International Journal of Statistical Sciences *ISSN 1683-5603*
Vol. 18, 2019, pp 65-76

In Silico Prediction of Protein Ubiquitination Sites by Using Binary Encoding on *Arabidopsis thaliana*

Md. Parvez Mosharaf[1], Fee Faysal Ahmed[1], Md. Mehedi Hassan[2], Samme Amena Tasmia[1] and Md. Nurul Haque Mollah[1*]

[1]Bioinformatics Lab., Department of Statistics, Rajshahi University, Rajshahi-6205, Bangladesh

[2]School of Life Sciences, the State and Key laboratory of Agrobiotechnology, The Chinese University of Hong Kong, N.T, Hong Kong

*Correspondence should be addressed to Md. Nurul Haque Mollah (mollah.stat.bio@ru.ac.bd)

[Received May 2, 2019; Revised July 29, 2019; Accepted November 15, 2019]

Abstract

Ubiquitination is one of the most important and significant protein post-translational modifications (PTMs), which can regulate the cellular functions. Therefore, identification of ubiquitination sites is an important task for understanding the cellular mechanisms based on ubiquitination. Several wet lab based experimental approaches are available for identifying ubiquitination sites in *Arabidopsis thaliana*. However, those experimental approaches are laborious, time consuming and costly. Dry lab based *in silico* prediction is an alternative and cost effective approach for identification of ubiquitination sites. In this paper, we proposed an *in silico* method for prediction of ubiquitination sites mapping on *A. thaliana* by using random forest classifier with binary encoding features, window size 25 and 1:1 ratio of positive and negative samples. We observed that the proposed prediction models perform better than the other candidate prediction models with both training and independent test sequence datasets. The proposed method achieves an AUC score 0.86 and 0.84 with the training and independent test dataset, respectively. The proposed model would be helpful computational resource in predicting ubiquitination sites mapping on Arabidopsis thaliana as well as others related species. .

Keywords: *Arabidopsis thaliana*, Ubiquitination site prediction, Binary encoding, Feature selection, Random forest classifier.

AMS Classification: 92B20.

1. Introduction

Ubiquitination is one of the most important and significant post-translational protein modifications of protein which can regulate the cellular functions. Ubiquitination (also known as ubiquitylation) is an enzymatic and post-translational modification (PTMs) process, in which ubiquitin (a small regulatory protein) is attached to a substrate protein (Welchman RL et al, 2005; Herrmann J et al, 2007; Tung CW et al, 2008). In the ubiquitination process, the small regulatory protein ubiquitin, either a single ubiquitin or chains of ubiquitin is bound to lysine (K) residues on the protein substrate. This process work in three steps, they are activation, conjugation and ligation which are performed by ubiquitin activating enzymes (E1s), ubiquitin conjugating enzymes (E2s), and ubiquitin ligases (E3s), respectively (Welchman RL et al, 2005; Herrmann J et al, 2007; Tung CW et al, 2008; Walsh I et al. 2014). It is known that protein post-translational modification on any cell is highly involved in lots of biological process and also intimately engaged with different kinds of diseases. As ubiquitination is one kind of post-translational modification that is why it plays a vital role in plants and animals. So ubiquitination is also very much related to various complex biological processes and diseases. Different kinds of significant regulatory functions and related diseases of ubiquitination have been found, such as hypersensitive response, proteasomal degradation and downregulation, DNA repair and transcription, signal transduction, and endocytosis and sorting, Alzheimers, infectious diseases, cancers etc, which are all important protein regulation functions in the biological processes (Tung CW et al, 2008; Walsh I et al. 2014; Kirkpatrick DS et al, 2005).

Due to ubiquitination's significant regulation roles in the biological system, extensive research has been conducted to further decipher the molecular mechanism of the ubiquitination process and its other regulatory roles in the biological processes. One of the initial and but challenging steps to understand more deeply about ubiquitination's molecular mechanism. For this purpose, various types of experimental methods have been devoted to purify ubiquitination proteins to determine ubiquitination sites, such as high-throughput Mass Spectrometry (MS) techniques (Kirkpatrick DS et al, 2005; Peng JM et al, 2003; Wagner SA et al, 2011; Xu G et al, 2010), ubiquitin antibodies and ubiquitin binding proteins (Xu G et al, 2010), and combinations of liquid chromatography

and mass spectrometry. However these experimental methods are very time-consuming, expensive and labor-intensive, because the ubiquitination process is dynamic, rapid and reversible (Walsh I et al. 2014). To reduce experiment cost and improve the effectiveness of ubiquitination site identification, different computational methods have been introduced and developed (Walsh I et al. 2014; Kirkpatrick DS et al, 2005; Chen Z et al, 2014; Hasan, M.M. et al, 2018) based on different classifier and encoding scheme. Nevertheless, there is still room for improvement in the performance of the predictors. On the other hand, there is no specific ubiquitination sites predictor yet for the model plant *A. thaliana.*

In this paper, we would like to introduce a new computational method for ubiquitination site prediction mapping on the model plant *A. thaliana*. The trial version of this article was presented in the "International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment" (Mosharaf et al. 2017) held on January, 2017. Here the updated and finalized version of the paper has been reported. We hope that it will be helpful to understand the biological implications for further decipher research about the ubiquitination of the model plant *A. thaliana*.

2. Materials and Methodology

2.1 Data description and computational pipeline

Experimentally validated 417 ubiquitinated protein sequences mapping on *A. thaliana*, were collected from the Uni-ProtKB/Swiss-Prot and NCBI protein sequence databases. The experimentally validated 522 lysine ubiquitinated sites were considered as positive samples (i.e. ubiquitinated sites), while all the remaining sites were considered as negative samples (i.e. non-ubiquitinated sites). We partitioned the protein sequence dataset into training and independent test datasets. The training dataset was consisted of 350 protein sequences, which contained 450 positive sites and around 4500 negative sites. On the other hand, the independent test dataset was consisted of 67 protein sequences, which contained 72 positive sites and around 700 negative sites. For each of positive and negative sites, we generated 4 windows of sizes 23, 25, 27 and 29 to select one of them for the improvement of prediction performance. Obviously, positive sample and negative samples were unbalanced, which leaded over estimation or under estimation of the parameters in the prediction model. To overcome these

problems, we considered 3 balanced datasets of ratios 1:1, 1:2 and 1:3 of positive and negative window samples. To convert each window sequence data into numeric data, we consider the binary encoding scheme. To develop a good computational method for prediction of ubiquitination sites, we trained RF classifier for each of ratio and window size conditions based on the training dataset. Then the best model was built by optimizing the performance scores (Sn, Sp, Ac, AUC). The working flowchart of this proposed prediction method is shown in Figure 1.

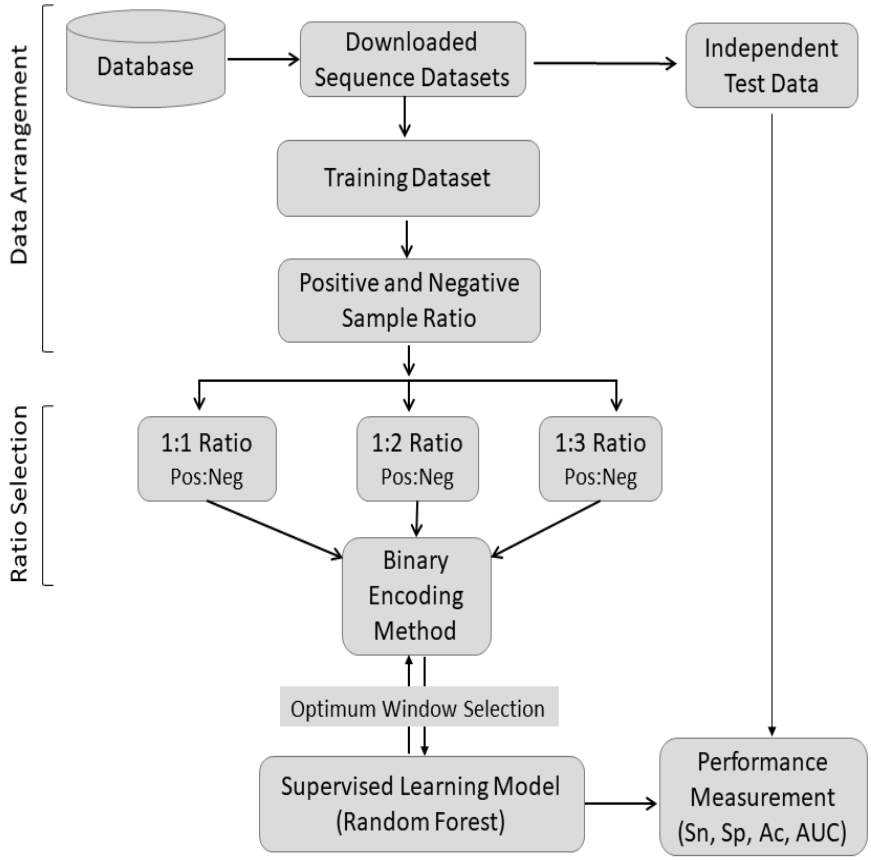

Figure 1: The working flowchart of *in silico* prediction method.

2.2 Sequence encoding

With the view of building an efficient prediction method, we encoded each sequence fragment into a numeric vector that was the ultimate step to present the classifier and ensemble architecture. Thus, for keeping the generated code compact in dimensionality a high-quality sequence encoding method was necessary. In this study, the binary sequence encodings was adopted, to represent the amino acid features.

2.3 Binary encoding

In order to make a robust predictor, binary amino acid encoding was considered to calculate the positional information from the corresponding sequence fragments. In this study, 21 (including gap (O)) amino acids were transformed into numeric vectors by adopting a binary vector. The 21 types of residues were ordered as ACDEFGHIKLMNPQRSTVWYO. For adopting binary vector, in query proteins, A was represented as 100000000000000000000 and C as 010000000000000000000, and so on. The selected window size of surrounding ubiquitinated sites was 25. For the query proteins of uubiquitination sites, the center position was always K. Thus, it was not considered to be taken into account. Finally, the feature vectors with a dimensionality (21×24) = 504. We're obtained from the binary encoding.

2.4 Random forest classifier

RF classifier is a collection of decision tree classifiers, wherein each tree is trained with a randomly selected subset of samples. The decision tree is grown as follows. Suppose N samples are randomly selected with replacement from the F features, then the best split node is selected from F features. Finally, the decision tree is grown as large as possible without pruning. In the construction of the forest, it is generalized based on most votes given by all the individual trees; within the post for the error estimate it does not produce bias. It is relatively robust to noise and outliers (Breiman L. 2001). As a supervised learning algorithm, it has been widely used in protein bioinformatics (Chen Z et al, 2014, Hassan MM et al, 2018, Hassan MM et al, 2018). The predicted result of the RF was decided by voting among the number of trees, which contains two classes, either positive samples

(ubiquitinated sites) or negative samples (non-ubiquitinated sites). In this study, the RF algorithm was implemented using the 'Random Forest' in Weka software.

2.5 Feature selection

As mentioned in ''Feature encoding'', each investigated ubiquitinated or non-ubiquitinated fragment was encoded into as a high dimensional vector. Therefore, they may not equally contribute to determine the surrounding ubiquitinated or non- ubiquitinated sites. To address this, we use a feature selection method (InfoGainAttributeEval) was adopted to distinguish them. Here we selected the first 1000 important features sequentially from 100, 200, 300, up to 1000. We analysed them and recorded their corresponding AUC scores. We observed that the AUC scores are almost nearest for each group of the features selected. It was an important issue for an ubiquitination sites prediction analysis.

2.6 Performance assessment measures

To observe the performance of the suggested ubiquitination sites predictor in silico method, we considered four widely used performance measures denoted as sensitivity (Sn), specificity (Sp), accuracy (Ac) and the Matthews correlation coefficient (MCC). To formulate these performance measures in our current context, let us consider a two-class prediction problem (binary classification), in which the outcomes (lysine ubiquitinated or non- ubiquitinated) are labelled either as positive (+) or negative (-). There are four possible outcomes from a binary classifier. If the predicted class is '+' and the actual value is also '+', then it is called a true positive (tp); however, if the actual value is '-', then it is said to be a false positive (fp). In contrast, a true negative (tn) is occurred when both the prediction outcome and the actual value are '-', and false negative (fn) is occurred when the prediction outcome is '-', while the actual value is '+'.

$$Accuracy\ (Ac) = \frac{tp + tn}{tp + fn + tn + fp}$$

$$Sensitivity\ (Sn) = \frac{tp}{tp + fn}$$

$$Specificity\ (Sp) = \frac{tn}{tn + fp}$$

$$Matthew\ correlation\ coefficient\ (MCC) = \frac{(tp \times tn) - (fp \times fn)}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fn) \times (tn + fp)}}$$

The values of all of these measurements lie between 0 and 1, and a higher value represents a better prediction. In addition, we also used the ROC (Receiver Operating Characteristics curve) and AUC (area under the ROC curve) measures to select the better predictor. An ROC space is defined by fp rate and tp rate as *x* and *y* axes, respectively, which depicts relative trade-offs between tp and fp. Since tp rate is equivalent to Sn and fp rate is equal to 1−Sp, the ROC graph is sometimes called the Sn vs (1-Sp) plot.

3. Result and Discussions

The performance of PTM site predictor depends on the data adequacy, ratio selection of positive and negative samples, and window size of each sample. Let us first discuss the adequacy of the datasets in the subsection 3.1, ratio selection in subsection 3.2, window selection in the subsection 3.3 and then the discussion of the proposed method based on random forest (RF) classifier with binary encoding features in subsection 3.3.

3.1 Adequacy of the dataset by two sample logo analysis

The two sample logo (Vacic V et al, 2006) of the ubiquitinated and non-ubiqutinated fragments of protein datasets (Figure 2) shows the amino acid residues combination around the lysine (k) residue. The two sample logo displayed a wide range of amino acids are surrounding around the centre position of the Lysine (K) residue. It has been observed that the Arginine (R) residues are the mostly surrounding the centre Lysine (K) residues in the dataset.

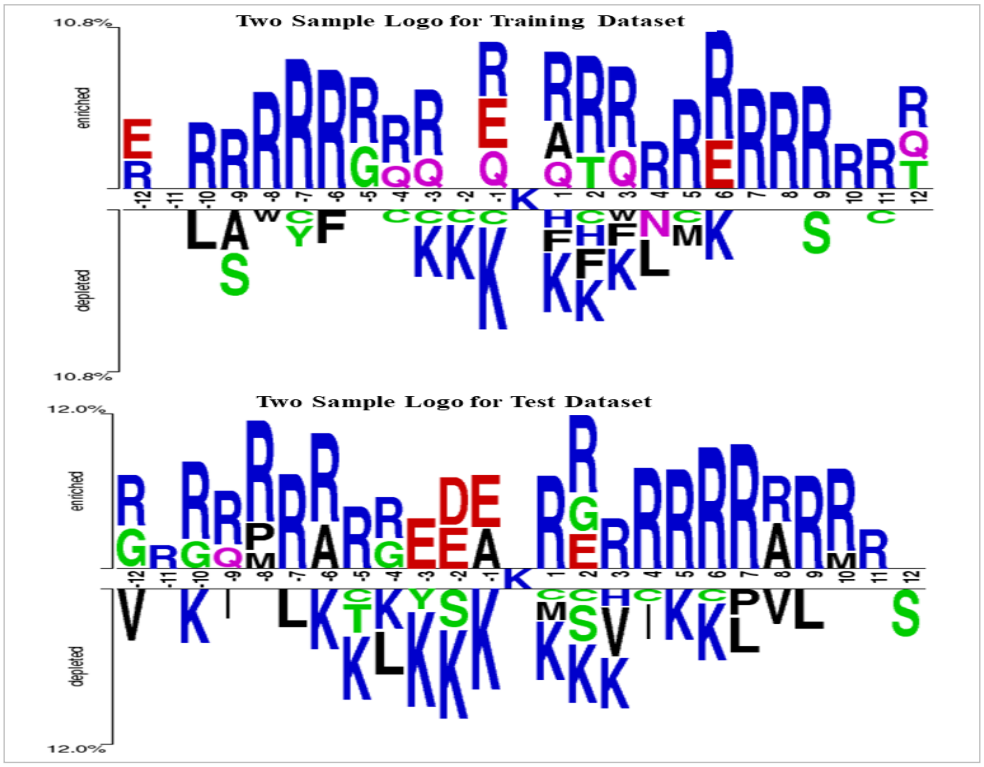

Figure 2: The amino acid propensities of surrounding ubiquitination sites compared to non-ubiquitination sites as displayed with the Two Sample Logos software. It also shows that the position between the compositional amino acids of the ubiquitinated and non-ubiquitinated peptides had a wide difference, especially those located in the positions from -12 to -1 and +1 to +12. Thus both training and independent test datasets are adequate for ubiquitination site prediction.

3.2 Optimum ratio selection to increase the prediction performance with the training dataset

In nature, the ubiquitination and non-ubiquitination datasets are highly unbalanced. The computational result's accuracy and efficiency are strongly affected due to the nature of the unbalanced datasets. To address this issue, many PTM site prediction studies employ a relatively balanced ratio between the positive and negative samples during the training of the dataset including the ubiquitination sites prediction as well (Chen Z et al, 2014, Hassan MM et al, 2018, Hassan MM et al, 2018). To select appropriate ratio to develop the predictor, we

computed different performance measures (Sn, Sp, Ac and AUC) with different ratios (1:1, 1:2, 1:3) of ubiquitination and non-ubiquitination peptides with different window sizes (23, 25, 27, 29) to develop a comparatively balanced training dataset. We observed that window size 25 produces higher performance scores with different ratio. For convenience of presentation, we displayed the performance score only for window size 25 with different ratios in table-1. From table-1, we observed that 1:1 ratio produces the highest performance scores 0.83, 0.85, 0.85, 0.86 and 0.70 of Sn, Sp, Ac, AUC and MCC, respectively. Therefore, ratio 1:1 was selected to develop the predictor.

Table 1: Performance comparison with different ratios of positive and negative samples

Ratio	Sn	Sp	Ac	AUC	MCC
1:1	**0.83**	**0.85**	**0.85**	**0.86**	**0.70**
1:2	0.81	0.85	0.83	0.84	0.63
1:3	0.80	0.85	0.81	0.82	0.60

3.3 Optimum window size selection to increase the prediction performance with the training dataset

Another issue is the optimal size of the sequence windows flanking the ubiquitination and non-ubiquitination sites. To select appropriate window size to develop the predictor fixing ratio at 1:1, we computed different performance measures (Sn, Sp, Ac and AUC) with different window sizes 23, 25, 27 and 29. We displayed the performance scores with different window sizes in table-2. From table-2, we observed that window size 25 produces the highest performance scores 0.83, 0.85, 0.85, 0.86 and 0.70 of Sn, Sp, Ac, AUC and MCC, respectively, as before. Therefore, window size 25 was selected to develop the predictor.

Table 2: Performance comparison with different window size.

Window size	Sn	Sp	Ac	AUC	MCC
23	0.82	0.85	0.81	0.85	0.62
25	**0.83**	**0.85**	**0.85**	**0.86**	**0.70**
27	0.81	0.85	0.81	0.85	0.62
29	0.80	0.85	0.83	0.84	0.50

3.4 Proposed prediction model

From the results in subsections 3.2 and 3.3 with the training dataset, our proposed prediction model consist of RF classifier with binary encoding features, 1:1 ratio of positive and negative samples with window size 25. Then we computed the performance scores with the independent test dataset. Table 3 shows the performance scores with independent test dataset. For convenience of comparison with the optimum results of training dataset, we also presented the performance scores of training dataset in table 3. Than we observed that the proposed predictor produces the performance scores 0.82, 0.85, 0.82, 0.84 and 0.67 of Sn, Sp, Ac, AUC and MCC, respectively, which is almost close to the performance with the training dataset. Thus the proposed predictor produces consistent results with both training and independent test datasets.

Table 3: Performance comparison with training and independent test datasets

Datasets	Sn	Sp	Ac	AUC	MCC
Test Data	0.82	0.85	0.82	0.84	0.67
Training Data	0.83	0.85	0.85	0.86	0.70

4. Conclusion

In this paper, we proposed a simple and efficient computational statistical method for prediction of ubiquitination sites mapping on the model plant *A. thaliana* by using random forest classifier with binary encoding features, window size 25 and 1:1 ratio of positive and negative samples. We observed that our proposed method performed better for both training and independent dataset. Moreover, we expect that our findings might be helpful for better understanding the important rules that underlie the ubiquitinated proteins. The data analysis results demonstrated that the proposed method might be helpful to understand ubiquitination as well as the mechanisms of protein ubiquitination. In our method we used R-programming, Perl programming, Weka software and web based software for calculation and analysis. Although our proposed method obtained a fairly good performance, there are still some spaces for improvement. In the future, we would like to pay more attention to make an organism specific prediction method or tool for improving the performance of ubiquitination sites prediction.

Acknowledgments: The authors thank and gratefully acknowledge the editor and referees for their valuable comments and positive critique that help us to improve the manuscript.

Reference

[1] The Nobel Prize in Chemistry (2004). Popular Information. Nobelprize.org. Nobel Media AB 2014. Available online at www.nobelprize.org/nobel_ prizes/chemistry/laureates/2004/popular.html. Accessed 26 Nov 2014.

[2] Welchman, R. L., Gordon, C. and Mayer, R. J. (2005). Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol. 6(8):599–609.

[3] Herrmann, J., Lerman, L. O. and Lerman, A. (2009). Ubiquitin and ubiquitin-like proteins inprotein regulation. Circ Res. 100(9):1276–91. [4] Bollerslev, T. (1986), Generalized autoregressive conditional heteroscedasticity. Journal of Econometrics, 31(3), 307-327.

[4] Tung, C. W. and Ho, S. Y. (2008). Computational identification of ubiquitylation sites from protein sequences. BMC Bioinformatics. 9:310.

[5] Walsh, I., Domenico, T. D., Tosatto, S. C. E. (2014). RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance. Amino Acids. 46:853–62.

[6] Kirkpatrick, D. S., Denison, C., Gygi, S. P. (2005). Weighing in on ubiquitin: the expanding role of massspectrometry-based proteomics. Nat Cell Biol. 7(8):750–7.

[7] Peng, J. M., Schwartz, D., Elias, J. E., Thoreen, C. C., Cheng, D., Marsischky, G., et al. (2003). A proteomics approach to understanding protein ubiquitination. Nat Biotechnol. 21:921–6.

[8] Wagner, S. A., Beli, P., Weinert, B. T., Nielsen, M. L., Cox, J., Mann, M. and Choudhary, C. (2011). A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics. 10(10):M111.013284.

[9] Xu, G., Paige, J. S. and Jaffrey, S. R. (2010). Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. Nat Biotechnol. 28:868–73.

[10] Chen, Z., Zhou, Y., Zhang, Z. and Song, J. (2014). Towards more accurate prediction of ubiquitination sites: a comprehensive review of current

methods, tools and features. Brief Bioinform, Advance Access, doi:10.1093/bib/bbu031

[11] Hasan, M. M. and Kurata, H. (2018). GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by Aggregating Multiple Sequence Features. PLoS One, 13(10): e0200283.

[12] Hasan, M. M., Khatun, M. S., Mollah, M. N. H., Yong, C. and Guo, D. (2018). NTyroSite: Computational Identification of Protein Nitrotyrosine Sites Using Sequence Evolutionary Features. Molecules 23, 1667.

[13] Breiman, L. (2001). "Random forests," Machine Learning, 45, 5-32.

[14] Walton, A., Stes, E., Cybulski, N., Van Bel, M. and Inigo, S. (2016). It's Time for Some "Site"-Seeing: Novel Tools to Monitor the Ubiquitin Landscape in Arabidopsis thaliana. The Plant Cell, 28 (1) 6-16

[15] Chen, Z., Chen, Y. Z., Wang, X. F., Wang, C., Yan, R. X., Zhang, Z. (2011). Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS One. 6(7):e22930. doi:10.1371/journal.pone.0022930

[16] Chen, K., Kurgan, L. A. and Ruan, J. (2007). Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs. BMC Struct Biol 7, 25 https://doi.org/10.1186/1472-6807-7-25

[17] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T. and Kanehisa M. (2008). AA-index: amino acid index database, progress report 2008. Nucleic Acids Res.;36(Database issue):D202-D205. doi:10.1093/nar/gkm998

[18] Vacic V., Iakoucheva L. M., and Radivojac P. (2006). "Two Sample Logo: A Graphical Representation of the Differences between Two Sets of Sequence Alignments." Bioinformatics, 22(12): 1536-1537.

[19] Mosharaf, M. P., Ahmed, F. F., Sutana, A., Reza, M. S., Ahmed, M. S., Khatun, M. S., Hasan, M. M., Ali, M. A. and Mollah. M. N. H. (2017). In silico prediction of protein ubiquitination sites mapping on Arabidopsis thaliana. Proceedings of International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment, Paper ID: 127, Pages: 586-593, 2017, ISBN: 978-984-34-0996-6, University of Rajshahi.