Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 530696, 6 pages
http://dx.doi.org/10.1155/2013/530696

Research Article

Naïve Bayes Classifier with Feature Selection to Identify Phage Virion Proteins

Peng-Mian Feng,[1] Hui Ding,[2] Wei Chen,[3] and Hao Lin[2]

[1] *School of Public Health, Hebei United University, Tangshan 063000, China*
[2] *Key Laboratory for Neuroinformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[3] *Department of Physics, School of Sciences, Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*

Correspondence should be addressed to Wei Chen; greatchen@heuu.edu.cn and Hao Lin; hlin@uestc.edu.cn

Received 10 March 2013; Revised 16 April 2013; Accepted 28 April 2013

Academic Editor: Yanxin Huang

Knowledge about the protein composition of phage virions is a key step to understand the functions of phage virion proteins. However, the experimental method to identify virion proteins is time consuming and expensive. Thus, it is highly desirable to develop novel computational methods for phage virion protein identification. In this study, a Naïve Bayes based method was proposed to predict phage virion proteins using amino acid composition and dipeptide composition. In order to remove redundant information, a novel feature selection technique was employed to single out optimized features. In the jackknife test, the proposed method achieved an accuracy of 79.15% for phage virion and nonvirion proteins classification, which are superior to that of other state-of-the-art classifiers. These results indicate that the proposed method could be as an effective and promising high-throughput method in phage proteomics research.

1. Introduction

Phage is a virus that infects and replicates within bacteria. Phages are widely distributed in locations populated by bacterial hosts, such as soil or the intestines of animals. A complete infectious phage viral particle (also, namely, phage virion) consists of an inner core of nucleic acid which gives the virus infectivity and a protein coat (called a capsid) which encases the nucleic acid and provides specificity, that is, determines which organisms the virus can infect.

The nucleic acid of phage virions is either RNA or DNA. Proteins of phage virions include structural proteins and nonstructural proteins. Structural proteins commonly termed "phage virion proteins" are essential materials of the infectious viral particles, including shell proteins, envelope proteins, and virus particle enzymes. Nonstructural proteins (namely, phage nonvirion proteins) refer to that encoded by the viral genome and play important roles in biological process of viral genome replication and expression, but they do not bind to phage virions. Due to the distinct functions between phage virion proteins and phage nonvirion proteins, knowledge about the protein composition of phage virions is an essential step to further understand the functions of phage virions.

Although the use of mass spectrometry (MS) for the identification of phage virion proteins has become popular [1], it has not kept pace with the explosive growth of protein sequences generated in the postgenomic age. Hence, it is highly desired to develop automated methods for timely and reliably classifying the protein composition of phage virions.

To the best of our knowledge, there is no computational system for the classification of phage virion proteins. In the current study, we propose a Naïve Bayes based computational model for predicting phage virion proteins using amino acid compositions and dipeptide compositions. The correlation-based feature subset selection algorithm [2] was introduced to find the optimal feature set. By using the optimized features, the proposed model was evaluated in a benchmark dataset

in the jackknife test. The performance demonstrates that this model could be a potentially useful tool for the annotation of the phage proteins.

According to some recent comprehensive reviews [3, 4] and demonstrated by a series of recent publications [5–10], to establish a really useful statistical predictor, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web server for the predictor that is accessible to the public. In the following, let us describe how to deal with these steps one by one.

2. Materials and Methods

2.1. Dataset. The raw datasets adopted in this research were extracted from the UniProt [11]. For the purpose of obtaining a reliable benchmark dataset, the following steps were considered. Firstly, only the experimentally confirmed phage virion and phage nonvirion protein sequences were included. Secondly, the sequences which are fragments of other proteins were dislodged. Thirdly, sequences containing nonstandard letters, that is, "B," "X," or "Z," were excluded as their meanings are ambiguous. After following the previous strict screening procedures, we obtained 121 phage virion protein sequences and 231 phage nonvirion protein sequences.

To prepare a high quality dataset, the CD-HIT program [12] was used to prune the data. By setting the cutoff of sequence identity to 40%, 307 sequences were remained in the final benchmark dataset, including 99 phage virion protein sequences and 208 phage nonvirion protein sequences.

2.2. Feature Vector. One of the most important parts for identifying protein attributes is to generate a set of proper informative parameters to encode the protein sequences. To avoid completely losing the sequence-order information, the pseudo amino acid composition (PseAAC) was proposed [13, 14] to replace the simple amino acid composition (AAC) for representing the sample of a protein. Since the concept of PseAAC was proposed in 2001 [13], it has been widely used to study various attributes of proteins, such as identifying bacterial virulent proteins [15], predicting supersecondary structure [16], predicting protein subcellular location [16–19], predicting membrane protein types [20], discriminating outer membrane proteins [21], identifying antibacterial peptides [22], identifying allergenic proteins [23], predicting metalloproteinase family [24], predicting protein structural class [25], identifying GPCRs and their types [26], identifying protein quaternary structural attributes [27], predicting protein submitochondria locations [28], identifying risk type of human papillomaviruses [29], identifying cyclin proteins [30], predicting GABA(A) receptor proteins [31], and classifying amino acids [32], among many others (see a long list of papers cited in the References section of [3]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [7, 9], as well as other biological samples (see, e.g., [33, 34]). Because it has been widely and increasingly used, recently two powerful softwares, called "PseAAC-Builder" [35] and "propy" [36], were established for generating various special Chou's pseudoamino acid compositions.

The amino acid composition and dipeptide composition are the general forms of PseAAC and the simplest parameters, which also have been widely applied in the realm of protein prediction [37–40]. Hence, every protein sequence in the benchmark dataset was encoded in a discrete vector as

$$\mathbf{F} = [f_1, f_2, \ldots, f_{420}]^T, \quad (1)$$

where f_i is the normalized occurrence frequencies of the 20 amino acids ($i = 1, 2, \ldots, 20$) and the 400 dipeptides ($i = 21, 22, \ldots, 420$) in the protein sequence, respectively. T is the transposing operator.

2.3. Feature Selection. Inclusion of redundant and noisy features in the model building process would cause poor predictive performance and increased computation. Feature selection is the process of removing irrelevant features and is extremely useful in reducing the dimensionality of the data and improving the predictive accuracy. To reduce the dimension of the feature space and improve the precision of phage virion and nonvirion protein classification, the filter method Correlation-based Feature Selection [2] combined with Best-first search strategy was used in the process of feature selection in the current work.

The process starts with an empty set of features and generates all possible single feature expansions. The subset with the highest accuracy is chosen and expanded in the same way by adding single features. If the accuracy does not maximize with the expansion of a subset, the search drops back to the next best unexpanded subset and continues from there until all features are added. The subset with the highest accuracy will be selected as the final optimized feature set [41].

2.4. Naïve Bayes. Naïve Bayes is an effective statistical classification algorithm [42] and has been successfully used in the realm of bioinformatics [43–46]. The basic theory of Naïve Bayes is similar to that of Covariance Determinant (CD) [47–52]. But for Naïve Bayes, it assumes the attribute variables to be independent from each other given the outcome. This assumption greatly simplifies the calculation of conditional probabilities and also overcomes the divergent problem when using the CD prediction engine to deal with those systems in which the components of constituent feature vectors are normalized.

In the Naïve Bayes framework, a classification problem can be seen as the problem of finding the outcome with maximum probability given a set of observed variables. Given a phage viral protein example, described by its feature vector $\mathbf{F} = (f_1, f_2, \ldots, f_n)$, we are looking for a class $\mathbf{C}$ that maximizes the likelihood $\mathbf{P}(\mathbf{F} \mid \mathbf{C}) = \mathbf{P}(f_1, f_2, \ldots, f_n \mid \mathbf{C})$.

Since the current work is intend to classify phage virion and nonvirion proteins, a binary class $\mathbf{C} \in \{0, 1\}$ was generated, where 1 denotes that the sample was predicted as a phage virion protein and 0 denotes phage nonvirion protein. For the binary classification, the class for the protein sample could be determined by comparing two posteriors as

$$\frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} = \frac{P\left(C = 1\right) \prod_{i=1}^{n} P_i\left(f_i \mid C = 1\right)}{P\left(C = 0\right) \prod_{i=1}^{n} P_i\left(f_i \mid C = 0\right)}. \tag{2}$$

Taking the logarithm of (2), we obtain

$$\log \frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} = \log \frac{P\left(C = 1\right)}{P\left(C = 0\right)} + \sum_{i=1}^{n} \log \frac{P_i\left(f_i \mid C = 1\right)}{P_i\left(f_i \mid C = 0\right)}. \tag{3}$$

Hence the sample will be predicted as 1 (phage virion protein) if

$$\log \frac{P\left(C = 1 \mid F = f_1, f_2, \ldots, f_n\right)}{P\left(C = 0 \mid F = f_1, f_2, \ldots, f_n\right)} \geq \theta \tag{4}$$

and 0 (phage nonvirion protein) for otherwise. θ is the threshold determining the trade-off between sensitivity and specificity and can be trained on the training dataset to maximize the prediction performance.

2.5. Performance Evaluation. The performance of the proposed model was evaluated using sensitivity (Sn), specificity (Sp), and accuracy (Acc), which are expressed as

$$\begin{aligned} \text{Sn} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Sp} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{Acc} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}. \end{aligned} \tag{5}$$

TP, TN, FP, and FN represent the number of the correctly recognized phage virion proteins, the number of the correctly recognized phage nonvirion proteins, the number of phage nonvirion proteins recognized as phage virion proteins, and the number of phage virion proteins recognized as phage nonvirion proteins, respectively.

As the performance of the current classifier depends on the threshold θ as given in (4), the threshold independent parameter, receiver operating characteristic curve, was employed as well. Therefore, the quality of a classifier can be objectively evaluated by measuring the area under the receiver operating characteristic curve (auROC). The value of auROC score ranges from 0 to 1, with a score of 0.5 corresponding to a random guess and a score of 1.0 indicating a perfect separation.

Table 1: Predictive performance of Naïve Bayes based on different features.

Feature dimensions	Sn (%)	Sp (%)	Acc (%)	auROC
420	53.54	83.17	75.57	0.758
38	75.76	80.77	79.15	0.855

3. Results and Discussion

Three cross-validation methods, namely, subsampling test, independent dataset test, and jackknife test, are often employed to evaluate the predictive capability of a predictor. Among the three methods, the jackknife test is deemed the most objective and rigorous one that can always yield a unique outcome as demonstrated by a penetrating analysis in a recent comprehensive review [53] and hence has been widely and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., [7, 19, 21, 30, 54–56]). Accordingly, the jackknife test was used to examine the performance of the model proposed in the current study. In the jackknife test, each sequence in the training dataset is in turn singled out as an independent test sample, and all the rule parameters are calculated without including the one being identified.

3.1. Prediction of Phage Virion Proteins. We trained the Naïve Bayes classifier using Waikato Environment for Knowledge Analysis (WEKA) [57] on the benchmark dataset. As shown in Table 1, an auROC score of 0.758 and an accuracy of 75.57% with an average sensitivity of 53.54% and an average specificity of 83.17% were obtained for the classification of phage virion and nonvirion proteins by using all the 420 features, that is, 20 amino acid compositions and 400 dipeptide compositions.

In order to identify prominent features that can distinguish between phage virion and nonvirion proteins, feature selection method as introduced in Section 2.3 was carried out to eliminate the redundant features using WEKA in a tenfold cross-validation approach on the benchmark dataset. We found that the proposed method achieved a maximum accuracy of 79.48% and auROC of 0.86 when the feature dimension reduced to 38 (i.e., V, T, A, H, K, E, R, S, LE, VT, VG, MK, TA, TS, AT, HI, KL, KI, KH, KN, KK, KD, KE, KW, KR, DK, EF, EL, EV, EK, EE, EW, CE, WK, RE, SG, GV, and GG). The jackknife test results of the Naïve Bayes classifier based on the 38 optimized features were listed in Table 1. As it can be seen from Table 1, the current method yielded a best auROC score of 0.855 and a predictive accuracy of 79.15% with an average sensitivity of 75.76% and an average specificity of 80.77% (Table 1). Both predictive accuracy and auROC are higher than that of the model based on the 420 features.

3.2. Comparison with Other Methods. To the best of our knowledge, there exists no theoretical method for phage virion and nonvirion protein classifications. Therefore, we cannot provide the comparison analysis with published results to confirm that the model proposed here is superior to

Table 2: Comparison of Naïve Bayes with other methods by using optimized features.

Classifier	Sn (%)	Sp (%)	Acc (%)	auROC
BayesNet	68.69	79.81	76.22	0.799
RBFnetwork	72.73	82.21	79.15	0.839
Random Forest	55.56	84.62	75.24	0.802
LogitBoot	52.53	85.10	74.59	0.795
SVM	63.64	86.54	79.15	0.836
J48	61.62	77.88	72.64	0.671
Naïve Bayes	75.76	80.77	79.15	0.855

other methods. However, the proposed Naïve Bayes classifier was compared with other state-of-the-art classifiers, that is, BayesNet, RBFnetwork, Random Forest, J48, Support Vector Machine (SVM), and LogitBoot. All the classifiers were compared on the benchmark dataset based on the optimized features (i.e., V, T, A, H, K, E, R, S, LE, VT, VG, MK, TA, TS, AT, HI, KL, KI, KH, KN, KK, KD, KE, KW, KR, DK, EF, EL, EV, EK, EE, EW, CE, WK, RE, SG, GV, and GG). Their best predictive results from jackknife test were shown in Table 2.

The predictive accuracy of Naïve Bayes is approximately 3%, 4%, 5%, and 7% higher than that of the BayesNet, Random Forest, LogitBoot, and J48 classifiers, respectively. Although the accuracies of RBFnetwork and SVM are equal to that of Naïve Bayes, their auROC scores are lower than that of Naïve Bayes. These results indicate that the proposed Naïve Bayes model can be effectively used to classify phage virion and nonvirion proteins.

4. Conclusions

In this study, the Naïve Bayes classifier with feature selection method is presented to identity phage virion proteins based on the primary sequence information. By using Correlation-based Feature Subset Selection algorithm, the feature dimensions were reduced, and 38 prominent features that could remarkably improve the predictive accuracies were obtained. However, the detailed analyses of the selected features are required to provide more information about their roles in biological activity. The accuracy for the classification of phage virion and nonvirion proteins reached 79.15% in the jackknife test, indicating that the proposed method is an effective tool for phage virion protein identification. It is expected that the presented model will provide novel insights into the research on phage proteomics. Since user-friendly and publicly accessible web servers represent the future direction for developing practically more useful predictors [58], we shall make efforts in our future work to provide a web server for the method presented in this paper.

Acknowledgments

The authors wish to express their gratitude to three anonymous reviewers whose constructive comments were very helpful in strengthening the presentation of this paper. This work was supported by the National Nature Scientific Foundation of China (nos. 61100092, 61202256), the Fundamental Research Funds for the Central Universities (ZYGX2012J113).

References

[1] R. Lavigne, P. J. Ceyssens, and J. Robben, "Phage proteomics: applications of mass spectrometry," *Methods in Molecular Biology*, vol. 502, pp. 239–251, 2009.

[2] M. A. Hall, *Correlation-Based Feature Selection for Machine Learning: Data Mining, Inference and Prediction*, Springer, Berlin, Germany, 2008.

[3] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[4] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, 2013.

[5] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins," *Molecular BioSystems*, vol. 9, no. 9, pp. 634–644, 2013.

[6] X. Xiao X, P. Wang, W. Z. Lin, J. H. Jia, and K. C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

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

[8] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular Biosystems*, vol. 8, no. 2, pp. 629–641, 2012.

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

[10] Y. Xu, J. Ding, L. Y. Wu, and K.-C. Chou, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[11] UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, 2012.

[12] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[13] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001.

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

[15] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[16] D. Zou, Z. He, J. He, and Y. Xia, "Supersecondary structure prediction using Chou's pseudo amino acid composition," *Journal of Computational Chemistry*, vol. 32, no. 2, pp. 271–278, 2011.

[17] S. W. Zhang, Y. L. Zhang, H. F. Yang, C. H. Zhao, and Q. Pan, "Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies," *Amino Acids*, vol. 34, no. 4, pp. 565–572, 2008.

[18] K. K. Kandaswamy, G. Pugalenthi, S. Möller et al., "Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 12, pp. 1473–1479, 2010.

[19] S. Mei, "Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning," *Journal of Theoretical Biology*, vol. 310, pp. 80–87, 2012.

[20] Y. K. Chen and K. B. Li, "Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 318, pp. 1–12, 2013.

[21] M. Hayat and A. Khan, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chou's PseAAC," *Protein and Peptide Letters*, vol. 19, no. 4, pp. 411–421, 2012.

[22] M. Khosravian, F. K. Faramarzi, M. M. Beigi, M. Behbahani, and H. Mohabatkar, "Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods," *Protein and Peptide Letters*, vol. 20, no. 2, pp. 180–186, 2012.

[23] H. Mohabatkar, M. M. Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[24] M. M. Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[25] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[26] R. Zia-Ur and A. Khan, "Identifying GPCRs and their types with Chou's pseudo amino acid composition: an approach from multi-scale energy representation and position specific scoring matrix," *Protein and Peptide Letters*, vol. 19, no. 8, pp. 890–903, 2012.

[27] X. Y. Sun, S. P. Shi, J. D. Qiu, S. B. Suo, S. Y. Huang, and R. P. Liang, "Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform," *Molecular BioSystems*, vol. 8, no. 12, pp. 3178–3184, 2012.

[28] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[29] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[30] 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.

[31] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[32] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[33] 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.

[34] T. Huang, J. Wang, Y. D. Cai, H. Yu, and K. C. Chou, "Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[35] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[36] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[37] L. Montanucci, P. Fariselli, P. L. Martelli, and R. Casadio, "Predicting protein thermostability changes from sequence upon multiple mutations," *Bioinformatics*, vol. 24, no. 13, pp. i190–i195, 2008.

[38] M. M. Gromiha and M. X. Suresh, "Discrimination of mesophilic and thermophilic proteins using machine learning algorithms," *Proteins*, vol. 70, no. 4, pp. 1274–1279, 2008.

[39] L. C. Wu, J. X. Lee, H. D. Huang, B. J. Liu, and J. T. Horng, "An expert system to predict protein thermostability using decision tree," *Expert Systems with Applications*, vol. 36, no. 5, pp. 9007–9014, 2009.

[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 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.

[42] T. M. Mitchell, *Machine Learning*, McGraw-Hill, New York, NY, USA, 1997.

[43] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad, "A naive Bayes model to predict coupling between seven transmembrane domain receptors and G-proteins," *Bioinformatics*, vol. 19, no. 2, pp. 234–240, 2003.

[44] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe, and M. K. Showe, "Naïve Bayes for microRNA target predictions—machine learning for microRNA targets," *Bioinformatics*, vol. 23, no. 22, pp. 2987–2992, 2007.

[45] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites," *Bioinformatics*, vol. 26, no. 15, pp. 1841–1848, 2010.

[46] F. Sambo, E. Trifoglio, B. Di Camillo, G. M. Toffolo, and C. Cobelli, "Bag of Naïve Bayes: biomarker selection and classification from genome-wide SNP data," *BMC Bioinformatics*, vol. 13, supplement 14, article S2, 2012.

[47] K. C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," *Proteins*, vol. 21, no. 4, pp. 319–344, 1995.

[48] G. P. Zhou, "An intriguing controversy over protein structural class prediction," *Journal of Protein Chemistry*, vol. 17, no. 8, pp. 729–738, 1998.

[49] K. C. Chou and D. W. Elrod, "Using discriminant function for prediction of subcellular location of prokaryotic proteins," *Biochemical and Biophysical Research Communications*, vol. 252, no. 1, pp. 63–68, 1998.

[50] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins*, vol. 44, no. 1, pp. 57–59, 2001.

[51] Y. X. Pan, Z. Z. Zhang, Z. M. Guo, G. Y. Feng, Z. D. Huang, and L. He, "Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach," *Journal of Protein Chemistry*, vol. 22, no. 4, pp. 395–402, 2003.

[52] G. P. Zhou and K. Doctor, "Subcellular location prediction of apoptosis proteins," *Proteins*, vol. 50, no. 1, pp. 44–48, 2003.

[53] K. C. Chou and H. B. Shen, "Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms," *Natural Science*, vol. 2, no. 10, pp. 1090–1103, 2010.

[54] M. Hayat and A. Khan, "MemHyb: predicting membrane protein types by hybridizing SAAC and PSSM," *Journal of Theoretical Biology*, vol. 292, pp. 93–102, 2012.

[55] X. Xiao, Z. C. Wu, and K. C. Chou, "iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *Journal of Theoretical Biology*, vol. 284, no. 1, pp. 42–51, 2011.

[56] K. C. Chou, Z. C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[57] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[58] 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.