Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 464093, 6 pages
http://dx.doi.org/10.1155/2014/464093

Research Article

Protein Binding Site Prediction by Combining Hidden Markov Support Vector Machine and Profile-Based Propensities

Bin Liu,[1,2] Bingquan Liu,[3] Fule Liu,[1] and Xiaolong Wang[1,2]

[1] *School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[2] *Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*

Correspondence should be addressed to Bin Liu; bliu@insun.hit.edu.cn and Bingquan Liu; liubq@insun.hit.edu.cn

Received 4 June 2014; Accepted 1 July 2014; Published 14 July 2014

Academic Editor: Wei Chen

Identification of protein binding sites is critical for studying the function of the proteins. In this paper, we proposed a method for protein binding site prediction, which combined the order profile propensities and hidden Markov support vector machine (HM-SVM). This method employed the sequential labeling technique to the field of protein binding site prediction. The input features of HM-SVM include the profile-based propensities, the Position-Specific Score Matrix (PSSM), and Accessible Surface Area (ASA). When tested on different data sets, the proposed method showed promising results, and outperformed some closely relative methods by more than 10% in terms of AUC.

1. Introduction

Prediction of protein binding sites provides valuable information for studying the function of proteins. The most efficient approaches are the computational methods. By using these approaches, the functionally important amino acid residues can be identified [1].

These computational methods used different features extracted from protein sequences, PSSM, or structure information. Hydrophobic and polar residues tend to occur in protein binding regions [2, 3]. The conservation scores of amino acid are often used as features, because the protein binding sites are more conserved than other surface residues [4]. Some kinds of conservation scores were proposed; a comprehensive evaluation of these scores was reported in [5]. One of the most widely used features is the Accessible Surface Area (ASA) [4], because the binding sites show higher ASA values than those of the other surface residues [6].

Some machine learning methods treated protein binding site prediction as a binary classification task, and some well-known machine learning techniques have been applied to this field, such as support vector machine [7, 8], neural network [1], Bayesian network [9], and hidden Markov model [10]. A comparison of these methods has been performed by Zhou and Qin [11].

In our previous study [12], we introduced a novel profile-level propensity for protein binding site prediction. Experimental results showed that this propensity can significantly improve the performance of the SVM based methods. Recently, we applied hidden Markov support vector machine (HM-SVM) to this field [13], which takes protein binding site prediction as a sequence-labeling task. The advantage of this method is that it is able to incorporate the sequence-order effects into the predictor. However, this method only uses two basic features (PSSM and ASA features) as input for protein binding site prediction. Therefore, it is interesting to explore whether the order profile propensity can improve the performance of HM-SVM based method or not. In this study, we proposed a computational method for protein binding site prediction by combining the hidden Markov support vector machine and the order profile propensity. When tested on six different data sets, the HM-SVM predictor using order profile propensity as an extra feature consistently outperformed the predictor only using two basic features (PSSM and ASA

Table 1: Summary of six data sets.

Data set	Chains	Res.	Surface res.	Interface res.
Heterocomplex I[a]	504	109829	92797	26085
Homocomplex I	620	172917	141295	38170
Mix[b] I	1124	282746	234092	64255
Heterocomplex II[c]	504	109829	92797	32386
Homocomplex II	620	172917	141295	45633
Mix II	1124	282746	234092	78019

[a]Type I data set with minor interface as negative samples.
[b]The mixed data set of heterocomplexes and homocomplexes.
[c]Type II data set with minor interface as positive samples.

features); in particular, in terms of AUC, the performance is improved by more than 10 percent, indicating that combining the order profile propensity and the HM-SVM is a suitable approach to improve the accuracy of protein binding site prediction.

2. Methods

2.1. Dataset Description. The datasets used in this study have been described in the study [13]. 1124 protein chains were selected from the Protein Data Bank (PDB) [14]. The chains were divided into six types of datasets according to homology of interacting chains and the definition of the interface. The information of the six datasets is shown in Table 1, and the process of dataset preparation is shown in the left part of Figure 1. The six datasets can be downloaded from http://bioinformatics.hitsz.edu.cn/HMSVM-OP.

2.2. Feature Description

2.2.1. Order Profile Propensity. The detailed information of how to calculate the order profile propensity was introduced in study [12]. Here we only briefly introduce this process. The order profile propensities were profile-based features, which extracted the evolutionary information from frequency profiles. The frequency profiles were calculated from the multiple sequence alignments outputted by running the PSI-BLAST software [13] searching against the nrdb90 database from EBI [15] with parameters of $j = 10$ and $e = 0.001$. The frequency profiles were converted into order profiles by combining the amino acids whose frequencies were higher than a given threadhold optimized on the benchmark dataset. Order profile can be viewed as a profile-based building block of proteins, which has been used for many tasks in the field of bioinformatics [12, 16].

The order profile propensity was based on the order profile occurrence differences between protein binding regions and other surface regions. The equations of how to calculate this feature were given by [12, Equations (3)–(5)].

2.2.2. Position-Specific Score Matrix (PSSM). PSSM was another profile-based feature, which was generated by using PSI-BLAST [13] with the parameters j and e set as 10 and 0.001, respectively.

2.2.3. Accessible Surface Area (ASA). We employed the DSSP program [17] to calculate the Accessible Surface Area (ASA) features, which were scaled by the nominal maximum area of each residue.

2.3. Hidden Markov Support Vector Machine. Hidden Markov support vector machine proposed by Altun et al. [15] was a sequential labelling model. In our previous study [13], it showed that when using the two basic features (PSSM and ASA features), the HM-SVM based method outperformed other machine learning methods, such as SVM, CRF, and ANN. In this study, we explored new features to improve the performance of HM-SVM based methods. For more information of HM-SVM, please refer to this paper [13].

The flowchart of the proposed computational method for protein binding site prediction was shown in Figure 1, in which the left part shows the process of dataset construction, and the right part shows the prediction process of the model based on HM-SVM.

In this paper, SVM^{hmm} toolkit (V3.10) was employed as the software of HM-SVM model with parameters c and e set as 0.1 and 1, respectively. This parameter combination was optimized with the training data. The input features of HM-SVM include order profile propensity, ASA, and PSSM. These features were extracted from the target residues and its 6 neighbouring residues in each direction.

2.4. Evaluation Methodology. The sensitivity (Sn), specificity (Sp), overall accuracy (Acc), F1 measure (F1), Matthews correlation coefficient MCC, and AUC can be, respectively, expressed as [18–22]

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}},$$

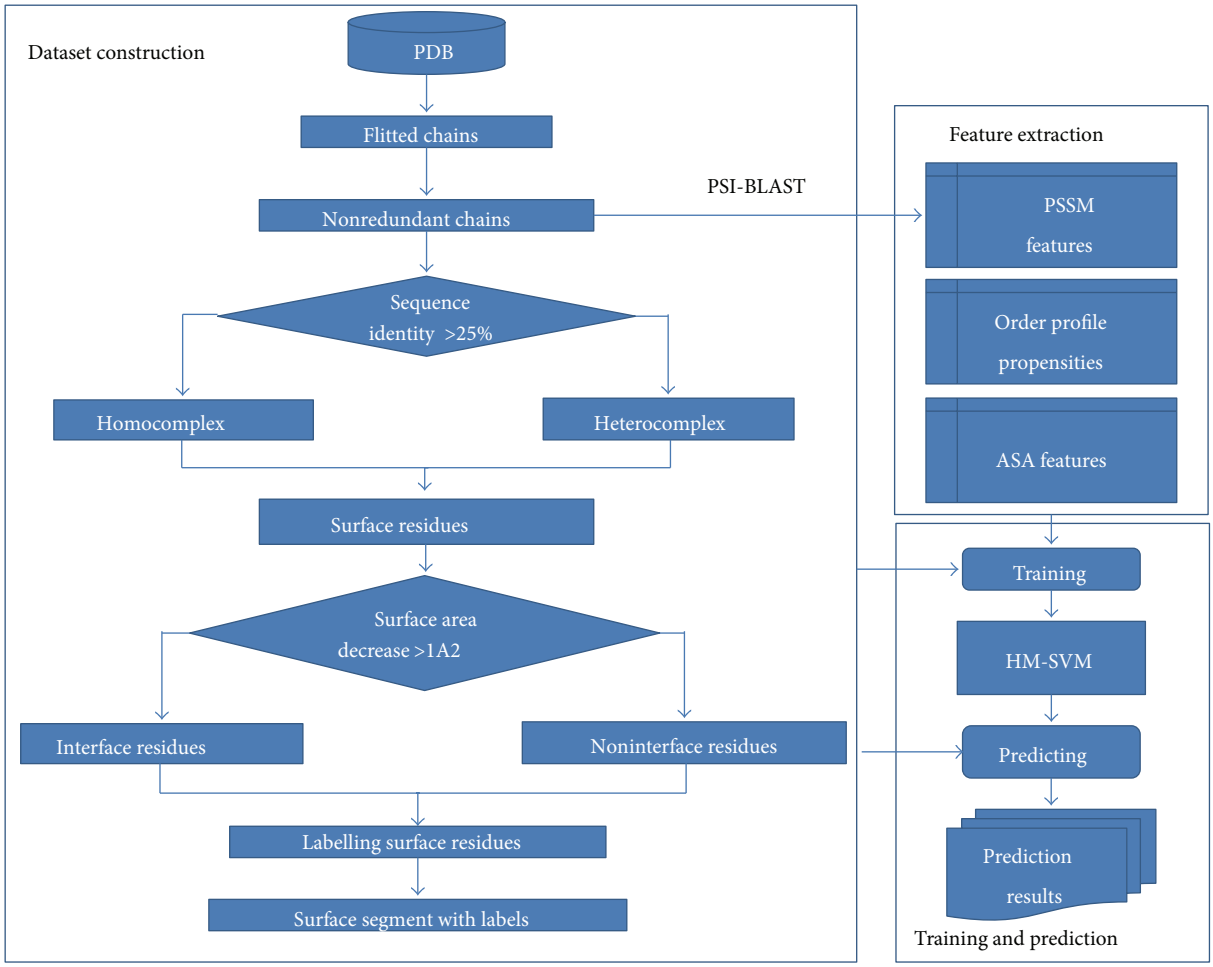

Figure 1: Overview of the proposed framework for protein binding site prediction.

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FN} * \text{FP})}{\sqrt{(\text{TP} + \text{FN}) * (\text{TN} + \text{FP}) * (\text{TP} + \text{FP}) * (\text{TN} + \text{FN})}},$$
$$\text{AUC}: \text{ the area under ROC cure,} \quad (1)$$

where TP represents the true positive, TN represents the true negative, FN represents the false negative, and FP represents the false positive.

3. Results

In order to validate whether the order profile propensities can improve the performance of the HM-SVM based methods or not, two HM-SVM predictors with different features were constructed. The first HM-SVM employed the PSSMs and ASA as input features. This predictor was treated as a baseline predictor. For the second HM-SVM predictor, order profile propensity is added as an extra feature to evaluate whether this feature can improve the performance or not. The performance of the two HM-SVM predictors was evaluated by fivefold cross-validation.

The results of the two HM-SVM predictors on the six datasets are shown in Table 2. It can be seen that the first HM-SVM predictor using the two basic features achieved the lowest performance. The second HM-SVM predictor using the order profile propensity as an extra feature achieved the best performance on all the six data sets, especially its AUC score being about 10% higher than that of the first HM-SVM predictor, indicating that order profile propensity can significantly improve the performance of the HM-SVM based methods. In our previous study [13], we showed that the first HM-SVM predictor outperformed some state-of-the-art methods, such as ANN, CRF, and SVM. The second HM-SVM predictor significantly outperformed the first HM-SVM predictor, indicating that the proposed computational method for protein binding site prediction is a good method in this field.

Šikić et al. [23] proposed a method based on random forest, which was evaluated on a heterocomplex data set, and achieved good performance (Sp = 76.45%, Sn = 38.06%, F1 = 50.82%, and Acc = 80.05%). Our method (results of heterocomplex II dataset) outperformed this method by 14.98% in terms of F1, which further confirms the better performance of our method than some state-of-the-art methods.

Table 2: Performance of HM-SVM based method with and without order profile propensities.

Dataset	Method	Sp %	Sn %	F1 %	Acc %	MCC	AUC %
Heterocomplex I	HM-SVM 1[a]	44.9	56.0	49.8	68.3	0.274	69.5
	HM-SVM 2[b]	**52.4**	**73.5**	**61.2**	**73.8**	**0.436**	**81.4**
Homocomplex I	HM-SVM 1	45.4	60.0	51.70	69.7	0.309	72.2
	HM-SVM 2	**54.5**	**74.6**	**62.9**	**76.3**	**0.474**	**83.6**
Mix I	HM-SVM 1	45.5	58.0	51.0	69.4	0.297	71.2
	HM-SVM 2	**53.5**	**74.0**	**62.1**	**75.0**	**0.455**	**82.5**
Heterocomplex II	HM-SVM 1	54.0	56.7	55.3	68.0	0.305	70.7
	HM-SVM 2	**60.8**	**71.7**	**65.8**	**74.0**	**0.454**	**81.2**
Homocomplex II	HM-SVM 1	53.3	60.1	56.5	70.1	0.340	73.4
	HM-SVM 2	**61.1**	**73.8**	**66.8**	**76.4**	**0.493**	**83.7**
Mix II	HM-SVM 1	53.6	58.6	56.0	69.3	0.326	72.4
	HM-SVM 2	**61.0**	**72.7**	**66.3**	**75.2**	**0.474**	**82.4**

[a]Results of HM-SVM 1 on the six data sets are obtained from [13]. HM-SVM 1 represents the HM-SVM predictor with the basic feature set using PSSM and ASA features; [b]HM-SVM 2 represents the HM-SVM predictor with the feature set using PSSM, ASA, and order profile propensity features.

4. Conclusion

In this study, we proposed a computational method for protein binding site prediction, which combines the order profile propensity and hidden Markov support vector machine. This method predicts the protein binding sites with a sequential labelling approach and uses a recently proposed feature to further improve the performance: order profile propensity, which contains the evolutionary information extracted from the sequence profiles. The main contribution of this study is that we validate the fact that order profile propensity can significantly improve the performance of the HM-SVM based method. The main advantage of the proposed method is that it treats the protein sequence as a whole and is able to use the label information of neighbour residues and the evolutionary information extracted from the frequency profiles. However, the order profile propensity was generated based on the frequency profiles, which require the computational expensive multiple sequences alignment process. It is the main disadvantage of the proposed method.

As noted by Li et al. [24], choosing proper features is a challenging task, especially for sequential labelling method, such as HM-SVM and conditional random field (CRF). In their experiments, the authors found that by simply adding some features into CRF cannot improve the performance of their method. Therefore, the obvious performance improvement when using order profile propensity as an extra feature will benefit our future studies, especially for the research on applying sequential method to this field. As pointed out in a comprehensive review and carried out in a series of recent publications [25–43], finding suitable features is the key step to improve the performance.

Furthermore, since user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors [44, 45], we shall make efforts in our future work to provide a web server for the method presented in this paper.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61300112, 61272383), the Natural Science Foundation of Guangdong Province (no. S2012040007390), the Scientific Research Innovation Foundation in Harbin Institute of Technology (Project no. HIT.NSRIF.201310b3), the Shanghai Key Laboratory of Intelligent Information Processing, China (Grant no. IIPL-2012-002), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, Strategic Emerging Industry Development Special Funds of Shenzhen (ZDSY20120613125401420 and JCYJ2012 0613151940045), and Key Basic Research Foundation of Shenzhen (JC201005260118A, JC201005260175A).

References

[1] A. Porollo and J. Meller, "Prediction-based fingerprints of protein-protein interactions," *Proteins: Structure, Function and Genetics*, vol. 66, no. 3, pp. 630–645, 2007.

[2] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, no. 2, pp. 89–102, 2001.

[3] W. L. DeLano, "Unraveling hot spots in binding interfaces: progress and challenges," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 14–20, 2002.

[4] H. Zhou and Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins: Structure, Function and Genetics*, vol. 44, no. 3, pp. 336–343, 2001.

[5] W. S. J. Valdar, "Scoring residue conservation," *Proteins: Structure, Function and Genetics*, vol. 48, no. 2, pp. 227–241, 2002.

[6] H. Chen and H. X. Zhou, "Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data," *Proteins: Structure, Function and Genetics*, vol. 61, no. 1, pp. 21–35, 2005.

[7] A. Koike and T. Takagi, "Prediction of protein-protein interaction sites using support vector machines," *Protein Engineering, Design and Selection*, vol. 17, no. 2, pp. 165–173, 2004.

[8] B. Wang, P. Chen, D. Huang, J. Li, T. Lok, and M. R. Lyu, "Predicting protein interaction sites from residue spatial sequence profile and evolution rate," *FEBS Letters*, vol. 580, no. 2, pp. 380–384, 2006.

[9] J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, "Insights into protein-protein Interfaces using a Bayesian network prediction method," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 365–386, 2006.

[10] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, "Modelling interaction sites in protein domains with interaction profile hidden Markov models," *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, 2006.

[11] H. Zhou and S. Qin, "Interaction-site prediction for protein complexes: a critical assessment," *Bioinformatics*, vol. 23, no. 17, pp. 2203–2209, 2007.

[12] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "Exploiting three kinds of interface propensities to identify protein binding sites," *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009.

[13] B. Liu, X. Wang, L. Lin, B. Tang, and Q. Dong, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[14] A. Kouranov, L. Xie, J. de la Cruz et al., "The RCSB PDB information portal for structural genomics," *Nucleic Acids Research.*, vol. 34, pp. D302–D305, 2006.

[15] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden markov support vector machines," in *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, pp. 3–10, August 2003.

[16] B. Liu, L. Lin, and X. Wang, "Protein remote homology detection using order profiles," in *Proceedings of the International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS '09)*, pp. 255–260, Shanghai, China, August 2009.

[17] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[18] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[19] B. Liu, X. Wang, L. Lin, and Q. Dong, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, article 510, 2008.

[20] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[21] B. Liu, J. Xu, Q. Zou et al., "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, p. S3, 2014.

[22] B. Liu, "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[23] M. Šikić, S. Tomić, and K. Vlahoviček, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000278, 2009.

[24] M. H. Li, L. Lin, X. Wang, and T. Liu, "Protein-protein interaction site prediction based on conditional random fields," *Bioinformatics*, vol. 23, no. 5, pp. 597–604, 2007.

[25] B. Liu, "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions," *BMC Genomics*, vol. 14, supplement 8, p. S3, 2013.

[26] W. Chen, P. Feng, and H. Lin, "Prediction of ketoacyl synthase family using reduced amino acid alphabets," *Journal of Industrial Microbiology and Biotechnology*, vol. 39, no. 4, pp. 579–584, 2012.

[27] H. Lin, W. Chen, and H. Ding, "AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes," *PLoS ONE*, vol. 8, no. 10, Article ID e75726, 2013.

[28] W. Chen, H. Lin, and P. M. Feng, "DNA physical parameters modulate nucleosome positioning in the *Saccharomyces cerevisiae* genome," *Current Bioinformatics*, vol. 9, no. 2, pp. 188–193, 2014.

[29] P.-M. Feng, H. Lin, and W. Chen, "Identification of antioxidants from sequence information using Naïve Bayes," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 567529, 5 pages, 2013.

[30] C. Ding, L. Yuan, S. Guo, H. Lin, and W. Chen, "Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions," *Journal of Proteomics*, vol. 77, pp. 321–328, 2012.

[31] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[32] P. M. Feng, W. Chen, H. Lin, and K. Chou, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[33] W. Chen, H. Lin, P. Feng, C. Ding, Y. Zuo, and K. Chou, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[34] S. H. Guo, E. Z. Deng, L. Q. Xu et al., "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[35] W. Chen, P. Feng, H. Lin, and K. Chou, "IRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, p. e68, 2013.

[36] P. Feng, H. Ding, W. Chen, and H. Lin, "Naïve bayes classifier with feature selection to identify phage virion proteins," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 530696, 6 pages, 2013.

[37] H. Ding, S. Guo, E. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[38] W. Chen and H. Lin, "Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information," *Biochemical and Biophysical Research Communications*, vol. 401, no. 3, pp. 382–384, 2010.

[39] H. Lin, H. Ding, F. Guo, and J. Huang, "Prediction of subcellular location of mycobacterial protein using feature selection techniques," *Molecular Diversity*, vol. 14, no. 4, pp. 667–671, 2010.

[40] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67–70, 2011.

[41] W. Chen, T. Y. Lei, D. C. Jin, H. Lin, and K. C. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, 2014.

[42] H. Lin, W. Chen, L. Yuan, Z. Li, and H. Ding, "Using over-represented tetrapeptides to predict protein submitochondria locations," *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[43] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 4, pp. 1038–1043, 2013.

[44] Q. Zou, X. Li, Y. Jiang, Y. Zhao, and G. Wang, "Binmempredict: a web server and software for predicting membrane protein types," *Current Proteomics*, vol. 10, no. 1, pp. 2–9, 2013.

[45] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.