doi 10.18699/bgrs2024-12.3-13

DeNovo sampling and sequence-derived structural descriptors based *in-silico* identification of host-pathogen protein interactions

Abbasi W.A.[1]*, Ali M.[1], Shabir A.[2], Andleeb S.[3], Bibi M.[1], Abbas S.A.[1]

[1] *Computaional Biology and Data Analysis Lab., Department of Computer Sciences & Information Technology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*

[2] *Department of Software Engineering, Capital University of Science & Technology (CUST), Islamabad, Pakistan*

[3] *Biotechnology Lab., Department of Zoology, King Abdullah Campus, University of Azad Jammu & Kashmir, Muzaffarabad, Pakistan*

** wajidarshad@ajku.edu.pk*

Key words: machine learning; protein interactions; infectious diseases; host-pathogen interactions; predictive modeling; protein sequence

Motivation and Aim: Infectious diseases pose an incessant socio-economic burden on our healthcare systems as we witnessed during the COVID-19 pandemic [1, 2]. Protein-protein interactions are a key mechanism in pathogen infections of hosts such as viral or bacterial diseases in humans [3]. Therefore, the prediction of host-pathogen protein-protein interactions (HPPPIs) is crucial for understanding the mechanisms of infectious diseases and developing effective therapies. Experimental methods for identifying these interactions are time-consuming and expensive, and there is a need for efficient computational methods to predict these interactions.

Methods and Algorithms: In this study, we propose a novel approach for predicting HPPPIs that combines sequence-derived structural features with de novo sampling for negative examples. Firstly, we extract structural features from the amino acid sequences of the host and pathogen proteins using a computational tool. These features include pseudo-amino acid compositions (PseAAC), autocorrelation features: normalized Moreau–Broto autocorrelation, Moran autocorrelation, Geary autocorrelation, composition, transition, distribution, and Sequence-order [4]. Then, we develop a machine learning model that combines these structural features with the DeNovo sampling of non-interacting proteins to predict HPIs. The DeNovo sampling technique and sequence-derived structural features increase the likelihood of identifying novel HPIs and improve the accuracy of the predictions. To evaluate the performance of our method, we apply it to a dataset of known HPIs and compare its results with those of other state-of-the-art methods. According to our findings, our method performs better than other methods in terms of accuracy and sensitivity.

Results: With leave-one-family-out (LOFO) cross-validation we have used three different classifiers (SVM, RF, and XGBoost) and three different types of feature representations (Composition, Substitution, and SDSD) to evaluate the performance of our proposed machine learning based model for host-pathogen protein interaction prediction both in terms of area under the ROC curve (AUC-ROC) and area under the PR curve (AUC-PR). For bacteria as host, our proposed model for host-pathogen protein interaction prediction achieved a maximum AUC-PR value of 76.55 % using XGB along

with sequence-derived structural feature representation. Similarly, for a virus as host, our proposed model for host-pathogen protein interaction prediction achieved a maximum AUC-PR value of 77.21 % using XGB along with sequence-derived structural feature representation. On an external validation dataset our proposed model for predicting host-pathogen interactions performed better, with AUC-PR values for viruses and bacteria of 77.77 % and 83.38 %, respectively. We have also used some benchmark datasets such as Zhou's H1N1 and Zhou's Ebola [5], Denovo Slim [6], and Barman [7] to compare our method (Proposed) with some state-of-the-art methods such as GENERALIZED [5], DOC2VEC [8], DENOVO [6], and MTT [1] by using F1 score.
Conclusion: The use of sequence-derived structural descriptors provides a more detailed representation of the proteins involved in the interaction, which is more informative than simply using the amino acid sequence. Our approach also includes negative samples generated using a de novo sampling approach, which is a valuable addition to the existing methods that rely on random negative examples.

References

1. Dong T.N., Brogden G., Gerold G., Khosla M. A multitask transfer learning framework for the prediction of virus-human protein–protein interactions. *BMC Bioinformatics*. 2021;22:572
2. Standl F., Jöckel K.-H., Brune B., Schmidt B., Stang A. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *Lancet Infect Dis*. 2021;21:e77
3. Acharya D., Dutta T.K. Elucidating the network features and evolutionary attributes of intra- and interspecific protein–protein interactions between human and pathogenic bacteria. *Sci Rep*. 2021;11:190
4. Cao D.-S., Xu Q.-S., Liang Y.-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*. 2013;29:960-962
5. Zhou X., Park B., Choi D., Han K. A generalized approach to predicting protein-protein interactions between virus and host. *BMC Genomics*. 2018;19:4924
6. Eid F.-E., ElHefnawi M., Heath L.S. DeNovo: virus-host sequence-based protein-protein interaction prediction. *Bioinformatics*. 2016;32(8):1144-1150. doi 10.1093/bioinformatics/btv737
7. Barman R.K., Saha S., Das S. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods. *PLoS One*. 2014;9:e112034
8. Yang X., Yang S., Li Q., Wuchty S., Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput Struct Biotechnol J*. 2020;18:153-161