COMPUTATIONAL METHODS FOR PROTEIN-PROTEIN INTERACTION IDENTIFICATION

by

Ziyun Ding

Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Biological Sciences
West Lafayette, Indiana
December 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Daisuke Kihara
    Department of Biological Sciences
    Department of Computer Science
Dr. Nicolas Carpita
    Department of Botany and Plant Pathology
Dr. Cynthia Stauffacher
    Department of Biological Sciences
Dr. Daoguo Zhou
    Department of Biological Sciences

Approved by:
    Dr. Jason Cannon
        Head of the Graduate Program

I would like to dedicate my thesis to my beloved grandparents

ACKNOWLEDGMENTS

First and foremost, I really appreciate my major professor, Professor Daisuke Kihara. With his help and advice, I have learned a lot and accomplished more than I could have imagined when I started graduate school.

I also want to thank my committee members, Professors Nicholas Carpita, Markus Lill, Daoguo Zhou, Cynthia Stauffacher. They lent me their expertise to provide context, guidance, and the occasional reality check as I pursued this research.

I am also grateful to the members of the Kihara lab during my PhD career: Tunde Aderinwale, Eman Alnabati, Charles Christoffer, Xusi Han, Aashish Jain, Xuejiao Kang, Ishita Khan, Hyung-Rae Kim, Sai Maddhuri, Lyman Monroe, Lenna Peterson, Woong-Hee Shin, Genki Terashi, Xiao Wang, Qing Wei, Mengmeng Zhu. Thanks for all the questions, answers, discussions, lunches, and so many hours spent together.

I also want to thank all my friends: Tong Qi, Xin Zhao, Jing Lang, Fei Jia, Linzi Chen, Dongdong Ma, Lei Nie, Xingjian Bai, Sherry Xu, Yongkai Chen, Chen Wei, Daiping Lu, Xiaowei Zhang, Dongqi Liu, Ruiyi Mao, Yi Yang, Hongxia Shi, Tian Zhou, Zheyu Jiang, and Zhi Li. I'm so grateful to know all of you as family at Purdue and we celebrated lots of events together.

Finally, I am so grateful to my parents for always supporting me and providing me the best educational resources. I will always remember my grandparents Shujiang Ding and Guifen Sun's spirit to study hard and work hard.

Last but not the least, I want to thank my kitten Sake for keeping company with me for the incredible two years. I am incredibly grateful to my fiancé, Changping Meng, for always supporting me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

3D three-dimensional

AA amino acid

ASA accessible surface area

CAS Co-occurrence Association Score

CNN convolutional neural network

DNA deoxyribonucleic acid

GO gene ontology

IAS interaction association score

LCA lowest common ancestor

LSTM long-short term memory

NMR nuclear magnetic resonance

PAS PubMed Association Score

PCC Pearson correlation coefficient

PDB protein data bank

PPI protein-protein interaction

PPIP protein-protein interaction prediction

RF random forest

RNN recurrent neural network

SVM support vector machine

ABSTRACT

Understanding protein-protein interactions (PPIs) in a cell is essential for learning protein functions, pathways, and mechanisms of diseases. This dissertation introduces the computational method to predict PPIs. In the first chapter, the history of identifying protein interactions and some experimental methods are introduced. Because interacting proteins share similar functions, protein function similarity can be used as a feature to predict PPIs. NaviGO server is developed for biologists and bioinformaticians to visualize the gene ontology relationship and quantify their similarity scores. Furthermore, the computational features used to predict PPIs are summarized. This will help researchers from the computational field to understand the rationale of extracting biological features and also benefit the researcher with a biology background to understand the computational work. After understanding various computational features, the computational prediction method to identify large-scale PPIs was developed and applied to Arabidopsis, maize, and soybean in a whole-genomic scale. Novel predicted PPIs were provided and were grouped based on prediction confidence level, which can be used as a testable hypothesis to guide biologists' experiments. Since affinity chromatography combined with mass spectrometry technique introduces high false PPIs, the computational method was combined with mass spectrometry data to aid the identification of high confident PPIs in large-scale. Lastly, some remaining challenges of the computational PPI prediction methods and future works are discussed.

CHAPTER 1. INTRODUCTION

1.1 History of Protein and Protein-Protein Interaction

The term “protein” was commonly assumed to be first originated in the letter between the Swedish chemist Jons Jacob Berzelius Berzelius and the Dutch Chemist Gerardus Johannes Mulder in 1838. Protein was described as the substance commonly presenting in the plant and animal albumin, silk, eggs, blood serum, and blood fibrin. At that time, the biological function of the protein and how the molecules made up the protein were not clear. Later, as the discovery of enzyme and fermentation process, it was shown that proteins play an important role in metabolism [1]. With the invention of ultracentrifugation, Svedberg did sedimentation experiments on hemocyanins and suggested that the protein consisted of small subunits in 1928 [2]. After Sanger and Thompson determined the complete amino acid sequence of insulin A and B in the 1950s, the consistent of protein were concluded as “each protein has its own unique (amino acid) arrangement; an arrangement that endows it with its particular properties and specificities and fits it for the function that it performs in nature” [3, 4].

One of the first identified regulatory protein-protein interaction is between trypsin and its inhibitor antitrypsin by a quantitative kinetic study of the trypsin activity. The activity of the enzyme-inhibitor complex was decreased slower than pure enzyme after dilution, which indicates that the dissociation of the enzyme-inhibitor complex results in the slower decreased activity [1, 5]. Later on, with more and more protein-protein interactions (PPIs) were identified, the importance of PPIs started to become appreciated.

1.2 Experimental Methods to Identify Protein-Protein Interactions

Identification of protein-protein interactions (PPIs) is important for understanding how proteins work together in a coordinated fashion in a cell to perform cellular functions. The experimental identification methods could be classified by biochemical and biophysical methods into two broader categories.

1.2.1 Biochemical methods

Several biochemical experimental methods are available for determining individual PPIs. Co-immunoprecipitation [6], is the most common method to identify PPIs *in vivo*. In this method, one protein binds with a specific antibody. If the other protein interacts with the antibody bond protein, the interacting partner is be pulled down together and identified by western plot. However, it is also possible that two proteins bind together via bridge proteins. Tandem affinity purification is an immunoprecipitation-based technique for identifying PPIs. The two proteins of interest are bound to two types of agarose beads so that they can be separated by purification. If the two proteins are co-purified, it indicates a PPI. Cross-linking is another biochemical method to create a chemical cross-linker between interacting proteins [7]. This method can help to stabilize the weak or transient protein interactions. Commonly used cross-linkers include non-cleavable NHS-ester cross-linker and the imidoester cross-linker dimethyl dithiobispropionimidate [8].

1.2.2 Biophysical methods

A type of biophysical method is called protein-fragment complementation assay [9]. Where a reporter protein is fragmented into two pieces and expressed together with two query proteins. If the query protein pairs interact with each other, the reporter gene can fuse together and fold into an active reporter. Bimolecular fluorescence complementation is one of this type of biochemical experimental methods when the reporter protein is fluorescent protein. Another very similar type to protein-fragment complementary assay is called fluorescence resonance energy transfer [10]. When the query protein pairs interact, fluorescence resonance energy transfer occurs between the pairs of fluorophore molecules bond with query protein pairs. Because environmental conditions can significantly affect the protein expression, posttranslational modifications, and protein folding and therefore interfere the protein interactions. Moreover, from late 1990s, another in vivo protein-fragment complementary assay called yeast two-hybrid was developed [11]. In this method, the binding domain of transcription factor with bait binds to upstream activating sequence. If the prey protein with activating domain of transcription factor binds to the bait, it will activate the downstream reporter gene. However, yeast-two hybrids assay yields to high false-positive rate [12].

Previous methods require the attachment of proteins to the query protein pairs. Surface plasmon resonance [13] is the most common label-free method to detect PPIs. In this method, one protein is immobilized on the metal surface of the biosensor. And its partner protein flows through the immobilized protein. If two proteins interact and bind together, the light reflection on the metal surface changes. The most quantitative biophysical experiment to identify PPIs is isothermal titration calorimetry [14]. It measures the binding affinity, the stoichiometry, and entropy of the PPIs in a single experiment. Ultimately, biophysical methods such as nuclear magnetic resonance spectroscopy (NMR) [15, 16], X-ray crystallography [17], and electron microscopy [18], can be used to determine the tertiary structure of protein complexes to obtain detailed atomic or molecular level information about how the proteins interact. Another large-scale identification method is affinity chromatography combined with mass spectrometry [19-22]. However, this method also generates false positive interactions. Because proteins can co-elute from the chromatography because of the similar protein size or charge, it doesn't necessarily indicate their physical interactions. In the next section, the limitations of the experimental methods will be further discussed.

1.3 Limitations of experimental methods

However, experimental methods have several shortcomings for detecting PPIs. First, these experimental methods are time-consuming and labor-intensive. Second, the applicability of experimental methods depends on how well assay protocols are established in target organisms. Also, a method might not work on some classes of proteins [23, 24]. Third, it is known that experimental methods often have difficulty in identifying weak interactions, which leaves out many transient interactions [25]. Fourth, it is discussed that results of large-scale methods often have a substantial disagreement with each other, which might be partly due to false positives and false negatives [26-28].

Consequently, PPIs have been identified only for a limited number of organisms; moreover, the coverage of PPI networks is very small for the majority of organisms. This is particularly true for plant species. Table 1.1 shows the statistics of experimentally identified PPIs in representative plant species taken from the BioGRID database [29]. Surprisingly, except for *Arabidopsis thaliana*, virtually no other plant species have experimentally determined PPI data available. Even for *Arabidopsis*, known PPI data cover interactions with only about 35% of proteins. Other representative plant species cover even less protein involved in known PPIs. Thus, it is apparent that plants are largely lagged behind from PPI studies (Table 1.1). In this omics era when various types of large-scale data are combined and used for formulating hypotheses and to interpret experimental data, PPI networks are fundamental reference data to have for studying an organism.

Table 1.1 Statistics of the number of experimentally determined PPIs in representative plant species.

Organism	Common Name	Number of Protein genes	Identified unique PPIs	Unique Proteins in PPIs	Fraction of proteins involved in known PPIs (%)
Arabidopsis thaliana	mousear cress	27,636	35,908	9,574	34.55
Zea mays	maize	37,376	13	21	0.06
Glycine max	soy bean	46,993	39	43	0.09
Oryza sativa (Japonica)	rice	35,679	90	72	0.20
Solanum lycopersicum	tomato	25,613	107	44	0.17
Solanum tuberosum	potato	28,463	2	3	0.01

Note: The statistics of PPIs were taken from the BioGRID database. The number of protein genes were taken from the KEGG database.

1.4 Contributions

The major contribution of this work is to develop a computational method to predict PPIs which can be applied to the large-scale prediction on other plant genome and can aid PPIs identification by size-exclusion chromatography combined with mass spectrometry (SEC-LC/MS) method. There are several barriers to predict PPIs with computational methods including the availability of some features such as annotated gene ontology (GO) terms. To fill this gap, the protein without annotated GO terms can be predicted by Protein Function Prediction (PFP) web server [30]. The PFP assigns the GO terms to query protein based on sequence similarity between the query protein sequence and proteins with annotated GO terms. Then the protein functional similarity can be quantified with a very convenient web application NaviGO. Chapter 2 introduces how to visualize the GO term relationship and quantify functional similarity.

There are several ways to extract useful features for predicting PPIs. Chapter 3 summarizes the common types of features which have been proven to be effective in prediction PPIs, including sequence-based, genomic-based, functional-based, co-expression based, protein tertiary structure-based, and PPI network topology based features.

Another problem faced by computational prediction of PPIs is the reliability of experimentally identified PPIs. Based on the statistics, BioGrid database, there are 1,345,800 PPIs are identified by experimental method [29]. 50.5% (679,718 out of 1,345,800) are physical interactions. When the known PPIs are used as training data to build the model, the reliability of the PPIs needs to be checked. The training data used in Chapter 4 only includes the physical PPIs, therefore our prediction focuses on physical interaction. The PPI prediction method was applied to common commercial plants including maize and soybean to increase the number of PPIs in these two species.

Finally, in Chapter 5, the computational PPI prediction method was applied with SEC-LC/MS to aid the identification of high-confident PPIs. In this work, I'm not only able to identify the commonly known protein complexes but also identified some interesting protein complexes with high probability score in the STRING database and high functional correlation of their subunits.

PPIs is very important to establish functionality in the cellular processes. Computational methods are an important complement to experimental methods to boost up the number of PPIs in the not very well studied species. Computationally predicted PPIs, as well as their predicted probabilities and feature scores, can serve as the testable hypothesis to guide biologists.

CHAPTER 2. COMPUTING AND VISUALIZING GENE AND PROTEIN FUNCTION SIMILARITY WITH NAVIGO[1]

2.1 Background

The Gene ontology (GO) is a widely-used vocabulary for representing gene functions across all species [33, 34]. It is maintained and updated by the Gene Ontology Consortium. Currently, GO terms are classified into three categories: Biological Process (BP), which describes pathway information of gene products such as cellular physiological process or signal transduction, Molecular Function (MF), which describes molecular level activities such as enzymatic activity, and Cellular Component (CC), which describe cellular localization of gene products. Currently, there are over 46,000 GO terms, which are organized in a hierarchical structure, a directed acyclic graph (DAG). GO is very useful, particularly for computational analysis of gene functions; however, the volume of the vocabulary and the complicated relationships often make analysis cumbersome.

NaviGO was developed to facilitate easy handling of GO terms, particularly for quantifying and visualizing relationships between GO terms [31]. NaviGO has four main functions: "GO Parents", "GO Set", "GO Enrichment", and "Protein Set". "GO Parents" maps and visualizes the hierarchical relationship of GO terms in an interactive fashion and "GO Set" calculates six functional similarity and association scores, and provides two visualization tools, a network and a multidimensional scaling visualization. For a list of proteins and associated GO terms, "GO Enrichment" performs GO enrichment analysis, while "Protein Set" identifies functionally related proteins. Compared with other related online tools [35, 36], NaviGO server has several advantages: first, it provides multiple similarity scores, which not only compare GO terms in the same GO category but also across GO categories. NaviGO provides biologists an intuitive and interactive tool to visualize parental relationships between GO terms. NaviGO is also integrated into the popular gene function prediction webservers PFP [30, 37] and ESG [38, 39].

---

[1] Portions of this chapter have been previously published 31. Wei, Q., et al., *NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology.* BMC Bioinformatics, 2017. **18**(1): p. 177, 32. Ding, Z., Q. Wei, and D. Kihara, *Computing and Visualizing Gene Function Similarity and Coherence with NaviGO*, in *Data Mining for Systems Biology*. 2018, Springer. p. 113-130.

2.2 Methods

2.2.1 Access to NaviGO

NaviGO can be freely accessed at http://kiharalab.org/web/navigo/. It is a web application and does not require any platform other than a web browser. The source codes of NaviGO and GOVisualizer, a tool for visualizing the hierarchy of GO terms, can be downloaded from Github at https://github.com/kiharalab/NaviGO and https://github.com/kiharalab/GOVisualizer under the terms of the GNU Lesser General Public License Ver. 2.1.

In order to use NaviGO, users need to provide a set of GO terms or a set of UniProt IDs of proteins and associated GO terms to be analyzed. These can be retrieved from the Gene Ontology Consortium website (http://www.geneontology.org/) [33] or from the UniProt database (http://www.uniprot.org/) [40], respectively.

2.2.2 GO similarity/association scores

In NaviGO, six scores can be used to quantify similarity or association relationship of GO terms. Three scores are for quantifying semantic similarity of GO terms: Resnik's, Lin's, and the relevant semantic similarity score. The other three scores, CAS, PAS, and IAS are for quantifying GO associations. Detailed explanation of the scores is provided in separate sections below.

To quantify the functional similarity of two genes, the *funsim* score [30, 41] is used. Funsim of two sets of terms, i.e. GO annotations of two genes, is calculated from an all-by-all similarity matrix, where each entry of the matrix is a similarity score of users' choice between a GO pair.

2.2.3 CAS and PAS

Previously two function association scores, Co-occurrence Association Score (CAS) and PubMed Association Score (PAS) were developed [42]. CAS quantifies frequency of co-occurring GO terms within the gene annotations in the GOA database while PAS takes consideration of co-occurrence of GO terms in PubMed abstracts. A characteristic differentiating the two methods from other methods is that the two scores can be defined cross-domain associations between GO terms, i.e. terms from Molecular Function (MF) and Biological Process (BP), those from MF and Cellular Component (CC), and those from BP and CC.

$$CAS(i,j) = \frac{\frac{c(i,j)}{\sum_{ij} c(i,j)}}{\left(\frac{c(i)}{\sum_k c(k)}\right)\left(\frac{c(j)}{\sum_k c(k)}\right)} \qquad \text{(Equation 2.1)}$$

where *C(i,j)* is the number of sequences in the database that contain both the GO terms *i* and *j*. Similarly, *C(i)* is the total number of sequences annotated with the GO term *i*, and so is the *C(j)*. The numerator of Eqn. 1, $\frac{c(i,j)}{\sum_{ij} c(i,j)}$, is essentially the fraction of sequences that are annotated with two particular GO terms, *i* and *j*, among all the sequences in the database. The denominator multiplies the fraction (probability) of sequences in the database that are annotated with GO term *i* and the fraction of sequences in the database that are annotated with GO term *j*. Thus, it is the expected fraction of sequences in the database with the two GO annotations, *i* and *j*, if i and j are randomly assigned to sequences. Using the numerator and the denominator, altogether CAS quantifies how often two GO terms *i* and *j* co-annotate sequences relative to the random chance. CAS = 1 means that the observation of co-annotation of *i* and *j* is the same as expected by the random chance, and a larger value indicates that *i* and *j* are correlated in gene annotation.

Similarly, PAS is defined as:

$$PAS(i,j) = \frac{\frac{Pub(i,j)}{\sum_{i,j} Pub(i,j)}}{\left(\frac{Pub(i)}{\sum_k Pub(k)}\right)\left(\frac{Pub(j)}{\sum_k Pub(k)}\right)} = \frac{Pub(i,j)}{Pub(i)Pub(j)} \cdot \frac{\left(\sum_k Pub(k)\right)^2}{\sum_{k,l} Pub(k,l)} \qquad \text{(Equation 2.2)}$$

Here, *Pub(i,j)* is the number of PubMed abstracts which contain both the GO terms *i* and *j*. Similarly, *Pub(i)* is the number of abstracts that contain GO term *i* and the same is applicable for *Pub(j)*. The numerator of Eqn. 2, $\frac{Pub(i,j)}{\sum_{ij} Pub(i,j)}$, is the fraction of abstracts in PubMed that mention two particular GO terms, *i* and *j*, among all the abstracts in the PubMed database. The denominator multiplies the fraction (probability) of abstracts in PubMed that mention GO term *i* and the fraction of abstracts that mention GO term *j*. Thus, it is the expected fraction of abstracts in the database with the two GO annotations, *i* and *j*, if i and j randomly show up in abstracts. Altogether, PAS quantifies how often two GO terms *i* and *j* are co-mentioned in PubMed abstracts relative to the random chance. PAS = 1 means that GO term *i* and *j* are not related, and a larger value indicates that *i* and *j* are related and frequently co-mentioned in biological contexts. Importantly, it is possible that GO terms that do not have a high functional similar scores (Resnik, Lin's, and Relevance Similarity scores) have a high CAS or PAS. High PAS and CAS implies that proteins with the GO term annotation are functionally related and play roles in the same biological context, e.g. pathways.

2.2.4 IAS

The Interaction Association Score (IAS) [43] captures the propensity of GO term pairs to occur in interacting proteins by counting the number of GO term pair that occur in interacting proteins normalized by random chance. Thus, high IAS between a protein pair indicates a high possibility that the protein pairs interact with each other. The GO_IAS for each GO term pair was computed as follows:

$$GO\_IAS(GOx, GOy) = \frac{\frac{N(GOx,GOy)}{\#T.Edges}}{\left(\frac{N(GOx)}{\#T.Nodes}\right)\left(\frac{N(GOy)}{\#T.Nodes}\right)} \qquad \text{(Equation 2.3)}$$

where *N(GOx-GOy)* is the number of times GO term pair *GOx* and *GOy* interact in PPI networks, *#T.Edges* is the total number of interactions (edges) in PPI networks, *N(GOx)* and *N(GOy)* are the number of times *GOx* and *GOy* independently occur in proteins the networks, and *#T.Nodes* is the total number of proteins in the PPI networks. Figure 2.9 shows an example of a small PPI network. This network has 5 edges between 5 proteins; 3 proteins are annotated with GO:1, and 2 proteins with GO:2. There are 2 edges that connects between GO:1 and GO:2 (P1 to P2 and P2 to P4). From this network, GO_IAS for GO:1 and GO:2 is computed as (2/5)/((3/5)(2/5)) = 1.67. Similar to PAS and CAS, IAS quantifies how often two GO terms *i* and *j* are observed in physically interacting proteins in a protein-protein interaction network relative to the expected number of observations by the random chance. If two proteins are annotated with GO terms that have high IAS, it suggests that the proteins might physically interact with each other.

Significant difference between CAS, PAS, and IAS from conventional GO functional similarity scores described in the next section is that the former three scores quantifies functional relevance of GO term pairs in biological contexts, co-annotation to genes (CAS), co-mention in PubMed abstracts (PAS), and interacting protein pairs (IAS). Due to the design, these scores are capable of identifying proteins in the same pathways (CAS, PAS) [42] and physical interacting proteins (IAS) [43]. Correlation of CAS/PAS/IAS to regular functional similarity scores (below) is not very high [42, 43], because proteins in the same pathway and physically interacting proteins are not necessarily having similar function.

2.2.5 Resnik, Lin's, and Relevance Similarity Scores

For quantifying GO term similarity, NaviGO provides three score options. The Resnik's [41] similarity score measures the semantic similarity of a GO term pair according to the lowest common ancestor (LCA) of the GO term pair, while the Lin's similarity is based on the information content of LCA and the GO term pair queried [42].

$$sim_{Resnik}(c_1, c_2) = \max_{c \in S(c_1, c_2)} (-\log p(c)) \qquad \text{(Equation 2.4)}$$

$$sim_{Lin}(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left(\frac{2 \cdot log\ p(c)}{log\ p(c_1) + log\ p(c_2)}\right) \qquad \text{(Equation 2.5)}$$

Here *p(c)* is the probability of a GO term c, which is defined as the fraction of the occurrence of *c* in the GO Database. *s(c1,c2)* is the set of common ancestors of the GO terms c1 and c2. The root of the ontology has a probability of 1.0.

The relevance semantic similarity score (sim_{Rel}) [41] for computing functional similarity of a pair of GO terms, c1 and c2:

$$sim_{Rel}(c_1, c_2) = \max_{c \in S(c_1, c_2)} \left(\frac{2 \cdot log\ p(c)}{log\ p(c_1) + log\ p(c_2)} \cdot (1 - p(c))\right) \qquad \text{(Equation 2.6)}$$

The first term considers the relative depth of the common ancestor c to the average depth of the two terms c1 and c2 while the second term takes into account how rare it is to identify the common ancestor c by chance.

2.2.6 Functional Similarity Score of Gene Pairs

To quantify the functional similarity of two annotated genes, the *funsim* score was used [30, 41]. The *funsim* score of two sets of terms, GO^{A} and GO^{B} for gene A and B, of a respective size of N and M, is calculated from an all-by-all similarity matrix s_{ij}.

$$s_{ij} = sim\left(GO_i^A, GO_j^B\right)_{\forall i \in \{1..N\}, \forall j \in \{1..M\}} \qquad \text{(Equation 2.7)}$$

For $sim(GO_i^A, GO_i^B)$, the relevance similarity score is usually used but other scores can be used, too. Since the relevance similarity score is defined only for GO pairs of the same category, a matrix is computed separately for the three categories, BP, CC, and MF:

$$GO_{score} = \max\left(\frac{1}{N}\sum_{i=1}^{N} \max_{1 \le i \le M} s_{ij}, \frac{1}{M}\sum_{i=1}^{M} \max_{1 \le j \le N} s_{ij}\right) \qquad \text{(Equation 2.8)}$$

GOscore will be any of the three category scores (MFscore, BPscore, CCscore). Finally, the *funsim* score is computed as

$$funsim = \frac{1}{3}\left[\left(\frac{MFscore}{\max(MFscore)}\right)^2 + \left(\frac{BPscore}{\max(BPscore)}\right)^2 + \left(\frac{CCscore}{\max(CCscore)}\right)^2\right] \quad \text{(Equation 2.9)}$$

where *max(GOscore)* = 1 (maximum possible GOscore) and the range of the *funSim* score is [0, 1].

2.2.7 Gene Ontology Enrichment Analysis

The probability of a GO term X being annotated to a protein in the cluster is computed by:

$$f(k; N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-\mathrm{k}}}{\binom{N}{n}} \quad \text{(Equation 2.10)}$$

where *k* is the number of proteins in the cluster annotated with X, *N* is the number of annotated proteins in the organism, *m* is the number of proteins in the organism annotated with X, and *n* is the number of annotated proteins in the cluster. To calculate a p-value for overrepresentation of a term, the following equation is used:

$$P_{hg}(X) = \sum_{i=k}^{n} f(i; N, m, \mathrm{k}) \quad \text{(Equation 2.11)}$$

2.3 Using NaviGO to Visualize GO Terms and Quantify Similarity

2.3.1 Overview of NaviGO

The NaviGO server has four main functions (Figure 2.1). Either a set of GO terms or a set of proteins (with their GO terms) can be analyzed. For a set of GO terms, the “GO Parents” tab visualizes input GO terms in the GO DAG, and the “GO Set” tab calculates the functional similarity and association scores and visualizes them. On the other hand, for a list of proteins with their GO terms, the “GO Enrichment” tab performs GO enrichment analysis and the “Protein Set” tab calculates functional similarity and association scores between proteins (Figure 2.1).

Throughout this tutorial, the following six proteins were used, which are involved in the light signaling pathway, as examples: phytochrome A (PHYA, UniProt ID: P14712), phytochrome B (PHYB, UniProt ID: P14713), phytochrome D (PHYD, UniProt ID: P42497), phytochrome-interacting factor 3 (PIF3, UniProt ID: O80536), pseudo-response regulator 7 (PRR7, UniProt ID: A0A1P8BCB0), and histone deacetylase 15 (HDA15, UniProt ID: Q8GXJ1) (Table 1). PHYA, PHYB, and PHYD are from the phytochrome family and mainly function as red and far-red photoreceptors. They have been experimentally verified to interact with each other [44]. Interaction of the transcription factor PIF3 with the phytochrome family causes phosphorylation and degradation of phytochrome [45, 46]. Interaction of PIF3 with HDA15, a transcriptional repressor, represses the chlorophyll biosynthesis and the photosynthesis [47]. PRR7 is one of the key components of molecular clock in Arabidopsis and involved in the phytochrome- mediated red light signal transduction pathway [48]. PRR7 interacts with phytochrome and PIF to regulate the red light signal transduction. Known physical interactions of the six proteins are summarized in Figure 2.2.

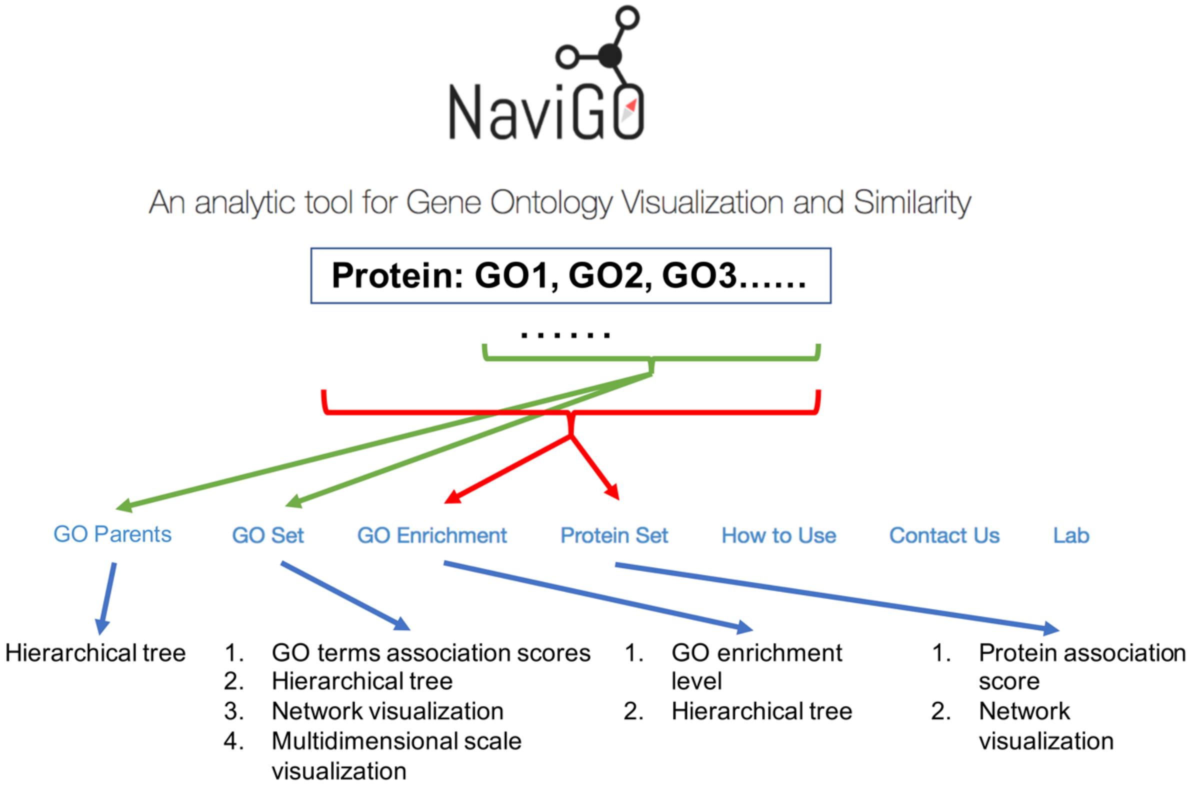


Figure 2.1 Overview of NaviGO functionality. Input to be analyzed can be either a set of GO terms or a set of protein

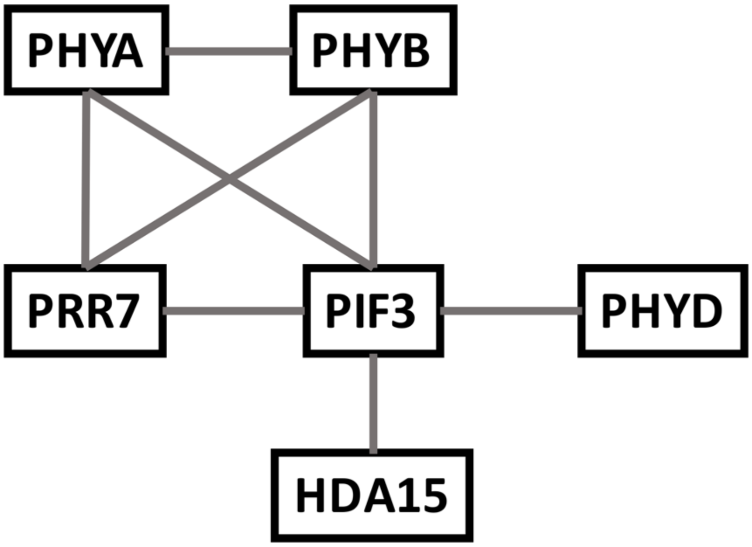


Figure 2.2 The interaction relationship of the six example proteins. These six proteins are involved in the light signaling pathway.

2.3.2 Quantification and visualization of GO term association and similarity

The “GO Set” tab computes six GO term similarity and association scores for all the pairs of input GO terms. The scores are Resnik’s, Lin’s, Relevance similarity score (RSS), the interaction association score (IAS), the PubMed association score (PAS), and the co-occurrence association score (CAS). The first three scores quantify similarity of a GO term pair of the same category. They are calculated based on the frequencies of two GO terms in the gene annotation database and their location in the GO DAG [41, 49, 50]. Among the three scores, RSS not only considers the relative depth of the common ancestor between the two GO terms, but also considers how rare the query GO terms are to identify the common ancestor. The last three semantic-based functional similarity scores were developed by our group. IAS quantifies the probability that two GO terms appear in physically interacting protein pairs [51]. PAS and CAS quantify the frequency with which two GO terms appear in the same PubMed abstract and in a single gene annotation, respectively [52].

To use the “GO Set” tab, please follow the steps described below:

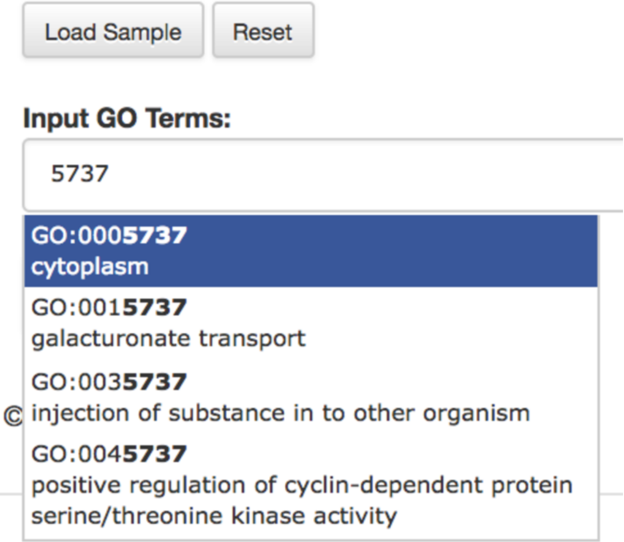
1. Enter your input in the box. The input format of the “GO set” tab is a list of GO terms separated by comma. Users can upload a formatted file or type in the GO term ID. As a

GO term ID is being typed, NaviGO will automatically recognize the GO term with the number and show candidates in a pull-down list. Thus, users can choose one from the list.

2. For example, “GO:0005737” can be retrieved after typing “5737” and clicking the first GO term in the pull-down list (Figure 2.3).
3. To empty inputs, click the “Reset button” located above the input box. To delete a single GO terms in the input box, click the “X” sign at the GO term.
4. Click the “Submit” button below the input box will start the analysis and show a result page when done.

At the top of the results page, query GO term scores are listed in colors that indicate categories: BP terms are in red, MF in blue, and CC in yellow (Figure 2.4, top). The numbers on the right side of GO terms are the counts of each GO term in the input. Clicking the BP/MF/CC Visualizer button below the query GO term list will open a new page that shows the GO terms of the category in the GO DAG (Figure 2.5). The color legends of GO terms are listed on the right side, and colors of GO term relationships are shown in the left upper corner of the page. The query GO terms are shown in a larger font in the DAG. Clicking a GO term in a graph will expend links to all the children GO terms. In the example shown in Figure 2.5, seven molecular function GO terms are mapped (Figure 2.5). We can see that the GO terms locate in two branches, photoreceptor activity and protein binding activity.

Pairwise GO term scores are calculated and listed in the table below the input GO term list (Figure 2.4, bottom). GO pairs in the table can be sorted by a score by clicking the title of the score column. If the members of a pair of GO terms do not belong to the same category or the score of the pair is not available, “n/a” is shown. The significance level of scores in each column are indicated in a color scale, from light pink to red as the significance level increases. Clicking the “+” in the “common parents” column expands the list of all common parents of the GO pair in the GO DAG. Clicking a GO term will take users to the AmiGO website, which provides more detailed information of the term. The results table can be also downloaded in a comma separated data file (a CSV format file) by clicking the “CSV file” button.



5737

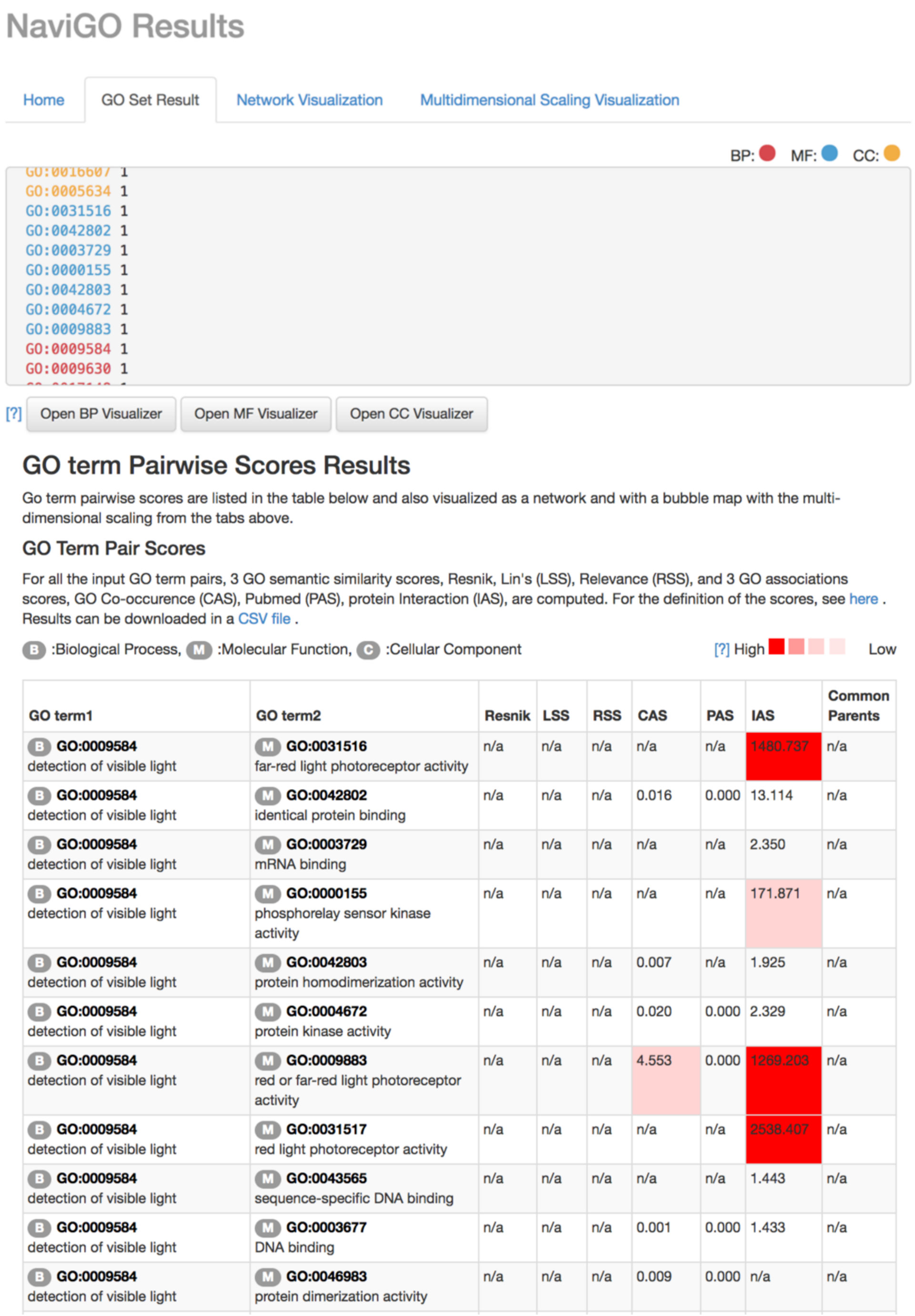
GO:0005737
cytoplasm

GO:0015737
galacturonate transport

GO:0035737
injection of substance in to other organism

GO:0045737
positive regulation of cyclin-dependent protein serine/threonine kinase activity

Figure 2.3 Example of inputting GO terms



```
GO:0016607 1
GO:0005634 1
GO:0031516 1
GO:0042802 1
GO:0003729 1
GO:0000155 1
GO:0042803 1
GO:0004672 1
GO:0009883 1
GO:0009584 1
GO:0009630 1
```

GO term Pairwise Scores Results

Go term pairwise scores are listed in the table below and also visualized as a network and with a bubble map with the multi-dimensional scaling from the tabs above.

GO Term Pair Scores

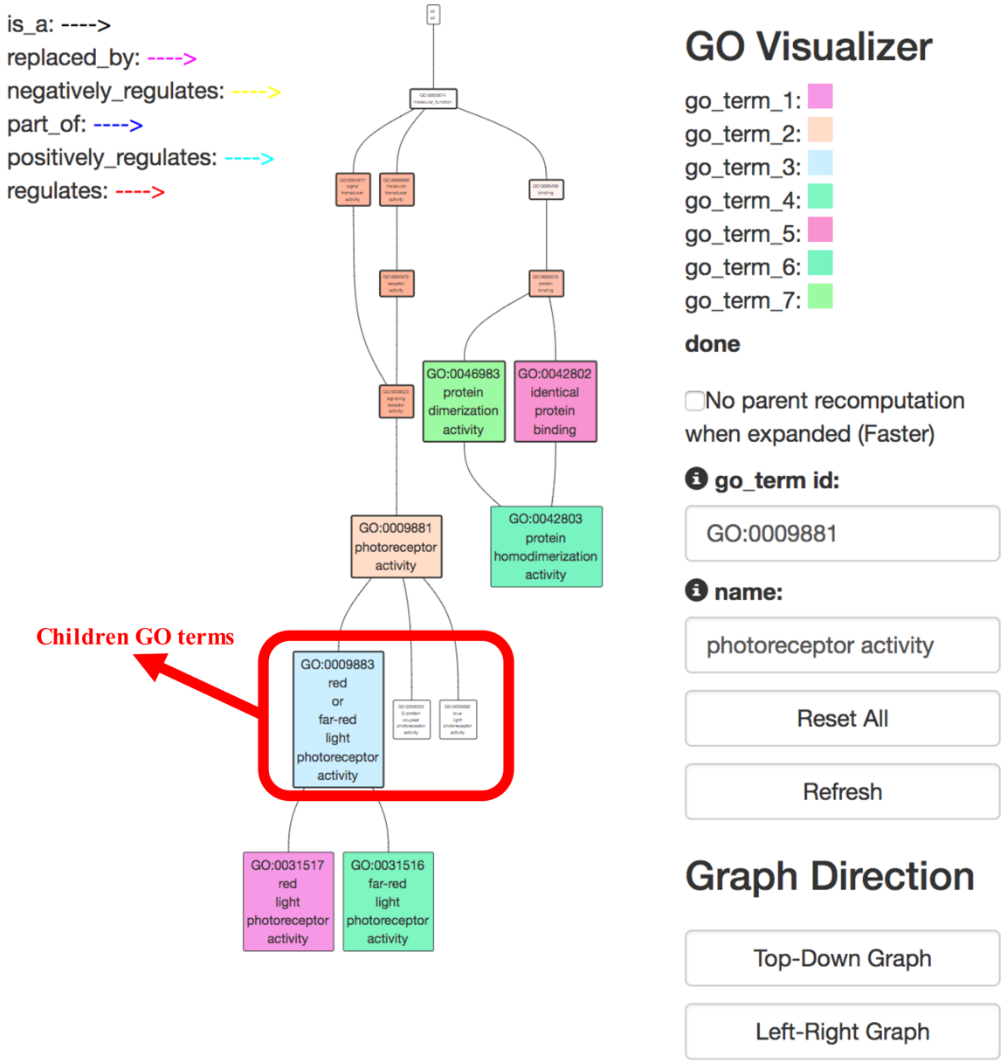
For all the input GO term pairs, 3 GO semantic similarity scores, Resnik, Lin's (LSS), Relevance (RSS), and 3 GO associations scores, GO Co-occurence (CAS), Pubmed (PAS), protein Interaction (IAS), are computed. For the definition of the scores, see here . Results can be downloaded in a CSV file .

B :Biological Process, M :Molecular Function, C :Cellular Component

[?] High ■ ■ ■ ■ Low

GO term1	GO term2	Resnik	LSS	RSS	CAS	PAS	IAS	Common Parents
B **GO:0009584** detection of visible light	M **GO:0031516** far-red light photoreceptor activity	n/a	n/a	n/a	n/a	n/a	1480.737	n/a
B **GO:0009584** detection of visible light	M **GO:0042802** identical protein binding	n/a	n/a	n/a	0.016	0.000	13.114	n/a
B **GO:0009584** detection of visible light	M **GO:0003729** mRNA binding	n/a	n/a	n/a	n/a	n/a	2.350	n/a
B **GO:0009584** detection of visible light	M **GO:0000155** phosphorelay sensor kinase activity	n/a	n/a	n/a	n/a	n/a	171.871	n/a
B **GO:0009584** detection of visible light	M **GO:0042803** protein homodimerization activity	n/a	n/a	n/a	0.007	n/a	1.925	n/a
B **GO:0009584** detection of visible light	M **GO:0004672** protein kinase activity	n/a	n/a	n/a	0.020	0.000	2.329	n/a
B **GO:0009584** detection of visible light	M **GO:0009883** red or far-red light photoreceptor activity	n/a	n/a	n/a	4.553	0.000	1269.203	n/a
B **GO:0009584** detection of visible light	M **GO:0031517** red light photoreceptor activity	n/a	n/a	n/a	n/a	n/a	2538.407	n/a
B **GO:0009584** detection of visible light	M **GO:0043565** sequence-specific DNA binding	n/a	n/a	n/a	n/a	n/a	1.443	n/a
B **GO:0009584** detection of visible light	M **GO:0003677** DNA binding	n/a	n/a	n/a	0.001	0.000	1.433	n/a
B **GO:0009584** detection of visible light	M **GO:0046983** protein dimerization activity	n/a	n/a	n/a	0.009	0.000	n/a	n/a

Figure 2.4 In Figure 2.4, a part of the result page for the example input in Table 2.1 is shown. IAS of the BP term GO:0009584 (related to detection of visible light) and the MF term GO:00031516 (far-red light photoreceptor activity) is highlighted in red, because the score 1480.737 is within the top 1% of scores relative to the score background distribution. Both GO terms annotate the phytochrome proteins, which are known to form heterodimers [44], so it is reasonable that the two GO terms annotating these interacting proteins have a very high IAS.


is_a: ---->
replaced_by: ---->
negatively_regulates: ---->
part_of: ---->
positively_regulates: ---->
regulates: ---->
GO:0046983 protein dimerization activity
GO:0042802 identical protein binding
GO:0042803 protein homodimerization activity
GO:0009881 photoreceptor activity
Children GO terms
GO:0009883 red or far-red light photoreceptor activity
GO:0031517 red light photoreceptor activity
GO:0031516 far-red light photoreceptor activity
GO Visualizer
go_term_1: 🟪
go_term_2: 🟧
go_term_3: 🟦
go_term_4: 🟩
go_term_5: 🟪
go_term_6: 🟩
go_term_7: 🟩
done
Reset All
Refresh
Graph Direction
Top-Down Graph
Left-Right Graph

☐ No parent recomputation when expanded (Faster)

go_term id: GO:0009881

name: photoreceptor activity

Figure 2.5 Hierarchical graph representation using GO Visualizer. Seven molecular function GO terms are visualized here, GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803, and GO:0046983. Clicking a GO term expands edges to all the children GO terms.

NaviGO provides two types of visualizations for GO pairwise score results. One is a network visualization available under the "Network Visualization" tab (Figure 2.6). In the network, functionally related GO terms are connected by edges. The score cutoff value to define edges can be controlled by a sliding the bar or by typing the value in a text box. The scores to visualize can be chosen at the upper left corner of the page. In the example in Figure 2.6, the same set of GO terms as Figure 2.5 was used. In the network with RSS (Figure 2.6, left), the GO terms were clustered into two groups, which is consistent with the two branches in the GO hierarchy shown in Figure 2.5. Using IAS, all GO terms are connected (Figure 2.6B). This is also reasonable because these terms are associated with light signaling proteins, and they are known to interact with each other as shown in Figure 2.2.

The second visualization is available at the "Multidimensional Scaling Visualization" tab. In this two-dimensional (2D) graph, GO terms are classified and mapped onto a 2D space with two scores selected by users. Placing the cursor over a GO term will show the normalized functional score of the GO term. In the example in Figure 2.7, the x-axis is RSS and the y-axis is PAS. The GO terms are largely classified in two groups, which is again consistent with the results in Figure 2.5 and Figure 2.6.

2.3.3 GO enrichment analysis

The goal of GO enrichment analysis is to find if any GO term appears more frequently in a set of proteins than would be expected from the background frequency of the term in the genome. The significance of protein is quantified by a *p-value.* A *p-value* of a GO term for a protein set is calculated by considering the number of proteins in the set, the number of proteins annotated with the GO term in the genome, and the total number of proteins in the organism. The smaller the *p-value* is, the more significant the GO term is.

The input format of the "GO Enrichment" tab is a list of UniProt IDs associated with GO terms, e.g. "A0A1P8BCB0 GO:0005634, GO:0000160". NaviGO automatically identifies the organism based on the UniProt ID of the first protein in the input.

In the results page, GO terms of input proteins are sorted by their p-value (Figure 2.8). The significant p-value (below 0.00005 or top 30) GO terms are highlighted in red. The total number of significantly enriched GO terms is counted in the box on the left from "Open GO Visualizer". The third column shows the number of input proteins that have the GO term. In the example shown in Figure 2.8, the most enriched GO term among the proteins from Arabidopsis is GO:0000155 *phosphorelay sensor kinase activity* with a p-value of 3.57E-10 and GO:0018298 *protein-chromophore linkage* with a p-value of 8.11E-9.

Enriched GO terms with p-value above 0.00005 (or the top 30 GO terms) can be mapped to the GO hierarchy by clicking "open GO Visualizer". The enriched GO terms are shown in a larger font and colored based on p-value from red to yellow indicating most to least significance. The figure can be downloaded by clicking "Download Figure Here". In the example in Figure 2.9, the significantly enriched GO terms are involved in the signal receptor activity such as GO: 0000155, "phosphorelay sensor kinase activity" with p-value of 3.57e-10 and GO:0009883, "red or far-red light photoreceptor" with *p-value* of 8.43e-8. Also, GO terms identified as enriched are involve in red or far-red light signaling pathway such as GO:0010161, "red light signaling pathway" with *p-value* of 8.43e-8, and detection of light stimulus such as GO:0009854, "detection of visible light" with *p-value* of 4.10e-8.

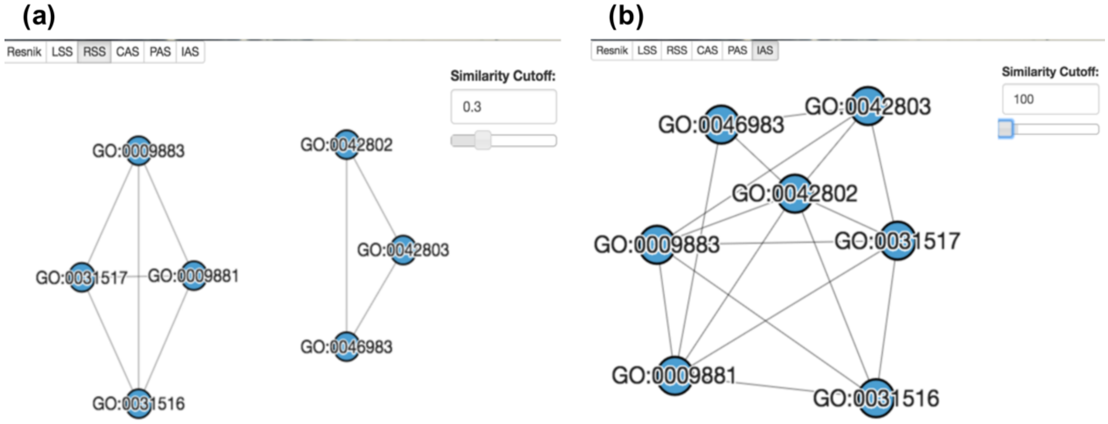
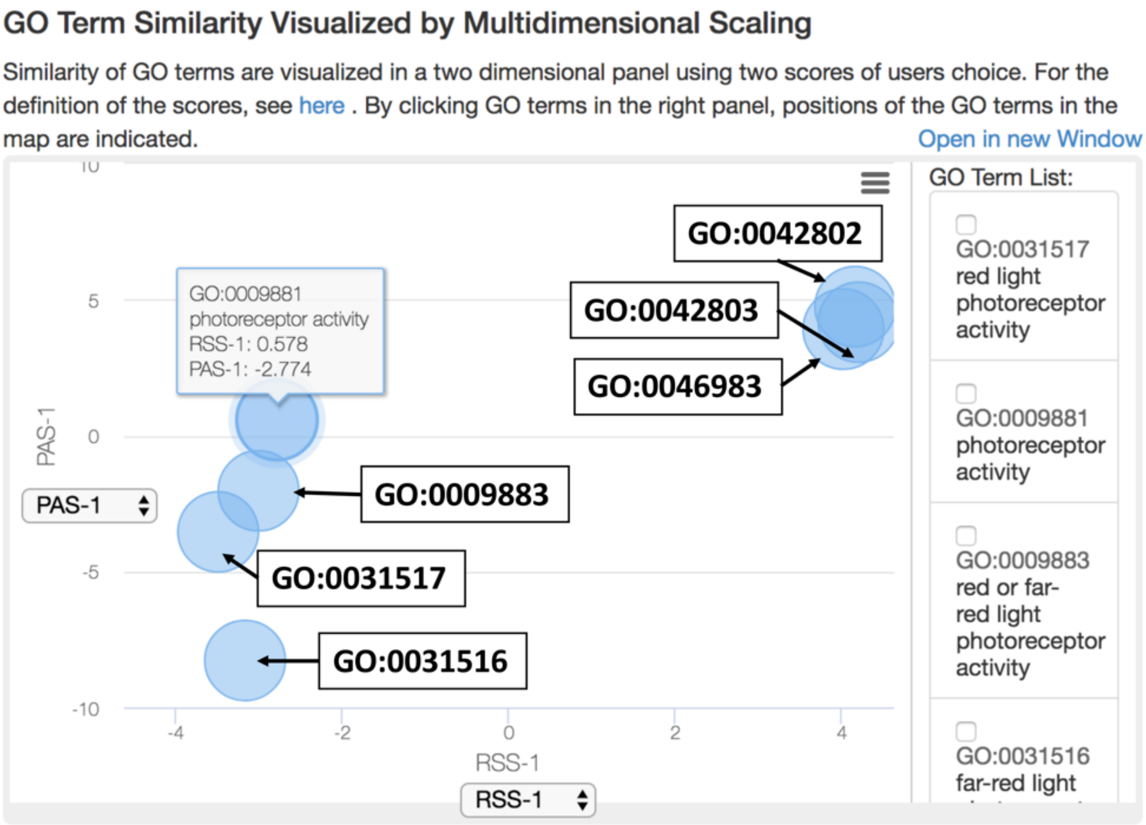


Figure 2.6 The network of functional association score of seven GO terms. The GO terms are GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803, and GO:0046983. (a) Network using RSS with a cut-off value of 0.3. (b) Network using IAS with a cut-off value of 100.



GO Term Similarity Visualized by Multidimensional Scaling

Similarity of GO terms are visualized in a two dimensional panel using two scores of users choice. For the definition of the scores, see here . By clicking GO terms in the right panel, positions of the GO terms in the map are indicated.

GO Term List:

☐ GO:0031517 red light photoreceptor activity

☐ GO:0009881 photoreceptor activity

☐ GO:0009883 red or far-red light photoreceptor activity

☐ GO:0031516 far-red light

Figure 2.7 The multidimensional scaling visualization of seven GO terms. The GO terms are GO:0031517, GO:0009881, GO:0009883, GO:0031516, GO:0042802, GO:0042803 and GO:0046983.

Table 2.1 List of the six example proteins and their associated GO terms.

Protein Name	Uniprot ID	CC	MF	BP
PHYA	P14712	GO:0005737, GO:0016604, GO:0016607, GO:0005634	GO:0031516, GO:0042802, GO:0003729, GO:0000155, GO:0042803, GO:0004672, GO:0009883	GO:0009584, GO:0009630, GO:0017148, GO:0009640, GO:0009638, GO:0018298, GO:0017006, GO:0010161, GO:0006355, GO:0046685, GO:0010201, GO:0010218, GO:0010203, GO:0006351
PHYB	P14713	GO:0005829, GO:0016604, GO:0016607, GO:0005634	GO:0031516, GO:0042802, GO:0000155, GO:1990841, GO:0042803, GO:0031517, GO:0009883, GO:0043565	GO:0009687, GO:0006325, GO:0010617, GO:0009584, GO:0009649, GO:0009630, GO:0009867, GO:0045892, GO:0009640, GO:0015979, GO:0009638, GO:0018298, GO:0017012, GO:0010161, GO:0031347, GO:2000028, GO:0010029, GO:0009409, GO:0010218, GO:0010244, GO:0010202, GO:0009266, GO:0010374, GO:0006351, GO:0010148
PRR7	A0A1P8BCB0	GO:0005634	NA	GO:0000160
PIF3	O80536	GO:0005634	GO:0003677, GO:0042802, GO:0046983, GO:0003700	GO:0009704, GO:0009740, GO:0031539, GO:0010017, GO:0009585, GO:0006355, GO:0009639, GO:0006351
PHYD	P42497	GO:0005634	GO:0042802, GO:0000155, GO:0009881, GO:0042803	GO:0018298, GO:0017006, GO:0009585, GO:0006355, GO:0006351
HDA15	Q8GXJ1	GO:0005634	GO:0046872, GO:0032041	GO:0006355, GO:0006351

2.3.4 Quantifying functional association of proteins

This function identifies protein pairs in a query protein set that have functional relevance. Using a GO pair score, functional relevance of a protein pair is evaluated by the funSim score [53], which is in essence the average GO pair scores of GO annotations of the two proteins. Eight different GO pair scores are used in NaviGO (Figure 2.10): "MF", "BP", and "CC" use RSS of the particular GO category, "BP+MF" is the funSim score using BP and MP, while "All" is the funSim using MF, BP, and CC. "PAS", "CAS", and "IAS" use the corresponding functional association scores to compute funSim.

For example, when studying whether proteins exist in the same cellular component, it would be interesting to check the CC funSim score. When studying whether proteins are involved in the same pathway or biological process, users would want to check "BP", "MF" or "BP+MF" columns.

The input data is a list of proteins and their GO annotations, the same as described in the GO enrichment analysis. In the results table (Figure 2.10), the significance levels of scores are shown in color-scale (red to pink for high to low). Since the significant cut-off is defined by the score distribution of a particular organism, there is a pull-down menu above the table to select the reference organism. In this example of six proteins, they all have the same RSS score of CC (Figure 2.10a), reflecting that all proteins are located in nucleus. PHYA (P14712) and PHYB (P14713) have a significantly high IAS of 1992.12, because they physically interact with each other [54].

Sometimes, it is difficult to see the functional association between proteins by looking at the score numbers in the output table. NaviGO provides a network visualization, which is available at "Open in new Window" (Figure 2.10b). In the network, proteins are connected if their association scores are above a cut-off value. In the example, only HDA15 is not connected in the association network with IAS cut-off value set to 100. This is consistent with the STRING database [55], where only HDA15 has low binding scores with all the other proteins in this network.

2.4 Function-Based PPI Prediction Methods

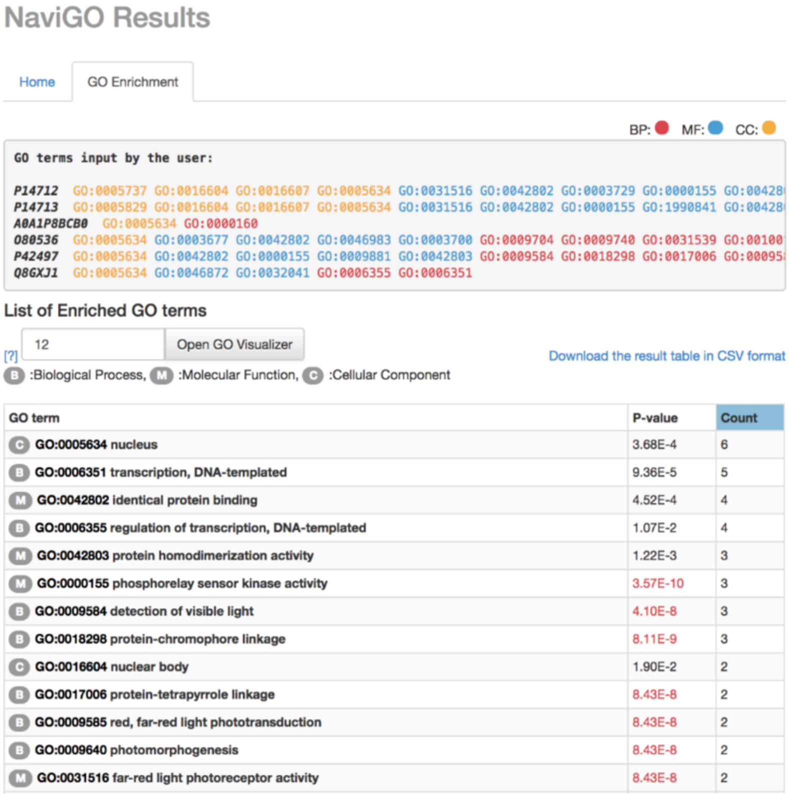
Because interacting proteins belong to the same pathway and share function, therefore, functional similarity of proteins can be a clue for predicting PPIs. Functional similarity of proteins are usually quantified by a similarity score of Gene Ontology (GO) terms [56] that annotate the proteins. Similarity of GO terms are defined by the closeness of the terms on the GO hierarchy tree and/or the frequency of the GO terms in gene annotations observed in an protein annotation database, e.g. UniProt [41, 49, 50] [57]. It was shown that including both common parental and children terms of GO terms, where common children terms are not used in the aforementioned scores that focus on quantifying similarity of parental terms in the GO hierarchy, improved PPI prediction accuracy [58]. Jain and Bader defined a GO similarity score by considering the distance to the leaf nodes in order to reduce the influence of imbalanced branch depths in the GO hierarchy [59].

GO term similarity (or relevance) can be also defined by counting frequency of co-occurrence of GO term pairs in biological contexts, in gene annotation or PubMed abstracts [52] or in known PPIs [31, 51].

Since PPI prediction is a suitable and handy application of GO term similarity scores, all the GO term scores above have been tested and compared for their performance of PPI predictions [51, 57, 59, 60]. Maetsche *et al.* showed that when using GO terms for PPI prediction in machine learning framework, induced GO term sets, e.g. common parental terms of annotated GO terms, performed better rather than using the original GO annotations of proteins [61].

2.5 Discussion

A web-based tool for analyzing GO terms and gene annotation was developed. Results are visualized by a user-friendly interactive panel, which provides intuitive understanding of gene function. A strength of NaviGO is that similarity or association of GO terms can be quantified in six different scores and it is equipped with real-time rendering of GO terms in the GO hierarchy. The unique feature of NaviGO should provide great convenience in functional analysis with GO for both bioinformatics researchers and biologists. Furthermore, NaviGO can also be used to compute the protein function similarity scores including the Relevance Semantic Similarity (RSS) score of 3 GO categories (MF, BP, CC), RSS of individual MF, BP or CC, RSS of BP and MR, IAS, PAS, and CAS. These functional similarity scores can be used as functional-based features to predict PPIs.



NaviGO Results

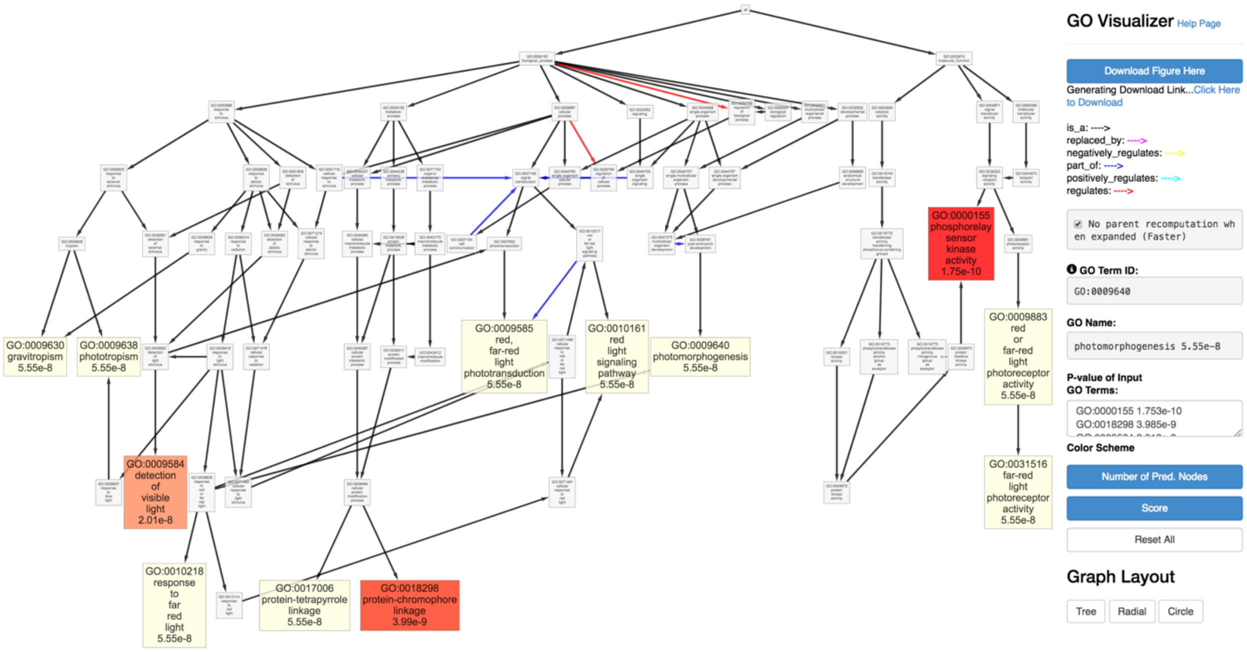
```
GO terms input by the user:

P14712   GO:0005737 GO:0016604 GO:0016607 GO:0005634 GO:0031516 GO:0042802 GO:0003729 GO:0000155 GO:00428
P14713   GO:0005829 GO:0016604 GO:0016607 GO:0005634 GO:0031516 GO:0042802 GO:0000155 GO:1990841 GO:00428
A0A1P8BCB0   GO:0005634 GO:0000160
O80536   GO:0005634 GO:0003677 GO:0042802 GO:0046983 GO:0003700 GO:0009704 GO:0009740 GO:0031539 GO:00100
P42497   GO:0005634 GO:0042802 GO:0000155 GO:0009881 GO:0042803 GO:0009584 GO:0018298 GO:0017006 GO:00095
Q8GXJ1   GO:0005634 GO:0046872 GO:0032041 GO:0006355 GO:0006351
```

List of Enriched GO terms

[?] 12 Open GO Visualizer
Download the result table in CSV format
B :Biological Process, M :Molecular Function, C :Cellular Component

GO term	P-value	Count
C **GO:0005634** nucleus	3.68E-4	6
B **GO:0006351** transcription, DNA-templated	9.36E-5	5
M **GO:0042802** identical protein binding	4.52E-4	4
B **GO:0006355** regulation of transcription, DNA-templated	1.07E-2	4
M **GO:0042803** protein homodimerization activity	1.22E-3	3
M **GO:0000155** phosphorelay sensor kinase activity	3.57E-10	3
B **GO:0009584** detection of visible light	4.10E-8	3
B **GO:0018298** protein-chromophore linkage	8.11E-9	3
C **GO:0016604** nuclear body	1.90E-2	2
B **GO:0017006** protein-tetrapyrrole linkage	8.43E-8	2
B **GO:0009585** red, far-red light phototransduction	8.43E-8	2
B **GO:0009640** photomorphogenesis	8.43E-8	2
M **GO:0031516** far-red light photoreceptor activity	8.43E-8	2

Figure 2.8 The GO enrichment analysis result of six proteins. The proteins are PHYA (P14712), PHYB (P14713), PRR7 (A0A1P8BCB0), PIF3 (O80536), PHYD (P42497), and HDA15 (Q8GXJ1).



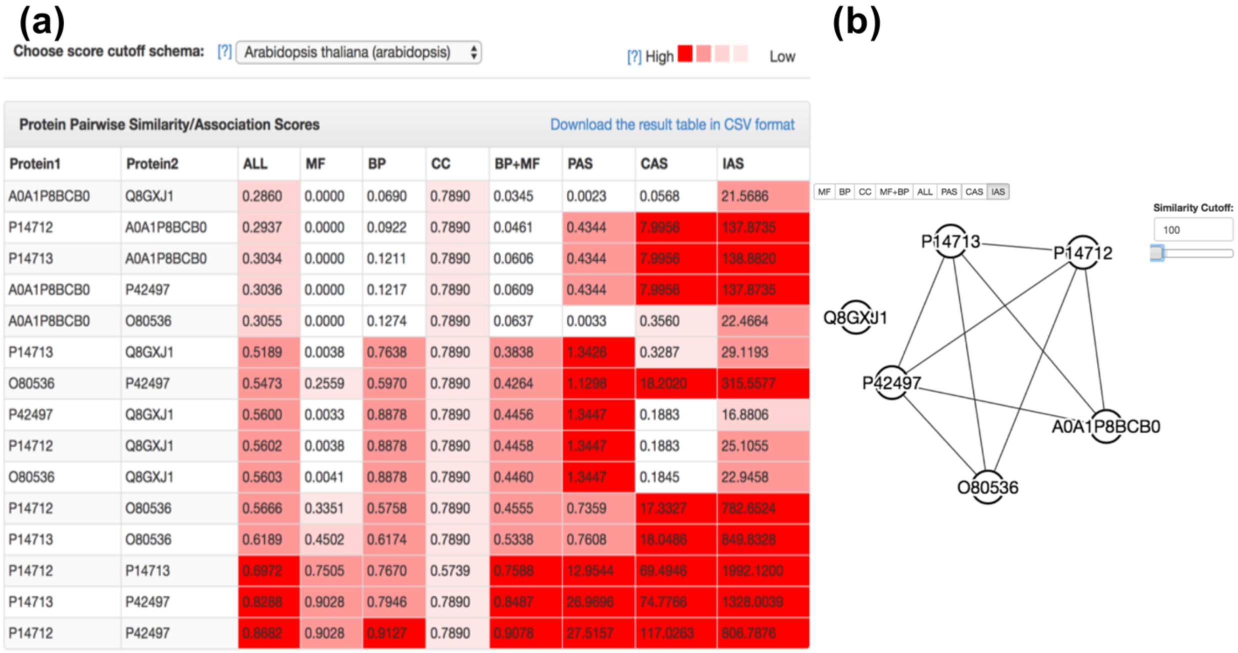
☑ No parent recomputation when expanded (Faster)

GO Term ID: GO:0009640

GO Name: photomorphogenesis 5.55e-8

P-value of Input GO Terms:
GO:0000155 1.753e-10
GO:0018298 3.985e-9

Figure 2.9 Visualization of twelve significantly enriched GO terms.



Choose score cutoff schema: [?] Arabidopsis thaliana (arabidopsis)

Protein Pairwise Similarity/Association Scores Download the result table in CSV format

Protein1	Protein2	ALL	MF	BP	CC	BP+MF	PAS	CAS	IAS
A0A1P8BCB0	Q8GXJ1	0.2860	0.0000	0.0690	0.7890	0.0345	0.0023	0.0568	21.5686
P14712	A0A1P8BCB0	0.2937	0.0000	0.0922	0.7890	0.0461	0.4344	7.9956	137.8735
P14713	A0A1P8BCB0	0.3034	0.0000	0.1211	0.7890	0.0606	0.4344	7.9956	138.8820
A0A1P8BCB0	P42497	0.3036	0.0000	0.1217	0.7890	0.0609	0.4344	7.9956	137.8735
A0A1P8BCB0	O80536	0.3055	0.0000	0.1274	0.7890	0.0637	0.0033	0.3560	22.4664
P14713	Q8GXJ1	0.5189	0.0038	0.7638	0.7890	0.3838	1.3426	0.3287	29.1193
O80536	P42497	0.5473	0.2559	0.5970	0.7890	0.4264	1.1298	18.2020	315.5577
P42497	Q8GXJ1	0.5600	0.0033	0.8878	0.7890	0.4456	1.3447	0.1883	16.8806
P14712	Q8GXJ1	0.5602	0.0038	0.8878	0.7890	0.4458	1.3447	0.1883	25.1055
O80536	Q8GXJ1	0.5603	0.0041	0.8878	0.7890	0.4460	1.3447	0.1845	22.9458
P14712	O80536	0.5666	0.3351	0.5758	0.7890	0.4555	0.7359	17.3327	782.6524
P14713	O80536	0.6189	0.4502	0.6174	0.7890	0.5338	0.7608	18.0486	849.8328
P14712	P14713	0.6972	0.7505	0.7670	0.5739	0.7588	12.9544	69.4946	1992.1200
P14713	P42497	0.8288	0.9028	0.7946	0.7890	0.8487	26.9696	74.7766	1328.0039
P14712	P42497	0.8682	0.9028	0.9127	0.7890	0.9078	27.5157	117.0263	806.7876

Figure 2.10 Example of Protein Set analysis. (a) Results of Ppairwise protein association scores. (b) The protein association network of six proteins PHYA (P14712), PHYB (P14713), PRR7 (A0A1P8BCB0), PIF3 (O80536), PHYD (P42497), and HDA15 (Q8GXJ1) with a cut-off value of 100.

CHAPTER 3. COMPUTATIONAL METHOD TO IDENTIFY PROTEIN- PROTEIN INTERACTIONS[2]

3.1 Background

To complement experimental methods for identifying PPIs, several computational methods have been developed [62]. These methods typically use a machine learning framework and consider various features of proteins as input. Protein features used for PPI prediction include occurrence of functional domains [63-65], short sequence patterns (e.g. n-grams, auto-covariation) [66-70], interlog (interaction inferred from homology) [71-75], codon usage [76], function [43, 77], similarity in phylogenetic trees [78-80], phylogenetic profiles [55], gene expression [81], and protein docking prediction [82-85]. Although many approaches were explored, there are not many works that applied developed methods to provide new proteomics-scale PPI predictions.

PPI prediction methods were classified into six large categories based on features of proteins considered as input information of the prediction. Protein sequence-based, comparative genomics-based, structure-based, PPI network topology-based methods, and methods using integrated features have been employed. Below ideas behind methods that fall into each category are discussed.

To develop a computational prediction method, one needs a dataset of known interacting protein pairs (a positive set) and a dataset of non-interacting protein pairs (a negative set), because the method needs to maximize its ability to distinguish between positive and negative datasets. A positive dataset is constructed from known PPIs stored in existing PPI databases (Table 3.1). On the other hand, constructing a negative dataset is not straightforward, because there are few collections protein pairs that are experimentally directly verified not to interact. To facilitate construction of a negative dataset, there is a database named Negatome, which collects protein pairs that are unlikely to interact by manual curation of literature and by analyzing protein complexes from the PDB [86]. Other commonly used strategies to construct a negative dataset is to pair proteins from different cellular locations or randomly pair proteins that appeared in the positive dataset excluding interacting pairs.

---

[2] This chapter have been previously published 62. Ding, Z. and D. Kihara, *Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features.* Current Protocols in Protein Science, 2018: p. e62.

Table 3.1 List of available protein-protein interaction databases.

Database	# of interactions	Description	Organisms	Website	Last update
BioGrid	1,110,310	Manually curated PPIs	62	https://thebiogrid.org/	Mar 2017
STRING	932,553,897	Protein associations including PPIs	2,031	http://string-db.org/	Jan 2017
DIP	81,731	Experimentally identified PPIs	834	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	Mar 2017
CORUM	6,375	Manually curated protein complexes in mammals	10	http://mips.helmholtz-muenchen.de/corum/	Dec 2016
IntAct	718,180	PPIs taken from literature and from user submissions	Model organisms including human, mouse, yeast, fruitfly, *C. elegans, E. coli, A. thaliana*	http://www.ebi.ac.uk/intact/	Mar 2017
MINT	125,464	Experimentally verified PPIs from literature	611	http://mint.bio.uniroma2.it/	Mar 2017
InnateDB	367,478	Manually curated PPIs for mammalian innate immune response	Human, mouse, *B. taurus*	http://www.innatedb.com/	Nov 2016
HPRD	41,327	PPI network of *H. sapiens*	human	http://www.hprd.org/	April 2010
EcoCyc	6,399	Manually curated PPIs in *E. coli* K-12 MG1655	*E. coli*	https://ecocyc.org/	Dec 2016
TAIR	8,826	Experimentally identified PPIs in *A. thaliana*	*A. thaliana*	https://www.arabidopsis.org/	Sep 2011

Note: References of databases: BioGrid: [29]; STRING: [87]; DIP: [88]; CORUM: [89]; IntAct: [90]; MINT: [91]; InnateDB: [92]; HPRD: [93]; EcoCyc: [94]; TAIR: [95].

3.2 PPI Prediction Methods

3.2.1 Sequence-Based Methods

Many methods have been developed that use the amino acid sequence information of target proteins. The obvious advantage of using sequence information is that it is available for all proteins in an organism as long as its genome sequence is available.

3.2.1.1 Motif/Domain-based approach

The most straightforward approach in this category is to predict that two proteins interact with each other if they possess known sequence patterns of interacting proteins in their amino acid sequences. For example, Becerra *et al.* predicted PPIs between human immunodeficiency virus 1 (HIV-1) and human cells by detecting sequence motifs of protein interacting regions that have disordered structures [96]. Sequence patterns of known functional regions including PPI sites, which are called motifs or domains depending on the sequence length, are stored in public databases, such as ELM [97], InterPro [98], PROSITE [99], PRINTS [100], Pfam [101], and ProDom [102, 103].

Instead of detecting specific motifs that are known as protein interaction sites, Sprinzak and Margalit computed the log-odds score of observing two motifs from the InterPro database in known interacting yeast protein pairs [104]. The log-odds value was computed as $\log_2(P_{ij}/P_iP_j)$, where P_{ij} is the observed frequency of motif pair *(i, j)* observed in interacting proteins, and P_i and P_j are the frequencies of motif i and j in the data, respectively. If a query protein pair contains at least one of motif pairs that have a log-odds value above a threshold, they are predicted as interacting. Later, essentially the same approach was taken to count motif pairs in interacting proteins in the DIP database [105]. Above methods consider only single motif pair from each protein pair. Chen and Liu extended the methods by considering contributions of all the possible pairs of 4293 Pfam domain combinations [106]. Each protein pair was represented with 4293 dimensional vectors with 0 indicating absence of a domain in neither of the proteins, 1 indicating one of the proteins contains the domain, and 2 indicating presence of the domain in the both proteins. Then protein pairs are predicted to interact or not to interact by classifying its feature vector using a machine learning method, random forest, which makes a prediction by voting from many decision trees.

Pitre *et al.* considered sequence similarity rather than detecting exact sequence patterns of interacting proteins [107]. The algorithm called Protein-Protein Interaction Prediction Engine (PIPE) they developed, considers the co-occurrence of all short subsequences. In this method, the query protein sequences A and B are fragmented into a_i and b_j using 20 amino acid-long sliding window. Then the fragment a_i is compared with fragments of proteins in a known PPI network using the PAM120 amino acid similarity matrix. Once matched fragment of known proteins similar to a_i are found, the known interacting partners to the matched proteins are compared with fragment b_j using the PAM120 matrix. Finally, two proteins A and B are predicted to interact if frequency of matched fragment pairs from known PPIs is above a threshold (set to 10). Another similar method called D-MIST adopted position-specific scoring matrix (PSSM) to evaluate the similarity of motifs in a query protein pair to binding motifs in known PPIs with solved tertiary structures [108].

3.2.1.2 Methods that capture sequence features

The motif/domain-based methods described in the previous section examine occurrence of known functional sequence motifs/domains in databases or in known interacting proteins. Sequence-based approaches can be extended to consider any sequence patterns including patterns that are not necessarily known to be involved in PPIs or in any function by simply extracting short sequences of a fixed length systematically from query protein sequences. A typical method in this category segments an amino acid sequence of a target protein into overlapping fragments (n-gram) by applying a small sliding window of a certain length (*n*), and to consider counts of sequence patterns of fragments as a feature vector of the protein (Figure 3.1). Then, a machine learning method is trained on a dataset of feature vectors of known interacting proteins and non-interacting protein pairs so that the method distinguishes between the two datasets [109, 110]. Instead of raw counts of sequence patterns, statistical significance of the counts relative to the background frequency of amino acids was also used [111]. Another variant of the n-gram approach was to consider sequence patterns that skip a certain number of sequence positions [112]. Martin *et al.* used a so-called signature molecular descriptor, which considers the frequency of adjacent (*i.e.* preceding and following) amino acids for each amino acid, which essentially captures sequence patterns of 3-grams [113]. Ding *et al.* considered both multivariate mutual information of 3-gram and mutual information of 2-gram, *i.e.*

$$I(a,b,c) = I(a,b) - I(a,b|c), \quad \text{(Equation 3.1)}$$

where $I(a,b,c)$ is the multivariate mutual information of 3-gram, $I(a,b)$ is the mutual information of 2-gram, a, b, c are amino acid classes, and $I(a,b|c)$ denotes the conditional mutual information of a and b given that c exists in the 3-gram [68]. Wong *et al.* considered amino acid pairs in a protein sequence (every pairs; including non-adjacent pairs) and represented it as an n*n matrix (n: the length of the protein) where each element is the sum of hydrophobicity value of every combination of two amino acids in the sequence [114]. An used PSSM to represent a protein sequence, which considers similarity of 19 other amino acids at each position of a sequence [115]. Using PSSM, 2-gram was represented as a 400-dimensional vector (=20*20), which was subject to the dimension reduction to 350 vectors.

The number of sequence combinations of *n*-grams will be quite large, for example, there are 20*20*20 = 8000 combinations for 3-grams for protein sequences which consist of 20 different amino acids. A large number of combinations will generate unnecessarily long feature vectors for proteins and will causes a data sparseness problem when some sequence patterns are not well sampled. Therefore, for computing *n*-grams, it is common to reduce the number of letters in sequences by clustering amino acids into a smaller number of groups. Shen *et al.* classified amino acids to seven classes considering their polarity and volume [110] and several later papers used the classification.

Besides using *n*-grams and its variants, there are several other ideas for capturing sequence patterns that were used for PPI prediction. To capture general characteristics of a protein sequence, three features, namely, the composition of amino acids, transition probabilities between two consecutive amino acids, and distribution describing the position of sequences from the N-terminus that contains the first, first 25%, 50%, 75%, and 100% of each amino acid (class) over the sequence were used [116]. The combination of these three sequence features is called the local descriptor [117] [118] [119] [120] (Figure 3.2).

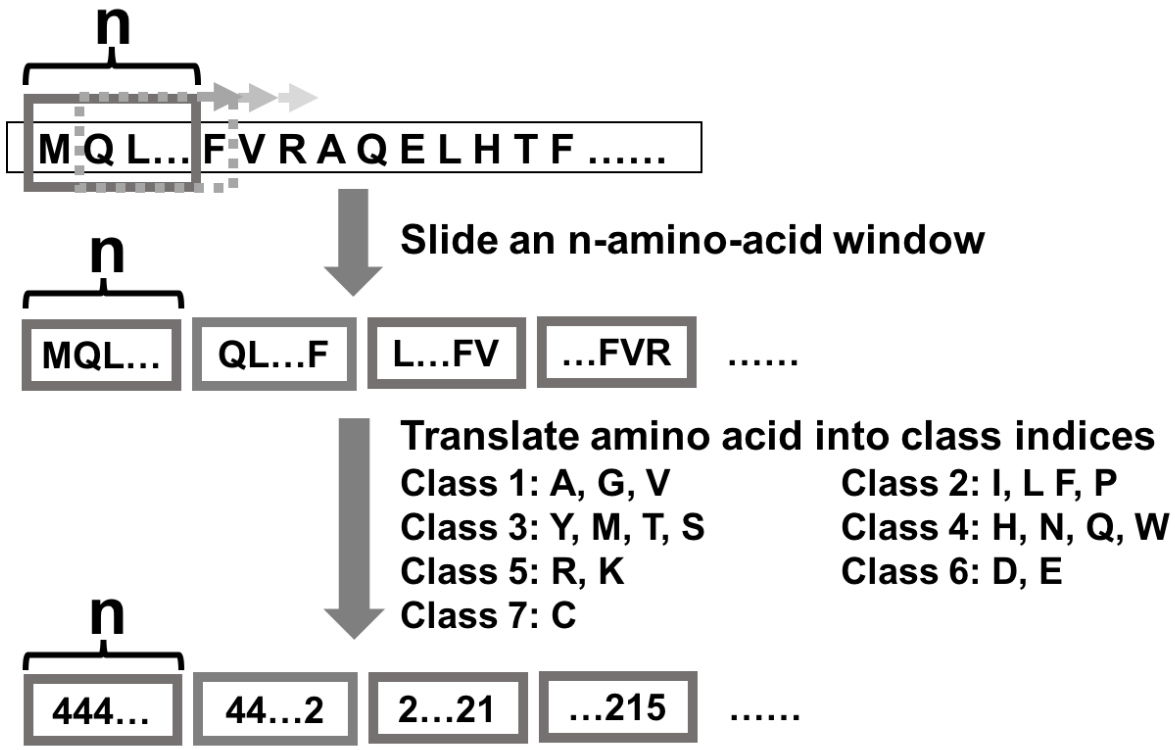


Figure 3.1 The n-gram features for a protein sequence. The 20 amino acids are clustered into seven classes based on their physicochemical properties. A window of length n (e.g. n=3) is sliding along the sequence and captures amino acid class patterns in the window. Then the occurrences of every combination of amino acid class are counted to generate a feature vector for the sequence. For example, when n equals 3, the total number of combinations of amino acid class is 7*7*7=343.

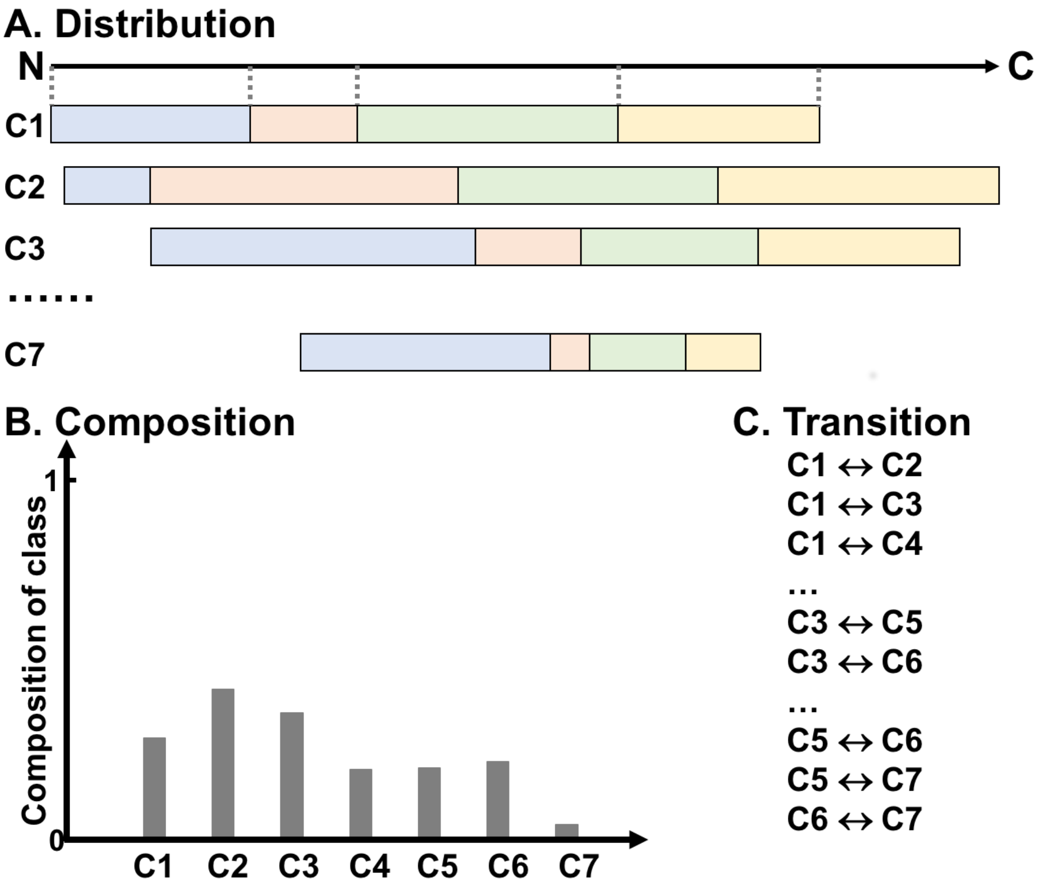


Figure 3.2 The local descriptors. Amino acids are clustered into seven classes (C1-C7). (A) The distribution of the lengths of sequences from the N-terminus that contain the first, first 25%, 50%, 75%, and 100% of each amino acid class in the protein sequence are represented in blue, pink, green, and yellow, respectively. The dotted line represents the position of the first, first 25%, 50%, 75%, and 100% of Class 1 in the local region. The number of distribution descriptor is 7 (classes) *5 (distribution values) =35 for a local region. (B) The composition of each amino acid class in a local region is considered. (C) The transition accounts for the frequency of the transition from one class to another. The number of transition descriptor is (7*6)/2=21. Therefore, each local region is represented by 35+7+21=63 descriptors.

Guo *et al.* used a feature called auto covariance (AC) for represent protein sequences [121]. AC is intended to capture the periodicity of physicochemical properties along a protein sequence (Figure 3.3). To compute AC of a protein sequence for a physicochemical property, amino acids are assigned with a property values, *e.g.* hydrophobicity, hydrophilicity, side-chain volume, polarity, solvent-accessible surface area, or the net charge index of side chain. Then, AC is defined as follows:

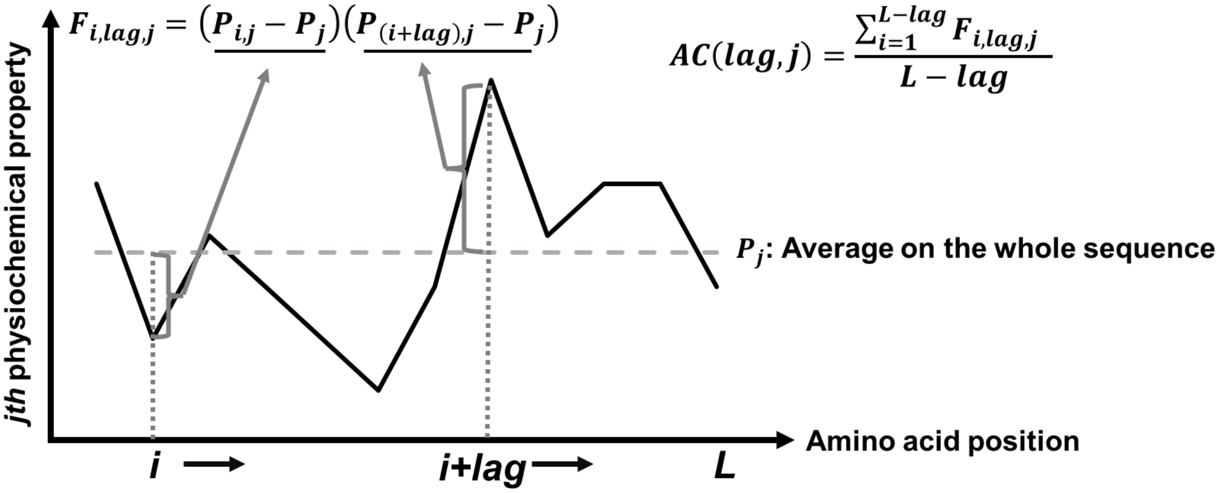
$$AC(lag,j) = \frac{\sum_{i=1}^{L-lag}\left(P_{i,j}-\frac{1}{L}\sum_{i=1}^{L}P_{i,j}\right)\times\left(P_{(i+lag),j}-\frac{1}{L}\sum_{i=1}^{L}P_{ij}\right)}{L-lag} \qquad \text{(Equation 3.2)}$$

where *lag* is the distance between covariant residues to consider, which ranges from 1 to 30, j is the *j-th* physiochemical descriptor, *i* is the position in the sequence, and *L* is the length of sequence. Thus, AC of a property with a certain *lag* length will be large if amino acids with a large (or small) property value appear periodically with an interval of *lag*. There is a similar value called Moran auto correlation (MAC), which is defined as

$$M_{AC}(d) = \frac{1}{N-d}\sum_{j=1}^{N-d}(P_j - \overline{P}) \times (P_{j+d} - \overline{P}) \,/\, \frac{1}{N}\sum_{j=1}^{N}\left(P_j - \overline{P}\right)^2 \qquad \text{(Equation 3.3)}$$

where *d* is the distance between covariant residues which ranges from 1 to 30, P_j *and* P_{j+d} are the physiochemical property of *j-th* and *(j+d)-th* amino acid, respectively, *N* is the length of the protein sequence, $\overline{P} = \frac{\sum_{j=1}^{N} P_j}{N}$ is the average value of the physiochemical property [120]. Thus, MAC is AC divided by variance of the physiochemical property, $\frac{1}{N}\sum_{j=1}^{N}\left(P_j - \overline{P}\right)^2$.

Intention behind computing the local descriptor, MC, and MAC is to capture global, long range sequence features of proteins, in contrast to *n*-gram and its variants that capture local patterns of sequences. As these features are complementary to each other, often both types were combined [68]. For example, in the method by You *et al.*, there were four components in the protein sequence feature representation [120]: 1) 3-grams. Amino acids were classified to seven classes and the frequency of 3-grams was considered as a feature of a protein. Thus, a protein pair is represented by a vector of 686 (= 2*7*7*7) features. 2) AC. Six physicochemical properties of amino acids were considered, which were hydrophobicity, side-chain volume, polarity, polarizability, solvent-accessible surface area, and the net charge of side chains. For each of the properties, AC was computed using 1 to 30 *lag* values following Eq. 2. Thus, the length of the vector for a protein pair was 360 (= 2*6*30). 3) MAC. Similar to AC, a 360-dimension vector was constructed for a protein pair. 4) Local descriptors. Amino acids were classified to seven classes and the local descriptor, the composition, the transition, and the distribution, were computed for each of the seven amino acid classes for ten local regions in a protein. Thus, a pair of proteins was represented by a vector of 1260 = 2* 10* (7 compositions + 21 transitions + 35 distributions) values. Overall, considering the all four features, a protein pair was represented by a vector of 2666 (= 686 + 360 + 360 + 1260) features.



$F_{i,lag,j} = (P_{i,j} - P_j)(P_{(i+lag),j} - P_j)$
$AC(lag,j) = \frac{\sum_{i=1}^{L-lag} F_{i,lag,j}}{L - lag}$
jth physiochemical property
P_j: Average on the whole sequence
Amino acid position
i
i+lag
L

Figure 3.3 Schematic view of calculating Auto-covariance (AC). The black line in the plot represents the value of the *j-th* physiochemical property along the amino acid sequence. The dashed grey line represents the average value of the *j-th* physiochemical property. The grey bracket regions are the difference from the average value of *i-th* and *(i+lag)-th* amino acid, respectively. AC is the average of $F_{i,lag,\,j}$.

With these sequence features prediction of PPIs was made using various machine learning algorithms. Algorithms used include support vector machine (SVM) [110, 113, 119, 121-123], relevance vector machine [115], random forest [68, 106, 117], rotation forest [114], linear discriminant classifier and cloud points [109], relaxed variable kernel density estimator (RVKDE) [111], an ensemble classifier [112], extreme learning machine (ELM) [120], and k-nearest neighbors (KNNs) [118].

These sequence-based methods reported surprisingly high accuracies. For example, Shen *et al.,* reported 83.90% accuracy on the HPRD dataset [93, 110]. Yang *et al.* reported 86.15%

accuracy on a yeast dataset [118]. Yu *et al.* achieved 93.7% accuracy on a highly unbalanced HPRD dataset where positive-to-negative ratio is 1:15 [111]. Zhu *et al.* reported over a 75% accuracy on five organisms including yeast, *C. elegans, E. coli,* human, and mouse [119]. Wong *et al.* achieved 93.92% on the *S. cerevisiae* dataset [114]. You *et al.* achieved 93.46% to 97.01% accuracy on six different organisms including yeast, *H. pylori, C. elegans, E. coli,* human, and mouse [117]. Ding *et al.* achieved 95.01% on the yeast dataset and 87.59% on the *H. pylori* dataset [68]. An *et al.* achieved 94.57% and 90.57% on the *S. cerevisiae* and the *H. pylori* dataset, respectively, also 97.15% accuracy on an imbalanced yeast dataset [115]. Wei showed over 81% accuracy using different features on the Negatome and the DIP dataset [86, 88, 112]. Although the reported accuracy values are high and encouraging, it needs to be noted that the datasets on which the methods are tested are limited to several organisms.

3.2.1.3 Using homology

So far the methods that use partial sequence patterns and statistical features in protein sequences were reviewed. In this section methods that use similarity of entire protein sequences are introduced. Many functionally important proteins in an organism are conserved across species, which is the rationale of sequence similarity search for annotating function of genes [30, 124, 125]. Several databases, such as KEGG Orthology [126], OrthoDB [127], OrthoMCL-DB [128], HomoloGene [129], and INPARANOID [130], contain lists of precomputed homologous genes in different species. As interactions with other proteins is a part of a protein's important function, it is known that PPIs are often conserved across species. These conserved interactions are noted as "interlogs" [131]. Matthew *et al.* mapped PPIs in the yeast interaction map to predict PPIs in *C. elegans*, and identified 257 potential interlogs [132]. Further experimental validation performed on 72 predicted interactions gave 19 positive results, which were roughly 25% among tested. The POINT web service provides human PPIs inferred from interlogs with mouse, fruitfly, yeast, and *C. elegans* [133]. Taking advantage of an increasing number of experimentally identified protein interactions, Lee *et al.* then expanded orthologous pairs to consider to those from 18 eukaryotic species [134]. The idea of interlogs were also applied to predict PPIs in plants, to *A. thaliana* by considering homologs with yeast, fruitfly, human, and *C. elegans* [135] [136] and to rice (*Oryza sativa*) by considering interlogs with the six species including the same four species with *E. coli* and *A. thaliana* [137]. Dutkowski *et al.* developed a statistical model, which represents specification and duplication events of genes along an evolutionary tree, on which known interacting protein pairs in seven eukaryotic organisms were mapped and used for predicting PPIs [138]. Interactome3D is a database that provides the tertiary structure models of protein complexes built based on known structure information of interlogs [139]. Wang *et al.* merged prediction results from an interlog-based method and a motif-based method to cover a larger number of predicted PPIs in the pig proteome [140].

3.2.1.4 Codon usage

Interestingly, it was shown that the codon usage of genes can be used to predict PPIs. Using difference of codon usage of protein pairs, Najafabadi *et al*. predicted PPIs in *E. coli,* yeast, and *Plasmodium falciparum* with reasonably good accuracy [141]. For a pair of genes i and j the

difference of usage of codon c among 64 codons is simply defined as

$$d_{ij}(c) = |f_i(c) - f_j(c)| \quad \text{(Equation 3.4)}$$

where $f_i(c)$ and $f_i(c)$ are the usage frequency of codon c of gene i and j repectively. Then, the difference of each codon usage (d_{ij}) is binned into 50 intervals, and the likelihood ratio of the fraction in interacting and non-interacting proteins in a training dataset was computed. A PPI prediction for a protein pair is performed with a naïve Bayes approach using the likelihood ratio. Zhou *et al*. used SVM with the codon usage difference and applied to the yeast genome [142]. One might wonder why codon usage is related to PPIs. But it is reasonable considering that codon usage is known to be correlated with gene expression levels [143] and also that neighboring genes have similar codon usage. As discussed later in this review, both gene expression level and conserved neighboring genes (gene order) have been successfully used to predict PPIs.

3.2.2 Comparative Genomics-Based Methods

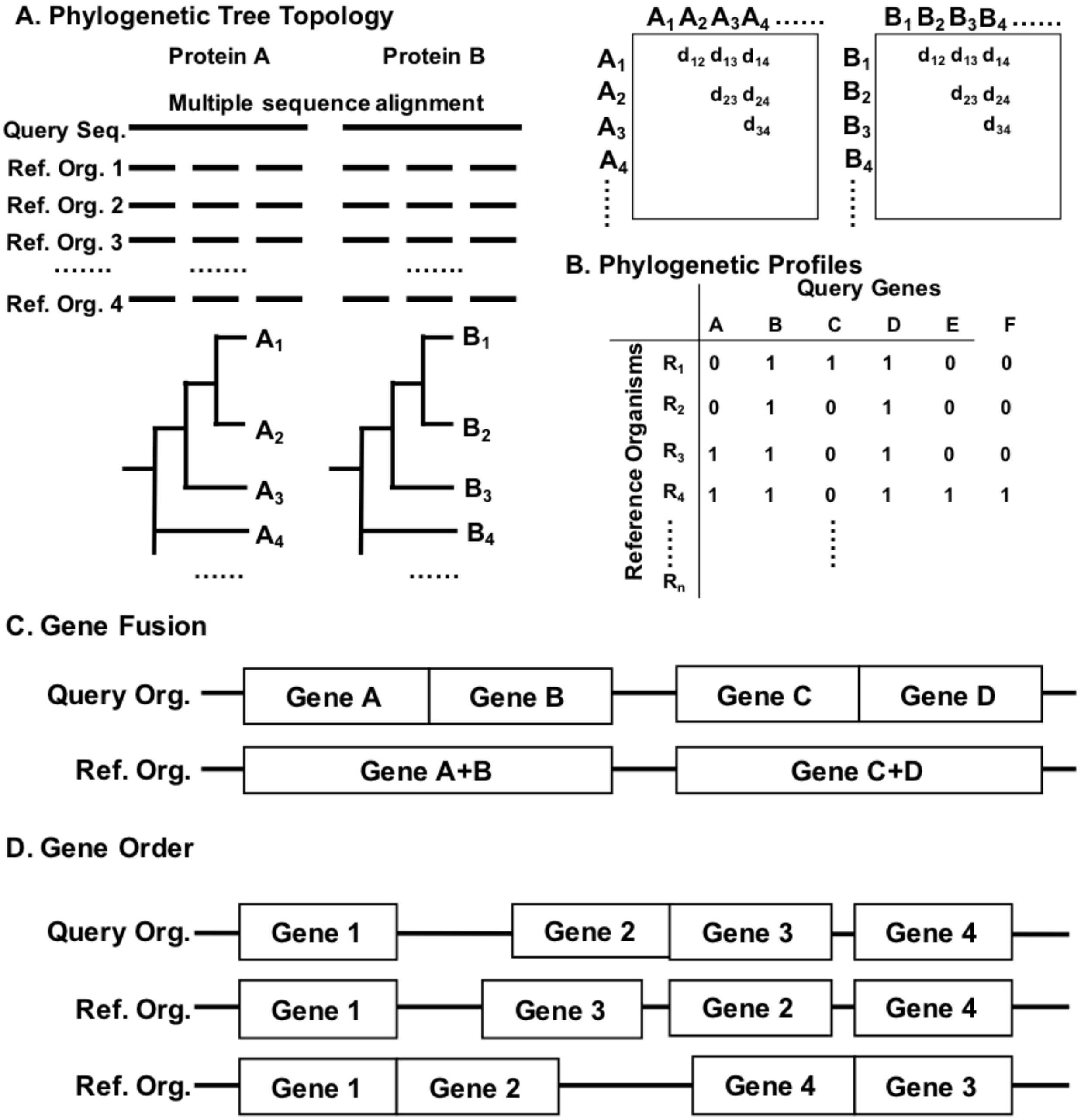
The last level of sequence information that can be used for PPI prediction is from genome sequences from various species. Since important features in a genome sequence including gene sequences, intergenic sequences, gene orders, are conserved during evolution, identifying such conserved features in genomes can be a clue for to identify proteins that are functionally related, which often also involve physical interactions between the proteins. Under this category, which we call the comparative genomics-based methods, the phylogenetic tree topology analysis, the phylogenetic profile, considering gene fusion events, and conserved gene orders (Figure 3.4). An important point to note is that these methods are not aiming for predicting physical PPIs directly but for identifying functionally related proteins, but quite often functionally related proteins do physically interact with each other. A strong advantage of the comparative genomics-based approaches is that, due to the increasing number of determined genome sequences, many proteins can now find related (and maybe interacting) proteins through these approaches [144].

3.2.2.1 Phylogenetic tree topology analysis

It has been observed that the phylogenetic trees between interacting proteins were more similar than a general divergence between the corresponding species [145-147]. The similarity between the phylogenetic trees of interacting proteins was explained as maintenance of the complex functionality and suffering similar evolutionary pressure. The sequence signal of the coevolution is strong at binding interface of proteins, but also come from other regions of proteins [148].

The tree topology similarity can be measured as the correlation between the evolutionary distance matrices used to build the trees. The algorithm to calculate similarity of distance matrices is called the mirror tree method [147]. It contains following steps (Figure 3.4 A): 1) To construct a multiple sequence alignment for each protein against a list of reference organisms; 2) To construct a phylogenetic tree for the proteins; 3) Then, for a pair of proteins in question, distances against orthologous proteins in different species are computed (distance matrices) and the correlation coefficient between two distance matrices is obtained. A protein pair is predicted to be interacting if the coefficient value is above a cut-off value, which is determined by known interacting and non-interacting proteins.

The mirror tree method was modified for improvement in several different ways. The method was extended to handle interacting protein families, such as a ligand family and a receptor family, to be able identify interacting specific protein pairs from the two families [149]. Sato *et al.* removed a background tree similarity that arises by the overall evolutionary distance of organisms from distance matrices of individual proteins, which made improvement of PPI prediction accuracy [150]. They further considered partial correlation of distance matrices that can more effectively remove background organism-level similarity from the tree similarity of a query protein pair, where the background organism-level similarity was represented by a linear combination of distance matrices of many proteins in the organisms [151]. Besides the background similarity of organisms, another source of noise in the mirror tree method is that a protein coevolves with multiple interacting proteins. Instead of evaluating tree similarity of a query protein pair, Juan *et al.* considered a network of similarities between all pairs of proteins simultaneously [152]. In the mirror tree method, selection of reference genomes is a key for successful prediction. Effective ways to select organisms for building trees were examined by Herman *et al.* [153]. Instead of using correlation coefficient, SVM was also used to make predictions from distance matrices [154].



	A	B	C	D	E	F
R_1	0	1	1	1	0	0
R_2	0	1	0	1	0	0
R_3	1	1	0	1	0	0
R_4	1	1	0	1	1	1
⋮			⋮			
R_n						

Figure 3.4 Comparative genomics-based methods. (A) Phylogenetic tree topology-based method. A pair of query sequences, A and B, are compared with *n* reference organisms and multiple sequence alignments are constructed. Based on the alignments, phylogenetic trees are computed, from which distances between all pairs of sequences are computed and stored in matrices. Then the similarity between two matrices is evaluated with a correlation coefficient. (B) Phylogenetic profiles. Genes in a query organism is compared with *n* reference organisms by BLAST search. 1 represents the presence of the gene in the reference organism 0 for the absence. The table can also be filled with real-values, such as BLAST bit scores. (C) Gene fusion. Two protein genes A and B are predicted as interacting if they fuse to form one protein gene AB in another organism. (D) Gene order. If protein gene orders of proteins are conserved among different species, they are predicted as interacting proteins with each other.

3.2.2.2 Phylogenetic profiles

Phylogenetically related and thus possibly interacting protein pairs can be identified in a simpler way of using comparative genomics. In the approach called the phylogenetic profiling co-presence and co-absence of orthologous proteins across organisms are examined rather than comparting phylogenetic trees of protein pairs as discussed in the previous section [155]. If PPIs are needed for realizing a certain biological function, an organism needs to possess both proteins if the function is required while both are not needed if it does not need the function. Coding one of the proteins only in its genome is meaningless.

There are three major steps to perform this method (Figure 3.4 B). The first step is to identify orthologous proteins for all the proteins in a query genome against other reference genomes by sequence similarity search. Then, construct a phylogenetic profile for each protein in the query genome, which has binary values with 1 indicating the presence of an orthologous gene and 0 for the absence of the ortholog in a reference genome. Thus, the dimension of the profile is the number of reference genomes used. Finally, protein pairs that have similar profiles are predicted to be interacting (more precisely, functionally related). Similar to the phylogenetic tree topology methods, the choice of reference genome is crucial for this approach [156]. Also, a threshold value (E-value) in sequence similarity search for detecting orthologous proteins strongly affects to profiles, and thus to the prediction performance of the method [157]. To accommodate the strong dependency of the performance to a threshold value in the similarity search, real value vectors of an alignment score was used for constructing profiles rather than binary values [158]. In the method by de Vienne and Azé a combination of the phylogenetic tree topology and profile was used as features in a machine learning framework [159].

3.2.2.3 Gene fusion events

A gene fusion refers to an event in the comparative genomics where two individual genes in one organism fuse as a continuous sequence in another organism [160] (Figure 3.4 C). Fused genes are usually functionally related and further implies physical interactions between the proteins [161-163]. Computationally, fused genes can be found by gene sequence similarity search between genomes. It was reported that metabolic enzymes are frequently involved in gene fusions [164].

3.2.2.4 Conserved gene orders

Through evolution genomes undergo various rearrangements and transfers, therefore locations of genes in a genome tend to be shuffled unless an evolutionary pressure keeps the order of some genes together [165] (Figure 3.4D). Thus, conservation of gene orders, i.e. common local clusters of genes in genomes, indicates that there is a requirement or an advantage to keep the gene order for the organisms, and in fact many cases genes in a conserved cluster are involved in the same function [166]. In bacterial and archaeal genomes, operon structures are conserved across many species, which code genes in the same pathways or complexes [167]. After initial findings of the conserved gene orders, more systematic studies have been done [168, 169]. Similar to the other comparative genomics-based methods, a key for successful application of this analysis is to choose an appropriate set of reference genomes, which should not be too evolutionary distant but not too close to each other, so that only clusters of functionally related genes can are conserved. A related work was done by Kihara & Kanehisa where transmembrane protein complexes were predicted from genomes by identifying gene clusters that have predicted transmembrane domains [170].

3.2.3 Gene Co-Expression-Based Methods

Gene co-expression data such as microarray and RNA-sequencing data are valuable experimental data that can be used to infer PPIs. Intuitively, interacting protein pairs are expected to have similar gene expression levels under variety conditions. Indeed significant correlation between the gene co-expression level and PPIs was shown in bacteriophage T7 [171], yeast [172, 173], human, mouse, and *E. coli* [174]. Fraser *et al.* showed that gene expression level of interacting proteins co-evolve using four closely related yeast species, where the expression level was estimated by the codon usage [175]. Databases that provides large-scale gene co-expression information includes GEO [176], ATTED-II [177], and COXPRESdb [178]. ATTED-II and COXPRESdb are pre-calculated gene co-expression databases of plant organisms and animal species, respectively.

Although gene expression is shown to have significant correlation to PPIs, a major challenge is that co-expression data is noisy due to various types of systematic and stochastic fluctuations. Soong *et al.* adopted principle component analysis (PCA) and independent component analysis (ICA) to filter out noise in microarray data before feeding the data to SVM classifier [179]. As we see later in the section for integrated methods, gene expression is used frequently as one of input features for proteins.

3.2.4 Protein Tertiary Structure-Based Methods

The tertiary (3D) structure of proteins can be an important information to predict PPIs if available, or if the structures can be computationally reliably modeled. Many computational methods are developed that "docks" two protein structures to provide the tertiary structures of a protein complex from individual protein structures, which include LZerD [180-183], GRAMM-X [184], ZDOCK [185], RosettaDock [186], HADDOCK [187], SwarmDock [188], HEX [189], and ClusPro [190]. These docking methods build structure models of a protein complex given individual protein structures, which provide structural insights of the PPI. However, these docking methods do not predict whether a protein pair actually interact or not.

Then how to use structure information for predicting PPIs? There are two approaches explored. The first approach is to detect energetic characteristics of interacting protein pairs observed in protein docking prediction. A protein docking program generates typically over tens of thousands of different docking poses for a pair of input protein structures. Wass *et al.* reported the score distribution of docking poses of interacting protein pairs can be distinguished from those of non-interacting proteins, because the former distribution is skewed toward favorable scores [191]. This is an observation, because a docking pose distribution include both near-native (i.e. almost correct) and incorrect poses, therefore, the report implies that even incorrect docking poses have relatively favorable scores (i.e. more favorable geometric complementary) in instances of interacting proteins. In MEGADOCK, a protein docking method aimed for fast large-scale protein docking screening, a protein pair is predicted as interacting if a pool of docking poses generated by the algorithm include clusters of similar poses that have significantly favorable docking scores in comparison with the rest of the poses [192].

The second approach to use protein structure information for PPI prediction is, for two query protein structure, to find similarity in known protein complexes. PRISM, developed by Keskin and his colleagues, is one of the first to take this approach [193, 194]. PRISM takes two query protein structures as input, and examines if surface shapes of the proteins have similarity to docking interfaces from known protein complexes structures. To perform this comparison, PRISM has a database of docking interface regions of known protein complexes extracted from the PDB database [195]. Identified potential interface regions in the two query proteins that are identified by comparison to known interface regions are examined for structural similarity to the template, sequence conservation, and the binding energy. Although the prediction power of PRISM relies on the coverage of template dataset, the method will be able identify interactions between proteins that are globally dissimilar but has similar local interface regions to known protein complexes. PrePPI takes a similar approach PRISM [196]. A difference is that PrePPI takes sequences of the query proteins and model their structures by homology modeling. Subsequently, the two structures are mapped to known protein complex structures, which are then evaluated by structure and sequence similarity scores to the known complex structures. Final prediction is made by a composite score that integrates five other features, gene co-expression, essentiality of the proteins, functional similarity, and the phylogenetic profile. Similarly, Coev2Net models a complex structure of two query proteins by mapping their sequences to a known complex structure with a threading method, and then evaluates the complex model by a logistic regression classifier that considers structural and sequence features taken from its interface [197]. In a recent method, InterPred, a similar approach is taken [198]: for a query protein sequence pair, structures are modelled, then known protein complexes are sought by structure comparison. Finally, the feasibility of the model is evaluated using a random forest classifier that considers interface structure and sequence features as well as overall structure similarity between individual models to the template complex structure.

Although protein structures can provide unique features for PPI prediction, a drawback is that not many proteins have known structures. In Table 3.2, the number of protein genes with GO terms, gene expression data, and experimentally determined/computationally-modelled protein structures for ten genomes are shown. Compared to GO terms and gene expressions, proteins with known structures are substantially fewer. This is more evident for genomes that are less studied. On the other hand, as shown in the right-most column, most of the protein structures can be computationally modelled [199]. Thus, there is a room for new structure-based approaches that use modelled protein structures.

Table 3.2 Lists of available information such as number of coding genes, number of genes with annotated GO terms, genes with co-expression information, number of solved structure, and number of redundant modeled structure, in each organism.

Organism	# of genes with protein products[a]	# of proteins with annotated GO terms[b]	# of genes with co-expression information[c]	# of proteins with a solved structure[d]	Fraction of modeled structure among all proteins (%)[e][200][200][200][203][203][203](Pieper et al., 2006)(Pieper et al., 2006)
Human	109,018	46,331	19,816	16620	82.21
Yeast	6,002	5,582	4,461	1340	90.25
Mouse	76,216	28,727	20,403	2891	84.07
A. thaliana	48,350	16,123	20,836	584	77.44
Fruitfly	30,482	6,886	13,099	875	88.92
Asian rice	28,555	100	20,625	1	NA[f]
X. tropicalis	39,662	1,998	11,095	10	91.06
D. rerio	46,451	2,723	10,112	26	NA
C. annuum	45,410	20	17,453	0	NA
P. persica	28,927	2	11	0	NA

Note: [a] Counted in the NCBI reference sequence (RefSeq) [201]. [b] Protein entries in RefSeq was mapped to UniProt, where the number of proteins with at least one annotated Gene Ontology term was counted [202]. [c] Gene expression data of human, yeast, mouse, fruitfly, and D. rerio were taken from COXPRESdb [178]. Data for A. thaliana and O. Sativa were taken from the ATTED-II database [177]. Expression data for the rest of the three organisms were taken from literature: X. tropicalis: [203]. C. annuum: [204]. P. persica: [205]. [d] After mapping RefSeq IDs of protein genes to UniProt, the number of proteins was counted with at *least one crosslinked PDB entry [202].* [e] *It is computed from the statistics provided in the ModBase base [200].* [f] *NA (0) indicates that no modeled structures provided on ModBase (but some proteins in a genome may be modelled by a standard modeling procedure.*

3.2.5 PPI Network Topology-Based Methods

Methods in this category start from an existing PPI network of an organism, and predict new interactions between proteins by evaluating their network topology features. In the IRAP method, a missing interaction is predicted if a protein pair has a high score that reflects the number of common neighbors between them in the current PPI network [206]. Another idea by Yu *et al.* is to predict a PPI if two proteins are neighbors of a clique, a fully-connected graph, in the PPI network of the organism and connecting them would complete a larger clique, because most probably the two proteins are subunits of a protein complex [207]. In the work by L. Wong and his colleagues, a prediction of a PPI is made using a combination of two scores, a score for capturing local network topology of proteins that is based on the number of common neighbors and a global topology-based score that accounts for the memberships of the proteins in protein groups where member proteins interact with each other [208]. Kuchaiev *et al.* applied Multi-Dimensional Scaling (MDS), a dimension reduction method in statistics, to a PPI network, where distances are based on edge distances between proteins [209]. New PPIs are predicted if proteins are closer than a threshold in the projected space by MDS. Lei and Ruan applied a random walk-based approach, where the probability of reaching each node from each of the other nodes in the network is computed by assuming a random walk [210]. The resulting probability matrix contains information of the topology of the PPI network. Based on the probability matrix, protein pairs are connected if they are similar in their probability vectors to reach the other nodes.

3.2.6 Integration of Multiple Features

PPIs can be predicted from different perspectives as discussed above. Naturally, there are methods that use multiple features to be able to combine strengths of different features and to increase the prediction confidence and coverage. Features can be combined using machine learning methods, such as random forest, Naïve Bayesian Network, artificial neural network, SVM, and logistic regression [211]. Table 3.3 summarized methods that use multiple features.

From the table, the most popular feature integrated was gene co-expression data (COX). The next most popular ones are GO functional similarity (GO), and homology (HOM). Several features in the table are not explained yet in this review. The physicochemical features (PCH) concerns features such as charge and aromaticity of amino acids in a protein sequence. The posttranslational modification feature (PTM) indicates that PTM motifs are found in UniProt and HPRD. The disordered region (DIS) is a protein structure feature, non-structured regions in a protein, which can be predicted from its sequence. Thus, besides obvious sequence-based features, DIS, PCH, and PTM are features that are predicted from protein sequences. Direct experimental data of PPIs (EXP) used by Qi *et al.* were yeast-two-hybrid and mass spectrometry data [211], and those used by Miller *et al.* were data from yeast two-hybrid system [212]. The protein functional class (CLA) in yeast are taken from the MIPS Protein Class Catalogue, which were determined by experiments [213]. Gene essentiality (ESN), synthetic lethality (SNL), and MIPS mutant phenotype (MUT) were determined by knockout mutants [211]. Text mining (TXT) counts co-mentions of two proteins in PubMed abstracts.

Regarding combinations of features, methods by Ben-Hur *et al.*, Xu *et al.* combine mostly sequence-based features [214, 215]. On the other hand, PrePPI [196], FpClass [216], and Taghipour *et al.* [217] are intended to combine different types of features. Turning our attention to algorithms used, naïve Bayes is the most frequently used among the multiple feature-based methods in Table 3. SVM was the next frequently used in the three methods. Qi *et al.* tested five integrating algorithms with different feature combinations
[211].

Table 3.3 Features and algorithms used in multiple-feature integrative methods.

	Feature	Ben-Hur *et al.* [a]	Miller *et al.* [b]	Qi *et al.* [c]	Scott *et al.* [d]	Xu *et al.* [e]	PAIR [f]	PrePPI [g]	FpClass [h]	STRING [i]	Taghi-pour *et al.* [j]	# of times used
Sequence	**MOT/DOM**	**X**		**X**	**X**	**X**	**X**		**X**			**6**
	NGM	**X**										**1**
	PCH								**X**			**1**
	HOM	**X**		**X**	**X**	**X**	**X**		**X**	**X**		**7**
	COD		**X**									**1**
	PHP			**X**		**X**	**X**	**X**		**X**		**5**
	FUS			**X**		**X**				**X**		**3**
	GNB			**X**		**X**				**X**		**3**
	PTM				**X**				**X**			**2**
Function	**GO**	**X**	**X**	**X**		**X**	**X**	**X**	**X**		**X**	**8**
	MIPS							**X**				**1**
Experiment	**COX**		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**9**
	XPI		**X**	**X**								**2**
	CLA			**X**								**1**
	ESN		**X**	**X**				**X**				**3**
	LOC		**X**		**X**		**X**					**3**
	COR										**X**	**1**
	SNL			**X**								**1**

	MUT			**X**								**1**
Structure	**STR**							**X**				**1**
	DIS				**X**							**1**
Network	**NET**	**X**	**X**		**X**				**X**		**X**	**5**
Literature	**TXT**									**X**		**1**
Integrating method		**LPK**	**SVM**	**RF, KNN, NB, DT, LR, SVM**	**NB**	**NB**	**SVM**	**NB**	**NOR**	**LNR**	**NB, MCL**	

3.3 Discussion

The identification of PPIs is vital for systems level understanding of molecular activity of living cells. To complement experimental approaches, many computational tools were developed, which use different types of protein features. Through writing this article, we felt that a wide variety of features were explored already, and development of novel computational approaches would need new types of experimental data. Also, large scale PPI networks are experimentally revealed only for a limited number of organisms, and thus many computational methods were developed and benchmarked on those organisms. Therefore, for further advancement of PPI prediction, proteomics-scale PPIs of many more organisms would be needed.

Current PPI networks constructed both experimental and computational methods only represent a static snapshot of interactions of proteins in a cell, which are dynamically changing over time, containing both transient and permanent interactions. Therefore, the next generation of PPI studies would aim to capture the time-dependent, dynamic aspects of PPIs. Computationally, this direction would eventually meet and integrated with other computational approaches, such as pathway simulations and molecular dynamics simulation of molecules in a cell.

CHAPTER 4. COMPUTATIONAL METHOD TO IDENTIFY PROTEIN-PROTEIN INTERACTIONS IN PLANT GENOME[3]

4.1 Background

Identification of protein-protein interactions (PPIs) is important for understanding how proteins work together in a coordinated fashion in a cell to perform cellular functions. PPIs are essential for individual protein functions, forming various cellular pathways, and are also involved in the development of diseases. PPI data is directly useful for identifying protein multimeric complexes[219, 220], identifying biological pathways as well as predicting protein function[221-223]. For more on the application side, PPIs are also important targets for drug design[224] and artificial design of protein complexes[225].

In this work, a computational method for PPI prediction was developed, named PPIP (PPI prediction for Plant genomes), and applied to three major plant proteomes, *Arabidopsis thaliana*, *Zea mays* (maize), and *Glycine max* (soybean). To capture different aspects of proteins that are relevant to PPIs, a combination of four features for predicting PPIs was used, i.e. protein sequence properties, protein functional similarity, co-expression patterns, and phylogenetic profile similarity. To provide a confidence level of predictions, two machine learning methods, support vector machine (SVM) and random forest (RF), were separately trained on different features, and commonly predicted PPIs by the two methods were considered to have high confidence. The machine learning methods were trained on known PPIs from *Arabidopsis*. The accuracy on the testing dataset of *Arabidopsis* achieved a high accuracy of over 90%. PPIP predicted 50,220, 13,175,414, and 13,527,834 confident PPIs in *Arabidopsis*, maize, and soybean, respectively.

Examples of predicted novel PPIs with high confidence are discussed. All confident predictions are provided on our lab website (http://kiharalab.org/PPIP_results/) so that they can be referenced by plant biologists.

---

[3] This chapter have been previously published 218. Ding, Z. and D. Kihara, *Computational identification of protein-protein interactions in model plant proteomes.* Scientific Reports, 2019. **9**(1): p. 8740.

4.2 Methods

4.2.1 Sequence-based prediction

Protein sequence features used in PPIP is explained. It is the left branch of the flowchart in Figure 4.2. To capture physicochemical properties of interacting proteins the following seven features are assigned to each amino acid of query protein sequences [226-232]: hydrophobicity, hydrophilicity, side-chain volumes, polarity, polarizability, solvent-accessible surface area, and net charge index (NCI) of side-chains. Then each query protein sequence is represented with auto-covariance (AC) using strings of the seven features as follows:

$$AC(lag,j) = \frac{\sum_{i=1}^{L-lag}\left(P_{i,j}-\frac{1}{L}\sum_{i=1}^{L}P_{i,j}\right)\times\left(P_{(i+lag),j}-\frac{1}{L}\sum_{i=1}^{L}P_{ij}\right)}{L-lag}, \qquad \text{(Equation 4.1)}$$

where *lag* is the distance between covariant residues to consider, which ranges from 1 to 30, j is the *j-th* physiochemical feature, i is the position in the sequence, P*i,j* is the value of the physicochemical feature j of amino acid position i, and L is the length of the sequence. Thus, AC of a physicochemical feature with a certain *lag* length will be large if amino acid with a large (or small) property value appears periodically with an interval of *lag*. AC is computed for each protein sequence, and thus a query protein pair is represented as a 2 (sequences) * 7 (features) * 30 (lag intervals) = 420-dimensional vector. The vector representation was used as input of SVM for predicting if protein pairs are interacting or not interacting. This approach was adopted because it was reported to be successful in a previous paper [121].

4.2.2 Gene expression features

On the other branch of PPIP (Figure 4.2), three features were used inlcuding gene co-expression, functional similarity, and phylogenetic profile similarity in the framework of RF to predicted PPI of a query protein pair. The three features in the following three subsections were dicussed .

If two proteins are upregulated or down-regulated simultaneously under various conditions, it is highly likely that the two proteins are involved in the same pathway and have a higher chance that they physically interact with each other. Thus, co-expression patterns can provide indirect evidence for predicting PPIs. The gene coexpression information were obtained from microarray experiments and RNA-seq experiments from the ATTED-II database (http://atted.jp/)[177]. It is a database of pre-calculated Pearson’s correlation coefficients (PCC) and the mutual rank (MR) of co-expressed genes. MR is defined as the geometric mean of the rank of the correlation gene A to gene B among proteins in the genome and the rank of gene B to gene A. The smaller MR is, the stronger the genes are co-expressed. Since gene expression data are provided in two sources, microarray and RNA-seq, four features (microarray MR, microarray PCC, RNA-seq MR, and RNA-seq PCC) were used to represent the co-expression profile of protein pairs.

4.2.3 Protein function features

The second feature used is protein functional similarity. Proteins with the same or similar biological functions are likely to physically interact because they might form permanent complexes or take part in the same pathway. Functional similarity of proteins was quantified by established similarity scores of Gene Ontology (GO) terms[233]. GO annotations were obtained from UniProt[40] and TAIR (for *Arabidopsis*). Three GO similarity/relevance scores, Interaction association score (IAS)[43], Co-Occurrence Association Score (CAS), and PubMed Association Score (PAS) was used [31, 32, 234]. IAS, CAS, and PAS quantify how significantly a pair of GO terms appear in physically interacting proteins, annotations of individual genes, and PubMed abstracts. Thus, they evaluate co-occurrence of GO terms in biological contexts and shown to be effective in identifying proteins that physically interact)[43] or in the same pathways[235].

4.2.4 Phylogenetic profile similarity

It has been observed that interacting proteins tend to coevolve [145]. The phylogenetic profile is used to exploit the evolutionary co-occurrence patterns of interacting proteins. The basic assumption of the phylogenetic profile method is that interacting proteins either co-present or co-absent across organisms [236]. The original phylogenetic profile [236] is a binary pattern of presence or absence of homologs in a set of reference genomes, but a modified version of the profile is applied that used BLAST bit score instead of binary values as follows [158]:

$$sim(i,j) = \frac{\sum_{k=1}^{n} R_{ik} \times R_{jk}}{\left[\left(\sum_{k=1}^{n} R_{ik}^2\right) \times \left(\sum_{k=1}^{n} R_{jk}^2\right)\right]^{1/2}} \qquad \text{(Equation 4.2)}$$

Where $R_{ik} = \frac{B_{ik}}{B_{ii}}$ (Equation 4.3)

The similarity of protein *i* and *j* are defined in Equation 4.2. *k* is the *k*-th reference genome, R_{ik} is the BLAST search bit score of homolog of protein *i* in the *k*-th genome divided by the BLAST search bit score of *i* in the query genome (Equation 4.3). 100 reference genomes (*n = 100*) was used (Figure 4.1). These genomes were selected in the following steps: BLAST searches from all *Arabidopsis* protein sequences against the UniProt database was performed using the default E-value cutoff of 10. Then, a phylogenetic tree was constructed for the genomes and manually selected the genomes from each branch of the tree so that the selected genomes are well distributed and represent the tree.

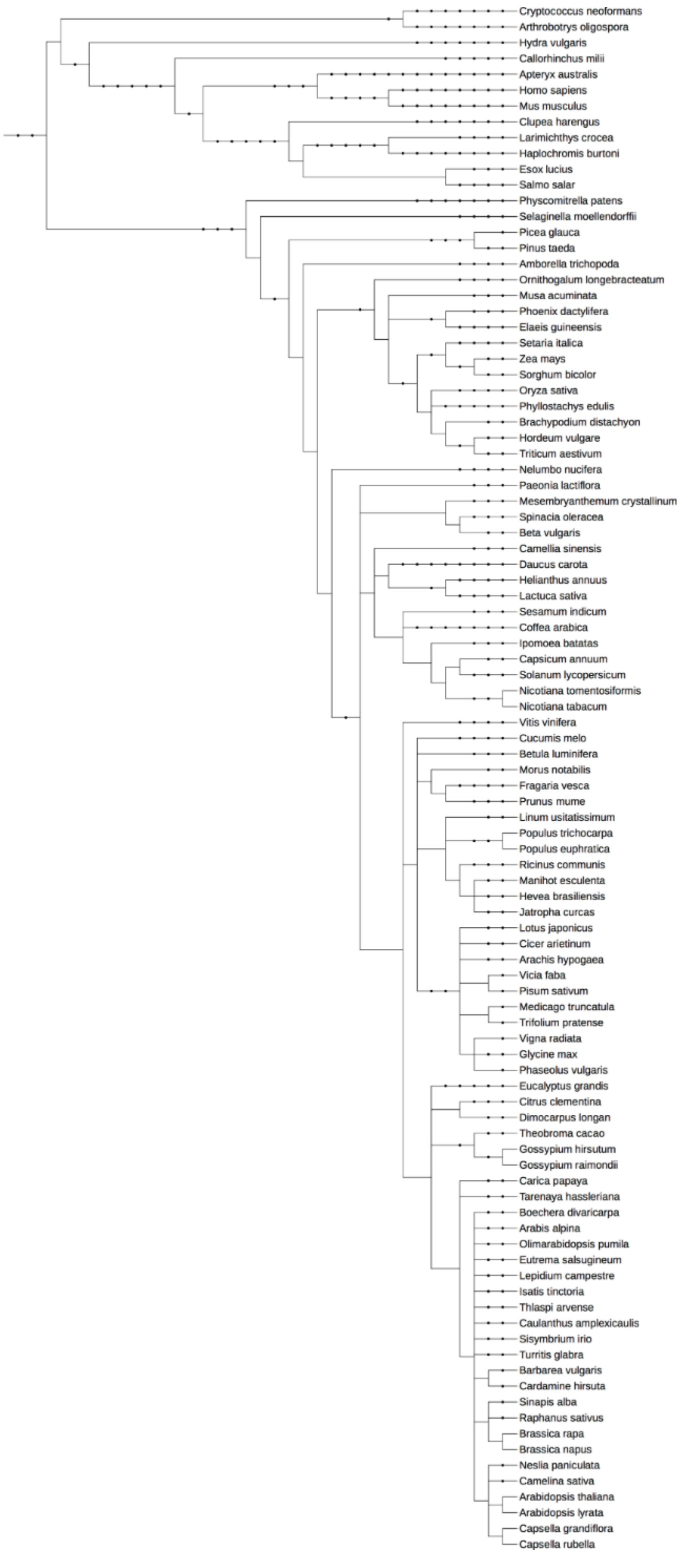


Figure 4.1 The phylogenetic tree of selected 100 reference organisms used to compute phylogenetic profile of proteins. The tree was generated with phyloT (http://phylot.biobyte.de/).

4.2.5 Machine learning methods

Two machine learning methods was used in PPIP, SVM for making predictions from sequence features and RF for predicting from a combination of four other features (Figure 4.2). For SVM, the software package libsvm 2.84 was used [237]. SVM uses a kernel function to transform input features and two hyper-parameters that need to be determined, a regularization parameter γ, which defines how far each training data influences the model and C, which controls the tradeoff of misclassification on training examples. For our kernel function, a radial kernel was used following previous works that predict PPI prediction from protein sequence features [70, 238-240]. The two hyper-parameter values, C and γ, were determined to be $log_2 C = 5$ and $log_2 \gamma = -1$ for SVM_{loc} as shown in Table 4.1 and $log_2 C = 1$ and $log_2 \gamma = 1$ for SVM_{rand} as shown in Table 4.2 by performing nested cross-validation [241, 242].

In parallel to the sequence-based prediction with SVM, RF was used to make an independent prediction from three features, functional similarity, gene co-expression, and phylogenetic profile similarity. RF is an ensemble learning method, which combines predictions made by a number of decision trees by a majority vote. RF can also determine important variables that contributed most in classification by calculating two metrics, the mean decrease of accuracy (MDA) and the mean decrease of Gini importance (MDGI) [243, 244]. MDA is the difference of the error rate of classification caused by permuting feature values with values of other data points in a dataset. MDGI tells how much less a particular feature is selected as a node in the random forest after permuting this feature. The larger MDA and MDGI of a certain feature are, the more important that feature is. Similarly to SVM, nested cross-validation was performed to determine three hyper-parameter values used in RF (Table 4.3 to Table 4.6).

Table 4.1 Prediction accuracy for validation sets from six-fold nested cross-validation on the *Arabidopsis* PPI_{loc} dataset using SVM with a radial kernel.

SVM_{loc}	**c**	**g**	**Accuracy**
1	log2c= 5	log2g= -1	0.934
2	log2c= 5	log2g= -1	0.932
3	log2c= 5	log2g= -1	0.937
4	log2c= 5	log2g= -1	0.930
5	log2c= 5	log2g= -1	0.929
6	log2c= 5	log2g= -1	0.933
Average			0.932

Note: To run SVM, two hyper-parameters, γ and C, need to be determined. From the Arabidopsis PPI dataset, short proteins whose length is less than 50 amino acids were excluded. The resulting dataset includes 4,759 interacting protein pairs and 4,759 non-interacting pairs. This dataset was separated into six subsets, where one subset was used for test and remaining five subsets were used for training and validation following the nested cross-validation procedure. By changing the test set among the six, the process was repeated six times. This corresponds to each row in the table. In nested cross-validation, the five subsets were used further for five-fold cross-validation. Four out of the five subsets were used for training and one subset was used for validation of the trained model under each hyper-parameter combination. This was repeated five times using a different subset for validation, and a hyper-parameter combination that gave the best average accuracy over the five validation subsets was selected. The accuracy is calculated as the total number of corrected prediction (true positive and true negative) divided by the total number of data (including true positive, true negative, false positive, and false negative). The table shows the best hype-parameter combination found and the average accuracy obtained by the hyper-parameter combination for each of the test – training/validation separation. $log_2 C$ *value was explored from -5 to 15 with a step size of 2 while* $log_2 \gamma$ *was changed from 3 to -15 with a step size of -2. Thus, in total 11 * 10 = 110 combinations were examined. For the* PPI_{loc} *dataset,* $log_2 C = 5$ *and* $log_2 \gamma = -1$ *were chosen because it was selected as the best for all six training/validation sets.*

Table 4.2 Prediction accuracy for validation sets from six-fold nested cross-validation on the *Arabidopsis* PPI_{rand} dataset using SVM with a radial kernel.

SVM_{rand}	c	g	Accuracy
1	log2c= 1	log2g= 1	0.768
2	log2c= 1	log2g= 1	0.765
3	log2c= 1	log2g= 1	0.775
4	log2c= 1	log2g= 1	0.766
5	log2c= 1	log2g= 1	0.760
6	log2c= 3	log2g= -1	0.774
Average			0.768

Note: For the PPI_{rand} *data set, the combination of* $log_2 C = 1$ *and* $log_2 \gamma = 1$ *was chosen because it was selected as the best in five out of six training/validation sets (a majority vote).*

Table 4.3 Results of a six-fold cross-validation on the PPI_{loc} dataset using random forest as the classifier. Eight features (RF_8) including the mutual rank and Pearson's correlation coefficient calculated for gene expression from microarray experiments and RNA-seq data, IAS, PAS, CAS, and the phylogenetic profile similarity score were used.

RF_{8loc}	ntree	mtry	cutoff	Accuracy
1	1400	5	0.5	0.796
2	1200	4	0.5	0.801
3	800	4	0.5	0.812
4	200	4	0.5	0.798
5	1400	3	0.5	0.790
6	600	5	0.5	0.798
Average				0.799

Note: To run random forest, three parameters, ntree, mtry, and cutoff need to be determined. Four co-expression features, the Pearson correlation coefficients and mutual ranks from microarray data and RNA-sequencing data, three functional association scores, IAS, PAS, and CAS, and the phylogenetic profile similarity score were used as features. Nested cross-validation was performed. Since, only 3,427 interacting protein pairs from the golden standard dataset have all eight features, 3,427 non-interacting protein pairs was randomly selected to balance the dataset. The ntree value was explored from 200 to 1400 with a step size of 200, mtry was explored from 3 to 6 with a step size of 1, and the cutoff value was explored from 0.3 to 0.7 with a step size of 0.2. For the PPI_{8loc} *data set, the combination of ntree = 800, mtry = 4, and cut-off = 0.5 were chosen because this combination achieved highest accuracy among six training/validation sets.*

Table 4.4 Results of a six-fold cross-validation on the PPI_{rand} dataset using random forest as the classifier. Eight features (RF_8) including the mutual rank and Pearson's correlation coefficient calculated for gene expression from microarray experiments and RNA-seq data, IAS, PAS, CAS, and the phylogenetic profile similarity score were used.

RF_{8rand}	**ntree**	**mtry**	**cutoff**	**Accuracy**
1	400	6	0.5	0.923
2	1200	6	0.5	0.923
3	400	4	0.5	0.924
4	400	6	0.5	0.910
5	400	6	0.5	0.915
6	1000	6	0.5	0.916
Average				0.919

Note: The combination of ntree = 400, mtry = 6, and cut-off = 0.5 were chosen because it was selected as the best in three out of six training/validation sets (a majority vote). In Table 4.10, these hyper-parameters were used to retrain the RF_{8rand} model and applied to the test set.

Table 4.5 Results of a six-fold cross-validation on the PPI_{loc} dataset using random forest as the classifier. Four features (RF_8) including IAS, PAS, CAS, and the phylogenetic profile similarity score were used.

RF_{4loc}	**ntree**	**mtry**	**cutoff**	**Accuracy**
1	400	2	0.5	0.793
2	200	2	0.5	0.794
3	800	2	0.5	0.808
4	1000	2	0.5	0.790
5	400	2	0.5	0.788
6	200	2	0.5	0.797
Average				0.795

Note: The nested cross-validation was performed as in Table 4.2. The ntree value was explored from 200 to 1000 with a step size of 200, mtry was explored from 2 to 4 with a step size of 1, and cutoff was explored from 0.3 to 0.7 with a step size of 0.2. The combination of ntree = 200, mtry = 2, and cut-off = 0.5 were chosen for RF_{4loc} because this combination was selected in two out of six training/validation sets and achieved higher accuracy than ntree = 400, mtry = 2, and cut-off = 0.5.

Table 4.6 esults of a six-fold cross-validation on the PPI_{rand} dataset using random forest as the classifier. Four features (RF_8) including IAS, PAS, CAS, and the phylogenetic profile similarity score were used.

RF_{4rand}	**ntree**	**mtry**	**cutoff**	**Accuracy**
1	400	2	0.5	0.931
2	200	2	0.5	0.930
3	600	3	0.5	0.928
4	1000	4	0.5	0.915
5	400	3	0.5	0.921
6	1000	2	0.5	0.923
Average				0.925

Note: From the results, the combination of ntree = 400, mtry = 2, and cut-off = 0.5 was chosen because this combination achieved highest accuracy among six training/validation sets.

4.3 Results

4.3.1 Constructing a benchmark dataset of known *Arabidopsis* PPIs

First, two machine learning prediction algorithms was tested in our prediction method, PPIP, namely, support vector machine (SVM) and random forest (RF) on the dataset of known *Arabidopsis* PPIs obtained from the TAIR database[95] (Additional file 1: Supplemental Table S1). These PPIs were determined by experiments including X-ray crystallography, affinity-capture mass spectrometry, co-immunoprecipitation, fluorescent resonance energy transfer, isothermal titration calorimetry, and surface plasmon resonance. The downloaded known *Arabidopsis* PPI dataset contained 4,908 PPIs, which were reduced to 4,759 PPIs after removal of short proteins of less than 50 amino acid residues and PPI identified by genetic experimental systems.

To train and test a machine learning method, a negative dataset is also needed, i.e. a dataset of protein pairs that do not interact. Negative sets were constructed in two different ways as mentioned by Guo[121]. One is to pair proteins from different cellular localizations and thus highly unlikely to interact with each other. The cellular localization information was downloaded from the TAIR database. The dataset with the positive set and the negative set constructed in this way is named as $PPI_{loc.}$ Another one is to randomly pair proteins in the positive set and then exclude pairs that are already in the list of positive interacting pairs. The dataset with the positive set and the negative set constructed in this way is named as $\mathrm{PPI}_{\mathrm{rand}}$. Both $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$ included the equal number of interacting and non-interacting pairs; thus there are 9,518 pairs in total.

For training and testing RF, protein pairs that lack co-expression information needed to be removed. This reduced the number of interacting pairs to 3,427. By adding the equal number of non-interacting protein pairs either from $\mathrm{PPI}_{\mathrm{loc}}$ or $\mathrm{PPI}_{\mathrm{rand}}$, the total number of the dataset for RF has become 6,854. The dataset was were split into a training, a validation, and a testing set, respectively, and a rigorous nested cross-validation evaluation was performed to evaluate the prediction accuracy of PPIP.

4.3.2 The design of PPIP

PPIP predicts if a pair of proteins is likely to have physical interaction or not in physiological condition from the proteins' sequence and proteomic features. As illustrated in Figure 4.2, for a query protein pair, their physical interaction is predicted from physiochemical property features of amino acid sequences of the protein pairs using SVM. The SVM protocol was named as $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{SVM}_{\mathrm{rand}}$ corresponding to the $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$ dataset used, respectively. The features used were hydrophobicity, hydrophilicity, side-chain volumes, polarity, polarizability, solvent-accessible surface area, and net charge index (NCI) of side-chains. In parallel, the complementary features of gene co-expression, functional similarity, and the phylogenetic profile[236] were used to make another independent prediction by RF. The RF protocol was named as $\mathrm{RF}_{\mathrm{loc}}$ and $\mathrm{RF}_{\mathrm{rand}}$ corresponding to refer to the $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$ dataset used, respectively. Predictions with RF were performed in two settings, one with all the features and the other without gene expression features (thus three functional similarity features and the phylogenetic profile) because gene expression data is currently not available for maize and soybean. See Methods for more details about the features.

4.3.3 Prediction Performance on the Known *Arabidopsis* PPIs

On the $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$ datasets of known PPIs and non-interacting protein pairs of *Arabidopsis*, parameters of SVM and RF were trained and tested using six-fold nested cross-validation. In this rigorous validation procedure, the dataset is split into six subsets, and prediction accuracy was measured on each of the subsets using parameters optimized on the rest of the five subsets. Using SVM, the overall prediction accuracies for the $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$ datasets were 91.9%

and 70.8%, respectively (Table 4.7 and Table 4.8). Thus, SVM performed better on the negative set with protein pairs from different cellular locations than on negative protein pairs that were randomly combined from the interacting pairs (Table 4.7 and Table 4.8). This is consistent with the conclusion in the paper by Guo[121], who performed a similar comparison of negative datasets.

On the other hand, RF_{rand} trained on PPI_{rand} performed better than RF_{loc}, which was trained on PPI_{loc} (Table 4.9 and Table 4.10). This order was consistently observed when the eight and the four features were used in RF. The accuracy of RF_{8rand} and RF_{4rand} were 92.0% and 92.6% accuracy, respectively. On the PPI_{loc}, the accuracies were lower, 80.0% and 79.6% for RF_{8loc} and RF_{4loc}, respectively (Table 4.9 and Table 4.10).

Table 4.7 The prediction results on the PPI_{loc} dataset of known *Arabidopsis* PPIs using SVM (SVM_{loc}).

SVM_{loc}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1586	685	777	16	108	0.922
2	1586	717	707	86	76	0.898
3	1586	701	716	77	92	0.893
4	1586	721	742	51	72	0.922
5	1586	739	790	3	54	0.964
6	1588	663	791	3	131	0.916
Average						0.919

Table 4.8 The prediction results on the PPI_{rand} dataset of known *Arabidopsis* PPIs using SVM (SVM_{rand}).

SVM_{rand}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1587	353	698	97	440	0.662
2	1586	557	649	144	236	0.760
3	1586	470	676	117	323	0.723
4	1586	377	692	101	416	0.674
5	1586	482	679	114	311	0.732
6	1586	428	674	119	364	0.695
Average						0.708

Table 4.9 The prediction results on the PPI_{loc} dataset using RF with the eight features (RF_{8loc}) and four features (RF_{4loc}), respectively.

With eight features

RF_{8loc}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1142	435	504	67	136	0.822
2	1142	430	460	111	141	0.779
3	1142	382	447	124	189	0.726
4	1142	438	483	88	133	0.806
5	1142	458	518	53	113	0.855
6	1144	476	453	119	96	0.812
Average						0.800

With four features

RF_{4loc}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1142	433	498	73	138	0.815
2	1142	421	470	101	150	0.780
3	1142	399	437	134	172	0.732
4	1142	439	483	88	132	0.807
5	1142	450	498	73	121	0.830
6	1144	466	459	113	106	0.809
Average						0.796

Table 4.10 The prediction results on the PPI_{rand} dataset using RF with the eight features (RF_{8rand}) and four features (RF_{4rand}), respectively.

With eight features						
RF_{8rand}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1142	482	545	26	89	0.899
2	1142	494	537	34	77	0.903
3	1142	479	543	28	92	0.895
4	1142	528	540	31	43	0.935
5	1142	549	535	36	22	0.949
6	1144	546	531	40	25	0.941
Average						0.920

With four features						
RF_{4rand}	**Number of PPIs**	**True Pos.**	**True Neg.**	**False Pos.**	**False Neg.**	**Test Set Accuracy**
1	1142	487	540	32	85	0.899
2	1142	493	536	36	79	0.901
3	1142	495	542	30	77	0.908
4	1142	534	550	22	38	0.950
5	1142	553	539	33	19	0.956
6	1144	544	536	36	28	0.944
Average						0.926

An advantage of random forest is that it can provide the importance of each feature in making correct classification using two metrics, the mean decrease of accuracy (MDA) and the mean decrease of Gini importance (MDGI) [243]. As shown in Table 4.11, two functional association scores, IAS, PAS, were found to be the two most important features for both RF_{8loc} and RF_{8rand} models. The IAS score was calculated based on the frequency of two GO terms annotating interacting proteins while the PAS score was calculated from the co-occurrence of two GO terms in PubMed abstracts. Therefore, it is reasonable that these two scores contribute largely to classifying interacting and non-interacting protein pairs, because they evaluate biological contexts of GO terms. The results in the table also showed that the phylogenetic profile was more informative than gene expression features.

Table 4.11 The importance of variables for classifying positive and negative data using RF.

RF_{8loc}	**Negative**	**Positive**	**MDA**	**MDGI**
micro_MR	38.6121	22.57167	49.13145	187.8184
micro_PCC	29.49857	20.7819	38.0066	179.3132
RNA_MR	51.68748	21.52674	60.84656	214.8184
RNA_PCC	35.877	21.73235	47.67577	194.7815
IAS	118.53192	149.28542	198.01382	783.0412
PAS	88.13659	145.18851	181.64305	1096.2835
CAS	-11.86095	62.29549	50.44961	396.5665
PHY	49.57318	73.07908	81.35968	373.8518

RF_{8rand}	**Negative**	**Positive**	**MDA**	**MDGI**
micro_MR	9.692238	7.019082	10.45622	58.2901
micro_PCC	11.20431	8.213752	13.44388	58.13418
RNA_MR	20.07403	7.774314	13.95462	60.91505
RNA_PCC	12.833527	8.10697	14.92287	56.76476
IAS	126.754812	118.383743	170.353	1908.49263
PAS	50.476012	111.477796	99.08722	750.90133
CAS	53.036897	39.316735	55.15384	433.29671
PHY	6.013217	13.609914	13.19855	99.69469

Note: micro_MR, micro_PCC, the mutual rank (MR) and the Pearson's correlation coefficient (PCC) from microarray data; RNA_MR, RNA_PCC, MR and PCC for RNA-sequencing data. IAS, PAS, CAS, three functional association scores. PHY, the phylogenetic profile similarity score. Negative, the mean decrease of the true negative rate in classifying non-interacting protein pairs; Positive, the mean decrease of the true positive rate in classifying interacting protein pairs. Instead of using decreased accuracy (DA), "positive" and "negative" are using decreased true positive rate (recall) and decreased true negative rate (specificity), respectively. MDA, the mean decrease of accuracy; MDGI, the mean decrease of Gini importance. The importance of each feature for making correct classification was measured with MDA and MDGI. The larger of the value is, the more important the feature is in classifying protein pairs. From these results, IAS and PAS are the two most important features.

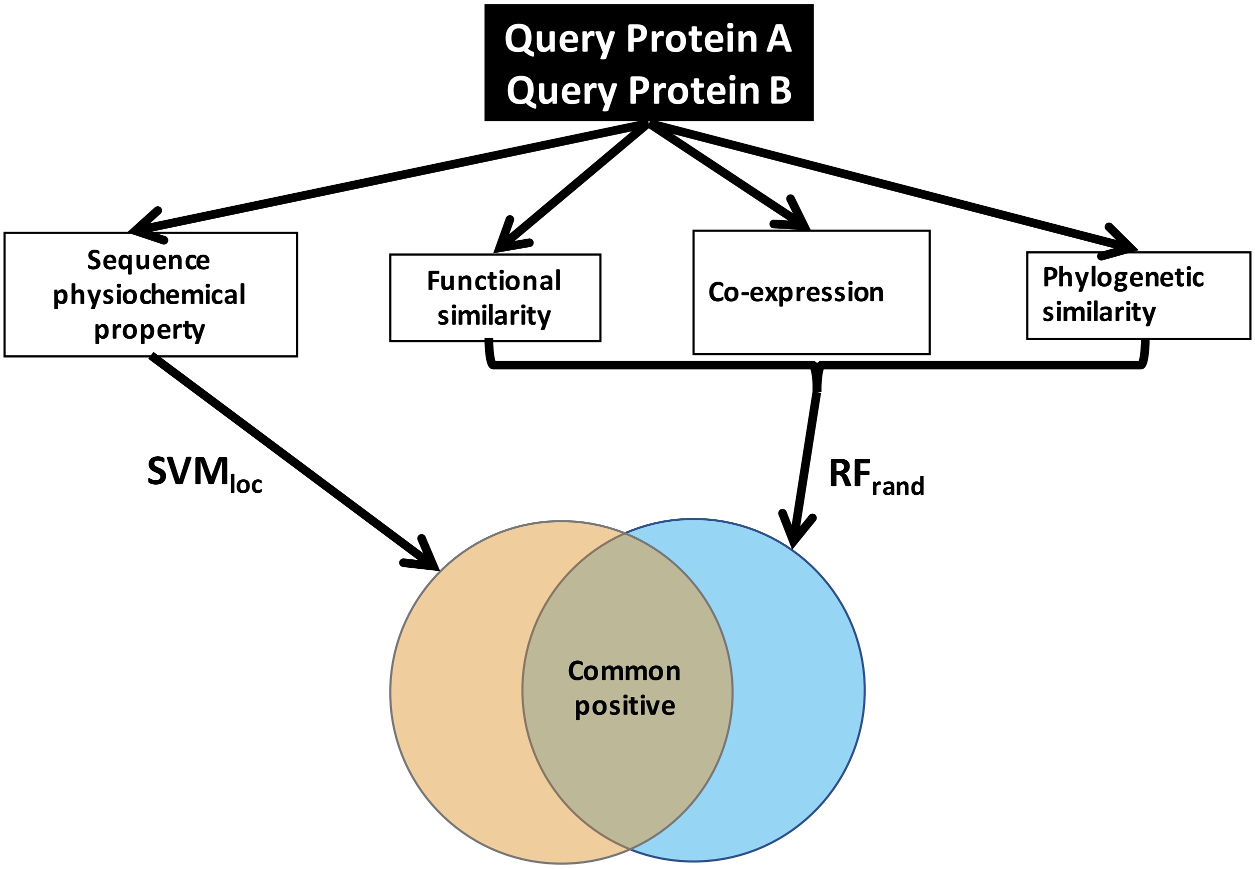


Figure 4.2 The schematic workflow of protein-protein interaction prediction by PPIP. Given a pair of protein A and B, physiochemical property features from their amino acid sequences are extracted and their interaction is predicted by support vector machine (SVM_{loc}). In parallel, functional features including functional similarity scores, gene co-expression scores, and phylogenetic profile similarity score are used to make an independent prediction of interaction by random forest (RF_{rand}).

4.3.4 Comparison with STRING confidence scores

Our prediction with the confidence score of protein-protein interactions was compared with the STRING database[55]. Since SVM used in the PPIP pipeline perform binary classifications and do not provide probability values that are comparable to the STRING scores, SVM regression was used for this comparison. Two types of SVM regression models were used, SVM-ϵ and SVM-υ[245]. SVM-ϵ controls the error tolerance in the training set whereas SVM-υ uses an additional parameter υ to control the proportion of the data points to be used as support vectors. These two SVM regression models was trained using the same hyperparameters as $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{SVM}_{\mathrm{rand.}}$ As for RF, the fraction of decision trees in RF that vote for protein-protein interaction as the probability was used.

First, I checked how consistent the predictions by SVM-ε and SVM-ν were with the SVM models in PPIP if a probability value of 0.5 in the SVM regressions was used to convert a probability value of SVM-ε and SVM-ν to binary prediction. On the $\mathrm{PPI}_{\mathrm{loc}}$ dataset, the prediction by $\mathrm{SVM}_{\mathrm{loc}}$-ε was consistent with $\mathrm{SVM}_{\mathrm{loc}}$ on 99.2% (9445 among 9518) of the cases while $\mathrm{SVM}_{\mathrm{loc}}$-ν was consistent with $\mathrm{SVM}_{\mathrm{loc}}$ on 97.8% of the cases (9312 among 9518). On the $\mathrm{PPI}_{\mathrm{rand}}$ dataset, $\mathrm{SVM}_{\mathrm{rand}}$-ε's results were consistent with $\mathrm{SVM}_{\mathrm{rand}}$ on 73.9% (7032 among 9518) of the cases while $\mathrm{SVM}_{\mathrm{rand}}$-ν's results agreed with $\mathrm{SVM}_{\mathrm{rand}}$ on 73.4% (6987 among 9518) of the proteins.

With these agreement results, the performance of SVM-ε, SVM-ν, and RF was compared with STRING. The two datasets, $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$, were used, and the comparison was made using Area Under the Curve (AUC) of Receiver operating characteristic (ROC) (Figure 4.3). On the $\mathrm{PPI}_{\mathrm{loc}}$ dataset (Figure 4.3A), two SVM models, $\mathrm{SVM}_{\mathrm{loc}}$-ε and $\mathrm{SVM}_{\mathrm{loc}}$-ν showed substantially higher AUC (0.98 and 0.97, respectively) than STRING. RF showed the same AUC value as STRING (0.88). On $\mathrm{PPI}_{\mathrm{rand}}$, $\mathrm{RF}_{\mathrm{rand}}$ performed the best with AUC of 0.98 and STRING came the second (AUC: 0.86) (Figure 4.3B). Thus, on the both datasets, at least one of our methods showed substantially better AUC than STRING.

Based on these results (Figure 4.3), $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{RF}_{\mathrm{rand}}$ was combined in the PPIP pipeline (Figure 4.2) since they achieved a very high AUC, substantially better than STRING, in the genome-scale prediction in *Arabidopsis*, maize, and soybean.

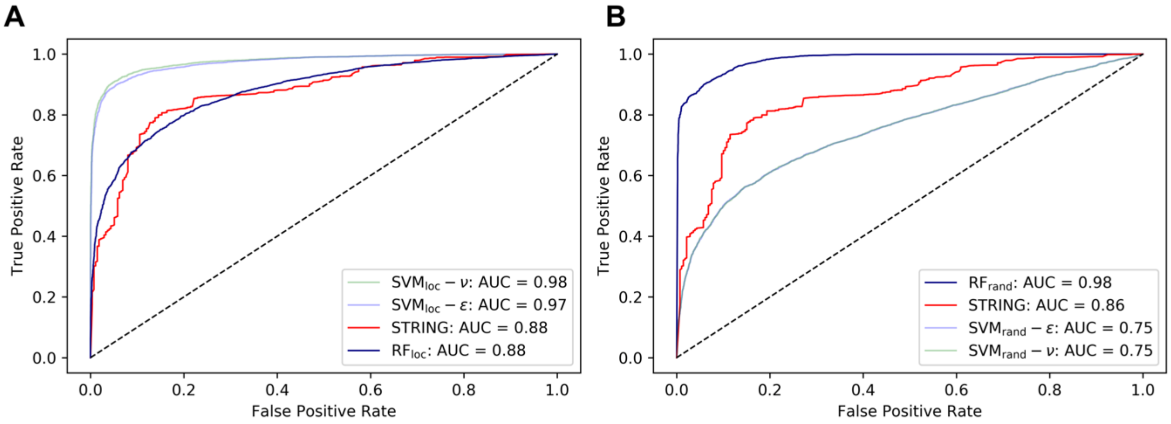


Figure 4.3 Comparison of PPI detection performance with STRING in identifying interacting protein pairs in *Arabidopsis.* Area Under the Curve (AUC) of Receiver operating characteristic (ROC) was used for comparison. A, Evaluation on the PPI_{loc} dataset; B, Evaluation on the PPI_{rand} dataset. $SVM_{loc/rand}$-ε and –ν are SVM regression models trained on the PPI_{loc} or PPI_{rand}, respectively. The probability for RF was computed as the fraction of votes from decision trees. For STRING, protein pairs were not included if they are not listed in STRING. A dashed line shows a random retrieval, which has an AUC of 0.5.

4.3.5 PPI detection by taking overlap between SVM and RF

In Table 4.12, PPI detection performance was examined by combining SVM and RF on the two datasets, $\mathrm{PPI}_{\mathrm{loc}}$ and $\mathrm{PPI}_{\mathrm{rand}}$. As shown in the table, precision improved by taking consensus: On $\mathrm{PPI}_{\mathrm{loc,}}$ while the precision of $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{RF}_{\mathrm{loc}}$ were 0.947 and 0.823, respectively, the combination of the two methods showed a higher value of 0.980. Similarly, On $\mathrm{PPI}_{\mathrm{rand,}}$ the combination of the two methods showed the perfect precision of 1.0. Note that the improvement of precision was a tradeoff with the accuracy, which decreased because apparently PPIs are missed if they are not detected by both SVM and RF. However, maintaining a high precision is more important when it comes to a genome-scale prediction because producing many false positive would be a serious concern.

The performance of $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{RF}_{\mathrm{rand}}$ were further tested on a large negative dataset of 2,038,222 protein pairs that are assembled by exhaustively pairing proteins from different cellular locations. The false positive rate of $\mathrm{SVM}_{\mathrm{loc}}$ on this dataset was 63.7%. $\mathrm{RF}_{\mathrm{rand}}$ was also tested after excluding pairs that do not have gene expression data (so that the RF model can run with gene expression features), which remained 1,048,575 pairs in the dataset. $\mathrm{RF}_{\mathrm{rand}}$ recorded a small false positive rate of 4.36%. Finally, as designed in the PPIP pipeline, the overlap between $\mathrm{SVM}_{\mathrm{loc}}$ and $\mathrm{RF}_{\mathrm{rand}}$ was taken, which yielded a very small false positive rate of 2.68%. The results confirm that the design of PPIP was effective in reducing false positives and that PPIP is suitable for a genome-scale prediction.

Table 4.12 PPI prediction by overlap with SVM and RF.

Model	Precision	Accuracy
SVM_{loc}	0.947 (4226/4426)	0.919 (8749/9518)
RF_{8loc}	0.823 (2619/3181)	0.800 (5484/6854)
SVM_{rand}	0.794 (2667/3359)	0.708 (6735/9518)
RF_{rand}	0.940 (3078/3273)	0.920 (6309/6854)
SVM_{loc} & RF_{8loc}	0.980 (1857/1894)	0.676 (4634/6854)
SVM_{rand} & RF_{rand}	1.0 (1260/1260)	0.713 (4884/6854)

Note: PPI detection performances by combining SVM and RF are examined on the two datasets, PPI_{loc} and PPI_{rand}. Precision is defined as the number of true interacting pairs predicted over the number of true interacting pairs and wrongly predicted interacting pairs (these numbers are shown in the parentheses of the precision column). Accuracy is defined as the number of corrected prediction (true positive and true negative) over the total number of data (including true positive, true negative, false positive, and false negative).

4.3.6 Prediction Performance on the BioGRID *Arabidopsis* Dataset

The prediction ability of our prediction method PPIP was further tested on a different *Arabidopsis* dataset, which was obtained from the BioGRID database[29]. BioGRID contained PPI data that were not included in the TAIR database, partly because it was updated more recently. In total, 3,280 *Arabidopsis* physical PPIs which have been verified by at least two experiments and not included in the TAIR-based dataset were found in BioGRID. The data were further pruned by the sequence identity cutoffs, 80%, 50%, and 30% (Table 4.13). The overlaps between the training dataset (PPI_{loc} and PPI_{rand}) were excluded. The sequence identity is a measurement of the protein sequence similarity from BLAST. Prediction by RF was applied for a smaller fraction of PPIs, only to those which have co-expression information. The data sets used by SVM and RF with different sequence identify cutoffs are provided in Supplemental Table S4.2. While an evaluation on over-prediction of our methods was extensively performed on a large negative dataset in the previous section, this benchmark provides an additional check of the SVM and RF models in terms of recall.

The recall values of SVM were somewhat lower than what was observed on the TAIR-based dataset (0.8880 for SVM_{loc} and 0.5606 for SVM_{rand}, which can be computed as the fraction of the sum of "True Positives" from the 1-6 test sets among the total of positives, i.e. the sum of True Positives" and "False Negatives" in Table 4.7 and Table 4.8). On the other hand, the recall values of RF were in the same range as the value observed on the TAIR-based dataset (0.7642 for RF_{8loc}, 0.7610 for RF_{4loc}, 0.8984 for RF_{8rand} and 0.9050 for RF_{4rand} from Table 4.9 and Table 4.10, which can be computed in the same way). RF's two predictions with different feature sets, the full features and the feature set without gene expression features, yielded almost identical recall. Among the two SVM models, SVM_{loc} showed a higher recall. When the two RF models were compared, RF_{rand} achieved a higher recall than the counterpart, $\text{RF}_{\text{loc.}}$ These results are consistent with the TAIR-based dataset (Table 4.9 and Table 4.10), which would justify our earlier choice of combining SVM_{loc} and RF_{rand} for the genome-scale PPI predictions to be discussed in the subsequent sections.

Table 4.13 The prediction accuracy (recall) on the BioGRID PPI dataset.

Seq. Identity cutoff	PPIs subject to prediction by SVM (RF)$^{a)}$	Recall by $\text{SVM}_{\text{loc}}^{b)}$	Recall by $\text{RF}_{\text{loc}}^{c)}$	Recall by $\text{SVM}_{\text{rand}}^{b)}$	Recall by $\text{RF}_{\text{rand}}^{c)}$
All PPIs	3280 (2468)	0.7466 (0.7057)	0.8440 (0.8327)	0.3250 (0.2565)	0.9444 (0.9444)
80%	1123 (797)	0.7266 (0.7114)	0.7804 (0.7604)	0.2654 (0.2597)	0.9448 (0.9448)
50%	937 (660)	0.7033 (0.6818)	0.7758 (0.7561)	0.2465 (0.2303)	0.9470 (0.9470)
30%	825 (585)	0.6909 (0.6667)	0.7641 (0.7453)	0.2364 (0.2239)	0.9470 (0.9470)

Note: The PPIs were clustered by the sequence identity cutoffs of 80%, 50%, and 30% to reduce similar sequences. The sequence identity of protein pairs was computed with the Needleman-Wunsch (global sequence alignment) algorithm implemented in the nwalign python library. On this dataset, recall was evaluated, i.e. the fraction of PPIs in the datasets that were correctly predicted as interacting protein pairs. SVM trained by PPI_{loc} or PPI_{rand} were named as SVM_{loc} or SVM_{rand}. RF trained by PPI_{loc} or PPI_{rand} were named as RF_{loc} or RF_{rand}. a) RF with the eight features was able to be applied only for PPIs that have gene co-expression data available. The numbers in the parentheses count such PPIs with expression data available. b) Recall is the fraction of PPIs that are correctly predicted. In the parentheses, SVM recall values measured on the PPIs with co-expression data i.e. the same dataset as used for prediction by RF with the eight features, are shown. c) The values show the recall of RF using the eight features including the gene expression features. Results with the four feature combinations that only use the functional association scores and the phylogenetic profile are provided in the parentheses.

4.3.7 PPI prediction for three plant genomes

Next, PPIP was applied to the three plant genomes, *Arabidopsis*, *Zea mays* (maize), and *Glycine max* (soybean). The genome sequences of the three plants were downloaded from the UniProt database[40]. Since the number of all possible protein pairs in the whole genome is too large, PPIP was applied only for protein pairs that are likely to co-locate in a cell, having a sufficient similarity in their Cellular Component (CC) Gene Ontology (GO) category terms[233], which describe the sub-cellular locations of proteins. Since many protein genes in maize and soybean do not have GO term annotations in UniProt, a function prediction method was used, PFP[30, 39, 246] to predict GO terms to supplement annotations to proteins. PFP is one of the top performing function prediction methods, which performs better than conventional methods, e.g. BLAST[247], as was also demonstrated in a community-wide function annotation assessment, CAFA[248, 249]. From PFP, only high confidence GO predictions with a score of over 10,000 were used[39]. The similarity of CC terms of two proteins was evaluated by the FunSim score, which essentially is the average pairwise similarity of CC GO terms[31, 250]. Protein pairs with a FunSim score of CC terms over 0.4 were subject to the prediction with PPIP. This cutoff was determined from the distribution of the CC-FunSim score of predicted *Arabidopsis* PPIs in three previous papers that made predictions based on the assumption that PPIs are conserved across species[135, 136, 138] (Figure 4.4). Proteins that do not have CC annotations even with PFP prediction were also discarded from the PPI prediction. Applying this pre-screening reduced the number of protein pairs to 21.36% (133,074,361 pairs), 13.54% (24,814,793 pairs), and 15.27% (54,814,995 pairs) of all possible protein pairs for *Arabidopsis*, maize, and soybean, respectively. This pre-screening process would likely miss some true PPIs that do not satisfy the CC similarity criteria, nevertheless, I decided to apply the process because having a common subcellular co-localization can serve as additional supporting evidence of PPIs.

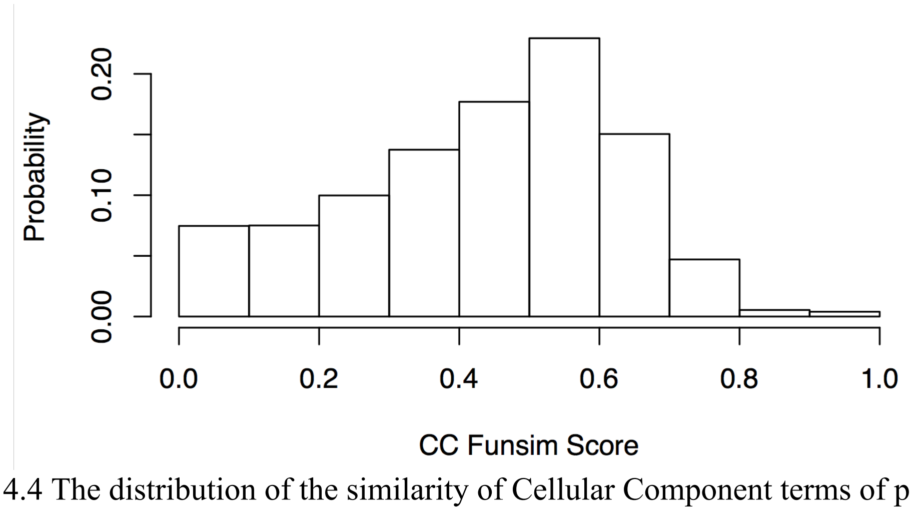


Figure 4.4 The distribution of the similarity of Cellular Component terms of predicted *Arabidopsis* interlogs.

Note: PPIs predicted in three papers, Geisler-Lee et al. [135], De Bodt et al. [136], Dutkowski et al [138] were analyzed. There were in total 50,949 PPIs predicted in the three papers, among which 2,786 pairs had CC GO term annotations. The FunSim score of only CC terms was computed. From this plot, 0.4 was used as a cutoff to pre-screen protein pairs for the genome-scale screening of the three genomes, Arabidopsis, soybean, and maize. 61.31% of PPIs in the distribution had the score above 0.4.

The numbers of predicted PPIs in the three genomes by PPIP are summarized in Table 3. For protein pairs that satisfied the CC GO term similarity, SVM_{loc} and RF_{rand} were independently applied, and commonly predicted PPIs by SVM_{loc} and RF_{rand} were selected. Among protein pairs with CC-FunSim > 0.4, SVM_{loc} selected about 10%, 56% and 56% as interacting pairs while RF_{rand} predicted 1.83%, 17.45%, and 18.06% as interacting, for *Arabidopsis*, maize, and soybean, respectively (Table 3). The final PPI predictions, which are the PPIs predicted commonly by SVM_{loc} and RF_{rand}, were 0.0081%, 7.19%, and 3.77% out of all the possible protein pairs for *Arabidopsis,* maize, and soybean, respectively. Compared to the fraction of known PPIs in several other well-studied organisms in Table 4.14, the fraction of *Arabidopsis*, maize, and soybean PPIs from the current study is at the same level. Particularly, the fraction of predicted PPIs for *Arabidopsis* (0.0081%) seems relatively small, but this fraction is consistent in Table S9. All the predicted PPIs for the three plant genomes are available as Supplemental data on our lab website (http://kiharalab.org/PPIP_results/).

Table 4.14 The percentage of known PPIs among all possible protein pairs in well-studied organisms.

Organism	**Number of Proteins**	**Number of PPIs**	**Percentage of experimentally identified PPIs over all possible protein pairs (%)**
Arabidopsis thaliana	*27,636*	35,896	0.00939%
Homo sapiens	*20,213*	332,829	0.163%
Escherichia coli (K12)	*4,140*	12,801	0.149%
Saccharomyces cerevisiae (S288c)	*6,002*	108,088	0.600%
Drosophila melanogaster	*13,931*	47,068	0.0485%
Mus musculus	*22,089*	38,587	0.0158%

Note: The number of proteins are taken from the KEGG database. The number of PPIs are taken from the BioGRID database.

It is known that the degree distribution of a PPI network of an organism follows a power-law distribution, i.e. the histogram of the number of interactions (called the degree) for each protein is well approximated with a power-law, $p(k) \sim k^{-\gamma}$ where k is the fraction of proteins with a certain number of interactions, and γ is a parameter called the degree exponent, which determines the slope of the distribution [53, 251, 252]. Figure 4.5 shows that the PPIs of the three plants detected in the current work follow the power-law, with γ being 1.362 in *Arabidopsis*, 0.204 in maize, and 0.401 in soybean Table 4.15. Smaller degree exponents for maize and soybean indicate that these two plants have more hub proteins that interact with many proteins.

Table 4.15 The Summary of the number of predicted PPIs in the three plant genomes.

Organism	**CC > 0.4**	**SVM_{loc}**	**RF_{rand}**	**Common**	**Degree exponent**
Arabidopsis	133,074,561 (21.36%$_{\text{all}}$)	13,682,168 (10.28%$_{\text{cc}}$)	2,440,139 (1.83%$_{\text{cc}}$)	50,220 (0.0081%$_{\text{all}}$)	1.362
maize	24,814,793 (13.54%$_{\text{all}}$)	13,902,459 (56.02%$_{\text{cc}}$)	23,223,947 (17.45%$_{\text{cc}}$)	13,175,414 (7.19%$_{\text{all}}$)	0.204
soybean	54,814,995 (15.27%$_{\text{all}}$)	30,844,273 (56.27%$_{\text{cc}}$)	24,031,016 (18.06%$_{\text{cc}}$)	13,527,834 (3.77%$_{\text{all}}$)	0.401

Note: CC> 0.4, the number and the percentage of protein pairs among all the possible protein pairs that satisfied the CC FunSim score criterion of over 0.4; (%$_{all}$); SVM_{loc} and RF_{rand}, predicted PPIs among pairs that satisfied the CC > 0.4 criterion by SVM_{loc} and RF_{rand}, respectively; Common, commonly predicted PPIs by SVM_{loc} and RF_{rand}; Degree exponent, the parameter value of the power-law distribution of PPIs (Figure 4.5). %all is the percentage relative to the all possible protein pairs of the organism while %$_{cc}$ is the percentage relative to the protein pairs that satisfied CC > 0.4.

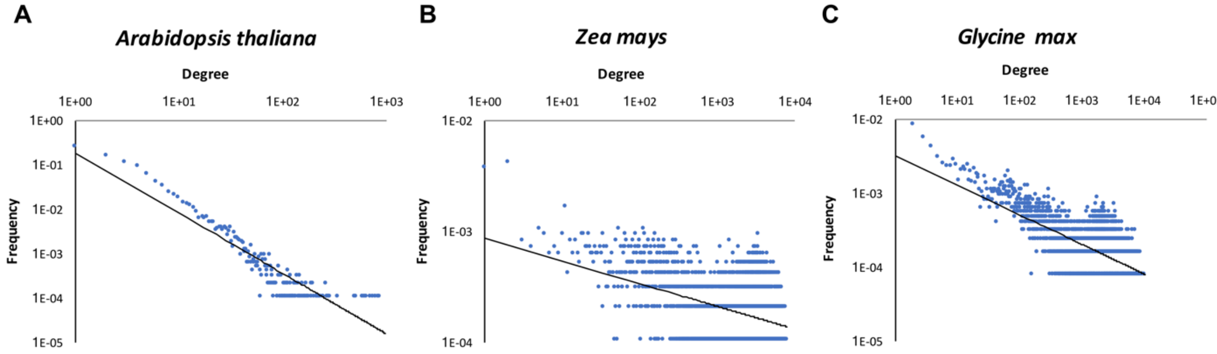


Figure 4.5 Degree distribution of proteins in the predicted protein-protein interaction network. The X-axis is the degree of proteins in the PPI network and the Y-axis is the frequency of proteins with a certain number of degrees. A log scale is used for both axes. A, *Arabidopsis thaliana*; B, *Zea mays* (maize); C, *Glycine max* (soybean). The exponents of power-law distribution are 1.362, 0.204, 0.401, respectively.

Next, our PPI prediction on *Arabidopsis* was compared with three existing genome-scale prediction results. These three works used a very different approach for prediction, the interlog concept[131], which assumes that interactions of orthologous proteins across different species are conserved. Geisler-Lee et al.[135] and De Bodt et al.[136] used the same reference organisms, *S. cerevisiae, C. elegans, D. melanogaster,* and *H. sapiens,* whereas Dutkowski et al.[138] used the same four organisms with two more organisms, *M. musculus,* and *R. norvegicus* (Table 4.16). Table 4.17 shows the number of PPIs predicted by the three works. All the three works predicted about the same number of PPIs, 14,009 to 19,974. The works by Geisler-Lee et al. and De. Bodt et al. made a very similar number of predictions, which is probably due to them using the same set of reference organisms. In Table 4.17, commonly predicted PPIs was compared by pairs of works including our method, PPIP. Geisler-Lee et al. and De. Bodt et al. had the largest number of common predictions, although the common predictions would be small considering the same approach and the reference organisms they used. Common predictions by other pairs including pairs with PPIP are roughly about the same numbers. Thus, PPIP has a similar level of agreement with the three previous works, although they took a very different approach from ours.

Table 4.16 The summary of predicted PPIs in *Arabidopsis* with the interlog concept.

Authors	**Number of predicted PPIs**	**Reference organisms**
Geisler-Lee et al.	19,979	*S. cerevisiae, C. elegans, D. melanogaster, H. sapiens*
De Bodt et al.	18,674	*S. cerevisiae, C. elegans, D. melanogaster, H. sapiens*
Dutkowski et al.	14,009	*S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, M. musculus, R. norvegicus*

Table 4.17 Commonly predicted PPIs by using interlogs and PPIP.

Method Combination	Number of commonly predicted PPIs
Geisler-Lee & De Bodt	934
Geisler-Lee & Dutkowski	294
De Bodt & Dutkowski	188
PPIP & Geisler-Lee	118
PPIP & De Bodt	208
PPIP & Dutkowski	47

4.3.8 Examples of predicted PPIs in *Arabidopsis*

From the predicted PPIs for the three plant genomes (http://kiharalab.org/PPIP_results/), here examples with three different levels of confidence are discussed. Supplemental Table S4.2-4.10 provide examples from *Arabidopsis*. All the listed PPIs in the tables were predicted consistently by the SVM_{loc} and RF_{rand} and the probability score of RF_{rand} was over 0.95. The difference of the confidence levels is based on the availability of additional supporting data.

The predictions are separated into three classes for each genome according to the availability of other evidence that supports the predictions. Supplemental Table S4.2 lists predicted PPIs with two more supporting evidence: a very high score of over 900 in the STRING database[55] and also satisfy at least one of the following three conditions: a) the two proteins are known to locate in the same pathway in the KEGG database[253], b) the two proteins are co-mentioned in a literature. The predicted PPIs with correlated elution profile in PPI detection using mass spectrometry (MS) [254, 255] are also included in Supplemental Table S4.2. STRING collects several different types of evidence for the functional association of protein pairs and provides a score that ranges from 0 to 1000 with 1000 as the most confident score. In the works by Aryal et al.[254, 255], proteins in *Arabidopsis* were fractionated using size exclusion chromatography, and abundance profiles across the column fractions were quantified using label-free precursor ion (MS1) intensity. Proteins with correlated profiles and clustered together among all detected proteins were more likely to form complexes (see the original papers for more details). Supplemental Table S4.3 lists predicted PPIs with at least one piece of additional evidence, either a common KEGG pathway or literature that co-mention the two proteins. Protein pairs listed in the Supplemental Table S4.4 do not have additional evidence because the proteins were not much studied before but were predicted with the highest score (1.0) by RF_{rand}.

The first half of Supplemental Table S4.2 lists predicted *Arabidopsis* PPIs with literature that describes evidence of their interaction. This list selected the most confident prediction in the *Arabidopsis*. Most of the interacting proteins are ribosomal proteins. Besides ribosomal proteins, several pairs of Sm-like proteins are predicted to interact, which are known as subunits of the heteroheptameric complexes and function in mRNA splicing and degradation[256, 257]. Another predicted PPI is plastid division protein PDV2 (AT2G16070) and protein accumulation and replication of chloroplasts 6 ARC6 (AT5G42480), which has been shown to interact in the intermembrane space to coordinate the division machinery of chloroplast membrane[258]. Our predicted PPIs also included some subunits of known complexes such as coatomer, RNA polymerase, DNA directed RNA polymerase, augmin, and adaptor complex 1 (AP-1).

The latter half of the table provides PPIs with similar MS elution profiles[254, 255]. For example, V-type proton ATPase subunit C (AT1G12840) and V-type proton ATPase subunit G2 (AT4G23710) are predicted to interact by RF and SVM. Additionally, they have highly correlated protein elution profile and are involved in the two common KEGG pathway including oxidative phosphorylation (ath00190), and phagosome (ath04145).

Supplemental Table S4.3 lists predicted PPIs in *Arabidopsis* with the next level of confidence, which has extra evidence but no information available in STRING or with a low STRING score (<400). An interesting example is asymmetric leaves 2 (AS2) (AT1G65620) and histone deacetylase 6 (HDA6) (AT5G63110). Although they have a very low STRING score a previous study showed that HDA6 functions with AS2 to regulate the leaf development and suggested that HDA6 may be the part of the AS1-AS2 repression complex to repress *KNOX* gene expression in *Arabidopsis*[259]. Another interesting example is protection of telomeres protein 1a POT1a (AT2G05210) and CST complex subunit TEN1 (AT1G56260). It is known that CTC1, STN1, and TEN1 consist of telomere complex and POT1a interplay with CST components to regulate the telomerase enzyme activity[260].

The last table for *Arabidopsis*, Supplemental Table S4.4, shows a list of PPIs with the highest RF probability score yet have no other known evidence. An interesting example in this table is pentatricopeptide repeat-containing protein (AT1G05600 and AT5G27270). They might function in RNA editing in chloroplast[261].

4.3.9 Examples of predicted PPIs in maize

Predictions for maize are shown in Supplemental Table S4.5, S4.6, S4.7, and Figure 4.6A. PPIs shown in these tables are predicted both by the SVM_{loc} and RF_{rand} with a high RF probability score of 0.9 or higher. As in the tables for *Arabidopsis*, Supplemental Table S4.5 lists predicted PPIs with two additional supporting pieces of evidence, and Supplemental Table S4.6 is for PPIs with a single existing piece of additional evidence, in this case having a common KEGG pathway, while Supplemental Table 4.7 includes the predicted PPIs with no existing evidence for interaction.

In Supplemental Table S4.5, 12 protein pairs out of 18 listed turned out to be NAD(P)H-quinone oxidoreductase subunits. This list also includes interaction between hydroxymethylglutaryl-CoA synthase (UniProt ID: B6U9M4) and acetyl-CoA acetyltransferase (UniProt ID: B4F9B2), both of which are involved in the four same pathways including terpenoid backbone biosynthesis (zma00900), valine leucine and isoleucine degradation (zma00280), butanoate metabolism (zma00650), and synthesis and degradation of ketone bodies (zma00072), and they have a very high database score in STRING.

Supplemental Table S4.6 shows predicted PPIs with the highest RF score (1.0) locating in at least one common KEGG pathway. For these PPIs, a STRING score was not available. As shown, most of the proteins are kinases in the same pathways, the plant hormone signal transduction (KEGG: zma04075) or the plant-pathogen interaction pathway (KEGG: zma04626). Thus, it is reasonable to conclude that they are interacting. In the table, protein functional association score (FunSim) with IAS scores are also provided. As mentioned in the Methods, IAS directly indicates the likelihood that proteins with the GO annotations interact. Since the IAS scores listed in the table are very high relative to the background distribution (within top 1 % for IAS and PAS for proteins in *Arabidopsis* protein pairs), these scores also support the PPI predictions.

The third list, Supplemental Table S4.7, includes PPIs that were predicted with a high confidence score (RF probability = 1) and high functional correlations (IAS>200 and PAS>20 and phylogenetic similarity > 0.9), but do not have other existing supporting information or protein annotations. When the 226 proteins involved in these PPIs were represented into a network by connecting protein pairs in the PPIs using Cytoscape[262], 224 of them (99.1%) are clustered into four subnetworks (Figure 4.6A). Since they have no functional annotations, PFP was used to predict their GO terms and performed GO enrichment analysis using NaviGO[31, 32] as shown in

Table 4.18. It turned out that each of the subnetworks has enriched GO terms that are common in the proteins in the network: The largest subnetwork involved in 101 proteins were predicted by PFP to be involved in flavonoid glucuronidation, flavonoid biosynthetic process, cellular glucuronidation, and quercertin 3-O-glucosyltransferase activity. In the second largest subnetwork, 57 out of 59 proteins were predicted to be involved in the RNA metabolic process and RNA secondary structure unwinding. All proteins in the third subnetwork were predicted to function in proteolysis and protein catabolic process while proteins in the last subnetwork were predicted to be involved in the regulation of the metabolic process, regulation in gene expression, and regulation of transcription DNA-templated. Thus, proteins in the predicted PPI networks have coherent biological functions.

4.3.10 Examples of predicted PPIs in soybean

The last plant genome to be analyzed was soybean. Soybean has much less available functional information in databases comparing with *Arabidopsis* and maize. Supplemental Table S4.8 selected a list of predicted PPIs using the same standard as Supplemental Table S4.5 for maize, i.e. PPIs supported by two additional evidence. This list includes subunits from known complexes, including ATP synthase, chalcone synthase, cytochrome, and NADH dehydrogenase.

The next table, Supplemental Table S4.9 shows PPIs without conclusive information in STRING. However, two proteins in each pair are found in the same KEGG pathway, and most of them have a similar function, judging from the name in KEGG or UniProt annotation. The predicted PPIs in this list include protein interaction between glycosyltransferase and protein kinase involved in plant hormone signal transduction, plant-pathogen interaction, zeatin biosynthesis, and carotenoid biosynthesis signaling pathway.

Supplemental Table S4.10 lists high confident PPIs (RF probability = 1, IAS>200, PAS>20 and phylogenetic similarity > 0.9), which do not have other existing supporting evidence and functional annotation in UniProt. As I did for the predicted PPIs in maize, in Figure 4.6B networks with 224 proteins was constructed that are involved in the PPIs in Table S10 and performed the functional enrichment analysis using predicted GO terms by PFP (Supplemental Table 4.13, the bottom half). 215 (96.0%) out of 224 proteins were included in four subnetworks. As observed in the subnetworks in maize (Figure 4.6A), proteins in each subnetwork are highly functionally relevant and would be reasonable to conclude that they are most likely to interact. All proteins in the largest subnetwork were predicted to be involved in the MAPK signaling pathway in response to stimuli. The second largest subnetwork with 41 proteins was predicted to be involved in the flavonoid biosynthetic process. Proteins in the third subnetwork are predicted to be involved in RNA processing and intracellular protein transport. The fourth subnetwork with 9 proteins is in a pathway for signal transduction and cell communication in response to the stimulus.

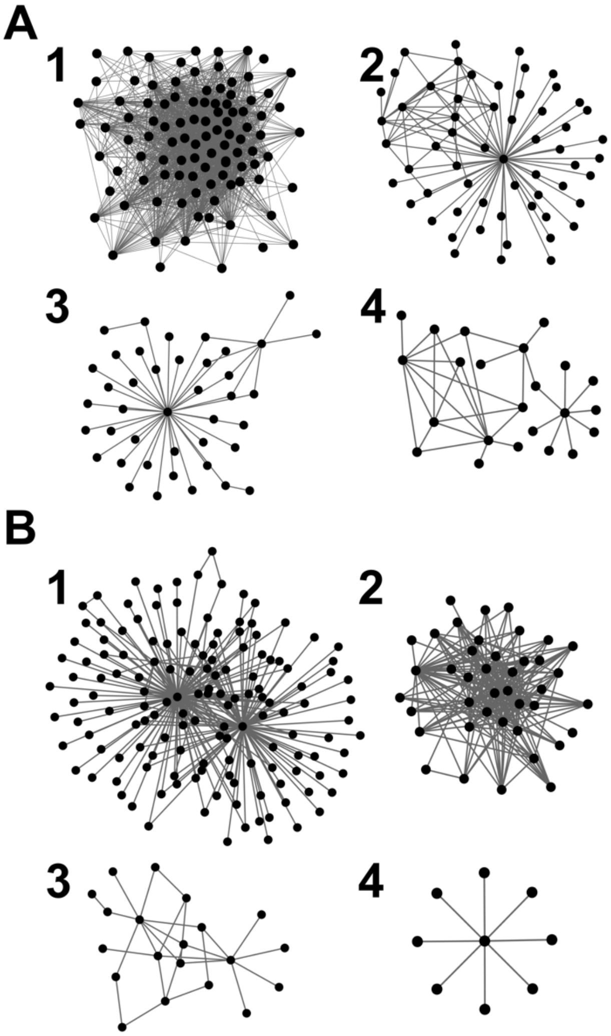


Figure 4.6 The networks constructed with predicted PPIs with the highest RF confidence scores (1.0) but do not have documented other supporting evidence. Connected PPIs are predicted by both SVM_{loc} and RF_{rand}. PPIs were further selected by high functional similarity scores: IAS-FunSim (> 200), PAS-FunSim (> 20), and the phylogenetic profile similarity (> 0.9). A, Predicted PPIs for maize. 224 out of 226 proteins in the PPIs qualified for the criteria are included in four subnetworks shown. The number of proteins in each network is 101, 59, 41, and 23 for the subnetwork 1 to 4, respectively. Table 4.18 provides the subnetwork index of each predicted PPI. B, predicted PPIs for soybean. 215 out of 224 proteins in the PPIs qualified for these criteria are included in four subnetworks shown. The number of proteins in each network is 145, 41, 20, and 9 for the subnetwork 1 to 4, respectively. Supplemental Table S4.10 provides the subnetwork index of each predicted PPI. See Table 4.18 for the results of the functional enrichment analysis of the subnetworks.

Table 4.18 Functional analysis of proteins in PPIs of maize and soy bean in Figure 4.6.

Sub-networks	**Number of proteins**	**Number of common GO terms**	**Number of common GO terms with P-value < 0.001**	**Function**
Maize 1	101	5	5	flavonoid glucuronidation, flavonoid biosynthetic process, cellular glucuronidation, and quercertin 3-O-glucosyltransferase activity
Maize 2	59	8*	6*	RNA metabolic process and RNA secondary structure unwinding
Maize 3	41	2	2	proteolysis and protein catabolic process
Maize 4	23	92	21	MAPK signaling, protein phosphorylation, neuron development, pathway in response to stimuli
Soybean 1	145	87	87	MAPK signaling pathway in response to stimuli
Soybean 2	41	5	5	flavonoid glucuronidation, flavonoid biosynthetic process, cellular glucuronidation, and quercertin 3-O-glucosyltransferase activity
Soybean 3	20	17	12	RNA processing and intracellular protein transport
Soybean 4	9	9	9	signal transduction and cell communication in response to the stimulus

Note: The "Number of common GO terms" column shows the predicted GO terms by PFP shared by all proteins involved in the subnetwork. For the maize subnetwork 2, there was no commonly shared GO term, but there are 8 commonly predicted biological process GO terms that are shared by 57 out of 59 proteins. The p-value indicates the rarity of the GO term in the protein set considering the number of proteins in the set, the number of proteins with that GO term in the organism, and the number of proteins in the organism (the function enrichment analysis).

4.4 Discussion

A PPI network is fundamental for understanding an organism's functional and structural units. For example, PPIs are very useful for predicting the function of individual proteins[124] as well as pathways of protein groups[263]. Although large-scale PPIs of several model organisms have been revealed by experimental methods[212, 264, 265] and by computational methods[74, 135, 138, 266], the works for plant PPIs were sparse. This work is intended to fill the gap for plant PPIs by providing PPI predictions with the method that was calibrated on known PPIs in *Arabidopsis*.

It is inevitable that a computational method often makes wrong PPI predictions. However, as discussed in the introduction, the situation would be similar in experimental methods, as it has been reported that independent experiments have substantial disagreements [26-28]. To reduce errors of a method, either computational or experimental, it would be useful to compare outputs from multiple methods. Having this idea in mind, PPIP was designed such that it combines two independent predictions, one using sequence-based features and the other with a combination of orthogonal features (Figure 4.2). This architecture sacrificed the recall rate but in return achieved a very low false positive rate, which is considered as a higher priority since all predicted PPIs are provided for reference information for biologists. Also, in the genome-scale predictions for the three genomes, additional evidence is provided from other sources whenever available. In the analysis, predicted PPIs with three levels of confidence are highlighted. All the PPIs in these three levels were predicted not only with high scores but with additional supporting evidence. PPIs in the first (best) confidence have direct literature information or multiple supporting data, including a high score in STRING and co-existence in the same pathway. In the second level, PPIs have at least one evidence including the co-existence in the same pathway. PPIs of the third level confidence are between proteins with functional coherence. While it is true that functional similarity between proteins does not necessarily indicate physical interaction between them, it is certainly highly related with each other as it is a common practice to verify experimentally detected PPIs by checking their functional similarity[212, 267, 268] and functional similarity is an informative feature of proteins for predicting PPIs[43, 77, 211, 269, 270]. PPI predictions made for the three plants are made available on our lab website (http://kiharalab.org/PPIP_results/). I hope they are used as a reference and found informative.

CHAPTER 5. COMBINATION OF COMPUTATIONAL METHOD AND MASS SPECTROMETRY DATA TO AID IDENTIFICATION OF PROTEIN-PROTEIN INTERACTIONS IN LARGE SCALE[4]

5.1 Background

Cyanobacteria are photosynthetic organisms that have played important roles in harvesting solar energy on a global scale and in the evolution of the oxygenic atmosphere [272]. They have great potential as a platform for carbon sequestration and biological energy production [273-276]. Flexible and diverse metabolic capabilities allow them to adapt to a wide range of environments. Among them, unicellular species such as *Cyanothece* 51142 can also fix atmospheric N_2, a process highly sensitive to oxygen [277]. This ability to carry out two opposite biological processes within the same cell makes it an interesting model system to investigate the fundamental processes of photosynthesis, respiration, biological N_2-fixation, and carbon sequestration [278, 279]. *Cyanothece* 51142 produces oxygen and stores photosynthetically fixed carbon in the form of glycogen granules during the day, and subsequently metabolizes stored carbon to produce excess energy and to create an O_2-limited intracellular environment [280, 281]. Respiratory electron transport scavenges oxygen to establish anaerobic intracellular conditions necessary for N_2-fixation. Thus cyanobacteria are known to perform substantially different metabolic processes during the light-dark periods. The diversity of metabolic pathways allows them to succeed in a wide variety of environments and provide a wealth of targets for metabolic engineering of energy-rich biomolecules. These diverse metabolic processes are governed not only by the expression and relative abundances of proteins but also by their association, localization, modifications as well as the spatial and temporal distribution of functionally active protein complexes. Protein oligomerization is a central feature of many cellular control mechanisms, and the changes in metabolic activities of these microbes between the light and dark periods must originate, in part, from the assembly and disassembly of protein complexes and cellular structure along the cycle. A thorough understanding of the biology of cyanobacteria requires in-depth knowledge of the composition and dynamics of multi-protein complexes, an area that has not been thoroughly investigated.

---

[4] This chapter has been previously published 271. Aryal, U.K., et al., *Proteomic analysis of protein complexes in photosynthetic cyanobacteria Cyanothece ATCC 51142.* Journal of Proteome Research, 2018. **Submitted**.

Cyanothece 51142 show distinct circadian rhythms of photosynthesis and N_2-fixation with peaks every 24 hours that are 12 hours out of phase from each other. The genome indicates a wealth of metabolic potential, in addition to very active photosynthesis and CO_2 uptake mechanisms [279]. Under N_2-fixing conditions, *Cyanothece* 51142 cells become filled with large granules between the photosynthetic membranes [280, 282]. These granules contain semi-amylopectin and are more similar to starch than to typical bacterial glycogen. The branching pattern of this starch-like material is quite different from glycogen, and *Cyanothece* 51142 has a series of branching and de-branching enzymes that might be involved. The composition and dynamics of assembly/disassembly enzyme complexes in *Cyanothece* 51142 for these glycogen granules are still outstanding. The sequencing of the genome [279] and the analysis of the transcriptome [283-285] and the proteome [278, 286-288] have uncovered many diurnal and circadian controlled genes and protein expressions. However, the oscillation of proteins were less pronounced compared to the transcripts [289], leaving us to speculate that the inventory of the genes and the proteins alone are not adequate to comprehend this organizational hierarchy. This led to the hypothesis that the molecular adaptation of *Cyanothece* 51142 occurs at a higher-level organization of protein complexes and protein-protein interactions (PPIs).

In recent years, there have been increasing efforts directed toward generating proteome-wide maps of PPIs [254]. The most commonly used high-throughput methods for the study of protein complexes are yeast-two-hybrid (Y2H) screens [290, 291] or affinity purification-mass spectrometry (AP-MS) [292-294]. The Y2H screens are expensive, time consuming, and incomplete [295]. The N- or C-terminal tagging in the AP-MS method can affect the expression and interaction of endogenous proteins [294, 296], and the application of an AP-MS method is also limited by the availability of tagged-constructs or antibodies.

An alternative size-based fractionation of native proteins via an SEC column combined with the high-resolution LC-MS/MS has been recently introduced [297]. SEC combined with LC-MS was applied to a non-N_2-fixing cyanobacterium *Synechococcus* elongates PCC 7942 [298], Arabidopsis cytosol [254, 299] and chloroplast [300] as well as human cell lysates [301, 302]. In this study, size fractionation of native proteins using Superdex 6 column was combined with label-free LC-MS profiling and bioinformatic analysis to identify subunits of protein complexes in *Cyanothece* 51142. Many proteins involved in key physiological processes including the capture of sunlight to produce energy and evolve O_2, the capture of N_2 to make fixed nitrogen, the capture of CO_2 for fixed carbon, the storage of large amounts of carbohydrates that represent potential energy, and ridding the cytoplasm of toxic oxygen, were identified as large protein complexes. The quality of the LC-MS profiling and complex prediction was evaluated by comparing two independent biological experiments in parallel, and by the identification of previously characterized protein complexes.

5.2 Methods

5.2.1 2.1. Cell growth and protein extraction

Cyanothece 51142 cells were maintained as previously described [277] in ASP2 medium with $NaNO_3$ at 30 °C and continuous illumination of white light at 50 μmol of photons m^{-2} s^{-1}. Cultures for this study were also grown in the same growth medium by inoculating 1/10 volume of the stock cell cultures and maintained at 30 °C under 12-h light/dark cycle for 7 days before harvesting at 6h into the light period. Cells were exposed to 50 μmol of photons m^{-2} s^{-1}white light during the light period. Cells were harvested by centrifugation at 14,000 rpm for 10 min at 4°C. Pelleted cells were gently washed 2× with ice-cold cell lysis buffer (20 mM Tris-HCl, pH 7.5, 5% glycerol, 50 mM KOAc, 2 mM $Mg(OAc)_2$, 1 mM EDTA, 1 mM EDTA, 0.5 mM DTT) followed by resuspension in 1 ml of the ice-cold lysis buffer. Cells were broken using a Precellys ® 24 Bead Mill Homogenizer (Bertin) at 6500 rpm for 3 cycles, each cycle lasting for 30s. Cell lysate was centrifuged at 14000 rpm for 20 min at 4 °C, and proteins in the supernatant were separated using size exclusion chromatography (SEC). Protein concentration was measured using a bicinchoninic acid (BCA) assay (Pierce Chemical Co., Rockford, IL) before being separating in the SEC column.

5.2.2 Size exclusion chromatography

The soluble fraction (0.5 ml, ~1 mg) was separated on a Superdex 200 10/300 GL column (GE Healthcare) using an ÄKTA FPLC system (Amersham Biosciences). Elution from the SEC column was performed with 20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 10 mM $MgCl_2$, and 5% glycerol at a flow rate of 0.2 ml/min, and absorbance was monitored at 280 nm. Two biological replicates were processed identically. The column was calibrated using protein standards (MWGF1000, Sigma-Aldrich, St. Louis, MO) covering a mass range from 29 kDa to 669 kDa.

The void volume was measured with blue dextran. SEC separation was performed at 6°C, and 20 SEC fractions of 500 μL were collected for mass spectrometry analysis as described below.

5.2.3 Sample preparation for LC-MS analysis

Sample preparation was carried out as described previously [286]. Briefly, proteins were denatured by adding 50 μl of 8 M urea for 1 h at room temperature, and the concentration in each fraction was determined by BCA assay. Proteins were reduced with 10 mM dithiothreitol (DTT), then cysteines were alkylated with IAA. Digestion was performed at 37˚C overnight using mass spec grade trypsin and Lys-C mix from Promega at a 1:25 (w/w) enzyme-to-substrate ratio. The digested peptides were desalted using Pierce C18 spin columns (Pierce Biotechnology, Rockford, IL). Peptides were eluted using 80% acetonitrile (ACN) containing 0.1% Formic Acid (FA) and dried in vacuum concentrator at room temperature. Dried clean peptides were re-suspended in 80 μl of the buffer containing 97% purified water, 3% ACN and 0.1% FA. Peptides were loaded to the LC column by equal volume (5 μl), not by equal amount or concentration. In an 80 μl solution, peptide concentration of the fraction that contained the highest protein amount (in this case fraction 21 in both the biological replicate) was 0.2 μg/μl.

5.2.4 LC-MS/MS data acquisition

Samples were analyzed by reverse-phase HPLC-ESI-MS/MS using the Dionex UltiMate 3000 RSLC nano System coupled to the Q-Exactive^{TM} High Field (HF) Hybrid Quadrupole Orbitrap^{TM} Mass Spectrometer (Thermo Scientific, Waltham, MA) and a Nano- electrospray Flex^{TM} ion source (Thermo Scientific). Purified peptides were loaded onto a trap column (300 μm ID × 5 mm) packed with 5 μm 100 Å PepMap C18 medium and washed using a flow rate of 5 μl/minute with 98% purified water/2% ACN /0.01% FA. The trap column was then switched in-line with the analytical column after 5 minutes. Peptides were separated using a reverse phase Acclaim^{TM} PepMap^{TM} RSLC C18 (75 μm x 15 cm) analytical column using a 120-min method at a flow rate of 300 nl/minute. The analytical column was packed with 2 μm 100 Å PepMap C18 medium (Thermo Scientific). Mobile phase A consisted of 0.01% FA in water and a mobile phase B consisted of 0.01 % FA in 80% ACN. The linear gradient started at 5% B and reached 30% B in 80 minutes, 45% B in 91 minutes, and 100% B in 93 minutes. The column was held at 100% B for the next 5 minutes before being brought back to 5% B and held for 20 minutes to equilibrate the column. Sample was injected into the QE HF through the Nanospray Flex™ Ion Source fitted with a stainless steel emission tip from Thermo Scientific. Column temperature was maintained at 35˚C. MS data was acquired with a Top20 data-dependent MS/MS scan method. The full MS spectra was collected over 300-1,650 *m/z* range with a maximum injection time of 100 milliseconds, a resolution of 120,000 at 200 *m/z*, and AGC target of 1 $\times 10^6$. Fragmentation of precursor ions was performed by high-energy C-trap dissociation (HCD) with the normalized collision energy of 27 eV. MS/MS scans were acquired at a resolution of 30,000 at m/z 200. The dynamic exclusion was set at 20 s to avoid repeated scanning of identical peptides. Instrument optimization and recalibration was carried out at the start of each batch run using the Pierce calibration solution. The sensitivity of the instrument was also monitored using an *E. coli* digest at the start of sample runs.

5.2.5 Data analysis

All LC-MS/MS data were analyzed using MaxQuant software (v. 1.5.3.28) [303-305] against the *Cyanothece* 51142 genome (http://img.jgi.doe.gov/cgi-bin/w/main.cgi) that contained 5,300 non-redundant protein sequences. MaxQuant includes common contaminants as a default. No external contaminants were added to the database. The minimal length of six amino acids was required in the database search. The database search was performed with the precursor mass tolerance set to 10 ppm and MS/MS fragment ions tolerance was set to 20 ppm. The database search was performed with the enzyme specificity for trypsin/Lys-C, allowing up to two missed cleavages. Oxidation of methionine (M) and phosphorylation of STY (pSTY) were defined as variable modifications, and carbamidomethylation of cysteine was defined as a fixed modification. MaxQuant search was performed as target-decoy, and the false discovery rate (FDR) of peptide spectral match (PSM) and protein identification was set at 0.01. After the search peptides without any identifiable peak (0 intensity) and with no MS/MS counts were removed from consideration. At the protein level, proteins with 0 intensity and with 1 MS/MS counts were also removed from consideration. The ‘unique plus razor peptides’ were used for peptide quantitation. Razor peptides are the non-unique peptides shared between the protein groups with the most other peptides. To increase the number of peptides that can be used for protein quantification and relative abundance profiling across SEC fractions, the “match between runs” function was enabled with a maximum retention time window of 1 min. This “match between runs” allows the transfer of peptide identification between fractions in the absence of peptide sequencing by MS/MS spectra, utilizing their accurate mass and aligned retention time [303]. The identified peptides and protein groups with their raw intensities were exported to Microsoft Access 2010 to perform subsequent analyses. The correlation coefficients between SEC fractions were calculated using Data Analysis and Extension Tool (DAnTE) [306].

5.2.6 Data normalization and clustering of protein profiles

In a protein elution profile, the peak is defined as the elution fraction with the largest abundance among all fractions in each SEC experiment. Since the SEC experiment was repeated twice independently, the two independent experiments should generate similar elution profiles for the same protein. To ensure the quality of the elution profiles, the difference of the index of peak fraction was checked between the two SEC experiments, and only proteins with a peak index shift within 2 fractions were selected for clustering analysis, which indicates the SEC experimental results are consistent. Since the experiments were performed independently, the elution profiles generated by the two independent experiments were normalized independently by dividing the corresponding maximum intensity among each experiment. The elution profiles of Bio1 and Bio2 were normalized separately by dividing the LFQ intensities by the maximum intensity among the twenty fractions. The normalized 20 fractions from Bio1 and 20 fractions from Bio2 were concatenated into 40 fraction, and clustered using the Euclidean distance measurement and the different combination of hierarchical methods such as average, complete, mcquitty and ward. For each clustering method, different number of clusters were applied by cutting the dendrogram tree at different distances to determine the optimum number of clusters. Clustering results were compared with some known protein complexes to determine the cluster quality and the optimal cluster numbers.

5.2.7 Sequence-based PPI prediction

For sequence-based pair-wise PPI prediction [62], the amino acid sequences of *Cyanothece* 51142 proteins were downloaded from CyanoBase (http://genome.annotation.jp/CyanoBase) [307]. The experimental results contained GeneBank protein IDs starting with “gi” and were converted into RefSeq ID following instructions on the GenBank webpage and the UniProt database [308]. For predicting PPI based on sequence information, seven physiochemical properties were considered including hydrophobicity, hydrophilicity, volumes of side chains of amino acids, polarity, polarizability, solvent-accessible surface area (SASA), and net charge index (NCI) of side chains of amino acids. The protein sequences were then represented as periodicity of each physicochemical property (Equation 5.1):

$$\mathrm{AC}(\mathrm{lag},\mathrm{j}) = \frac{\sum_{\mathrm{i}=1}^{\mathrm{L}-\mathrm{lag}}\left(\mathrm{P}_{\mathrm{i},\mathrm{j}}-\frac{1}{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{L}}\mathrm{P}_{\mathrm{i},\mathrm{j}}\right)\times\left(\mathrm{P}_{(\mathrm{i}+\mathrm{lag}),\mathrm{j}}-\frac{1}{\mathrm{L}}\sum_{\mathrm{i}=1}^{\mathrm{L}}\mathrm{P}_{\mathrm{ij}}\right)}{\mathrm{L}-\mathrm{lag}} \qquad \text{(Equation 5.1)}$$

where *lag* is the distance between covariant residues to consider, which ranges from 1 to 30, *j* is the *j*-th physiochemical descriptor, *i* is the position in the sequence, and *L* is the length of sequence. Each protein pair was transformed into 420 dimensional vectors [70]. Then support vector machine (SVM) (the software libsvm 2.84 http://www.csie.ntu.edu.tw/~cjlin/libsvm/) [309] was used to predict PPIs. SVM is a supervised learning method which uses the kernel function to transform the nonlinear features into linearly separable data. A total of 4908 experimentally verified non-redundant protein interactions in Arabidopsis were used as a training dataset for the SVM. A radial basis function (RBF) was chosen as the kernel function with regularization parameters C and kernel parameter γ optimized as 32 and 0.5 because of the highest cross validation accuracy.

5.2.8 Quantify gene co-expression

Next, the current protein complex profiles and the computationally predicted PPI were compared with previously published gene [285] and protein expression [287] data sets. The mRNA gene expression data set by *Stockel et al.* [285] includes 1443 genes of *Cyanothece* 51124. 572 out of 1443 proteins overlap with our experiment. The protein expression data set by *Aryal et al.* [287] was collected under day and night period and includes 976 proteins. 561 out of 976 proteins overlap with our experiment. Co-expression level of protein pairs were evaluated by the Pearson's correlation coefficient (PCC) (Equation 5.2):

$$\mathrm{PCC} = \frac{\mathrm{cov}(\mathrm{A},\mathrm{B})}{\sigma_{\mathrm{A}}\sigma_{\mathrm{B}}} \quad , \qquad \text{(Equation 5.2)}$$

where $\mathrm{cov}(\mathrm{A},\mathrm{B})$ is a covariance of protein A and B, σ_{A} and σ_{B} are the standard deviation of protein A and B, respectively. In Table S3, the PCC of the day, night expression and the average of the two (overall PCC) was provided. For the protein co-expression data [287], the mutual Rank (MR) of co-expression strength was also computed as following:

$$\mathrm{MR} = \sqrt{\mathrm{R}_{\mathrm{A}\to\mathrm{B}} * \mathrm{R}_{\mathrm{B}\to\mathrm{A}}} \quad , \qquad \text{(Equation 5.3)}$$

which is the geometric mean of the correlation rank of gene A to gene B ($R_{A\rightarrow B}$) and of gene B to gene A ($R_{B\rightarrow A}$) (Eq. 3). A small MR correlates to a stronger co-expression of the gene. MR is useful in evaluating co-expression when some genes weakly co-expressed with all other genes and have spurious PCC values. In Table S3, PCC and MR for the protein expression data were provided [287].

5.3 Results

5.3.1 SEC fractionation and LC-MS reproducibility

Native *Cyanothece* 51142 proteins were separated into 20 SEC fractions (Figure 5.1, Supplemental Figure S5.1). The void volume was determined based on the elution peak of blue dextran (Supplemental Figure S5.1A). The molecular weight of proteins eluting in each SEC fraction was determined based on calibration curve (Supplemental Figure S5.1B). Two independent SEC fractionations were performed (Supplemental Figures S5.1C and S5.1D). Accuracy of label-free protein quantitation is limited if peptide intensity measurement is inconsistent. The reproducibility of peptide signal intensity and peptide retention time on the LC column was tested by analyzing three technical replicates from one of the fractions (F9 of Bio2). Of the total 1170 peptides and 335 proteins, 971 (83%) peptides and 298 (89%) proteins overlapped in all the 3 technical replicates (Supplemental Figures S5.2A and S5.2B), which is a good indication of LC-MS reproducibility for protein identification. The average coefficient of variation (CV) of MS1 intensity was ~15.1% and the CV of the peptide retention time was <1.0% (Supplemental Figures S2C and S2D), which also indicated good reproducibility for intensity-based label free quantitation.

5.3.2 Global analysis of the expressed proteome

In total, 1,567 proteins in Bio1 and 1,436 proteins in Bio2 were identified, of which 1386 proteins (88% of Bio1 and 96% of Bio2) were common (Figure 5.2A). Pearson's correlation coefficient of 1386 protein intensities as a function of SEC fraction numbers (Figure 5.2B) showed the highest correlation coefficients along the diagonal, which indicated that protein elution peaks were reproducible between the biological replicates. However, the high correlation of signal intensities expanded to several adjacent fractions for high molecular weight protein complexes. This is because molecular weight of these proteins were beyond the size limit of the SEC column.

The box plots in Figure 5.2C further confirmed that quantitation were consistent across column fractions. Reproducibility of protein elution peaks in SEC column between the replicates is important to predict protein complexes based on their apparent mass (size). To check the reproducibility, the shift in the elution peak fraction (global maximum) of all the identified proteins between Bio1 and Bio2 was compared (Figure 5.2D). 55% of the proteins were identified without any peak shift (0 fraction shift) and >90% of the proteins were identified within 0-2 fraction shift, confirming good SEC reproducibility.

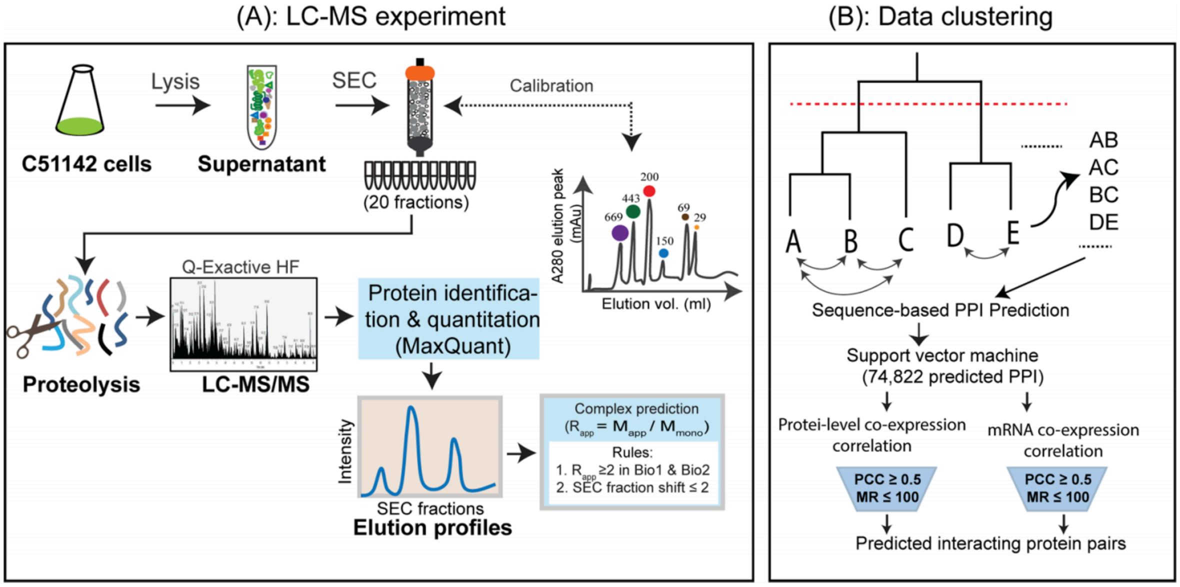


Figure 5.1 Experimental workflow. (A) Proteins extracted under native condition were fractionated by SEC, and analyzed by Q Exactive Orbitrap HF mass spectrometer. Data were analyzed using MaxQuant [303-305] for protein identification and label free MS1 quantitation. Peak elution fraction of each identified protein, M_{app}, and R_{app} were determined as described previously [254]. M_{app}, apparent molecular mass; M_{mono}, predicted molecular mass of monomer. R_{app}, the ratio of the M_{app} to the M_{mono} (M_{app} / M_{mono}). Proteins with an $R_{app} \geq 2$ in both the replicate were considered to be in a complex.

5.3.3 Hierarchical clustering of protein elution profiles

Proteins with a similar elution profile were clustered and further subjected to bioinformatics predictions of PPI. Proteins interacting within complexes should display similar SEC elution profiles and belong to the same cluster. The results of different clustering methods (see method for details) were compared using several known protein complexes such as PSI, PSII, light harvesting complex, ribosomal proteins and others, and the method which assigned most of the known protein complex subunits within the same cluster were selected. Since these known protein complexes stably exist in the *Cyanothece* 51142, the clustering results did not differ much with different combinations of clustering methods. Because the computational method was used to further filter out the false interacting pairs with similar elution profiles, the smaller number of clusters with more proteins within each cluster was adopted in order to generate more protein pairs subject to prediction within the same cluster. The average linkage hierarchical clustering method with 30 clusters was used. The heat map of the Euclidean distance of elution profiles was plotted in the Figure 5.6A. The heat map of elution throughout the SEC fractions shows that a significant number of proteins peaked at the high molecular weight fractions, which indicates that many proteins are migrating through the SEC column as complexes. To roughly estimate the proportion of proteins that migrate as stable complexes, the peak elution fraction (global max) of each protein was determined and used that global peak fraction to estimate the size or apparent (native) molecular weight (M_{app}). Many proteins eluted in high mass fractions suggesting that they remained intact during SEC separation.

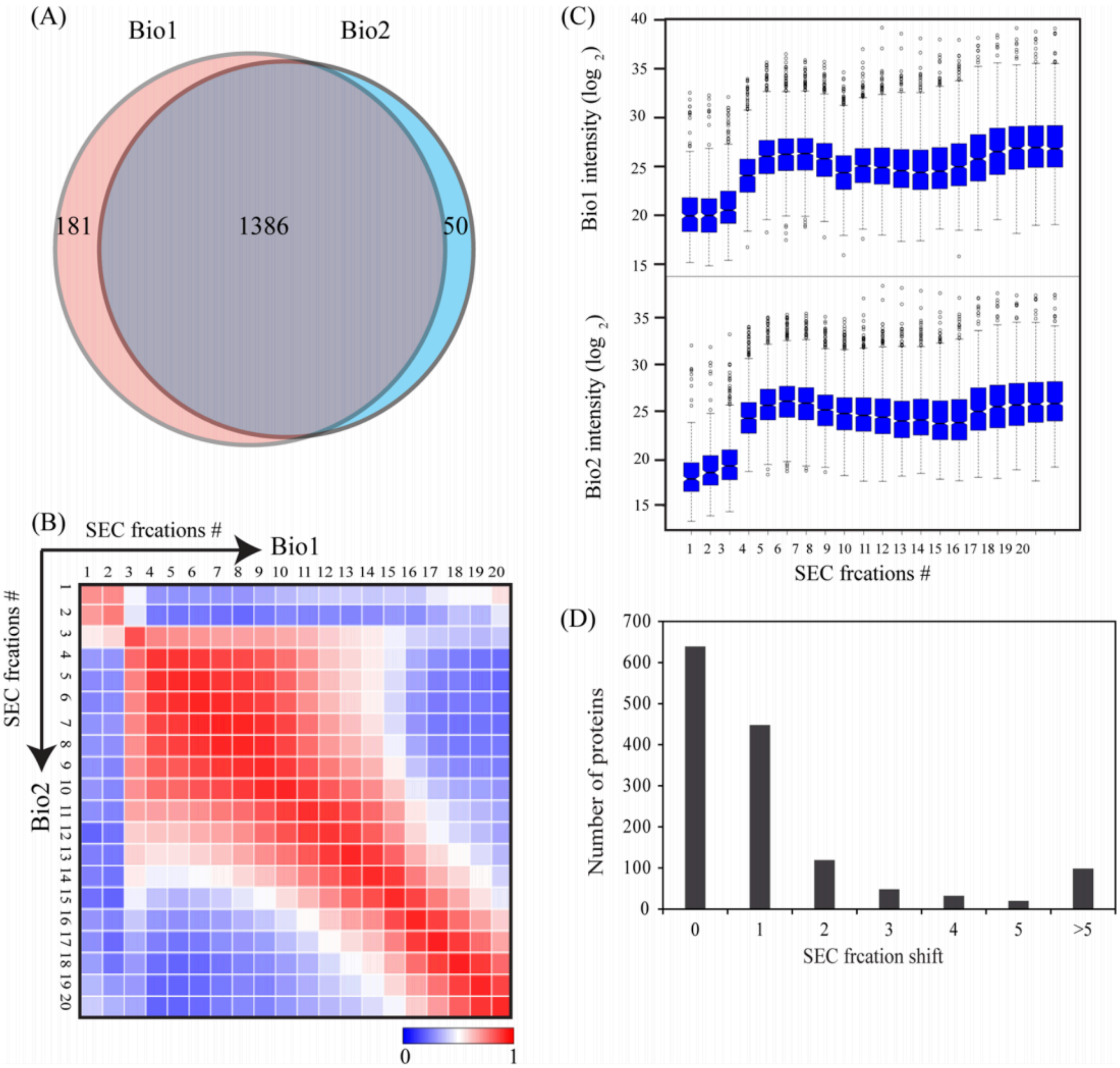


Figure 5.2 LC-MS reproducibility. (A) Venn diagram showing the overlap of proteins identified between two biological replicates. (B) Heat map showing the Pearson's correlation coefficients (PCC) of protein abundances (MS1 intensity) across SEC fractions. The correlations coefficients were calculated using Data Analysis and Extension Tool (DAnTE) [306]. (C) Box plot showing the median distribution of protein intensities. (D) Shift in peak elution fraction of proteins in two SEC separations. ~90% proteins were identified within 0-2 fractions shift indicating good SEC reproducibility.

5.3.4 Determination of protein complexes

Figure 5.3B shows the distribution of the monomer (M_{mono}) and the M_{app} of proteins. The M_{mono} is concentrated in the lower molecular weight ranges and M_{app} is concentrated in high molecular weight ranges. Previously, R_{app} (apparent ratio = M_{app} divided by M_{mono}) have been used [254, 310, 311] to define a protein complex as those having an R_{app} value of 2 or higher in both the biological replicates. Despite several limitations, R_{app} is a useful metric to globally predict putative protein complexes. Figure 5.6C shows the R*app* distribution of proteins in the two biological replicates. The circles along the solid line represent proteins eluting in exactly the same fraction, thus the same R_{app} values, in both the biological replicates (0 fraction shift in elution peak). ~55% of the proteins fell in this category. Circles along the dotted lines indicate proteins with 1 fraction shift, and ~30% of the proteins had 1 fraction shift in their elution peaks. Our R_{app} predictions agreed well with the oligomerization state of several known protein complexes. For example, the M_{app} of PSI complex subunits ranged from ~376-550 kDa (Supplemental Table S2, row 565-578) in agreement with the previous report [312]. Enolase peaked in fraction 11 with an M_{app} of ~105 kDa, close to the known dimeric structure [313]. Enolase also peaked in a fraction with M_{app} of ~105 kDa in our previous analysis using Arabidopsis [254]. Another glycolytic enzyme, phosphoenolpyruvate carboxylase (Ppc; cce_3822), was identified with R_{app} 4.6 in both the replicates, close to the known tetrameric structure of this enzyme [314]. Arabidopsis PEPC (PEPC1 and PEPC2) were also detected with R_{app} of ~4 in our previous study [254, 299]. Using R_{app} values, 64% (946 out of 1386) of the proteins detected in both the biological replicates were predicted as complexes. The protein complexes were functionally diverse including those involved in translation, carbohydrate metabolism, photosynthesis, respiration, ion transport, folding, and ATP and metal ion binding (Figures 5.3A and 5.3B). Despite our mild lysis buffer, the protein list included both cytosolic and membrane proteins (Figure 5.3C). Our membrane protein list included many cytoplasmic and thylakoid membrane proteins, and both cytoplasmic (hydrophilic) and membrane (hydrophobic) domain proteins. However, cytoplasmic domain proteins were detected with higher relative abundances than membrane domain proteins indicating that they are more accessible for solubilization during extraction. It is important to mention here that PsaA, PsaB, PsaC, PsbB, PsbC and PsbA2 were detected, and all are known to be hydrophobic.

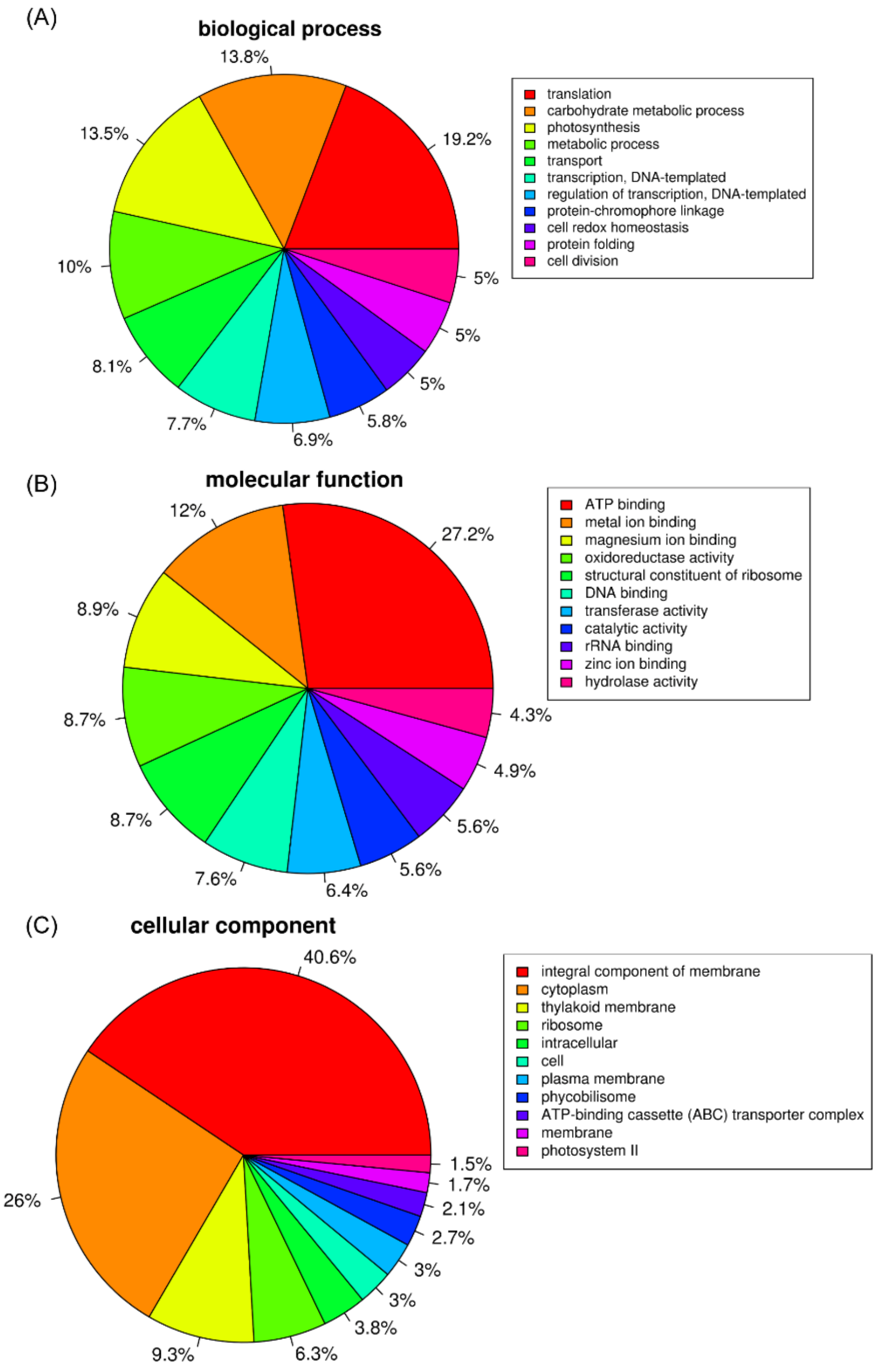


Figure 5.3 Pie charts showing distribution of proteins into diverse biological processes (A), molecular functions (B) and cellular components (C).

Protein sizes were also diverse ranging from ~20 kDa to ~800 kDa. About 50% of those putative complexes eluted either in the void or high molecular weight (> 600 kDa) fractions, including many 30S and 50S ribosomal proteins, PSI and PSII proteins (Figure 5.4), phycobilisomes, thioredoxins, ferroredoxins, glutaredoxins, NDH-1 complex (Figure 5.6), elongation factors and many unknown or hypothetical proteins. One-third of the proteins eluting in the void were unknown or hypothetical proteins. Many of these unknown proteins showed highly correlated elution profiles with other known protein complexes and also were predicted as interacting pairs by computational method. For example, unknown protein cce_4744 showed correlated elution profile with cytochrome f (PetC1; cce_2958) (Figure 5.5A); another unknown protein cce_0494 co-eluted with PSII reaction center protein PsbB (cce_1837) and PsbC (cce_0659) (Figure 5.5B). In addition, uncharacterized proteins cce_1749, cce_3678, and cce_3430 have highly correlated elution profiles with the protein involved in disulfide bond formation (cce_1972) (Figure 5.5C), and their protein-level expression are highly correlated (Supplemental Table S5.3). These and several other evidences suggest that many novel and apparently large protein complexes that are currently characterized as unknown have been uncovered.

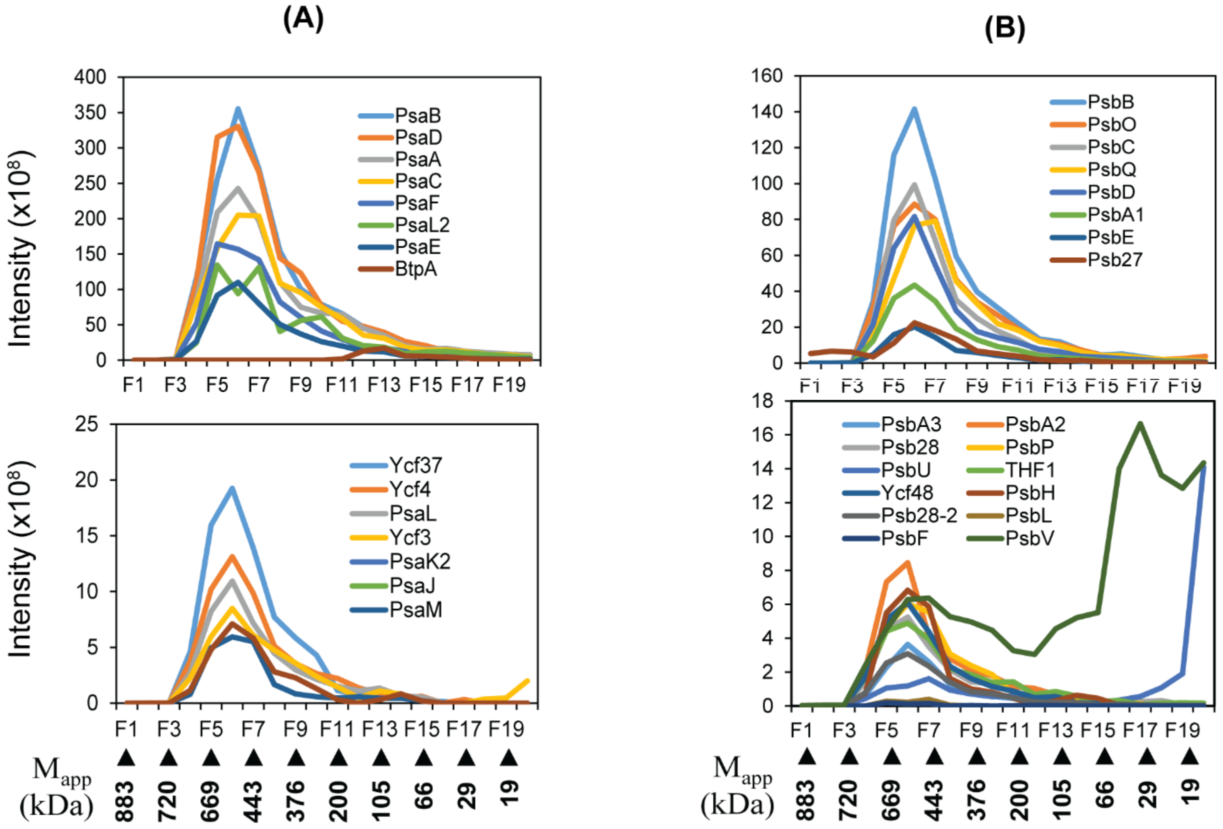


Figure 5.4 SEC elution profiles of PSI (A) and PSII (B) polypeptides. PSI and PSII polypeptides eluted in high molecular weight (669 kDa) fraction.

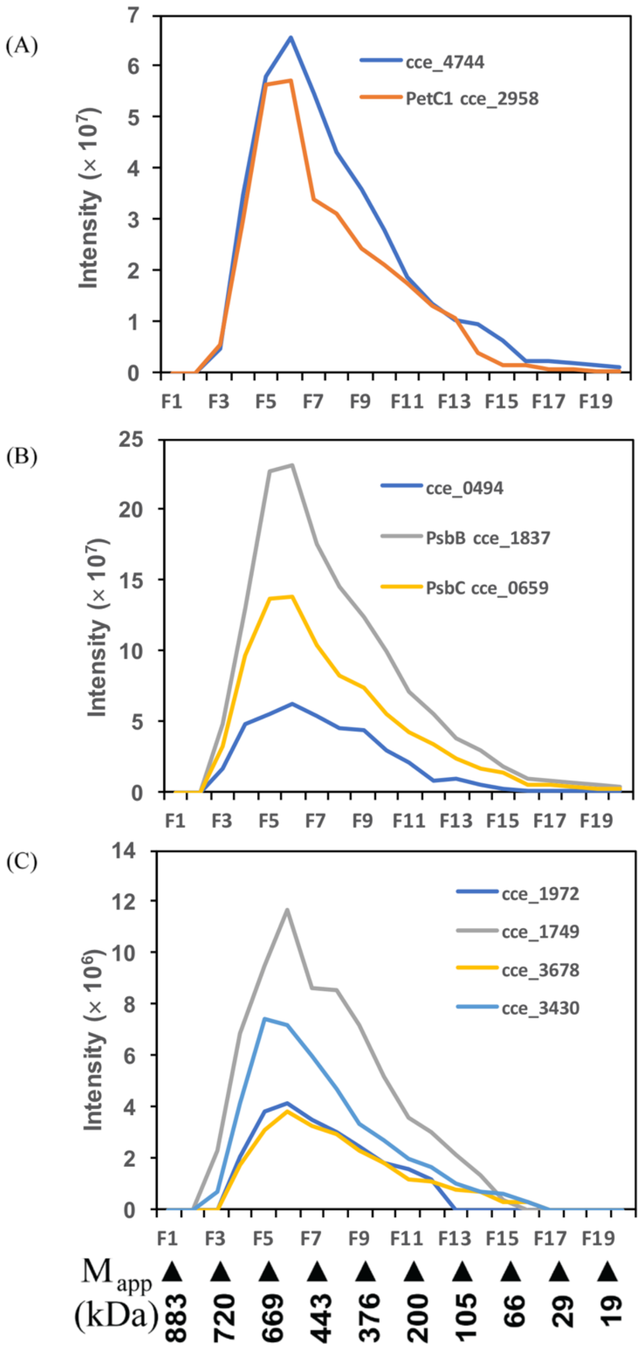


Figure 5.5 Correlated elution of unknown proteins with known protein complexes. (A) Unknown protein cce_4744 eluting with PetC1, (B) cce_0494 eluting with PsbB and PsbC, and (C) protein involved in disulfide bind formation cce_1972 eluting with unknown protein cce_1749, cce_3678, and cce_3430.

Of the 1386 proteins, ~400 proteins were annotated as unknown and ~70 proteins were classified as hypothetical proteins. Two-third of these proteins (~300) have $R_{app} \geq 2$ in both the biological replicates (Supplemental Table S2). This suggests that many protein complexes have been detected, whose function is currently unknown, and highlights the significant challenge ahead for functional characterization of these unknown proteins, as in general, >40% of the proteome in prokaryotes and >50% in eukaryotes are not characterized [315].

Our experimental system also detected proteins that are partitioned between the cytosol and the cytoplasmic and/or thylakoid membrane; indeed, there are a number of proteins with known membrane localization that were detected as apparent subunits of large complexes. Most of those detected membrane proteins are abundant proteins such as light harvesting phycobilisomes proteins (Figures 5.4A and 5.4B), subunits of NDH-1 complex (Figure 5.9), PSI and PSII complexes (Supplemental Figure S5.4)), and the ATP synthases (Supplemental Figure S6). It appears that subunits of these complexes are easily accessible for solublization during cell lysis due to cytoplasmic domain localization.

Key enzymes of glycolysis (GlgP1, Pgi1, Pgi2, PfkA1, Fda, Gap, Pgk, Eno1, Eno2, Ppc), TCA cycle (GltA, AcnB, SucC, SdhB, FumC), pentose phosphate (PP) pathways (Zwf, Gnd, TalA, Rpe, Pkt), and amino acid biosynthesis (AroQ, IlvN, TrpD, AroK, CysK, LeuB) (Supplemental Table S2) eluted as stable complexes. Proteins involved in glycogen synthesis, GlgA1 (cce_3396) and GlgA2 (cce_0890) were identified as large protein complexes with M_{app} of 466 kDa and R_{app} >5 in both the replicates (Figure 5.6C). Of the three circadian clock (Kai) proteins, KaiB (cce_0423) and KaiC (cce_0422; cce_4716) were identified, and eluted with multiple but consistent elution peaks in Bio1 and Bio2 (Supplemental Table S2, rows 236-238). The first elution peak corresponding to fraction 5 represents approximately 466 kDa in both the replicates (Figure 5.7F). In cyanobacteria, KaiA and KaiB work together to modulate the activity of KaiC in a phosphorylation dependent manner [316]. The link between metabolic activity and the circadian behavior has previously been reported [316], and this link might be important in *Cyanothece* 51142 as these microbes are typically dependent on photosynthesis as an energy source.

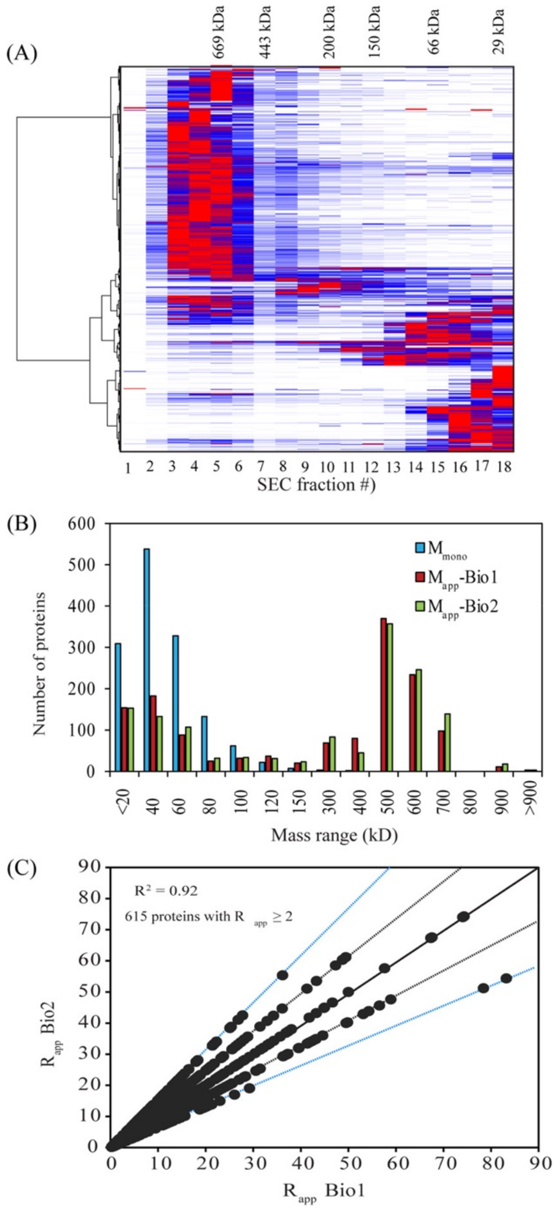


Figure 5.6 Determination of protein oligomerization states. (A) Hierarchical clustering of protein elution profiles. Proteins were clustered using Euclidean distance and average linkage hierarchical clustering method. In this plot, each row represents a protein and each column represents the index of protein elution fraction. Numbers on the top show molecular masses of protein standards, and the peak elution fraction for each of the standard was used to determine the M_{app} of proteins. (B), Histogram showing the distribution of the monomeric (blue) and experimentally determined apparent masses (green and red) of proteins that were identified in both the biological replicates. (C), Scatter plots showing R_{app} distribution of proteins between the two biological replicates. Each circle represents R_{app} values for Bio1 and Bio 2. Circles along the black solid line represent proteins without any fraction shift in elution peak (same R_{app} values) in both the replicates. Circles along the black dotted lines represent proteins with 1 fraction shift and circles along the blue dotted lines represent proteins with 2 fraction shifts between the replicates. Bio1; biological replicate 1, Bio2; biological replicate 2.

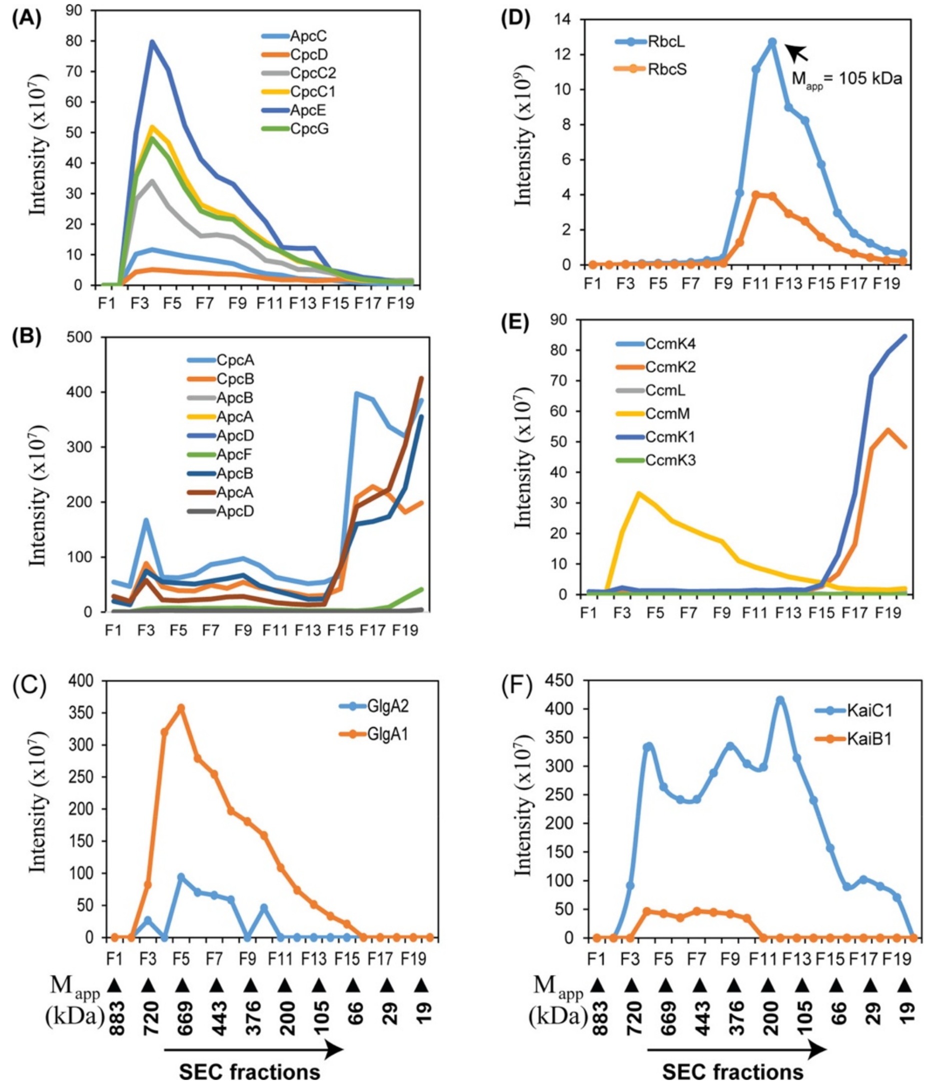


Figure 5.7 Elution profiles of phycobilisomes (PBS) and other complexes. (A, B), Elution profiles of phycocyanin (Cpc) and allophycocyanin (Apc) subunits. Elution profiles varied among the individual polypeptide. (C), Elution profiles of Rubisco large (RbcL) and small (RbcS) subunits. Both RbcL and RbcS peaked at fraction 12 with calculated M$_{app}$ of 105 kDa. (D), Elution profiles of CO$_2$ concentrating mechanism (Ccm) proteins. CcmM showed major elution peak as a complex while others showed major peaks as monomers.

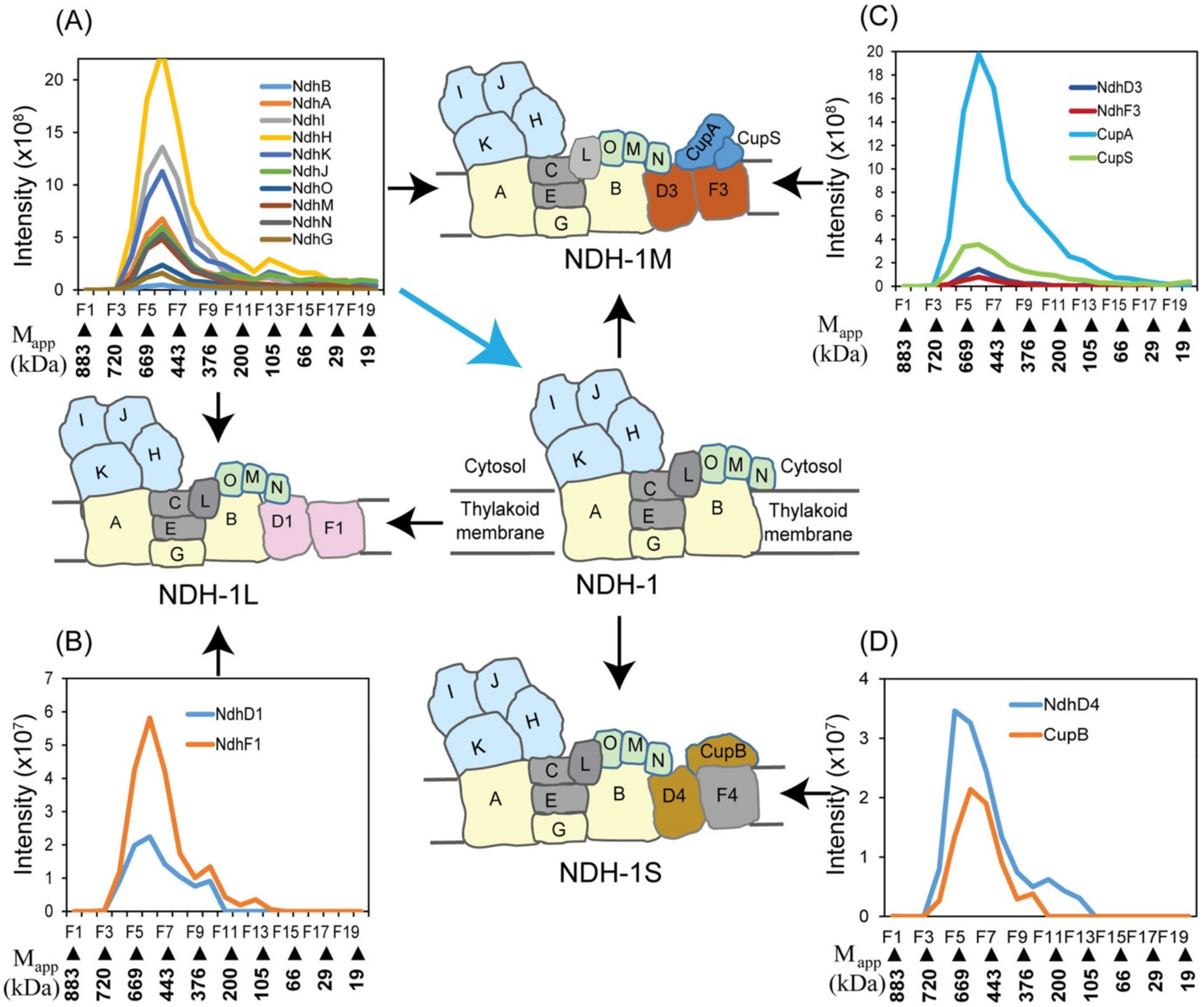


Figure 5.8 NDH-1 complex. (A-D), Elution profiles and structure of multiple forms of NDH-1 complex subunits. All the subunits eluted in high molecular weight (669 kDa) fraction. The existence of NDH-1L (respiratory), and NDH-1MS and NDH-1MS' (CO_2 uptake) forms of NDH-1 complexes were determined by comparing SEC co-elution profiles and known functional and structural multiplicity in the literature [317, 318]. Hydrophilic domain subunits showed higher abundance than the membrane domain subunits. Both hydrophilic (I, J, K, H) and hydrophobic domain subunits (A, B, C, D1, F1, D3, F3, D4) as well as Oxygenic-Photosynthesis-Specific (OPS)-domain subunits (O, M, N) were identified. Results show the existence of functional multiplicity of NDH-1 complexes in Cyanothece 51142 cells that are responsible for a variety of functions including respiration, cyclic electron flow and CO_2 uptake.

5.3.5 Computational protein-protein interaction prediction

Pair-wise sequence-based PPI prediction [121] identified 74,822 putative PPI pairs among all the 1386 proteins, of which 561 proteins have been found in the previously published protein expression data by *Aryal et al.*[287], and 572 genes overlap with mRNA expression data by *Stockel et al.*[285]. To further select predicted PPI pairs with high confidence, I referred to these protein-level and mRNA-level co-expression information. In Table S5.3, predicted PPI pairs among the 561 proteins are selected that have a protein co-expression correlation [287] above 0 (Supplemental Table S5.3, column F) or mutual rank below 100 (Supplemental Table S5.3, column G), and with at least one of mRNA co-expression correlation [285] above 0 (Supplemental Table S5.3, column C)). There were in total 2,461 such protein pairs. These proteins are plotted in Figure 5.10 with the Euclidean distance of protein elution profiles and Pearson's correlation coefficient of the mRNA-level co-expression information. If protein pairs are both annotated, the number of common GO terms of the protein pairs is indicated in a color scheme with a darker color for stronger function similarity. The figure shows such pairs with functional similarity mainly locate at the top left of the plot, which indicates that they have a higher co-expression correlation and similar elution profiles with each other. Thus the plot implies that the similarity of elution profile and high expression correlation indeed capture physically interacting protein pairs.

The predicted PPI list (Supplemental Table S5.3, Figure 5.9) includes pairs of obviously similar function such as PSI and PSII proteins, ribosomal proteins, cytochrome b6f complex, ATP synthases, NADPH- related proteins, chaperones, amino acid synthesis and carbohydrate metabolism. Figure 5.9 visualizes the interaction network of the pairs using Cytoscape [262]. For example, PBS complex subunits ApcA and ApcB, which co-elute together (Figure 5.4A), have a very high co-expression correlation at both protein and mRNA levels and share four GO terms (Supplemental Table S3), and were predicted as interacting proteins (Figure 5.9A). CcmK1 and CcmK2 were also predicted as interacting pairs with a very high protein level and mRNA level co-expression (Figure 5.9B), as well as AtpE and AtpB1 proteins (Supplemental Table S5.3) and PSI and PSII polypeptides (Supplemental Table S5.3, Figures 5.9C and 5.9D). PsaB and PsaA (MR=2.83, PCC=0.98) and CpcG, ApcE and CpcC2 (Figure 5.4A) were predicted as interacting pairs with strong co-expression (Supplemental Table S5.3). The NDH-1 complex subunits NdhO and NdhM (Figure 5.8) were predicted as interacting pairs with high protein co-expression score (MR: 39.87, PCC: 0.81) and 3 common GO terms (Supplemental Table S3).

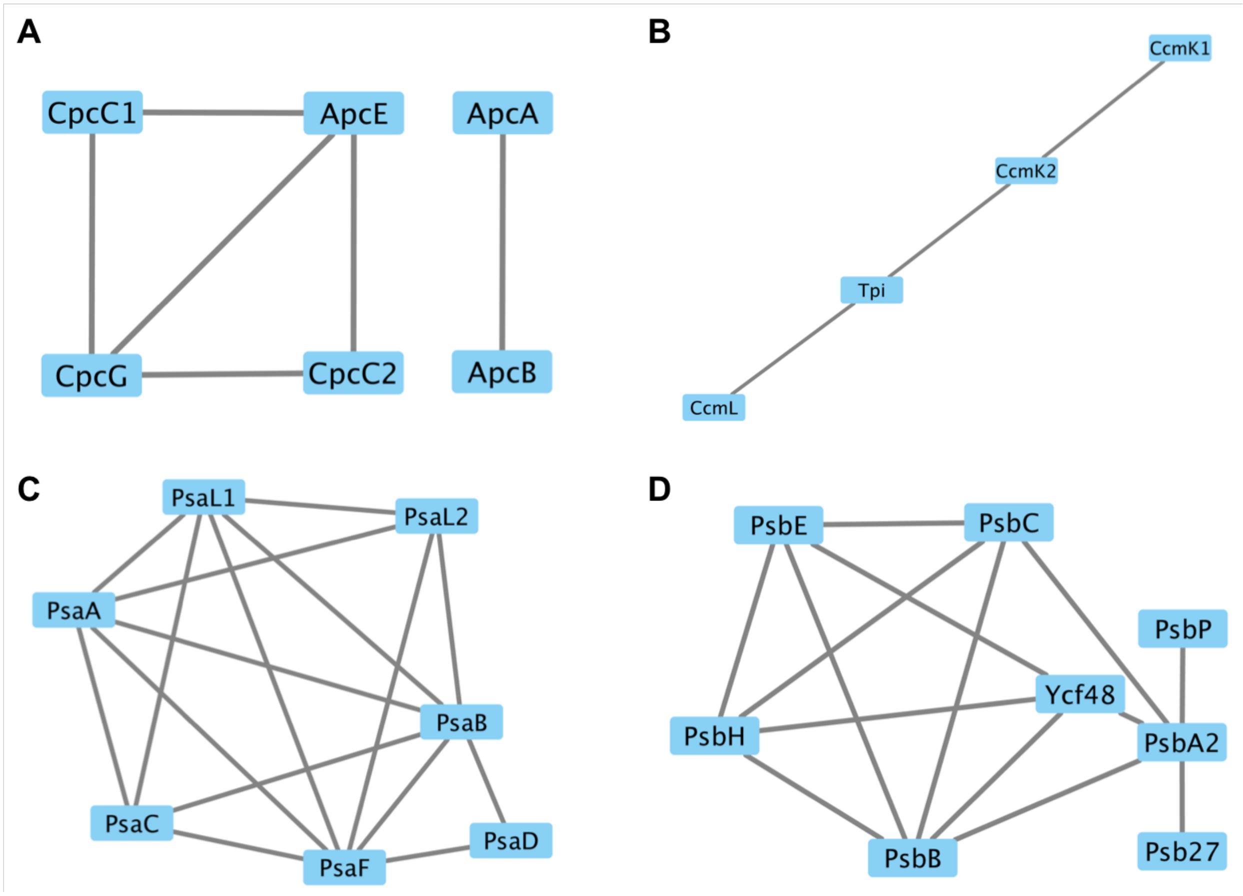


Figure 5.9 Sequence-based prediction of protein-protein interaction (PPI) network. The edge represents the predicted interaction between two proteins. The predicted network is drawn by Cytoscape.

Additionally, I have referred to the STRING score for the predicted protein pairs in the Supplemental Table S3. STRING is a database which provides various data that indicate functional and physical interactions of protein pairs in over 2,000 organisms [55]. The plausibility of interactions are indicated by a score, which ranges from 0 to 1000, with 1000 for the most confident interaction. Thus, STRING provides further additional support of identified interacting protein pairs. Among the predicted PPIs with STRING combined score above 900, four interesting examples are discussed.

Putative homologs of Glucose-1-phosphate adenylytransferase (GlgC2; cce_2658) and phosphoglucomutase (cce_0770) are both involved in the glycogen biosynthesis pathway in ten other organisms. Since proteins involved in the same pathway have a higher probability to interact, it is highly possible that these two proteins interact. Another example is the type IV pilus assembly protein (PilM; cce_1578) and hypothetical protein (cce_1579). Their genes are coded in the vicinity on the *Cyanothece* genome within only 4 bp intergenic distance, and they also co-occur across multiple organisms. Studies of protein interactions show that genes encoding interacting proteins are kept close to each other on the genome [166, 167], and thus these two proteins have high probability of interacting. The third example is uroprophyrinogen decarboxylase (HemE; cce_2966) and corproporphyrinogen III oxidase (HemF; cce_3201). They are both involved in the heme biosynthesis pathway and porphyrin chlorophyll metabolism pathway not only in *Cyanothece* 51142 but also in other 4 *Cyanothece* strains. Also, their putative homologous proteins are found to have correlated expression patterns in other organism. The fourth example is the pyrroline-5-carboxylate reductase (ProC; cce_2615) and bifunctional proline dehydrogenase (PutA; cce_1595). They are both involved in the arginine and proline metabolism pathway. Furthermore, it has been shown in *Thermus thermophilus* HB27 that PutA catalyzes the conversion of proline to prroline-5-carboxylate, which is the target of ProC [319]. Therefore, it is highly possible that these two homologous proteins also interact in *Cyanothece* 51142. Overall, many large and apparently novel protein complexes were identified in *Cyanothece* 51142 and further discuss several more complexes below, which are highlighted in yellow in Supplemental Table S5.3.

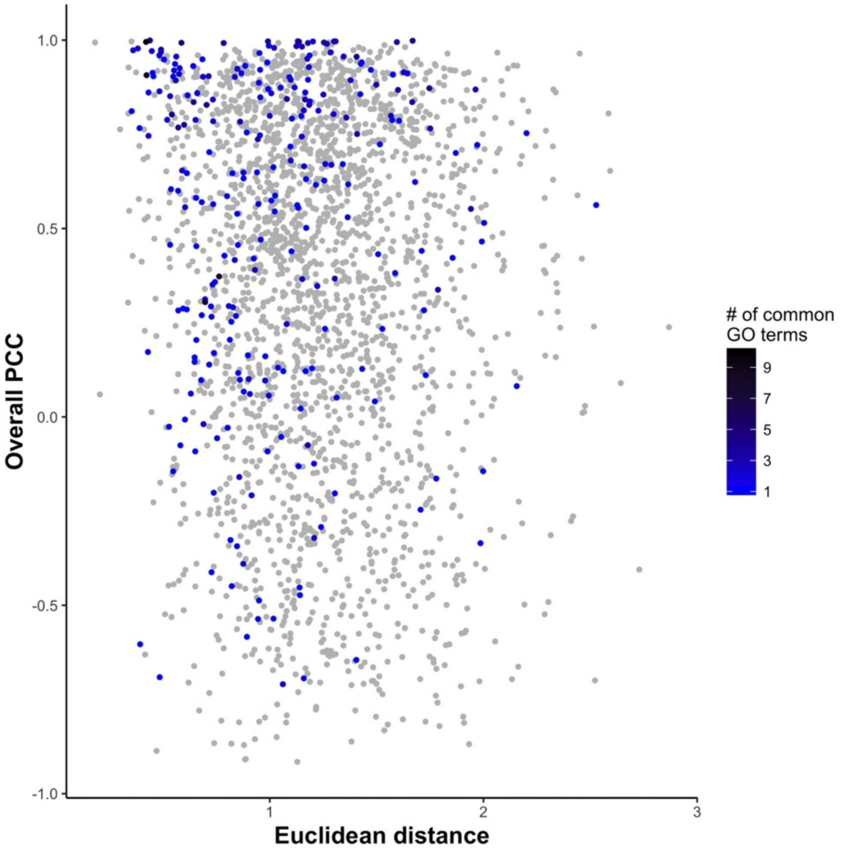


Figure 5.10 The plot of Euclidean distance of protein elution profiles vs. Pearson's correlation coefficient of the mRNA-level co-expression information [285]. The dots are colored based on the number of common GO terms. Grey indicates no common GO terms. Blue to black color indicates the number of common GO terms is from 1 to 10.

5.4 Discussion

The physiology of unicellular *Cyanothece* 51142 is diverse. An understanding of its physiology requires the analysis of the full complement of proteins and the way they are organized and regulated in the cell. I started this by analyzing the *Cyanothece* 51142 protein complexes using the combination of SEC fractionation and quantitative LC-MS/MS profiling. This technique opens up the possibility for systems-wide studies of protein complex dynamics and interactions in cyanobacteria under various physiological conditions. I note that while this technique is very suitable for mapping stable complexes, transient or weak complexes have a higher chance to dissociate during lysis (and dilution) and SEC fractionation and consequently, missed from the detection. Therefore there is still a great need to develop a method that can better discover transient PPIs. Nonetheless, a number of protein complexes were successfully identified that are involved in key metabolic processes, which indicates the validity of our approach, and furthermore, many other known and unknown interacting pairs were identified (Supplemental Table S3), which can serve as valuable reference for future biological works.

In this work, bioinformatics analysis was used to follow up the experiments to provide further supporting evidence of detected PPIs. Since the protein clusters with their elution profiles from the SEC fractionation only provides sets of proteins that have similar profiles, bioinformatics analysis is necessary to actually identify interacting pairs. As a future direction of this work, tertiary structure of protein complexes can be further constructed using protein docking programs [320-322] to provide residue- and atom-level information of protein complexes.

In conclusion, this work represents the first comprehensive analysis on large-scale protein complex study in *Cyanothece* 51142. So far, differential and quantitative proteomic analysis of soluble and membrane proteins of this strain has been well established, and wealth of information of proteins involved in major metabolic pathways is known. However, how these proteins assemble into complexes and function was largely unknown. Here, I was able to add protein complex information with other qualitative and quantitative information, and established an isolation procedure and analytical platform for future studies to reveal how these protein complexes assemble and disassemble as a function of diurnal and circadian rhythms.

CHAPTER 6. DISCUSSION AND SUMMARY

6.1 Remaining Challenges

Current computational methods to predict PPIs mainly relies on the knowledge of the experimentally verified PPIs. These are several disadvantages of predicting PPIs based on the signatures of the experimentally verified PPIs.

First, lots of PPIs are identified by large-scale experimental methods, such as two-yeast hybrid and mass spectrometry. These experimental methods bear high false-positive rate. If the data is used as training data to build the prediction model, it leads to the mistakenly labeling of the interacting pairs and non-interacting pairs in the training data. The prediction model built on these PPIs can introduce some false predictions and lower the reliability of the prediction results.

Second, current classifiers try to detect the underlying patterns of the known PPIs and make the prediction on the unknown protein pairs. Therefore, if some types of PPIs have not been discovered before, it will not be included in the prediction model. This situation would result in the loss of some true PPIs while I am making the predictions. Only the protein pairs with similar patterns as the previously identified PPIs are predicted as PPIs.

Another challenge of computational PPI prediction is feature extraction. Protein sequences can be easily accessed through UniProt or NCBI databases, so lots of sequence-based computational methods were developed. Functional similarity calculated from gene ontology terms requires the query protein with annotated GO terms. However, for some not well-studied species, only very few proteins have annotated GO terms. To overcome this challenge, PFP can be used to predict the GO terms [39]. When gene/protein coexpression level is used as a feature, genome-scale experiment on the query species is required in order to quantify the co-expression level. However, only several model species have large-scale gene co-expression level stored in the databases such as ATTED-II and COXPRESdb [323, 324]. Otherwise, small-scale experiments on the genes/proteins of interest with a limited number of conditional variants can be used as a filter to select the high confident PPIs. However, this type of experiment can only indicate the co-expression association under such conditional variants. The structure-based feature is a very powerful feature in predicting PPIs and getting more insights about how proteins interact. Only a limited number of protein structures are solved and available on PDB and EMDB. To extract the protein structure-based features, structural models need to be first generated. This process is not only time consuming but also have a deviation from the native structure. Therefore, more reliable protein modeling methods are required in order to use the structure-based feature for PPI predictions.

6.2 Future work

Nowadays, lots of deep learning methods are applied to solve biological questions. One strategy is to extract the features the same way as the previous strategies and input them into a neural network. An example is to extract the same physiochemical auto-covariant features based on amino acid sequences, but input into the stacked autoencoder. The accuracy has been improved to 97.19% [325]. Another example is to extract various sequence-based features including amino acid composition, dipeptide composition, amino acid transition, amino acid distribution, and amphiphilic pseudo-amino acid composition and input into a deep recurrent neural network combined with the embedding techniques [326]. It also achieved higher accuracy than You's method who used similar features but input into the SVM model [117].

Another strategy is to develop innovated ways to extract features from protein sequences. One example is transferring the method, that how computational linguists encode the sentences into numerical vectors, to encode the protein sequences using a type of recurrent neural network called long-short term memory (LSTM). In the natural language processing field, there is a very famous LSTM structural called Siamese neural network. It was used to detect the semantic similarity between two sentences. After applying it to predict the PPIs, it also achieved very high accuracy [327]. This model can be improved by considering the physiochemical proprieties of amino acid while encoding the sequences.

Besides the improving the PPI prediction methods by adopting deep learning methods, PPI structural models are also needed to better understand the mechanism of interaction including the shape complementary, interacting residues, and the functionality of PPIs. PPI structural modeling methods will generate thousands of models. Quality assessment is necessary to select the one closest to the native structure. Commonly used quality assessment methods including calculation of protein model energy scores such as DFIRE, GOAP, ITScore, and Soap [328-331]. These scores provide lookup tables with distance and angle probability distributions of pairwise atom type, which are calculated based on the statistics of the known protein structures. The close-to-native protein structures should have lower energy scores. A very naïve but efficient way to select the best model is called RankSum. It selects the model with the smallest sum of the rank of each energy score. Other than calculating the energy scores, several deep learning methods using the 3D convolutional neural network (CNN) are also been applied to assess the quality of the single protein structural models [332-334]. In Derevyanko's method, the kernel of the CNN is just simply checking the presence of the 11 atom types [335]. This method can be improved by subdividing the atom types or changing the kernel function to the energy score probability distribution function.

6.3 Outlook

Previous work has used structural information from the homologous modeled protein complex to predict the PPIs [336]. In the long term, I also hope that the structure of PPIs can be accurately modeled. In this field of work, computational docking can help to locate the interface [180]. Direct coupling analysis has been proven to successfully predict hot-spot residues [337]. This analysis is based on the assumption that the structure of the protein complex is conserved to conserve the functionality during the evolution. Therefore, structural conservation constrains the sequence variability and forces the interacting residues to coevolve. Understanding the interface and hot-spot residues are important and they will guide biologist to design mutagenesis experiments, as well as to gain further understanding of the functionality and interacting mechanisms. The modeled protein structures can also give some insights to the drug discovery targeting PPIs [338]

REFERENCES

1. Hedin, S., *Trypsin and antitrypsin.* Biochemical Journal, 1906. **1**(10): p. 474.
2. Svedberg, T. and E. Chirnoaga, *The molecular weight of hemocyanin.* Journal of the American Chemical Society, 1928. **50**(5): p. 1399-1411.
3. Sanger, F. and E. Thompson, *The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates.* Biochemical Journal, 1953. **53**(3): p. 353.
4. Sanger, F. and E. Thompson, *The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates.* Biochemical Journal, 1953. **53**(3): p. 366.
5. Northrop, J.H., *THE INACTIVATION OF TRYPSIN: II. THE EQUILIBRIUM BETWEEN TRYPSIN AND THE INHIBITING SUBSTANCE FORMED BY ITS ACTION ON PROTEINS.* The Journal of general physiology, 1922. **4**(3): p. 245.
6. *Identification of associated proteins by coimmunoprecipitation.* Nat Meth, 2005. **2**(6): p. 475-476.
7. Sinz, A., *Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions.* Mass spectrometry reviews, 2006. **25**(4): p. 663-682.
8. Fancy, D.A. and T.J.P.o.t.N.A.o.S. Kodadek, *Chemistry for the analysis of protein–protein interactions: rapid and efficient cross-linking triggered by long wavelength light.* 1999. **96**(11): p. 6020-6024.
9. Barnard, E., et al., *Detection of protein-protein interactions using protein-fragment complementation assays (PCA).* Current Proteomics, 2007. **4**(1): p. 17-27.
10. Kenworthy, A.K., *Imaging protein-protein interactions using fluorescence resonance energy transfer microscopy.* Methods, 2001. **24**(3): p. 289-296.
11. Young, K.J.B.o.r., *Yeast two-hybrid: so many interactions,(in) so little time….* 1998. **58**(2): p. 302-311.
12. Deane, C.M., et al., *Protein interactions: two methods for assessment of the reliability of high throughput observations.* Molecular & Cellular Proteomics, 2002. **1**(5): p. 349-356.
13. Nikolovska-Coleska, Z., *Studying protein-protein interactions using surface plasmon resonance.* Protein-Protein Interactions: Methods and Applications, 2015: p. 109-138.
14. Pierce, M.M., C. Raman, and B.T. Nall, *Isothermal titration calorimetry of protein–protein interactions.* Methods, 1999. **19**(2): p. 213-221.
15. Vinogradova, O. and J. Qin, *NMR as a unique tool in assessment and complex determination of weak protein–protein interactions*, in *NMR of Proteins and Small Biomolecules*. 2011, Springer. p. 35-45.
16. Zuiderweg, E.R., *Mapping protein− protein interactions in solution by NMR spectroscopy.* Biochemistry, 2002. **41**(1): p. 1-7.
17. Kobe, B., et al., *Crystallography and protein–protein interactions: biological interfaces and crystal contacts*. 2008, Portland Press Limited.
18. Dudkina, N.V., et al., *Imaging of organelles by electron microscopy reveals protein–protein interactions in mitochondria and chloroplasts.* FEBS letters, 2010. **584**(12): p. 2510-2515.

19. Boeri Erba, E. and C. Petosa, *The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes.* Protein Science, 2015. **24**(8): p. 1176-1192.
20. Dunham, W.H., M. Mullin, and A.C. Gingras, *Affinity-purification coupled to mass spectrometry: Basic principles and strategies.* Proteomics, 2012. **12**(10): p. 1576-1590.
21. Morris, J.H., et al., *Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions.* Nature protocols, 2014. **9**(11): p. 2539-2554.
22. Guruharsha, K., et al., *A protein complex network of Drosophila melanogaster.* Cell, 2011. **147**(3): p. 690-703.
23. Rao, V.S., et al., *Protein-protein interaction detection: methods and analysis.* International journal of proteomics, 2014. **2014**.
24. Piehler, J., *New methodologies for measuring protein interactions in vivo and in vitro.* Current opinion in structural biology, 2005. **15**(1): p. 4-14.
25. Wetie, N., et al., *Investigation of stable and transient protein–protein interactions: past, present, and future.* Proteomics, 2013. **13**(3-4): p. 538-557.
26. Huang, H. and J.S. Bader, *Precision and recall estimates for two-hybrid screens.* Bioinformatics, 2009. **25**(3): p. 372-378.
27. Serebriiskii, I.G. and E.A. Golemis, *Two-Hybrid System and False Positives: Approahes to Detection and Elimination.* Two-Hybrid Systems: Methods and Protocols, 2001: p. 123-134.
28. Gingras, A.-C., et al., *Analysis of protein complexes using mass spectrometry.* Nature reviews Molecular cell biology, 2007. **8**(8): p. 645-654.
29. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2017 update.* Nucleic Acids Res, 2017. **45**(D1): p. D369-D379.
30. Hawkins, T., et al., *PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data.* Proteins: Struct, Funct, Bioinf, 2009. **74**.
31. Wei, Q., et al., *NaviGO: interactive tool for visualization and functional similarity and coherence analysis with gene ontology.* BMC Bioinformatics, 2017. **18**(1): p. 177.
32. Ding, Z., Q. Wei, and D. Kihara, *Computing and Visualizing Gene Function Similarity and Coherence with NaviGO*, in *Data Mining for Systems Biology*. 2018, Springer. p. 113-130.
33. Consortium, G.O., *Gene Ontology annotations and resources.* Nucleic acids research, 2013. **41**(D1): p. D530-D535.
34. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The gene ontology consortium.* Nat Genet, 2000. **25**.
35. Carbon, S., et al., *AmiGO: online access to ontology and annotation data.* Bioinformatics, 2009. **25**.
36. Binns, D., et al., *QuickGO: a web-based tool for Gene Ontology searching.* Bioinformatics, 2009. **25**(22): p. 3045-3046.
37. Hawkins, T., S. Luban, and D. Kihara, *Enhanced automated function prediction using distantly related sequences and contextual association by PFP.* Protein Sci, 2006. **15**.
38. Chitale, M., et al., *ESG: extended similarity group method for automated protein function prediction.* Bioinformatics, 2009. **25**.
39. Khan, I.K., W. Qing, and D. Kihara, *PFP/ESG: automated protein function prediction servers enhanced with gene ontology visualization tool.* Bioinformatics, 2015. **31**.

40. Pundir, S., M.J. Martin, and C. O'Donovan, *UniProt Protein Knowledgebase.* Methods Mol Biol, 2017. **1558**: p. 41-55.
41. Schlicker, A., et al., *A new measure for functional similarity of gene products based on gene ontology.* BMC Bioinf, 2006. **7**.
42. Meghana, C., P. Shriphani, and K. Daisuke, *Quantification of protein group coherence and pathway assignment using functional association.* BMC Bioinf, 2011. **12**.
43. Yerneni, S., et al., *IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks.* IEEE/ACM Trans Comput Biol Bioinform, 2018. **15**(4): p. 1247-1258.
44. Dieterle, M., et al., *A new type of mutation in phytochrome A causes enhanced light sensitivity and alters the degradation and subcellular partitioning of the photoreceptor.* The Plant Journal, 2005. **41**(1): p. 146-161.
45. Nito, K., et al., *Tyrosine phosphorylation regulates the activity of phytochrome photoreceptors.* Cell Reports, 2013. **3**(6): p. 1970-1979.
46. Al-Sady, B., et al., *Photoactivated phytochrome induces rapid PIF3 phosphorylation prior to proteasome-mediated degradation.* Molecular cell, 2006. **23**(3): p. 439-446.
47. Liu, X., et al., *PHYTOCHROME INTERACTING FACTOR3 associates with the histone deacetylase HDA15 in repression of chlorophyll biosynthesis and photosynthesis in etiolated Arabidopsis seedlings.* The Plant Cell, 2013. **25**(4): p. 1258-1273.
48. Ito, S., et al., *Genetic linkages between circadian clock-associated components and phytochrome-dependent red light signal transduction in Arabidopsis thaliana.* Plant and cell physiology, 2007. **48**(7): p. 971-983.
49. Resnik, P., *Using information content to evaluate semantic similarity in a taxonomy.* arXiv preprint cmp-lg/9511007, 1995.
50. Lin, D. *An information-theoretic definition of similarity*. in *ICML*. 1998. Citeseer.
51. Yerneni, S., et al., *IAS: Interaction specific GO term associations for predicting Protein-Protein Interaction Networks.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015.
52. Chitale, M., S. Palakodety, and D. Kihara, *Quantification of protein group coherence and pathway assignment using functional association.* BMC bioinformatics, 2011. **12**(1): p. 373.
53. Hawkins, T., M. Chitale, and D. Kihara, *Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP.* Bmc Bioinformatics, 2010. **11**(1): p. 265.
54. Clack, T., et al., *Obligate heterodimerization of Arabidopsis phytochromes C and E and interaction with the PIF3 basic helix-loop-helix transcription factor.* The Plant Cell, 2009. **21**(3): p. 786-799.
55. Szklarczyk, D., et al., *The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.* Nucleic Acids Res, 2017. **45**(D1): p. D362-D368.
56. Consortium, G.O., *Expansion of the Gene Ontology knowledgebase and resources.* Nucleic acids research, 2017. **45**(D1): p. D331-D338.
57. Wu, X., et al., *Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and IC-based hybrid method.* PloS one, 2013. **8**(5): p. e66745.

58. Zhang, S.-B. and Q.-R. Tang, *Protein–protein interaction inference based on semantic similarity of Gene Ontology terms.* Journal of theoretical biology, 2016. **401**: p. 30-37.
59. Jain, S. and G.D. Bader, *An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology.* BMC bioinformatics, 2010. **11**(1): p. 562.
60. Guo, X., et al., *Assessing semantic similarity measures for the characterization of human regulatory pathways.* Bioinformatics, 2006. **22**(8): p. 967-73.
61. Maetschke, S.R., et al., *Gene Ontology-driven inference of protein–protein interactions using inducers.* Bioinformatics, 2012. **28**(1): p. 69-75.
62. Ding, Z. and D. Kihara, *Computational Methods for Predicting Protein-Protein Interactions Using Various Protein Features.* Current Protocols in Protein Science, 2018: p. e62.
63. Chen, X.W. and M. Liu, *Prediction of protein-protein interactions using random decision forest framework.* Bioinformatics, 2005. **21**(24): p. 4394-400.
64. Sprinzak, E. and H. Margalit, *Correlated sequence-signatures as markers of protein-protein interaction.* J Mol Biol, 2001. **311**(4): p. 681-92.
65. Pitre, S., et al., *PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.* BMC Bioinformatics, 2006. **7**: p. 365.
66. Shen, J., et al., *Predicting protein-protein interactions based only on sequences information.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4337-4341.
67. Nanni, L. and A. Lumini, *An ensemble of K-local hyperplanes for predicting protein-protein interactions.* Bioinformatics, 2006. **22**(10): p. 1207-10.
68. Ding, Y., J. Tang, and F. Guo, *Predicting protein-protein interactions via multivariate mutual information of protein sequences.* BMC Bioinformatics, 2016. **17**(1): p. 398.
69. You, Z.H., K.C. Chan, and P. Hu, *Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest.* PLoS One, 2015. **10**(5): p. e0125811.
70. Guo, Y., et al., *Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences.* Nucleic Acids Res, 2008. **36**(9): p. 3025-30.
71. Walhout, A.J., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116.
72. Huang, T.W., et al., *POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome.* Bioinformatics, 2004. **20**(17): p. 3273-6.
73. Lee, S.A., et al., *Ortholog-based protein-protein interaction prediction and its application to inter-species interactions.* BMC Bioinformatics, 2008. **9 Suppl 12**: p. S11.
74. De Bodt, S., et al., *Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression.* BMC Genomics, 2009. **10**: p. 288.
75. Gu, H., et al., *PRIN: a predicted rice interactome network.* BMC Bioinformatics, 2011. **12**: p. 161.
76. Najafabadi, H.S. and R. Salavati, *Sequence-based prediction of protein-protein interactions by means of codon usage.* Genome Biol, 2008. **9**(5): p. R87.
77. Zhang, S.B. and Q.R. Tang, *Protein-protein interaction inference based on semantic similarity of Gene Ontology terms.* J Theor Biol, 2016. **401**: p. 30-7.

78. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein-protein interaction.* Protein Eng, 2001. **14**(9): p. 609.
79. Juan, D., F. Pazos, and A. Valencia, *High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.* Proc Natl Acad Sci U S A, 2008. **105**(3): p. 934-9.
80. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-92.
81. Soong, T.T., K.O. Wrzeszczynski, and B. Rost, *Physical protein-protein interactions predicted from microarrays.* Bioinformatics, 2008. **24**(22): p. 2608-14.
82. Wass, M.N., et al., *Towards the prediction of protein interaction partners using physical docking.* Mol Syst Biol, 2011. **7**: p. 469.
83. Ohue, M., et al., *MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data.* Protein Pept Lett, 2014. **21**(8): p. 766-78.
84. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM.* Nat Protoc, 2011. **6**(9): p. 1341-54.
85. Mirabello, C. and B. Wallner, *InterPred: A pipeline to identify and model protein-protein interactions.* Proteins, 2017. **85**(6): p. 1159-1170.
86. Blohm, P., et al., *Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis.* Nucleic Acids Res, 2014. **42**(Database issue): p. D396-400.
87. Szklarczyk, D., et al., *STRING v10: protein–protein interaction networks, integrated over the tree of life.* Nucleic acids research, 2014: p. gku1003.
88. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic acids research, 2002. **30**(1): p. 303-305.
89. Wong, P., et al., *An evolutionary and structural characterization of mammalian protein complex organization.* Bmc Genomics, 2008. **9**(1): p. 629.
90. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic acids research, 2004. **32**(suppl 1): p. D452-D455.
91. Licata, L., et al., *MINT, the molecular interaction database: 2012 update.* Nucleic Acids Res, 2012. **40**(Database issue): p. D857-61.
92. Breuer, K., et al., *InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1228-33.
93. Prasad, T.K., et al., *Human protein reference database—2009 update.* Nucleic acids research, 2009. **37**(suppl 1): p. D767-D772.
94. Keseler, I.M., et al., *The EcoCyc database: reflecting new knowledge about Escherichia coli K-12.* Nucleic Acids Research, 2016: p. gkw1003.
95. Garcia-Hernandez, M., et al., *TAIR: a resource for integrated Arabidopsis data.* Funct Integr Genomics, 2002. **2**(6): p. 239-53.
96. Becerra, A., V.A. Bucheli, and P.A. Moreno, *Prediction of virus-host protein-protein interactions mediated by short linear motifs.* BMC bioinformatics, 2017. **18**(1): p. 163.
97. Dinkel, H., et al., *ELM--the database of eukaryotic linear motifs.* Nucleic Acids Res, 2012. **40**(Database issue): p. D242-51.
98. Finn, R.D., et al., *InterPro in 2017-beyond protein family and domain annotations.* Nucleic Acids Res, 2017. **45**(D1): p. D190-D199.

99. Sigrist, C.J., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D161-6.
100. Attwood, T.K., et al., *The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012.* Database (Oxford), 2012. **2012**: p. bas019.
101. Finn, R.D., et al., *The Pfam protein families database: towards a more sustainable future.* Nucleic Acids Res, 2016. **44**(D1): p. D279-85.
102. Corpet, F., J. Gouzy, and D. Kahn, *The ProDom database of protein domain families.* Nucleic Acids Res, 1998. **26**(1): p. 323-6.
103. Bru, C., et al., *The ProDom database of protein domain families: more emphasis on 3D.* Nucleic Acids Res, 2005. **33**(Database issue): p. D212-5.
104. Sprinzak, E. and H. Margalit, *Correlated sequence-signatures as markers of protein-protein interaction.* Journal of molecular biology, 2001. **311**(4): p. 681-692.
105. Kim, W.K., J. Park, and J.K. Suh, *Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair.* Genome Informatics Series, 2002: p. 42-50.
106. Chen, X.-W. and M. Liu, *Prediction of protein–protein interactions using random decision forest framework.* Bioinformatics, 2005. **21**(24): p. 4394-4400.
107. Pitre, S., et al., *PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs.* BMC bioinformatics, 2006. **7**(1): p. 365.
108. Betel, D., et al., *Structure-templated predictions of novel protein interactions from sequence information.* PLoS Comput Biol, 2007. **3**(9): p. e182.
109. Nanni, L., *Fusion of classifiers for predicting protein–protein interactions.* Neurocomputing, 2005. **68**: p. 289-296.
110. Shen, J., et al., *Predicting protein–protein interactions based only on sequences information.* Proceedings of the National Academy of Sciences, 2007. **104**(11): p. 4337-4341.
111. Yu, C.-Y., L.-C. Chou, and D.T.-H. Chang, *Predicting protein-protein interactions in unbalanced data using the primary structure of proteins.* BMC bioinformatics, 2010. **11**(1): p. 167.
112. Wei, L., et al., *Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier.* Artif Intell Med, 2017.
113. Martin, S., D. Roe, and J.-L. Faulon, *Predicting protein–protein interactions using signature products.* Bioinformatics, 2005. **21**(2): p. 218-226.
114. Wong, L., et al. *Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel PR-LPQ descriptor*. in *International Conference on Intelligent Computing*. 2015. Springer.
115. An, J.Y., et al., *Improving protein–protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model.* Protein Science, 2016. **25**(10): p. 1825-1833.
116. Dubchak, I., et al., *Prediction of protein folding class using global description of amino acid sequence.* Proceedings of the National Academy of Sciences, 1995. **92**(19): p. 8700-8704.
117. You, Z.-H., K.C. Chan, and P. Hu, *Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest.* PLoS One, 2015. **10**(5): p. e0125811.

118. Yang, L., J.-F. Xia, and J. Gui, *Prediction of protein-protein interactions from protein sequence using local descriptors.* Protein and Peptide Letters, 2010. **17**(9): p. 1085-1090.
119. Zhou, Y.Z., Y. Gao, and Y.Y. Zheng, *Prediction of protein-protein interactions using local description of amino acid sequence*, in *Advances in Computer Science and Education Applications*. 2011, Springer. p. 254-262.
120. You, Z.-H., et al., *Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis.* BMC bioinformatics, 2013. **14**(8): p. S10.
121. Guo, Y., et al., *Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences.* Nucleic acids research, 2008. **36**(9): p. 3025-3030.
122. Bock, J.R. and D.A. Gough, *Predicting protein–protein interactions from primary structure.* Bioinformatics, 2001. **17**(5): p. 455-460.
123. Liu, X., et al., *SPPS: a sequence-based method for predicting probability of protein-protein interaction partners.* PloS one, 2012. **7**(1): p. e30938.
124. Hawkins, T. and D. Kihara, *Function prediction of uncharacterized proteins.* Journal of bioinformatics and computational biology, 2007. **5**(01): p. 1-30.
125. Chitale, M., et al., *ESG: extended similarity group method for automated protein function prediction.* Bioinformatics, 2009. **25**(14): p. 1739-1745.
126. Tanabe, M. and M. Kanehisa, *Using the KEGG database resource.* Curr Protoc Bioinformatics, 2012. **Chapter 1**: p. Unit1 12.
127. Waterhouse, R.M., et al., *OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs.* Nucleic acids research, 2013. **41**(D1): p. D358-D365.
128. Chen, F., et al., *OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.* Nucleic acids research, 2006. **34**(suppl 1): p. D363-D368.
129. NCBI, R.C., *Database resources of the National Center for Biotechnology Information.* Nucleic acids research, 2016. **44**(D1): p. D7.
130. Sonnhammer, E.L. and G. Ostlund, *InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic.* Nucleic Acids Res, 2015. **43**(Database issue): p. D234-9.
131. Walhout, A.J., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116-122.
132. Matthews, L.R., et al., *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome research, 2001. **11**(12): p. 2120-2126.
133. Huang, T.-W., et al., *POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome.* Bioinformatics, 2004. **20**(17): p. 3273-3276.
134. Lee, S.-A., et al., *Ortholog-based protein-protein interaction prediction and its application to inter-species interactions.* BMC bioinformatics, 2008. **9**(12): p. S11.
135. Geisler-Lee, J., et al., *A predicted interactome for Arabidopsis.* Plant Physiology, 2007. **145**(2): p. 317-329.
136. De Bodt, S., et al., *Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression.* BMC genomics, 2009. **10**(1): p. 288.
137. Gu, H., et al., *PRIN: a predicted rice interactome network.* BMC bioinformatics, 2011. **12**(1): p. 161.

138. Dutkowski, J. and J. Tiuryn, *Phylogeny-guided interaction mapping in seven eukaryotes.* BMC bioinformatics, 2009. **10**(1): p. 393.

139. Mosca, R., et al., *Towards a detailed atlas of protein–protein interactions.* Current opinion in structural biology, 2013. **23**(6): p. 929-940.

140. Wang, F., et al., *Prediction and characterization of protein-protein interaction networks in swine.* Proteome science, 2012. **10**(1): p. 2.

141. Najafabadi, H.S. and R. Salavati, *Sequence-based prediction of protein-protein interactions by means of codon usage.* Genome biology, 2008. **9**(5): p. R87.

142. Zhou, Y., et al., *Can simple codon pair usage predict protein–protein interaction?* Molecular BioSystems, 2012. **8**(5): p. 1396-1404.

143. Jansen, R., H.J. Bussemaker, and M. Gerstein, *Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models.* Nucleic acids research, 2003. **31**(8): p. 2242-2251.

144. Huynen, M.A., et al., *Function prediction and protein networks.* Current opinion in cell biology, 2003. **15**(2): p. 191-198.

145. Goh, C.-S., et al., *Co-evolution of proteins with their interaction partners.* Journal of molecular biology, 2000. **299**(2): p. 283-293.

146. Goh, C.-S. and F.E. Cohen, *Co-evolutionary analysis reveals insights into protein–protein interactions.* Journal of molecular biology, 2002. **324**(1): p. 177-192.

147. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein–protein interaction.* Protein engineering, 2001. **14**(9): p. 609-614.

148. Kann, M.G., et al., *Correlated evolution of interacting proteins: looking behind the mirrortree.* Journal of molecular biology, 2009. **385**(1): p. 91-98.

149. Ramani, A.K. and E.M. Marcotte, *Exploiting the co-evolution of interacting proteins to discover interaction specificity.* Journal of molecular biology, 2003. **327**(1): p. 273-284.

150. Sato, T., et al., *The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.* Bioinformatics, 2005. **21**(17): p. 3482-3489.

151. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-2492.

152. Juan, D., F. Pazos, and A. Valencia, *High-confidence prediction of global interactomes based on genome-wide coevolutionary networks.* Proceedings of the National Academy of Sciences, 2008. **105**(3): p. 934-939.

153. Herman, D., et al., *Selection of organisms for the co-evolution-based study of protein interactions.* BMC bioinformatics, 2011. **12**(1): p. 363.

154. Craig, R.A. and L. Liao, *Improving Protein–Protein Interaction Prediction Based on Phylogenetic Information Using a Least-Squares Support Vector Machine.* Annals of the New York Academy of Sciences, 2007. **1115**(1): p. 154-167.

155. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(8): p. 4285-4288.

156. Sun, J., Y. Li, and Z. Zhao, *Phylogenetic profiles for the prediction of protein-protein interactions: how to select reference organisms?* Biochem Biophys Res Commun, 2007. **353**(4): p. 985-91.

157. Sun, J., et al., *Refined phylogenetic profiles method for predicting protein–protein interactions.* Bioinformatics, 2005. **21**(16): p. 3409-3415.
158. Lin, T.-W., J.-W. Wu, and D.T.-H. Chang, *Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins.* PloS one, 2013. **8**(9): p. e75940.
159. de Vienne, D.M. and J. Azé, *Efficient prediction of co-complexed proteins based on coevolution.* PloS one, 2012. **7**(11): p. e48728.
160. Snel, B., P. Bork, and M. Huynen, *Genome evolution.* Trends in genetics, 2000. **16**(1): p. 9-10.
161. Yanai, I., A. Derti, and C. DeLisi, *Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.* Proceedings of the National Academy of Sciences, 2001. **98**(14): p. 7940-7945.
162. Enright, A.J., et al., *Protein interaction maps for complete genomes based on gene fusion events.* Nature, 1999. **402**(6757): p. 86-90.
163. Marcotte, E.M., et al., *Detecting protein function and protein-protein interactions from genome sequences.* Science, 1999. **285**(5428): p. 751-753.
164. Tsoka, S. and C.A. Ouzounis, *Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion.* Nature Genetics, 2000. **26**(2): p. 141-142.
165. Suyama, M. and P. Bork, *Evolution of prokaryotic gene order: genome rearrangements in closely related species.* Trends in Genetics, 2001. **17**(1): p. 10-13.
166. Tamames, J., et al., *Conserved clusters of functionally related genes in two bacterial genomes.* Journal of molecular evolution, 1997. **44**(1): p. 66-73.
167. Dandekar, T., et al., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends in biochemical sciences, 1998. **23**(9): p. 324-328.
168. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proceedings of the National Academy of Sciences, 1999. **96**(6): p. 2896-2901.
169. Fujibuchi, W., et al., *Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping.* Nucleic acids research, 2000. **28**(20): p. 4029-4036.
170. Kihara, D. and M. Kanehisa, *Tandem clusters of membrane proteins in complete genome sequences.* Genome research, 2000. **10**(6): p. 731-743.
171. Grigoriev, A., *A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae.* Nucleic acids research, 2001. **29**(17): p. 3513-3519.
172. Ge, H., et al., *Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae.* Nature genetics, 2001. **29**(4): p. 482-486.
173. Jansen, R., D. Greenbaum, and M. Gerstein, *Relating whole-genome expression data with protein-protein interactions.* Genome research, 2002. **12**(1): p. 37-46.
174. Bhardwaj, N. and H. Lu, *Correlation between gene expression profiles and protein–protein interactions within and across genomes.* Bioinformatics, 2005. **21**(11): p. 2730-2738.
175. Fraser, H.B., et al., *Coevolution of gene expression among interacting proteins.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(24): p. 9033-9038.
176. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic acids research, 2013. **41**(D1): p. D991-D995.

177. Aoki, Y., et al., *ATTED-II in 2016: A Plant Coexpression Database Towards Lineage-Specific Coexpression.* Plant Cell Physiol, 2016. **57**(1): p. e5.
178. Okamura, Y., et al., *COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems.* Nucleic acids research, 2014: p. gku1163.
179. Soong, T.-t., K.O. Wrzeszczynski, and B. Rost, *Physical protein–protein interactions predicted from microarrays.* Bioinformatics, 2008. **24**(22): p. 2608-2614.
180. Esquivel-Rodriguez, J., et al., *Pairwise and multimeric protein-protein docking using the LZerD program suite.* Methods Mol Biol, 2014. **1137**: p. 209-34.
181. Venkatraman, V., et al., *Protein-protein docking using region-based 3D Zernike descriptors.* BMC bioinformatics, 2009. **10**(1): p. 407.
182. Esquivel-Rodríguez, J., Y.D. Yang, and D. Kihara, *Multi-LZerD: Multiple protein docking for asymmetric complexes.* Proteins: Structure, Function, and Bioinformatics, 2012. **80**(7): p. 1818-1833.
183. Peterson, L.X., et al., *Modeling disordered protein interactions from biophysical principles.* PLOS Computational Biology, 2017. **13**(4): p. e1005485.
184. Tovchigrechko, A. and I.A. Vakser, *GRAMM-X public web server for protein-protein docking.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W310-4.
185. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics, 2014. **30**(12): p. 1771-3.
186. Lyskov, S. and J.J. Gray, *The RosettaDock server for local protein-protein docking.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W233-8.
187. Geng, C., et al., *Information-Driven, Ensemble Flexible Peptide Docking Using HADDOCK.* Modeling Peptide-Protein Interactions: Methods and Protocols, 2017: p. 109-138.
188. Torchala, M. and P.A. Bates, *Predicting the structure of protein–protein complexes using the SwarmDock web server.* Protein Structure Prediction, 2014: p. 181-197.
189. Ritchie, D.W. and G.J. Kemp, *Protein docking using spherical polar Fourier correlations.* Proteins: Structure, Function, and Bioinformatics, 2000. **39**(2): p. 178-194.
190. Kozakov, D., et al., *The ClusPro web server for protein-protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.
191. Wass, M.N., et al., *Towards the prediction of protein interaction partners using physical docking.* Molecular systems biology, 2011. **7**(1): p. 469.
192. Ohue, M., et al., *MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data.* Protein and peptide letters, 2014. **21**(8): p. 766-778.
193. Tuncbag, N., et al., *Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM.* Nature protocols, 2011. **6**(9): p. 1341-1354.
194. Aytuna, A.S., A. Gursoy, and O. Keskin, *Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces.* Bioinformatics, 2005. **21**(12): p. 2850-2855.
195. Rose, P.W., et al., *The RCSB protein data bank: integrative view of protein, gene and 3D structural information.* Nucleic Acids Res, 2017. **45**(D1): p. D271-D281.
196. Zhang, Q.C., et al., *Structure-based prediction of protein-protein interactions on a genome-wide scale.* Nature, 2012. **490**(7421): p. 556-560.

197. Hosur, R., et al., *A computational framework for boosting confidence in high-throughput protein-protein interaction datasets.* Genome biology, 2012. **13**(8): p. R76.
198. Mirabello, C. and B. Wallner, *InterPred: a pipeline to identify and model protein-protein interactions.* Proteins, 2017.
199. Kihara, D. and J. Skolnick, *Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(2): p. 464-473.
200. Pieper, U., et al., *MODBASE: a database of annotated comparative protein structure models and associated resources.* Nucleic acids research, 2006. **34**(suppl 1): p. D291-D295.
201. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic acids research, 2005. **33**(suppl 1): p. D501-D504.
202. Boutet, E., et al., *UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View.* Methods Mol Biol, 2016. **1374**: p. 23-54.
203. Yanai, I., et al., *Mapping gene expression in two Xenopus species: evolutionary constraints and developmental flexibility.* Developmental cell, 2011. **20**(4): p. 483-496.
204. Sharma, A., *Computational gene expression profiling under salt stress reveals patterns of co-expression.* Genomics data, 2016. **7**: p. 214-221.
205. Tong, Z., et al., *Selection of reliable reference genes for gene expression studies in peach using real-time PCR.* BMC molecular biology, 2009. **10**(1): p. 71.
206. Chen, J., et al., *Increasing confidence of protein interactomes using network topological metrics.* Bioinformatics, 2006. **22**(16): p. 1998-2004.
207. Yu, H., et al., *Predicting interactions in protein networks by completing defective cliques.* Bioinformatics, 2006. **22**(7): p. 823-829.
208. Liu, G., J. Li, and L. Wong, *Assessing and predicting protein interactions using both local and global network topological metrics.* Genome Informatics, 2008. **21**: p. 138-149.
209. Kuchaiev, O., et al., *Geometric de-noising of protein-protein interaction networks.* PLoS Comput Biol, 2009. **5**(8): p. e1000454.
210. Lei, C. and J. Ruan. *A random walk based approach for improving protein-protein interaction network and protein complex prediction*. in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. 2012. IEEE.
211. Qi, Y., 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, 2006. **63**(3): p. 490-500.
212. Miller, J.P., et al., *Large-scale identification of yeast integral membrane protein interactions.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(34): p. 12123-12128.
213. Mewes, H.-W., et al., *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic acids research, 2004. **32**(suppl 1): p. D41-D44.
214. Ben-Hur, A. and W.S. Noble, *Kernel methods for predicting protein–protein interactions.* Bioinformatics, 2005. **21**(suppl 1): p. i38-i46.
215. Xu, F., et al., *Global protein interactome exploration through mining genome-scale data in Arabidopsis thaliana.* BMC genomics, 2010. **11**(2): p. S2.
216. Kotlyar, M., et al., *In silico prediction of physical protein interactions and characterization of interactome orphans.* Nat Methods, 2015. **12**(1): p. 79-84.

217. Taghipour, S., et al., *Improving protein complex prediction by reconstructing a high-confidence protein-protein interaction network of Escherichia coli from different physical interaction data sources.* BMC Bioinformatics, 2017. **18**(1): p. 10.
218. Ding, Z. and D. Kihara, *Computational identification of protein-protein interactions in model plant proteomes.* Scientific Reports, 2019. **9**(1): p. 8740.
219. Habibi, M., C. Eslahchi, and L. Wong, *Protein complex prediction based on k-connected subgraphs in protein interaction network.* BMC Syst Biol, 2010. **4**: p. 129.
220. King, A.D., N. Przulj, and I. Jurisica, *Protein complex prediction via cost-based clustering.* Bioinformatics, 2004. **20**(17): p. 3013-20.
221. Hawkins, T. and D. Kihara, *Function prediction of uncharacterized proteins.* J. Bioinform. Comput. Biol., 2007. **5**(1): p. 1-30.
222. Hawkins, T., M. Chitale, and D. Kihara, *New paradigm in protein function prediction for large scale omics analysis.* Mol Biosyst, 2008. **4**(3): p. 223-31.
223. Khan, I.K. and D. Kihara, *Genome-scale prediction of moonlighting proteins using diverse protein association information.* Bioinformatics, 2016. **32**(15): p. 2281-2288.
224. Shin, W.H., C.W. Christoffer, and D. Kihara, *In silico structure-based approaches to discover protein-protein interaction-targeting drugs.* Methods, 2017. **131**: p. 22-32.
225. King, N.P., et al., *Computational design of self-assembling protein nanomaterials with atomic level accuracy.* Science, 2012. **336**(6085): p. 1171-1174.
226. Tanford, C., *Contribution of hydrophobic interactions to the stability of the globular conformation of proteins.* Journal of the American Chemical Society, 1962. **84**(22): p. 4240-4247.
227. Hopp, T.P. and K.R. Woods, *Prediction of protein antigenic determinants from amino acid sequences.* Proceedings of the National Academy of Sciences, 1981. **78**(6): p. 3824-3828.
228. Krigbaum, W. and A. Komoriya, *Local interactions as a structure determinant for protein molecules: II.* Biochimica et biophysica acta, 1979. **576**(1): p. 204-248.
229. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. **185**(4154): p. 862-864.
230. Charton, M. and B.I. Charton, *The structural dependence of amino acid hydrophobicity parameters.* Journal of theoretical biology, 1982. **99**(4): p. 629-644.
231. Rose, G.D., et al., *Hydrophobicity of amino acid residues in globular proteins.* Science, 1985. **229**(4716): p. 834-838.
232. Zhou, P., et al., *Genetic algorithm-based virtual screening of combinative mode for peptide/protein.* ACTA CHIMICA SINICA-CHINESE EDITION-, 2006. **64**(7): p. 691.
233. Consortium, G.O., *Gene Ontology Consortium: going forward.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56.
234. Chitale, M., S. Palakodety, and D. Kihara, *Quantification of protein group coherence and pathway assignment using functional association.* BMC Bioinformatics, 2011. **12**: p. 373.
235. Chitale, M., I.K. Khan, and D. Kihara, *Missing gene identification using functional coherence scores.* Scientific reports, 2016. **6**: p. 31725.
236. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci USA, 1999. **96**(8): p. 4285-4288.
237. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology (TIST), 2011. **2**(3): p. 27.

238. You, Z.H., et al., *Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis.* BMC Bioinformatics, 2013. **14 Suppl 8**: p. S10.
239. An, J.Y., et al., *Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model.* Protein Sci, 2016. **25**(10): p. 1825-33.
240. Huang, Y.A., et al., *Using Weighted Sparse Representation Model Combined with Discrete Cosine Transformation to Predict Protein-Protein Interactions from Protein Sequence.* Biomed Res Int, 2015. **2015**: p. 902198.
241. Varma, S. and R. Simon, *Bias in error estimation when using cross-validation for model selection.* BMC bioinformatics, 2006. **7**(1): p. 91.
242. Aliferis, C.F., A. Statnikov, and I. Tsamardinos, *Challenges in the analysis of mass-throughput data: a technical commentary from the statistical machine learning perspective.* Cancer Informatics, 2006. **2**: p. 117693510600200004.
243. Louppe, G., et al. *Understanding variable importances in forests of randomized trees*. in *Advances in neural information processing systems*. 2013.
244. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
245. Chang, C.-C. and C.-J. Lin, *Training v-support vector regression: theory and algorithms.* Neural computation, 2002. **14**(8): p. 1959-1977.
246. Hawkins, T. and D. Kihara, *PFP:Automatic annotation of protein function by relative GO association in multiple functional contexts.* The 13th Annual International Conference on Intelligent Systems for Molecular Biology, 2005: p. 117.
247. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-410.
248. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nat Methods, 2013. **10**(3): p. 221-7.
249. Jiang, Y., et al., *An expanded evaluation of protein function prediction methods shows an improvement in accuracy.* Genome Biol, 2016. **17**(1): p. 184.
250. Hawkins, T., M. Chitale, and D. Kihara, *Functional enrichment analyses and construction of functional similarity networks with high confidence function prediction by PFP.* BMC Bioinformatics, 2010. **11**: p. 265.
251. Barabasi, A.-L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nature reviews. Genetics, 2004. **5**(2): p. 101.
252. Clauset, A., C.R. Shalizi, and M.E. Newman, *Power-law distributions in empirical data.* SIAM review, 2009. **51**(4): p. 661-703.
253. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-D361.
254. Aryal, U.K., et al., *A proteomic strategy for global analysis of plant protein complexes.* Plant Cell, 2014. **26**(10): p. 3867-82.
255. Aryal, U.K., et al., *Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling.* Journal of Proteomics, 2017.
256. Perea-Resa, C., et al., *LSM proteins provide accurate splicing and decay of selected transcripts to ensure normal Arabidopsis development.* The Plant Cell, 2012: p. tpc. 112.103697.
257. Golisz, A., et al., *Arabidopsis thaliana LSM proteins function in mRNA splicing and degradation.* Nucleic acids research, 2013. **41**(12): p. 6232-6249.

258. Glynn, J.M., J.E. Froehlich, and K.W. Osteryoung, *Arabidopsis ARC6 coordinates the division machineries of the inner and outer chloroplast membranes through interaction with PDV2 in the intermembrane space.* The Plant Cell, 2008. **20**(9): p. 2460-2470.
259. Luo, M., et al., *Histone deacetylase HDA6 is functionally associated with AS1 in repression of KNOX genes in Arabidopsis.* PLoS genetics, 2012. **8**(12): p. e1003114.
260. Renfrew, K.B., et al., *POT1a and components of CST engage telomerase and regulate its activity in Arabidopsis.* PLoS genetics, 2014. **10**(10): p. e1004738.
261. Kotera, E., M. Tasaka, and T. Shikanai, *A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts.* Nature, 2005. **433**(7023): p. 326.
262. Smoot, M.E., et al., *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics, 2010. **27**(3): p. 431-432.
263. Khan, I.K., et al., *Prediction of protein group function by iterative classification on functional relevance network.* Bioinformatics, 2018.
264. Arifuzzaman, M., et al., *Large-scale identification of protein–protein interaction of Escherichia coli K-12.* Genome research, 2006. **16**(5): p. 686-691.
265. Sato, S., et al., *A large-scale protein–protein interaction analysis in Synechocystis sp. PCC6803.* DNA research, 2007. **14**(5): p. 207-216.
266. Li, Z., et al., *Large-scale identification of human protein function using topological features of interaction network.* Scientific Reports, 2016. **6**: p. 37179.
267. Rual, J.-F., et al., *Towards a proteome-scale map of the human protein–protein interaction network.* Nature, 2005. **437**(7062): p. 1173-1178.
268. Rajagopala, S.V., et al., *The binary protein-protein interaction landscape of Escherichia coli.* Nature biotechnology, 2014. **32**(3): p. 285-290.
269. Zhang, J., et al., *An improved approach to infer protein-protein interaction based on a hierarchical vector space model.* BMC bioinformatics, 2018. **19**(1): p. 161.
270. Bandyopadhyay, S. and K. Mallick, *A new feature vector based on gene ontology terms for protein-protein interaction prediction.* IEEE/ACM transactions on computational biology and bioinformatics, 2017. **14**(4): p. 762-770.
271. Aryal, U.K., et al., *Proteomic analysis of protein complexes in photosynthetic cyanobacteria Cyanothece ATCC 51142.* Journal of Proteome Research, 2018. **Submitted**.
272. Elvitigala, T., et al., *Effect of continuous light on diurnal rhythms in Cyanothece sp. ATCC 51142.* BMC Genomics, 2009. **10**: p. 226.
273. Dutta, D., et al., *Hydrogen production by Cyanobacteria.* Microb Cell Fact, 2005. **4**: p. 36.
274. Ghirardi, M.L., et al., *Microalgae: a green source of renewable H(2).* Trends Biotechnol, 2000. **18**(12): p. 506-11.
275. Nozzi, N.E., J.W. Oliver, and S. Atsumi, *Cyanobacteria as a Platform for Biofuel Production.* Front Bioeng Biotechnol, 2013. **1**: p. 7.
276. Zehr, J.P., et al., *Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean.* Nature, 2001. **412**(6847): p. 635-8.
277. Reddy, K.J., et al., *Unicellular, aerobic nitrogen-fixing cyanobacteria of the genus Cyanothece.* J Bacteriol, 1993. **175**(5): p. 1284-92.
278. Stockel, J., et al., *Diurnal rhythms result in significant changes in the cellular protein complement in the cyanobacterium Cyanothece 51142.* PLoS One, 2011. **6**(2): p. e16680.
279. Welsh, E.A., et al., *The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle.* Proc Natl Acad Sci U S A, 2008. **105**(39): p. 15094-9.

280. Schneegurt, M.A., D.M. Sherman, and L.A. Sherman, *Composition of the carbohydrate granules of the cyanobacterium, Cyanothece sp. strain ATCC 51142.* Arch Microbiol, 1997. **167**(2-3): p. 89-98.
281. Sherman, L.A., P. Meunier, and M.S. Colon-Lopez, *Diurnal rhythms in metabolism: A day in the life of a unicellular, diazotrophic cyanobacterium.* Photosynthesis Research, 1998. **58**(1): p. 25-42.
282. Schneegurt, M.A., et al., *Oscillating behavior of carbohydrate granule formation and dinitrogen fixation in the cyanobacterium Cyanothece sp. strain ATCC 51142.* J Bacteriol, 1994. **176**(6): p. 1586-97.
283. Bandyopadhyay, A., et al., *Novel metabolic attributes of the genus cyanothece, comprising a group of unicellular nitrogen-fixing Cyanothece.* MBio, 2011. **2**(5).
284. Toepel, J., et al., *Differential transcriptional analysis of the cyanobacterium Cyanothece sp. strain ATCC 51142 during light-dark and continuous-light growth.* J Bacteriol, 2008. **190**(11): p. 3904-13.
285. Stockel, J., et al., *Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes.* Proc Natl Acad Sci U S A, 2008. **105**(16): p. 6156-61.
286. Aryal, U.K., et al., *Proteomic profiles of five strains of oxygenic photosynthetic cyanobacteria of the genus Cyanothece.* J Proteome Res, 2014. **13**(7): p. 3262-76.
287. Aryal, U.K., et al., *Dynamic proteomic profiling of a unicellular cyanobacterium Cyanothece ATCC51142 across light-dark diurnal cycles.* BMC Syst Biol, 2011. **5**: p. 194.
288. Aryal, U.K., et al., *Dynamic proteome analysis of Cyanothece sp. ATCC 51142 under constant light.* J Proteome Res, 2012. **11**(2): p. 609-19.
289. McDermott, J.E., et al., *Defining the players in higher-order networks: predictive modeling for reverse engineering functional influence networks.* Pac Symp Biocomput, 2011: p. 314-25.
290. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-26.
291. Jansen, R., et al., *A Bayesian networks approach for predicting protein-protein interactions from genomic data.* Science, 2003. **302**(5644): p. 449-53.
292. Dunham, W.H., M. Mullin, and A.C. Gingras, *Affinity-purification coupled to mass spectrometry: basic principles and strategies.* Proteomics, 2012. **12**(10): p. 1576-90.
293. Altelaar, A.F., J. Munoz, and A.J. Heck, *Next-generation proteomics: towards an integrative view of proteome dynamics.* Nat Rev Genet, 2013. **14**(1): p. 35-48.
294. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nat Biotechnol, 1999. **17**(10): p. 1030-2.
295. Du, C., et al., *Dinitrogenase reductase ADP-ribosyl transferase and dinitrogenase reductase activating glycohydrolase in Gloeothece.* Biochem Soc Trans, 1994. **22**(3): p. 332S.
296. Wodak, S.J., et al., *Challenges and rewards of interaction proteomics.* Mol Cell Proteomics, 2009. **8**(1): p. 3-18.
297. Dong, M., et al., *A "tagless" strategy for identification of stable protein complexes genome-wide by multidimensional orthogonal chromatographic separation and iTRAQ reagent tracking.* J Proteome Res, 2008. **7**(5): p. 1836-49.

298. Guerreiro, A.C., et al., *Monitoring light/dark association dynamics of multi-protein complexes in cyanobacteria using size exclusion chromatography-based proteomics.* J Proteomics, 2016. **142**: p. 33-44.
299. Aryal, U.K., et al., *Analysis of protein complexes in Arabidopsis leaves using size exclusion chromatography and label-free protein correlation profiling.* J Proteomics, 2017. **166**: p. 8-18.
300. Olinares, P.D., L. Ponnala, and K.J. van Wijk, *Megadalton complexes in the chloroplast stroma of Arabidopsis thaliana characterized by size exclusion chromatography, mass spectrometry, and hierarchical clustering.* Mol Cell Proteomics, 2010. **9**(7): p. 1594-615.
301. Kirkwood, K.J., et al., *Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics.* Mol Cell Proteomics, 2013. **12**(12): p. 3851-73.
302. Kristensen, A.R., J. Gsponer, and L.J. Foster, *A high-throughput approach for measuring temporal changes in the interactome.* Nat Methods, 2012. **9**(9): p. 907-9.
303. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
304. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
305. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
306. Polpitiya, A.D., et al., *DAnTE: a statistical tool for quantitative analysis of -omics data.* Bioinformatics, 2008. **24**(13): p. 1556-8.
307. Fujisawa, T., et al., *CyanoBase: a large-scale update on its 20th anniversary.* Nucleic Acids Res, 2017. **45**(D1): p. D551-D554.
308. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2012. **40**(Database issue): p. D48-53.
309. Chang, C.C. and C.J. Lin, *LIBSVM: A Library for Support Vector Machines.* Acm Transactions on Intelligent Systems and Technology, 2011. **2**(3).
310. Liu, X.P., et al., *Toward chromatographic analysis of interacting protein networks.* Journal of Chromatography A, 2008. **1178**(1-2): p. 24-32.
311. Gao, Q., et al., *Coupling protein complex analysis to peptide based proteomics.* J Chromatogr A, 2010. **1217**(49): p. 7661-8.
312. Tucker, D.L. and L.A. Sherman, *Analysis of chlorophyll-protein complexes from the cyanobacterium Cyanothece sp. ATCC 51142 by non-denaturing gel electrophoresis.* Biochim Biophys Acta, 2000. **1468**(1-2): p. 150-60.
313. Pancholi, V., *Multifunctional alpha-enolase: its role in diseases.* Cell Mol Life Sci, 2001. **58**(7): p. 902-20.
314. O'Leary, B., et al., *Bacterial-type phosphoenolpyruvate carboxylase (PEPC) functions as a catalytic and regulatory subunit of the novel class-2 PEPC complex of vascular plants.* J Biol Chem, 2009. **284**(37): p. 24797-805.
315. Dhanyalakshmi, K.H., et al., *An Approach to Function Annotation for Proteins of Unknown Function (PUFs) in the Transcriptome of Indian Mulberry.* PLoS One, 2016. **11**(3): p. e0151323.

316. Rust, M.J., S.S. Golden, and E.K. O'Shea, *Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator.* Science, 2011. **331**(6014): p. 220-3.
317. Battchikova, N., M. Eisenhut, and E.M. Aro, *Cyanobacterial NDH-1 complexes: novel insights and remaining puzzles.* Biochim Biophys Acta, 2011. **1807**(8): p. 935-44.
318. Battchikova, N. and E.M. Aro, *Cyanobacterial NDH-1 complexes: multiplicity in function and subunit composition.* Physiol Plant, 2007. **131**(1): p. 22-32.
319. Kosuge, T. and T. Hoshino, *Construction of a proline-producing mutant of the extremely thermophilic eubacterium Thermus thermophilus HB27.* Applied and environmental microbiology, 1998. **64**(11): p. 4328-4332.
320. Peterson, L.X., et al., *Modeling disordered protein interactions from biophysical principles.* PLoS Comput Biol, 2017. **13**(4): p. e1005485.
321. Esquivel-Rodriguez, J., Y.D. Yang, and D. Kihara, *Multi-LZerD: multiple protein docking for asymmetric complexes.* Proteins, 2012. **80**(7): p. 1818-33.
322. Venkatraman, V., et al., *Protein-protein docking using region-based 3D Zernike descriptors.* BMC Bioinformatics, 2009. **10**: p. 407.
323. Obayashi, T., et al., *COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference.* Nucleic acids research, 2018. **47**(D1): p. D55-D62.
324. Obayashi, T., et al., *ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index.* Plant and Cell Physiology, 2017. **59**(1): p. e3-e3.
325. Sun, T., et al., *Sequence-based prediction of protein protein interaction using a deep-learning algorithm.* BMC bioinformatics, 2017. **18**(1): p. 277.
326. Du, X., et al., *DeepPPI: boosting prediction of protein–protein interactions with deep neural networks.* Journal of chemical information and modeling, 2017. **57**(6): p. 1499-1510.
327. Chen, M., et al., *Multifaceted protein–protein interaction prediction based on Siamese residual RCNN.* Bioinformatics, 2019. **35**(14): p. i305-i314.
328. Zhou, H. and J. Skolnick, *GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction.* Biophysical journal, 2011. **101**(8): p. 2043-2052.
329. Zhang, C., S. Liu, and Y. Zhou, *Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential.* Protein science, 2004. **13**(2): p. 391-399.
330. Huang, S.Y. and X. Zou, *An iterative knowledge-based scoring function for protein–protein recognition.* Proteins: Structure, Function, and Bioinformatics, 2008. **72**(2): p. 557-579.
331. Dong, G.Q., et al., *Optimized atomic statistical potentials: assessment of protein interfaces and loops.* Bioinformatics, 2013. **29**(24): p. 3158-3166.
332. Pagès, G., B. Charmettant, and S. Grudinin, *Protein model quality assessment using 3D oriented convolutional neural networks.* bioRxiv, 2018: p. 432146.
333. Hou, J., R. Cao, and J. Cheng, *Deep convolutional neural networks for predicting the quality of single protein structural models.* bioRxiv, 2019: p. 590620.
334. Cao, R., et al., *DeepQA: improving the estimation of single protein model quality with deep belief networks.* BMC bioinformatics, 2016. **17**(1): p. 495.
335. Derevyanko, G., et al., *Deep convolutional networks for quality assessment of protein folds.* Bioinformatics, 2018. **34**(23): p. 4046-4053.

336. Singh, R., et al., *Struct2Net: a web service to predict protein–protein interactions using a structure-based approach.* Nucleic acids research, 2010. **38**(suppl_2): p. W508-W515.

337. Morcos, F., et al., *Direct coupling analysis for protein contact prediction*, in *Protein structure prediction*. 2014, Springer. p. 55-70.

338. Rosell, M. and J. Fernández-Recio, *Hot-spot analysis for drug discovery targeting protein-protein interactions.* Expert opinion on drug discovery, 2018. **13**(4): p. 327-338.

A. SUPPLEMENTAL INFORMATION FOR CHAPTER 4

Supplemental Table S4.1 (in a separate Excel file) The experimentally verified protein-protein interactions (PPIs) downloaded from TAIR database. Only the 4,776 PPIs identified by physical experimental systems are included. The PPIs only identified by genetic experimental systems are discarded and they are listed at the right side of the table highlighed in grey. After removing the protein length less than 50 residues, only 4,759 PPIs are used for training. The number in each cell indicates the number of the same type of experiment which identifies such PPI.

Supplemental Table S4.2, S4.3, S4.4 (in a separate Excel file) Predicted PPIs in *Arabidopsis thaliana* with different levels of confidence.

S4.2: PPI predictions with two additional evidence in *Arabidopsis thaliana*.
S4.3: PPI predictions with additional evidence in *Arabidopsis thaliana.*
S4.4: PPI predictions with no known evidence in *Arabidopsis thaliana.*

Supplemental Table S4.5, S4.6, S4.7 (in a separate Excel file) Predicted PPIs in *Zea mays* (maize) with different levels of confidence.

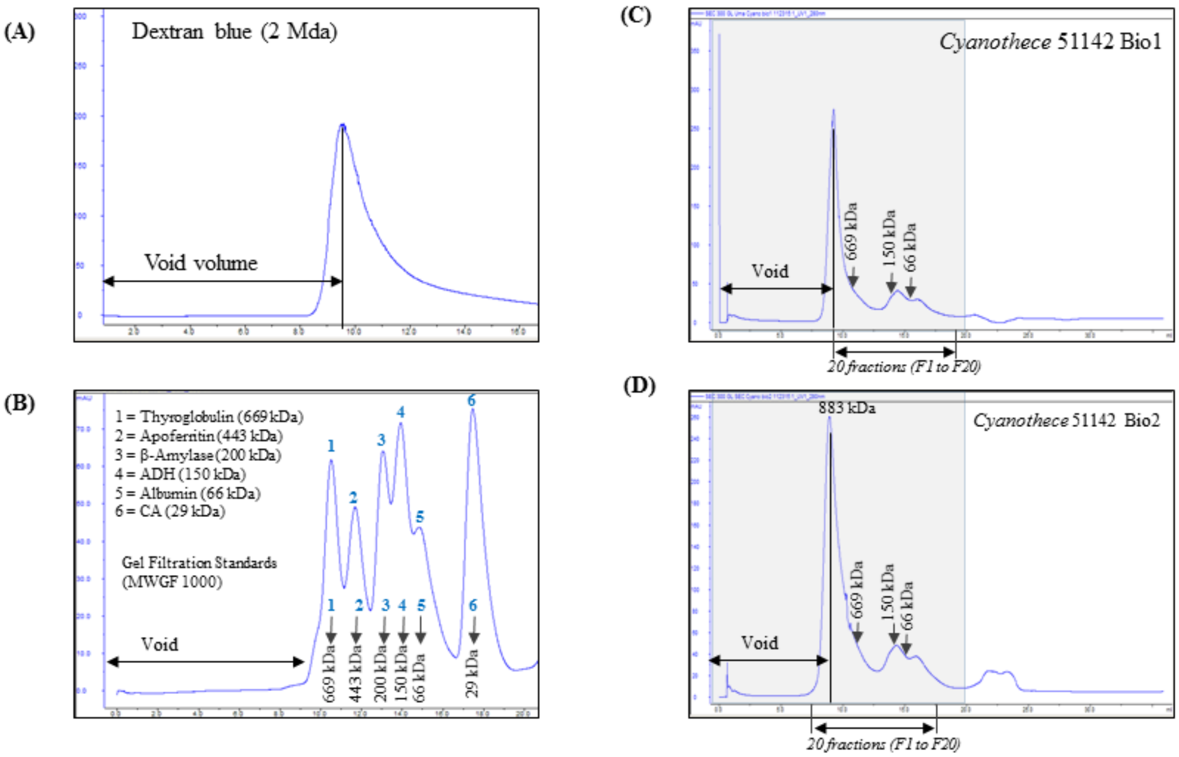
S4.5: PPI predictions with two additional evidence in *Zea mays* (maize).
S4.6: Predicted PPIs with additional evidence in *Zea mays* (maize).
S4.7: Predicted PPIs with no known evidence in *Zea mays* (maize).

Supplemental Table S4.8, S4.9, S4.10 (in a separate Excel file) Predicted PPIs in *Glycine max* (soybean) with different levels of confidence.

S4.8: PPI predictions with two additional evidence in *Glycine max* (soybean).
S4.9: Predicted PPIs with an additional evidence in *Glycine max* (soybean).
S4.10: Predicted PPIs with no other known evidence in *Glycine max* (soybean).

B. SUPPLEMENTAL INFORMATION FOR CHAPTER 5

Supplementary Figure S5.1 Size Exclusion Chromatography (SEC) protein elution profiles. (A) Dextran blue elution peak. (B) Elution peaks of six protein standards. (C and D) Elution profiles of *Cyanothece* 51142 native proteins in biological replicate 1 and 2, respectively.



Supplementary Table S5.1 (in a separate Excel file) List of peptides commonly identified in duplicate biological runs with matched proteins/protein groups, intensity and MS/MS counts.

Supplementary Table S5.2 (in a separate Excel file) List of proteins commonly identified in duplicate biological runs with their intensity profiles, M_{app} and R_{app} values.

Supplementary Table S5.3 (in a separate Excel file) List of computationally predicted protein-protein interactions.

VITA

Ziyun Ding was born and raised in Xinjiang. She attended Hua Shan high school, where she expressed interests in math and physics. She went to China Agricultural University at Beijing China for freshman and sophomore, and transferred to Purdue University studying plant biology. She did junior year summer research with Dr. Wanqing Liu and did senior year research with Dr. Nicholas Carpita and Dr. Maureen McCann. Right after her undergraduate, she began her PhD studies at Purdue University in August 2013. Meanwhile, she also finished Applied Statistic Master Program at Purdue University.