Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 283129, 7 pages
http://dx.doi.org/10.1155/2013/283129

Research Article

Identification of Protein Pupylation Sites Using Bi-Profile Bayes Feature Extraction and Ensemble Learning

Xiaowei Zhao,[1,2] Jian Zhang,[1] Qiao Ning,[1] Pingping Sun,[1] Zhiqiang Ma,[1] and Minghao Yin[2]

[1] *College of Computer Science and Information Technology, Northeast Normal University, 2555 Jingyue Street, Changchun 130117, China*
[2] *Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun 130117, China*

Correspondence should be addressed to Zhiqiang Ma; zhiqiang.ma967@gmail.com and Minghao Yin; minghao.yin197@gmail.com

Received 13 July 2013; Accepted 1 August 2013

Academic Editor: William Guo

Pupylation, one of the most important posttranslational modifications of proteins, typically takes place when prokaryotic ubiquitin-like protein (Pup) is attached to specific lysine residues on a target protein. Identification of pupylation substrates and their corresponding sites will facilitate the understanding of the molecular mechanism of pupylation. Comparing with the labor-intensive and time-consuming experiment approaches, computational prediction of pupylation sites is much desirable for their convenience and fast speed. In this study, a new bioinformatics tool named EnsemblePup was developed that used an ensemble of support vector machine classifiers to predict pupylation sites. The highlight of EnsemblePup was to utilize the Bi-profile Bayes feature extraction as the encoding scheme. The performance of EnsemblePup was measured with a sensitivity of 79.49%, a specificity of 82.35%, an accuracy of 85.43%, and a Matthews correlation coefficient of 0.617 using the 5-fold cross validation on the training dataset. When compared with other existing methods on a benchmark dataset, the EnsemblePup provided better predictive performance, with a sensitivity of 80.00%, a specificity of 83.33%, an accuracy of 82.00%, and a Matthews correlation coefficient of 0.629. The experimental results suggested that EnsemblePup presented here might be useful to identify and annotate potential pupylation sites in proteins of interest. A web server for predicting pupylation sites was developed.

1. Introduction

As the firstly identified posttranslational small protein modifier in prokaryotes, prokaryotic ubiquitin-like protein (Pup) in *Mycobacterium tuberculosis* (Mtb) is an important signal for the selective degradation of proteins [1]. Pup attaches to substrate lysine via isopeptide bonds in a manner reminiscent of ubiquitin (Ub) and ubiquitin-like modifier (Ubl) conjugation to proteins in eukaryotes [2]. Although pupylation and ubiquitylation are functional similarity, the enzymology of pupylation and ubiquitylation is different [3]. Generally, there are three-step reaction and three kinds of enzymes participating in the eukaryotic ubiquitylation process, including ubiquitin-activating enzymes, ubiquitin-conjugating enzymes, and ubiquitin ligases [4, 5], but only two-step reaction and two kinds of enzymes participating in the prokaryotic pupylation process. Firstly, the Pup-GGQ C-terminal is deamidated to -GGE by deamidase of Pup [6], and then the proteasome accessory factor A (PafA) attaches the deamidated Pup to specific lysine residues of substrates [7].

Since identification of protein pupylation sites is of fundamental importance to understand the molecular mechanism of pupylation in biological systems, much interest has focused on this field and large-scale proteomics technology has been applied to identify pupylation proteins and pupylation sites [8–10]. However, the experimental determination of exact modified sites of pupylated substrates is labor intensive and time consuming, especially for large-scale data sets. In this regard, the computation approaches which could effectively and accurately identify the pupylation sites are urgently needed. Liu et al. had constructed the first online predictor, GPS-PUP, for the prediction of pupylation sites [11]. In their method, 127 experimentally identified pupylation sites in 109 prokaryotic proteins had been utilized as the training dataset, with an accuracy of 0.789 and an MCC of 0.286. However, there is significant room for improvement of the prediction performance.

Table 1: Number of pupylation and non-pupylation sites in each dataset.

	Pupylation proteins sequences	Positive sites	Negative sites
Dataset 1	153	183	2288
Dataset 2	109	127	1405

In this study, the prediction performance of pupylation sites has been improved by using a new encoding scheme, Bi-profile Bayes feature extraction (BPB), which has been widely used to deal with diverse prediction topics in the field of bioinformatics [12–15]. Since the new constructed pupylation sites dataset was highly imbalanced: the number of pupylation sites was much smaller than the number of nonpupylation sites, the ensemble learning method was adopted here to deal with the imbalanced data classification problem. The performance of EnsemblePup was measured with a sensitivity of 79.49%, a specificity of 82.35%, an accuracy of 85.43%, and a Matthews correlation coefficient of 0.617 using the 5-fold cross validation on the training dataset. When compared with other existing methods on a benchmark dataset, the EnsemblePup provided better predictive performance, with a sensitivity of 80.00%, a specificity of 83.33%, an accuracy of 82.00%, and a Matthews correlation coefficient of 0.629. The experimental results suggested that EnsemblePup presented here might be useful to identify and annotate potential pupylation sites in proteins of interest. A web server for predicting pupylation sites was developed and was available at http://210.47.24.217:8080/EnsemblePup/.

The organization of this paper is as follows. Section 2 introduces the dataset for establishing the predictor, the vector encoding schemes, and the proposed prediction model. Section 3 shows the experimental results, discusses the performance of the proposed predictor, and compares the proposed predictor with other methods. Finally Section 4 gives the conclusions.

2. Materials and Methods

2.1. Dataset. The pupylated proteins used in this study were extracted from PupDB [3]. Protein sequences with less than 50 amino acids were excluded because they may be just fragments [16, 17]. Protein sequences including nonstandard amino acids, such as "B," "J," "O," "U," "X," and "Z" were excluded as well. As a result, there were 182 pupylated proteins with 215 known pupylation sites. After a homology-reducing screening procedure by using CD-HIT [18, 19] to remove those proteins that had 40% sequence identity to any other, we finally got 153 pupylated proteins with 183 positive sites, which constructed the nonredundant training dataset named as Dataset 1 in this study (see Supporting Information Text S1 available online at http://dx.doi.org/10.1155/2013/283129). In order to fairly compare our proposed method EnsemblePup with a previously developed method GPS-PUP, the dataset collected by Liu et al. [11] was also adopted here. We named it as Dataset 2 in this work, and the details of Dataset 2 were listed in Table 1.

Subsequently, similar to the development of other PTM site predictors [20, 21], the sliding window strategy was utilized to extract positive and negative samples. In order to ensure the peptides (sequence fragments) with a unified length, a nonexisting residue coded by "-" was used to fill the corresponding position. Peptides with pupylation lysine as the middle residue were regarded as positive samples, and the remaining peptides with nonpupylation lysine as the middle residue were regarded as negative samples.

2.2. Vector Encoding Schemes. In this study, the Bi-profile Bayes feature extraction (BPB) based encoding scheme was used. For details on this encoding scheme, readers are advised to refer to Shao et al. [14]. Briefly, let $S = s_1, s_2, \ldots, s_n$ represent a sequence fragment, where s_j denotes one amino acid and n stands for the length of the sequence fragment. S belongs to two categories C_1 and C_{-1}, where C_1 and C_{-1} represent pupylation sites and nonpupylation sites, respectively. Then, a feature vector can be described as

$$\vec{P} = (p_1, p_2, \ldots, p_n, p_{n+1}, \ldots, p_{2n}), \quad (1)$$

where $p_1, p_2, \ldots, p_n$ represent the posterior probability of each amino acid at each position for the sequence fragment of pupylation sites (category C_1) and $p_{n+1}, \ldots, p_{2n}$ represent the posterior probability of each amino acid at each position for the sequence fragment of nonpupylation sites (category C_{-1}), which is the so-called Bi-profile. In this paper, the posterior probability is estimated by the occurrence of each amino acid at each position in the training datasets [14].

The binary encoding scheme was also carried out here to be compared with the BPB encoding scheme. As it is known to all, there are 20 types of amino acids in protein sequences, which are given as ACDEFGHIKLMNPQRSTVWY. Therefore, each amino acid is represented by a 20-dimensional binary vector; that is, A corresponds to (10000000000000000000), C corresponds to (01000000000000000000), and Y corresponds to (00000000000000000001). For each sequence fragment with length n, the total dimension of the binary feature vector is $20 \times (n - 1)$, since the central amino acid is always K, which is not necessary to be considered.

2.3. Support Vector Machine Learning and Imbalanced Data. Support vector machine (SVM) is a popular machine learning algorithm mainly used in dealing with binary classification problems. SVM looks for a rule that best maps each member of training set to the correct classification [22, 23], and it has been widely used in bioinformatics community. In this paper, LIBSVM package [24] with radial basis kernels (RBF) is used, where the kernel width parameter γ represents how the samples are transformed to a high dimensional space. Grid search strategy based on 5-fold cross-validation is utilized to find the optimal parameters C and $\gamma \in \{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$, so that a total number of 256 grids are evaluated.

Since the training dataset was imbalanced, in which the number of pupylation sites was much smaller than the

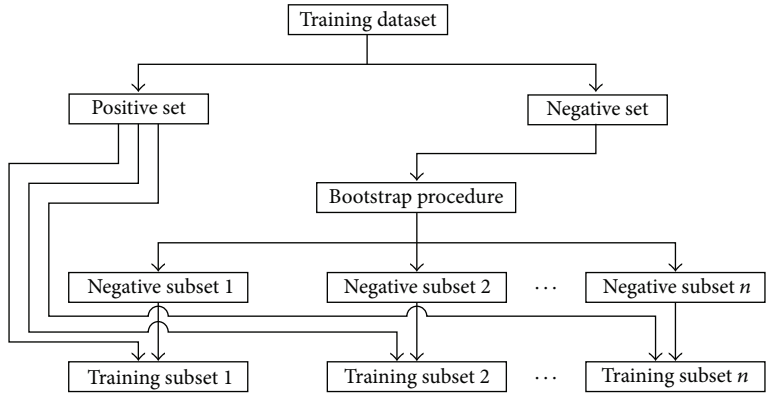

Figure 1: The bootstrap procedure for the imbalanced dataset.

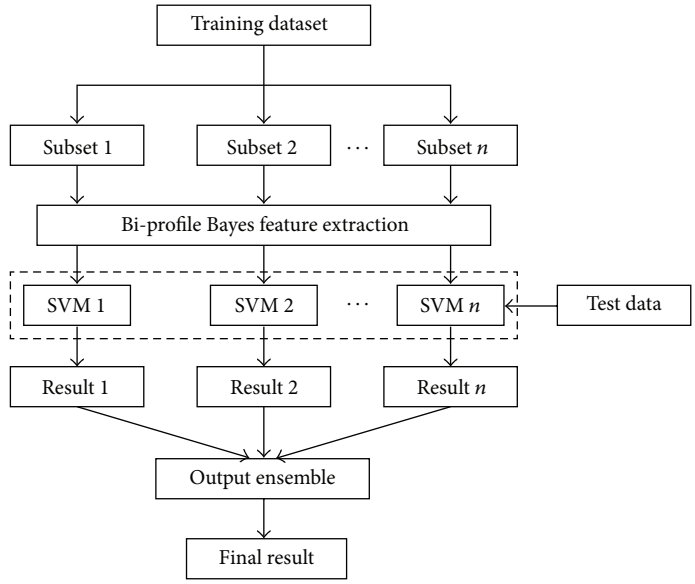

Figure 2: The entire schematic diagram for the prediction of pupylation sites.

number of nonpupylation sites, the bootstrap procedure was used to deal with this situation. As shown in Figure 1, we obtained n training subsets using the bootstrap procedure, where n represented the times of data sampling. In this study, the bootstrap procedure was implemented by WEKA package [25] and the parameter n was set as the ratio of the number of positive samples divided by the number of negative samples.

2.4. The Ensemble Model for Pupylation Sites Identification. Since ensemble learning methods have unique advantages in dealing with high-dimensional and complicated data, there is an increasing use of it in the field of bioinformatics [26–30]. In this study, the ensemble model was established by a collection of SVM classifiers, each was trained on a subset of the original training dataset (obtain by the bootstrap procedure in Figure 1). Figure 2 showed the entire schematic diagram for the prediction of pupylation sites. As shown in Figure 2, the final result was computed from the prediction result of the individual SVM classifier. For example, when given a new unlabeled test data x, the jth SVM classifier returned a probability P_j of x belonged to the positive class, where $j = 1, 2, \ldots, n$. The collection estimated probability was obtained by $P_{\text{Ensemble}} = (1/n) \sum_{j=1}^{n} P_j$.

2.5. Performance Assessment. In this study, 5-fold cross validation and jackknife cross validation tests were chosen for evaluating the proposed predictor. More details about these two methods can be found in two recent papers [31, 32]. In order to evaluate the proposed predictor, four measurements are used: sensitivity (Sn), specificity (Sp), accuracy (Ac), and Matthews correlation coefficient (MCC). They are defined by the following formulas:

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

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

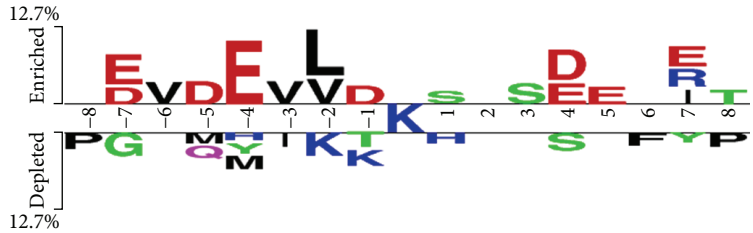

Figure 3: The Two-Sample-Logo of the position-specific residue composition surrounded the pupylation sites and nonpupylation sites. This logo was generated using the web server http://www.twosamplelogo.org/ and only residues significantly enriched and depleted surrounding pupylation sites (t-test, $P < 0.1$) were shown.

Table 2: Results of the SVM prediction on Dataset 1.

SVM classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	A_{ROC}
BPB-SVM15	78.79	67.50	72.60	0.462	0.791
BPB-SVM17	75.76	72.50	73.97	0.480	0.807
BPB-SVM19	69.23	67.65	68.49	0.368	0.738
BPB-SVM21	65.79	80.00	72.60	0.463	0.741
BPB-SVM23	63.16	80.00	71.23	0.436	0.788
Binary-SVM17	55.50	51.52	53.42	0.065	0.527

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}},$$
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}, \quad (2)$$

where TP, TN, FP, and FN stand for the number of true positive, true negative, false positive, and false negative, respectively. In addition, the receiver operating characteristic (ROC) curves and the area under the curve (AUC) values are also carried out.

3. Results and Discussion

3.1. Determination of the Best Window Size. We firstly analyzed the position-specific propensities of the residues surrounding pupylation sites and nonpupylation sites using Two-Sample-Logo, which generated the graphical sequences logo for the relative frequency of the corresponding amino acid at each position around pupylation sites and nonpupylation sites. As shown in Figure 3, we found that the characteristics of the residues had significant differences between pupylation sites and nonpupylation sites. To encapsulate the position-specific propensities of residues for computational prediction, we established SVM prediction models of different lengths (represented as BPB-SVM15, BPB-SVM17, BPB-SVM19, BPB-SVM21, and BPB-SVM23) trained on a balanced training dataset (constructed by sampling a number of nonpupylation sites equal to the number of pupylation sites) using the Bi-profile Bayes feature extraction (BPB) method. As shown in Table 2, after a preliminary evaluation, the optimal window size was 17 in this paper (BPB-SVM17), with 8 residues located upstream and 8 residues located downstream of the pupylation sites in the protein sequence.

Table 3: The comparison of predictive performance between single SVM and ensemble of SVMs using the 5-fold cross validation on Dataset 1.

SVM classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	A_{ROC}
SinglePup	75.76	72.50	73.97	0.480	0.807
EnsemblePup	79.49	82.35	80.82	0.617	0.862

However, when the Bi-profile Bayes feature extraction encoding scheme was replaced by the binary encoding scheme of window size of 17 (represented as Binary-SVM17), the binary encoding scheme showed mediocre prediction performance, the prediction accuracy was 20.55% lower than that of the Bi-profile Bayes feature extraction encoding scheme, which indicated that the Bi-profile Bayes feature extraction encoding scheme has an advantage over the binary encoding scheme in predicting pupylation sites. Therefore, we adopted Bi-profile Bayes feature extraction encoding scheme in this study.

3.2. Comparison of EnsemblePup with a Single SVM Classifier. In order to enhance the prediction performance of the pupylation sites predictor, ensemble learning was used, and the final results were obtained by combining the outputs of different single SVM classifier. Here, we compared the performance of the ensemble of SVM classifiers with that of a single SVM classifier. All experiments were performed and reported the Sn, Sp, Ac, and MCC. The comparison results of the two prediction models by 5-fold cross validation test on the Dataset 1 were shown in Table 3 we can see the ensemble predictor got the accuracy of 80.82%, higher than the result obtained by using a single SVM classifier with 73.97%, and the AUC value was 0.55 higher than that of the a single

Table 4: The comparison of predictive performance between our method and GPS-PUP using the leave-one-out cross validation on Dataset 2.

Prediction method	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	A_{ROC}
GPS-PUP	63.78	80.21	78.85	0.286	0.708
SinglePup	73.91	78.95	76.19	0.526	0.825
EnsemblePup	80.00	83.33	82.00	0.629	0.873

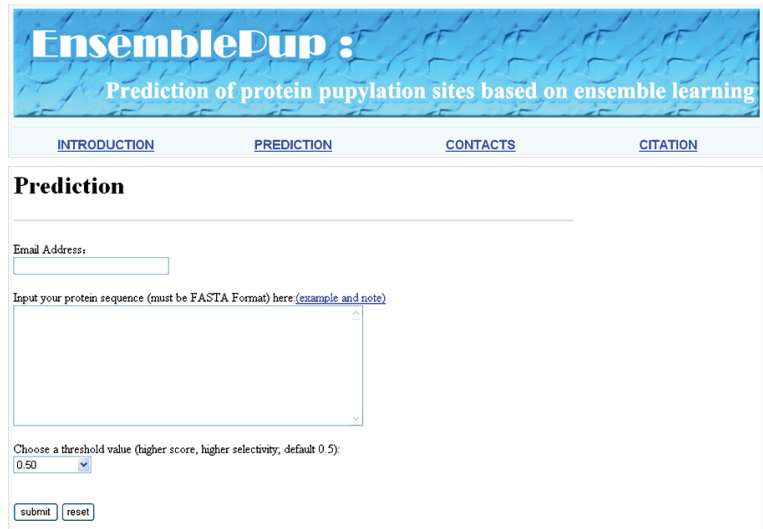

Figure 4: The prediction page of the EnsemblePup web server at http://210.47.24.217:8080/EnsemblePup/.

SVM classifier. In summary, the ensemble learning had an advantage in predicting pupylation sites.

3.3. Comparison of EnsemblePup with Other Methods. We have demonstrated that EnsemblePup could achieve a promising prediction performance in the 5-fold cross validation on Dataset 1. To objectively evaluate our proposed predictor, we further compared the EnsemblePup predictor with GPS-PUP [11]. Liu et al. searched PubMed with the keywords of "pupylation" and "prokaryotic ubiquitin" and collected 127 experimentally identified pupylation sites in 109 prokaryotic proteins; we named the data from Liu et al. as Dataset 2 in this work (the details were listed in Table 1). The compared results were shown in Table 4. As can be seen from the table, the EnsemblePup predictor proposed in this study obtained an accuracy of 82.00%, higher than the GPS-PUP predictor with the accuracy of 80.21%, and MCC of EnsemblePup was 0.343 greater than that of GPS-PUP.

3.4. The EnsemblePup Web Server. The EnsemblePup was implemented in Java and hosted on Windows platform. For the convenience of experimental scientists, we gave a step-by-step guide on how to use it to get the desired results as follows. (i) Open the web server at http://210.47.24.217:8080/EnsemblePup/ and you can see the prediction page on your computer screen, as shown in Figure 4. You must input your email address since the prediction process may take a long time. (ii) Input your query protein sequence to the text box in Figure 4. Note that the input protein sequence must be in the FASTA format. The FASTA format sequence consists of a single initial line beginning with a greater-than symbol (">"), followed by lines of amino acid sequence. You can click on the "example and note" button to see the example protein sequence. (iii) Choose a threshold value in the drop-down list. For prediction with high confidence (less probability of false positive prediction), high threshold should be chosen. (iv) Click on the submit button to see the predicted result. For example, if you use the first sequence in the example page, the prediction results will be ">A0QNF6 K147 0.7450251 yes," which means that the lysine on the position of 147 is a pupylation site with the probability of 0.7450251. Generally, it takes about 50 seconds to predict the pupylation site for a protein sequence shorter than 1000 amino acids before the predicted result appears.

4. Conclusion

Prediction of pupylation sites is important to understand the molecular mechanism of pupylation in biological systems. Though some researchers have focused on this problem, the accuracy of prediction is still not satisfied. In this study, we have presented a new predictor EnsemblePup for the prediction of pupylation sites based on Bi-profile Bayes feature extraction encoding scheme. Since the new constructed pupylation sites dataset was highly imbalanced: the number of pupylation sites was much smaller than the number of nonpupylation sites, the ensemble learning

method was adopted here to deal with the imbalanced data classification problem. The performance of EnsemblePup was measured with a sensitivity of 79.49%, a specificity of 82.35%, an accuracy of 85.43%, and a Matthews correlation coefficient of 0.617 using the 5-fold cross-validation on the training dataset. When compared with other existing methods on a benchmark dataset, the EnsemblePup provided better predictive performance, with a sensitivity of 80.00%, a specificity of 83.33%, an accuracy of 82.00%, and a Matthews correlation coefficient of 0.629. Experimental results have shown that our method is very promising and may be a useful supplement tool to existing methods. Due to the considerable performance, we have made EnsemblePup freely available as a web server. Although the results obtained here were very promising, further investigation was needed to further clarify the mechanism of pupylation process.

Acknowledgments

Jian Zhang and Qiao Ning collected data, wrote codes, and developed the web server. Zhiqiang Ma and Minghao Yin participated in the research design, method assessment, and preparation of the paper. Xiaowei Zhao directed the research and wrote the paper. All authors read and approved the final paper. This research is partially supported by the Science Foundation for Young Teachers of Northeast Normal University (no. 12QNJJ005) and the Natural Science Foundation of JiLin Province (nos. 20101506 and 20110104).

References

[1] R. A. Festa, F. McAllister, M. J. Pearce et al., "Prokayrotic ubiquitin-like protein (Pup) proteome of *Mycobacterium tuberculosis*," *PLoS ONE*, vol. 5, no. 9, Article ID e8589, 2010.

[2] K. H. Darwin, "Prokaryotic ubiquitin-like protein (Pup), proteasomes and pathogenesis," *Nature Reviews Microbiology*, vol. 7, no. 7, pp. 485–491, 2009.

[3] C.-W. Tung, "PupDB: a database of pupylated proteins," *BMC Bioinformatics*, vol. 13, article 40, 2012.

[4] R. L. Welchman, C. Gordon, and R. J. Mayer, "Ubiquitin and ubiquitin-like proteins as multifunctional signals," *Nature Reviews Molecular Cell Biology*, vol. 6, no. 8, pp. 599–609, 2005.

[5] A. Hershko and A. Ciechanover, "The ubiquitin system," *Annual Review of Biochemistry*, vol. 67, pp. 425–479, 1998.

[6] F. Striebel, F. Imkamp, M. Sutter, M. Steiner, A. Mamedov, and E. Weber-Ban, "Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes," *Nature Structural and Molecular Biology*, vol. 16, no. 6, pp. 647–651, 2009.

[7] E. Guth, M. Thommen, and E. Weber-Ban, "Mycobacterial ubiquitin-like protein ligase PafA follows a two-step reaction pathway with a phosphorylated Pup intermediate," *Journal of Biological Chemistry*, vol. 286, no. 6, pp. 4412–4419, 2011.

[8] F. A. Cerda-Maira, F. McAllister, N. J. Bode, K. E. Burns, S. P. Gygi, and K. H. Darwin, "Reconstitution of the *Mycobacterium tuberculosis* pupylation pathway in *Escherichia coli*," *EMBO Reports*, vol. 12, no. 8, pp. 863–870, 2011.

[9] C. Poulsen, Y. Akhter, A. H. Jeon et al., "Proteome-wide identification of mycobacterial pupylation targets," *Molecular Systems Biology*, vol. 6, article 386, 2010.

[10] J. Watrous, K. Burns, W.-T. Liu et al., "Expansion of the mycobacterial 'pUPylome'," *Molecular BioSystems*, vol. 6, no. 2, pp. 376–385, 2010.

[11] Z. Liu, Q. Ma, J. Cao, X. Gao, J. Ren, and Y. Xue, "GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins," *Molecular BioSystems*, vol. 7, no. 10, pp. 2737–2740, 2011.

[12] L. J. Wee, D. Simarmata, Y.-W. Kam, L. F. Ng, and J. C. Tong, "SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction," *BMC Genomics*, vol. 11, article S21, supplement 4, 2010.

[13] C. Jia, T. Liu, A. K. Chang, and Y. Zhai, "Prediction of mitochondrial proteins of malaria parasite using bi-profile Bayes feature extraction," *Biochimie*, vol. 93, no. 4, pp. 778–782, 2011.

[14] J. Shao, D. Xu, S.-N. Tsai, Y. Wang, and S.-M. Ngai, "Computational identification of protein methylation sites through Bi-profile Bayes feature extraction," *PLoS ONE*, vol. 4, no. 3, Article ID e4920, 2009.

[15] J. L. Shao, D. Xu, L. Hu et al., "Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation," *Molecular BioSystems*, vol. 11, no. 8, pp. 2964–2973, 2012.

[16] K.-C. Chou and H.-B. Shen, "Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization," *Biochemical and Biophysical Research Communications*, vol. 347, no. 12, pp. 150–157, 2006.

[17] K.-C. Chou and H.-B. Shen, "Large-scale plant protein subcellular location prediction," *Journal of Cellular Biochemistry*, vol. 100, no. 3, pp. 665–678, 2007.

[18] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, Article ID btq003, pp. 680–682, 2010.

[19] 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.

[20] L.-L. Hu, Z. Li, K. Wang et al., "Prediction and analysis of protein methylarginine and methyllysine based on Multisequence features," *Biopolymers*, vol. 96, no. 5, pp. 763–771, 2011.

[21] 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. 6, pp. 8347–8361, 2011.

[22] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[23] C.-W. Tung and S.-Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, vol. 9, article 310, 2008.

[24] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machine," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 1, pp. 1–27, 2011.

[25] H. Mark, F. Eibe, H. Geoffry, P. Bernhard, and R. Peter, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 361–369, 2009.

[26] J. Hu, Y. D. Yang, and D. Kihara, "EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences," *BMC Bioinformatics*, vol. 7, article 342, 2006.

[27] M. Netzer, G. Millonig, M. Osl et al., "A new ensemble-based algorithm for identifying breath gas marker candidates in liver disease using ion molecule reaction mass spectrometry," *Bioinformatics*, vol. 25, no. 7, pp. 941–947, 2009.

[28] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, Article ID btp630, pp. 392–398, 2009.

[29] L. Deng, J. Guan, Q. Dong, and S. Zhou, "Prediction of protein-protein interaction sites using an ensemble method," *BMC Bioinformatics*, vol. 10, article 426, 2009.

[30] W. Zhang, Y. Niu, Y. Xiong, M. Zhang, R. Yu, and J. Liu, "Computational prediction of conformational B-cell epitopes from antigen primary structures by ensemble learning," *PLoS ONE*, vol. 7, no. 8, Article ID e43575, 2012.

[31] K. C. Chou and C. T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 5, pp. 275–349, 1995.

[32] 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.