Carbonylation sites prediction in proteins using support vector machine

1st Subrota Sarkar
CSE
United International University
Dhaka, Bangladesh
subrotasarkar114@gmail.com

2nd Trisha Barman
CSE
United International University
Dhaka, Bangladesh
trishabarman12@gmail.com

***Abstract*—The carbonylation is found as an irreversible post-translational modification and considered a biomarker of oxidative stress. It plays major role not only in orchestrating various biological processes but also associated with some diseases such as Alzheimer's disease, diabetes, and Parkinson's disease. However, since the experimental technologies are costly and time-consuming to detect the carbonylation sites in proteins, an accurate computational method for predicting carbonylation sites is an urgent issue which can be useful for drug development. In this study, a novel computational tool termed predCar-Site has been developed to predict protein carbonylation sites by (1) incorporating the sequence-coupled information into the general pseudo amino acid composition, (2) balancing the effect of skewed training dataset by Different Error Costs method, and (3) constructing a predictor using support vector machine as classifier. This predCar-Site predictor achieves an average AUC (area under curve) score of 0.9959, 0.9999, 1, and 0.9997 in predicting the carbonylation sites of K, P, R, and T, respectively. All of the experimental results along with AUC are found from the average of 5 complete runs of the 10-fold crossvalidation and those results indicate significantly better performance than existing predictors. A userfriendly web server of predCar-Site is available at**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

troduction The structural and functional diversities of proteins as well as plasticity and dynamics of living cells are significantly dominated by the post-translational modifications (PTMs). Not only that, PTMs are also responsible for expanding the genetic code and for regulating cellular physiology as well.A variety of PTMs such as hydroxylation, nitration, sulfhydrylation, carbonylation and glutathionylation have been induced from Oxidative stress which is the direct result of imbalance in the production and degradation of reactive oxygen species (ROS) and reactive nitrogen species (RNS). Oxidative stress may occur when an excess production of reactive oxygen species (ROS) has surpassed the detoxification ability of cells and weakened the damage-repairing ability.

In this context, it is highly demanded to use computational approaches to identify the carbonylated sites effectively and accurately. Recently various types of computational classifiers have been developed to identify carbonylation sites through different types of machine learning algorithms.However, in order to meet the current demand to produce efficient high-throughput tools, additional effort are required to further improve the prediction quality. In the development of computational classifier, one of the major challenges is to handle imbalance dataset problem, as it is found in most of the dataset for this kind of prediction, the number negative subset is much larger than the corresponding positive subset. As the real world picture is that the noncarbonylation sites are always the majority compared with the carbonylation ones, so naturally the predictor should be biased to the non-carbonylation sites. Here the problem is that, for this type of predictors may interpret many carbonylation sites as noncarbonylation sites.But, the information about the carbonylation sites is mostly desired than non-carbonylation sites. As a result, it is crucial to find an effective solution to balance this kind of bias consequence.

II. Material and methods

A. Benchmark dataset

iCar-PseCp's benchmark dataset set has been used in this study. iCar-PseCp's dataset was derived from the 230 carbonylated protein sequences from human and 20 carbonylated protein sequences from Photobacterium and Escherichia coli.

III. Feature extraction

- The appropriate features of protein sequences or samples plays very important roles for the prediction of carbonylation site, as a result it draws much attention of scientist that how to select the core and essential features of protein samples. As most existing machine learning algorithm can handle only vector but not sequence sample, one of the critical problem in bioinformatics is how to extract vector from biological sequence with keeping considerable sequence characteristics. In this paper, to avoid completely losing the sequence pattern information of protein, the Chou's general PseAAC has been adopted to extract feature from peptide segment using sequence-coupling model which has been described briefly below.

IV. SVM CLASSIFICATION

V. IMBALANCE DATASET PROBLEM MANAGEMENT

VI. MEASURING METRICS

VII. EXPERIMENTAL SETTING

VIII. RESULTS AND DISCUSSION

A. Model selection for SVM

During the development of the project we researched the feasibility in different fields, especially software and hardware. The feasibility study is shown below.

B. Comparison with the existing methods

CONCLUSION

In this article, we have designed a simple and efficient predictor predCar-Site for predicting carbonylation sites. Experimental results show that our method is very promising and can be a useful tool for prediction of carbonylation sites. The predCar-Site has achieved remarkably higher success rates in comparison with the existing predictors in this area. In addition to it, we have established a user-friendly web server and provided a step-by-step guide for the convenience of the experimental scientists. It provides as easier way to obtain the desired results without knowing the mathematical details. We have projected that the predCar-Site will become a very useful and higher throughput tool for predicting of protein carbonylation sites

FUTURE SCOPE

-

- REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.