Crimson Publishers
Wings to the Research

Short Communication

Efficient Prediction of Protein Malanylation Sites Using NLP and Machine Learning

Hananeh Rajabiun, Mohammad Ghasemzadeh* and Masroor Hassan

Computer Engineering Department, Yazd University, Yazd, Iran

ISSN: 2832-4463

***Corresponding author:** Mohammad Ghasemzadeh, Computer Engineering Department, Yazd University, Yazd, Iran

Submission: May 22, 2023
Published: June 08, 2023

Volume 3- Issue 2

How to cite this article: Hananeh Rajabiun, Mohammad Ghasemzadeh* and Masroor Hassan. Efficient Prediction of Protein Malanylation Sites Using NLP and Machine Learning. COJ Rob Artificial Intel. 3(2). COJRA. 000558. 2023.
DOI: 10.31031/COJRA.2023.03.000558

Abstract

This research fills a scientific gap by addressing the challenge of identifying the site of Malanylation in proteins. It highlights the importance of efficient solutions that reduce execution time and improve output accuracy. The study introduces a novel framework for extracting informative features from protein functional domains. Multiple classifiers are utilized for prediction and experimental results indicate that the CRF-Mal method outperforms other approaches. Notably, the XG Boost classifier demonstrates superior performance compared to alternative classifiers.

Keywords: Malanylation; Machine learning; Natural language processing; Feature extraction

Introduction

Research in the field of protein function and Post-Translational Modifications (PTMs) has highlighted their crucial role in regulating biological processes. There are over 600 identified types of PTMs, including diverse chemical groups and peptides. Malanylation, a recently discovered PTM, plays a significant role in cellular processes and dynamic regulation. However, conventional methods for identifying alanylation sites encounter challenges. To address this, machine learning techniques are being employed to accurately predict these sites and reduce the reliance on time-consuming and labor-intensive experimental approaches [1]. In this research, various methods for predicting alanylation sites in proteins are discussed. These include the Mal-Lys [2] method, Malo Pred [3], feature extraction using sequence and structural features, a hybrid feature selection method and the use of machine learning classifiers such as random forest, support vector machine, K nearest neighbors, logistic regression and Light Gradient Boosting Machine. Additionally, deep learning approaches such as Deep Mal and Mal site-Deep have been utilized for accurate prediction [4,5]. The use of natural language processing techniques and protein domain information has also been proposed. The CRF-Mal framework, which extracts features from amino acid sequences using Term Frequency Category Relevance and employs the Fisher's score for feature selection is introduced as a novel approach. The framework is evaluated and compared with existing tools and machine learning classifiers.

Material and Methods

The article describes a study that analyzed three protein data sets of E. coli, H. sapiens, and M. musculus using a cross-validation strategy. The data sets were selected to reduce sequence similarity and homology and were prepared from Uniports and CD-HIT databases [6]. The protein sequences were shortened to 25 amino acids with lysine in the center. The model was trained in 10 iterations, optimizing the parameters based on the training sets, and averaging the results of 10 repetitions. The study provides a systematic approach for analyzing protein sequences from different organisms and can help identify similarities and differences in protein structure and function across species. In the TFCRF method, two factors, Positive RF and Negative RF, are used to accurately weigh the features. Positive RF represents

the ratio of the number of amino acids in a protein sequence that contains a specific property to the total number of amino acids in the sequence. Negative RF, on the other hand, represents the ratio of the total number of amino acids in other protein sequences without that property to the total number of amino acids in those sequences. These factors help in determining the significance of the features for prediction purposes [7].

Result

The proposed framework for detecting alanylation sites was evaluated using 10-fold cross-validation on three datasets: E. coli, M. musculus, and H. sapiens. The results showed that the XG Boost algorithm achieved the highest accuracy among the classifiers, while SVM had the lowest reported accuracy. The analysis of the area under the ROC curve confirmed that XG Boost outperformed the other classifiers in terms of generalization ability and prediction performance for identifying alanylation and non-alanylation sites. Error analysis further supported the superiority of XG Boost, demonstrating lower error rates and greater consistency compared to DNN, RF and SVC algorithms. In this study, the performance of the proposed method for predicting alanylation sites was compared to other previous methods. The comparison was based on evaluation criteria such as ACC, SN, SP and MCC, using the XG Boost classification results. The results showed that the proposed method outperformed other methods in all tested datasets, with higher accuracy, sensitivity, and other specific features. Specifically, in the E. coli dataset, the proposed method achieved significantly higher accuracy compared to other methods. Moreover, in the H. sapiens and M. musculus datasets, the proposed method also achieved higher accuracy and other evaluation metrics compared to previous methods.

Conclusion

This study presents a machine learning and natural language processing approach for the detection of malanylation sites in proteins. The TF-CRF technique is employed to extract functional domain information, and the most effective features are selected to prevent overfitting of the model. The cross-validation results demonstrate that the XG Boost classifier performs better than other classifiers when using the selected and extracted features. Furthermore, the results indicate that the features extracted by the proposed method exhibit the best performance.

References

1. Islam S, Mugdha SBS, Dipta SR, Arafat ME, Shatabda S, et al. (2022) MethEvo: An accurate evolutionary information-based methylation site predictor. Neural Computing and Applications pp: 1-12.
2. Xu Y, Ding YX, Ding J, Wu LY, Xue Y (2016) Mal-Lys: Prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection. Sci Rep 6: 38318.
3. Liu X, Wang L, Li J, Hu J, Zhang X (2020) Mal-Prec: Computational prediction of protein Malonylation sites via machine learning based feature integration. BMC Genomics 21(1).
4. Wang M, Xiaowen Cui, Shan Li, Xin HY, Anjun M, et al. (2020) Deepmal: Accurate prediction of protein malonylation sites by deep neural networks. Chemometrics and Intelligent Laboratory Systems 207: 104175.
5. Wang M, Song L, Zhang, Y, Gao H, Yan L, et al. (2022) Malsite-deep: Prediction of protein malonylation sites through deep learning and multi-information fusion based on Nearmiss-2 strategy. Knowledge Based Systems 240: 108191.
6. Hananeh R, Mahdis MH, Hadi Z, Mehdi D (2022) A hybrid feature selection method for predicting lysine malonylation sites in proteins via machine learning. Chemometrics and Intelligent Laboratory Systems 222: 104496.
7. Maleki M, Abdollahzadeh A (2007) TFCRF: A novel feature weighting method based on class information in text categorization. International Conference on Computer, Information and Systems Science and Engineering, Bangkok.