Journal of Computer Science and Cybernetics, V.40, N.4 (2024), 315-325
DOI: 10.15625/1813-9663/19626

CLW-SUMO: A HYBRID DEEP LEARNING MODEL FOR PREDICTING PROTEIN SUMOYLATION SITES

THI-XUAN TRAN[1], THI-THU-HUONG TRAN[2],
NGUYEN-QUOC-KHANH LE[3,4], VAN-NUI NGUYEN[5,*]

[1] *University of Economics and Business Administration, Tan Thinh Ward, Thai Nguyen City, Viet Nam*
[2] *Thai Binh University, Tan Binh Ward, Thai Binh City, Viet Nam*
[3] *Professional Master Program in Artificial Intelligence in Medicine, Taipei Medical University, Yuantong Road., Zhonghe District., Taipei City, Taiwan*
[4] *Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Yuantong Road., Zhonghe District., Taipei City, Taiwan*
[5] *University of Information and Communication Technology, Z115 Street, Quyet Thang Commune, Thai Nguyen City, Viet Nam*

Abstract. Protein SUMOylation is one of the most important post-translational modifications in Eukaryotes species and plays significant roles in many biological processes. The mechanism underlined the SUMOylation process will be an important cause leading to many common serious diseases, such as breast cancer, cardiac, Parkinson's, Alzheimer's disease, etc. Due to the very important roles regulated by SUMOylation, the demand for an in-depth understanding of SUMOylation and its mechanism is currently a hot topic that interests many scientists. In this study, we propose a novel approach, called CLW-SUMO, for predicting SUMOylation sites using a hybrid deep learning model that combines convolutional neural networks (CNN) and long short-term memory (LSTM), using Word2Vec as the word embedding technique. The 10-fold cross-validation demonstrates that our proposed model achieves the best performance with an accuracy of 82.33%, MCC of 0.589 and AUC of 0.829. Besides, the independent testing also shows that our proposed model obtains the highest performance, reaching an accuracy of 90.03%, MCC of 0.773 and AUC of 0.889. Furthermore, when compared to several existing predictors of SUMOylation using an independent dataset, our proposed model exhibits the highest performance with an ACC value of 90.03% and an MCC value of 0.773. We hope that our findings will provide effective suggestions and greatly help researchers in their studies related to protein SUMOylation identification.

Keywords. SUMOylation, prediction, convolutional neural networks, long short-term memory, natural language processing, Word2Vec.

*Corresponding author.
E-mail addresses: tranxuantbhd@tueba.edu.vn (T.X. Tran); tranhuongdhcn@gmail.com (T.T.H. Tran); khanhlee@tmu.edu.tw (N.Q.K. Le); nvnui@ictu.edu.vn (V.N. Nguyen).

1. INTRODUCTION

Most proteins undergo post-translational modifications (PTMs) throughout their lifetimes, which modulate their functions by altering the structure, dynamics, subcellular locations, and interactions of the modified proteins, resulting in a broader functional repertoire of the proteome. Among these PTMs, SUMOylation is highly significant in eukaryotic cells, where small ubiquitin-like modifiers (SUMOs) covalently attach to specific lysine (K) residues of target proteins in a reversible manner. SUMO proteins are widely expressed and evolutionarily conserved in eukaryotes, underscoring their functional importance. Initially recognized for its role in binding nuclear proteins, SUMO has since been found to participate in various activities, including transcription regulation, chromatin remodeling, DNA repair, and the control of cell cycle progression. Alterations in the SUMOylation process have also been linked to several diseases, including Alzheimer's, cancer, and autoimmune diseases [1–3].

The covalent attachment of SUMO to its target involves a series of reactions facilitated by SUMO-activating enzymes (E1), SUMO-conjugating enzyme E2 (Ubc9), and various SUMO E3 ligases. To reverse the SUMOylation process, SUMO-specific proteases cleave the bond between SUMO and its substrate. Mass spectrometry (MS)-based proteomics is a powerful tool for detecting SUMOylated proteins and SUMOylation sites in a high-throughput manner. However, the transient nature of SUMOylation, the low stoichiometry of SUMO, and the small fraction of SUMOylated proteins present technical challenges in studying SUMOylation events. Consequently, computational methods have been proposed to complement experimental efforts by narrowing down potential SUMOylation sites.

In recent years, the amount of interest in the prediction of proteins based on computational approaches has been increasing rapidly. Another widely utilized tool is GPS-SUMO in [4], which represents the latest iteration in a series of previously developed tools, including SUMOsp [5] and SUMOsp 2.0 [6], GPS-SUMO [7] was created using the group-based phosphorylation scoring algorithm. Additionally, pSumo-CD [8] is another tool that employs a covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC for its predictions. SUMO-Forest [9] utilizes bi-gram and k-skip-bi-gram representations of the input peptide (with k=1,2) and relies on an ensemble technique called Cascade Forest on imbalanced data. C-iSUMO [10] uses an Adaboost classifier, which depends on features derived from structural properties, such as the accessible surface area of the protein site and backbone torsion angles between residues. ResSUMO [11] uses the convolution neural network (CNN) model integrated with residue structure. Furthermore, several other methods rely on various machine learning and deep learning models [12, 13].

In this work, we introduce a learning architecture for predicting protein SUMOylation sites, called CLW-SUMO. This CLW-SUMO architecture has been built on a hybrid Deep Learning by incorporating CNN and LSTM, using Word2Vec as the word embedding technique. To assess the performance of the CLW-SUMO model, we have applied the 10-fold cross-validation and independent testing approaches. As a result, the CLW-SUMO model achieves the highest performance on both two evaluation approaches, reaching an accuracy of 82.33% and MCC of 0.589 on 10-fold cross-validation evaluation, an accuracy of 90.03% and MCC of 0.773 on independent testing evaluation. Furthermore, we also compare the proposed CLW-SUMO against other machine learning methods. The obtained result reveals that our proposed model of CLW-SUMO outperforms existing predictors. This demonstrates that our proposed model of CLW-SUMO could provide a promising approach for the predic-

tion of protein SUMOylation sites.

2. MATERIAL AND METHODS

2.1. Data collection and pre-processing

The experimentally verified SUMOylation sites are collected from various open resources and published in the literature, such as dbPTM3.0, JASSA, SUMOhydro, pSumo-CD, HseSUMO, GPS-SUMO, ResSUMO, etc. [4, 8, 11, 13–17].

In total, a dataset of 3639 proteins with 8838 SUMO-sites has been collected. After doing some technical steps to remove duplicated or redundant proteins, we obtain the final non-redundant dataset containing 3000 unique proteins. In order to prepare for independent testing, we randomly select 1/3 proteins from the non-redundant dataset to serve as the independent testing dataset. The remaining data is then considered as a training dataset. As a result, our final training dataset contained 2000 unique proteins, and the final independent testing dataset contained 1000 unique proteins.

Table 1: Data statistics of experimentally verified SUMOylation sites

Resources	SUMOylated proteins	SUMO-sites
dbPTM 3.0	1432	5191
SUMOsp	197	332
seeSUMO	247	377
GPS-SUMO	510	912
JASSA	505	877
pSUMO-CD	510	755
SUMOhydro	238	394
Total	3639	8838
Combined NR data	3000	7982
Training dataset	2000	5890
Testing dataset	1000	2092

In this study, we focus on the sequence-based characterization of SUMOylation sites with substrate specificities. Therefore, to generate the positive data (SUMO-data), we use a window length of $2n+1$ to extract sequence fragments that center at the experimentally verified SUMOylated lysine (K) residue, as well as containing n upstream and n downstream flanking amino acids. With a given a number of experimentally verified SUMOylated proteins, the sequence fragments containing window length of $2n + 1$ amino acids and centering at lysine residue without the annotation of SUMOylation were regarded as the negative training data (non-SUMO data). Based on previous studies [18–22] and our preliminary evaluation of various window sizes, the window size of 13 ($n = 6$) is found to be optimal in the identification of SUMOylation sites. Therefore, the training dataset consisted of 5890 positive training sequences and 15260 negative training sequences. However, as some negative data (non-SUMO data) may be identical to positive data (SUMO data) the predictive model's performance may be overestimated in both the training and testing datasets. To prevent this, we applied the CD-HIT program [23] to remove homologous data. After filtering out

sequences with 40% [7, 11] sequence identity, the training dataset contained 4985 positive training sequences and 9967 negative training sequences (Table 2).

To generate data for independent testing, the positive and negative independent testing datasets are constructed using the same approach as applied to the training dataset. As a result, the independent testing dataset contains 1245 positive and 2870 negative data (Table 2).

Table 2: Training dataset and testing dataset to use in this study

	Positive sites	Negative sites	Total
Training dataset	4985	9967	14952
Testing dataset	1245	2870	4115

2.2. Feature extraction and encoding

In comparison with traditional machine learning and statistical computation methods, the deep learning approach can automatically extract features from amino acid sequences, eliminating the need for manual feature engineering. Therefore, it is important to convert protein peptide sequences into quantification vectors for the application of deep learning-based models.

In this study, we utilize an embedding encoding technique built from the Word2Vec model [24]. Word2Vec, a widely used word embedding model in natural language processing (NLP) developed by Tomas Mikolov at Google, is not an individual algorithm, rather it comprises two learning models: Skip-gram and Continuous Bag of Words (CBOW).

By inputting text data into one of these learning models, Word2Vec generates word vectors that can represent a large piece of text or even the entire article. In NLP-based encoding techniques, words in a sentence are treated as real numbers. In our case, each protein is treated as a sentence and its residues as words. The protein sequences are represented as a collection of counts of n-grams, in which n adjacent amino acids are recognized as words. Inspired by the idea of Hamid and Friedberg [25], the length of the grams of 1, 2, and 3 have been tested in our work. The results show that the $n = 3$ appeared to be optimal in this evaluation, leading to $213 = 9261$ trigrams (20 common types of amino acids forming a protein and a special 'X' wildcard to replace missing amino acid in the fragment peptide). Figure 1 shows the representation learning for trigrams with skip-gram training. For each protein sequence, we create two sequences by starting the sequence from the first and second amino acids so that we can consider all the overlapping bigrams for a protein sequence. We generate the training instances using a context window of size ± 5, where we take the central word as input and use the surrounding words within the context window as outputs. The neural network architecture for training is used on all instances, and then a 9261-dimensional vector for each trigram is generated by the neural network. The trained hidden layer weights are transferred as the initial parameters of the embedding layer in the proposed SUMOylation site prediction model. The embedding matrix of Word2Vec acts as an embedding neural network. For details, see the data preprocessing phase in Figure 1.

2.3. Model construction, learning and evaluation

To build a model for the prediction of protein SUMOylation sites, we do the following steps:

Firstly, we used two different deep learning algorithms to develop a classifier for predicting SUMOylation sites. As depicted in Figure 1, our proposed method takes the raw protein fragments as input. Through the process of feature extraction and encoding, we create an embedding matrix Word2Vec. This matrix is embedded in the first layer of the deep-learning networks.

Secondly, we employ two parallel deep learning models. Each model learns its own set of features. Model 1 utilizes a CNN network with two convolutional layers and two max pooling layers. Model 2 employs an LSTM (128) network commonly used in natural language processing.

Finally, these two models are combined and then fed into a dense network with 64 nodes. The output layer contains a single neuron and ends with sigmoid activation to calculate the output x of this layer.

During the training of the CLW-SUMO models, the dropout units (the drop rate is set to 0.3) are added after each max pooling layer in the convolutional layer and before and after the LSTM (128) layer, which are usually required for generalization on unseen data and to avoid overfitting.

The proposed model is achieved through the Keras framework under the Python language. In an attempt to improve training speeds in deep neural networks and reach convergence quickly, we have utilized the Adam optimization method with an initial learning rate of 0.001, a batch size of 128, the number of epochs of 100. The early stop mechanism is applied to each training. All the experiments have been performed on a server equipped with Google Colab Pro.

In order to evaluate the performance of the predictive models, the 10-fold cross-validation approach has been performed to assess the classifying power of the predictive models. The following measurements are commonly used to evaluate the performance of the constructed models: Sensitivity (SEN), Specificity (SPE), Accuracy (ACC), Matthew's Correlation Coefficient (MCC), $Recall$, and $F1-score$, the area under the curve (AUC).

$$SEN = \frac{TP}{TP+TN}, \tag{1}$$

$$SPE = \frac{TN}{TN+FP}, \tag{2}$$

$$ACC = \frac{TP+TN}{TP+FN+TN+FP}, \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{4}$$

$$Recall = \frac{TP}{TP+FN}, \tag{5}$$

$$Precision = \frac{TP}{TP+FP}, \tag{6}$$

$$F1-Score = 2\frac{Precision \times Recall}{Precision + Recall}. \tag{7}$$

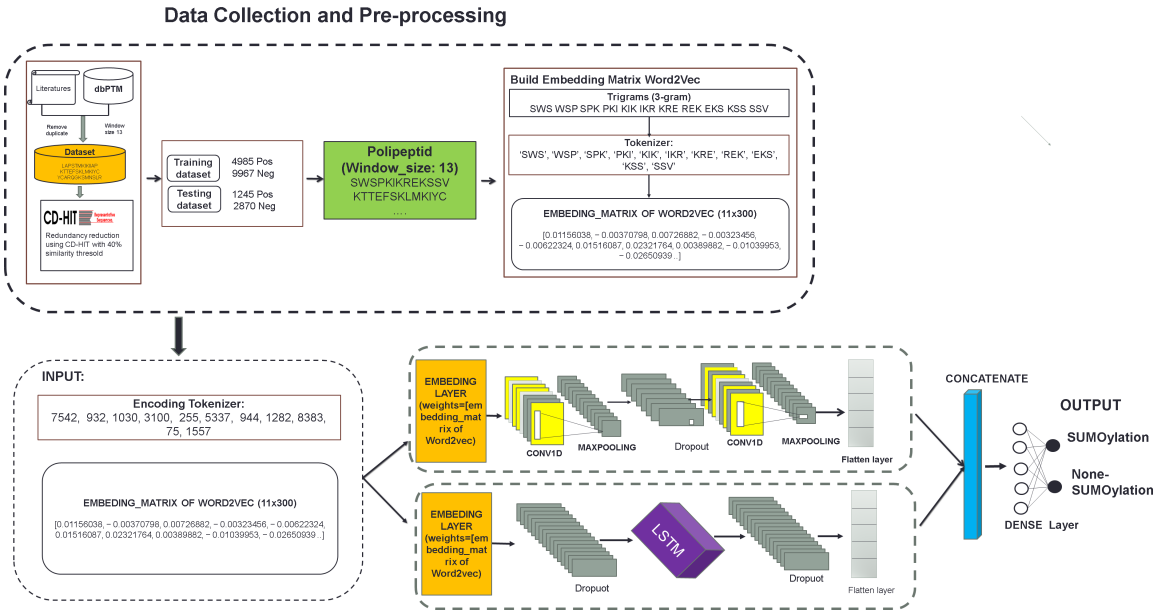

Figure 1: The architecture of the proposed model

Herein, TP, TN, FP, and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

In this study, we apply a k-fold cross-validation approach to evaluate the ability of the predictive models. After running a 10-fold cross-validation process, the predictive model with the highest values of MCC and accuracy is selected as the optimal model for identifying SUMOylation protein. Moreover, an independent testing approach was carried out to evaluate the ability of the selected model in the real case.

3. RESULTS AND DISCUSSIONS

3.1. Dataset analysis

To examine the position-specific amino acid composition for SUMOylation sites, WebLogo [26] is applied to generate the graphical sequence logo for the relative frequency of the corresponding amino acid at positions surrounding SUMOylation sites. Using WebLogo, the flanking sequences of substrate sites (at position 0) could be graphically visualized in the entropy plots of the sequence logo. Through the identified motif, we can easily observe the conservation of the amino acids around the SUMOylation sites. The identified motif is subsequently evaluated for its ability to distinguish SUMOylation from non-SUMOylation using 10-fold cross-validation.

The investigation of differences between the amino acid composition surrounding SUMOylation (positive data) and those of non-SUMOylation (negative data) shows that the overall trends are similar with slight variations. As shown in Figure 2a, the most four prominent amino acid residues include Glutamate (E), Serine (S), Lysine (K) and Leucine (L); whereas Tryptophan (W), (Cysteine C), and Methionine (M) are three of the least significant amino acid residues. Besides, the sequence logo displays the most enriched residues surrounding

the SUMOylation. As shown in Figure 2b, it also visualizes the most conserved amino acid residues including Glutamate (E), Leucine (L), Proline (P), Lysine (K), and Serine (S). In addition, Two sample logo [27] is used to visualize the difference between SUMO-sites and non-SUMO sites. As displayed in Figure 2c, the enriched residues appear to be Glutamate (E), Valine (V), and Lysine (K) while the depleted amino acid residues include arginine (R), Aspartic acid (D), and Leucine (L).

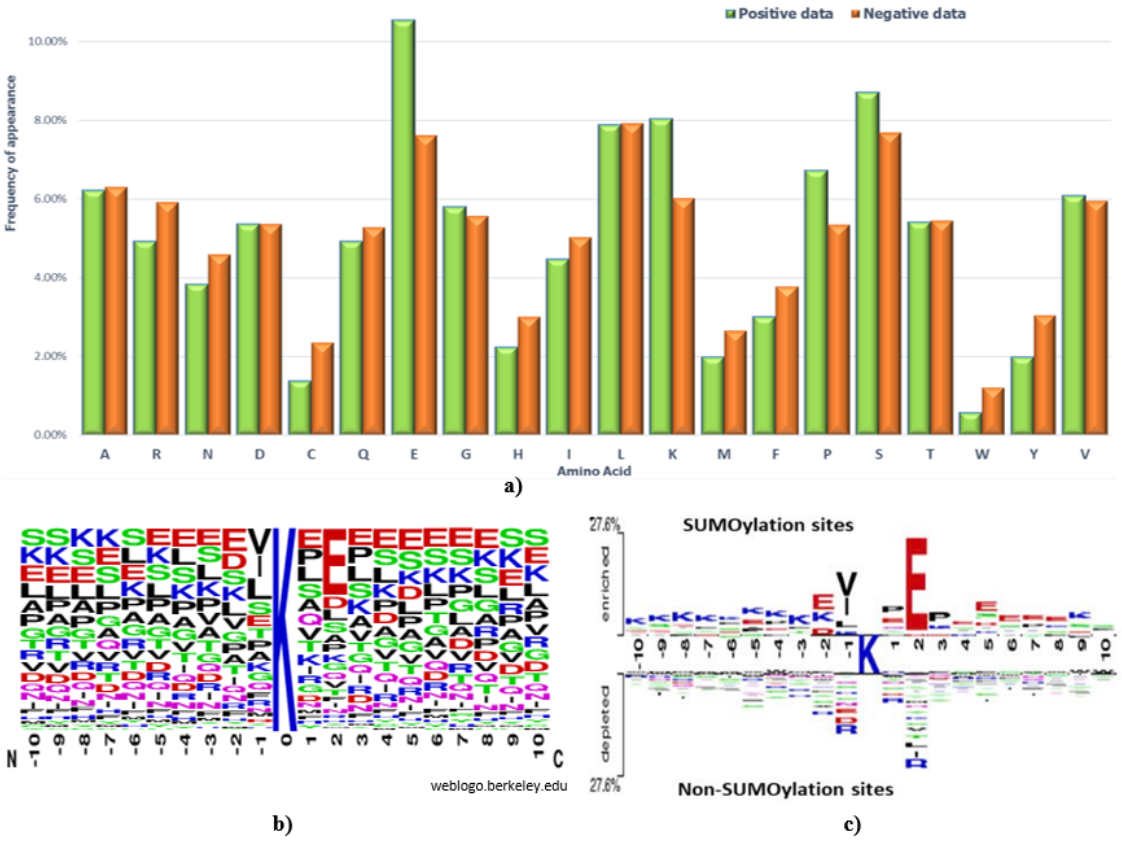

Figure 2: Frequency of the amino acid composition surrounding the SUMOylation sites (Figure 2a. The frequency of amino acids appeared in the positive training dataset; Figure 2b. The sequence logo - a graphical representation of the sequence conservation of amino acids in positive training dataaset); Figure 2c. Two sample logo - a visualization of differences between positive training and negative training dataset)

3.2. Performance evaluation by cross-validation approach

In Section 2, the 10-fold cross-validation approach is utilized to evaluate the ability of the predictive models trained on the Word2Vec embedding-based feature. To assess the performance of the predictive models, we have investigated the performance of single CNN and LSTM models. As displayed in Table 3, the CNN and LSTM models achieve quite high performance, obtaining accuracy of 81.30% and 80.31%, reaching MCC values of 0.564 and 0.537, respectively.

In general, it is straightforward and very beneficial to combine two or more different models to exploit advantages from them. Therefore, in an attempt to further improve the performance of the predictive model, we try to incorporate the power of the two separate CNN and LSTM models into a combined model, called CLW-SUMO, to identify protein SUMOylation sites. Fortunately, the proposed model, CLW-SUMO, achieves the highest performance, reaching 82.33% and 0.589 on *ACC* and *MCC*, respectively (Table 3). These results indicate that the proposed model of CLW-SUMO is highly effective for the prediction of protein SUMOylation.

Table 3: Performance evaluation by 10-fold cross-validation

Model	ACC(%)	SEN(%)	SPE(%)	MCC	Recall	Precision	F1-score
CNN	81.30	77.9	82.5	0.564	0.779	0.613	0.686
LSTM	80.31	80.3	80.3	0.537	0.803	0.542	0.647
CLW-SUMO	**82.33**	**79.7**	**83.3**	**0.589**	**0.797**	**0.631**	**0.704**

3.3. Performance evaluation by Independent testing approach

Cross-validation is a technique that combines the results from multiple local models to validate the global model, hence it cannot guarantee the ability of the model in the real case. Therefore, it is necessary to use an independent dataset, which is distinct from the training dataset, to examine the power and general ability of the model in the real case. Table 4 presents the detailed performance of the proposed model using the independent testing dataset. Fortunately, our proposed model obtains the highest performance with an accuracy reaching 90.03% and *MCC* value of 0.773. This result indicates that our proposed model of CLW-SUMO is a promising approach and could provide effective support for the prediction of protein SUMOylation.

Table 4: Performance evaluation by Independent testing

Model	ACC(%)	SEN(%)	SPE(%)	MCC	Recall	Precision	F1-score
CNN	88.89	85.5	90.5	0.747	0.855	0.803	0.828
LSTM	87.18	84.0	88.6	0.706	0.840	0.761	0.798
CLW-SUMO	**90.03**	**88.0**	**90.9**	**0.773**	**0.880**	**0.812**	**0.844**

3.4. Performance comparison with previous existing predictors

To evaluate the performance and practicality of the proposed model, in this study, we perform the comparison between our proposed model and several previous existing predictors using the same independent testing dataset. As displayed in Table 5, our proposed model is compared with five previous predictors (GPS-SUMO2.0 [4], seeSUMO2.0 [17], JASSA [14], ResSUMO [11], and RXS-SUMO [22]) and the result shows that our proposed model outperforms others, reaching accuracy at 90.03% and *MCC* value at 0.773. This strongly convinces us that our proposed model of CLW-SUMO is a promising approach and could provide effective support for the prediction of protein SUMOylation.

Table 5: Independent dataset comparison of CLW-SUMO with existing predictors

Predictor	Threshold	ACC(%)	SEN(%)	SPE(%)	MCC
GPS-SUMO2.0	Low	87.74	88.42	87.5	0.715
	Medium	79.39	69.37	83.87	0.525
	High	87.74	88.42	87.5	0.715
seeSUMO2.0	Low	85.52	82.83	86.54	0.661
	Medium	76.88	64.41	82.99	0.475
	High	83.57	79.0	85.33	0.615
JASSA	Low	76.05	73.92	76.31	0.343
	Medium	84.36	50.03	84.02	0.422
	High	87.04	45.32	86.42	0.423
	Very High	89.02	20.91	88.51	0.388
ResSUMO	-	72.67	80.91	64.9	0.456
RXS-SUMO	-	88.60	83.48	91.0	0.743
CLW-SUMO	-	**90.03**	**88.0**	**90.9**	**0.773**

4. CONCLUSION

Protein SUMOylation is one of the most important post-translational modifications in Eukaryotes species and plays significant roles in many biological processes. The mechanism underlined the SUMOylation process will be an important cause leading to many common serious diseases, such as breast cancer, cardiac, Parkinson's, Alzheimer's disease, etc.. In this study, we propose a novel approach for the prediction of protein SUMOylation sites using a hybrid deep learning model that combines convolutional neural networks (CNN) and long short-term memory (LSTM), using Word2Vec as the word embedding technique. Experimental results demonstrate the ability and power of our proposed model in the prediction of protein SUMOylation sites. We hope that our findings will provide effective suggestions and support to researchers in their studies related to the determination of protein SUMOylation sites.

ACKNOWLEDGMENT

The authors sincerely thank TUEBA for partly financially supporting this research under the TNU-level protject ID: DH2023-TN08-05.

REFERENCES

[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] R. T. Hay, "SUMO: a history of modification," *Molecular Cell*, vol. 18, no. 1, pp. 1–12, 2005.

[3] S. Müller, C. Hoege, G. Pyrowolakis, and S. Jentsch, "SUMO, ubiquitin's mysterious cousin," *Nature Reviews Molecular Cell Biology*, vol. 2, no. 3, pp. 202–210, 2001.

[4] Q. Zhao, Y. Xie, Y. Zheng, S. Jiang, W. Liu, W. Mu, Z. Liu, Y. Zhao, Y. Xue, and J. Ren, "GPS-

SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs," *Nucleic Acids Research*, vol. 42, no. W1, pp. W325–W330, 2014.

[5] Y. Xue, F. Zhou, C. Fu, Y. Xu, and X. Yao, "SUMOsp: a web server for sumoylation site prediction," *Nucleic Acids Research*, vol. 34, no. suppl_2, pp. W254–W257, 2006.

[6] J. Ren, X. Gao, C. Jin, M. Zhu, X. Wang, A. Shaw, L. Wen, X. Yao, and Y. Xue, "Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0," *Proteomics*, vol. 9, no. 12, pp. 3409–3412, 2009.

[7] Y. Gou, D. Liu, M. Chen, Y. Wei, X. Huang, C. Han, Z. Feng, C. Zhang, T. Lu, D. Peng *et al.*, "GPS-SUMO 2.0: an updated online service for the prediction of sumoylation sites and sumo-interacting motifs," *Nucleic Acids Research*, p. gkae346, 2024.

[8] 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.

[9] Y. Qian, S. Ye, Y. Zhang, and J. Zhang, "SUMO-Forest: a cascade forest based method for the prediction of SUMOylation sites on imbalanced data," *Gene*, vol. 741, p. 144536, 2020.

[10] Y. Lopez, A. Dehzangi, H. M. Reddy, and A. Sharma, "C-iSUMO: a sumoylation site predictor that incorporates intrinsic characteristics of amino acid sequences," *Computational Biology and Chemistry*, vol. 87, p. 107235, 2020.

[11] Y. Zhu, Y. Liu, Y. Chen, and L. Li, "ResSUMO: A deep learning architecture based on residual structure for prediction of lysine sumoylation sites," *Cells*, vol. 11, no. 17, p. 2646, 2022.

[12] H. Lv, F.-Y. Dao, H. Zulfiqar, and H. Lin, "DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach," *Briefings in Bioinformatics*, vol. 22, no. 6, p. bbab244, 2021.

[13] A. Sharma, A. Lysenko, Y. López, A. Dehzangi, R. Sharma, H. Reddy, A. Sattar, and T. Tsunoda, "HseSUMO: Sumoylation site prediction using half-sphere exposures of amino acids residues," *BMC Genomics*, vol. 19, pp. 1–7, 2019.

[14] G. Beauclair, A. Bridier-Nahmias, J.-F. Zagury, A. Saïb, and A. Zamborlini, "ASSA: a comprehensive tool for prediction of SUMOylation sites and sims," *Bioinformatics*, vol. 31, no. 21, pp. 3483–3491, 2015.

[15] Y.-Z. Chen, Z. Chen, Y.-A. Gong, and G. Ying, "SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties," *PloS One*, vol. 7, no. 6, p. e39195, 2012.

[16] C.-T. Lu, K.-Y. Huang, M.-G. Su, T.-Y. Lee, N. A. Bretana, W.-C. Chang, Y.-J. Chen, Y.-J. Chen, and H.-D. Huang, "DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications," *Nucleic Acids Research*, vol. 41, no. D1, pp. D295–D305, 2013.

[17] S. Teng, H. Luo, and L. Wang, "Predicting protein sumoylation sites from sequence features," *Amino Acids*, vol. 43, pp. 447–455, 2012.

[18] V.-N. Nguyen, H.-K. Do, T.-X. Tran, N.-Q.-K. Le, A.-T. Le, and T.-Y. Lee, "Exploiting two-layer support vector machine to predict protein sumoylation sites," in *Advances in Engineering Research and Application: Proceedings of the International Conference, ICERA 2018*. Springer, 2019, pp. 324–332.

[19] V.-N. Nguyen, K.-Y. Huang, C.-H. Huang, T.-H. Chang, N. A. Bretaña, K. R. Lai, J. T.-Y. Weng, and T.-Y. Lee, "Characterization and identification of ubiquitin conjugation sites with e3 ligase recognition specificities," in *BMC Bioinformatics*, vol. 16. Springer, 2015, pp. 1–11.

[20] V.-N. Nguyen, K.-Y. Huang, C.-H. Huang, K. R. Lai, and T.-Y. Lee, "A new scheme to characterize and identify protein ubiquitination sites," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 2, pp. 393–403, 2016.

[21] V.-N. Nguyen, H.-M. Nguyen, and T.-X. Tran, "An approach by exploiting support vector machine to characterize and identify protein sumoylation sites," *The 20th National Symposium of Selected ICT problems*, 2017.

[22] T.-X. Tran, V.-N. Nguyen, and N. Q. K. Le, "Incorporating natural language-based and sequence-based features to predict protein sumoylation sites," in *Conference on Information Technology and its Applications*. Springer, 2023, pp. 74–88.

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

[24] T. Mikolov, "Efficient estimation of word representations in vector space," *ArXiv Preprint ArXiv:1301.3781*, 2013.

[25] H. Fu, Y. Yang, X. Wang, H. Wang, and Y. Xu, "DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins," *BMC Bioinformatics*, vol. 20, pp. 1–10, 2019.

[26] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[27] V. Vacic, L. M. Iakoucheva, and P. Radivojac, "Two sample logo: a graphical representation of the differences between two sets of sequence alignments," *Bioinformatics*, vol. 22, no. 12, pp. 1536–1537, 2006.

Received on December 11, 2023
Accepted on September 22, 2024