This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Mutli-features Predction of Protein Translational Modification Sites

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

Abstract— The 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 [1-3]. 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]. 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 for regulating cellular physiology is the PTM in the field of bioinformatics and machine learning [1, 2]. Considering the importance of PTM in basic biological research and drug development, a great deal of efforts have been made with the aim of predicting various modificaton sites.

Recently many researches and large-scale biology experiments 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 [7]. 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 [8]. 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 [9-10]. 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 [7], Random Forests [8], Conditional Ran-

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 (Corresponding author).

Chang-An Yua is 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

K. Han, School of Computer Science and Engineering Inha University Incheon South Korea. E-mail: khan@inha.ac.kr

Y.H. Zhang is with School of Information and Computer, Anhui Agricultural University, Changjiang West Road 130, Hefei, Anhui, China. E-mail:zhangyh@ahau.edu.cn

A.K. Nandi is with Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, UB8 3PH, United Kingdom. Email: Asoke.Nandi@brunel.ac.uk

B. Honig is with Center for Computational Biology and Bioinformatics, 1130 St. Nicholas Avenue, Room 815, New York, NY, 10032, USA. Email: bh6@columbia.edu.

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

dom Field [9] 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 [11-12].

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 [14]. 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 [15]. 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 [16]. It has been utilized in studying and researching enzyme specificity (ES) [17] as well as protein-protein binding sites (PPB) [23-24].

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 [25]. 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. So the detailed information about these two datasets can be found in the Table 1.

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 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.

TABLE 2 SECOND BENCHMARK DATASETS

Dataset	*Positive Samples*	*Negative Samples*	*All Samples*
H.sapiens	14078	14078	20144
M.musculus	2622	2622	5244
S.cerevisiaes	5242	5242	10484

Fig. 1.1. Species of Second Benchmark Datasets

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

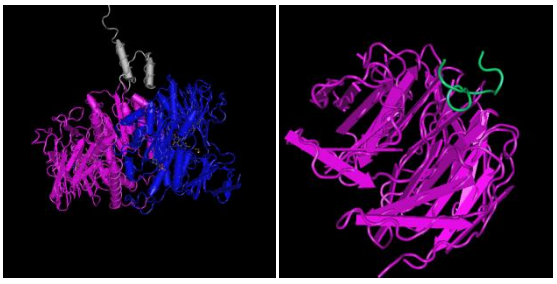

Fig. 1.2 Positive Samples of Protein Structure

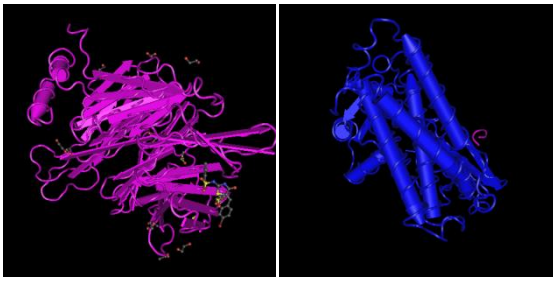

Fig. 1.3 Negative Samples of Protein Structure

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 [14]. 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 [15]. 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 [16]. It has been utilized in studying and researching enzyme specificity (ES) [17], signal peptide/ amino acid residues' cleavage sites (AACS) [18], hydroxyproline and hydroxylysine sites (H2S) [19] methylation sites [20], nitrotyrosine sites (NiS) [21], protein-protein interaction (PPI) [22], and protein-protein binding sites (PPB) [23-24].

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 [26]. 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) [34]. PSIPRED's developers try to predict the special tendency with the method of neural network technology in the protein sequence [36].

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.

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*

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.

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*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

11 *Ave_In_EH*

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 [27-29]. 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 [47-49].

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		

2.3 Classification

Classification is very important and is often used in the field of bioinformatics [25]. 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 [26].

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 [41-42]. Flexible neural network tree has been introduced and designed by Chen [reference]. The model constributes to an alternative neural network structure, in which both depth and width can be employed in the model [43-44]. 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 is shown in Fig. 2.

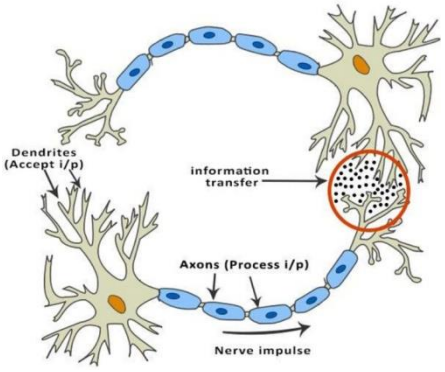

Fig. 2.1. The Structure of Nerve Cells

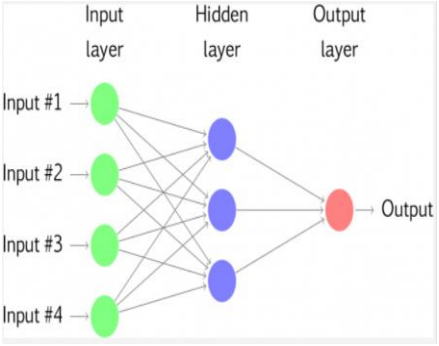

Fig. 2.2. The Structure of Neural Network

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 modi-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

fied 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):

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

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

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

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 6-10. 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.

TABLE 6 COMPARISON 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 COMPARISON 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 COMPARISON 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 COMPARISON 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 COMPARISON 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2752703, IEEE/ACM Transactions on Computational Biology and Bioinformatics

4 Conclusions

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 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.

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.

Acknowledgment

This work was supported by the grants of the National Science Foundation of China, Nos. 61520106006, 31571364, U1611265, 61532008, 61672203, 61402334, 61472282, 61472280, 61472173, 61572447, 61373098 and 61672382, China Postdoctoral Science Foundation Grant, Nos. 2016M601646. De-Shuang Huang is the corresponding author of this paper.

References

[1] Zhao S, Xu W, Jiang W, et al. Regulation of cellular metabolism by protein lysine acetylation[J]. Science, 2010, 327(5968): 1000-1004.

[2] Cheng, Alice, et al. "MoMo: Discovery of post-translational modification motifs." bioRxiv (2017): 153882.

[3] Khan, Amina S., 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." Journal of General Virology 98.4 (2017): 754-768.

[4] Yang, Jing, et al. "Post-Translational Modification of the Membrane Type 1 Matrix Metalloproteinase (MT1-MMP) Cytoplasmic Tail Impacts Ovarian Cancer Multicellular Aggregate Dynamics." Journal of Biological Chemistry (2017): jbc-M117.

[5] Liu X, Wang L, Zhao K, et al. The structural basis of protein acetylation by the p300/CBP transcriptional coactivator[J]. Nature, 2008, 451(7180): 846-850.

[6] Schwer B, Eckersdorff M, Li Y, et al. Calorie restriction alters mitochondrial protein acetylation[J]. Aging cell, 2009, 8(5): 604-606.

[7] Roth G J, Stanford N, Majerus P W. Acetylation of prostaglandin synthase by aspirin[J]. Proceedings of the National Academy of Sciences, 1975, 72(8): 3073-3076.

[8] Bode, Ann M., and Zigang Dong. "Post-translational modification of p53 in tumorigenesis." Nature reviews. Cancer 4.10, 2004: 793.

[9] Quivy V, Van Lint C. Regulation at multiple levels of NF-κB-mediated transactivation by protein acetylation[J]. Biochemical pharmacology, 2004, 68(6): 1221-1229.

[10] Guan K L, Xiong Y. Regulation of intermediary me-tabolism by protein acetylation[J]. Trends in biochemical sciences, 2011, 36(2): 108-116.

[11] Caron C, Boyault C, Khochbin S. Regulatory cross‐talk between lysine acetylation and ubiquitination: role in the control of protein stability[J]. Bioessays, 2005, 27(4): 408-415.

[12] Pesericо A, Simone C. Physical and functional HAT/HDAC interplay regulates protein acetylation bal-ance[J]. BioMed Research International, 2010, 2011.

[13] Watala C, Pluta J, Golanski J, 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]. Journal of Molecular Medicine, 2005, 83(2): 148-158.

[14] Hu L I, Lima B P, Wolfe A J. Bacterial protein acetylation: the dawning of a new age[J]. Molecular microbiology, 2010, 77(1): 15-21.

[15] Baeza J, Smallegan M J, Denu J M. Mechanisms and dynamics of protein acetylation in mitochondria[J]. Trends in biochemical sciences, 2016, 41(3): 231-244.

[16] Khoury, George A., Richard C. Baliban, and Christodoulos A. Floudas. "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database." Scientific reports 1 (2011): 90.

[17] Tsuda M, Kinugawa S, Takada S, et al. Mitochondrial Protein Acetylation in Skeletal Muscle Was Associated With Exercise Intolerance in Heart Failure After Myocardial Infarction in Mice[J]. 2016.

[18] Minton K. Gene expression: Reading protein acetyla-tion[J]. Nature Reviews Molecular Cell Biology, 2016.

[19] Chatterjee N, Tian M, Spirohn K, et al. Keap1-Independent Regulation of Nrf2 Activity by Protein Acetylation and a BET Bromodomain Protein[J]. PLoS Genet, 2016, 12(5): e1006072.

[20] Galdieri L, Gatla H, Vancurova I, et al. Activation of AMP-activated Protein Kinase by Metformin Induces Protein Acetylation in Prostate and Ovarian Cancer Cells[J]. Journal of Biological Chemistry, 2016, 291(48): 25154-25166.

[21] Shindo Y, Komatsu H, Hotta K, et al. An Artificial Reaction Promoter Modulates Mitochondrial Functions via Chemically Promoting Protein Acetylation[J]. Scien-tific Reports, 2016, 6.

[22] Barjaktarovic Z, Merl-Pham J, Azimzadeh O, et al. Low-dose radiation differentially regulates protein acetylation and histone deacetylase expression in human coronary artery endothelial cells[J]. International Journal of Radiation Biology, 2016: 1-9.

[23] Bharti S K, Brosh Jr R M. Fine-tuning DNA repair by protein acetylation[J]. Cell Cycle, 2016, 15(15): 1952-1953.

[24] Kuczyńska-Wiśnik D, Moruno-Algara M, Stojowska-Swędrzyńska K, et al. The effect of protein acetylation on the formation and processing of inclusion bodies and endogenous protein aggregates in Escherichia coli cells[J]. Microbial Cell Factories, 2016, 15(1): 189.

[25] Xie M, Hill J A. Cardiac Autophagy and Its Regula-tion by Reversible Protein Acetylation[M]//Epigenetics in Cardiac Disease. Springer International Publishing, 2016: 231-262.

[26] Baeza J, Smallegan M J, Denu J M. Mechanisms and dynamics of protein acetylation in mitochondria[J]. Trends in biochemical sciences, 2016, 41(3): 231-244.

[27] Li Y, Wang M, Wang H, et al. Accurate in silico iden-tification of species-specific acetylation sites by integrat-ing protein sequence-derived and functional features[J]. Scientific reports, 2014, 4: 5765.

[28] Hou T, Zheng G, Zhang P, et al. LAceP: lysine acety-lation site prediction using logistic regression classifiers[J]. PloS one, 2014, 9(2): e89575.

[29] Verdin E, Ott M. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond[J]. Nature reviews Molecular cell biology, 2015, 16(4): 258-264.

[30] Weinert B T, Satpathy S, Hansen B K, et al. Accurate quantification of site-specific acetylation stoichiometry reveals the impact of sirtuin deacetylase CobB on the E. coli acetylome[J]. Molecular & Cellular Proteomics, 2017: mcp. M117. 067587.

[31] Baeza J, Smallegan M J, Denu J M. Site-specific reac-tivity of nonenzymatic lysine acetylation[J]. ACS chemical biology, 2015, 10(1): 122-128.

[32] Pejaver V, Hsu W L, Xin F, et al. The structural and functional signatures of proteins that undergo multiple events of post‐translational modification [J]. Protein Science, 2014, 23(8): 1077-1093.

[33] Masui K, Tanaka K, Ikegami S, et al. Glucose-dependent acetylation of Rictor promotes targeted cancer therapy resistance [J]. Proceedings of the National Academy of Sciences, 2015, 112(30): 9406-9411.

[34] Kurotani A, Tokmakov A A, Kuroda Y, et al. Correlations between predicted protein disorder and post-translational modifications in plants [J]. Bioinformatics, 2014, 30(8): 1095-1103.

[35] Xu Y, Chou K C. Recent progress in predicting post-translational modification sites in proteins [J]. Current topics in medicinal chemistry, 2016, 16(6): 591-603.

[36] Wright P E, Dyson H J. Intrinsically disordered pro-teins in cellular signalling and regulation[J]. Nature Re-views Molecular Cell Biology, 2015, 16(1): 18-29.

[37] Mitchell L, Huard S, Cotrut M, et al. mChIP-KAT-MS, a method to map protein interactions and acetylation sites for lysine acetyltransferases[J]. Proceedings of the National Academy of Sciences, 2013, 110(17): E1641-E1650.

[38] Jia J, Liu Z, Xiao X, et al. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensem-ble random forest approach[J]. Journal of theoretical biol-ogy, 2016, 394: 223-230.

[39] Cobbold S A, Santos J M, Ochoa A, et al. Proteome-wide analysis reveals widespread lysine acetylation of major protein complexes in the malaria parasite[J]. Scien-tific reports, 2016, 6.

[40] Dou Y, Yao B, Zhang C. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine[J]. Amino acids, 2014, 46(6): 1459-1469.

[41] Minguez P, Letunic I, Parca L, et al. PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins[J]. Nucleic acids research, 2013, 41(D1): D306-D311.

[42] Qiu W R, Xiao X, Lin W Z, et al. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model[J]. Journal of Biomolecular Structure and Dynamics, 2015, 33(8): 1731-1742.

[43] Xiong Y, Peng X, Cheng Z, et al. A comprehensive catalog of the lysine-acetylation targets in rice (Oryza sativa) based on proteomic analyses[J]. Journal of prote-omics, 2016, 138: 20-29.

[44] Jia C, Lin X, Wang Z. Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition[J]. International journal of molecular sciences, 2014, 15(6): 10410-10423.

[45] Pougovkina O, te Brinke H, Ofman R, et al. Mito-chondrial protein acetylation is driven by acetyl-CoA from fatty acid oxidation[J]. Human molecular genetics, 2014: ddu059.

[46] T.L. Zhang, Y.S. Ding and K.C. Chou, Prediction Protein Structural Classes with Pseudo Amino Acid Composition: Approximate Entropy and Hydrophobicity Pattern, J. Theor. Bi-Ol. 250, 186-193. (2008)

[47] Hong-Jie Yu, D.S.Huang, "Graphical representation for DNA sequences via joint diagonalization of matrix pencil," IEEE Journal of Biomedical and Health Informat-ics, vol.17, no.3, pp.503-511, 2013

[48] Berezovsky I N, Kilosanidze G T, Tumanyan V G, et al. Amino acid composition of protein termini are biased in different manners [J]. Protein engineering, 1999, 12(1): 23-30.

[49] Andreeva A., Howorth D., Chandonia J.M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. Data Growth And Its Impact On The SCOP Database: New Development,(2007).

[50] D.S.Huang, Lei Zhang, Kyungsook Han, Suping Deng, Kai Yang, Hongbo Zhang, " Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," Current Protein & Peptide Science, vol. 15, no. 6: 553-560, 2014.

[51] D.S.Huang, Hong-Jie Yu, "Normalized feature vec-tors: A novel alignment-free sequence comparison meth-od based on the numbers of adjacent amino acids," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.10, no.2, pp.457-467, 2013

[52] Ding, C.H.Q., Dubchak, I. Multi-Class Protein Fold Recognition Using Support Vector Machines And Neural Networks. Bioinformatics (2001) 17 (4), 349-358.

[53] K. Chen, L.A. Kurgan, J.S. Ruan, Prediction of protein structural class using novel evolutionary colloca-tion-based sequence representation, J. Comput. Chem. 29(2008) 1596-1604.

[54] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195-202.

[55] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search pro-grams, Nucleic Acids Res. 25 (1997) 3389-3402.

[56] L.A. Kurgan, T. Zhang, H. Zhang, S. Shen, J. Ruan, Secondary structure-based assignment of the protein structural classes, Amino Acids 35 (2008)551-564.

[57] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate pre-diction of protein structural class for sequences of twi-light-zone similarity with predicting sequences, BMC Bioinform. 9 (2008) 226.

[58] T. Liu, C. Jia, A high-accuracy protein structural class prediction algorithm using predicted secondary structural information, J. Theor. Biol. 267 (2010) 272-275.

[59] Lempel, Ziv, On the complexity of finite sequences, IEEE Trans. Inf. Theory 22(1976) 75-81.

[60] Shuyan Ding, Shengli Zhang, Yang Li, Tianming Wang. A novel protein structural classes prediction method based on predicted secondary structure. Bio-chimie 94 (2012) 1166-1171

[61] Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, Y.Z. Chen, PROFEAT: A Web Server for Computing Structural and Physicochemical Features Of Proteins and Peptides from Amino Acid Sequence, Nucleic Acids Res(2006). 34 ,W32-W37.

[62] Chun-Hou Zheng, D.S.Huang, Lei Zhang, and Xiang-Zhen Kong, "Tumor clustering using non-negative matrix factorization with gene selection," IEEE Transactions on Information Technology in Biomedicine, vol. 13, no.4, pp 599-607, 2009.

[63] Lin Zhu, Zhu-Hong You, D.S.Huang and Bing Wang, "t-LSE:A Novel Robust Geometric Approach for Modeling Protein-Protein Interaction

Networks , " PLOS ONE, 8(4): e58368. doi:10.1371/journal.pone.0058368,2013.

[64] H.B. Rao, F. Zhu, G.B. Yang, Z.R. Li, Y.Z. Chen, Up-date Of PROFEAT: A Web Server for Computing Struc-tural and Physicochemical Features of Proteins And Pep-tides From Amino Acid Sequence, Nucleic Acids Res(2011). 39,W385-W39.

[65] Chun-Hou Zheng, Lei Zhang, Vincent To-Yee Ng, Simon Chi-Keung Shiu, and D.S.Huang, "Molecular pattern discovery based on penalized matrix decomposition," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol.8, no.6, pp.1592-1603, 2011.

[66] Chatterjee P, Basu S, Nasipuri M. Improving Predic-tion Of Protein Secondary Structure Using Physicochemi-cal Properties of Amino Acids[C]//Proceedings Of The 2010 International Symposium on Biocomputing (ISB 10). New York, NY, USA:ACM.

[67] B. Yang, Y.H. Chen and M.Y. Jiang, Reverse Engi-neering of Gene Regulatory Networks Using Flexible Neural Tree Models, Neurocomputing 99, 458-466. (2013)

[68] Chen Y, Yang B, Dong J. Evolving flexible neural networks using ant programming and PSO algo-rithm[M]//Advances in Neural Networks–ISNN 2004. Springer Berlin Heidelberg, 2004: 211-216.

[69] K.C. Chou, C.T. Zhang, Predicting of protein struc-tural class, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275-349.

[70] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1-16.

[71] H.B. Shen, K.C. Chou, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition, Biochem. Biophys. Res. Commun. 337 (2005) 752-756.

[72] Y.S. Ding, T.L. Zhang, K.C. Chou, Prediction of pro-tein structure classes with pseudo amino acid composi-tion and fuzzy support vector machine network, Protein Pept. Lett. 14 (2007) 811-815.

Wenzheng Bao received the master degree in computer science from the University of Jinan in 2015. He currently study for a PhD of computer science in the the School of Electronic and Information Engineering, Tongji University, Shanghai, China. His research interests include bioinfor-matics and machine learning.

ChangAn Yuan received the Ph.D. degree in Com puter Application Technology from the Sichuan U niversity, China, in 2006. He is a professor at Guan gxi Teachers Education University. His research in terests include Computational intelligence and Da ta mining.

Youhua Zhang, Male, Born Zhang Youhua, Male, Born in 1966.11, Ph.D. of University of Science and Technology of China, Professor, Master's tutor. Dean of school of information and computer of Anhui Agricultural University, executive council member of Computer Applications Branch of China Agricultural Society, executive council member of Anhui Association of Agricultural Information. Member of National Bee modern industrial technology and product quality and safety standards system. expert advisory committee of Anhui provincial human resources and Social Security Department of information technology. Executive Committee of the Hefei branch of the CCF. mainly engaged in teaching and research in computer applications and logistics engineering. In charge or participate in Provincial science and technology research, National Science and Technology Support Program, 863 project, National Natural Science Foundation, won the First Prize of scientific and technological progress in Anhui Province.

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

Asoke K. Nandi received the degree of Ph.D. in Physics from the University of Cambridge (Trinity College), Cambridge (UK). He held academic positions in several universities, including Oxford (UK), Imperial College London (UK), Strathclyde (UK), and Liverpool (UK) as well as Finland Distinguished Professorship in Jyvaskyla (Finland). In 2013 he moved to Brunel University London (UK), to become the Chair and Head of Electronic and Computer Engineering. Professor Nandi is a Distinguished Visiting Professor at Tongji University (China) and an Adjunct Professor at University of Calgary (Canada).

His current research interests lie in the areas of signal processing and machine learning, with applications to communications, gene expression data, functional magnetic resonance data, and biomedical data. He has authored over 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.

Professor Nandi 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 (USA) 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 (UK) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (UK) in 1998.

Barry Honig, tenured professor at Columbia University, the US Academy of Sciences. Director of Biology and Bioinformatics computing center of Columbia University. Editorial Board of PNAS, Journal of Molecular Biology, Structure, Biochemistry and Current Opinion in Structural Biology. Research interests include computational biology and bioinformatics. As of 2013, the international high-level SCI journals published more than 300 papers. SCI cited more than 15,600 times, Nature published 17, Science 3, Cell 3, PNAS 22. 2004 was elected to the National Academy of Sciences. In 2007 the US National Academy of Sciences Alexander Hollaender Award

De-Shuang Huang received the B.Sc., M.Sc. and Ph.D. degrees all in electronic engineering from 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. During 1993-1997 period he was a postdoctoral research fellow respectively in Beijing Institute of Technology and in National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In Sept, 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 TongjiUniversity as Chaired Professor. From Sept 2000 to Mar 2001, he worked as Research Associate in Hong Kong PolytechnicUniversity. From Aug. to Sept. 2003, he visited the GeorgeWashingtonUniversity as visiting professor, WashingtonDC, USA. From July to Dec 2004, he worked as the University Fellow in Hong KongBaptistUniversity. From March, 2005 to March, 2006, he worked as Research Fellow in ChineseUniversity of Hong Kong. From March to July, 2006, he worked as visiting professor in Queen's University of Belfast, UK. In 2007, 2008, 2009, he worked as visiting professor in Inha University, Korea, respectively. At present, he is the director of Institute of Machines Learning and Systems Biology, Tongji University. Dr. Huang is currently Fellow of International Association of Pattern Recognition (IAPR Fellow), senior members of the IEEE and International Neural Networks Society. He has published over 180 journal papers. Also, in 1996, he published a book entitled "Systematic Theory of Neural Networks for Pattern Recognition" (in Chinese), which won the Second-Class Prize of the 8th Excellent High Technology Books of China, and in 2001 & 2009 another 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. His current research interest includes bioinformatics, pattern recognition and machine learning.