2017

Towards Elucidating the Structural Principles of Host-Pathogen Protein-Protein Interaction Networks: A bioinformatics survey

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

Jiangning Song
Monash University, jiangning.song@monash.edu

Geng Sun
University of Wollongong, gs147@uowmail.edu.au

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

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

Publication Details

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. 2017 IEEE 6th International Congress on Big Data (pp. 177-184).

Towards Elucidating the Structural Principles of Host-Pathogen Protein-Protein Interaction Networks: A bioinformatics survey

Abstract
The ultimate goal of systems biology research area is to accurately predict the behavior of biological systems through the construction of computational models, using the related molecular-level data as the input, especially when the structural information of such biological system is available. Combining the three-dimensional (3D) structural information of the cohort of macromolecules underpinning the biological system, the researchers are poised with an unprecedented opportunity to gain a full understanding on how the molecules interact with each other, particularly for an interaction network, e.g. protein-protein interaction networks. Specifically, there are currently a limited number of studies focused on the reconstruction and modelling of the structural interaction networks (SIN) between hosts-pathogens protein-protein interaction networks. In this paper, we will survey the SIN on protein-protein interactions network, in which we focus on the interactions between pathogen and host species (PHPPI). As one of the most important component of inter-species PPI study, in-depth study of PHPPI at atomic-resolution level would reveal novel insights into the underlying principles of the organization and complexity of host-pathogen PPI networks. Several related sub areas are discussed, and the related typical Big Data methods including machine learning methodologies and statistics models will also be discussed. This paper contributes to a new, yet challenging, research area in applying data analytic and machine learning technologies in bioinformatics.

Keywords
structural, principles, towards, survey, protein-protein, elucidating, host-pathogen, interaction, networks:, bioinformatics

Disciplines
Engineering | Science and Technology Studies

Publication Details
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. 2017 IEEE 6th International Congress on Big Data (pp. 177-184).

Towards Elucidating the Structural Principles of Host-Pathogen Protein-Protein Interaction Networks: A bioinformatics survey

Huaming Chen[1], Jiangning Song[2], Geng Sun[1], Jun Shen[1], Lei Wang[1]
1-School of Computing and Information Technology, University of Wollongong, Wollongong, NSW, Australia
2-Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Victoria, Australia
Email: hc007@uowmail.edu.au, jiangning.Song@monash.edu, gs147@uowmail.edu.au, {jshen, leiw}@uow.edu.au

***Abstract*— The ultimate goal of systems biology research area is to accurately predict the behavior of biological systems through the construction of computational models, using the related molecular-level data as the input, especially when the structural information of such biological system is available. Combining the three-dimensional (3D) structural information of the cohort of macromolecules underpinning the biological system, the researchers are poised with an unprecedented opportunity to gain a full understanding on how the molecules interact with each other, particularly for an interaction network, e.g. protein-protein interaction networks. Specifically, there are currently a limited number of studies focused on the reconstruction and modelling of the structural interaction networks (SIN) between hosts-pathogens protein-protein interaction networks. In this paper, we will survey the SIN on protein-protein interactions network, in which we focus on the interactions between pathogen and host species (PHPPI). As one of the most important component of inter-species PPI study, in-depth study of PHPPI at atomic-resolution level would reveal novel insights into the underlying principles of the organization and complexity of host-pathogen PPI networks. Several related sub areas are discussed, and the related typical Big Data methods including machine learning methodologies and statistics models will also be discussed. This paper contributes to a new, yet challenging, research area in applying data analytic and machine learning technologies in bioinformatics.**

Keywords: host-pathogen protein-protein interactions; structural interaction network; bioinformatics

I. INTRODUCTION

Owing to the development of advanced high-throughput technologies, an overwhelming avalanche of experimental data has been rapidly accumulated in recent years and accordingly this phenomenon has propelled hypothesis-driven biomedical research into the 'big data' driven era. The availability of large-scale multi-omics data, including proteomics data from The European Bioinformatics Institute (EBI) [1, 2, 3] and genomics data from The Cancer Genome Atlas (TCGA) [4], provides an unprecedented opportunity to transform the biomedical research onto system-level, mechanistic studies aimed at a comprehensive and holistic understanding of biological systems [5]. The combination of experimental data and systems biology techniques present a more promising and more precise modeling alternative option for researchers. Although there are still challenges for systems biology, e.g., specialized domain knowledge and data issues, this data-driven work to gain deep understanding of biological systems from huge amount of raw data is currently in the spotlight of both the academia and industry [6]. In this paper, we focus on the proteomics data, specifically on host-pathogen protein-protein interactions data, to present a comprehensive survey towards structural principles analytics.

Given a set of interacting molecules, systems biology aims to understand and further predict the behavior of biological systems [7]. Thus, systems biology consists of studies on functional genomics and molecular biology. There are several researches focusing on genomics data since a nearly complete map of human and other species had been provided with the development of genome-sequencing projects [7]. These studies provided the insights towards understanding gene-related networks. Basically, a full understanding of how the set of molecules interact with each other requires heterogeneous data [8]. Among these data, three-dimensional (3D) structures of these molecules are the most critical ones.

Proteomics is an important area in bioinformatics, in which the interaction network and structural information researches remain as hot topics for decades. However, due to the limited availability of proteomics data, most of the researches were carried out within the same species, which is called "intra-species PPIs". Recently, several studies have shown their improvements in PPIs between different species, which are concerning "inter-species PPIs". This kind of PPIs offers important information for further analysis of infectious mechanisms between different species. In this paper, we focus on the PPIs between the host and pathogen, in which we benefit from the identified data collected via open databases [9]. These PHPPIs are experimentally verified and manually recorded in systems. They include the information of the infection pathways in their interactions network and they can reveal much more information in the infection mechanisms between hosts and pathogens.

In one of our recent work [9], a basic sequence information based survey of PHPPI was presented to exploit the online available and experimentally verified PHPPIs data. Beyond classifying pairs of proteins as interacting or not, in this paper we further reach out to a comprehensive study on building structural interaction network for PHPPIs, since systems biology might provide a highly convincing network analysis and also bring trustworthy statistics in cooperation with the corresponding structural information and domain data, on top of the atomic resolution level networks.

Structural interaction network (SIN) is an atomic-resolution protein-protein interaction network with structural

detail by combining the structural information of each of the proteins [10]. The structural information of proteins is another main experimentally determined 3D structural data that has been published already. Since there are few studies based on 3D structural detail to provide an atomic mechanism view of PHPPIs, we hope to take stock of the progress that biologists have made in bioinformatics area, including the well-maintained 3D structural databases and analysis based on these structural information, and further help readers navigate through the gap between biology analysis and computational model building.

Thus, this paper contributes to a comprehensive survey on:

1) Review on current protein structure prediction task targeting on secondary and teritary structure and also the domain-domain interaction prediction task based on machine learning technologies;

2) Review on structural interaction network, including the building process and statistical analysis.

The rest of this paper is organized as follows: Section II describes the 3D structure and domain information of proteins; Section III introduces the related public databases; Section IV discusses a variety of machine learning algorithms that have been developed and applied in protein 3D structure and domain prediction, while Section V describes a detailed process to layer curated 3D structural models on top of traditional interaction network, and also provides the linking domain knowledge between model and analysis; latter the challenges for building structural interaction model are discussed in Section VI. We conclude this paper in Section VII.

II. PROTEIN STRUCTURE

Since both structure and domain data are usually difficult for bioinformatics researchers to fetch, they are currently two hot topics and remains much for future researches. To build SIN, a good and complete understanding on domain-domain interactions is also important. In this section, we first present the biological meaning for both structural information and domain-domain interactions.

A. Structural Information

It is well known that amino acids are the basic units to build the protein. Their direct concatenate string becomes the sequence information of proteins. In [9], we give a detailed discussion of the 20 different proteinogenic kinds of amino acids and the sequence information of proteins. However, we have identified that there are 25 different expressions of amino acids existing in the human and pathogens protein sequence information in our PHPPI researches. The other five expressions of amino acids are Sec (Selenocysteine/U), Pyl (Pyrrolysine/O), Asx (Aspartate or Asparagine/B), Glx (Glutamate or Glutamine/Z) and an unknown (X). There are 20 different kinds of amino acids from [11].

There are four distinct structural stages for protein sequence, which are primary structure, secondary structure, tertiary structure and quaternary structure respectively. Since the protein sequences have various lengths, for those which are composed of less than 50 amino acids, regularly only the primary level information available. This kind of protein sequence is called polypeptide. For the secondary structure, it is recognized as regions in which the sequence forms the most common structures: alpha helices (α-helix) and beta sheets (β-strand). Another structure is called random coil (C) which is not a secondary structure. But it is also included as one of the features to present the absence of regular secondary structure for proteins. Upon folding, a secondary structure subunit transforms into a tertiary structure. For some proteins, they consist of more than one polypeptide, which means there is more than one tertiary structure. This context information of how these polypeptides fit together along their subunits is called quaternary structure.

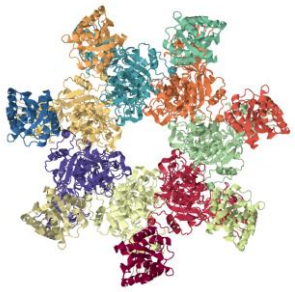

Figure 1. The 3D structure of the Protective Antigen (UniProt ID: Protein: P13423)

Among these four distinct structural stages, each stage is highly related to its prior stage. Understanding protein structures is critical for protein analysis. Meanwhile the ongoing experiments to determine these structures keep increasing the known protein structures for biologist and biochemists. However, these experiments, which are normally conducted by using X-ray crystallography, NMR spectroscopy and even cryo-electron microscopy, are extremely time-consuming and expensive. It is reported that so far only about less than 0.5% of all sequencing proteins structures have been known due to these limitations of biological experiments methods [12]. Hence, the researches on protein structures took place first on secondary structure prediction decades ago. Because the secondary structure could be analyzed with the efficient sequence information from primary structure, it has been a hot topic till now. As shown in Figure 1, it is an illustration of secondary structure for Protective Antigen protein (the UniProt ID is 'P13423'). As mentioned earlier, the secondary structure is pre-defined with three types: α-helix, β-strand and coli, which is called Q3 accuracy in the prediction task [13, 14, 15, 16]. Ranging from statistics models to machine learning methods, Q3 accuracy has been intensively improved from 65% to 80%. Recently a more challenging problem targeting on eight categories prediction (Q8) for secondary structure is drawing the researchers' attention. The eight categories refine the secondary structure to more elements: 3_{10}-helix, α-helix, π-helix, β-strand, β-bridge, β-turn, bend and loop/irregular [17, 18].

To achieve a result with better accuracy on secondary structure, it requires not only an efficient model but also

sufficient feature representation from sequence information. The involved models will be introduced in Section IV. The key challenge to predict secondary structure is the prediction for those proteins which have no close homologs, which in turn have experimental verified 3D structures.

To achieve sufficient feature representations for the secondary structure prediction, most studies introduce the protein sequence information, amino acid profile information, local and global information of sequence [14,16,19,20]. In this paper, we first focus on eight categories secondary structure prediction, which has been intensively studied recently due to its complexity.

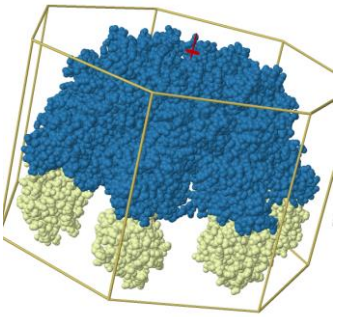

Figure 2. Tertiary structure of the Protective Antigen (UniProt ID: Protein: P13423)

Figure 2 provides an example of the tertiary structure of the Protective Antigen protein (UniProt ID: P13423). Aside from secondary structure prediction, prediction for this level structure normally falls on homology modeling method [21]. The homology modeling is also known as comparative modeling, in which the main result candidate comes from the amino acid sequence alignment by mapping the amino acid between different sequences. The reason on introducing homology modeling method into tertiary structure prediction is that, the evolutionary results show similar protein in amino acids sequence share similar tertiary structure to accomplish related biological function [22].

The structure information is requisite for structural interaction network since they provide the atom level information of protein sequences. In section III, we will detail the related databases to acquire such information.

B. Domain- Domain Interactions

Given a protein sequence, protein domains are distinct functional or structural subunits. Most of the protein domains build independently stable and folded 3D structures, with which the domains could be combined into different arrangements to form a unique protein with different functions [23]. Therefore, the binary PPI networks can be further considered at the domain level, especially when the interacting protein has an extremely long length. Although most proteins consist of multi domains, a pair of protein-protein interactions often involves only one pair of domain-domain interaction.

The domain level interaction provides a global view of the binary PPIs network. For PHPPIs researches, it reveals the actual interacting location for pathological interactions and can help to facilitate the drug development targeting on infectious diseases. To acquire the comprehensive understanding of how interactions between domains are mediated, the primary method is to analyze every single interacting protein with their experimentally determined 3D structures. However, this kind of information remains only a small fraction for proteins, which means the domain level PPIs interaction data are not readily fully accessible.

There are several existing databases, i.e. 3did [24] and iPfam [25]. They provide domain-domain interactions by identifying them based on experimentally determined 3D structures. Also, there are other databases providing combined interactions, in which part of them are from experimentally determined data and the rest are from computational predicted result. For example, DOMINE [26] includes both 3D structure-based and predicted domain-domain interactions datasets. Moreover, DOMINE indicates the predicted domain-domain interactions with three different levels, namely 'High', 'Middle' and 'Low'. Two primary methods, which are association method [26] and maximum likelihood estimation [27], are introduced in this domain-domain interaction prediction task. The essential information utilized in these models includes the domain information from protein sequence and binary protein-protein interaction information.

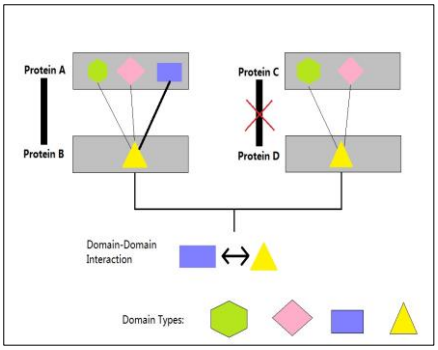

Figure 3. Domain-Domain Interaction

In order to provide a general understanding of domain-domain interactions associated with binary protein-protein interactions, Figure 3 shows a basic diagram for domain-domain interaction prediction task from [28]. 'Protein A' is interacting with 'Protein B' while 'Protein C' is not interacting with 'Protein D'. Several different domains types are identified using the related databases; mostly we would choose Protein Data Bank (PDB) [29] as most of the literature suggested so. Later we will compare the difference between these two groups of domain-domain relationships to identify the exactly interacting domains between two different proteins, in this example they are the purple box and yellow triangle.

III. RELATED DATABASES

Ranging from protein sequence information to their structure data, several different databases are available on the Web and they are also well maintained. These databases include host-pathogen protein-protein interactions databases,

structure databases, protein families and domain databases, as well as domain-domain interactions databases.

A. Host-Pathogen Protein-Protein Interactions Databases

Although several different standardized formats for the host-pathogen protein-protein interactions are published by different organizations, these databases contain the most important binary information for PHPPIs researches. Some popular repositories are initially built by universities, which include HPRD by Johns Hopkins University and the Institute of Bioinformatics, PATRIC by University of Chicago, PHISTO by Boğaziçi University, VirHostNet by Université de Lyon. The highly credible positive PHPPIs pairs are manually recorded in these systems and updated periodically. The details of these databases could be found in [9].

B. Structure Databases

Protein Data Bank (PDB) [29] is the primary database for the structural information of proteins, which is managed by the worldwide Protein Data Bank (wwPDB) international collaboration.. The PDB database contains all experimentally determined protein structure ranging in different resolutions and different detection methods.

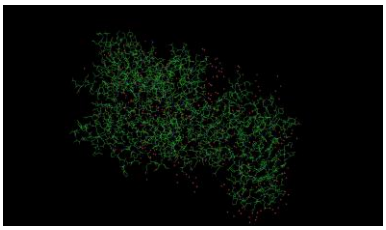

Figure 4. 3D Virsulization of of the Protective Antigen (UniProt ID: Protein: P13423)

PDB is currently updated weekly. It 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. Beside the critical information, abbreviation content about the corresponding literatures is mandatory in PDB file, which is listed as Header. Several other specific columns are: HEADER (The ID NO., date of publication), OBSLTE (mark for obsolete or not), TITLE (details about the related experimental methodology), COMPND (molecular components of the complexes), SOURCE (the source of the complexes), EXPDTA (the experimental method for determining the structure), AUTHOR (the authors), SPRSDE (the modification and revocation records), and REMARK (including the related literatures, the maximum resolution and other statistic).

To illustrate a lively picture for the corresponding PDB file in 3D vision, we present a simple example of the Protective Antigen (UniProt ID: P13423) using PyMOL [30].

However, it costs lots of efforts and time to acquire an experimental determined structure for protein, and currently not every protein has its corresponding structural information available. How to determine those proteins without PDB data is crucial for building SIN.

C. Protein Families and Domain Databases

Acting as an important database of protein domains and families, Pfam provides a complete map for protein domains and families searching [31, 32]. It is regularly updated and the latest version is Pfam 31.0 released in March 2017. It contains more than 16,712 protein families.

Although amino acids are the elements to compose a protein sequence, the actual function execution takes place with multi sequential amino acids, which is called domain. Different combinations of domains result in various functions of proteins. Identifying these domains in proteins would give deep details and insights for their function mechanism.

With structural information, the bond of interactions between proteins is more concrete along the sequences than the binary PHPPIs network provided in PHPPIs databases. Therefore, iPfam is introduced in SIN study to acquire domain-domain interactions between proteins [25]. iPfam is developed by Howard Hughes Medical Institute, and currently it hosts more than 9,500 domain-domain interactions. iPfam is based on two continuously updating databases, PDB and Pfam. Both are well established for their 3D structure and domain information purposes. Most of the structural information in PDB also contains multiple domains. 3did is another domain-domain interaction databases for 3D interacting domains between proteins. It is a collection of protein interactions from which high-resolution 3D structures are known [24, 33].

By introducing iPfam and 3did to achieve domain level resolution of PHPPIs, SIN considers proteins in their precise spatial relationships by layering domain-domain interactions on the top of conventional protein-protein interactions network. As protein sequence information are accumulated in a staggering rate, these data depict its characteristics with high volume, high velocity, high variety, high value and high veracity (5V). It brings a joint possibility by adopting big data analytics, including machine learning technologies, to tackle the structural and domain-domain interaction prediction problems. In the next section, we will introduce the related computational models or methods for SIN construction, among them machine learning methodologies are mostly utilized recently.

IV. MACHINE LEARNING METHODOLOGIES

Before we can layer the domain-domain interactions upon the traditional PHPPIs network, the structural information of corresponding proteins is requisite. However, only a few proteins have experimentally determined structure, specifically with high resolution scale. Thus, we herein present the related studies for structure prediction, and also domain-domain interactions prediction in this section. Firstly, to input the related information, which are mainly from sequence information, proper data processing for protein is required. Given a protein sequence denoted as $X = x_1, x_2, x_3, \ldots, x_n$, x_i presents each amino acid in this sequence.

1) Sequence Information: The amino acid information is essential to protein sequence. Normally, x_i could be a l-dimensional feature vectors which is one-hot sparse vector (l

denotes the amino acid types considered in the related projects). In our previous study [9], the protein sequence is decomposed into a 25*n dimension vector. In [18], the utilized database was from PISCES Cull PDB server and the ultimate datasets was filtered based on non-homologous principles. Their sequence information is with 22*n dimension while in [16] the dimension is 21*n. This one-hot sparse vector representation method is widely used in recent structure prediction researches. Another type of feature representation based on sequence information is the evolutionary information as position-specific scoring matrix (PSSM). PSI-BLAST[34] is one of the most frequently used tools to derive PSSM from protein sequences. The generated matix is also with $b * n$ dimension, in which b is the types of amino acids considered in the protein sequence.

2) Global/Local Information: The global information from the whole protein sequence is also crucial to improve the accuracy [18], although protein sequence information is considered as the main feature for secondary structure prediction. Before folding to build a tertiary structure, the secondary structure remains in a two-dimensional space, which is stabilized by hydrogen bonds between different amino acids located in different locations in the protein sequence. The local information could also be generated by dividing protein sequence into several segments. Thus capturing this kind of global/local information is resonable and widely believed to improve the eight categories accuracy.

Beyond the data processing, different data query procedures are required to collaborate with the specific domain knowledge. However, most of the datasets are built with the sequence information. To deal with these datasets, mostly statistics analytics and machine learning methods are utilized. In the following we will present the most relevant machine learning methods through their typical sample applications in detail.

A. Bayesian Statistics

The earliest studies on protein secondary structure prediction mainly focused on Bayesian statistics method [35,36,37]. Basically, the Bayesian statistics described this problem by:

$$\mathrm{I}(\mathrm{S};\mathrm{R}) = \log[P(S|R)/P(S)] \qquad (1)$$

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

In such methods, the cooperation with information theory to project the known twenty amino acids types for each specific secondary structure could achieve a Q3 accuracy of 73.5%. Specifically, in [36] the GOR method (Garnier-Osguthorpe-Robson) is based on the information theory, which uses a 17-amino-acid sequence window to extract properties from protein sequence. The GOR method in [36] presented the observed frequencies of single, then pairs of residues on a local sequence of 17 residues to build the Bayesian model, then to estimate the probabilities for the Q3 structures. This method increased the accuracy from 55% up to 64.4%.

B. Support Vector Machine (SVM)

The debut to predict protein secondary structure was firstly introduced in 2001 [38], though support vector machine was proposed in 1995 [39]. It is not the first machine learning approach for protein secondary structure prediction, yet by then it achieved the best performance overall on Q3 task by its first use of the SVM approach.

Similar to earlier research with neural network based method [40], the encoding scheme for input layer is called local coding scheme. It denotes every amino acid with a 21-dimensional orthogonal binary vector as follows:

$(1,0,\dots,0) or\ (0,1,\dots,0), etc$

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

[38] considered SVM as a superior model by then with its attractive characteristics, including the effective avoidance of overfitting and the ability to handle large feature spaces. In details, the authors [38] selected the radial basis function (RBF) as the kernel function to train the SVM. Their result on Q3 task is 73.5%.

C. Artificial Neural Network

To the best of our knowledge, artificial neural network was first introduced in protein secondary structure prediction in [40] with the fully connected three-layer network. The learning algorithm is Back-Propagation algorithm. Later, the authors in [41] used a two-tier architecture to deploy neural network for prediction. However, the improvement for Q3 accuracy has been stalled since then.

Recently, Q8 accuracy has come into the spotlight of academia and industry, which aims to apply deep learning techniques to improve the performance. In [42], probabilistic graphical models, which combine conditional neural fields (CNFs) with neural network, were deployed to improve the Q8 accuracy. The features are extracted from PSSM (position-specific score matrix) and the physico-chemical property of the amino acids. According to [42], 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.

In [18], generative stochastic networks (GSN) model was utilized to learn a generative model of data distribution without explicitly specifying a probabilistic graphical model. Specifically, this supervised extension of GSN is deployed via learning a Markov chain to sample from a conditional distribution for training on protein structure prediction task. They presented this model with deep learning techniques to tackle Q8 problem for protein secondary structure prediction. The empirical design for the data preprocessing step in their work was to choose 700 lengths as the cutoff threshold value

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 [16], which proposed a deep convolutional and recurrent neural network. The feature to encoding the protein sequence remains partially the same as 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. Later on, multi CNN layers and stacked bidirectional RNN layers were included to learn both local context information and global context information from the denser matrix. A fully connected and softmax layers were layered on the top of the model to build the classifier for prediction task.

D. Random Forests

Apart from predicting secondary structure, domain-domain interaction is also crucial to build our SIN. Random forests model was introduced to build multi classifiers to vote a final decision for a dataset with 1050-dimensional feature [43]. Also in [44] an ensemble model of random forests and SVM was presented to predict the domain interacting sites.

Derived from decision trees model, random forest leverages the power of randomization to increase the model performance [45,46]. Random forest is able to deal with imbalanced data problems via the voting mechanism, whilst its random feature selection method may benefit the model from high dimensional data.

Various models have been discussed in this section for these problems; however we would mainly aim to stack these different types of data on the top of traditional PHPPIs network to achieve a structural principles analysis. In the next section, we will discuss the structural interaction network.

V. STRUCTURAL INTERACTION NETWORK

Since the principles analysis of protein interactions between pathogen and host still remain poorly understood, an ensemble network of traditional binary PHPPIs network and structural information gives an efficient option for mining these knowledges with a systems biology approach.

In [47], altogether 3,949 genes, 62,663 mutations and 3,453 associated disorders are analyzed based on a three-dimensional, structurally resolved human interactome network. Integrating the data from iPfam, 3did and Human Gene Mutation Database (HGMD) [48], authors of [47] successfully built a high-quality binary PPIs network with the atomic-resolution interfaces. This network provides some deep insights including in-frame mutations locations and the disease specificity for different mutations of the same gene, which could not be acquired on a low-resolution network. The original interactions network obtained from literature-curated databases in [47] have 82,823 pairs; however, after filtering out the proteins without experimentally determined structures, only 4,222 structurally resolved interactions between 2,816 proteins are kept. To build a structural interaction network still requires much more efforts on the structure experimental determination or computational prediction since only a tiny fraction of these binary PPIs can be analyzed with their corresponding structure information.

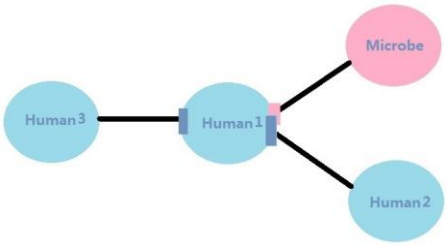

Figure 5. An Example for Domain-Domain Interactions Analysis [49]

Fortunately, in [10] we have witnessed that several possible structural principles analyses had been obtained within the human-virus protein-protein interaction network. The SIN approach in human-virus PPIs network reveals atomic resolution, mechanistic patterns, and gives systematic comparison with the human's endogenous interactions. Figure 5 is an example from [10, 49] to details how to layer the structure and domain-domain interaction information on top of the traditional PPIs network.

Figure 5 reveals a high-resolution relationship between the protein "Microbe" and the protein "Human2" as some overlapping area is detected. This kind of information could not be observed in the binary PPI network. Further analysis reveals that protein "Microbe" is mimicking the action of protein "Human2".

The experimental host-pathogen PPI networks provide not only specific pathogen protein functions but also global analyses, which reveal the critical proteins in the networks [49]. Although Figure 5 provides some essential mappings via domain-domain interactions, annotating the experimental host-pathogen PPI networks with 3D structural information may provide further information, because the protein-protein interactions can be interacted between two globular domains and also between one short linear motif (a short functional segment considered on secondary structure) and globular domains.

Several methods to assemble structural information with binary PHPPIs network could be:

1) Using only the experimentally determined structural information

2) Using both the experimentally determined and computationally predicted structural information

3) Using only the computationally inferred structural information

In [10] the computationally predicted structural information mainly comes from the homology modeling method which is widely used in bioinformatics area, because it is widely understood that the structure and function of protein are mostly determined by their sequence information.

Typically, for host-pathogen protein-protein interactions, hypothesis always exhibits that imitating the binding actions

between proteins is the main infectious mechanism. Given a SIN, there are several statistics data could help us to propose and support this hypothesis. As a specific example between virus and host PPI networks, [10] analyzed the exogenous and endogenous interactions in the human-virus SIN. Meanwhile, the overlapping ratio of protein interactions involved in exogenous interface and protein interactions involved in endogenous interface indicate the potential infectious targets, though the mapping of endogenous interfaces in [10] is not guaranteed to be complete.

To achieve a better understanding of the mimicry mechanism, which provides possible explanation for virus infectious procedure, a similarity statistical analysis can be carried out by z-score [50] and E-value [51] level. Since the mimicry action occurs between host protein and pathogen protein, similarity statistics may help to bring up insightful findings.

Overall, SIN on the top of binary protein-protein interactions exhibits many advantage with precise analysis based on the statistics from 3D structural and domain information.

VI. CHALLENGES

While the boom if big data research looks promising, when dealing with both the structural information and domain-domain interactions, there are several challenges to build SIN for PHPPI.

A. Feasible and Efficient Feature Representation

For the computational model, especially for protein sequence researches, feature representation remains a hot and challenge topic. Various methods for feature representation already exist [9, 13, 15, 17, 19, 20, 34]. One reason for us to reconsider this problem is that currently more and more protein sequence information are experimentally determined. Meanwhile, more and more models based on deep learning techniques present end-to-end frameworks for learning from big data sets. The automatic feature extraction process could be a promising option for protein sequence researches.

Prior to inputting data into machine learning models, several traditional feature representation methods, which include one-hot vector method, PSSM feature, and global/local information transformation method, are widely used. Recently, deep learning techniques are also first introduced in protein secondary structure prediction task in [16, 18]. In terms of feature representation, deep learning techniques could harness the power of a rich and high dimensional data in large volume. This could be a good opportunity for us to obtain more feature information and further improve the model performance.

B. Imbalanced Data

For both structure prediction and domain-domain interaction problems, the imbalanced ratio between different classes is also crucial to improve the models performance. In [43], the ratio of non-interface interactions to interface interactions is about 9:1. In structure prediction task, the ratios in both Q3 and Q8 tasks are also different and imbalanced between different protein families.

With the continuous expansion of structural information and domain data being available, the imbalanced data issue in biology area becomes more intensive. A possible solution could be either from consideration at data level or via innovative algorithm design.

VII. CONCLUSIONS

In this paper, we present a review as for building structural interaction network (SIN) for host-pathogen protein-protein interactions to analyze the network in a systems biology approach. Several multidisciplinary but related areas are reviewed, including protein structure prediction, domain-domain interaction prediction and machine learning methods applied in these prediction tasks.

For PHPPI researches, building SIN with the atomic level data can provide insights on the high-resolution interactions based on protein structures and further present high-quality analysis of interactions targeting the infection mechanisms. To the best of our knowledge, currently there are still a lot of efforts to be accomplished in this area.

ACKNOWLEDGMENTS

This work is supported by the scholarship from the China Scholarship Council (CSC), while the first author pursues his PhD degree in the University of Wollongong.

REFERENCES

[1] Orchard, S., Hermjakob, H. and Apweiler, R., 2004. Proteomics and data standardisation. Drug Discovery Today: Biosilico, 2(3), pp.91-93.

[2] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. and Martin, M.J., 2004. UniProt: the universal protein knowledgebase. Nucleic acids research, 32(suppl 1), pp.D115-D119.

[3] Vaudel, M., Verheggen, K., Csordas, A., Ræder, H., Berven, F.S., Martens, L., Vizcaíno, J.A. and Barsnes, H., 2016. Exploring the potential of public proteomics data. Proteomics, 16(2), pp.214-225.

[4] McLendon, R., Friedman, A., Bigner, D., Van Meir, E.G., Brat, D.J., Mastrogianakis, G.M., Olson, J.J., Mikkelsen, T., Lehman, N., Aldape, K. and Yung, W.A., 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 455(7216), pp.1061-1068.

[5] Alyass, A., Turcotte, M. and Meyre, D., 2015. From big data analysis to personalized medicine for all: challenges and opportunities. BMC medical genomics, 8(1), p.33.

[6] Min, S., Lee, B. and Yoon, S., 2016. Deep learning in bioinformatics. Briefings in Bioinformatics, p.bbw068.

[7] Aloy, P. and Russell, R.B., 2006. Structural systems biology: modelling protein interactions. Nature Reviews Molecular Cell Biology, 7(3), pp.188-197.

[8] Libbrecht, M.W. and Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), pp.321-332.

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

[10] Franzosa, E.A. and Xia, Y., 2011. Structural principles within the human-virus protein-protein interaction network. Proceedings of the National Academy of Sciences, 108(26), pp.10538-10543.

[11] Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. and Jiang, H., 2007. Predicting protein–protein interactions based only on

sequences information. Proceedings of the National Academy of Sciences, 104(11), pp.4337-4341.

[12] Zamani, M. and Kremer, S.C., 2015, August. Protein secondary structure prediction using an evolutionary computation method and clustering. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on (pp. 1-6). IEEE.

[13] Floudas, C.A., Fung, H.K., McAllister, S.R., Mönnigmann, M. and Rajgaria, R., 2006. Advances in protein structure prediction and de novo protein design: A review. Chemical Engineering Science, 61(3), pp.966-988.

[14] Aydin, Z., Altunbasak, Y. and Borodovsky, M., 2006. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. BMC bioinformatics, 7(1), p.178.

[15] Spencer, M., Eickholt, J. and Cheng, J., 2015. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(1), pp.103-112.

[16] Li, Z. and Yu, Y., 2016. Protein Secondary Structure Prediction Using Cascaded Convolutional and Recurrent Neural Networks. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp.2560-2567

[17] Yaseen, A. and Li, Y., 2014. Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features. BMC bioinformatics, 15(8), p.S3.

[18] Zhou, J. and Troyanskaya, O.G., 2014, March. Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction. In ICML (pp. 745-753).

[19] Qian, N. and Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), pp.865-884.

[20] Dor, O. and Zhou, Y., 2007. Achieving 80% ten‐fold cross‐validated accuracy for secondary structure prediction by large‐scale training. Proteins: Structure, Function, and Bioinformatics, 66(4), pp.838-845.

[21] Jayaram, B., Dhingra, P., Mishra, A., Kaushik, R., Mukherjee, G., Singh, A. and Shekhar, S., 2014. Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. BMC bioinformatics, 15(16), p.S7.

[22] Kaczanowski, S. and Zielenkiewicz, P., 2010. Why similar protein sequences encode similar three-dimensional structures?. Theoretical Chemistry Accounts, 125(3-6), pp.643-650.

[23] Yellaboina, S., Tasneem, A., Zaykin, D.V., Raghavachari, B. and Jothi, R., 2011. DOMINE: a comprehensive collection of known and predicted domain-domain interactions. Nucleic acids research, 39(suppl 1), pp.D730-D735.

[24] Stein, A., Panjkovich, A. and Aloy, P., 2009. 3did Update: domain–domain and peptide-mediated interactions of known 3D structure. Nucleic acids research, 37(suppl 1), pp.D300-D304.

[25] Finn, R.D., Miller, B.L., Clements, J. and Bateman, A., 2014. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic acids research, 42(D1), pp.D364-D373.

[26] Shoemaker B, Panchenko A (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3:e43

[27] Khor, S., 2014. Inferring domain-domain interactions from protein-protein interactions with formal concept analysis. PloS one, 9(2), p.e88943.

[28] Zhao, X.M., Chesi, G. and Chen, L., Computational Systems Biology: Understanding Biological Systems from the Perspective of Networks and Dynamics.

[29] Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O. and Abola, E.E., 1998. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallographica Section D: Biological Crystallography, 54(6), pp.1078-1084.

[30] DeLano, W.L., 2002. The PyMOL molecular graphics system.

[31] Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J. and Sonnhammer, E.L., 2013. Pfam: the protein families database. Nucleic acids research, p.gkt1223.

[32] 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, G.A. Salazar, J. Tate, A. Bateman, 2016, The Pfam protein families database: towards a more sustainable future, Nucleic Acids Research, Database Issue 44:D279-D285

[33] Mosca, R., Céol, A., Stein, A., Olivella, R. and Aloy, P., 2013. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic acids research, p.gkt887.

[34] Bhagwat, M. and Aravind, L., 2008. PSI-blast tutorial. Comparative Genomics, pp.177-186.

[35] Stolorz, P., Lapedes, A. and Xia, Y., 1992. Predicting protein secondary structure using neural net and statistical methods. Journal of Molecular Biology, 225(2), pp.363-377.

[36] Garnier, J., Gibrat, J.F. and Robson, B., 1996. [32] GOR method for predicting protein secondary structure from amino acid sequence. Methods in enzymology, 266, pp.540-553.

[37] Sen, T.Z., Jernigan, R.L., Garnier, J. and Kloczkowski, A., 2005. GOR V server for protein secondary structure prediction. Bioinformatics, 21(11), pp.2787-2788.

[38] Hua, S. and Sun, Z., 2001. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. Journal of molecular biology, 308(2), pp.397-407.

[39] Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp.273-297.

[40] Qian, N. and Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), pp.865-884.

[41] Rost, B., 1996. [31] PHD: Predicting one-dimensional protein structure by profile-based neural networks. Methods in enzymology, 266, pp.525-539.

[42] Wang, Z., Zhao, F., Peng, J. and Xu, J., 2011. Protein 8‐class secondary structure prediction using conditional neural fields. Proteomics, 11(19), pp.3786-3792.

[43] Chen, X.W. and Jeong, J.C., 2009. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics, 25(5), pp.585-591.

[44] Wei, Z.S., Han, K., Yang, J.Y., Shen, H.B. and Yu, D.J., 2016. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. Neurocomputing, 193, pp.201-212.

[45] Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8), pp.832-844.

[46] Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

[47] Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M. and Yu, H., 2012. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nature biotechnology, 30(2), pp.159-164.

[48] Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S. and Cooper, D.N., 2009. The human gene mutation database: 2008 update. Genome medicine, 1(1), p.13.

[49] Franzosa, E.A., Garamszegi, S. and Xia, Y., 2012. Toward a three-dimensional view of protein networks between species. Frontiers in microbiology, 3, p.428.

[50] Holm, L. and Sander, C., 1993. Protein structure comparison by alignment of distance matrices. Journal of molecular biology, 233(1), pp.123-138.

[51] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), pp.3389-3402.