Bioinformatic characterization of plant networks

Jason McDermott[1] and Ram Samudrala[2]

[1]*Computational Biology and Bioinformatics Group*

Pacific Northwest National Laboratory, MSIN K7-90

Richland, Washington 99352

[2]*Department of Microbiology, Box 357242*

University of Washington

Seattle, Washington 98195

Phone: 1-206-732-6122

FAX: 1-206-732-6055

Corresponding author: ram@compbio.washington.edu

Abstract

Cells and organisms are governed by networks of genetic, physical and metabolic interactions between proteins and other their substrates, such as DNA, RNA, and other biologically important molecules. Large scale experimental studies of interactions between components of biological systems have been performed for a variety of eukaryotic organisms. However, there is a dearth of such data for plants. Computational methods for prediction of relationships between proteins, primarily based on comparative genomics, i.e., using homology to transfer information between and across organisms, provide useful molecular and systems level views of cellular function and can be used to integrate and extend information available about other eukaryotes to plants. We have predicted protein-protein and protein-DNA networks for six proteomes of *Oryza sativa, Arabidopsis thaliana*, and several plant pathogens using the Bioverse framework (http://bioverse.compbio.washington.edu), that relates three-dimensional atomic level detail of information regarding single molecules to systems, cellular, and organismal biology. We show that our predictions are similar to experimentally derived interactions, but provide greater coverage for a larger number of proteins. Predicted interaction networks for plants can also be used to provide novel functional annotations and predictions about plant biochemical pathways to aid in rational engineering, either by genetic modification and/or marker-assisted breeding, to improve human heath and quality of life.

Introduction

The cell can be envisioned as a complicated machine . Parts of the machine correspond to biological components such as proteins, nucleic acids, and small molecules such as metabolites and ions that serve as enzymatic substrates, inhibitors, and cofactors. These parts function as members of complexes and pathways that are interconnected in a large network. The structure and dynamics of this network allow the cell to function both internally and as a part of a larger organism as it interacts with and adapts to its constantly changing environment. The advent of high-throughput experimental methodologies to generate interaction data has spawned a paradigm shift regarding biological networks or pathways. Whereas it was only possible to determine the local structure of a portion of the network (for example, proteins that function together in a single complex), it has become

now possible to consider the networks on a more holistic level that includes thousands of proteins and millions of interacting components.

Two of the most established high-throughput methods for interaction determination are the yeast two-hybrid (Y2H) method and tandem affinity purification (TAP) . Partial protein interaction networks for a few eukaryotic organisms such as *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans* have been experimentally determined by combining results from a variety of such methods . A number of databases have been established to catalog these protein-protein interactions including the Biomolecular Interaction Network Database (BIND;), the Database of Interacting Proteins (DIP;), and the Human Proteome Research Database (HPRD;). Although protein interaction networks are being determined for higher eukaryotes, very few experiments in plants have shed light on protein-protein interactions for more than a few interactions at a time. Several experimental methods were combined to elucidate complex membership for around 2,500 proteins in *Oryza sativa* (rice) and a similar approach was used to characterize several hundred proteins complexes in *Triticum* spp. (wheat) . Interactions for several hundred proteins involved in stress response, seed development, and cyclin-related networks in rice were studied using a combination of two-hybrid genetic screens and expression techniques . Finally, a method using two-dimensional gel electrophoresis to screen chromosomal deletions was used to analyze protein expression correlation in wheat and the results were extrapolated to predict protein-protein interactions . Transcriptomics and metabolomics have been more common approaches applied to plants (see for a recent review), but these studies do not elucidate protein-protein interactions directly, and thus provide large gaps in understanding the mechanistic basis, or "wiring diagrams" of plant interactomes.

In the Bioverse framework , we have used a number of computational techniques to predict protein-protein interaction networks for several plants and plant pathogens by the use homology-based transfer of experimentally determined interactions from other organisms. We have demonstrated that these interaction networks are useful for functional annotation of proteins with predicted interaction but no known function , and provide a mechanistic detail of plant interactomes.

Methods

Experimental characterization of interaction networks is expensive and requires a large amount of time and effort. The focus of these investigations has been on model organisms, or on particular pathways or complexes of interest. Therefore, computational methods have been developed to predict functional relationships and physical interactions in novel organisms using comparative genomics techniques . These methods exploit evolutionary relationships between organisms to either directly infer relationships (co-conservation of pairs of proteins, for instance) or to extend experimental by homology-based transfer data from different organisms to a target organism. Predictions can then be used to focus the efforts of bench experimentalists, with a significant reduction in cost and time taken.

For plants, large scale computational predictions of protein interactions and relationships have been generated by several groups. STRING provides computational predictions of protein associations for some plant species. VisANT/Predictome provides computational predictions of protein associations for a number of organisms, but contains only experimentally determined interactions for plants. AraCyc , for *Arabidopsis thaliana*, and KEGG use a combination of orthologous relationships and experimental data, when available, to assign proteins from plant species to metabolic and signaling pathways (see for a recent review of pathway resources). Additionally, the phenylpropanoid pathway in *Arabidopsis thaliana* was analyzed using orthologous mapping and gene co-expression was used to predict networks for barley .

The Bioverse uses several methods to predict physical protein-protein interactions including phylogenetic profile correlation , domain fusion and extension of experimental results using evolutionary relationships (i.e., homology-based transfer). If two proteins that have been experimentally shown to interact in one organism both have orthologs in another organism an interaction, or interolog, can be predicted between the orthologs . In the interolog method as implemented by the Bioverse, sequence similarity between all protein sequences from a target organism and sequences in several databases of protein interactions is determined using PSI-BLAST . The organism is then examined for all occurrences of pairs of proteins that are orthologs of respective partners from a known interaction. The

predicted relationship in the target organism is the interolog of the experimental interaction and an interolog score (IS) is assigned as the product of both similarity measures. Source databases for experimental interactions used by the Bioverse include the Biomolecular Interaction Network Database (BIND;) and BIND's dataset of interactions derived from crystallized structures MMDBBind , the Database of Interacting Proteins (DIP;), and the Human Proteome Research Database (HPRD;). They contain interactions determined from many types of experimental methods, the most prevalent being Y2H and TAP, and cover a large number of diverse organisms. Further details of the method are given elsewhere [17].

Results and Discussion

Several protein interaction networks from plants and plant pathogens have been predicted in the Bioverse and are listed in Table 1. It is clear from this table that plants have far fewer predictions than for some other higher eukaryotes. This is because most of the predicted interactions are based on homology to proteins with experimentally determined interactions, and very few such interactions have been determined for plants. As more experimental interaction data becomes, the accuracy and coverage of the predicted interaction networks will increase. Even with current coverage rates, the predicted interaction networks in plants can still be used to provide novel annotations for a significant number of proteins (Table 1).

In addition to a general overview, our analysis provides more detailed predictions of interaction pathways that can be used to form hypotheses and guide experimental investigation. A portion of the predicted protein interaction network from *O. sativa* is shown in Figure 1. The network is centered on those proteins annotated by Interpro as being plant defense proteins. This figure shows how predicted interaction networks can be used to examine potentially novel interactions between components of known and unknown function. The figure also shows those proteins that had no annotation by Interpro but were able to annotated by our network annotation method outlined in green. It should be noted that the Bioverse framework has been used to predict structures, functions, and interactions of nearly 500,000 proteins from more than 50 proteomes, with over 3 million predicted interactions. Among these, nearly 200,000 molecules are from rice, with over 500,000

predicted interactions (see http://bioverse.compbio.washington.edu for a full list of organisms and detailed information about the structure, function, and interaction of each protein molecule)..

Conclusions

Predicted protein interaction networks provide a starting point for experimental investigation in organisms with little experimental evidence. This is especially important in plants where the amount of experimental information available in public resources is minimal, and it is unlikely that such information will be available shortly. The Bioverse has predicted interaction networks for a number of plants and plant pathogens. These networks have been used to provide functional annotations for proteins which were not amenable to annotation by standard computational approaches. Further, these network provide a mechanistic basis of plant biochemical pathways that can be used for reengineering of plants, particularly those that are important food sources, to provide a full range of bioavailable nutrients.

Table I. Prediction details of protein interactions for selected proteomes in the Bioverse.

Organism	Number of proteins	Proteins in network	Predicted interactions	Coverage	Network-annotated
A. thaliana	**27,833**	**699**	**2,959**	**3%**	**322**
O. sativa indica	**40,925**	**2,756**	**28,003**	**7%**	**250**
O. sativa japonica	**36,658**	**2,677**	**31,557**	**7%**	**245**
A. tumefaciens	5,396	569	1,357	10%	153
C. familiaris	16,817	5,785	39,302	35%	1,436
D. melanogaster	16,475	13,290	405,812	80%	326
M. grisea	11,042	2,248	12,261	20%	1,730
Total (54 proteomes)	**486,530**	**145,390**	**3,410,936**	**31%**	**42,482**

Shown for each organism is the **number of proteins** in the proteome; the **number of proteins in the predicted interaction network** with interolog score (IS) greater than 0.15; the **number of predicted protein interactions** with IS greater than 0.15; the **coverage** or percentage of proteins in the proteome with predicted interactions; and the number of proteins with no significant functional annotation that could be annotated using our **network annotation** protocol [16] with an estimated accuracy of 30% or more. Plant proteomes are listed on top in bold and a summary line for all 54 proteomes analyzed in the Bioverse is also provided (this includes six rice proteomes consisting of nearly 200,000 molecules with over 500,000 predicted interactions).

Figure 1. Proteins involved in defense in *Oryza sativa* (rice). A portion of the protein interaction network predicted using the Bioverse framework for *Oryza sativa* (rice). Proteins are depicted as nodes and predicted protein interactions as edges. Proteins are colored by general functional category as shown in the legend. Proteins with no preexisting functional annotation but capable of being annotated using our network-based functional annotation algorithm are outlined in bold green. This example illustrates how the functions of key pathways and networks in important food crops such as rice can be elucidated using the Bioverse framework, and how these networks can be reengineered to provide greater robustness.