Mutli-Features Prediction of Protein Translational Modification Sites

Wenzheng Bao, Chang-An Yuan, Youhua Zhang, Kyungsook Han, Asoke K. Nandi, Barry Honig, and De-shuang Huang

Abstract—Post translational modification plays a significiant role in the biological processing. The potential post translational modification is composed of the center sites and the adjacent amino acid residues which are fundamental protein sequence residues. It can be helpful to perform their biological functions and contribute to understanding the molecular mechanisms that are the foundations of protein design and drug design. The existing algorithms of predicting modified sites often have some shortcomings, such as lower stability and accuracy. In this paper, a combination of physical, chemical, statistical, and biological properties of a protein have been ulitized as the features, and a novel framework is proposed to predict a protein's post translational modification sites. The multi-layer neural network and support vector machine are invoked to predict the potential modified sites with the selected features that include the compositions of amino acid residues, the E-H description of protein segments, and several properties from the AAIndex database. Being aware of the possible redundant information, the feature selection is proposed in the propocessing step in this research. The experimental results show that the proposed method has the ability to improve the accuracy in this classification issue.

Index Terms—Post translational modification, protein, classification, prediction

1 Introduction

Post translation modifications (PTMs) are of pivotal importance for understanding protein functionalities in the field of bioinformatics and machine learning [1], [2], [3]. PTMs lie in the crucial functional regions of protein; they maintain the stability of protein-protein interactions and other protein functions [4]. The prediction of post translational modificiation sites in protein sequences is one of the main challenges and research directions in the field of molecular biology [5]. An increasing number of modified information of protein sequences have been found and stored in the various bioinformatics databases. Yet, a large amount of such information seems to be either unavailable or redundant [6]. This gives rise to the extreme difficulty of identifying modified sites directly from the protein sequences [7], [8], [9]. However, protein sequence residues based analysis can enormously help to reveal the formation mechanism of the potential modified sites in the target protein sequences.

According to the latest reseaech, one of the most efficient biological mechanisms for expanding the genetic code and regulating cellular physiology is the PTM in the field of bioinformatics and machine learning [10], [11], [12]. Considering the importance of PTM in basic biological research and drug development, a great deal of efforts has been made with the aim of predicting various modificaton sites.

Recently many researches and large-scale biology ex-periments show that PTM sites of proteins seem to be not evenly distributed over the whole protein sequences. Only a small group of neighbor residues contributes a disproportionately large amount to the potential protein PTM sites of a protein and a ligand [13]. In addition, the modified sites in protein primary sequences can be generated by using the up/down stream residues which are from the protein sequence databases [14]. Therefore, we will work towards identifying the modification sites from different potential target protein fragments based on the physical, chemical or biological characters of amino acid residues. Traditional methods seem to be time consuming, expensive, and less efficient in predicting such sites from large numbers of PTMs data. The development of high quality predictive models and analysis algorithms, which are used by machine learning methods is a challenge and yet constitutes an essential task in the field of bioinformatics and computational biology. Therefore, various computational models have been proposed to predict post translational modification sites by silicon means

- *W.Z. Bao and D.-S. Huang are with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China. E-mail: baowz55555@126.com, dshuang@tongji.edu.cn.*
- *C.-A. Yuan is with the Science Computing and Intelligent Information Processing of GuangXi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning, Guangxi 530001, China. E-mail: yca@gxtc.edu.cn.*
- *Y. Zhang is with the School of Information and Computer, Anhui Agricultural University, Changjiang West Road 130, Hefei, Anhui 230031, China. E-mail: zhangyh@ahau.edu.cn.*
- *K. Han is with the School of Computer Science and Engineering, Inha University, Incheon 22212, South Korea. E-mail: khan@inha.ac.kr.*
- *A.K. Nandi is with the Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, United Kingdom. E-mail: Asoke.Nandi@brunel.ac.uk.*
- *B. Honig is with the Center for Computational Biology and Bioinformatics, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032. E-mail: bh6@columbia.edu.*

Manuscript received 30 Mar. 2017; revised 20 July 2017; accepted 28 Aug. 2017. Date of publication 28 Sept. 2017; date of current version 5 Oct. 2018. (Corresponding author: De-Shuang Huang.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2752703

[15], [16], [17], [18], [19]. Given the importance of the topic as well as the urgency of more powerful high-throughput tools in this area, further efforts are definitely needed to enhance the prediction quality.

Currently, a variety of computational methods based on manchine learning have been developed and explored to identify other protein modification sites with a considerable number of machine learning algorithms, such as Support Vector Machines [20], Random Forests [21], Conditional Random Field [22] and other machine learning algorithms. Aforementioned machine learning methods are usually more appropriate for balanced datasets with supervised learning methods. Unformately, in such imbalanced datasets, the size of positive samples is far smaller than the negative ones. So, one of the serious challenges is that searching and selecting the distinguishing features play important role in the imbalanced classification problems. Considering the experimental errors and other unknowns, some sample labels are wrong or missing. So, several samples seem to be false, which have the ability to contribute to an increased false negative prediction. The preporcessing, which is the step to delete the ambiguous samples from the datasets, seems to be necessary. With such a step, the false positive samples and the true negative samples will be reduced to some degree. With the help of the machine learning technology, the novel supervised inference of novel post modification sites may meet the need of prediction [23], [24], [25], [26], [27].

To deal with the above mentioned issues, the novel classificiation framework based on machine learning methods have been adopted to extract the key features and classification potential modification site effectively and quickly.

Particularly interesting are the segements which are formed by protein potential modified sites in the spatial segements' structures. Our research made contributions to the prediction of protein segements' potential modification spatial sites in protein sequences. Nevertheless, the research of prediction modification sites in protein sequence seems to be a very important and quite a difficult challenge. It is necessary to enhance further the accuracy and the coverage of prediction methods.

In this paper, we propose a novel framework to predict the modified sites at protein segements based on some fundamental features that contribute to physical, chemical, and biological types. The experimental results provide accurately the protein segements about the predicted modification sites in protein sequences.

2 Methods

2.1 Data Set

As is well known, the protein function is contributed by spatial conformation of proteins. Therefore, the protein segement's spatial structure may be helpful to analyze and find out the characterisitics of potential modifciation sites.

The original data set is the benchmark data set in the field of prediction PTM. The first selected dataset was derived from CPLM that is a famous database in the area of protein post translational modifcation [28]. The database, which contains more than 2,500 lysine succinylated sites treated as the positive samples and 24,000 non-succinylated sites treated as the negative samples, has been extracted from 896 protein sequences [29]. All the above mentioned protein segements and polypeptides sequences have been derived from the UniProt, which is the well-known protein database in the field of bioinformatics [30]. It has been utilized in studying and researching enzyme specificity (ES) [31] as well as protein-protein binding sites (PPB) [32], [33], [34] [35], [36], [37].

TABLE 1
First Benchmark Datasets

Dataset	Positive Samples	Negative Samples
CPLM	2,521	24,128
K-PTM	1,169	5,225
Dataset	**Protein**	**Total Samples**
CPLM	896	26,649
K-PTM	521	6,394

The next testing dataset utilized to train and test the framework for predicting the modified sites of multiple K-PTM types in protein sequences that contains 6,394 potential modified sites treated as samples from 27-tuple peptides [38]. The detailed information on this dataset can be found in the following. There are 1,750 samples not belonging to any of the four K-PTM types, 3,895 samples belonging to one type of K-PTM, 740 samples to two PTM types, 9 samples to three PTM types, and none to all the four types [39], [40], [41], [42]. So the detailed information about these two datasets can be found in the Table 1.

The following data set contains various spieces data about post translation modification. The data set on lysine acetylation site for three species, include Homo sapiens, Mus musculus and Saccharomyces cerevisiae from several sources including PhosphoSite, UniProtKB/Swiss-Prot, UbiProt and SCUD, which are the well-known databases in the field proteomics. Because of exceptions, ubiquitin seems to be attached to lysine residues of proteins in the degree. So, we merely considered lysine ubiquitylation in the above mention three species in the work. The raw dataset includeed 11,547 protein sequences covering different species; of these sequences, more than 8,000 are from H.sapiens, about 3,300 are from M.musculus and more than 4,500 are from S.cerevisiae. After removing the redundant protein segements of three kinds of samples, we have extracted and get several samples of three species, which include 6,323 samples of H. sapiens, 2,342 samples of M.musculus and 7,863 samples of S.cerevisiaes, respectively. Afterwards, 20 proteins haven been randomly selected from each of the datasets of three species to form the independent test sets, and the remaining 6,303, 2,322, and 7,843 <please check these numbers> proteins were used to construct the training set, respectively.

The original data set is the benchmark data set in the field of prediction PTM. The first selecting dataset hae been ulitilized in this research was derived from CPLM, which is a famous and world-renowned lysine modification database in the field of protein post translational modifcation [43]. The database, which contains more than 2,500 lysine succinylated sites treated as the positive samples and 24,000 non-succinylated sites treated as the negative samples, has been extracted from 896 protein sequences [44]. All the above mentioned protein segements and polypeptides sequences have been derived from the UniProt, which is the well-known protein

TABLE 2
Second Benchmark Datasets

Dataset	*Positive Samples*	*Negative Samples*	*All Samples*
H.sapiens	14,078	14,078	20,144
M.musculus	2,622	2,622	5,244
S.cerevisiaes	5,242	5,242	10,484

database in the field of bioinformatics [45]. It has been utilized in studying and researching enzyme specificity (ES) [46], signal peptide/ amino acid residues' cleavage sites (AACS) [47], hydroxyproline and hydroxylysine sites (H2S) [19] methylation sites [48], nitrotyrosine sites (NiS) [49], protein-protein interaction (PPI) [50], and protein-protein binding sites (PPB) [51], [52], [53], [54].

2.2 Feature Description

Generally speaking, the kind of protein features could reach more than 40,000. Those various types of features, including amino acid compositions model (AAC) pseudo amino acid compositions model (PseAAC) and other related information of protein characteristics [55], [56], [57]. Those features, however, could hardly meet the need of effectively and accurately description of the interactions among predicted modified site and neighbor amino acid residues. Therefore, a typical and special feature, which has the ability to describe the segement of protein peptide, has been introduced in this work.

First of all, when it comes to the amino acid residues' composition, a great many of researchers could not help taking advantage of the statistical information of protein sequences in the field of bioinformatics and computational biology. Those features merely described the potential modified segements in the statistical aspect. Of course, the selection of key feature may be treated as a difficult task in this kind of feature sets.

It was found that 20 kinds of amino acid residues have the tendency to be gourped in the 3 types of special structure elements: Helix, Strand and Coil. Such features are selected from PSIPRED (version 2.6) [58]. PSIPRED's developers try to predict the special tendency with the method of neural network technology in the protein sequence [59].

Considering the distributions of α-helices and β-strands effectively, we have denoted the predicted protein segements by E-H sequence description. The next table contains several features by the E-H's description.

From the above mentioned features both the basic feature and the novel feature may describe the statistical information of E and H type that describe the predicted modified segements. According to the Ding's work [40], all of the above mentioned features contain some redundant information and noise. So, the selected features are shown in the following Table 4.

Fig. 1. Species of second benchmark datasets.

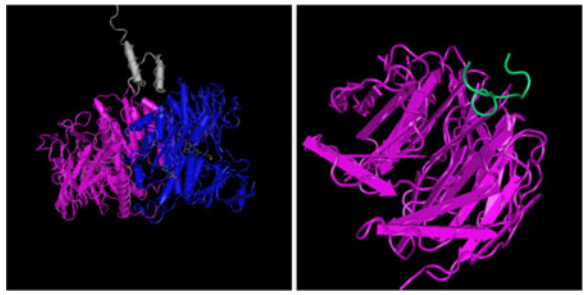


Fig. 2. Positive samples of protein structure.

The most popular and well-known amino acids' feature index is the AAindex, which is a website database of numerical indices including various biological, physical and chemical properties of the amino acid residues and other forms of protein sequences' features. Meanwhile, AAindex contains three types protein properties information: AAindex1, AAindex2 and AAindex3 [60], [61], [62]. So, several types of amino acids' features have been employed in this research. The more detailed information have been shown in Table 2. The selected properties of AAindex database is shown in Table 7 [63], [64], [65].

2.3 Classification

Classification is very important and is often used in the field of bioinformatics [66]. Due to post modification sites consisting of potential modified residues, the up/down stream amino acid residues should be treated as a feature vector. In this paper, a feature-based classification method is proposed to detect the modified residues in protein segements.

In this paper, support vector machine (SVM) classification model is created to identify the post modification residues in the field of proteomics. Currently, some good use of SVM has been made for bioinformatics and computational biology. Such model has been introduced and proposed by Vapnik for classification and regression, which are a set of related supervised learning methods in the field of machine learning. It is a well-known classifier used to validate the application of the successful classification phosphorylation sites [67], [68], [69], [70].

Developments of artificial intelligence and neural networks can be traced back to the 1950s. Currently, the deep learning, which is more complex with deeper structures, seems to be at the forefront of the current topics in the field of machine learning. It was pointed that neural network can be widely used in various fields [71], [72], [73]. Flexible neural network tree has been introduced and designed by Chen

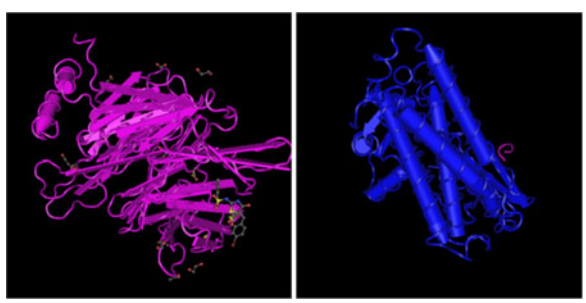


Fig. 3. Negative samples of protein structure.

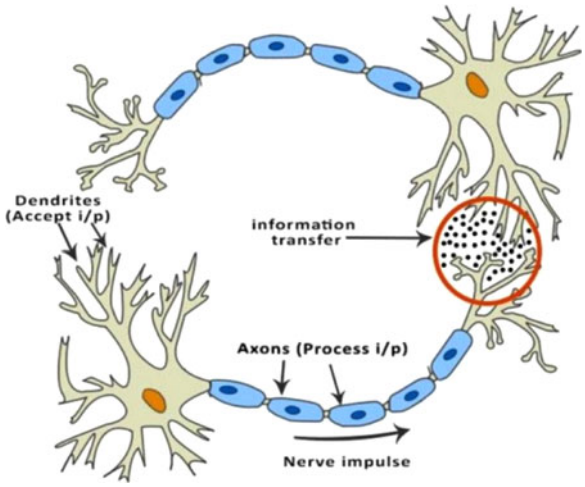


Fig. 4. The structure of nerve cells.

[74], [75], [76]. It is noted that the flexible structure could be regarded as the prototype of deep neural network. So the main steps of such neural network model are shown in Fig. 4.

In this paper, SVM and multi-layer neural networks have been adopted to learn from the training set, which can classify the potential modified segement in the protein sequence.

In order to make sure that the parameter evaluation of support vector machines is thoroughly independent of the data set, the original data set of potential modified sites has been grouped into two sets. One part is used to optimize the parameters of SVM and multi-layer neural networks as a separate validation set which includes one tenth of the whole protein segements. When classifying sites as potential center amino acid residues and non-modified amino acid residues, the ensemble model is trained by all positive labels with the modified sites and all negative labels with the non-modified sites.

In this paper, the ten-fold cross validation method has been utilized to validate this classification framework. The sample set is divided into ten parts, 90% of which is treated as the training set and the remaining 10% subset has been regarded as the test set. Afterwards, the prediction results from the two classifiers are integrated as input vectors of the classifier model and the prediction result of the whole model is the finial result.

2.4 Performance Measures

To evaluate the performance of modified sites prediction, the several measures are used. The true positive of the modified prediction is the number of modified in predicted modified and also in natural modified sites. The false positive of the modified prediction is the number of modified in predicted modified but not in natural modified sites. The false negative of the modified prediction is the number of modified that are not in predicted modified but in natural modified sites.

To evaluate the predicition model, the Root Mean Square (RMS) [45], which has been used as evaluation function of features, has been employed in this reseach. The overall accuracy (OA) means the computing for each dataset in the prediction model. At the same time, the next several performances have been utilized in evaluating the prediction accuracy, namely, Sensitivity (Sens) and Specificity (Spec). Explicitly, they are described by the formulation (1)-(3):

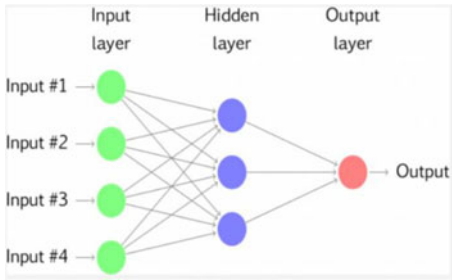


Fig. 5. The structure of neural network.

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \quad OA \in [0, 1] \tag{1}$$

$$Sens = \frac{TP}{TP + FN} \quad Sens \in [0, 1] \tag{2}$$

$$Spec = \frac{TN}{FP + TN} \quad Spec \in [0, 1]. \tag{3}$$

3 Results

In order to further assess performance of this model, comparison has been carried out for our proposed ensemble classification model and other existing means. The proposed UbiProber predictor trained and tested several data sets, which include of H.sapiens, M.musculus and S.cerevisiae, based on the combined features and proposed ensemble prediction model. To evaluate the performance of UbiProber for species-specific modification sites prediction, the 10-fold cross-validation test has been performed in each species.

First of all, several acetylation prediction softwares have been developed in the website resources. However, some of them had broken internet links, so they could hardly be tested in this model. In fact the predictors, which employed EnsemblePail, PHOSIDA, PLMLA and PSKAcePred, were included in the comparison tables. The comparison results are shown in Tables 6, 7, 8, 9, and 10. In terms of sensitivity and specificity, the proposed method achieved relatively high performance compared to the other compared methods. On the contrary, there was a great divergence between sensitivity and specificity in the data sets of PHOSIDA, PLMLA and PSKAcePred. When it comes to the predicition accuracy, the value from the proposed method could almost reach ideal values, which overwhelmed all other methods. Compared to state-of-the-art methods, it is worth pointing out that the proposed method demonstrates a fairly good capability to predict modification sites.

4 Conclusion

In this paper, we propose the machine learning algorithm, with the features of amino acid residues, to predict the potential modified sites. First of all, the machine learning method is carried out to delete the redunant potential samples. Subsequently, SVM and multi-layer neural network models are created to predict the modified sites and non-modified sites based on the features selected. Finally, the potential modified

TABLE 3
The E-H Features

No.	Description
1	Ratio_EC
2	Ratio_HC
3	Appearance_Seg_H
4	Appearance _Seg_E
5	Appearance H/L_segment
6	Appearance E/L_segment
7	Var_seg_H/L_segment
8	Var_seg_E/L_segment
9	Com_Moment_segment_EH
10	Com_Moment_segment_HE
11	Com_Moment_EH
12	Com_Moment_HE
13	Var_Pos_E_segment
14	Var_Pos_H-segment
15	Var_Pos_E
16	Var_Pos_H
17	f_{EH}
18	f_{HE}
19	Appearance_Seg_E
20	Appearance_Seg_ H
21	LZ_seq
22	LZ_E&H
23	Ave _E&H

TABLE 4
The Selected E-H Features

No.	Description
1	Appearance(H)
2	Appearance (H)
3	Max_segment_H/L_EH
4	Max_segment_E/L_EH
5	Avg_segment_H
6	Avg_segment_E
7	Com_Moment_E
8	Count_Segment_E
9	Ave_In_E
10	Ave_In_HE
11	Ave_In_EH

TABEL 5
The Selected AAIndex Properties

No.	AAindex ID	No.	AAindex ID
1	CHOP780207	9	KLEP840101
2	DAYM780201	10	KRIW710101
3	EISD860102	11	KRIW790102
4	FAUJ880108	12	NAKH920103
5	FAUJ880111	13	QIAN880101
6	FINA910103	14	QIAN880139
7	JANJ780101	15	RACS820114
8	KARP850103		

residues are clustered to different modified types, which represent different sets of modified sites where different sets are dissimilar from each other. Our method chooses the similarity as a measure of local neighbor residuues discovery. One of the future researches seems to consider the modified residues conservations and different energy contributions to each other, which are still very necessary and important.

TABLE 6
Comparisons of Performances on Proposed and Existing Methods (H.sapiens)

Method	*Sn (%)*	*Sp (%)*	*Acc (%)*
SVM	75.33	82.67	72.07
NN	72.33	79.33	70.33
FNT	78.92	84.29	72.64
Previous Method	80.21	87.74	74.59
Proposed Method	85.94	61.37	79.53

TABLE 7
Comparisons of Performances on Proposed and Existing Methods (M.musculus)

Method	*Sn (%)*	*Sp (%)*	*Acc (%)*
SVM	75.47	83.67	72.38
NN	72.74	75.84	71.75
FNT	79.12	85.29	73.71
Previous Method	81.21	86.74	78.65
Proposed Method	83.23	79.74	81.49

TABLE 8
Comparisons of Performances on Proposed and Existing Methods (S.cerevisiae)

Method	*Sn (%)*	*Sp (%)*	*Acc (%)*
SVM	83.00	72.34	77.67
NN	77.60	69.54	73.57
FNT	81.57	71.89	76.73
Previous Method	80.38	68.94	74.66
Proposed Method	86.31	69.87	78.09

TABLE 9
Comparisons of Performances on Proposed and Existing Methods (CPLM)

Method	*Sn (%)*	*Sp (%)*	*Acc (%)*
SVM	79.51	72.37	77.76
NN	74.74	69.87	73.87
FNT	79.12	72.74	77.82
Previous Method	79.37	81.52	79.42
The Method	79.93	82.87	80.92

TABLE 10
Comparisons of Performances on Proposed and Existing Methods (K-PTM)

MethodS	*Sn (%)*	*Sp (%)*	*Acc (%)*
EnsemblePail	49.36	62.68	56.04
PHOSIDA	42.37	92.35	67.36
PLMLA	78.90	44.20	61.55
PSKAcePred	72.24	49.66	60.95
LA+FNT	61.38	75.40	68.39
Previous Method	73.47	74.35	73.91
Proposed Method	74.21	74.87	74.54

From the above analysis and discussion, it can be concluded that features of amino acid residues, especially the neighbor residues of the center potential modification sites, appear to play a critical role in this prediction issues. Therefore, the assumption of relationship between the upstream/downstream resideus and the center modified sites could be

regarded as the combination feature type of amino acid residues' interaction. At the same time, the multi-layer neural network and support vector machine essemble model have integrated both complex features combination and the kernel technology in this prediction issue. The modified sites' prediction seems to be a classical two-classification issue in the field of machine learning and bioinforamtics. Nevertheless, several challenges still have to be resolved in such field. So, in the future work, various types of protein post translational modification in the PTM process needs to be more clearly explained or described in detail in the field of biology. Finally, the special structure information seems to be employed as the novel types of prediction features.

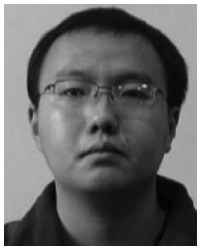
Acknowledgments

This work was supported by the grants of the National Science Foundation of China, Nos. 61732012, 61520106006, 31571364, U1611265, 61532008, 61672382, 61772370, 61402334, 61472282, and 61472173, and China Postdoctoral Science Foundation [Grant No. 2015M580352, 2017M611619, and 2016M601646.

References

[1] S. Zhao, et al., "Regulation of cellular metabolism by protein lysine acetylation," *Sci.*, vol. 327, no. 5968, pp. 1000–1004, 2010.
[2] A. Cheng, et al., "MoMo: Discovery of post-translational modification motifs," *Bioinf.*, vol. 35, no. 1, pp. 121–124, 2017
[3] A. S. Khan, et al., "High-throughput screening of a GlaxoSmithKline protein kinase inhibitor set identifies an inhibitor of human cytomegalovirus replication that prevents CREB and histone H3 post-translational modification," *J. Gen. Virology*, vol. 98, no. 4, pp. 754–768, 2017.
[4] J. Yang, et al., "Post-translational modification of the membrane type 1 matrix metalloproteinase (MT1-MMP) cytoplasmic tail impacts ovarian cancer multicellular aggregate dynamics," *J. Biol. Chem.*, vol. 284, no. 30, pp. 19791–19799, 2009.
[5] S.-P. Deng, L. Zhu, and D. S. Huang, "Predicting hub genes associated with cervical cancer through gene co-expression networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 1, pp. 27–35, Jan./Feb. 2016.
[6] L. Zhu, W. Guo, S.-P. Deng, and D. S. Huang, "ChIP-PIT: Enhancing the analysis of ChIP-Seq data using convex-relaxed pair-wise interaction tensor decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 1, pp. 55–63, Jan./Feb. 2016.
[7] B. Schwer, et al., Calorie restriction alters mitochondrial protein acetylation," *Aging Cell*, vol. 8, no. 5, pp. 604–606, 2009.
[8] G. J. Roth, N. Stanford, and P.W. Majerus, "Acetylation of prostaglandin synthase by aspirin," *Proc. Nat. Acad. Sci.*, vol. 72, no. 8, pp. 3073–3076, 1975.
[9] A. M. Bode and Z. Dong, "Post-translational modification of p53 in tumorigenesis," *Nature Rev. Cancer*, vol. 4, no. 10, pp. 793–805, 2004.
[10] V. Quivy and C. Van Lint, "Regulation at multiple levels of NF-κB-mediated transactivation by protein acetylation," *Biochem. Pharmacology*, vol. 68, no. 6, pp. 1221–1229, 2004.
[11] L. Zhu, D.-S. Ping, and D. S. Huang, "A two stage geometric method for pruning unreliable links in protein-protein networks," *IEEE Trans. NanoBioscience*, vol. 14, no. 5, pp. 528–534, Jul. 2015.
[12] S.-P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *BMC Genomics*, 2015, vol. 16, no. Suppl 3, Art. no. S4.
[13] K. L. Guan and Y. Xiong, "Regulation of intermediary metabolism by protein acetylation," *Trends Biochem. Sci.*, vol. 36, no. 2, pp. 108–116, 2011.
[14] C. Caron, C. Boyault, and S. Khochbin, "Regulatory cross-talk between lysine acetylation and ubiquitination: Role in the control of protein stability," *Bioessays*, vol. 27, no. 4, pp. 408–415, 2005.
[15] A. Peserico and C. Simone, "Physical and functional HAT/HDAC interplay regulates protein acetylation balance," *BioMed. Res. Int.*, p. 371832, 2011.
[16] S.-P. Deng and D. S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods*, vol. 69, no. 3, pp. 207–212, 2014.
[17] C. Watala, et al., "Increased protein glycation in diabetes mellitus is associated with decreased aspirin-mediated protein acetylation and reduced sensitivity of blood platelets to aspirin," *J. Mol. Med.*, vol. 83, no. 2, pp. 148–158, 2005.
[18] L. I. Hu, B. P. Lima, and A. J. Wolfe, "Bacterial protein acetylation: The dawning of a new age," *Mol. Microbiol.*, vol. 77, no. 1, pp. 15–21, 2010.
[19] J. Baeza, M. J. Smallegan, and J. M. Denu, "Mechanisms and dynamics of protein acetylation in mitochondria," *Trends Biochem. Sci.*, vol. 41, no. 3, pp. 231–244, 2016.
[20] M. Tsuda, S. Kinugawa, S. Takada, et al., "Mitochondrial protein acetylation in skeletal muscle was associated with exercise intolerance in heart failure after myocardial infarction in mice," vol. 35, no. 2, pp. 541–552, 2015.
[21] K. Minton, "Gene expression: Reading protein acetylation, *Nature Rev. Mol. Cell Biol.*, vol. 17, pp. 676–677, 2016.
[22] N. Chatterjee, M. Tian, K. Spirohn, M. Boutros, and D. Bohmann, "Keap1-independent regulation of Nrf2 activity by protein acetylation and a BET bromodomain protein," *PLoS Genetics*, vol. 12, no. 5, 2016 Art. no. e1006072.
[23] L. Galdieri, H. Gatla, I. Vancurova, and A. Vancura, "Activation of AMP-activated protein kinase by metformin induces protein acetylation in prostate and ovarian cancer cells," *J. Biol. Chem.*, vol. 291, no. 48, pp. 25154–25166, 2016.
[24] Y. Shindo, H. Komatsu, K. Hotta, K. Ariga, and K. Oka, "An artificial reaction promoter modulates mitochondrial functions via chemically promoting protein acetylation," *Sci. Rep.*, vol. 6, 2016, Art. no. 29224.
[25] Z. Barjaktarovic, et al., "Low-dose radiation differentially regulates protein acetylation and histone deacetylase expression in human coronary artery endothelial cells," *Int. J. Radiat. Biol.*, vol. 93, no. 2, 2016, pp. 156–164.
[26] S. K. Bharti and R. M. Brosh Jr., "Fine-tuning DNA repair by protein acetylation," *Cell Cycle*, vol. 15, no. 15, pp. 1952–1953, 2016.
[27] D. Kuczyńska-Wiśnik, M. Moruno-Algara, K. Stojowska-Swędrzyńska, and E. Laskowska, "The effect of protein acetylation on the formation and processing of inclusion bodies and endogenous protein aggregates in Escherichia coli cells," *Microbial Cell Factories*, vol. 15, no. 1, 2016, Art. no. 189.
[28] M. Xie and J. A. Hill, "Cardiac autophagy and its regula-tion by reversible protein acetylation," in *Epigenetics in Cardiac Disease*. Cham, Switzerland: Springer Int. Publishing, 2016, pp. 231–262.
[29] J. Baeza, M. J. Smallegan, and J. M. Denu, "Mechanisms and dynamics of protein acetylation in mitochondria," *Trends Biochem. Sci.*, vol. 41, no. 3, pp. 231–244, 2016.
[30] Y. Li, et al., "Accurate in silico iden-tification of species-specific acetylation sites by integrat-ing protein sequence-derived and functional features," *Sci. Rep.*, vol. 4, 2014, Art. no. 5765.
[31] T. Hou, et al., "LAceP: Lysine acety-lation site prediction using logistic regression classifiers," *PloS One*, vol. 9, no. 2, 2014 Art. no. e89575.
[32] E. Verdin and M. Ott, "50 years of protein acetylation: From gene regulation to epigenetics, metabolism and beyond," *Nature Rev. Mol. Cell Biol.*, vol. 16, no. 4, pp. 258–264, 2015.
[33] B. T. Weinert, S. Satpathy, B. K. Hansen, D. Lyon, L. J. Jensen, and C. Choudhary, "Accurate quantification of site-specific acetylation stoichiometry reveals the impact of sirtuin deacetylase CobB on the E. coli acetylome," *Mol. Cell. Proteomics*, vol. 16, pp. 759–769, 2017.
[34] J. Baeza, M. J. Smallegan, and J. M. Denu, "Site-specific reac-tivity of nonenzymatic lysine acetylation," *ACS Chem. Biol.*, vol. 10, no. 1, pp. 122–128, 2015.
[35] V. Pejaver, W. L. Hsu, F. Xin, A. K. Dunker, V. N. Uversky, and P. Radivojac, "The structural and functional signatures of proteins that undergo multiple events of post-translational modification," *Protein Sci.*, vol. 23, no. 8, pp. 1077–1093, 2014.
[36] K. Masui, et al., "Glucose-dependent acetylation of Rictor promotes targeted cancer therapy resistance," *Proc. Nat. Acad. Sci.*, vol. 112, no. 30, pp. 9406–9411, 2015.
[37] A. Kurotani, A. A. Tokmakov, Y. Kuroda, Y. Fukami, K. Shinozaki, and T. Sakurai, "Correlations between predicted protein disorder and post-translational modifications in plants," *Bioinf.*, vol. 30, no. 8, pp. 1095–1103, 2014.
[38] Y. Xu and K. C. Chou, "Recent progress in predicting post-translational modification sites in proteins," *Current Top. Medicinal Chem.*, vol. 16, no. 6, pp. 591–603, 2016.

[39] P. E. Wright and H. J. Dyson, "Intrinsically disordered pro-teins in cellular signalling and regulation," *Nature Rev. Mol. Cell Biol.*, vol. 16, no. 1, pp. 18–29, 2015.
[40] L. Mitchell, et al., "mChIP-KAT-MS, a method to map protein interactions and acetylation sites for lysine acetyltransferases," in *Proc. Nat. Acad. Sci.*, vol. 110, no. 17, pp. E1641–E1650, 2013.
[41] J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensem-ble random forest approach," *J. Theor. Biol.*, vol. 394, pp. 223–230, 2016.
[42] S. A. Cobbold, J. M. Santos, A. Ochoa, et al., "Proteome-wide analysis reveals widespread lysine acetylation of major protein complexes in the malaria parasite," *Sci. Rep.*, vol. 6, pp. 19722–19736, 2016.
[43] Y. Dou, B. Yao, and C. Zhang, "PhosphoSVM: Prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino Acids*, vol. 46, no. 6, pp. 1459–1469, 2014.
[44] P. Minguez, I. Letunic, L. Parca, and P. Bork, "PTMcode: A database of known and predicted functional associations between post-translational modifications in proteins," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D306–D311, 2013.
[45] W. R. Qiu, X. Xiao, W. Z. Lin, and K. C. Chou, "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," *J. Biomol. Struct. Dyn.*, vol. 33, no. 8, pp. 1731–1742, 2015.
[46] Y. Xiong, X. Peng, Z. Cheng, W. Liu, and G. L. Wang, "A comprehensive catalog of the lysine-acetylation targets in rice (Oryza sativa) based on proteomic analyses," *J. Proteomics*, vol. 138, pp. 20–29, 2016.
[47] C. Jia, X. Lin, and Z. Wang, "Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition," *Int. J. Mol. Sci.*, vol. 15, no. 6, pp. 10410–10423, 2014.
[48] O. Pougovkina, et al., "Mitochondrial protein acetylation is driven by acetyl-CoA from fatty acid oxidation," *Human Mol. Genetics*, vol. 23, pp. 3513–3522, 2014.
[49] D. S. Huang, "Systematic theory of neural networks for pattern recognition (in Chinese)," Publishing House of Electronic Industry of China, May 1996.
[50] D. S. Huang, "Radial basis probabilistic neural networks: Model and application," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 13, no. 7, pp. 1083–1101, 1999.
[51] H.-J. Yu and D. S. Huang, "Graphical representation for DNA sequences via joint diagonalization of matrix pencil," *IEEE J. Biomed. Health Inf.*, vol. 17, no. 3, pp. 503–511, May 2013
[52] I. N. Berezovsky, G. T. Kilosanidze, V. G. Tumanyan, and L. L. Kisselev, "Amino acid composition of protein termini are biased in different manners," *Protein Eng.*, vol. 12, no. 1, pp. 23–30, 1999.
[53] A. Andreeva, et al., "Data growth and its impact on the SCOP database: New development," *Nucleic Acids Res.*, vol. 36, pp. D419–D425, 2007.
[54] D. S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Curr. Protein Peptide Sci.*, vol. 15, no. 6, pp. 553–560, 2014.
[55] D. S. Huang and H.-J. Yu, "Normalized feature vectors: A novel alignment-free sequence comparison method based on the numbers of adjacent amino acids," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 2, pp. 457–467, Mar./Apr. 2013
[56] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinf.*, vol. 17, no. 4, pp. 349–358, 2001.
[57] K. Chen, L. A. Kurgan, and J. S. Ruan, "Prediction of protein structural class using novel evolutionary collocation-based sequence representation," *J. Comput. Chem.*, vol. 29, pp. 1596–1604, 2008.
[58] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, pp. 195–202, 1999.
[59] S. F. Altschul, et al., "Gapped BLAST and PSI-BLAST: A new generation of protein database search pro-grams," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.
[60] L. A. Kurgan, T. Zhang, H. Zhang, S. Shen, and J. Ruan, "Secondary structure-based assignment of the protein structural classes," *Amino Acids*, vol. 35, pp. 551–564, 2008.
[61] L. Kurgan, K. Cios, and K. Chen, "SCPRED: accurate pre-diction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinf.*, vol. 9, 2008, Art. no. 226.
[62] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," *J. Theor. Biol.*, vol. 267, pp. 272–275, 2010.
[63] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inf. Theory*, vol. IH-22, no. 1, pp. 75–81, Jan. 1976.
[64] S. Ding, S. Zhang, Y. Li, and T. Wang, "A novel protein structural classes prediction method based on predicted secondary structure.," *Biochimie*, vol. 94, pp. 1166–1171, 2012
[65] D. S. Huang and C.-H. Zheng, "Independent component analysis based penalized discriminant method for tumor classification using gene expression data," *Bioinf.*, vol. 22, no. 15, pp. 1855–1862, 2006.
[66] C.-H. Zheng, D.S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using non-negative matrix factorization with gene selection," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 4, pp. 599–607, Jul. 2009.
[67] L. Zhu, Z.-H. You, D. S. Huang, and B. Wang, "t-LSE: A novel robust geometric approach for modeling protein-protein interaction networks," *Plos One*, vol. 8, no. 4, 2013, Art. no. e58368, doi:10.1371/journal.pone.0058368.
[68] H. B. Rao, F. Zhu, G. B. Yang, Z. R. Li, and Y. Z. Chen, "Update of PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Res.*, vol. 39, pp. W385–W390, 2011.
[69] C.-H. Zheng, L. Zhang, V. T.-Y. Ng, S. C.-K. Shiu, and D. S. Huang, "Molecular pattern discovery based on penalized matrix decomposition," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 6, pp. 1592–1603, Nov./Dec. 2011.
[70] D. S. Huang and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Trans. Neural Netw.*, vol. 19, no. 12, pp. 2099–2115, Dec. 2008.
[71] B. Yang, Y. H. Chen, and M. Y. Jiang, "Reverse engineering of gene regulatory networks using flexible neural tree models, *Neurocomput.*, vol. 99, pp. 458–466, 2013.
[72] Y. Chen, B. Yang, and J. Dong, "Evolving flexible neural networks using ant programming and PSO algorithm," in *Advances in Neural Networks*. Berlin, Germany: Springer, 2004, pp. 211–216.
[73] D. S. Huang and W. Jiang, "A general CPL-AdS methodology for fixing dynamic parameters in dual environments," *IEEE Trans. Syst., Man Cybern. - Part B*, vol. 42, no. 5, pp. 1489–1500, Oct. 2012
[74] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Anal. Biochem.*, vol. 370, pp. 1–16, 2007.
[75] H. B. Shen and K. C. Chou, "Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition," *Biochem. Biophys. Res. Commun.*, vol .337, pp. 752–756, 2005.
[76] Y. S. Ding, T. L. Zhang, and K. C. Chou, "Prediction of pro-tein structure classes with pseudo amino acid composi-tion and fuzzy support vector machine network," *Protein Pept. Lett.*, vol. 14, pp. 811–815, 2007.



Wenzheng Bao received the master's degree in computer science from the University of Jinan, in 2015. He is currently working toward the PhD degree in computer science in the the School of Electronic and Information Engineering, Tongji University, Shanghai, China. His research interests include bioinformatics and machine learning.

Chang-An Yuan received the PhD degree in computer application technology from Sichuan University, China, in 2006. He is a professor with Guangxi Teachers Education University. His research interests include computational intelligence and data mining

Youhua Zhang received the PhD degree from the University of Science and Technology of China. He is a professor, the master's tutor, and the dean of the School of Information and Computer, Anhui Agricultural University. He is an executive council member of the Computer Applications Branch of the China Agricultural Society, an executive council member of the Anhui Association of Agricultural Information, a member of the National Bee modern Industrial Technology and Product Quality and Safety Standards System, an Expert Advisory Committee of the Anhui Provincial Human Resources and Social Security Department of Information Technology, and an Executive Committee member of the Hefei branch of the CCF. He is mainly engaged in teaching and research in computer applications and logistics engineering. He is in charge or participates in Provincial Science and Technology Research, the National Science and Technology Support Program, 863 project, and the National Natural Science Foundation. He received the First Prize of Scientific and Technological Progress in Anhui Province.

Kyungsook Han received the BS degree cum laude from Seoul National University, in 1983, the MS degree cum laude in computer science from KAIST, in 1985, the MS degree in computer science from the University of Minnesota at Minneapolis, in 1989, and the PhD degree in computer science from Rutgers University, in 1994. She is a professor in the Department of Computer Science and Engineering, Inha University, Korea. Her research interests include bioinformatics, visualization, and data mining.

Asoke K. Nandi received the PhD degree in physics from the University of Cambridge (Trinity College), Cambridge, United Kingdom. He held academic positions at several universities, including Oxford (United Kingdom), Imperial College London (United Kingdom), Strathclyde (United Kingdom), and Liverpool (United Kingdom) as well as a Finland distinguished professorship in Jyvaskyla (Finland). In 2013, he moved to Brunel University London (United Kingdom), to become the chair and head of electronic and computer engineering. He is a distinguished visiting professor at Tongji University (China) and an adjunct professor at the University of Calgary (Canada). His current research interests include areas of signal processing and machine learning, with applications to communications, gene expression data, functional magnetic resonance data, and biomedical data. He has authored more than 550 technical publications, including 220 journal papers as well as four books, entitled *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Automatic Modulation Recognition of Communications Signals* (Springer, 1996), and *Blind Estimation Using Higher-Order Statistics* (Springer, 1999). Recently, he published in *Blood*, *BMC Bioinformatics*, *IEEE TWC*, *NeuroImage*, *PLOS ONE*, *Royal Society Interface*, and *Signal Processing*. The h-index of his publications is 67 (Google Scholar) and his ERDOS number is 2. He is a fellow of the Royal Academy of Engineering and a fellow of seven other institutions including the IEEE and the IET. Among the many awards he received are the Institute of Electrical and Electronics Engineers Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (United Kingdom) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (United Kingdom), in 1998.

Barry Honig is a tenured professor with Columbia University, the US Academy of Sciences. He is the director of the Biology and Bioinformatics Computing Center of Columbia University, and an editorial board member of *PNAS*, *Journal of Molecular Biology*, *Structure*, *Biochemistry*, and *Current Opinion in Structural Biology*. His research interests include computational biology and bioinformatics. As of 2013, he has published more than 300 papers in international high-level SCI journals. His papers have been SCI cited more than 15,600 times. He has published 17 papers in *Nature*, three in *Science*, three in *Cell*, and 22 in *PNAS*. In 2004, he was elected to the National Academy of Sciences. In 2007, he received the US National Academy of Sciences Alexander Hollaender Award.

De-Shuang Huang received the BSc, MSc, and PhD degrees all in electronic engineering from the Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China and XidianUniversity, Xian, China, in 1986, 1989, and 1993, respectively. From 1993 to 1997, he was a postdoctoral research fellow, respectively, in the Beijing Institute of Technology and in the National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In September 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the recipient of "Hundred Talents Program of CAS". In September 2011, he entered into Tongji University as a chaired professor. From September 2000 to March 2001, he worked as a research associate at Hong Kong Polytechnic University. From August to September 2003, he visited George Washington University in Washington DC as a visiting professor. From July to December 2004, he worked as the University fellow at Hong Kong Baptist University. From March, 2005 to March, 2006, he worked as a research fellow at the Chinese University of Hong Kong. From March to July, 2006, he worked as a visiting professor at Queen's University of Belfast, United Kingdom. In 2007, 2008, and 2009, he worked as a visiting professor at Inha University, Korea, respectively. At present, he is the director of the Institute of Machines Learning and Systems Biology, Tongji University. His current research interest includes bioinformatics, pattern recognition, and machine learning. He has published more than 180 journal papers. Also, in 1996, he published a book entitled *Systematic Theory of Neural Networks for Pattern Recognition* (in Chinese), which received the Second-Class Prize of the 8th Excellent High Technology Books of China, and in 2001 & 2009 two books entitled *Intelligent Signal Processing Technique for High Resolution Radars* (in Chinese) and *The Study of Data Mining Methods for Gene Expression Profiles* (in Chinese), respectively. He is currently fellow of the International Association of Pattern Recognition (IAPR fellow), and a senior member of the IEEE and International Neural Networks Society.