Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 861269, 6 pages
http://dx.doi.org/10.1155/2013/861269

Research Article

The Prediction of Calpain Cleavage Sites with the mRMR and IFS Approaches

Wenyi Zhang, Xin Xu, Longjia Jia, Zhiqiang Ma, Na Luo, and Jianan Wang

School of Computer Science and Information Technology, Northeast Normal University, Changchun 130117, China

Correspondence should be addressed to Na Luo; luon110@nenu.edu.cn and Jianan Wang; wangjn@nenu.edu.cn

Received 11 July 2013; Accepted 12 August 2013

Academic Editor: William Guo

Calpains are an important family of the Ca^{2+}-dependent cysteine proteases which catalyze the limited proteolysis of many specific substrates. Calpains play crucial roles in basic physiological and pathological processes, and identification of the calpain cleavage sites may facilitate the understanding of the molecular mechanisms and biological function. But traditional experiment approaches to predict the sites are accurate, and are always labor-intensive and time-consuming. Thus, it is common to see that computational methods receive increasing attention due to their convenience and fast speed in recent years. In this study, we develop a new predictor based on the support vector machine (SVM) with the maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). And we concern the feature of physicochemical/biochemical properties, sequence conservation, residual disorder, secondary structure, and solvent accessibility to represent the calpain cleavage sites. Experimental results show that the performance of our predictor is better than several other state-of- the-art predictors, whose average prediction accuracy is 79.49%, sensitivity is 62.31%, and specificity is 88.12%. Since user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors, here we have provided a web-server for the method presented in this paper.

1. Introduction

Calpains are an important family of the Ca^{2+}-dependent cysteine proteases which catalyze the limited proteolysis of many specific substrates [1, 2]. Probably, 16 known calpain isoform genes are founded in humans. Then, 14 genes encoded proteins have cysteine protease domains, and the other 2 genes that encode some regulatory proteins are associated with some catalytic subunits forming heterodimeric proteases [3, 4]. Calpains play crucial roles in basic physiological and pathological processes, including the regulation of gene expression, signal transduction, cell death and apoptosis, remodeling cytoskeletal attachments during cell fusion or motility, and cell cycle progression [3–5]. Moreover, calpain aberrancies frequently lead to a variety of diseases and cancers [6]. As we know, traditional experimental identification and characterization of calpain cleavage sites are labor-intensive and expensive. Recently, calpain cleavage sites prediction attracts more and more attention, and more and more studies have understood its regulatory roles and molecular mechanisms of calpain cleavage.

In recent years, many computational methods were developed to predict calpain cleavage sites. In the paper [7], Tompa et al. selected 49 calpain substrates with a total of 106 sequentially identified cleavage sites from the literature. They determined the amino acid preferences around the cleavage bond with 11-mer peptide, and they synthesized a short peptide of TPLKSPPPSPR to be a superior substrate of calpain. Then, Boyd et al. developed PoPS online tool to predict protease specificity [8, 9]. And the site prediction based on the frequency and substitution matrix scoring strategy predicted Calpain 1 and 2 specific cleavage sites [10]. Recently, Liu et al. developed a new computational program for the prediction of calpain cleavage sites. With the previously released algorithm of GPS (Group-based Prediction System), they designed a novel software package of GPS-CCD (Calpain Cleavage Detector) for prediction of calpain cleavage sites [6]. Although aforementioned predictors were effective, we

should make more efforts to improve the performance of calpain cleavage sites prediction.

In this study, we developed a new predictor based on the support vector machine (SVM) with the maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). And we concerned the features of physicochemical/biochemical properties, sequence conservation, residual disorder, secondary structure, and solvent accessibility to represent the calpain cleavage sites. Experimental results showed that the performance of our predictor was better than several other state-of-the-art predictors, whose average prediction accuracy was 79.49%, sensitivity was 62.31%, and specificity was 88.12%. Since user-friendly and publicly accessible web servers represented the future direction for developing practically more useful predictors [11], here we have provided a web server for the method presented in this paper at http://202.198.129.219:8080/calpain_cleavage/.

2. Materials and Method

2.1. Data Sets. Here, we selected 130 unique substrates for calpain cleavage sites. And all the proteins were extracted from Uniprot/Swiss-Prot (Jul 20, 2012), by searching the "calpain" in the field "Sequence annotation" with experimental verification. We defined a calpain cleavage peptide (m, n) as the cleavage bond flanked by m residues upstream and n residues downstream, where m and n were equal to 10. Similar to [6], all experimentally verified cleavage sites were regarded as positive samples, and the other noncleavage sites in the same substrates were taken as the negative samples. With the threshold of 40% identity by CD-HIT, the training dataset contained 368 positive samples.

2.2. Protein Features and Vector Encoding. The first feature we select is the position specific scoring matrix (PSSM) of each calpain cleavage peptide. All biological species have developed starting out from a very limited number of ancestral species. Their evolution involves changes of single residues, insertions and deletions of several residues [12], gene doubling, and gene fusion. With these changes accumulated for a long time, many similarities between initial and resultant amino acid sequences are gradually eliminated, but the corresponding proteins may still share some equal functions and action mechanisms. Accordingly, evolutionary conservation may play important roles in biological analysis [13]. We used PSI-BLAST [14] to generate the scoring of specific residues. PSSM profile for each peptide can be represented as a matrix of $M \times 20$ dimensions, and M is the length of peptide; 20 dimensions mean a measure of residue conservation of 20 different standard amino acids [15].

The second feature we exploit is the feature of amino acid factors derived from AAIndex [16], which is a famous database including various physicochemical and biochemical properties. Due to native 20 amino acids having their own specific properties, the composition of these properties of different residues will affect the structure and function of the protein [13]. Atchley et al. [17] performed multivariate statistical analyses on AAIndex and replaced amino acid properties with five pattern scores (polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge) [13]. We use the five pattern scores to represent each amino acid.

We also consider other features to make full use of protein sequence and prior knowledge, including disorder score, secondary structure, and solvent accessibility. Therefore, the information of disorder score is involved with protein structure and function. In this study, we use VSL2 [18] to calculate the disorder score of each amino acid peptide. Moreover, we use SSpro4 [19] to predict the secondary structural property of each amino acid of a given protein sequence as "helix," "stand," or "other" which are encoded with "100," "010," and "001". So, we construct a series of $K \times 3$ matrix, K is the length of the chain peptide. The predictor SSpro4 also can predict solvent accessibility of each amino acid as "buried" or "exposed," which is encoded with "10" and "01;" then, $K \times 2$ matrix is formed; K is also the length of chain peptide.

2.3. The Feature Space. As mentioned previously, for each amino acid of a given peptide, the following 31 features are needed: 20 PSSM conservation score features, 5 amino acid factors features, 1 disorder feature, 3 secondary structure features, and 2 solvent accessibility features. The length of given peptide is 21; there are total of $31 \times 21 = 651$ features. According to (6) of [20], the feature vector for any protein, peptide, or biological sequence is none but a general form of pseudo amino acid composition or PseAAC [21, 22] that can be formulated as

$$P = \left[\psi_\psi \psi_2 \cdots \psi_\mu \cdots \psi_\Omega\right]^T, \tag{1}$$

where $\Omega = 651$, T is the transpose operator, and ψ_μ $(\mu = 1, 2, \ldots, 651)$ represents the μ feature.

2.4. The mRMR Method. We use the mRMR method to rank the importance of the 651 features based on minimal redundancy and maximal relevance [23]. The ranked feature with the smaller index indicates that it has a better trade-off between the maximum relevance and minimum redundancy. The mutual information is used for reflecting the dependence of vector x and vector y:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} dx\, dy, \tag{2}$$

where x and y are two random vectors, $p(x, y)$ is the joint probabilistic density, and $p(x)$ and $p(y)$ are the marginal probabilistic densities.

Suppose that the set M is the already-selected feature set containing m features, and the set N is the to-be-selected feature set containing n features. D denotes the relevance between the feature f in N and the class c:

$$D = I(f, c). \tag{3}$$

And R denotes the relevance between the feature f in N and all features in M, and R can be calculated by

$$R = \frac{1}{m} \sum_{f_i \in M} I(f, f_i). \tag{4}$$

So the feature f_i in the set N with the maximum relevance and minimum redundancy can be calculated by

$$\max_{f_j \in N} \left[I(f_j, c) - \frac{1}{m} \sum_{f_i \in M} I(f_j, f_i) \right], \quad (j = 1, 2, \ldots, n). \tag{5}$$

We can use the mRMR method to find the feature set S, and each feature in S has the index indicating its importance; the more important the feature is, the smaller the index is.

2.5. Support Vector Machine. SVM belongs to the family of margin-based classifier and is very powerful to deal with prediction, classification, regression problems [24, 25]. Therefore, SVM is widely used for all kinds of problems. SVM looks for optimal hyperplane which maximizes the distance between the hyperplane and the nearest sample from each of the two classes. Formally, given a training vector $x_i \in R^n$ and their class values $y_i \in (-1, 1)$, $i = 1, \ldots, N$, SVM solves the following optimization problems:

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{2}\omega^T \cdot \omega + C \sum_{i=1}^{N} \xi_i, \\ &\text{subject to} \quad y_i \left(\omega^T \cdot x_i + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \end{aligned} \tag{6}$$

where ω is a normal vector perpendicular to the hyperplane and ξ_i is slake variables for allowing misclassifications. Here C (>0) is the penalty parameter which balances the trade-off between the margin and training error. In the work, LIBSVM package [26, 27] with radial basis kernel function is used. Two parameters, the regularization parameter C and kernel width parameter γ, are optimized based on 5-fold cross validation using a grid search strategy.

2.6. Evaluation. In statistical prediction, three cross validation tests are often used to evaluate the performance of predictors: subsampling test, independent dataset test, and jackknife test [28, 29]. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in [30] and demonstrated by (28)–(30) in [21]. Accordingly, the jackknife test has been increasingly and widely used by investigators to examine the quality of various predictors (see, e.g., [31–34]). However, to reduce the computational time, we adopted the 5-fold cross validation in this study as done by many investigators with SVM as the prediction engine.

5-fold cross validation [29] is used in this work. The dataset is randomly divided into five equal sets, out of which four sets are used for training and the remaining one for testing. This procedure is repeated five times, and the final prediction result is the average accuracy of the five testing sets.

Four parameters, sensitivity (Sn), specificity (Sp), accuracy (Ac), and matthews' correlation coefficient (MCC) are used to measure the performance of our model. They are defined by the following formulas:

$$\begin{aligned} \text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Sp} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Ac} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \\ \text{MCC} &= \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FN})}}, \end{aligned} \tag{7}$$

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively. For a given dataset, all these values can be obtained from the decision function with fixed cutoff.

2.7. Incremental Feature Selection (IFS). With the mRMR method, we can rank the importance of the 651 features, and then, we can use Incremental Feature Selection (IFS) [35–38] to determine the optimal number of features. We can create the features set by the features importance rank, such as

$$S_i = \{f_1, f_2, \ldots, f_i\}, \quad (1 \leq i \leq N). \tag{8}$$

We can use SVM to predict the performances of each feature set and evaluate the set with the 5-fold cross validation; thus, the optimal feature set can be yielded [39–41].

3. Result and Discussion

3.1. The mRMR Result and IFS Result. In the Supporting Information S5, the mRMR feature table listed the ranked 651 features with the maximum relevance and minimum redundancy to the class of samples. The list of ranked feature was to be used in the following IFS procedure for the optimal feature set selection.

In IFS test (Support Information S5, see Supplementary Material available online at http://dx.doi.org/10.1155/2013/861269), we added the feature one by one and built about 651 predictors. In Figure 1, the MCC reached their maximum value when 284 features were used. The accuracy, sensitivity, specificity, and MCC were 0.7949, 0.6231, 0.8812, and 0.5249. Figure 1 showed the MCC plot based on Supporting Information S5.

3.2. Analysis of the Optimal Feature Set. In the IFS procedure, we selected 284 optimal features (Supporting Information S5). In the result, 154 belonged to the PSSM conservation score, 43 to the amino acid factors, 21 to the disorder, 63 to the

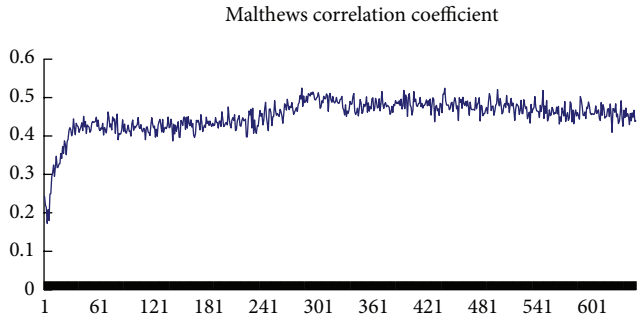

Figure 1: Plot of the MCC of different number features.

secondary structure, and 3 to the solvent accessibility. It indicated that PSSM conservation score, amino acid factors and secondary structure played important roles to predict calpain cleavage sites. The optimal feature distribution revealed that site 1, site 7 to site 8, site 10, and site 12 to site 14 played the most important roles in prediction of calpain cleavage sites. Moreover, the features close to calpain cleavage site were more important than central site and the site far from the calpain cleavage site.

3.2.1. PSSM Conservation Score Feature Analysis. As mentioned previously, there were 154 PSSM conservation features, and we found that the conservation against mutations of the 20 amino acids had different impacts on the prediction of calpain cleavage sites. We measured the number of each kind of amino acids for the PSSM features (Figure 2(a)) and found that different mutations of amino acid had different roles in prediction of calpain cleavage sites. Mutation of the amino acid Valine (V), Leucine (L) and Phenylalanine (P) were important in predicting the calpain cleavage sites. In the mRMR feature list, the 2 to 5 ranked features were PSSM features at site 11, site 8, site 6, and site 4 against transition to amino acid Leucine (L). This indicated that the conservation of Valine, Leucine, and Phenylalanine were the keys to determining whether or not it was calpain cleavage site. We also measured the PSSM feature number of each amino acid site (Figure 2(b)). The result revealed that site 5, site 8, site 12, and site 13 were more important in predicting calpain cleavage sites than other sites (shown in Figure 2(b)).

3.2.2. Amino Acid Factor Analysis. We investigated the number of each type of amino acid factor features (Figure 3(a)) and the number of amino acid factors at each site (Figure 3(b)). As a result, the secondary structure was the most important features in predicting the calpain cleavage sites. And the codon diversity was the second important feature to predict calpain cleavage sites. In Figure 3(b), site 10, site 11, and site 21 had relatively more effects on the calpain cleavage sties. Moreover, in the Supporting Information S5, the first feature was the polarity feature, and the polarity feature at site 10, site 12, and site 15 played more roles in the calpain cleavage sites prediction. This indicated that the polarity of the residues located more close to the calpain cleavage site has a critical role in predicting the calpain cleavage site.

3.2.3. Disorder, Secondary Structure, and Solvent Accessibility Feature Analysis. With the final optimal feature set, there were 63 secondary structure features; 21 disorder features and a reasonable explanation was that the feature of secondary structure and disorder encoding were sensitive for predicting calpain cleavage sites. And in the Supporting information S5, we could show that first index of secondary structure was 189 in site 1, and it was the "stand" feature; also the first index of disorder was 143 at site 1. There were 3 solvent accessibility features in the optimal feature set; they were in sites 1 and 2; however the index was 281, 282, and 284.

3.3. Comparison with Existing Method. According to the mRMR and IFS procedure, the performance of predictor was the best when we selected 284 features. And the accuracy, specificity, sensitivity, Matthews' correlation coefficient was 0.7949, 0.8812, 0.6231, and 0.5249. And we made a comparison with GPS 2.0, GPS 1.0, PoPS, site prediction 1, and site prediction 2. For MCC, our predictor was obviously improved than all other predictors. It indicated that our predictor has excellent performance in predicting positive samples. And in the same value of specificity (~0.90), our sensitivity was 0.6231, higher than the other predictor. But the accuracy of our predictor was slightly worse than those of the others. Since we did not know the ratio of positive and negative in their training set, we built the predictor based on a training set in which the negative samples were two times than positive samples. The sensitivity of our method was 0.6231, and the sensitivity GPS 2.0 was 0.6087, GPS1.0 was 0.5000, PoPS was 0.5245, site prediction 1 was 0.4130, and site prediction 2 was 0.3967, when the threshold was medium.

More importantly, the reasonably good performance of our method reflects that the physicochemical/biochemical properties can effectively capture the information of around calpain cleavage sites. But there still exists some limits in our method. For example, our predictor only considers the amino acid sequence information but does not consider the protein structure features. Therefore, in the future, we should focus on development of amino acid encoding schema and development of a predictor to maximize the prediction performance of calpain cleavage sites.

4. Conclusion

Calpains are an important family of the Ca^{2+}-dependent cysteine proteases which catalyze the limited proteolysis of many specific substrates. Calpains play crucial roles in basic physiological and pathological processes. In the paper, we developed a new predictor based on the support vector machine (SVM) with the maximum relevance minimum redundancy (mRMR) method followed by incremental feature selection (IFS). And we concerned the feature of physicochemical/biochemical properties, sequence conservation, residual disorder, secondary structure, and solvent accessibility to represent the calpain cleavage sites. And we selected 284 optimal features; these features were central to predict the calpain cleavage sites, and with the optimal features set, the accuracy of our predictor was 0.7949, and the sensitivity and

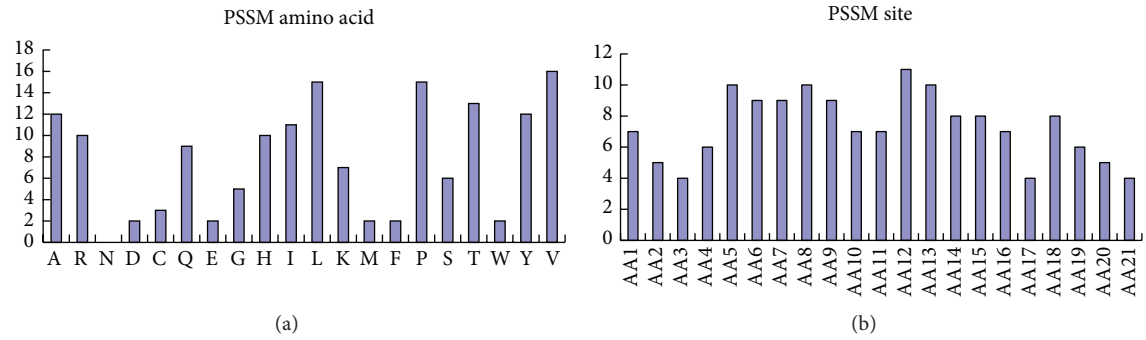

Figure 2: Bar plots to show the distribution in the final optimal feature set for (a) the PSSM score and (b) the corresponding specific site score. It was shown from panel (a) that mutations of amino acid Valine (V) played most important role in prediction of calpain cleavage sites; followed by Leucine (L) and Phenylalanine (P). And it was shown from panel (b) that conservation in site 5, site 8, site 12, and site 13 was more important in determining the calpain cleavage sites.

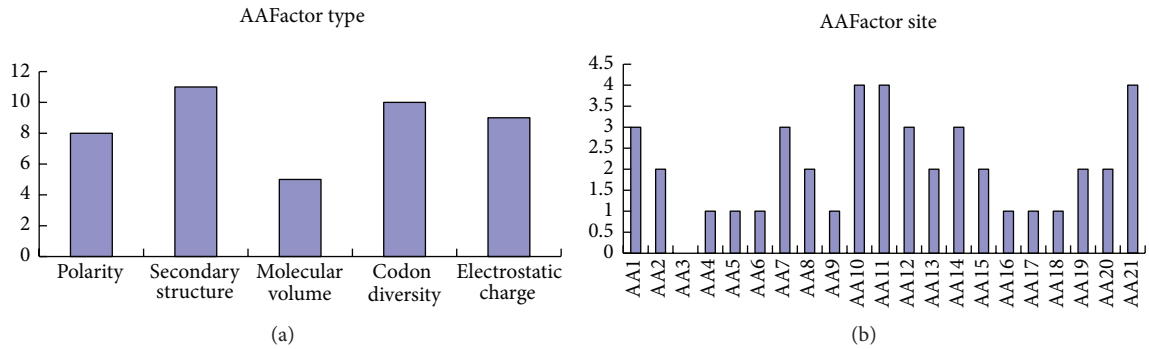

Figure 3: Bar plots to show the distribution in the final optimal feature set for (a) the amino acid factor features, and (b) the corresponding specific site score. It can be seen from panel (a) that the secondary structure and codon diversity were the most important one for predicting the calpain cleavage sites. It can be seen from panel (b) that the residues at site 10, site 11, and site 21 played more roles in the calpain cleavage sites prediction.

MCC were especiallyhigher than other predictor. Further, remaining feature would contain more information of calpain cleavage and are needed more analysis in the feature.

Acknowledgments

The work is fully supported by the National Natural Science Foundation of China under Grant no. 60803102, Natural Science Foundation of Jilin Province under Grant nos. 201215006 and 201101004, Fundamental Research Funds for the Central Universities under Grant no. 11CXPY010, and Opening Fund of Top Key Discipline of Computer Software and Theory in Zhejiang Provincial Colleges at Zhejiang Normal University of China under Grant no. ZSDZZZZXK37.

References

[1] S. J. Storr, N. O. Carragher, M. C. Frame, T. Parr, and S. G. Martin, "The calpain system and cancer," *Nature Reviews Cancer*, vol. 11, no. 5, pp. 364–374, 2011.

[2] I. Bertipaglia and E. Carafoli, "Calpains and human disease," *Subcellular Biochemistry*, vol. 45, pp. 29–53, 2007.

[3] S. J. Franco and A. Huttenlocher, "Regulating cell migration: calpains make the cut," *Journal of Cell Science*, vol. 118, no. 17, pp. 3829–3838, 2005.

[4] Y. X. Fan, Y. Zhang, and H. B. Shen, "LabCaS: labeling calpain substrate cleavage sites from amino acid sequence using conditional random fields," *Proteins*, vol. 81, no. 4, pp. 622–634, 2013.

[5] D. E. Croall and K. Ersfeld, "The calpains: modular designs and functional diversity," *Genome Biology*, vol. 8, no. 6, article 218, 2007.

[6] Z. Liu, J. Cao, X. Gao, Q. Ma, J. Ren, and Y. Xue, "GPS-CCD: a novel computational program for the prediction of calpain cleavage sites," *PLoS One*, vol. 6, no. 4, Article ID e19001, 2011.

[7] P. Tompa, P. Buzder-Lantos, A. Tantos et al., "On the sequential determinants of calpain cleavage," *Journal of Biological Chemistry*, vol. 279, no. 20, pp. 20775–20785, 2004.

[8] S. E. Boyd, M. G. De La Banda, R. N. Pike, J. C. Whisstock, and G. B. Rudy, "PoPS: a computational tool for modeling and predicting protease specificity," in *Proceedings of the IEEE Computational Systems Bioinformatics Conference (CSB '04)*, pp. 372–381, August 2004.

[9] S. E. Boyd, R. N. Pike, G. B. Rudy, J. C. Whisstock, and M. G. De La Banda, "Pops: a computational tool for modeling and predicting protease specificity," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 3, pp. 551–585, 2005.

[10] J. Verspurten, K. Gevaert, W. Declercq, and P. Vandenabeele, "SitePredicting the cleavage of proteinase substrates," *Trends in Biochemical Sciences*, vol. 34, no. 7, pp. 319–323, 2009.

[11] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 2, pp. 63–92, 2009.

[12] K.-C. Chou, "The convergence-divergence duality in lectin domains of selectin family and its implications," *FEBS Letters*, vol. 363, no. 1-2, pp. 123–126, 1995.

[13] 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.

[14] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[15] Y. N. Zhang, D. J. Yu, S. S. Li, Y. X. Fan, Y. Huang, and H. B. Shen, "Predicting protein-ATP binding sites from primary sequence through fusing bi-profile sampling of multi-view features," *BMC Bioinformatics*, vol. 13, article 118, 2012.

[16] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, article 374, 2000.

[17] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6395–6400, 2005.

[18] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein in intrinsic disorder," *BMC Bioinformatics*, vol. 7, article 208, 2006.

[19] J. Cheng, A. Z. Randall, M. J. Sweredoski, and P. Baldi, "SCRATCH: a protein structure and structural feature prediction server," *Nucleic Acids Research*, vol. 33, no. 2, pp. 72–76, 2005.

[20] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, pp. 236–247, 2011.

[21] K.-C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *Proteins*, vol. 43, pp. 246–255, 2001, Erratum in *Proteins*, vol. 44, pp. 60, 2001.

[22] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[23] 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.

[24] X. Zhao, Z. Ma, and M. Yin, "Using support vector machine and evolutionary profiles to predict antifreeze protein sequences," *International Journal of Molecular Sciences*, vol. 13, no. 2, pp. 2196–2207, 2012.

[25] X.-W. Zhao, Z.-Q. Ma, and M.-H. Yin, "Predicting protein-protein interactions by combing various sequence-derived features into the general form of chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 19, no. 5, pp. 492–500, 2012.

[26] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machine," *CM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2001.

[27] LIBSVM, 2012, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[28] X. Zhao, X. Li, Z. Ma, and M. Yin, "Prediction of lysine ubiquitylation with ensemble classifier and feature selection," *International Journal of Molecular Sciences*, vol. 12, no. 12, pp. 8347–8361, 2011.

[29] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[30] K.-C. Chou and H.-B. Shen, "Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms," *Nature Protocols*, vol. 3, no. 2, pp. 153–162, 2008.

[31] H. Mohabatkar, "Prediction of cyclin proteins using chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[32] X. T. Li and M. H. Yin, "Application of differential evolution algorithm on self-potential data," *Plos One*, vol. 7, no. 12, Article ID E51199, 2012.

[33] X. W. Zhao, W. Y. Zhang, X. Xu, Z. Q. Ma, and M. H. Yin, "Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs," *Plos One*, vol. 7, no. 10, Article ID E46302, 2012.

[34] X. T. Li, J. Zhang, and M. H. Yin, "Animal migration optimization: an optimization algorithm inspired by animal migration behavior," *Neural Computing and Applications*, 2013.

[35] 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.

[36] T. Huang, W. Cui, L. Hu, K. Feng, Y.-X. Li, and Y.-D. Cai, "Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles," *PLoS One*, vol. 4, no. 12, Article ID e8126, 2009.

[37] X. T. Li and M. H. Yin, "An opposition-based differential evolution algorithm for permutation flow shop scheduling based on diversity measure," *Advances in Engineering Software*, vol. 55, pp. 10–31, 2013.

[38] X. Li, J. Wang, J. Zhou, and M. Yin, "A perturb biogeography based optimization with mutation for global numerical optimization," *Applied Mathematics and Computation*, vol. 218, no. 2, pp. 598–609, 2011.

[39] D. B. Cai and M. H. Yin, "On the utility of landmarks in SAT based planning," *Knowledge-Based Systems*, vol. 36, pp. 146–154, 2012.

[40] J. P. Zhou, P. Huang, M. H. Yin, and C. G. Zhou, "Phase transitions of EXPSPACE-complete problems," *International Journal of Foundations of Computer Science*, vol. 21, no. 6, pp. 1073–1088, 2010.

[41] J. P. Zhou, M. H. Yin, X. T. Li, and J. Y. Wang, "Phase transitions of EXPSPACE-complete problems: a further step," *International Journal of Foundations of Computer Science*, vol. 23, no. 1, pp. 173–184, 2012.