A Multimodal Deep Architecture for Large-Scale Protein Ubiquitylation Site Prediction

1st Fei He
School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University
Changchun, China
hef740@nenu.edu.cn

2nd Lingling Bao
School of Information Science and Technology
Northeast Normal University
Changchun, China
baoll601@nenu.edu.cn

3rd Rui Wang
School of Information Science and Technology
Northeast Normal University
Changchun, China
wangr921@nenu.edu.cn

4th Jiagen Li
School of Information Science and Technology
Northeast Normal University
Changchun, China
lijg803@nenu.edu.cn

5th Dong Xu
Department of Electrical Engineering and Computer Science
Christopher S. Bond Life Sciences Center
University of Missouri
Columbia, MO 65211, USA
xudong@missouri.edu

6th Xiaowei Zhao*
School of Information Science and Technology
Institute of Computational Biology
Northeast Normal University
Changchun, China
zhaoxw303@nenu.edu.cn

***Abstract*—In eukaryotes, protein ubiquitylation is an important type of post-translation modification, in which the ubiquitin conjugates to a substrate protein. To have a better insight of the mechanisms underlying ubiquitylation, a key step is to identify protein ubiquitylation sites. Many existing computational methods are based on feature engineering, which may lead to biased and incomplete features. Deep learning provides multiple-layer networks and non-linear mapping operations to detect potential complex patterns in a data-driven way, especially for large-scale data. It provides a promising new method to predict ubiquitylation sites. In this paper, we proposed a multimodal deep architecture for protein ubiquitylation sites prediction. First, we designed different multiple layers to extract hidden informative patterns from three modalities, namely protein fragments, physico-chemical properties, and sequence profiles. Then, the deep representations corresponding to three modalities were merged to implement the classification. On the available largest scale protein ubiquitylation site database PLMD, the performance of our proposed method was measured with 66.7% sensitivity, 66.4% specificity, 66.43% accuracy, and 0.221 MCC value. A range of comparative experiments also showed that our proposed architecture outperformed several popular protein ubiquitylation site prediction tools. Our source code is freely available at https://github.com/jiagenlee/deepUbiquitylation.**

Keywords—Protein Ubiquitylation Site Prediction, Multiple Modalities, Deep Learning, Convolution Neural Network, Deep Neural Network

I. INTRODUCTION

Ubiquitin is a small protein consists of 76 amino acids [1, 2]. The conjugation of ubiquitin to a substrate protein on a particular lysine is an important post-translation modification, called ubiquitylation [3, 4]. Protein ubiquitylation involves three types of enzymes (activating enzymes, ligases and conjugating enzymes). It plays an important role in various cellular functions, such as signal transduction, apoptosis and cell proliferation [5, 6]. The conventional experimental techniques such as CHIP-CHIP analysis and mass spectrometry are usually time-consuming, laborious and expensive to detect protein ubiquitylation sites. Thus, the computational approaches that could effectively and accurately identify protein ubiquitylation sites are urgently needed.

Many computational methods have been developed for the identification of protein ubiquitylation sites. Huang et al. established a predictor called UbiSite, which used a two-layer machine learning method with substrate motifs to predict protein ubiquitylation sites [7]. Nguyen et al. proposed a scheme to characterize and identify protein ubiquitylation sites using three features including amino acid composition, evolutionary information and amino acid pair composition. Additionally, the motif discovery tool, MDDLogo, was also used in their predictor [8]. Qiu et al. selected sequence evolutionary information and gray system model to construct a protein ubiquitylation site predictor named iUbiq-Lys [9]. UbiProber presented by Chen et al. is another protein ubiquitylation site prediction tool, which combined sequence information, physico-chemical properties and amino acid composition to Support Vector Machine (SVM) for identifying potential protein ubiquitylation sites [10]. Wang et al. developed an improved protein ubiquitylation site predictor named ESA-UbiSite using an evolutionary screening algorithm (ESA) [11].

These existing machine-learning approaches perform effectively on small-scale data; however, some shared challenges on large-scale protein ubiquitylation site prediction still need to address [12]: (1) Weakness of handcrafted protein features. The traditional feature engineering relied on expert knowledge usually leads to biased and incomplete feature vectors. (2) Heterogeneity among different shallow representations. Most protein ubiquitylation site prediction tools combined multimodal features to improve their accuracies, but neglect the intrinsic heterogeneity among such shallow representations. (3) Unbalanced distributions between positive and negative samples [13]. Only a small number of lysine residues can be attached to ubiquitin in the whole proteome, and existing methods cannot function well to accurately identify potential protein ubiquitylation sites under such an extreme unbalanced circumstance. Deep learning, as a cutting-edge machine learning technique for big data, has been

* Xiaowei Zhao is the corresponding author

considered promising to tackle these problems. It provides multiple-layer networks and non-linear mapping operations to detect potential complex patterns from raw input signals, and generates homogenous deep representations for classification tasks [12]. The deep-learning framework simultaneously generates novel features and conducts the classification according to the input raw signals in a data-driven way, which could avoid some biases from feature engineering and reduce the mismatch between feature extraction and classifier. Diverse types of deep-learning networks have been successfully utilized to genomic and proteomic analyses and researches [14-16]; however, there has been no report on applying deep-learning technique to protein ubiquitylation site prediction.

In this paper, we proposed a multimodal deep architecture fusing three different categories of protein modalities for large-scale protein ubiquitylation site prediction, i.e. raw protein sequence fragment, selected physico-chemical properties of amino acids, and its corresponding position-specific scoring matrix (PSSM). In the deep architecture, we employed multiple convolution layers as the feature extractor to generate protein sequence representations, and brought several stacked fully connected layers to combine the physico-chemical properties of amino acids, and used other multiple convolution layers as a detector to discover the evolutionary profile around the potential ubiquitylation site. These multiple modalities were transformed into more compatible and abstract representations by our deep architecture. Finally, we integrated these hidden layers in the network to a softmax layer for predicting protein ubiquitylation sites. To the best of our knowledge, this is the first deep architecture for identifying protein ubiquitylation sites. In the comparisons with several recent state-of-the-art protein ubiquitylation site prediction tools, our approach exhibits more encouraging performance.

II. Material and Methods

A. Large-Scale Dataset Collection

For large-scale protein ubiquitylation site prediction, we collected 25,103 proteins with 12,1742 ubiquitylated sites from version 3.0 of Protein Lysine Modification Database (PLMD), which is a comprehensive dataset for 20 types of protein lysine modifications, and was extended from the CPLA 1.0 dataset and CPLM 2.0 dataset. As far as we know, PLMD is the available largest scale protein ubiquitylation site database, and is never mentioned in any other protein ubiquitylation site prediction research. In order to avoid over-estimation caused by homologous sequences, we used CD-HIT program [17] to filter the homologous sequences with 40% sequence similarity in all data, and obtained 17,406 proteins with 60,879 annotated protein ubiquitylation sites. These protein sequences were divided into the training dataset and the testing dataset by a random partition. The training dataset comprised 12,100 protein sequences with 54,586 ubiquitylation sites while the independent testing dataset consisted of 1345 proteins with 6293 ubiquitylation sites. According to these annotated information and protein sequences, we extracted 427,305 and 46,080 non-annotated ubiquitylation sites regarded as negative samples from the training dataset and the independent testing dataset, respectively. To construct the training and testing samples, we intercepted a protein fragment with the central lysine residue and fixed the window length of $2n+1$ for considering n upstream and downstream flanking amino acids around the targeting lysine residue as a sample. Furthermore, to prevent the interference that some negative training samples may be homologous to positive training samples, the tool CD-HIT-2D was utilized to remove the negative samples with 50% similarity to positive samples [7]. In order to achieve unbiased models, we extracted 30% of training samples as validation samples by random sampling in each training iteration. Finally, we obtained the experimental datasets as Table I summarized.

TABLE I. Brief description of collected protein ubiquitylation site data

Data set	Description			
	Number of sequences	***Number of positive data***	***Number of negative data***	***Note***
Training	12100	38211	224,059	Random partitioning in each training iteration
Validation		16375	96,024	
Testing	1345	6293	46,080	Reserved

B. Encoding of Protein Segments

In this paper, three types of quantized biological descriptors are employed to encode all involving protein samples.

1) One hot vector: each sample contained m amino acids is represented as a $m \times k$ 2-dimensional (2D) matrix, which uses a k dimensional zero vector with a one in the index corresponding to the amino acid in the protein sequence. When the left or right neighboring amino acids cannot fit the window size, a dash will be filled in these positions and be encoded to 0.05. In such encoded scheme, every protein fragment will be mapped to an exclusive and sparse coding, which quantifies amino acids and maintains their relative positions.

2) Physico-chemical Proverties: Some researches indicate that there is a strong connection between physico-chemical proverties of amino acids and ubiquitylation sites [13] [18]. And physico-chemical proverties have been widely used in many types of protein post-translation modification such as phosphorylation, acetylation and sulfation [10]. Such physico-chemical proverties of each amino acid can be found in an AAindex database [19]. Among the 544 physico-chemical metrics recorded in AAindex, we only selected top 13 physico-chemical proverties that have been validated by comparing the prediction accuracy of all physico-chemical proverties in light of the literature [10], and thus formed a $m \times 13$ 2D matrix as another encoding modality for each sample. The details of the selected physico-chemical properties are given in Table II.

TABLE II. The selected physico-chemical properties

Physico-chemical property	Description

Physico-chemical property	Description
EISD860102	Atom-based hydrophobic moment
ZIMJ680104	Isoelectric point
HUTJ700103	Entropy of formation
KARP850103	Flexibility parameter for two rigid neighbors
JANJ780101	Average accessible surface area
FAUJ880111	Positive charge
GUYH850104	Apparent partition energies calculated from Janin index
JANJ780103	Percentage of exposed residues
JANJ790102	Transfer free energy
PONP800102	Average gain in surrounding hydrophobicity
CORJ870101	NNEIG index
VINM940101	Normalized flexibility parameters, average
OOBM770101	Average non-bonded energy per atom

3) PSSM Profile: PSSM was also employed here to represent the evolutionary profile of the protein sequence. We set the non-redundant Swiss-Prot as the search database, and generate the raw PSSMs of all involving protein sequences using BLAST with the parameter "-j 3 -h 0.001"[12]. In a raw PSSM, it sets 20 dimensional vector to demonstrate the preference of 20 types of amino acids at each position of a protein seqence. For the purpose of focusing on the potential ubiquitylation sites, we extract the PSSM fragment corresponding to the window size *m* from the PSSM result of whole protein sequence which indicates the position-specific evolutionary profile of amino acids neighboring potential ubiquitylation sites. Thus, we obtain a *m*×20 2D matrix as PSSM modality.

C. Multimodal Deep Architecture Construction

Figure 1 presents the deep architecture, including three parts of sub-nets to separately deal with the above mentioned three kinds of input modalities, and then merge their output hidden states for classification.

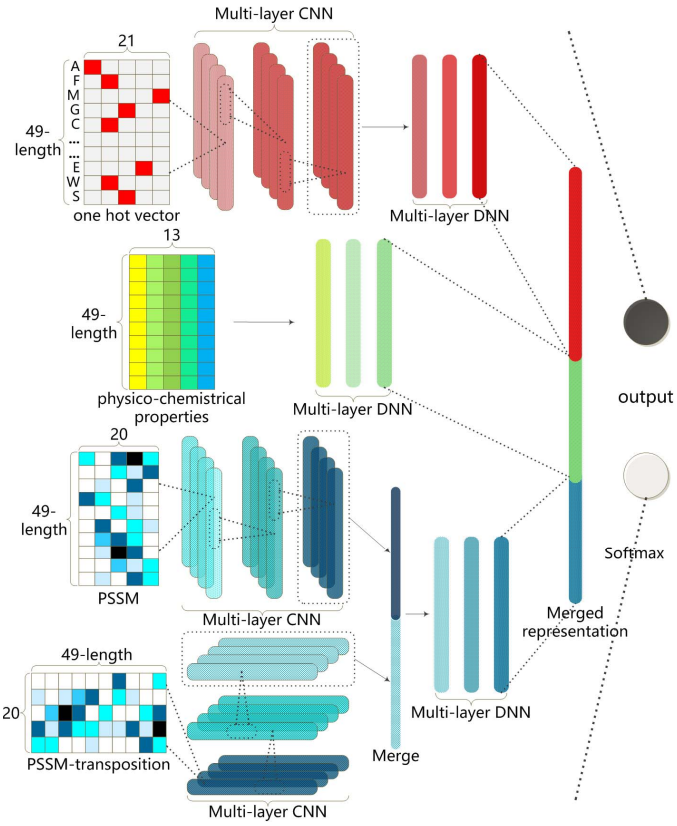

Fig. 1. The structure of our proposed deep architecture

For a given one hot vector, a one dimensional Convolution Neural Network (1D CNN) with three hidden layers is designed to extract its implicit local features. Due to the natural data sparsity of one hot vector, the locally connected convolution layers encode it into a range of feature maps, which represents subtle structural features hidden in raw protein sequence. After the hierarchical convolution, all feature maps are merged together and generated lower dimensional states by three fully connected hidden layers. Such a sub-net can detect informative sequential representations in its hidden states.

For the physico-chemical properties of corresponding amino acids, we introduced a Deep Neural Network (DNN) with three hidden layers to generate their deep representations. Since their components describe the characteristic of potential ubiquitylation sites from different viewpoints, the fully connected DNN structure could interconnect all factors for their joint effect in its hidden states.

For the input PSSM, we also employed 1D CNN with three hidden layers to detect potential informative descriptions among amino acids through evolution to the protein fragment. Different from the sub-net of one hot vector, the trans-positioned PSSM vector is then sent into another 1D CNN with three hidden layers to obtain deep evolutionary characterization among different sequence positions. The feature maps from the two 1D CNNs are merged together to produce the complete PSSM representations by three following fully connected hidden layers.

Subsequently, the output layer states of three sub-nets are merged into a mixed representation for fusing the three deep representations of input multiple modalities at the higher level. The merged layer is fully connected to a two-state output layer for implementing binary classification by a softmax function. These deep representations eliminate mutual heterogeneity among their raw shallow representations; therefore, they are more readily used for fusion. The weights between the merged layer and the output layer may be considered as the contributions of the three deep representations. The source code of our architecture is freely available at https://github.com/jiagenlee/deepUbiquitylation.

In this study, we introduce a training trick to accelerate the training procedure of the proposed multi-modal deep architecture [20]. Considering the multi-modal subnets, we separately trained each subnet to guarantee the optimality of their weights, and then reloaded these trained weights to the whole multi-modal deep architecture as its initialization. In the training process of the whole network, these weights and the weights of last merged layer would be fine tuned. Meanwhile, to eliminate the influence caused by the extremely unbalanced distribution of positive and negative samples, we implemented the training procedure of the whole deep architecture and subnets following the bootstrapping strategy. Let *pos* and *neg* denote the numbers of positive and negative samples respectively. Owing to the relatively small size of positive samples, we randomly chose *pos* negative samples to form a balanced training dataset with all positive samples in each bootstrapping iteration. Therefore, all negative samples were divided into $N = \lfloor neg / pos \rfloor$ bins, and the deep architecture is trained N times for modeling a classifier. Such a bootstrapping strategy can involve as many as training samples in classification model on the premise of unbiasedness [12]. The early stop rule [21] was adopted to control epoch numbers here, and the training procedure stops automatically by the time the validation accuracy has been stable for the default epoch iterations (we set 50 here).

We built this deep architecture using Theano 0.9 and keras 1.1.0, and ran on a graphic processing units (GPU) GTX1080Ti. Taking advantage of GPU computations, we can obtain a trained deep model in 180 minutes but predict general ubiquitylation sites of an unseen protein in seconds only.

III. Results and Discussion

A. Performance of our Mutimodal Deep Architecture

First, we would like to report our experiments of different window sizes of protein fragments. Owing to its direct effect on the available information involving in the prediction algorithm, the best suited window size to our deep architecture should be determined. Some researches adopted empirical value directly; however, different representations and classifiers prefer different window sizes [22]. Thus, we conducted a series of tests using the window length *m* from 7 to 61 (*n* is from 3 to 30). For each window length, we encoded all training protein fragments into three kinds of input modality and trained their corresponding subnet. The trained subnets were designated to predict the three types of input modality from the validation samples separately. The performance of different window sizes on one hot vector, physico-chemical properties and PSSM can be observed in Figure 2.

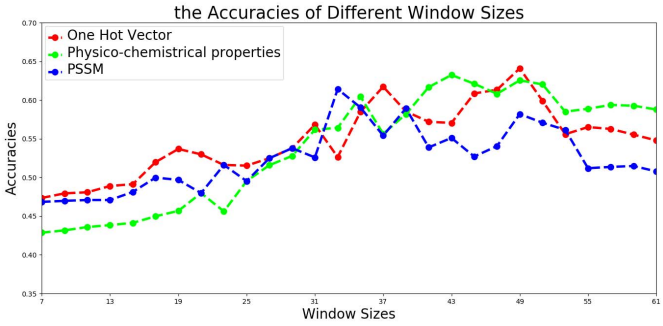

Fig. 2. The accuracy of validation samples using different window sizes on three modalities

Figure 2 suggests that when the window length is 49, the three types of representations achieved better or comparable accuracies to other window sizes. This conclusion was inconsistent with some existing studies [7, 10], which indicated that our deep architecture needed longer distance sequence fragments to introduce more raw information for further detecting deep features.

Next, the whole multi-modal network was trained using one hot vector, physico-chemical property and PSSM profile inputs simultanously. The trained whole network and trained subnets were tested on independent testing set. Their generative ROC (receiver operating characteristic) curves and precision-recall curves were plotted in Figure 3.

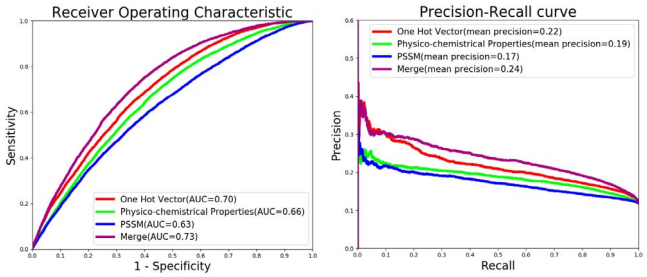

Fig. 3. ROC and precision-recall curves comparing between our multi-modal network and subnets of uni-modality

Figure 3 indicates some improvements of the whole multi-modal network on ROC and precision recall curves over subnets of uni-modality. Its AUC (area under the ROC curves) and mean precision (area under the precision-recall curves) reached 0.73 and 0.24, which might benefit from the data-driven combination way. In the training process of the whole deep architecture, we pre-loaded the weights of trained subnets to ensure that the subnets generate optimal deep representations of one hot vector, physico-chemical property and PSSM profile. And then, a supervised finetune was started to modify the weights of the merged layer adaptively. Such a process continued until the deep architecture made all input

modalities at full capacity. Figure 3 also manifests that one hot vector performed the best among the three input modalities. It can be inferred that a proper deep learning network may detect underlying informative expressions from raw protein sequence fragments.

We visualized the states in the original input layer and the merged layer of the whole model using t-SNE[23], to observe the overlap of positive samples and negative samples in the independent testing set. The visualization results are shown in Figure 4, where the positive samples and negative samples were in mixture, which is challenging for classification. As the samples were processed layer by layer in our deep architecture and the distinctive features were detected, and then the two classes of samples tended to separate. It implied that our proposed deep architecture may generate deep representations with more discriminative ability than input multiple modalities.

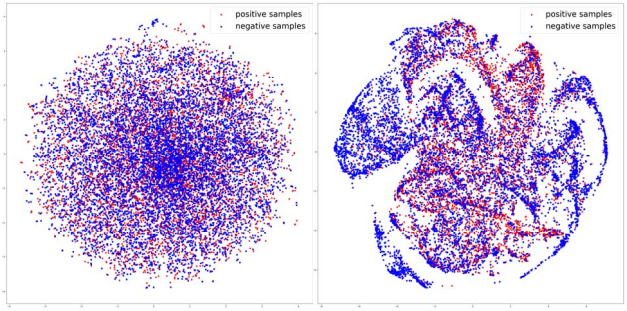

Fig. 4. t-SNE visualization of (a) input layers and (b) merged layer

B. *Comparisons with Other Protein Ubiquitylation Site Prediction Tools*

We compared our architecture with some popular protein ubiquitylation site prediction tools supporting batch sample mode, namely Ubisite [7], UbiProber [10], iUbiq-Lys [9], and ESA-UbiSite [11]. We submitted our testing protein sequences to their websites, and calculated all comparative metrics of involving tools as shown in Table V.

TABLE III. COMPARATIVE RESULTS WITH OTHER PROTEIN UBIQUITYLATION SITE PREDICTION TOOLS

Tool	Metrics			
	Accuracy	***Sensitivity***	***Specificity***	***MCC***
ESA-Ubisite	61.26%	46.14%	63.34%	0.064
UbiProber	55.06%	62.40%	54.05%	0.107
iUbiq-Lys	84.63%	3.35%	96.88%	0.005
Ubisite	73.63%	29.62%	79.64%	0.073
Our deep architecture	66.43%	66.67%	66.40%	0.221

In Table V, because the websites of UbiProber and iUbiq-Lys only returned the predicted decisions but not predicted scores, we computed their predicted metrics according to the classification results of the four tools for direct comparisons. From Table V, it can be found that our deep architecture performed excellent in most estimators, reaching 66.7% sensitivity, 66.4% specificity, 66.43% accuracy, and 0.221 MCC value with a 0.5 decision threshold. Even though the accuracy of our model cannot compare with Ubisite and iUbiq-Lys, their exorbitant specificities implied that they classified most of testing samples into non-ubiquitylation sites. That matched the unbalanced negative distribution of testing samples, and led to higher accuracy. However, they may not be as effective in predicting potential ubiquitylation sites according to their lower sensitivities. Overall, our model achieved simultaneous improvement in both sensitivity and specificity, especially obtained highest sensitivity among all tools. These demonstrated that our deep architecture was more effective and robust than the existing tools. Moreover, with the predicted scores Ubisite and ESA-Ubisite provided, we also plotted the ROC and precision-recall curves with AUC and mean precision of the two tools and our model in Figure 5.

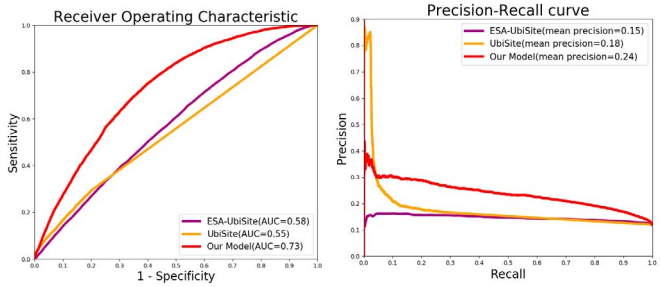

Fig. 5. The ROC and precision-recall curves comparing proposed deep architecture and two other protein ubiquitylation site prediction tools

Figure 5 exhibited that our deep architecture performed at a higher sensitivity under most certain specificities, and obtained better AUCs and mean precisions than other tools in most cases. It validated high confidence of our architecture on large-scale protein ubiquitylation site data. It is worth noting that under a certain minor recall, Ubisite achieved higher precision among the three methods,probably because Ubisite introduced more prior knowledge from positive training samples to its classification model. It divided positive training samples into 12 subgroups according to the clustered results of significant substrate motifs using the MDDLogo tool [24]. And then it trained 12 sub-models using the 12 subgroups of positive training samples and the same number of negative samples to implement a boosting classification. Such classification models emphasized the feature patterns of positive samples, and guided to detect potential homologous protein fragments with high similarity to its positive training samples. Consequently, it resulted in better precision than that of our deep architecture only when the recall was less than 3.89%. Nevertheless, our deep architecture has evident overall advantages in term of ROC and precision-recall carves.

Although our deep-learning architecture has promoted the performance of protein ubiquitylation site prediction on large scale data, there is still room for improvement. In the future, we would like to continue studying the optimization strategy for guiding the selection of deep learning hyper-parameters, and cooperate with biologists to upgrade the model more biologically interpretable and reliable.

IV. Conclusion

In this study, we proposed a multimodal deep architecture for large scale protein ubiquitylation sites prediction. Three kinds of modalities including one hot vector, physico-chemical properties and PSSM, which have been demonstrated to be associated with ubiquitylation, were firstly used to encode each input protein fragment. Then a multimodal deep architecture fusing these encoding modalities was established for robust classification. Experimental results on the available largest scale protein ubiquitylation site dataset have proved the effectiveness of the proposed method to deal with the large-scale data. The t-SNE visualization results also indicated that our deep architecture may generate more discriminative features from multiple modalities. The comparative experiments validated that our model outperformed several popular protein ubiquitylation site prediction tools. The success of our method is mainly due to the data-driven feature detection in deep learning, the multimodal fusion of deep representations, and the bootstrapping algorithm.

Acknowledgment

This research is partially supported by National Natural Science Foundation of China (61403077 and 61402098), the China Postdoctoral Science Foundation funded project (2015T80285), the Scientific and Technological Development Program of Jilin Province (20170520058JH), the Natural Science Foundation of the Education Department of Jilin Province (2016-505), and US National Institutes of Health grant R01-GM100701.

References

[1] G. Goldstein, M. Scheid, U. Hammerling, D. H. Schlesinger, H. D. Niall, and E. A. Boyse, "Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells," Proceedings of the National Academy of Sciences of the United States of America, vol. 72, no. 1, pp. 11, 1975.

[2] K. D. Wilkinson, "The Discovery of Ubiquitin-Dependent Proteolysis," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 43, pp. 15280, 2005.

[3] C. M. Pickart, and M. J. Eddins, "Pickart CM, Eddins MJ.. Ubiquitin: structures, functions, mechanisms. Biochim Biophys Acta 1695: 55-72," vol. 1695, no. 1-3, pp. 55-72, 2004.

[4] R. L. Welchman, C. Gordon, and R. J. Mayer, "Ubiquitin and ubiquitin-like proteins as multifunctional signals," Nature Reviews Molecular Cell Biology, vol. 6, no. 8, pp. 599, 2005.

[5] J. H. Hurley, S. Lee, and G. Prag, "Ubiquitin-binding domains," Biochemical Journal, vol. 6, no. 8, pp. 610, 2005.

[6] J. Peng, D. Schwartz, J. E. Elias, C. C. Thoreen, D. Cheng, G. Marsischky, J. Roelofs, D. Finley, and S. P. Gygi, "Peng, J. et al. A proteomic approach to understanding protein ubiquitination. Nature Biotech. 21, 921-926," Nature Biotechnology, vol. 21, no. 8, pp. 921-6, 2003.

[7] C. H. Huang, M. G. Su, H. J. Kao, J. H. Jhong, S. L. Weng, and T. Y. Lee, "UbiSite: incorporating two-layered machine learning method with substrate motifs to predict ubiquitin-conjugation site on lysines," Bmc Systems Biology, vol. 10 Suppl 1, no. Suppl 1, pp. 6, 2016.

[8] 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 & Bioinformatics, vol. 14, no. 2, pp. 393-403, 2017.

[9] W. R. Qiu, X. Xiao, W. Z. Lin, and K. C. Chou, "iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model," Journal of Biomolecular Structure & Dynamics, vol. 33, no. 8, pp. 1731, 2015.

[10] X. Chen, J. D. Qiu, S. P. Shi, S. B. Suo, S. Y. Huang, and R. P. Liang, "Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites," Bioinformatics, vol. 29, no. 13, pp. 1614, 2013.

[11] J. R. Wang, W. L. Huang, M. J. Tsai, K. T. Hsu, H. L. Huang, and S. Y. Ho, "ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives," Bioinformatics, vol. 33, no. 5, pp. 661, 2017.

[12] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs."

[13] C. W. Tung, and S. Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," Bmc Bioinformatics, vol. 9, no. 1, pp. 310, 2008.

[14] H. Y. Xiong, B. Alipanahi, L. J. Lee, H. Bretschneider, D. Merico, R. K. C. Yuen, Y. Hua, S. Gueroussov, H. S. Najafabadi, and T. R. Hughes, "The human splicing code reveals new insights into the genetic determinants of disease," Science, vol. 347, no. 6218, pp. 1254806, 2015.

[15] J. Zhou, and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," Nature Methods, vol. 12, no. 10, pp. 931, 2015.

[16] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," Nature Biotechnology, vol. 33, no. 8, pp. 831, 2015.

[17] 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, 2010.

[18] P. Radivojac, V. Vacic, C. Haynes, R. R. Cocklin, A. Mohan, J. W. Heyen, M. G. Goebl, and L. M. Iakoucheva, "Identification, analysis, and prediction of protein ubiquitination sites," Proteins-structure Function & Bioinformatics, vol. 78, no. 2, pp. 365-380, 2010.

[19] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: Amino Acid Index Database," Nucleic Acids Research, vol. 27, no. 1, pp. 368, 1999.

[20] X. Pan, and H.-B. Shen, "RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach," BMC bioinformatics, vol. 18, no. 1, pp. 136, 2017.

[21] Y. Yao, L. Rosasco, and A. Caponnetto, "On Early Stopping in Gradient Descent Learning," Constructive Approximation, vol. 26, no. 2, pp. 289-315, 2007.

[22] T. Chun-Wei, "Prediction of pupylation sites using the composition of k-spaced amino acid pairs," Journal of Theoretical Biology, vol. 336, no. 25, pp. 11-17, 2013.

[23] V. D. M. Laurens, G. Hinton, and V. D. M. Hinton, Geoffrey, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 2605, pp. 2579-2605, 2008.

[24] T. Y. Lee, Z. Q. Lin, S. J. Hsieh, N. A. Bretaña, and C. T. Lu, "Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences," Bioinformatics, vol. 27, no. 13, pp. 1780, 2011.