2020

A Computational Framework for Host-Pathogen Protein-Protein Interactions

Huaming Chen
University of Wollongong

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

A Computational Framework for Host-Pathogen Protein-Protein Interactions

Huaming Chen

This thesis is presented as part of the requirements for the conferral of the degree:

Doctor of Philosophy

Supervisor:
Associate Professor Jun Shen

Co-supervisors:
Associate Professor Lei Wang
Associate Professor Jiangning Song
Professor Chi-Hung Chi

The University of Wollongong
School of School of Computing and Information Technology

June, 2020

Declaration

I, *Huaming Chen*, declare that this thesis is submitted in partial fulfilment of the requirements for the conferral of the degree *Doctor of Philosophy*, from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Huaming Chen

June, 2020

ABSTRACT

Infectious diseases cause millions of illnesses and deaths every year, and raise great health concerns world widely. How to monitor and cure the infectious diseases has become a prevalent and intractable problem. Since the host-pathogen interactions are considered as the key infection processes at the molecular level for infectious diseases, there have been a large amount of researches focusing on the host-pathogen interactions towards the understanding of infection mechanisms and the development of novel therapeutic solutions. For years, the continuously development of technologies in biology has benefitted the wet lab-based experiments, such as small-scale biochemical, biophysical and genetic experiments and large-scale methods (for example yeast-two-hybrid analysis and cryogenic electron microscopy approach). As a result of past decades of efforts, there has been an exploded accumulation of biological data, which includes multi omics data, for example, the genomics data and proteomics data.

Thus, an initiative review of omics data has been conducted in Chapter 2, which has exclusively demonstrated the recent update of 'omics' study, particularly focusing on proteomics and genomics. With the high-throughput technologies, the increasing amount of 'omics' data, including genomics and proteomics, has even further boosted. An upsurge of interest for data analytics in bioinformatics comes as no surprise to the researchers from a variety of disciplines. Specifically, the astonishing rate at which genomics and proteomics data are generated leads the researchers into the realm of 'Big Data' research. Chapter 2 is thus developed to providing an update of the omics

background and the state-of-the-art developments in the omics area, with a focus on genomics data, from the perspective of big data analytics.

Even though the host-pathogen interactions (HPI) systems have been a hot research topic, the study of HPI is still in its early stage. One of the dominant reasons is that, the identification of host-pathogen interactions takes a huge amount of experimental resources and will demand lots of time, which has significantly limited the progress in studying different HPI systems. Alternatively, computational models, as a cost-effective approach, will facilitate the process for analysis and predictions of HPI systems with the basis of the experimental data. Considering some specific issues existed in host-pathogen protein-protein interactions (HP-PPIs) area, including the databases and computational models, this thesis aims to propose a computational framework for HP-PPIs prediction and study the structural interaction network to further reach an in-depth analysis. In details, this thesis has made the following contributions.

A comprehensive review targeting on host-pathogen interactions resources published in the last decades is firstly conducted. Since the prevailing application of high-throughput sequencing and interaction detection methods has improved the production of inter-species interaction data, numerous host-pathogen interactions resources have been released. Ranging from various aspects of available 'omics' data, theses host-pathogen interactions (HPI) resources are accumulated in a fast speed, in which one of the dominant sources is the protein-protein interactions. However, some of the published data may only relate to specific human-pathogen interactions system, for example, the interactions between human and HIV virus. These databases may be of special interests to a sub-group of researchers. Moreover, there is still lacking a comprehensive overview of these host-pathogen interactions resources with fingerposts delivering for particular research issues in the present, including the goal towards data analysis and prediction.

Secondly, a systematic evaluation of machine learning-based models for prediction of HP-PPI is presented. Although several literature reviews have been published by introducing the machine learning-based models and some applications in the HPI domain,

little research with empirical evaluations of the performance of HPI predictions based on machine learning models has been conducted. Meanwhile, most studies of protein-protein interactions prediction have been conducted based on a hypothesis of evaluating the predictor with a balanced and small dataset, in which the numbers of positive and negative PPIs are equivalent. In this thesis, a more extensively empirical evaluation considering different categories of sequence feature representation algorithms and numerous traditional machine learning models is delivered.

Given the systematic evaluation performance of machine learning prediction models for HP-PPIs in Chapter 4, the thesis subsequently focuses on developing novel machine learning-based models to improve the performance for discovery of interactions of HP-PPIs.

The third contribution in this thesis is a heterogeneous information mining and ensembling (HIME) model for discovery of interactions of HP-PPIs. In presence of heterogeneous information based on sequence data, the HIME model is designed to harness the power of the heterogeneous information and to benefit from various weak machine learning models. The studied six different HP-PPIs datasets are included to evaluate the performance and the extensive experiments show that HIME model is highly effective and efficient.

The fourth contribution in this thesis is a two-layer machine learning-based model for discovery of interactions of HP-PPIs. The two-layer machine learning-based model, which is entitled 'APEX2S' model, is proposed to alleviate the latent imbalanced characteristics of HP-PPIs dataset. In details, the two-layer model consists of two essential modules, which are XGBoost and SVM. Herein, XGBoost is included to reduce the imbalanced ratio of the data and SVM is utilised to enhance the prediction performance. SMOTE technology is as well incorporated as a key component in the model to alleviate the bias of imbalanced ratio. The curated dataset human-Shigella protein-protein interactions dataset in Chapter 3 is utilised as the independent benchmarking dataset, and the results have shown an enhancement of the overall performance.

The fifth contribution in this thesis is an advanced bidirectional long short-term memory-based (Bi-LSTM-based) model for discovery of interactions of HP-PPIs. The bidirectional LSTM-based model is a variant deep learning model of long short-term memory model, which has demonstrated superior performance in domains such as natural language processing, transportation and action recognition. However, the direct incorporation of traditional Bi-LSTM model will cause the model explicitly suffering from the conventional vanishing gradient problem for the prediction of HB-PPI data. Thus, in this chapter, a novel bidirectional LSTM-based (Bi-LSTM-based) model is designed to yield results quite smoothly when the ratio changes. Meanwhile, the Bi-LSTM-based model also shows a strong capability in dealing with the imbalanced issue. In comparison with the evaluation models and the literature methods, Bi-LSTM-based model has demonstrated a better performance in our study.

The sixth contribution in this thesis includes an unsupervised deep learning model for discovery of interactions of HP-PPIs. Since deep learning method has shown powerful performance in many areas, such as computer vision and nature language processing, this thesis presents an unsupervised deep learning model for HP-PPIs prediction task, which is based on stacked denoising autoencoders to capture higher level features regarding the sequence information. The achieved performance indicates a superior capability of the unsupervised deep learning model in dealing with the host-pathogen protein-protein interactions scenario.

Lastly, this thesis concludes the contributions with a detailed effort for the reconstruction of a HP-PPIs structural interaction network (SIN) utilizing structure information of proteins. Besides sequence information, structure information of protein is another main published, experimentally determined three-dimensional (3D) structural data. It is an atomic resolution macromolecular information for protein. However, the missing data problem would hamper the acquirement of the structure information. A mapping tool, which is *BLAST/PRISM*, is thus considered. Since the domain interactions are considered as the solid evidence between proteins, *iPfam/3did* databases would be also

utilized to filter this SIN to validate this network. There is a scarcity of studies based on 3D structural data to provide an atomic mechanistic and high-resolution view of HP-PPIs. In this chapter, we have demonstrated that SIN would be an alternative solution revealing more mechanistic patterns of host-pathogen interactions which will be an essential part for the future research. Lastly, we have concluded the thesis by summarizing the advanced machine learning-based models, which will include inventive feature representation algorithms and novel deep learning models, in the future work to enhance the effectiveness for predictions of HP-PPIs.

ACKNOWLEDGMENTS

A long journey with happiness and frustration, with smiles and tears, with sweat and joy, to pursue the Ph.D would be impossible for me without the help from many people whom I sincerely wish to appreciate.

Foremost, I would like to express my deepest gratitude to my supervisor, also my life mentor, A/Prof. Jun Shen, for his continuous support, guidance, encouragement and patience during my study. It has been a great pleasure and honor to have his supervision. His constant and consistant support to me during my most depressed time has enlightened my life. It has been my great privilege to accomplish the journey with his motivation.

My great acknowledgements are also for my cosupervisors A/Prof. Lei Wang, A/Prof. Jiangning Song and Prof. Chi-Hung Chi. Their consistent guidance and encouragement have provided me a strong support for the study.

This long journey would not be possible from the great and precious help and advice from many people. I would like to thank Prof. Yaochu Jin (Surrey), Prof. Yaoqi Zhou (Griffith), Prof. Jie Lu (UTS), Prof. Jinyan Li (UTS), Prof. Geoff Webb (Monash), Prof. William Guo (CQU), Dr. Xianjun Dong (Harvard), Prof. Fang Dong (SEU), Pro. Nam Ling (SCU), Dr. Luping Zhou (USYD), A/Prof. Jia Zhang (CMU), Dr. Shiping Chen (CSIRO), Dr. Ruimin Li (TfNSW), Dr. Tingru Cui (Melbourne), Dr. Qiang He (Swinburne), Dr. Geng Sun (UOW), Prof. Qingguo Zhou (LZU), Dr. Xuyun Zhang (Macquarie), A/Prof. Robert Clark (ANU), A/Prof. Ping Yu (UOW), Dr. Rongmao Chen (NUDT), Dr. Jianjia Zhang (UTS), Dr. Yangguang Tian (SUTD), Dr. Pichao Wang

(Alibaba), Dr. Zhimin Gao (UOW), Dr. Fuyi Li (Monash), Dr. Binbin Yong (LZU), Dr. Lijuan Wang (Xidian), Dr. Da-wei Zhang (Karolinska), Dr. Jack Yang (UOW), Dr. Yan Wang (SCU), Dr. Guo Yu (Surrey), Dr. Hua Zhang (LinkingMed) for their insightful advices, valuable research communications and collaborations from which I am still benefitting today.

I would like to thank Fucun Li, Jianqing Wu, Jiayi Lin, Yunshu Zhu, Ting Song, Fei Xie, Fanghui Jia, Zhuoling Tian, Dr. Qing Zhang, Jiayi Yang, Weishi Shi, Zhihao Zhang, Yang Li, Lei Qi, Yuying Qi, Rui Han, Ruoqi Zhao, Yu Ding, Dr. Chen Zu, Dr. Dan Yuan, Prof. Weihua Li and so on, for their valuable friendships.

Lastly and most importantly, I would express my greatest gratitude to my family. I could not complete this journey without your support. Thank you all for your precious understanding and endless love and support throughout my life.

My deepest love to my little man. Thank you to come to our family, Mr. Chen.

PUBLICATIONS

This thesis is developed based on the first-authored papers in the following publications/manuscripts. The co-authored papers are not included in the main body of this thesis. All the listed publications/manuscripts are published during the Ph.D course.

Accepted&Published Papers

1. ***Chen, H.***, Wang, L., Chi, H., Shen, J., 2020, "A Two-Layer Machine Learning Model for Discovery of Interactions on Cloud-based Multi-omics Data", pp. 1-16, *Concurrency & Computation: Practice & Experience*, https://doi.org/10.1002/cpe.5846

2. ***Chen, H.***, Li, F., Shen, J., Song, J., Wang, L., Jin, Y., Chi, H., 2020, "Systematic Evaluation of Sequence-based Machine Learning Prediction Models for Human-Bacterium Protein-Protein Interactions', pp. 1-21, *Briefings in Bioinformatics*, bbaa068, https://doi.org/10.1093/bib/bbaa068

3. ***Chen, H.***, Wang, L., Jin, Y., Chi, H., Shen, J., 2020, "HIME: Mining and Ensembling Heterogeneous Inforamtion for Protein-Protein Interaction Prediction', pp. 1–8, *The 2020 International Joint Conference on Neural Networks (IJCNN 2020)*

4. ***Chen, H.***, Shen, J., Wang, L., Jin, Y., 2020, 'Towards A More Effective Bidirectional LSTM-based Learning Model for Human-Bacterium Protein-Protein Interactions', pp. 1-10, *14th International Conference on Practical Applications of*

Computational Biology and Bioinformatics

5. ***Chen, H.***, Shen, J., Wang, L., Song, J., 2020, 'A framework towards data analytics on host–pathogen protein–protein interactions', pp. 1-13, *Journal of Ambient Intelligence and Humanized Computing*

6. ***Chen, H.***, Wang, L., Jin, Y., Chi, H., Li, F., Chu, H., Shen, J., 2019, 'Hyper-parameters Estimation in SVM with GPU Acceleration for Prediction of Protein Interactions', pp. 2197–2204, *IEEE International Conference on Big Data*

7. ***Chen, H.***, Wang, L., Chi, C. H., & Shen, J., 2019, September. 'Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions', In 2019 Seventh International Conference on Advanced Cloud and Big Data (CBD) (pp. 133-138). IEEE.

8. ***Chen, H.***, Wang, L., Chi, H., Shen, J., 2019, 'Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions', pp. 133-138, *The Seventh Intrnational Conference on Advanced Cloud and Big Data*

9. ***Chen, H.***, Guo, W., Shen, J., Wang, L. & Song, J., 2018, ' Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey', *IEEE Access*, 6:11760-11771.

10. ***Chen, H.***, Shen, J., Wang, L., Song, J. & Chi, C., 2018, 'Towards Biological Sequence Data Service with Insights', *Proceedings of IEEE International Conference on Big Data, 2018*, pp. 2847–2854

11. ***Chen, H.***, Shen, J., Wang, L. & Song, J., 2017. 'Leveraging Stacked Denoising Autoencoder for prediction of PHPPI', *IEEE International Congress on Big Data*, pp. 368-375

12. ***Chen, H.***, Song, J., Sun, G., Shen, J., & Wang, L., 2017. 'Towards Elucidating the Structural Principles of Host-Pathogen Protein-Protein Interaction Networks: A

bioinformatics survey', *IEEE International Congress on Big Data*, pp. 177-184

13. ***Chen, H.***, Shen, J., Wang, L. & Song, J., 2016. 'Collaborative Data Analytics towards Prediction on Pathogen-Host Protein-Protein Interactions', *CSCWD*, pp. 269-274

14. ***Chen, H.***, Shen, J., Wang, L. & Song, J., 2016. 'Towards Data Analytics of Pathogen-Host Protein-Protein Interaction: A survey'. *IEEE International Congress on Big Data*, pp. 377-388

15. Yong, B., Shen, J., Liu, X., Li, F., ***Chen, H.***, Zhou, Q., 2019, 'An Intelligent Blockchain-based System for Safe Vaccine Supply and Supervision', 52(2020), pp.102024: 1-12, *International Journal of Information Management*

16. Brown, P., ***RELISH Consortium***, Zhou, Y., 'Large expert-curated database for benchmarking document similarity detection in biomedical literature search', pp 1-66, *Database*

17. Li, R., Rose, G., ***Chen, H.*** and Shen, J., 2018. 'Effective long-term travel time prediction with fuzzy rules for tollway'. Neural Computing and Applications, 30(9), pp.2921-2933.

18. Yong, B., Xu, Z., Shen, J., ***Chen, H.***, Wu, J., Zhou, Q. & Li, F., 2018, 'General Vector Machine for Electricity Forecasting', International Journal of High Performance Computing and Networking, invited paper from AusPDC, in press

19. Li, F., Wu, J., Dong, F., Lin, J., Sun, G., ***Chen, H.*** & Shen, J., 2018, 'Ensemble Machine Learning Systems for the Estimation of Steel Quality Control', *Proceedings of IEEE International Conference on Big Data, 2018*, pp.2245–2252

20. Zhou, Q., Chen, C., Zhang, G., ***Chen, H.***, Chen, D., Yan, Y., Shen, J. & Zhou, R., 2018, 'Real-time Management of groundwater resource based on wireless sensor networks', *Journal of Sensor and Actuator Networks*, 4(1): 1-11

21. Yong, B., Shen, J., Shen, Z., ***Chen, H.***, Wang, X & Zhou, Q., 2018, 'GVM Based Intuitive Simulation Web Application for Collision Detection', *Neurocomputing*, 276:63-73

22. Sun, G., Cui, T., Xu, D., ***Chen, H.***, Chen, S. & Shen, J., 2017, 'Assisting Open Education Resource Providers and Instructors to Deal With Cold Start Problem in Adaptive Micro Learning: a Service Oriented Solution', *14th IEEE International Conference on Services Computing*, pp. 196-203

23. Yong, B., Shen, J., Sun, H., ***Chen, H.*** & Zhou, Q., 2017, 'Parallel GPU-Based Collision Detection of Irregular Vessel Wall for Massive Particles'. *Cluster Computing: The Journal of Networks, Software Tools and Applications*, online first pp. 1-13.

24. Yong, B., Xu, Z., Shen, J., ***Chen, H.***, Tian, Y. & Zhou, Q., 2017. 'Neural Network Model with Monte Caro Algorithm for Electricity Demand Forecasting in Queensland'. *Australasian Symposium on Grid Computing and e-Research (now AusPDC)*, pp. 47:1-47:7

25. Zhou, Q., ***Chen, H.***, Zhao, H., Zhang, G., Yong, J. & Shen, J., 2016. 'A local field correlated and Monte Carlo based shallow neural network model for non-linear time series prediction'. *Scalable Information Systems*, 3(8): e5

26. Yong, B., Zhang, G., ***Chen, H.*** and Zhou, Q., 2016. 'Intelligent monitor system based on cloud and convolutional neural networks'. *The Journal of Supercomputing*, pp.1-17.

27. Wang, L., Shen, J., Zhou, Q., Shang, Z., ***Chen, H.*** & Zhao, H., 2016, 'An evaluation of the dynamics of diluted neural network'. *International Journal of Computational Intelligence Systems*, 9 (6), pp. 1191-1199.

Book Chapters&Posters&Patents

28. ***Chen, H.***, Song, J., Shen, J. and Wang, L., 2016. 'Big Data in Genomics'. Big Data Management and Processing. Taylor Francis Group: CRC Press, pp. 363-384.

29. Zhou, Q., ***Chen, H.***, Di, C. and Zhou, R., Granted Patent: 'A method and device of Line Tracking and Navigation for Robots'. Application Number: 201310623873.3, issued in 2017.05

30. ***Chen, H.***, Zhao, H., Wang, L., Song, J. and Shen, J., 2017, 'A Comparison Study for Supervised Machine Learning Models in Cancer Classification'. *16th International Conference on Bioinformatics (InCoB 2017)*, selected as Oral Presentation.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Background

Immersing in many disciplines, such as computer vision, economics, natural language processing, online learning, bioinformatics and so on, big data, which terms data with characteristics of high volume, high velocity and high variety, is impacting every aspect of our research and life. More and more researches focus on data mining and machine learning for predicting and uncovering the related domain knowledge. Particularly, the extraordinarily expanding pace of data volume, variety and value characteristics is bringing more attention on research towards the advancement of biology science. The adoption of big data in bioinformatics has become a hot research topic not only in genomics and proteomics areas [1], but also in biomedical medicine and biomedical image areas [2].

Omics data, image data and signal data are dominant in biomedical research whilst providing insights and research opportunities for biologists. These accumulated data are deemed essential for transformation from experiments to valuable knowledge [3]. With the development of advanced high-throughput technologies, enormous amounts of data are being generated by biologists. The availability of large-scale multi-omics data, including proteomics data from The European Bioinformatics Institute (EBI) [4–6] and genomics data from The Cancer Genome Atlas (TCGA) [7], provides an unprecedented opportunity to transform the biomedical research onto system-level, mechanistic studies

aimed at a comprehensive and holistic understanding of biological systems [8].

It is witnessed that, the data accumulated in a large scale via *wet lab*-based and *in vitro*-based methods for biology science is booming to challenge the traditional data analysis area of both computational and biological methods. Ranging from genomic sequencing experiments to images of physiological structures, biologists are starting to grapple with tremendous data sets, encountering challenges in processing and analysing information that were once considered only with specific domain knowledge [9]. The analysis of biological data is now becoming a *data-driven* work that helps biologists designing the future experiments. The direct benefit for data analytics in biology is that, with the huge amounts of data we have obtained nowadays, the hypothesis and phenomena behind these biology researches could be generated based on data, which was summarized via vast amount of experiments.

Herein, this chapter begins with the definitions of biological terminologies used in the context of this thesis.

Proteins are considered as the basics of living organisms and the interactions between different proteins are the basics of the biological functions, including immune response, signal transduction and other essential functions [10].

* **Definition 1 (Amino Acids)**: Amino acids are the structural units (monomers) that make up proteins. As the important organic compounds, amino acids consist of carbon, hydrogen, oxygen and nitrogen. Basically, a composition of hundreds or even thousands of amino acids residues defines the primary structure of a protein. There are 20 different kinds of amino acids. Shown as below Figure 1.1 is a diagram for these amino acids. There are also several other rarely existing amino acids, such as Selenocysteine (Sec) and Pyrrolysine (Pyl).

* **Definition 2 (Protein)**: As amino acids are the monomer units of proteins, proteins are composed of at least one chain of amino acids which is called a polypeptide. A polypeptide is a linear chain which is directly defined by the composition of amino acids. With folding, at least one polypeptide bends, twists and forms a unique,

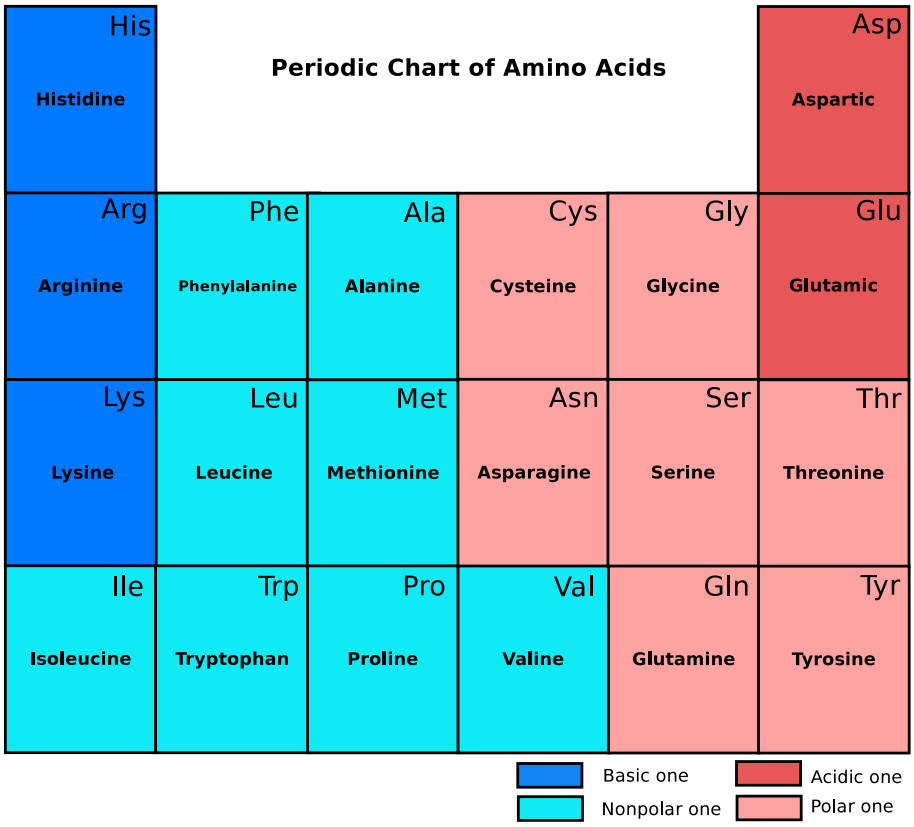


Figure 1.1: Twenty Basic Proteinogenic Types of Amino Acids [11]

complicated and stable three-dimensional structure. Most polypeptides shorter than about forty amino acids in length do not fold. A determination of structure is far harder than the composition of amino acids. Figure 1.2 and Figure 1.3 illustrate the basic formation process of protein.

Proteomics is a main branch in computational biology, since proteins are considered as the basis of living organisms and the interactions between different proteins result in the biological functions, including immune response, signal transduction and other essential functions [10]. Given a proteome is a collection of functional and non-functional proteins existing in an organism, biological system, even biological context, proteomics considers the proteomes study in a large scale [14]. Figure 1.4 shows the included areas of proteomics, in which data will be collected from three different properties of location, abundance/turnover and post-translational modifications. Either directly utilising these data or inferring additional information from these data, the study of proteomics provides

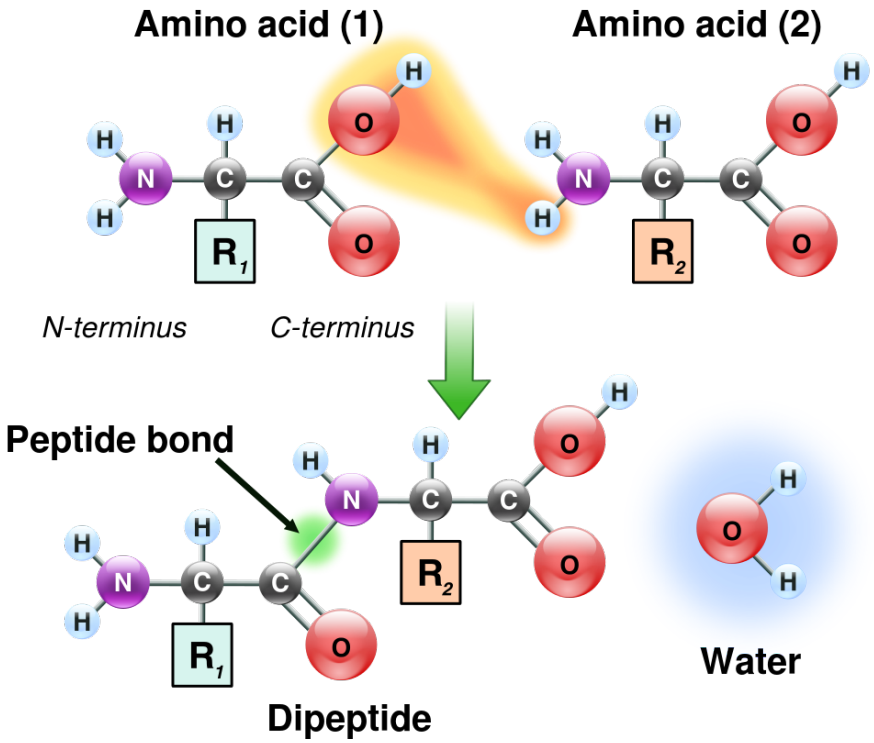


Figure 1.2: Amino Acids to Polypeptide [12]

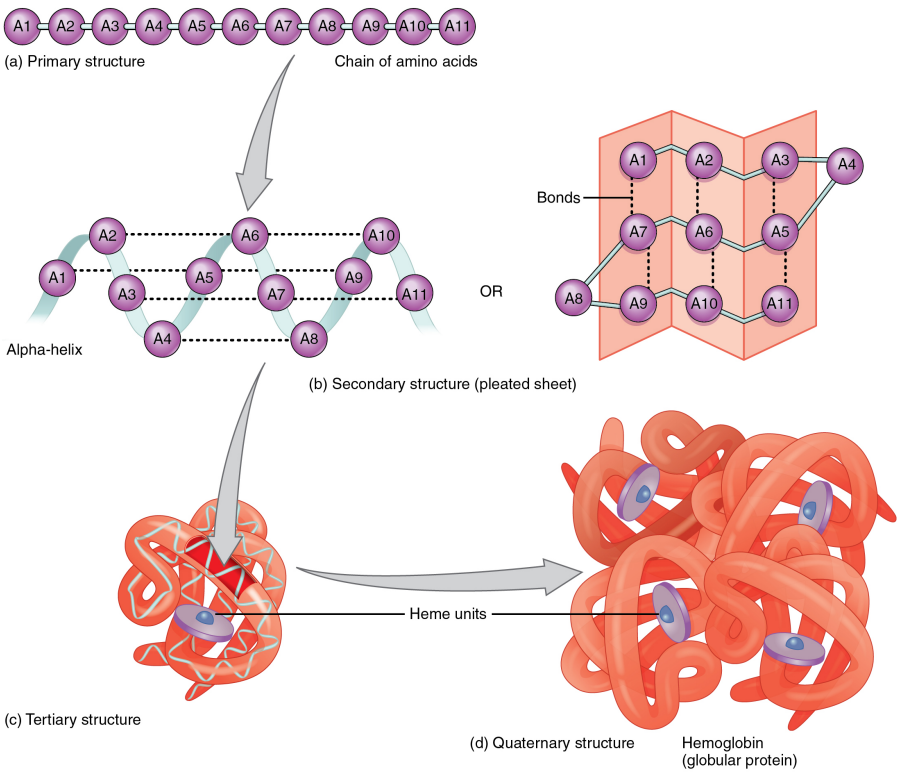


Figure 1.3: Protein Structure [13]

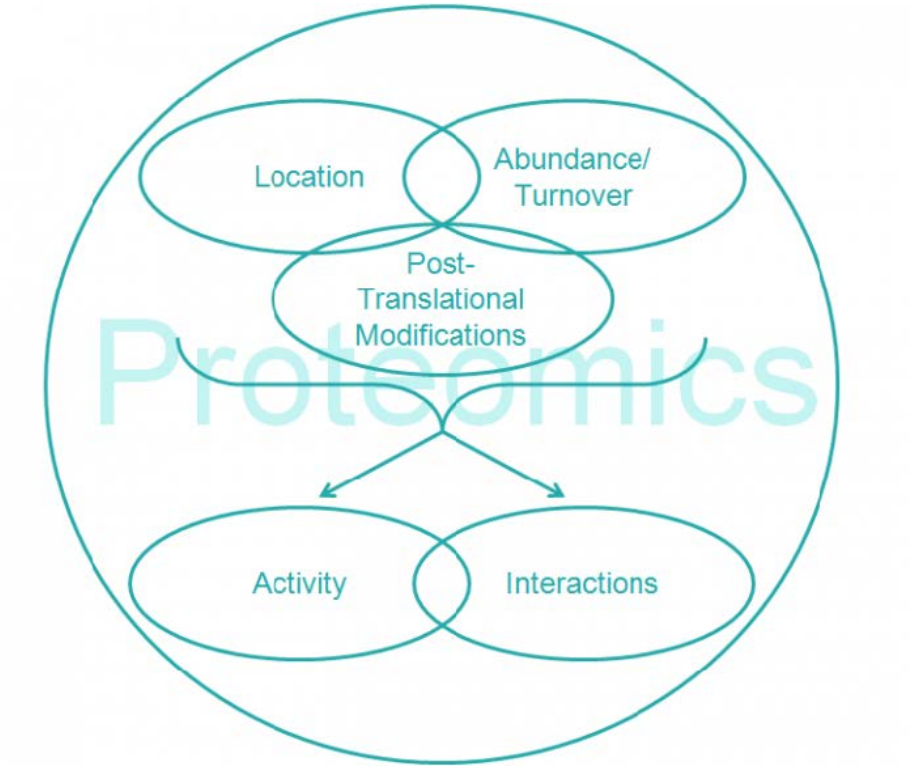


Figure 1.4: The Diagram of Proteomics[14]

substantial biological information and benefits the study of many biological problems. 'Proteomics' was originally coined in 1995.

* **Definition 3 (Proteomics) [15–17]**: Proteomics is the large-scale study of proteins, usually by biochemical methods, which goal is to define the large-scale characterization of the entire protein complement of a cell line, tissue and organism by focusing on the five central pillars of proteomics research – mass spectrometry-based proteomics, proteome-wide biochemical assays, systematic structural biology and imaging techniques, proteome informatics, and clinical applications of proteomics.

As one of the major topics in proteomics, the studies of protein-protein interactions mostly utilise large-scale and small-scale experimental methods, such as affinity purification, yeast two-hybrid assay, affinity purifications-mass spectrometry (AP-MS) method, nuclear magnetic resonance (NMR) and mass spectrometry methods. It is a key question about a protein, which is as important as the questions concerning when and where the protein is expressed. Protein-protein interactions tell with which other proteins the protein

interacts [15]. The tremendous value of a comprehensive protein-protein interaction map of a biological system, such as a cell, presents a precious opportunity to understand the mechanism of the biological system. Figure 1.5 presents a protein-protein interactions network for schizophrenia genesis.

* **Definition 4 (Protein-Protein Interactions) [18, 19]**: Protein–protein interactions (PPIs) are the physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by electrostatic forces including the hydrophobic effect. Commonly, the physical contacts with molecular association/docking between chains occur in a cell or in a living organism in a specific biomolecular context. Specifically, two aspects are elaborated for the definition of protein-protein interactions, which are firstly the interaction interface should be intentional and not accidental and secondly the interaction interface should be non-generic. Protein-protein interactions do not imply a static or permanent status for the physical contacts between proteins.

1.2 Motivation and Goals

As for proteomics is a main branch in bioinformatics, a natural benefit for data analytics in proteomics is to facilitate the understanding and prediction of the knowledge for proteins, specifically for protein-protein interactions. Protein-protein interactions play a crucial role by conducting basic biological functions in most biological processes. Mostly, PPIs can be referred to either ‘intra-species PPI’ or ‘inter-species PPI’. Intra-species PPI is the interaction between two proteins from the same species, while inter-species PPI means the interaction occurs between two proteins from two different species. In this thesis, how to get a better understanding and prediction of inter-species PPI, exactly between the host and pathogen, is the research objective.

How to identify a PPI is essential for understanding the whole biological functions. Since PPIs are essential to the majority of most cellular functions, many innovative

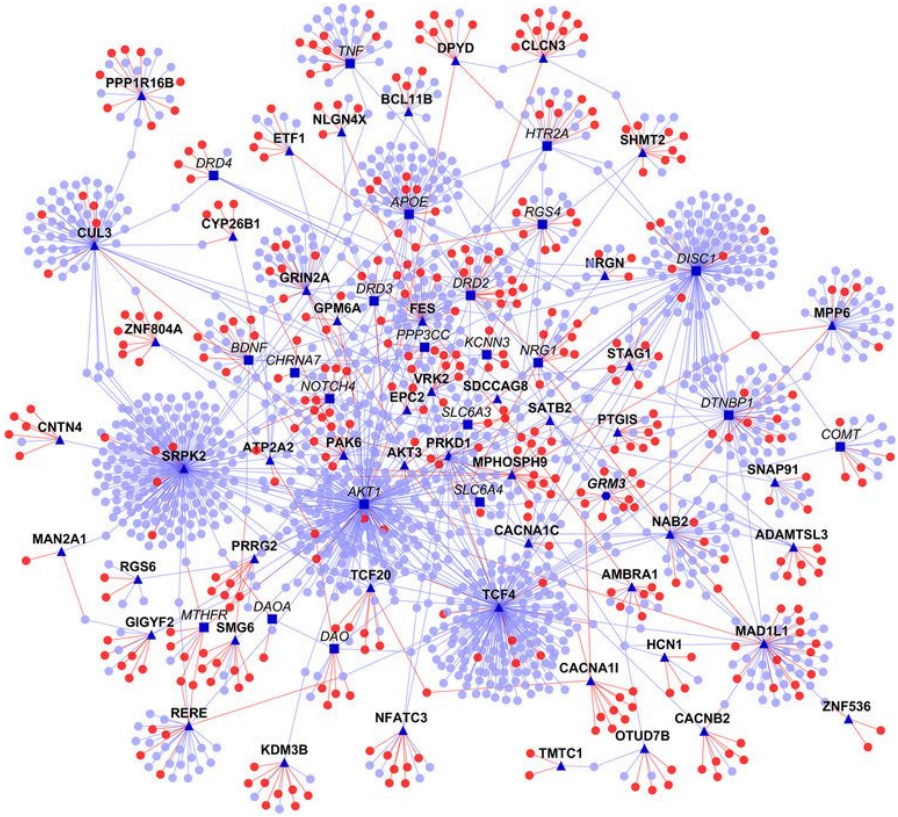


Figure 1.5: Schziophrenia Protein-Protein Interactions Network [20]

techniques and systems for identifying protein interactions have been developed [21]. Classifying pairs of proteins as interacting or not has been the subject of intense research in recent years both in computational and experimental areas [22]. By far, numerous supervised machine learning models have been adopted to PPIs' prediction task.

For determining and studying PPIs, the most reliable biological methods are conducted by *wet lab*-based experiments, which are deemed to be time consuming and resource costing. Including low-throughput and high-throughput technologies, false positive data and false negative data will be highly possible produced in one-time shot experiment. Thereby, a reliable positive and negative data always require more repetitive experiments, which will consume much more efforts, lab resources and time. It is also very difficult for investigation with the prohibitively large set of possible host-pathogen protein-protein interactions [23]. It has been reported that the unavailability of experimental methods for large-scale detection of interactions between host and pathogen organisms is the obstacles

[24] and also the false positive rate of available both computational predicted and high-throughput experimentally verified interaction data are high [25].

Considering infectious disease as the major worldwide health concerns, they have caused millions of illnesses and deaths every year. Figure 1.6 is an example from Leibniz Institute for Natural Product Research and Infection Biology, which research is to illustrate the interaction of Aspergillus fumigatus with the human innate immune system. Host-pathogen interactions are subsequently studied as they play as the key infection processes at the molecular level. Most diseases, which occur between hosts and pathogens, can be analysed by groups of infection mechanisms. For host-pathogen protein-protein interactions (HP-PPIs), including information of infection pathways, it reveals much more in the infection mechanisms between host and pathogen. Thus, analysing and understanding protein-protein interactions is of great importance and presents huge values to the study on infectious diseases, especially for inter-species interactions. This thesis focuses on the protein-protein interactions between human and pathogens [26, 27], termed as host-pathogen protein-protein interactions (HP-PPIs) in the following sections, which has been one of the hot topics towards the mechanism study of infectious diseases. Concerning infectious diseases are still one of the dominant diseases causing death, the research of HP-PPIs generally solicits data from different perspectives to examine the hypothesis and propose potential therapeutics. Vast researches have been conducted with a long time of development and examination.

Since host-pathogen PPIs are the key to either the mechanisms of infection or medicine treatment, how to obtain a better understanding and prediction of inter-species PPI, specifically between hosts and pathogens, is a hot topic for biology research. As a result of decade efforts of *wet lab*-based experiments in biology, the production of biological data, e.g. protein interactions data, has exploded. Even though there are still substantial data to be further experimented, the collected data has benefitted the research on disease mechanisms though to a limited extent. One of the earliest studies was on the symptom of anthrax, which was identified as primarily caused by the protein interactions between

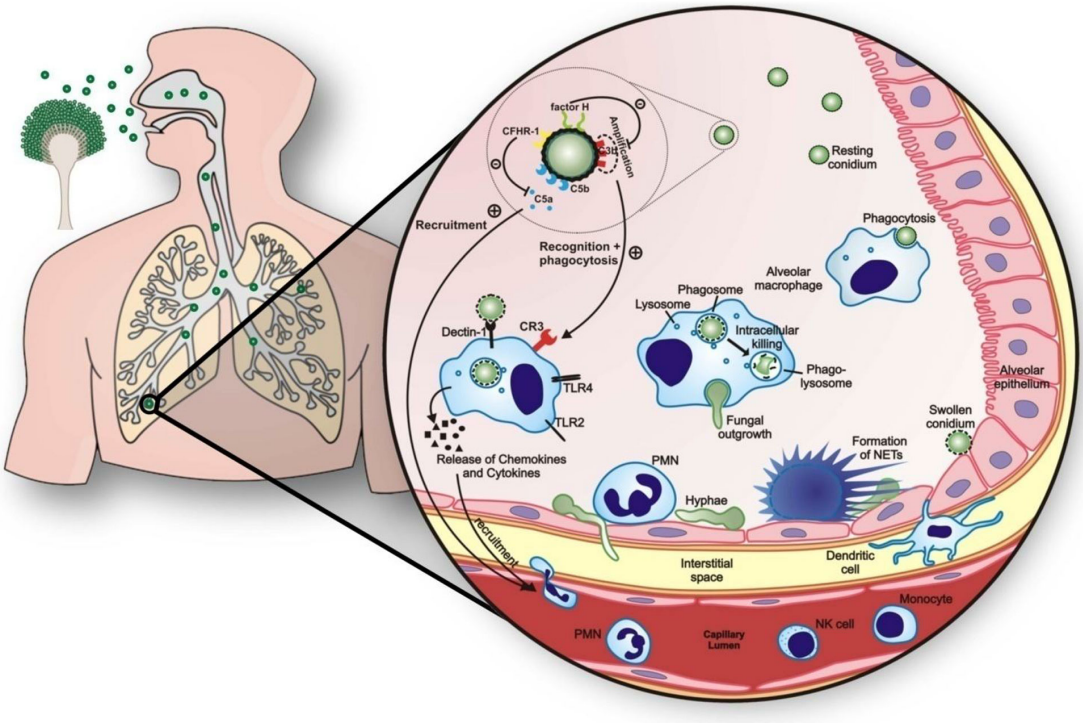


Figure 1.6: Interaction of A. fumigatus with the human innate immune system, from Leibniz Institute for Natural Product Research and Infection Biology with link of *https://www.leibniz-hki.de/en/virulence-of-aspergillus-fumigatus.html*

human and Bacillus anthracis. Bacillus anthracis is a type of bacterium pathogens, where people want to fully understand mechanisms with the protein interactions map between Bacillus anthracis and Homo sapiens (the host).

Meanwhile, as protein-protein interactions take charge of almost every biological processes, systems biology-based approaches also study infectious diseases by analysing the interactions between the host species and the pathogen organisms [23]. Different from traditional protein-protein interactions, it has been reported that the unavailability of experimental methods for large-scale detection of interactions between host and pathogen organisms is one of the main obstacles [24]. On the other hand, the false positive rate of the available computational and high-throughput experimental interaction data sets remains high [25]. Specifically, for HP-PPIs, less experimentally identified data than intra-species PPIs are available and multi high skewed data distribution should be further investigated with regard to computational model constructions.

Currently, there is little effort on delivering a comprehensive study of the experimentally identified HP-PPIs data, as well as a systematic evaluation of the computational models for the prediction task of host-pathogen protein-protein interactions. Even though, there is an abundant information of protein with the details available for HP-PPIs. Different kinds of data regarding protein characteristics, such as sequence information, the homology between proteins, structural information of each proteins, and annotations, are available for data analysis and can be utilised to build computational model in a positive way. How to deal with the existing data and furthermore to incorporate these data with machine learning models to predict the potential host-pathogen protein-protein interactions is the main proposed research problem in this thesis. With the development of advanced machine learning models, how to construct robust and effective computational models dealing with different characteristics of the HP-PPIs data is the primary task in the following chapters. At the same time, the in-depth knowledge of the potential HP-PPIs lays behind the structural interaction network (SIN) [28]. To achieve this goal, this thesis also initiates an investigation to evaluate the reconstruction of SIN for HP-PPIs. The investigation includes the study of protein structural information from Protein Data Bank (PDB [29]), the interacting interface and domain information from iPfam [30] and 3did [31] databases. SIN is designed to offer an atomic resolution understanding of host-pathogen interactions.

Thus, the main goals of this research include:

1. Review the host-pathogen interactions databases published in the past decade in a comprehensive way;

2. Evaluate machine learning-based computational models for prediction of host-pathogen protein-protein interactions in a systematic manner. Several characteristics that may affect the performance of the computational models are identified, which include the singular information availability, model complexity and imbalanced data set issue.

3. Propose novel machine learning-based computational framework to better improve the

prediction performance of host-pathogen protein-protein interactions. Four computational models are proposed in this thesis, which include conventional and deep learning models, supervised and unsupervised models, for HP-PPIs prediction regarding the aforementioned different characteristics of HP-PPIs data set.

4. Review the state-of-the-art of the SIN reconstruction, which could offer an atomic resolution analysis on host-pathogen interactions.

1.3 Contributions of the Thesis

This thesis is to address the aforementioned goals with the purpose to provide a computational framework for HP-PPIs research. By conducting the researches, the main contributions of this thesis are summarized as follows.

1. This thesis fills the void of a comprehensive review of published databases regarding pathogens study. A broad investigation of published databases regarding pathogens study is presented. The investigation including the analysis of their corresponding data source, target objects, current status and statistical analysis. A detailed statistic analysis regarding selected databases for human-bacterial interactions (HBI) systems is delivered, which involve a cross-check with their biological information. With this regard, we focus the information primarily from the protein aspect since HBI mostly happen between large molecular systems.

2. This thesis revisits the prediction task of HP-PPIs, specifically of human-bacterium protein-protein interactions (HB-PPIs), which is a first endeavour in this area by covering different aspects of HB-PPIs to report a systematic evaluation of different machine learning-based computational models. A broader and deeper experimental framework is designed to tackle the challenges, which explores a variety of feature representation algorithms and different computational models to learn from the data and perform predictions.

3. This thesis proposes novel computational models for the prediction of host-pathogen protein-protein interactions. To build a computational model to predict the potential HP-PPIs, data and the corresponding methodologies are the main problems we first need to figure out. In this thesis, a data set including positive and negative pairs is curated from the verified experiments with selected features. Different information of proteins, including sequence information, gene ontology, human interactome graph and gene expression, are included at the first stage of investigations. Since for general protein-protein interactions researches, most of them are based on sequence information by the reason of its abundant availability, and as a result these researches indicated they showed relatively good performance on a balanced data set. However, for a HP-PPIs data set, a high skewed data distribution sometimes reflect the true scenario presenting in the HPI system thus will be the main issue.

   As machine learning models have become popular on dealing with PPI data[25, 32–35], and nowadays the deep learning technology has also shown its ability to handle the 'Big Data' [36] including the imbalanced ones [37], in this thesis, four computational models are proposed to tackle the HP-PPIs prediction task.

4. A review of the reconstruction of SIN process. Besides sequence information, structure information of protein sequence is another main published, experimental determined three-dimensional (3D) structural data. There is a scarcity of studies based on 3D structural data to provide an atomic mechanistic and high-resolution view of HP-PPIs. In this thesis, a primary goal is to deliver a review of the reconstruction of HP-PPIs SIN process, which could be of potential to reveal more mechanistic patterns of host-pathogen interactions in the future work.

With this research, the ultimate goal of this thesis is to deliver a deliberate and dedicated computational framework to facilitate the study of HP-PPIs research. The bioinformatics researchers would be the first group to benefit from our research. The proposed computational models will be designed in accordance with the interested aspects in data

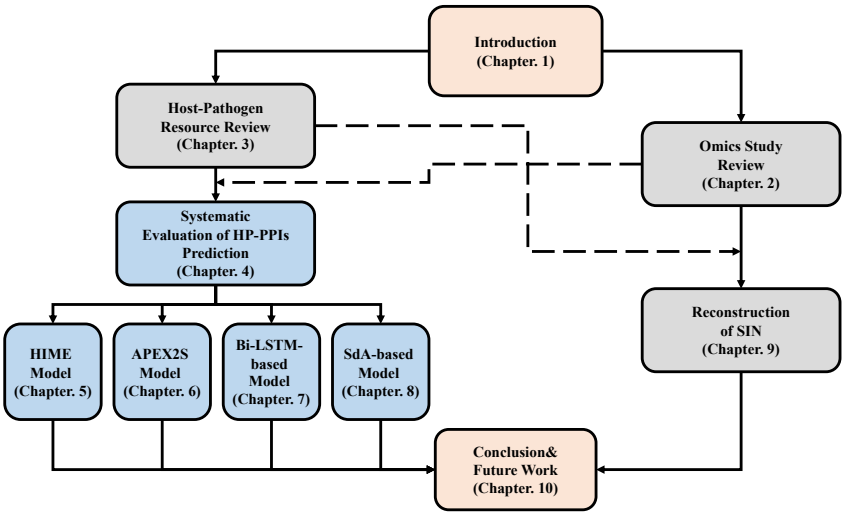


Figure 1.7: The Structure of the Thesis

analytics, which could provide a better insights for molecular level interactions study. Thus, including data scientists in applications modelling, the biologists in proteomics and computer aided medicine area, this research would be beneficial to them.

1.4 Structure and Organization of the Thesis

The remainder of the thesis is structured in accordance with Figure 1.7.

Two major resources review contributions concerning omics data and host-pathogen interactions resources are conducted in Chapter 2 and Chapter 3 respectively. Following in Chapter 4, a systematic evaluation contribution with regard to computational models for host-pathogen protein-protein interactions prediction is presented with the discussion of the latent issues. Chapter 5 elaborates a two-layer model named APEX2s to handle the imbalanced ratio issue for HP-PPIs data set. Chapter 6 propose a novel ensemble model entitled HIME, which is capable of harnessing the power of the heterogeneous information and benefitting from various weak machine learning models. In Chapter 7, an unsupervised deep learning model which is based on stacked denoising autoencoders to capture higher level features regarding the sequence information is presented for discovery of interactions of HP-PPIs data sets. Chapter 8 further explore the power of deep learning model, in which the bidirectional long short-term memory-based (LSTM-

based) model is studied. Several novel designs are implemented in the proposed Bi-LSTM-based model to yield performance results quite smoothly and effectively for HP-PPIs data set. Lastly, in Chapter 9, a conclusion with a preliminary review effort for the reconstruction of structure interaction network is provided, and the future work is discussed.

Chapter 2

BIG DATA IN OMICS DATA RESEARCH

2.1 Introduction of the Omics Data

In recent years, bioinformatics has drawn much attentions from the academia and industry, which demonstrates a strong vision to understand the internal and correlated meanings of different mechanisms of the molecular systems on the Earth, with many advanced tools and in-depth analyses. With the high-throughput technologies, the increasing amount of 'omics' data, including proteomics and genomics, has even further boosted. An upsurge of interest for data analytics in bioinformatics comes as no surprise to the researchers from a variety of disciplines. Specifically, the astonishing rate at which genomics and genetic data are generated leads the researchers into the realm of 'Big Data'. This chapter is dedicated to providing an update of the omics background, particularly focusing on the state-of-the-art developments in the genomics area from the perspective of big data analytics.

2.1.1 History of the Omics Data

The research of omics data is developed for a number of different areas in biology, which is widely studied with the advanced 'omic' technologies for the universal detection of genes, mRNA, protein and metabolites [17, 38, 39]. The omics data research shares a novel vision for analysis of the genome and molecule level data of the biological systems, which is in contrast to conventional biological technology, for example, the genetics [40].

From the World Health Organization (WHO) definition, the genomics data present a more complex, more complicated and more comprehensive view towards the biological systems by scrutinizing the functions and compositions of all genes and studying their inter relationships, while the genetics focus on singe gene [40].

As two important categories of omics data, proteomics and genomics have gained lots of attention in life science. While genomics is the study of the functions and composition of genes, proteomics is dedicated to the research of the functions of all expressed proteins [17]. Proteomics is considered as important as genomics, since the sharing and integrations of proteomics and genomics data will yield substantial improvements and meaningful reference for both gene and protein functions and properties [38, 41, 42].

In this section, we will start with the study of genomics data. The study of genomics started in the early 1990s when the Human Genome Project (HGP) launched its research on a complete sequence of all three billion base pairs in the human genome. The experimental genomics data, which provides the veracious data of life at the molecular level, promises to revolutionize the way in which cells and cellular processes have been studied [43]. The Human Genome Project was designed as a three-step program to produce genetic maps, physical maps and then the complete nucleotide sequence map of the human chromosomes [44]. Besides the sequencing and genotyping technologies development in the past decades, computational biology has become intrinsic to modern biological research [45].

The dominant contribution of HGP is the generation of large, publicly available and comprehensive genomics data [45]. On April 14th, 2003, the USA's National Human Genome Research Institute (NHGRI), the Department of Energy (DOE) and their partners in the International Human Genome Sequencing Consortium announced the successful completion of the Human Genome Project within the state-of-the-art technology [46]. Not only the human beings, but also other species are being sequenced. In 1995, the first bacterium genome sequence was completed, namely Haemophilus influenza. The second species being completely sequenced was Saccharomyces cerevisiae, one kind of beer

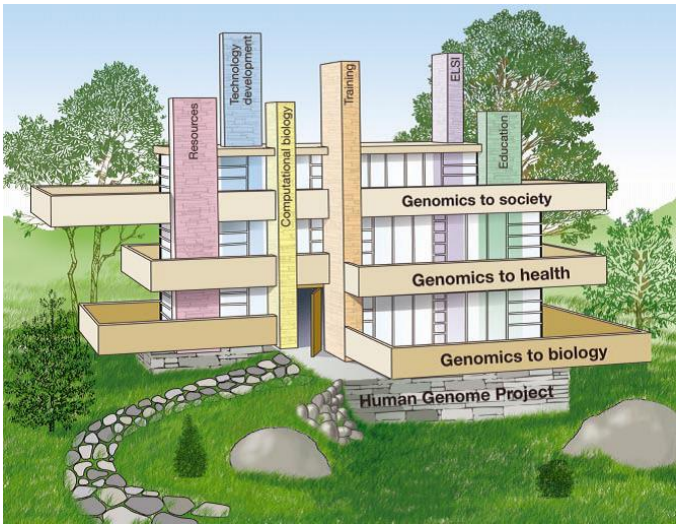


Figure 2.1: The future of Genomics rests on the foundation of the Human Genome Project [45]

yeast, in 1996. In 2000, Drosophila Melanogaster, a famous fruit fly, has its full genome sequence completely finished. The latest sequenced species in records is Zebrafish, which was finished its sequencing results in 2013.

So far, more and more different types of life on Earth are being sequenced, which means more and more proteomics and genomics data have been recorded. As shown in Figure 2.1, it details the future of genomics firmly resting on the foundation of HGP [45]. Three themes are presented: the genomics to biology, the genomics to health and the genomics to society. There are six critically important components relevant to the themes, which are resources, technology development, computational biology, training, ELSI (ethical, legal and social implications) and education.

It has been a promising research area which integrates computational and experimental technology components [45]. The emergent availability of massive biological data has demanded to involve a bunch of computational technologies, including the big data analytic tools, data mining and machine learning, to cooperatively handle these data. How to address the computational technology towards developing data-driven decision support systems, in order to help biologists either design further experiments or conduct data analysis, is the key issue in the next generation biomedical research.

2.1.2 Genomics Data

With the impressive cost drop in high-throughput instruments, now there are many biology laboratories being able to produce the data as quickly and vastly as they want. Comparing genomics data with other major areas of 'Big Data', such as proteomics data, astronomy data, particle physics, website resources (such as YouTube, Twitter) and so on, it is very critical to have an insightful view about genomics via big data analytics [47], as the genomics data is being produced at an extraordinary speed and has its specific domain knowledge.

Every year over 25 zeta bytes (ZB) is being produced in astronomy area [47]. Same phenomenon happens in particle physics which produces massive quantities of raw data. However, only very few data is kept for storing and further analysis after data cleansing and preprocessing.

With regard to genomics data, around 1 ZB data is generated annually. There are more than 7,000 recorded high-throughput instruments all over the world. These instruments are located in nearly 1,000 sequencing centers [47]. It is estimated that over the next ten years, the sequencing genomics data of over 1.2 million reported species of plants and animals would be encompassed.

Genomics data refers to the genome and DNA/RNA data of the organism. Typically, it is the representation in an alphabet array for every sequence. It is a chemical and mechanical process essentially to 'digitizes' the information present in DNA and RNA. Beside these data, other available omics data, which include transcriptomics, methylomics and metabolomics data, could be integrated hierarchically to improve our further understanding from the genotype to the phenotype [48]. Either for considering individual data type for specific domain study or integrating related data types for knowledge discovery between different domains, a data-driven framework built upon a comprehensive representation of biology is desired to ease the upsurge of data and facilitate the bioinformatics research.

For example, in one of our work, considering the proteomics data is publicly avail-

able and is an expression of genomics data, we had drilled the big data analytics into proteomics area to facilitate the experimental research of biologists. To be specific, among the proteomics research, direct benefit from proteomics would be infectious diseases. Thus, in this work, where host-pathogen protein-protein interactions (HP-PPIs) is considered as the key infection process at the molecular level, a proper representation of the proteomics data would introduce high dimensionality issue, while the highly skew ratio between positive and negative HP-PPIs exist in a big dataset [49]. The highly skew ratio is normally set to be 1:100. Considering the variety of infectious diseases and the rising number of proteomics data, a powerful and comprehensive model is desired in this area to help biologist to analyze these proteomics data.

These omics data, including proteomics, genomics and so on, have revolutionized the system biology for a better understanding of biological mechanisms [8]. Bottlenecks and opportunities are posed by a growing gap between the abilities in generating and interpreting these data. The cost and difficulties in quantitative experiment have been relatively controllable nowadays, whereas the challenges are further extended in data analysis stage, which involves the process of data management, data integration, data analysis and data interpretation [50]. Now, it has become even more challenging, as recently precision medicine is gaining intensive attentions, the cooperation of big data analytics with researchers on personalized medicine has also becoming very promising.

2.1.3 Challenges Ahead

While the extensive specialized analyses are required when data is becoming extremely large, different big data areas have different domain knowledge. The interpretation of genomics sequences and analysis of DNA expression, and the research of mutations and developments at the molecular level are the main vision of the genomics [47]. Incorporated with big data analytics technologies, an integration of biology domain expertise, data science, machine learning and even an infrastructure with powerful computation capability are demanded to achieve these goals.

There is no clear consensus among and within biologist and bioinformatics researchers nowadays to best describe the process of leveraging the available omics data to interpret such a domain knowledge, which could be either discovering previously unknown insights or looking for specific patterns [51], such as recognizing the locations of transcription start sites [52]. Today, many research institutions and companies are utilizing their specialty domain knowledge to define and explore their own big data solutions for analyzing these omics data for a further research and application [53, 54].

Since profiling the genomics data is no longer a bottleneck for biology study, an efficient framework for data storage, transfer, and analysis is desired. Unlike the traditional dataset, a single genome sequencing file could be several gigabytes, meanwhile the worldwide distribution of the high-throughput instruments would have facilitated the research on formulating a fast and qualified system for cooperation. These specifications in genomics areas call for more considerations in data acquisition, data transfer, data storage and data analysis.

Next section would provide an in-depth view of the genomics area and its knowledge delivered by cooperation with Big Data analytics technology. In the third section we will detail the current research on data science in genomics area.

2.2 Domain Knowledge Driven by Genomics Data: In-Depth View

The general definition of 'Big data' falls in using inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large dataset to reveal relationships, dependencies, and to perform predictions of outcomes and behaviors. By now, the DNA data deluge comes from thousands of sources. More than 7,000 sequencing instruments are dispersed around the world generating genomics data and sooner or later there will be tens of thousands of the profiling instruments. As a consequence, both the storage and computation burden have been increasing dramatically. In spite of these challenges, how to narrow the gap and

build an efficient connection between the genomics data and the domain knowledge we want to discover is an urgent research problem. Precision medicine and cancer genomics are two major sub areas, which we would like to discuss in this section.

2.2.1 The Knowledge for Precision Medicine

As the genomics data piles up with an extraordinary speed and volume, biomedicine area is increasingly turning into cross disciplines of data science [51, 55, 56]. Specifically, it delivers a promising fortune towards precision and personalized research, which means a P4 medicine: predictive, preventive, participatory and personalized [57].

On January 20, 2015, US President Barack Obama announced a speech to launch a new Precision Medicine Initiative, which brings a closer look to curing diseases like cancer and diabetes. The ultimate goal is to generate a medical solution according to the personalized information to keep the human body healthy. According to the definition of precision medicine in [58], besides the other biological databases, it is important to consider individual information to pose a possible precaution and treatment solution against diseases. Even though the development of high-throughput technologies has lowered the cost of data acquisition, the development of electronic medical system is still on its early stage for data acquisition. Currently, there are two main components being discussed in precision medicine: a short-term goal in personalized therapeutic solution for specific disease and a long-term goal in knowledge extraction for better health [54]. A basic framework of personalized medicine, as shown in Figure 2.2 was proposed.

The accumulated genomics data also stimulates the development of system biology, which is an integrative research strategy for tackling the complexity of biological systems and interpreting their behavior and interactions across all organization levels [59]. The precision medicine benefits from the overwhelming medical data, which establishes a new link between genes, biologic functions and the related diseases [59–62]. Analogous to the proteomics area, assembling genomics data in system biology could deliver a trustful graphical representation of biological interaction maps, and further compute a predictive

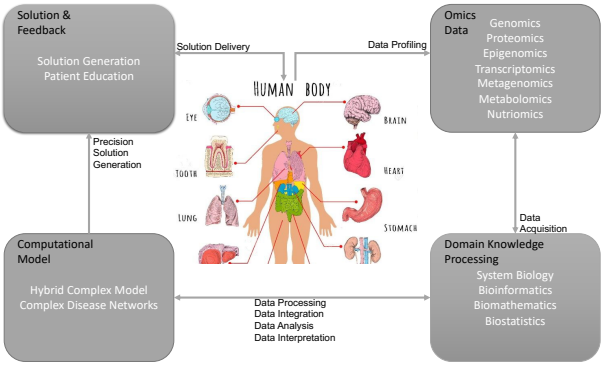


Figure 2.2: A basic framework of personalized medicine [8]

and dynamic model of organisms and diseases. The advancement in identifying the interactions between proteins reduced the false positive rate and improved the quality of curated data sets [63, 64]. In cooperation with genomics data, a study which utilized machine learning methods to recognize the locations of transcription start sites in a genome sequence [52] has been a great start. Similar studies are supposed to be conducted on splice sites, promoters, or positioned nucleosomes identification [65–68].

The genomics related medicine research has been known as 'genomics medicine' [69], which has a consensus definition —- 'using an individual patient's genotypic information in their clinical care [70]'. However, the approach to an effective precision medicine solution is currently on its very early stage of development incorporating with the genomics data. The private protocol issues would be a hindrance in both the electronic medical system development and genomics data sequencing stage.

Towards a precision medicine solution, not only genomics data would be involved, but also other omics data, especially the electronic medical records. This particular vision provides a hierarchical framework as the physiology and pathophysiology do, in which there is a belief that 'genetic can be used to definitely explain features that our genome might accurately indicate the individual risk of developing diseases' [71]. Some specific examples in therapy related study have been done, such as the discussion of the relevance of CYP2D6 in breast cancer tamoxifen therapy decision [72], which tried to interpret the genotype-phenotype association of cancer.

A rational scheme of precision medicine would require each person's genomics profile, which raises not only ethical or legal issues, but also the modeling, computing and analyzing ability problems. Even though almost 2,000 clinical conditions are achieved with genetic testing nowadays, the effective electronic health records (EHRs) still need to be further developed, in an efficient way, which would accordingly produce a comprehensive and individual-specific data [73]. The ultimate goal for precision medicine would be aiming to deliver an exactly right treatment at a right dose at a right time, meanwhile with minimum illness consequences and maximum efficacy [74, 75].

2.2.2 The Knowledge for Cancer Genomics

Among the overwhelming amounts of genomics data, the big data analytics provides a novel paradigm to retrieve information into the related domain knowledge. Besides the precision medicine area, several other research areas, such as functional traits research [76], rice genome project [77], and plant genome annotation and function prediction [78], have been raised associating with the boosting genomics data. In this section, we will discuss about another major area: cancer genomics, which covers the study of cancer mechanism, mutation prevention and detection, and cancer treatment. As an important step towards precision medicine, cancer genomics study is one of the most important discovery science areas [79]. A proposed paradigm from cancer genomics to precision medicine is shown in Figure 2.3. The gap between the cancer genomics and precision medicine is wide, and bridging this gap is far from straightforward. The major ethical proof, data profiling and annotation, the integration of domain knowledge are the first layer hurdles. Proper patient consents are required to proceed to data generation and computational analyses. Furthermore, an efficient knowledge based system to process data to achieve functional and mechanistic studies is desired.

Since cancer is considered as a disease of the genome mutation, the more the biologists learn from the cancer tumors the more they put the belief in the finding that each single cancer tumor is a representation of one specific set of genome changes. Even though

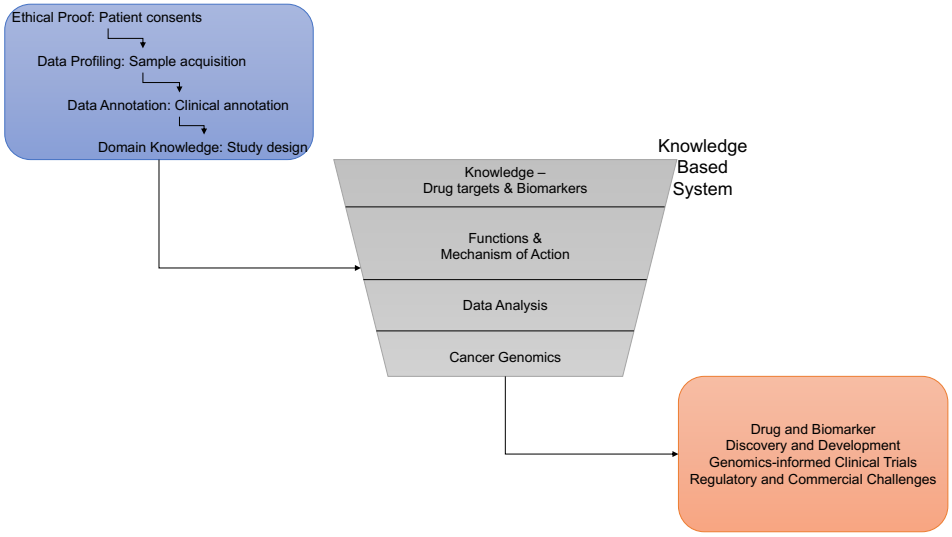


Figure 2.3: From Cancer Genomics to Personalized Medicine [79]

its effects in clinic is currently limited because of the gap between cancer study and therapeutic decision, the cancer genomics is considered to affect every corner of cancer research and would be extended as a critical link for personalized cancer medicine [80, 81].

Most of the data science researches on cancer genomics area are currently conducted on pattern detection problems. Our previous work once aimed to achieve a fast and accurate cancer subtype classification on the genomics dataset. Machine learning technology is the most popular method in classification. Specifically, extreme learning machines (ELM), support vector machine (SVM), general vector machine (GVM) and the state-of-the-art deep learning methods have been deployed to tackle the gene expression data classification problem [82, 83]. In the classification problem of cancer genomics dataset, the small quantity of samples and high dimensionality are two main hindrances for learning model development. As the cancer genomics data piling, a relatively big dataset with high dimension would appear in the near future, which is supposed to be an important but also challenging branch of machine learning application in big data area.

There are two major consortia in the cancer genomics area, which are The Cancer Genome Atlas (TCGA) Research Network and the International Cancer Genome Consortium (ICGC). Both tumor and healthy cells over one thousand patients have been

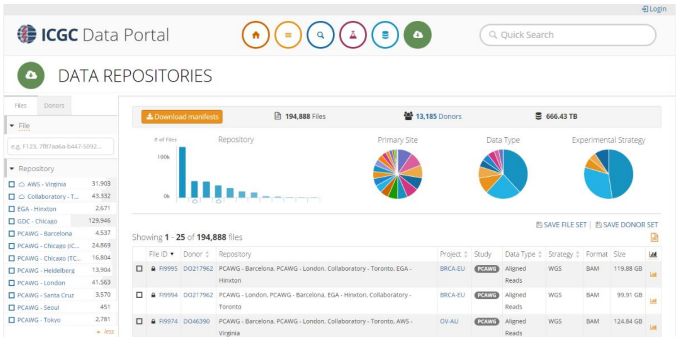


Figure 2.4: A Statistic Diagram from ICGC Data Portal

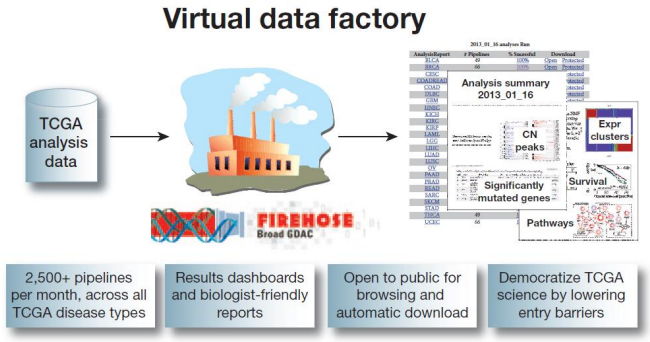


Figure 2.5: Broad's Genome Data Analysis Center: Firehose [53]

sequenced and the molecular differences have been recorded in TCGA across 34 cancer types. These data are currently held at the Cancer Genomics Hub at the University of California, Santa Cruz (UCSC). Also for ICGC, more than 666 terabytes of data has been profiled. The recent ICGC data release is version 21, which contains 68 different cancer projects covering 18,677 donors. These data are housed on separate repositories, such as the European Genome-phenome Archive (EGA-Hinxton), Pan-Cancer Analysis of Whole Genomes (PCAWG), Genomic Data Commons in the University of Chicago (GDC) and so on. As a benefit of the cloud computing technologies, now more and more data are being transferred to Amazon Web Services (AWS). Shown in Figure 2.4 is a statistics diagram of ICGC. Meanwhile the Broad's Genome Data Analysis Center (GDAC) is another genome data center which process TCGA data through their computational framework to generate analysis reports. This pipeline shown in Figure 2.5 in the computational framework is called Firehose.

However most of the ongoing work still focuses on data acquisition and storage. Espe-

cially for some controlled data, the ethical and legal policies still need more consolidation efforts and a proper protocol to process. An in-depth analysis, such as a specific discovery which is previously unreported loss-of-function mutations in HLA-A gene in over 170 squamous cell lung cancers by 'The Cancer Genome Atlas Research Network' (TCGA) [84], has shown the power and importance of network collaboration. Beyond TCGA, these data would need to be more publicly available to researchers all over the world to facilitate the analysis.

With the benefits from high-throughput technologies, cancer genomics is able to compare the genomics sequences, epigenomics profiles and even the transcriptomics data between tumor cells and normal cells [85]. As the increasing researches on genomics aberrations inspire to target on the ultimate goal, i.e. personalized cancer medicine, the future focus of cancer genomics falls on the identification of new genetic aberrations [86], which is the critical aspect in revealing the cancer mechanism. Specifically, as cancer is mostly occurred due to somatic mutations in genome with additional contributions from epigenetic and transcriptomics alterations, one of the in-depth analyses is mainly focused on the somatic mutations in cancer genomics data. This awareness focusing on somatic mutations research has promised us within reach of personalized cancer medicine [87], in which three main challenges are considered as the key hurdles. The first issue is to identify the somatic mutations from the short sequence reads, the second issue is to distinguish the responsible but small somatic mutation for the development and progression of cancer, and the last one is to determine the developing biological pathways and processes which are expressed by these somatic mutations [86].

Along with the studies on cancer mechanism via cancer genomics, the research on cancer treatments is another main area in cancer genomics. Through the enhanced understanding of molecular mechanisms of cancer, it is meaningful to translate the genomics data to improve cancer prevention, early detection, diagnosis, and treatment [88]. This would also be the link between cancer genomics and precision medicine, especially the personalized cancer medicine, in modern oncology. Since very tiny changes

in DNAs and RNAs could possibly introduce large-scale effects on the phenotype [89], the more we know by extracting from cancer genomics, the deeper and closer we are able to get a precision treatment.

Early stage research is ongoing on the area of associating the high or low levels of gene expression with profiles of increased sensitivity or resistance to specific compounds [80, 90]. As TCGA and ICGC are generating an overwhelming amount of cancer genomics data, both whole genome sequencing and targeted genome sequencing are promising to reveal individual genomics variants information [81, 86]. The research on cancer treatments associated with genomics aims to detect the molecularly targeted therapies based on the genomics alterations in patient's tumor, from the perspectives of initiation and progression of cancer [91, 92]. A specific research based on integration of analyzing complex cancer genomics and clinical profiles is introduced in [93]. Focusing on visualization and analysis multidimensional cancer genomics data, [93] provides a portal, namely cBioPortal, to process the overwhelming surge of multidimensional genomics data. Currently the users are able to view some basic patterns in gene alterations across samples in a cancer study, even to link the patterns to clinical outcomes when the related data is available. Yet the future direction for cBioPortal is to include more genomics data types and clinical attributes. The related genomics data types include somatic mutations, DNA copy-number alterations (CNAs), mRNA and micro RNA (miRNA) expression, DNA methylation, and proteomics data. The feature of batch download of complete data sets is also anticipated.

The gap between the study of precision medicine and cancer genomics is wide. Currently, the research strength on translating genomics data from genotype to phenotype could not yet narrow the research gap and bring these two areas together to generate better knowledge discovery. This intrigues the introduction of data science, especially big data related research, into this domain. Focusing on the early stage of big data analytics in genomics area, we would give a discussion about the data management and analysis in genomics data in next section.

2.3 Emerging Big Data Landscape in Genomics

As discussed in the aforementioned sections, to adapt big data analytics technologies in genomics area, a scientific community, which consists of bioinformatics, biomathematics and biostatistics, would be requisite to transfer the genomics data to its biological meaning, which targets on both precision medicine and cancer genomics areas [8]. At the turning points towards a data intensive research in bioinformatics area, we are able to decipher the potential clues on the mechanisms underlying disease initiation and progression, as well as providing further novel strategies for efficient prevention and treatment [8, 50, 94]. Insides these expectations, the efforts in drilling the big data analytic technology into genomics data entails many challenges and future research directions. Although there are very few studies to reveal and establish a general or specific model on discovering inner value out of genomics data for further study of disease mechanism, interventions and treatments, the bottleneck has been shifted from the genomics data profiling to data management, which includes acquisition, transfer and storage.

A basic 'life cycle' of a data set encompasses data acquisition, data transfer, data storage and data analysis. In bioinformatics, the typical initialized data set size was about 2.5 gigabyte in the year 2000, which was publicly available on the file transfer protocol site of the University of California, Santa Cruz [95]. In 2012 the data set size was reported approximately 170 terabyte in the Cancer Genomics Hub (CGHub) [96, 97]. Beyond the size of data set, the computational infrastructure and software tools need to meet the requirement of the analysis tasks. Comparing with the data in astronomy, the data in genomics is much more heterogeneous [47], which brings more challenges when considering that even a single human sequencing genome is around 140 gigabyte in size nowadays.

Utilizing and optimizing the technologies in big data area for genomics require special expertise and experiences in data sciences. As mentioned, data is the key factor to interpret these inner meanings. In this section, the emerging big data landscape in

genomics would introduce several novel ideas to overcome the challenges in dataset transfer, storage and computation.

2.3.1 Data Acquisition

According to the facilities recorded in [88], currently there are 7389 high-throughput 'next-generation' sequencing machines situated in 1027 centers, in which the most machines are situated in the United States (5492 machines). These machines are the main data acquisition access of genomics. Since most machines are located in the United States, these sequencing data are mostly archived in Sequence Read Archive (SRA) maintained by the United State National Institutes of Health National Center for Biotechnology Information (NIH/NCBI). Besides these direct sequencing data, the TCGA and ICGC also archive the cancer genomics data from both tumor and healthy cells. The genomics data are heterogeneous and the research focus of these centers differs with each other. Currently the genomics data are highly distributed and stored in different satellite sites as a consequence of the location distribution.

For the highly distributed data sites, a comprehensive dataset repository in one single site seems to be impossible in a short term. Beside the data transfer to AWS, there is also an ongoing project in ICGC that transfers data from different satellite sites to a single controllable repository, which is considered as a much more efficient way to maintain and distribute the data [53]. However, for other big data areas, the data acquisition accesses and acquisition differ a lot [47]. In the astronomy area the astronomical data is acquired by limited specialist facilities [98, 99], while in the video area most of the video data comes from YouTube streaming clips under several standard protocols. The fMRI (functional magnetic resonance imaging) images are collected with controllable converted formats by some centralized facilities.

Data quality control is an important aspect in these area and genomics data, since these data are generally unaligned and noisy, even missing. The electronic internal fluctuations of the instruments result in a non-consistent performance across the profiling process.

Considering the published data set, the Genomics of Drug Sensitivity in Cancer project, it contains 639 cancer cell lines which are described by a set of genomics features [80]. However, the data missing problem reduce the available training data set from 639 to 608, which results in less data samples. To uncover the knowledge beneath these data, a simple target towards data analysis is not enough since the data consist of multiple levels for their own corresponding meanings: including DNA sequencing data, RNA expression data, miRNA data and so on.

To accommodate these problems, completing the missing data via data analytics method and designing a rational data integration model from multi levels are demanded. A hybrid understanding on these data is critical in the data acquisition stage and may leads to a more meaningful and better knowledge discovery.

2.3.2 Data Transfer

It becomes more and more challenging for a single facility to host its own data on a single machine since the upsurge speed of data is exceeding the Moore's Law. Over the next ten years, the sequencing speed and capacity are expected to grow continually. As collaborations are more common nowadays, the data in TCGA and ICGC are deposited in the corresponding portal and also every collaborator houses their own data. Considering the heterogeneity in omics data, the various communities supported by different foundering agency also generate their own omics data [38]. An increasing motivation to share and transfer the data from the data portal to scientists in a fast speed has been significantly raised.

As a starting maneuver, some ICGC data are deposited in the European Genome Phoneme Archive [53]. Meanwhile each ICGC collaboration country (since PCAWG is distributed by countries) and AWS also house their own data. Yet the network issues have been occasionally occurring and brought the inconvenience for scientists. Thus, now, a centralized database is being built to host all the interpreted data. This centralized database is chosen to be located in the Ontario Institute for Cancer Research (OICR).

With such a strategy that centralized administering data by one single portal site, a faster and more stable connectivity is critical in data transfer. Currently, the Beijing Genomics Institute (BGI) in Shenzhen, China, is able to generate 6 terabytes of genomics data per day. BGI can transfer about 1 terabyte per day to its customer. By exploring a variety of technologies for data transfer over internet, BGI has a vision that their transferred ability could reach 24 gigabyte every 30 seconds when transferring data from China to University of California, San Diego (UCSD) [9]. However this technology, namely fasp, also demands the operators maintaining an extremely large bandwidth which makes the transfer of data an expensive cost in genomics area.

An improvement on internet protocol itself would be a direct solution for big data transfer in genomics, such as Internet2 [100]. Aside from protocol technology, data compression on the DNA sequence reads, specifically in the FASTQ format is another aspect to speed the data transfer [97, 101–104]. FASTQ format is a standard format for storing both a biological sequence and its corresponding quality scores. Another method to boost the data transfer speed would be realized via the efficient data distribution [97, 104, 105].

However, data transfer could be one of the less critical bottlenecks to apply big data analytics in genomics, while data storage strategy is supposed to significantly affect the performance of data processing. Since a single genome data file could be several gigabytes and also the data is highly distributed all over the world, the data analysis neither on the cloud nor the local storage in a raw data format could be limited. This introduces the discussion of the genomics data storage.

2.3.3 Data Storage

Peta byte level storage management is required nowadays to tackle the storage demands in many big data areas. In genomics area, the huge demand for storage mainly comes from the raw genomics data. Since the storage issue has been identified and shifted from the physical storage issues to the data itself, nowadays shipping is still the main method

to transfer large quantities of sequence data [106]. Thus, an efficient method to store the genomics data remains a major challenge for genomics data.

A method which encodes the difference between a logging genome sequence and a recorded reference genome sequence was introduced in [107]. Considering that a single human genome which might occupy three gigabytes of storage, it would be 150 terabyte when it might reach a 50,000 human genomes [106]. Different from traditional data compression algorithms, the bioinformatics utilizes a referential data compression algorithm to avoid a huge decompression time consumption and keep the absolute fidelity of the raw sequence data [47, 108, 109]. A simple example for referential compression sequence is shown in Figure 2.6. A developed algorithm based on this compression schema could reach an evolutionary compression rate of 400:1 or even higher [106, 107]. Shown as Figure 2.6, the reference sequence is set to be 'GCAAAACAAAGT' while normally we used the Revised Cambridge Reference Sequence (rCRS). It is represented by its coordinate positions. For the uncompressed sequence, 'AAAGGCAAAATA', the matches (7,4) and (0,6) indicate the segments of 'AAAG' and 'GCAAAA' by the start position and the length of the segments. The last segment, which is 'TA', is stored in its raw data format since there is no good matching in the reference sequence.

To achieve an optimal compression algorithm and develop it into a standard is a promising effort to facilitate the storage of genomics data efficiently. However, using the compression strategy on genomics data to resolve the data storage problem remains open and challenge for researchers [106]. A balance between compression speed and compression rate is one of the critical issues. Another issue is after the data compression about how to utilize these compressed sequences directly. Despite the data compression aspect in storage, data reduction is also a main aspect in data storage, which introduces great opportunities for a direct understanding of the raw genomics data. As soon as the real-time abstraction method becomes mature, these raw data will be redundant and no longer needs to be stored in their raw representation method.

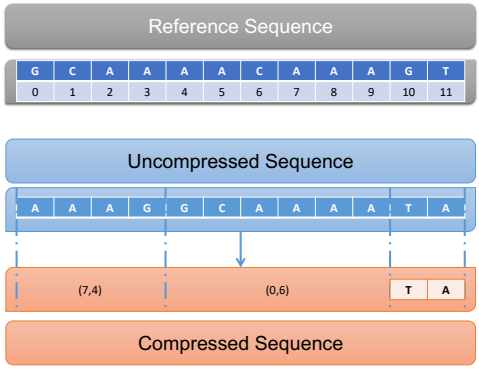


Figure 2.6: A Referential Compression Sequence [106]

2.3.4 Data Analysis

Data analysis is the final stage which matters the most. It is the primary challenge when the researchers aim to learn knowledge from the massive genomics data. A functional data analysis comprises data visualization, data relationship network mapping, data relationship rules extraction and data prediction. Genomics data is heterogeneous and high dimensional, it fits perfectly with the 4 'Vs' definition of big data: which are high volume, high velocity, high variety and high veracity [110–112].

As now data science is flourishing with the overwhelming data, various frameworks and tools have been developed. Taking TCGA as an example, every two weeks the Broad's Genome Data Analysis Center (GDAC) would process the TCGA data by the computational framework Firehose and release a brief analysis report, profiling the significant alterations, and correlating methylation status with clinical features and mutated genes. Meanwhile, another framework, namely SeqWare, takes consideration of a small portion of ICGC data to release a report.

One important aspect of data analysis is data visualization. In the knowledge extraction phase, a useful and important step is offering an intuitive visualization of the genomics data to display the different types of alterations. As long as the visualization techniques are employed in many areas, several tools such as Circos, Gitools, the UCSC Cancer Genomics Browser, the Cancer Genome Workbench, and the cBio Cancer Genomics Portal are developed [85, 113–117]. The visualization techniques offer a visual explo-

ration mostly for the cancer genomics area, in which the concerned data could reveal the cancer initialization genes and pathways. Several examples have been visualized in [78], which are distinguishing the alterations in cancer-driven genome data in tumors, studying the cause-effect relationships between different alteration types data in tumor samples, stratifying the tumor samples based on clinical annotations data, and mapping the global alteration profile patterns on the rearrangement large chromosomal regions data. Visualization of cancer genomics data is critical to translate knowledge of cancer genomics data into a possible personalized cancer medicine, which provides challenges and opportunities for the complex genomics data.

Since machine learning methods have been extensively employed in almost every scientific and engineering area, it has been considered as the next powerful toolbox to interpret the genomics data and act as an important piece of precision medicine [118–121]. An example utilizing machine learning in genomics is to learn to recognize the locations of transcription start sites (TSSs) in a genome sequence [52, 65]. As a blend of machine learning and bioinformatics, it develops into several special learning models considering the application situations in genomics area, including supervised learning and unsupervised learning.

As quoted from the 'No free lunch theorems' [122], there is not an exactly perfect machine learning algorithm working for all applications. In bioinformatics area, especially in the genomics area, the various types of biology knowledge at hand are critical in selecting a proper model. However, mostly it is implicit in mapping the prior knowledge into the framing of the machine learning problem [65]. For example, there was a study to quantitatively link the genomics data with its functional traits by utilizing the whole genome sequence data from the related microbial communities [76]. In [119] both the multi-layer perceptron (MLP) and radial basis function neural networks (RBFNN) have been employed to predict the probability of membership of one individual in a phenotypic class of interest using genomics and phenotypic data.

Along with several other issues, such as handling of heterogeneous data [123–128],

feature selection, imbalanced data sets and the missing data considering different data sources, using the machine learning methods to provide a comprehensive analysis and prediction in genomics area remains challenging, yet promising [129, 130].

In a nutshell, the ultimate goal for big data analytics in genomics area is to be able to interpret genomics sequence, and explain the relationship between genotypes and phenotypes using data. To accomplish this goal, a hybrid understanding and cooperation from different domains, including the data science, computer science, genomics specialist and so on [38, 131–134] are required. In the next section, we would dive into two major projects: ENCODE project and CGHub project, to show how the big data analytics could facilitate genomics research.

2.4 Cases in Genomics Analytics and Bioinformatics

Several researches have achieved inspiring and interesting results from the analyses of big data in genomics. In this section, we will review some state-of-the-art achievements. One is the ENCODE project [131], and the other is the CGHub project [96, 97].

2.4.1 ENCODE

ENCODE (the encyclopedia of DNA Elements) project aims to project all the human genome to their corresponding functional elements. Launched in 2003, ENCODE involved more than 400 leading scientists and processed more than 11,972 files, with a size of more than 15 terabyte. The National Human Genome Research Institute (NHGRI) established a worldwide research consortium.

Started with two phases simultaneously: a pilot phase and a technology development phase, currently ENCODE is on its third phases: the production phase. The pilot phase tested and compared existing methods to rigorously analyze a defined portion of the human genome sequence, while the technology development phase scaled the ENCODE project to a production phase on the entire genome along with additional pilot-scale studies. The report of the pilot phase was published in June 2007 [135]. The findings

highlighted the success of the project to identify and characterize functional elements in the human genome. The technology development phase has also been a success with the promotion of several new technologies to generate high throughput data on functional elements.

The successes of the pilot phase and technology development phase stimulate the NHGRI to fund more studies in order to scale the ENCODE project to a production phase. Meanwhile the production phase starts to include a Data Coordination Center, which is located in the University of California, Santa Cruz, to offer a storage, analysis and service of the ENCODE data. Currently there are over 440 scientists from 32 laboratories participating in the ENCODE project and the tasks are also assigned over different sub groups in the ENCODE Consortium, namely Production Centers, Data Coordination Center, Data Analysis Center, Computational Analysis Awards, Technology Development Effort.

The pilot phase targeted to identify gaps in current tools and data for detecting functional sequences, and also evaluate the efficiency of the available methods in a large-scale scenario. This phase involved both computational and experimental methods to annotate the human genome. The findings promoted the knowledge of human genome functions [135]. The targeted 1% of the human genome were studied from multiple and diverse experiments. The genome transcribed process, transcriptional regulation, a sophisticated view of chromatin structure, and data integration for new mechanistic and evolutionary insights of human genome functions, were reported. The pilot phase helps defining a more comprehensive pathway to understand the functional elements of the human genome.

Since September 2007, the Production Phase was initiated in ENCODE project. As a benefit from the pilot phase and technology development phase, an organized framework for genomics study was established, in which raw sequence data acted as the bottom layer with the annotation layers above [137]. The data model has facilitated the research on knowledge mining of the human genome [131, 138–143]. As the data is continually

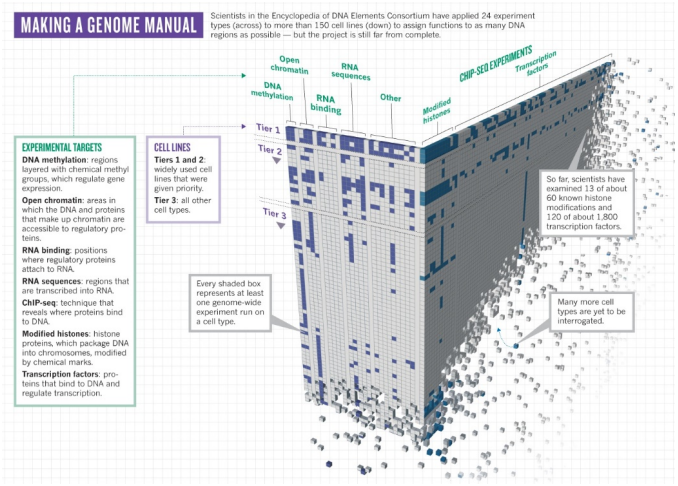


Figure 2.7: A diagram of ENCODE Project [136]

accumulated, the real improvements start when the various data sets are layered together [136] to tackle much more complex genome mechanisms and diseases. Figure 2.7 shows a diagram of ENCODE project. Currently, 13 of 60 known histone modifications and 120 of 1800 transcription factors are examined, which benefits a lot for the complex genome mechanisms study about the genotype-phenotype relationships. The view of genomics data from biologists side has been changed and revolutionized towards a data intensive research when various data are tiered together in ENCODE project.

As the ENCODE project is currently on its high way to the discovery of the functional elements of the human genome, the sub group ENCODE Data Coordination Center (DCC) plays a key role in this project. A well organized, data transfer capability and well developed data visualization tool are the basic demands in the ENCODE consortium. An available ENCODE data site on UCSC Genome Center is http://genome.ucsc.edu/ENCODE/. For cancer genomics research, another site named Cancer Genomics Hub in UCSC have already imposed massive impact towards overcoming the cancer through the power of torrential data [96, 97].

2.4.2 CGHub

Under a contract with the National Cancer Institute (NCI), the Cancer Genomics Hub (CGHub) is an online repository of the sequence data, including the Cancer Genomics

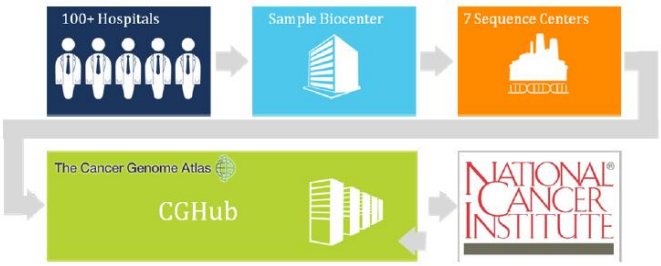


Figure 2.8: General TCGA data flow in CGHub [96]

Atlas (TCGA), the Cancer Cell Line Encyclopedia (CCLE) and the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) projects. Among the repository there are more than 1.4 petabyte data.

Shown in Figure 2.8 is the general TCGA data flow. Cancer genomics is the main focused area in CGHub. Considering data acquisition and data transfer issues mentioned in section 3, a specially enhanced protocol and well-designed data organization method have been developed.

To achieve a higher and better network service, CGHub utilized the Annai GeneTorrent (GT) protocol. It is an enhanced version of the BitTorrent (BT) protocol. Combined with the IBM General Purpose Filesystem (GPFS), CGHub is able to transfer data in a highly parallel and secure mode.

Since the data storage on CGHub is mostly patient-derived cancer genomics data, it is highly confidential. Only the authorized researchers are able to access the data. In the system design phase, CGHub deployed a separate authentication and authorization component solution which is a single-sign-on (SSO) architecture, and the full authorization is under control of the NCI appointed Data Access Committee (DAC).

To be a secure repository for the cancer genomics data, both the storage and transmission need to be encrypted. In CGHub, the SHA-1 (160 bits) hash and encryption are implemented for each single genomics sequence file. The genomics data are stored under the definition of the Sequence Read Archive Metadata XML schema, which is popular in the cancer genomics community. Including the available commands and interfaces, CGHub is an integrated system to provide confidential and interact service for cancer

genomics researchers. As an extension of future development on CGHub, the expansion of data acquisition and storage issues are the promising research areas. Besides these, more help will come from the efforts on data transfer, such as deploying the INTERNET2 technology to increase the internet speed [53].

However, to address a possible solution on either precision medicine or cancer genomics, not a single site or single technology would be able to achieve them all [144, 145]. DISSECT is now able to analyze a wide range of genomics data using the distributed-memory parallel computational architectures of computer clusters [144]. Even though the data are under restricted conditions, DISSECT shows an ability of achieving same performance on large sample sizes. From the data sharing aspect, an omics data sharing mechanism is inevitably needed in the long run [146]. The genomics data are stored worldwide in many data centers. To reveal the genotype-phenotype relationships, the BD2K architecture is proposed to combine the separate genomics data repositories and deliver an open source software stack [146]. A cohesive genomics informatics ecosystem is desired and developing very quickly.

2.5 Summary

To utilize the big genomics data is challenging for our life and also research from every aspects. The life science, biomedicine and health care sectors are currently at a turning point into a data intensive science with the benefit from the overwhelmingly available data. When we are talking about big data analytics, the vision is not only about a research output but also the economic outcome and other benefits, specifically concerning the human life. The genomics data leads us to a new era to play with heterogeneous data and domain knowledge in order to extract insightful knowledge for improving a better life.

As an emerging big data area, the knowledge discovery process of genomics data not only requires abundant data but also leverages the corresponding domain knowledge. In this chapter, two main concerning areas are discussed: precision medicine and cancer

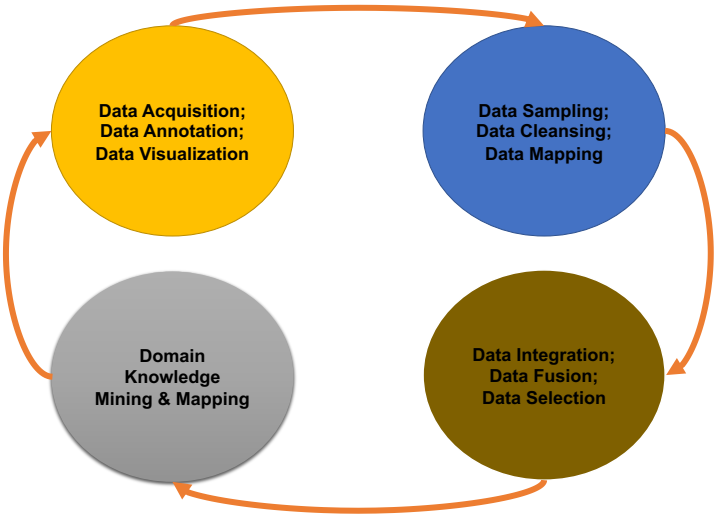


Figure 2.9: Proper Framework for Knowledge Discovery in Genomics

genomics. There is a scarcity of studies on the well-designed framework by now, which is both time-consuming and costly. A hybrid education and cooperation is highly demanded to leverage the data. Figure 2.9 shows a basic framework for data science application. Several aspects must be considered during the research development, which are interpretability (being able to interpret the data clearly), reproducibility (could be mirrored to other researches), simplicity (ease to deploy), affinity (efficient utilization of the computation power).

Besides the domain knowledge involved in this chapter, we have reviewed the current international efforts in the big data analytics in genomics data. In the big data analytic, data matters the most, which introduces the issues of acquisition, storage, transfer and analysis. As long as an urgent desire for efficient data operations before the specific analysis, the data operation problem is considered from several aspects: data acquisition, data transfer and data storage. The highly distributed and heterogeneous characters of genomics data result in the specific requirement for data integration. Since both structured and unstructured data exist in genomics area, an analysis either on the cloud side or in the local system involves a hybrid understanding of the cross-disciplines areas.

We have also introduced some of our work [49, 82] in big data analytics on genomics and proteomics. The ENCODE project and CGHub system were presented to give an

understanding about how we take care of genomics data and how the data is revolutionizing our understanding of life. Technically, the legal and ethical issues are the first to be considered in the genomics area. Beyond the further research of the genomics data, a basic pipeline to deal with the data operation issues (focusing on data acquisition, transfer and storage) and also a general framework towards data analysis are desired to facilitate the international cooperation and research.

We have just reached a turning point towards the data intensive life and research. Among these complex and unknown data, big data analytics has the potential to deliver a better understanding and improvement of our life. As in a nascent stage, the combination of big data analytics technologies and the surge of veracious data entail a lot of challenges and research visions.

Chapter 3

LITERATURE REVIEW OF HOST-PATHOGEN INTERACTIONS RESOURCES

As an important research topic towards the understanding of infectious diseases mechanisms study, the study of host-pathogen interactions has been a hot topic for decades. In this chapter, the goal is to conduct a comprehensive literature review related to host-pathogen interactions, particularly focusing on the resources which are collectively published in last two decades. A background of the host-pathogen interactions resources and a summary of the contributions is presented in Chapter.3.1. A wide range of topics of host-pathogen interactions will be included in the review of the resources in Chapter.3.2. Furthermore, Chapter.3.3 will introduce several standards and tools published in the aim of facilitating proteomics research and development. Later on in Chapter.3.4 and Chapter.3.5, both the statistic report of the curated human-pathogen interactions database and the primary categories of bioinformatics tasks of host-pathogen interactions study will be reported to give the details of the current status of human-pathogen interactions resources by collectively analysing the selected databases.

3.1 Introduction

3.1.1 Background

The study of host-pathogen interactions has been a hot research topic dedicating to the researches of infectious diseases mechanisms, which result in millions of illnesses and

death worldwidely [147–149]. Ranging from various aspects of available 'omics' data, theses host-pathogen interactions (HPI) are accumulated in an extraordinary speed with the development of high-throughput detection methods, in which one of the dominant sources is protein interactions. It presents both opportunities and challenges towards infectious mechanisms study with the benefit of enormous data being generated by biologists.

Since pathogens may vary from fungi pathogen, to virus pathogen and bacterium pathogen, the host-pathogen interactions include interactions between proteins, nucleic acid sequences, metabolites and small ligands [150–153]. The continuing researches elucidating the response to invading pathogens or the cause of concomitant and antagonistic processes with host immune-defence systems show complex and dynamic interaction networks between host and pathogens [154, 155]. Recently, both *in vivo* and *in silico* methods have particularly examined the protein-protein interactions between host and pathogens (HP-PPI) and revealed that the outset of HP-PPI governs the infectious mechanism of most host-pathogen interactions system [24, 148, 156–158].

The computational analysis of interactomes is of critical meaning to model the host-pathogen space, which consolidates the prediction of possible pathogen interactors (e.g. effector proteins) and knowledge generation of prospective host binding strategies [159]. Although more and more host-pathogen interactions data have been verified by experiments, the high cost of *in vivo* and *in vitro* experimental approaches and their high false-positives rate determine a fact that bioinformatics approaches towards obtaining and understanding host-pathogen interactions is deemed essential.

There are three major components in the construction of whole life-cycle study of HPI, including accessible databases, designed bioinformatics approaches and statistic analytic strategy for hypothesis examination and knowledge extraction. We, in this chapter, anticipate to contribute the inter-disciplinary studies with specific interest and focus on understanding of HPI from both computer science and biology sides. Numerous bioinformatics approaches designed for HPI will be discussed in this overview, whilst the

analytics strategy for HPI is also included.

3.1.2 Contribution

With regard to building accessible HPI databases, there have been efforts from the academia researchers, among which several HPI resource systems have been actively updated, such as The Pathosystems Resource Integration Center (PATRIC) [160], the pathogen-host interaction search tool (PHISTO) [161] and so on. Some of them contains only experimentally verified data and some others may include mixing results from both literature and computational prediction. The computational prediction approaches for HPI are generally categorized based on two different ways. One is by the study objects, which include the protein-protein interactions (PPI), domain-domain interactions (DDI), and mRNA-peptide interactions. Another one is the bioinformatics approaches, which include the machine learning-based method, text mining-based method and so on. In this way, the contributions of this chapter are summarized as below:

- A broad investigation of published databases focusing on the topic of pathogen study is presented. The investigation including the analysis of their corresponding data sources, pathogen types, the database current status and the statistic analysis and so on.

- A detailed statistic analysis regarding selected databases for our subsequent research topic is delivered. A general analysis concerning the host-pathogen interactions human-bacterial interactions (HBI) systems is delivered, which also involve a cross-check with their biological information. This chapter focuses the information primarily from the protein aspect since HBI mostly happen between large molecular systems.

- Bioinformatics approaches for HPI study, including task requirements and different prediction strategies towards prediction and analysis, are also included in this chapter. It is anticipated that this part would be helpful on designing computational

methodologies towards a completing analysis for HPI in the future.

This overview is structured as follows. Firstly, the resources of currently available host-pathogen interactions databases will be summarized, and the discussion will focus on the specification of each databases and report the statistic analysis to extract a potential solution for future HPI databases design. Second, a set of bioinformatics approaches for HPI studies is elaborated, which includes homology-based methods (i.e. for bacterial transport-systems) and machine learning-based methods (i.e. from sequence information). The gap to constructing full map between biology experiments and computational approaches is discussed in this part. In third part of this chapter, the focus will be on the analytics strategy for HPI which shows the inherent source to stimulate HPI network, atop of which how to integrate various data to complement the HPI network is presented.

3.2 Host-pathogen Interaction Resources

3.2.1 History of HPI Resources

To encompass the study of HPI, the efforts of initial development of online HPI-specific databases and repositories are being continuously conducted by the researchers. Though the interests of each HPI-specific resources vary a lot, the development of the resources facilitates HPI studies and allows multidisciplinary collaboration [162]. There are numerous HPI resources published in the literature (Table 3.1). These resources were filtered and manually examined with the 'Abstract' from the first 400 results provided by the NCBI PubMed searching engine with *best relevance* ranking out of more than 4,000 returning result items, which were searched with the keywords 'pathogen' and 'database'.

These efforts and developments mostly benefited from the results of a strategic plan initialized by the National Institute of Allergy and Infectious Diseases (NIAID), which focused on biodefense research to define the 'Priority Pathogens' and to develop a subsequent watch list of genera [160, 163]. There have been several initial developments wholly

or partially funded by NIAID, including the pathogen interaction gateway (PIG [164]), BioHealthBase [150], The Pathosystems Resource Integration Center (PATRIC [160]), The Virus Pathogen Database and Analysis Resource (ViPR [165]), VectorBase [166], The Eukaryotic Pathogen Database (EuPathDB [167, 168]). These efforts consolidate and facilitate the understanding of host-pathogen ranges [169] to elaborates local defence mechanisms [27] and a spectrum of diverse discordance in outcomes [170]. The host-pathogen ranges contain a set of species such as eukaryotic pathogens, fungi, virus, protozoa and bacteria. These ranges are somehow identified as the specific contributions from these developed resources.

3.2.2 Review of HPI Resources

This section herein start with reviewing these public databases in Table 1. The web-based database with massive annotated records for pathogen research can be firstly found in the Ecological Database of the World's Insect Pathogens (EDWIP) [169]. As a searchable database majoring in insect pathogens, EDWIP has a foundation of association records of infection between a single host specie and a single pathogen specie. The one-to-one interaction relationship is defined as an association record, which summed up as a result of over 9,400 records between 4,454 host species and 2,285 pathogen species when EDWIP was released. Though it is now no longer available, it shows a particular interest for pathologists and ecologists presenting literature records more dynamically and more precisely. The data in EDWIP are dominantly taken from literature and reports, including books, journals, dissertations from various sources.

MvirDB [171] is termed as a microbial database for protein toxins, virulence factors for biodefence systems. MvirDB solicited most of the data resources from eight public-access databases, which comprise the known protein toxins, virulence factors and antibiotic resistance genes. It is a centralized resource gearing with extensive functions, such as allowing user to search for entries in MvirDB for similar sequence. The data in MvirDB are synchronized weekly from these eight databases and annotated with a developed

parser.

The Host Pathogen Interaction Database (HPIDB) [10] and HPIDB 2.0 [172] refer to two iterated versions of HPI databases pointing at one same hyper URL address http://agbase.msstate.edu/hpi/main.html. Both of them feature the service to provide unified resource for host-pathogen interactions. This data resources were firstly implemented with downloading and parsing several public-access databases. One major update in HPIDB 2.0 is the inclusion of manual biocuration of HPI from literature. It expands the scope from simply looking into existing databases to developing a community annotation data system, which allows a more comprehensive integration of HPI data from a wide range of hosts and pathogens.

Viral Protein Structural Database (VPDB) [173] summarizes the viral proteins with the related structures. Its warehouse maintains viral proteins structures annotating with detailed binding interaction information. Its motivation was to deliver a comprehensive dataset with both sequence, structure and interactions information. As of its release date, it hosted more than 1670 viral protein structures.

The Pathosystems Resource Integration Center (PATRIC) [160] targeted on all bacterial data types in its current incarnation for all NIAID priority pathogenic genera. The related data types include PPIs, genomics, transcriptomics, three-dimensional protein structures and sequence data. This relational database jointly integrates analytic and visualization tools, such as BLAST (the Basic Local Alignment Search Tool), to allow experts and computationally 'naïve' users to obtain metadata with interests. The data in PATRIC dominantly come from a number of public-access repositories and are automatically updated monthly following the PSI Common Query Interface (PSICQUIC) [174] service. It was initially built upon several other public archival databases, such as MINT [175], IntAct [176], BioGRID [177] and DIP [178]. The pathogen-host interaction search tool (PHISTO) [161] is another Web-accessible platform for HPI resources. The goal was to access a complete coverage of HPI data. The database is updated monthly.

The virulence factor database (VFDB) [179] provides up-to-date knowledge of viru-

lence factors in bacterial pathogens. It is one of the most important repository for bacterial virulence factors. The latest generation of VFDB hosts both experimentally verified and predicted virulence factors, which are delivered as one core dataset and another full dataset. It is dedicated to facilitating the aid and development from big data analytics.

BioHealthBase [150] is another host-pathogen interaction resources in the context of influenza virus. It was built upon a wide range of host species and influenza virus strains, which includes data imported from both public-access databases and computational algorithms derived data.

The Pathogen Interaction Gateway (PIG) [164] is integrated from a number of public resources, including all experimentally verified and manually curated HP-PPIs. It serves as a centralized database for easy-to-use aim. Each entry in PIG leads the hyperlink to relevant database of interest, such as UniProt database, functional annotations to the Gene Ontology, etc.

EuPathDB [167, 168] originated from ApiDB and expanded to include dominant database resources for several eukaryotic pathogens of different genera. It encompasses both apicomplexan-specific databases and non-apicomplexan pathogens databases to direct an interactive portal for users as well as to generate across-genera orthology research of interests.

VirusMINT [180] specifies virus protein and its interactions with host as the collection. It accommodates all host-viral protein interactions reported in literature based on a structural format following PSI-MI standards. The curation process also solicits data from some other databases: MINT, which also adopted PSI-MI standards as the data management policy.

VirusMentha [152] was established as an update generation of VirusMINT. It established the data collection within IMEx databases and were regularly and automatically updated weekly by capturing the interactions data via PSICQUIC service. VirusMentha captured all published host-virus interactions without considering specific virus strains and host species. In this regard, it achieved a larger coverage of 24 viral families than

VirusMINT.

The pathogen-host interaction search tool (PHISTO) [161] is another web-accessible platform for HPI resources. Its goal is to deliver an available resource to access a complete coverage of HPI data, which is based on monthly data update strategy. It utilised PSICQUIC service to access and extract HPI data from other nine developed databases, which included all data with and without experimental method detection annotation. Currently, it focus on human as the host.

The host-Pseudomonas and Coxiella interaction database (HoPaCI-DB) [181] is another database resource targeting on bacterial infectious diseases. Its data curation and system development are based on the experimentally validated interactions between molecules, bioprocesses and cellular structures. The dominant data sources come from the pathogenic bacteria P. aeruginosa and C. burnetii. HoPaCI-DB consolidates the collection and finding by comprising comprehensive information extracted from the scientific literature. This process is as well processed with the help of experienced biocurators.

The pathogen-host interactions database (PHI-base) [153, 182] is a long-term maintained resource with expertly curated molecular and biological information on genes proven for literature-reported host-pathogen interactions. It covers the information for more than 4,000 genes from over 200 pathogens interacting with 176 host species. Both prokaryotic and eukaryotic pathogens are included equally.

Recently, researchers have distilled the knowledge from related host-pathogen system resources to conquer specific pathogen research issues. Among these, *Penicillium-crop protein-protein interactions* (PCPPI) [183] encompasses the experimentally determined orthologous interactions from available pathogen-plant systems to curate the database. It was established with an initial collection of 439, 904 non-redundant PPIs between *P. expansum* and seven crops including apple, kiwifruit, maize, pear, rice, strawberry and tomato. These interactions were subsequently verified thoroughly with both interolog mapping and domain-domain interactions supporting. As of this collection, it contained

9,911 proteins from *P. expansum* and seven host species.

Particularly, the database resources focusing on host-pathogen interaction are discussed, whilst there are still a number of similar databases. In Table 1, 45 different databases published between year 2002 and 2017 are included. These databases are evaluated from different aspects, which include the data sources, targeting object information, storing data type, the released website link and the corresponding status. Concerning the status of these databases, 29 out of the 45 are still operational. From the development path of database 'DIP' and 'EDWIP' to 'PHI-base', the database is becoming more interactive for the users and the related information is growing abundant as both biological sequencing technologies and computational resources are evolving fast. These databases concern mostly on pathogens systems, which include eukaryotic pathogens and viruses pathogens.

Among these information, one of the most important factors to build a trustable database is the data sources. In summary, there are several different sources. One of the major ways is from literature and domain expert manual verification. Several databases, such as DIP [178], BIND [184] and PHI-base [153, 182], collect the data primarily via this method. Another major way to collect data is from public archival databases. From the literature, we have identified that several databases are dominantly using the public archival databases as the source. Alternatively, several databases use the submission from users as part of the data source while the rest also include novel derived/predicted data as the data source, such as PHIDIAS [185] and PCPPI [183].

In Table 1, a summary for the relationship between different databases is also collected in the last column. The dispersion of data source motivates the ongoing development of new database to offer wider coverage of data information by integrating heterogeneously curated data [186]. From Table 1, a database with relationship 'None' identifies itself as self-sourcing database, which collects data without other public archival databases. As a result of cross-checking of 'Maintenance' and 'Related Databases' information, the following operational databases are selected as our referred databases for curating the HPI

dataset for following research. These databases include DIP [178], Reactome [187], APID [188], IntAct [176], MINT [175], InnateDB [189], PHISTO [161], PATRIC [160], Mentha [186], HPIDB [10, 172], BioGRID [177]. In following section, the statistic regarding these databases will be reported.

Table 3.1: Host-pathogen Interaction Resources (sorted by published date). The information posted in this table were collected in September 2018.

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
DIP [178]	2002	Literature and domain expert manual verification	Target on interactions for major organism and humans	Protein-protein interactions	http://dip.doe-mbi.ucla.edu	Operational	None
BIND [184]	2003	Literature	Target on biomolecular interactions	Biomolecular interactions, complex and pathway information	http://bind.ca	Retired	None
EDWIP [169]	2003	Reports in the worldwide literature	Target on infectious pathogens	Host-pathogen interactions and bibliographical references	http://insectweb.inhs.uiuc.edu/Pathogens/EDWIP	Retired	None
VIDIL [169]	2003	Literature	Target on viral diseases on insects host	Annotated literature for viral diseases of insects	http://insectweb.inhs.uiuc.edu/Pathogens/VIDIL	Retired	None
PathoPlant [190, 191]	2004	Literature	Target on plant-pathogen interaction systems	Plant-pathogen interactions, proteins, microarray gene expression data	http://www.pathoplant.de/expression{_}analysis.php	Operational	None
Reactome [187]	2005	Literature and domain expert manual verification	Target on Homo sapiens	Data portal for pathway and its analysis	http://www.reactome.org	Operational	None

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
APID [188]	2006	Public archival databases	Target on protein-protein interactions	An interactive platform for collecting and analyzing protein-protein interactions	http://bioinfow.dep.usal.es/apid/	Operational	BIND, DIP, HPRD, IntAct, MINT
MPact [192]	2004	Literature and domain expert manual verification	Target on yeast proteins	Data portal for yeast protein interactions	http://mips.gsf.de/genre/proj/mpact	Operational	None
I2D [193]	2006	Public archival databases and novel derived/predicted data	Target on protein-protein interactions	Organism protein-protein interaction network	http://ophid.utoronto.ca/ophidv2.204/	Operational	OPHID (An earlier version of I2D)
MvirDB [171]	2006	Public archival databases	Target on pathogens in bio-defense fields	publicly available, organized sequences representing known toxins, virulence factors and antibiotic resistance genes	http://mvirdb.llnl.gov/	Retired	VFDB, ToxProt, SCORPRION, Prints, TVFac, Islander, VIDA, ARGO

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
PHIDIAS [185]	2006	Literature and novel derived/predicted data	Extensions of BBP (Brucella Bioinformatics Portal) target on 100 pathogens for bacteria, virus, parasite and fungus	Host-pathogen interactions, genome sequences, conserved domains, gene expression data	http://www.phidias.us	Operational	None
MPIDB [194]	2008	Public archival databases and literature	Target on microbial interactions	Known physical microbial interactions	https://www.jcvi.org/mpidb/	Retired	BIND, DIP, IntAct, MINT, MPI-EXP, MPI-LIT
BioHealthBase [150]	2007	Public archival databases, literature and novel derived/predicted data	Target on specific bio-defense and public health pathogen systems	Biological data related to influenza virus physiology and pathogenesis	www.biohealthbase.org	Retired	N/A
VirusMINT [180]	2008	Public archival databases and literature	Target on host-virus interactions, mostly between human proteins and proteins encoded by some of the most medically relevant viruses, following IMEx standards	Host-virus protein-protein interactions	http://mint.bio.uniroma2.it/virusmint	Retired	MINT, IntAct, HIV-1

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
PIG [164]	2008	Public archival databases	Target on most available human-pathogen interaction systems	Host-pathogen protein-protein interactions	http://pig.vbi.vt.edu	Retired	MINT, DIP, BIND, Reactome, Mpact, HPRD, MvirDB
Proteopathogen [195]	2009	Literature	Target on Candida-macrophage interactions	Host-fungi interactions	http://proteopathogen.dacya.ucm.es	Retired	None
EuPathDB [167]	2009	Public archival databases	Target on eukaryotic pathogen systems	Genome sequence, annotation, functional genomics data, pathway and metadata	http://EuPathDB.org	Operational	BIND
HPRD [196]	2003	Literature and extensive experiments	Target on Human	Protein information of human	http://www.hprd.org/	Operational	None
bioDBnet [197]	2008	Public archival databases	Target on presenting ways to work with various databases	Integrating a vast number of biological databases	https://biodbnet-abcc.ncifcrf.gov	Operational	As a node connecting 153 databases for all aspects of biology

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
PID [198]	2008	Public archival databases and manual review	Target on human molecular event and key cellular process	Signaling and regulatory pathways	http://pid.nci.nih.gov	Retired	None
AquaPathogen X [199]	2010	Public archival databases	Target on aquatic pathogens	Aquatic pathogens	http://wfrc.usgs.gov	Operational	None
HCVpro [200]	2011	Public archival databases and literature	Target on interactions between Hepatitis C virus and human	Hepatitis C virus-virus and virus-human protein interactions	http://www.cbrc.kaust.edu.sa/hcvpro/	Operational	MINT, BIND
VPDB [173]	2011	Public archival databases	Target on viral proteins	Viral proteins and 3D structures	http://www.vpdb.bicpu.edu.in	Retired	None
VectorBase [166]	2011	Public archival databases and community submission	Target on invertebrate vectors of human pathogens	Genome sequence, structural/functional annotations and reference, etc	http://www.vectorbase.org	Operational	None
ViPR [165]	2011	Public archival databases, direct submission and novel derived/predicted data	Target on human pathogenic viruses belonging to specific families	Sequence records, gene and protein annotations, 3D protein structures, immune epitope locations, etc.	www.ViPRbrc.org	Operational	None

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
IntAct [176]	2004	Public archival databases and literature	Target on protein-protein interaction data	Molecular interaction database	http://www.ebi.ac.uk/intact	Operational	N/A
MINT [175]	2011	Literature	Target on protein-protein interaction	Protein-protein interaction	https://mint.bio.uniroma2.it/mint/	Operational	None
vHoT [201]	2011	Novel derived/predicted data	Target on the interaction between viral microRNA and host genomes	Viral microRNA and host genomes interactions	http://dna.korea.ac.kr/vhot	Retired	None
InnateDB [189]	2008	Literature	Target on mammalian innate immunity systems	Mammalian innate immunity networks, pathways and genes	http://www.innatedb.com	Operational	None
PHISTO [161]	2012	Public archival databases	Target on human as the host specie	Host-pathogen interactions and human intra-species protein-protein interactions	http://www.phisto.org	Operational	MINT, IntAct, DIP, APID, iRefIndex, STRING, MPIDB, BIND, Reactome
PATRIC [160]	2013	Public archival databases	Target on bacterial pathogen systems	Data portal for bacterial pathogens	http://www.patricbrc.org	Operational	MINT, IntAct, BioGRID, DIP
HoPaCI-DB [181]	2013	Literature	Target on Pseudomonas aeruginosa and Coxiella burnetii pathogens	Host-pathogen interactions	http://mips.helmholtz-muenchen.de/HoPaCI	Operational	None

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
HIV1-HPID [202, 203]	2008	Literature	Target on all HIV-1 and human protein interactions	HIV-1-human protein-protein interctions	https://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/	Operational	None
VirHostNet [204]	2009	Public archival databases and literature and domain expert manual verification	Target on virus-virus, virus-host and host-host interaction networks	Virus-host interaction study with extensive functionality	http://virhostnet.prabi.fr/	Operational	BIND, MINT, IntAct, HPRD, DIP, BioGRID, Reactome, Generif, Networkin
MatrixDB [205]	2009	Public archival databases and literature	Target on interactions for matrix proteins, proteoglycans and polysaccharides	A database for interactions established by extracellular matrix proteins, proteoglycans and polysaccharides	http://matrixdb.univ-lyon1.fr/	Operational	IntAct, MINT, DIP, InnateDB
Mentha [186]	2013	Public archival databases	Target on interactions between proteins	A database including comprehensive resource archiving all published protein-protein interactions	https://mentha.uniroma2.it/	Operational	MINT, IntAct, DIP, MatrixDB, BioGRID

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
VirusMentha [152]	2014	Public archival databases and literature	Target on host-virus and virus-virus interaction systems	Host-pathogen interactions	http://virusmentha.uniroma2.it	Operational	MINT, IntAct, BioGRID, VirusMINT, DIP, MatrixDB, APID
VFDB [179]	2005	Public archival databases and literature	Target on bacterial pathogen systems	The virulence factors of bacterial pathogens	http://www.mgc.ac.cn/VFs	Operational	None
HPIDB 2.0 [10, 172]	2010	Public archival databases and literature	Target on most available PHI systems	Host-pathogen interactions	http://hpidb.igbb.msstate.edu/index.html	Operational	MINT, IntAct, BioGRID, DIP, Reactome, MPIDB, VirHostNet, I2D, InnateDB
SugarBindDB [206]	2015	Literature and domain expert manual verification	Target on carbohydrate sequences binding with pathogenic organisms	Glycan binding of human pathogen lectins and adhesins, functional annotation, 3D structure and binding patterns	http://sugarbind.expasy.org/	Operational	None

Continuation of Table 3.1

Databases	Release Year	Sources	Object	Data Type	URL	Maintenance	Related Databases
DenHunt [207]	2016	Literature	Target on Dengue virus and human interaction system	Dengue-human interactions, genes, pathways	http://proline.biochem.iisc.ernet.in/DenHunt/	Operational	None
STRING [208]	2000	Public archival databases and novel derived/predicted data	Target on protein-protein interaction data	A platform for collecting and integrating protein-protein interactions as well as delivering network analysis	https://string-db.org/cgi/input.pl	Operational	BioGRID and databases organized in IMEx consortium
PCPPI [183]	2016	Public archival databases and novel derived/predicted data	Target on interactions between P. expansum and crops	Penicillium-crop interactions, gene ontology, sequence records, DDI	http://bdg.hfut.edu.cn/pcppi/index.html	Operational	GDR, KIR, maizeGDB, HPIDB
BioGRID [177]	2015	Literature	Target on interctions for major organism species and humans	A comprehensive data portal for protein, genetic and chemical interactions	https://thebiogrid.org/	Operational	None
PHI-base [153, 182]	2016	Literature and domain expert manual verification	Target on most available HPI systems	Host-pathogen interactions, genome information, referred literature annotation and phenotype	http://phi-base.org	Operational	None

End of Table 3.1

Table 3.1 Host-pathogen Interaction Resources (sorted by published date). The information posted in this table were collected in September 2018.

3.3 Available Proteomics Standards and Tools

Despite the traditional published publication which continues acting as the most practical method for disseminating experiment results and conclusions, the experimentally derived scientific contributions are now generating substantial value by acting as important references for building publicly accessible database. Mostly, this process will allow a rich and centralized resource and data portal for researchers [209]. Although the databases are published online and are mostly developed with specific interests, the primary goal is to ease the downloading/searching of data, and to facilitate the communication between biologists. In this sense, the researchers have strived to identify the requirement of the creation of data standards and interchange formats for the database, which is considered to benefit the storage and distribution of proteomics data [209, 210], particularly for our study.

The Human Proteome Organization Proteomics Standards Initiative (HUPO PSI) is one of the voluntary organizations which has developed the HUPO PSI-MI XML as one of the widely adopted data format standards [211]. Meanwhile, with the efforts of community researchers, recent examples of themed curation projects, such as BioGRID [177] and Mentha [186], have taken advantages of the establishment of the International Molecular Exchange (IMEx) consortium (http://www.imexconsortium.org/) [212]. IMEx consortium has released a single joint data curation manual by 2005. In this section, we briefly introduce the available standards with the file format for representing molecular interactions data and the utilised tools in several databases.

- Data Format:

  HUPO PSI-MI XML is a data format initially established by HUPO PSI in 2004. Its generation encompasses the incremental needs of high-quality interaction datasets for biologists [213]. It has taken extensive update from version 1.0 in 2004 to version 2.5 in 2007 and the latest version is PSI-MI XML3.0 in 2018. The updates of PSI-MI XML represent the continuing changes of standard data interchange

format between data producers, data users, tool developers and databases providers.

The scope of PSI-MI XML schema is expanding from the inclusion of simple protein interaction (version 1.0) to a rich description of molecule features (version 3.0). Along the development of PSI-MI XML2.5, the molecular interaction tabular format MITAB2.5 is a simpler data format which provides less detailed molecular interaction data. Currently, both PSI-MI XML2.5 and MITAB2.5 are also extensively utilised by users and data providers. The data format contributes to a systematically description ability for important biological events in molecular interaction data. It will facilitate the data curation strategy and the service development.

- Data Curation Strategy:

  Acting as an international collaboration community of databases providers, IMEx has further extended the accessibility of data based on the common data format HUPO PSI-MI XML. Since the scarce public funding opportunity and different curation strategies, IMEx is framed as a long-term coordination for curation of dataset and avoiding redundant work on same data [212] on a single website (www.imexconsortium.org).

  The data curation strategy is thus designed to align the worldwide databases to a same identifier, which allows user to trace data from both the original resource and IMEx website resource. IMEx website also encompasses the access function of PSICQUIC service.

- Data Service:

  PSICQUIC retains the standard PSI-MI XML format and is designed to be a common computational access to multiple molecular interaction databases. PSICQUIC is jointly developed with PSI confidence scoring system (PSISCORE), which extends the system ability with retrieving confidence scores of molecular interactions [174].

As of this writing, there are over 30 databases supporting PSICQUIC service from http://www.ebi.ac.uk/Tools/webservices/psicquic/view/home.xhtml. Via PSIC-QUIC registry, there are totally over 7,015,614 accessible interactions and the PSICQUIC service could also easily help to cluster and filter these interactions. Since the PSICQUIC service is developed to be programmatic, it can be integrated with other applications with their stylish manner.

3.4 Statistical Analysis of HPI Resources

For a purpose of evaluating the accessible pathogen databases, in this section, an extensive literature review of the databases published in the last two decades has been conducted. The resources were filtered and manually examined with the 'Abstract' from the first 400 results ranking by *best relevance* out of more than 4,000 returning result items, which were searched by NCBI PubMed search engine with keywords 'pathogen' and 'database'. In Table 3.2, partial details of the selected databases is listed.

These databases are evaluated from different aspects, which include the data sources, targeting object information, storing data type and the corresponding status. Concerning the status of these databases, 29 databases are still operational. From the development path of database, such as 'DIP' [178] and 'EDWIP' [169] to 'PHI-base' [182], the database is becoming more interactive for the users and the related information is growing abundant as both biological sequencing technologies and computational resources are evolving fast. These databases concern mostly on pathogens systems, which include eukaryotic pathogens and virus pathogens.

Databases	Release Year	Object	Data Type	Maintenance
DIP [178]	2002	Target on interactions for major organism and humans	Protein-protein interactions	Operational
Reactome [187]	2005	Target on Homo sapiens	Data portal for pathway and its analysis	Operational
APID [188]	2006	Target on protein-protein interactions	An interactive platform for collecting and analyzing protein-protein interactions	Operational
IntAct [176]	2004	Target on protein-protein interaction data	Molecular interaction database	Operational
MINT [175]	2011	Target on protein-protein interaction	Protein-protein interaction	Operational
InnateDB [189]	2008	Target on mammalian innate immunity systems	Mammalian innate immunity networks, pathways and genes	Operational
PHISTO [151]	2012	Target on human as the host specie	Host-pathogen interactions and human intra-species protein-protein interactions	Operational
PATRIC [160]	2013	Target on bacterial pathogen systems	Data portal for bacterial pathogens	Operational
Mentha [152, 186]	2013	Target on interactions between proteins	A database including comprehensive resource archiving all published protein-protein interactions	Operational
HPIDB [10, 172]	2010	Target on most available PHI systems	Host-pathogen interactions	Operational
BioGRID [177]	2015	Target on interctions for major organism species and humans	A comprehensive data portal for protein, genetic and chemical interactions	Operational

Table 3.2: The Resource of Pathogen Databases

Among the information, one of the most important factors to build a trustable database is the data sources, which indicates how is the data derived. In summary, there are several different sources. One of the major ways is from literature and domain expert manual verification. Several databases, such as DIP [178], collect the data primarily via this method. Another major way to collect data is from public archival databases. From the literature, we have identified that several databases are dominantly using the public archival databases as the source. Alternatively, several databases use the submission from users as part of the data source while the rest also include novel derived/predicted data as the data source, such as PHIDIAS [185] and PCPPI [183].

To collectively build a basic pathogen database, normally it is better to have more databases involved in the curation stage. The reason is that, mostly the databases are developed with different specification and they serve for various research interests of pathogens study. However, according to our literature review, it is clearly to see that, computational prediction interactions are as well included in some databases. This work focuses on the experimentally verified interactions, which limits the usage of the databases among those with only experimentally verified interactions, as shown in Table 3.2.

DIP [178], Reactome [187], APID [188], IntAct [176], MINT [175], InnateDB [189], PHISTO [161], PATRIC [160], Mentha [152, 186], HPIDB [10, 172] and BioGRID [177] are included as the databases in our study to build the database and present the inceptive data analysis. All the databases were downloaded on the date of 2018-August-31th.

Since host species are mostly limited within several species including plants and human, pathogen species could be referred to many, such as bacteria, fungi, protozoa, helminths and viruses. Figure 3.1 illustrates the coverage of the pathogen types and the amount of proteins. It is easy to see that, Mentha database has housed the most pathogen types as well as the proteins, whereas the result of IntAct database is the second. In Figure 3.2, the example of pathogen interactions with Homo sapiens (taxonomy ID: 9606) is diagrammed. The data are counted for human proteins in inter-species and intra-species interactions respectively. Most databases provide intra-species interactions for

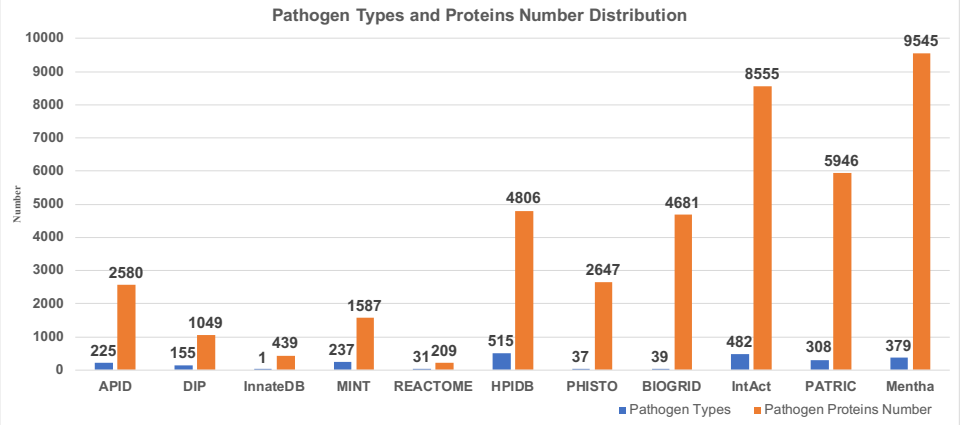


Figure 3.1: The Pathogen Proteins Distributions in Databases

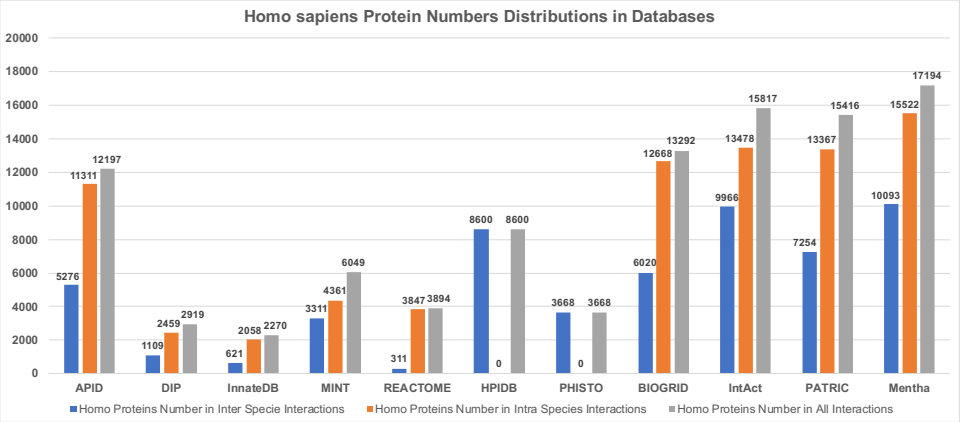


Figure 3.2: The Homo Sapiens Protein Numbers Distributions in Databases

human as the inclusion for researchers. From Figure 3.1 and Figure 3.2, the Mentha database has hosted the most pathogen proteins number as well as the Homo sapiens protein numbers. It is interesting to note that, for some databases such as HPIDB and PHISTO, only the inter-species interactions between human and pathogens are reported. These two databases have a focus on the study of host-pathogen interactions.

The related statistic in Figure 3.3 indicates that, Mentha database covers most of the Homo sapiens interaction information, including the inter species interactions as well as the intra species interactions. Although PHISTO and PATRIC are two databases focusing on Homo sapiens inter species interactions, it will be a good supplementary of Mentha database.

Furthermore, the corresponding pathogen categories within the different human-pathogen

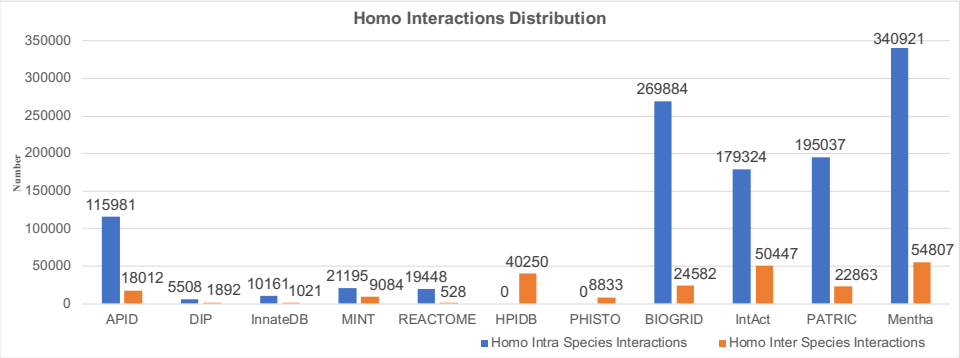


Figure 3.3: The Homo sapiens Interactions Distributions in Databases

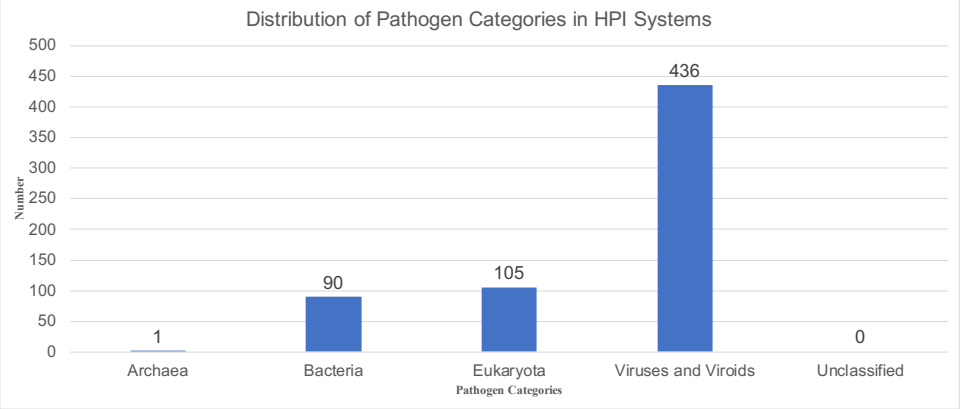


Figure 3.4: Distribution of Pathogen Categories in HPI systems

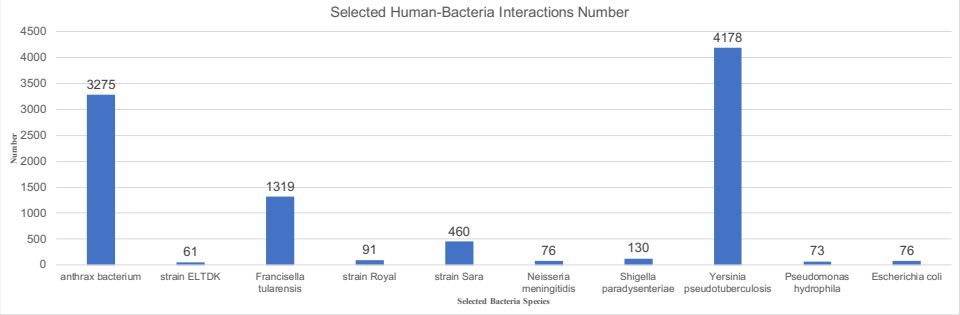


Figure 3.5: Selected Human-Bacteria Interactions from Databases

interaction (HPI) systems are analysed. By combining the 11 databases as one, we show the result in Figure 3.4. There are totally 436 different types of Viruses and Viroids in the database, and it is one of the most studied pathogen categories ranging from these 11 databases. The species number of pathogens in the combined database is 649, which consist 502,635 different intra species interactions for Human-pathogen interactions systems.

With species-specific interest, it is also possible to consider one pathogen types as the analysed subject. Herein, we take bacteria as the selected pathogen category. The bacteria species containing more than 50 interaction pairs with homo sapiens are reported in Figure 3.5. They are collected distinctly with their corresponding taxonomy ID from the database files. This information could help researchers in designing future biological and computational experiments with regard to analyse the internal relationship for pathogenic mechanism studies.

3.5 Bioinformatics Approaches for HPI study

There are two featuring bioinformatics tasks in the pathogenic mechanism studies. One is the secreted system effector proteins and another is the complete pathogen interactions network completion. In this section, we focus on discussing the issues and solutions for the prediction task of host-pathogen interactions network. In Figure 3.6, the experimentally verified host-pathogen protein-protein interactions are collected via literature review, and the diagram is generated with the software of Cytoscape [214]. The green

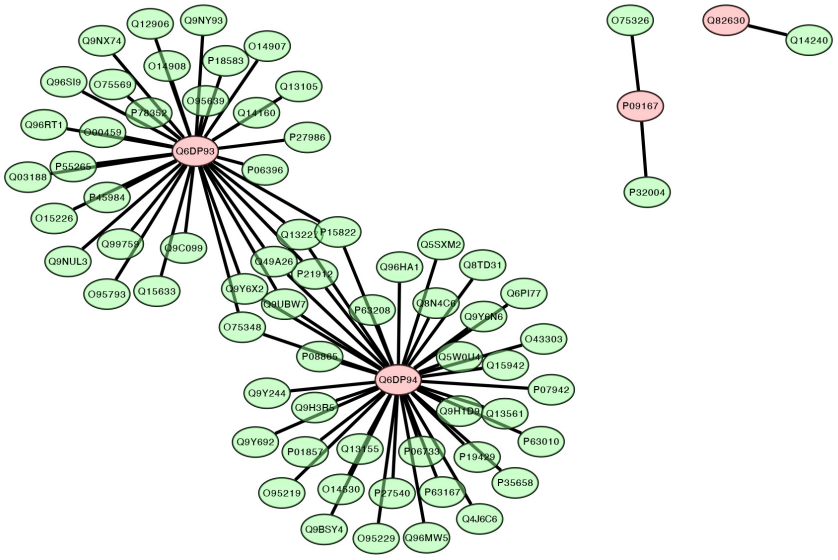


Figure 3.6: The Protein-protein Interactions Network between Human and Influenza A virus

nodes represent the proteins from human, while the pink nodes are the proteins from pathogen of Influenza A virus. It represents a small set of the overall protein-protein interactions network between human and Influenza A virus. The rest interactions network, which is not illustrated in Figure 3.6, could be either unknown or un-experimentally verified. Thus, the prediction task of host-pathogen interactions network will be critical for the researchers to understand the holistic interactions mechanisms between host and pathogen. Generally, this prediction task is dealt with two different methods, which are template-based method and machine learning-based method.

Template-based Method for Prediction

For host-pathogen interactions, it is of biological meaning to predict the interactions with template-based methods, which mostly utilise the homology similarity, structure similarity and domain interactions relationships for prediction. For template-based method, the homologous proteins with known experimentally verified structures and other properties are firstly identified by searching among a number of databases, such as the Protein Data Bank (PDB) [29]. One prominent advantage of template-based method is that, the relevant biological meaning is easy to interpret if the indexed homologous protein

has been systematically studied [215].

However, several shortcomings have been identified as well. One of them is the limited resources with experimentally verified data. In some cases, the proteins might not be able to detect the homologous proteins [24]. The limitation of accessible database has shorten the application of such kind of method [216]. Another drawback of template-based method is the sources of databases. Even though most of the databases deposit the experimentally verified data, their sources are from different experiments circumstances. With the different reliability of data, the template-based method would probably generate different hypothesis and conclusions, which will further demand more other methods for verification.

Machine Learning-based Method for Prediction

Another important category of methods for host-pathogen interactions prediction is the machine learning-based method. Building a data-driven model to predict HPI in a broader range is the motivation of applying machine learning models in prediction task of HPI, since there may be only a small number of templates with biological experiments support and the relation between host and pathogen has been roughly studied.

Applying machine learning model has shown the effectiveness for predicting novel HP-PPIs. Most of the machine learning models, such as Bayesian statistics [217], random forest [218], support vector machine [34, 156] and so on, have been utilised as the primary computational model to learn the internal relationship from protein information and curated dataset. Various sources of protein information have been considered to represent the protein in the curated dataset, while the selected machine learning model would be different due to the studied pathogen species and dataset. The studies of machine learning-based method have also been performed for general intra-species PPIs predictions, which indicates its scalability and effectiveness, with regard to the significant challenges impairing the experiments to develop proteome-wide interactions network. Besides the identified literatures for the prediction task of HP-PPIs, there are numerous

works focusing on the feature representation algorithms for other protein related topics, such as structure, folding topics and so on. To present a comprehensive literature review with regard to the machine learning-based method for prediction, the details of systematic evaluation is included in Chapter. 4.

3.6 Summary

In this chapter, a comprehensive literature review related to host-pathogen interactions resources, which are collectively published in last two decades, was conducted. The resources reviewed in this chapter cover a wide range of topics of host-pathogen interactions in Chapter.3.1 and Chapter.3.2. Furthermore, several standards and tools published in the aim of facilitating proteomics research and development were reviewed in Chapter.3.3. Later on, a brief statistic report of the curated human-pathogen interactions database and the primary categories of bioinformatics tasks of host-pathogen interactions study were elaborated in Chapter.3.4 and Chapter.3.5 repectively, which give the details of the current status of human-pathogen interactions resources by collectively analysing the selected databases.

Following in next section, the research will focus on the task of evaluating the machine learning-based computational models, which covers a broad range of machine learning models and model from literatures, for the prediction task of HP-PPIs.

Chapter 4

SYSTEMATIC EVALUATION OF SEQUENCE-BASED MACHINE LEARNING PREDICTION MODELS FOR HUMAN-PATHOGEN PROTEIN-PROTEIN INTERACTIONS

Developing machine learning models in the predictions task for HP-PPIs has been studied with the interests of its efficiency and accuracy. However, how to select and determine the best model requires a systematic evaluation of different predictors for HP-PPIs. In this chapter, a wide and deep overview on currently available resources and computational tools is reported in Chapter. 4.2. In Chapter. 4.3, a dedicated data curation process will be implemented and a pipeline for HB-PPI studies will be summarized which includes numerous sequential feature-representation algorithms and machine learning models. In Chapter. 4.4, the experimental results of different ratios of benchmark datasets, different feature-representation algorithms and different machine-learning models will presented.

4.1 Introduction

4.1.1 Background

Infectious diseases are predominantly caused by many pathogenic species, such as bacteria, fungi and viruses and so on. These infectious species actively interact with their hosts in a variety of ways, which place the host-pathogen interactions (HPI) in a

complicated, but also critical, role in the study of infectious-diseases mechanisms. In most cases, the host-pathogen system is studied from different perspectives to further our understanding of infectious mechanisms [219]. A major approaches is studying the interactions of inter-species proteins, in which one protein is from the host and the other is from the pathogen.

While protein interactions occur extraordinarily between human and bacterium pathogens, one of the earliest studies illustrated the importance of human-bacterium interactions (HBI) in relation to the symptom cause by anthrax *Bacillus anthracis*. In this study, *Bacillus anthracis* was conclusively demonstrated as the primary cause of anthrax [220]. Additionaly studies of *Bacillus anthracis* were conducted, aimed at fully understanding the mechanisms of a complete protein interaction network between *Bacillus anthracis* (the bacterium pathogen) and *Homo sapiens* (the host) [221, 222]. These studies encouraged researchers to study a broad range of infectious diseases by exploring the human-bacterium protein-protein interactions (HB-PPI).

However, the investigation of HBIs consumes lots of time, money and resources in determining the complete interaction network and understanding their mechanisms. Currently, investigations of the interactions between host and pathogens are still very limited. Even though large-scale biomedical technologies, such as yeast two-hybrid assay and the affinity purifications-mass spectrometry (AP-MS) method, have allowed us to detect the interactions (positive or negative) in a faster and more accurate way, the amount of possible human-bacterium protein-protein interactions is large. Other small-scale technologies, like nuclear magnetic resonance (NMR), are often labor-intensive and time-consuming. Thus, it is critical to formulate a computational model for the prediction of HB-PPIs.

Several reviews studied current computational approaches [32, 151] as well as researches on applying machine learning-based models to predict host-pathogen protein-protein interactions (HP-PPIs) [23, 34, 156, 223, 224]. In particular, how to deploy machine learning-based models as a generic approach in predicting novel human-

bacterium interactions based on sequence information is considered as an important category of research, which involves many challenges and opportunities. However, there is currently not comprehensive evaluation study that has focused on machine learning-based model as the primary computational method and further comparatively evaluated their corresponding performances across a wide range of HBI systems.

4.1.2 Contributions

In this chapter, we follow the previous study of host-pathogen resource review to implement an evaluation protocol for human-bacterium protein-protein interactions study. This study is based on literature reviews by firstly collecting human-bacterium interactions systems data from the mentioned wide range of host-pathogen databases. The systematic evaluation is subsequently achieved from two aspects. The first considered the application of feature representation algorithms to the protein data, while the other was related to different machine learning-based models. Meanwhile, the literature methods on topic of host-pathogen protein-protein interactions is reported.

We summarize the contributions of this study as follows:

- A review on currently available data sources and computational tools is presented. This chapter is based on the investigation of the reviewed databases, while the performance evaluation is carried out among different computational tools and methods from the literature.

- A systematic evaluation of machine learning-based computational prediction is delivered. Although there have been several existing studies reporting the performance of traditional machine learning-based methods on the specific HPI prediction task separately, such as support vector machine, random forest, decision trees and so on, we anticipate to cover a comprehensive study of machine learning models and the feature representation algorithms in this chapter. The evaluation is conducted by reporting multiple metrics and comparing the performance in a substantial manner.

- A pipeline for human-bacterium interactions study is summarized whilst the datasets are also curated for further researches. By building the pipeline for HBI study, we anticipate to answer the following questions:
  - How do machine learning models perform on the prediction task of human-bacterium protein-protein interactions;
  - How do the feature representation algorithms based on sequence information affect the model performance;
  - Do the ratios between positive and negative interactions have impact on the model performances;
  - What are the key issues to be addressed in order to build a robust and effective machine learning-based method for human-bacterium protein-protein interactions prediction.

4.2 Overview of Predictors for Host-Bacterium Protein-protein Interactions

4.2.1 The Overview of Predictors for HB-PPIs Study

Although there has been a long history of research on protein-protein interactions prediction, so far there are only a small number of publications that have focused on host-pathogen interactions reviews [32, 151, 162, 225]. A broad search has resulted in four major review papers, and Table 4.1 summarizes the reviews.

The studies by [162] and [151] have a wide coverage on host-pathogen interactions, which include prediction as well as analyses, while the reviews by [151] and [225] focused on the computational prediction of host-pathogen interactions. These reviews aimed at describing the progress of host-pathogen interactions, without anchors of naming pathogens, and they collectively reported on potential computational methods, such as homology-based approaches, structure-based approaches, domain and motif interactions-based approaches and machine learning-based approaches. Furthermore, no systematic

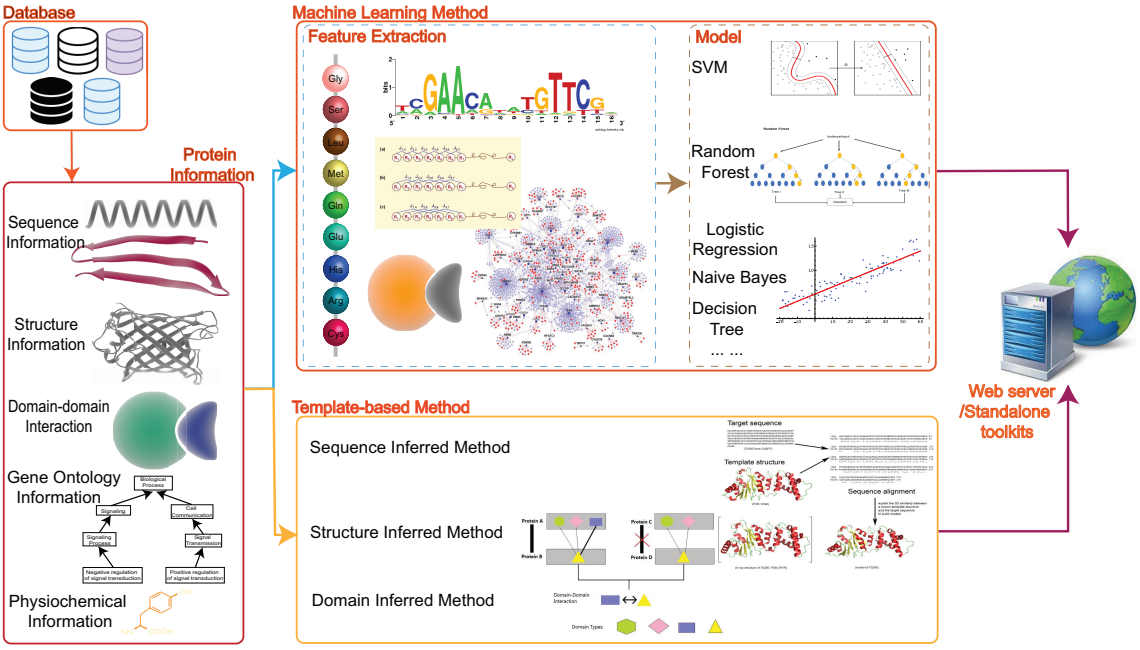


Figure 4.1: A General Computational Framework for Host-Pathogen Protein-Protein Interactions Prediction

evaluation with details was implemented or reported in these reviews. Recently, [226] conducted a sequence-based predictors review, however they focused on the prediction of protein-binding residues via single-sequence methods.

Adapted from these reviews, we subsequently collected all published predictors that focused on host-bacterium protein-protein interactions and host-pathogen protein-protein interactions, which are summarized in Table 4.2. The frameworks of the two different types of computational models for predicting HP-PPIs, including machine learning-based models and template-based models, are shown in Figure 4.1.

A template-based model utilises different types of protein information to build the prediction model, including sequence information, structure information and domain information [227–229]. Template-based models use different protein information to detect high score homology which might yield similar functions. However, template-based models may fail to predict whether the remote homology will interact with known proteins or not. Another type of computational model is based on machine-learning models. The protein information is first vectorized as the input to learn their inherent relationships automatically, which are thus used to build the model and predict the interactions.

Review	Reviewed Methods for HP-PPI	Reviewed Database
[162] (2013)	Homology-based approaches, structure-based approach, domain and motif interaction-based approach, machine-learning-based approach	VirusMINT, PHI-base, MINT, VirHostNet, BioGRID, IntAct, APID, PATRIC and so on
[151] (2015)	N/A	Web-Based Databases (HCVPro, PATRIC, HPIDB, PHISTO and so on)
[32] (2015)	Machine learning and data mining based approaches, homology based approaches, structure based approaches, domain and motif based approaches	None
[225] (2016)	Homology-based prediction, structure-based prediction, domain/motif interaction-based prediction, machine learning-based predictions of host-pathogen interactions	PATRIC, PIG, VirHostNet, HPIDB, VirusMINT and so on

1. For the released datasets in the reviews, they are not available at the time of our review;
2. For the evaluation of methods, including general methods, independent data methods and performance measurement, the quantitative reports are not available in the reviews.

Table 4.1: Overview of the reviews for host-pathogen protein-protein interactions

Specifically, for PPIs, the relevant protein information can be sequence information, gene ontology information, domain information, gene expression information and interaction network information.

As indicated in Table 4.2, numerous feature-representation algorithms for sequence information are incorporated with different machine-learning models for predicting host-pathogen protein-protein interactions. In this regard, we first grouped the sequential feature-representation algorithms into three different types: amino acid composition, pseudo-amino acid composition and evolutionary information. It should be noted that, not only the reported algorithms in Table 4.2, but also the related sequential-representation algorithms from other protein sequence-specific topics, such as protein structure, protein folding topics, are included in this section. The models from [218] and [34], which are shown in Table 4.2, were selected as the representative models regardless of the pathogen species.

Table 4.2: Computational Approaches for Prediction of Host-pathogen Protein-Protein Interactions (sorted by published year)

Refere-nce	Predictor	Pathogen	Data Source	Training Data		Independent Data	Protein Information	Sequence representation algorithms	Algorithm/ Model	Stand-alone software/ platform	Web server	Performance
				Positive Pairs	Negative Pairs							
[217]	N/A	Parasite	BIND, DIP, IntAct, Reactome	39207 human, 18412, 2643 Plasmodium falciparum intra-species interactions	N/A	N/A	Gene Ontology	N/A	Bayesian statistics	N/A	N/A	N/A
[227]	N/A	Parasite	DIP	N/A	N/A	N/A	Sequence	PSSM	Remote homology detection	N/A	N/A	N/A
[218]	N/A	Parasite	MINT, IntAct, Reactome, HPRD	1112	1136	N/A	Sequence	***CTM variation***	***Random Forest***	N/A	N/A	ROC curve
[156]	N/A	Virus	HPRD, MINT, BIND, DIP, IntAct, Reactome	1028	1:25, 1:50, 1:100 ratio comparing with positive	N/A	Domain, sequence, interaction network	k-mers	Support vector machine (linear kernel)	N/A	N/A	AUC value
[34]	N/A	Virus	I-MAP	500	500	N/A	Sequence	***CTM variation***	***Support vector machine (radial basis function kernel)***	N/A	N/A	Sensitivity, Specificity and Accuracy

Continuation of Table 4.2

Refere-nce	Predictor	Pathogen	Data Source	Training Data		Independent Data	Protein Information	Sequence representation algorithms	Algorithm/ Model	Stand-alone software/ platform	Web server	Performance
				Positive Pairs	Negative Pairs							
[23]	N/A	Bacterium (B. anthracis, F. tularensis, Y. pestis, S. typhi)	PHISTO	655 B. anthracis, 491 F. tularensis, 839 Y. pestis, 62 S. typhi	1:100 ration comparing with postive	N/A	Sequence, gene ontology, gene expression, interaction network	k-mers	Multitask learning	Yes	N/A	Precision, Recall and F1 score
[230]	PWEN-TLM	Virus	RefSeq	3638	3638	Holding subcatalog PPI dataset	Gene Ontology	N/A	Transfer learning	N/A	N/A	F1 score and ROC curve
[223]	N/A	Virus	IntAct	657	2910	N/A	Sequence, interaction network, tissue information, post-translational modifications	AAC, PAAC, PSSM	Ensemble learning	N/A	N/A	Sensitivity, Specificity and Accuracy
[228]	N/A	Bacterium (Mycobac-terium tuberculo-sis)	N/A	-	N/A	N/A	Sequence, motif	N/A	Homologous method	N/A	Databse: protein interactions of M.tuberculosis and human	N/A
[231]	N/A	Bacterium (Bacillus anthracis)	PATRIC	554	N/A	N/A	Sequence, graph properties	CTM variation, quadruples of consecutive amino acids	Four layers neural network	Yes	N/A	Accuracy and F1

End of Table 4.2

Table 4.2 Computational Approaches for Prediction of Host-pathogen Protein-Protein Interactions (sorted by published year)

4.2.2 Host-Pathogen Interactions Databases

There has been continuous effort spent on developing online HPI databases and repositories by many researchers. These developments mostly benefited from the National Institute of Allergy and Infectious Diseases (NIAID), which initialized a strategic plan to focus on biodefense research. Several 'priority pathogens' were defined. Several initial developments, including pathogen interaction gateway (PIG [164]), BioHealthBase [150] and the Pathosystems Resource Integration Center (PATRIC [160]), were wholly or partially funded by the NIAID.

The first web-based database with massive annotated records for pathogen research was the Ecological Database of the World's Insect Pathogens (EDWIP) [169]. EDWIP uses a one-to-one interaction relationship, which records the infection between a single host species and a single pathogenic species. This strategy resulted in 9,400 records between 4,454 host species and 2,285 pathogen species when it was first released in 2003. PIG was designed as a collection of a number of public resources, which focussed on experimentally verified and manually curated HP-PPIs. This centralized database served as an easy-to-use database which transfers search results to the relevant database, such as the UniProt [5] database. Another important host-pathogen interaction database is the Pathogen-Host Interaction Search TOol (PHISTO) [161]. This tool aims to provide researchers with a complete coverage of HPI data via monthly updates. Proteomics Standards Initiative Common Query InterfaCe (PSICQUIC) [174] service was installed to allow access to and extraction of HPI data the other web-based databases.

Following Chapter 2, numerous publicly available databases were reviewed, which were returned by searching specific keywords in the NCBI PubMed search engine. We manually examined the abstracts of the first 400 results ranked by 'best relevance' out of more than 4,000 returned items based on the keywords 'pathogen' and 'database'. As such, in this paper, a selection of 11 databases is reviewed and evaluated based on their contents. The selection is followed by the review of Chapter. 3, in which a cross-checking of 'Maintenance' and 'Related Databases' information has resulted in a subset of the 11

operational databases as the referred databases for curating the HPI dataset. Meanwhile, the 11 databases were mainly collected with the data sources coming from literature, domain expert manual verification and public archival databases, which are with high confidence. Details are provided in the following sections.

4.2.3 Sequence Representation Algorithms

To encode proteins as feature vectors, several different features have been included in this study to predict protein-protein interactions between *Homo sapiens* and bacterium pathogens, which are: (1) protein amino acid composition information [11, 33, 232]; (2) protein pseudo-amino acid composition information [233–235]; (3) protein evolutionary information feature [236, 237]. We discuss the related feature encoding algorithms below.

Amino acid composition

* **Conjoint Triad Method**

  It was proposed in [11] to classify the 20 amino acids into seven groups according to each amino acids dipole scale and volume scale, which are their electrostatic and hydrophobic properties. We briefly describe the physicochemical information in Table 4.3. There are afterwards several variations of encoding algorithms for sequence representation based on this table.

  Among these, one popular approach is to consider the relationship of the properties of one amino acid and its vicinal amino acids as a descriptor [11], which is named the conjoint triad method (CTM). The conjoint triad information of several adjacent amino acids makes it easy to represent every single protein sequence into a class-based feature with the same length, which is also called its k-mer features.

  Each amino acid type is indicated as a number ranging from $1-7$ according to its group. A detailed diagram for illustration of how k-mer features work is shown in Figure 4.2. The frequency of three conjoint triad data (3-mer) of a sequence is calculated. In total, there will be a combinations set including $\{(1,1,1), (1,2,1), \ldots,$

Group Index	Dipole	Volume	Amino Acids
1st	-	-	Ala(A), Gly(G), Val(V)
2nd	-	+	Ile(I), Leu(L), Phe(F), Pro(P)
3rd	+	+	Tyr(Y), Met(M), Thr(T), Ser (S)
4th	++	+	His(H), Asn(N), Gln(Q), Tpr(W)
5th	+++	+	Arg(R), Lys(K)
6th	+'+'+'	+	Asp(D), Glu(E)
7th	+‘	+	Cysc(C)

Table 4.3: Seven Groups of 20 Basic Amino Acids [11]

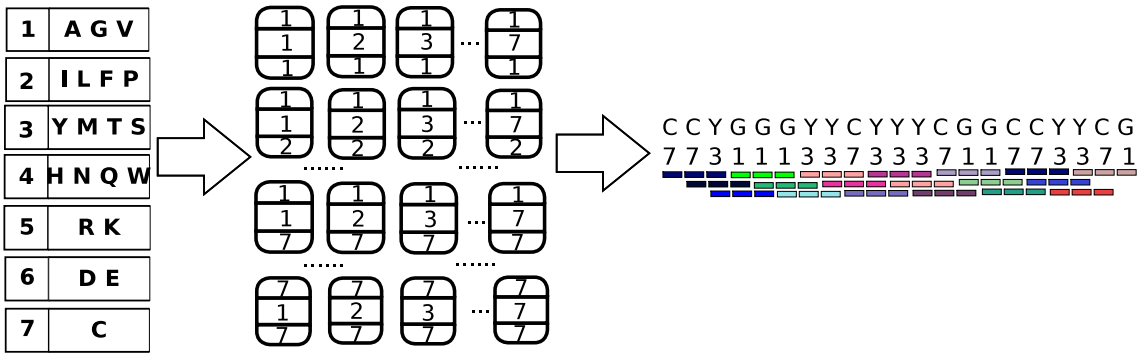


Figure 4.2: Basic Process of CTM [11]

(1,7,1), . . . ,(1,7,7), . . . , (7,7,7)}. As a result, *3-mer* features will encode a sequence to a vector of 343 dimensions. For other *2-mer*, *4-mer* and *5-mer* features, the features number would be 49, 2401 and 16807, respectively.

* **Auto Covariance**

The auto csovariance (AC) relationship among the amino acids based on the order of the sequence information was utilised in another feature representation algorithm by [33]. It is a popular transformation algorithms used to adopt numerical vectors to uniform matrices by analyzing sequences in the auto cross covariance (ACC) information.

Between two different vectors, there are two covariance relationships, which are cross covariance (CC) and auto cross covariance (ACC). Only ACC variables are calculated [33]. The basic idea is to derived the physicochemical properties of the amino acid, which include hydrophobicity (H), volumes of side chains of amino acids (VSCs), polarity (P1), polarizability (P2), solvent-accessible surface area

Name	H1	H2	Vsc	P1	P2	SASA	NCISC
A	0.62	-0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	-1	44.6	5.5	0.128	1.461	-0.03661
D	-0.9	3	40	13	0.105	1.587	-0.02382
E	-0.74	3	62	12.3	0.151	1.862	0.006802
F	1.19	-2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0	0	9	0	0.881	0.179052
H	-0.4	-0.5	79	10.4	0.23	2.025	-0.01069
I	1.38	-1.8	93.5	5.2	0.186	1.81	0.021631
K	-1.5	3	100	11.3	0.219	2.258	0.017708
L	1.06	-1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	-1.3	94.1	5.7	0.221	2.034	0.002683
N	-0.78	2	58.7	11.6	0.134	1.655	0.005392
P	0.12	0	41.9	8	0.131	1.468	0.239531
Q	-0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	-2.53	3	105	10.5	0.291	2.56	0.043587
S	-0.18	0.3	29.3	9.2	0.062	1.298	0.004627
T	-0.05	-0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	-1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	-3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	-2.3	117.3	6.2	0.298	2.368	0.023599

Table 4.4: Physicochemical Properties for Amino Acids [33]

(SASA) and the net charge index of the side chains (NCISC). These properties of 20 types of amino acids are reported in Table 4.4.

In the auto covariance method, each single protein sequences is first translated into a numerical value corresponding to seven different physicochemical properties. Since the ranges of these seven physicochemical properties vary a lot from each other, a first step of performing normalization for the numerical values is required. These values were hence normalized to a distribution, whose mean is zero and the standard deviation is one. The normalization equation is shown in Equa. 4.1.

$$\overline{p_{i,j}} = \frac{p_{i,j} - mean_j}{sd_j} (i = 1,2,3,...,20; j = 1,2,3,4,5,6,7) \tag{4.1}$$

where $p_{i,j}$ represents the jth property value of the ith amino acid, $mean_j$ is the mean value of the jth property over the 20 amino acids. sd_j is the standard deviation of jth property over the 20 amino acids. Via this operation, every protein sequence is

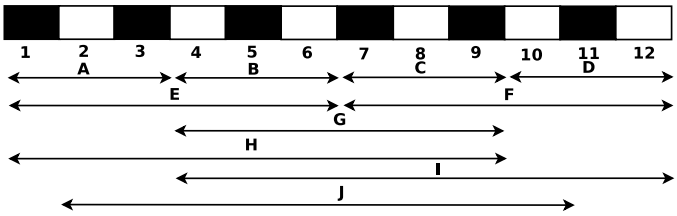


Figure 4.3: Dividing Protein Sequence into 10 Regions [232]

translated into a $N * M$ with zero mean and a standard deviation of unity in each column. With a proper range of these numerical values for each single protein sequence, auto covariance can be used to represent them into a uniform matrix. Based on Equa. 4.2, a matrix of $lg * 7$ is calculated, where lag is the distance between two amino acids, and $0 < lg \leq lag$.

$$AC(lag, j) = \frac{1}{N - lag} \sum_{i=1}^{N-lag} (P_{i,j} - \frac{1}{N} \sum_{i=1}^{N} P_{i,j}) * (P_{i+lag,j} - \frac{1}{N} \sum_{i=1}^{N} P_{i,j}) \tag{4.2}$$

For z properties chosen out of the seven physicochemical properties, the length of AC is $lag * z$. $P_{i,j}$ orresponds to the value from $\{p_{i,j}\}$. Here, N is the length of the protein sequence. After ACC transformation, a representation of protein-protein interaction is a concatenation of these two AC transform calculations results.

* **Local Descriptor**

Another sequenced-based feature representation method is a local descriptor [232]. The most important feature of an HP-PPI is that the interaction often occurs in some specific intermittent fragments. To better extract this continuous or discrete knowledge from sequence information, [232] proposed using region descriptors to firstly divide a protein sequence into 10 regions. As shown in Figure 4.3, a protein sequence is divided into four quarter regions (A-D), two half regions (E, F), the central 50% region (G), the first 75% region (H), the last 75% region (I) and the central 75% region (J).

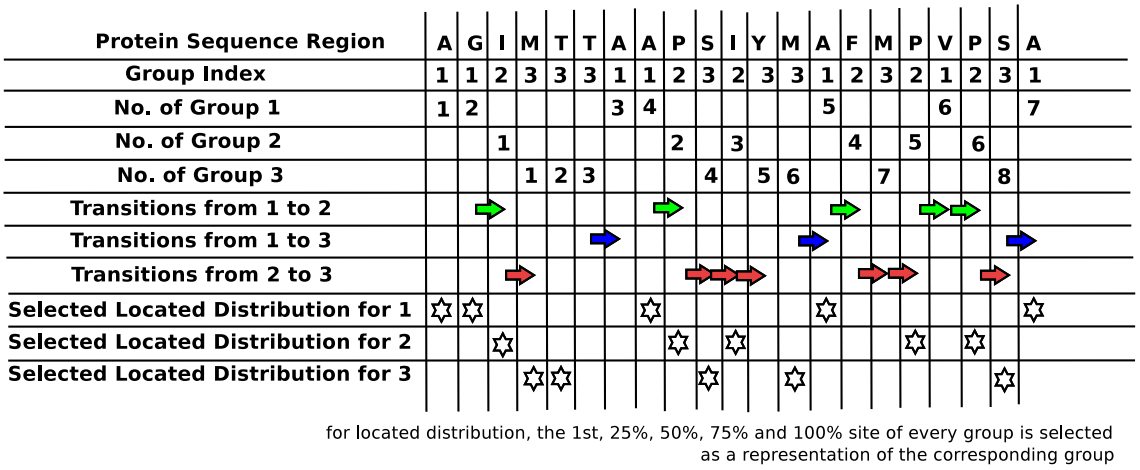


Figure 4.4: Local Descriptor for Protein Sequence adapted from [232]

With these 10 regions, a local descriptor is utilised to transform the region sequence into three related descriptors [232]. These three descriptors are composition (C), transition (T) and distribution (D). Composition is the composition ratio of each group of amino acid within a separate region. Transition represents the percentage of which amino acid group is followed by another amino acid group. Distribution describes the specific location information obtained by selecting the first, 25%, 50%, 75% and last one of each amino acid group. Figure 4.4 shows more details of C, T and D on a protein region sequence with 21 amino acids.

When using a local descriptor, the extracted feature vector contains 7 features for composition, 21 features for transition and 35 features for distribution. When multiplied by 10 different local regions, the local descriptor method generate 630 features for a single protein sequence. For a HB-PPI pair, this local descriptor contains 1260 features.

There are also some other schemes that can be used to extract different types of features of a protein sequence, for example Moran Autocorrelation Score [238] and the amino acid triplet [34]. As protein sequence information is directly linked to protein-protein interactions, a further novel representation of protein-protein interactions, especially for human-bacterium protein-protein interactions, might include any other information related to the specific host species and pathogenic

species, which may be a better alternative for prediction of host-pathogen protein-protein interactions [23].

Pseudo-amino acid composition

* **PseAAC**

Directly converting a protein sequence to a vectorized feature according to the amino acid composition (AAC) might result in sequence-order information loss. The pseudo-amino acid composition (PseAAC) method was proposed as a novel protein sequence representation of a discrete model, which has remarkable improvement in prediction performance as an important feature representation algorithm [233, 239, 240].

Various modes of PseAAC have been introduced in the literature. The key is to combine the sequence order correlation information from the protein sequence. In the work of [233], the original version of PseAAC was introduced, as shown in Equa. 4.3.

$$\begin{aligned}
\theta_1 &= \frac{1}{T-1}\sum_{i=1}^{T-1}\Theta(S_i,S_{i+1}) \\
\theta_2 &= \frac{1}{T-2}\sum_{i=1}^{T-2}\Theta(S_i,S_{i+2}) \qquad \lambda < T \\
&\cdots \\
\theta_\lambda &= \frac{1}{T-\lambda}\sum_{i=1}^{T-\lambda}\Theta(S_i,S_{i+\lambda})
\end{aligned} \tag{4.3}$$

Here, the Θ function is calculated by Equa. 4.4:

$$\begin{aligned}
\Theta(S_i,S_j) = \frac{1}{3}\{&[H_1(R_j)-H_1(R_i)]^2+[H_2(R_j)-H_2(R_i)]^2+ \\
&[M(R_j)-M(R_i)]^2\}
\end{aligned} \tag{4.4}$$

where $H_1(R_i)$, $H_2(R_i)$ and $M(R_i)$ are the corresponding phycial-chemical properties of the amino acid residue R_i. Equa. 4.4 produces a λ-dimensional vector.

Evolutionary information

* **Position-Specific Scoring Matrix (PSSM)**

  By scanning a unique sequence against a reference database, the compilation of a set of alignment profiles results in a position-specific scoring matrix (PSSM) of the sequence, which indicates the probability of the corresponding positions of the amino acid types [236]. The position-specific scoring matrix is returned as a $T \times 20$ matrix for a given protein sequence by position-specific iterated BLAST (PSI-BLAST). Here, T denotes the length of the corresponding protein sequence. Transformation of the PSSM, which involves highly and broadly homologous sequences information, has been widely used in sequence-related studies [237, 241–246]. These studies indicated that, including evolutionary information for feature representation helps to improve prediction model performance.

  In detail, given a protein sequence as $S = S_1S_2S_3S_4 \ldots S_T$, where T is the length of the protein sequence, the corresponding PSSM $P = \{P_{m,n}\}, m = 1, \ldots T; n = 1, \ldots, 20$ is calculated based on the amino acid similarity matrix. The matrix used can be either point-accepted mutation (PAM, such as Dayhoff's mutation matrix [247]) or position-weight matrix (PWM, such as the block substitution matrix BLOSUM [248]). The value is calculated according to Equa. 4.5:

$$P_{m,n} = \sum_{k=1}^{20} w(m,k) \times \theta(n,k) \tag{4.5}$$

  where $w(m,k)$ is the probability that the k_{th} amino acid appears at position m, and $\theta(n,k)$ is the value of the position of (n,k) in the similarity matrix.

  In this study, PSI-BLAST was employed to create PSSM with three iterations, where the e-value was set to 0.001. Accordingly, the various lengths of the protein sequences resulted in matrices with different dimension, which introduces different encoding features based on the PSSM profile. The following parts present several PSSM-based feature representation algorithms.

- **Pse-PSSM**

  The pseudo Position-Specific Score Matrix (Pse-PSSM) is firstly introduced in the task of predicting an uncharacterized protein to be membrane protein or not [239]. It extends the idea of corrupting PSSM descriptor vertically as a mean value, as shown in Equa. (7), though the value of PSSM is firstly processed by a standardization procedure horizontally by rows in Equa. 4.6 and Equa. 4.7. The concept of pseudo amino-acid composition is to generate correlation information between different amino acid locations.

$$P'_{m,n} = \frac{P_{m,n} - \frac{1}{20}\sum_{k=1}^{20} P_{m,k}}{\sqrt{\frac{1}{20}\sum_{k=1}^{20}(P_{m,k} - \frac{1}{20}\sum_{k=1}^{20} P_{m,k})^2}} \tag{4.6}$$

$$\overline{P_n} = \frac{1}{T}\sum_{m=1}^{T} P'_{m,n} \qquad (n = 1,2,\ldots,20) \tag{4.7}$$

  Thus, the original PSSM profile is converted to a 20-dimensional vector, $\overline{P} = \{\overline{P_n}, n = 1,\ldots,20\}$. This derived feature focuses on representing the average score of each amino acid types according to the reference database, which loses the sequence order information of the protein. Thus, [239] proposed considering supplementary information from the pseudo amino acid composition, which slices the PSSM profile according to Equa. 4.8.

$$Pse_n = \frac{1}{T-c}\sum_{m=1}^{T-c}[P'_{m,n} - P'_{(m+c),n}]^2 \qquad (n = 1,2,\ldots,20; c < T) \tag{4.8}$$

  This process generates a 40-dimensional vector $P_{se} = \{\overline{P_1}, \overline{P_2}, \ldots, \overline{P_{20}}, Pse_1, Pse_2, \ldots, Pse_{20}\}$ while $0 < c < min(T)$. For a given set of protein sequences, the upper bound of c should be smaller than the shortest length of the protein sequences.

- **Block-PSSM**

  By considering the PSSM profile in a dimension format of $T \times 20$, [249] proposed dividing the whole sequence into 20 equal blocks, where each represents five percent of the total sequence. Each block generate a 20-dimensional vector, which

is finally combined as a $20 \times 20 = 400$ dimension vector in total.

The ith block is calculated according to following Equa. 4.9.

$$Pblock_{i,j} = \frac{1}{B_i} \sum_{i=1}^{B_i} P_{i,j} \qquad i = 1,2,\ldots,20; j = 1,2,\ldots,20 \tag{4.9}$$

where i represents the block number. Since each five percent of a sequence is considered as a block, i ranges from 1 to 20. j is the number of amino acid types. In short, $Pblock_i$ is extracted as a 1×20 vector, thus $Pblock = Pblock_1, Pblock_2, \ldots, Pblock_{20}$ is calculated as the Block-PSSM feature in a form of 1×400 vector feature.

- **AAC-PSSM & DPC-PSSM**

  Another variation of PSSM-based features was proposed in [250]. The original PSSM profile is scaled to the range from 0 to 1 by following a sigmoid function shown in Equa. 4.10:

$$P''_{m,n} = \frac{1}{1 + e^{-P_{m,n}}} \tag{4.10}$$

  $P''_{m,n}$ is also used in the transition probability composition (TPC) PSSM [250]. AAC-PSSM extracts the corresponding amino acid composition information from $P = \{P''_{m,n}, m = 1,\ldots,T; n = 1,\ldots,20\}$. The vector from Equa. 4.11 represents an average mutation score of the amino acid types in the protein during the evolution process, namely AAC-PSSM. This calculation generates a 20-dimensional feature vector.

  As a supplementary, traditional dipeptide composition (DPC) from the protein sequence is extended [250], which is then named DPC-PSSM. The calculation is based on the covariance between two adjacent amino acid residues, denoted by Equa. 4.12. This process produces a 400-dimensional feature vector.

$$Paac_n = \frac{1}{T} \sum_{m=1}^{T} P''_{m,n} \qquad m = 1,\ldots,T; n = 1,\ldots,20 \tag{4.11}$$

$$Pdpc_{i,j} = \frac{1}{T-1}\sum_{k=1}^{T-1} P''_{k,i} \times P''_{(k+1),j} \qquad i,j = 1,\ldots,20 \tag{4.12}$$

- **TPC-PSSM & DP-PSSM**

  The transition probability composition (TPC) PSSM [251] and directional property (DP) PSSM [252] are two variants of PSSM-based feature algorithms from DPC-PSSM and Pse-PSSM, respectively.

  TPC-PSSM is defined as a 400-dimensional feature vector $Ptpc = \{Ptpc_{i,j}, i,j = 1,\ldots,20\}$, and it is calculated by following Equa. (13).

$$Ptpc_{i,j} = \frac{\sum_{k=1}^{T-1} P''_{k,i} \times P''_{(k+1),j}}{\sum_{j=1}^{20}\sum_{k=1}^{T-1} P''_{(k+1),j} \times P''_{k,i}} \qquad i,j = 1,\ldots,20 \tag{4.13}$$

  DP-PSSM takes the standardization procedure from Pse-PSSM feature and expands the extraction of information from both positive and negative terms [252]. It consists of two parts, in which one is from individual amino acid composition and the other is from the dipeptide composition. Pdp could be illustrated as the following Equa. 4.14.

$$\begin{aligned} Pdp &= [T', G'] \\ T' &= [T_1^P, T_1^N, \ldots, T_{20}^P, T_{20}^N] \\ G' &= [G_1, G_2, \ldots, G_{20}] \\ G_j &= [\Delta_{1,j}^P, \Delta_{1,j}^N, \ldots, \Delta_{\alpha,j}^P, \Delta_{\alpha,j}^N] \end{aligned} \tag{4.14}$$

  In Equa. 4.15 and Equa. 4.16, the superscripts P and N represent the positive terms and negative terms according to following equations.

$$\begin{aligned} T_j^P &= \frac{1}{NP_j}\sum P'_{i,j} \qquad if \quad P'_{i,j} \geq 0 \\ T_j^N &= \frac{1}{NN_J}\sum P'_{i,j} \qquad if \quad P'_{i,j} < 0 \\ & \qquad\qquad j = 1,2,\ldots,20 \end{aligned} \tag{4.15}$$

$$\begin{aligned}
\Delta^{P}_{k,j} &= \frac{1}{NDP_j}\sum[P^{'}_{i,j} - P^{'}_{i+k,j}]^2 \qquad if \quad P^{'}_{i,j} - P^{'}_{i+k,j} \geq 0 \\
\Delta^{N}_{k,j} &= \frac{1}{NDN_j}\sum[P^{'}_{i,j} - P^{'}_{i+k,j}]^2 \qquad if \quad P^{'}_{i,j} - P^{'}_{i+k,j} < 0 \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 1,2,\ldots,\alpha
\end{aligned} \tag{4.16}$$

$T^{'}$ contributes 40 dimensions and $G^{'}$ contains another $40 \times \alpha$-dimensional feature. Totally, protein sequence is represented by a $(40 + 40 \times \alpha)$-dimensional feature vector.

4.2.4 Machine Learning Models for Prediction

Applying computational approaches for prediction of bioinformatics tasks is considered as an important supplementary method for identifying specific targets and high-fidelity interactions in experiments. Recently, we have witnessed numerous applications focusing on the domains containing an abundance of unknown data, which require hypothesis verification [26, 32, 159, 162].

In Table 4.2, the predictors from [34, 218], which are based on machine learning model and protein sequence information, were selected for our following study. The machine learning models include support vector machine (SVM) and random forest (RF).

In this section, we will first briefly review most of the potential machine learning models that can be utilized for host-pathogen interactions prediction in Table 4.2, which include logistic regression (LR), the Na$^{'}$ive Bayes (NB) model, decision tree (DT) model, random forest (RF) model, support vector machine (SVM) model and gradient boosting machine (GBM) model. These models have demonstrated their capability in other applications for protein structure prediction; however, this is the first time they have been presented in an overall performance evaluation in relation to different feature-representation algorithms for HB-PPIs.

Support Vector Machine

Support vector machine (SVM) model is one of the most widely used models in the literature, which was originally developed by [253]. The introduced structural risk minimization theory ensures the performance of SVM to be widely and successfully applied to many classification and regression tasks in computational biology. SVM with a Radial Basis Functions (RBF) kernel is firstly deployed given a task of classifying HP-PPI pairs [34, 224]. Given a dataset of HB-PPI denoted as $\{x_i,y_i\}$, i=1,2,...,N, where $x_i \in R^n$ and $y_i \in \{+1,-1\}$, y_i is calculated in the following Equa. 4.17 in SVM:

$$y(x) = sign[\sum_{i=1}^{N} y_i \alpha_i * K(x,x_i) + b] \tag{4.17}$$

where $K(x_i,x_j) = exp(-\gamma \|x_i - x_j\|^2)$ stands for the RBF kernel, and α_i contains the parameters from a convex quadratic programming problem.

Decision Tree

The decision tree (DT) model is designed as a non-parametric supervised model [254]. It uses a tree-like graph to predict an incoming instance based on learnt decision rules from given data samples and represented features. Decision trees are simple to understand and interpret, and they are also capable of handling both numerical and categorical data.

Random Forests

Derived from the decision tree model, random forests (RF) adopts random learning method to construct a combination of decision trees [255]. It presents superior performance compared with other machine learning models for classification task, regression task and so on. Technically, it is an ensemble learning model based on the tree bagging method, which builds a bunch of random decision trees to avoid the latent problem caused by potentially biased data.

In this study, we implement random forest using *scikit*-learn toolkit [256] in Python.

Logistic Regression

Logistic regression is an important machine learning model, which targets modelling y_i between 0 and 1 given unseen data x_i. Accordingly, the logistic regression returns results by Equa. 4.18:

$$\begin{aligned} P(y_i = 1|x_i) &= h_\theta(x_i) = 1/(1+exp(-\theta^T * x_i)) \\ P(y_i = 0|x_i) &= 1 - P(y_i = 1|x_i) = 1 - h_\theta(x_i) \end{aligned} \tag{4.18}$$

where θ is the combination of the model parameters, and the optimization of θ is solved with either the cross-entropy function J_1 or the mean square error loss function J_2, which is shown in Equa. 4.19:

$$\begin{aligned} J_1(\theta) &= -\sum_i (y_i log(h_\theta(x_i)) + (1-y_i) log(1-h_\theta(x_i))) \\ J_2(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_i))^2 \end{aligned} \tag{4.19}$$

Naïve Bayes Model

Based on the Bayes‘ theorem [257, 258], the naïve Bayes model consists of a probabilistic classifier and considers features as independent variables between each other when the class label is given. Given $X = (x_1, x_2, \ldots, x_n)$, x_i is the i_{th} feature, the probability of being in category y_k is calculated by Equa. 4.20:

$$p(y_k|X) = \frac{p(y_k)}{p(X)} \prod_{i=1}^{n} p(x_i|y_k) \tag{4.20}$$

In this study, we selected the Gaussian Naïve Bayes (GNB) model to deal with the continuous data projected from the various feature representation algorithms. The distribution of the data was assumed to be a Gaussian distribution, which follows Equa. 4.21.

$$p(x_i|y_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_i-\mu_k)}{2\sigma_k^2}} \tag{4.21}$$

In Equa. 4.21, μ_k is the mean of X and σ_k^2 is the corresponding variance.

Gradient Boosting Machine

Gradient boosting machine (GBM) was firstly developed as a greedy optimization model [259] for both regression and classification tasks. Among the variants of gradient boosting machine, gradient tree boosting is a frequently used model integrated with the decision trees model. Given a $X = (x_1, x_2, \ldots, x_n)$, in which x_i is related to label y_i, gradient tree boosting builds an ensemble of trees sequentially by distilling the gradient descent algorithm into the process of new tree construction. A new tree is constructed under the discrepancy between target function $f(x)$ and current model, in which $f(x_i) = y_i$. The discrepancy between target function $f(x)$ and the current model is also called residual of gradient boosting machine.

4.3 Host-pathogen Interactions Materials

4.3.1 Human-bacterium Interaction Resources

In this section, we firstly collected and reviewed 11 public databases, as summarized in Table 4.5: the Database of Interacting Proteins (DIP) [178], Reactome [187], the Agile Protein Interaction DataAnalyzer (APID) [188], IntAct [176], the Molecular Interaction Database (MINT) [260], the InnateDB [189], the pathogen-host interaction search tool (PHISTO) [161], the Pathosystems Resource Integration Center (PATRIC) [160], Mentha [186], the Host Pathogen Interaction Database (HPIDB) [172], the Biological General Repository for Interaction Datasets (BioGRID) [177].

As humans are one of the primary host species among infectious diseases, the human-pathogen interaction resources are considered as the preliminary investigation subjects from all these databases. The column 'HPI number' indicates the corresponding recorded interaction number from the databases, which contain both inter-species interactions and intra-species interactions. These 11 databases were selected because their data sources mainly come from literature, domain expert manual verification and public archival databases, which are with high confidence of the presented data.

Taking database PATRIC [160] as an example, the data source was built upon several public archival databases, such as MINT [260], IntAct [176], BioGRID [177], and DIP [178]. The cross-archived databases have extended the availability of host-pathogen interactions resources, however there would also be some duplicates which inevitably occur during the combination of these 11 databases. Thus, we followed the traditional data collection and cleansing method from the literature [23, 34, 156].

Database	Data Source	Data Type	HPI number
DIP [178]	Literature and domain expert manual verification	Protein-protein interactions	76,882
Reactome [187]	Literature and domain expert manual verification	Comprehensive data portal including pathway and analysis	1,016,953
APID [188]	Public archival databases	Protein-protein interactions	133,994
IntAct [176]	Public archival databases and literature	Molecular interaction database	857,826
MINT [260]	Literature	Protein-protein interactions	123,892
InnateDB [189]	Literature	Mammalian innate immunity networks, pathways and genes	24,077
PHISTO [161]	Public archival databases	Host-pathogen and human intraspecies protein-protein interactions	90453
PATRIC [160]	Public archival databases	Comprehensive data portal for bacterium pathogens	618,737
Mentha [186]	Public archival databases	Protein-protein interactions	1,272,096
HPIDB [10, 172]	Public archival databases and literature	Host-pathogen interactions	62,783
BioGRID [177]	Literature	Comprehensive data portal for protein, genetic and chemical interactions	1,568,115

Table 4.5: The Human-Pathogen Interaction Resources

4.3.2 Data Curation

In this section, we briefly describe the major statistics for our 'golden dataset' curation, which will be thoroughly surveyed in following sections.

Positive Interactions

Six different types of bacterium pathogens were selected and the related data were pre-processed from the available databases. We identified these bacterium by mapping the taxonomy IDs according to the NCBI Taxonomy database. In Table 4.6, the corresponding information, including taxonomy ID, organism name, total pair number from the databases and the number after cleansing, are presented. These 11 databases were accessed and downloaded in September, 2018.

Despite the redundant ID information appearing in the databases, the collected protein sequence information from Swiss-Prot/UniProtKB is also involved at this stage with the assistance of CD-HIT tool [261]. Herein, CD-HIT is a popular tool to cluster highly homologous sequences (in this paper the threshold of sequence identity is set as 70%) to reduce the redundancy of database. It also helps to identify the clusters with representative protein. The redundancy between sequences is deemed to bring potential bias in the trained models.

In Table 4.6, the statistics refer to the results of the representative proteins. Meanwhile, any proteins with less than 50 amino acids were removed since these proteins may be non-functional fragments. The protein sequence information was primarily from the SwissProt/UniProtKB database [5].

Negative Interactions

How to select feasible negative PPIs remains an active topic for prediction of protein-protein interaction. Currently, there is not a standard protocol defining both the negative pairing strategy and the ratio to positive interactions. In most cases, building a negative interaction dataset by randomly selecting protein pairs from a set of unknown interacting

relationship between protein pairs is utilised. This heuristic approach works well in practice as the interaction ratio (i.e. the number of positives in a large, random set of protein airs) is expected to be very low, which in the work of [156] was defined as 25, 50, and 100 times as many negative examples as positive examples. In the study by [23], the ratio was set as 1/100. The assumption in this approach is that the probability that the selected negatives contain true positives is negligible.

Thus, we follow the traditional approaches from the literature [23, 156, 223, 231]. A random pairing for a negative protein-protein interaction was firstly undertaken between different proteins sets, which in this study was between the chosen bacterium pathogens (listed in Table 4.6) and Homo sapiens proteins (taxonomy ID: 9606). Then, we randomly selected a subset from this random pairing set to be the negative dataset. The negative interactions were selected with different ratio, which are 1:1, 1:25, 1:50 and 1:100.

Protein Information

When building machine learning models for prediction of protein-protein interactions, it requires the research subject HB-PPI to utilise the diverse protein information, which can be divided into three groups: structure-based, domain-based and sequence-based protein information.

Numerous studies have utilised and examined different information in the prediction of specific host-pathogen protein-protein interactions [156, 262, 263]. Particularly, domain-domain and structure-structure interaction methods are two main approaches to complement existing high-confidence interactions [156, 263]. Also, the structural similarity, which refers to a result of homology-based modelling, is an important alternative for detecting proteins with a homogeneous structure based on experimentally verified host-pathogen protein-protein interactions [262].

Although structure-based and domain-based information have some benefits for exploring the host-pathogen interactions [28, 264], it limits the scale of the study of HP-PPI to specific genre and species, such as HIV-1, HCV, Ebola viruses and so on [22, 26, 156,

223, 265, 266]. One dominant reason is the limited amount of available experimentally determined structures and domain information, particularly for bacteria. Imputation remains a core technology to compensate for the dearth of protein information and helps to address the challenge of interaction prediction [263]. Imputation for missing data also have impacts on the prediction performance since it brings putative information, which might not be accurate. Thus, utilizing structure-based and domain-based information limits the availability and scalability to a wide range of studies of HB-PPIs.

Alternatively, there has been a research trend of predicting PPIs from sequence-based protein information [11, 267]. Sequence-based protein information is one of the most abundant protein information, which has stimulates ongoing research to improve the prediction performance of novel feature representation and machine learning models [34, 226, 231, 268, 269]. The sequence-based methods enable the models to be applied on larger dataset and various species and genres.

Independent Datasets

To help understanding each dataset's information, in Table 4.7, all the proteins numbers related to the different subsets were included. This information, which was related to the reviewed sequence information from UniProtKB database [5], was last updated on 30th Oct., 2018. In total, we collected 18,181 *Homo sapiens* protein sequence information, and the corresponding protein numbers for each taxonomy ID are reported in Table 4.7.

The evaluation of models requires a careful preparation of independent datasets. Generally, cross-validation shows better performance than the independent-testing model for an unseen dataset. To give a general performance evaluation, we followed [34] when we built the independent datasets. The difference was that we further built five-fold independent datasets, which helped us to better measure the means and variations of the machine-learning models.

The independent datasets were not used during the training, and various measurements were included to evaluate the performance of different models based on the independent

datasets. Thus, we first randomly select one-fifth of the PPIs from both positive and negative interactions to be the independent dataset. The remaining PPIs of positive and negative interactions were then combined as the training set. We assembled the negative interactions with a random sampling method, where random sampling of the negative interactions was conducted five times, which allowed us to evaluate the different models with statistic means and variations to reduce the bias caused by negative interactions. The involved proteins number for *Homo sapiens* and corresponding bacterium pathogen taxonomy IDs are reported in details in Table 4.7. We have reported the number of utilised proteins for each species for different ratio settings. We anticipate that this experimental setting and details will help to provide more information to build novel machine learning methods in future work.

The framework of our evaluation study is presented in Figure 4.5. In Figure 4.5, a clear process procedure from databases to training and independent datasets, followed by the feature representation algorithms and machine learning model evaluations, are mapped in a coherent line. The best model selection and prediction are given as the main outcome of this framework.

Taxonomy ID	Bacterium Pathogen	Total number from Databases	After cleansing
1491	Clostridium botulinum	61	57
644	Aeromonas hydrophila	73	73
623	Shigella paradysenteriae	118	105
177416	Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4)	1319	1207
1392	Bacillus anthracis bacterium	3275	2810
632	Yersinia pseudotuberculosis subsp. pestis (Lehmann and Neumann 1896) Bercovier et al. 1981	4114	3528

Table 4.6: Selected bacterium Species Positive Interactions

Taxonomy ID	Whole Proteome Information		Positive Information		Positive Pairs	Total no. of HB-PPI	Negative Pairs			
	Human	Bacterium	Human	Bacterium			1:1	1:25	1:50	1:100
1491	18181	524	9	7	57	9.5 *M*	57	1425	2850	5700
644	18181	511	66	4	73	9.3 *M*	73	1825	3650	7300
623	18181	1724	75	60	105	31.3 *M*	105	2625	5250	10500
177416	18181	550	889	306	1207	10.0 *M*	1207	30175	60350	120700
1392	18181	1501	1537	844	2810	27.3 *M*	2810	70250	140500	281000
632	18181	1893	1866	1092	3528	34.4 *M*	3528	88200	176400	352800

Table 4.7: Overview of the Protein Information for the Datasets Preparation Process. Note: Only the proteins from the positive interactions which are processed by CD-HIT [261] are kept and counted in this table; *M* is short for 'million'. For each human-bacterium PPI dataset, the number of pathogen proteins, the size of the dataset and other such statistics are shown.

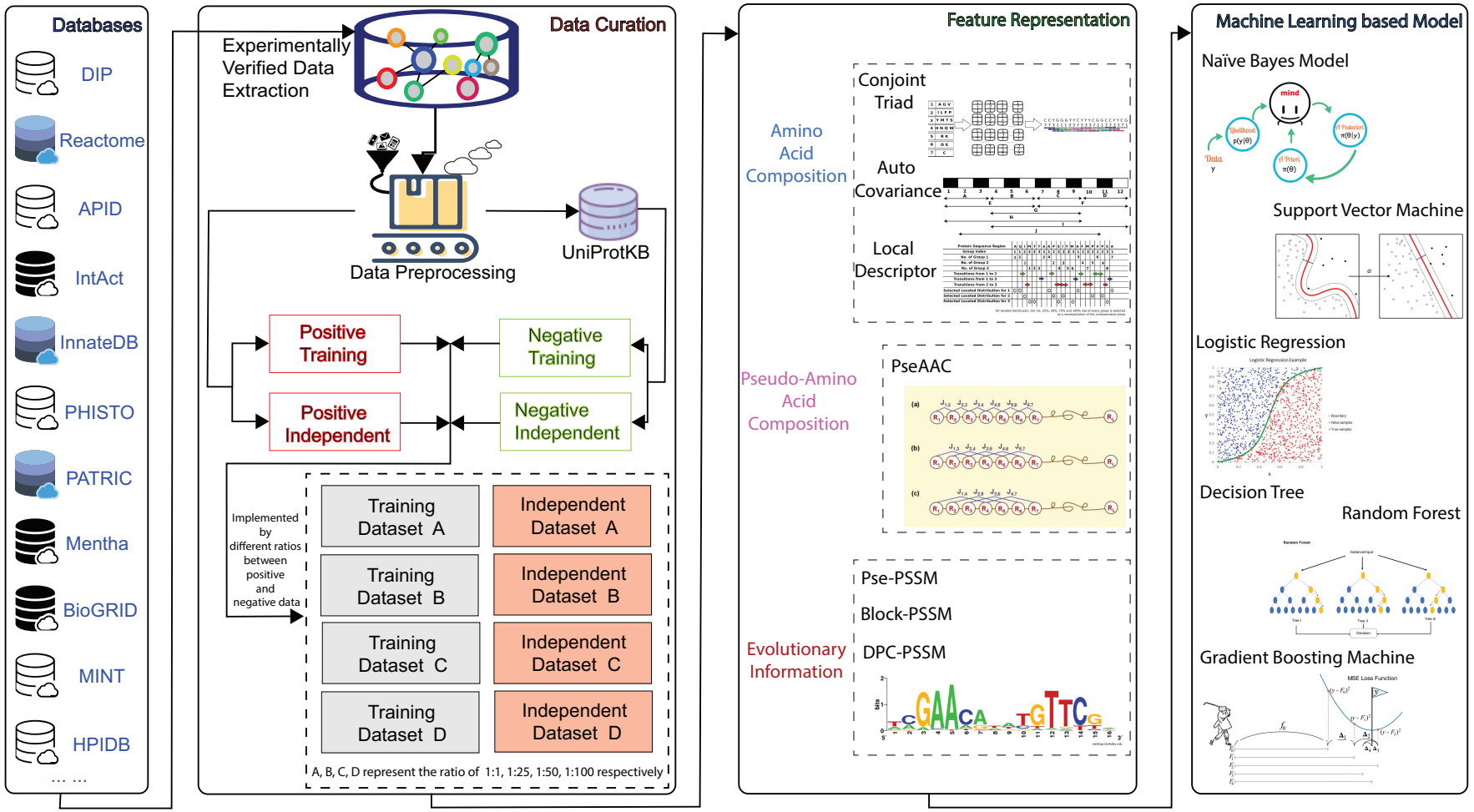


Figure 4.5: Designed Framework of Human-Bacterium Protein-protein Interaction Prediction

4.4 Evaluation Results

4.4.1 Evaluation Metrics

A set of six popular performance evaluation metrics, including precision (Pre), accuracy (Acc), sensitivity (Sn), specificity (Sp), F1-score and Matthew's correlation coefficient (MCC) score were applied to evaluate the overall prediction performance of the models. The measurements are defined as following Equa. 4.22.

$$\begin{aligned} Pre &= \frac{TP}{TP+FP} \\ Acc &= \frac{TP+TN}{TP+FP+TN+FN} \\ Sn &= \frac{TP}{TP+FN} \\ Sp &= \frac{TN}{TN+FP} \\ F1 &= \frac{2\times Pre}{Pre+Rec} \\ MCC &= \frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \end{aligned} \tag{4.22}$$

where TP, FP, TN and FN means the number of true positives, false positives, true negatives and false negatives respectively. Also, the receive operating characteristic (ROC) curve and the area under the curve (AUC) are included to quantify the model performance.

4.4.2 Performance Evaluation and Discussion

In this section, the holistic performance evaluation is presented with regard to the prediction task of human-bacterium protein-protein interactions, including the details of numerous feature representation algorithms, different ratios between positive and negative interactions, different machine learning models.

Performance Evaluation Based on Different Class Ratios

One major evaluation of this study was the ratio impact of different predictors, which was the ratio between the positive and negative protein interactions. Herein, we present the F1 score and Acc value from our measurements for feature 'ACC' for the evaluation discussion in the main body of this thesis, while the more details of the metrics are reported in the appendix. The mean value and deviation of each of the five independent tests were calculated in terms of different bacterial species and building ratio settings between the positive and negative pairs. In general, the ability to predict positive interactions as negative pairs decreases both the F1 and Acc results. Here, we found that the Acc was as high as 0.990099 when all the test data were predicted as negative interactions for a ratio of 1 : 100 between the positive and negative interactions. For ratios of 1 : 25, 1 : 50 and 1 : 100 between the positive and negative interactions, the datasets were considered as imbalanced datasets. Thus, F1 score was more suitable for measuring the performance of imbalanced datasets.

From Figure 4.6, it is easy to see that the F1 scores present a trend of getting worse as the dataset becomes bigger and more complicated, which means more protein nodes and edges are involved in the dataset. For example, when the positive to negative ratio was 1 : 1, a 1.0 ± 0.0 F1 score was found for the RF algorithm and the taxonomy ID is "1491". However, the F1 score became 0.96 ± 0.0 with RF for ID "644", 0.817555 ± 0.029558 with LR for ID "623", 0.730386 ± 0.005192 with RF for ID "177416", 0.770171 ± 0.007703 with RF for ID "1392" and 0.752226 ± 0.006632 with RF for ID "632".

In Figure 4.7 and Figure 4.8, feature representation algorithms 'PseAAC' and 'BlockPSSM' from the evolutionary information method are included with different ratios. The performance comparison between these two different sequence based features also indicate the impact of the ratio upon the F1 and Acc results.

From Figure 4.8, we can see that all the predictors have worse performance for all datasets when the ratio increases from 1 : 1 to 1 : 25, 1 : 50 and 1 : 100, especially when the dataset is with more than one hundred thousand samples. For example, for taxonomy

Figure 4.6: Accuracy and F1 Score of Different Machine Learning-based Models for 'Auto Covariance' Feature Representation Algorithm in Predictions of HB-PPIs

Figure 4.7: Accuracy and F1 Score of Different Machine Learning-based Models for 'PseAAC' Feature Representation Algorithm in Predictions of HB-PPIs

Figure 4.8: Accuracy and F1 Score of Different Machine Learning-based Models for 'BlockPSSM' Feature Representation Algorithm in Predictions of HB-PPIs

ID "632", the F1 score was 0.752226 ± 0.006632 for a $1:1$ ratio, however, the F1 scores dropped to 0.312530 ± 0.010944 for a $1:25$ratio, 0.243679 ± 0.012883 for a ratio of $1:50$ and 0.154535 ± 0.012569 for the $1:100$ ratio. These results were all achieved with the RF algorithm.

In Figure 4.9 and Figure 4.10, the results of the existing available methods from literature are included. Figure 4.9 contains the Acc, F1 and MCC scores for IDs "1491", "644" and "623", and Figure 4.10 contains the results for IDs "177416", "1392" and "632". Both Figure 4.9 and Figure 4.10 indicate the performance variation when the dataset changes from taxonomy ID "1491" to "644" and "623", which becomes worse for taxonomy IDs "177416", "1392" and "632". Even though the existing methods in Figure 4.9 and Figure 4.10 have incorporated several novel sequential feature representation algorithms, their performance has not improved.

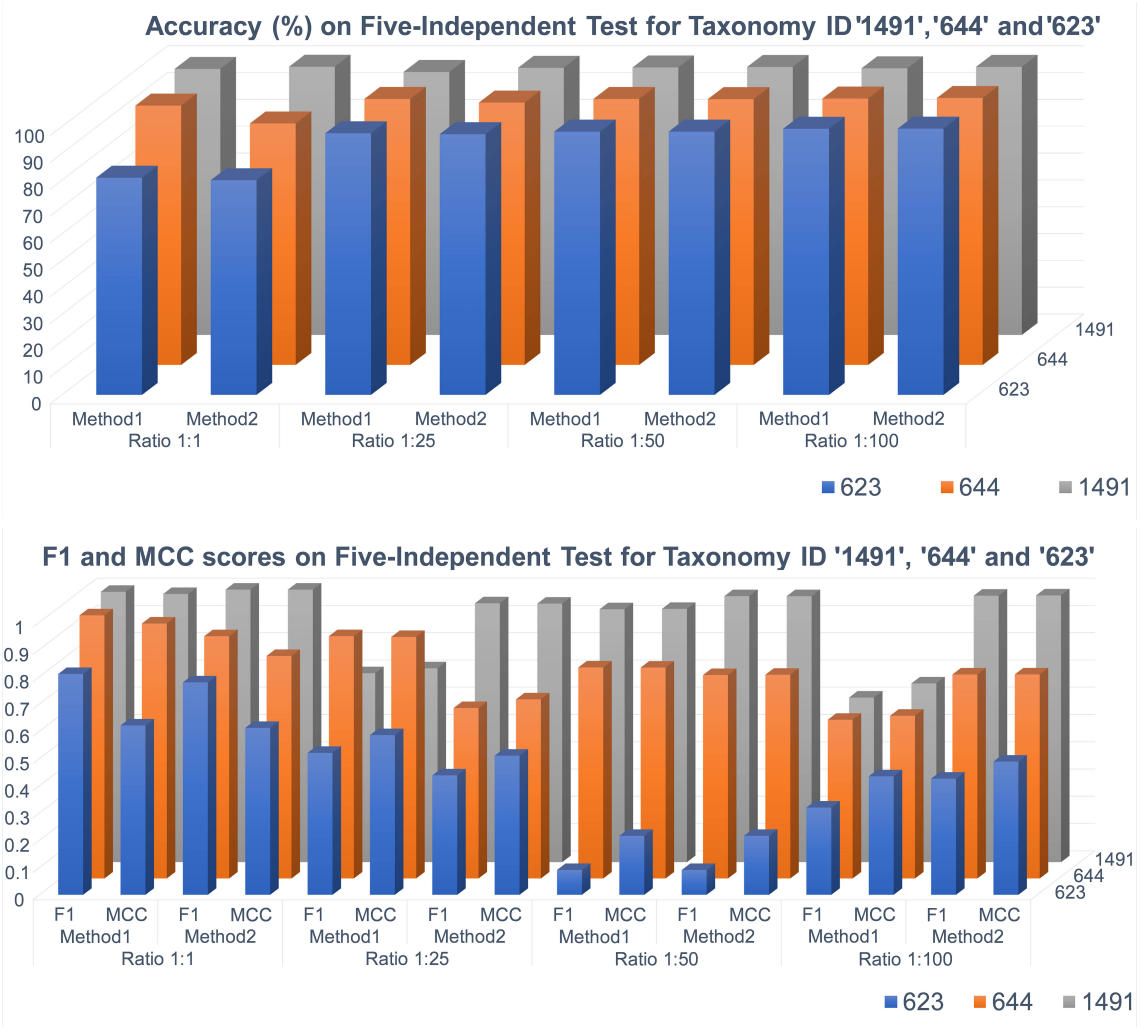


Figure 4.9: Accuracy, F1 Score and MCC Value of Methods from Literature for 'Clostridium botulinum', 'Aeromonas hydrophila' and 'Shigella paradysenteriae'

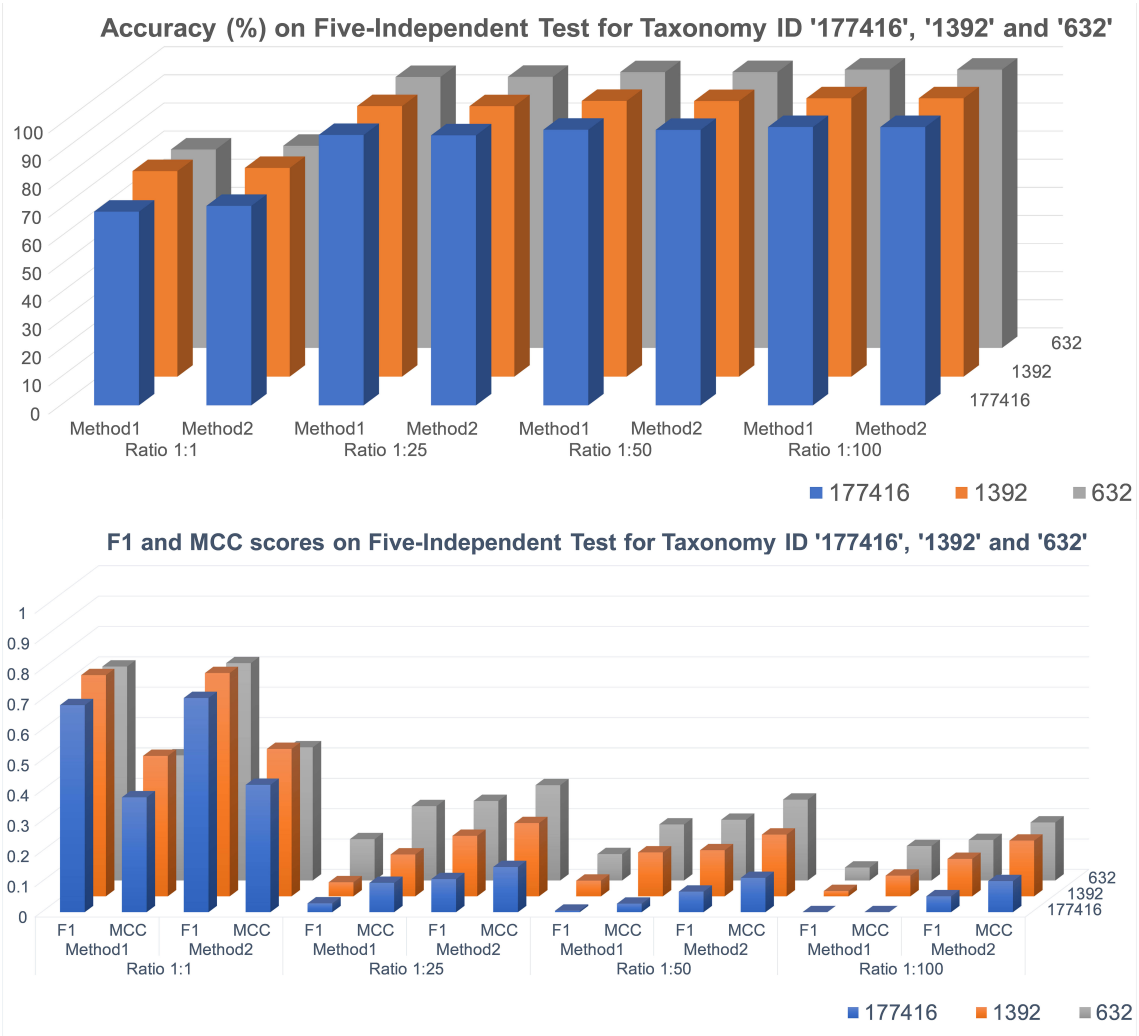


Figure 4.10: Accuracy, F1 Score and MCC Value of Methods from Literature for 'Francisella tularensis', 'Bacillus anthracis' and 'Yersinia pseudotuberculosis'

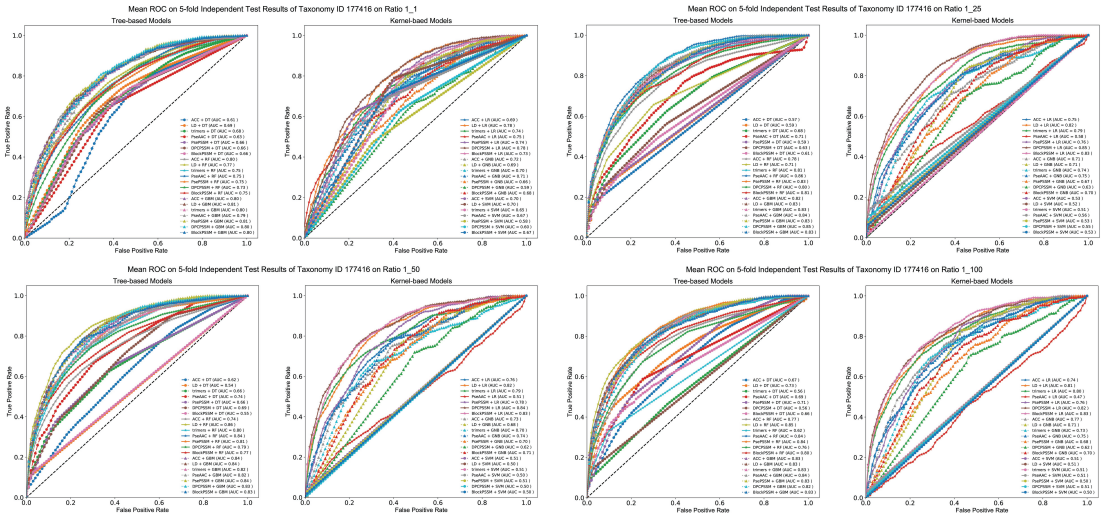


Figure 4.11: The ROC Curve for 'Francisella tularensis'

Performance Evaluation of Different Machine Learning Models

In Figure 4.11 and Figure 4.12, the ROC curves for taxonomy IDs "177416" and "644" are illustrated, respectively. For each figure, we have listed the six evaluated machine learning models as two groups for the convenience of following analysis. One group is called tree-based models which are mostly based on decision trees, while another group is called kernel-based models which are not based decision trees instead involving complex optimization algorithms. The tree-based models contain decision tree (DT), random forest (RF) and gradient boost machine (GBM). The kernel-based models include support vector machine (SVM), logistic regression (LR) and Gaussian Naïve Bayes model. The performances are presented as mean ROC curves from five-fold independent test results for different ratios.

As there are 1207 positive interaction pairs for taxonomy ID "177416", the dataset size is 121907 for a ratio of 1 : 100, which is larger than that of taxonomy ID "644". Somehow, the predictors performance became worse for the larger dataset, for both the two groups of models. One major outcome is that, the tree-based models appears to perform better for the prediction task in comparison with the kernel-based models. Although the tree-based models still outperformed the kernel-based models for each dataset, the overall performance was not stable across the different host-bacterium systems.

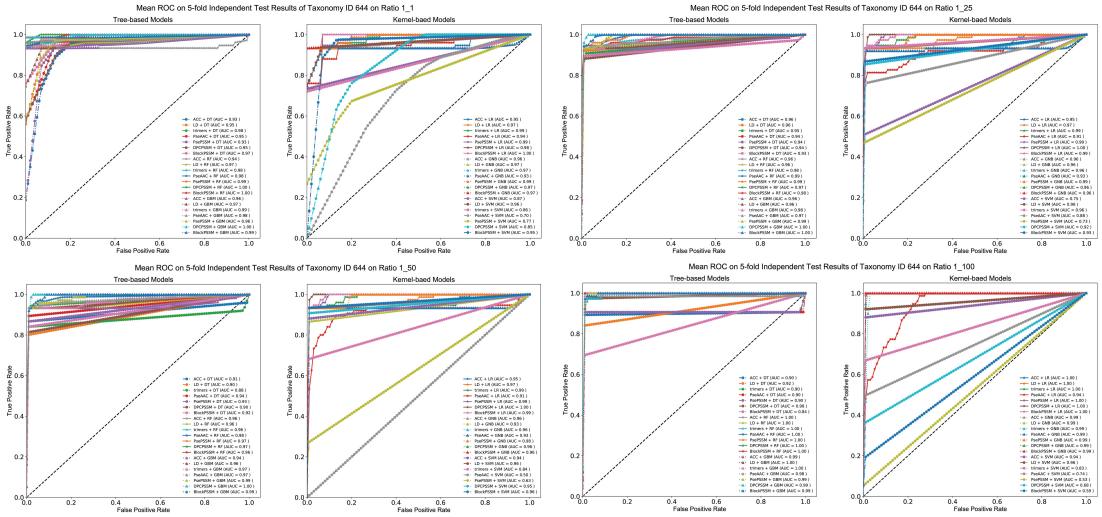


Figure 4.12: The ROC Curve for 'Aeromonas hydrophila'

Performance Evaluation of Different Feature Representation Algorithms

In the following tables, the results of accuracy value, F1 score, and Matthew's correlation coefficient value are reported. Since the results of each value are still of large amount, which include the performance for the combination set of six different machine learning model and seven feature representation algorithms, the best machine learning models with each feature representation algorithms are selected for the tables.

In Table 4.8, Table 4.9 and Table 4.10, the best results of all the predictors are listed accordingly for taxonomy ID '632'. For example, for the AC feature representation algorithm dataset, the best results of for ratios of $1:1$, $1:25$, $1:50$ and $1:100$ were all achieved by RF model with accuracies of 0.757082 ± 0.008000, 0.967350 ± 0.000365, 0.982521 ± 0.000128, and 0.990674 ± 0.000043, respectively. The tree-based models, including DT, RF, and GBM, have demonstrated a strong generalization ability in terms of providing effective and efficient performance. The other models, such as kernel-based model, including SVM, Gaussian Naïve Bayes (GNB) model and the LR model, however, are less robust compared with the tree-based models. Meanwhile, the training time was in higher demand than for the tree-based models. Taking CTM as the feature representation algorithm, the time spent training GBM for the dataset of ratio $1:100$ on taxonomy ID "632" was over 1,500 seconds. However, the time spent training the SVM model was

more than 23,000 seconds.

Following, an extension of discussion for future directions, which identifies the key issues and suggestions to build a robust and effective machine learning-based model, is presented.

Table 4.8: Results of Accuracy on 'Yersinia pseudotuberculosis'

Model	Accuracy			
	1:1	1:25	1:50	1:100
Auto Covariance (420D)	0.757082±0.008000 (RF)	0.96735±0.000365 (RF)	0.982521±0.000128 (RF)	0.990674±0.000043 (RF)
Local Descriptor (1260D)	0.720963±0.016687 (GBM)	0.965377±0.000487 (RF)	0.981676±0.000091 (RF)	0.990444±0.000060 (RF)
Conjoint Triad Method (686D)	0.700283±0.010306 (GBM)	0.965039±0.000311 (RF)	0.98176±0.000208 (SVM)	0.990523±0.000051 (SVM)
PseAAC (110D)	0.718697±0.014061 (GBM)	0.964374±0.000203 (RF)	0.981415±0.000113 (RF)	0.990391±0.000145 (GBM)
PsePSSM (80D)	0.709632±0.005540 (GBM)	0.966216±0.000450 (RF)	0.982049±0.000120 (RF)	0.990624±0.000044 (RF)
DPCPSSM (800D)	0.734278±0.009506 (GBM)	0.966053±0.000333 (RF)	0.98206±0.000208 (RF)	0.990585±0.000061 (RF)
Block-PSSM(800D)	0.729037±0.008095 (GBM)	0.965279±0.000464 (RF)	0.981698±0.000293 (GBM)	0.990551±0.000078 (RF)

Table 4.9: Results of F1 Score on 'Yersinia pseudotuberculosis'

Model	F1 Score			
	1:1	1:25	1:50	1:100
Auto Covariance (420D)	0.752226±0.006632 (RF)	0.31253±0.010944 (RF)	0.243679±0.012883 (RF)	0.154535±0.012569 (RF)
Local Descriptor (1260D)	0.727218±0.013162 (GBM)	0.255139±0.009452 (RF)	0.177423±0.010255 (DT)	0.173899±0.010245 (GBM)
Conjoint Triad Method (686D)	0.700275±0.006187 (GBM)	0.18578±0.010755 (RF)	0.180318±0.006771 (SVM)	0.129115±0.010062 (RF)
PseAAC (110D)	0.724976±0.010361 (GBM)	0.23077±0.011551 (SVM)	0.148855±0.011666 (GBM)	0.130497±0.010625 (GBM)
PsePSSM (80D)	0.720259±0.004842 (GBM)	0.256988±0.009757 (RF)	0.191165±0.012116 (RF)	0.143488±0.008093 (RF)
DPCPSSM (800D)	0.742534±0.008470 (GBM)	0.259213±0.010695 (RF)	0.205636±0.012626 (RF)	0.154714±0.013181 (DT)
Block-PSSM(800D)	0.739192±0.007676 (GBM)	0.207103±0.009433 (RF)	0.175258±0.011638 (GBM)	0.157700±0.004793 (GBM)

Table 4.10: Results of MCC Value on 'Yersinia pseudotuberculosis'

Model	MCC Value			
	1:1	1:25	1:50	1:100
Auto Covariance (420D)	0.514740±0.016240 (RF)	0.389241±0.010464 (RF)	0.335746±0.010207 (RF)	0.253434±0.010138 (RF)
Local Descriptor (1260D)	0.442457±0.032907 (GBM)	0.328314±0.013210 (RF)	0.256948±0.008037 (RF)	0.233817±0.012761 (GBM)
Conjoint Triad Method (686D)	0.400747±0.020562 (GBM)	0.297864±0.012409 (RF)	0.249646±0.012510 (RF)	0.219050±0.009724 (SVM)
PseAAC (110D)	0.437930±0.027559 (GBM)	0.270015±0.008225 (RF)	0.241630±0.013639 (GBM)	0.212697±0.018621 (GBM)
PsePSSM (80D)	0.420486±0.010975 (GBM)	0.348116±0.013818 (RF)	0.294669±0.010245 (RF)	0.242970±0.007387 (RF)
DPCPSSM (800D)	0.469595±0.018729 (GBM)	0.344422±0.010812 (RF)	0.300172±0.014612 (RF)	0.240708±0.009831 (RF)
Block-PSSM (800D)	0.459515±0.016168 (GBM)	0.321348±0.005442 (SVM)	0.274589±0.018411 (RF)	0.227935±0.004565 (GBM)

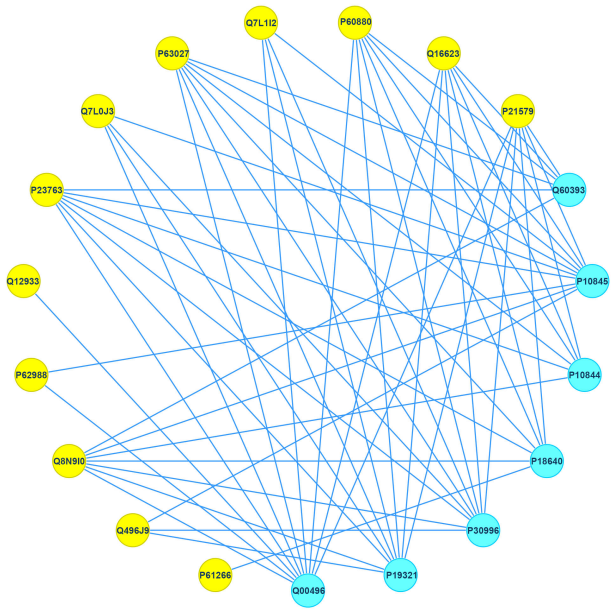


Figure 4.13: Protein Interaction Map between *Homo Sapiens* and Clostridium botulinum (ID: 1491)

4.4.3 Further Discussion

Given different PPI networks, such as the HB-PPI between *Homo sapiens* and Clostridium botulinum (ID: 1491), and the interaction between Homo sapiens and Yersinia pseudotuberculosis subsp. pestis (ID: 632), the positive interactions networks have presented different complexities. As we can see, it still requires huge amounts of work towards the completeness of human-bacterium protein-protein interactions network. They have indicated different pathways between the different species. Figure 4.13 and Figure 4.14 show diagrams of two different interaction networks for taxonomy IDs 1491 and 632, respectively.

To accomplish a robust performance of predicting HB-PPIs, the relationship between positive and negative protein interactions requires further consideration. There have been several methods dedicated to one-class classification tasks, such as semi-supervised learning [270–272], to leverage the power of singularly labelled data and unlabelled data. This may help to improve the performance of protein interaction prediction regardless of the ratio between the positive and negative protein interactions. Meanwhile, since

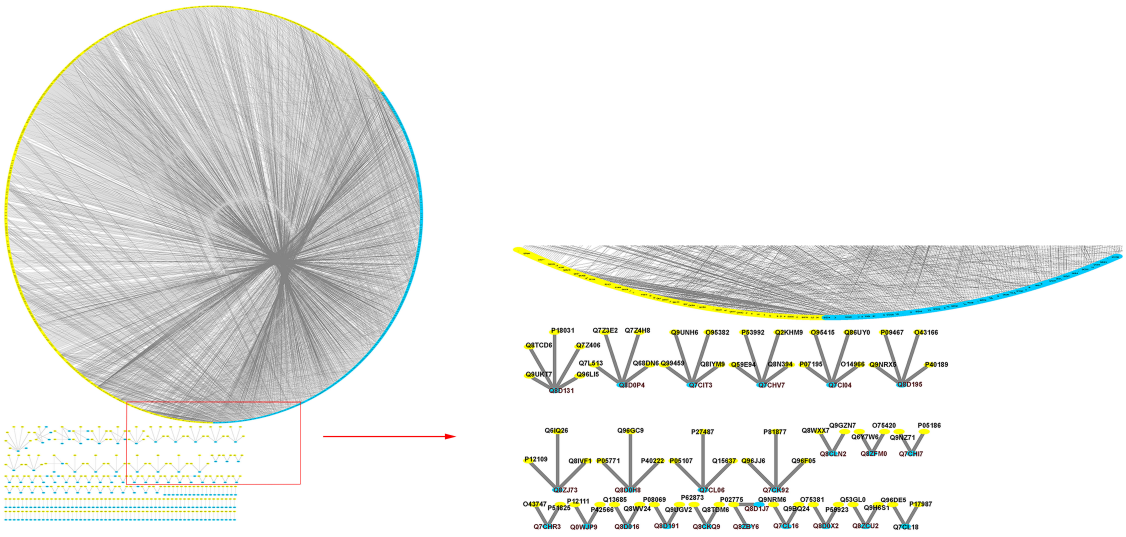


Figure 4.14: Protein Interaction Map between *Homo Sapiens* and Yersinia pseudotuberculosis subsp. pestis (ID: 632)

sequential feature-representation algorithms have been an active and challenging area, a better feature representation algorithm is needed to help build a sequence based end-to-end machine learning model [3, 273, 274] for predicting HB-PPIs.

4.5 Summary

In this chapter, we have evaluated the predictions task for HP-PPIs in a systematic manner. The focus was on leveraging machine learning-based models as the primary computational method. We first presented a wide and deep review on currently available resources and computational tools. As noted in the literature review in Chapter. 4.2, to evaluate the computational tools developed for prediction tasks of HP-PPIs, a dedicated data curation process was implemented and a pipeline for HB-PPI studies was summarized in Chapter. 4.3, which included numerous sequential feature-representation algorithms and machine-learning models. Several other computational methods concerning HB-PPIs were also evaluated.

Given the study of HP-PPIs, we have tried to determine the impacts caused by different ratios of benchmark datasets, different feature-representation algorithms and different machine-learning models. The experimental results in Chapter. 4.4 indicated that to better

utilise machine learning models and harness the power of accumulated protein interaction data, a more robust and more powerful computational model is required to achieve better performance across different HB-PPI prediction tasks.

In following chapters, the design and details with regard to develop novel machine learning-based models with novel feature representation algorithms are presented, which has greatly improved the performance for discovery of interactions of HP-PPIs.

Chapter 5

HETEROGENEOUS INFORMATION MINING AND ENSEMBLING MODEL FOR DISCOVERY OF HP-PPIS

Research on protein-protein interactions (PPIs) data is of critical meaning towards the understanding of the infectious mechanisms of diseases. In previous chapters, reviews with regard to host-pathogen interactions resources and computational models evaluation have been broadly conducted. However, it remains a challenge to improve the prediction performance of PPIs of inter-species, particularly between host and pathogen. In this chapter, a novel framework for HP-PPIs prediction based on *H*eterogeneous *I*nformation *M*ining and *E*nsembling (HIME) process to effectively learn from the interaction data. In particular, the proposed approach introduces an ensemble process together with substantial features that generate better performance of HP-PPIs prediction task. The performance of the proposed framework is validated on the curated protein interactions datasets. The extensive experiments show that HIME achieves higher performance over all existing methods reported in literature so far.

In this chapter, a brief introduction and review work will be reported in Chapter. 5.1 and 5.2 to build the context for discovery of HP-PPIs. The detail of proposed HIME model and experiment settings will be presented in Chapter. 5.3. In Chapter. 5.4, a comprehensive comparison against different machine-learning models will debriefed.

5.1 Introduction

Analyzing and understanding protein-protein interactions (PPIs) for inter-species interactions is of great importance, such as the interactions between human and pathogens [26, 27]. One of the earliest studies was on the symptom of anthrax, which was identified as primarily being caused by the protein interactions between human and Bacillus anthracis. Bacillus anthracis is a type of bacterium pathogens, where people want to fully understand mechanisms with the protein interactions map between Bacillus anthracis and Homo sapiens (the host).

However, the experiment results to investigate protein-protein interactions are still very limited. The identification of protein-protein interactions is traditionally conducted by *in vitro* and *in vivo* methods, which are deemed cost-sensitive task for both time and resources. To effectively generate high-fidelity PPIs prior to biology experiments, there has been numerous studies introducing computational methods to facilitate the process. One major category is to build machine learning-based model with different protein data, such as protein sequence data [34], gene ontology data [275], and protein structure data [276], for the prediction of protein interactions.

Among these, sequence information is considered as the main protein information because of its substantial accumulation in a large scale. Specifically, the proteins have been determined uniquely by the sequence information as for their physical and biochemical characteristics. By analyzing the protein sequence information hosted by the Universal Protein Resource (UniProt), the past studies had indicated that combining machine learning-based models with protein sequence data mining would benefit the prediction and analysis of protein interactions task [26, 32, 162]. More recently, Soyemi et al. [277] have reviewed the relevant data of inter-species/host-parasite protein interaction in a comprehensive manner, though the quantitative evaluation is still void. Inspired from the idea in [26, 277], a systematic evaluation of machine learning-based models, include the methods from literature focusing on the prediction of HP-PPIs, was conducted in

Chapter 4.

Given the void of systematic evaluation of machine learning-based HP-PPIs prediction models, the first of this kind of evaluation show that there is plenty of room for improvements to achieve a robust and efficient machine learning-based model. In this chapter, an ensemble machine learning-based model is proposed through mining the heterogeneous information of protein data. The proposed framework demonstrates its robustness and accuracy based on Heterogeneous Information Mining and Ensembling (HIME) prediction model to harness the power of heterogeneous information, thereby greatly improving the prediction performance. The experimental results indicate that the HIME model achieves the best and most robust performance for prediction of HP-PPIs in comparison with the state-of-the-art.

5.2 Review and Motivation of HIME Study

There have been a large body of research on protein-protein interactions, aiming at developing cost-effective methods for prediction of protein interactions [278–281]. Since there are different characteristics presented by protein, the methods include text mining method, network analysis method, kernel-based method, machine learning-based method and so on. However, these methods are presented as feasible and effective methods in a combination with the corresponding protein characteristics, such as sequence data, gene ontology data, gene expression data.

In recent years, protein sequence data has prevailed in numerous research areas of protein, for example protein structure prediction, protein function prediction and as in our study, PPIs prediction. In [282], development of Pups (pupylation site predictor) involved the utilization of protein sequences and machine learning model, in which the pseudo-amino acid composition information was particularly employed. To deal with the avalanche of newly sequenced protein data, the feature representation methods of protein sequence data were well designed as one of the important components for machine learning-based PPIs prediction models [33, 34, 156, 278]. Because sequence data was the

most abundant data benefiting from high-throughput technology development, it would be beneficial to understand the performance in computational models and develop a more efficient model for HP-PPIs prediction.

In Chapter. 3 and Chapter. 4, a comprehensive review with regard to the HPI resources was conducted by manually examining with 'Abstract' from the first 400 returning items ranking by best relevance out of more than 4,000 papers. A huge number of databases were reviewed prior to be included in the study. The selected eleven public databases were utilised in this chapter. With the reported performance of numerous feature representation algorithms and different machine learning models, how to mine the most of protein sequence information to enhance the prediction performance is the goal. In following chapter, HIME model is presented to harness the heterogeneous information from protein sequence and it has presented a better performance than the others.

5.3 The HIME Model

5.3.1 Material Brief

For the collected data, only positive protein interactions data are available from the databases. Two steps are conducted to process the data. One is to reduce the ID information redundancy, as there may be duplicate entries when combining data from different databases. Another is related to sequence length. The proteins with less than 50 amino acids are discarded since they may be non-functional fragments. In Table 5.1, the statistic of the collected positive human-bacterium protein-protein interactions is presented, which includes the species of 'Clostridium botulinum', 'Aeromonas hydrophila', 'Shigella paradysenteriae', 'Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4)', 'Bacillus anthracis bacterium' and 'Yersinia pseudotuberculosis subsp. pestis (Lehmann and Neumann 1896) Bercovier et al. 1981'. In most of the literature, building a negative interaction dataset by randomly pairing proteins from the set of unknown interacting PPIs is utilized [34, 156, 218], since none standard protocol defines the

Table 5.1: Selected Human-Pathogens Interactions Systems' Datasets

Taxonomy ID	Bacterium Pathogens	Positive Interactions	Negative Interactions	Interactions Number of Training Dataset	Interactions Number of Independent Dataset
1491	Clostridium botulinum	57	57	90	24
644	Aeromonas hydrophila	73	73	116	30
623	Shigella paradysen-teriae	105	105	168	42
177416	Francisella tularensis subsp. tularensis	1207	1207	1930	484
1392	Bacillus anthracis bacterium	2810	2810	4496	1124
632	Yersinia pseudotu-berculosis subsp. pestis	3528	3528	5644	1412

negative pairing strategy.

Following Chapter. 4, protein sequence data, which is dominantly published by UniProtKB database, was utilised. The information helps building the negative HP-PPIs as well as building the independent datasets. To obtain an extensive evaluation, a dedicated preparation of independent datasets is applied, which datasets should not be used during the training and will be reported with different measurements to evaluate the model performance.

Thus, a randomly selection of one-fifth HP-PPIs from both positive and negative interactions as the independent dataset is conducted. The rest PPIs of positive and negative interactions were combined as the training set. Since the construction of the negative interactions is achieved by a random sampling method, the random sampling for the

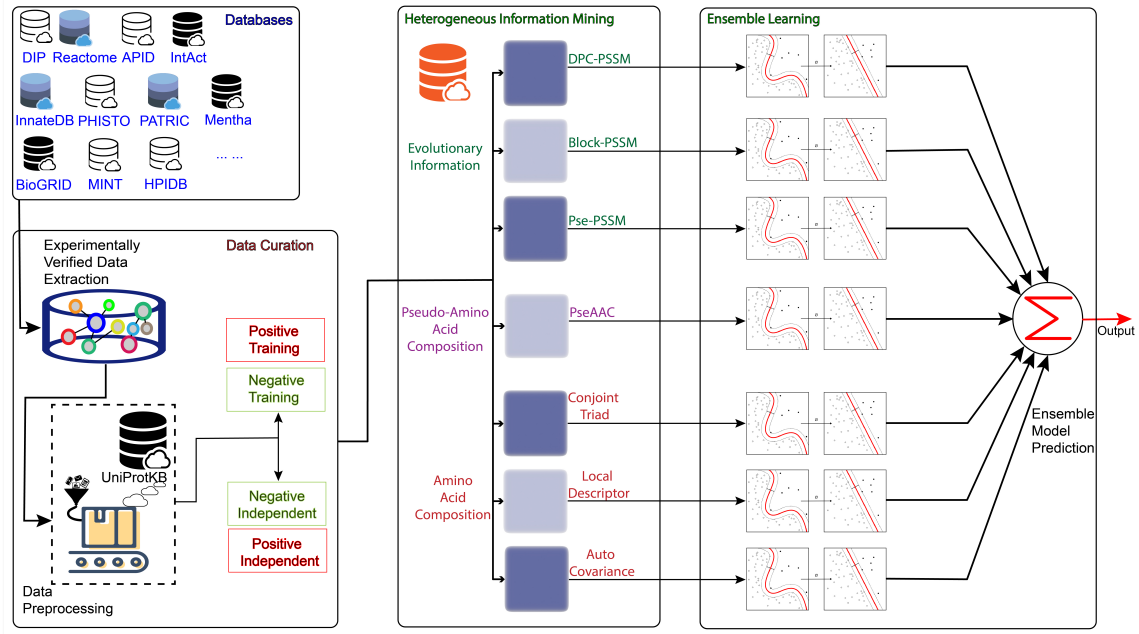


Figure 5.1: The Framework of HIME Model

negative interactions was applied five times and the evaluation was measured with statistic means and variations to reduce the bias caused by negative interactions.

5.3.2 The HIME Model

This chapter firstly introduce the HIME model, then the details of each part of HIME model will be explained.

The proposed heterogeneous information mining and ensembling (HIME) model is shown in Figure 5.1, which leverages the mining and ensembling process of heterogeneous information of sequence data, and also includes the learning process. HIME model is a sequence-based model, since the protein sequence data is considered as one of the most abundant data. The overwhelming sequence data has exclusively stimulated the ongoing research to improve the prediction performance based on novel feature representation algorithms of sequence data and machine learning models. It helps to generalize the computational models on a larger dataset and various species and genres.

HIME model tackle the heterogeneous information of sequence data in three different types, as shown in Figure 5.1, which are amino acid composition information, pseudo-

amino acid composition information and evolutionary information. Multiple training models are produced for different information, and HIME model subsequently utilises ensemble learning techniques to make the prediction with high performance for different human-pathogen interactions systems.

Heterogeneous Information of Sequence Data

Encoding sequence data as feature vectors is the first step in building computational model for prediction [34, 218]. Three different types of heterogeneous information of sequence data are explored in our proposed model, which helps to build a robust and efficient model. Since the information was reported in details in Chapter. 4, they will be briefly reported in this chapter.

Amino acid composition information Amino acid composition information is dominantly inferred by the amino acids order of protein sequence data. There are several different methods converting this information into feature vectors. One was considering several adjacent amino acids as one region in the sequence, which was also called conjoint triad method feature or k-mer [283]. It considered the protein in segments to be functional between different proteins, which firstly classified the 20 different types of amino acids into seven groups according to their physiochemical characteristics. This encoded the sequence data into a 343-dimension vector. The flexibility of this method allows the region to be two, four, and other length adjacent amino acids.

Another approach based on amino acid composition information is to discover the auto covariance relationship among amino acids [33]. Auto covariance method considered each amino acid with its seven physicochemical properties. For different properties, the auto covariance relationship was calculated for two different locations of amino acids given the maximum distance *Dis*. The dimension of feature vector generated via auto covariance method would be $Dis * 7$, when all seven properties are employed.

The last popular method for amino acid composition information is local descriptor [232], which has divided the protein sequence information into 10 regions of six different

types, including by quarter division, half division, central 50% region, first 75% region, last 75% region and central 75% region. Local descriptor specifically defined three different descriptors for each region, including composition, transition and distribution. This generated seven features for composition, 21 features for transition and 35 features for distribution. Totally with the 10 regions, local descriptor generated 630-dimension feature vector for single protein sequence.

Pseudo-amino acid information Even though amino acid composition information takes consideration of sequence order to some extent, there is still some information loss when directly encoding sequence data based on composition information. Thus, pseudo-amino acid information is discovered as an important type of information of sequence data [233].

Evolutionary information Another important information of sequence data is the evolutionary information, which represents the continuous change and evolution trends in a given reference protein database. The information is referred as a scoring matrix to indicates the probability of related amino acid types in corresponding position. It is commonly derived by aligning a set of sequence, which is considered to be functionally related. One important matrix firstly derived is called the position-specific scoring matrix (PSSM), which is a T*20 matrix for a given protein sequence. T represents the length of its corresponding protein sequence. Several algorithms have been developed to generate feature vector for single protein sequence. The first one is pseudo position-specific score matrix (Pse-PSSM), which combines the idea of pseudo-amino acid composition [239]. Pse-PSSM represented the original PSSM by compressing the matrix values vertically into their corresponding mean value. This means, after transformation, PSSM becomes a 20-dimension Pse-PSSM vector. Another one is called Block-PSSM by dividing sequence data into 20 equal blocks [249]. Each block represents five percent of a sequence. For each block, a 20-dimension vector is extracted. This generates a 20*20=400-dimension vector totally with 20 blocks. The last one is the traditional dipeptide composition PSSM (DPC-

PSSM) [251]. It calculated the covariance of two adjacent amino acid and represented the information in a 400-dimension feature vector.

The heterogeneous information of sequence data have been categorized in three different types, as shown in Figure 5.1. Different algorithms including conjoint triad method (CTM) [283], auto covariance (ACC) [33], local descriptor (LD) [232], PseAAC [233], pseudo position-specific score matrix (Pse-PSSM) [239], transition dipeptide composition PSSM (DPCPSSM) [251] and block PSSM (BlockPSSM) [249] algorithms, are subsequently incorporated in HIME model.

Ensemble Learning

Machine learning-based models have been widely applied for prediction of bioinformatics tasks recently. Mostly, the models are compared and the best of the models is selected as the applied computational model.

Ensemble learning model is designed with multiple machine learning models, which are called 'base learner' for same task [284]. Typically, ensemble learning model benefits from the integration of individual base learners to achieve a robust and superior performance. Even though there are different categories of ensemble learning model, various applications have shown that none of them could be outstanding consistently [285–287].

Generally, the ensemble learning model can be deployed either vertically or horizontally [287]. To avoid building a single strong machine learning model in the task, HIME model leverages the heterogeneous information and plenarily exerts the various base learners in a horizontal way. lightGBM [288], one of the recently popular tree-based models, is selected as the base learner in the model to build HIME for prediction of human-pathogen protein-protein interactions.

Algorithms 1 illustrates the procedure of HIME model. Our model not only leverages the precision and diversity from base learner, but also emphasises the diversity from the heterogeneous information mining process. As a result, the proposed HIME model

Algorithm 1 Heterogeneous Information Ensembling Process

Input: Dataset $D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$

- Heterogeneous information feature representation algorithms $\mathfrak{R}_1, \mathfrak{R}_2, ..., \mathfrak{R}_T$
- Base learner algorithms $\mathscr{L}_1, \mathscr{L}_2, ..., \mathscr{L}_T$
- Ensemble learner $\mathscr{L}$

Output: $H(x)$

Process:

for $t = 1$ to T **do** *Heterogeneous information mining*
    $\mathscr{D}_t = \mathfrak{R}_t()$ %Mining heterogeneous information
        %and applying the different feature
        %representation algorithms
end for
for $t = 1$ to T **do**
    $h_t = \mathscr{L}_t(\mathscr{D}_t)$ %Training a base learner algorithm h_t
        %by applying the base leanrner
        %algorithm $\mathscr{L}_t$ to the dataset $\mathscr{D}_t$
end for
$\mathscr{D}' = \varnothing$ %Collect the base learners
for $i = 1$ to m **do**
    for $t = 1$ to T **do**
        $z_{it} = h_t(x_i)$ %Use h_t to classify the Dataset D
    end for
    $\mathscr{D}' = \mathscr{D}' \cup \{((z_{i1}, z_{i2}, ..., z_{iT}), y_i)\}$
end for
$h' = \mathscr{L}(\mathscr{D}')$
Output: $H(x) = h'(h_1(x), ..., h_T(x))$.

is capable to enhance the performance fueled by the designed information mining and ensembling procedure.

5.3.3 Baseline Models

In this study, different methods, such as [218] and [34] from literature, and traditional machine learning models including random forest, support vector machine, logistic regression model, Gaussian naïve Bayes, decision tree and gradient boosting machine, are used in the prediction task of HP-PPIs. These models explicitly demonstrate different capabilities on different tasks, such as classification task and time series regression task. Since these models are traditionally used in different tasks, as mentioned in Chapter. 4, the performance of different groups of feature representation algorithms and machine learning models is included. This results in 42 different combinations as the first group baseline models. The hyperparameters are subsequently obtained by 5-fold cross validation test for each classifier according to the dataset.

Meanwhile, two methods from literature, which are [34] and [218] were included. In [218], random forests model was selected as the ensemble model to learn from the host-parasite protein-protein interactions. A variant version of amino acid triplets algorithm was used as the feature representation algorithm. [34] applied SVM as the computational model with the proposed protein sequence representation algorithm to predict the human-pathogen protein-protein interactions.

5.3.4 Performance Measurements

To evaluate the performance of HIME model, numerous metrics are compared, including the accuracy, precision, recall, specificity, F1-score, the area under curve (AUC) value and Matthew's correlation coefficient (MCC) score. The receiver operating characteristic curves (ROC) is also collected. The definition can be reffered to Equa. 4.22.

5.4 Results and Discussion

The results of a 5-fold independent test of the six different taxonomy IDs datasets were collected to present the performances with both the mean values and the deviations.

5.4.1 Baseline Models

The evaluations on traditional machine learning models, including decision tree (DT), random forest (RF), gradient boosting machine (GBM), logistic regression (LR), Naïve Bayesian and support vector machine (SVM) will be discussed firstly. Seven different feature representation algorithms of sequence data are included and the corresponding models are built upon six traditional machine learning models, which result in 42 different models. Table 5.2 includes the accuracy and F1 score for all the evaluated models, including HIME model. The performances of traditional models, 'Model_1' and 'Model_2', share a same fluctuation trend concerning different datasets, which worst performances are all observed with 'HB_6'. HIME model has shown its enhanced performance by improving the results of accuracy, in which multi feature representation algorithms are utilised to mine the heterogeneous information. The proposed Algorithm 1 has further improved the performance by combining the horizontal ensemble procedure for the heterogeneous information. For both accuracy and F1 score, HIME model has demonstrates a best performance in comparison with the others. Following, we will show more details with regard to the ROC curves.

5.4.2 HIME Model Performance and Comparison

In Table 5.2, the best models are indicated in bold fonts. We can clearly observe that for five prediction tasks, which are 'HB_1', 'HB_3', 'HB_4', 'HB_5' and 'HB_6', the best performances are all achieved by our proposed HIME model. This indicates that mining and ensembling heterogeneous information of sequence data indeed help boosting the model performance.

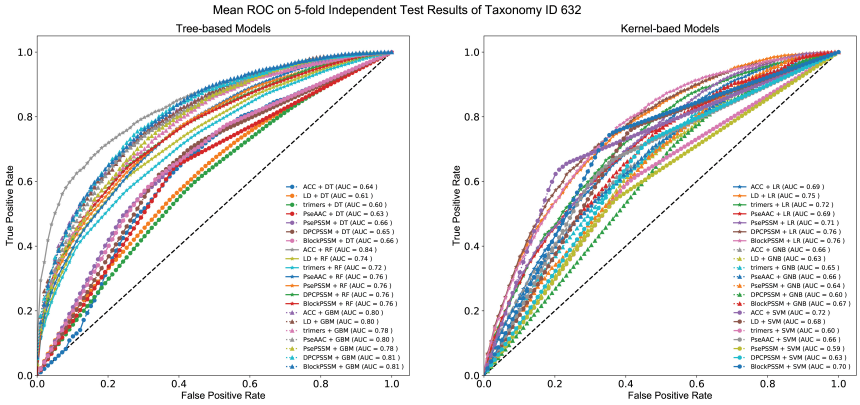


Figure 5.2: The ROC Curves for 'HB_6' of Traditional Models

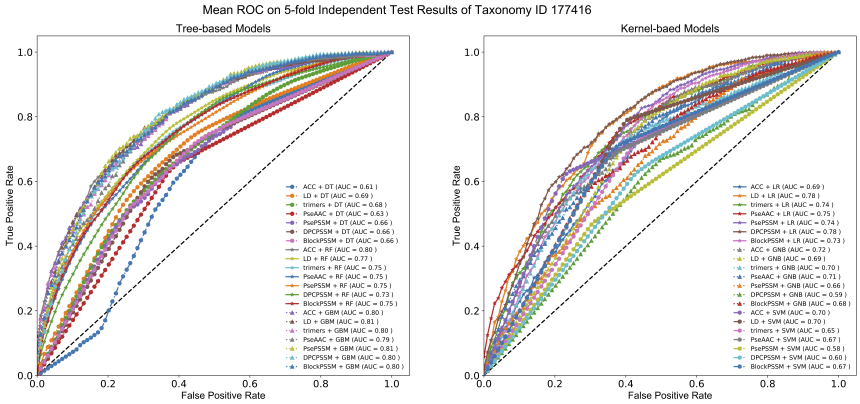


Figure 5.3: The ROC Curves for 'HB_4' of Traditional Models

In Figure 5.2 and Figure 5.4, we have shown partial results of the ROC curves for discussion due to the limited space. The ROC curves show that, different types of protein sequence information generate diverse learners, which generate different performance. One particularly selected information may not be sufficient to produce a robust model. Moreover, the performance will become worse when the dataset is larger.

In comparison with Figure 5.2, the ROC curves for five-times independent test of 'HB_6' with HIME model is illustrated in Figure 5.4. Since the proposed HIME model utilizing heterogeneous information, the model obtains a more robust and accurate performance than the other baseline models. From Table 5.2, it is easy to see the proposed HIME model has a better prediction capability than the other methods. Out of the six different types of dataset, it has achieved five of the best performance, in which each dataset may have a different second-best model. In Figure 5.3 and Figure 5.5, the ROC curves for 'HB_4' are also illustrated with the conclusion that HIME model has achieved a better performance. Given the performance metrics including Specificity, MCC and AUC values, we have also observed the same performance comparison results, in which HIME model outperforms the others. The performance comparison demonstrates that, the proposed HIME model outperforms most of the predictor compared in this study for different human-pathogen PPIs prediction tasks. Hence, the heterogeneous information mining and ensembling strategy benefits the performance improvement in this work.

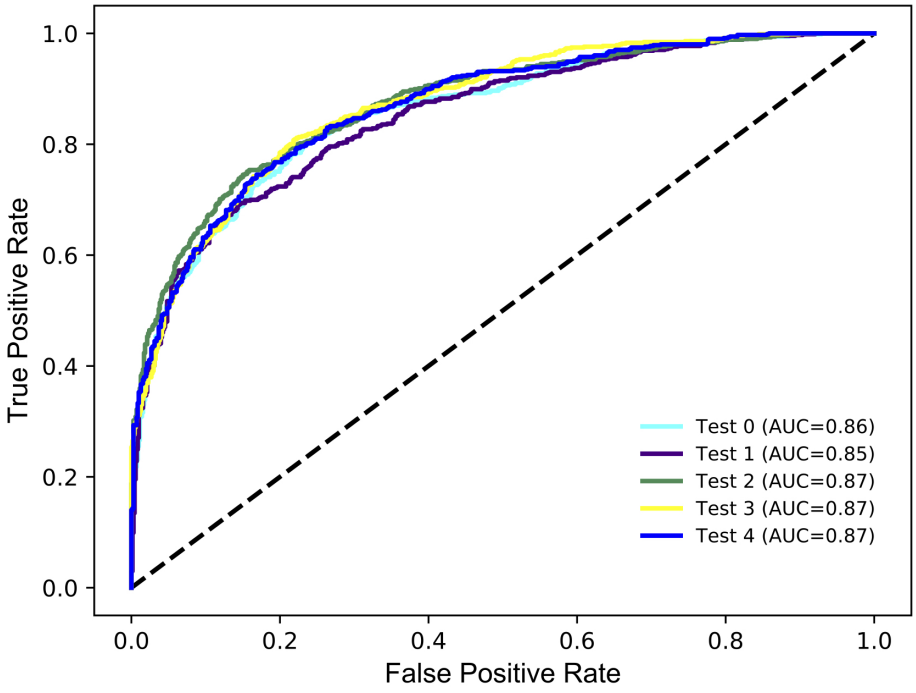


Figure 5.4: The ROC Curves for 'HB_6' of HIME Model

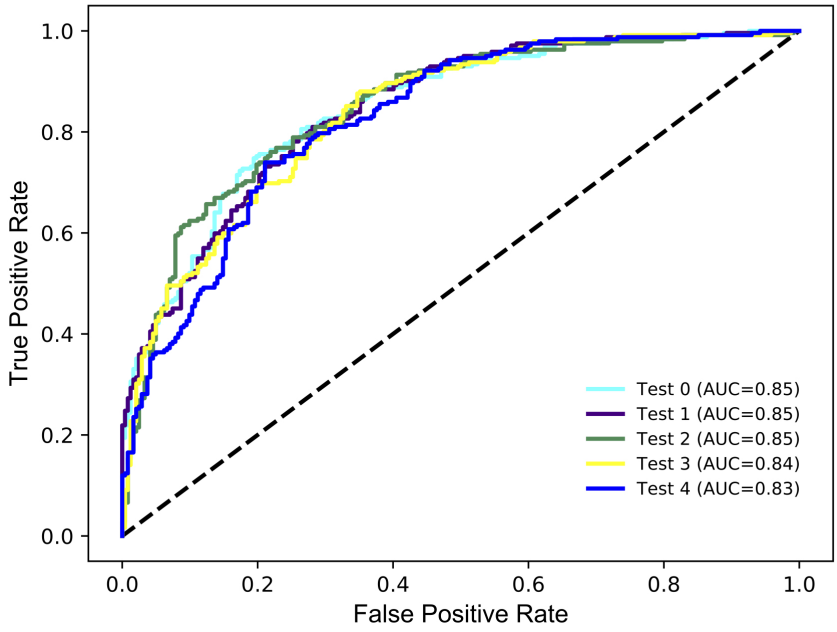


Figure 5.5: The ROC Curves for 'HB_4' of HIME Model

Table 5.2: Results of Accuracy and F1 Score for Models

Model		Accuracy						F1 Score					
		HB_1[a]	HB_2	HB_3	HB_4	HB_5	HB_6	HB_1	HB_2	HB_3	HB_4	HB_5	HB_6
RF	$\Re_1$[b]	1.000±0.000	0.967±0.000	0.824±0.036	0.725±0.008	0.773±0.011	0.757±0.008	1.000±0.000	0.966±0.000	0.811±0.040	0.730±0.005	0.770±0.011	0.752±0.007
	$\Re_2$	1.000±0.000	0.967±0.000	0.757±0.038	0.696±0.007	0.689±0.010	0.661±0.015	1.000±0.000	0.966±0.000	0.771±0.034	0.715±0.011	0.710±0.009	0.691±0.012
	$\Re_3$	0.975±0.033	0.967±0.000	0.752±0.053	0.686±0.011	0.670±0.015	0.651±0.012	0.974±0.036	0.966±0.000	0.768±0.049	0.707±0.009	0.696±0.014	0.680±0.011
	$\Re_4$	1.000±0.000	0.967±0.000	0.795±0.039	0.682±0.017	0.701±0.005	0.684±0.016	1.000±0.000	0.966±0.000	0.807±0.033	0.705±0.016	0.723±0.006	0.711±0.012
	$\Re_5$	1.000±0.000	0.973±0.013	0.876±0.035	0.671±0.014	0.680±0.004	0.683±0.009	1.000±0.000	0.972±0.014	0.878±0.038	0.696±0.008	0.704±0.004	0.711±0.006
	$\Re_6$	1.000±0.000	0.973±0.013	0.814±0.063	0.679±0.010	0.690±0.015	0.678±0.009	1.000±0.000	0.974±0.013	0.831±0.049	0.709±0.012	0.712±0.011	0.707±0.008
	$\Re_7$	1.000±0.000	**0.993±0.013**	0.838±0.066	0.678±0.014	0.687±0.011	0.676±0.006	1.000±0.000	**0.993±0.014**	0.852±0.057	0.707±0.010	0.709±0.010	0.709±0.004
SVM	$\Re_1$	1.000±0.000	0.867±0.000	0.800±0.024	0.700±0.013	0.653±0.015	0.719±0.012	1.000±0.000	0.846±0.000	0.775±0.016	0.674±0.014	0.656±0.013	0.697±0.012
	$\Re_2$	0.975±0.033	0.960±0.013	0.676±0.053	0.696±0.012	0.708±0.016	0.676±0.007	0.977±0.031	0.959±0.013	0.705±0.041	0.722±0.012	0.701±0.017	0.703±0.005
	$\Re_3$	1.000±0.000	0.860±0.033	0.790±0.046	0.651±0.008	0.696±0.007	0.597±0.009	1.000±0.000	0.835±0.046	0.792±0.036	0.678±0.009	0.702±0.007	0.599±0.007
	$\Re_4$	1.000±0.000	0.700±0.101	0.752±0.049	0.666±0.019	0.604±0.025	0.661±0.018	1.000±0.000	0.762±0.061	0.741±0.040	0.670±0.020	0.539±0.125	0.657±0.014
	$\Re_6$	1.000±0.000	0.767±0.060	0.729±0.058	0.583±0.008	0.665±0.007	0.588±0.005	1.000±0.000	0.734±0.062	0.722±0.028	0.531±0.010	0.682±0.005	0.567±0.002
	$\Re_6$	0.992±0.017	0.853±0.086	0.648±0.035	0.601±0.010	0.642±0.016	0.635±0.007	0.992±0.016	0.877±0.064	0.701±0.031	0.615±0.008	0.644±0.026	0.663±0.005
	$\Re_7$	1.000±0.000	0.947±0.027	0.900±0.035	0.673±0.011	0.635±0.047	0.699±0.009	1.000±0.000	0.948±0.026	0.908±0.032	0.683±0.006	0.665±0.090	0.713±0.008
LR	$\Re_1$	0.942±0.033	0.967±0.000	0.819±0.032	0.635±0.018	0.656±0.012	0.645±0.006	0.946±0.031	0.966±0.000	0.818±0.030	0.642±0.020	0.661±0.011	0.654±0.004
	$\Re_2$	0.983±0.033	0.953±0.016	0.695±0.051	0.709±0.017	0.709±0.013	0.686±0.010	0.985±0.031	0.953±0.016	0.691±0.039	0.719±0.018	0.720±0.014	0.696±0.006
	$\Re_3$	0.983±0.020	0.960±0.013	0.829±0.028	0.681±0.013	0.692±0.010	0.659±0.010	0.984±0.020	0.959±0.013	0.828±0.029	0.684±0.015	0.698±0.007	0.659±0.011
	$\Re_4$	0.975±0.033	0.887±0.045	0.843±0.061	0.673±0.017	0.676±0.014	0.643±0.015	0.974±0.036	0.892±0.041	0.840±0.065	0.673±0.011	0.664±0.007	0.624±0.024
	$\Re_5$	1.000±0.000	0.953±0.016	0.857±0.052	0.678±0.008	0.662±0.008	0.648±0.008	1.000±0.000	0.953±0.016	0.856±0.055	0.692±0.006	0.670±0.005	0.659±0.006
	$\Re_6$	0.992±0.017	0.960±0.025	0.876±0.038	0.712±0.018	0.706±0.016	0.694±0.013	0.992±0.016	0.961±0.024	0.881±0.035	0.723±0.016	0.713±0.012	0.702±0.011
	$\Re_7$	1.000±0.000	0.973±0.025	0.895±0.029	0.667±0.015	0.702±0.017	0.695±0.014	1.000±0.000	0.973±0.025	0.899±0.028	0.675±0.015	0.711±0.016	0.707±0.013
Naïve Bayes	$\Re_1$	1.000±0.000	0.967±0.000	0.767±0.061	0.661±0.007	0.625±0.002	0.608±0.019	1.000±0.000	0.966±0.000	0.794±0.043	0.714±0.005	0.672±0.003	0.653±0.013
	$\Re_2$	1.000±0.000	0.967±0.000	0.667±0.050	0.658±0.007	0.634±0.008	0.597±0.015	1.000±0.000	0.966±0.000	0.697±0.030	0.708±0.006	0.687±0.006	0.646±0.009
	$\Re_3$	1.000±0.000	0.967±0.000	0.733±0.049	0.621±0.016	0.639±0.011	0.609±0.013	1.000±0.000	0.966±0.000	0.724±0.034	0.685±0.006	0.667±0.003	0.614±0.026
	$\Re_4$	1.000±0.000	0.947±0.040	0.619±0.054	0.638±0.022	0.580±0.009	0.574±0.006	1.000±0.000	0.947±0.036	0.524±0.088	0.622±0.039	0.341±0.032	0.380±0.011
	$\Re_5$	1.000±0.000	0.967±0.000	0.833±0.040	0.609±0.008	0.592±0.014	0.589±0.009	1.000±0.000	0.966±0.000	0.828±0.039	0.574±0.005	0.539±0.016	0.545±0.007
	$\Re_6$	1.000±0.000	0.960±0.013	0.619±0.045	0.571±0.015	0.565±0.005	0.586±0.007	1.000±0.000	0.958±0.015	0.673±0.033	0.618±0.010	0.602±0.013	0.654±0.003
	$\Re_7$	1.000±0.000	0.967±0.000	0.752±0.061	0.627±0.014	0.610±0.003	0.632±0.007	1.000±0.000	0.966±0.000	0.771±0.052	0.647±0.010	0.616±0.001	0.665±0.005
GBM	$\Re_1$	0.892±0.090	0.933±0.021	0.771±0.089	0.719±0.009	0.744±0.013	0.724±0.011	0.900±0.079	0.934±0.020	0.758±0.097	0.727±0.010	0.747±0.013	0.733±0.009
	$\Re_2$	0.975±0.020	0.953±0.016	0.814±0.038	0.728±0.013	0.753±0.017	0.721±0.017	0.976±0.020	0.953±0.016	0.817±0.037	0.731±0.013	0.754±0.015	0.727±0.013
	$\Re_3$	0.867±0.085	0.960±0.013	0.795±0.039	0.714±0.016	0.735±0.006	0.700±0.010	0.860±0.098	0.959±0.013	0.794±0.040	0.719±0.017	0.737±0.006	0.700±0.006
	$\Re_4$	1.000±0.000	0.953±0.034	0.833±0.058	0.720±0.020	0.737±0.013	0.719±0.014	1.000±0.000	0.954±0.032	0.842±0.054	0.724±0.022	0.741±0.011	0.725±0.010
	$\Re_5$	1.000±0.000	0.933±0.021	0.886±0.059	0.728±0.015	0.720±0.010	0.710±0.006	1.000±0.000	0.935±0.020	0.893±0.052	0.737±0.014	0.722±0.011	0.720±0.005
	$\Re_6$	0.992±0.017	0.987±0.016	0.824±0.072	0.725±0.011	0.738±0.007	0.734±0.010	0.992±0.016	0.987±0.016	0.830±0.068	0.735±0.010	0.738±0.009	0.743±0.008
	$\Re_7$	1.000±0.000	0.967±0.030	0.910±0.038	0.719±0.011	0.743±0.007	0.729±0.008	1.000±0.000	0.967±0.029	0.915±0.035	0.729±0.009	0.748±0.005	0.739±0.008
DT	$\Re_1$	0.900±0.077	0.927±0.033	0.676±0.061	0.628±0.015	0.604±0.017	0.624±0.013	0.905±0.077	0.929±0.030	0.681±0.072	0.655±0.009	0.640±0.012	0.656±0.011
	$\Re_2$	0.942±0.057	0.953±0.016	0.729±0.081	0.650±0.028	0.593±0.003	0.590±0.021	0.936±0.070	0.953±0.016	0.722±0.092	0.653±0.028	0.587±0.055	0.591±0.038
	$\Re_3$	0.883±0.085	0.980±0.016	0.695±0.087	0.640±0.018	0.609±0.015	0.576±0.009	0.875±0.099	0.979±0.017	0.712±0.079	0.634±0.029	0.611±0.047	0.571±0.023
	$\Re_4$	1.000±0.000	0.953±0.034	0.705±0.094	0.633±0.010	0.644±0.019	0.629±0.008	1.000±0.000	0.954±0.032	0.685±0.105	0.640±0.010	0.651±0.017	0.636±0.010
	$\Re_5$	0.983±0.020	0.933±0.021	0.829±0.065	0.632±0.011	0.644±0.006	0.632±0.012	0.984±0.020	0.934±0.020	0.834±0.061	0.639±0.006	0.650±0.009	0.641±0.014
	$\Re_6$	0.992±0.017	0.947±0.054	0.710±0.070	0.640±0.023	0.640±0.013	0.634±0.015	0.992±0.016	0.949±0.051	0.735±0.049	0.648±0.028	0.648±0.011	0.648±0.015
	$\Re_7$	1.000±0.000	0.960±0.033	0.771±0.072	0.632±0.018	0.631±0.008	0.630±0.012	1.000±0.000	0.960±0.031	0.764±0.096	0.639±0.022	0.639±0.005	0.639±0.014
$Model_1$[c]		1.000±0.000	0.900±0.000	0.800±0.029	0.710±0.012	0.742±0.004	0.719±0.019	1.000±0.000	0.889±0.000	0.779±0.028	0.705±0.011	0.735±0.004	0.716±0.016
$Model_2$[d]		0.992±0.017	0.967±0.000	0.810±0.030	0.689±0.021	0.731±0.015	0.706±0.008	0.992±0.016	0.966±0.000	0.810±0.026	0.682±0.018	0.728±0.013	0.705±0.007
HIME (proposed)		**1.000±0.000**	0.967±0.000	**0.929±0.037**	**0.757±0.009**	**0.801±0.009**	**0.783±0.010**	**1.000±0.000**	0.966±0.000	**0.931±0.032**	**0.763±0.009**	**0.798±0.007**	**0.783±0.008**

[a] HB_1–HB_6 represent the six built dataset in the order from Table 5.1, which are '1491', '644', '623', '177416', '1392', '632', respectively;
[b] $\Re_1$–$\Re_7$ are the different feature representations algorithms, representing ACC, LD, CTM, PseAAC, PsePSSM, DPCPSSM, BlockPSSM;
[c] $Model_1$ is the method from [34];[d] $Model_2$ is the method from [218]

5.5 Summary

In this chapter, Chapter. 5.1 and Chapter. 5.2 have firstly presented a short review and reported the motivation of HIME model in this chapter. Generally, a machine learning-based model with robust performance is desired to achieve for different HPI systems. Through mining the heterogeneous information of sequence data, HIME model was proposed leveraging the abundant information and the details were included in Chapter. 5.3. The horizontal ensemble procedure with heterogeneous information has greatly exerted the base learners to boost the performance in the prediction task. The performances were evaluated on six different datasets indicating HIME model outperforms the others in Chapter. 5.4.

Chapter 6

APEX2S: A TWO-LAYER MACHINE LEARNING MODEL FOR DISCOVERY OF HP-PPI

In this chapter, with the focus for host-pathogen protein-protein interactions study, developing novel machine learning techniques for learning the interactions data and making predictions is the goal. This chapter follows a brief introduction and review work reported in Chapter 6.1 and 6.2 to build the context for HP-PPIs study. Meanwhile, a general workflow to harness multi-omics data is discussed in 6.2. Given the foundation of the review, a novel two-layer machine learning model, namely APEX2S, is proposed to deal with the imbalanced issue, which has discussed in Chapter 4. The model will be discussed in 6.3. A vanilla version of APEX2S model was initialised as the effort to illustrate the effectiveness of two-layer model, which is indicated as $Model_3$ in Chapter 6.4. The advanced APEX2S model is thus compared with other twenty different traditional machine learning models and $Model_3$ with regard to various performance metrics. The results are comprehensively illustrated in Chapter 6.4, showing that APEX2S model can better learn and predict from the accumulated host-pathogen protein-protein interactions.

6.1 Introduction

The continuous development of biology technology contributes a substantial accumulation of biological interactions data. Different subareas of computational biology are

investigated, such as protein-protein binding prediction [289], protein complexes study [290], protein-protein interactions predictions [291] and sequence analysis [292], and have continuously drawn the focus of research topics towards the mechanisms of infectious diseases [149, 293]. In particular, the researches on pathogens causing the infectious diseases solicit a complete proteins and genomes interactions data collection from the hosts, pathogens and host-pathogen interactions to elucidate the infection rationale and develop effective therapy. One of the major research challenges between the practice and idealist is that, the host-pathogen interactions data are not yet complete and ready for a genome-wide level study, among which host-pathogen protein-protein interactions (HP-PPIs) data are one of the major objects [49]. Most of the host-pathogen protein-protein interactions have remained unknown, since the wet-lab experiments to determine whether the relationship should be negative or positive are deemed to be both time and cost sensitive. While positive host-pathogen protein-protein interactions data indicate that there are physical and chemical interactions between different proteins from hosts and pathogens separately, there are also a huge amount of negative HP-PPIs. Meanwhile, the number of HP-PPIs is huge given the nature of proteins number in hosts and pathogens.

As one dominant alternative, computational biology seeks to develop computational models to be cost-efficient and outcome-reliable to facilitate the study. Several studies have indicated that allocating computational resources will benefit the modelling and predicting phases, in which recently machine learning techniques are mostly involved to accelerate the generation progress of high-fidelity biological hypothesis candidates. These candidates, which represent a small amount of all interactions data, will be subsequently verified by wet-lab experiments. However, for HP-PPIs study, the research gap concerning the available omics data and computational model construction still exists.

There are two research questions of the study of HP-PPIs. The first one is related to the data, which has been addressed in previous chapters. Given the abundance of biological interactions data, the researchers are expected to delve into the data to learn from their different natures. The other one is about how to further enhance

the prediction performance of computational model by incorporating different machine learning models and feature representation algorithms. Particularly, some datasets may present the imbalance issue among the positive and negative interactions. As Chapter 3 and Chapter 4 have reported, there remains a hot topic on improving the prediction performance of HP-PPIs. Although several studies have been presented [34, 218], how to address the imbalance issue in a dedicated manner remains a problem.

In this chapter, a two-layer machine learning-based model is proposed in a more compatible manner to achieve a best performance for prediction of HP-PPIs. APEX2S model is designed as a two-layer model to alleviate the imbalanced characteristics of HP-PPIs dataset. The comparison against the traditional models and literature-based models indicates that APEX2S model achieves the best performance.

6.2 Review and Motivation of APEX2S

As for the research of infectious diseases, HP-PPIs data is considered as one of the dominant data sources in host-pathogen interactions. Particularly, the development of wet-lab techniques, such as high-throughput sequencing and interaction detection methods, has contributed to the accumulation of HP-PPIs data, which has been published across different organizations. This results in many available database resources targeting on specific scientific interests and topics.

This chapter firstly reviews a general workflow for HP-PPIs from multi-omics perspectives. An overview of the workflow is presented in Figure. 6.1.

The workflow in Figure.6.1 includes five consecutive steps, which starts from the evaluation of host-pathogen interactions (HPI) databases, to the consideration of multi-omics databases to pre-process the HPI databases for a curated HP-PPIs dataset. Given the large number of databases, both the host-pathogen interactions databases and the multi-omics databases are being extensively studied with the assistance of ‘in silico’ and ‘in vitro’ methods. In Figure.6.1, the HPI databases are firstly examined with the huge amount of accumulated interactions data, which may include intra-species

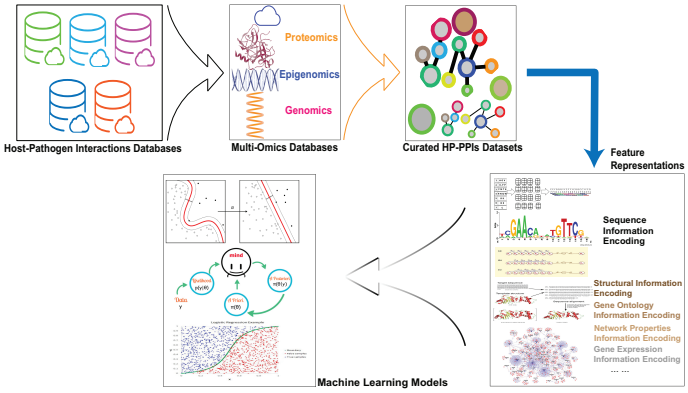


Figure 6.1: General Workflow for Host-Pathogen Protein-Protein Interactions

interactions, inter-species interactions and so on. Meanwhile, the interactions data may have sources from different methodologies, such as the wet-lab experimentally verified data and computational predictions. In this regard, an extensively review and selection of HPI databases is required to filter the untrusted and unrelated data. Thus, the trustworthy data with annotations will be output for the multi-omics databases step, which is designed as attributing the information for the interaction data. This step presents the abundance information from different omics studies, including proteomics, epigenomics and genomics, for the interactions. By implementing these two steps, a curated HP-PPIs dataset is subsequently achieved with both annotations and information for the following steps study. The details regarding the HPI databases and multi-omics databases are discussed in following sections.

The following steps involve the information encoding depending on the selection of multi-omics databases, and eventually the machine learning models are constructed to learn and predict from the HP-PPIs datasets. Technically, the mismatch between the multi-omics databases will results in missing data issue, which can hinder the generation of subsequently curated HP-PPIs datasets. To alleviate the issue, the collection of host-pathogen interactions databases will be firstly processed by the consideration of multi-omics databases. The retained data will later be utilised to generate the HP-PPIs datasets. At this stage, the datasets will only hold the proteins ID information. For feature representation, the corresponding information from the multi-omics databases are again

introduced for the encoding phase, such as the sequence information encoding scheme and the gene expression information encoding scheme.

Multi-omics study has started since 1990s when the biology technology has been continuously developed in a rapid pace. There are some relevant definitions of omics study referring to a large-scale experimental analysis giving credits to the living organisms study, which include phenomics, transcriptomics and so on. Since proteins perform a great amount of functions with organisms and host-pathogen protein-protein interactions are essential to biology functions between hosts and pathogens, HP-PPIs are known to correlate with various diseases. Particularly, HP-PPIs have associations with several omics studies, such as proteomics for proteins, epigenomics for epigenomes and genomics for genome. However, as noted in [41, 294], it is still difficult to coordinate a harmonious environment for multi-omics datasets, especially when the datasets are produced by different laboratories.

Including the early Protein Data Bank (PDB) [295], UniProt [5], and the recently constructing ENCODE [296] and so on, the database systems have allowed a better sharing for biologists, with which researchers are presented an easy access to heterogeneous datasets to build workflow for the analyses. UniProt hosts most of the proteins sequence information, which are determined by the sequencing technology. Sequence information retains the basic information of protein in a composition of hundreds or thousands of amino acid residues. By folding and binding different amino acid residues, the sequence information is developed into a unique corresponding protein structure, which is mostly archived in PDB database. For gene ontology, GO annotations are fundamentally defined by the literature research to justify three distinct aspects of biological domain including molecular function, cellular component and biological process. Entitled with the three important properties, each single protein has its own GO terms. Gene expression data is another category of process information providing a gene to regulate the synthesis of a functional gene product, which are mostly proteins [297]. They are collectively available in GEO database. Mostly, they are presented in two ways, namely microarray

and RNA-seq data. Concerning protein-protein interactions between hosts and pathogens for diseases, the interest mostly occurs for human as the host. With this regard, HPRD is included as potential omics database for HP-PPIs prediction task. It is manually curated by biologists for most human proteins. Three different properties are subsequently annotated by human protein interaction network, which are graph degree, between-ness centrality and clustering coefficient.

Although the coordination for multi-omics databases is somehow hampered and the amount of accumulated data between different databases are not level, the multi-omics databases have shown some benefits on building powerful computational models towards the analysis of infectious diseases and improving the performance of protein related prediction task [23, 218, 298–300]. Thus, the prospects of using multi-omics databases for HP-PPIs prediction task in Figure.6.1 is designed, which solicits future work from different disciplines to acquire more data.

In this chapter, we have followed the systematic review from Chapter 4 to consider the imbalance issue of the prediction task of HP-PPIs. The initial effort was conducted on a vanilla version of APEX2S model, which demonstrate that the construction of two-layer model can achieve an improved performance. The vanilla version of APEX2S model considers upsampling technique to balance the dataset in the second layer, while most hard negative and all positive are output by the first layer training model. Based on this strategy, APEX2S model takes a further consideration of numerous feature representation algorithms to improve the performance. The model design and experiments evaluation are included in Chapter 6.3 and Chapter 6.4.

6.3 The Two-Layer APEX2S Model

In this section, the proposed APEX2S model, which is a novel two-layer machine learning model based on the preliminary model [291], will be developed to enhance the prediction performance comparing with the other traditional models. The HP-PPIs workflow from Figure.6.1 is applied by considering the sequence information from proteomics

data as the primary information for HP-PPIs study. Since the missing data issue of valuable information are inevitable when incorporating different types of omics data for HP-PPIs prediction task based on machine learning models, which will subsequently cause removal and discards of the data in the following studies, the focus of this task is within different sequence feature representation algorithms and different machine learning models. Following, the sequence feature representation algorithms utilised in APEX2S model will be firstly debriefed, and the design of APEX2S model will be subsequently reported.

6.3.1 Sequence Feature Representation Algorithms

Feature representation algorithms are important for the construction of computational models. In APEX2S model, three different encoding schemes for sequence information, which are amino-acid composition method, pseudo-amino-acid composition method and evolutionary information method, are applied. Traditionally, only one of the methods is utilised to represent the sequence information, as introduced in Chapter 3 and 4.

For amino acid composition method, local descriptor algorithm [232] has been utilized to encode sequence information as the dominant feature vector, which takes regional amino acid order into accounts [291]. The protein sequence information is considered in ten regions, by which the regional amino acid order information could be retained and calculated. The method firstly divides the sequence into ten regions. Within these regions, three different descriptors, which are composition descriptor (C), distribution descriptor (D) and transition descriptor (T), are calculated. In a HP-PPIs pair, a vector of 1260 features is eventually generated by the local descriptor algorithm.

For pseudo-amino acid composition method, it was developed with the consideration of the loss of potential sequence order information when directly encoding protein sequence with amino-acid composition method.

For evolutionary information method, it processes the protein sequence against a given reference protein database. The evolutionary information is captured by constructing a

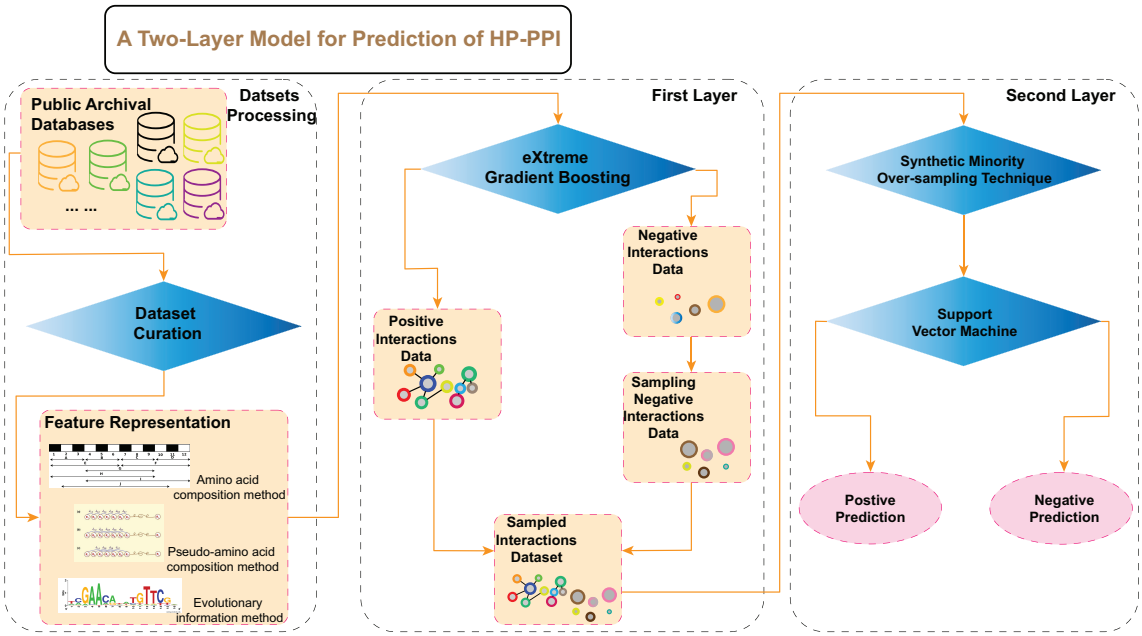


Figure 6.2: The Two-layer APEX2S Model for Host-Pathogen Protein-Protein Interactions

scoring matrix to record the probability of related amino acid types in different position. The position-specific scoring matrix (PSSM) is one direct output, which is a $T * 20$ matrix for a given protein sequence [236]. T is the length of the given protein sequence. PSSM is then processed by the developed algorithm, one of which is called Block-PSSM [249]. Block-PSSM divides sequence data into 20 equal blocks, and each block represents five percentage a sequence. As the outcome, a feature of $20 * 20 = 400$ dimension vector will be generated.

6.3.2 Proposed Two-Layer Model

The design of APEX2S will be elaborated in this section. Both the discussion of model learning stage and algorithm design will be discussed. XGBoost, which is short for eXtreme gradient boosting, is embedded as the first layer with the sampling scheme, and support vector machine (SVM) is the final classifier in the second layer with the synthetic minority over-sampling technique (SMOTE). Figure.6.2 illustrates a global diagram of the proposed two-layer model APEX2S.

eXtreme Gradient Boosting Machine

XGBoost machine is a scalable tree boosting system. Its applications in many areas have proved its ability as a powerful and efficient gradient boosting framework library [301]. Benefiting from the boosting algorithm, XGBoost has substantially extended the gradient boosting decision tree (GBDT) parallelly to achieve an efficient and accurate result.

Since XGBoost is an implementation of 'extreme gradient boosting machine' for tree ensemble models, APEX2S model applies it firstly to learn the imbalanced dataset and make classification. During the training phase, the predicted true negatives are removed from the dataset. A sub-dataset which consists of predicted positives and predicted false negatives are kept for the training of next layer. A random sampling process is then conducted on the removed negative dataset to generate a sampled negative data in the scenario that positive data in the sub-dataset is much more than negative data. This scenario will limit the performance of SMOTE and SVM of the second layer. In the algorithm, the performance of preliminary experiments indicates that when negative data is less than half of the positive data, sampling the predicted true negatives for an amount of appending the negative data to be half of the positive data in the sub-dataset, particularly will maximise the performance of SMOTE and SVM.

The final outcome of the first layer for the training phase will be the collection of the sub-dataset and the sampling data, which is to be input into the second layer. For the testing phase, the predicted negative interactions data by XGBoost is directly output as the predicted negative data, and only the rest predicted positive interactions data are further dealt within SVM.

Synthetic Minority Over-sampling Technique

In most real-world cases, the datasets are imbalanced concerning the 'irrelevant' examples and 'relevant' examples. The machine learning model performance may be fluctuated due to the imbalanced ratio between different classes, which will fail to yield desired prediction. The situation will be worse especially when the ratio becomes as high as

1:50 even 1:100 for binary classification tasks. Thus, two different types of sampling algorithms, including down-sampling the majority class [302, 303] and over-sampling the minority class [304, 305], have been proposed to address this issue.

In the proposed two-layer model APEX2S, SMOTE is applied to alleviate the imbalanced affect caused by the positive and negative HP-PPIs data. SMOTE over-samples the minority class by generating 'synthetic' examples [304]. These 'synthetic' examples present the model with more training data of the minority class by operating in 'feature space'. SMOTE has been proved as a better option than the naive over-sampling method which uses replacement data in 'data space'.

In the two-layer model APEX2S, SMOTE is utilised for the output data from first layer in the case that negative and positive data are not balanced. SVM is designed as the final classifier in the second layer to learn from the balanced dataset. APEX2S benefits from SVM's ability to map raw data into higher-dimension space, and thus the prediction performance is enhanced to finally achieve a better result.

Overall APEX2S Model

Overall, the two-layer APEX2S model is described in Algorithm 1 and the flowchart is shown in Figure.6.2.

6.4 Experimental Evaluation and Discussion

The design and selection of HP-PPIs datasets and the performance metrics is firstly reported. Then, the performance of different models will be reported.

6.4.1 Experiment Evaluation

Concerning the imbalanced ratio of HP-PPIs dataset [156], the negative HP-PPIs data are as important as the positive ones to build the HP-PPIs dataset. As mentioned earlier in Chapter 3 and 4, a thorough investigation has been conducted for 11 public archival databases to collect the positive HP-PPIs data. A shared meaningful character is identified

Algorithm 2 Training APEX2S Model for Prediction of HP-PPIs

Require: Dataset $M = \{v_i, o_i\}$,

- v_i is the vector of input, a concatenated representation of protein-protein pair with the corresponding ***A***mino-acid-composition, ***P***seudo-amino-acid-composition and ***E***volutionary-information features;
- $o_i \in \{+1, -1\}$ represents positive and negative interactions;

Ensure: Output of the APEX2S model, p_i;

Initializing e***X***treme gradient boosting machine (XGBoost);
Inputting the training dataset M into XGBoost model;
Saving the first layer training model X;
Giving the first layer prediction results $M_{XGBoost}$ by X giving M;
Obtaining an interactions dataset N by comparing $M_{XGBoost}$ with M, in which the data except predicted true negatives are kept;
Saving the set O, which is the true negatives predicted by X;
Defining O_1=Ø;
if N(Neg)<N(Pos)

- N(Neg) is the negative interactions data in N
- N(Pos) is the negative interactions data in N

then
    $\lambda = N(Pos)/2 - N(Neg)$;
    if $\lambda > 0$ **then**
        randomly sampling λ negative interactions data from O, as O_1;
    end if
end if
Obtaining an interactions dataset as $Q = N + O_1$;
if $Q(Pos) < Q(Neg) or Q(Pos) > Q(Neg)$

- Q(Pos) is the positive interactions data in Q
- Q(Neg) is the negative interactions data in Q

then
    Balancing the dataset Q by the ***S***yenthetic minority over-sampling technique (SMOTE);
    Obtaining a sub-sampling balanced interactions dataset $\bar{M}$;
end if
Training ***S***upport vector machine (SVM) giving $\bar{M}$;
Saving the second layer training model S;
Indexing X and S for performance evaluation.

among the databases, which is the resources of the HP-PPIs data are highly trustworthy. All of the data are by verification of literature or domain experts. The collectded data are then carefully processed to remove the redundant HP-PPIs data and the highly homologous proteins. The redundancy of the HP-PPIs datasets was reduced by this step, so as the bias in the training models was reduced. Once the positive HP-PPIs data was collected, the different ratios on positive HP-PPIs data are applied to build the negative interaction data, which is of 1:25, 1:50 and 1:100 following the procedure from [23, 156].

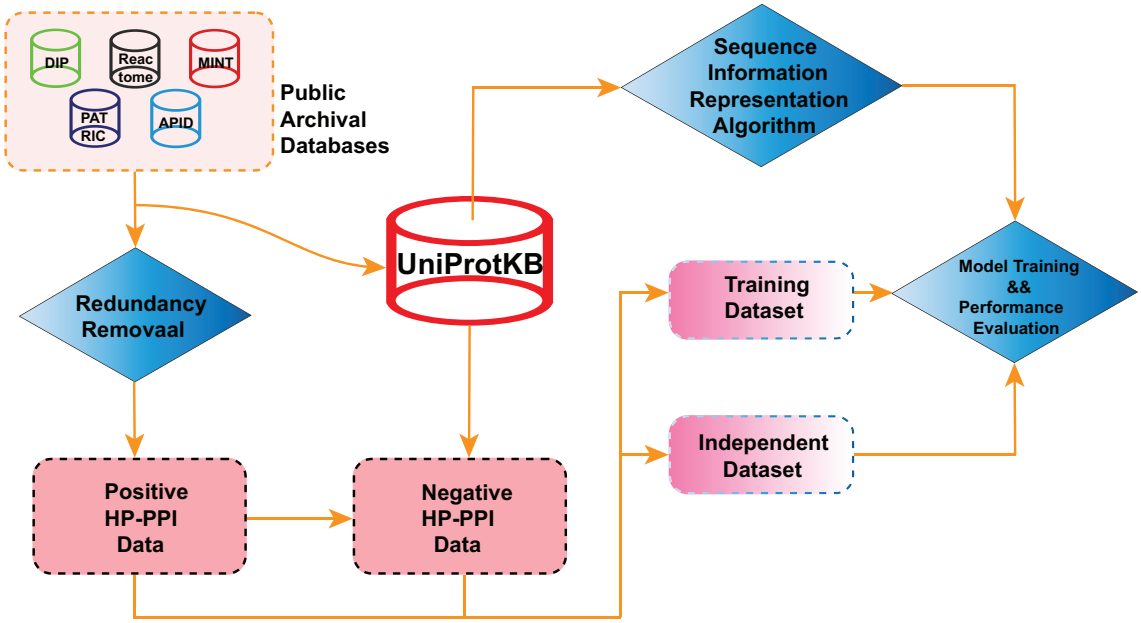


Figure 6.3: The Dataset Curation Protocol for Host-Pathogen Protein-Protein Interactions

Both the training dataset and the independent test dataset are required for evaluation and comparison of computational models. Figure.6.3 briefly demonstrate the diagram of the applied curation protocol from previous chapters. One-fifth HP-PPIs data from positive and negative HP-PPIs data are randomly selected to build the independent test datasets. These datasets are held till the model is trained and are unseen until the model makes all the predictions. The rest data will be the training dataset. Combining one independent test dataset and one training dataset is called a complete curated HP-PPIs dataset. In this way, the HP-PPIs datasets are built five times to avoid the bias causing by random sampling method. All five curated HP-PPIs datasets are used to train and test the models.

Table 6.1: The Statistics of Training Datasets

Pathogen Name	ID	Positive Interactions Number	Ratio 1:25		Ratio 1:50		Ratio 1:100	
			Training	Independent Testing	Training	Independent Testing	Training	Independent Testing
Shigella paradysen-teriae	623	105	2184	546	4284	1071	8484	2121

The performance results are collected with the mean and deviation results with regard to different measurement metrics.

6.4.2 Datasets

To verify both the HP-PPIs workflow from Figure.6.1 and proposed APEX2S model from Figure.6.2 to be applicable for the prediction task, the experimental HP-PPIs dataset is selected to consist of the protein interactions between homo sapiens (taxonomy ID 9606) as host species and Shigella paradysenteriae as the bacterium pathogen (taxonomy ID 623). Table.6.1 shows the final statistics of the curated datasets. The datasets has 118 pairs of the positive HP-PPIs data, and a total number of 2184, 4284, 8484 for different ratios of 1:25, 1:50 and 1:100 for negative HP-PPIs data. Among the interactions between homo sapiens and Shigella paradysenteriae, there are 75 different proteins from homo sapiens and 60 different proteins from the pathogen. Given its relative high protein nodes number for both human and pathogens, the dataset is considered as the examplar dataset for the evaluation of the proposed APEX2S model to study the impact of different high skewed ratios, which are 1:25, 1:50 and 1:100.

6.4.3 Performance Metrics

In this study, numerous performance metrics have been included. The accuracy is usually not accurate at comparing models in a full scale for an imbalanced dataset. Especially when the ratio is 1:100, the accuracy value would still be very high whereas the difference between different models would be negligible in the worst case when giving all predictions to be majority class. Thus, other measurement metrics, such as precision,

recall, F1-score and Matthew's correlation coefficient (MCC) score, are included. The deinition of the metrics can be reffered to Equa. 4.22.

6.4.4 Results and Discussion

In the experiments, the results are collected on the curated HP-PPIs dataset, which is the protein-protein interaction between human and Shigella paradysenteriae pathogen. By conducting each experiments for a five-independent test, both standard and deviation results are recorded with regard to accuracy, precision, recall, F1 and MCC. For the execution environment of the experiment, the working system is built with 64GB memory and a core CPU of Intel i7-6700K. The working operating system is Ubuntu 16.04, and all the implementations were written in Python, partially with the support of open source package 'scikit-learn' [256].

Table.6.2 shows the accuracy, precision and recall results of numerous models, including the traditional machine learning models with different feature representation algorithms, three different models from literature [34, 218, 291] and the proposed APEX2S model. For the symbol of 'A', 'P' and 'E' in Table.6.2, they represent the amino acid composition method, the pseudo-amino acid composition method and the evolutionary information method.

It is clearly to observe that, APEX2S model has achieved the best performance on all the different datasets. However, the improvement on accuracy may not be outstanding due to the high imbalanced ratio of the HP-PPIs dataset. For the dataset with the ratio of 1:25, APEX2S model achieves a result of 0.982051 ± 0.004546 while the following results in the performance ladder are 0.981685 ± 0.001638 by SVM and 0.980586 ± 0.002484 by Model_3. Also, for the dataset with the ratio of 1:100, the result of APEX2S model is 0.993022 ± 0.000625, however the results of the second and third places are 0.992834 ± 0.000693 and 0.992551 ± 0.000189 achieved by Model_3 and RF respectively.

Thus, the evaluation has further compared the precision, recall and F1-score results. Table.6.3 contains the results of precision and recall, and Table.6.4 shows the results of

F1-score and MCC. Both the best and second best results are illustrated in the bond and italic font. All the results are presented in Table.6.2, Table.6.4 and Table.6.3with the mean values and deviation values for the five-independent tests experiments.

For F1-score and MCC values, the closer the value is to 1.0 indicates the better the trained model is. In TABLE 6.4, APEX2S model has shown a best performance of F1 on datasets with the ratio of 1:25 and 1:100, but only achieved a second best performance on dataset with the ratio of 1:50. The best model for dataset with the ratio of 1:50 is presented by the logistic regression (LR) model. For dataset with the ratio of 1:25, the second best model is achieved by SVM, while the third best model is Model_3 [291]. It shows a comparative performance between APEX2S model and SVM model for dataset with the ratio 1:50. However, the proposed APEX2S model achieves a much better performance for the other datasets with the ratio 1:25 and 1:100.

The best performance regarding MCC value for datasets with the ratio 1:25, 1:50 and 1:100 is 0.735440 ± 0.063093, 0.469571 ± 0.036759 and 0.543188 ± 0.051758, respectively. The results of MCC values also indicate that, APEX2S model can achieve a best performance on all datasets.

For the execution times, the average costing times are 21.0856, 47.9768 and 100.1876 seconds for APEX2S model accordingly. The comparison between different models is illustrated in Figure.6.4. In Figure.6.4, Model_1 - Model_3 are the models from [218], [34] and [291] respectively. Since APEX2S model consists of two layers and has included SMOTE technique to enhance the performance, its consumption of time has also been the most. However, as it can be observed from Figure.6.4, the time cost for models training are mostly around minutes.

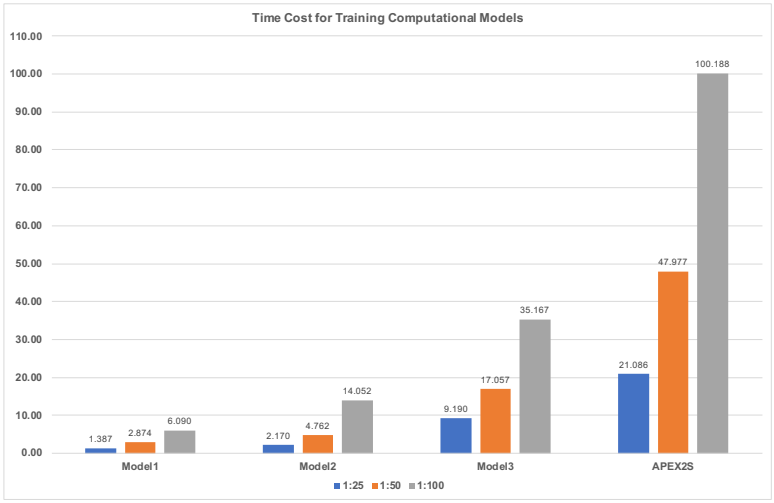


Figure 6.4: The Time Cost for Computational Model Training

Table 6.2: Results of Accuracy for Models

Model		Accuracy		
		1:25	1:50	1:100
RF	***A***	0.970330±0.001371	0.981139±0.000373	0.991985±0.000298
	P	0.978022±0.003065	0.981326±0.001446	0.992456±0.000298
	E	0.978755±0.001868	0.981699±0.000747	0.992551±0.000189
SVM	***A***	0.979121±0.001465	0.980952±0.000457	0.992268±0.000971
	P	0.975458±0.005507	0.980392±0.000000	0.990948±0.000550
	E	***0.981685±0.001638***	***0.981886±0.000457***	0.991513±0.000000
LR	***A***	0.971795±0.002741	0.980766±0.000747	0.991702±0.000377
	P	0.958608±0.004719	0.977218±0.001267	0.988967±0.000640
	E	0.969963±0.003589	0.979085±0.002547	0.989062±0.000754
Naïve Bayes	***A***	0.677289±0.015341	0.694304±0.017995	0.680717±0.009091
	P	0.820147±0.020572	0.766760±0.011319	0.836115±0.003708
	E	0.797436±0.021252	0.823903±0.012225	0.798680±0.007286
GBM	***A***	0.971429±0.004719	0.978711±0.001811	0.988213±0.002109
	P	0.975458±0.004719	0.980205±0.002241	0.991042±0.000596
	E	0.979121±0.003771	0.980952±0.001923	0.986327±0.000516
DT	***A***	0.952381±0.007141	0.971242±0.001712	0.988685±0.001606
	P	0.961905±0.005102	0.979645±0.001239	0.988967±0.000640
	E	0.961172±0.004546	0.968067±0.004023	0.985196±0.001418
$Model_1$[218]		0.975092±0.089700	0.981326±0.000000	0.991985±0.000000
$Model_2$[34]		0.971795±0.000897	0.981326±0.000000	0.992362±0.000462
$Model_3$[291]		0.980586±0.002484	0.981513±0.001089	***0.992834±0.000693***
APEX2S		***0.982051±0.004546***	***0.984500±0.000952***	***0.993022±0.000625***

[a] ***A***, ***P*** and ***E*** represent the three feature representation algorithms of amino acid composition, pseudo-amino acid composition and evolutionary information methods.

[b] RF, SVM, LR, GBM and DT are acronyms for random forests, support vector machine, logistic regression, gradient boosting machine and decision tree, respectively;

[c] $Model_1$ is the method from [218],[d] $Model_2$ is the method from [34] and $Model_3$ is the method from [291].

Table 6.3: Results of Precision and Recall for Models

Model		Precision			Recall		
		1:25	1:50	1:100	1:25	1:50	1:100
RF	***A***	0.695922±0.045611	0.900000±0.200000	0.960000±0.080000	0.419048±0.087287	0.047619±0.000000	0.200000±0.019048
	P	0.907692±0.113053	0.850000±0.300000	0.966667±0.066667	0.485714±0.019048	0.076190±0.023328	0.247619±0.019048
	E	0.901732±0.057326	0.783333±0.194365	1.000000±0.000000	0.504762±0.038095	0.104762±0.019048	0.247619±0.019048
SVM	***A***	0.907143±0.068760	0.800000±0.244949	0.877778±0.173561	0.514286±0.019048	0.047619±0.000000	0.266667±0.038095
	P	0.753875±0.119560	0.000000±0.000000	0.612857±0.073426	0.552381±0.048562	0.000000±0.000000	0.247619±0.063174
	E	0.939377±0.054579	1.000000±0.000000	1.000000±0.000000	0.561905±0.019048	0.076190±0.023328	0.142857±0.000000
LR	***A***	0.805000±0.074833	0.766667±0.290593	0.860000±0.127192	0.352381±0.048562	0.047619±0.000000	0.200000±0.019048
	P	0.459524±0.131406	0.276032±0.066228	0.332143±0.100661	0.238095±0.000000	0.104762±0.046657	0.133333±0.063174
	E	0.594326±0.047306	0.466190±0.068087	0.438730±0.042999	0.714286±0.030117	0.400000±0.064594	0.361905±0.064594
Naïve Bayes	***A***	0.094328±0.004231	0.044737±0.002568	0.026024±0.000730	0.857143±0.000000	0.714286±0.000000	0.857143±0.000000
	P	0.080587±0.004092	0.045398±0.002598	0.026684±0.001396	0.352381±0.048562	0.542857±0.023328	0.438095±0.019048
	E	0.145761±0.015513	0.092781±0.005800	0.044068±0.001609	0.866667±0.019048	0.904762±0.000000	0.933333±0.023328
GBM	***A***	0.750999±0.153070	0.406926±0.090796	0.406117±0.208510	0.409524±0.023328	0.152381±0.019048	0.200000±0.035635
	P	0.725199±0.097395	0.521111±0.143776	0.609207±0.085156	0.600000±0.048562	0.200000±0.035635	0.295238±0.019048
	E	0.813650±0.073419	0.533492±0.143428	0.277746±0.027269	0.600000±0.088320	0.180952±0.069985	0.238095±0.030117
DT	***A***	0.397069±0.083398	0.213095±0.030152	0.383333±0.178263	0.409524±0.038095	0.171429±0.023328	0.133333±0.035635
	P	0.508844±0.049335	0.362338±0.204123	0.150000±0.133333	0.609524±0.092337	0.114286±0.088320	0.028571±0.023328
	E	0.514949±0.141427	0.215148±0.062349	0.279321±0.031932	0.209524±0.048562	0.228571±0.055533	0.304762±0.048562
$Model_1$[218]		1.000000±0.000000	1.000000±0.000000	1.000000±0.000000	0.352381±0.023328	0.047619±0.000000	0.190476±0.000000
$Model_2$[34]		0.942857±0.069985	1.000000±0.000000	0.847619±0.106053	0.285714±0.000000	0.047619±0.000000	0.285714±0.000000
$Model_3$[291]		0.905505±0.090918	0.643333±0.124544	0.908333±0.130171	0.561905±0.035635	0.133333±0.035635	0.314286±0.023328
APEX2S		0.869937±0.105470	0.821429±0.111677	0.950000±0.100000	0.638095±0.038095	0.276190±0.019048	0.314286±0.038095

[a] ***A***, ***P*** and ***E*** represent the three feature representation algorithms of amino acid composition, pseudo-amino acid composition and evolutionary information methods.
[b] RF, SVM, LR, GBM and DT are acronyms for random forests, support vector machine, logistic regression, gradient boosting machine and decision tree, respectively;
[c] $Model_1$ is the method from [218],[d] $Model_2$ is the method from [34] and $Model_3$ is the method from [291].

Table 6.4: Results of F1 Score and MCC for Models

Model		F1-score			MCC		
		1:25	1:50	1:100	1:25	1:50	1:100
RF	***A***	0.515472±0.057686	0.090119±0.001581	0.330462±0.027493	0.522309±0.044992	0.202909±0.026521	0.435610±0.031072
	P	0.630859±0.038081	0.138530±0.043473	0.393732±0.025949	0.653953±0.050929	0.247093±0.080549	0.486779±0.028365
	E	0.645855±0.033990	0.182899±0.029017	0.396581±0.023932	0.665099±0.034396	0.278747±0.041975	0.495406±0.018609
SVM	***A***	0.654747±0.015513	0.089328±0.001936	0.406553±0.058855	0.673402±0.021537	0.189648±0.032481	0.479377±0.074487
	P	0.635962±0.071484	0.000000±0.000000	0.346947±0.067800	0.632491±0.079112	0.000000±0.000000	0.381936±0.057596
	E	***0.702562±0.023684***	0.140711±0.040663	0.250000±0.000000	***0.718527±0.027769***	0.269979±0.043936	0.376355±0.000000
LR	***A***	0.488988±0.056970	0.088603±0.003047	0.322872±0.025335	0.520864±0.056991	0.183660±0.040926	0.410708±0.034654
	P	0.308802±0.026074	0.149518±0.058623	0.188561±0.081085	0.308530±0.048173	0.158566±0.056588	0.204576±0.081694
	E	0.647466±0.029745	***0.428223±0.058702***	0.393653±0.046146	0.635593±0.029990	***0.420129±0.059519***	0.391535±0.044933
Naïve Bayes	***A***	0.169922±0.006850	0.084184±0.004546	0.050514±0.001375	0.212762±0.008699	0.122152±0.007389	0.113248±0.002763
	P	0.130786±0.006891	0.083776±0.004604	0.050304±0.002596	0.098283±0.011443	0.102841±0.008675	0.074663±0.005712
	E	0.249316±0.023371	0.168247±0.009552	0.084157±0.002969	0.303128±0.026605	0.256394±0.010482	0.177718±0.005867
GBM	***A***	0.527137±0.054758	0.219974±0.030379	0.255816±0.052325	0.540288±0.073108	0.238834±0.039974	0.271733±0.075399
	P	0.654161±0.057202	0.284501±0.047494	0.395112±0.017083	0.645980±0.062210	0.311523±0.060472	0.418660±0.026424
	E	0.685798±0.066382	0.268066±0.093979	0.255978±0.027031	0.686201±0.063461	0.301726±0.098554	0.250090±0.027065
DT	***A***	0.401192±0.057632	0.189447±0.024179	0.190145±0.046456	0.377599±0.061403	0.176362±0.024848	0.215932±0.066885
	P	0.550339±0.050891	0.165606±0.118433	0.047481±0.038882	0.535429±0.053898	0.188720±0.122450	0.061233±0.054063
	E	0.292852±0.064972	0.220097±0.056711	0.288767±0.029965	0.308496±0.073374	0.204790±0.057824	0.282994±0.030010
$Model_1$[218]		0.520690±0.025340	0.090909±0.000000	0.320000±0.000000	0.585766±0.019551	0.216169±0.000000	0.434680±0.000000
$Model_2$[34]		0.438095±0.007776	0.090909±0.000000	0.426032±0.014395	0.510286±0.020527	0.216169±0.000000	0.488573±0.031846
$Model_3$[291]		0.690496±0.032247	0.219316±0.053920	***0.465441±0.038502***	0.703311±0.038889	0.285516±0.061405	***0.530828±0.052146***
APEX2S		***0.734050±0.056882***	***0.411436±0.025322***	***0.470881±0.049072***	***0.735440±0.063093***	***0.469571±0.036759***	***0.543188±0.051758***

[a] ***A***, ***P*** and ***E*** represent the three feature representation algorithms of amino acid composition, pseudo-amino acid composition and evolutionary information methods.
[b] RF, SVM, LR, GBM and DT are acronyms for random forests, support vector machine, logistic regression, gradient boosting machine and decision tree, respectively;
[c] $Model_1$ is the method from [218],[d] $Model_2$ is the method from [34] and $Model_3$ is the method from [291].

6.5 Summary

Machine learning and data analytics techniques can be a good leverage to boost the study of biological interactions data, especially when more and more data have been available. In this chapter, the HP-PPIs prediction problem is studied and a detailed investigation concerning the multi-omics data for HP-PPIs is conducted firstly in Chapter 6.1. Presented by the abundant multi-omics data, a comprehensive and practical workflow is subsequently designed in Chapter 6.2, which has elaborated the usage of machine learning techniques in a preliminary stage. More importantly, an improved two-layer model APEX2S for the prediction task of HP-PPIs is presented in Chapter 6.3. In Chapter 6.4, a practice of the model in real case concerning the protein-protein interactions between human and Shigella infections pathogen is reported to evaluate the performance of various machine learning models, which include the traditional machine learning models and our two-layer model. The comparison against traditional models and literature-based models has indicated the better prediction ability and higher efficiency of APEX2S model

Chapter 7

TOWARDS A MORE EFFECTIVE BIDIRECTIONAL LSTM-BASED LEARNING MODEL FOR HUMAN-PATHOGEN PROTEIN-PROTEIN INTERACTIONS

In this chapter, the deep learning model will be examined for the prediction task of HP-PPIs. Particularly, a bidirectional LSTM-based model will be presented for the prediction task, which demonstrates a more effective performance in comparison with the others. The imbalance issue is considered as the main research task in this chapter, for which we have include the imbalanced ratios of 1:25, 1:50 and 1:100. Particularly, we have designed a novel feature representation algorithm for protein information to be input in the LSTM model. For the Bi-LSTM model, a novel loss function is introduced to enhance the performance of the deep learning model. To evaluate the performance, we have conducted the comparison with numerous traditional machine learning models as well as the methods from literature. In this chapter, we will start with a brief introduction and related work in Chapter 7.1 and 7.2. The details of the Bi-LSTM-based model will be presented in Chapter 7.3. Chapter 7.4 will report the performance evaluation while Chapter 7.5 summarise the chapter.

7.1 Introduction

Monitoring and curing the infectious diseases for human are still prevalent and intractable problems, while there have been substantial researches focusing on the understanding of infectious mechanisms and the development of novel therapeutic solutions. This solicits great efforts in revealing the biological interactions between human and different pathogens [147, 306, 307]. However, research on identification of interactions is still in its early stage. Some published data may focus on particular human-pathogen interactions (HPI) system, for example between human and HIV virus, which may be of special interest to a small group of researchers. Meanwhile, the identification of interactions takes huge amount of experimental resources and consumes lots of time. This has significantly limited the progress in studying different HPI systems.

Although several literature reviews have been published by introducing the machine learning-based methods and some applications in the HPI domain, little research on empirical evaluations of the performance of HBI predictions based on machine learning models has been ever conducted [32, 225], and no work focusing on the prediction of human-bacterium interactions has been reported. Meanwhile, most studies of PPI predictions have been conducted based on a hypothesis on evaluating the predictor with a balanced and small dataset, in which the numbers of positive and negative PPIs are the same.

As a cost-effective approach, in Chapter. 3 and 4, the computational models for analysis and predictions of HPI systems have been broadly investigated to learn the HP-PPIs data in a comprehensive manner. Moreover, the prediction performance has shown distinct fluctuation on HP-PPI datasets using different machine learning models. The evaluations of various traditional machine learning methods and models found in the literature review have revealed that, current techniques could not render a robust performance and could not generalise well for the HP-PPIs dataset.

Thus, how to design a novel robust and effective model for the prediction task remains

challenging. In this chapter, a bidirectional long short-term memory-based model is proposed, jointly learning with the designed multi-channel feature representation algorithm, tree-based feature selection algorithm and synthetic minority over-sampling technique (SMOTE), for the prediction of HP-PPIs dataset. The proposed model achieves a more robust and effective performance on the HP-PPIs datasets of three different HPI systems, which demonstrates a superior performance over the others. The details of design will be discussed in Section 3. The proposed model indicates a promising research direction of studying big HP-PPIs dataset with deep learning model.

7.2 Review and Motivation

There have been substantial research interests in applying machine learning methods for prediction of protein interactions [156, 218, 225, 267, 278, 283]. A similarity between all these works was to have successfully applied machine learning methods in a given positive protein interactions data, whilst their work focused on a balanced protein interactions dataset by building negative protein interactions data with a same number of the positives.

For the prediction of HP-PPIs, a wide coverage of host-pathogen interactions can be found in [225], [151] and [277], which includes the prediction as well as analysis, while research on computational prediction of host-pathogen interactions was discussed in [32] and [162]. Since these reviews aimed at describing the progress in prediction of host-pathogen interactions without anchors of naming pathogens, they have collectively listed potential computational methods. The computational methods include a homology-based approach, a structure-based approach, and a motif interaction-based approach and machine learning-based approach. Furthermore, no systematic evaluation with sufficient details has been implemented and reported in these reviews.

Following the systematic review from Chapter 3, 11 databases are chosen to curate the dataset of different HPI systems. More details could be referred to Chapter 3 and 4. To better address the imbalance issue of the prediction task of HP-PPIs, we at first time introduce deep learning model for training. Particularly, we consider the bidirectional

LSTM model. The reason is that, LSTM model has illustrated a better capability for sequence-based task, such as natural language processing and protein structure prediction. Based on Bi-LSTM model, a novel feature representation algorithm is demanded to translate the sequence information as three-dimensional data. Thus, in this chapter, a dedicated Bi-LSTM-based model is proposed and has achieved a superior performance in comparison with the other models.

7.3 Proposed Bi-LSTM-based Model

7.3.1 Our Model

Fig. 7.4 illustrates the Bi-LSTM-based model. It includes five layers to learn from the raw data, which are feature representation algorithm layer, SMOTE layer, a multi-channel feature representation layer, the Bi-LSTM layer and a full-connected layer. In this chapter, the details of two important layers, the Bi-LSTM layer and the multi-channel feature representation layer will be discussed.

7.3.2 Bidirectional LSTM

The Bidirectional LSTM (Bi-LSTM) is the critical component of the model, which is a variant deep learning model of LSTM proposed by [308, 309]. LSTM model and its variant version Bi-LSTM have demonstrated superior performance in domains such as natural language processing, transportation and action recognition [310, 311].

Figure. 7.1 demonstrates the block of vanilla LSTM model. Figure. 7.2 is the vanilla LSTM model. Briefly, the block of vanilla LSTM model is built with four gates and the mainline on top of the block connects the state C_{t-1} and C_t. The four gates are forget gate, two input gates and one output gate.

For forget gate, the main output is f_t, which is defined by Equa. 7.1.

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \tag{7.1}$$

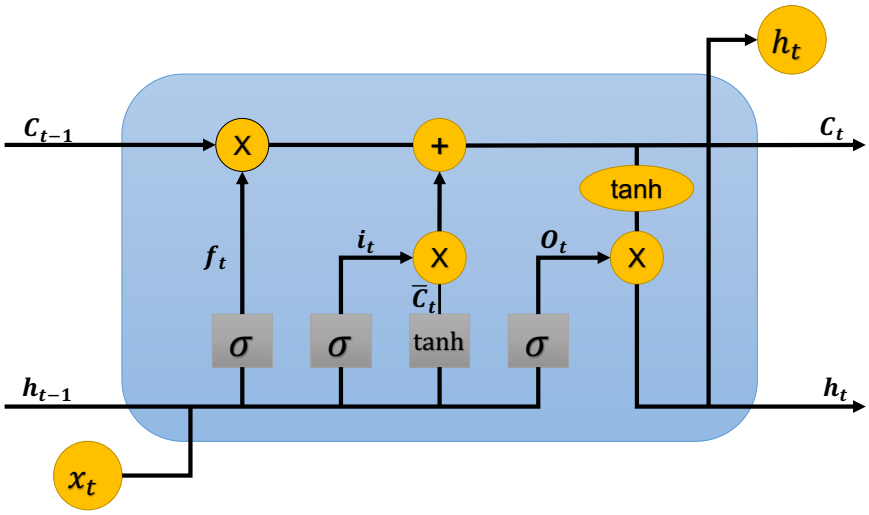


Figure 7.1: The Block of Long Short-Term Memory-based Model

For the input gates, the definition is given in Equa. 7.2.

$$\begin{aligned} i_t &= \sigma(W_i * [h_{t-1}, x_t] + b_i) \\ \overline{C_t} &= tanh(W_c * [h_{t-1}, x_t] + b_c) \\ C_t &= f_t * C_{t-1} + i_t * \overline{C_t} \end{aligned} \tag{7.2}$$

For the output gate, the definition is given in Equa. 7.3.

$$\begin{aligned} O_t &= \sigma(W_o * [h_{t-1}, x_t] + b_o) \\ h_t &= O_t * tanh(C_t) \end{aligned} \tag{7.3}$$

Several variations of LSTM block connect the internal signals by different mechanisms, including the peephole connections LSTM, coupled LSTM and gated recurrent unit (GRU). In Bi-LSTM model, two layers, namely forward and backward layers, are designed to converge into a single layer. The details can be found in

However, the traditional Bi-LSTM model explicitly suffers from the conventional vanishing gradient problem for the prediction of HB-PPI data. In most cases, the cross entropy loss is applied as the cost function for binary classification, as following Equa. 7.4.

$$CE(p, y) = \begin{cases} -log(p) & if\ y = 1 \\ -log(1-p) & otherwise \end{cases} \tag{7.4}$$

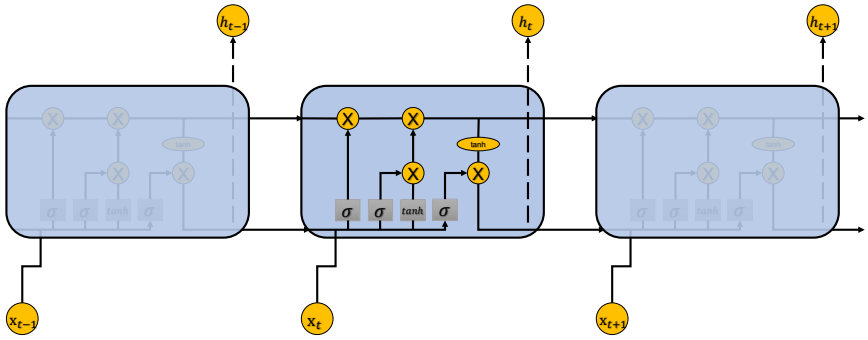


Figure 7.2: The Long Short-Term Memory-based Model

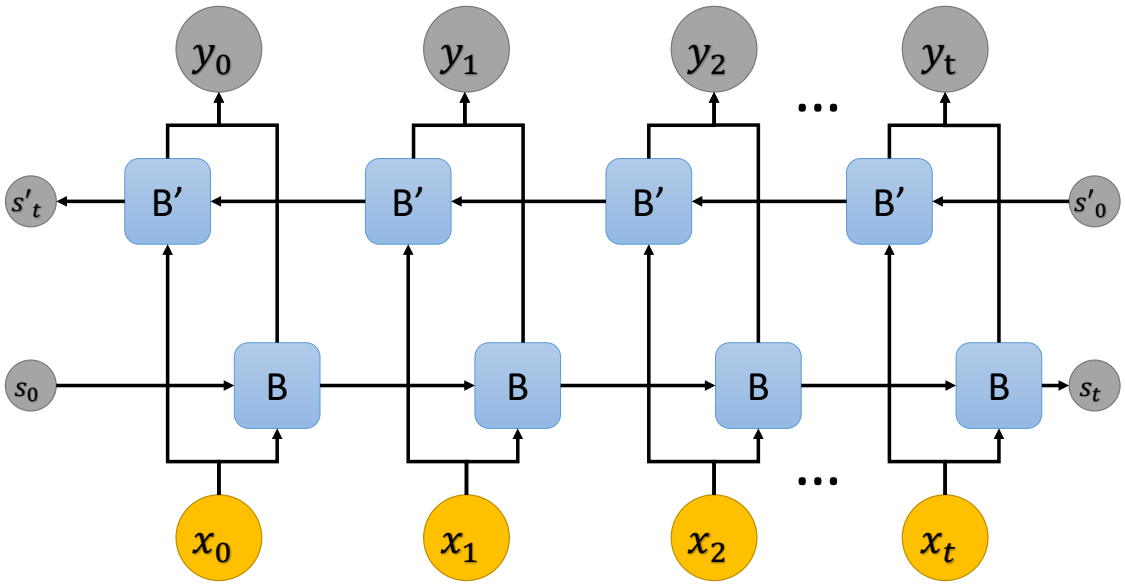


Figure 7.3: The Bidirectional Long Short-Term Memory-based Model

In the definition, $y \in +1, -1$ which is the ground-truth class. $p \in [0,1]$ is the estimated probability for the class while $y = 1$. For notational efficiency, p_t is defined as:

$$p_t = \begin{cases} p & if\ y = 1 \\ 1-p & otherwise \end{cases} \tag{7.5}$$

Thus, Equa. 7.4 can be rewritten as Equa. 7.6.

$$CE(p,y) = CE(p_t) = -log(p_t) \tag{7.6}$$

In the Bi-LSTM-based model, the focal loss function as the cost function Δ to resolve gradient vanishing problem, which is defined in Equq. 7.7 [312]. α_t and γ are the parameters.

$$\Delta(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{7.7}$$

Besides the machine learning model, the feature representation algorithm is another critical factor that will contribute to the performance improvement. Next, the details of several utilised sequence feature representation algorithm will be firstly debriefed, and the design of the multi-channel feature will be introduced for the machine learning model.

7.3.3 Interpreting the Sequence Information

Since utilizing protein sequence information has become a research trend due to its availability of abundant information, it also solicits novel feature representation algorithms to the ongoing protein researches to improve the prediction performance [34, 226, 268]. In this work, sequence information is selected as the primary information. It is anticipated that the study can be potentially extended to other related research topics. Thus, mapping the sequence information according to the selected feature representation algorithms is the first step.

Because every different protein possesses different length of amino acid combinations, it will be difficult to directly input the sequence information into the machine learning

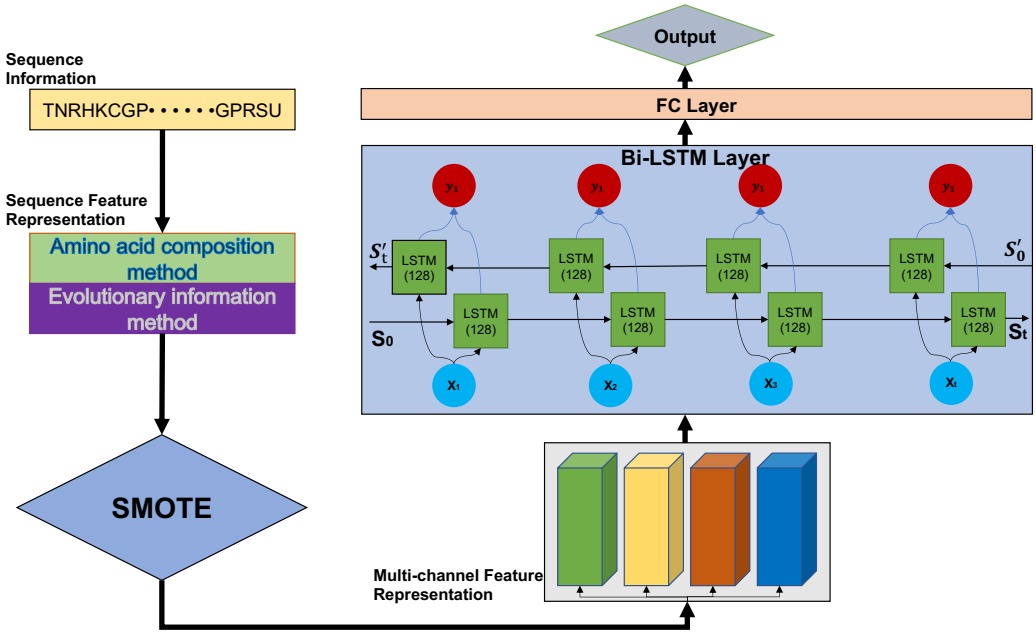


Figure 7.4: The Bi-LSTM-based Model for HP-PPIs

methods. This raises a great interest for us to develop an efficient and powerful algorithm to retain the identity of proteins. Two different categories are included in this work, namely, amino acid composition methods and evolutionary information methods.

Amino acid composition methods consider the feature representation according to the amino acid combination of a given protein sequence information in different ways, such as their grouping based on different physicochemical characteristics and their order of sequence information. This results in two different popular algorithms, namely the conjoint triad method [283] and auto covariance algorithm [33]. The conjoint triad method divides twenty different types of amino acid into seven groups according to their different physicochemical characteristics. The auto covariance algorithm calculates the auto covariance relationship using the order of amino acids in the sequence information. The basic idea is to consider the difference between proteins according to their frequency in amino acid combinations, for example three adjacent amino acids (3-mer). The combinations set will be $\{(1,1,1),(1,2,1),(1,3,1),...,(1,7,1),...,(1,7,7),...,(7,7,7)\}$. $1-7$ represent the seven groups, which results in 343 different combinations for 3-mer features. It also utilizes seven different physicochemical properties to represent the amino

acids, which results in a $N * 7$ matrix, where N is the length of the sequence.

Evolutionary information methods refer to a process of protein alignment against a reference protein sequence database, which produces a position-specific scoring matrix (PSSM) to indicate the appearing probability of each amino acid types for corresponding position. PSSM is a T*20 matrix for a protein sequence by PSI-BLAST. T denotes the length of the protein sequence. In our work, we apply two different methods, which are Pseudo Position-Specific Score Matrix (Pse-PSSM) [239] and Block-PSSM [249]. Pse-PSSM is a 40-dimensional vector, which generates a direct and joint amino acids relationship from the original PSSM. Block-PSSM firstly divides PSSM profile and protein sequence into 20 equal blocks. For each block, a 20-dimensional vector is calculated according to amino acid information. Since each block generates a 20-dimensional vector, Block-PSSM will produce a 20*20 = 400-dimensional vector feature for the whole protein sequence.

Based on the selected feature representation algorithms, a novel three-dimension tensor data as the feature representation algorithm, which is a multi-channel feature in this study. The design of the multi-channel feature benefits from various sequence-based feature representation algorithms. The tree-based feature selection algorithm is employed at first to reduce the abundant features. Once the features are processed, the data will be learnt by SMOTE technique to ease the imbalanced ratio. The output of the SMOTE model will be subsequently stacked horizontally to build the multi-channel feature data, which is then input to Bi-LSTM.

Totally, the proposed model is designed with the consideration of the distinct feature of protein sequence information and the imbalanced issue of the HB-PPI datasets as illustrating in Fig. 7.4. In next section, a complete evaluation performance as well as the proposed model results with regard to F1-score will be presented, which is a suitable measurement for the evaluation in this research.

Table 7.1: The Statistics of Datasets

Taxonomy ID[a]	Positive Interactions Number	Ratio 1:25		Ratio 1:50		Ratio 1:100	
		Training	Independent Testing	Training	Independent Testing	Training	Independent Testing
1491	57	1185	297	2325	582	4605	1151
177416	1207	25105	6277	49245	12312	97525	24382
1392	2810	58448	14612	114648	28662	227048	56762

[a] '1491' represents Clostridium botulinum, '177416' is Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4), and '1392' is Bacillus anthracis bacterium

7.4 Evaluation and Discussion

In this work, the study have dedicated to focus on three different HPI systems, for the reason of their sufficiently available protein information to constitute both small and big dataset for the subsequent evaluation and comparison with the proposed model. The others could be used for further research or repeated verification, but are not within the scope of this chapter.

7.4.1 The HB-PPI Dataset

The details of the curated HB-PPI dataset are shown in Table. 7.1. The taxonomy IDs are listed as the specific bacterium pathogen species selected after data pre-processing. They correspond to three different bacterium pathogens actively interacting with human host. The selected bacterium pathogens are Clostridium botulinum (taxonomy ID: 1491), Francisella tularensis subsp. tularensis (strain SCHU S4 / Schu 4) (taxonomy ID: 177416) and Yersinia pseudotuberculosis subsp. pestis (taxonomy ID: 632). To alleviate the impact of randomness in sampling, this process is repeated for five times, which resulted in a five-fold independent tests for the evaluation.

7.4.2 Machine Learning based Methods

It is crucial to select feasible machine learning methods to perform the HPI prediction task. In this paper, we evaluate several popular machine learning models, including support vector machine (SVM), random forests (RF), logistic regression (LR), naïve

Bayes model (GNB), decision tree (DT) and gradient boosting machine (GBM). These machine learning models are still more predominant than deep learning methods in protein interaction related studies, because they usually require less data and have a simpler architecture, yet achieving a reasonable performance, in contrast to computer vision or other AI problems. Meanwhile, two sequence-based machine learning models [34, 218] are included for comparisons.

7.4.3 Evaluation Metrics

Different metrics, including precision (Pre), recall (Rec), accuracy (Acc), F1-score and Matthew's correlation coefficient (MCC) score, are adopted to evaluate the overall prediction performance of these models under comparison, since the dataset is highly imbalanced. The measurements are defined in Chapter. 4.22.

7.4.4 Evaluation and Discussion

Since the datasets exhibit a highly skewed characteristics for the classes, the F1-score is the first one considered to evaluate the performance. The F1-score results of pathogens with taxonomy ID '1491' and '1392' are collectively included in Table. 7.2 and Table. 7.3, and the result of '177416' is included in Table. 7.4. For the performance in Table. 7.2, Table. 7.3 and Table. 7.4, all first two best performances of each column are indicated by bold font.

It is easy to observe that, the performance of different machine learning models for the different dataset vary a lot. For example, the best models by applying different machine learning models on the feature representation algorithms for '1491' are different for the different ratios. For the ratio of '1:25', the best model is achieved by applying auto covariaance feature representation algorithm with SVM model. However, for the ratio of '1:50' and '1:100' in Table. 7.2, the best model are random forests, although the feature representation algorithms are different. It is not easy to identify which one would achieve the best in a combination with an appropriate feature representation algorithm, even for

Table. 7.3 and Table. 7.4.

For the proposed Bi-LSTM-based model, it achieves a more stable and better performance than the others for HPI systems of ID '1392' and '177416'. These two datasets are much bigger than the one of ID '1491', for which Bi-LSTM-based model has not been the best. However, it still yields results quite smoothly when the ratio changes. For the best performance on '177416', Bi-LSTM-based model delivers the results of 0.244036 ± 0.011577, 0.186221 ± 0.014773 and 0.135071 ± 0.014663 for ratio of '1:25', '1:50' and '1:100', respectively. For the rest models, the second best models are decision tree with PsePSSM feature, decision tree with BlockPSSM feature and logistic regression with BlockPSSM feature for ratio of '1:25', '1:50' and '1:100' accordingly. Their performances are 0.164183 ± 0.017133, 0.106201 ± 0.009550 and 0.055682 ± 0.004960. The performance improvements are substantial. In comparison with the evaluation models and the literature methods, Bi-LSTM-based model has demonstrated a better and more stable performance in the three tables, which dataset sizes range from thousands to hundred thousand. It is considered that deep learning model will be more powerful for bigger dataset. In the meantime, Bi-LSTM-based model has shown a stronger capability in dealing with the imbalanced issue.

Table 7.2: Results of F1 score for 'Clostridium botulinum' (ID '1491')

Model		'1491'[a]		
		1:25	1:50	1:100
RF	$\Re_1$[b]	0.956522±0.000000	**0.992000±0.016000**	0.984000±0.019596
	$\Re_2$	0.941414±0.075156	0.959224±0.024432	0.924522±0.083396
	$\Re_3$	0.984615±0.030769	0.968615±0.028956	0.982609±0.021300
	$\Re_4$	0.955429±0.052297	0.992000±0.016000	**1.000000±0.000000**
SVM	$\Re_1$	**1.000000±0.000000**	0.992000±0.016000	0.984000±0.019596
	$\Re_2$	0.968615±0.028956	0.991304±0.017391	1.000000±0.000000
	$\Re_3$	1.000000±0.000000	0.984000±0.019596	0.956522±0.000000
	$\Re_4$	1.000000±0.000000	0.984000±0.019596	0.956522±0.000000
LR	$\Re_1$	0.666667±0.000000	0.406032±0.070505	0.278095±0.009331
	$\Re_2$	0.968615±0.028956	0.992000±0.016000	0.956522±0.000000
	$\Re_3$	0.953846±0.037684	0.939009±0.037895	0.832094±0.100316
	$\Re_4$	0.984615±0.030769	0.984615±0.030769	0.984000±0.019596
Naïve Bayes	$\Re_1$	0.883028±0.024547	0.758788±0.082664	0.649003±0.070794
	$\Re_2$	0.911173±0.043226	0.858851±0.037809	0.771614±0.076227
	$\Re_3$	0.852107±0.030032	0.710083±0.093410	0.508655±0.071947
	$\Re_4$	0.852063±0.028789	0.708051±0.098742	0.534819±0.071385
GBM	$\Re_1$	0.940580±0.019525	0.955429±0.052297	0.911363±0.044262
	$\Re_2$	0.920551±0.051773	0.984000±0.019596	0.828641±0.120346
	$\Re_3$	0.937862±0.055289	0.939348±0.047797	0.876161±0.050595
	$\Re_4$	0.915481±0.091261	0.961231±0.034407	0.856216±0.056709
DT	$\Re_1$	0.870437±0.016061	0.867342±0.076270	0.859912±0.069821
	$\Re_2$	0.767804±0.095732	0.885468±0.081738	0.804467±0.062837
	$\Re_3$	0.934857±0.065113	0.902340±0.063049	0.890957±0.027656
	$\Re_4$	0.892971±0.075027	0.933333±0.054433	0.955429±0.052297
Model_1[c]		0.693333±0.065741	0.928063±0.023236	0.603922±0.031373
Model_2[b]		0.949913±0.038801	0.976000±0.019596	0.977778±0.044444
Proposed Model		0.939009±0.037895	0.925206±0.043780	0.968615±0.028956

[a] '1491' represents the taxonomy ID for the related bacterium pathogen species, details can be found in Table. 7.1;
[b] $\Re_1$–$\Re_4$ are the different feature representations algorithms, representing ACC, CTM, PsePSSM and BlockPSSM;
[c] Model_1 is the method from [218];[d] Model_2 is the method from [34] .

Table 7.3: Results of F1 score for 'Bacillus anthracis' (ID '1392')

Model		'1392'		
		1:25	1:50	1:100
RF	$\Re_1^b$	0.169574±0.009818	0.139858±0.008719	0.067787±0.007116
	$\Re_2$	0.102595±0.019776	0.078685±0.006405	0.056184±0.012903
	$\Re_3$	0.206946±0.009279	0.166349±0.003645	0.092209±0.013757
	$\Re_4$	0.198025±0.016039	0.173696±0.008468	0.104019±0.003440
SVM	$\Re_1$	0.000000±0.000000	0.000000±0.000000	0.000000±0.000000
	$\Re_2$	0.000000±0.000000	0.000000±0.000000	0.000000±0.000000
	$\Re_3$	0.000000±0.000000	0.000000±0.000000	0.000000±0.000000
	$\Re_4$	0.047702±0.029081	0.000000±0.000000	0.002773±0.005546
LR	$\Re_1$	0.021057±0.005808	0.000000±0.000000	0.007090±0.000005
	$\Re_2$	0.050513±0.005512	0.011968±0.002819	0.007072±0.003149
	$\Re_3$	0.030629±0.002586	0.000000±0.000000	0.000000±0.000000
	$\Re_4$	0.108181±0.004233	0.042454±0.005122	0.016177±0.002795
Naïve Bayes	$\Re_1$	0.105464±0.001593	0.057432±0.000310	0.030130±0.000167
	$\Re_2$	0.108749±0.000841	0.067130±0.000972	0.029825±0.000117
	$\Re_3$	0.115154±0.001251	0.060245±0.001256	0.038347±0.000314
	$\Re_4$	0.117401±0.001404	0.062892±0.000381	0.033565±0.000154
GBM	$\Re_1$	0.158191±0.004589	0.117939±0.004031	**0.141740±0.017368**
	$\Re_2$	0.152236±0.007431	0.118625±0.010773	0.093061±0.012445
	$\Re_3$	0.114896±0.009166	0.096469±0.022617	0.090649±0.008828
	$\Re_4$	0.156120±0.012877	0.114349±0.011817	0.100959±0.018261
DT	$\Re_1$	**0.237893±0.013778**	0.039335±0.013057	0.011078±0.016093
	$\Re_2$	0.084730±0.021789	0.034857±0.006656	0.017363±0.007127
	$\Re_3$	0.235041±0.016372	0.072505±0.009355	0.005604±0.011208
	$\Re_4$	0.035104±0.033722	**0.186819±0.013715**	0.079502±0.018047
Model_1^c		0.046355±0.004581	0.051664±0.004274	0.017488±0.002201
Model_2^b		0.199300±0.011655	0.151864±0.005336	0.123265±0.015311
Proposed Model		**0.281453±0.010696**	**0.243263±0.016143**	**0.194048±0.010940**

[a] '1491' and '1392' represent the taxonomy IDs for the related bacterium pathogen species, details can be found in Table. 7.1;
[b] $\Re_1$–$\Re_4$ are the different feature representations algorithms, representing ACC, CTM, PsePSSM and BlockPSSM;
[c] Model_1 is the method from [218];[d] Model_2 is the method from [34] .

Table 7.4: Results of F1 score for 'Francisella tularensis' (ID '177416')

Model		'177416'[a]		
		1:25	1:50	1:100
RF	$\Re_1$[b]	0.039731±0.013702	0.003259±0.003991	0.000000±0.000000
	$\Re_2$	0.028668±0.015407	0.006518±0.003259	0.008085±0.005142
	$\Re_3$	0.068694±0.014250	0.014582±0.006053	0.004878±0.003983
	$\Re_4$	0.042995±0.012560	0.008071±0.008799	0.001619±0.003239
SVM	$\Re_1$	0.126970±0.014229	0.052459±0.006285	0.027375±0.006314
	$\Re_2$	0.022791±0.005985	0.041411±0.012721	0.052282±0.010144
	$\Re_3$	0.121876±0.011250	0.040185±0.007599	0.000000±0.000000
	$\Re_4$	0.106272±0.014042	0.019592±0.006398	0.000000±0.000000
LR	$\Re_1$	0.008210±0.000027	0.000000±0.000000	0.000000±0.000000
	$\Re_2$	0.062458±0.007364	0.011429±0.003999	0.000000±0.000000
	$\Re_3$	0.000000±0.000000	0.000000±0.000000	0.000000±0.000000
	$\Re_4$	0.145415±0.010038	0.082374±0.010049	**0.055682±0.004960**
Naïve Bayes	$\Re_1$	0.115664±0.001010	0.063485±0.000650	0.035815±0.000347
	$\Re_2$	0.113069±0.001109	0.055818±0.000839	0.028822±0.000140
	$\Re_3$	0.123359±0.002777	0.076426±0.001523	0.039672±0.000415
	$\Re_4$	0.119222±0.001129	0.066671±0.000397	0.034570±0.000356
GBM	$\Re_1$	0.076012±0.008883	0.074200±0.024640	0.040672±0.007316
	$\Re_2$	0.102556±0.023905	0.044597±0.005541	0.037121±0.008202
	$\Re_3$	0.111070±0.007061	0.091760±0.009066	0.047583±0.007300
	$\Re_4$	0.121684±0.017300	0.081683±0.007212	0.051077±0.012051
DT	$\Re_1$	0.153325±0.023107	0.017464±0.011643	0.000000±0.000000
	$\Re_2$	0.035766±0.036349	0.020309±0.014948	0.006439±0.005997
	$\Re_3$	**0.164183±0.017133**	0.049342±0.005803	0.014460±0.011808
	$\Re_4$	0.001639±0.003279	**0.106201±0.009550**	0.020142±0.014128
Model_1[c]		0.029041±0.010670	0.004918±0.004016	0.000000±0.000000
Model_2		0.108861±0.015895	0.067577±0.011019	0.052312±0.012894
Proposed Model		**0.244036±0.011577**	**0.186221±0.014773**	**0.135071±0.014663**

[a] '177416' represents the taxonomy ID for the related bacterium pathogen specie, details can be found in Table. 7.1;
[b] $\Re_1 - -\Re_4$ are the different feature representations algorithms, representing ACC, CTM, PsePSSM and BlockPSSM;
[c] Model_1 is the method from [218];[d] Model_2 is the method from [34]

7.5 Summary

In this chapter, a Bi-LSTM-based model achieving a more robust and effective performance was proposed. A multi-channel feature representation algorithm, which is based on tree-based feature selection algorithm and synthetic minority over-sampling technique (SMOTE), was firstly designed. Later, the bidirectional LSTM model was introduced as the learning model. Given the scenario of imbalanced issue, the focal loss function was subsequently employed as a novel cost function for the training. The prediction performance of HP-PPIs dataset has indicated that Bi-LSTM-based model has obtained the best results.

Chapter 8

UNSUPERVISED DEEP LEARNING MODEL FOR DISCOVERY OF HP-PPIS

This chapter will investigate the host-pathogen protein-protein interactions with an unsupervised deep learning model based on stacked denoising autoencoders. Given the various feature representation algorithms and the different characterisitcs exhibited by the HP-PPIs dataset, we firstly introduce and briefly review the current state-of-the-art techniques for prediction task of HP-PPIs in Chapter 8.1 and Chapter 8.2. The goal of this chapter is to propose an unsupervised deep learning model, which is capable of mining latent protein information for model training to improve the performance. Thus, the proposed model based on stacked denoising autoencoder will be discussed in Chapter 8.3. The stacked denoising autoencoder further extends the capability of mining higher level information from protein sequence in the model, and the designed multi-layer model has subsequently obtained the advantage in the training phase. The experiment evaluation is discussed in Chapter 8.4. The achieved performance indicates a superior capability of the unsupervised deep learning model in dealing with the host-pathogen protein interactions scenario among all of these models.

8.1 Introduction

Given the high volume and variety of data, many researches are being conducted in data analytics to predict and uncover information and knowledge concerning related domains, including computer vision, economics, online resources and bioinformatics. Based on the availability of data, computational biology methods, including omics fields, biomedical imaging, and biological signal processing [3], have grown in importance, with pilot studies having been previously conducted in areas such as genomics and proteomics areas [1], and biomedical medicine and imaging areas [2].

Proteomics is an important branch of system biology in the post-genomics era, with data analytics playing a vital role in understanding and predicting biological knowledge for proteins. Proteomics research focuses on utilising existing experimental data related to the protein interactions in order to elucidate high-fidelity interaction networks for future biological experiments. Predicting protein-protein interactions remains an active research area in bioinformatics [22]. Among the protein interactions, inter-species protein-protein interactions (HP-PPIs) are one type of interactions observed within the same species. Thus, it is motivated to study inter-species PPIs to reveal interactions between proteins from different species. Specifically, host-pathogen interactions (HPI) are considered as key infection processes at the molecular level with the associated infectious diseases representing major worldwide health concerns, which have caused millions of illnesses annually.

There has been an accumulation of experimentally verified PPI data generated through *in vitro* methods, including small-scale biochemical, biophysical, and genetic experiments, as well as large-scale methods, such as yeast-two-hybrid analysis. However, these methods are time consuming and require substantial biomedical resources. Additionally, many of the methods exhibit high false positive rates, and the occasional large number of potential interactions hinders the deployment of some *in vitro* methods.

Here, this chapter is designed to describe the development of a new method for

HP-PPIs prediction. Since host-pathogen protein-protein interactions reveal substantial information concerning HP-specific infection mechanisms, a better understanding on HP-PPIs and the application of computational methods to promote their prediction will assist *in vitro* experimental design.

In this chapter, the development of an unsupervised deep learning model is designed to handle the HP-PPIs datasets, and the comparison against various supervised machine learning models indicates that unsupervised deep learning model achieves a best performance, particularly when the HP-PPIs datasets present both small and large scales. Meanwhile, a highly skewed ratio between different classes exhibits a significant challenge for model learning.

8.2 Related Work

As PPIs offer insights into molecular interactions and disease genes identification [313] for a specific species, such as yeast [314], biological experiments are being carried out to reveal or determine the interaction-specific relationships between proteins. In this regard, HP-PPIs could further assist revealing the information concerning infection pathways and providing additional insight from the interactions between host and pathogens [49].

A previous review [49] detailed the research vision for HP-PPIs and it highlighted the importance of database construction. Several databases, including *HPIDB* [10], *PATRIC* [160], *PHISTO* [161], *VirHostNet* [315] and *VirusMentha* [152], represent the most relevant PPI repositories. Owing to these earlier research efforts, these databases provide well sorted and experimentally verified HP-PPIs information. Nevertheless, these manually updated databases currently represent only a small quantity of all PPIs.

There have been several recent studies on host-pathogen protein-protein interactions [23, 157, 316, 317], with each testing a biological hypothesis that 'similar pathogens target the same critical biological processes in the host' through the use of learning models. These studies constructed a common structure using the pathway information to compute the similarities between different types of pathogens, with human considered

as the primary host. One of these studies constructed a pairwise level multi-task model to combine two different tasks. A potential solution for combining more tasks in the multi-task model has been proposed in [23], where the term 'Task' describes a computational model used to predict interactions between a specific pathogen and host.

Since supervised machine learning models have been widely applied for diverse topics of biological data, such as the decision tree for lung carcinoma cancer prediction model [318] and an lung cancer diagnosis system based on support vector machine [319], the traditional supervised machine learning models have been utilized to facilitate PPI research. A previous study used two pathogen-human datasets as source tasks and a third one as a target task to build a transfer learning model. Two other studies described extreme learning machine (ELM) models, which aimed at obtaining faster training speeds and higher degrees of accuracy [21, 320]. Such a model was deployed via using a balanced intra-species PPI dataset. Additionally, one method using Naïve Bayes classification model was described in [215] and the results for a comprehensive study and prediction of PPIs on yeast and humans via three-dimensional structural information were presented. The algorithm (PrePPI) uses Bayesian statistics to derive relationships between structural information and other functional clues. This method yields over 30,000 high-confidence interactions for yeast and over 300,000 for humans [215].

Given the potential in utilizing computational models, especially machine learning models, to facilitate the HP-PPIs task, possible solutions have been widely discussed in [225] and [277]. Without positioning verified databases and specific pathogens, a collection of traditional machine learning models has been assessed, including support vector machine, decision tree, Naïve Bayes and so on. Deep learning models, which have shown great power in protein structure prediction task [321, 322], have also been included as very important categories of machine learning models for prediction of HP-PPIs. However, a comprehensive framework with detailed artefacts to illustrate data analytics and machine learning models for HP-PPIs is still needed. Meanwhile, how to leverage deep learning model to improve the performance comparing with traditional

machine learning models is also lacking.

8.3 Unsupervised Deep Learning Model

Given the large number of databases, data analytics and learning models will contribute to HP-PPIs research. Following previous chapters' fashion, a complete framework for HP-PPIs research involves data pre-processing, feature representation, and learning model. In this section, the learning model will be firstly discussed, which mainly details the proposed unsupervised deep learning-based model.

8.3.1 Unsupervised Deep Learning Model

Deep learning models have achieved good performance on both classification and regression tasks, suggesting their generalized utility for learning relationships from data [3, 321–324]. These models have shown that, deep learning models are capable of learning protein structure prediction task in a more efficient way, and can achieve better performance than the other models.

There is another group of unsupervised deep learning model, namely denoising autoencoder (dA), which represents features via a deep neural network. Denoising autoencoder [325] is a training model used for unsupervised learning. It is motivated from general autoencoder and is capable of reconstructing original input from corrupted input. Additionally, the denoising autoencoder could be stacked as stacked denoising autoencoders (SdA) to build a multi-layer network [323].

As a primary unsupervised learning model, a stacked denoising autoencoders can construct higher level features to allow for a better initial state in the deep learning model. Herein, an SdA model is applied as the unsupervised model to learn from the curated datasets comprising three different bacterial species, whereas at the top layer, logistic regression (LR) [326] is chosen as the classification model. The network is subsequently fine-tuned to achieve better performance than simply training the network in two separate stages [327].

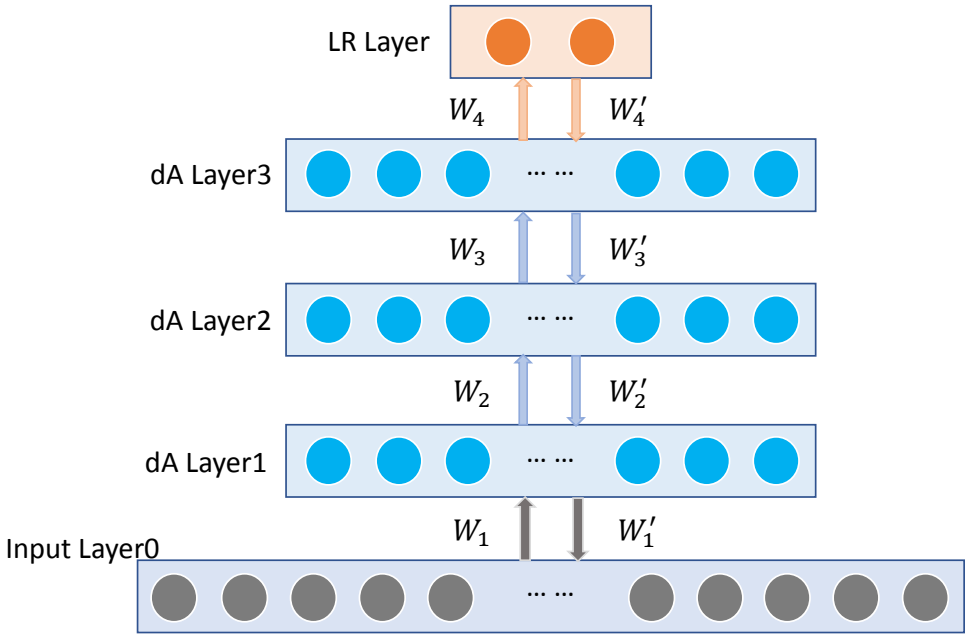


Figure 8.1: The Whole Model based on SdA

Technically, the input will be corrupted by adding small amounts of noise, in which both *Gaussian* noise and *'mask'* noise are feasible. The integrated model is depicted in Fig. 8.1.

This four-layer network is applied to learn and predict from the HP-PPIs datasets. It has a similar architecture as that of another work described in [327]; however, the network is fine-tuned following initial training using *LR Layer*. The architecture of this network is as follows: *input layer (420 input nodes)* $\rightarrow$ *dA layer1 (210 neurons)* $\rightarrow$ *dA layer2 (210 neurons)* $\rightarrow$ *dA layer3 (210 neurons)* $\rightarrow$ *LR layer (1 output node)*.

In Fig. 8.2, the details of construction of the denoising autoencoder layer is described. In Fig. 8.2, the $\ddot{X}$ is the corrupted input data from X. For the experiments, it ended up with choosing only *Gaussian* noise as it achieved better performance over $\ddot{X}$ wit *'mask'* noise. The encoding process and decoding process is given as:

$$\begin{aligned} Y &= W * \ddot{X} + b_x \\ \tilde{X} &= W' * Y + b_h \end{aligned} \tag{8.1}$$

The dA layer trains each layer as an individual component first, followed by output of

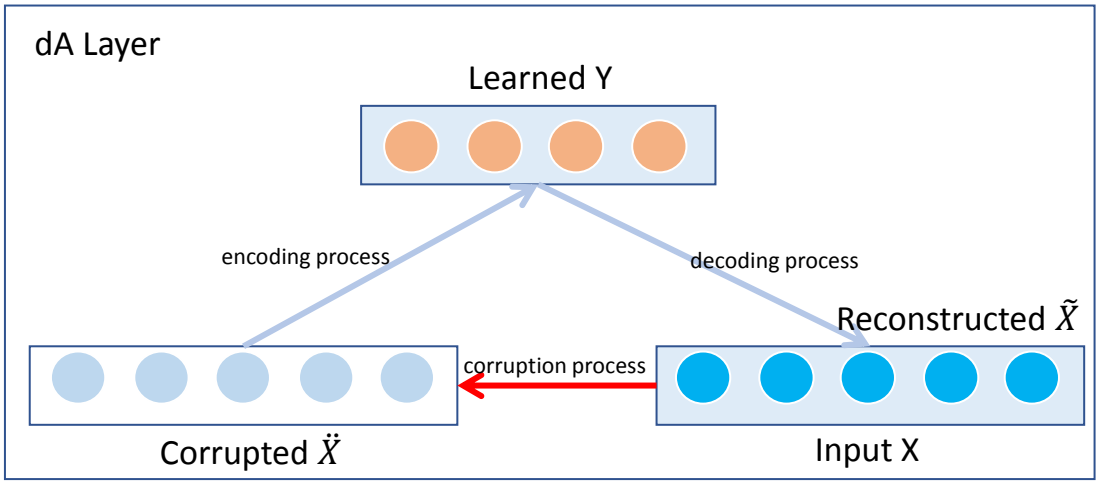


Figure 8.2: The Denoising Autoencoder Layer

the learned data, Y, to subsequent layers. The learned parameters, W, are maintained and will be applied to the entire network during subsequent fine-tuning steps. Each layer is pre-trained using the same process.

The logistic regression layer is the final classification layer. For a binary classification problem, $y_i = 0, 1$, where i represents the i_{th} example, the LR model returns the result according to the following:

$$\begin{aligned} P(y_i = 1|x_i) &= h_\theta(x_i) = 1/(1 + exp(-\theta^T * x_i)) \\ P(y_i = 0|x_i) &= 1 - P(y_i = 1|x_i) = 1 - h_\theta(x_i) \end{aligned} \tag{8.2}$$

Here, θ represents the model parameters. The cost function applied in logistic regression model is:

$$J(\theta) = -\sum_i (y_i log(h_\theta(x_i)) + (1 - y_i) log(1 - h_\theta(x_i))) \tag{8.3}$$

After pre-training the different layers, the overall network using a back propagation algorithm is fine-tuned.

8.3.2 Traditional Learning Algorithms and Models

The deep learning model is designed as the primary model for model learning and prediction. Meanwhile, several traditional supervised learning models are also implemented for comparison, including a linear-kernel support vector machine (SVM), extreme learning machine (ELM), naïve Bayes and decision tree models.

Besides the traditional learning models, to input information related to each unique protein interaction into a learning model, feature representation is required. Since sequence information includes most information of the corresponding protein and is protein specific, in this study, we primarily use sequence information for feature representation. Following the collection of positive protein-protein interactions from various database repositories, the negative protein-protein interactions are also curated by a random sampling strategy to build the supervised machine learning model.

As a result of the data pre-processing, a HP-PPIs dataset will be ready, which indicate only the identities of interacting proteins between host and pathogen.

For different protein properties, it is required to represent the properties into a numerical form. In the past, numerous studies related to feature representation have been conducted for sequence information ([23, 33, 232, 283, 328]). The feature representation remains a hot and ongoing research area for bioinformatics researchers. The unique information include the different types of amino acids in different combination and various lengths. As said in 'The amino acid sequence of a protein determines its three-dimensional structure' [329], it also provides a widely adopted view that knowledge of the sequence information would be adequately feasible to represent a protein.

In this chapter, auto covariance algorithm (ACC) [33] is selected as the first step of features representation methodology. As one of the popular feature representation algorithms, ACC is capable of transforming numerical vectors to uniform matrices based on sequence information. The representing matrices are having a same dimension after ACC transformation regardless of protein sequence length. For the details of ACC algorithm, it could be referred to Chapter 4. In this study, the length of each vector

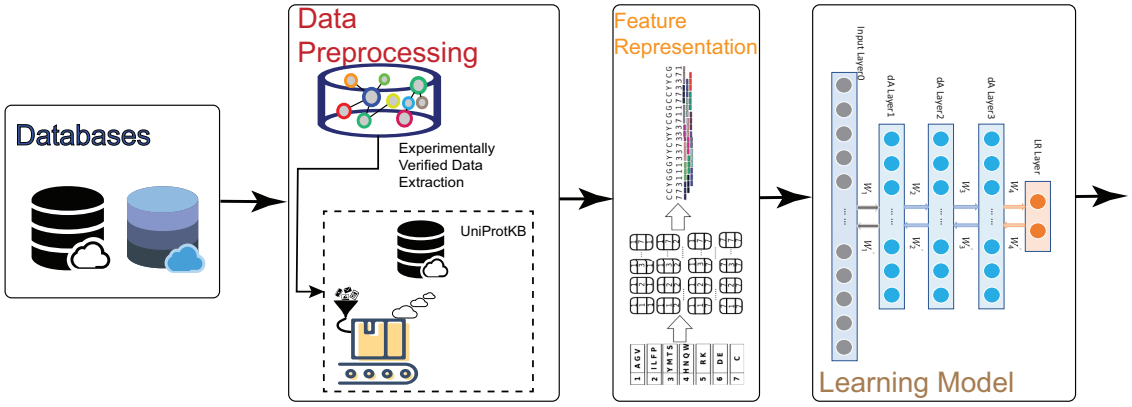


Figure 8.3: The Overall Framework of the Stacked Denoising Autoencoder-based Model

was set to 210 for each protein, resulting in a pair-wise feature vector of 420 dimensions for each HP-PPIs pair.

By this step, a curated HP-PPIs dataset is ready to be fed into the learning model. In Figure. 8.3 overall framework of the unsupervised learning-based model is illustrated, which is capable of constructing higher level features and initiate a deep neural network in a better state. The stacked denoising autoencoders is deployed to achieve a boost performance. In next section, the model will be evaluated on several HP-PPIs datasets in comparison with the other traditional machine learning models.

8.4 Evaluation and Discussion

To evaluate the feasibility of the framework discussed in section 3, a detailed practice is presented in this section. Specifically, two HP-PPIs database repositories, PATRIC and PHISTO, were used for construction of the HP-PPIs datasets. The benefit from these two databases is that, the hosted positive data are manually extracted and uploaded based on biological literature.

Table. 8.1 shows the statistics associated with the bacterium pathogen species used for construction of the datasets and used for model learning. For the data redundancy, we have conducted two different ways to remove the redundant interaction pairs, which are manual redundancy and CD-HIT redundancy removal. The manual redundancy removal

Table 8.1: Processing of HP-PPIs Dataset

Species	Positive Pairs	Manual Redundancy	CD-HIT Redundancy Removal	Ratio 1:100
Clostridium difficile	56	53	52	5252
Escherichia coli	168	104	98	9898
Bacillus anthracis	6073	3138	3035	306535

aims at the repeated interaction pairs in the original databases, which might have been reported more than once by different researchers. For the CD-HIT redundancy removal, it compares the sequence similarity between different proteins and removes the redundant ones with high similarity. These two steps ensure the dataset with less redundancy from both interactions IDs and protein sequence similarity. After the data redundancy analysis, three different bacterium pathogen species were retained containing enough samples for model training and also in the best interest of infectious diseases for human. The HP-PPIs datasets are corresponding to *Clostridium difficile*, *Escherichia coli*, and *Bacillus anthracis* in the study, as shown in Table. 8.1, with the positive protein pairs numbers decreasing after redundancy analyses. Here, *Clostridium difficile* is the primary cause of the inflammation of the colon, *Escherichia coli* causes both minor and severe intestines illness and *Bacillus anthracis* is the etiologic agent of anthrax.

As shown in Table. 8.1, the relatively small datasets that included 56 and 168 pairs of positive HP-PPIs are utilised in this chapter, meanwhile, the large size dataset with 6073 pairs of positive HP-PPIs is also exploited. The ratio of positive and negative pairs is set at 1:100 to align with experiment scenarios, which is normally considered to yield less bias in predictions (Table. 8.1).

We further evaluated the learning models by 10-fold cross validation after dividing the HP-PPIs datasets into training and test datasets. Details are listed in Table. 8.2.

Table 8.2: Statistics of HP-PPIs Dataset

Species	Training size	Test size
Clostridium difficile	4545	707
Escherichia coli	8181	1717
Bacillus anthracis	275427	31108

8.4.1 Evaluation Metrics

To evaluate the performance and robustness of the models, the experiments are conducted using 10-fold cross validation. The evaluation results are presented as the mean and variance in terms of precision, recall values, F1 score, and accuracy. It should be noted that the accuracy measurement might not fully reflect the performance of these models, because the datasets are highly skewed. However, we have reported these results for completeness. The precision value represents the fraction of retrieved information relevant to the result, whereas the recall value represents the ratio of successful retrievals by the learning model. These are critical factors necessary to determine system performance, specifically on an imbalanced dataset. The precision and recall values are further used to calculate a harmonic average, which is subsequently termed as F1 score to provide a final measurement for a given model. Normally, the F1 score is ranging between 0 and 1. It reaches the best performance at 1 while worst at 0. The definition formulations can be referred to Equa. 4.22

8.4.2 Evaluation and Discussion

Although supervised machine learning model is considered as the dominant classification model, the unsupervised deep learning model is introduced in this chapter to build a complementary feature representation, which also helps tuning multi-layer supervised model. As for learning models for comparison, several traditional supervised machine learning models are simultaneously built, including support vector machine (SVM), extreme learning machine (ELM) and Naïve Bayes Model, among others.

In this chapter, the SdA, SVM, ELM, decision tree, naïve Bayes and also logistic regres-

sion models are implemented with support partially from 'Tensorflow' ([330]), 'libsvm' ([331]), 'hpelm' ([332]) and 'scikit-learn' ([256]). Furthermore, training deep learning model on big datasets highly relies on specific structures, such as GPU/TPU/FPGA, to decrease the running time and finalise the parallel processing tasks. In this regard, the computing resources system is built upon 'NVIDIA GTX 1080Ti' GPU and 64GB RAM, which allowed efficient parallelization computing. The working operating system is Ubuntu 16.04. In this study, all framework implementations were written in Python.

Table 8.3: Precision Result of Models (%)

Species	Gaussian NB	LR	SVM	DT	ELM	SdA
Clostridium difficile	78.53±11.37	97.50±0	96.25±5.73	84.88±9.48	97.50±5.0	**100±0.00**
Escherichia coli	2.52±0.55	50.30±9.99	62.86±14.95	49.16±11.13	20.00±40.00	**87.00±6.52**
Bacillus anthracis	1.65±0.04	92.48±7.97	70.00±45.83	60.25±1.33	10.00±30.00	**92.49±2.04**

Table 8.4: Recall Result of Models (%)

Species	Gaussian NB	LR	SVM	DT	ELM	SdA
Clostridium difficile	**100±0**	98.57±4.29	98.57±4.29	95.71±6.54	94.29±7.00	98.57±4.29
Escherichia coli	**71.76±14.11**	35.88±10.00	29.41±11.16	70.59±10.85	1.18±2.35	51.18±8.34
Bacillus anthracis	**79.83±2.27**	4.42±1.28	0.39±0.32	66.72±2.90	0.03±0.10	48.83±2.86

Table 8.5: F1 Result of Models

Species	Gaussian NB	LR	SVM	DT	ELM	SdA
Clostridium difficile	0.8752±0.307	0.9790±0.0322	0.9723±0.0340	0.8954±0.0571	0.9559±0.362	**0.9923±0.0230**
Escherichia coli	0.486±0.106	0.4097±0.0899	0.3939±0.1295	0.5775±0.1126	0.222±0.444	**0.6382±0.0649**
Bacillus anthracis	0.323±0.009	0.0841±0.0238	0.0077±0.0063	0.6330±0.0175	0.006±0.019	**0.6387±0.0278**

Table 8.6: Accuracy Result of Models (%)

Species	Gaussian NB	LR	SVM	DT	ELM	SdA
Clostridium difficile	99.70±0.18	99.96±0.06	99.94±0.07	99.77±0.13	99.90±0.09	**99.99±0.04**
Escherichia coli	71.88±2.57	98.99±0.18	99.13±0.15	98.95±0.37	98.98±0.09	**99.44±0.07**
Bacillus anthracis	52.57±0.27	99.05±0.01	99.01±0.00	99.23±0.03	99.01±0.00	**99.45±0.03**

Table 8.7: The Area Under Curve Value of Models

Species	Gaussian NB	LR	SVM	DT	ELM	SdA
Clostridium difficile	0.9985±0.001	0.9991±0.0026	0.9926±0.0214	0.9776±0.0326	**0.9997±0.0005**	0.9985±0.0045
Escherichia coli	0.7182±0.0756	0.9413±0.0204	0.6462±0.0559	0.8491±0.0553	**0.9448±0.0276**	0.9431±0.0318
Bacillus anthracis	0.6607±0.01	0.7675±0.0125	0.5019±0.0016	0.8314±0.0145	0.8157±0.0099	**0.9250±0.0112**

Primary Results

The precision and recall values for all of the models are firstly collected. Table. 8.3 shows the statistics associated with precision results, Table. 8.4 for the recall results, Table. 8.5 for the F1 results and Table. 8.6 for the accuracy results. In these tables, 'SVM' refers to linear-kernel SVM, 'ELM' represents to extreme learning machine while 'SdA' is the stacked denoising autoencoders model, 'Gaussian NB' indicates Gaussian Naïve Bayes, 'DT' refers to decision tree model and 'LR' is logistic regression model.

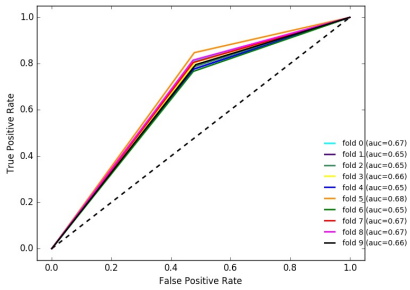
The results of receiver operating characteristic (ROC) and the area under ROC curve (AUC) value analysis for 'Bacillus anthracis' are shown in Fig. 8.4. The ROC results illustrate the classification ability of binary HP-PPIs prediction according to various discrimination thresholds. It was plotted based on different settings of TP rates against FP rates. The AUC value ranges between 0 and 1 with higher values indicating a better classification performance.

Moreover, it is worth noting that ELM model achieves better AUC value on smaller datasets based on the comprehensive results from Table.8.7. It achieves AUC values of 0.9997 for C. difficile and 0.9448 for E. coli. However, across all three tasks, the SdA model presents a more stable performance (0.9985 for C. difficile, 0.9431 for E. coli and 0.9250 on B. anthracis). From Table. 8.7, it is observed that the performance of SdA model on B. anthracis specie is much better than the others, including the followings from decision tree model (0.8314) and ELM model (0.8157).

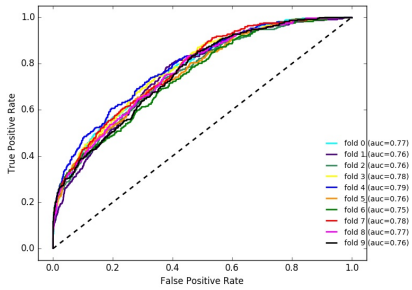
Discussion

From Table. 8.3 and Table. 8.4, the proposed SdA model has illustrated a strongest capability of precision result. However, for the recall result, the best model is achieved by Gaussian NB model, though the performance of SdA model has been moderate. Overall, the F1 score reported in Table. 8.5 indicates that the prediction performance of SdA model is the best.

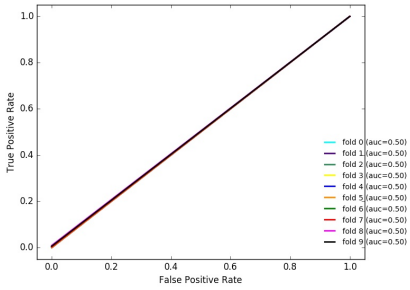
According to these measurements, the SdA model achieved the best performance on



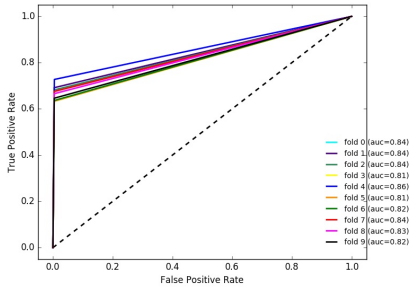
(a) Gaussian Naïve Bayes (0.6607±0.01)



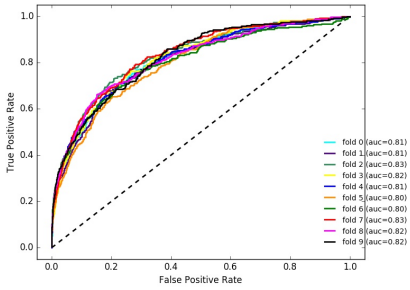
(b) Logistic Regression (0.7675±0.0125)



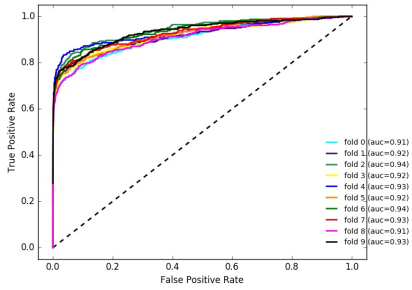
(c) SVM (0.5019±0.0016)



(d) Decision Tree (0.8314±0.0145)

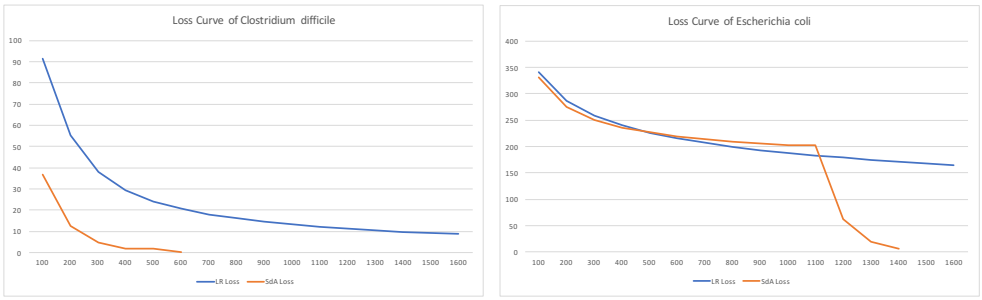


(e) ELM (0.8157±0.0099)



(f) SdA (0.9250±0.0112)

Figure 8.4: Learning Models ROC Curve on Bacillus anthracis



(a) Clostridium difficile (b) Escherichia coli

Figure 8.5: Convergence Curve

F1 score as well as accuracy for HP-PPIs prediction for Clostridium difficile, Escherichia coli and Bacillus anthracis. Specifically, the SdA model outperformed the LR model in terms of F1 score and accuracy, indicating that the unsupervised learning model presented a better feature learning capability and resulted in an improved predictive performance. Although model performances on different datasets are varied, the SdA model retains the best performance among all the models. It is witnessed that, for both small and big datasets, SdA model has benefitted from the unsupervised learning model, which generates a higher-level feature. The four-layer model has obtained a best performance for all three different datasets with regard to the accuracy result in Table. 8.6.

Furthermore, we have considered the training time, which may have been a big challenge for training deep learning model. Regarding learning and convergence curve, the related comparison results are presented in Fig. 8.5. The convergence curve represents the relationship between the training epoch and global loss, with a lower global loss suggesting the closeness of the model to the optimal state.

Fig. 8.5 shows the convergence curves for logistic regression and SdA model, with pre-training step for the SdA model initially applied in the SdA layers, after which the output of the last SdA layer is used as input for the logistic regression layer. Our results indicated that the training iterations needed for the SdA model for C. difficile and E. coli HP-PPIs prediction were much less than those needed for training the LR model. Retaining the parameters from the pre-training step in the SdA layers improved the convergence speed and aided the efficient realization of the optimal state.

8.5 Summary

A well-designed framework of HP-PPIs study will facilitate the exploration and understanding of HP-PPIs networks, and offer critical insights of infectious mechanisms between host and pathogen. In this chapter, a SdA-based deep learning model for HP-PPIs datasets is presented and the comparison of the SdA model with other models indicated its superiority for this application. From the evaluation result of this chapter, the unsupervised SdA model is optimal for the highly skewed and big datasets and is better at feature representation if compared to other models. Additionally, model convergence speed has benefited from the unsupervised learning technique and the usage of GPU. The results suggested that, the deep learning model was capable of dealing with big HP-PPIs datasets.

Chapter 9

STRUCTURAL PRINCIPLES ANALYSIS OF HOST-PATHOGEN PROTEIN-PROTEIN INTERACTIONS: A STRUCTURAL BIOINFORMATICS SURVEY

Computational-intelligence methods in bioinformatics and systems biology show promising potential for leveraging abundant, large-scale molecular data. These methods can facilitate analysis and prediction of the principles of biological systems through construction of statistical and visualised models. Specifically, structural data from exogenous and endogenous protein-protein interactions are of vital significance in this context, encompassing primarily three-dimensional (3D) structural information for a cohort of macromolecules underpinning the biological system.

This chapter anticipates to further survey the main methodologies and algorithms for the reconstruction and modelling of the structural-interaction networks (SINs) of host-pathogen protein-protein interactions (HP-PPIs), regarding how the protein domains interact with each other to constitute a SIN. Surveying the pattern and organisation of the SIN delivers a state-of-the-art view of HP-PPIs and illustrates prospective future research directions. In addition to the binary PPI network, the relevant data sources into several branching research areas will be distilled and the discussions will be further extended into computational-intelligence methods to shed light on effective method design. In particular, atomic resolution level investigations can reveal novel insights

into the underlying principles of the organisation and complexity of HP-PPIs networks. Combining data analytics and machine learning technologies, it is anticipated that this systematic overview will serve as a useful guide for interested researchers to carry out related studies on this exciting and challenging research topic in system biology in the future.

9.1 Introduction

In this chapter, the main goal is to discover how the computational-intelligence methods can help solve key problems and the dominant mechanisms involved in proteomics research. Considering proteomics represent the large-scale study of proteins, proteomics relies upon the investigation of several aspects, including when, where, and how proteins function, and how proteins interact with each other. Recently, an abundance of experimental data has accumulated, propelling hypothesis-driven biomedical research into the big-data era.

Given the continuous growth and availability of large-scale multi-omics data, both the protein-protein interaction (PPI) networks and structural analyses involving proteomics remain hot topics. Exploration of proteomics data sources, such as those from the European Bioinformatics Institute [4, 6, 333], promotes research in transforming biomedical research at system-level, mechanistic studies aimed at a comprehensive and holistic understanding of biological systems [8]. Although challenges, such as specialised domain knowledge and data issues, might hinder proteomics researches, this data-driven work to obtain extensive information about systems from large amounts of raw data is currently popular in both academia and industry [3].

Systems biology [334] represents the comprehensive study of presenting a holistic view and analysis of biological processes. Specifically, systems biology aims to understand and further predict the behaviour of biological systems [335] and includes studies on functional genomics and proteomics. There are several studies focusing on genomics data, mostly from The Cancer Genome Atlas (TCGA) [7], given that a nearly complete map

for human and other species had been provided along with the development of genome-sequencing projects [335]. These studies provided insights into gene-related networks and a fuller understanding of how a set of molecules interacts with each other [65]. Three-dimensional (3D) structures of these molecules are the most critical for deriving relationships.

This chapter is focused on proteomics, and specifically on HP-PPIs. Considering the prevalence of protein interactions between species, most early studies were performed within the same species due to the limited availability of proteomics data at that time [33, 283]. Several recent studies demonstrated improvements in PPI between different species, which were referred as 'interspecies PPI', and that offered important information for further analysis of infectious mechanisms [327, 335]. However, beyond the interaction between these PPIs, their structural information is vital to their discovery. We anticipate that study of the identified data collected via open databases [49] would present a comprehensive survey towards structural principles concerning the PPI identified between the host and pathogen. These HP-PPIs are experimentally verified and manually recorded in systems and include information regarding infection pathways in their interaction networks and are able to reveal much more information regarding the infectious mechanisms between hosts and pathogens. We first investigated a previous HP-PPIs study [49] and expanded our work based on the preliminary sequence information [327, 336] to exploit the online available and experimentally verified HP-PPIs data. However, these studies simply focused on binary protein interactions prediction.

In addition to these studies, we expect to leverage the structural information of the HP-PPIs data for building structural-interaction networks (SINs) with respect to simply classifying pairs of proteins as interacting or not. The structural information of the HP-PPIs represents various protein properties, from which systems biology might extract a highly convincing network-analysis result and introduce trustworthy statistics in cooperation with the corresponding structural information and domain data, as well as the atomic resolution-level networks.

Therefore, the structural-principle analysis of HP-PPIs networks is discussed and surveyed in the following sections, which covers most branches closely associate with the protein structural information. This analysis was achieved by SIN, an atomic-resolution PPI network [28]. Protein structural information is another experimentally determined set of 3D data previously described. It mainly contains several protein properties, including domain information, family annotation, secondary/tertiary structure.

Because there are few 3D-specific studies offering an atomic view of HP-PPIs, we provide an overview of progress made by biologists in relation to bioinformatics, including 3D structural databases and analysis based on the structural information. It is anticipated that the efforts will help to navigate gaps between biological analysis and computational modelling. This includes: 1) Protein secondary/tertiary structure prediction; 2) Domain-domain interaction prediction. These provide the basics for reconstruction of a SIN.

9.2 Preliminary Concepts

The two main predictive tasks associated with proteomics related to computational biology are the protein structure and the domain-domain interaction. Both sets of data are usually difficult for bioinformatics researchers to obtain; however, building a SIN requires a complete understanding of both protein structure and domain features. In this section, we present the biological meaning for both the structural information and the domain-domain interactions, and also introduce the modelling process necessary for completing the prediction of both tasks.

9.2.1 Sequence Information

Proteins are comprised of various numbers of amino acids as their basic building blocks. The concatenated string of amino acids forming the folded protein represents its primary sequence information. Typically, there are 20 different proteinogenic amino acids [283], although five additional amino acids exist in the human and pathogen protein sequences [49], including selenocysteine/U, pyrrolysine/O, aspartate or Asparagine/B, glutamate,

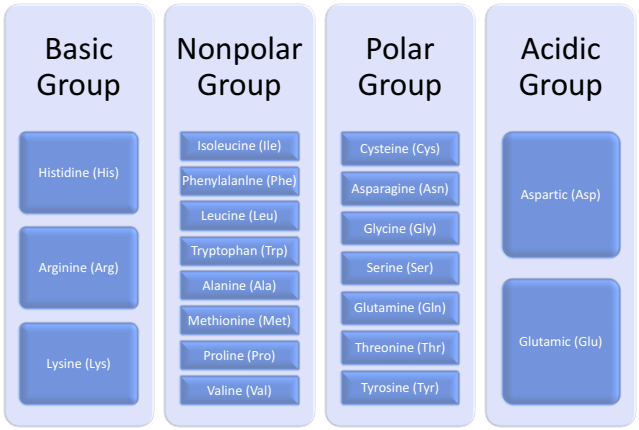


Figure 9.1: Amino Acids Groups

and glutamine/Z. Figure 9.1 shows the 20 different amino acids.

As a preprocessing step for inputting sequence data into computational model built for protein classification and regression tasks, transformation of efficient and effective data into the model is necessary. Sequence representation is a vital preprocessing step for efficiently and effectively feeding data to any computational model for protein classification and regression analysis. In Table 9.1, several mainstream algorithms concerned with sequence representation is listed, where the protein sequence is denoted as $X = x_1, x_2, ..., x_n$. For a complete list of the feature representation algorithm review, please refer to Chapter. 4.

These different sequence-representation algorithms provide as much information as possible to the computational model in different vector lengths. Because the sequence information is easier to obtain via the high-throughput technology, it is primarily utilised for both protein structure prediction and interaction prediction.

Table 9.1: Protein Sequence Representation Algorithms

Algorithm	Reference	Definition	Prefix	Equation	Feature Dimension
Amino acid composition	[337]	Each feature represents the frequency of the corresponding amino acid type in the protein	aa_i is one of the 20 types of amino acids $aa_1, aa_2, ..., aa_{20}$	$f_i = \frac{counts_{aa_i}}{n}$	$1 * 20$
Conjoint triad method	[283]	Considering the properties of one amino acid and its vicinal amino acids as a pattern f_i, the frequency of f_i represents one feature. The concatenation of these f_i defines a unique feature vector.	For the amino acids that have been catalogued into seven classes, $F = f_1, f_2, \ldots, f_{343}$. $D = d_1, d_2, \ldots, d_{343}$	$d_i = \frac{f_i - min(f_1, f_2, ..., f_{343})}{max(f_1, f_2, ..., f_{343})}$	$1 * 343$
Auto covariance	[33]	Projecting the amino acids with their specific seven kinds of physiochemical properties, auto covariance formalizes the sequence information into a uniform matrix	$P_{i,j}$ is the jth property of the ith amino acid, while the protein has n amino acids. lag is defined as the distance between two amino acids and lg is the maximum value of lag.	$AC(lag, j) = \frac{1}{n-lag} \sum_{i=1}^{n-lag} (P_{i,j} - \frac{1}{n} \sum_{i=1}^{n}) * (P_{i+lag,j} - \frac{1}{n} \sum_{i=1}^{n} P_{i,j})$	$lg * 7$
Local descriptor	[232]	Segmenting a protein sequence into several individual regions, i.e. 10 regions in [232], three descriptors are used to describe each region, including Composition (C), Transition (T) and Distribution (D).	The basis to group amino acids is considered by different biology schemes, i.e. three functional groups (hydrophobic (CVLIMFW), neutral (GASTPHY) and polar (RKEDQN)), seven physiochemical groups.	$i = 1, 2, \ldots, 7$; $c_i = \frac{counts_{C_i}}{n}$; $t_i = \frac{counts_{T_i}}{n-1}$; $d_i = \frac{loc(T_i)}{n}$, $loc(T_i)$ denote the location index of i	$1 * 630$
Position-Specific Scoring Matrix (PSSM)	[249]	The defined matrix, P, is in $n * 20$ dimensions, where $P(i, j)$ indicates the possibility of the jth amino acid appears at i position. PSI-BLAST [338] is one of the most frequently used tools. PSI-BLAST [236] is one of the most frequently used tools.	The protein sequence is divided into 20 blocks while its length is n.	$P_{ij} = \sum_{k=1}^{20} w(i,k) * Y(j,k)$; $F_j = \frac{1}{B_j} \sum_{i=1}^{B_j} P_i^{(} j)$	$20 * 20$
One-hot sparse vector	[339, 340]	Each amino acid is defined in a one-hot sparse vector. The length, M, of vector is dependent upon the number of the amino acid types, i.e. 25 in [49], 22 in [339] and 21 in [340]	Normally, a balance cut-off value should be defined before preprocessing. *700* is mostly used.	Each row only has one position with value '1'	$[700 * 20]$

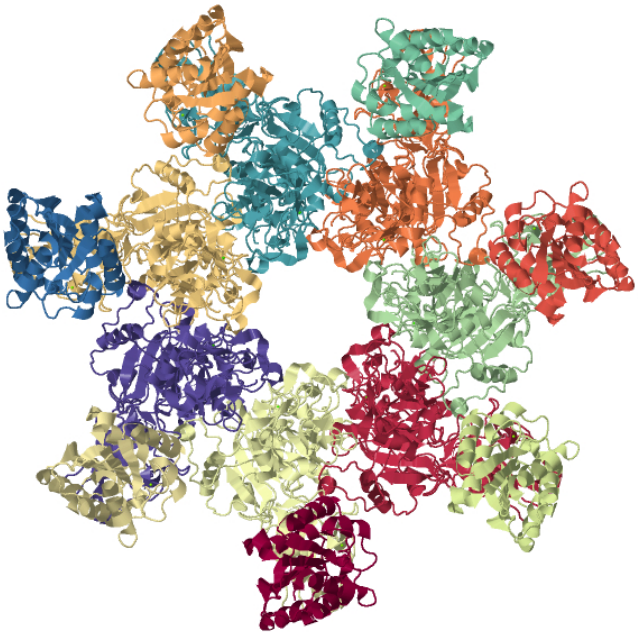


Figure 9.2: The 3D structure of the protective antigen (Uniprot ID: 'P13423')

9.2.2 Structural Information

Because protein sequences exhibit various lengths, those with < 50 amino acids are generally referred to as polypeptides and contain only primary level information. For secondary structure, folding forms common structures, such as $\alpha - helices$ and $\beta - sheets$ (from $\beta - strands$). Another structure is referred to as a random coli. Upon folding, a secondary structure subunit transforms into tertiary structure. For some proteins, their structure consist of more than one polypeptide, suggesting multiple tertiary structures. This context information is subsequently referred to as quaternary structure. The 3D structure for protective antigen (UniProt ID: 'P13423') is illustrated in Figure 9.2.

Because the wet lab is the site of protein-structure determination by X-ray crystallography, NMR spectroscopy or cryo-electron microscopy, these methods are extremely time-consuming and expensive. Therefore, an *ab initio* method based on computational modelling is a current focus of academic and industrial research. Only $< 0.5\%$ of all sequenced protein structures have solved structures according to the limitations of biological experiments methods [341].

Study of secondary structure prediction creates a dictionary of protein secondary structure (DSSP), which is better defined and clearer than tertiary structure and quaternary structure. Additionally, secondary structure can be analysed using efficient sequence information from primary structure. The secondary structure is predefined with three types of motifs: α-helix, β-strand and coli, allowing Q3 accuracy [340, 342–344]. Statistical models and machine learning methods have extensively improved Q3 predictive accuracy from 65% to 80%. Recently a more challenging problem targeting on eight-category prediction (Q8) defined in DSSP for secondary structure prediction was described. These eight categories describe the secondary structure based on additional elements: 3_{10}-helix, α-helix, π-helix, β-strand, β-bridge, β-turn, bend and loop/irregular [339, 345]. To achieve more accurate results on secondary structure, these methods require not only an efficient model but also sufficient feature representations from the sequence information. The involved models will be introduced in Section 4. The key challenge to predicting secondary structure involves prediction of those proteins having no close homologs and that have not experimentally verified 3D structures.

To achieve sufficient feature representations for secondary structure prediction, most studies introduced the protein-sequence information, amino acid profile information, local and global sequence information [340, 343, 346, 347]. Herein, the focus is firstly on the eight categories for secondary structure prediction task.

Figure 9.3 provides an example of a tertiary structure of the protective antigen protein (UniProt ID: P13423). Prediction for this level of structure normally involves homology modelling [348], which is also known as comparative modelling, where the main resulting candidate is derived from amino acid sequence alignment by mapping amino acids between different sequences. Introduction of homology modelling method into tertiary structure prediction allows evolutionary results to reveal proteins harboring similar amino acid sequences based on their shared similar tertiary structure to accomplish related biological function [349]. The structure information is a requisite for structural interaction networks, given that they provide atom level information.

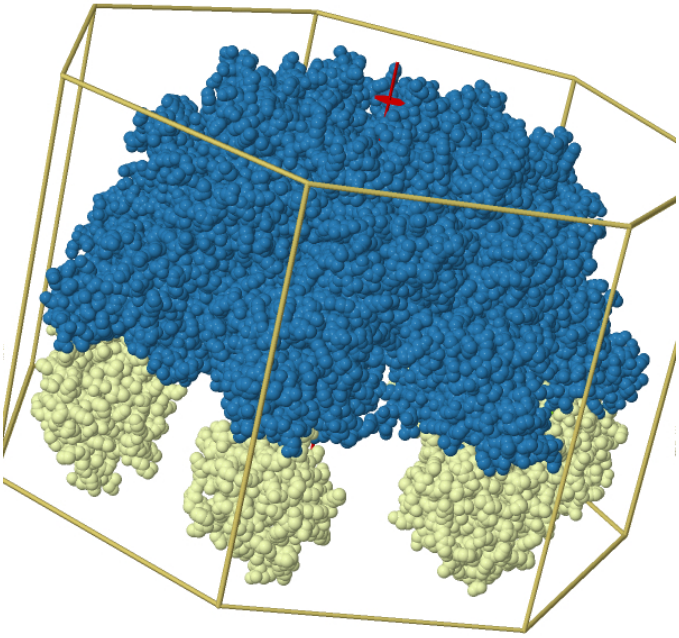


Figure 9.3: Tertiary Structure of Protein Protective Antigen (Uniprot ID 'P13423')

9.2.3 Domain-Domain Interactions

Given a protein sequence, protein domains are distinctive functional or structural subsegments. Most protein domains build independently stable and folded 3D structures, with which the domains combined into different arrangements to form a unique protein with different functions [350]. Therefore, binary PPI networks can be further considered at the domain level, especially when the interacting protein is large. Although most proteins consist of multiple domains, a pair of PPIs often involves only one pair of domain-domain interaction focusing on the actual binding site.

Domain-level interactions provide a global view of the binary PPIs network. For HP-PPIs investigations, this reveals interaction location or pathological interactions and can help facilitate drug-development targeting for infectious diseases. To acquire a comprehensive understanding of how domain interactions are mediated, the primary method involves analysis of individual interactions using experimentally determined 3D structures. However, this information is available for only a small fraction of proteins, indicating the domain-level PPI data not readily accessible.

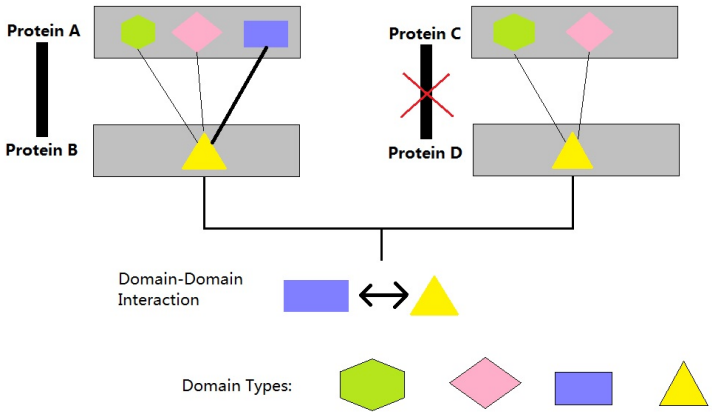


Figure 9.4: Domain-domain Interaction

There are several existing databases, including 3did [31] and iPfam [351], that provide domain-domain interactions by identifying these based on experimentally determined 3D structures. Other databases provide combined interactions, in which data are derived experimentally and the rest is computationally predicted. DOMINE [352] includes both 3D-structure-based and predicted domain-domain interactions and shows the predicted domain-domain interactions at three different levels, namely 'High', 'Middle' and 'Low'. Two primary methods, association [352] and maximum-likelihood estimation [353], are introduced in this domain-domain interaction-prediction task. The essential information utilised in these models includes domain information from protein sequence and binary PPI information.

To provide a general understanding of domain-domain interactions associated with binary PPIs, Figure 9.4 shows a basic diagram for domain-domain-interaction prediction [354]. 'Protein A' interacts with 'Protein B' while 'Protein C' does not interact with 'Protein D'. Several different domains types are identified using the related databases. Mostly, Protein Data Bank (PDB) [355] is applied as suggested. Next, the differences between these two groups of domain-domain relationships to identify the interacting domains between two different proteins will be compared.

9.3 RELATED DATABASES

Ranging from protein-sequence information to their structure data, several different databases are currently available and well maintained, including host-pathogen PPI databases, structure databases, protein families and domain databases, and also domain-domain-interactions databases.

9.3.1 Host-Pathogen Interactions Databases

Although several different standardized formats for the host-pathogen PPIs are published by different organizations, these databases contain the most important binary information for HP-PPIs researches. Some popular repositories are initially built by universities, such as HPRD by Johns Hopkins University and the Institute of Bioinformatics, PATRIC by University of Chicago, PHISTO by Boğaziçi University, VirHostNet by Université de Lyon. Highly credible positive HP-PPIs pairs are manually recorded in these systems and updated periodically. The details of several selected popular databases are listed in Table 9.2. For a complete survey regarding the HPI databases, please refer to Chapter. 3.

Table 9.2: Partial Host-Pathogens PPIs Database

HP-PPI Database	Contents
HPRD [196]	A database manually extracted from literature, is built by Johns Hopkins University, includes more than 39,000 interaction pairs.
BIND [184]	It belongs to Biomolecular Object Network Databank, and is maintained by University of Toronto. It provides more than 200,000 interaction pairs.
DIP [178]	It includes several sources, i.e. Yeast Protein Database, Kyoto Encyclopedia of Genes and Genomes.
PATRIC [160]	Continuously updated by University of Chicago, this database is built upon a combination of several public repositories.
PHISTO [161]	Currently it stores over 23,000 interaction pairs and these data are imported from several PPI databases using PSICQUIC tool.

9.3.2 Structure Databases

The Protein Data Bank (PDB) [355] is the primary database housing structural information for proteins and is managed by the worldwide Protein Data Bank (wwPDB) international collaboration. The PDB contains all experimentally determined protein structures ranging from different resolutions and detection methods.

The PDB is currently updated weekly and has its own file format standard, which is strictly defined to provide protein and nucleic acid structure details. A standard PDB file should contains atomic coordinates, observed sidechain rotamer, secondary structure assignments and atomic connectivity information. Apart from the critical information, abbreviation content about the corresponding literatures is also mandatory in PDB file, which is listed as Header. Several other specific columns include the ID number, date for publication, obsolete status, details about the related experimental methodology, molecular components of the complexes, the source of the complexes, the experimental method used to determine the structure, the authors, modification and revocation records, and related literature, the maximum resolution, and other statistics.

A simple example of the protective antigen protein (UniProt ID: P13423) using PyMOL [356, 357] is shown in Figure 9.3. It requires substantial time and effort to acquire an experimentally determined protein structure, and currently, not every protein has its corresponding structural information available. Determination of this information for these proteins is critical for building a SIN.

9.3.3 Protein Families and Domain Databases

As an important database of protein domains and families, Pfam provides a complete map for protein domains and families [358, 359]. It is regularly updated, with the latest version being Pfam 31.0 released in March 2017 for instance and containing $>$16,712 protein families.

Although amino acids are the elements comprising a protein sequence, functions occur in multi-sequential regions which are called domains. Identifying these domains provides

details and insights regarding the functional mechanism of the protein.

Structural information allows bond information detailing interactions between proteins, which is more concrete than binary HP-PPIs network provided in HP-PPIs databases. Therefore, iPfam is used in SIN studies to identify domain-domain interactions between proteins [351]. iPfam was developed by Howard Hughes Medical Institute, and currently harbors $>$ 9,500 domain-domain interactions.

iPfam is based on two continuously updating databases, PDB and Pfam, both of which are well established for their 3D structure and domain-information purposes. Most of the structural information in the PDB also contains multiple domains. The 3did is another domain-domain interaction databases for 3D-interacting domains between proteins, and is a collection of protein interactions from which high-resolution 3D structures are known [31, 360].

By using iPfam and 3did to achieve domain-level resolution of HP-PPIs, SIN considers proteins in their precise spatial relationships by layering domain-domain interactions on top of the conventional PPI networks. As protein-sequence information accumulates at a staggering rate, these data depict its characteristics with high volume, high velocity, high variety, high value and high veracity (5V). This, along with big-data analytics, including machine learning technologies, allows addressing structural and domain-domain-interaction prediction problems. The following sections will introduce the related computational models or methods for SIN construction, including machine learning methodologies.

9.4 Computational Models

SIN is designed to layer high-confidence 3D models on top of PPIs. Before layering the structural information on the binary HP-PPIs network, the structural information of corresponding proteins is requisite. However, only a few proteins have experimentally determined structure, specifically with high-resolution scale. Therefore, herein we present related studies outlining structure prediction and domain-domain-interactions prediction.

The review in this section is considered as an important step in jointly studying protein structural information while supplementing the structural interaction network.

9.4.1 Bayesian Statistics

The earliest studies on protein secondary structure prediction mainly focused on the use of Bayesian statistics [361–363]. Basically, Bayesian statistics describes this problem as following Equa. 9.1.

$$I(S;R) = log[\frac{P(S|R)}{P(S)}] \tag{9.1}$$

where $P(S|R)$ is the conditional probability for observing a conformation S, when a residue (amino acid) R is present, and $P(S)$ is the probability of observing S. According to the conditional probabilities definition, $P(S|R) = P(S,R)/P(R)$. $P(S,R)$ is the joint probability of S and R. Through the use of Eq. (1), an estimation of $I(S;R)$ from a database of known protein sequences and corresponding secondary structures can be achieved.

Specifically, a previous study [362] showed that the the Garnier-Osguthorpe-Robson (GOR) method based on information theory used a 17-amino-acid sequence window to extract properties from protein sequences. The GOR method presented the observed frequencies of singletons, then in pairs of residues on a local sequence of 17 residues to build the Bayesian model, followed by estimation of the probabilities for the Q3 structures. This method increased the accuracy from 55% to 64.4%. Later, in [363], combined with information theory, GOR V algorithm projects the known twenty amino acids types for each specific secondary structure to achieve a Q3 accuracy of 73.5%.

9.4.2 Support Vector Machine (SVM)

Using SVMs to predict protein secondary structure was firstly introduced in 2001 [364], with the first SVM proposed in 1995 [253]. It is not the first machine learning approach used for protein secondary structure prediction, yet by then, it achieved the best performance overall on Q3 task.

Similar to earlier researches using neural network based methods [346], the encoding

scheme for the input layer is called a local-coding scheme and denotes every amino acid with a 21-dimensional orthogonal binary vector as Equa. 9.4.2.

$$(1,0,...,0) or (0,1,...,0), etc$$

In the output layer, the Q3 task was first considered as a binary classifier later combined into a tertiary classifier.

A previous study [364] considered the SVM as a superior model based on its ability to effectively avoid overfitting and to handle large feature spaces. In details, the authors [364] selected the radial basis function as the kernel function to train the SVM, resulting in a Q3 task of 73.5%.

9.4.3 Random Forests

Apart from predicting secondary structure, domain-domain interaction is also critical to the SIN. The random forest model was introduced to build multi-classifiers to determine a decision for a dataset with 1050-dimensional features [365]. Additionally, another study [366] showed an ensemble model of random forests and SVMs were able to predict the domain-interacting sites.

Derived from decision trees model, random forest leverages the power of randomisation to increase model performance [255, 367]. It is able to deal with imbalanced data problems via the voting mechanism while its random feature selection benefits the model in case of high-dimensional data.

9.4.4 Artificial Neural Networks

To the best of our knowledge, artificial neural networks were first introduced in protein secondary structure prediction using a fully connected three-layer network in [346], with a learning algorithm involving back propagation. Later, the authors of [368] used a two-tier architecture to deploy neural networks for prediction; however, the improvement in Q3 accuracy has since stalled.

Recently, Q8 accuracy has been the focus of academia and industry, aiming to apply deep learning techniques to improve performance. In [369], probabilistic graphical models, which combine conditional neural fields (CNFs) with neural network, were deployed to improve Q8 accuracy. The features are extracted from position-specific score matrix (PSSM) and the physico-chemical properties of the amino acids. Both the complex relationship between sequence and secondary structure information, and the interdependency relationship among secondary structure types of adjacent amino acids were studied using the CNFs model [369].

Generative stochastic networks (GSNs) were utilised to learn a generative model of data distribution without explicitly specifying a probabilistic graphical model [339], . Specifically, this supervised extension of GSNs is deployed via learning a Markov chain to sample from a conditional distribution for training on a protein structure prediction task. This model was presented with deep learning techniques to tackle Q8 problem for protein secondary structure prediction. The empirical design for the data preprocessing step involved choosing 700 lengths as the cut-off threshold to balance the efficiency and coverage of protein sequence. The main features extracted included the evolutionary information (PSSM feature) and the sequence information (one-hot binary vector feature). The model achieved 66.4% accuracy on Q8 problem.

The most recent result on Q8 accuracy task was reported in [340], which proposed a deep convolutional and recurrent neural network. The feature encoding the protein sequence remained partially similar to the local-coding scheme. In this network model, a feature embedding layer was deployed to map sequence information and profile feature (by PSI-BLAST) to a denser matrix. Multiple convolutional neural network layers and stacked bidirectional relational neural network layers were included to learn both local context information and global context information from the denser matrix. Fully connected and softmax layers were layered on the top of the model to build the classifier for the prediction task.

Considering the different properties of protein structure, an iterative use of predicted

features, including the backbone angles and dihedrals based on C_α atoms, improves secondary structure prediction accuracy [370]. Stacked sparse auto-encoders with three hidden layers were introduced. The hidden layers were all with 150 neuron nodes. The method achieved an accuracy 80.8% in secondary structure prediction in the recent CASP targets[a] [370].

Various models have been discussed in this section; however, our goal is to stack these different data types atop the binary HP-PPIs network to achieve structural principles analysis. In the following section, the structural interaction network will be discussed.

9.5 STRUCTURAL INTERACTION NETWORK

Since principles analysis of protein interactions between host and pathogens still remains poorly understood, an ensemble network of binary HP-PPIs networks and structural information would provide an efficient option for mining this knowledge using a systems biology approach.

A previous study used 3,949 genes, 62,663 mutations and 3,453 associated disorders for analysis using a 3D structurally resolved human interactome network [371]. By integrating data from iPfam, 3did and the Human Gene Mutation Database (HGMD) [372], a high-quality binary PPIs network with the atomic-resolution interfaces was successfully built [371], providing key insights to in-frame mutations, locations, and disease specificity for different mutations in the same gene, which had not been possible to be acquired on a low-resolution network. The original interaction network obtained from literature-curated databases [371] contained 82,823 pairs; however, after filtering out the proteins without experimentally determined structures, only 4,222 structurally resolved interactions between 2,816 proteins remained. To build a structural interaction network still requires more efforts on experimental determination of a structure or computational prediction, because only a tiny fraction of these binary PPIs can be analysed with their corresponding structure information.

[a]http://predictioncenter.org/casp11/index.cgi

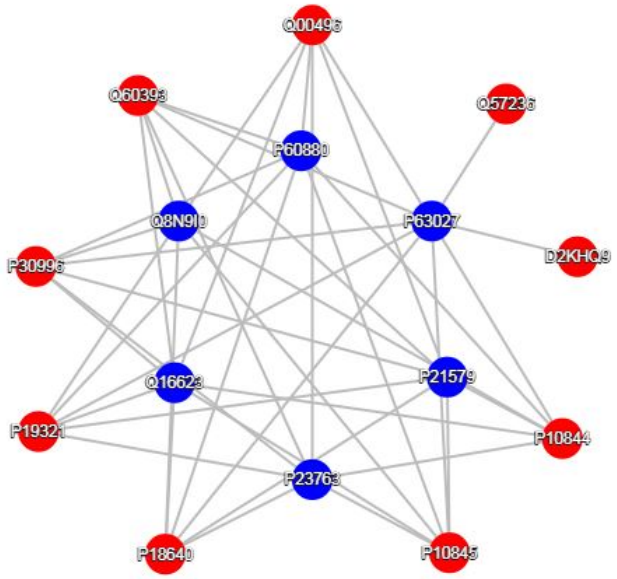


Figure 9.5: Binary PPI Network of Clostridium botulinum

9.5.1 Construction of SIN

Figure 9.5 shows six primary human proteins interacting with nine *Clostridium botulinum* proteins, resulting in 44 HP-PPIs connections derived from the PHISTO database. These interactions are considered as exogenous interactions. To further analyse interactions from the PPI network, this information with structural information is embedded. There are two classes of protein-protein interaction in physical interactions: interactions mediated by two domains and that between short motifs and domains.

It can be observed that, several possible structural principles analyses were obtained within the human-virus protein-protein interaction network [28]. The SIN approach in human-virus PPIs network reveals atomic resolution, mechanistic patterns, and allows systematic comparison with human endogenous interactions.

Figure 9.6 shows an example detailing how to layer the structure and domain-domain interaction information on top of the binary PPIs network [28, 264].

Figure 9.6 reveals the overlapping interfaces between the 'Pathogen Protein-Host Protein2' and the 'Host Protein3-Host Protein2', which determine the interaction. This type of information could not be observed in the binary PPI network. Further analysis

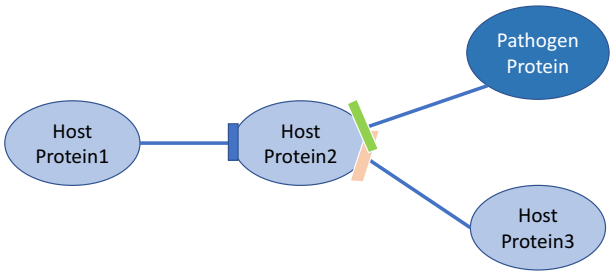


Figure 9.6: Structure Interaction Network [264]

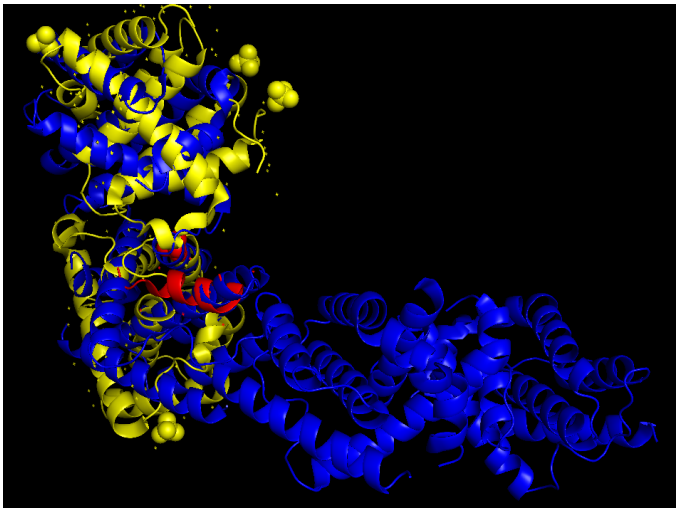


Figure 9.7: The Overlapping Structure Interaction: The red string is the human protein Beclin-1, which is annotated with *5EFM* as its PDB id. The compound (in yellow), which is interacted by human protein 'Beclin-1' and Gamma Herpesvirus protein 'v-Bcl2', is associated with the compound (in blue) by human protein 'Beclin-1' and human protein 'BCL-XL'. The 3D structure of yellow compound can be fetched by PDB id *4MI8* while the blue is *2P1L* [373].

revealed that 'Pathogen protein' is mimicking the action of 'Host Protein3'. Layering the 3D structural information to illustrate the details of the protein interaction allows derivation of two different classes of protein interactions (Figure 9.7 and Figure 9.8) [373]. The results are generated by PyMOL [357].

The illustration examples present the non-overlapping protein-protein interactions by 3D structures *1F5Q-1BUH*, and overlapping protein-protein interaction by *4MI8-2P1L* [373]. Here, *1F5Q*, *1BUH*, *4MI8* and *2P1L* are their PDB id.

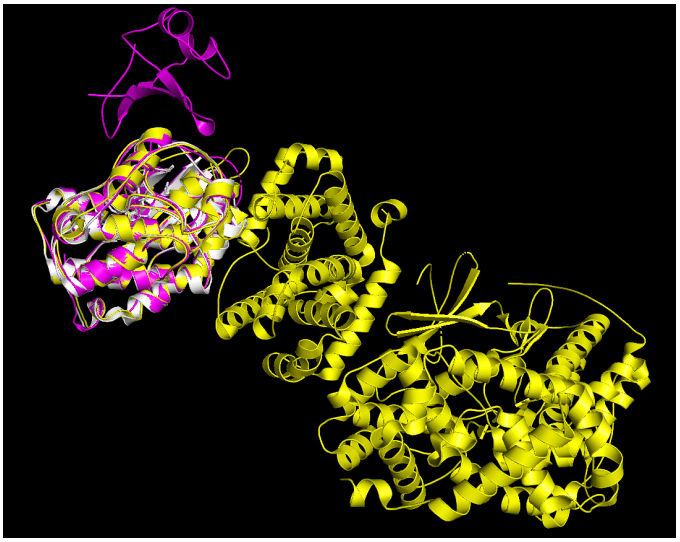


Figure 9.8: The Non-overlapping Structure Interaction: The interaction is linked by the human protein 'CDK2'. The PDB id is *5MHQ*. The yellow compound is the interaction between Gama Herpervirus 'Cyclin' and human protein 'CDK2'. The purple compound is by human protein 'CKS1' and 'CDK2' [373].

9.5.2 Highlights of SIN

The host-pathogen PPI networks provide specific pathogen protein functions and the global analyses on this network help revealing critical proteins in the networks [264]. Although Figure 9.7 provides essential mappings via the overlapping interfaces, annotating the experimental HP-PPIs networks with 3D structural information will provide further information, because the PPIs can be combined between two globular domains and also between one short linear motif (a short functional segment considered on secondary structure) and globular domains. Superimposing structures of the HP-PPIs can help to visually reveal the details.

Several methods to assemble structural information with binary HP-PPI network include:

- Using only the experimentally determined structural information. Both proteins in the HP-PPIs network could be mapped along with the determined structural information;

- Using both the experimentally determined and computationally predicted structural

information. One of the proteins in the HP-PPIs could not be mapped with its determined structural information;

- Using only the computationally inferred structural information. Both proteins in the HP-PPIs could not be mapped with its determined structural information. The homology modelling method is widely used for searching for homologous proteins with having determined structure according to the BLAST E-value.

Computationally predicted structural information mainly comes from homology modelling, which is widely used in bioinformatics, provided that protein structure and function are primarily determined according to their sequence information [28].

Typically, for host-pathogen protein-protein interactions, most researches hypothesised that imitating the binding activities between proteins would allow insight into primary mechanism associated with infections. Given a SIN, there are several types of statistics data that may help us propose and support this hypothesis. As a specific example between virus and host-PPI networks, a previous study [28] analysed the exogenous and endogenous interactions in the human-virus SIN model.

Meanwhile, the overlapping ratio of protein interactions involved in exogenous interface to those involved in endogenous interface indicates potential infectious targets, although the mapping of endogenous interfaces is not guaranteed to be complete [28].

To achieve a better understanding of the mimicry mechanism that possibly explains virus-infectious procedure, similarity statistical analysis can be performed according to z-score [374] and E-value [236] levels. Since the mimicry action occurs between host protein and pathogen protein, similarity statistics might help elucidate potential activities.

Overall, SIN, combined with binary protein-protein interactions, has many advantages for precise analysis based on statistics associated with 3D structure and domain information.

9.6 Challenges

While the boom of big data analytics appears promising, when dealing with both the structural information and domain-domain interactions, there still remains several challenges in the areas of SIN and HP-PPIs network development.

9.6.1 Feasible and Efficient Feature Representation

For computational models, especially protein sequences, feature representation remains a challenging topic. Various methods for feature representation currently exist [49, 327, 336, 338, 344–347]. Previous studies have indicated that, various representational methods yielded different performances across several species, although additional protein sequence information is being experimentally generated.

Additional models based on deep learning techniques present a more effective framework for learning from big data sets. The automatic feature extraction process could be a promising option for protein sequence research. For example, in previous chapters, the unsupervised learning model, which applies the stacked denoising autoencoder as the model to extract high-level feature for model learning, has shown a promising vision [327]. The result showed a potential direction for introducing deep learning neural networks.

Prior to inputting data into learning models, several traditional feature representation methods, including one-hot vector method, PSSM feature, and other statistic methods shown in Table 9.1, were widely used. Additionally, deep learning techniques are also first introduced in protein secondary structure prediction [339, 340] and HP-PPIs prediction tasks [327]. In terms of feature representation, deep learning techniques could harness the power of high-dimensional data in large volumes, enabling acquisition of large volumes of feature information to further improve model performance.

9.6.2 Imbalanced Data

Another challenging issue is the imbalanced ratio among different classes of the structural information, such as the eight categories of protein secondary structure. For structure prediction, domain-domain interaction and host-pathogen protein-protein interaction problems, the imbalanced ratio between different classes is important in improving model performance.

The ratio of non-interface interactions to interface interactions is about 9:1 [365]. In structure prediction task, the ratios in both Q3 and Q8 tasks are also different and imbalanced between different protein families. Specifically, for Q8 tasks, some structures are barely observable in the protein structures. In a previous study, the interacting pairs and non-interacting pairs were defined with 1:100 ratio, which is a highly skewed number [49].

With the continuous expansion and availability of structural information and domain data, the issues involving imbalanced data biological areas intensifies.

9.7 Summary

This chapter is designed as a survey describing the building of structural interaction network (SIN) for host-pathogen protein-protein interactions to analyse the resulting network using a systems biology approach. This chapter is focused on structural information and also SIN analysis. Several multidisciplinary and interdisciplinary areas were reviewed, including protein feature representation, protein structure prediction, domain-domain interaction prediction and machine learning methods applied for these prediction tasks.

For HP-PPIs researches, building SIN using atomic level data can provide insights into high-resolution interactions based on protein structures and offer high-quality analyses of interactions targeting infectious mechanisms. As a survey result of the state-of-the-art methods, multiple areas still need to be addressed in this research direction. It is anticipated that, this survey will benefit future proteomics studies, as well as the

computational method design.

Chapter 10

CONCLUSION AND FUTURE WORK

This chapter will summarise the contributions of this thesis, and it will then discuss the potential directions of future work.

10.1 Contributions

The main focus of this thesis is to deliver a comprehensive study of host-pathogen protein-protein interactions, particularly there has been little effort on delivering a systematic work of the computational models for the prediction task of host-pathogen protein-protein interactions. Although great achievements have been made in biology and public health areas around the world, it is still very important, also inspiring, to find novel methods other than traditional purely medical and biological lab experiments. The traditional methods are still expensive and slow-going, to uncover or predict mechanisms of viral and bacterial infectious diseases. Considering the recent panic caused by the outbreak of SARS-CoV-2, it has again brought great attention to viruses' invading mechanism. Little was known about the 'novel' coronavirus so that little therapeutic plan was ready at once to handle it, even though similar viruses, such as SARS-CoV/MERS-CoV and HIV/HPV, have been extensively studied for many years. On the other hand, benefiting from the advanced development of high-throughput experimental and sequencing technologies, increasingly tremendous and complicated omics data has been accumulated, which poses great opportunities for computational biologist to find clues from protein interactions and

omics data for carrying out system biology study. The knowledge learnt and shared from such data can improve the understanding of the diseases and expedite the development of effective therapeutic measures.

Since the study of HP-PPIs is critical to the understanding of infecitous diseases and presents great values for the mechanism study, this thesis focus on building a delibarate computational framework for discovery of HP-PPIs, which solicits an in-depth research of HP-PPIs resources as well as feasible computational models. Thus, the thesis has studied these two aspects by designing four distinct goals: 1) reviewing the host-pathogen interactions databases published in the past decades in a comprehensive way; 2) evaluateing machine learning-based computational models for discovery of host-pathogen protein-protein interactions in a systematic manner; 3) developing novel machine learning-based computational framework to better improve the discovery performance of host-pathogen protein-protein interactions; 4) reviewing the state-of-the-art of the SIN reconstruction, which could offer an atomic resolution analysis on host-pathogen interactions. In details, following conclusions reports the achieved tasks.

In Chapter. 3, a comprehensive literature review related to host-pathogen interactions resources, which are collectively published in last two decades, is conducted. The resources reviewed in this chapter cover a wide range of topics of host-pathogen interactions in Chapter.3.1 and Chapter.3.2. Furthermore, several standards and tools published in the aim of facilitating proteomics research and development are reviewed in Chapter.3.3. Later on, a brief statistic report of the curated human-pathogen interactions database and the primary categories of bioinformatics tasks of host-pathogen interactions study are elaborated in Chapter.3.4 and Chapter.3.5 repectively, which give the details of the current status of human-pathogen interactions resources by collectively analysing the selected databases.

In Chapter. 4, an systematic evaluation of the predictions task for HP-PPIs is conducted. Different computational methods are included for evaluation, among which we have presented a wide and deep review on currently available resources and computational

tools. As noted in the literature review in Chapter. 4.2 to evaluate the computational tools developed for prediction tasks of HP-PPIs, a dedicated data curation process is implemented and a computational pipeline for HP-PPIs studies is summarized in Chapter. 4.3, which includes numerous sequence feature representation algorithms and machine learning models. Also, the computational methods concerning HP-PPIs from literature are also elaborated. Given the evaluation of HP-PPIs, we have strived to quantitatively determine the impacts caused by different ratios of benchmark datasets, different feature representation algorithms and different machine learning models. The experimental results in Chapter. 4.4 indicate that, to better utilise machine learning models and harness the power of accumulated protein interaction data, a more robust and more powerful computational model is required to achieve better performance across different HP-PPI prediction tasks.

In Chapter. 5, a novel framework for HP-PPIs prediction based on *H*eterogeneous *I*nformation *M*ining and *E*nsembling (HIME) process to effectively learn from the interaction data is proposed. Since a robust performance of the prediction model is desired to achieve for different HPI systems, HIME model leverages the abundant information through mining the heterogeneous information of sequence data, and the details are included in Chapter. 5.3. The horizontal ensemble procedure with heterogeneous information has greatly exerted the base learners to boost the performance in the prediction task. The performances are evaluated on different datasets, which has indicated HIME model outperforms the others in Chapter. 5.4.

In Chapter. 6, given the foundation of the systematic review in Chapter. 4, a novel two-layer machine learning model, namely APEX2S, is proposed to deal with the imbalanced issue. In this chapter, the HP-PPIs prediction problem is further studied and a detailed investigation concerning the multi-omics data for HP-PPIs in a broader scale is discussed firstly in Chapter 6.1. Presented by the abundant multi-omics data, a comprehensive and practical workflow is subsequently designed in Chapter 6.2, which has elaborated the usage of machine learning techniques in a preliminary stage. More importantly

for Chapter. 6, a novel two-layer model APEX2S for the prediction task of HP-PPIs is presented in Chapter 6.3. In Chapter 6.4, a practice of the model in the dataset concerning PPIs between human and Shigella infections pathogen is reported to evaluate the performance of the computational models, which include the traditional machine learning models and the two-layer APEX2S model. The comparison result has indicated the better prediction ability and higher efficiency of APEX2S model

In Chapter. 7, the deep learning model is introduced to build a novel machine learning model for the prediction task of HP-PPIs. Particularly, a bidirectional LSTM-based model is presented for the prediction task, which demonstrates a more effective performance in comparison with the others. In details, a multi-channel feature representation algorithm, which is based on tree-based feature selection algorithm and synthetic minority over-sampling technique (SMOTE), is firstly desined. Later, we discuss the bidirectional LSTM model. Due to the scenario of imbalanced issue, the focal loss function is subsequently employed as a novel cost function for Bi-LSTM model. The prediction performance of HP-PPIs dataset has indicated that Bi-LSTM-based model has obtained the best results.

In Chapter. 8, we report the investigation of the host-pathogen protein-protein interactions with an unsupervised deep learning model based on stacked denoising autoencoders. A SdA-based deep learning model for HP-PPIs datasets is presented and the comparison of the SdA model with other models indicated its superiority for this application. From the evaluation result of this chapter, the unsupervised SdA model is optimal for the highly skewed and big datasets and is better at feature representation comparing with other models. The results suggested that, the deep learning model is capable of dealing with big HP-PPIs datasets.

In Chapter. 9, we have further surveyed the main methodologies and algorithms for the reconstruction and modelling of the structural-interaction networks (SINs) of host-pathogen protein-protein interactions (HP-PPIs), regarding how the protein domains interact with each other to constitute a SIN. This chapter is focused on structural infor-

mation and also SIN analysis. Several multidisciplinary and interdisciplinary areas were reviewed, including protein feature representation, protein structure prediction, domain-domain interaction prediction and machine learning methods applied for these prediction tasks. As a survey outcome of the state-of-the-art methods, multiple areas will need to be addressed in this research direction. It is anticipated that, this survey will benefit our future work, for the future proteomics studies and the computational method design.

10.2 Future Work

In light of the research contents of this thesis, the following research directions could be further explored in the future:

- Advanced Deep Learning Models: Deep learning models have been studied and two models based on Bi-LSTM model and SdA model are proposed in Chapter. 7 and 8. The results have shown benefits by constructing deeper model for HP-PPIs task. Meanwhile, advanced deep learning models, including the adversarial model and attention model, also demonstrate capbilities of learning raw data automatically in other research areas, such as computer vision and natural language processing. The advanced deep learning model will be investigated in the future work. First of all, the adversarial model aims to explore the data by generative model and discriminative model. It could better utilise the unannotated data which has widely exisited in proteomics area. Secondly, the attention model builds the model with human-like attention mechanism. It can be better integrated in the sequence feature representation algorithms to improve the model performance, for the reason that the protein-protein interactions indeed occur between domains which is a functional segmentation of the protein sequence. Thus, the advanced deep learning models are expected to decipher the code of protein information in a better way and thus to deliver more effective and efficient frameworks.

- Structural-Interaction Networks (SINs): A structural interaction etwork is of cru-

cial significance for understanding the protein-protein interaction network at the systems level. With the atomic level resolution, the structural information of the HP-PPI network will be further interrogated. Based on Chapter. 9, the SIN will be investigated based on various heterogeneous sources of structure data in the future work. This network will be an essential component for systems biology to better discover the biological functions and infectious mechanisms that underlie many infectious diseases caused by pathogens. To reconstruct the structural interaction network, it is important to annotate the protein interactions network with 3D structural information, along with the protein family and domain data. However, it still requires adequate efforts to be expended on the experimental determination of structure or computational prediction, because only a tiny fraction of these binary PPIs can be analysed with their structure information. Currently, there are several feasible methods to assemble structural information with host-pathogen protein interaction network, including: 1) using the experimentally determined structural information only; 2) using both the experimentally determined and computationally inferred structural information; 3) using the computationally inferred structural information only. In the future work, both the computationally inferred method and experimentally determined method to assemble structural information will be investigated.

- Host-pathogen Interactions Network: Different protein interaction networks typically exhibit different characters due to the nodes representing proteins and edges connecting proteins that can interact. For host-pathogen interactions network, it plays a central role in biology function which regulates the mechanisms related to healthy and diseased states in organisms. Meanwhile, there have been several studies focusing on protein interaction networks alignment, by either local alignment or global alignment approach leading to new discoveries of protein complexes, infectious pathways and functional orthologs. Herein, how to distill the alignment task of protein interaction networks between host and pathogen could generate sub-

stantial value to transfer knowledge between species. Most alignments of protein interaction networks are achieved at pairwise level since the main approaches are built upon either local network alignment or global network alignment. In the future work, we anticipate to leverage the heterogeneous information to align the pairwise level protein interaction network. From computational perspective, the evolutionary algorithm is of great advantage to solve the similarity calculation problem between interaction networks, which is NP-complete, in an efficient. It is inspired by biological evolution and the computational complexity largely depend on the fitness approximation method. It is expected that these future work could benefit the computation of alignment task of HPI network and generate more knowledge.

BIBLIOGRAPHY

(1) C. S. Greene, J. Tan, M. Ung, J. H. Moore and C. Cheng, "Big data bioinformatics", *Journal of Cellular Physiology*, 2014, **229**, 1896–1900.

(2) N. Savage, "Bioinformatics: big data versus the big C", *Nature*, 2014, **509**, S66–S67.

(3) S. Min, B. Lee and S. Yoon, "Deep learning in bioinformatics", *Briefings in bioinformatics*, 2017, **18**, 851–869.

(4) S. Orchard, H. Hermjakob and R. Apweiler, "Proteomics and data standardisation", *Drug Discovery Today: Biosilico*, 2004, **3**, 91–93.

(5) U. Consortium, "UniProt: a worldwide hub of protein knowledge", *Nucleic Acids Research*, 2018, **47**, D506–D515.

(6) M. Vaudel, K. Verheggen, A. Csordas, H. Ræder, F. S. Berven, L. Martens, J. A. Vizcaíno and H. Barsnes, "Exploring the potential of public proteomics data", *Proteomics*, 2016, **16**, 214–225.

(7) C. G. A. R. Network et al., "Comprehensive genomic characterization defines human glioblastoma genes and core pathways", *Nature*, 2008, **455**, 1061–1068.

(8) A. Alyass, M. Turcotte and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities", *BMC Medical Genomics*, 2015, **8**, 33.

(9) V. Marx, "Biology: The big challenges of big data", *Nature*, 2013, **498**, 255–260.

(10) R. Kumar and B. Nanduri, "HPIDB–a unified resource for host-pathogen interactions", *BMC Bioinformatics*, 2010, **11**, S16.

(11) J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, "Predicting protein-protein interactions based only on sequences information", *Proc. Natl. Acad. Sci.*, 2007, **104**, 4337–4341.

(12) Wikipedia, *Amino acid - Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Amino_acid, [Online; accessed 21-April-2016], 2016.

(13) Wikipedia, *Protein - Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/Protein, [Online; accessed 21-April-2016], 2016.

(14) S. Orchard, "Proteomics: An introduction to EMBL-EBI resources".

(15) A. Pandey and M. Mann, "Proteomics to study genes and genomes", *Nature*, 2000, **405**, 837.

(16) P. R. Graves and T. A. Haystead, "Molecular biologist's guide to proteomics", *Microbiol. Mol. Biol. Rev.*, 2002, **66**, 39–63.

(17) M. Tyers and M. Mann, "From genomics to proteomics", *Nature*, 2003, **422**, 193.

(18) Wikipedia, *Protein–protein interaction — Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/w/index.php?title=Protein%E2%80%93protein%20interaction&oldid=931773323, [Online; accessed 30-December-2019], 2019.

(19) J. De Las Rivas and C. Fontanillo, "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks", *PLoS Computational Biology*, 2010, **6**, e1000807.

(20) M. K. Ganapathiraju, M. Thahir, A. Handen, S. N. Sarkar, R. A. Sweet, V. L. Nimgaonkar, C. E. Loscher, E. M. Bauer and S. Chaparala, "Schizophrenia interactome with 504 novel protein–protein interactions", *npj Schizophrenia*, 2016, **2**, 16012.

(21) Z.-H. You, S. Li, X. Gao, X. Luo and Z. Ji, "Large-scale protein-protein interactions detection by integrating big biosensing data with computational model", *BioMed research international*, 2014, **2014**.

(22) Y. Qi, O. Tastan, J. G. Carbonell, J. Klein-Seetharaman and J. Weston, "Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins", *Bioinformatics*, 2010, **26**, i645–i652.

(23) M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, "Multitask learning for host–pathogen protein interactions", *Bioinformatics*, 2013, **29**, 217–226.

(24) O. Krishnadev and N. Srinivasan, "Prediction of protein–protein interactions between human host and a pathogen and its application to three pathogenic bacteria", *International journal of biological macromolecules*, 2011, **48**, 613–619.

(25) M. S. Scott and G. J. Barton, "Probabilistic prediction and ranking of human protein-protein interactions", *BMC Bioinformatics*, 2007, **8**, 239.

(26) A. K. Halder, P. Dutta, M. Kundu, S. Basu and M. Nasipuri, "Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions", *Briefings in functional genomics*, 2017, **17**, 381–391.

(27) K. Asehnoune, J. Villadangos and R. Hotchkiss, "Understanding host–pathogen interaction", *Intensive care medicine*, 2016, **42**, 2084–2086.

(28) E. A. Franzosa and Y. Xia, "Structural principles within the human-virus protein-protein interaction network", *Proceedings of the National Academy of Sciences*, 2011, **108**, 10538–10543.

(29) H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The protein data bank", *Nucleic Acids Research*, 2000, **28**, 235–242.

(30) R. D. Finn, M. Marshall and A. Bateman, "iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions", *Bioinformatics*, 2005, **21**, 410–412.

(31) A. Stein, A. Panjkovich and P. Aloy, "3did Update: domain–domain and peptide-mediated interactions of known 3D structure", *Nucleic Acids Research*, 2009, **37**, D300–D304.

(32) E. Nourani, F. Khunjush and S. Durmuş, "Computational approaches for prediction of pathogen-host protein-protein interactions", *Frontiers in Microbiology*, 2015, **6**, 94.

(33) Y. Guo, L. Yu, Z. Wen and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences", *Nucleic Acids Research*, 2008, **36**, 3025–3030.

(34) G. Cui, C. Fang and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model", *BMC Bioinformatics*, 2012, **13**, S5.

(35) G. Xiaolong, J. Yan and Q. Lu, "Study of Decision Tree in the Application of Predicting Protein-protein Interactions", *Journal of Biomedical Engineering*, 2013, **5**, 009.

(36) M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald and E. Muharemagic, "Deep learning applications and challenges in big data analytics", *Journal of Big Data*, 2015, **2**, 1–21.

(37) Y. Yan, M. Chen, M.-L. Shyu and S.-C. Chen, Multimedia(ISM 2015), 2015 IEEE International Conference on, 2015, pp. 483–488.

(38) D. Field, S.-A. Sansone, A. Collis, T. Booth, P. Dukes, S. K. Gregurick, K. Kennedy, P. Kolar, E. Kolker, M. Maxon et al., "'Omics data sharing", *Science*, 2009, **326**, 234–236.

(39) R. P. Horgan and L. C. Kenny, "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics", *The Obstetrician & Gynaecologist*, 2011, **13**, 189–195.

(40) W. H. Organization et al., "Genomics and world health: Report of the Advisory Committee on Health Research", 2002, 1–241.

(41) B. Berger, J. Peng and M. Singh, "Computational solutions for omics data", *Nature Reviews Genetics*, 2013, **14**, 333–346.

(42) D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa and J. Tegnér, "Data integration in the era of omics: current and future challenges", *BMC Systems Biology*, 2014, **8**, I1.

(43) D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and DNA arrays", *Nature*, 2000, **405**, 827–836.

(44) E. S. Lander, "The new genomics: global views of biology", *Science*, 1996, **274**, 536.

(45) F. S. Collins, E. D. Green, A. E. Guttmacher and M. S. Guyer, "A vision for the future of genomics research", *Nature*, 2003, **422**, 835–847.

(46) *Human Genome Project Completion: Frequently Asked Questions*, https://www.genome.gov/11006943/, Accessed: 2016-08-15.

(47) Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha and G. E. Robinson, "Big data: astronomical or genomical?", *PLoS Biol*, 2015, **13**, e1002195.

(48) M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions", *Nature Reviews Genetics*, 2015, **16**, 85–97.

(49) H. Chen, J. Shen, L. Wang and J. Song, Big Data (BigData Congress), 2016 IEEE International Congress on, 2016, pp. 377–388.

(50) E. R. Mardis, "The $1,000 genome, the $100,000 analysis?", *Genome medicine*, 2010, **2**, 1.

(51) A. Holzinger, M. Dehmer and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions", *BMC Bioinformatics*, 2014, **15**, 1.

(52) U. Ohler, G.-c. Liao, H. Niemann and G. M. Rubin, "Computational analysis of core promoters in the Drosophila genome", *Genome biology*, 2002, **3**, 1.

(53) V. Marx, "Drilling into big cancer-genome data", *Nature methods*, 2013, **10**, 293–297.

(54) F. S. Collins and H. Varmus, "A new initiative on precision medicine", *New England Journal of Medicine*, 2015, **372**, 793–795.

(55) V. Dhar, "Data science and prediction", *Communications of the ACM*, 2013, **56**, 64–73.

(56) W. Burke and B. M. Psaty, "Personalized medicine in the era of genomics", *Jama*, 2007, **298**, 1682–1684.

(57) L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine", *Nature Reviews Clinical Oncology*, 2011, **8**, 184–187.

(58) N. R. C. (C. on A Framework for Developing a New Taxonomy of Disease et al., *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*, National Academies Press (US), 2011.

(59) C. Auffray, Z. Chen and L. Hood, "Systems medicine: the future of medical genomics and healthcare", *Genome medicine*, 2009, **1**, 1.

(60) C. Giallourakis, C. Henson, M. Reich, X. Xie and V. K. Mootha, "Disease gene discovery through integrative genomics", *Annu. Rev. Genomics Hum. Genet.*, 2005, **6**, 381–406.

(61) J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross et al., "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease", *science*, 2006, **313**, 1929–1935.

(62) M. Liang, "Integrative pathway knowledge bases as a tool for systems molecular medicine", *Physiological genomics*, 2007, **30**, 209–212.

(63) P. Beltrao, C. Kiel and L. Serrano, "Structures in systems biology", *Current opinion in structural biology*, 2007, **17**, 378–384.

(64) J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou et al., "Towards a proteome-scale map of the human protein–protein interaction network", *Nature*, 2005, **437**, 1173–1178.

(65) M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics", *Nature Reviews Genetics*, 2015, **16**, 321–332.

(66) S. Degroeve, B. De Baets, Y. Van de Peer and P. Rouzé, "Feature subset selection for splice site prediction", *Bioinformatics*, 2002, **18**, S75–S83.

(67) P. Bucher, "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences", *Journal of Molecular Biology*, 1990, **212**, 563–578.

(68) E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang and J. Widom, "A genomic code for nucleosome positioning", *Nature*, 2006, **442**, 772–778.

(69) D. Roden and R. Tyndale, "Genomic medicine, precision medicine, personalized medicine: what's in a name?", *Clinical Pharmacology & Therapeutics*, 2013, **94**, 169–172.

(70) T. A. Manolio, R. L. Chisholm, B. Ozenberger, D. M. Roden, M. S. Williams, R. Wilson, D. Bick, E. P. Bottinger, M. H. Brilliant, C. Eng et al., "Implementing genomic medicine in the clinic: the future is here", *Genetics in Medicine*, 2013, **15**, 258–267.

(71) J. H. Coote and M. J. Joyner, "Is precision medicine the route to a healthy world?", *The Lancet*, 2015, **385**, 1617.

(72) J. M. Rae, M. Regan, B. Leyland-Jones, D. F. Hayes and M. Dowsett, "CYP2D6 genotype should not be used for deciding about tamoxifen therapy in post-menopausal breast cancer", *Journal of Clinical Oncology*, 2013, **31**, 2753–2755.

(73) M. T. Scheuner, H. de Vries, B. Kim, R. C. Meili, S. H. Olmstead and S. Teleki, "Are electronic health records ready for genomic medicine?", *Genetics in Medicine*, 2009, **11**, 510–517.

(74) R. Mirnezami, J. Nicholson and A. Darzi, "Preparing for precision medicine", *New England Journal of Medicine*, 2012, **366**, 489–491.

(75) S. J. Bielinski, J. E. Olson, J. Pathak, R. M. Weinshilboum, L. Wang, K. J. Lyke, E. Ryu, P. V. Targonski, M. D. Van Norstrand, M. A. Hathcock et al., Mayo Clinic Proceedings, 2014, vol. 89, pp. 25–33.

(76) W. Zhang, E. Zeng, D. Liu, S. Jones and S. Emrich, Computational Advances in Bio and Medical Sciences (ICCABS), 2012 IEEE 2nd International Conference on, 2012, pp. 1–6.

(77) J.-Y. Li, J. Wang and R. S. Zeigler, "The 3,000 rice genomes project: new opportunities and challenges for future rice research", *GigaScience*, 2014, **3**, 1.

(78) C. Ma, H. H. Zhang and X. Wang, "Machine learning for Big Data analytics in plants", *Trends in plant science*, 2014, **19**, 798–808.

(79) L. Chin, J. N. Andersen and P. A. Futreal, "Cancer genomics: from discovery science to personalized medicine", *Nature medicine*, 2011, **17**, 297–303.

(80) M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares et al., "Systematic identification of genomic markers of drug sensitivity in cancer cells", *Nature*, 2012, **483**, 570–575.

(81) B. Yngvadottir, D. G. MacArthur, H. Jin and C. Tyler-Smith, "The promise and reality of personal genomics", *Genome biology*, 2009, **10**, 1.

(82) H. Chen, H. Zhao, J. Shen, R. Zhou and Q. Zhou, Big Data (BigData Congress), 2015 IEEE International Congress on, 2015, pp. 134–141.

(83) R. Fakoor, F. Ladhak, A. Nazi and M. Huber, Proceedings of the ICML Workshop on the Role of Machine Learning in Transforming Healthcare. Atlanta, Georgia: JMLR: W&CP, 2013.

(84) C. G. A. R. Network et al., "Comprehensive genomic characterization of squamous cell lung cancers", *Nature*, 2012, **489**, 519–525.

(85) M. P. Schroeder, A. Gonzalez-Perez and N. Lopez-Bigas, "Visualizing multidimensional cancer genomics data", *Genome medicine*, 2013, **5**, 1.

(86) B. Tran, J. E. Dancey, S. Kamel-Reid, J. D. McPherson, P. L. Bedard, A. M. Brown, T. Zhang, P. Shaw, N. Onetto, L. Stein et al., "Cancer genomics: technology, discovery, and translation", *Journal of Clinical Oncology*, 2012, **30**, 647–660.

(87) B. J. Raphael, J. R. Dobson, L. Oesper and F. Vandin, "Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine", *Genome medicine*, 2014, **6**, 1.

(88) *Omicsmap. Next Generation Genomics: World Map of High-throughput*, http://omicsmaps.com/, Accessed: Sequencers. 2015.

(89) P. S. Meltzer, "Cancer genomics: small RNAs with big impacts", *Nature*, 2005, **435**, 745–746.

(90) M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties", *PLoS one*, 2013, **8**, e61318.

(91) R. Sameek and A. M. Chinnaiyan, "Translating genomics for precision cancer medicine", *Annual review of genomics and human genetics*, 2014, **15**, 395–415.

(92) D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation", *cell*, 2011, **144**, 646–674.

(93) J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson et al., "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal", *Science signaling*, 2013, **6**, pl1.

(94) L. Hood and M. Flores, "A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory", *New biotechnology*, 2012, **29**, 613–624.

(95) W. J. Kent and D. Haussler, "Assembly of the working draft of the human genome with GigAssembler", *Genome Research*, 2001, **11**, 1541–1548.

(96) C. Wilks, D. Maltbie, M. Diekhans and D. Haussler, "CGHub: Kick-starting the Worldwide Genome Web", *Proceedings of the Asia-Pacific Advanced Network*, 2013, **35**, 1–13.

(97) C. Wilks, M. S. Cline, E. Weiler, M. Diehkans, B. Craft, C. Martin, D. Murphy, H. Pierce, J. Black, D. Nelson et al., "The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data", *Database*, 2014, **2014**, bau093.

(98) P. E. Dewdney, P. J. Hall, R. T. Schilizzi and T. J. L. Lazio, "The square kilometre array", *Proceedings of the IEEE*, 2009, **97**, 1482–1496.

(99) C. Kiddle, A. Taylor, J. Cordes, O. Eymere, V. Kaspi, D. Pigat, E. Rosolowsky, I. Stairs and A. Willis, Proceedings of the 2011 ACM workshop on Gateway computing environments, 2011, pp. 65–72.

(100) R. Summerhill, 6th GLIF Meeting, 2006.

(101) S. Deorowicz and S. Grabowski, "Compression of DNA sequence reads in FASTQ format", *Bioinformatics*, 2011, **27**, 860–862.

(102) J. K. Bonfield and M. V. Mahoney, "Compression of FASTQ and SAM format sequencing data", *PLoS One*, 2013, **8**, e59190.

(103) Z. Zhu, Y. Zhang, Z. Ji, S. He and X. Yang, "High-throughput DNA sequence data compression", *Briefings in bioinformatics*, 2013, bbt087.

(104) A. J. Cox, M. J. Bauer, T. Jakobi and G. Rosone, "Large-scale compression of genomic sequence databases with the Burrows–Wheeler transform", *Bioinformatics*, 2012, **28**, 1415–1419.

(105) M. G. Langille and J. A. Eisen, "BioTorrents: a file sharing service for scientific data", *PLoS One*, 2010, **5**, e10071.

(106) S. Wandelt, A. Rheinländer, M. Bux, L. Thalheim, B. Haldemann and U. Leser, "Data management challenges in next generation sequencing", *Datenbank-Spektrum*, 2012, **12**, 161–171.

(107) M. C. Brandon, D. C. Wallace and P. Baldi, "Data structures and compression algorithms for genomic sequence data", *Bioinformatics*, 2009, **25**, 1731–1738.

(108) M. H.-Y. Fritz, R. Leinonen, G. Cochrane and E. Birney, "Efficient storage of high throughput DNA sequencing data using reference-based compression", *Genome research*, 2011, **21**, 734–740.

(109) P.-R. Loh, M. Baym and B. Berger, "Compressive genomics", *Nature biotechnology*, 2012, **30**, 627–630.

(110) M. C. Schatz, "Computational thinking in the era of big data biology", *Genome biology*, 2012, **13**, 1.

(111) O. Trelles, P. Prins, M. Snir and R. C. Jansen, "Big data, but are we ready?", *Nature Reviews Genetics*, 2011, **12**, 224–224.

(112) A. Golden, S. G. Djorgovski and J. M. Greally, "Astrogenomics: big data, old problems, old solutions?", *Genome biology*, 2013, **14**, 1.

(113) M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra, "Circos: an information aesthetic for comparative genomics", *Genome research*, 2009, **19**, 1639–1645.

(114) C. Perez-Llamas and N. Lopez-Bigas, "Gitools: analysis and visualisation of genomic data using interactive heat-maps", *PLoS One*, 2011, **6**, e19541.

(115) J. Zhu, J. Z. Sanborn, S. Benz, C. Szeto, F. Hsu, R. M. Kuhn, D. Karolchik, J. Archie, M. E. Lenburg, L. J. Esserman et al., "The UCSC cancer genomics browser", *Nature methods*, 2009, **6**, 239–240.

(116) J. Zhang, R. Finney, M. Edmonson, C. Schaefer, W. Rowe, C. Yan, R. Clifford, S. Greenblum, G. Wu, H. Zhang et al., "The Cancer Genome Workbench: identifying and visualizing complex genetic alterations in tumors", *NCI Nature Pathway Interaction Database*, 2010, **10**.

(117) E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson et al., "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data", *Cancer discovery*, 2012, **2**, 401–404.

(118) R. S. Kim, N. Goossens and Y. Hoshida, "Use of big data in drug development for precision medicine", *Expert Review of Precision Medicine and Drug Development*, 2016, **1**, 245–253.

(119) J. M. González-Camacho, J. Crossa, P. Pérez-Rodriguez, L. Ornella and D. Gianola, "Genome-enabled prediction using probabilistic neural network classifiers", *BMC Genomics*, 2016, **17**, 1.

(120) M. Galli, I. Zoppis, A. Smith, F. Magni and G. Mauri, "Machine learning approaches in MALDI-MSI: clinical applications", *Expert Review of Proteomics*, 2016, **13**, 685–696.

(121) B. E. Huang, W. Mulyasasmita and G. Rajagopal, "The path from big data to precision medicine", *Expert Review of Precision Medicine and Drug Development*, 2016, **1**, 129–143.

(122) D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization", *IEEE Transactions on Evolutionary Computation*, 1997, **1**, 67–82.

(123) N. Long, D. Gianola, G. Rosa, K. Weigel and S. Avendano, "Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers", *Journal of Animal Breeding and Genetics*, 2007, **124**, 377–389.

(124) L. Peña-Castillo, M. Tasan, C. L. Myers, H. Lee, T. Joshi, C. Zhang, Y. Guan, M. Leone, A. Pagnani, W. K. Kim et al., "A critical assessment of Mus musculus gene function prediction using integrated genomic evidence", *Genome biology*, 2008, **9**, 1.

(125) T. Wasson and A. J. Hartemink, "An ensemble model of competitive multi-factor binding of the genome", *Genome research*, 2009, **19**, 2101–2112.

(126) R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad and J. K. Pritchard, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data", *Genome research*, 2011, **21**, 447–455.

(127) I. Walsh, G. Pollastri and S. C. Tosatto, "Correct machine learning on protein sequences: a peer-reviewing perspective", *Briefings in bioinformatics*, 2015, bbv082.

(128) L. M. Breckels, S. B. Holden, D. Wojnar, C. M. Mulvey, A. Christoforou, A. Groen, M. W. Trotter, O. Kohlbacher, K. S. Lilley and L. Gatto, "Learning from heterogeneous data sources: an application in spatial proteomics", *PLoS Computational Biology*, 2016, **12**, e1004920.

(129) L. Ornella, P. Pérez, E. Tapia, J. González-Camacho, J. Burgueño, X. Zhang, S. Singh, F. Vicente, D. Bonnett, S. Dreisigacker et al., "Genomic-enabled prediction with classification algorithms", *Heredity*, 2014, **112**, 616–626.

(130) I. Sant'Anna, R. Tomaz, G. Silva, M. Nascimento, L. Bhering and C. Cruz, "Superiority of artificial neural networks for a genetic classification procedure", *Genet Mol Res*, 2015, **14**, 9898–906.

(131) E. P. Consortium et al., "An integrated encyclopedia of DNA elements in the human genome", *Nature*, 2012, **489**, 57–74.

(132) M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects", *Science*, 2015, **349**, 255–260.

(133) X. Wu, X. Zhu, G.-Q. Wu and W. Ding, "Data mining with big data", *IEEE Transactions on Knowledge and Data Engineering*, 2014, **26**, 97–107.

(134) A. O'Driscoll, J. Daugelaite and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics", *Journal of Biomedical Informatics*, 2013, **46**, 774–781.

(135) E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman et al., "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project", *Nature*, 2007, **447**, 799–816.

(136) B. Maher, "ENCODE: The human encyclopaedia.", *Nature*, 2012, **489**, 46.

(137) M. Gerstein, "Genomics: ENCODE leads the way on big data", *Nature*, 2012, **489**, 208–208.

(138) S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger et al., "Landscape of transcription in human cells", *Nature*, 2012, **489**, 101–108.

(139) R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot et al., "The accessible chromatin landscape of the human genome", *Nature*, 2012, **489**, 75–82.

(140) A. Sanyal, B. R. Lajoie, G. Jain and J. Dekker, "The long-range interaction landscape of gene promoters", *Nature*, 2012, **489**, 109–113.

(141) M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander et al., "Architecture of the human regulatory network derived from ENCODE data", *Nature*, 2012, **489**, 91–100.

(142) S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson et al., "An expansive human regulatory lexicon encoded in transcription factor footprints", *Nature*, 2012, **489**, 83–90.

(143) K. Y. Yip, C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder et al., "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors", *Genome biology*, 2012, **13**, 1.

(144) O. Canela-Xandri, A. Law, A. Gray, J. A. Woolliams and A. Tenesa, "A new tool called DISSECT for analysing large genomic data sets using a Big Data approach", *Nature communications*, 2015, **6**.

(145) E. R. Holzinger, S. M. Dudek, A. T. Frase, B. Fridley, P. Chalise and M. D. Ritchie, European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 2012, pp. 134–143.

(146) B. Paten, M. Diekhans, B. J. Druker, S. Friend, J. Guinney, N. Gassner, M. Guttman, W. J. Kent, P. Mantey, A. A. Margolin et al., "The nih bd2k center for big data in translational genomics", *Journal of the American Medical Informatics Association*, 2015, **22**, 1143–1147.

(147) R. König, Y. Zhou, D. Elleder, T. L. Diamond, G. M. Bonamy, J. T. Irelan, C.-y. Chiang, B. P. Tu, P. D. De Jesus, C. E. Lilley et al., "Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication", *Cell*, 2008, **135**, 49–60.

(148) S. D. Tekir, T. Çakir and K. Ö. Ülgen, "Infection strategies of bacterial and viral pathogens through pathogen-human protein-protein interactions", *Frontiers in Microbiology*, 2012, **3**, 1–11.

(149) E. Guven-Maiorov, C.-J. Tsai, B. Ma and R. Nussinov, *Prediction of Host–Pathogen Interactions for Helicobacter pylori by Interface Mimicry and Implications to Gastric Cancer*, Elsevier, 2017, vol. 429, pp. 3925–3941.

(150) B. Squires, C. Macken, A. Garcia-Sastre, S. Godbole, J. Noronha, V. Hunt, R. Chang, C. N. Larsen, E. Klem, K. Biersack and R. H. Scheuermann, "BioHealthBase: Informatics support in the elucidation of influenza virus host–pathogen interactions and virulence", *Nucleic Acids Research*, 2008, **36**, D497–D503.

(151) S. Durmuş, T. Çakır, A. Özgür and R. Guthke, "A review on computational systems biology of pathogen–host interactions", *Frontiers in Microbiology*, 2015, **6**, 235.

(152) A. Calderone, L. Licata and G. Cesareni, "VirusMentha: a new resource for virus-host protein interactions", *Nucleic Acids Research*, 2014, **43**, D588–D592.

(153) M. Urban, R. Pant, A. Raghunath, A. G. Irvine, H. Pedro and K. E. Hammond-Kosack, "The Pathogen-Host Interactions database (PHI-base): additions and future developments", *Nucleic Acids Research*, 2014, **43**, D645–D655.

(154) T. van der Poll and S. M. Opal, "Host–pathogen interactions in sepsis", *The Lancet infectious diseases*, 2008, **8**, 32–43.

(155) E. Gómez-Dı**********az, M. Jordà, M. A. Peinado and A. Rivero, "Epigenetics of host–pathogen interactions: the road ahead and the road behind", *PLoS Pathogens*, 2012, **8**, e1003007.

(156) M. D. Dyer, T. Murali and B. W. Sobral, "Supervised learning and prediction of physical interactions between human and HIV proteins", *Infection, Genetics and Evolution*, 2011, **11**, 917–923.

(157) S. Schleker, M. Kshirsagar and J. Klein-Seetharaman, "Comparing human–Salmonella with plant–Salmonella protein–protein interaction predictions", *Frontiers in Microbiology*, 2015, **6**, 45.

(158) A. Wallqvist, V. Memišević, N. Zavaljevski, R. Pieper, S. V. Rajagopala, K. Kwon, C. Yu, T. A. Hoover and J. Reifman, "Using host-pathogen protein interactions to identify and characterize Francisella tularensis virulence factors", *BMC Genomics*, 2015, **16**, 1106.

(159) R. Arnold, K. Boonen, M. G. Sun and P. M. Kim, "Computational analysis of interactomes: Current and future perspectives for bioinformatics approaches to model the host–pathogen interaction space", *Methods*, 2012, **57**, 508–518.

(160) A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon et al., "PATRIC, the bacterial bioinformatics database and analysis resource", *Nucleic Acids Research*, 2014, **42**, D581–D591.

(161) S. Durmuş Tekir, T. Çakır, E. Ardıç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür et al., "PHISTO: pathogen–host interaction search tool", *Bioinformatics*, 2013, **29**, 1357–1358.

(162) H. Zhou, J. Jin and L. Wong, "Progress in computational studies of host–pathogen interactions", *Journal of Bioinformatics and Computational Biology*, 2013, **11**, 1230001 (1–26).

(163) A. S. Fauci, N. A. Touchette and G. K. Folkers, "Emerging infectious diseases: a 10-year perspective from the National Institute of Allergy and Infectious Diseases", *International Journal of Risk & Safety in Medicine*, 2005, **17**, 157–167.

(164) T. Driscoll, M. D. Dyer, T. Murali and B. W. Sobral, "PIG—the pathogen interaction gateway", *Nucleic Acids Research*, 2008, **37**, D647–D650.

(165) B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu et al., "ViPR: an open bioinformatics database and analysis resource for virology research", *Nucleic Acids Research*, 2011, **40**, D593–D598.

(166) K. Megy, S. J. Emrich, D. Lawson, D. Campbell, E. Dialynas, D. S. Hughes, G. Koscielny, C. Louis, R. M. MacCallum, S. N. Redmond et al., "VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics", *Nucleic Acids Research*, 2011, **40**, D729–D734.

(167) C. Aurrecoechea, J. Brestelli, B. P. Brunk, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges et al., "EuPathDB: a portal to eukaryotic pathogen databases", *Nucleic Acids Research*, 2009, **38**, D415–D419.

(168) C. Aurrecoechea, A. Barreto, E. Y. Basenko, J. Brestelli, B. P. Brunk, S. Cade, K. Crouch, R. Doherty, D. Falke, S. Fischer et al., "EuPathDB: the eukaryotic pathogen genomics database resource", *Nucleic Acids Research*, 2016, **45**, D581–D591.

(169) S. Braxton, D. Onstad, D. Dockter, R. Giordano, R. Larsson and R. Humber, "Description and analysis of two internet-based databases of insect pathogens: EDWIP and VIDIL", *Journal of Invertebrate Pathology*, 2003, **83**, 185–195.

(170) G. Silvestri and P. A. Barry, "Editorial overview: Host pathogens: New paradigms and tools to decipher and deconstruct the host–pathogen interaction", *Current opinion in immunology*, 2015, **36**, v–viii.

(171) C. Zhou, J. Smith, M. Lam, A. Zemla, M. D. Dyer and T. Slezak, "MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications", *Nucleic Acids Research*, 2006, **35**, D391–D394.

(172) M. G. Ammari, C. R. Gresham, F. M. McCarthy and B. Nanduri, "HPIDB 2.0: a curated database for host–pathogen interactions", *Database : the journal of biological databases and curation*, 2016, **2016**, 1–9.

(173) O. P. Sharma, A. Jadhav, A. Hussain and M. S. Kumar, "VPDB: viral protein structural database", *Bioinformation*, 2011, **6**, 324.

(174) B. Aranda, H. Blankenburg, S. Kerrien, F. S. Brinkman, A. Ceol, E. Chautard, J. M. Dana, J. De Las Rivas, M. Dumousseau, E. Galeota et al., "PSICQUIC and PSISCORE: accessing and scoring molecular interactions", *Nature methods*, 2011, **8**, 528.

(175) L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico et al., "MINT, the molecular interaction database: 2012 update", *Nucleic Acids Research*, 2012, **40**, D857–D861.

(176) S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz et al., "The IntAct molecular interaction database in 2012", *Nucleic Acids Research*, 2012, **40**, D841–D846.

(177) A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam et al., "The BioGRID interaction database: 2017 update", *Nucleic Acids Research*, 2017, **45**, D369–D379.

(178) I. Xenarios, L. Salwínski, X. J. Duan, P. Higney, S.-M. Kim and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions", *Nucleic Acids Research*, 2002, **30**, 303–305.

(179) L. Chen, D. Zheng, B. Liu, J. Yang and Q. Jin, "VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on", *Nucleic Acids Research*, 2015, **44**, D694–D697.

(180) A. Chatr-Aryamontri, A. Ceol, D. Peluso, A. Nardozza, S. Panni, F. Sacco, M. Tinti, A. Smolyar, L. Castagnoli, M. Vidal et al., "VirusMINT: a viral protein interaction database", *Nucleic Acids Research*, 2008, **37**, D669–D673.

(181) S. Bleves, I. Dunger, M. C. Walter, D. Frangoulidis, G. Kastenmüller, R. Voulhoux and A. Ruepp, "HoPaCI-DB: host-Pseudomonas and Coxiella interaction database", *Nucleic Acids Research*, 2013, **42**, D671–D676.

(182) M. Urban, A. Cuzick, K. Rutherford, A. Irvine, H. Pedro, R. Pant, V. Sadanadan, L. Khamari, S. Billal, S. Mohanty et al., "PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database", *Nucleic Acids Research*, 2016, **45**, D604–D610.

(183) J. Yue, D. Zhang, R. Ban, X. Ma, D. Chen, G. Li, J. Liu, M. Wisniewski, S. Droby and Y. Liu, "PCPPI: a comprehensive database for the prediction of Penicillium–crop protein–protein interactions", *Database*, 2017, **2017**, baw170: 1–9.

(184) G. D. Bader, D. Betel and C. W. Hogue, "BIND: the biomolecular interaction network database", *Nucleic Acids Research*, 2003, **31**, 248–250.

(185) Z. Xiang, Y. Tian and Y. He, "PHIDIAS: a pathogen-host interaction data integration and analysis system", *Genome biology*, 2007, **8**, R150.

(186) A. Calderone, L. Castagnoli and G. Cesareni, "Mentha: a resource for browsing integrated protein-interaction networks", *Nature methods*, 2013, **10**, 690–691.

(187) G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews et al., "Reactome: a knowledgebase of biological pathways", *Nucleic Acids Research*, 2005, **33**, D428–D432.

(188) C. Prieto and J. De Las Rivas, "APID: agile protein interaction DataAnalyzer", *Nucleic Acids Research*, 2006, **34**, W298–W302.

(189) K. Breuer, A. K. Foroushani, M. R. Laird, C. Chen, A. Sribnaia, R. Lo, G. L. Winsor, R. E. Hancock, F. S. Brinkman and D. J. Lynn, "InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation", *Nucleic Acids Research*, 2013, **41**, D1228–D1233.

(190) L. Bülow, M. Schindler, C. Choi and R. Hehl, "PathoPlant®: a database on plant-pathogen interactions", *In silico biology*, 2004, **4**, 529–536.

(191) J. C. Bolívar, F. Machens, Y. Brill, A. Romanov, L. Bülow and R. Hehl, "'In silico expression analysis', a novel PathoPlant web tool to identify abiotic and biotic stress conditions associated with specific cis-regulatory sequences", *Database*, 2014, **2014**, DOI: 10.1093/database/bau030.

(192) U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes and V. Stümpflen, "MPact: the MIPS protein interaction resource on yeast", *Nucleic Acids Research*, 2006, **34**, D436–D441.

(193) K. R. Brown and I. Jurisica, "Unequal evolutionary conservation of human protein interactions in interologous networks", *Genome biology*, 2007, **8**, R95.

(194) J. Goll, S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb and P. Uetz, "MPIDB: the microbial protein interaction database", *Bioinformatics*, 2008, **24**, 1743–1744.

(195) V. Vialas, R. Nogales-Cadenas, C. Nombela, A. Pascual-Montano and C. Gil, "Proteopathogen, a protein database for studying Candida albicans–host interaction", *Proteomics*, 2009, **9**, 4664–4668.

(196) T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal et al., "Human protein reference database—2009 update", *Nucleic Acids Research*, 2009, **37**, D767–D772.

(197) U. Mudunuri, A. Che, M. Yi and R. M. Stephens, "bioDBnet: the biological database network", *Bioinformatics*, 2009, **25**, 555–556.

(198) C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay and K. H. Buetow, "PID: the pathway interaction database", *Nucleic Acids Research*, 2009, **37**, D674–D679.

(199) E. Emmenegger, E. Kentop, T. Thompson, S. Pittam, A. Ryan, D. Keon, J. Carlino, J. Ranson, R. Life, R. Troyer et al., "Development of an aquatic pathogen database (AquaPathogen X) and its utilization in tracking emerging fish virus pathogens in North America", *Journal of Fish Diseases*, 2011, **34**, 579–587.

(200) S. K. Kwofie, U. Schaefer, V. S. Sundararajan, V. B. Bajic and A. Christoffels, "HCVpro: hepatitis C virus protein interaction database", *Infection, Genetics and Evolution*, 2011, **11**, 1971–1977.

(201) H. Kim, S. Park, H. Min and S. Yoon, "vHoT: a database for predicting interspecies interactions between viral microRNA and host genomes", *Archives of virology*, 2012, **157**, 497–501.

(202) R. G. Ptak, W. Fu, B. E. Sanders-Beer, J. E. Dickerson, J. W. Pinney, D. L. Robertson, M. N. Rozanov, K. S. Katz, D. R. Maglott, K. D. Pruitt et al., "Cataloguing the HIV type 1 human protein interaction network", *AIDS research and human retroviruses*, 2008, **24**, 1497–1502.

(203) D. Ako-Adjei, W. Fu, C. Wallin, K. S. Katz, G. Song, D. Darji, J. R. Brister, R. G. Ptak and K. D. Pruitt, "HIV-1, human interaction database: current status and new features", *Nucleic Acids Research*, 2014, **43**, D566–D570.

(204) T. Guirimand, S. Delmotte and V. Navratil, "VirHostNet 2.0: surfing on the web of virus/host molecular interactions data", *Nucleic Acids Research*, 2014, **43**, D583–D587.

(205) G. Launay, R. Salza, D. Multedo, N. Thierry-Mieg and S. Ricard-Blum, "MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities", *Nucleic Acids Research*, 2014, **43**, D321–D327.

(206) J. Mariethoz, K. Khatib, D. Alocci, M. P. Campbell, N. G. Karlsson, N. H. Packer, E. H. Mullen and F. Lisacek, "SugarBindDB, a resource of glycan-mediated host–pathogen interactions", *Nucleic Acids Research*, 2015, **44**, D1243–D1250.

(207) P. Karyala, R. Metri, C. Bathula, S. K. Yelamanchi, L. Sahoo, S. Arjunan, N. P. Sastri and N. Chandra, "DenHunt-A Comprehensive Database of the Intricate Network of Dengue-Human Interactions", *PLoS neglected tropical diseases*, 2016, **10**, e0004965.

(208) D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork et al., "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible", *Nucleic Acids Research*, 2016, gkw937.

(209) C. F. Taylor, H. Hermjakob, R. K. Julian Jr, J. S. Garavelli, R. Aebersold and R. Apweiler, "The work of the human proteome organisation's proteomics standards initiative (HUPO PSI)", *Omics: a journal of integrative biology*, 2006, **10**, 145–151.

(210) H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. Von Mering et al., "The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data", *Nature biotechnology*, 2004, **22**, 177.

(211) L. Martens, S. Orchard, R. Apweiler and H. Hermjakob, "Human Proteome Organization Proteomics Standards Initiative Data Standardization, a View on Developments and Policy", *Molecular & Cellular Proteomics*, 2007, **6**, 1666–1667.

(212) S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman, G. Cesareni et al., "Protein interaction data curation: the International Molecular Exchange (IMEx) consortium", *Nature Methods*, 2012, **9**, 345–350.

(213) M. Sivade, D. Alonso-López, M. Ammari, G. Bradley, N. H. Campbell, A. Ceol, G. Cesareni, C. Combe, J. De Las Rivas, N. Del-Toro et al., "Encompassing new use cases-level 3.0 of the HUPO-PSI format for molecular interactions", *BMC Bioinformatics*, 2018, **19**, 134.

(214) P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks", *Genome research*, 2003, **13**, 2498–2504.

(215) Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter et al., "Structure-based prediction of protein-protein interactions on a genome-wide scale", *Nature*, 2012, **490**, 556–560.

(216) Y. Liu, X. Wang and B. Liu, "A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction", *Briefings in Bioinformatics*, 2017, 1–17.

(217) M. D. Dyer, T. M. Murali and B. W. Sobral, "Computational prediction of host-pathogen protein-protein interactions", *Bioinformatics*, 2007, **23**, i159–i166.

(218) S. Wuchty, "Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens", *PLoS One*, 2011, **6**, e26960.

(219) K. Prashanthi and N. Chandra, in *Encyclopedia of Systems Biology*, ed. W. Dubitzky, O. Wolkenhauer, K.-H. Cho and H. Yokota, Springer New York, 2013, pp. 904–908.

(220) M. Mock and A. Fouet, "Anthrax", *Annual Reviews in Microbiology*, 2001, **55**, 647–671.

(221) A. W. Maresso, G. Garufi and O. Schneewind, "Bacillus anthracis secretes proteins that mediate heme acquisition from hemoglobin", *PLoS Pathogens*, 2008, **4**, e1000132.

(222) M. D. Dyer, C. Nef, M. Dufford, C. G. Rivera, D. Shattuck, J. Bassaganya-Riera, T. M. Murali and B. W. Sobral, "The Human-Bacterial pathogen protein interaction networks of Bacillus anthracis, Francisella tularensis, and Yersinia pestis", *PLoS ONE*, 2010, **5**, DOI: 10.1371/journal.pone.0012089.

(223) A. Emamjomeh, B. Goliaei, J. Zahiri and R. Ebrahimpour, "Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method", *Molecular Biosystems*, 2014, **10**, 3147–3154.

(224) F.-E. Eid, M. ElHefnawi and L. S. Heath, "DeNovo: virus-host sequence-based protein-protein interaction prediction", *Bioinformatics*, 2016, **32**, 1144–1150.

(225) R. Sen, L. Nayak and R. K. De, "A review on host–pathogen interactions: classification and prediction", *European Journal of Clinical Microbiology & Infectious Diseases*, 2016, **35**, 1581–1599.

(226) J. Zhang and L. Kurgan, "Review and comparative assessment of sequence-based predictors of protein-binding residues", *Briefings in Bioinformatics*, 2018, **19**, 821–837.

(227) O. Krishnadev and N. Srinivasan, "A data integration approach to predict host-pathogen protein-protein interactions: application to recognize protein interactions between human and a malarial parasite", *In Silico Biol*, 2008, **8**, 235–250.

(228) T. Huo, W. Liu, Y. Guo, C. Yang, J. Lin and Z. Rao, "Prediction of host - pathogen protein interactions between Mycobacterium tuberculosis and Homo sapiens using sequence motifs", *BMC Bioinformatics*, 2015, **16**, 1–9.

(229) H. Hwang, F. Dey, D. Petrey and B. Honig, "Structure-based prediction of ligand–protein interactions on a genome-wide scale", *Proceedings of the National Academy of Sciences*, 2017, **114**, 13685–13690.

(230) S. Mei, "Probability weighted ensemble transfer learning for predicting interactions between HIV-1 and human proteins", *PLoS ONE*, 2013, **8**, 1–13.

(231) I. Ahmed, P. Witbooi and A. Christoffels, "Prediction of human-Bacillus anthracis protein–protein interactions using multi-layer neural network", *Bioinformatics*, 2018, **34**, 4159–4164.

(232) M. N. Davies, A. Secker, A. A. Freitas, E. Clark, J. Timmis and D. R. Flower, "Optimizing amino acid groupings for GPCR classification", *Bioinformatics*, 2008, **24**, 1980–1986.

(233) K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition", *Proteins: Structure, Function, and Bioinformatics*, 2001, **43**, 246–255.

(234) H. B. Shen and K. C. Chou, "PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition", *Analytical Biochemistry*, 2008, **373**, 386–388.

(235) K.-C. Chou, "Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology", *Current Proteomics*, 2009, **6**, 262–274.

(236) 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", *Nucleic Acids Research*, 1997, **25**, 3389–3402.

(237) S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins", *BMC Bioinformatics*, 2005, **6**, 1–6.

(238) J.-F. Xia, K. Han and D.-S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor", *Protein and Peptide Letters*, 2010, **17**, 137–145.

(239) K.-C. Chou and H.-B. Shen, "MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM", *Biochemical and biophysical research communications*, 2007, **360**, 339–345.

(240) K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition", *Journal of Theoretical Biology*, 2011, **273**, 236–247.

(241) J. Zahiri, J. H. Bozorgmehr and A. Masoudi-Nejad, "Computational Prediction of Protein-Protein Interaction Networks: Algorithms and Resources.", *Curr. Genomics*, 2013, **14**, 397–414.

(242) J. Wang, B. Yang, A. Leier, T. T. Marquez-Lago, M. Hayashida, A. Rocker, Y. Zhang, T. Akutsu, K.-C. Chou, R. A. Strugnell et al., "Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors", *Bioinformatics*, 2018, **34**, 2546–2555.

(243) M. R. Uddin, A. Sharma, D. M. Farid, M. M. Rahman, A. Dehzangi and S. Shatabda, "EvoStruct-Sub: An accurate Gram-positive protein subcellular localization predictor using evolutionary and structural features", *Journal of Theoretical Biology*, 2018, **443**, 138–146.

(244) Y. E. Göktepe and H. Kodaz, "Prediction of Protein-Protein Interactions Using An Effective Sequence Based Combined Method", *Neurocomputing*, 2018, **303**, 68–74.

(245) B. Zhang, J. Li and Q. Lü, "Prediction of 8-state protein secondary structures by a novel deep learning architecture", *BMC Bioinformatics*, 2018, **19**, 1–13.

(246) Y.-B. Wang, Z.-H. You, L.-P. Li, D.-S. Huang, F.-F. Zhou and S. Yang, "Improving Prediction of Self-interacting Proteins Using Stacked Sparse Auto-Encoder with PSSM profiles", *International Journal of Biological Sciences*, 2018, **14**, 983–991.

(247) M. O. Dayhoff, "A model of evolutionary change in proteins", *Atlas of protein sequence and structure*, 1972, **5**, 89–99.

(248) S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.", *Proceedings of the National Academy of Sciences*, 1992, **89**, 10915–10919.

(249) J. cheol Jeong, X. Lin and X.-w. Chen, "On position-specific scoring matrix for protein function prediction", *IEEE/ACM transactions on computational biology and bioinformatics*, 2010, **8**, 308–315.

(250) T. Liu, X. Zheng and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile", *Biochimie*, 2010, **92**, 1330–1334.

(251) S. Zhang, F. Ye and X. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM", *Journal of Biomolecular Structure and Dynamics*, 2012, **29**, 1138–1146.

(252) E. Y. Juan, W. Li, J. Jhang and C. Chiu, "Predicting Protein Subcellular Localizations for Gram-Negative Bacteria Using DP-PSSM and Support Vector Machines", *2009 International Conference on Complex, Intelligent and Software Intensive Systems*, 2009, **101**, 836–841.

(253) C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, 1995, **20**, 273–297.

(254) Wikipedia, *Decision tree*, [Online; accessed 23-Oct-2018], 2017.

(255) L. Breiman, "Random forests", *Machine learning*, 2001, **45**, 5–32.

(256) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

(257) Wikipedia, *Naive Bayes classifier*, [Online; accessed 23-Oct-2018], 2017.

(258) H. Zhang, "The optimality of naive Bayes", *AA*, 2004, **1**, 3.

(259) J. H. Friedman, "Greedy Function Approximation : A Gradient Boosting Machine", *The Annals of Statistics*, 2001, **29**, 1189–1232.

(260) L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, E. Santonico et al., "MINT, the molecular interaction database: 2012 update", *Nucleic Acids Research*, 2011, **40**, D857–D861.

(261) W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences", *Bioinformatics*, 2006, **22**, 1658–1659.

(262) F. P. Davis, D. T. Barkan, N. Eswar, J. H. McKerrow and A. Sali, "Host–pathogen protein interactions predicted by comparative modeling", *Protein Science*, 2007, **16**, 2585–2596.

(263) R. Mariano and S. Wuchty, "Structure-based prediction of host–pathogen protein interactions", *Current Opinion in Structural Biology*, 2017, **44**, 119–124.

(264) E. A. Franzosa, S. Garamszegi and Y. Xia, "Toward a three-dimensional view of protein networks between species", *Frontiers in Microbiology*, 2012, **3**, 1–6.

(265) O. Tastan, Y. Qi, J. G. Carbonell and J. Klein-Seetharaman, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2009, p. 516.

(266) N. Tyagi, O. Krishnadev and N. Srinivasan, "Prediction of protein–protein interactions between Helicobacter pylori and a human host", *Molecular bioSystems*, 2009, **5**, 1630–1635.

(267) S. M. Gomez, W. S. Noble and A. Rzhetsky, "Learning to predict protein–protein interactions from protein sequences", *Bioinformatics*, 2003, **19**, 1875–1881.

(268) L. Zhang, International Conference on Intelligent Computing, 2012, pp. 334–341.

(269) S. Yang, H. Li, H. He, Y. Zhou and Z. Zhang, "Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods", *Briefings in Bioinformatics*, 2017, 1–14.

(270) L. M. Manevitz and M. Yousef, "One-class SVMs for document classification", *Journal of Machine Learning Research*, 2001, **2**, 139–154.

(271) B. Chidlovskii and M. Hovelynck, *Multi-modality classification for one-class classification in social networks*, US Patent 8,386,574, 2013.

(272) L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller and M. Kloft, International conference on machine learning, 2018, pp. 4393–4402.

(273) P. Perera and V. M. Patel, "Learning deep features for one-class classification", *IEEE Transactions on Image Processing*, 2019, **28**, 5450–5463.

(274) J. G. Greener, S. M. Kandathil and D. T. Jones, "Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints", *Nature communications*, 2019, **10**, 1–13.

(275) H. Chen, W. Guo, J. Shen, L. Wang and J. Song, "Structural principles analysis of host-pathogen protein-protein interactions: A structural bioinformatics survey", *IEEE Access*, 2018, **6**, 11760–11771.

(276) E. A. Franzosa and Y. Xia, in *Biocomputing 2012*, World Scientific, 2012, pp. 287–298.

(277) J. Soyemi, I. Isewon, J. Oyelade and E. Adebiyi, "Inter-Species/Host-Parasite Protein Interaction Predictions Reviewed", *Current bioinformatics*, 2018, **13**, 396–406.

(278) A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein–protein interactions", *Bioinformatics*, 2005, **21**, i38–i46.

(279) N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou and I. Iliopoulos, "Protein–protein interaction predictions using text mining methods", *Methods*, 2015, **74**, 47–53.

(280) U. Kuzmanov and A. Emili, "Protein-protein interaction networks: probing disease mechanisms using model systems", *Genome medicine*, 2013, **5**, 37.

(281) R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt and M. Gerstein, "A Bayesian networks approach for predicting protein-protein interactions from genomic data", *Science*, 2003, **302**, 449–453.

(282) Y.-X. Fan and H.-B. Shen, "Predicting pupylation sites in prokaryotic proteins using pseudo-amino acid composition and extreme learning machine", *Neurocomputing*, 2014, **128**, 267–272.

(283) J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, "Predicting protein–protein interactions based only on sequences information", *Proceedings of the National Academy of Sciences*, 2007, **104**, 4337–4341.

(284) Z.-H. Zhou, "Ensemble learning", *Encyclopedia of biometrics*, 2015, 411–416.

(285) Á. Győrfi, L. Kovács and L. Szilágyi, 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2019, pp. 909–914.

(286) S. Akodad, S. Vilfroy, L. Bombrun, C. C. Cavalcante, C. Germain and Y. Berthoumieu, 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5.

(287) F. Fahiman, S. M. Erfani and C. Leckie, 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

(288) G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Advances in Neural Information Processing Systems, 2017, pp. 3146–3154.

(289) K. Yugandhar and M. M. Gromiha, "Protein-protein binding affinity prediction from amino acid sequence.", *Bioinformatics (Oxford, England)*, 2014, **30**, 3583–3589.

(290) N. Nakajima, M. Hayashida, J. Jansson, O. Maruyama and T. Akutsu, "Determining the minimum number of protein-protein interactions required to support known protein complexes", *PLoS ONE*, 2018, 1–17.

(291) H. Chen, L. Wang, C.-H. Chi and J. Shen, "Leveraging SMOTE in A Two-Layer Model for Prediction of Protein-Protein Interactions", *2019 International Conference on Advanced Cloud and Big Data (CBD)*, 2019, 133–138.

(292) B. Liu, "BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches", *Briefings in Bioinformatics*, 2017, 1–15.

(293) S. Blasche, S. Arens, A. Ceol, G. Siszler, M. A. Schmidt, R. Häuser, F. Schwarz, S. Wuchty, P. Aloy, P. Uetz, T. Stradal, M. Koegl, M. A. Schmid, R. Häuser, F. Schwarz, S. Wuchty, P. Aloy, P. Uetz, T. Stradal and M. Koegl, "The EHEC-host interactome reveals novel targets for the translocated intimin receptor", *Scientific Reports*, 2014, **4**, 22–26.

(294) F. Tordini, "A Cloud Solution for Multi-omics Data Integration", *Proceedings - 13th IEEE International Conference on Ubiquitous Intelligence and Computing, 13th IEEE International Conference on Advanced and Trusted Computing, 16th IEEE International Conference on Scalable Computing and Communications, IEEE Internationa*, 2017, 559–566.

(295) H. M. Berman, P. E. Bourne, J. Westbrook and C. Zardecki, in *Protein Structure*, CRC Press, 2003, pp. 394–410.

(296) E. P. Consortium et al., "The ENCODE (ENCyclopedia of DNA elements) project", *Science*, 2004, **306**, 636–640.

(297) X. Chen, M. Li, R. Zheng, S. Zhao, F.-X. Wu, Y. Li and J. Wang, "A novel method of gene regulatory network structure inference from gene knock-out expression data", *Tsinghua Science and Technology*, 2019, **24**, 446–455.

(298) C. Wu, J. Zhu and X. Zhang, "Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes", *BMC Bioinformatics*, 2012, **13**, 1.

(299) M. Kshirsagar, J. G. Carbonell, J. Klein-Seetharaman and K. Murugesan, "Multi-task matrix completion for learning protein interactions across diseases", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, **9649**, 53–64.

(300) Z. Ding and D. Kihara, "Computational identification of protein-protein interactions in model plant proteomes", *Scientific Reports*, 2019, **9**, 1–13.

(301) T. Chen and C. Guestrin, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

(302) J. Laurikkala, Conference on Artificial Intelligence in Medicine in Europe, 2001, pp. 63–66.

(303) I. Mani and I. Zhang, Proceedings of workshop on learning from imbalanced datasets, 2003, vol. 126.

(304) N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, 2002, **16**, 321–357.

(305) H. He, Y. Bai, E. A. Garcia and S. Li, 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008, pp. 1322–1328.

(306) S. Durmus Tekir, T. Cakir and K. Ulgen, "Infection strategies of bacterial and viral pathogens through pathogen–human protein–protein interactions", *Frontiers in Microbiology*, 2012, **3**, 46.

(307) H. R. Ahmed and J. I. Glasgow, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 2938–2945.

(308) S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, 1997, **9**, 1735–1780.

(309) M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks", *IEEE Transactions on Signal Processing*, 1997, **45**, 2673–2681.

(310) Y. Yao and Z. Huang, International Conference on Neural Information Processing, 2016, pp. 345–353.

(311) J. Wu, L. Zhou, C. Cai, F. Dong, J. Shen and G. Sun, 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019, pp. 3482–3487.

(312) T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

(313) M. M. D. Masood, D. Manjula and V. Sugumaran, "Identification of new disease genes from protein–protein interaction network", *Journal of Ambient Intelligence and Humanized Computing*, 2018, 1–9.

(314) T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proceedings of the National Academy of Sciences*, 2001, **98**, 4569–4574.

(315) V. Navratil, B. de Chassey, L. Meyniel, S. Delmotte, C. Gautier, P. André, V. Lotteau and C. Rabourdin-Combe, "VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks", *Nucleic Acids Research*, 2009, **37**, D661–D668.

(316) M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, "Multisource transfer learning for host-pathogen protein interaction prediction in unlabeled tasks", *NIPS Work. Mach. Learn. Comput. Biol.*, 2013, 3–6.

(317) M. Kshirsagar, S. Schleker, J. Carbonell and J. Klein-Seetharaman, "Techniques for transferring host-pathogen protein interactions knowledge to new tasks.", *Front. Microbiol.*, 2015, **6**, 36.

(318) R. Varadharajan, M. Priyan, P. Panchatcharam, S. Vivekanandan and M. Gunasekaran, "A new approach for prediction of lung carcinoma using back propogation neural network with decision tree classifiers", *Journal of Ambient Intelligence and Humanized Computing*, 2018, 1–12.

(319) M. Prabukumar, L. Agilandeeswari and K. Ganesan, "An intelligent lung cancer diagnosis system using cuckoo search optimization and support vector machine classifier", *Journal of Ambient Intelligence and Humanized Computing*, 2019, **10**, 267–293.

(320) Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis", *BMC Bioinformatics*, 2013, **14**, 1.

(321) M. Gao, H. Zhou and J. Skolnick, "DESTINI: A deep-learning approach to contact-driven protein structure prediction", *Scientific reports*, 2019, **9**, 3514.

(322) B. Panda and B. Majhi, "A novel improved prediction of protein structural class using deep recurrent neural network", *Evolutionary Intelligence*, 2018, 1–8.

(323) Y. LeCun, Y. Bengio and G. Hinton, "Deep learning", *Nature*, 2015, **521**, 436–444.

(324) C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles", *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**, 284–295.

(325) P. Vincent, H. Larochelle, Y. Bengio and P.-A. Manzagol, Proceedings of the 25th international conference on Machine learning, 2008, pp. 1096–1103.

(326) J. M. Hilbe, *Logistic regression models*, CRC press, 2009.

(327) H. Chen, J. Shen, L. Wang and J. Song, Big Data (BigData Congress), 2017 IEEE International Congress on, 2017, pp. 368–375.

(328) Z. Du, L. Li, C.-F. Chen, S. Y. Philip and J. Z. Wang, "G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery", *Nucleic Acids Research*, 2009, gkp463.

(329) J. M. Berg, J. L. Tymoczko, L. Stryer et al., *Biochemistry*, 2002.

(330) M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin et al., "TensorFlow: Large-scale machine learning on heterogeneous systems, 2015", *Software available from tensorflow. org*, 2015, **1**.

(331) C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011, **2**, 27.

(332) A. Akusok, K.-M. Björk, Y. Miche and A. Lendasse, "High-performance extreme learning machines: a complete toolbox for big data applications", *IEEE Access*, 2015, **3**, 1011–1025.

(333) U. Consortium et al., "UniProt: the universal protein knowledgebase", *Nucleic Acids Research*, 2017, **45**, D158–D169.

(334) R. Breitling, "What is systems biology?", *Frontiers in Physiology*, 2010, **1**.

(335) P. Aloy and R. B. Russell, "Structural systems biology: modelling protein interactions", *Nature Reviews Molecular Cell Biology*, 2006, **7**, 188–197.

(336) H. Chen, J. Shen, L. Wang and J. Song, International Conference on Computer Supported Cooperative Work in Design, 2017, pp. 269–274.

(337) K. Li, C. Xu, J. Huang, W. Liu, L. Zhang, W. Wan, H. Tao, L. Li, S. Lin, A. Harrison et al., "Prediction and identification of the effectors of heterotrimeric G proteins in rice (Oryza sativa L.)", *Briefings in bioinformatics*, 2016, **18**, 270–278.

(338) M. Bhagwat and L. Aravind, "PSI-blast tutorial", *Comparative Genomics*, 2008, 177–186.

(339) J. Zhou and O. Troyanskaya, International Conference on Machine Learning, 2014, pp. 745–753.

(340) Z. Li and Y. Yu, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016, pp. 2560–2567.

(341) M. Zamani and S. C. Kremer, Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on, 2015, pp. 1–6.

(342) C. Floudas, H. Fung, S. McAllister, M. Mönnigmann and R. Rajgaria, "Advances in protein structure prediction and de novo protein design: A review", *Chemical Engineering Science*, 2006, **61**, 966–988.

(343) Z. Aydin, Y. Altunbasak and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models.", *BMC Bioinformatics*, 2006, **7**, 178–178.

(344) M. Spencer, J. Eickholt and J. Cheng, "A deep learning network approach to ab initio protein secondary structure prediction", *IEEE/ACM transactions on computational biology and bioinformatics*, 2015, **12**, 103–112.

(345) A. Yaseen and Y. Li, "Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features", *BMC Bioinformatics*, 2014, **15**, S3.

(346) N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models", *Journal of Molecular Biology*, 1988, **202**, 865–884.

(347) O. Dor and Y. Zhou, "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training", *Proteins: Structure, Function, and Bioinformatics*, 2007, **66**, 838–845.

(348) B. Jayaram, P. Dhingra, A. Mishra, R. Kaushik, G. Mukherjee, A. Singh and S. Shekhar, "Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins", *BMC Bioinformatics*, 2014, **15**, S7.

(349) S. Kaczanowski and P. Zielenkiewicz, "Why similar protein sequences encode similar three-dimensional structures?", *Theoretical Chemistry Accounts*, 2010, **125**, 643–650.

(350) S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari and R. Jothi, "DOMINE: a comprehensive collection of known and predicted domain-domain interactions", *Nucleic Acids Research*, 2010, **39**, D730–D735.

(351) R. D. Finn, B. L. Miller, J. Clements and A. Bateman, "iPfam: a database of protein family and domain interactions found in the Protein Data Bank", *Nucleic Acids Research*, 2013, **42**, D364–D373.

(352) B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners", *PLoS Computational Biology*, 2007, **3**, e43.

(353) S. Khor, "Inferring domain-domain interactions from protein-protein interactions with formal concept analysis", *PLoS One*, 2014, **9**, e88943.

(354) X.-M. Zhao, G. Chesi and L. Chen, "Computational Systems Biology: Understanding Biological Systems from the Perspective of Networks and Dynamics", *IEEE Systems, Man and Cybernetics Society: eNewsletter*, March 2009, **26**.

(355) J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, E. Abola et al., "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules", *Acta Crystallographica Section D Biological Crystallography*, 1998, **54**, 1078–1084.

(356) W. L. DeLano, "The PyMOL molecular graphics system", *accessed via: http://pymol.org*, 2002.

(357) Schrödinger, LLC, "The PyMOL Molecular Graphics System, Version 1.8", *accessed via: https://pymol.org/2/*, 2015.

(358) R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry et al., "Pfam: the protein families database", *Nucleic Acids Research*, 2013, **42**, D222–D230.

(359) R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas et al., "The Pfam protein families database: towards a more sustainable future", *Nucleic Acids Research*, 2016, **44**, D279–D285.

(360) R. Mosca, A. Céol, A. Stein, R. Olivella and P. Aloy, "3did: a catalog of domain-based interactions of known three-dimensional structure", *Nucleic Acids Research*, 2013, **42**, D374–D379.

(361) P. Stolorz, A. Lapedes and Y. Xia, "Predicting protein secondary structure using neural net and statistical methods", *Journal of Molecular Biology*, 1992, **225**, 363–377.

(362) J. Garnier, J.-F. Gibrat and B. Robson, "[32] GOR method for predicting protein secondary structure from amino acid sequence", *Methods in enzymology*, 1996, **266**, 540–553.

(363) T. Z. Sen, R. L. Jernigan, J. Garnier and A. Kloczkowski, "GOR V server for protein secondary structure prediction", *Bioinformatics*, 2005, **21**, 2787–2788.

(364) S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach", *Journal of Molecular Biology*, 2001, **308**, 397–407.

(365) X.-w. Chen and J. C. Jeong, "Sequence-based prediction of protein interaction sites with an integrative method", *Bioinformatics*, 2009, **25**, 585–591.

(366) Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen and D.-J. Yu, "Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests", *Neurocomputing*, 2016, **193**, 201–212.

(367) H. Hu, J. Li, H. Wang, G. Daggard and M. Shi, Proceedings of the 2006 workshop on Intelligent systems for bioinformatics-Volume 73, 2006, pp. 35–38.

(368) B. Rost, "[31] PHD: Predicting one-dimensional protein structure by profile-based neural networks", *Methods in enzymology*, 1996, **266**, 525–539.

(369) Z. Wang, F. Zhao, J. Peng and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields", *Proteomics*, 2011, **11**, 3786–3792.

(370) R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang and Y. Zhou, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning", *Scientific reports*, 2015, **5**, 11476.

(371) X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin and H. Yu, "Three-dimensional reconstruction of protein networks provides insight into human genetic disease", *Nature Biotechnology*, 2012, **30**, 159–164.

(372) P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas and D. N. Cooper, "The human gene mutation database: 2008 update", *Genome medicine*, 2009, **1**, 13.

(373) E. Guven-Maiorov, C.-J. Tsai and R. Nussinov, "Structural host-microbiota interaction networks", *PLoS Computational Biology*, 2017, **13**, e1005579.

(374) L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices", *Journal of Molecular Biology*, 1993, **233**, 123–138.