SUMOylation Sites Prediction by Machine Learning Approaches

Chi-Wei Chen[13], Chin-Hau Tu[1] and Yen-Wei Chu[12*]

[1] Institute of Genomics and Bioinformatics, National Chung Hsing University, Kuo Kuang Rd., Taichung, Taiwan

[2] Biotechnology Center, Agricultural Biotechnology Center, Institute of Molecular Biology, Graduate Institute of Biotechnology, National Chung Hsing University, Kuo Kuang Rd., Taichung, Taiwan

[3] Department of Computer Science and Engineering, National Chung-Hsing University, Kuo Kuang Rd., Taichung, Taiwan

* Contact emails: ywchu@nchu.edu.tw

***Abstract*—Post-translational modification (PTM) influence still does not considered by current sumoylation prediction tools. Therefore, this study developed a sumoylation prediction system based on machine learning approach employing SVM (support vector machine) and related information. In the feature coding, we encoded binary code and protein properties based on amino acid sequence. Besides, we encoded other PTM distribution as functional feature and secondary information as structure feature. In addition, we analyzed the number of the post-modification distributions under the central lysine and window size 21 rules, and we provided some of our findings and recommended post-modification types that could be considered. Finally, this study developed a new sumoylation prediction algorithm called SUMOdig. The prediction system of Matthew's correlation coefficient achieves to 0.504.**

I. Introduction

Post-translational modification (PTM) refers to the covalent and generally enzymatic modification of proteins during or after protein biosynthesis. Small Ubiquitin-like Modifier (SUMO) proteins are a family of small proteins that are covalently attached to and detached from other proteins in cells to modify their function. Sumoylation is a post-translational modification involved in various cellular processes, such as gene expression, DNA repair, chromosome assembly, and cellular signaling [1]. Along with the accumulating research on its biological functions, there are abundant evidences that the aberrance of SUMO regulation is highly associated with various diseases, such as neurodegenerative diseases, congenital heart defects, diabetes, and cancers. Thus, identified sumoylation sites is valued at the research on several of disease and biological mechanism.

Machine learning is one of the common methods to predict sumoylation sites [2,3]. In this study, we developed a sumoylation prediction system based on machine learning approach employing SVM (support vector machine) and also updated sumoylation consensus motif and related information.

Recently, most sumoylation prediction tools are using algorism, protein physicochemical and biochemical properties or consensus motif to predict modification sites. But those tools rarely mention to the effect of other post translational modification (PTM) on sumoylation prediction. To our knowledge, a study suggested that acetylation may influence sumoylation and it is an entry point for sumoylation research. According to the entry point, we guess not only acetylation but also other PTMs may influence the process of sumoylation. Furthermore, secondary structure has reported as an important feature from a related paper [19]. Thus, in this study, we develop a sumoylation prediction system based on PTMs distribution to investigate the effect of other PTMs. Besides, other studies usually use a large number of protein properties as features from AAindex. Instead of those protein properties, we use ten recommended protein properties to simplify the process of selecting useful features. Finally, we develop a sumoylation prediction system which considered the effect of other PTMs.

II. Materials and Methods

A. Data collection

The experimental datasets used in this study were obtained from UniProt, dbPTM, and PhosphoSitePlus. Table I shows the details information from the three data.

TABLE I
THE INFORMATION OF DOWNLOAD SUMOYLATION DATA

Protein database	Protein number	SUMOylation site
Uniprot	1779	3615
dbPTM	434	1029
PhosphoSitePlus	492	912

The training and testing data were removed duplicated data based on protein number. After determining the training data was sent to CD-HIT to remove the similarity sequence. The cut-off values set in this study were 0.3, and the results of the removed similarity sequences. All of the lysine on the amino acid fragment in this study were the target. In contrast to positive data, negative data is lysine but not sumoylation. Under this definition, the statistics of positive and negative number in training and testing data are shown in Table II.

TABLE II
THE NUMBER OF POSITIVE AND NEGATIVE NUMBER IN TRAINING AND TESTING DATA

Training data set		Testing data set	
Positive	Negative	Positive	Negative
869	20903	867	18825

B. Binary encoding

The vector is a common formation using in machine learning. Thus, we encoded 20 amino acid including gap with 20 dimensions. The dimension occupied by the amino acid is 1. So the 20 amino acids by binary encoding into 0 and 1 constitute the 20 different combinations of the series. If the value is null, all 0 are used.

C. Physicochemical and biochemical properties

Previous studies have used physic-chemical properties codes and most of which used AAindex (Amino acid index database). William *et al.* [4] simplified AAindex's amino acid

characteristics according to its similarity, classified as polarity, secondary structure, molecular size or volume, codon diversity and electrostatic charge. In addition, Mathura *et al.* [5] summarized five characteristics of Hydrophobicity, Side chain length, α-helix propensity, Number of codons and β-strand propensity from the literature. Thus, the results of the amino acid characteristics summarized in the two studies were coded.

D. Structure based features

This study uses the NetSurfP website to predict protein structure surface accessibility, which provides information of relative surface accessibility, absolute surface accessibility, Z-score for relative surface accessibility. We encoded amino acid with those information, but in the buried and exposed column, if amino acid is buried that we encoded it 10 and exposed encoded it 100. While the eighth to tenth columns are divided into three major classes of α-helix, β-strand and Coil. Those secondary structure probability scores were encoded as secondary structure information.

E. PTM distribution

ModPred is a tool that simultaneously predicts 23 post-translational modifications of proteins. The coding method is following: sites with low confidence were coded as 10, medium confidence as 50, and high confidence as 100. And the number of sites without post-modification was zero. Based on the post-translational modification of lysine-related proteins [30], six proteins were selected for post-translational modification, including Acetylation, Hydroxylation, Methylation, Phosphorylation, SUMOylation and Ubiquitination.

F. Training model

We used LIBSVM with radial basis function (RBF) kernel to predict sumoylation sites, which could solve non-linear problems. The SVM running files contains the positive data and the ratio of negative data. All positive data were added to SVM running file. While negative data were extracted based on ratio by random sampling because of the larger data number than positive data number. In this study, the ratio of positive data set is 1, the number of negative data set is 1, 1.5, and 2. The method of selecting negative data is random sampling. Each consensus motif type of each ratio would be evaluated 30 times with 5-fold cross-validation. The final result would take the average of the respective items of Matthews's correlation coefficient results.

III. Result and Discussion

Fig. 1. show the Methylation distribution of window size 21. The colors blue, red and green in the graph represent confidence in ModPred, which is low, medium, and high, respectively. In this feature encoding, those are coded as 1, 50,100. Methylation has a significant number of PTM at position twenty-first, which is a suggestion for further research. And they may be used as a clue for the future sumoylation prediction development.

A. Comparison results of other SUMOylation predict tools

In this study, we developed a prediction system and was named SUMOdig. To investigate our prediction system ability, we compared with other predictive tools, including SUMOsp2.0, JASSA, and PCI-SUMO. The results are shown in Table III. The results indicated that SUMOdig showed the highest prediction accuracy and its average MCC value reached 0.504 which is better than SUMOsp2.0, JASSA and PCI-SUMO.

TABLE III
COMPARISON RESULTS OF SUMOYLATION PREDICTIVE TOOLS IN THE DP TESTING SET

Tool name	Sn	Sp	Acc	MCC
SUMOdig	0.475	0.981	0.953	0.504
SUMOsp2.0	0.287	0.983	0.913	0.396
JASSA	0.391	0.984	0.941	0.476
PCI-SUMO	0.075	0.978	0.616	0.128

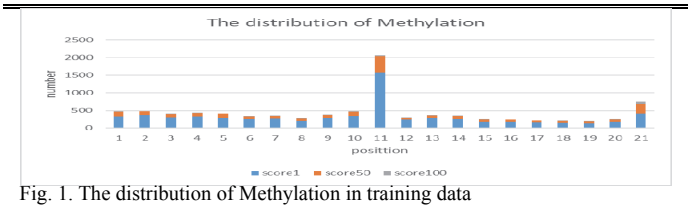

Fig. 1. The distribution of Methylation in training data

IV. Conclusion

In this study developed SUMOdig, a sumoylation predictive algorithm based on machine learning. And we investigated the model without other PTM distribution. The result shows that other PTM distribution affect the result of sumoylation prediction. Furthermore, our prediction, SUMOdig, has a better stability and MCC result of testing data.

In this study, we also calculated the distribution of the PTM. Moreover, we suggest that Methylation on 21th position is a candidate feature. We can study it and add it in future development.

Acknowledgment

This research was supported by Ministry of Science and Technology, Taiwan, R.O.C. under grant number 106-2221-E-005-077-MY2.

Examples of Reference Styles

[1] R. Geiss-Friedlander and F. Melchior, "Concepts in sumoylation: a decade on," Nature reviews Molecular cell biology, vol. 8, no. 12, pp. 947-956, 2007.

[2] A. S. Yavuz and O. U. Sezerman, "Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder," BMC genomics, vol. 15, no. 9, p. S18, 2014.

[3] J. Jia, L. Zhang, Z. Liu, X. Xiao, and K.-C. Chou, "pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC," Bioinformatics, vol. 32, no. 20, pp. 3133-3141, 2016.

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

[5] M. S. Venkatarajan and W. Braun, "New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties," *Journal of Molecular Modeling,* vol. 7, no. 12, pp. 445-453, 2001.