2018

Towards Biological Sequence Data Service with Insights

Huaming Chen
University of Wollongong, hc007@uowmail.edu.au

Jun Shen
University of Wollongong, jshen@uow.edu.au

Lei Wang
University of Wollongong, leiw@uow.edu.au

Chi-Hung Chi
CSIRO

Towards Biological Sequence Data Service with Insights

Abstract

Testable prediction outcomes generated by computational models based on available databases are the primary sources helping to design biological experiments. Although numerous databases have been designed by collecting data either only from literature manually or together with prediction outcomes from computational models, there is currently not a comprehensive data service framework delivering better insights for these results. In this paper, we introduce a biological sequence data service towards delivering deeper insights and helping better biological experiments design. The service includes following major components: a comprehensive database for storing biological data, data analytics tools for analysing biological data, and computational models for delivering testable prediction outcomes. Specifically, we present this service in a framework for studies on host-pathogen interactions. The design of this framework aims to improve the understanding of host-pathogen interactions. The relationships of hierarchical databases and their working mechanism, specifically between PPIs and DDIs, are also presented in this framework. Finally, the preliminary and practical experiences of building computational model for prediction is discussed.

Disciplines

Engineering | Science and Technology Studies

Publication Details

Chen, H., Shen, J., Wang, L. & Chi, C. (2018). Towards Biological Sequence Data Service with Insights. 2018 IEEE International Conference on Big Data (Big Data) (pp. 2847-2854). United States: IEEE.

Towards Biological Sequence Data Service with Insights

Huaming Chen[1], Jun Shen[1], Lei Wang[1], Chi-Hung Chi[2]
[1]*School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia*
[2]*Data61, CSIRO, Australia*
Email: hc007@uowmail.edu.au,{jshen,leiw}@uow.edu.au, chihung.chi@data61.csiro.au

***Abstract*—Testable prediction outcomes generated by computational models based on available databases are the primary sources helping to design biological experiments. Although numerous databases have been designed by collecting data either only from literature manually or together with prediction outcomes from computational models, there is currently not a comprehensive data service framework delivering better insights for these results. In this paper, we introduce a biological sequence data service towards delivering deeper insights and helping better biological experiments design. The service includes following major components: a comprehensive database for storing biological data, data analytics tools for analysing biological data, and computational models for delivering testable prediction outcomes. Specifically, we present this service in a framework for studies on host-pathogen interactions. The design of this framework aims to improve the understanding of host-pathogen interactions. The relationships of hierarchical databases and their working mechanism, specifically between PPIs and DDIs, are also presented in this framework. Finally, the preliminary and practical experiences of building computational model for prediction is discussed.**

***Keywords*-biological data service ; HP-PPI ; data mining ; machine learning**

I. Introduction

Considering data service as a popular tool in most scientific scenarios, the expectation from such a service is nowadays shifting from merely data searching to intensively data analysis. The extraordinarily expanding pace of data in volume, variety and value characteristics is bringing more attention on research towards the advancement of biology science [1], [2]. With regard to data service as a public online service for managing and analysing the scientific data, we particularly focus on the biological sequence data service, which specifically defines the framework for study on host-pathogen interactions (HPI).

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]. Focusing in omics data, recently both the online and offline data services are experiencing major changes meeting the demand of better data acquisition, storage, distribution and analysis [2]. These changes are mostly raised accordingly since the high-throughput technologies have been deployed in multiple omics areas.

It is witnessed that sequence data accumulated in a large scale via *in vitro* method from biology science is booming to challenge the traditional data service in cloud, not only from data volume aspect but also from data variety characteristic [2], [4]. The volume size benefits the analysis process while the variety of data sources deepens the understanding from these raw biological data [5] in proteomics area, where we consider both the protein-protein interactions (PPIs) and domain-domain interactions (DDI) as the extension components of biological sequence data service [6].

The transformation to knowledge stage delivers crucial insights in the omics study, specifically when focusing on host-pathogen protein interactions [7], [6]. In the post-genomics era, the omics study in host-pathogen interactions mostly relies on protein interactions analyses, which is named 'proteomics' [8]. As the basic functional units in living organisms, protein-protein interactions have directly or indirectly yielded biological functions, via generating the response of immune systems and transduction of biological signals and so on [7]. Targeting this purpose in conjunction with complete data service from cloud is the impulse to strengthen and distribute research in biology science. The host-pathogen interactions study will help to provide new insights in the pathogenesis, particularly for immune system response between host and pathogens.

These studies highly depend on the domain knowledge for host-pathogen interactions, from which we firstly expect building computational models to predict latent interactions. The testable prediction outcomes generated by the models are the primary sources to help biologists design biological experiments, which reduces the time- and labour-intensive experiments cost. Although there are several databases having been implemented as primary sources to provide data service for host-pathogen protein-protein interactions (HP-PPI) study[9], [10], these data services result in limited functions without offering opportunity in gaining insights from raw data and generating prediction outcomes.

In this paper, we extend the vision to introduce a biological sequence data service jointly with analytics for HP-PPI study. In order to capture the full scope of host-pathogen interactions, we present this data service with a framework covering data sources, data processing, database construction, computational modelling and data analytics for host-pathogen interactions. To cover a wider scale in

the framework, we also include the knowledge from not only the protein interactions but also the related domain interactions. By collecting the data sources from numerous verified databases, we also report the practical experiences of building computational model.

The rest of this paper is organized as follows. In section II, the related work is discussed. The framework of biological sequence data service is introduced in section III, while in section IV the practical experience of building computational models for prediction is reported. Section V is the discussion of the framework as well as the conclusion of this paper.

II. Related Work

Since data is playing an essential role in most of the research communities, there have been several leading researches taking the data benefit further into service, such as data as a service (DaaS) [11] and database as a service (DBaaS) [12]. The researches on these areas benefit from the development of data explosion and computational intelligence. Meanwhile, the online services are also booming for our daily life and research, including everything as a service (XaaS)[13], software as a service (SaaS) [14], micro learning as a service (MLaaS) [15] and so on.

On the other hand, the accumulation of data from science, specially from biology, is challenging the DaaS. Specialized in host-pathogen interactions area, the data are of critical meaning contributing to the understanding of biological functions, immune systems response and biological signals transduction. Above all, these data, among which the sequence data is dominant, present hierarchical meanings in terms of their characteristics.

Meanwhile, although the high-throughput technologies in biology have advanced the generation of omics data, the potential pairs number is huge, and the biological experiments are costly in terms of time and experiment resources to identify the protein interactions relationship between host and pathogens. Especially, the high false positives rate under biology experiments conditions should also be considered.

With regard to DBaaS applications in biology, there are currently some related databases for host-pathogen interactions. In [9], an integrated platform, named 'Pathogen Interaction Gateway' (PIG), was proposed as an integrated platform for host-pathogen PPIs. It includes experimentally verified HP-PPI and is manually updated. Some computational tools have also been integrated to strengthen the HPI study, primarily focusing on known interactions with visualization interfaces. Later on, in [10], the 'pathogen-host interaction search tool' (PHISTO) was built. It argues that the development of PHISTO contributes as an up-to-date and functionally enhanced source of host-pathogen interactions database, since the only available databases from [16] does not provide enhanced features for analysis. Another one is the Pathosystems Resource Integration Center (PATRIC), which delivers a variety of ways for researchers to access and store data [17]. It currently archives more than 1300 experimentally characterized HP-PPIs from various public repositories.

To further explore the power of data in supporting the scope of host-pathogen interaction study, one major challenge is to provide testable prediction outcomes from the computational models. Afterwards, the understanding of the host-pathogen interactions could be achieved by combining these interactions, including prediction outcomes and experimentally verified ones, with other databases, which we will focus on domain-domain interactions database as an example in this paper.

In this way, a comprehensive biological sequence data service to strengthen the study is required to harness the hierarchical meanings of sequence data, as well as the computational models outcomes, to be integrally studied in the framework.

III. Biological Sequence Data Service

In this section, the complete framework of biological sequence data service is presented. We will describe each of the components, and last part includes the details of three related services to functionally deliver our biological sequence data service - data as a service, software as a service and knowledge as a service. The key components of the framework are illustrated in Figure 1.

A. The Framework

We define the framework with several sequential stages. The various sources of data as the input of the framework are aligned, in which the host-pathogen protein interactions databases and protein characteristics databases are particularly involved as the scenario in our following study. Meanwhile, the whole proteome sequences information, which is considered as the primary information, are collected from UniProt [18] and filtered by the identified host and pathogen species. Upon these databases, the computational modelling step is carefully designed according to the biological sequence data. The preliminary goal of the computational model is to generate testable prediction outcomes, for which we have to delicately define the dataset, feature representation method and learning model. Lastly the results are jointly processed via the third step to deliver knowledge. To achieve these goals, we will have three important and related components to distribute the biological sequence data service, including database as a service, software as a service and knowledge as a service.

B. Database as a Service

As an initial service in this framework, a well-studied and comprehensive database is very important. To cover the most of host and pathogens interactions, several experimentally verified and manually updated databases should be included and manipulated with the redundancy analysis. As the first

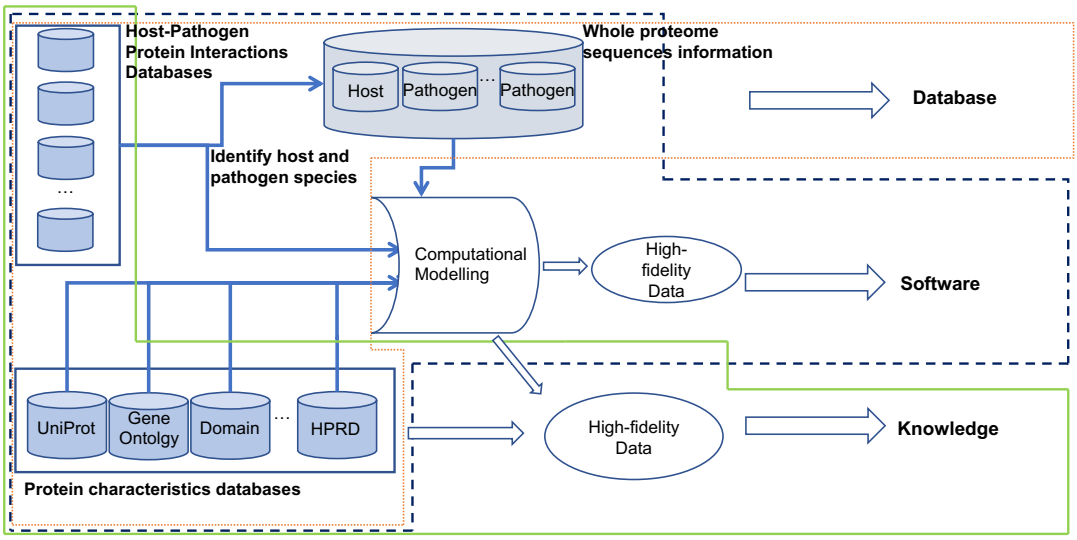


Figure 1: A Framework of Biological Sequence Data Service

step, the overlapping interactions from different databases and the interactions with high-confidence homology proteins are removed to build the whole database. After this preprocessing step, different protein characteristics databases are linked to the interactions databases to provide the details of each proteins.

There are several data sources for protein characteristics, including 'domain-domain interactions', 'human protein reference database' (HPRD), 'UniProt' and 'gene expressions ontology' (GO) databases. These databases can help us to design the computational model, and also to act as supplementary information to provide information for our data analytics.

The host-pathogen protein interactions databases only contain interacting proteins IDs, and this requires other data sources for protein characteristics to be replenished. On the one hand, researchers would be interested in looking into the relevant protein characteristics to understand the binding interfaces and interactions. On the other hand, the protein characteristics provide unique and important information for computational models construction.

C. *Software as a Service*

Simply providing database as an interaction tool for researchers is not enough. We further consider a computational model to help the experiments design on detecting potential host-pathogen interactions. The output of the model is thus an important source.

In details, the model has to incorporate different databases to construct an effective data representation algorithm. Thus, the first issue for the computational model is to curate the dataset. Even though we have clearly cleaned the data from various databases, they currently only represent the positive interactions IDs between host and pathogens.

To validate the feasibility in providing the software service to output predictions, Figure 2 details the steps required to formulate the service. A structured dataset, a well-defined feature representation algorithm and a learning model are the essential components. For host-pathogen interactions, the positive protein interactions from different databases will be firstly collected. Following the database as a service, a data pool containing the whole proteome sequences information is included regarding the host and pathogen species. The data pool contains all identified proteins IDs from related host and pathogen species. This pool is to define the unlabelled protein pairs, which will be handled with the data sampling method to build a discriminative dataset. The discriminative dataset is later used in the computational model, which we mostly consider supervised learning model as the primary approach. Ultimately, there should be both positive interactions and negative interactions in this discriminative dataset.

To learn from these datasets, the interactions data will be input into the model, which will require the IDs information being transformed into vectorized data incorporating the biological sequence data. Typically, the biological sequence data is the dominant information in the study of host-pathogen interactions. The well-known biology hypothesis tells us that 'amino acid sequence specifies the structure of protein then structure dictates their corresponding function' [19]. Thus, in this stage, the sequence data for proteins would be efficient for feature representation.

The prediction output from the computational model is of great interest for biologists as it provides testable data to further evaluate with experiments. As another important source to complete the whole proteome scale interactions between host and pathogens, it also offers great value on identifying the latent relationship between host and pathogens, and locating the infectious meanings for biology study.

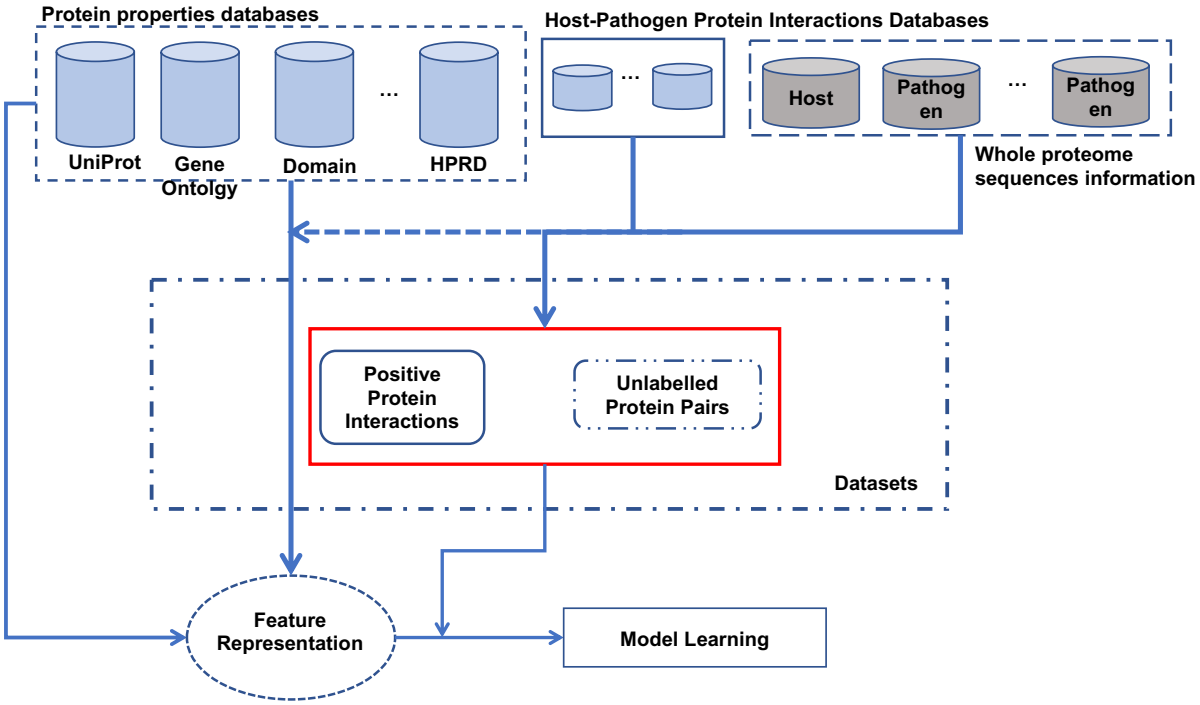


Figure 2: Computational Modelling for Software as a Service

D. *Knowledge as a Service*

Combining the testable prediction outcomes and experimentally verified host-pathogen interactions as a whole proteome interactions map, an interacting network on protein atom level could be finally reserved and presented. To expand the study between host and pathogens, the framework involves the system analyses with the exploration of other proteomics data. The goal is to engender a holistic view for researchers, whilst the analysis with specific domain knowledge is also considered. For host-pathogen interactions, we primarily consider the domain-domain interactions as they specify the exact physically interacting location between proteins.

In Figure 3, the domain-domain interactions are included as important supplementary data. Although the biology assumptions always propose that imitating the binding actions between proteins could be the primary infectious mechanism between host and pathogens, the general and possibly evolved principles for these HP-PPIs may be different and have not been well studied. The biological data mining, from both the whole proteome interactions map and domain interactions information, synthesizes the underlying statistic analysis.

As the final stage of data analytics for the predictions outcomes and experimentally verified data, the understanding from a fine-grained level interaction will help to investigate

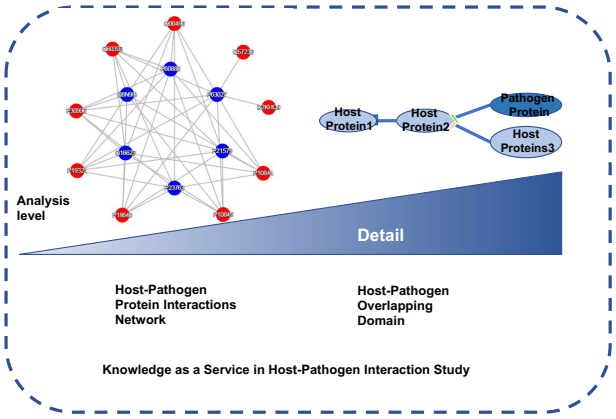


Figure 3: Insights Integration with Overlapping Domain for Knowledge as a Service

the infectious mechanism and assist the experiment design.

IV. Computational Model for Host-Pathogen Interactions

Though the framework delivers a system-level diagram on presenting specific biological sequence data with their domain knowledge, a core task beyond system engineering is to embed related computational model for prediction from the experimental data relationship. As the experimental data

is currently a small portion of the whole landscape for host-pathogen interactions, obtaining the testable prediction outcomes is still an important task, which result is dedicated to the software level and knowledge level service in the framework.

In details, Figure 2 includes data collection, feature representation and model learning as for a computational model. Particularly, we select *Yersinia pestis* as our exemplar study object in this paper. The associated data is accessed from the well-maintained databases *PATRIC* and *PHISTO*, which are manually uploaded as discussed in Section II.

Following we firstly incorporate the data statistics for *Yersinia pestis* [20]. Both the redundancy and homology of the raw data are considered during the data cleansing step. Redundancy of raw data exists within the different well-maintained databses while the homology of raw data were determined by the similarity of the corresponding sequence information.

These data indicate the interacting protein IDs, which does not contain any non-interacting protein data as well as sequence information by far. In most cases, supervised learning models are built to learn the data [20], [21], [22], which demands a discriminative dataset. Thus, data sampling in the unlabelled protein pairs is adopted. There are several studies discussing different sampling methods may have impacts on the prediction results. Some studies use ranking the unlabelled data to build the negative dataset as the data sampling method [23]. The ranking mechanism of unlabelled data is normally based on the information of sequence, gene ontology terms and the human protein reference database. Another one is random sampling [24], [25], which is the primary data sampling method in this regards. During the data sampling step, the sequence homologies analysis is also included.

Feature Representation The protein ID refers to a unique item in the protein databases, i.e. *UniProt*. Protein is a sequential amino acid combination, which includes 20 different types of amino acids in different lengths, i.e Arginine (Arg), Phenylalanine (Phe), Glycine (Gly) and Glutamic (Glu) and so on. These 20 types of amino acids exhibit different physicochemical properties. In terms of the feature representation algorithm for the protein, we consider these sequence information as the primary data and introduce the 'conjoint triad method' [22] to represent sequence information into a vectorized data, termed as '*k*-mer' feature.

As Table II indicates, these 20 types of amino acids are assigned to seven different groups according to their different physiochemical properties. Grouping the amino acids into seven groups facilitates a more efficient sequence coding method, which considers adjacent amino acids ranking and calculates the frequency of corresponding conjoint triad information. Figure 4 shows a basic diagram for *k*-mer feature encoding.

In Figure 4, a segment from human Histone H1 protein is chosen to illustrate the *k*-mer feature representation method. Normally, k is set to three though two- or four-mer may also be feasible. In this paper, we choose three to achieve a relatively effective feature representation method to avoid high-dimensional and sparse features. In Figure 4, $f_i(i = 1, 2, ..., 343)$ represents the derived frequency of different combinations within three adjacent amino acids. Sliding the three adjacent amino acids as a window through the whole sequence, f_i is calculated. However, some of the f_i may be zero as the corresponding combination does not appear in the sequence. In this segment, we could calculate $f_3, f_{10}, f_{16}, ..., f_{229}$.

So far, we have curated the dataset[1] and have included feature representation method, building a computational model for prediction is the next step.

Learning Model In our preliminary experiments, we firstly deploy random forest as our computational model, which was originally proposed in [26]. Derived from decision forests techniques, it further adopted random learning method to build a combination of decision forests. It is an ensemble learning model for classification, regression and so on. Based on the tree bagging method, it builds a bunch of random decision trees to avoid the bias problem occurring in singular decision tree.

The reason we choose random forest in this study is that, we witness the advantage of random forest in dealing with datasets which are curated by random sampling method [27], [28]. In this study, we utilise random sampling method to pair unlabelled data to build the negative interactions. The mechanism of bagging with replacement allows random forest to decreases the variance of model and achieve a better performance. Especially, when the experimentally verified data indicate the interacting proteins from host and pathogen are limited and far less than the proteins pool of host and pathogen, random forest would be an efficient and effective option for model learning.

In this study, we implement random forest by *sklearn* toolkit [29] in Python language. Since we have collected the dataset for pathogen *Yersinia pestis*, the training and testing data are divided from the dataset and the 10-fold cross validation method is included to evaluate the performance. Table III shows the statistics for *Yersinia pestis*.

Performance Evaluation As for the statistics for *Yersinia pestis*, an imbalanced ratio between positive pairs and negative pairs is chosen as $1:100$. It is designed to meet the biology scenario and request for a set of comprehensive performance evaluation metrics. In this study, we involve several metrics to evaluate the computational model, including precision, recall values, F1 score and accuracy result. The calculation for precision and recall values is as follows:

$$Precision = TP/(TP + FP) \tag{1}$$

[1]Dataset can be downloaded from: https://drive.google.com/drive/folders/1yz1Nc6qBrQ0ABnk5ZpMiTiBlucGx5e6r?usp=sharing

Pathogen	All Proteins	Experimental PPIs	Valid Positive PPIs	Invoved Proteins
Yersinia pestis	20226	4118	4045	1208

Table I: Data Statistics for *Yersinia pestis*

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

Table II: Group of 20 Basic Amino Acids [22]

Pathogen	Positive Pairs	Negative Pairs	Training Data	Testing Data
Yersinia pestis	4045	404500	367236	41309

Table III: Dataset for *Yersinia pestis*

$$Recall = TP/(TP + FN) \quad (2)$$

In Equation (1-2), 'TP' represents true positive number and 'FP' is false positive number. 'FN' means false negative number and 'TN' is true negative number. Following we give the definition of accuracy as well as F1 score. As for F1 score, it is normally measured based on precision and recall values.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

$$F1 = 2 * Precision/(Precision + Recall) \quad (4)$$

To better demonstrate the performance on classification task, especially for imbalanced HP-PPI dataset, receiver operating characteristic (ROC) is usually included by plotting a curve with different settings of true positive rate against false positive rate. These different settings refer to varied discrimination thresholds from the results. Meanwhile, AUC value can be calculated as the area under ROC curve, which means the value should be between 0 and 1. Normally, we define a better classifier based on a higher AUC value. We also show the results on ROC curve and the area under ROC curve (AUC) value in Figure 5(a) and Table IV respectively.

In Table 3, the results on these metrics are depicted as *0.9932* for precision, *0.6377* for recall, *0.7766* for F1 score and *0.9964* for accuracy. The overall AUC value is *0.9313*, which is relatively high in this HP-PPI task. These results

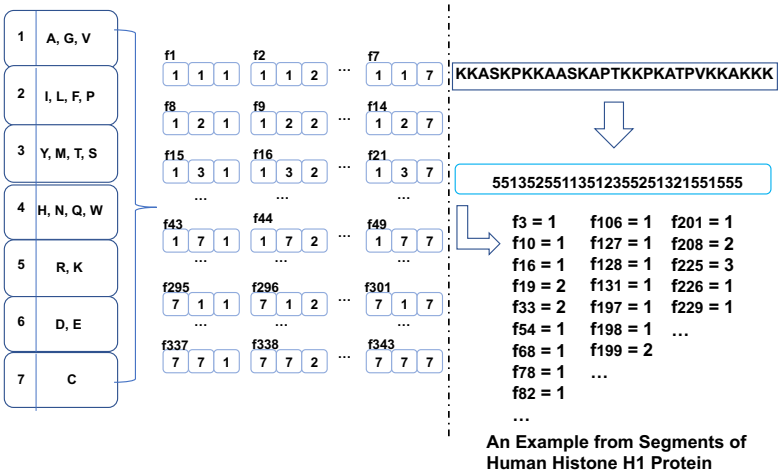


Figure 4: k-mer Feature Encoding ($k = 3$)

Table IV: Metric Results

Model	Precision	Recall	F1	Accuracy	AUC
Random Forest	0.9932±0.0054	0.6377±0.0142	0.7766±0.0100	0.9964±0.0001	0.9313±0.0066

are collected from the 10-fold cross validation experiments. We also report the variance in Table IV.

As random forest is chosen as the model, we further study the influence of the features for the prediction task. The ranking of features is conducted to achieve the importance score based on the contributions of each feature. Figure 5(b) illustrates the result of top 50 important features while their ID are listed in x-axis. An interesting finding regarding the feature ranking is that the top 50 features out of the 686-dimension vector feature are all contributed by pathogen proteins.

V. Conclusion

Soliciting the Internet database resources to amplify the data analytics and knowledge mining is the core of the study of host-pathogen interaction. In this paper, we present the framework considering the biological sequence data as a service, which formulate the insights from database manipulation, software development and knowledge discovery. With the contributions in composing the computational model and building the differential analysis level, the framework strives to deliver a comprehensive service towards understanding the host-pathogen interaction network and their internal infectious mechanisms.

However, as for future work, building DDI on top of the HP-PPI network to conclude the statistics analysis is also important to further strengthen the study, whilst a comprehensive computational model is desired on delivering testable prediction outcomes on different levels. In this study, random forest has demonstrated a powerful and efficient performance on *Yersinia pestis* dataset. We also anticipate the result of feature ranking, which shows a bias on one side proteins, may infer a more efficient feature representation method to derive a better computational model. We will further evaluate the different properties from the biological sequence data and develop machine learning based method towards better and more interpretable results.

Acknowledgment

This work is supported by the scholarship from the China Scholarship Council (CSC) and Faculty Strategic Investments Grant for DP 2019 development, while the first author pursues his PhD degree in the University of Wollongong.

References

[1] E. Williams, J. Moore, S. W. Li, G. Rustici, A. Tarkowska, A. Chessel, S. Leo, B. Antal, R. K. Ferguson, U. Sarkans, A. Brazma, R. E. Carazo Salas, and J. R. Swedlow, "Image Data Resource: A bioimage data integration and publication platform," *Nature Methods*, vol. 14, no. 8, pp. 775–781, 2017.

[2] 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 biology*, vol. 13, no. 7, p. e1002195, 2015.

[3] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[4] 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*, vol. 12, no. 5, pp. 1–26, 2016.

[5] S. Durmus, T. Çakir, A. Özgür, and R. Guthke, "A review on computational systems biology of pathogen-host interactions," *Frontiers in Microbiology*, vol. 6, pp. 1–19, 2015.

[6] 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*, vol. 6, pp. 11 760–11 771, 2018.

[7] 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*, vol. 16, no. 1, p. 1106, 2015.

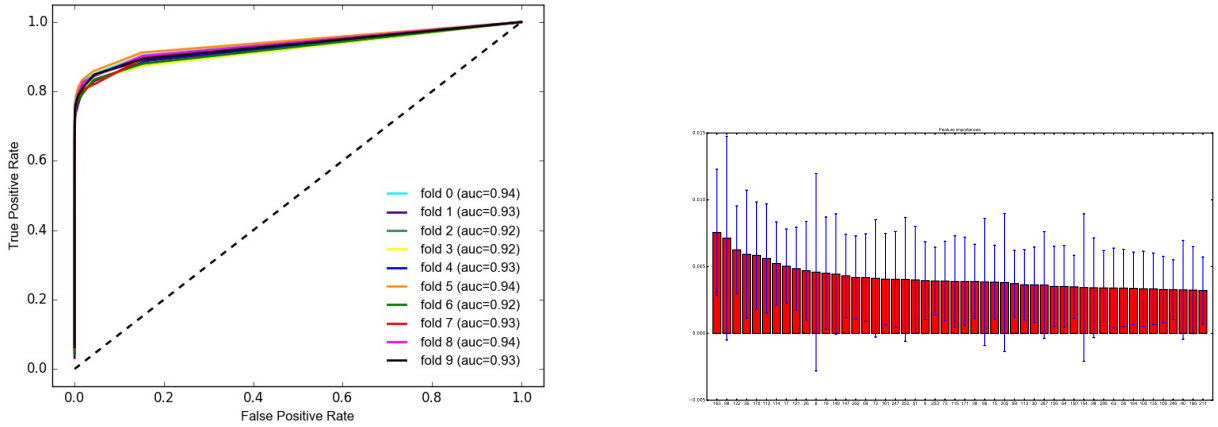
[8] M. R. Wilkins, R. D. Appel, J. E. Van Eyk, M. Chung, A. Görg, M. Hecker, L. A. Huber, H. Langen, A. J. Link, Y.-K. Paik *et al.*, "Guidelines for the next 10 years of proteomics," *Proteomics*, vol. 6, no. 1, pp. 4–8, 2006.

[9] T. Driscoll, M. D. Dyer, T. M. Murali, and B. W. Sobral, "PIG - The pathogen interaction gateway," *Nucleic Acids Research*, vol. 37, no. suppl_1, pp. 647–650, 2008.

[10] S. Durmuş Tekir, T. Çakir, E. Ardiç, A. S. Sayilirbaş, G. Konuk, M. Konuk, H. Sariyer, A. Uğurlu, I. Karadeniz, A. Özgür, F. E. Sevilgen, and K. Ö. Ülgen, "PHISTO: Pathogen-host interaction search tool," *Bioinformatics*, vol. 29, no. 10, pp. 1357–1358, 2013.

[11] J. Zhang, B. Iannucci, M. Hennessy, K. Gopal, S. Xiao, S. Kumar, D. Pfeffer, B. Aljedia, Y. Ren, M. Griss *et al.*, "Sensor data as a service–a federated platform for mobile data-centric service development and sharing," in *Services Computing (SCC), 2013 IEEE International Conference on*. IEEE, 2013, pp. 446–453.

[12] W. Lehner and K.-U. Sattler, "Database as a service (dbaas)," in *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. IEEE, 2010, pp. 1216–1217.



(a) ROC Curve

(b) Top 50 Features Ranking

Figure 5: Results of Random Forest on *Yersinia pestis*

[13] P. Banerjee, R. Friedrich, C. Bash, P. Goldsack, B. Huberman, J. Manley, C. Patel, P. Ranganathan, and A. Veitch, "Everything as a service: Powering the new information economy," *Computer*, vol. 44, no. 3, pp. 36–43, 2011.

[14] T. Kwok, T. Nguyen, and L. Lam, "A software as a service with multi-tenancy support for an electronic contract management application," in *Services Computing, 2008. SCC'08. IEEE International Conference on*, vol. 2. IEEE, 2008, pp. 179–186.

[15] G. Sun, T. Cui, J. Yong, J. Shen, and S. Chen, "Mlaas: A cloud-based system for delivering adapative micro learning in mobile mooc learning," *IEEE Transactions on Services Computing*, pp. 1–14. doi: 10.1109/TSC.2015.2 473 854, online first on 27 August, 2015.

[16] R. Kumar and B. Nanduri, "Hpidb-a unified resource for host-pathogen interactions," in *BMC bioinformatics*, vol. 11, no. 6. BioMed Central, 2010, p. S16.

[17] 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*, vol. 42, no. D1, pp. D581–D591, 2013.

[18] U. Consortium, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2016.

[19] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*. Garland Science, New York, 2008.

[20] H. Chen, J. Shen, L. Wang, and J. Song, "Towards data analytics of pathogen-host protein-protein interaction: a survey," in *Big Data (BigData Congress), 2016 IEEE International Congress on*. IEEE, 2016, pp. 377–388.

[21] J. Song, F. Li, A. Leier, T. T. Marquez-Lago, T. Akutsu, G. Haffari, K.-C. Chou, G. I. Webb, and R. N. Pike, "Prosperous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy," *Bioinformatics*, vol. 34, no. 4, pp. 684–687, 2018.

[22] 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*, vol. 104, no. 11, pp. 4337–4341, 2007.

[23] S. Mei and H. Zhu, "A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks." *Scientific reports*, vol. 5, p. 8034, 2015.

[24] F. E. Eid, M. Elhefnawi, and L. S. Heath, "DeNovo: Virus-host sequence-based protein-protein interaction prediction," *Bioinformatics*, vol. 32, no. 8, pp. 1144–1150, 2016.

[25] G. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," *BMC Bioinformatics*, vol. 13, no. Suppl 7, p. S5, 2012.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC medical informatics and decision making*, vol. 11, no. 1, p. 51, 2011.

[28] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, 2015.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.