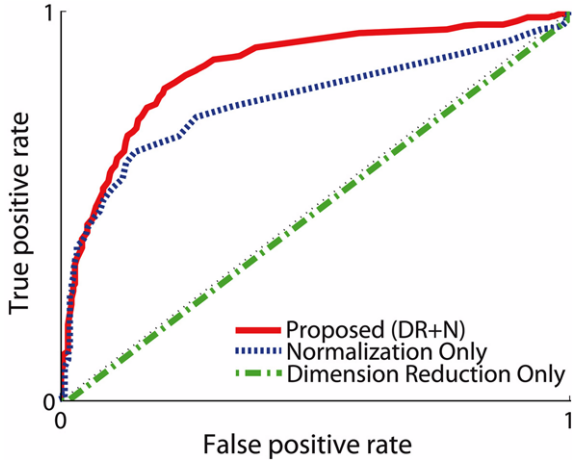

Fig. 3. Effects of preprocessing methods. The site sample named PROTEIN_KINASE_STDR was chosen to compare test results (DR: Dimensionality reduction; N: Normalization).

normalization. Note that AUC is one of the best indicators of the diagnostic accuracy of a classifier (Rosner, 2000). Also note that the effect of normalization was greater than that of dimensionality reduction and that dimensionality reduction could be made more effective by combining it with normalization.

We found out that the normalization process is critical to obtain robust results. This is probably due to the fact that the FEATURE vectors used have various types of attributes, namely integers, floating-point numbers, binary variables and categorical values. Furthermore, the dynamic ranges of many attributes are wide, which results in a need for an effective normalization scheme.

Dimensionality reduction

Another issue we had was that the FEATURE vectors were 264 dimensional, while the number of available site samples were 203. Thus, in order to resolve the problem of solving the underdetermined system caused by the lack of enough data to give a unique solution, we performed dimensionality reduction by using the Wilcoxon rank-sum test. We combined each feature within vectors from sites and non-sites, sorted the combined data set into an ascending order, and summed the ranks for observations from site samples. We then obtained the P-value corresponding to the test statistic which was the sum of ranks from site samples. In order to decide the P-value threshold which can maximally reduce dimensionality without losing representative characteristics of sites, we tested various P-value thresholds and found that the optimal P-value was 0.05. We then built up the list of over-represented properties which had greater numerical values and whose P-values were lower than 0.05 by using the rank-sum test, and only accepted the over-represented ones for constructing modified vectors with a reduced number of dimensions.

Fig. 4 represents the results of the Wilcoxon rank-sum test. Black squares indicate over-represented features and white squares indicate under-represented features or statistically equivalent features between

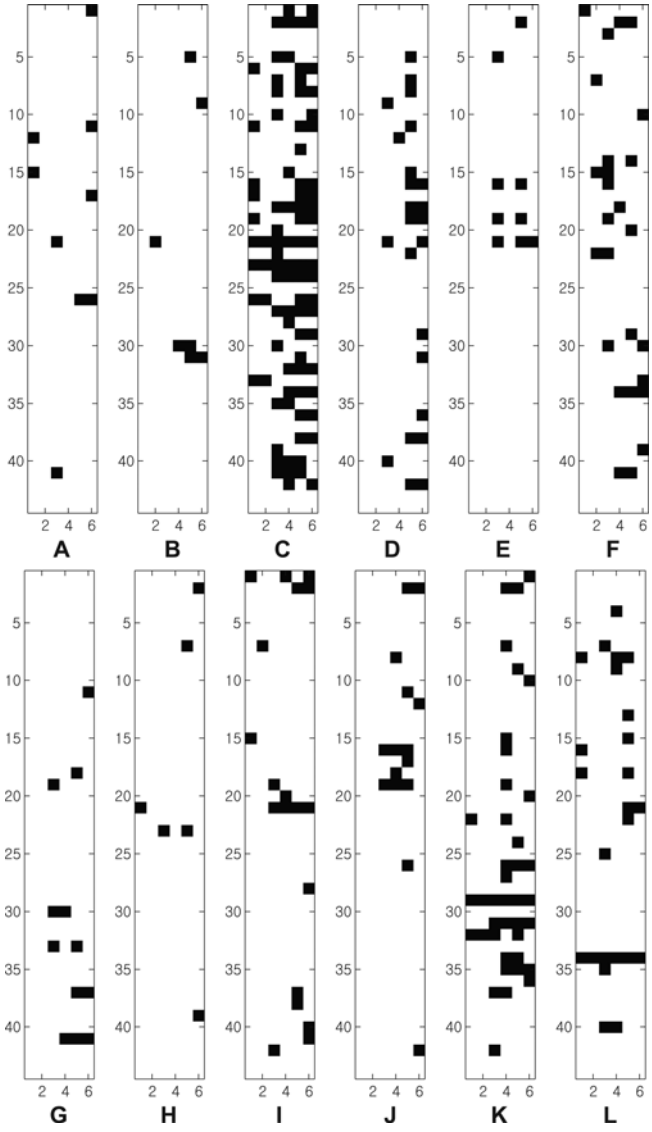

Fig. 4. Fingerprints of results after the Wilcoxon rank-sum test. The horizontal axis represents 6 shells and vertical axis represents 44 properties. Black squares show over-represented properties in positive samples, and white squares represent under-represented or statistically non-significant, equivalent properties. (**A**) 4HBCOA_THIOESTERASE, (**B**) 5_NUCLEOTIDEASE_1, (**C**) SUBTILASE_SER, (**D**) GLUCOAMYLASE, (**E**) PROLINE_PEPTIDASE, (**F**) ADENYLSUCCIN_SYN_1, (**G**) ADOHCYASE_1, (**H**) ASN_GLN_ASE_1, (**I**) ASP_GLU_RACEMASE_1, (**J**) ATPASE_GAMMA, (**K**) IDH_IMDH, (**L**) PHOSPHORYLASE

sites and non-sites. As a result of dimensionality reduction through the rank-sum test, positive and negative samples were better-separated after the normalization and dimensionality reduction processes (Fig. 2C) as compared with the result from the normalization alone (Fig. 2B), let alone the raw data that were not preprocessed (Fig. 2A). Furthermore, when dimensionality reduction and normalization were done simultaneously, the AUC value for ROC was increased more than that with normalization alone (Fig. 3), suggesting that our proposed preprocessing method is effective.

An interesting observation is that the over-represented features play a key role of contrasting sites from non-sites, whereas the included under-represented features rather blur the boundary between sites and non-sites and are not helpful in the classification task. This suggests that we need to focus on finding

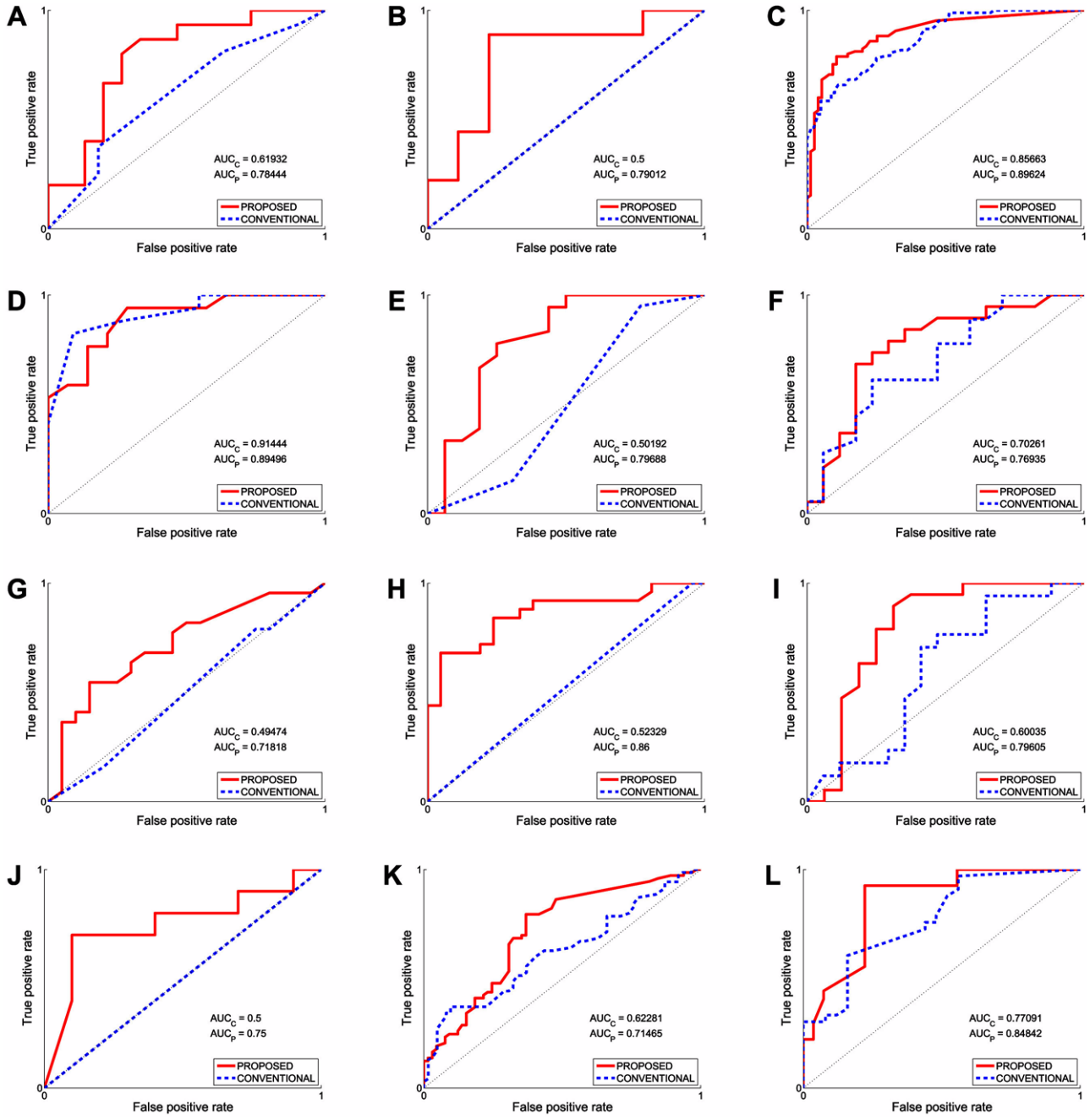

Fig. 5. Comparison of receiver operating characteristic (ROC) curves between the proposed and conventional method. The proposed method shows higher true positive rates and lower false positive rates than the original FEATURE implementation. The annotations used are identical to those described in Fig. 4.

the over-represented features to devise a powerful framework for automated protein function prediction. Moreover, considering the over-represented features only are computationally favorable due to the reduced number of dimensions.

Performance comparison

While the original FEATURE method used the Naïve Bayes classifier for classification of functional sites and non-sites, herein we exploited the support vector machine for more accurate and robust classification. The effectiveness of SVM lies in the selection of its kernel which maps the vectors in a lower-dimensional space into a new higher-dimensional space in which they are linearly separable. We tested the homogeneous polynomial and Gaussian radial basis kernel functions (Bishop, 2007) and decided to use the homogeneous polynomial method kernel due to its simplicity and efficiency. 30% of positive and negative sample data were used for training SVM, and the rest 70% were used for validation. For comparison of our proposed method and the conventional method, we analyzed the data sets with each method and assessed their performances in terms of four widely-used performance measures such as true positive rate (TPR), false positive rate (FPR), recall and precision (Witten and Frank, 2005).

Fig. 5 shows the ROC curves of the proposed and conventional methods. The results demonstrate that the proposed method shows higher true positive rates and lower false positive rates than the conventional FEATURE method. For example, the average of the true positive rates of the proposed method is 80.86% and that of the false positive rates is 29.06% which are better than the values that were obtained from the original FEATURE implementation by 20.13% and 13.42% respectively (Fig. 6). In addition, the AUCs of the proposed method were much higher than those of the conventional method, indicating that the proposed method outperforms the original FEATURE method in terms of the accuracy of classification.

Moreover, we plotted precision versus recall curves (PRC; Fig. 7) which are known to be much more informative when data are highly skewed (Davis et al., 2006). The results display that the proposed method shows higher AUCs for PRC than the conventional method. Additionally, the averages of recall and precision of the proposed method are 80.71% and 79.49% which are greater than those acquired by the conventional method by 19.98% and 17.92%, respectively (Fig. 7).

Lastly, in Fig. 6, we compared the performance of the proposed method to with that of the conventional method in terms of 7 measures including accuracy

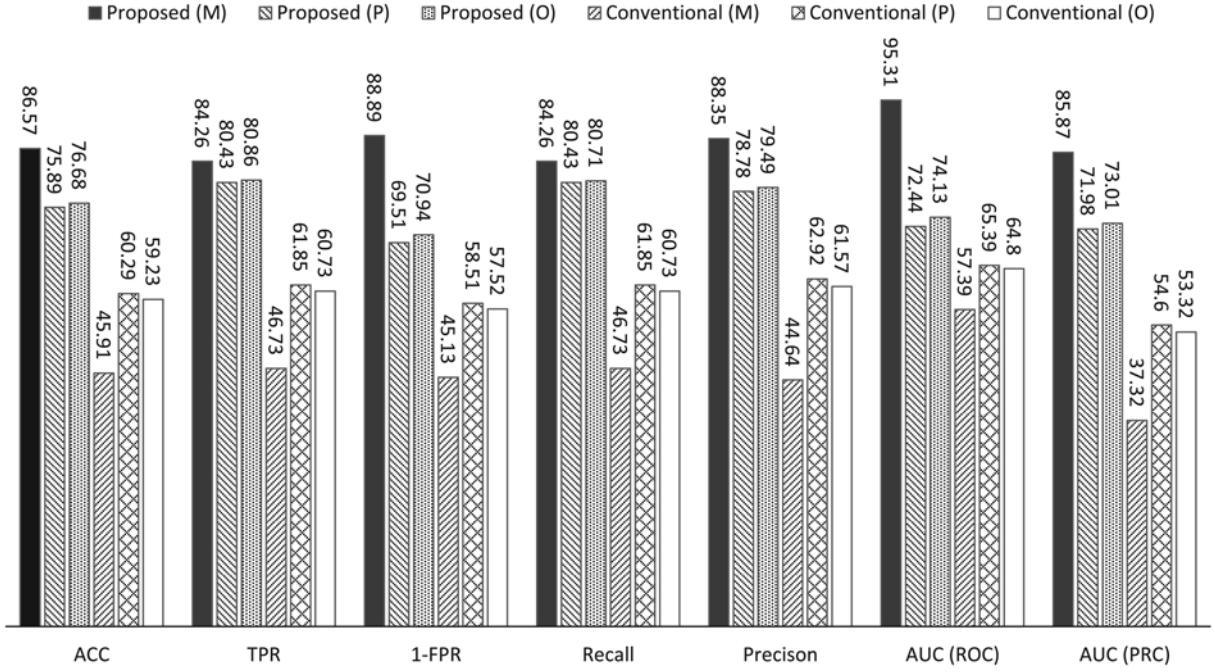

Fig. 6. Comparison of the proposed method with the conventional method in terms of 7 measures. In every measure, the proposed method shows better performance. Accuracy (ACC) = (TP+TN)/(TP+FN+TN+FP); True positive rate (TPR) = TP/(TP+FN); False positive rate (FPR) = FP/(FP+TN); Recall = TP/(TP+FN); Precision = TP/(TP+FP); AUCs for ROC and PRC; M: manually created; P: PROSITE annotated; O: overall

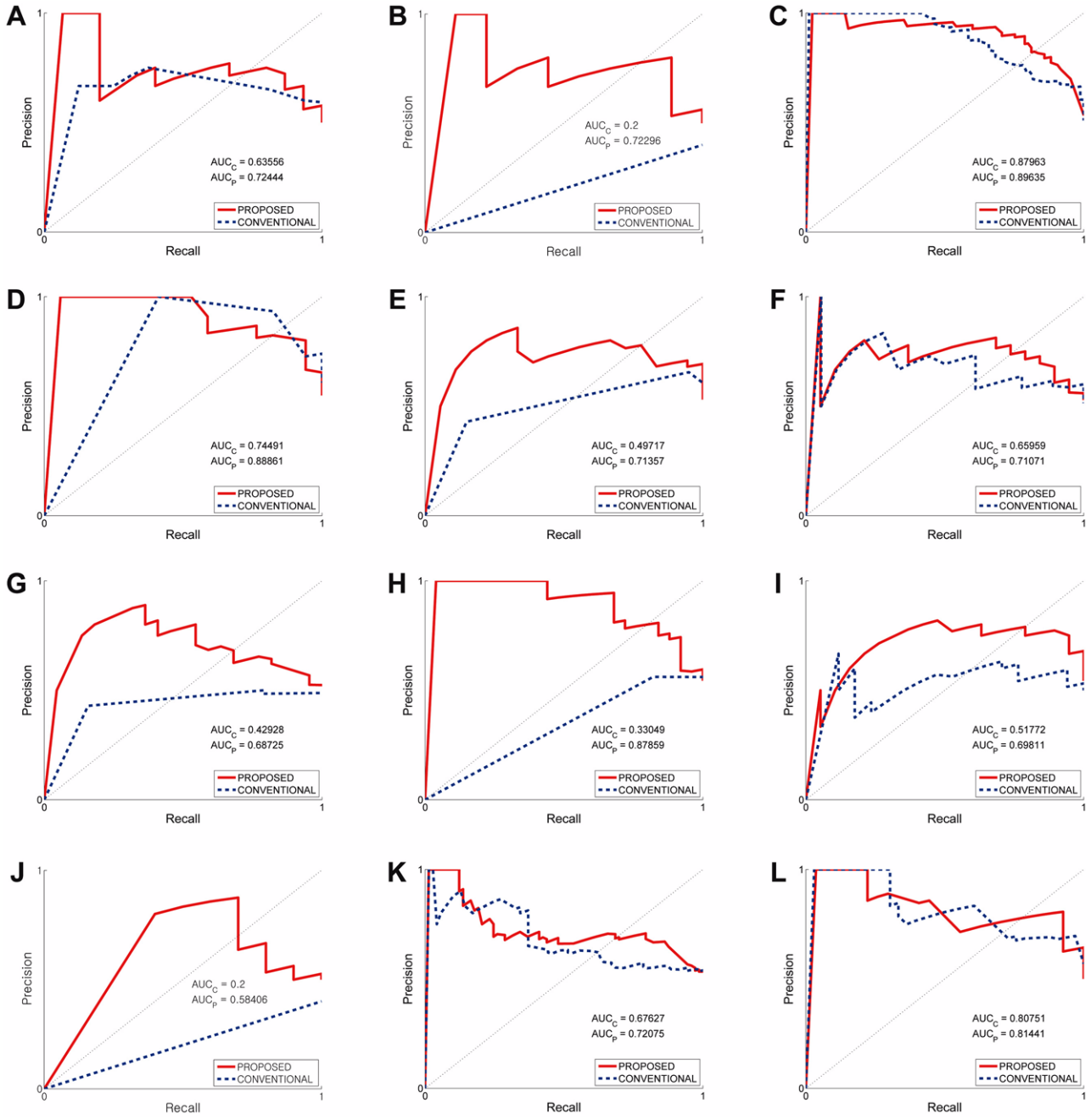

Fig. 7. Precision versus recall curves of the proposed and conventional method. The proposed method shows higher recall and precision than the conventional method. The annotations used are identical to those described in Fig. 4.

(ACC), the true positive rate (TPR), the false positive rate (FPR), recall, precision, and AUCs for ROC and PRC. In every measure, the proposed method shows better performance regardless of the fact whether the site types are manually selected or not (i.e., PROSITE annotated).

CONCLUSION

We have suggested a robust computational method to predict protein functions. This method is based on the FEATURE method, an effective tool for characterizing protein environments. We improved the performance of the original method by employing a statistical preprocessing technique and by replacing the Naïve Bayes method used in the original study

with the support vector machine. As a result, the averages of true and false positive rates of the proposed method are 80.86% and 29.06%, respectively. These values are better than the values obtained from the FEATURE method by 20.13% and 13.42%, respectively. In addition, the proposed approach outperformed the original tool by 19.98% and 17.92% with respect to recall and precision, respectively. Taken together, the results indicate that our approach surpasses the conventional protein function prediction method and would satisfy the great demands of structural genomics.

ACKNOWLEDGEMENTS

This work was supported in part by National Research Foundation (NRF) funded by Korean Government Ministry of Education, Science and Technology (MEST) (No. 2010-0000631 and 2010-0005856).

REFERENCES

Ausiello, G., Via, A., and Helmer-Citterich, M., Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, 6 (Suppl 4), S5-10 (2005).

Baker, J. and Thornton, J., An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. *Bioinformatics*, 19, 1644-1649 (2003).

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. -D., and Zardecki, C., The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, 58 (Pt 6 No 1), 899-907 (2002).

Bishop, C. M., Pattern Recognition and Machine Learning, Springer, Heidelberg, (2007).

Davis, J. and Goadrich, M., The relationship between precision-recall and ROC curves, In Proceedings of the 23rd international conference on Machine learning, ACM New York, pp. 233-240, (2006).

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L., The Mahalanobis distance. *Chemometr. Intell. Lab.*, 50, 1-18 (2000).

Di Gennaro, J., Siew, N., Hoffman, B., Zhang, L., Skolnick, J., Neilson, L., and Fetrow, J., Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.*, 134, 232-245 (2001).

Domingos, P. and Pazzani, M., Beyond independence: Conditions for the optimality of the simpleBayesian classifier. *Machine Learning*, 29, 103-130 (1997).

Frank, E., Trigg, L., Holmes, G., and Witten, I. H.,Naive Bayes for regression. *Mach. Learn.*, 41, 5-15 (2000).

Friedberg, I., Automated protein function prediction-the genomic challenge. *Brief. Bioinformatics*, 7, 225-242 (2006).

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., Pagni, M., and Sigrist, C. J., The PROSITE database. *Nucleic Acids Res.*, 34 (Database Issue), D227-D230 (2006).

Jambon, M., Imberty, A., Deleage, G., and Geourjon, C., A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins*, 52, 137-145 (2003).

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N., ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.*, 33 (Web Server Issue), W299-W302 (2005).

Liang, M., Banatao, D. R., Klein, T. E., Brutlag, D. L., and Altman, R. B., WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.*, 31, 3324-3327 (2003).

Redfern, O. C., Dessailly, B. H., Dallman, T. J., Sillitoe, I., and Orengo, C. A., FLORA: a novel method to predict protein function from structure in diverse superfamilies, *PLoS Comput. Biol.*, 5, e1000485 (2009).

Rosner, B., Fundamentals of biostatistics, Fifth ed., Pacific Grove, California, Duxbury, (2000).

Witten, I. H. and Frank, E., Data mining: practical machine learning tools andtechniques, 2nd ed., Morgan Kaufmann, California, (2005).

Yoon, S., Ebert, J. C., Chung, E. Y., De Micheli, G., and Altman, R. B., Clustering protein environments for function prediction: finding PROSITE motifs in 3D. *BMC Bioinformatics*, 8 (Suppl 4), S10 (2007).