Network-Based Analysis for Biological Knowledge Discovery

Lokesh P. Tripathi[1,2], Rodolfo S. Allendes Osorio[3,2], Yoichi Murakami[4], Yi-An Chen[2] and Kenji Mizuguchi[5,2]
[1]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan.
[2]AI Center for Health and Biomedical Research (ArCHER), National Institutes of Biomedical Innovation, Health and Nutrition, Settsu, Osaka, Japan.
[3]WPI Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, Suita, Osaka, Japan.
[4]Tokyo University of Information Sciences, Wakaba-ku, Chiba, Japan.
[5]Institute for Protein Research, Osaka University, Suita, Osaka, Japan.

Abstract

Genome-scale interaction networks involving correlational, physical, or regulatory associations between key biomolecules such as genes and proteins are key to the functioning of the cell; analysis of such gene networks facilitates a deeper understanding of gene function and the underlying biological processes. Protein-protein interactions (PPIs) are fundamental to many cellular processes and living systems. PPI dysfunctions have been implicated in multiple diseases and hence understanding PPI mechanisms and events leading to their dysregulation is significantly useful in disease biology research. In the post-genomics era, the emergence of improved experimental technologies has enabled the characterization and construction of PPI networks (PPINs) on a proteome-wide scale. Here, we briefly discuss how PPINs inferred from experimentally characterized PPI data have been utilized for understanding cellular organizations, disease mechanisms, and genotype-phenotype relationships. We also discuss how bioinformatics methods for PPI prediction can facilitate PPIN-based biological research. Despite the rapid advances in the field, our understanding of protein interactomes is rather limited. We, therefore, briefly discuss future efforts in the field and how subsequent developments will facilitate the researchers to better leverage the PPINs and prioritize physiologically and therapeutically relevant proteins.

Keywords

Bottlenecks, Experimental PPI mapping, Genotype-phenotype relationships, Hubs, Interactome, Interlogue, Knowledge discovery, Network analysis, Network topology, PPI Networks (PPIN), PPI prediction, Protein-protein interactions (PPIs) and Proteomics

Key point/ objectives box

- An overview of methods used to experimentally generate Proteome-Scale Interaction Maps.
- A summarized review of the network methods that can be used for the analysis of PPINs.
- A (brief) survey on the current trends and advancements of *in silico* methods for prediction and assessment of PPIs.
- A perspective on likely future developments in the field, together with their likely significance in biological and clinical settings.

Introduction

Biological processes are complex systems, involving manifold interactions among elementary units of a living system such as DNA, RNA, proteins, lipids, and small molecule metabolites. To describe such processes, a common representation is a network model in which the participating biomolecules are represented as nodes and the connections between them as edges. Proteins are the most important biological building blocks, and they carry out their functions in the cells by interacting with each other. Therefore, it is not surprising that the largest amount of biomolecular interaction data is available for PPIs and consequently, a substantial chunk of biological network analysis encompasses the construction and analysis of protein-protein interaction (PPI) networks (PPINs). PPIs are crucial to the formation of macromolecular structures and enzymatic complexes that form the basis of nearly every cellular process ranging from signal transduction and cellular transport to catalyzing metabolic reactions, activating or inhibiting other proteins and biomolecular synthesis. PPIs are thus essential to homeostasis and their dysregulation typically leads to cellular dysfunction and is often associated with various diseases. A systematic mapping of protein interactomes, i.e., the entirety of PPIs in a cell or an organism, is necessary to gain a deeper understanding of the roles of PPIs and PPINs in fundamental cellular processes. It also enables a better understanding of the genotype-phenotype relationships and the perturbations that are involved with the onset of complex diseases.

Owing to their high specificity, PPIs are also promising targets to develop drugs that are attuned to specific disease-related pathways (Jubb *et al.*, 2015; Wells and McClendon, 2007; Murakami *et al.*, 2017). However, in this review, we will focus on the reconstruction and analysis of PPINs and their applications in interpreting available biological data to gain a deeper understanding of cellular processes and disease mechanisms.

The remainder of the chapter is organized as follows: In the first section, we discuss how experimentally defined PPI data have been generated and harnessed for knowledge discovery. Next, we will discuss how and why *in silico* methods for PPI characterization are important in PPIN-based biological research. We will conclude with how future mapping efforts centered around a more dynamic analysis of PPINs will continue to shape the field.

Experimental Methods to Generate Proteome-Scale Interaction Maps

A variety of powerful experimental techniques are now available to characterize PPIs. Initially, however, interactions between protein pairs were described by independent studies that employed small-scale biochemical or genetic experiments (Koh *et al.*, 2012). The steady improvements in experimental methodologies and the development of new technologies that were more amenable to PPI mapping on a larger scale, such as yeast two-hybrid system (Y2H) (Fields and Song, 1989) (see below), have allowed rapidly increasing amounts of PPIs to be characterized. However, the advent of high-throughput genome sequencing technologies and the genomics breakthrough at the turn of the millennium was a milestone that paved the way for the PPI characterization on a proteome-wide scale. The appearance of the first draft model organism genome sequences and the accompanying collection of genome-wide open reading frames (ORFs), coupled with the availability of robust, high-throughput PPI detection methods allowed the PPI mapping to truly take off (Luck *et al.*, 2017). Thus, proteome-scale interaction maps have been generated for different proteomes using available experimental techniques that are amenable to large-scale interactome mapping (Vidal *et al.*, 2011; Huttlin *et al.*, 2015; Rolland *et al.*, 2014).

When considering proteome-scale interaction maps, Johnson and colleagues suggested the classification of large PPI mapping into three categories: one-to-one, one-to-many, and many-to-many approaches (Johnson *et al.,* 2021). The first group leverages the parallelization and automation of Y2H assays. The Y2H system is one of the most widely used methods to map binary PPIs. Y2H is an *in vivo* method based on the reconstitution of a functional transcription factor (TF) following an interaction between two proteins and the subsequent activation of reporter genes controlled by the TF (Fields and Song, 1989). It is a scalable and relatively inexpensive method that is well suited to detecting binary interactions between proteins and therefore facilitates the characterization of physiologically relevant PPIs. Unsurprisingly, Y2H has been the method of choice for generating proteome-wide binary interaction maps for many model organisms such as E. coli (Rajagopala *et al.*, 2014), Yeast (Uetz *et al.*, 2000; Ito *et al.*, 2001; Yu *et al.*, 2008; Vo *et al.*, 2016), C. elegans (Li *et al.*, 2004), Drosophila (Formstecher *et al.*, 2005), Arabidopsis thaliana (Arabidopsis Interactome Mapping Consortium, 2011) and human (Luck *et al.* 2020; Vidal *et al.*, 2011; Huttlin *et al.*, 2015; Rolland *et al.*, 2014). However, Y2H suffers from notable shortcomings; it is less amenable to capturing PPIs involving extracellular or membrane proteins, PPIs that require proper folding as a part of protein complex subunits, or PPIs that require post-translational modifications (PTMs). To overcome the limitations of Y2H and to study different types of PPIs, several Y2H variants such as the mammalian cell-based two-hybrid assay (Luo *et al.*, 1997), the membrane-anchored two-hybrid assay (Snider *et al.*, 2010), and the three-hybrid assay (Maruta *et al.*, 2016) have been developed.

Included in the one-to-many approaches is the mapping of PPIs by Affinity Purification – Mass Spectrometry (AP-MS). This approach involves biochemical purification of the epitope-tagged target proteins from the cells, followed by the identification of the components of the purified protein complexes (including proteins interacting with the target protein) using mass-spectrometry analysis (Dunham *et al.*, 2012). AP-MS method has been widely used to characterize protein complexes on a large scale in different species including yeast, Drosophila and human (Guruharsha *et al.*, 2011; Krogan *et al.*, 2006; Ewing *et al.*, 2007). To overcome the non-specific detection of co-purified proteins, two-step tandem affinity protein purification systems have been developed (Burckstummer *et al.*, 2006). This approach allows the preparation of a substantially pure target protein complex and reduces the background signals. The quantitative mass-spectrometry analysis also has been used to identify different contaminants (Trinkle-Mulcahy *et al.*, 2008). However, AP-MS data may not always detect binary interactions and often reflects only steady-state PPI dynamics, thereby, potentially missing weak and transient interactions.

As for many-to-many approaches, we may consider the use of co-fractionation and mass spectrometry to characterize protein complexes. In this method, cellular extracts are subject to intense co-fractionation by using biochemical separation methods such as chromatography, and a precise co-elution of proteins is used to determine PPIs. A distinct advantage of this method over AP-MS is that it allows for a mapping of dynamic PPIs and the determination of the size of the protein complexes due to the use of size-exclusion chromatography (Yang *et al.*, 2015). Thus, this method has been used to map protein complexes on a proteome-wide scale in different organisms such as yeast and humans (Doerr, 2012; Havugimana *et al.*, 2012; Phanse *et al.*, 2016). More recently, PROPER-seq, a method that leverages the use of high throughput DNA sequencing for the mapping of many-to-many non-binary PPIs has also been introduced. Taking a group of cells as input, this method first generates a barcode for each protein that conjugates the protein itself with its mRNA. Next, the barcodes are grouped into two libraries, a bait library that is kept immobilized and a prey library that is not. Once these libraries are combined, it is possible to identify the interactions in the form of chimeric sequences of interacting mRNA barcodes (Johnson *et al.*, 2021).

Network Based PPIN Analysis

PPINs assembled from the experimentally characterized PPIs are crucial to understanding cellular organization and complex diseases. PPI data extracted from the proteomic literature and compiled within the expert-curated resources are highly useful in uncovering functional PPINs and guiding subsequent research. However, such data are scattered across multiple databases that differ in scope and content, i.e., the type and number of PPIs they contain, the number of organisms that are covered, and the experimental and computational methods that were used for PPI characterization. Therefore, combining different PPI maps is necessary to obtain a complete view of protein interactomes (Razick *et al.*, 2008; Chen *et al.*, 2011; Chen *et al.*, 2016; Chen *et al.*, 2019).

However, the combined PPI datasets will likely be noisy and beset with false positives that are inherent in experimentally characterized PPIs and therefore, they must be carefully assessed before being used for PPIN analyses. A relatively simple and commonly used approach is to consider only those PPIs that are determined by at least two different experimental methods or are reported in the literature in two different publications (Chen *et al.*, 2016). Biophysical data from the experimentally determined structures of interacting proteins can be useful since they offer detailed insights into how PPIs are formed at the atomic scale (De Las Rivas and Fontanillo, 2010; Erijman *et al.*, 2014; Moal *et al.*, 2011, 2013), but such data are available for very few protein complexes. It is also important to view and analyze the PPIN data in the proper spatiotemporal context such as cellular/tissue specificity, protein subcellular localization, gene expression patterns, homologous associations, and PTMs (Schaefer *et al.*, 2013). Several studies have employed PPIN analysis to probe a broad spectrum of biomolecular processes and seek answers to key biological questions (Fig. 1).

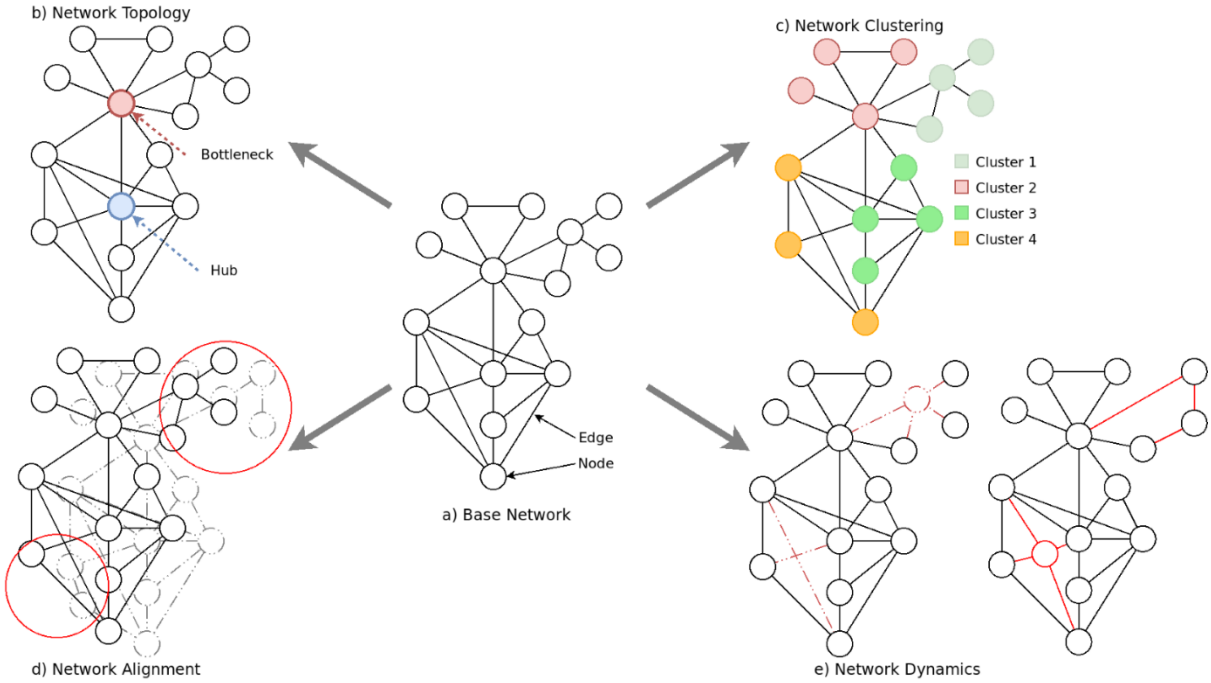

Figure 1. Multiple types of analysis can be performed on a base PPI network.

Network Topology

A typical PPIN is an undirected graph with each protein represented as a node and each interaction between two proteins represented as an edge (Fig. 1(a)). This wiring or the connectivity of the

different proteins within a PPIN is referred to as network topology. There is a strong correlation between the topological properties of a network and its functioning. Therefore, graph theory concepts such as node degree distribution, betweenness centrality, and shortest path length have been used to pinpoint key determinants of network function (Raman, 2010). Network 'hubs' are highly connected proteins with many PPIs (that is, they have a high node degree); they are therefore likely to have a greater influence on network functioning via multiple interactions. Network 'bottlenecks' are proteins with high betweenness centrality; they regulate the flow of signaling information across the network and therefore represent key nodes for communication (Yu *et al.*, 2007) (Fig. 1(b)). Thus, analyzing network topologies can be a means of new discoveries such as identifying novel biomarkers and potential drug targets (Csermely *et al.*, 2013; Kotlyar *et al.*, 2012; Charitou *et al.*, 2016; Gebicke-Haerter, 2016; Hakes *et al.*, 2008). Network topologies have been employed to identify novel disease-associated genes (Vidal *et al.*, 2011; Feldman *et al.*, 2008; Sarajlic *et al.*, 2013) to better understand the organization of localized cellular networks. For example, Gupta et al. (2015) mapped the centriole-cilium protein interaction landscape by generating a PPIN consisting of >7000 interactions and using network topology analysis, which led to the discovery of novel insights into human centrosome and cilia biology.

Network Clustering

It is reasonably understood that proteins group together to form complexes to perform biological functions such as transcription, translation, and cell growth. The clustering of nodes, namely the identification of groups of proteins within a PPIN-based on its intrinsic properties and associated information, is thus a commonly used method for the characterization of protein complexes (Fig 1(c)). In Computer Science literature, clustering algorithms are typically classified into *partitional, hierarchical* and *density*-based methods. Partitional approaches, as the name suggests, generate partitions of the initial data based on the minimization of the difference between the points in a given cluster, expressed usually as some sort of distance. Hierarchical approaches are rooted in the generation of dendrograms that represent the nesting of different elements. Finally, density-based approaches relate to the concept of determining the density of a given region (Jain *et al*., 1999). All of these methods can be potentially used for the clustering of PPINs.

For example, Yu and Zheng used clustering approaches for the identification of complexes in PPIN that have been constructed with weight information from Gene Ontology and van Mering data (Yu and Zheng, 2019); whilst Ranjani Rani and colleagues applied the commonly known Markov Clustering algorithm in conjunction with optimization methods, for the detection of dynamic protein complexes (Ranjani Rani *et al.*, 2019). Also, novel clustering algorithms, specifically developed for the identification of complexes have also been described (Shirmohammady *et al.*, 2021). The effectiveness of different clustering methods in the identification of complexes was examined by Brohée and van Helden, who using a test network of complexes based on the MIPS database, evaluated the sensitivity of various algorithms to the setting of different parameters, and their robustness to alterations in the graph (Brohée and van Helden, 2006).

Another important outcome of using PPIN clustering to investigate disease mechanisms is the disease module hypothesis, which is based on the observation that genes associated with the same disease preferentially interact with each other and tend to form well-connected clusters in the same network neighborhood (Barabasi *et al.*, 2011; Ideker and Sharan, 2008; Menche *et al.*, 2015). The disease module hypothesis has attracted much interest from the researchers since a given set of disease-causing genes can provide a deeper insight into related diseases. This is carried out by collating other disease-causing genes, defining tightly interconnecting communities (Girvan and Newman, 2002) of

functionally related or disease-related proteins, and then retrieving the uncharacterized neighboring genes connected with the initial "seed" genes by shortest paths. For instance, Huttlin and colleagues constructed BioPlex, a network of experimentally derived human PPIs, and defined many protein communities and subnetworks that enabled functional characterization of poorly characterized human proteins, including many with novel roles in human diseases such as cancer and hypertensive disease (Huttlin *et al.*, 2015, 2017). Rolland and co-workers also demonstrated that known cancer-associated genes are highly interconnected in the human protein interactome (Rolland *et al.*, 2014).

Clustering can also be used as a complement or in conjunction with other techniques to extract biological insights. While studying the role of the Rho-GDI signaling pathway in the progression of non-small cell lung cancer, Gupta and colleagues (Gupta *et al.*, 2022) initially used feature selection strategies to reduce the original list of over 10 thousand genes originally sampled from patients, to just over 400. Using this scaled-down list of genes, they constructed a PPIN that, when clustered, could be used to find PPI cliques that were potentially relevant to cancer progression.

Network Alignment

Network alignment allows the discovery of similar parts between molecular systems, particularly those that are evolutionarily conserved between species. The alignment places together the sections of the PPIN that remain constant together, thus clearly identifying the interactions that can be understood as conserved across species; whilst at the same time it highlights the areas that are different across different networks, i.e., thus that should be considered to be species' specific. The alignment is made considering both topological and functional properties of the networks (Ma and Liao, 2020) (Fig 1(d)).

Various approaches have been developed for the alignment of PPI networks; for example, Mahdipour and Ghasemzadeh (2021) introduced a deep learning approach that starts by using different sequence and topological properties of the networks to define embeddings for each of them; these embeddings are then processed by a recurrent neural network in order to predict the alignment of the nodes in the two networks. Alternatively, Menor-Flores and Vega-Rodriguez (2022) focused on improving the alignment results by jointly considering both topological and biological features.

In addition to the comparison across species, network alignment could also play a role when considering PPI data from different cells and tissues. Complex diseases are usually very site-specific and mostly impact specific cells and/or tissues (Goh *et al.*, 2007; Magger *et al.*, 2012). Therefore, it is necessary to examine the PPI data, in cellular and tissue context, such as cell/tissue-specific expression of proteins, the relative abundance of alternatively spliced isoforms and PTMs, and their impact on the interactome of different cells and tissues. Magger and co-workers, for instance, observed that using tissue-specific PPINs greatly enhanced the prioritization of candidate disease-causing genes compared with generic PPINs and highlighted novel tissue-disease associations (Magger *et al.*, 2012).

For more details on current developments of algorithms and metrics for PPI alignment, readers are also encouraged to see the review on the subject from Ma and Liao (2020).

Network Dynamics

Although the interactions that give rise to PPI networks can in many cases be stable, there are instances where the interactions could be temporary, to allow changes to different regulatory processes in response to prevailing conditions (Fig 1(e)). Various efforts thus can also be found when

it comes to the study of the dynamic nature of PPIs. For example, by incorporating expression data into the genome-scale interactome of cassava, Thanasomboom and colleagues were able to rewire the interactions under various conditions, such as drought stress or virus infection (Thanasomboom *et al.*, 2020). Using a similar strategy, Li and colleagues leveraged gene expression to generate aging-specific dynamic PPINs and examined whether these were better suited to predict age-related genes than their static counterparts (Li *et al.*, 2021).

The rapid proliferation of high-quality sequence data generation using low-cost Next Generation Sequencing (NGS) experimental platforms coupled with speedy bioinformatics methods have contributed to the mapping of scores of sequences and structural variants associated with clinically relevant phenotypes and diseases. Although there is a limited understanding of causal relationships between various mutations and diseases, it has been well established that functional variants may often impact overall protein functions including PPIs (Shameer *et al.*, 2016). Comparative PPIN analysis, therefore, offers a promising avenue to examine the genotype-phenotype relationships underlying key biological processes and the causative mechanisms of disease-causing mutations emanating from the gain and/or loss of specific PPIs. Different studies involving the analysis of global PPINs in humans have highlighted mutation-induced network perturbation and loss of specific PPIs that can be reliably linked with specific diseases (Rolland *et al.*, 2014; Sahni *et al.*, 2015).

PPIN rewiring has also been examined in the context of evolution and conservation of PPIs and interactions across species (Fig. 1(e)). Vo and co-workers (Vo *et al.*, 2016) constructed a high-quality binary protein interactome for S. pombe and implemented a framework to compare the organization and evolution of PPINs across yeast and humans. Their findings revealed extensive species-specific network rewiring and novel paradigms on network co-evolution and conservation of interacting proteins.

Despite the wealth of knowledge emerging from the analysis of increasingly available large-scale PPI data, the known protein interactomes are incomplete. This is not only due to the challenges associated with the experimental determination of PPIs, but it was also speculated that to obtain a comprehensive coverage of the human protein interactome, ~200 million protein pairs would need to be experimentally tested (Rolland *et al.*, 2014). Moreover, because of systemic bias, well-studied genes and proteins are screened more frequently than others and are thereby disproportionately represented in literature and PPI databases. This lopsidedness has led to other proteins, potentially the causative agents of diseases, remaining under-represented (Edwards *et al.*, 2011). This issue is particularly visible in model organisms such as rats and mice (Murakami *et al.*, 2017) that are key for biomedical research and it is estimated that only ~10% of the human protein interactome has been characterized so far (Kotlyar *et al.*, 2015). Therefore, to generate a complete interactome, it is necessary to develop computational methods for PPI prediction to expand the coverage of PPI space and mine the protein interactomes for knowledge discovery.

In Silico Prediction of PPIs for PPI Network Analysis and Assessment of PPI Quality

A wide range of *in silico* methods for predicting PPIs have been proposed as complementary to experimental methods and to assess the quality of existing PPIs using their associated features obtained from known PPIs, such as gene co-localization, phylogenetic profiling, gene fusion, domain-domain interactions (DDIs), homologous interactions, contextual information of amino acid (AA) residues, and also using various computational methods such as text mining and machine learning (ML) (Murakami *et al.*, 2017; Peng *et al.*, 2017; Keskin *et al.*, 2016). *In silico* methods can be

broadly classified into two types: Low-resolution methods that offer a simple binary classification to determine whether a given pair of proteins interact or not, and high-resolution methods that can predict the detailed interatomic interactions between proteins (Vakser, 2014). The former can swiftly predict many PPIs as compared with experimental methods and may also be applied to the assessment of known PPIs. The latter can predict PPIs based on their structural and physicochemical complementarities, i.e., protein docking (Tuncbag *et al.*, 2009; Keskin *et al.*, 2016), and require protein structural information, and therefore, are less suitable for characterizing the entire interactome.

Although recent advances in docking methodologies have yielded robust protein complex models, docking proteins with large conformational changes and/or without prior knowledge of the PPI sites (*ab initio* docking) remains a non-trivial task (Janin *et al.*, 2003; Janin and Wodak, 2007). To achieve this task, molecular dynamics (MD) simulations, which take into consideration the physical movements of atoms in proteins, have been used to elucidate the precise positions of atoms involved in the interaction; MD simulations, however, are computational resource intensive and therefore, unsuitable for whole interactome modeling. Moreover, in recent years, it has become possible to predict the structures of monomeric proteins with physical and biological knowledge about protein structure using approaches such as AlphaFold 2.0 (Jumper *et al.*, 2021), and also to predict protein docking models with high accuracy with approaches such as AlphaFold-Multimer (Evans *et al.*, 2022), which is an extension of AlphaFold 2.0. For their efforts, AlphaFold 2.0 chief developers Demis Hassabis and John M. Jumper were awarded one half of the 2024 Nobel Prize in Chemistry. However, those approaches are resource intensive, and the docking model accuracy is yet to be validated fully.

Consequently, most of the existing *in silico* methods applicable to interactome modeling and PPI assessment lean heavily on information obtained from known PPIs, especially their sequence information, which is more widely available than structural information. In addition, *in silico* methods based on only sequence information are useful for predicting PPIs involving proteins for which either structures are yet undetermined, or which are inherently disordered. Below, we discuss the underlying principles of different *in silico* PPI prediction methods for PPIN analysis.

Interolog-Based Methods

Orthologous proteins are descended from a common ancestral gene as a consequence of speciation, and they are believed to retain similarity in structure and function (Fitch, 2000; Koonin, 2005; Watson *et al.*, 2005; Webber and Ponting, 2004), including PPIs. Interologue-based methods predict PPIs and assess the quality of existing PPIs based on the biological principle of orthologous PPIs (interologue) across different species, that is, if two or more proteins are known to interact in a species A and if they have identifiable orthologs in species B, the orthologous proteins may also potentially interact in species B. This approach is similar to those used for gene function annotation, where a gene function is inferred from the function of homologous genes in other species. A large amount of PPI data is available in public databases (Salwinski *et al.*, 2004; Licata *et al.*, 2012; Aranda *et al.*, 2010; Szklarczyk *et al.*, 2015; Chatr-Aryamontri *et al.*, 2017), where orthologous PPIs can be identified. This approach is useful in transferring the annotation of PPIs from one species to another species of interest; for example, it has been applied to predict PPIs in human cancer proteins (Jonsson and Bates, 2006). However, the accuracy of this approach depends on the reliability of the interactions, so it is considered to be inappropriate for the prediction of transient interactions because such interactions are poorly conserved across species (Keskin *et al.*, 2016). Thus, other features, such as domain co-occurrences, gene co-expression or functional similarity, can be integrated into this approach to assess PPIs and the co-localization of the proteins predicted to interact, as implemented in BIPS (Garcia-

Garcia *et al.*, 2012), I2D (Brown and Jurisica, 2005) and PSOPIA (Murakami and Mizuguchi, 2014) (Table 1).

Domain-Based Methods

Protein domains are independent evolutionary units, which define protein function. Multiple studies have demonstrated that DDIs are useful for predicting PPIs since domains are directly involved in intermolecular interactions (Memisevic *et al.*, 2013; Shoemaker and Panchenko, 2007). Domain-based methods can identify PPIs without relying on homologous interactions that exist in public databases, unlike the interologue-based approaches. To use DDIs for the prediction of new PPIs, most methods annotate protein sequences using domain databases such as Pfam (Finn *et al.*, 2016), SCOP2 (Andreeva *et al.*, 2014) and CATH (Sillitoe *et al.*, 2015). There are two types of domain-based approaches (Shoemaker and Panchenko, 2007). The first type consists of the association approach-based methods that are based on the idea that certain domains are frequently observed in interacting proteins and therefore can be used as markers to predict new PPIs (Sprinzak and Margalit, 2001). However, this approach does not consider the relationships of all possible domain pairs in interacting pairs, and the missing domain pairs not observed in known interacting pairs. The second type is the Bayesian network approach, where the interaction probabilities of all possible domain pairs are estimated using the Maximum Likelihood Estimation (Burger and van Nimwegen, 2008; Deng *et al.*, 2002). The accuracy of this approach depends on the reliability of the domain assignments, so sufficient coverage of domain databases is necessary to obtain sufficient true positives and negatives. This approach can also be used to assess the quality of PPIs since an interaction can be deemed more reliable if it contains domain pairs found in known PPIs in the database (Ng *et al.*, 2003).

In recognition of the limitations of the domain-based approaches, a new set of PPI prediction methods has been developed, which are based on the principle of short co-occurring polypeptide regions as mediators of PPIs (Pitre *et al.*, 2012; Schoenrock *et al.*, 2014). A distinct advantage of these methods is that unlike the classical domain-based approaches, they are designed to predict PPIs solely based on primary sequence and are thus, not handicapped by the absence of characterized protein domains; these methods are therefore useful for large-scale PPI prediction. For instance, Schoenrock and colleagues (Schoenrock *et al.*, 2014) designed a tool to predict human PPINs and validated their prediction results experimentally. They further employed their computationally predicted human PPINs for the prediction of gene functions and formations of PPI complexes in human diseases, some of which were validated by follow-up experimental assays (Schoenrock *et al.*, 2014).

Gene Neighborhood-Based Methods

Gene co-localization-based methods are centered on the idea that two proteins are more likely to interact when their genes are in the same region of the genome (Tamames *et al.*, 1997). This approach requires several genome sequences to predict and assess PPIs using information about the conservation of gene locations, and the confidence increases with increasing genome sequences. Although this approach can predict new PPIs without relying on known PPIs reported in the literature or available in databases, it would not be applicable to the eukaryotic genomes, since there is no tangible evidence that two genes that encode for interacting proteins are always co-localized within a genome. Although this approach is simple in comparison with other *in silico* approaches, it often fails to detect interactions between distantly located genes and often generates many false negatives in the eukaryotic genomes (Zahiri *et al.*, 2013).

Phylogenetic Similarity-Based Methods

There are two types of phylogenetic similarity-based approaches. One is the phylogenetic tree-based approach (also known as the mirror tree approach), which is based on the underlying principle that interacting proteins tend to co-evolve through the interaction and thus have similar topological phylogenetic tree profiles (Craig and Liao, 2007; Sato *et al.*, 2005; Pazos and Valencia, 2001; Goh and Cohen, 2002), as implemented in MirrorTree (Ochoa and Pazos, 2010) (Table 1). However, when a pair of proteins co-evolve through the speciation events even if they do not interact, many false positives are created due to the generation of similar mirror trees (Ochoa and Pazos, 2014). The elimination of non-specific tree similarities has been attempted by different methods, for example, the 16S rRNA tree is used as a representation of the speciation process to normalize the non-specific similarities by subtracting their phylogenetic distances from the distance matrices for a pair of proteins (Sato *et al.*, 2005; Pazos *et al.*, 2005). The second type is the phylogenetic profile-based approach that assumes that functionally related proteins tend to be inherited together during evolution. A phylogenetic profile represents the conservation of a certain protein in various species. Thus, if two proteins are functionally related, they are more likely to have similar phylogenetic profiles (Juan *et al.*, 2008). However, the outcomes of this approach are largely dependent on the number of species used to construct phylogenetic profiles. Thus, this approach is not very suitable for eukaryotic proteins since there are comparatively fewer eukaryotes with complete genomic sequences than prokaryotes (Muley and Ranjan, 2012).

Gene Fusion-Based Methods

Gene fusion (domain fusion)-based approaches are based on the genetic observation that independent genes can combine or "fuse" together to form a single chimeric gene, known as the "Rosetta Stone". This method is based on the observation that two separate proteins are functionally related and are likely to interact if certain proteins in a given species consist of co-localized domains that are otherwise mapped to different proteins in another species (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Chia and Kolatkar, 2004). Although gene fusion is an informative feature of the functional relationships between different proteins, it requires a mapping of domain architecture across different genomes and is usually only applicable to proteins corresponding to well-characterized protein domain families.

Function Annotation-Based Methods

Function annotation-based methods are based on the observation that interacting proteins tend to significantly share function annotations since they are involved in the same biological processes (Peng *et al.*, 2017). This method is often used to assess the quality of existing PPIs and evaluate the reliability of different sources with experimentally determined PPIs, for example, the reliability is evaluated by computing the faction of interacting proteins that have at least one identical function (Nabieva *et al.*, 2005). Gene Ontology (GO) can be used to define functional similarity between two proteins to assess the quality of existing PPIs (Cho *et al.*, 2007). However, this approach cannot reliably evaluate the quality of PPIs where the interacting proteins are annotated with many different GO terms (Peng *et al.*, 2017).

Text Mining-Based Methods

Text mining-based approaches use grammatical rules to retrieve the co-occurrence of predefined entities, i.e., in biology, genes, or proteins, and the relationship between these entities in repositories such as literature and various databases (Papanikolaou *et al.*, 2015). An interaction between two

proteins (A, B) can be ascertained if the grammatical rules, such as “A interaction verb B” or “interaction between A and B”, are used in the repositories. Although such approaches cannot retrieve PPIs not described in the repositories, they may be able to infer potentially novel PPIs in a given species based on homologous PPIs in another species. For example, this approach has been used to automatically extract host-pathogen interactions from the biomedical literature (Thieu *et al.*, 2012) and is used in the STRING database to retrieve predicted interactions from the literature (Szklarczyk *et al.*, 2017).

Machine Learning-Based Methods

Machine learning-based methods train a classifier on a set of known PPI data to predict whether a given pair of proteins is likely to interact or not. Many different supervised ML techniques, such as Support Vector Machine (SVM), Naïve Bayes (NB), Neural Networks (NN), k-Nearest Neighbors (kNN), Random Forest (RF), and Deep Learning (DL) have been used to imbibe the informative protein features that can distinguish between true and false interactions. For example, the NB integrating protein domain data, gene expression data, and functional annotation data has been applied to the human interactome network analysis to identify PPIs and subnetworks relevant to human cancer (Rhodes *et al.*, 2005). The kNN has been applied to identify human hereditary disease-gene based on topological features, which describe a protein in PPINs (Xu and Li, 2006). Specifically, DL has been applied to predict human-SARS-CoV-2 PPIs (Yang *et al.*, 2021; Liu-Wei *et al.*, 2021) and to predict PPIs containing cell and tumor information in PPIN prediction based on one-core and crossover network (Li *et al.*, 2022). Furthermore, an ensemble learning approach, which combines multiple scores obtained from different classifiers trained on different ML techniques can be effective (Peng *et al.*, 2017). For example, an ensemble learning method that utilizes four classifiers trained using RF, NB, SVM, and multilayer perceptron (MLP, a type of NN with multi-layers), has been applied to the prediction of PPIs between humans and the hepatitis C virus (HCV) proteins (Emamjomeh *et al.*, 2014). The high quality of the training dataset, i.e., informative and unbiased, is crucial for accurate assessments and predictions, as well as for the evaluation of the ML models. The testing dataset, for example, is classified into three types by examining if the interacting partner proteins in the dataset are similar to the proteins in the training dataset or not (Hamp and Rost, 2015; Park and Marcotte, 2012). Such a classification offers an effective mechanism not only to evaluate the models but also to prepare high-quality datasets. In supervised ML, the quality and quantity of a set of non-PPI data significantly impact the predictions. Non-PPIs can be generated by randomly pairing any proteins or proteins found in different subcellular locations and ignoring the actual interactions and are generally sampled by having a 1:1 or 1:10 ratio of PPIs to non-PPIs. However, data imbalance is an issue that needs to be suitably resolved.

Furthermore, in this approach, various types of protein features (descriptors) are combined and used to train the prediction models; these descriptors include the positions of amino acids (AA), the localization of proteins, domains within proteins, phylogenetic profiles, the degree of the conservation and the physicochemical characteristics of AA, protein sequence profiles (evolutionary profiles), protein sequence embedding, and so on. Various physicochemical properties of AA are available in the AAindex (https://www.genome.jp/aaindex/) database (Kawashima *et al.*, 2008). Protein sequence profiles are a list of preferences for each AA at each position in a multiple sequence alignment (MSA), i.e., a position-specific scoring matrix (PSSM). Protein sequence embedding captures semantic information on AA residues in entire sequences. The widely used embedding methods, such as Word2Vec (Mikolov *et al.*, 2013) and Doc2Vec (Le and Mikolov, 2014), were originally developed in the field of natural language processing (NLP) to obtain the distributed representation of words and documents. In the context of biological sequences, a sequence is regarded as a sentence

and represented by multiple k consecutive AA (k-mer) used to train Word2Vec or Doc2Vec models. These methods were recently applied to the prediction of human and virus protein interactions, showing that they learned the protein features well and enabled a robust prediction of human-virus PPIs (Yang *et al.*, 2020; Tsukiyama *et al.*, 2021). Recently, approaches that combine protein embeddings with dimensionality reduction and transfer learning, an ML technique that repurposes knowledge learned from one task to boost prediction in a related task, have been increasingly used for predicting protein structure and function, including PPIs (Dallago *et al*., 2021).

In addition, although some methods are more suitable for prokaryotes than eukaryotes, the confidence scores assigned to existing PPIs or to potentially interacting protein pairs can be useful to ascertain the reliability of the inferred interactomes for the subsequent PPIN analysis. A list of *in silico* prediction web servers that are useful for PPIN analysis is shown in Table 1.

Conclusions and Future Prospects

A proteome-wide mapping of PPIs and leveraging them in PPIN-based analyses can help to gain knowledge of the genotype-phenotype relationships and the functioning of complex biological systems. Publicly available PPI data are expected to grow substantially and more comprehensive interactome maps are likely to become available in the near future. Consequently, the accuracy and efficacy of the various *in silico* PPI prediction and scoring methods will also likely improve with the increasing amounts of genomic and proteomics data that are likely to become available in the near future. As PPI mapping transitions from capturing steady state associations to dynamic interactions, it will become increasingly necessary to take additional parameters such as protein isoforms, spatiotemporal expression, localization and interaction, and perhaps even the “strength” of the PPIs into consideration to maximize the robustness of biological insights obtained from the PPIN-based analyses.

See also

Algorithms for Graph and Network Analysis: Clustering and Search of Motifs in Graphs. Algorithms for Graph and Network Analysis: Graph Alignment. Algorithms for Graph and Network Analysis: Graph Indexes/Descriptors. Algorithms for Graph and Network Analysis: Traversing/Searching/Sampling Graphs. Natural Language Processing Approaches in Bioinformatics. Network-Based Analysis of Host-Pathogen Interactions.

References

Andreeva, A., et al. (2014). SCOP2 prototype: A new approach to protein structure mining. Nucleic Acids Res. 42 (Database issue), D310–D314.

Arabidopsis Interactome Mapping Consortium. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science. 333 (6042), 601–607.

Aranda, B., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res. 38 (Database issue), D525–D531.

Barabasi, A.L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. Nat. Rev. Genet. 12 (1), 56–68.

Brohée, S., & van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 7(488).

Brown, K.R., & Jurisica, I. (2005). Online predicted human interaction database. Bioinformatics, 21 (9), 2076–2082.

Brown, K.R., & Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 8 (5), R95.

Burckstummer, T., et al. (2006). An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. Nat. Methods. 3 (12), 1013–1019.
Burger, L., & van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol. Syst. Biol. 4, 165.
Charitou, T., Bryan, K., & Lynn, D.J. (2016). Using biological networks to integrate, visualize and analyze genomics data. Genet. Sel. Evol. 48, 27.
Chatr-Aryamontri, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. 45 (D1), D369–D379.
Chen, Y.A., Tripathi, L.P., & Mizuguchi, K. (2011). TargetMine, an integrated data warehouse for candidate gene prioritization and target discovery. PLOS ONE, 6 (3), e17844.
Chen, Y.A., Tripathi, L.P., & Mizuguchi, K. (2016). An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework. Database (Oxf.), 2016:baw009.
Chen, Y.A., Tripathi, L.P., Fujiwara, T., *et al*. (2019). The TargetMine Data Warehouse: Enhancement and Updates. Front Genet., 10:934.

Chia, J.M., & Kolatkar, P.R. (2004). Implications for domain fusion protein-protein interactions based on structural information. BMC Bioinform. 5, 161.
Cho, Y.R., et al. (2007). Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinform. 8, 265.
Craig, R.A., & Liao, L. (2007). Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. BMC Bioinform. 8 (6), 1-12.
Csermely, P., et al. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacol. Ther. 138 (3), 333–408.
Dallago, C., et al. (2021). Learned embeddings from deep learning to visualize and predict protein sets. Current Protocols, 1, e113.
De Las Rivas, J., & Fontanillo, C. (2010). Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. PLOS Comput. Biol. 6 (6), e1000807.
Deng, M., et al. (2002). Inferring domain-domain interactions from protein-protein interactions. Genome Res. 12 (10), 1540–1548.
Doerr, A. (2012). Interactomes by mass spectrometry. Nat. Methods. 9 (11), 1043.
Dunham, W.H., Mullin, M., & Gingras, A.C. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. Proteomics. 12 (10), 1576–1590.
Edwards, A.M., et al. (2011). Too many roads not taken. Nature. 470 (7333), 163–165.
Emamjomeh, A., et al. (2014). Predicting protein-protein interactions between human and hepatitis C virus via an ensemble learning method. Mol. Biosyst. 10 (12), 3147–3154.
Enright, A.J., et al. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature. 402 (6757), 86–90.
Erijman, A., Rosenthal, E., & Shifman, J.M. (2014). How structure defines affinity in protein-protein interactions. PLOS ONE. 9 (10), e110085.
Ewing, R.M., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol. 3, 89.
Evans, R., et al. (2022). Protein complex prediction with AlphaFold-Multimer. bioRxiv. doi: https://doi.org/10.1101/2021.10.04.463034.
Feldman, I., Rzhetsky, A., & Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. Proc. Natl. Acad. Sci. USA. 105 (11), 4323–4328.
Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature. 340 (6230), 245–246.
Finn, R.D., et al. (2016). The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 44 (D1), D279–D285.
Fitch, W.M. (2000). Homology: a personal view on some of the problems. Trends Genet. 16 (5), 227–231.
Formstecher, E., et al. (2005). Protein interaction mapping: A Drosophila case study. Genome Res. 15 (3), 376–384.
Garcia-Garcia, J., et al. (2012). BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. Nucleic Acids Res. 40 (Web Server issue), W147–W151.
Gebicke-Haerter, P.J. (2016). Systems psychopharmacology: A network approach to developing novel therapies. World J. Psychiatry. 6 (1), 66–83.
Girvan, M. & Newman, M.E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12), 7821-7826.
Goh, C.S., & Cohen, F.E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. J. Mol. Biol. 324 (1), 177–192.

Goh, K.I., et al. (2007). The human disease network. Proc. Natl. Acad. Sci. USA. 104 (21), 8685–8690.
Gupta, G.D., et al. (2015). A dynamic protein interaction landscape of the human centrosome-cilium interface. Cell. 163 (6), 1484–1499.
Gupta, S., Vundavilli, H., Allendes Osorio, R.S., *et al*. (2022). Integrative network modeling highlights the crucial roles of Rho-GDI signaling pathway in the progression of non-small cell lung cancer. IEEE Journal of Biomedical and Health Informatics. 26 (9), 4785--4793.
Guruharsha, K.G., et al. (2011). A protein complex network of Drosophila melanogaster. Cell. 147 (3), 690–703.
Hakes, L., et al. (2008). Protein-protein interaction networks and biology – What's the connection?. Nat. Biotechnol. 26 (1), 69–72.
Hamp, T., & Rost, B. (2015). Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics. 31 (12), 1945–1950.
Havugimana, P.C., et al. (2012). A census of human soluble protein complexes. Cell. 150 (5), 1068–1081.
Huttlin, E.L., et al. (2015). The BioPlex network: A systematic exploration of the human interactome. Cell. 162 (2), 425–440.
Huttlin, E.L., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature. 545 (7655), 505–509.
Ideker, T., & Sharan, R. (2008). Protein networks in disease. Genome Res. 18 (4), 644–652.
Ito, T., et al. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA. 98 (8), 4569–4574.
Jain, A.K., Murty, M.N. & Flynn, P.J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.
Janin, J., et al. (2003). CAPRI: A Critical Assessment of PRedicted Interactions. Proteins. 52 (1), 2–9.
Janin, J., & Wodak, S. (2007). The third CAPRI assessment meeting Toronto, Canada, April 20–21, 2007. Structure. 15 (7), 755–759.
Johnson, K.L., Qi, Z., Yan, Z., *et al*. (2021). Revealing protein-protein interactions at the transcriptome scale by sequencing. Molecular Cell. Vol 81, Issue 19, Pages 4091-4103.E9, October 07, 2021.
Jonsson, P.F., & Bates, P.A. (2006). Global topological features of cancer proteins in the human interactome. Bioinformatics. 22 (18), 2291–2297.
Juan, D., Pazos, F., & Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. Proc. Natl. Acad. Sci. USA. 105 (3), 934–939.
Jubb, H., Blundell, T.L., & Ascher, D.B. (2015). Flexibility and small pockets at protein-protein interfaces: New insights into druggability. Prog. Biophys. Mol. Biol. 119 (1), 2–9.
Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 596 (7873), 583-589.
Kawashima, S., et al. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36 (Database issue), D202-205.
Keskin, O., Tuncbag, N., & Gursoy, A. (2016). Predicting protein-protein interactions from the molecular to the proteome level. Chem. Rev. 116 (8), 4884–4909.
Koh, G.C., et al. (2012). Analyzing protein-protein interaction networks. J. Proteome Res. 11 (4), 2014–2031.
Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. 39, 309–338.
Kotlyar, M., et al. (2015). In silico prediction of physical protein interactions and characterization of interactome orphans. Nat. Methods. 12 (1), 79–84.
Kotlyar, M., Fortney, K., & Jurisica, I. (2012). Network-based characterization of drug-regulated genes, drug targets, and toxicity. Methods. 57 (4), 499–507.
Krogan, N.J., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 440 (7084), 637–643.
Le, Q.V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, PMLR, 32 (2), 1188-1196.
Li, Q., Newaz, K., & Milenkovic, T. (2021). Improved supervised prediction of aging-related genes view weighted dynamic network analysis. BMC Bioinformatics. 22 (5120).
Li, S., et al. (2004). A map of the interactome network of the metazoan C. elegans. Science. 303 (5657), 540–543.
Li, X., Han, P., et al. (2022). SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction. BMC Genomics, 23 (474), 1-14.
Licata, L., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 40 (Database issue), D857–D861.
Liu-Wei, W., et al. (2021). DeepViral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes. Bioinformatics 37 (17), 2722-2729.

Luck, K., et al. (2017). Proteome-scale human interactomics. Trends Biochem. Sci. 42 (5), 342–354.
Luck, K., Kim, D.K., Lambourne, L., et al. (2020). A reference map of the human binary protein interactome. Nature 580, 402–408. https://doi.org/10.1038/s41586-020-2188-x.
Luo, Y., et al. (1997). Mammalian two-hybrid system: A complementary approach to the yeast two-hybrid system. Biotechniques. 22 (2), 350–352.
Ma, C.Y., & Liao, C.S. (2020). A review of protein-protein interaction network alignment: From pathway comparison to global alignment. Computational and Structural Biotechnology Journal, 18, 2647–2656.
Magger, O., et al. (2012). Enhancing the prioritization of disease-causing genes through tissue-specific protein interaction networks. PLOS Comput. Biol. 8 (9), e1002690.
Mahdipour, E., & Ghasemzadeh, M. (2021). The protein-protein interaction network alignment using recurrent neural network. Medical & Biological Engineering & Computing. 59, 2263—2286.
Marcotte, E.M., et al. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science. 285 (5428), 751–753.
Maruta, N., Trusov, Y., & Botella, J.R. (2016). Yeast three-hybrid system for the detection of protein-protein interactions. Methods Mol. Biol. 1363, 145–154.
Memisevic, V., Wallqvist, A., & Reifman, J. (2013). Reconstituting protein interaction networks using parameter-dependent domain-domain interactions. BMC Bioinform. 14, 154.
Menche, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science. 347 (6224), 1257601.
Menor-Flores, M., & Vega Rodriguez, M.A. (2021). Decomposition-based multi-objective optimization approach for PPI network alignment. Knowledge-Based Systems 243, 108527.
Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space. In 1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings, 1-12. https://doi.org/10.48550/arXiv.1301.3781.
Moal, I.H., et al. (2013). Scoring functions for protein-protein interactions. Curr. Opin. Struct. Biol. 23 (6), 862–867.
Moal, I.H., Agius, R., & Bates, P.A. (2011). Protein-protein binding affinity prediction on a diverse set of structures. Bioinformatics. 27 (21), 3002–3009.
Muley, V.Y., & Ranjan, A. (2012). Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. PLOS ONE. 7 (7), e42057.
Murakami, Y., et al. (2017). Network analysis and in silico prediction of protein-protein interactions with applications in drug discovery. Curr. Opin. Struct. Biol. 44, 134–142.
Murakami, Y., & Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. BMC Bioinform. 15, 213.
Nabieva, E., et al. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 21 (Suppl 1), i302–i310.
Ng, S.K., et al. (2003). InterDom: A database of putative interacting protein domains for validating predicted protein interactions and complexes. Nucleic Acids Res. 31 (1), 251–254.
Ochoa, D., & Pazos, F. (2010). Studying the co-evolution of protein families with the Mirrortree web server. Bioinformatics. 26 (10), 1370–1371.
Ochoa, D., & Pazos, F. (2014). Practical aspects of protein co-evolution. Front. Cell Dev. Biol. 2, 14.
Papanikolaou, N., et al. (2015). Protein-protein interaction predictions using text mining methods. Methods. 74, 47–53.
Park, Y., & Marcotte, E.M. (2012). Flaws in evaluation schemes for pair-input computational predictions. Nat. Methods. 9 (12), 1134–1136.
Pazos, F., et al. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352 (4), 1002–1015.
Pazos, F., & Valencia, A. (2001). Similarity of phylogenetic trees as an indicator of protein-protein interaction. Protein Eng. 14 (9), 609–614.
Peng, X., et al. (2017). Protein-protein interactions: Detection, reliability assessment, and applications. Brief Bioinform. 18 (5), 798–819.
Phanse, S., et al. (2016). Proteome-wide dataset supporting the study of ancient metazoan macromolecular complexes. Data Brief. 6, 715–721.
Pitre, S., et al. (2012). Short co-occurring polypeptide regions can predict global protein interaction maps. Sci. Rep. 2, 239.
Rajagopala, S.V., et al. (2014). The binary protein-protein interaction landscape of Escherichia coli. Nat. Biotechnol. 32 (3), 285–290.
Raman, K. (2010). Construction and analysis of protein-protein interaction networks. Autom. Exp. 2 (1), 2.

Ranjani Rani, R., Ramyachitra, D., & Brindhadevi, A. (2019). Detection of dynamic protein complexes through Markov Clustering based on Elephant Herd Optimization Approach. Scientific Reports 9 (11106).
Razick, S., Magklaras, G., & Donaldson, I.M. (2008). iRefIndex: A consolidated protein interaction database with provenance. BMC Bioinform. 9, 405.
Rhodes, D.R., et al. (2005). Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. 23 (8), 951–959.
Rolland, T., et al. (2014). A proteome-scale map of the human interactome network. Cell. 159 (5), 1212–1226.
Sahni, N., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell. 161 (3), 647–660.
Salwinski, L., et al. (2004). The Database of interacting proteins: 2004 update. Nucleic Acids Res. 32 (Database issue), D449–D451.
Sarajlic, A., et al. (2013). Network topology reveals key cardiovascular disease genes. PLOS ONE. 8 (8), e71537.
Sato, T., et al. (2005). The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics. 21 (17), 3482–3489.
Schaefer, M.H., et al. (2013). Adding protein context to the human protein-protein interaction network to reveal meaningful interactions. PLOS Comput. Biol. 9 (1), e1002860.
Schoenrock, A., et al. (2014). Efficient prediction of human protein-protein interactions at a global scale. BMC Bioinform. 15, 383.
Shameer, K., et al. (2016). Interpreting functional effects of coding variants: Challenges in proteome-scale prediction, annotation and assessment. Brief. Bioinform. 17 (5), 841–862.
Shirmohammady, N., Izadkhah, H., & Isazadeh, A. (2021). PPI-GA: A novel clustering algorithm to identify protein complexes within protein-protein interaction networks using genetic algorithm. Complexity 2021 (2132516).
Shoemaker, B.A., & Panchenko, A.R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLOS Comput. Biol. 3 (4), e43.
Sillitoe, I., et al. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res. 43 (Database issue), D376–D381.
Snider, J., et al. (2010). Detecting interactions with membrane proteins using a membrane two-hybrid assay in yeast. Nat. Protoc. 5 (7), 1281–1293.
Sprinzak, E., & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. J. Mol. Biol. 311 (4), 681–692.
Szklarczyk, D., et al. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43 (Database issue), D447–D452.
Szklarczyk, D., et al. (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 45 (D1), D362–D368.
Tamames, J., et al. (1997). Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. 44 (1), 66–73.
Thanasomboon, R., Kalapanulak, S., Netrphan, S., & Saithong, T. (2020). Exploring dynamic protein-protein interactions in cassava through the integrative interactome network. Scientific Reports 10 (6510).
Thieu, T., et al. (2012). Literature mining of host-pathogen interactions: Comparing feature-based supervised learning and language-based approaches. Bioinformatics. 28 (6), 867–875.
Trinkle-Mulcahy, L., et al. (2008). Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. J. Cell Biol. 183 (2), 223–239.
Tuncbag, N., et al. (2009). A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. Brief Bioinform. 10 (3), 217–232.
Tsukiyama, S., et al. (2021). LSTM-PHV: prediction of human-virus protein-protein interactions by LSTM with word2vec. Brief Bioinform, 22 (6), 1-9.
Uetz, P., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature. 403 (6770), 623–627.
Vakser, I.A. (2014). Protein-protein docking: From interaction to interactome. Biophys. J. 107 (8), 1785–1793.
Vidal, M., Cusick, M.E., & Barabasi, A.L. (2011). Interactome networks and human disease. Cell. 144 (6), 986–998.
Vo, T.V., et al. (2016). A Proteome-wide fission yeast interactome reveals network evolution principles from yeasts to humans. Cell. 164 (1–2), 310–323.

Watson, J.D., Laskowski, R.A., & Thornton, J.M. (2005). Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. 15 (3), 275–284.
Webber, C., & Ponting, C.P. (2004). Genes and homology. Curr. Biol. 14 (9), R332–R333.
Wells, J.A., & McClendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. Nature. 450 (7172), 1001–1009.
Xu, J., & Li, Y. (2006). Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 22 (22), 2800–2805.
Yang, J., Wagner, S.A., & Beli, P. (2015). Illuminating spatial and temporal organization of protein interaction networks by mass spectrometry-based proteomics. Front. Genet. 6, 344.
Yang, X., et al. (2020). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Comput Struct Biotechnol J, 18, 153-161.
Yang, X., et al. (2021). Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction. Bioinformatics. 37 (24), 4771-4778.
Yu, H., et al. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. PLOS Comput. Biol. 3 (4), e59.
Yu, H., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. Science. 322 (5898), 104–110.
Yu, H., & Zheng, Z. (2019). Protein complex identification based on weighted PPI network with multi-source information. Journal of Theoretical Biology, 477, 77-83.
Zahiri, J., Bozorgmehr, J.H., & Masoudi-Nejad, A. (2013). Computational prediction of protein-protein interaction networks: Algorithms and resources. Curr. Genom. 14 (6), 397–414.

Figures and Tables

Figures

Figure 1 Applications of PPIN-based analysis using protein interactome maps. (a) A base PPIN is usually illustrated with proteins represented as nodes (circles) and their interactions represented as edges (solid lines). Multiple types of analysis can be performed on a base PPIN: (b) identify topological features of the network and their biological significance; (c) finding clusters of tightly related groups of proteins; (d) Determining commonalities across two or more PPINs using alignment and (e) Investigating changes in PPIN over time with network dynamics.

Tables

Table 1: A selection of *in silico* prediction web servers that are useful for PPIN analysis. Indicated are the web address where the service is available, the strategy used for the prediction of interactions, and the reference to the publication associated to the method that can be used for further details.

Web server	**Method, URL**
MirrorTree Server (Ochoa and Pazos 2010)	Phylogenetic similarity
	http://csbg.cnb.csic.es/mtserver/
BIPS**:** Biana Interolog Prediction *Serv*er (Garcia-Garcia *et al*., 2012)	Interolog (across multiple species) / domain / functional annotation
	http://sbi.imim.es/web/index.php/research/servers/bips
I2D**:** Interolog Interaction Database (Brown and Jurisica, 2007)	Interologs (across seven species; human, rat, mouse, fly, worm, yeast, and hhv8)
	http://ophid.utoronto.ca/ophidv2.204
PSOPIA: Prediction Server of Protein-Protein Interactions	Homologs (within the human genome) / domain / the shortest path between two homologous proteins

(Murakami and Mizuguchi, 2014)	https://psopia.mizuguchilab.org/PSOPIA[a]
InterSPPI-HVPPI (Yang *et al.*, 2020)	RF/protein sequence embedding / k-mers / Doc2Vec
	http://zzdlab.com/hvppi/
LSTM-PHV (Tsukiyama *et al.*, 2021)	DP / protein sequence embedding / k-mers / Word2Vec
	http://kurata35.bio.kyutech.ac.jp/LSTM-PHV/

[a] These servers accept only a single protein pair per submission.