Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 674215, 7 pages
http://dx.doi.org/10.1155/2013/674215

Research Article

Prediction of Substrate-Enzyme-Product Interaction Based on Molecular Descriptors and Physicochemical Properties

Bing Niu,[1] Guohua Huang,[2,3] Linfeng Zheng,[4] Xueyuan Wang,[1] Fuxue Chen,[1] Yuhui Zhang,[5] and Tao Huang[6]

[1] *Shanghai Key Laboratory of Bio-Energy Crops, School of Life Science, Shanghai University, 333 Nancheng Road, Shanghai 200444, China*
[2] *Institute of Systems Biology, Shanghai University, Shanghai, China*
[3] *Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Shanghai 200444, China*
[4] *Department of Radiology, First People's Hospital, Shanghai Jiaotong University, Shanghai 200080, China*
[5] *Department of Neurosurgery, Changhai Hospital, Second Military Medical University, Shanghai 200433, China*
[6] *Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

Correspondence should be addressed to Yuhui Zhang; gong_chang2008@126.com and Tao Huang; tohuangtao@126.com

Received 2 November 2013; Accepted 30 November 2013

Academic Editor: Yudong Cai

It is important to correctly and efficiently predict the interaction of substrate-enzyme and to predict their product in metabolic pathway. In this work, a novel approach was introduced to encode substrate/product and enzyme molecules with molecular descriptors and physicochemical properties, respectively. Based on this encoding method, KNN was adopted to build the substrate-enzyme-product interaction network. After selecting the optimal features that are able to represent the main factors of substrate-enzyme-product interaction in our prediction, totally 160 features out of 290 features were attained which can be clustered into ten categories: elemental analysis, geometry, chemistry, amino acid composition, predicted secondary structure, hydrophobicity, polarizability, solvent accessibility, normalized van der Waals volume, and polarity. As a result, our predicting model achieved an MCC of 0.423 and an overall prediction accuracy of 89.1% for 10-fold cross-validation test.

1. Introduction

With the completion of gene sequencing projects, scientific focus is shifting from the investigation of the proteomics to metabonomics which is of chemical processes involving metabolites. Metabolism consists of almost all of the chemical-chemical reactions or chemical-macromolecules reactions that generally take place within metabolic pathway [1]. Above linked individual interactions form the whole metabolic pathway and interaction network which produce more new complex and higher order structure [2]. Metabolic pathways are sequences of metabolic steps forming highly regulated networks of interacting enzymes and substrates. In metabolic pathways, the substrate is transformed through a series of steps into another chemical, by a sequence of enzymes. Given a substrate and an enzyme, people may wonder whether they can interact with each other or what is the product. Herein, network of interaction of substrate-enzyme-product can provide assistance in R&D of drug. For example, based on interaction of substrate-enzyme-product, maybe people can discover some candidate drug from nature product, and can even predict its potential side effect [3]. Besides this, network of interaction of substrate-enzyme-product can also be applied in evaluating the safety of research of Genetically Modified Food (GMF). By using the network of substrate-enzyme-product, the potential toxicity of product derived from GMF could be predicted. Hence, the interaction network of substrate-enzyme-product will provide us further knowledge and information beyond metabolic pathway.

Due to the complexity of metabolic pathways, it is both time-consuming and costly to determine the interaction of substrate-enzyme-product by experiments. It is in urgent to develop a quick, reliable, and effective approach to predict the interactions among substrate, enzyme, and product.

In this study, we reported a computational approach for predicting the network of substrate-enzyme-product triads based on K-nearest neighbor (KNN) [4–6] algorithm combined with mRMR-IFS feature selection method.

2. Methods and Materials

2.1. Methods

2.1.1. mRMR. Minimum Redundancy Maximum Relevance (mRMR), proposed by Peng et al., is an effective feature-selection method for evaluating the worth of an attribute by considering the minimum redundancy between attributes and the maximum relevance between attributes and targets [7]. More information of mRMR selection algorithm can be found in [7] and related studies [8–19].

2.1.2. KNN. K-nearest neighbors (KNN) is the most basic instance-based machine learning technique classifying objects based on cluster theory [4–6]. KNN recognizes a sample's class according to the label on the K-nearest neighbors. The nearest neighbors of an instance are defined by the Euclidean distance [4]. KNN has been widely applied in the field of biological sciences [20–24]. More details about KNN can be referred to in [25, 26].

2.1.3. Incremental Feature Selection (IFS). First, construct N feature subset by incrementally adding features to D as follows:

$$\begin{aligned} D_0 &= \{f_0\}, \\ D_1 &= \{f_0, f_{i1}\}, \\ &\vdots \\ D_i &= \{f_0, f_1 \ldots, f_i\}, \\ &\vdots \\ D_{N-1} &= \{f_0, f_1, \ldots, f_{N-1}\} \end{aligned} \tag{1}$$

(f_i is the ith feature added into feature subset D).

Second, use KNN method to build the prediction model based on subset D_i and evaluate the model by cross-validation. Then, a classification accuracy curve called IFS curve is attained.

2.2. Materials

2.2.1. Data Preparation. In this study, 14,229 compounds derived from database KEGG (http://www.genome.jp/kegg/) (release 42 in 2006) [27] were collected. After removing the compounds which do not participate in any metabolic reactions which have been supported by experiments, 1326 compounds and 939 enzyme molecules of the human genome participating metabolic reaction were obtained (please refer to Supplemental Material available online at http://dx.doi.org/10.1155/2013/674215).

In metabolic pathway, each substrate binds to one or more enzymes, but the production may not be different. Therefore, substrates and enzymes are subject to be involved in a network of interactions. In this study, substrate, enzyme, and product in each interaction are defined as a positive sample; and those that cannot interact with each other or those interactions that cannot attain the product are defined as negative samples. Triads in the positive set are termed as networking triads, and those in the negative set as nonnetworking triads. These networking triads are supported by solid experiments with 100% credibility by KEGG. As a result, 14,592 networking triads were obtained. To generate the negative datasets, firstly, we built a dataset by randomly combining two small molecules and an enzyme together; then, we removed the 14,592 networking triads. It should be mentioned that although some nonnetworking triads may not be true nonnetworking triads by chance in negative database set, the chance is small. Therefore, the credibility of the negative dataset is also very high. To reflect that the number of networking triads is much less than that of the nonnetworking triads, the negative samples of training set were generated 50 times as many as the positive ones. As a result, the final training dataset contains 14,592 networking triads and 729,600 nonnetworking triads (please refer to supplemental material II and III for the data).

2.2.2. Representation of Compounds. In developing a method for predicting drug-protein interaction, the first problem is how to describe this networking triad correctly as input for the prediction program. It is obvious that the performances of prediction model depend mostly on the features used to describe the molecular structures. In this study, molecular descriptors were applied to reflect the physicochemical and geometric properties of substrates and products which have been applied in our previous studies [28–30]. The values of these molecular descriptors were calculated by program ChemAxon which is available for computing the molecular descriptors [31, 32] (see supplemental material IV). As some molecular descriptors cannot be calculated for some compounds, finally totally 79 molecular descriptors are used in building the model. Before calculating molecule descriptors, the compounds' three-dimensional structures were optimized by using MM+ force field with the Polak-Ribiere algorithm until the root-mean-square gradient became less than 0.1 Kcal/mol. Then, the descriptors were calculated under stable conformation of each molecule based on AM1 semiempirical molecular orbital method at the restricted Hartree-Fock level with no configuration interaction.

2.2.3. Representation of Enzymes. As each protein has its own physicochemical properties, like hydrophobicity, polarizability, and so vent accessibility, it is a good method to describe a protein sequence, and it has been employed for predicting various protein attributes. In this paper, the enzymes are encoded by 132 physicochemical descriptors (amino acid composition, predicted secondary structure, hydrophobicity, polarizability, solvent accessibility, normalized van der Waals volume, and polarity) [33–38] (see supplemental material V) due to its effective and selective ability in the prediction of protein characteristics. More details can be seen in reference [33–38] or our previous study [39].

2.3. Accuracy Measure. Generally speaking, the prediction performance of different discriminative methods is commonly evaluated by the function of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). In this study, we employed sensitivity (SN = TP/[TP + FN]), specificity (SP = TN/[TN + FP]), overall accuracy (ACC = [TP + TN]/[TP + TN + FP + FN]), and Matthew's correlation coefficient (MCC) to measure the prediction. The MCC can be represented as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TP} + \text{FP})}}. \tag{2}$$

3. Results

In the recent years, many efforts have been made in feature selection [40–46]. In this study, mRMR method was applied to search for a subset with optimal features. After mRMR calculation, two tables are attained (see supplemental material VI). One is called MaxRel feature table that ranks the features based on their relevance to the class of samples and the other is called mRMR feature table that lists the ranked features by the maximum relevance and minimum redundancy to the class of samples.

Then, IFS method is applied based on mRMR feature table. From Figure 1, it can be found that while adding new feature continually, the value of MCC increased, although during this process, the value of MCC decreased at some point. While the number of features reaches 160, the value of MCC is 0.423, the highest point. Then, the value of MCC begins to decrease. Hence, the subset containing these 160 features is considered as an optimal subset which is derived from original data set containing 290 features. These features selected are irrelevant to each other but relevant to the target.

Based on the 160 features, predicting model of network of substrate-enzyme-production interaction could be built.

Ten folds cross-validation test, which is applied in many other applications [36, 47–52], is adopted in this study to validate the model's prediction accuracy. During 10-fold cross-validation test, the datasets are divided into 10-folds, a model is built with N-1 fold samples and the 10th fold data are treated as unseen data, which is used for the prediction as the testing data. Each fold is left out from building the model and predicted in turn. The predictive ability is evaluated by averaging the correct prediction rates of the 10-fold data. Table 1 lists the prediction results while using KNN method.

To evaluate our feature selection method, we compared the prediction results generated by final optimal subset and the original data set with 10-folds cross validation test (see Table 1). Table 1 shows that the prediction results of the 10-folds cross-validation test improved after applying feature selection. This demonstrates that maybe some features are redundant and interfering to each other in the original dataset; hence, it is better to remove some of them. Furthermore, the number of features in the final subsets is 55% of the original feature set. This result suggests that mRMR feature selection approach could make a good optimization and improve the accuracy of prediction for substrate-enzyme-product interaction.

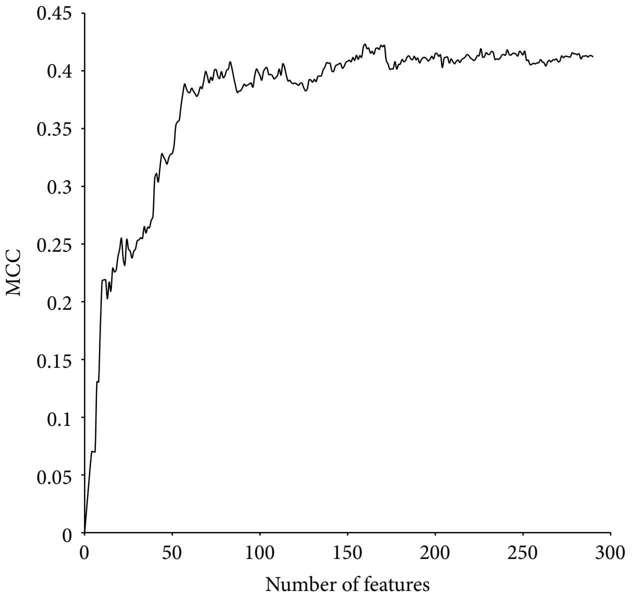

Figure 1: The curve of the 290 prediction models using IFS.

Table 1: Prediction accuracies of different dataset with KNN.

Dataset	10-folds cross-validation test			
	SN (%)	SP (%)	ACC (%)	MCC
Original dataset	53.71	92.4	88.9	0.412
Optimal dataset	55.2	92.4	89.1	0.423

4. Discussion

The selected 160 features in the final subset can be clustered into the following ten categories: elemental analysis, geometry, chemistry, amino acid composition, predicted secondary structure, hydrophobicity, polarizability, solvent accessibility, normalized van der Waals volume, and polarity (see Figure 2). The former three kind features are molecular descriptors which are of substrate and product, and the left seven kind features are of enzyme.

According to the distribution of features of compounds (substrate and product) and enzymes, it shows that enzymes contribute more to the interaction process. Further calculating the proposition of the selected features to the original features, it is found that the proposition of enzyme feature (92/132 = 0.70) is higher than the proposition of compound

Table 2: Top 80 features rank according to their correlation to target.

No.	Name	Categories	No.	Name	Categories
1	Polarity	Polarity	41	Amino Acids Composition Cys	Amino acids composition
2	Substrate_Polarizability	Chemical	42	Polarizability	Polarizability
3	Solvent accessibility	Solvent accessibility	43	Polarizability	Polarizability
4	Solvent accessibility	Solvent accessibility	44	Amino Acids Composition Ile	Amino acids composition
5	Secondary structure	Secondary structure	45	Hydrophobicity	Hydrophobicity
6	Normalized Van Der Waals volume	Normalized Van Der Waals volume	46	Secondary structure	Secondary structure
7	Normalized Van Der Waals volume	Normalized Van Der Waals volume	47	Substrate_Stereo DoubleBondCount	Geometry
8	Secondary structure	Secondary structure	48	Normalized Van Der Waals volume	Normalized Van Der Waals volume
9	Secondary structure	Secondary structure	49	Substrate_Smallest RingSystemSize	Geometry
10	Substrate_Log$P$	Chemical	50	Substrate_Smallest RingSize	Geometry
11	Substrate_CComposition	Elemental analysis	51	Substrate_Rotatable BondCount	Geometry
12	Amino Acids Composition Asn	Amino acids composition	52	Substrate_H Composition	Elemental analysis
13	Polarity	Polarity	53	Amino Acids Composition Thr	Amino acids composition
14	Hydrophobicity	Hydrophobicity	54	Polarizability	Polarizability
15	Substrate_MinZ	Geometry	55	Amino Acids Composition Leu	Amino acids composition
16	Solvent accessibility	Solvent accessibility	56	Amino Acids Composition His	Amino acids composition
17	Polarity	Polarity	57	Substrate_CarboAliphatic RingCount	Geometry
18	Hydrophobicity	Hydrophobicity	58	Product_HComposition	Elemental analysis
19	Substrate_VanDerWaals SurfaceArea	Chemical	59	Polarizability	Polarizability
20	Amino Acids Composition Asp	Amino acids composition	60	Normalized Van Der Waals volume	Normalized Van Der Waals volume
21	Hydrophobicity	Chemical	61	Amino Acids Composition Gln	Amino acids composition
22	Substrate_OComposition	Elemental analysis	62	Normalized Van Der Waals volume	Normalized Van Der Waals volume
23	Solvent accessibility	Solvent accessibility	63	Polarizability	Polarizability
24	Secondary structure	Secondary structure	64	Amino Acids Composition Lys	Amino acids Composition
25	Amino Acids Composition Ser	Amino acids composition	65	Polarizability	Polarizability
26	Substrate_Water AccessibleSurface Area Negative	Chemical	66	Amino Acids Composition Tyr	Amino acids composition
27	Secondary structure	Secondary structure	67	Amino Acids Composition Arg	Amino acids composition
28	Hydrophobicity	Hydrophobicity	68	Secondary structure	Secondary structure
29	Substrate_FusedRingCount	Geometry	69	Polarizability	Polarizability
30	Substrate_Carbo RingCount	Geometry	70	Normalized Van Der Waals volume	Normalized Van Der Waals volume

Table 2: Continued.

No.	Name	Categories	No.	Name	Categories
31	Amino Acids Composition Glu	Amino acids composition	71	Polarity	Polarity
32	Hydrophobicity	Hydrophobicity	72	Normalized Van Der Waals volume	Normalized Van Der Waals volume
33	Polarizability	Polarizability	73	Product_NComposition	Elemental analysis
34	Polarity	Polarity	74	Solvent accessibility	Solvent accessibility
35	Normalized Van Der Waals volume	Normalized Van Der Waals volume	75	Product_Hetero AliphaticRingCount	Geometry
36	Substrate_Fused Aliphatic RingCount	Geometry	76	Substrate_CarboAromatic RingCount	Geometry
37	Polarizability	Polarizability	77	Substrate_PComposition	Elemental analysis
38	Secondary structure	Secondary structure	78	Hydrophobicity	Hydrophobicity
39	Substrate_RingCount	Geometry	79	Product_CComposition	Elemental analysis
40	Amino Acids Composition Pro	Amino acids composition	80	Normalized Van Der Waals volume	Normalized Van Der Waals volume

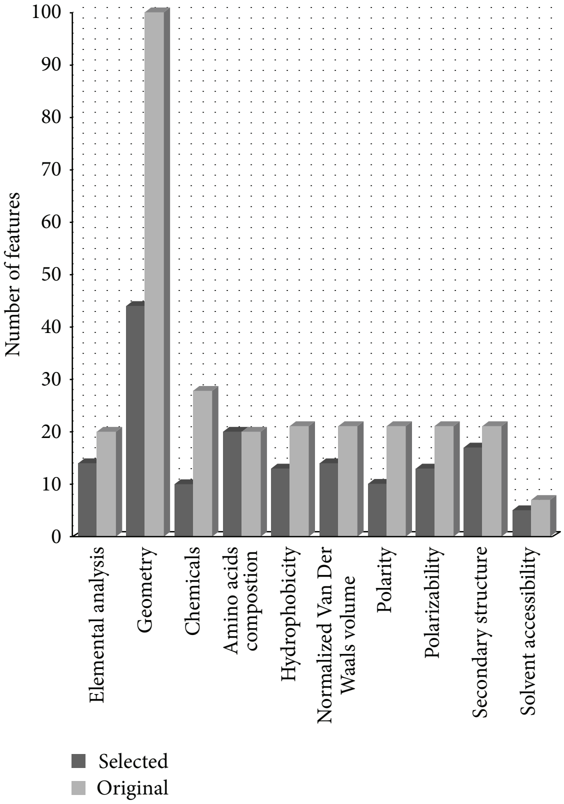

Figure 2: Feature distribution.

feature (70/158 = 0.44). Table 2 also shows that several enzyme features are in the top ten and top twenty features. This result suggests that enzyme-centric features make more contributions to our proposed interactions network of substrate-enzyme-product.

From Table 2, it can be further found that for compound features, there are much less features of product than features of substrate and enzyme in the top fifty features. This is because during the interaction of substrate-enzyme-product, substrate and enzyme determine the products, and changing substrate or enzyme could result in a different product.

According to the distribution of features in Figure 2, it can be found that the number of geometry features is more than that of the other kind features. In this regard, geometry features have great effect and contribute to the substrate-enzyme-product interaction not only in substrate features but also in product features. However, from MaxRel feature table, it can be found that there are not many geometry features appearing in the top ten features. Therefore, we feel interesting of this problem. Actually, the order of geometry features is not incompatible with its distribution. Geometry features contain information of the structure of a molecule like the volume, size, and shape which leads to steric hindrance and steric resistance. These factors are of great importance in substrate-enzyme-product interaction. Only correctly three-dimensional size and shape molecule can interact with enzyme according to the Lock and Key Theory. Meanwhile, steric hindrance or steric resistance affect the substrate-enzyme-products' interaction as some big functional groups like aromatic ring prevent interaction. On the other hand, these functional groups also provide key interactive force to enzyme like heteroaromatics ring's π-π stacking interaction to enzyme's functional site. The substrates and products are varied and diverse greatly in structure. And it is difficult to describe their structure with only one or two descriptors. Hence, more geometry features could better extract the information of compounds' structure. This is why though single geometry feature has no strong relevance to the interaction, the overall contribution of the forty-four geometry feature can often be crucial to the interaction.

Figure 2 also shows that amino acid compositions and second structure occupied important propositions among the ten types' features. Amino acid composition in the binding site contributes a lot in substrate-enzyme-product interaction because it could affect the state energy. Some experiments have verified the importance for amino acid compositions in protein related interaction [53–55]. For example, Tyr265 plays a central role in enzyme alanine racemase's binding to L-alanine and pyridoxal 5-phosphate [54]. Hence, for a unique structure, the amino acid composition plays the essential role in the interactions. Secondary structure is considered as an important property in many protein related problems, since the shape and biological function of a protein are mainly determined by its secondary structures. Secondary structure features reflect the steric structure of protein. According to the Lock and Key Theory, the size and shape of substrate were rigid and restricted by enzyme. Accordingly, secondary structure has relatively more impact on the determination of substrate and product.

5. Conclusion

In this paper, a feature selection method called mRMR combined with IFS was applied to dataset of substrate-enzyme-product interaction which is encoded with molecular descriptors of substrate/product and 132 physicochemical protein descriptors. As a result, we find that enzymes are essential in substrate-enzyme-product interaction; 160 important features were abstracted from 290 features. Based on the above findings, we also used KNN method to build a prediction model of substrate-enzyme product interaction. Based on the prediction results, it is expected that molecular descriptors and 132 physicochemical protein descriptors can be served as an efficient coding method for network of substrate-enzyme-product interaction.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China (20973108, 81271384), the Natural Science Foundation of Shanghai Science and Technology Commission (12ZR1424900), and the Shanghai Key Laboratory of Bio-energy Crops (13DZ2272100) for their financial support. We also acknowledge ChemAxon for their excellent products.

References

[1] J. A. Papin, N. D. Price, S. J. Wiback, D. A. Fell, and B. O. Palsson, "Metabolic pathways in the post-genome era," *Trends in Biochemical Sciences*, vol. 28, no. 5, pp. 250–258, 2003.

[2] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[3] P. Reichard, "Ribonucleotide reductases: the evolution of allosteric regulation," *Archives of Biochemistry and Biophysics*, vol. 397, no. 2, pp. 149–155, 2002.

[4] C. J. Huberty, *Applied Discriminant Analysis*, vol. 297, John Wiley & Sons, New York, NY, USA, 1994.

[5] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: consistency properties," USAF School of Aviation Medicine: Randolph Field, pp. 261-279, San Antonio, Tex, USA, 1951.

[6] R. A. Johnson and D. W. Wichern, *Applied MultiVariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, NJ, USA, 5th edition, 1982.

[7] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[8] B. Niu, L. Lu, L. Liu et al., "HIV-1 protease cleavage site prediction based on amino acid property," *Journal of Computational Chemistry*, vol. 30, no. 1, pp. 33–39, 2009.

[9] Y. Cai, J. He, X. Li et al., "Prediction of protein subcellular locations with feature selection and analysis," *Protein and Peptide Letters*, vol. 17, no. 4, pp. 464–472, 2010.

[10] Y. Cai, Z. He, X. Shi, X. Kong, L. Gu, and L. Xie, "A novel sequence-based method of predicting protein DNA-binding residues, using a machine learning approach," *Molecules and Cells*, vol. 30, no. 2, pp. 99–105, 2010.

[11] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, 2012.

[12] L. Chen, Z.-S. He, T. Huang, and Y.-D. Cai, "Using compound similarity and functional domain composition for prediction of drug-target interaction networks," *Medicinal Chemistry*, vol. 6, no. 6, pp. 388–395, 2010.

[13] L.-L. Hu, S. Niu, T. Huang, K. Wang, X.-H. Shi, and Y.-D. Cai, "Prediction and analysis of protein hydroxyproline and hydroxylysine," *PLoS One*, vol. 5, no. 12, Article ID e15917, 2010.

[14] B. Li, K. Feng, L. Chen, T. Huang, and Y. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS One*, vol. 7, Article ID e43927, 2012.

[15] B.-Q. Li, L.-L. Hu, S. Niu, Y.-D. Cai, and K.-C. Chou, "Predict and analyze S-nitrosylation modification sites with the mRMR and IFS approaches," *Journal of Proteomics*, vol. 75, no. 5, pp. 1654–1665, 2012.

[16] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS One*, vol. 7, no. 4, Article ID e33393, 2012.

[17] K. Lin, Z. Qian, L. Lu et al., "Predicting miRNA's target from primary structure by the nearest neighbor algorithm," *Molecular Diversity*, vol. 14, no. 4, pp. 719–729, 2010.

[18] S. Niu, T. Huang, K. Feng, Y. Cai, and Y. Li, "Prediction of tyrosine sulfation with mRMR feature selection and analysis," *Journal of Proteome Research*, vol. 9, no. 12, pp. 6490–6497, 2010.

[19] Y. Yuan, X. Shi, X. Li et al., "Prediction of interactiveness of proteins and nucleic acids based on feature selections," *Molecular Diversity*, vol. 14, no. 4, pp. 627–633, 2010.

[20] Y.-D. Cai and K.-C. Chou, "Predicting subcellular localization of proteins in a hybridization space," *Bioinformatics*, vol. 20, no. 7, pp. 1151–1156, 2004.

[21] Y.-D. Cai and A. J. Doig, "Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition," *Bioinformatics*, vol. 20, no. 8, pp. 1292–1300, 2004.

[22] K.-C. Chou and Y.-D. Cai, "Predicting protein-protein interactions from sequences in a hybridization space," *Journal of Proteome Research*, vol. 5, no. 2, pp. 316–322, 2006.

[23] Z. Qian, Y.-D. Cai, and Y. Li, "A novel computational method to predict transcription factor DNA binding preference," *Biochemical and Biophysical Research Communications*, vol. 348, no. 3, pp. 1034–1037, 2006.

[24] J. Song, "Prediction of homo-oligomeric proteins based on nearest neighbour algorithm," *Computers in Biology and Medicine*, vol. 37, no. 12, pp. 1759–1764, 2007.

[25] T. M. Cover and P. E. Hart, "Nearst neighbor pattem classlfication," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[26] J. H. Friedman, F. Baskett, and L. J. Shustek, "An algorithm for finding nearest neighbors," *IEEE Transactions on Computers*, vol. C-24, no. 10, pp. 1000–1006, 1975.

[27] S. Goto, T. Nishioka, and M. Kanehisa, "LIGAND: chemical database for enzyme reactions," *Bioinformatics*, vol. 14, no. 7, pp. 591–599, 1998.

[28] B. Niu, L. Gu, C. R. Peng, J. Ding, X. C. Yuan, and W. C. Lu, "Small molecules' multi-metabolic pathways prediction using physico-chemical features and multi-task learning method," *Current Bioinformatics*, vol. 8, no. 5, pp. 564–568, 2013.

[29] C.-R. Peng, W.-C. Lu, B. Niu, M.-J. Li, X.-Y. Yang, and M.-L. Wu, "Predicting the metabolic pathways of small molecules based on their physicochemical properties," *Protein & Peptide Letters*, vol. 19, no. 12, pp. 1250–1256, 2012.

[30] C.-R. Peng, W.-C. Lu, B. Niu, Y.-J. Li, and L.-L. Hu, "Prediction of the functional roles of small molecules in lipid metabolism based on ensemble learning," *Protein & Peptide Letters*, vol. 19, no. 1, pp. 108–112, 2012.

[31] L. Weber, "JChem Base—ChemAxon," *Chemistry World-Uk*, vol. 5, no. 10, pp. 65–66, 2008.

[32] F. Csizmadia, "JChem: java applets and modules supporting chemical database handling from web browsers," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 2, pp. 323–324, 2000.

[33] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S.-H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: Structure, Function and Genetics*, vol. 35, no. 4, pp. 401–407, 1999.

[34] C. Chothia and A. V. Finkelstein, "The classification and origins of protein folding patterns," *Annual Review of Biochemistry*, vol. 59, pp. 1007–1039, 1990.

[35] D. Frishman and P. Argos, "Seventy-five percent accuracy in protein secondary structure prediction," *Proteins: Structure, Function and Genetics*, vol. 27, no. 3, pp. 329–335, 1997.

[36] M. H. Mucchielli-Giorgi, S. Hazout, and P. Tufféry, "PredAcc: prediction of solvent accessibility," *Bioinformatics*, vol. 15, no. 2, pp. 176–177, 1999.

[37] I. Dubchak, I. Muchnik, S. R. Holbrook, and S.-H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 19, pp. 8700–8704, 1995.

[38] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. H. Kim, "Recognition of a protein fold in the context of the structural classification of proteins (SCOP) classification," *Proteins*, vol. 35, no. 4, pp. 401–407, 1999.

[39] B. Niu, Y. Jin, L. Lu et al., "Prediction of interaction between small molecule and enzyme using AdaBoost," *Molecular Diversity*, vol. 13, no. 3, pp. 313–320, 2009.

[40] Z. He, J. Zhang, X.-H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS One*, vol. 5, no. 3, Article ID e9603, 2010.

[41] L. Hu, T. Huang, X. Shi, W.-C. Lu, Y.-D. Cai, and K.-C. Chou, "Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties," *PLoS One*, vol. 6, no. 1, Article ID e14556, 2011.

[42] T. Huang, X.-H. Shi, P. Wang et al., "Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks," *PloS One*, vol. 5, no. 6, Article ID e10972, 2010.

[43] P. Wang, L. Hu, G. Liu et al., "Prediction of antimicrobial peptides based on sequence alignment and feature selection methods," *PLoS One*, vol. 6, no. 4, Article ID e18476, 2011.

[44] L. Chen, K.-Y. Feng, Y.-D. Cai, K.-C. Chou, and H.-P. Li, "Predicting the network of substrate-enzyme-product triads by combining compound similarity and functional domain composition," *BMC Bioinformatics*, vol. 11, article 293, 2010.

[45] L. Chen, T. Huang, X.-H. Shi, Y.-D. Cai, and K.-C. Chou, "Analysis of protein pathway networks using hybrid properties," *Molecules*, vol. 15, no. 11, pp. 8177–8192, 2010.

[46] K. C. Chou and H.-B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 2, pp. 63–92, 2009.

[47] T. E. Creighton, *Proteins: Structures and Molecular Properties*, W. H. Freeman, New York, NY, USA, 2nd edition, 1993.

[48] G. E. Tusnády and I. Simon, "Principles governing amino acid composition of integral membrane proteins: application to topology prediction," *Journal of Molecular Biology*, vol. 283, no. 2, pp. 489–506, 1998.

[49] Y. Freund, Y. Mansour, and R. E. Schapire, "Generalization bounds for averaged classifiers," *Annals of Statistics*, vol. 32, no. 4, pp. 1698–1722, 2004.

[50] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[51] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[52] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Machine Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[53] I. A. Yudushkin, A. Schleifenbaum, A. Kinkhabwala, B. G. Neel, C. Schultz, and P. I. H. Bastiaens, "Live-cell imaging of enzyme-substrate interaction reveals spatial regulation of PTP1B," *Science*, vol. 315, no. 5808, pp. 115–119, 2007.

[54] M. J. Ondrechen, J. M. Briggs, and J. A. McCammon, "A model for enzyme-substrate interaction in alanine racemase," *Journal of the American Chemical Society*, vol. 123, no. 12, pp. 2830–2834, 2001.

[55] C. Sadasivan and V. C. Yee, "Interaction of the factor XIII activation peptide with α-thrombin. Crystal structure of its enzyme-substrate analog complex," *Journal of Biological Chemistry*, vol. 275, no. 47, pp. 36942–36948, 2000.