International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Identifying Protein SUMOylation Sites based on the Combination of Amino Acid Composition and k-Spaced Amino Acid Pairs

Van-Nui Nguyen[1], Hong-Tan Nguyen[2]

[1,2]University of Information and Communication Technology (ICTU), Quyet Thang, Thai Nguyen, Vietnam

Abstract: ***SUMOylation has been known as one of the most important post-translational modification in Eukaryotes species, which has significant roles in many biological processes and cellular functions. The mechanism underlined SUMOylation process will affect many biological processes and functions, leading to many common serious diseases, such as: breast cancer, cardiac, Parkinson's and Alzheimer's disease. Because of its very inportant roles, the demand on extensively understanding of SUMOylation and its mechanism is one of the most hottest isusse that interested many researchers nowadays. In this work, we will present an approach combinating of amino acid composition and informative k-spaced amino acid pairs to identify protein SUMOylation sites.***

Keywords: SUMOylation, support vector machine (SVM), amino acid composition, k-spaced amino acid composition

1. Introduction

Protein SUMOylation is a kind of very important post-translational modification (PTM) that plays significant roles in many biological processes and cellular functions. The machinery of SUMOylation process will affect many biological processes and functions, and then leading to many common serious diseases [1, 2]. Due to the important roles regulated by SUMOylation, the demand on extensively understanding of SUMOylation and its mechanism is one of the most hottest isusse that interested many researchers nowadays. So far, there is an increasing number of researches proposed for the identification of protein SUMOylation [3-8]. Besides, various predictors have been developed to support scientist identifying protein SUMOylation sites [9-13].

Althought there are many of researches has been proposed for identifying protein SUMOylation sites [9-18], however the number is still not meet our demand to have extensively understanding of protein SUMOylation and its mechanism. Therefore, in this work we will present an approach incorporating of amino acid composition and k-spaced amino acid pairs to identify protein SUMOylation sites. The results has demonstrated that our proposed approach could be efficiently used for identifying the potential protein SUMOylation sites

2. Data Preparation and Model Learning

2.1. Data preparation

In this work, the experimentally verified SUMOylation sites has been collected from many different resources, including: SUMOsp [9], GPS-SUMO-Ver 3.0 [10], JASSA [11], pSumo-CD [12], SUMOhydro [13] and dbPTM-2019 [19]. After the process of removing duplicated or redundant data, we obtained a total of 1160 uniques proteins (having 2109 SUMOylation sites) for this work. Of these 1160 uniques proteins, we have randomly selected 160 proteins (containing 289 SUMOylation sites) to be utilized as independent testing dataset. The remaining data (1000 unique proteins, having 1820 SUMOylation sites) has been used as training dataset.

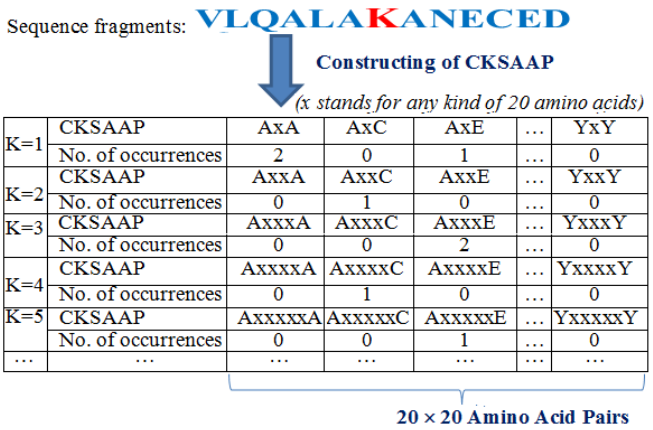
In this work, we analyze the characterization of substrate site specificity of SUMOylated protein in-term of sequence-based. So, applying the same approach from previous works [14-18, 20] in extracting data being used for model training, the window size of 13 has been selected to extract 13-mer fragment sequence (-6 to +6) with the Lysine (K) at the central of the sequence. With the 1000 experimentally verified SUMOylated proteins of the training dataset, the total fragments that has extracted using window size of 13 containing 1820 postitive fragments and 37222 negative fragments. As the binary classification problem, the performance of the predictive models may be overestimated or underestimated due to the fact of homologous fragments in the positive and negative dataset. Thus, the CD-HIT program [21] has been applied to remove homologous fragments. With the use of 40% of fragment identify, the training dataset aftered filtered out consists of 745 positive training fragment and 7450 negative training fragments.

To find out the best model, firstly the cross-validaion approach is adopted to evaluate the performance of the various predictive models. Then, the best predictive model with the highest accuracy and MCC value is selected. After choosing of the best predictive model, it is neccessary to perform an independent testing to assess the real case of the chosen model. As presented above, the independent testing dataset contains 160 proteins. Applying the same approach of extracting training fragment, the final independent testing dataset containing 117 positive and 1170 negative fragments.

2.2. Features Encoding and Transformation

In this study, various sequence-based features have been investigated, including: Amino Acid Composition (AAC), Amino Acid Pairwise Composition, Positional Weighted Matrix (PWM) and Evolutionary information (PSSM, Position-Specific Scoring Matrix). The AAC, AAPC and PSSM features have been extracted and encoded by applying same approach with previous studies [14-18]. The PWM feature has been built by referring the SulfoSite method [22]. The PWM was determined by calculating the occurrence rate of twenty types of amino acids surrounding a substrate SUMOylation sites, and was utilized in encoding for the sequence fragment. Each sequence fragment as represented by a matrix of $(2n+1)\times w$ elements, where w stands for 21 elements including 20 types of amino acids and one for the non-existing residue. In the viewpoint of protein sequence evolution, several amino acid residues of a protein can be mutated without changing its score structure or functional domain.

Besides, the *k-spaced* amino acid pairs (CKSAAP) encoding using CKSAAP scheme [23-25] been anlayzed also. This study has examined the CKSAAPs with *k* ranging from 1 to 5, as displayed in Figure 1.



K=1	CKSAAP	AxA	AxC	AxE	…	YxY
	No. of occurrences	2	0	1	…	0
K=2	CKSAAP	AxxA	AxxC	AxxE	…	YxxY
	No. of occurrences	0	1	0	…	0
K=3	CKSAAP	AxxxA	AxxxC	AxxxE	…	YxxxY
	No. of occurrences	0	0	2	…	0
K=4	CKSAAP	AxxxxA	AxxxxC	AxxxxE	…	YxxxxY
	No. of occurrences	0	1	0	…	0
K=5	CKSAAP	AxxxxxA	AxxxxxC	AxxxxxE	…	YxxxxxY
	No. of occurrences	0	0	1	…	0
…	…	…	…	…	…	…

Figure 1: The construction of CKSAAP feature

Given 20×20 amino acid pairs and five values for *k*, the total of 5×20×20 = 20000 attributes are used to train the predictive model. Due to the fact that the higher dimensions of features vectors could induce a lower efficiency of model learning and evaluation. Thus, all of these 2000 CKSAAP features should be tuned to achieve the optimal CKSAAPs for providing better predictive performance. In this work, to extract informative features prior constructing predictive model, each CKSAAPs attribute is examined based on the index score calculated by the minimum redundancy-maximum relevance (mRMR) algorithm [26]. According the findings in [26, 27], the CKSAAP attribute having maximum relevance and minimum redundancy will contain the best discriminating power between positive and negative instances.

2.3. Model learning and performance evaluation

It has been common known that support vector machine (SVM) is a well-known machine learning method and widely utilized for solving the parttern identification problem with clear connection to the underlying statistical learning theory. With purpose of identifying potential protein SUMOylation site is positive or not, it comes to meet and suitable with the problem of the binary classification using SVM method. Herein, LibSVM [28], a public SVM library proposed by Chang C. C. and Lin C.J, is adopted to contruct the predictive models to discriminate the SUMOylation sites from non-SUMOylation sites.

To evaluate the performance of the predictive models, the 5-fold cross-validation approaches has been performed to assess the classifying power of the constructed SVM-based models. The following measurements are common used to evaluate the performance of the constructed models:

The common measures: Sensitivity (SEN), Specificity (SPE), Accuracy (ACC), and Matthews Correlation Coefficient (MCC):

$$SEN = \frac{TP}{TP+FN}; SPE = \frac{TN}{TN+FP}; ACC = \frac{TP}{TP+FN};$$
$$MCC = \frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)(TP+FP)(TN+FN)}}$$

Wherein the measurements were explained as belows:

+ *TP* (True Positive), *TN* (True Negative) represented the number of positive and negative sites that are correctly predicted.
+ *FP (False Positive) and FN (False Negative) indicated the number of positive and negative sites that are falsely predicted.*
+ *SEN (Sensitivity) and SPE (Specificity) measured the proportion of positives and negatives that are correctly identified.*
+ *MCC is an import measurement that has been used to reflect the balance quality in case of the numbers of negative and positive data are significant imbalance.*

After running 5-fold cross-validation process, the constructed model containing highest values of MCC and accuracy has been selected as the optimal model for identifying potential protein SUMOylation sites. Moreover, the independent testing approach has also been carried out to evaluate the ability of selected model, in the real case.

3. Results and Discussion

Based on the analysis of amino acid composition on the substrate protein, the frequency of occurrence of twenty amino acid residues surrounding the substrate sites could be determined to find the potential consensus motifs for the identifying SUMOylation sites. As displayed in
Table ***1***, the total of 5 single features (AAC, AAPC, PWM, PSSM, CKSAAP) have been investigated for the identification of protein SUMOylation sites.

As displayed in
Table ***1***, the total of 5 single features (AAC, AAPC, PWM, PSSM, CKSAAP) have been investigated for the identification of protein SUMOylation sites. Additionally, as binary classification between SUMOylation and non-SUMOylation sites, it is feasible to combine two or more

different feature to generate hybrid features to be used for the model learning. Therefore, based on single features, we have constructed 4 hybrid features to be analyzed for the identification of SUMOylation sites.

Table ***1*** displayed in detail the performance of constructed models when evaluated using the *five-fold cross-validation*. The hybrid feature of "PSSM+CKSAAP" is appeared to be the optimal feature for contructing the predictive model, reaching the accuracy value at 78,44% and MCC value is at 0,380.

Table 1: Performance evaluation by Five-Fold Cross-Validation

Feature	Five-fold Cross-Validation			
	SEN	SPE	ACC	MCC
AAC	60.40%	67.11%	66.50%	0.165
AAPC	67.11%	67.11%	67.11%	0.205
PWM	66.22%	67.11%	67.03%	0.199
PSSM	73.83%	67.11%	67.72%	0.244
CKSAAP (K = 1)	73,15%	67,11%	67,66%	0,240
CKSAAP (K = 2)	74,63%	67,11%	67,80%	0,249
CKSAAP (K = 3)	75,17%	67,11%	67,85%	0,252
CKSAAP (K = 4)	74,50%	66,85%	67,54%	0,246
CKSAAP (K = 5)	66,44%	67,11%	67,05%	0,201
AAC+CKSAAP	66,17%	67,11%	67,03%	0,199
AAPC+CKSAAP	81,21%	73,56%	74,25%	0,339
PWM+CKSAAP	78,52%	75,56%	75,82%	0,338
PSSM+CKSAAP	81,21%	78,17%	78,44%	0,380

Moreover, the independent testing has been performed to assess the performance of the predictive model for the real case. Table 2 displayed in detail the performance of the predictive model using independent testing approach. Luckily, the results indicated that the hybrid feature of "***PSSM+CKSAAP"*** was also the best feature that could help to yield the highest performance, reaching the accuracy value at 73,91% and MCC value is at 0,324.

Table 2: Performance evaluation by Independent Testing

Feature	Independent Testing			
	SEN	SPE	ACC	MCC
AAC	62.39%	61.97%	62.00%	0.143
AAPC	65.81%	62.39%	62.70%	0.165
PWM	64.10%	62.31%	62.47%	0.155
PSSM	70.09%	70.51%	70.47%	0.248
CKSAAP (K = 1)	72,65%	70,54%	70,73%	0,263
CKSAAP (K = 2)	72,65%	72,33%	72,36%	0,278
CKSAAP (K = 3)	73,45%	71,82%	71,96%	0,275
CKSAAP (K = 4)	75,22%	73,10%	73,29%	0,296
CKSAAP (K = 5)	75,22%	72,59%	72,82%	0,291
AAC+CKSAAP	72,57%	73,19%	73,13%	0,281
AAPC+CKSAAP	77,88%	73,27%	73,68%	0,313
PWM+CKSAAP	76,99%	73,10%	73,44%	0,306
PSSM+CKSAAP	79,65%	73,36%	73,91%	0,324

4. Conclusion

SUMOylation has been known as one of the most important post-translational modification in Eukaryotes species. It plays a very important roles in many biological processes, cellular functions, as well as being a key factor that leads to many common serious diseases nowadays. In this work, we have presented an approach that combinates amino acid composition and informative k-spaced amino acid pairs to identify protein SUMOylation sites. Evaluation by cross-validation and independent testing approach, the proposed model has been demonstrated its strength and ability in the purpose of identifying the potential protein SUMOylation sites.

5. Acknowledgement

The authors sincerely thank to ICTU for partly financial supported this research under the TNU-level project ID: DH2018-TN07-01.

References

[1] Yang Y, He Y, Wang X, Liang Z, He G, Zhang P, Zhu H, Xu N, Liang S: **Protein SUMOylation modification and its associations with disease**. *Open biology* 2017, **7**(10).
[2] Sarge KD, Park-Sarge OK: **Sumoylation and human disease pathogenesis**. *Trends in biochemical sciences* 2009, **34**(4):200-205.
[3] Ramazi S, Zahiri J, Arab SS, Parandian Y: **Computational Prediction of Proteins Sumoylation: A Review on the Methods and Databases**. *Journal of Nanomedicine Research* 2016, **3**(5).
[4] Gong L, Qi R, Li DW: **Sumoylation Pathway as Potential Therapeutic Targets in Cancer**. *Current molecular medicine* 2016.
[5] Chen Z, Lu W: **Roles of ubiquitination and SUMOylation on prostate cancer: mechanisms and clinical implications**. *International journal of molecular sciences* 2015, **16**(3):4560-4580.
[6] Yavuz AS, Sezerman OU: **Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder**. *BMC genomics* 2014, **15 Suppl 9**:S18.
[7] Osula O, Swatkoski S, Cotter RJ: **Identification of protein SUMOylation sites by mass spectrometry using combined microwave-assisted aspartic acid cleavage and tryptic digestion**. *Journal of mass spectrometry : JMS* 2012, **47**(5):644-654.
[8] Galisson F, Mahrouche L, Courcelles M, Bonneil E, Meloche S, Chelbi-Alix MK, Thibault P: **A novel proteomics approach to identify SUMOylated proteins and their modification sites in human cells**. *Molecular & cellular proteomics : MCP* 2011, **10**(2):M110 004796.
[9] Xue Y, Zhou F, Fu C, Xu Y, Yao X: **SUMOsp: a web server for sumoylation site prediction**. *Nucleic acids research* 2006, **34**(Web Server issue):W254-257.
[10] Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, Liu Z, Zhao Y, Xue Y, Ren J: **GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs**. *Nucleic acids research* 2014, **42**(Web Server issue):W325-330.

[11] Beauclair G, Bridier-Nahmias A, Zagury JF, Saib A, Zamborlini A: **JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs**. *Bioinformatics* 2015, **31**(21):3483-3491.
[12] Jia J, Zhang L, Liu Z, Xiao X, Chou KC: **pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC**. *Bioinformatics* 2016, **32**(20):3133-3141.
[13] Chen YZ, Chen Z, Gong YA, Ying G: **SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties**. *PloS one* 2012, **7**(6):e39195.
[14] Nguyen VN, Huang KY, Huang CH, Chang TH, Bretana N, Lai K, Weng J, Lee TY: **Characterization and identification of ubiquitin conjugation sites with E3 ligase recognition specificities**. *BMC bioinformatics* 2015, **16 Suppl 1**:S1.
[15] Nguyen VN, Huang KY, Weng JT, Lai KR, Lee TY: **UbiNet: an online resource for exploring the functional associations and regulatory networks of protein ubiquitylation**. *Database : the journal of biological databases and curation* 2016, **2016**.
[16] Nguyen VN, Huang KY, Huang CH, Lai KR, Lee TY: **A New Scheme to Characterize and Identify Protein Ubiquitination Sites**. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 2017, **14**(2):393-403.
[17] Nguyen VN, Tran TX, Nguyen HM, Nguyen HT, Lee TY: **A new schema to identify S-farnesyl cysteine prenylation sites with substrate motifs**. *Advances in Information and Communication Technology ICTA 2016 Advances in Intelligent Systems and Computing* 2017, **53**.
[18] Nguyen VN, Bui VM: **The prediction of Succinylation site in protein by analyzing amino acid composition**. *Advances in Information and Communication Technology ICTA 2016 Advances in Intelligent Systems and Computing* 2017, **538**.
[19] Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, Chen YJ, Chen YJ, Huang HD: **DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications**. *Nucleic acids research* 2013, **41**(Database issue):D295-305.
[20] Nguyen V-N, Do H-K, Tran T-X, Le N-Q-K, Le A-T, Lee T-Y: **Exploiting two-layered suport vector machine to predict protein sumoylation sites**. *Advances in Engineering Research and Application: Proceedings of the International Conference, ICERA 2018* 2019, **63**:9.
[21] Huang Y, Niu B, Gao Y, Fu L, Li W: **CD-HIT Suite: a web server for clustering and comparing biological sequences**. *Bioinformatics* 2010, **26**(5):680-682.
[22] Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD, Pan RL: **Incorporating support vector machine for identifying protein tyrosine sulfation sites**. *Journal of computational chemistry* 2009, **30**(15):2526-2537.
[23] Chen Z, Zhou Y, Song J, Zhang Z: **hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties**. *Biochimica et biophysica acta* 2013, **1834**(8):1461-1467.
[24] Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z: **Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs**. *PloS one* 2011, **6**(7):e22930.
[25] Wang XB, Wu LY, Wang YC, Deng NY: **Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs**. *Protein engineering, design & selection : PEDS* 2009, **22**(11):707-712.
[26] Ding C, Peng H: **Minimum redundancy feature selection from microarray gene expression data**. *Journal of bioinformatics and computational biology* 2005, **3**(2):185-205.
[27] Lv H, Han J, Liu J, Zheng J, Liu R, Zhong D: **CarSPred: a computational tool for predicting carbonylation sites of human proteins**. *PloS one* 2014, **9**(10):e111478.
[28] Lin C-CCaC-J: **LIBSVM: a library for support vector machines**. *Acm Transactions on Intelligent Systems and Technology* 2011

Author Profile

Van-Nui Nguyen was born in Hai Duong province, Vietnam. He obtained his PhD degree in Department of Computer Science & Engineering from Yuan Ze University, Taiwan. His research interests include computer science, bioinformatics, computational proteomics and data mining, machine learning and deep learning.

Hong-Tan Nguyen was born in Bac Giang province, Vietnam. He obtained his master degree in University of Information and Communication technology. His research interests include computer science, machine learning, software engineering and software testing and assessment.