Application of Genomics and Bioinformatics in Plants
Eds. G.P. Rao, C. Wagner & R.K. Singh
Studium Press LLC, Texas, USA, 2008, pp. 207-238.

CHAPTER

9

Prediction and Annotation of Plant Protein Interaction Networks

JASON MCDERMOTT[1,2], JUN WANG[4], JUN YU[4], GANE KA-SHU WONG[3,4] & RAM SAMUDRALA[2★]

ABSTRACT

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, provide a useful systems-level view of cellular functioning and can be used to extend information about other eukaryotes to plants. We have predicted networks for *Arabidopsis thaliana*, *Oryza sativa indica* and *japonica* and several plant pathogens using the Bioverse (http://bioverse.compbio.washington.edu) and show that they are similar to experimentally-derived interaction networks. Predicted interaction networks for plants can be used to provide novel functional annotations and predictions about plant phenotypes and aid in rational engineering of biosynthesis pathways.

Key Words : Protein interaction, computational methods, networks, comparative genomics, plant pathways

[1]*Computational Biology and Bioinformatics, Pacific Northwest National Laboratory, MSIN K7-90, 902 Battelle Boulevard, P.O. Box 999, Richland, WA 99352*
[2]*Department of Microbiology Mail 357242*
[3]*Department of Biological Sciences and Department of Medicine, University of Alberta, Edmonton AB, T6G 2E9, Canada*
[4]*Beijing Institute of Genomics of the Chinese Academy of Science, Beijing Genomics Institute, Beijing Proteomics Institute, Beijing, China. James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou Genomics Institute, Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou, China.*
J.M. : Jason.McDermott@pnl.gov *J.W. : wangj@genomics.org.cn*
J.Y. : junyu@genomics.org.cn *G.KS.W. : gane@ualberta.ca*
★*E-mail : R.S. : ram@compbio.washington.edu*

1. INTRODUCTION

The genetic engineering of crops for increased yield, pest and disease resistance, and for increased nutritional value has been occurring for approximately 10,000 years (Paterson *et al.*, 2003). Advances in the biological sciences have accelerated this process and made it more knowledge-based, allowing a greater range of possibilities (Tzfira and White, 2005). In the post-genomic era, understanding of cellular systems on a global scale is becoming increasingly important for biologists. Recent advances have made this problem more approachable as methodology enabling large-scale experimental studies of eukaryotes has been developed (Uetz *et al.*, 2000; Ho *et al.*, 2002; Bhalla *et al.*, 2005; Dong *et al.*, 2005; Rensink and Buell, 2005). These methods have generated data of different types for a large number of organisms but data covering more than just a few complexes or pathways is limited to a handful of model organisms. As a result, computational bioinformatics methods have been developed to integrate this data and to extrapolate from it to provide predictions for proteins and organisms not yet experimentally well characterized (Date and Marcotte, 2003; Morett *et al.*, 2003; Strong *et al.*, 2003; McDermott and Samudrala, 2004; Yu *et al.*, 2004; Wichadakul *et al.*, 2007).

The cell functions as a kind of machine (Alberts, 1998); all components of the cell, small molecules, DNA, RNA, proteins, are parts of the machine and the ways that these parts work together determine how the machine functions as a whole. A single part of the machine doesn't function by itself: The function of the part is dictated by how it works with the other parts of the machine. Protein complexes and interactions form a network which comprises the complicated inner workings of this machine and is essential to the cell's ability to walk the thermodynamic line between order and disorder.

Many types of relationships exist between components and various bioinformatics resources have been created to organize and use the data related to these relationships (Frishman *et al.*, 2001; Mewes *et al.*, 2004; Birkland and Yona, 2006). Our own project, the Bioverse (http://bioverse.compbio.washington.edu), is a computational framework for the organization, representation and integration of molecular, cellular and organismal worlds (McDermott and Samudrala, 2003; Guerquin *et al.*, 2007). We have used it to provide detailed functional annotations for a number of organisms including *Arabidopsis thaliana* and several strains of rice (Kikuchi *et al.*, 2003; Yu *et al.*, 2005). We have also used it to investigate relationships between components and to extend analysis to largely uncharacterized organisms, as discussed in this review.

2. DETERMINATION OF PROTEIN INTERACTIONS

Traditional biochemical and genetic approaches have focused on a few specific interactions pertaining to a single protein, complex or pathway. The impetus of proteomics efforts has produced high-throughput methods for characterization of large numbers of proteins and interactions at once. Partial protein interaction networks for several organisms have been experimentally determined by combining results from a variety of such methods (Ge *et al.*, 2001; Li *et al.*, 2004). Two of the most established high-throughput methods for interaction determination are the yeast two-hybrid method (Chien *et al.*, 1991) and tandem affinity purification (TAP) (Ho *et al.*, 2002). Microarray analysis of mRNA levels can also provide an indication of association between genes and proteins, and is relatively easy to perform (Teichmann and Babu, 2002; Rensink and Buell, 2005). These approaches have been used to characterize specific pathways (Drees *et al.*, 2001; Rivas *et al.*, 2002) and to derive large scale protein interaction networks for the several eukaryotic organisms (Uetz *et al.*, 2000; Walhout *et al.*, 2000; Ho *et al.*, 2002; Giot *et al.*, 2003).

Few experiments in plants have shed light on protein-protein interactions for more than a few interactions at a time. Several proteomics methods were combined to provide an overview of approximately 2500 proteins involved in complexes in rice (Koller *et al.*, 2002) and a similar approach was used to characterize several hundred complexed proteins in wheat amyloplasts (Andon *et al.*, 2002). Stress response and seed development (approx. 200 proteins), and cyclin-related networks (approx. 150 proteins) in rice were studied with two-hybrid and expression techniques (Cooper *et al.*, 2003; Cooper *et al.*, 2003). Finally, a genetic technique involving screening of chromosomal deletions coupled with MS or 2D electrophoresis to analyze protein expression correlation has been used to suggest protein-protein interactions in wheat (Islam *et al.*, 2003). Metabolomics, experimental analysis of metabolic networks, and expression-based experiments have been more common for plants (see (Bhalla *et al.*, 2005) for a recent review), but these studies do not elucidate protein-protein interactions directly.

High-throughput methods are necessarily error-prone. Various factors including protein promiscuity (the tendency of some proteins to interact with many proteins in a non-functional manner), variations in experimental conditions and inherent systematic noise in each method all contribute to this error. Examination of the correspondence between data derived using different methodologies has shown that

the false positive error rate of high-throughput interaction data is as high as 50% (von Mering *et al.*, 2002). Examining the overlap between different experimental data (von Mering *et al.*, 2002) and integrating various genomic features (Lu *et al.*, 2005) can improve the accuracy of these methods but comes with a corresponding decrease in number of interactions represented in the high-confidence data sets. The yeast network is the most well-characterized and 60-70% of proteins in yeast have been shown to be involved in at least one interaction by at least one experimental interaction (Bork, 2002; von Mering *et al.*, 2002).

2.1. Applications of protein interaction networks

Protein interaction and metabolic networks have particular structure dictated by evolutionary processes. In the most essential sense they must function to ensure the survival of the organism with which they are associated. But how to best accomplish this goal? It appears that biological networks are fairly error tolerant. That is, if a component of the network (a node) is removed the structure of the network remains largely unchanged (Albert *et al.*, 2000; Jeong *et al.*, 2000), in general. This seems to be accomplished through a scale-free organization. That is, there are many components in the network with very few connections and only a few components with many connections in a power-law distribution (e.g. Figure 3) (Girvan and Newman, 2002; Ravasz *et al.*, 2002). This architecture protects against removal of a component by random processes since in all likelihood this will be a non-essential component. The networks also seem to be modular in nature. There are fairly discrete regions of the network which perform a particular function or range of related functions (Snel *et al.*, 2002; Rives and Galitski, 2003; Spirin and Mirny, 2003) and these regions may be linked through a small number of connections.

Consideration of biological information at the systems level has been applied to elucidate different aspects of cellular biology and evolution. Even with the partial networks currently available, characteristics of a protein's interacting neighbors can be informative. The protein's position in the network, say as the sole link between two subgraphs or pathways, and more global properties of the network have been investigated as well. Each of these network features may be correlated with some characteristic of the protein(s) involved or with a phenotype expressed at the cellular or organismal levels.

One study reported that the connectivity of a protein in the yeast network (i.e. how many interacting partners it has) correlates well with the importance of the protein as judged by the likelihood that it will be lethal if removed by mutagenesis (Jeong *et al.*, 2001). Another study

in yeast found that more connected proteins evolved more slowly than those that had fewer interactions (Fraser *et al.*, 2002). Both these correlations were also reported in the *C. elegans* and *D. melanogaster* protein interaction networks (Hahn and Kern, 2005). These studies also showed that global network features such as 'betweeness' and 'closeness' were well correlated with lethality and evolutionary rates. Betweeness is a measure of to what degree a protein is a bottleneck for information flow in a network. It is calculated as the percentage of times that a protein appears in the shortest path between all pairs of proteins in the network. Closeness is calculated as the average number of proteins separating a protein and all other proteins in the network. One conclusion drawn from these studies is that more highly connected proteins evolve more slowly not only since they are more important for organism survival but also because more of the sequence of such proteins is devoted to interacting with other proteins (Fraser *et al.*, 2002).

Proteins involved in the process of aging from yeast, fly and worm were also shown to be more highly connected than non-aging related proteins (Pletcher, 2004; Promislow, 2004; Ferrarini *et al.*, 2005). Additionally, a recent report showed that the same network properties hold true for toxicity-modulating proteins in yeast. The study, which examined the effects of DNA-damaging agents on yeast single-gene deletion strains covering the 4,733 nonessential proteins, found that proteins involved in recovery from DNA-damaging agents were more highly connected and more globally central than proteins that were not, similar to essential proteins (Said *et al.*, 2004). This study is particularly exciting since, if extended to plants, these results could be used to design crops with better resistance to herbicides, or to design novel herbicides with better targeting.

Annotated protein interaction networks have also been used to associate proteins with known pathways. In some cases these proteins were previously uncharacterized and had no known functional annotations. Since membership in a known pathway implies a function for a protein, protein interactions can be used to provide functional annotations. In network-wide studies it was found that interacting proteins were more likely to share a functional category than those that do not interact (Schwikowski *et al.*, 2000; von Mering *et al.*, 2002). This observation has been expanded into several general methods for functional prediction. The simplest of these methods, the majority-rule method, predicts a function for a protein based on the most commonly occurring functions of its interacting partners. These methods have been applied to the yeast protein interaction network giving accuracies

as high as 90%. To expand the utility of these methods beyond the scope of well-characterized organisms we have extended this approach to computationally predicted protein interaction networks including those for several plants (McDermott *et al.*, 2005), discussed in detail below.

3. COMPARATIVE METHODS OF PREDICTING PLANT PROTEIN INTERACTION NETWORKS

Experimental investigation is the best way to approach elucidation of protein interaction networks. However, the amount of effort required to determine even a small portion of an interaction network for a single organism is considerable. A number of methods have been developed to computationally predict physical interactions or functional relationships between proteins (Valencia and Pazos, 2003; Shi *et al.*, 2005). A functional relationship indicates that the proteins work in concert in some way but does not imply a physical interaction between them. Examples of functional interactions for which there might be no

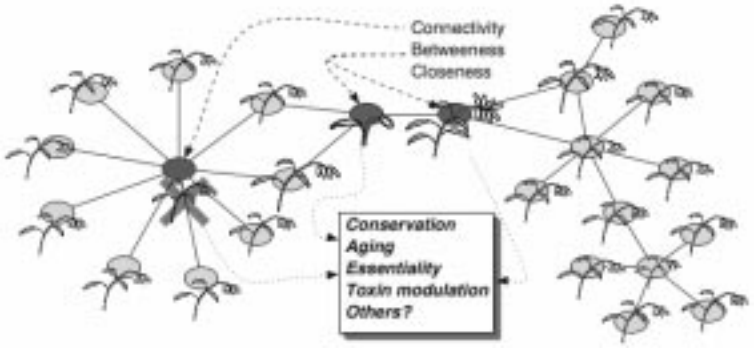

Fig. 1 : Correlation of network properties and protein properties/ phenotype. A number of studies have reported the correlation of various network properties and evolutionary conservation, essentiality/lethality, involvement in aging and senescence and modulation of response to toxins. Connectivity is calculated as the number of interacting neighbors of a protein (k). The red colored protein has the highest connectivity in the network pictured. Betweeness is the percentage of times a protein appears in the shortest paths between all other proteins in the network. Closeness is calculated as the average distance from a protein to all other proteins in the network. Proteins with maximal betweeness and minimal closeness are shown in green in the pictured network. Ovals represent proteins and lines the interactions between them. Possible phenotypic outcomes of mutations in these proteins are depicted.

physical interaction would be proteins involved in successive steps in a metabolic or signal transduction pathway, and two proteins in multiprotein complex that do not interact directly.

One method to predict relationships between proteins is by the physical proximity of the genes that code for those proteins. In prokaryotes a high correlation between membership in an operon and presence of functional or physical interactions has been found and used to predict interaction (Dandekar *et al.*, 1998; Huynen *et al.*, 2000; Snel *et al.*, 2002; Teichmann and Babu, 2002). A relationship between gene proximity and co-expression has been demonstrated in eukaryotes (Cohen *et al.*, 2000; Teichmann and Babu, 2002; Fukuoka *et al.*, 2004). Co-expression has also been shown to correlate with protein interaction in eukaryotes (Jansen *et al.*, 2002). None of these features is predictive of protein interaction by itself so they are usually approaches which integrate many different features to give a prediction (Lu *et al.*, 2005).

Other experimental and genomic feature information can be indicative of a functional relationship or physical interaction but is generally not used to make predictions by itself. The observation that proteins that have similar functions are more likely to interact is used to predict functions from known interactions (see ***Network-based annotation of plant interaction networks***, below). This observation has also been used in integrative approaches for predicting a relationship between the proteins and is commonly used as a way to verify, or at least support, predictions made by other methods (Ge *et al.*, 2001; Lin *et al.*, 2004; Zhang *et al.*, 2004; Lu *et al.*, 2005).

Most methods predicting interactions or functional associations use evolutionary relationships between proteins, also called comparative genomics approaches. It has been reported that proteins that interact are more conserved than those that do not interact (Matthews *et al.*, 2001; Fraser *et al.*, 2002). Phylogenetic profiling examines patterns of orthologs over many different organisms and predicts relationships between proteins if the orthologs appear in a correlated manner (Figure 2)(Pellegrini *et al.*, 1999; Yanai and DeLisi, 2002; Date and Marcotte, 2003). Phylogenetic profiling works well for prokaryotes and has been extended to eukaryotes with some success.

Another method employs protein domain structure across organisms. The domain fusion method (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Yanai *et al.*, 2001) predicts that proteins interact if they represent two domains that appear in a single protein in another organism. The prediction is based on the idea that two domains appearing in one protein must be working together to accomplish a function. If they are

separated by evolutionary processes into two separate proteins, they must interact to accomplish the same function.

The "interolog" method of (Walhout *et al.*, 2000; Matthews *et al.*, 2001; Yu *et al.*, 2004) predicts an interaction between two proteins in a novel proteome by looking for orthologs which are known to interact (see Figure 2). 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 (Altschul *et al.*, 1997). The organism is then examined for all occurrences of two 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; (Bader *et al.*, 2003)) and BIND's dataset of interactions derived from crystallized structures MMDBBind (Salama *et al.*, 2001), the Database of Interacting Proteins (DIP; (Xenarios *et al.*, 2002)), and the Human Proteome Research Database (HPRD; (Peri *et al.*, 2004)). They contain interactions determined from many types of experimental methods, the most prevalent being two-hybrid and TAP, and cover a large number of diverse organisms. Due to the paucity of high-throughput experimental data, plants are underrepresented in experimental interaction databases. BIND, for example, lists about 385 protein

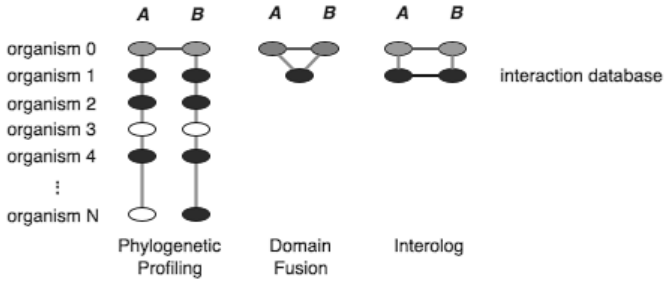

Fig. 2 : A. Comparative methods of interaction prediction. Phylogenetic profiling predicts a functional link between two proteins if their orthologs appear in a correlated manner over a number of organisms. The domain fusion method predicts an interaction between two proteins if they appear as one protein in another organism. The interolog method predicts an interaction between two proteins if their orthologs are known to interact. **B.** Accuracy of the interolog method in *Drosophila.* The accuracy of the interolog method as reported in XXX is shown.

interactions from rice compared to 38,000 for human. The high-throughput interaction determination methods developed in fungi and animals have not been widely applied to plants. Therefore predicted networks must be used to provide preliminary models which can generate hypotheses and lead future computational and experimental investigation.

For plants large-scale computational predictions of protein interactions and relationships have been generated by several groups (Table I). STRING provides computational predictions of protein associations for some plant species. VisANT/Predictome (Mellor *et al.*, 2002) provides computational predictions of protein associations for a number of organisms, but contains only experimentally determined interactions for plants. Finally, AraCyc (Zhang *et al.*, 2005), for *Arabidopsis*, and KEGG (Kanehisa and Goto, 2000) use a combination of orthologous relationships and experimental data, when available, to assign proteins from plant species to metabolic and signaling pathways (see (Lange and Ghassemian, 2005) for a recent review of pathway resources). Additionally, the phenylpropanoid pathway in *Arabidopsis* was analyzed using orthologous mapping (Costa *et al.*, 2003) and gene co-expression was used to predict networks for barley (Faccioli *et al.*, 2005).

The Bioverse has interactions predicted by the interolog method for different organisms, including *Arabidopsis*, *indica* and *japonica* rice, and the plant pathogens *Agrobacterium tumefaciens* and *Magnaporthe grisea* (rice blast)(McDermott and Samudrala, 2003, 2004; McDermott *et al.*, 2005). Table II shows the size of predicted networks for these organisms, as well as two other eukaryotes for comparison, using an IS cutoff of 0.2 and shows the coverage of the predicted networks relative to the total number of proteins in the organism. The coverage for the plants examined is much lower than for *D. melanogaster*, an organism with a large amount of high-throughput experimental data accumulated for it. Coverage is also higher in *C. familiaris*, which does not have an abundance of experimental information but is more closely evolutionarily related to those organisms that do. Nonetheless, predicted plant networks appear to have a very similar structure to experimentally-derived interaction networks. For instance, the scale-free organization of interaction networks is observed in networks predicted using the interolog method (Figure 3), as well as those predicted using other methods (Snel *et al.*, 2002). Importantly, the networks predicted for both *Arabidopsis* and rice encompass many more proteins than do experimentally determined interactions for those organisms.

Table 1 : Resources for protein interactions in plants

Resource	Predicted interactions?	URL	Notes
AraCyc	Yes	http://arabidopsis.org/tools/aracyc/	Orthologous associations for *Arabidopsis* only
BIND	No	http://www.bind.ca/	-
Bioverse	Yes	http://bioverse.compbio.washington.edu	Interolog-based predictions
DIP	No	http://dip.doe-mbi.ucla.edu/	Few plant interactions
KEGG	Yes	http://www.genome.jp/kegg/	Orthologous associations
STRING	Yes	http://string.embl.de	-
VisANT	No	http://visant.bu.edu	Experimental only for plants

Resources in bold type include predicted protein relationships for plants.

Table 2 : Network comparison.

Organism	Total	Network proteins	Predicted interactions	Coverage	Network annotated
A. tumefaciens	5396	569	1357	10.5%	153
A. thaliana	27833	699	2959	2.5%	322
C. familiaris	16817	5785	39302	34.4%	1436
D. melanogaster	16475	13290	405812	80.7%	326
M. grisea	11042	2248	12261	20.4%	1730
O. sativa indica	40925	2756	28003	6.7%	250
O. sativa japonica	36658	2677	31557	7.3%	245

Total, number of proteins in proteome; Network proteins, number of proteins in predicted network considering those interactions with interolog score (IS) greater than 0.15; Predicted interactions, number of interactions with IS greater than 0.15; Coverage, percentage of proteins in organism that have predicted interactions; Network-annotated, number of proteins with no significant functional annotation that could be annotated using the neighborhood-weighting method with a confidence of 30% or better.

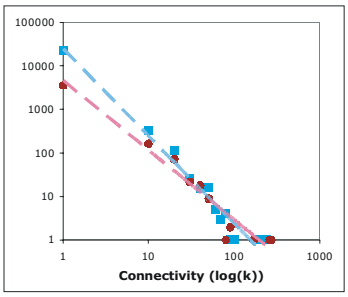

Fig. 3 : Distribution of connectivity in predicted plant networks. The log of the number proteins (N) is plotted against the log of the number of connections per protein (k) with bin size of 10 connections. Dashed lines are fitted to a power-law distribution. The interolog-predicted network from *Arabidopsis* (red circles) is compared with the experimentally-determined yeast network (blue squares) showing that the scale-free nature of interaction networks is preserved in predicted plant networks.

Networks were constructed by considering all proteins as nodes (circles) and all predicted interactions between them as edges (lines) in the network. A portion of the network from *Arabidopsis* is shown in Fig. 4. Proteins are colored by the highest scoring broad GO annotation category (indicated in the legend) and edges between them are the predicted interactions colored according to their IS. It is clear even from this limited subnetwork that proteins with similar functions are likely to have predicted interactions with each other. A subset of more specific GO annotations and PO categories from TAIR appear as square nodes in the network. This allows proteins annotated as, for example, chloroplast proteins, to be associated spatially and shows clearly functions of different regions of the network. Sequence similarity between proteins in the network and *O. sativa japonica* (Syngenta) was calculated and those *Arabidopsis* proteins with PIDs above 80% to rice are shown as a light green circle. This indicates which proteins and regions of the network the most conserved with rice and which are more distinctive to *Arabidopsis*. Red-colored proteins in the network have no strong functional association (computational or manual) and thus represent proteins that would benefit from further investigation (see ***Network-based annotation of plant interaction networks*** section below).

4. ANNOTATION OF PLANT INTERACTION NETWORKS

Providing accurate, high-resolution functional annotations for large numbers of proteins in an organism requires a large amount of

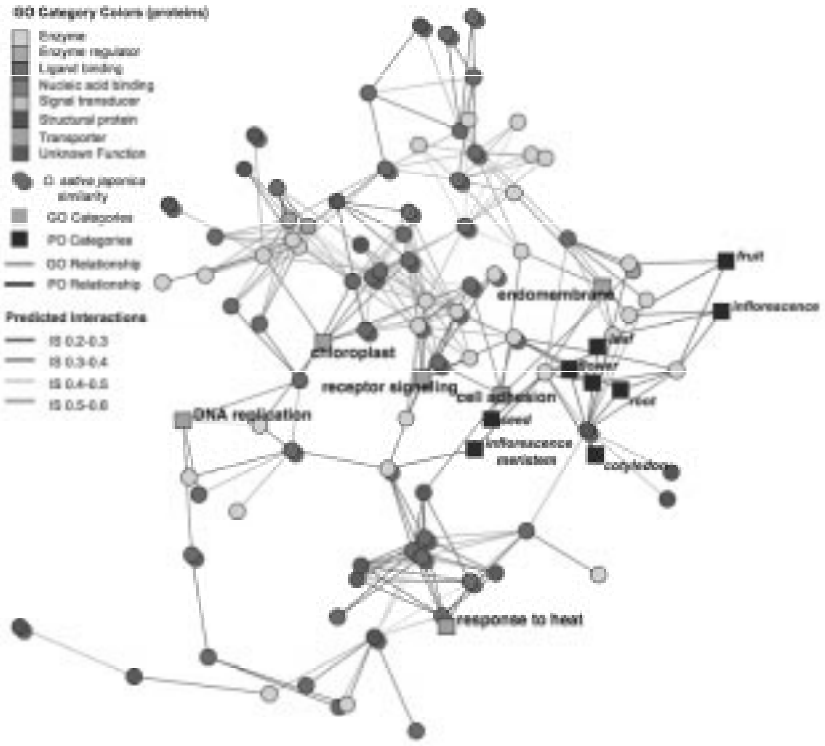

Fig. 4 : Annotated network from *Arabidopsis thaliana*. A network was generated for *Arabidopsis* by predicting interactions (lines) between proteins (circles) as described in the text. Computationally predicted gene ontology (GO) annotations from the Bioverse and manual annotations from TAIR were used to color proteins according to broad categories. Additionally, more specific selected GO categories (light green squares) and plant ontology (PO) categories (blue squares) were associated with proteins and used to enhance the layout of the network. Proteins which are highly conserved (greater than 80% sequence identity) relative to proteins in *O. sativa japonica* (Syngenta) are indicated by a overlapped light green node. Red-colored proteins have no strong functional association and thus represent cases for which network-based annotation will be most useful.

experimental research and manual curation of annotations. Manual curation is normally based on careful reading of literature regarding the protein in question and assignment of functional labels based on the results described therein (Lewis *et al.*, 2000; Iliopoulos *et al.*, 2003; Haas *et al.*, 2005) or by novel experimental investigation targeting numbers of proteins for the purpose of annotation. It can also be a

result of critical evaluation of computational annotations by a biologist familiar with the organism in question. These cases take a significant amount of time and effort and are limited by the availability of experimental information for the protein and/or computational predictions for the protein.

A number of organism-specific projects that perform manual annotation exist. The *Arabidopsis* Information Resource (TAIR (Rhee *et al.*, 2003)), Gramene (http://www.gramene.org; (Ware *et al.*, 2002)) and the Institute for Genome Research (TIGR; http://www.tigr.org) maintain databases for several plants including *Arabidopsis* (Haas *et al.*, 2005), rice (Yuan *et al.*, 2003) and wheat. Due to its suitability for genetic and biochemical analysis *Arabidopsis* has established itself as a model organism and so has the most complete functional annotation of any plant genome. As of 2005, TAIR reports that 75% of the organism's approximately 25,000 proteins could be assigned to at least one functional category using a combination of manual annotation and computational methods (Berardini *et al.*, 2004).

Structured vocabularies such as the gene ontology (GO; (The Gene Ontology Consortium, 2001)) have been developed to provide universal functional descriptions that range in specificity and emphasis. GO was originally designed with Metazoa and Fungi in mind as the first member organisms were fruit fly, mouse and yeast. Recently, GO has been expanded to include more plant-specific functional categories or variations on existing categories (designated by the "*sensu*" label) that reflect differences in plant physiology or biochemistry (Clark *et al.*, 2005). Figure 5 gives some examples of functional categories in the context of the GO structure relevant to plants. A property of this structure is that a protein annotated with a particular category is also associated with all of that category's 'ancestors' in the GO as well.

Additionally, other biological ontologies that provide more specific descriptions of plants and plant physiology have been developed. Gramene, which provides a comparative genomics resource for grasses, has led the development of the Plant Ontology (PO), Growth Stage Ontology (GRO), and Trait Ontology (TO) (Ware *et al.*, 2002; Yamazaki and Jaiswal, 2005). The PO provides descriptions of morphologies and developmental stages of flowering plants. GRO is specific to cereals and provides categories to compare growth stages, presently only for a limited number of plants (Yamazaki and Jaiswal, 2005). Finally, TO provides phenotypic descriptions and information about the methodology used to gather the phenotypes.

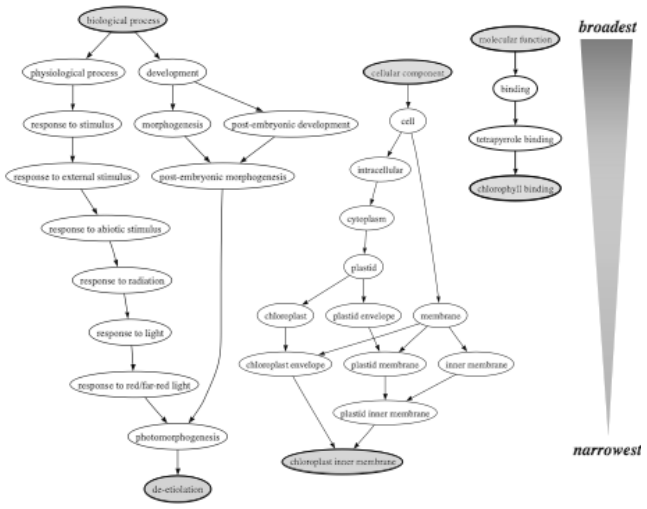

Fig. 5 : Example Structures from the Gene Ontology (GO). Example categories (pink) and their 'path' structures are shown from each of the three main branches of GO, biological process, cellular component and molecular function (green). The structure of GO is a directed acyclic graph (DAG) meaning that categories can have multiple children (as in 'membrane' from cellular component) as well as multiple parents (as in 'photomorphogenesis' from biological process). Categories closer to the root (top) of this structure are broader and those further down the structure are more narrow and convey increasingly specific function.

Since manually-assigned annotations are incomplete, even for the well-covered *Arabidopsis*, other methods can be used to provide functional annotations by extrapolation based on evolutionary comparison, structure-based annotation and novel methods based on the organization of protein interaction networks.

4.1. Sequence-based annotation of plant interaction networks

Functional annotations can be computationally predicted to generate a preliminary model for manual annotation and make predictions to guide direction of experimental investigations to areas of greatest need. Most methods of computational annotation are based on sequence comparison and assignment of function by transferring known functions to the target protein based directly on inferred evolutionary relationships. This type of method is known as transitive functional annotation and can be divided into two major groups: those methods which transfer function based on similarity to other sequences and those which transfer

function based on similarity with conserved sequence elements such as families, domains and motifs. The latter group is more rigorous, since the conserved sequence elements have been well characterized and curated, and can provide better coverage, since many of these elements are modular and can be arranged in different ways for different proteins. Results from a variety of resources for conserved sequence features (e.g. Pfam (Bateman *et al.*, 2000) and Superfamily (Gough and Chothia, 2002)) can be combined into common Interpro categories (Apweiler *et al.*, 2001), which can then be used to derive GO categories for the protein.

Approaches like this have been used by a number of groups including ours to provide functional annotations for plant proteomes (The Arabidopsis Genome Initiative, 2000; Frishman *et al.*, 2001; Goff *et al.*, 2002; Kikuchi *et al.*, 2003; McDermott and Samudrala, 2003; Yuan *et al.*, 2003; Berardini *et al.*, 2004). Large-scale functional annotation has shown that certain functional categories are overrepresented or underrepresented in plants. Proteins involved in RNA processing, protein kinases, the disease resistance-associated proteins LRR (leucine rich repeats) and TIR (Toll/IL-1R), and RING zinc finger and F-box proteins are all overrepresented in *Arabidopsis* and rice (The Arabidopsis Genome Initiative, 2000; Kikuchi *et al.*, 2003; Yuan *et al.*, 2003). In all, about 150 protein families are unique to plants (The Arabidopsis Genome Initiative, 2000).

In *Arabidopsis* 69% of the proteins could be assigned at least one GO annotation using computational means alone (75% including manual annotations) (Berardini *et al.*, 2004). Coverage is lower in rice, between 42-49% (Goff *et al.*, 2002; Yuan *et al.*, 2003). Currently rice, and other plant genomes which are in the process of being sequenced such as potato, soybean and tomato and the cereals sorghum, wheat and maize (Paterson *et al.*, 2005), has far fewer manual annotations than *Arabidopsis*. However, several studies have found that sequence-based computational annotation of whole genomes and/or proteomes can be nearly as accurate as manual annotation efforts (Iliopoulos *et al.*, 2003; Mi *et al.*, 2003), but this has not been demonstrated in plants.

Functional annotation of predicted protein interaction networks provides a context for the functions of individual proteins. Context of a protein in the network can reiterate known mechanisms for protein function, point out straightforward associations that have not been previously reported, and provide novel predictions for functional associations. Shown in Figure 6 is an annotated subnetwork from *japonica* rice (Syngenta) which can be interactively explored using the Bioverse Integrator (http://bioverse.compbio.washington.edu/

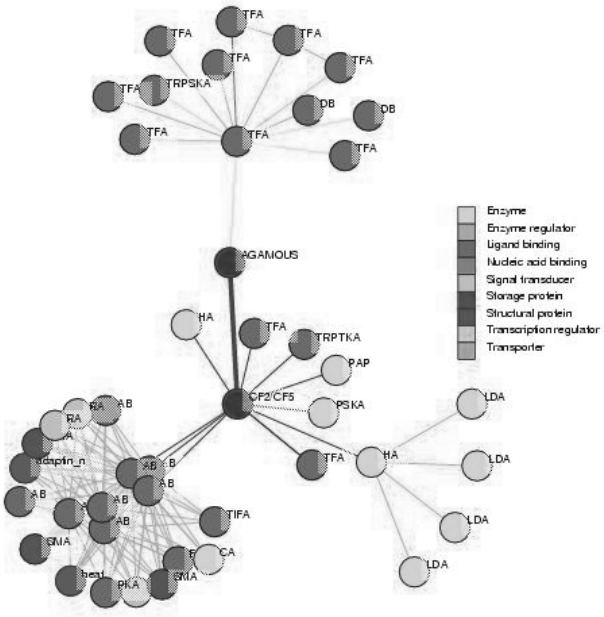

Fig. 6 : Predicted interaction between the transcriptional factor, AGAMOUS, and a Cf2/Cf5 defense protein in rice. AGAMOUS and Cf2/Cf5 are shown in purple, other proteins are colored according to the broad functional categories shown in the legend. Predicted interactions are colored by interolog score (see legend in Figure 4). Although there is evidence to support the existence of this interaction, it has not yet been investigated experimentally.

integrator; (Chang *et al.*, 2005; Chang *et al.*, 2006)). Highlighted in the figure is a novel predicted association between a putative defense response protein, Cf-2/Cf-5, and a transcriptional factor (TF), AGAMOUS. The organization of functional domains in both proteins is evident from their sequence-based annotation. AGAMOUS is a C-class MADS box protein involved in various aspects of floral development (Theissen, 2001; Yamaguchi *et al.*, 2006). Although not directly implicated in disease resistance AGAMOUS-deletion mutants have been shown to be less susceptible to certain pathogens (Urban *et al.*, 2002). Additionally, the Cf-2/Cf-5 homolog is a member of the leucine-rich repeat (LRR) receptor-like protein (RLP) family which has been implicated in both disease resistance and development (Fritz-Laylin *et al.*, 2005). Several interesting directions are suggested by this association: The first is that response to disease-causing agents by Cf-2/Cf-5 makes use of or affects developmental pathways through

AGAMOUS. Alternatively, AGAMOUS may respond to developmental cues from Cf-2/Cf-5. Neither of these possibilities have been investigated experimentally, though in *Arabidopsis* AGAMOUS was identified in a signaling complex with a receptor-like kinase (RLK) that is closely related to the RLP family (Fritz-Laylin *et al.*, 2005; Karlova *et al.*, 2006) and AGAMOUS was also found to be complexed with FLOR1, an LRR protein of unknown function (Gamboa *et al.*, 2001). This example shows how the integration of functional annotation with predicted protein interaction networks can provide novel and useful predictions about the context and mechanism of those functions.

4.2. Structural annotation of plant interaction networks

Determination of function by comparison of structure offers another approach to the annotation problem (Teichmann *et al.*, 2001). Sequence and structure are often conserved in a concerted fashion, that is proteins with similar sequences often have the same general structure, called the fold, and vice-versa. However, structure is more conserved than sequence, leading to proteins with similar structures and functions but unrelated sequences. So, knowing the structure of a protein allows approaches to provide more detailed functional and mechanistic information about proteins and functional determination for proteins for which sequence-based methods don't work well (Samudrala *et al.*, 2000; Pal and Eisenberg, 2005). Structural comparison can also allow better sequence alignments to be constructed from distant orthologs. These alignments can be used to derive functional information by looking for conserved active site residues or other conserved features (Wang and Samudrala, 2005).

Generally, structure-based approaches are limited by the availability and coverage of protein structures for an organism. For plants this is especially relevant since relatively few protein structures have been determined, as mentioned previously. Plant proteins account for only about 1.5% (approximately 600) of proteins with experimentally determined structures listed in the protein data bank (PDB;(Berman *et al.*, 2002)). Comparative modeling offers one way to address this problem: Using closely related sequences with known structures as templates can provide high-resolution structural models (Samudrala and Levitt, 2002). A survey of the plant proteomes in the Bioverse shows that only 15-20% of proteins have sequence similarity to proteins with known structure sufficient to produce moderate quality comparative models (above 20% percentage identity). Proteomics efforts are rapidly expanding the number of proteins and fold families that have known structures and this will continue to improve the utility of structure-based methods.

Protein-protein interactions can be predicted on the basis of structural annotation by identifying protein complexes that have been structurally characterized. In the Bioverse these predictions are implicitly included by using structurally-based protein-protein interactions from MMDBBind for comparison in the interolog method. Structural information and comparisons in the context of metabolic networks have been used to characterize biosynthesis pathways in plants (Noel *et al.*, 2005) and can be used to help engineer pathways for natural product production in plants (Dixon, 2005). More complete structural characterization will greatly enhance the value of protein-protein interaction networks in plants and other organisms.

4.3. Network-based annotation of plant interaction networks

Methods of annotation based on sequence similarity are limited by the existence of characterized sequences with significant sequence similarity and by methods to detect remote similarity. Methods for identification of remote homologs are steadily improving but there are still proteins for which no reliable functional annotations can be assigned. Another way of assigning function to proteins is by integration of existing genomic information of different kinds.

Yeast has the best characterized protein interaction network as well as the most extensively functionally annotated genome of any model organism. These qualities have been exploited to provide annotations for some of the remaining proteins which don't have well characterized functions. Interacting proteins tend to have correlated function as well as expression, evolution and cellular location. Interactions can be used to generate novel functional predictions based on this observation using network-based annotation. The simplest network annotation method is the "majority-rule" method in which the functions from all proteins interacting with a particular protein, its network neighbors, are tabulated and the function or functions with the highest frequency are assigned to the protein in question (Schwikowski *et al.*, 2000). For yeast, the accuracy of the majority-rule method is about 70% for proteins covered, about 30% of the total proteins. Approaches using Markov random field analysis (Letovsky and Kasif, 2003; Deng *et al.*, 2004; Reichmann *et al.*, 2005), global network connectivity (Vazquez *et al.*, 2003) and clustering (Brun *et al.*, 2003; Samanta and Liang, 2003) have all been applied to the yeast protein interaction network with equal or better success. These have improved accuracy, up to the 80-85% range, though the number of functional categories assigned varies.

Until recently, this approach hadn't been applied to plants since experimental protein interaction networks have not been determined.

This fact, as mentioned above, severely limits the utility of the approach, and its applications to less well-characterized organisms. To address this issue we developed a network-based annotation method using predicted interaction networks (McDermott and Samudrala, 2004; McDermott *et al.*, 2005). The method uses interactions predicted by the interolog method and IS scores to generate a list of annotations for each protein in the network and a confidence measure which reflects the probability that the prediction is correct.

Table II shows the number of proteins with no existing automated functional annotation that could be assigned a network-based annotation with a moderate confidence for several organisms. An interesting point is that even though *Arabidopsis* and rice have much lower network coverage than the fly network, the number of useful network annotations produced for these organisms was similar. The method is most useful on the rice blast fungus, *Magnaporthe grisea*, primarily because the proteome is poorly annotated by conventional sequence-based methods. This illustrates the utility of predicted networks for annotation of proteins, even when the predicted networks are quite limited.

5. CORRELATION OF PREDICTED NETWORK CHARACTERISTICS WITH PHENOTYPE

To be useful, predicted interaction networks must be informative about member proteins, or about the organism itself. It has been shown that predicted networks have scale-free and modular structures which are similar to experimental networks (Pellegrini *et al.*, 1999; Snel *et al.*, 2002; Yanai and DeLisi, 2002). Predicted networks can also be used to functionally annotate proteins with no previously known function (McDermott *et al.*, 2005). As discussed above characteristics of experimentally-determined protein interaction networks have been correlated with protein and organismal phenotypes (Figure 1). However, it remains unclear how similar these predicted networks are to experimental networks in terms of this phenotypic correlation.

Our preliminary results indicate that networks predicted by the interolog method can have similar properties to experimentally determined networks in this regard. Analysis of the predicted *Arabidopsis* network from the Bioverse revealed that proteins with more predicted interactions are more evolutionarily conserved than those with fewer predicted interactions, on average (McDermott and Samudrala, unpublished findings). This is not too surprising given that the interolog method will work better on more highly conserved

proteins. Additionally, we have found a similar ratio of connectivity in the predicted network to protein essentiality in the predicted fly network, even when experimental fly data was eliminated, as that found in the corresponding experimental network (McDermott and Samudrala, unpublished findings). Studies are underway to further characterize this relationship in the predicted plant protein interaction networks.

For plants the potential applications of this approach are obvious. Predicted protein interaction networks could be used to identify proteins involved in senescence, toxin-response or other pertinent phenotypes. Due to the limited number of manual annotations available and the limited coverage of the predicted networks for plants, thorough evaluation of these networks is difficult at this time. Increased availability of experimental interactions and more extensive annotation of plant proteomes will allow us to better address this issue.

6. CONCLUSION

Computational methods based on comparative genomics can be used to predict protein interaction networks for previously uncharacterized organisms. Building on evolutionary relationships and the burgeoning amount of experimental data covering protein-protein interactions for a small number of organisms, methods extrapolating interactions to largely uncharacterized organisms have been developed. A number of groups in this field have generated predicted interaction networks for many organisms, but very few have focused on prediction of plant networks (Table I). Using our computational biology framework, the Bioverse, we have applied the interolog method to generate predicted interaction networks in *Arabidopsis* and the *japonica* and *indica* cultivars of rice. Although the networks generated by this approach cover only a small percentage of proteins in their respective organisms, they represent a valuable starting point for exploration of plant protein interaction networks. Evolutionary relationships between plants and other organisms preserve a core interaction network for the plants examined.

Well-characterized protein interaction datasets can be used to verify and calibrate predictive methods described. Such calibration is essential to provide robust estimates of accuracy for predictions made. An assumption of the work described here is that accuracy estimates for predictive methods made using gold standard sets from animals, fungi and bacteria, will apply to predictions made for plant proteins. Based on results from other studies (e.g. extrapolating from fungi to *C. elegans*) it seems that protein interactions can be transferred across appreciable evolutionary distances but only increases in the number of

experimentally determined interactions for plants will allow rigorous examination of the accuracy of these predictions.

Clearly, predicted networks such as those described in this chapter can not substitute for experimental determination of protein interactions in the organism of interest. Rather, these networks represent prototype models for biological systems. The value of such models is still being assessed but these types of networks have been proven to provide valuable information in terms of novel functional assignments (Date and Marcotte, 2003; McDermott *et al.*, 2005). Based on our preliminary investigations they also seem to provide useful information in terms of correlations of network structure and organismal phenotypes. These findings indicate that the predicted networks reflect the underlying real protein interaction networks but it is possible that the predictive methods generate interactions and networks that merely have some similar properties of real ones. For instance, the observation that proteins in the predicted *Arabidopsis* network with greater connectivity also are more evolutionarily conserved is inextricable from the method used to make the predictions, which is based on evolutionary conservation. However, this issue doesn't render the models useless, merely points out the importance of caution in interpretation of results. This is the same caution which should exist when considering networks determined by high-throughput methods since they may well be *less* accurate, considered independently (von Mering *et al.*, 2002), than predicted models such as interolog networks with appropriate calibration (McDermott and Samudrala, manuscript in preparation; (Yu *et al.*, 2004)).

Proteins must be subjected to extensive experimental investigation to achieve confident, high-resolution functional characterization, but computational methods can provide a good starting point for such investigation. Rigorous predictive methods, especially those that provide coverage for the greatest number of proteins with unknown function, can identify proteins that most need experimental characterization, and can place proteins in a functional context that can provide valuable information for experimentalists.

Computational predictions can be used to further plant biology. The predictive methods described here, as well as prediction of protein structure, provide useful hypotheses for experimental investigation. It is clear that the prediction of interaction networks for plants is still in its infancy. Further experimental characterization of protein interaction networks from plants and other organisms will allow more accurate and complete prediction of networks.

7. REFERENCES

Albert, R., Jeong, H. and Barabasi, A. L. 2000. Error and attack tolerance of complex networks. Nature, 406(6794): 378-382.

Alberts, B. 1998. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell, 92(3): 291-294.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25(17): 3389-3402.

Andon, N. L., Hollingworth, S., Koller, A., Greenland, A. J., Yates, J. R., 3rd and Haynes, P. A. 2002. Proteomic characterization of wheat amyloplasts using identification of proteins by tandem mass spectrometry. Proteomics, 2(9): 1156-1168.

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. and Zdobnov, E. M. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res, 29(1): 37-40.

Bader, G. D., Betel, D. and Hogue, C. W. 2003. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res, 31(1): 248-250.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. 2000. The Pfam protein families database. Nucleic Acids Res, 28(1): 263-266.

Berardini, T. Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L. A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. and Rhee, S. Y. 2004. Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol, 135(2): 745-755.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. and Zardecki, C. 2002. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr, 58(Pt 6 No 1): 899-907.

Bhalla, R., Narasimhan, K. and Swarup, S. 2005. Metabolomics and its role in understanding cellular responses in plants. Plant Cell Rep, 24(10): 562-571.

Birkland, A. and Yona, G. 2006. BIOZON: a hub of heterogeneous biological data. Nucleic Acids Res, 34(Database issue): D235-242.

Bork, P. 2002. Comparative analysis of protein interaction networks. Bioinformatics, 18 Suppl 2: S64.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A. and Jacq, B. 2003. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol, 5(1): R6.

Chang, A. N., McDermott, J., Frazier, Z., Guerquin, M. and Samudrala, R. 2006. INTEGRATOR: interactive graphical search of large protein interactomes over the Web. BMC Bioinformatics, 7: 146.

Chang, A. N., McDermott, J. and Samudrala, R. 2005. An enhanced Java graph applet interface for visualizing interactomes. Bioinformatics, 21(8): 1741-1742.

Chien, C. T., Bartel, P. L., Sternglanz, R. and Fields, S. 1991. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc Natl Acad Sci U S A, 88(21): 9578-9582.

Clark, J. I., Brooksbank, C. and Lomax, J. 2005. It's all GO for plant scientists. Plant Physiol, 138(3): 1268-1279.

Cohen, B. A., Mitra, R. D., Hughes, J. D. and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet, 26(2): 183-186.

Cooper, B., Clarke, J. D., Budworth, P., Kreps, J., Hutchison, D., Park, S., Guimil, S., Dunn, M., Luginbuhl, P., Ellero, C., Goff, S. A. and Glazebrook, J. 2003. A network of rice genes associated with stress response and seed development. Proc Natl Acad Sci U S A, 100(8): 4945-4950.

Cooper, B., Hutchison, D., Park, S., Guimil, S., Luginbuhl, P., Ellero, C., Goff, S. A. and Glazebrook, J. 2003. Identification of rice (*Oryza sativa*) proteins linked to the cyclin-mediated regulation of the cell cycle. Plant Mol Biol, 53(3): 273-279.

Costa, M. A., Collins, R. E., Anterola, A. M., Cochrane, F. C., Davin, L. B. and Lewis, N. G. 2003. An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in *Arabidopsis thaliana* and limitations thereof. Phytochemistry, 64(6): 1097-1112.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci, 23(9): 324-328.

Date, S. V. and Marcotte, E. M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol, 21(9): 1055-1062.

Deng, M., Tu, Z., Sun, F. and Chen, T. 2004. Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics, 20(6): 895-902.

Dixon, R. A. 2005. Engineering of plant natural product pathways. Curr Opin Plant Biol, 8(3): 329-336.

Dong, Q., Kroiss, L., Oakley, F. D., Wang, B. B. and Brendel, V. 2005. Comparative EST analyses in plant systems. Methods Enzymol, 395: 400-418.

Drees, B. L., Sundin, B., Brazeau, E., Caviston, J. P., Chen, G. C., Guo, W., Kozminski, K. G., Lau, M. W., Moskow, J. J., Tong, A., Schenkman, L. R., McKenzie, A., 3rd, Brennwald, P., Longtine, M., Bi, E., Chan, C., Novick, P., Boone, C., Pringle, J. R., Davis, T. N., Fields, S. and Drubin, D. G. 2001. A protein interaction map for cell polarity development. J Cell Biol, 154(3): 549-571.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C. and Ouzounis, C. A. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature, 402(6757): 86-90.

Faccioli, P., Provero, P., Herrmann, C., Stanca, A. M., Morcia, C. and Terzi, V. 2005. From single genes to co-expression networks: extracting knowledge from barley functional genomics. Plant Mol Biol, 58(5): 739-750.

Ferrarini, L., Bertelli, L., Feala, J., McCulloch, A. D. and Paternostro, G. 2005. A more efficient search strategy for aging genes based on connectivity. Bioinformatics, 21(3): 338-348.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. and Feldman, M. W. 2002. Evolutionary rate in the protein interaction network. Science, 296(5568): 750-752.

Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A. and Mewes, H. W. 2001. Functional and structural genomics using PEDANT. Bioinformatics, 17(1): 44-57.

Fritz-Laylin, L. K., Krishnamurthy, N., Tor, M., Sjolander, K. V. and Jones, J. D. 2005. Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis. Plant Physiol, 138(2): 611-623.

Fukuoka, Y., Inaoka, H. and Kohane, I.S. 2004. Inter-species differences of co-expression of neighboring genes in eukaryotic genomes. BMC Genomics, 5(1): 4.

Gamboa, A., Paez-Valencia, J., Acevedo, G. F., Vazquez-Moreno, L. and Alvarez-Buylla, R. E. 2001. Floral transcription factor AGAMOUS interacts in vitro with a leucine-rich repeat and an acid phosphatase protein complex. Biochem Biophys Res Commun, 288(4): 1018-1026.

Ge, H., Liu, Z., Church, G. M. and Vidal, M. 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat Genet, 29(4): 482-486.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., Jr., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J. and Rothberg, J. M. 2003. A protein interaction map of *Drosophila melanogaster*. Science, 302(5651): 1727-1736.

Girvan, M. and Newman, M. E. 2002. Community structure in social and biological networks. Proc Natl Acad Sci U S A, 99(12): 7821-7826.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W. L., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science, 296(5565): 92-100.

Gough, J. and Chothia, C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res, 30(1): 268-272.

Guerquin, M., McDermott, J., Frazier, Z. and Samudrala, R. (2007). The Bioverse API and Web Application. Computational Systems Biology. J. Mcdermott, R. Ireton, K. Montgomery, R. Bumgarner and R. Samudrala, Humana Press: [in press].

Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith, R. K., Jr., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., White, O. and Town, C. D. 2005. Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release. BMC Biol, 3(1): 7.

Hahn, M. W. and Kern, A. D. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol Biol Evol, 22(4): 803-806.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D. and Tyers, M. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature, 415(6868): 180-183.

Huynen, M., Snel, B., Lathe, W. and Bork, P. 2000. Exploitation of gene context. Curr Opin Struct Biol, 10(3): 366-370.

Iliopoulos, I., Tsoka, S., Andrade, M. A., Enright, A. J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C., Hamodrakas, S., Tamames, J., Yagnik, A. T., Tramontano, A., Devos, D., Blaschke, C., Valencia, A., Brett, D., Martin, D., Leroy, C., Rigoutsos, I., Sander, C. and Ouzounis, C. A. 2003. Evaluation of annotation strategies using an entire genome sequence. Bioinformatics, 19(6): 717-726.

Islam, N., Tsujimoto, H. and Hirano, H. 2003. Proteome analysis of diploid, tetraploid and hexaploid wheat: towards understanding genome interaction in protein expression. Proteomics, 3(4): 549-557.

Jansen, R., Greenbaum, D. and Gerstein, M. 2002. Relating whole-genome expression data with protein-protein interactions. Genome Res., 12(1): 37-46.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. 2001. Lethality and centrality in protein networks. Nature, 411(6833): 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. 2000. The large-scale organization of metabolic networks. Nature, 407(6804): 651-654.

Kanehisa, M. and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 28(1): 27-30.

Karlova, R., Boeren, S., Russinova, E., Aker, J., Vervoort, J. and de Vries, S. 2006. The Arabidopsis SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1 Protein Complex Includes BRASSINOSTEROID-INSENSITIVE1. Plant Cell, 18: 626-638.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, C. J., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuki, Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizuno, K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashidume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konno, H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M. and Hayashizaki, Y. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science, 301(5631): 376-379.

Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., Haynes, P. A., Hays, L., Schieltz, D., Ulaszek, R., Wei, J., Wolters, D. and Yates, J. R., 3rd 2002. Proteomic survey of metabolic pathways in rice. Proc Natl Acad Sci U S A, 99(18): 11969-11974.

Lange, B. M. and Ghassemian, M. 2005. Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry, 66(4): 413-451.

Letovsky, S. and Kasif, S. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics, 19 Suppl 1: I197-I204.

Lewis, S., Ashburner, M. and Reese, M. G. 2000. Annotating eukaryote genomes. Curr Opin Struct Biol, 10(3): 349-354.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez,

M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. and Vidal, M. 2004. A map of the interactome network of the metazoan *C. elegans*. Science, 303(5657): 540-543.

Lin, N., Wu, B., Jansen, R., Gerstein, M. and Zhao, H. 2004. Information assessment on predicting protein-protein interactions. BMC Bioinformatics, 5: 154.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M. 2005. Assessing the limits of genomic data integration for predicting protein networks. Genome Res, 15(7): 945-953.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. and Eisenberg, D. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science, 285(5428): 751-753.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. and Vidal, M. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. Genome Res, 11(12): 2120-2126.

McDermott, J., Bumgarner, R. and Samudrala, R. 2005. Functional annotation from predicted protein interaction networks. Bioinformatics, 21(15): 3217-3226.

McDermott, J. and Samudrala, R. 2003. Bioverse: functional, structural and contextual annotation of proteins and proteomes. Nucleic Acids Res, 31(13): 3736-3737.

McDermott, J. and Samudrala, R. 2004. Enhanced functional information from predicted protein networks. Trends Biotechnol, 22(2): 60-62.

Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J. and DeLisi, C. 2002. Predictome: a database of putative functional links between proteins. Nucleic Acids Res, 30(1): 306-309.

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. and Ruepp, A. 2004. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res, 32(Database issue): D41-44.

Mi, H., Vandergriff, J., Campbell, M., Narechania, A., Majoros, W., Lewis, S., Thomas, P. D. and Ashburner, M. 2003. Assessment of genome-wide protein function classification for Drosophila melanogaster. Genome Res, 13(9): 2118-2128.

Morett, E., Korbel, J. O., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B. and Bork, P. 2003. Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat Biotechnol, 21(7): 790-795.

Noel, J. P., Austin, M. B. and Bomati, E. K. 2005. Structure-function relationships in plant phenylpropanoid biosynthesis. Curr Opin Plant Biol, 8(3): 249-253.

Pal, D. and Eisenberg, D. 2005. Inference of protein function from protein structure. Structure, 13(1): 121-130.

Paterson, A. H., Bowers, J. E., Peterson, D. G., Estill, J. C. and Chapman, B. A. 2003. Structure and evolution of cereal genomes. Curr Opin Genet Dev, 13(6): 644-650.

Paterson, A. H., Freeling, M. and Sasaki, T. 2005. Grains of knowledge: genomics of model cereals. Genome Res, 15(12): 1643-1650.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96(8): 4285-4288.

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T. K., Chandrika, K. N., Deshpande, N., Suresh, S., Rashmi, B. P., Shanker, K., Padma, N., Niranjan, V., Harsha, H. C., Talreja, N., Vrushabendra, B. M., Ramya, M. A., Yatish, A. J., Joy, M., Shivashankar, H. N., Kavitha, M. P., Menezes, M., Choudhury, D. R., Ghosh, N., Saravana, R., Chandran, S., Mohan, S., Jonnalagadda, C. K., Prasad, C. K., Kumar-Sinha, C., Deshpande, K. S. and Pandey, A. 2004. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res, 32(Database issue): D497-501.

Pletcher, S. D. 2004. Vital connections. Sci Aging Knowledge Environ, 2004(19): pe19.

Promislow, D. E. 2004. Protein networks, pleiotropy and the evolution of senescence. Proc Biol Sci, 271(1545): 1225-1234.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. 2002. Hierarchical organization of modularity in metabolic networks. Science, 297(5586): 1551-1555.

Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O. and Schreiber, G. 2005. The modular architecture of protein-protein binding interfaces. Proc Natl Acad Sci U S A, 102(1): 57-62.

Rensink, W. A. and Buell, C. R. 2005. Microarray expression profiling resources for plant genomics. Trends Plant Sci, 10(12): 603-609.

Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J. and Zhang, P. 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res, 31(1): 224-228.

Rivas, S., Mucyn, T., van den Burg, H. A., Vervoort, J. and Jones, J. D. 2002. An approximately 400 kDa membrane-associated complex that contains one molecule of the resistance protein Cf-4. Plant J, 29(6): 783-796.

Rives, A. W. and Galitski, T. 2003. Modular organization of cellular networks. Proc Natl Acad Sci U S A, 100(3): 1128-1133.

Said, M. R., Begley, T. J., Oppenheim, A. V., Lauffenburger, D. A. and Samson, L. D. 2004. Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. Proc Natl Acad Sci U S A, 101(52): 18006-18011.

Salama, J. J., Donaldson, I. and Hogue, C. W. 2001. Automatic annotation of BIND molecular interactions from three-dimensional structures. Biopolymers, 61(2): 111-120.

Samanta, M. P. and Liang, S. 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci U S A, 100(22): 12579-12583.

Samudrala, R. and Levitt, M. 2002. A comprehensive analysis of 40 blind protein structure predictions. BMC Struct Biol, 2(1): 3.

Samudrala, R., Xia, Y., Levitt, M., Cotton, N. J., Huang, E. S. and Davis, R. 2000. Probing structure-function relationships of the DNA polymerase alpha-associated zinc-finger protein using computational approaches. Pac Symp Biocomput: 179-190.

Schwikowski, B., Uetz, P. and Fields, S. 2000. A network of protein-protein interactions in yeast. Nat Biotechnol, 18(12): 1257-1261.

Shi, T. L., Li, Y. X., Cai, Y. D. and Chou, K. C. 2005. Computational methods for protein-protein interaction and their application. Curr Protein Pept Sci, 6(5): 443-449.

Snel, B., Bork, P. and Huynen, M. A. 2002. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A, 99(9): 5890-5895.

Spirin, V. and Mirny, L. A. 2003. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A, 100: 12123-12128.

Strong, M., Mallick, P., Pellegrini, M., Thompson, M. J. and Eisenberg, D. 2003. Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach. Genome Biol, 4(9): R59.

Teichmann, S. A. and Babu, M. M. 2002. Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol, 20(10): 407-410; discussion 410.

Teichmann, S. A., Murzin, A. G. and Chothia, C. 2001. Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol, 11(3): 354-363.

The Arabidopsis Genome Initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408(6814): 796-815.

The Gene Ontology Consortium 2001. Creating the gene ontology resource: design and implementation. Genome Res, 11(8): 1425-1433.

Theissen, G. 2001. Development of floral organ identity: stories from the MADS house. Curr Opin Plant Biol, 4(1): 75-85.

Tzfira, T. and White, C. 2005. Towards targeted mutagenesis and gene replacement in plants. Trends Biotechnol, 23(12): 567-569.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature, 403(6770): 623-627.

Urban, M., Daniels, S., Mott, E. and Hammond-Kosack, K. 2002. Arabidopsis is susceptible to the cereal ear blight fungal pathogens *Fusarium graminearum* and *Fusarium culmorum*. Plant J, 32(6): 961-973.

Valencia, A. and Pazos, F. 2003. Prediction of protein-protein interactions from evolutionary information. Methods Biochem Anal, 44: 411-426.

Vazquez, A., Flammini, A., Maritan, A. and Vespignani, A. 2003. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 21(6): 697-700.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature, 417(6887): 399-403.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N. and Vidal, M. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science, 287(5450): 116-122.

Wang, K. and Samudrala, R. 2005. FSSA: a novel method for identifying functional signatures from structural alignments. Bioinformatics, 21(13): 2969-2977.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., McCouch, S. and Stein, L. 2002. Gramene: a resource for comparative grass genomics. Nucleic Acids Res, 30(1): 103-105.

Wichadakul, D., McDermott, J. and Samudrala, R. (2007). Prediction and integration of regulatory and protein-protein interactions. Computational Systems Biology. J. Mcdermott, R. Ireton, K. Montgomery, R. Bumgarner and R. Samudrala, Humana Press: [in press].

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. and Eisenberg, D. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res, 30(1): 303-305.

Yamaguchi, T., Lee, D. Y., Miyao, A., Hirochika, H., An, G. and Hirano, H. Y. 2006. Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in *Oryza sativa*. Plant Cell, 18(1): 15-28.

Yamazaki, Y. and Jaiswal, P. 2005. Biological ontologies in rice databases. An introduction to the activities in Gramene and Oryzabase. Plant Cell Physiol, 46(1): 63-68.

Yanai, I. and DeLisi, C. 2002. The society of genes: networks of functional links between genes from comparative genomics. Genome Biol, 3(11): research0064.

Yanai, I., Derti, A. and DeLisi, C. 2001. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc Natl Acad Sci U S A, 98(14): 7940-7945.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. 2004. Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs. Genome Res, 14(6): 1107-1118.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Huang, X., Su, Z., Tong, W., Tong, Z., Ye, J., Wang, L., Lei, T., Chen, C., Chen, H., Huang, H., Zhang, F., Li, N., Zhao, C., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Hu, W., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wong, G. K. and Yang, H. 2005. The Genomes of Oryza sativa: A History of Duplications. PLoS Biol, 3(2): e38.

Yuan, Q., Ouyang, S., Liu, J., Suh, B., Cheung, F., Sultana, R., Lee, D., Quackenbush, J. and Buell, C. R. 2003. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res, 31(1): 229-233.

Zhang, L. V., Wong, S. L., King, O. D. and Roth, F. P. 2004. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics, 5: 38.

Zhang, P., Foerster, H., Tissier, C. P., Mueller, L., Paley, S., Karp, P. D. and Rhee, S. Y. 2005. MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol, 138(1): 27-37.