Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2010, Article ID 924529, 15 pages
doi:10.1155/2010/924529

Review Article

From Experimental Approaches to Computational Techniques: A Review on the Prediction of Protein-Protein Interactions

Fiona Browne,[1] Huiru Zheng,[1] Haiying Wang,[1] and Francisco Azuaje[2]

[1] *Faculty of Computing and Engineering, University of Ulster Jordanstown Campus, Shore Road, Newtownabbey, Co. Antrim BT37 0QB, UK*
[2] *Laboratory of Cardiovascular Research, Public Research Centre for Health (CRP-Santé), 120, route d'ArlonL-1150, Luxembourg*

Correspondence should be addressed to Huiru Zheng, h.zheng@ulster.ac.uk

Received 15 September 2009; Revised 13 November 2009; Accepted 6 January 2010

Academic Editor: Daniel Berrar

A crucial step towards understanding the properties of cellular systems in organisms is to map their network of protein-protein interactions (PPIs) on a proteomic-wide scale completely and as accurately as possible. Uncovering the diverse function of proteins and their interactions within the cell may improve our understanding of disease and provide a basis for the development of novel therapeutic approaches. The development of large-scale high-throughput experiments has resulted in the production of a large volume of data which has aided in the uncovering of PPIs. However, these data are often erroneous and limited in interactome coverage. Therefore, additional experimental and computational methods are required to accelerate the discovery of PPIs. This paper provides a review on the prediction of PPIs addressing key prediction principles and highlighting the common experimental and computational techniques currently employed to infer PPI networks along with relevant studies in the area.

1. Introduction

Proteins are involved in many essential processes within the cell such as metabolism, cell structure, immune response and cell signaling [1]. Although advances have been made within the realm of genome biology and bioinformatics, the function of a large proportion of sequenced proteins remains uncharacterised [2]. Uncovering the function of proteins is a complex process as one protein may perform more than one function and many proteins may have undiscovered functionality [3]. Research in [4] has suggested that the functionality of unknown proteins could be identified from studying the interaction of unknown proteins with a known protein target with a known function. Thus, the determination of protein-protein interactions (PPIs) is an important challenge currently faced in computational biology [5]. Interaction patterns among proteins can suggest novel drug targets aiding in the design of new drugs by providing a clearer picture of the biological pathways in the neighbourhoods of the potential drugs targets [6].

Large-scale high-throughput experiments have assisted in defining PPIs within the interactome (all possible PPIs in a cell). However, data generated by these experiments often contain false positives, false negatives, missing values with little overlap observed between experimentally generated datasets [3]. This may suggest that the data are erroneous, incomplete or both [3]. Previous studies have estimated that 50% of the yeast PPI map and only 10% of the human PPI network have been characterised [7].

Due to the limitations of experimental data and the need to determine PPIs, additional methods both experimental and computational are required to accelerate the discovery of PPIs. Computational methods (for example, statistical and machine learning techniques) have been applied at various stages in the inference of PPI networks, for instance, the integration of diverse heterogeneous datasets, the prediction of potential PPIs, the evaluation of predictions, and the analysis of inferred PPI networks [8–11].

The aim of this paper is to provide a review on the prediction of PPI networks focusing on the application of computational techniques to infer PPIs. The remainder of this paper is organised as follows. Section 2 describes PPI prediction tasks and principles, followed by a description on how PPIs are constructed from experimental data. Section 4 presents an overview of data sources previously employed to infer PPIs. Section 5 reviews the prediction of PPIs

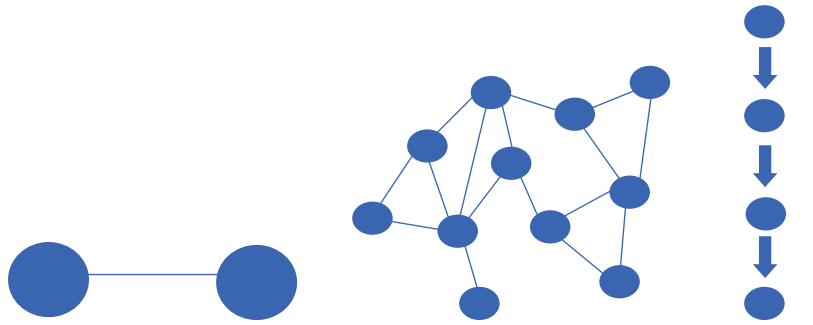

Figure 1: A graphical depiction PPI predictive tasks. (a) Direct binary interacting between two proteins; (b) proteins interact directly or interact indirectly through functional associations; and (c) Pathways represent logically linked connections for instance signal transduction. The nodes represent proteins; solid lines between the nodes represent direct interactions while dashed lines represent indirect functional associations.

using computational methods and recent studies. The paper concludes with a summary and future research.

2. Protein-Protein Interactions

Although a small percentage of proteins may operate in isolation, many proteins perform their functions by interacting with other proteins in PPI networks [9]. A protein interaction implies a specific physical contact between proteins which contributes to the formation of a biologically active protein complex. PPIs signal transduction, protein folding, cell cycle control, DNA replication and transport [10]. For instance, in signal transduction PPIs are involved in relaying signals from the cell exterior to the interior of the cell [10]. Furthermore, a protein may modify another protein through interaction. A common example of protein modification is the phosphorilation process. A kinase (a modifier protein) requires a physical contact with the target protein to add it a phosphate group. The modification of proteins can alter protein-protein interactions [9]. PPIs are involved in virtually all functions within a cell, however, a large proportion of PPIs still remain unknown [9]. This highlights the requirement to enhance our understanding of PPIs. It has been suggested that PPI patterns may aid in discovering new drug targets, and support the development of novel drugs. This is because PPI patterns illustrate biological pathways surrounding potential drugs targets [11].

2.1. Protein Interactions Prediction. The prediction of PPIs can be viewed as a binary classification problem whereby the aim is to identify pairs of proteins as either interacting or noninteracting [9, 12, 13]. There are various PPI prediction tasks including.

(1) Direct PPI prediction which involves the inference of direct physical interactions between proteins. Studies in [14, 15] have applied this predictive task to infer PPIs.

(2) Direct PPI and indirect functional association prediction whereby an interacting protein pair may not necessarily have direct physical contact but may indirectly interact through for example, complex formation. Protein scaffolding involves proteins which are important regulators in key signalling pathways. Scaffolding proteins interact with other proteins within a signalling pathway, tethering them into complexes. The study in [11] applied this principle in suggesting that proteins from the same subcellular complex may be considered "interacting" even if they do not directly physically interact with one another, but are connected through other proteins within the complex. Furthermore, the studies in [9–11, 16, 17] have employed this predictive task when inferring PPI.

(3) Pathway membership prediction whereby interactions occur in logical order (for instance, a signalling pathway). The study in [18, 19] applied this predictive task. Interactions within the pathways are often transient and may occur under specific conditions. Therefore, interactions may be difficult to measure using large-scale techniques [20].

These predictive tasks are summarised in Figure 1.

2.2. Protein-Protein Interaction Principles. PPI networks can be constructed by applying the principles of pair-wise (PW) interaction prediction or module-based (MB) interaction prediction. This review paper will focus on the prediction of PW interaction prediction as the majority of studies [9–13, 16, 21, 22] inferring PPIs apply the PW interaction prediction principle.

The aim of PW interaction prediction is to infer if two proteins are located in same protein complex [11]. The prediction of PW interaction deals with the prediction of the direct contact between two proteins. This interaction might occur between proteins appearing in the same cellular compartment by participation in the same protein complex. By contrast, the prediction of MB interactions deals with

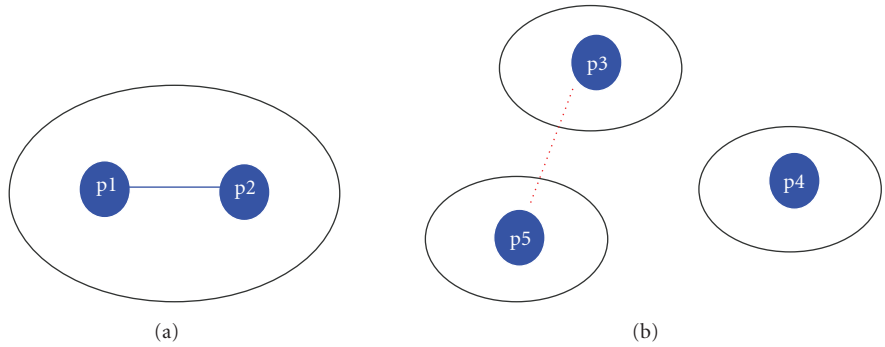

Figure 2: Graphic (a) illustrates an interaction between a pair of proteins (positive case). Graphic (b) illustrates a (negative case) noninteracting protein pair.

interactions of group of proteins, although in this case a direct contact between proteins is not required [23, 24]. Both the PW and MB prediction approaches aim to classify protein pairs or groups of proteins as either "interacting" or "noninteracting". PW and MB predictions can be used to construct a PPI network.

The concept of a positive PW interaction is graphically depicted in Figure 2(a) whereby one protein (p1) is connected (in an abstract sense) to protein (p2) for example, within the same subcellular complex. A noninteracting PW interaction is represented in Figure 2(b), whereby protein pairs in different clusters are considered to be unconnected. For instance proteins p4 and p5 are said to be noninteracting as they are in different protein complexes. Although a physical contact can be possible (as indicated by the dashed red line in (b)), an actual interaction is improbable due to these proteins belong to different protein compartments.

A graphical representation of a PPI network is illustrated in Figure 3, in which the nodes graphically represent proteins and edges represent binary interactions between proteins. This graph describes all 237 binary interactions associated with tumour suppressor proteins P53 (TP53) which has the highest degree found in the July 2009 release of HPRD database.

Limited research has been performed in the area of supervised MB PPI network prediction. The MB approach applied aims to detect whether (or not) a group of proteins (rather than a pair of proteins) belongs to the same protein complex. MB interaction prediction aims to predict various "modules" (that can vary in module size) of interacting proteins. A module can consist of a group of interacting proteins. This group may represent a protein complex. Publicly available sources, for instance, the Munich database of Interacting Proteins (MIPS) Complex Catalogue [25] contains definitions on known protein complexes and proteins within these complexes for different organisms. Figure 4 graphically illustrates the MB prediction task: (a) illustrates a group of proteins (p1, p2, p3, p4, p5, p6) found within the same complex representing a positive case and (b) the proteins p4, p8, p9 can be defined as a negative case as these proteins are found in different subcellular complexes.

Groups of genes are involved in many cellular activities. These genes behave in a coordinated fashion to perform specific biological processes [24]. Publically available high-throughput large-scale data contain a wealth of information to uncover PPI networks. The vast majority of this data is currently used for the prediction of PW interactions. However, the full potential of these data may not be fully utilised. These data could be further exploited to discover MB PPI networks [24]. Initial research suggests that modular-specific interaction predictions are an important area in predicting PPIs [24].

3. Experimental Data

Data relating to PPIs have been generated through the application of small-scale and large-scale high-throughput experimental methods. Using these data, efforts have been made to map PPIs on a proteomic-wide scale [26, 27].A review of experimental methods employed to detect PPIs including an outline of their advantages and limitations is presented in Table 1.

3.1. Small-Scale Experimental Methods. Small-scale methods focus upon specific bio-chemical or bio-physical properties of protein complexes [3]. Experimentalists often investigate several or one PPI at a time. Small-scale experiments are often applied for the detection and selection of proteins which bind to other proteins. This could be performed via affinity measurement of binding partners [3]. Small-scale experiments can be performed in vitro or in vivo. In vitro experiments are done outside of a living organism in a controlled environment and may provide valuable insights into PPIs [4]. In contrast, in vivo experiments are performed inside an organism. A selection of experimental methods is described in Table 1.

3.2. Large Scale Experimental Methods. Large-scale experiments are used to screen a vast number of proteins within the cell (i.e., across the whole proteome) [3]. Thousands of PPIs are produced which can be used to construct PPI networks. To increase the speed of discovery of PPIs, large-scale high-throughput experimental techniques have been developed to detect PPIs on a proteomic-wide scale, resulting in the production of a vast amount of interaction data [3].

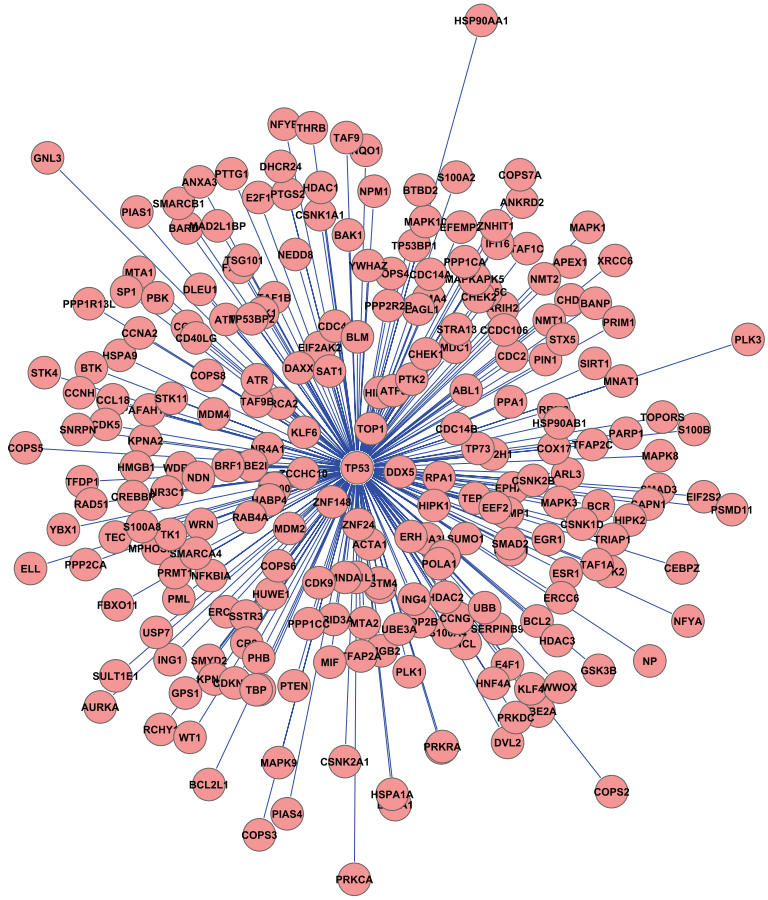

Figure 3: A graphical representation of proteins interacting with the tumour suppressor protein TP53 highlighted in yellow. All red circles represent its interaction partners found in the July 2009 release of HPRD database. This representation was performed using Cytoscape software.

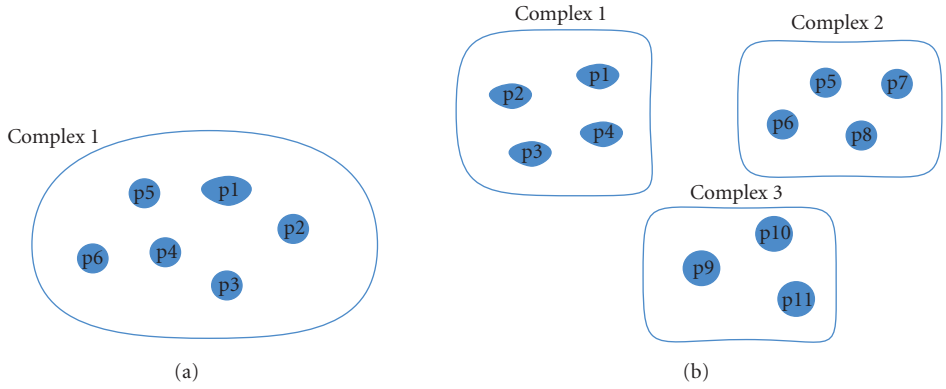

Figure 4: Graphic (a) illustrates a positive MB interaction between a group of proteins. Graphic (b) illustrates a (negative case) noninteracting group of proteins as some belong to different subcellular complexes.

A number of different experimental methods are usually required to determine, characterise and validate PPIs [3]. Common large-scale detection techniques include the Yeast Two-Hybid (Y2H) [33] and Mass Spectrometry and Tandem Affinity Purification (MS TAP) [34] which directly measure protein interactions and synthetic lethality [35], and gene coexpression [36] which indirectly provide evidence of PPIs. Descriptions of these techniques are presented in Table 2.

3.3. Construction of PPI Networks from Experimental Data. Efforts to map PPIs on a proteomic-wide large-scale have been made across different organisms including yeast [26, 27,

Table 1: A description of small-scale experimental methods applied to identify PPIs. A description and application of each method and reference to each technique are described below.

Technique	Description	Reference
Co-immunoprecipitations	To determine if two or more proteins are interacting, a purification procedure is applied to identify unknown or novel interactions	[28]
Surface Plasmon Resonance	A bait protein is attached to a gold surface where a laser light is reflected to measure small changes related to protein binding to identify unknown or novel interactions	[29]
Nuclear Magnetic Resonance (NMR)	Provides a dynamic view of PPIs when in a solution to investigate PPIs at the atomic level (i.e., smallest particle)	[30]
X-ray Crystallography	Aids in defining the structure of the interaction through crystallisation of the PPIs to investigate PPIs at an atomic level	[31]
Label Transfer	A known protein is "tagged". Interactions with this protein are obtained from detecting the tag to verify predicted or known PPIs	[32]
FRET	Proteins are labelled with fluorescence to detect interacting proteins to verify predicted or known PPIs	[32]
Far Western Blot	Proteins are applied to the blot to detect proteins of interest to verify predicted or known PPIs	[31]

33, 38, 44], fruit fly [45, 46], worm [47–49] and human [2, 7, 50, 51] through the use of experimental high-throughput technologies. Among these, yeast is perhaps one of the most investigated organisms [52]. PPI networks for yeast have been produced using various experimental techniques including Y2H, MS, and Tandem Affinity Purification (TAP) [44, 53]. Pioneering studies carried out by Schwikowski et al. [54], Ito et al. [38], Uetz et al. [33] and Gavin et al. [44] performed a comprehensive analysis of PPIs in yeast. For instance, Ito et al. [38] and Uetz et al. [33] applied the Y2H approach to infer PPI networks. Although there are limitations to the Y2H approach, it has been estimated that Y2H projects [44] have increased the amount of potential PPI data available [38].

Recent studies reported by Gavin et al. [26] and Krogan et al. [27] have utilised the experimental methods TAP and MS to construct PPI networks in yeast. Krogan et al. [27] produced a dataset consisting of 7,123 PPIs using 2,708 yeast proteins and obtained a greater coverage and accuracy in comparison to other high-throughput methods. In their study coverage was enhanced by applying rigorous computational procedures to assign confidence values to the predictions [27]. The related study in [26] produced a PPI network of the proteome averaged over all phases of the cell cycle.

The recognised significance of PPI networks has triggered huge efforts to construct PPI networks for more complex organisms. For instance, the study by Lehner and Fraser [55] developed the first draft of the human PPI map. In their study, Lehner and Fraser [55] applied the hypothesis "protein functions are usually conserved between species". Experimental data was obtained from other organisms such as yeast and integrated to produce a PPI network for human. The completed PPI network predicted interactions for one third of human genes [55]. A study by Bunescu et al. [56] produced a PPI network for human by extracting data from Medline abstracts using natural language processing and literature-mining algorithms techniques [56]. A total of 6580 interactions were identified among 3,737 human proteins and a network consisting of 31,609 interactions among 7,748 human proteins was produced through the integration of functional "omic" datasets [56].

Similar work has been performed using the organisms fruit fly and worm [45]. Formstecher et al. [57] and Giot et al. [46] both constructed a PPI network for the fruit fly uncovering 4,679 proteins and 4,780 interactions.

3.4. Limitations of Experimental Methods. The development and application of large-scale high-throughput technologies have resulted in the generation of vast amounts of data on PPI. This has contributed to the identification of PPIs [3]. However, data obtained by large-scale experimental methods are often noisy, incomplete and contradictory (i.e., weak predictive data sources) with thousands or tens of thousands interactions yet unknown [3]. Experimental methods can only identify a subset of the interactions that occur in an organism, therefore coverage (i.e., the area of the proteome covered by protein pairs) of the interactome is limited [58]. Furthermore, high-throughput studies are difficult to reproduce [3]. Methods such as the Y2H system exhibit high false positive and false negative interaction rates [3]. Traditional methods (e.g., small scale manual experiments) to infer PPIs may produce more accurate results compared to single source high-throughput methods. However, they are expensive and time consuming [4]. Furthermore different experimental conditions applied in different laboratories protocols makes it difficult to compile this information in a meaningful way. Therefore the use of a uniform method which is occurring in the large-scale approach facilitates the comparison. Due to inadequacies exhibited by both the small and large-scale experimental methods, advancements in computational methods are needed in the prediction of PPIs [8].

Table 2: An overview of large-scale experimental methods applied for the detection of PPIs on an interactome-wide scale. The first column states the name of the experiment followed by a description and a reference.

Technique	Description	Reference
Y2H	The Y2H employs a "two-plasmid" approach in yeast. The yeast protein GAL4 is a transcriptional activator consisting of two domains. The domains must be in close proximity to start the transcription process. The two plasmids are placed into a cell. A physical interaction occurs if the GAL4 binding and activation domains come together if physical interaction occurs, demonstrating that the "bait" and "target" bind [37]. This technique provides evidence of direct physical interactions between proteins.	[33, 38, 39]
MS TAP	Affinity tags are attached to a protein of interest (target), systematic precipitation of bait proteins is performed. Proteins are separated according to their mass to uncover purified protein complexes. Proteins are removed from a gel and analysed by MS techniques [40]. This technique provides evidence of direct physical interactions between proteins.	[34]
Gene coexpression	Gene expression profiles can be obtained from cell cycle experiments and the measurement of gene expression levels when the cell is under different conditions [41]. Gene expression similarity values may be calculated as the Pearson correlation co-efficient between expression levels of two proteins Protein pairs that are coexpressed are more likely to be interacting proteins [35, 42]. This technique provides indirect evidence of interactions between proteins.	[35, 42]
Synthetic lethality	This method involves the deletion or mutation of two genes which are viable alone, but cause lethality when combined in a cell under specific conditions [36, 43]. Synthetic interactions may detect PPIs between gene products, their occurrence in a pathway or participation in a function [40]. This technique provides indirect evidence of interactions between proteins.	[36, 43]

4. Data Sources

Data obtained from large-scale high-throughput experiments and "omic" information can be employed to support large-scale prediction of PPI networks [11]. However, individually these data are often limited in terms of accuracy and interactome coverage [6]. For example, estimated error rates of high-throughput experimental PPI datasets range 41–90% [6]. Studies in [10, 16, 17, 58, 59] have suggested that the integrating heterogeneous biological data using supervised machine learning methods can improve both the interactome coverage and predictions of PPIs. For example, Jansen et al. [11] integrated four features: (1) mRNA coexpression correlation, (2) MIPS functional similarity, (3) GO annotations, and (4) coessentiality using a Naïve Bayesian (NB) approach to infer PPIs in yeast. An increase in interactome coverage and predictive performance was observed when these features were integrated in comparison to the application of individual features alone [11]. Rhodes et al. [60] inferred PPIs in human by combining biological features within a probabilistic framework. These features included (1) homologous PPIs, (2) mRNA coexpression correlations, (3) functional similarity based on GO annotations, and (4) enriched domain pairs. By integrating these diverse heterogeneous features, ~40,000 human PPIs were predicted.

In this section, a brief description of a sample of data sources employed in the prediction of PPIs are presented.

mRNA Coexpression (COE). Based on the assumption that proteins which are coexpressed are more likely to interact than protein that are not-coexpressed, the COE information has been widely employed for the predictive task of inferring PPIs. For example, in yeast, the COE has been constructed from publicly-available expression data which represent the "time course of expression fluctuations during the yeast cell cycle and the Rosetta compendium" [61]. The data consists of expression profiles from 300 deletion mutants and cells which have undergone various chemical treatments. Pearson's correlation values were calculated for each protein pair in the data set.

MIPS Functional Similarity (FunCat). The FunCat data source is based on the assumption that proteins found within the same biological process are more likely to interact in comparison to proteins from different biological processes. Protein pairs are defined as interacting if they both belong to the same biological process or noninteracting if they belong to different biological processes (as defined by the Functional Catalogue). In the study published by Jansen et al. [11],

the FunCat was constructed by calculating similarity values between protein pairs annotated in the MIPS Functional Catalogue.

Coessentiality (ESS). The construction of the ESS dataset for the prediction of PPI is based on the assumption that proteins can be experimentally characterised as either essential (EE) or non-essential (NN), which may be used an indicator that the proteins are both members of the same complex. A protein can be classified as essential or non-essential, based on the viability of the cell when the gene is knocked out [11]. If two proteins exist in the same complex they are either essential or non-essential but not both.

The ESS dataset used in [11] is derived from the MIPS complex catalogue, transposon and gene deletion experiments [25].

Absolute Protein Abundance (APA). APA has been employed as a predictive feature to infer PPIs in yeast based on the hypothesis that an interacting protein pair should be present in stoichiometrically similar amounts (that is, the calculation of reactants and products in a chemical reaction) [10]. In one of the pioneering research published by Jansen and his colleagues [11], protein abundance is calculated by counting the number of proteins within a cell. APA values have been obtained from a number of experimental methods including gel-electrophoresis and mass spectrometry which have been scaled and merged by Greenbaum et al. [62].

Domain (DOM). The DOM has been employed as a predictive feature to infer PPIs in human. PPIs involve the physical interaction between domains (of proteins), therefore, PPIs could be inferred by identifying domain pairs enriched by known PPIs [63]. Hyper geometric distribution values between protein pairs were calculated in [59] to provide DOM feature values.

Phylogenetic Profiles. Pairs of non-homologous proteins that are either absent or present together in different organisms are more likely to have co-evolved [64]. Co-evolution has been observed between interacting proteins, such as chemokine and its receptors [64]. The study by Pellegrini et al. [65] examined co-occurrence or absence of genes across multiple genomes inferring functional relatedness.

Interologs. Interolog mapping involves the transfer of interaction annotation from one organism to another using comparative genomics [11]. This approached was used in the study by Yu et al. [66] to assess the degree to which interologs can be reliably transferred between species as a function of the sequence similarity between the corresponding interacting proteins.

Synthetic Lethality. This method involves the deletion or mutation of two genes which are viable alone, but cause lethality when combined in a cell under specific conditions. As the mutations are lethal, they should be synthetically generated. Synthetic interactions may detect PPIs between gene products, their occurrence in a pathway or participation in a function [40]. For instance, the application of synthetic lethality experiment discovered that the unknown function of the gene "YLL049W" belonged to the pathway dynein-dynactin [67].

4.1. Availability of Data. Various databases store information relating to PPIs (e.g., direct physical PPIs or data relating to protein complex membership) for different organisms. These data have been extracted from manually curated data or by data-mining literature. A list of popular databases containing PPIs is provided in Table 3.

4.2. Gold Standards. Gold Standards (GS) contain known interacting (positive) and noninteracting (negative) protein pair cases and can be employed to: (1) train classifiers for the predictive task of PPI inference or (2) evaluate computationally predicted PPIs. Furthermore, the quality of statistical and machine learning methods will depend upon the relevance and validity of the GSs to the prediction problem under study [11]. The study by Jansen et al. [11] suggested that a GS should be (1) generated independently from the data sources applied to infer PPI, (2) contain a sufficient number of protein pairs to provide reliable statistics, and (3) to be free of systematic bias. However, the selection of a GS for the prediction of PPIs can be problematic. For example, selecting a GS with adequate coverage of the interactome and defining what the GS specifically measures (i.e., complex membership, direct physical interactions) can be a difficult task. High quality positive GSs (GSP) are often assembled from interactions generated from small-scale manually curated experiments [2].

The construction of a negative GS (GSN) is also difficult as there are no "gold standard" noninteractions. Two methods to construct GSNs have been described in the literature: (1) studies in [8, 9, 11, 35] have suggested that high quality noninteractions can be generated by selecting pairs of proteins from different subcellular compartments, as they are more likely to be prevented from participating within biologically relevant interactions [8]; (2) other studies in [71, 72] have selected noninteracting pairs uniformly at random from a set of all proteins pairs that are not known to interact. Both of these two methods have limitations. For example, proteins selected from different cellular compartments may interact (for example proteins in the nucleus and cytoplasm) [72]. Moreover, due to the incompleteness of PPI networks, a GSN constructed by randomly selecting protein pairs may contain undiscovered true positive protein pairs, and thus may counteract the successful prediction of those [71].

GSs employed for the predictive task of PPI inference are often highly unbalanced with many more noninteracting pairs than interacting pairs. This is because the number of true biological PPIs is a rare phenomena among all possible protein pairs in the interactome [8]. For instance, yeast has ~6000 proteins resulting in ~18 million protein pairs. Estimates place the number of interacting protein pairs in yeast around 10,000–20,000 [6].

Table 3: A list of popular databases containing PPI information for organisms.

Database	Data Type	Organisms	URL	Reference
BioGRID	Experimental and manually curated data	22 organisms including: yeast, fruit fly, worm, human	http://bind.ca	[68]
Database of Interacting Proteins (DIP)	Experimental and structural data	274 organisms including: yeast, human, rat, mouse, fly and worm	http://dip.doe-mbi.ucla.edu/	[15]
Munich Database of Interacting Proteins (MIPS)	Experimental, functional predictions and manually curated	Mouse, human, yeast	http://www.helmholtz-muenchen.de/en/ibis	[25]
Saccharomyces genome database (SGD)	Experimental and manually curated	Yeast	http://www.yeastgenome.org/	
IntACT	Experimental and manually curated	Includes the organisms: yeast, human, rat, mouse, fly and worm	http://www.ebi.ac.uk/intact/site/	[69]
Human Protein Reference Database	Experimental and manually curated	Human	http://www.hprd.org/	
MINT	Experimental and manually curated	30 organisms including: yeast, human, rat, mouse, fly and worm	http://mint.bio.uniroma2.it/mint/	[70]

The web-based system GRIP (Gold Reference dataset constructor from Information on Protein complexes) outlined in the study by Browne et al. [73] provides researchers with the functionality to create reference datasets for PPI prediction in yeast. GRIP integrates the functionality for constructing reference datasets, protein complex membership matching and protein complex matching. Recent research by [10, 11] demonstrated that the generation of reference datasets are critical for the verification of computationally-inferred PPI networks. A study by [74] implemented reference datasets constructed using GRIP to demonstrate that supervised statistical and machine learning techniques can be successfully applied to PW and MB interaction prediction.

5. Computational Prediction of PPIs

The prediction of PPIs can be defined as a classification problem. For instance, a statistical or machine learning technique can be applied to the predictive task of determining whether a pair of proteins are interacting or noninteracting [9]. However, the prediction of PPIs is a complex task. For example, the datasets are highly skewed (i.e., there are more noninteracting PPIs than interacting PPIs) [17] and may be noisy and contain missing values [11]. Therefore, the selection of an appropriate classification technique is an important task. Classifiers that perform well in other problem domains may not perform as well within the realm of PPI prediction [75]. It is essential to assess available classification models for inferring PPIs [75]. This section will provide an overview of statistical and machine learning techniques and their application to PPI inference.

5.1. Statistical and Machine Learning Techniques. Computational methods (for example, statistical and machine learning techniques) have been applied at various stages in the inference of PPI networks. For instance, the integration of diverse heterogeneous datasets; the prediction of potential PPIs; the evaluation of predictions and the analysis of inferred PPI networks [8–11]. A summary of statistical and machine learning techniques including (1) K-Nearest Neighbour (KNN), (2) Naïve Bayesian (NB), (3) Support Vector Machine (SVM), (4) Artificial Neural Networks (ANN), (5) Decision Tree (DT), and (6) Random Forest (RF) are presented in Table 4. These techniques have been selected as they have previously been employed for the predictive task of inferring PPI networks.

5.2. Review of Current Studies. A number of studies have combined both direct and indirect experimental information in a supervised learning framework to predict PPIs [9, 11, 59]. These studies focus on the prediction of PPIs in yeast and human. The Yeast is an important experimental organism for the prediction of PPIs as it has been extensively characterised and the genome is fully sequenced [83]. Furthermore, yeast displays many features of higher eukaryotes (such as human). This is important as cellular processes are often conserved between eukaryote species [83]. Relatively few studies have been performed to computationally predict PPIs in human. Compared to yeast, the human interactome is considered more complex due to a larger number of proteins, post-translational modifications, splice isoforms and dynamic regulations [59]. Mapping human PPIs could provide a framework to improve understanding of protein function in complex diseases such as cancer [60]. Table 5 provides a summary of these studies.

5.2.1. PPI Prediction for Yeast. The study by von Mering et al. [3] was one of the first studies to discuss the issues

Table 4: A summary of machine and statistical learning approaches applied to the predictive task of inferring PPI. Advantages and limitations for each approach are presented along with a reference to the studies where they have been applied.

Classifier	Description	Reference
NB	Ability to integrate diverse heterogeneous data. Can handle missing data. Assumes conditional independence between datasets. Performance of NB deteriorates when dependencies between features exist.	[10, 11]
KNN	Classification method which has been considered "simple but powerful" providing competitive performance compared with other classifiers [76]. Classifier performance may deteriorate if many variables are used or if the GS is not balanced.	[17, 74]
SVM	Can handle non-linearly separable datasets. Can incorporate prior information.	[77]
RF	Can handle missing values. Can integrate diverse heterogeneous data	[78, 79]
ANN	Ability to recognise complex patterns	[80–82]

of computationally predicting PPI using experimental data. Data such as: Y2H, MS, mRNA gene-expression, gene fusion, gene neighbourhood and phylogenic profiles were employed in their study. Results obtained highlighted a low overlap between the various data sources. This suggests that experimental methods: (1) may not have reached saturation; (2) methods produce high false positives; (3) methods identify different interactions. von Merring et al. [3] suggested high-throughput experimental data could be integrated to improve the confidence of PPI predictions. The integration of diverse heterogeneous data in their study lead to a reduction in the number of false positives, however the coverage of the interactome was limited [3]. For example, only ~2,400 of a possible 80,000 protein interactions in yeast were supported by more than one method [3].

Jansen et al. [11] applied a Bayesian Network (BN) approach to predict PPIs using four features: gene coexpression, GO biological process similarity, MIPS functional similarity, essentiality. The MIPS Complex Catalogue [25] was employed as a GS. Individually, the datasets were weak predictors of PPIs. However, when the datasets were integrated via BN, accurate PPI networks were produced providing a comprehensive view of the yeast interactome [11]. Troyanskaya et al. [13] also applied a BN approach to combine diverse data sources for the inference of PPIs in yeast. The data sources employed included: gene coexpression and physical associations. The GS was constructed from information extracted from the GO [84]. The study in [14] employed a confidence measure for predictive PPIs using a Logistic Regression approach. Their study produced a high-confidence PPI network for over one third of the yeast proteome. Lin et al. [9] repeated the experiments by [11] and employed the classifiers NB, Random Forest (RF) and Logistic Regression to infer PPIs. Using only a subset of the integrated datasets with no missing values, Lin et al. [9] discovered that the MIPS and GO functional datasets were the most dominant features.

The study by Browne et al. [85] investigated the integration of functional genomic data for the prediction of PPI in yeast. A Bayesian classifier was employed to reassess the limits of genomic integration using seven genomic features ranging from coexpression to essentiality. Assessment methods such as true positive/false positive (TP/FP) rate and sensitivity were applied as comparative predictive measures to the ROC curve. A clear increase in predictive performance was observed using the measures TP/FP and sensitivity when the features were integrated.

A RF classification method was employed by Qi et al. [78] for the prediction a PPI network in yeast. The RF classifier predicted PPIs with an average sensitivity of ~80% and a specificity below 65%. Additionally, Qi et al. [78] demonstrated how selection and encoding of datasets has an impact upon the PPI predictive performance. Various classification techniques such as RF, RF integrated with KNN, NB, DT, Logistic Regression and SVM were applied. It was discovered that the RF classifier performed robustly in inferring PPIs.

Lu et al. [10] extended a study in [11] to evaluate the predictive limits of "omic" integration using a NB approach. Sixteen diverse datasets ranging from: synthetic lethality to MIPS functional similarity was integrated to predict PPIs. Compared to the previous study in [11], relatively high predictive accuracies were obtained. However, the addition of "weaker" datasets provided only marginal improvement in terms of predictive performance. This is in comparison to the integration of seven "strong" datasets. The NB classifier assumes conditional independence between datasets, Lu et al. [10] provided evidence of only marginal dependencies between the datasets employed in the study. However, as high-throughput technologies continue to emerge, datasets produced will present more potential dependencies. Therefore, the NB classifier may not be the optimal computational approach to predict PPIs. Dependencies between datasets may possibly cause the predictive accuracy obtained by NB to decrease [10].

Myers et al. [24] constructed a system entitled "bioPIXIE" to provide integration, analysis and visualisation of PPI predictions in yeast. This system used a BN approach; the PPIs predicted were validated by recovering networks for 31 known biological processes in yeast. Their study outlined critical issues when evaluating functional "omic" data. These include (1) bias and inconsistencies of GS, (2) the selection of negative GS, (3) number of proteins pairs in the GS. The GS employed in their study was constructed based on expert curation [24].

5.2.2. PPI Prediction for Human. The human proteome is considered more complex in comparison to the yeast proteome. This is due to a larger number of proteins, dynamic regulations, and post-translational modifications in human [2]. Moreover, more data sources are available for yeast in comparison to human [2]. This has resulted

Table 5: A summary of related work in inferring PPI networks. The first column presents the study, this is followed by advantages and limitations of the study.

Related Work	Advantages	Limitations	Ref
A Bayesian networks approach for predicting protein-protein interactions from genomic data	Pioneering study which applied a Bayesian approach to infer PPI in yeast by integrating diverse genomic data	Only four features were integrated. By integrating more features an improvement in interactome coverage and classification predictive performance may be achieved	[11]
A Bayesian networks approach for predicting protein-protein interactions from genomic data	Sixteen diverse features were integrated using a NB classifier to predict PPI in yeast.	The NB classifier approach was applied—this classifier assumes feature independence. Subtle dependencies between features may have an adverse affect on the NB performance. ROC curves were the only assessment method applied to measure the predictive performance of the classifier.	[10]
Information assessment on predicting protein-protein interactions	Application of RF, NB and logistic regression for the prediction of PPI in yeast. Discovered MIPS and GO annotation data were dominant features	Only used subset of data. Missing data was removed.	[9]
Probabilistic model of the human protein-protein interaction network	One of the first studies to integrate diverse "omic" data for the prediction of PPI in human.	A NB approach was employed to infer PPI. Three gene coexpression datasets will employed however only the maximum likelihood ratio per gene coexpression data source per protein pair was considered	[60]

in a limited number of studies which have computationally inferred PPIs for human.

Rhodes et al. [60] provided an integrated analysis of human PPIs using a NB approach. The data employed consisted of homology, gene coexpression, shared biological process and domain data. Information extracted from the Human Protein Reference Database (HPRD) [63] was used as the GS to evaluate PPI predictions. Experimental methods confirmed protein interactions predicted by the framework.

Xia et al. [86] integrated 27 heterogeneous data sources using a probabilistic approach to infer PPIs for human. An integrated network database was constructed and provides the functionality of prediction and visualisation of genes of interest. Scott and Barton [2] constructed a probabilistic framework to integrate diverse features including: gene coexpression, localisation information, domain-domain interactions. A total of 37,606 PPIs were predicted, 80% of which are not found in other human PPI databases.

A recent study by Qi et al. [59] addressed the limitation of missing data and feature redundancy in inferring PPIs in human. A "mixture-of-features" framework was applied to predict PPIs. They employed obtained Precision-Recall curves to evaluate the predictive performance of classifiers including: NB, SVM and RF. In their study, 18 potentially novel interacting protein pairs were identified.

Browne et al. [73] applied a fully connected BN approach to integrate diverse "omic" features for the inference of disease-specific PPI networks. The case study integrated three gene coexpression datasets relevant to human heart failure along with other datasets to reconstruct a PPI network relevant to the development of dilated cardiomyopathy. By modelling relationships between multiple datasets of the same "omic" type, an improvement in prediction performance was achieved in terms of partial AUC and the ratio of TP/FP by the fully connected BN approach in comparison to the maximum likelihood ratio and NB approaches.

The studies highlighted above for prediction of PPIs in human and yeast share commonality in the types of data sources that were employed and in some cases the predictive computational methods employed. A commonly

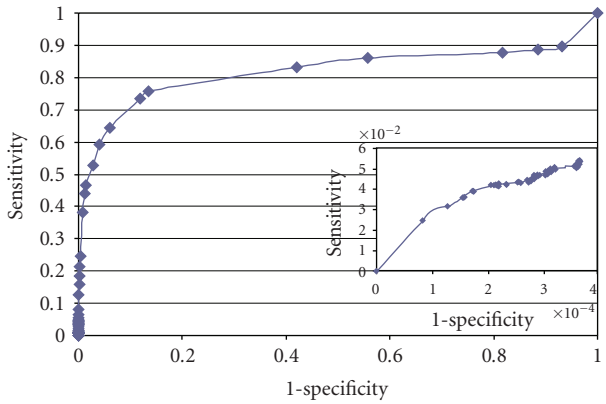

Figure 5: Plotting of a ROC curve when the integration scheme including seven features in applied. The sensitivity is plotted against the 1-specificity for different likelihood thresholds. Various likelihood thresholds have been highlighted on the ROC to illustrate the different AUC thresholds. The inset graph illustrates the AUC when the likelihood of 600 and greater is selected.

applied computational predictive approach in these studies was the Bayesian classifier. This classifier can handle diverse heterogeneous data types and missing values which is advantageous when inferring PPIs as the data is often obtained from different sources and may suffer from missing values. The studies differ in the data sources employed for the prediction of PPIs, selection of GSs and the evaluation methods employed. Therefore it is difficult to obtain a comparative view of the different computational methods in predicting PPIs. The study by Browne et al. [75] and Qi et al. [17] performed a comparative review of different computational techniques when inferring PPIs using a selection of supervised learning approaches. In this study the same data sources, GSs and evaluation methods were applied to provide a comprehensive comparison of computational approaches when inferring PPIs in yeast.

5.3. Limitations of Computational PPI Prediction. Despite the relative success of the computational methods applied to infer PPIs, no approach can accurately predict all PPIs within an interactome. A number of computational limitations outlined below need to be addressed for this to become reality. For example, computational efficiency of the classifier needs to be addressed. For instance, classifiers such as KNN have been found to be time consuming and processor intensive [17]. Statistical and machine learning methods are known to exhibit systematic bias [75]. A computational technique may produce solutions that favour a limited number of specific situations or circumstances [75]. Computational classification techniques make assumptions, such as the NB which assumes dataset independence [10]. A number of studies have applied different predictive models to predict PPIs in yeast [8, 12, 17, 21, 87, 88]. However, there is difficulty when comparing and contrasting results from these studies due to differences in the predictive models, features, GS and predictive tasks applied. For relatively simple organisms, such as yeast, more datasets are available for the prediction of PPIs compared to more complex organisms such as human [2]. As organisms increase in complexity the data obtained and the task of PPI prediction also increase in complexity [2]. Datasets obtained for organisms such as human are sparse and suffer from high rates of false positives and false negatives with little coverage of the interactome [2]. Computational docking in protein folding may be employed as a local prediction method to computationally infer PPIs. However this method has only been successful when used on a small-scale [89].

5.4. Overview of Predictive Performance Measurement Techniques. The performance of a supervised machine learning framework is evaluated in terms of predictive quality and potential significance of PPI predictions. The selection of a measurement approach is essential in determining the predictive performance of a supervised learning approach. Various studies employ different predictive quality measures making it difficult to compare classification performance. A selection of assessment methods previously applied to evaluate the predictive performance of classifiers when inferring PPIs are presented below.

5.4.1. Cross Validation (CV). To estimate the performance of a predictive model CV can be applied. In n fold CV the dataset is partitioned into segments, analysis is performed on one segment (called the training set), one segment is left out for validation (called the test set). To reduce variability CV are preformed with different partitions with the validation results averaged over the different CVs.

5.4.2. ROC Curves. ROC curves have been commonly used to illustrate classification performance when predicting PPIs [10, 75]. In ROC analysis, the accuracy by which a model can separate positive from negative instances is investigated [19]. ROC curves plot in a single graph the sensitivity against 1-specificity over a range of different thresholds. The graph consists of a set of points each computed for a different threshold. For each point, the vertical co-ordinate represents the sensitivity and the horizontal co-ordinate the 1-specificity. Therefore, the predictive quality of a classifier is assessed by measuring the sensitivity and 1-specificity. The counts of the: TP, TN, FP and FN are obtained from the CV analysis. The formula used to calculate sensitivity and specificity are detailed below:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$1 - \text{Specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{2}$$

As illustrated in Figure 5, a predictive dataset will produce a ROC curve that rises steeply to the left hand side of the graph and has a large area under the curve. The AUC is a measurement of the area under the ROC Curve. A perfect classifier will have an AUC value of 1.0. A prediction model based on random assignments of pairs of proteins to classes would give an AUC equal to 0.5.

The majority of the AUC of a ROC curve when inferring PPI in yeast may not represent biologically informative results. For example, Figure 5 illustrates a ROC curve plotted when 7 features were integrated using the NB classifier to infer PPIs in yeast [85]. Various likelihood thresholds have been highlighted to illustrate how the majority of the AUC relates to a likelihood threshold which is less than or equal to 1. Therefore, the AUC of the ROC curve is not considered biologically meaningful as a threshold greater than or equal to 600 is required to predict a positive interaction (posterior odds of an interacting protein pair in yeast). The threshold of 600 is suboptimal for the trade-off between the TP and FP rate highlighted in Figure 5, from this it can be observed that relatively little of the total AUC is represented by a threshold of 600 and above.

These results highlight the importance of selecting an adequate assessment method for the quality testing to assess the quality of a prediction model. In the study by Browne et al. [73] and Jansen et al. [11] alternative representative methods: True Positive (TP)/False Positive (FP) rate and TP/Positive (P) have been employed as alternative representative measures to assess the performance of prediction model. These are detailed below.

5.4.3. TP/FP Ratio and TP/P. The TP/FP ratio is plotted against the threshold (TH) of likelihood ratio as a measure of the probability of a real interaction. This measure has previously been employed in the study by Jansen et al. [11]:

$$\left.\frac{\mathrm{TP}}{\mathrm{FP}}\right|_{L=\mathrm{TH}} = \sum_{L=\mathrm{TH}} \frac{N_{\mathrm{pos}}(L)}{N_{\mathrm{neg}}(L)}. \tag{3}$$

The $N_{\mathrm{pos}}(L)$ and $N_{\mathrm{neg}}(L)$are the number of interacting and noninteracting protein pairs in the GS with a given likelihood ratio of L.

The TP/P ratio is applied as a measure of coverage whereby P represents the number of positives in the GS.

5.4.4. Partial ROC Curve. Rather than measuring the AUC under the entire ROC curve, it may be more informative to consider the area under a portion of the curve. This is referred to as the Partial ROC curve AUC which has previously been employed in the study by Browne et al. [73] to illustrate the number of true positives identified by the Bayesian classifier against specified likelihood cut-off rates which represent thresholds of biologically meaningful predictions.

Partial ROC curves have been applied as evaluation measures in recent studies [2, 90]. In these studies, the partial ROC plots the AUC whereby the false positive rate is low (for instance, measuring the AUC until 50 negative predictions have been reached) [90]. The partial curve applied in the study by Browne et al. [73] differs from previous studies as the area of the ROC whereby the predictions exceeding a selected threshold is measured. For yeast the threshold selected is 600 and for human 400. These thresholds are based upon the prior odds of an interacting protein in yeast and human, respectively. The partial ROC measures are referred to as ROC_{600} for yeast and ROC_{400} for human.

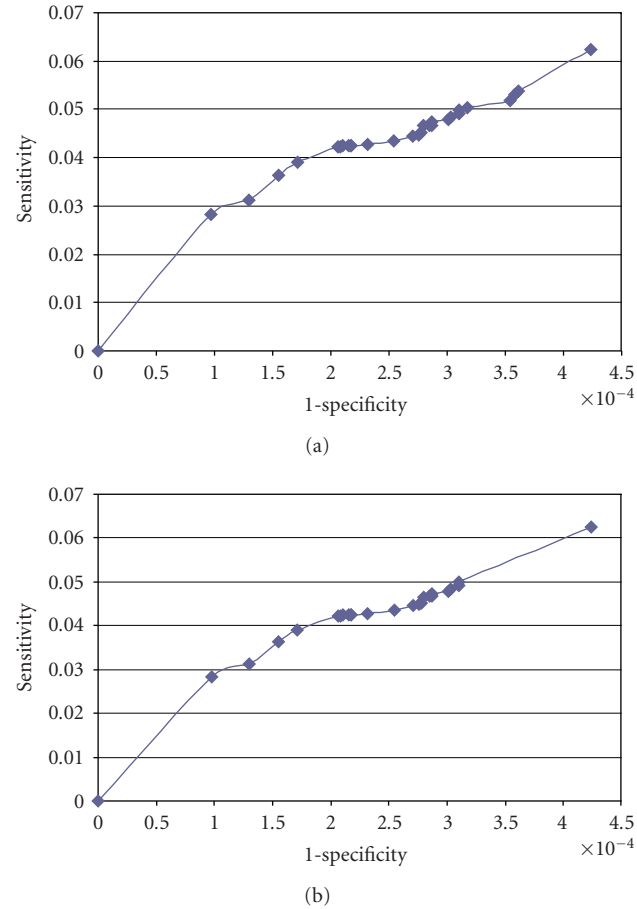

Figure 6: Graphical representation of a Partial ROC curves. (a) represents a portion of the ROC curve plotted when predicting PPIs in yeast (the threshold is greater than 600); (b) represents a portion of the ROC curve plotted when inferring PPI predictions in human where the threshold is greater than 400.

ROC_{600} and ROC_{400} measure high quality predictions. Figure 6 illustrates examples of partial ROC curves, (a) a portion of the ROC curve plotted representing PPI predictions in yeast whereby the threshold is greater than 600; (b) a portion of the ROC curve plotted representing PPI predictions in human whereby the threshold is greater than 400.

6. Conclusions and Future Trends

PPIs play an important role in many biological functions and diseases [7]. A wealth of biological data has been provided though the advent of experimental high-throughput technologies [3]. Data obtained from large-scale high-throughput experiments and "omic" information (e.g., essentiality and functional information) can be employed to support large-scale prediction of PPI networks [11]. However, individually, these data are often limited in terms of accuracy and interactome coverage [6]. For example, estimated error rates of high-throughput experimental PPI datasets range from 41–90% [6]. Studies in [10, 16, 17, 58, 59] have suggested that the integration heterogeneous

biological data using supervised machine learning methods can improve both the interactome coverage and predictions of PPIs.

PPI networks can be constructed using a number of prediction principles including PW interaction prediction and MB interaction prediction.

Statistical and machine learning techniques can be applied in the computational prediction of PPI [10, 11, 74]. These techniques are required for the integration of heterogeneous features and the inference of PPI networks. However, computational techniques may make assumptions and as of yet, there is no standard machine learning technique within the area of PPI prediction [75]. Further investigation is required to assess the predictive performance of different statistical and machine learning techniques employed to integrate diverse features for the prediction of PPIs.

AUC values from the ROC curves are commonly employed as the evaluation method to assess the predicative performance of the classifiers when inferring PPIs [10, 75]. However, this method may not be the most optimal approach to evaluate the predictive performance of classifiers when inferring PPIs. The study by Browne et al. [85] has demonstrated that the additional application of other assessment techniques such as partial AUC values from ROC curves, TP/FP rates, and sensitivity could be employed as comparative predictive measures to the ROC curve approach when evaluating the classification performance for the predictive task of PPI inference.

The computational inference of PPI networks is still a relatively new research area. Future research in inferring PPI networks may be performed in the areas including the recovery of PPIs between proteins [80, 88], identification of protein complexes [23, 91, 92], investigating network topology of PPI networks [67], defining and modelling pathways (for instance, signalling and metabolic pathways) [93].

References

[1] B. Alberts, *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, Garland, New York, NY, USA, 1998.

[2] M. S. Scott and G. J. Barton, "Probabilistic prediction and ranking of human protein-protein interactions," *BMC Bioinformatics*, vol. 8, article 239, 2007.

[3] C. von Mering, R. Krause, B. Snel, et al., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[4] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiological Reviews*, vol. 59, no. 1, pp. 94–123, 1995.

[5] P. Bork, L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee, and E. M. Marcotte, "Protein interaction networks from yeast to human," *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 292–299, 2004.

[6] J. Yu and F. Fotouhi, "Computational approaches for predicting protein-protein interactions: a survey," *Journal of Medical Systems*, vol. 30, no. 1, pp. 39–44, 2006.

[7] G. T. Hart, A. K. Ramani, and E. M. Marcotte, "How complete are current yeast and human protein-interaction networks?" *Genome Biology*, vol. 7, no. 11, article 120, 2006.

[8] R. Jansen and M. Gerstein, "Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction," *Current Opinion in Microbiology*, vol. 7, no. 5, pp. 535–545, 2004.

[9] N. Lin, B. Wu, R. Jansen, M. Gerstein, and H. Zhao, "Information assessment on predicting protein-protein interactions," *BMC Bioinformatics*, vol. 5, article 154, 2004.

[10] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," *Genome Research*, vol. 15, no. 7, pp. 945–953, 2005.

[11] R. Jansen, H. Yu, D. Greenbaum, et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[12] C. L. Myers and O. G. Troyanskaya, "Context-sensitive data integration and prediction of biological networks," *Bioinformatics*, vol. 23, no. 17, pp. 2322–2330, 2007.

[13] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 14, pp. 8348–8353, 2003.

[14] J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant, "Gaining confidence in high-throughput protein interaction networks," *Nature Biotechnology*, vol. 22, no. 1, pp. 78–85, 2004.

[15] 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*, vol. 30, no. 1, pp. 303–305, 2002.

[16] Y. Xia, L. J. Lu, and M. Gerstein, "Integrated prediction of the helical membrane protein interactome in yeast," *Journal of Molecular Biology*, vol. 357, no. 1, pp. 339–349, 2006.

[17] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 63, no. 3, pp. 490–500, 2006.

[18] Y. Yamanishi, J.-P. Vert, and M. Kanehisa, "Protein network inference from multiple genomic data: a supervised approach," *Bioinformatics*, vol. 20, supplement 1, pp. i363–i370, 2004.

[19] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, no. 5701, pp. 1555–1558, 2004.

[20] A. Ghavidel, G. Cagney, and A. Emili, "A skeleton of the human protein interactome," *Cell*, vol. 122, no. 6, pp. 830–832, 2005.

[21] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.

[22] C. L. Myers, D. Robson, A. Wible, et al., "Discovery of biological networks from diverse functional genomic data," *Genome Biology*, vol. 6, no. 13, article R114, 2005.

[23] S. Asthana, O. D. King, F. D. Gibbons, and F. P. Roth, "Predicting protein complex membership using probabilistic network reliability," *Genome Research*, vol. 14, no. 6, pp. 1170–1175, 2004.

[24] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya, "Finding function: evaluation methods for functional genomic data," *BMC Genomics*, vol. 7, article 187, 2006.

[25] H. W. Mewes, C. Amid, R. Arnold, et al., "MIPS: analysis and annotation of proteins from whole genomes," *Nucleic Acids Research*, vol. 32, pp. D41–D44, 2004.

[26] A.-C. Gavin, P. Aloy, P. Grandi, et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[27] N. J. Krogan, G. Cagney, H. Yu, et al., "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[28] M.-H. Kuo and C. D. Allis, "In vivo cross-linking and immunoprecipitation for studying dynamic protein: DNA associations in a chromatin environment," *Methods*, vol. 19, no. 3, pp. 425–433, 1999.

[29] J. Homola, S. S. Yee, and G. Gauglitz, "Surface plasmon resonance sensors: review," *Sensors & Actuators, B*, vol. 54, no. 1, pp. 3–15, 1999.

[30] A. Moser and C. Detellier, "Nuclear magnetic resonance spectroscopy," in *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York, NY, USA, 2004.

[31] Y. Wu, Q. Li, and X. Z. Chen, "Detecting protein-protein interactions by far western blotting," *Nature Protocols*, vol. 2, no. 12, pp. 3278–3284, 2007.

[32] K. D. Pfleger and K. A. Eidne, "Illuminating insights into protein-protein interactions using bioluminescence resonance energy transfer (BRET)," *Nature Methods*, vol. 3, no. 3, pp. 165–174, 2006.

[33] P. Uetz, L. Glot, G. Cagney, et al., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[34] G. D. Bader and C. W. V. Hogue, "Analyzing yeast protein-protein interaction data obtained from different sources," *Nature Biotechnology*, vol. 20, no. 10, pp. 991–997, 2002.

[35] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating whole-genome expression data with protein-protein interactions," *Genome Research*, vol. 12, no. 1, pp. 37–46, 2002.

[36] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[37] D. T. Suzuki, A. J. F. Griffiths, and R. C. Lewontin, *An Introduction to Genetic Analysis*, WH Freeman, New York, NY, USA, 7th edition, 2000.

[38] 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 of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[39] T. Ito, K. Tashiro, S. Muta, et al., "Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 3, pp. 1143–1147, 2000.

[40] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein-protein interactions—part I: experimental techniques and databases," *PLoS Computational Biology*, vol. 3, no. 3, article e42, 2007.

[41] L. Zhou, X. Ma, and F. Sun, "The effects of protein interactions, gene essentiality and regulatory regions on expression variation," *BMC Systems Biology*, vol. 2, article 54, 2008.

[42] O. G. Troyanskaya, "Putting microarrays in a context: integrated analysis of diverse biological data," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 34–43, 2005.

[43] H. B. Fraser, A. E. Hirsh, D. P. Wall, and M. B. Eisen, "Coevolution of gene expression among interacting proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 24, pp. 9033–9038, 2004.

[44] A.-C. Gavin, M. Bösche, R. Krause, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[45] M. Middendorf, E. Ziv, and C. H. Wiggins, "Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 9, pp. 3192–3197, 2005.

[46] L. Giot, J. S. Bader, C. Brouwer, et al., "A protein interaction map of *Drosophila melanogaster*," *Science*, vol. 302, no. 5651, pp. 1727–1736, 2003.

[47] W. Zhong and P. W. Sternberg, "Genome-wide prediction of C. elegans genetic interactions," *Science*, vol. 311, no. 5766, pp. 1481–1484, 2006.

[48] S. Li, C. M. Armstrong, N. Bertin, et al., "A map of the interactome network of the metazoan C. elegans," *Science*, vol. 303, no. 5657, pp. 540–543, 2004.

[49] A. J. Walhout, R. Sordella, X. Lu, et al., "Protein interaction mapping in C. elegans using proteins involved in vulval development," *Science*, vol. 287, no. 5450, pp. 116–122, 2000.

[50] R. M. Ewing, P. Chu, F. Elisma, et al., "Large-scale mapping of human protein-protein interactions by mass spectrometry," *Molecular Systems Biology*, vol. 3, article 89, 2007.

[51] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

[52] H. Yu, P. Braun, M. A. Yildirim, et al., "High-quality binary protein interaction map of the yeast interactome network," *Science*, vol. 322, no. 5898, pp. 104–110, 2008.

[53] Y. Ho, A. Gruhler, A. Heilbut, et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[54] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[55] B. Lehner and A. G. Fraser, "A first-draft human protein-interaction map," *Genome Biology*, vol. 5, no. 9, article R63, 2004.

[56] R. Bunescu, R. Ge, R. J. Kate, et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[57] E. Formstecher, S. Aresta, V. Collura, et al., "Protein interaction mapping: a Drosophila case study," *Genome Research*, vol. 15, no. 3, pp. 376–384, 2005.

[58] R. Jansen, N. Lan, J. Qian, and M. Gerstein, "Integration of genomic datasets to predict protein complexes in yeast," *Journal of Structural and Functional Genomics*, vol. 2, no. 2, pp. 71–81, 2002.

[59] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "A mixture of feature experts approach for protein-protein interaction prediction," *BMC Bioinformatics*, vol. 8, supplement 10, article S6, 2007.

[60] D. R. Rhodes, S. A. Tomlins, S. Varambally, et al., "Probabilistic model of the human protein-protein interaction network," *Nature Biotechnology*, vol. 23, no. 8, pp. 951–959, 2005.

[61] R. J. Cho, M. J. Campbell, E. A. Winzeler, et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[62] D. Greenbaum, R. Jansen, and M. Gerstein, "Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts," *Bioinformatics*, vol. 18, no. 4, pp. 585–596, 2002.

[63] S. Peri, J. D. Navarro, R. Amanchy, et al., "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Research*, vol. 13, no. 10, pp. 2363–2371, 2003.

[64] C.-S. Goh and F. E. Cohen, "Co-evolutionary analysis reveals insights into protein-protein interactions," *Journal of Molecular Biology*, vol. 324, no. 1, pp. 177–192, 2002.

[65] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[66] H. Yu, N. M. Luscombe, H. X. Lu, et al., "Annotation transfer between genomes: protein-protein interrologs and protein-DNA regulogs," *Genome Research*, vol. 14, no. 6, pp. 1107–1118, 2004.

[67] H. N. Chua, W.-K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623–1630, 2006.

[68] B.-J. Breitkreutz, C. Stark, T. Reguly, et al., "The BioGRID interaction database: 2008 update," *Nucleic Acids Research*, vol. 36, database issue, pp. D637–D640, 2008.

[69] H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, et al., "IntAct: an open source molecular interaction database," *Nucleic Acids Research*, vol. 32, pp. D452–D455, 2004.

[70] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, et al., "MINT: the molecular interaction database," *Nucleic Acids Research*, vol. 35, database issue, pp. D572–D574, 2007.

[71] A. Ben-Hur and W. S. Noble, "Choosing negative examples for the prediction of protein-protein interactions," *BMC Bioinformatics*, vol. 7, supplement 1, 2006.

[72] J. Guo, X. Wu, D.-Y. Zhang, and K. Lin, "Genome-wide inference of protein interaction sites: lessons from the yeast high-quality negative protein-protein interaction dataset," *Nucleic Acids Research*, vol. 36, no. 6, pp. 2002–2011, 2008.

[73] F. Browne, H. Wang, H. Zheng, and F. Azuaje, "GRIP: a web-based system for constructing Gold Standard datasets for protein-protein interaction prediction," *Source Code for Biology and Medicine*, vol. 4, article 2, 2009.

[74] F. Browne, H. Wang, H. Zheng, and F. Azuaje, "Supervised statistical and machine learning approaches to inferring pairwise and module-based protein interaction networks," in *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE '07)*, pp. 1365–1369, 2007.

[75] F. Browne, H. Wang, H. Zheng, and F. Azuaje, "An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions," *Journal of Integrative Bioinformatics*, vol. 3, 2006.

[76] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2000.

[77] S. L. Lo, C. Z. Cai, Y. Z. Chen, and M. C. M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005.

[78] Y. Qi, J. Klein-Seetharaman, and Z. Bar-Joseph, "Random forest similarity for protein-protein interaction prediction from multiple sources," *Pacific Symposium on Biocomputing*, pp. 531–542, 2005.

[79] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[80] Z. Ma, C. Zhou, L. Lu, Y. Ma, P. Sun, and Y. Cui, "Predicting protein-protein interactions based on BP neural network," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW '07)*, pp. 3–7, 2007.

[81] E. Keedwell and A. Narayanan, "Discovering gene networks with a neural-genetic hybrid," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 231–242, 2005.

[82] P. Fariselli, A. Zauli, M. Finelli, P. Martelli, and R. Casadio, "A neural network method to improve prediction of protein-protein interaction sites in heterocomplexes," in *Proceedings of the 13th IEEE Workshop on Neural Networks for Signal Processing (NNSP '03)*, pp. 33–41, September 2003.

[83] F. M. Ausubel, R. Brent, R. Kingston, et al., *Current Protocols in Molecular Biology*, vol. 3, John Wiley & Sons, New York, NY, USA, 2008.

[84] M. Ashburner, C. Ball, and J. Blake, "Gene ontology: tool for the unification of biology. The gene ontology consortium database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 34, 2006.

[85] F. Browne, H. Wang, H. Zheng, and F. Azuaje, "Reassessing the genomic data integration limits for the prediction of protein-protein interactions in saccharomyces cerevisiae," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 28–35, 2008.

[86] K. Xia, D. Dong, and J.-D. J. Han, "IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model," *BMC Bioinformatics*, vol. 7, article 508, 2006.

[87] R. J. P. van Berlo, L. F. A. Wessels, D. de Ridder, and M. J. T. Reinders, "Protein complex prediction using an integrative bioinformatics approach," *Journal of Bioinformatics and Computational Biology*, vol. 5, no. 4, pp. 839–864, 2007.

[88] P. Prusis, S. Uhlen, R. Petrovska, M. Lapinsh, and J. E. S. Wikberg, "Prediction of indirect interactions in proteins," *BMC Bioinformatics*, vol. 7, article 167, 2006.

[89] S. Grosdidier and J. Fernández-Recio, "Identification of hot-spot residues in protein-protein interactions by computational docking," *BMC Bioinformatics*, vol. 9, article 447, 2008.

[90] S. R. Collins, P. Kemmeren, X.-C. Zhao, et al., "Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae," *Molecular and Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.

[91] H. Zheng, H. Wang, and D. H. Glass, "Integration of genomic data for inferring protein complexes from global protein-protein interaction networks," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 38, no. 1, pp. 5–16, 2008.

[92] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12128, 2003.

[93] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005.