Review

Special Issue: Systems Biology

Towards revealing the functions of all genes in plants

Seung Yon Rhee[1] and Marek Mutwil[2]

[1] Carnegie Institution for Science, Department of Plant Biology, 260 Panama St, Stanford, CA 94305, USA
[2] Max Planck Institute for Molecular Plant Physiology, 14476 Potsdam, Germany

The great recent progress made in identifying the molecular parts lists of organisms revealed the paucity of our understanding of what most of the parts do. In this review, we introduce computational and statistical approaches and omics data used for inferring gene function in plants, with an emphasis on network-based inference. We also discuss caveats associated with network-based function predictions such as performance assessment, annotation propagation, the guilt-by-association concept, and the meaning of hubs. Finally, we note the current limitations and possible future directions such as the need for gold standard data from several species, unified access to data and tools, quantitative comparison of data and tool quality, and high-throughput experimental validation platforms for systematic gene function elucidation in plants.

How little we know

The elucidation of the genome sequence of many organisms, one of the outstanding achievements of our generation, confirmed what most biologists already suspected – that we know little about what most genes do. For example, approximately 40% of *Arabidopsis* (*Arabidopsis thaliana*, thale cress) and 1% of rice (*Oryza sativa*) protein-coding genes have had some aspect of their functions annotated based on experimental evidence (Figure 1) [1,2]. Moreover, we know about the biochemical activity, subcellular location, and biological role of only ~5% of *Arabidopsis* genes based on experimental evidence. It is difficult to determine the number of experimentally characterized genes in public databases for any plant species other than for *Arabidopsis* and rice. This paucity and disparity in the level of functional annotation in different plant species is a bottleneck for understanding how biological processes are organized, how they function, and how they evolved in plants.

Because empirical elucidation of gene function and extraction of such information from the literature are time-consuming processes, researchers have been turning to *in silico* methods for assistance in elucidating and annotating gene function. Fortunately, the past decade has seen a revolution in omics technologies (see Glossary) that have generated copious amounts of data useful for *in silico* function prediction. In this review, we examine the different types of omics data that are being generated and

Glossary

Bayesian network: a graphical representation of the conditional dependencies of nodes.
Cluster compactness: a measure for determining the degree of similarity of nodes in a cluster.
Cluster completeness: a measure of how many nodes with the same property are assigned to the same cluster.
Cluster connectedness: a measure of the density of the links between nodes in a cluster.
Cluster purity or homogeneity: a measure of the homogeneity of the characteristics of the nodes in a cluster.
Cluster stability: a measure of the degree of conservation of a cluster with respect to the composition of the nodes when different parameters or datasets are used to generate it.
Decision tree: a model that uses a tree-like graph of decisions and their possible consequences.
Evidence code: a type of evidence supporting the assertion of the annotation. Experimental evidence codes include: inferred from direct assays (IDA), inferred from expression patterns (IEP), inferred from genetic interactions (IGI), inferred from mutant phenotypes (IMP), and inferred from physical interactions (IPI). More information about GO evidence codes can be found online (http://www.geneontology.org/GO.evidence.shtml).
Evolutionary context: the co-gain or loss of genes through evolution. Also called phylogenomic or phylogenetic profiling.
Gene fusion: an evolutionary event where two proteins in a species have been fused into one protein in another species.
Genomic context: physical proximity of genes belonging to the same pathway or process on the chromosome.
Gold standard data: data that have been experimentally validated and published in primary research articles.
Granularity: specificity of a term in an ontology, often represented as the distance from the root term.
Guilt-by-association: in function prediction, this is a conjecture that genes of related functions share similar characteristics.
Machine learning: a branch of artificial intelligence dealing with learning from data, often used for classification.
Network neighbors: nodes that are connected by a link in a network.
Neural network: a model based on the human neuron perception system.
Omics technologies: high-throughput experimental techniques that are applied genome-wide.
Ontologies: controlled vocabulary systems with an explicit definition of meaning and relationship with other terms in the system.
Predictive power: a measure of the accuracy of a prediction method.
Support vector machine: a computational method used for optimally separating data into categories by drawing a hyperplane in a multidimensional data space.
Weighted co-function network: a network where nodes represent genes and links represent functional associations between those genes. The links are assigned weights to represent the probability of two genes being functionally associated.

Corresponding authors: Rhee, S.Y. (srhee@carnegiescience.edu);
Mutwil, M. (mutwil@mpimp-golm.mpg.de).

Keywords: function prediction; omics; big data; networks; co-expression; co-function.

1360-1385/$ – see front matter
http://dx.doi.org/10.1016/j.tplants.2013.10.006

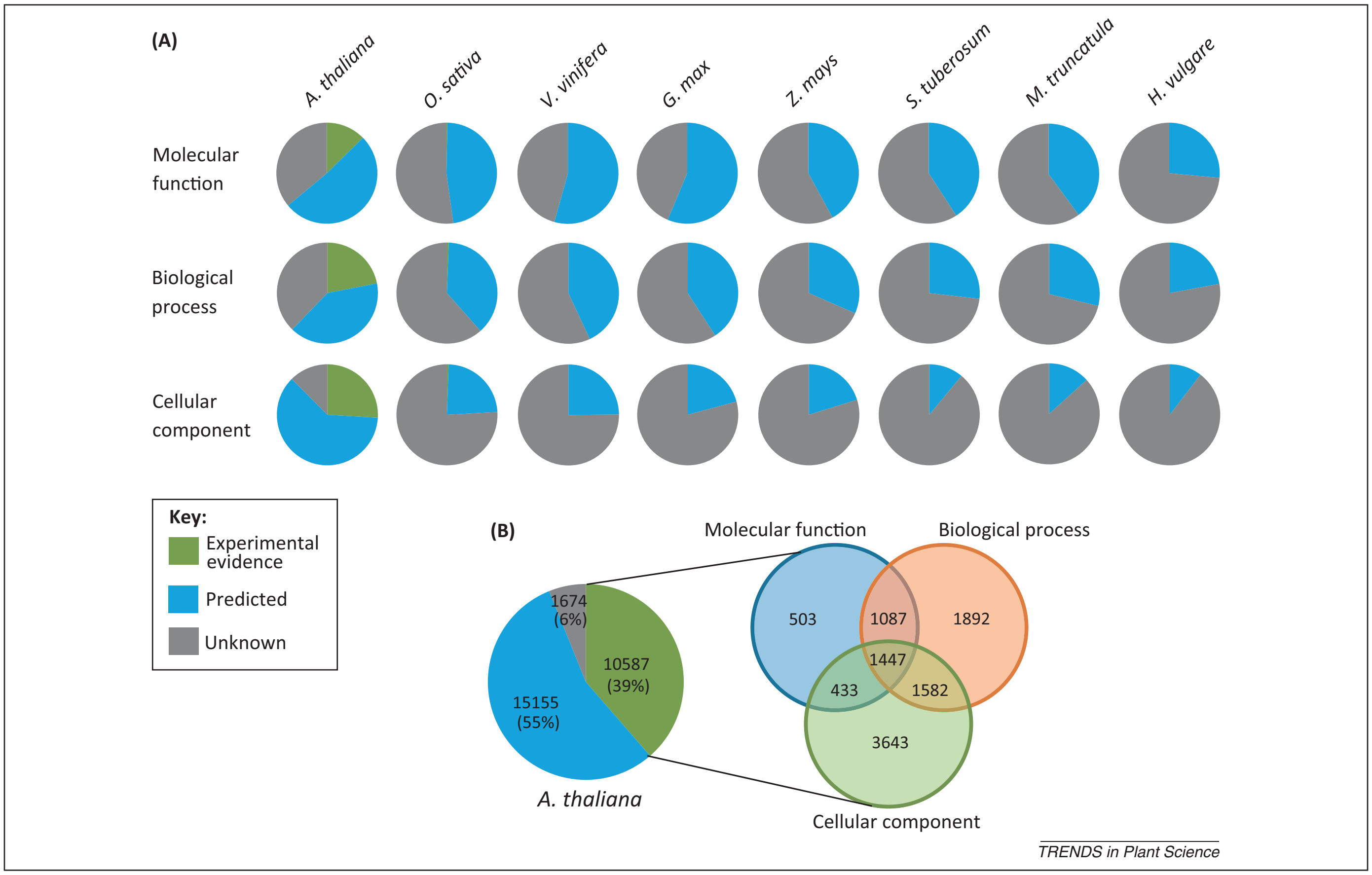

Figure 1. Status of gene function elucidation and annotation in plants: *Arabidopsis thaliana*, rice (*Oryza sativa*), grapevine (*Vitis vinifera*), soybean (*Glycine max*), maize (*Zea mays*), potato (*Solanum tuberosum*), *Medicago truncatula*, and barley (*Hordeum vulgare*). **(A)** Each pie chart shows the proportion of genes that are annotated to a domain of Gene Ontology (GO), molecular function, biological process, or cellular component, based on experimental evidence (green), computational predictions (blue), or uncharacterized or unannotated (gray). GO annotations were downloaded from GRAMENE (http://www.gramene.org) on June 17, 2013 using BioMart. **(B)** Completeness of gene annotation for *A. thaliana*. The pie chart shows the number and proportion of genes annotated to at least one GO domain. The Venn diagram shows the number of genes annotated to each domain of GO based on experimental data. GO evidence codes [11] were used to distinguish experimentally derived annotations from computationally predicted ones.

the methods that can be used to infer the molecular function, biological process, or cellular component of a gene product.

What is in a function?

Gene function can mean different things to different people. Therefore, it is important to use controlled vocabularies for defining the function explicitly [3]. It is also helpful to use the same vocabularies for describing functions to maximize comparability across species. The Open Biological Ontologies consortium provides a set of guidelines for creating and improving ontologies and a forum for sharing them [4]. The Gene Ontology (GO) vocabulary system exemplifies the minimal information necessary to define gene function by using three domains: cellular component (subcellular components where the gene product acts), molecular function (biochemical activities of the gene product), and biological process (goals of the activities of the gene product) [5]. For example, using GO, we can state that the large subunit of the ribulose-1,5-bisphosphate carboxylase oxygenase complex (RBCL) is involved in ‘carbon fixation’ (GO:0015977, biological process) and works in the ‘chloroplast thylakoid membrane’ (GO:0009535, cellular component) where it has ‘ribulose-bisphosphate carboxylase activity’ (GO:0016984, molecular function). Other commonly used ontologies in plant research include the Enzyme Commission nomenclature for describing catalytic reactions [6], Transporter Classification for transporters [7], Plant Ontology for plant growth stages and anatomical structures [8,9], and Mapman ontologies for biological processes [10]. An important characteristic of these vocabulary systems is that they are organized into hierarchical structures that enable groupings, comparisons, and inferences to be made at different granularities of function [11]. A disadvantage of ontologies is that the multiple parent–child relationships make visualization and maintenance of the ontologies non-trivial. An annotation of gene function using these ontologies should be accompanied by explicit evidence types and confidence measures and linked to primary sources supporting the evidence [11].

What is in a network?

Just as a function can have different meanings, a network can also have different meanings and purposes in biology. Molecular networks that have been generated can be grouped into three categories: associational, informational, and mechanistic. Associational networks are akin to social networks such as Facebook or LinkedIn. We can guess things about a gene (or person) based on other genes (or people) it is connected to. For example, properties of genes can be identified from omics data and used to link the genes that share the properties, resulting in a co-function network [12]. How co-function networks are used to infer gene function is described in more detail in this review. Informational networks process information (e.g., the Ethernet or Internet). Signal transduction networks and genome-wide metabolic networks where the nodes represent individual functions are so far represented as informational networks in plant biology [13]. Although information about the nature of how the functions are linked in these networks can be used to predict genes that may perform the functions, we will not discuss these approaches further here. Mechanistic networks (e.g., the electronic circuit of a computer or the lac operon) describe a system quantitatively and mechanistically [14]. A grand challenge in biology is to uncover these mechanistic networks at scales of increasing depth and breadth. The elucidation of all functions encoded by a genome will be an important step for tackling this challenge.

Omics data used in inferring gene function

Omics data can help elucidate functions of gene products either by direct measurement or usage in inference programs. Typically, a particular type of omics data is useful for elucidating functions in a particular GO domain. For example, sequencing peptides from isolated subcellular compartments is useful for assigning gene products to cellular components [15], but is less valuable for assigning to biological processes. In addition, similarity between protein sequences enables molecular functions to be inferred [16]. Finally, analyses of high-throughput interactomes can help infer biological processes because physically interacting proteins tend to be involved in the same biological process [17].

The different types of omics data and methods that can be used to predict the molecular function, biological process, or cellular component of a gene product are described below. Public repositories and the amount of available plant omics data are listed in Table 1.

Genomic data

Advances in genome sequencing and assembly have made it possible to sequence the genomes of more than 40 plant species and nearly 7000 species from the other kingdoms [18,19]. Genomic data can be used to infer molecular function and biological process. For example, protein sequence similarity is commonly used to transfer molecular function annotation from one protein to another [20]. Molecular function annotation by sequence comparison is commonly performed using programs such as BLAST [20] and InterProScan [21]. Typically, sequence identity of more than 60% can predict enzyme function with at least 90% accuracy [16]. A major pitfall of sequence-based inference is that high sequence similarity does not always guarantee the same function [22] and, conversely, lack of sequence similarity does not preclude similar function [23]. Furthermore, a substantial proportion of plant genomes lack sequence similarity to any characterized genes [1], making sequence similarity based molecular function inference inapplicable.

Genomic data can be used to infer not only molecular function but also biological process. For example, gene fusion events have been used to infer biological processes in prokaryotic and eukaryotic organisms, including plants [24]. Genomic context information can also be used to predict biological processes in prokaryotes. However, eukaryotes show this tendency to a much weaker degree than prokaryotes, although fungi and plants have been reported to contain clustered genes for specialized metabolism [25,26]. Finally, evolutionary context analysis [27] was applied to *Escherichia coli* and *Arabidopsis* genomes and assigned 19 gene families of unknown function to metabolic processes [28].

Transcriptomic data

Transcriptomic data capture changes in gene expression levels of all genes in an organism and represent a rapidly growing resource (Table 1). Co-expression analysis is based on the observation that functionally related genes often have similar expression profiles across different experiments, and has become a powerful tool for reverse genetics [29]. It is generally used to infer biological processes, although it has also been used recently to support the inference of cellular components [30]. Co-expression analysis has been successfully used to study many processes in plants, including secondary cell wall biosynthesis, fatty acid biosynthesis, and specialized metabolism [31–33]. In addition, transcriptional regulation appears to be conserved across species to a degree and co-expression analysis across species can enhance the predictive power to

Table 1. Availability and source of omics data relevant for predicting gene function

Plant species	Genomes (year published)[a]	Transcriptome RNA samples[b]		Protein interactions[c]	Genetic interactions[c]	3D structures[d]
		Microarray	RNAseq			
Arabidopsis thaliana	2000	26 747	391	16 697	171	2135
Oryza sativa	2002	5464	85	0	0	129
Vitis vinifera	2007	1064	7	0	0	19
Hordeum vulgare	2009	2558	4	0	0	111
Medicago truncatula	2009	1184	0	0	0	15
Zea mays	2009	3275	58	1	0	182
Glycine max	2010	1565	67	0	0	136
Solanum tuberosum	2011	1061	16	0	0	32

[a] Data from Genomes OnLine database (http://www.genomesonline.org/).

[b] Data from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/): downloaded June 16, 2013.

[c] Data from BioGRID (http://thebiogrid.org/).

[d] Data from NCBI (http://www.ncbi.nlm.nih.gov/structure/).

detect functional homologs [34–36]. However, the most commonly used microarray platforms for plants are missing ~40% of genes, which can lead to many false negatives. Luckily, data from RNA sequencing are becoming abundant enough to generate co-expression networks in *Arabidopsis* [37].

Interactomic data

Many biological processes such as photosynthesis and protein synthesis use multimeric protein complexes to perform their function. Therefore, the interaction of two proteins generally implicates that they act in the same biological process and cellular component [17]. High-throughput interaction data, often obtained through yeast two-hybrid [38] or tandem affinity purification coupled to mass spectrometry (TAP-MS) [39], are now available for several model organisms and have been used in biological process prediction [40,41]. Although a substantial amount of protein–protein interaction (PPI) data are available for *Arabidopsis*, only approximately 2% of its interactome has been tested [42]. There are virtually no PPI data for any other plant species in public repositories (Table 1). High-throughput TAP-MS is currently not applicable to plants in a high-throughput manner because it requires efficient transformation. PPI data often have false positive (spurious interactions) and false negative (missing relevant interactions) rates that can reach 45% and 50%, respectively [43]. Validating the interaction data with at least two independent studies has been suggested to improve the confidence and coverage of the interactions [44,45].

Genetic interactions

Genetic interaction (GI) measures the extent to which the phenotypes of one mutation in a gene are influenced by a mutation in another gene. GI screens look for pairs of genes that exhibit either suppression or enhancement of a phenotype when both genes are mutated, which would imply that both genes are involved in the same biological process. High-throughput detection of GIs enabled the rapid identification of biological pathways in yeast and predicted the roles of uncharacterized genes in *E. coli*, yeast, and mammals [46–48]. Unfortunately, only a limited number of GI mapping studies have been performed and there are little GI data available in public repositories [49], particularly from plants (Table 1).

3D structures

Structure-based function prediction is motivated by the observation that structurally related proteins often share a similar molecular function [50–52]. However, despite several structural genomics efforts, the number of available structures is limited in plants (Table 1).

Process of systematic gene function elucidation

Most of the omics data can be used to build co-function networks that are useful for inferring biological processes. The process of biological process inference using co-function networks can be broken down into seven steps, as shown in Figure 2. Typically, function inference uses the guilt-by-association concept that tries to find similarity between characterized and uncharacterized genes based on some shared feature, and transfers the annotation from knowledge donor (gene of known function) to knowledge acceptor (gene of unknown function) [53]. Function prediction begins with generating or obtaining omics data from public repositories (Table 1, Figure 2A), linking genes using a similarity measure (Figure 2B), and producing a gene co-function network (Figure 2C), where genes are represented as nodes and functional associations as links (also called edges in network analysis) between the genes [12,54]. Weighting the links and integrating the co-function networks generate integrated co-function networks (Figure 2D) [55] that can be used for gene function prediction and experimental validation of the predictions (Figure 2E,F) [12]. Experimentally validated gold standard data (Figure 2G) are crucial for assessing the quality of and integrating co-function networks [12,56]. A growing number of useful co-function and co-expression tools using plant data, some of which enable functional inference, are available online (Table 2).

Integrating co-function networks

There are two advantages in integrating different types of omics data to construct co-function networks [57]. First, one type of data often reveals only one domain of gene function. Therefore, combining data types can increase prediction coverage. Second, a predicted functional association between two genes is more likely to be true if it is supported by multiple, independent data sources. Various data types have been integrated and used for biological process prediction in *Arabidopsis* [12,54]. Integrating different data types has been shown to outperform single data type based co-function networks [12].

There are several ways of integrating different data to build co-function networks, which have been reviewed extensively [58]. Functional linkage is a binary classification problem (gene A is or is not linked to gene B) for which many machine learning and statistical algorithms exist. Popular machine learning algorithms include support vector machines, Bayesian network, decision trees, and neural networks (reviewed in [59]). The use of support vector machines has been applied to integrate *Arabidopsis* co-expression networks and protein sequences to improve cellular component prediction [30]. Bayesian network approaches are used extensively in function prediction for *Arabidopsis* using multiple omics data (Table 2). Decision trees have been trained to predict the function of yeast, mouse, and *Arabidopsis* genes by combining sequence and expression data [60]. Finally, neural networks have been used to combine protein sequence features to predict molecular function [61].

These methods can either predict gene function directly or be used to construct weighted co-function networks for inferring function using network properties. In addition, global properties of co-function networks can help improve the accuracy of function prediction. For example, normalizing the links in common between two genes against all the links the two genes have to other genes in the network can significantly increase the performance of network-based inferences [62].

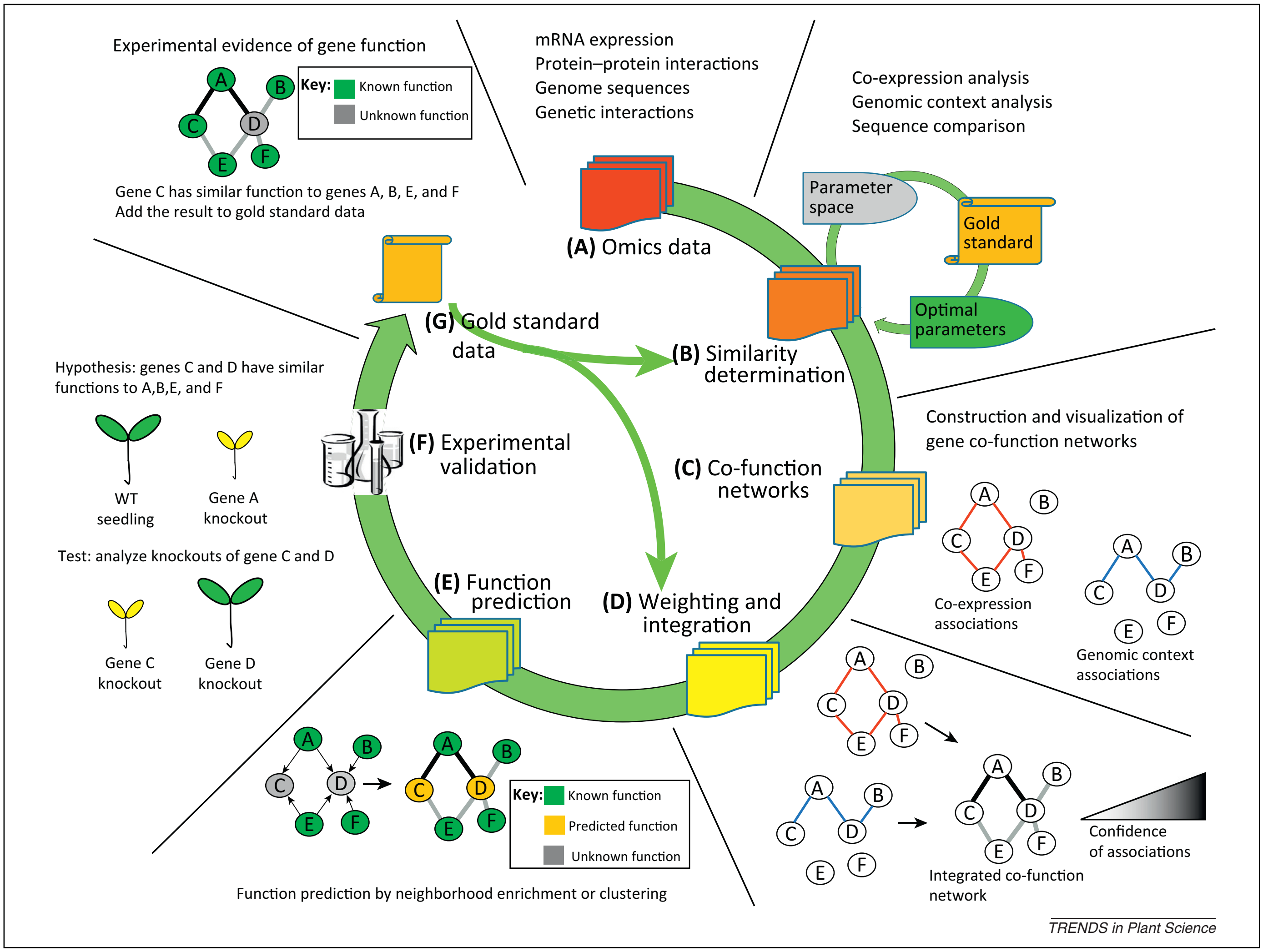
Experimental evidence of gene function
Key: 🟩 Known function
▨ Unknown function
A
B
C
D
E
F
Gene C has similar function to genes A, B, E, and F
Add the result to gold standard data
mRNA expression
Protein–protein interactions
Genome sequences
Genetic interactions
Co-expression analysis
Genomic context analysis
Sequence comparison
(A) Omics data
Parameter space
Gold standard
Optimal parameters
(G) Gold standard data
(B) Similarity determination
Hypothesis: genes C and D have similar functions to A,B,E, and F
WT seedling
Gene A knockout
(F) Experimental validation
(C) Co-function networks
Construction and visualization of gene co-function networks
Co-expression associations
Genomic context associations
Test: analyze knockouts of gene C and D
Gene C knockout
Gene D knockout
(E) Function prediction
(D) Weighting and integration
Key: 🟩 Known function
🟨 Predicted function
▨ Unknown function
Function prediction by neighborhood enrichment or clustering
Integrated co-function network
Confidence of associations
TRENDS in Plant Science

Figure 2. Process of systematic function elucidation using omics data. **(A)** Omics datasets are generated. **(B)** The data are analyzed by appropriate methods to determine similarities between genes. Gold standard data are used to find optimal parameter values and to train machine learning algorithms. **(C)** Each method generates a matrix of gene–gene associations and can be visualized as a network. **(D)** To integrate results from the different methods, gold standard data are used to weight the gene–gene associations. Associations consistent with gold standard data are assigned more weight, whereas poorly performing associations are assigned less weight. Weighted associations from the different data sources are integrated into one co-function network. **(E)** The network can now be used to predict gene function by using neighborhood enrichment, clustering, or other methods. **(F)** Predictions are used for focused reverse genetic testing, where uncharacterized genes can be associated with a biological process of interest. **(G)** New evidence from experimental validation is used to expand the gold standard data, which can be used to train and improve future predictions.

Besides integrating multiple data sources, integrating various inference methods can dramatically boost performance. The Dialogue on Reverse Engineering Assessment and Methods (DREAM) project assessed the performance of 35 transcriptional regulation inference methods in *E. coli* [63]. Although there was no clear winner among the inferences, integrating the methods produced an ensemble prediction that outperformed all the individual methods. Ensemble prediction methods have been developed for subcellular localization and transporters in plants [64–66].

Predicting and validating gene function

The biological processes of genes can be inferred from co-function networks either by using the enriched (statistically overrepresented) functions of network neighbors or by clustering genes and identifying the enriched or majority functions of the genes within a cluster [62,67] (Figure 2E). Clustering techniques largely fall into three categories: hierarchical, partitioning, and density-based (reviewed in [68]). Popular algorithms include Markov Cluster Algorithm (MCL) for its efficiency and performance (S. van Dongen, PhD thesis, University of Utrecht, 2000, http://micans.org/mcl/) and CFinder for its ability to find overlapping clusters [69]. There are various function enrichment techniques that can be used to help characterize the clusters (reviewed in [70]).

The quality of clusters can be assessed using various measures, including the compactness, connectedness, spatial separation, predictive power, and stability of the clusters [68]. In addition, clusters can be validated using external, 'gold standard' datasets and measures, such as purity, completeness, and similarity among the clusters. The performance of functional inference from clusters can be improved by using multiple lines of evidence supporting the links between genes [71,72]. Various clustering methods may work differently for different structures of networks. Therefore, systematic analyses of the effect of network properties on different prediction algorithms

Table 2. Tools useful for inferring the biological processes of plant genes

Tool	Network visualization	New member identification	Function prediction	Data types use[m]	Data integration scheme	Confidence of association or prediction shown[m]	Tool performance published[m]	Refs
AraNet[a]	Y	Y	Y	SS, CE, PPI, GI, GC	Bayesian	Y	Y	[12]
AFMSD[b]	N	Y	Y	SS, CE, PPI, GI, GC	Clustering	Y	Y	[92]
ATTED-II[c]	Y	Y	N	SS, CE	None	Y	Y	[93]
BMRF[d]	N	Y	Y	SS, CE, PPI	Bayesian	N	Y	[94]
CoP[e]	Y	Y	N	SS, CE	None	Y	N	[95]
GeneMANIA[f]	Y	Y	Y	SS, CE, PPI, GI	Support vector machine	Y	Y	[55]
GO-At[g]	N	Y	Y	SS, CE, PPI, GC	Bayesian	N	Y	[96]
PlaNet[h]	Y	Y	N	SS, CE	None	Y	Y	[97]
SCoPNet[i]	Y	Y	Y	CE	None	N	N	[98]
StarNet2[j]	Y	Y	N	SS, CE	None	Y	N	[99]
STRING[k]	Y	Y	N	SS, CE, PPI, GI, GC	None	Y	N	[100]
VirtualPlant[l]	Y	Y	N	SS, CE	None	N	N	[101]

[a] http://www.functionalnet.org/aranet/.

[b] http://bioinformatics.psb.ugent.be/cig_data/plant_modules/.

[c] http://atted.jp/.

[d] http://www.ab.wur.nl/bmrf/.

[e] http://webs2.kazusa.or.jp/kagiana/cop0911/.

[f] http://www.genemania.org/.

[g] http://www.bioinformatics.leeds.ac.uk/goat.

[h] http://aranet.mpimp-golm.mpg.de/.

[i] http://bree.cs.nott.ac.uk/arabidopsis/neighbor/network.php.

[j] http://vanburenlab.medicine.tamhsc.edu/starnet2.html.

[k] http://string-db.org/.

[l] http://www.virtualplant.org/.

[m] Abbreviations: CE, co-expression; GC, genomic context; GI, genetic interaction; N, no; PPI, protein–protein interaction; SS, sequence similarity; Y, yes.

could help determine the most appropriate clustering method to use on a given network.

Predictions can be followed up by focused reverse genetic approaches to validate the functions of uncharacterized genes (Figure 2F). Typically, this is initially done by reducing the expression of the candidate gene by using various methods described in [73]. Experimental validation can expand the gold standard data, which in turn can be used to train and improve any future prediction methods (Figure 2G).

Systematic gene function elucidation is not likely to progress linearly nor one gene at a time. The approaches described here enable the prediction of functions of hundreds to thousands of genes in a genome using properties such as network neighbors, linkage distributions, and clustering. The elucidation of a substantial proportion of the functions encoded in a genome will enable the generalization of properties of the genome.

Caveats for gene function prediction

There are some caveats involved in function inference, which we will briefly discuss to help scientists identify which predictions are likely to be less reliable, and indicate the areas that are likely to be the focus of future research.

Performance assessment

An assessment of the performance of prediction methods is rarely conducted as rigorously as it should be. For example, the granularity of functional inference is generally not considered when reporting the performance of inference methods. A method could have a high level of performance if it is restricted to predicting high-level functions. However, functional inference to a general, high-level term in the ontology (e.g., 'binding' GO:0005488) is not useful when trying to determine the role of a gene product. Therefore, the granularity of predictability should be considered when comparing different programs. In addition, most inference methods use all annotations, including those derived from other computational predictions, as a part of benchmarking data. This may lead to circular prediction problems or inaccurate assessment of true performance. Furthermore, gene product features that are tightly linked to function (e.g., catalytic residues in protein sequences or specific microarray experiments that boost co-expression prediction) are often not trivial to identify [74] and inappropriate features can lead to incorrect annotations [75]. Finally, it is important to use a community accepted set of gold standard data as an external standard when comparing the performance of different methods [53].

Annotation propagation

Generally, one type of omics data is useful for inferring gene function in a particular domain of GO. For example, sequence similarity can be used to infer molecular function, but not necessarily biological process or cellular compartment. However, common knowledge is sometimes used to infer other types of gene function using sequence similarity. For example, proteins containing predicted DNA-binding domains are inferred not only to have 'DNA binding' activity (GO:0003677, molecular function) but also to be located in the 'nucleus' (GO:0005634, cellular component) and involved in 'regulation of transcription' (GO:0006355, biological process). These secondary annotations are more often subject to the 'fallacy of converse accident' (e.g., if A then B, therefore if B then A) than the primary annotations [75] and should be avoided or used with caution.

Guilt-by-association

Guilt-by-association is a logical fallacy (e.g., Sue is a scientist. Sue has black hair. Therefore, all people with black hair are scientists). However, when constrained by knowledge in biology, the 'guilt-by-association' principle, which states that genes that have similar functions will share similar properties, has been successfully used to infer gene function [53]. Inferences made using 'guilt-by-association' are best suited for discovering novel members of known pathways. However, it may not be the best approach for discovering new pathways because it relies on assumptions made using our current knowledge of biology (e.g., relationships depicted in the GO system). Systematic discovery of novel mechanisms and pathways remains a big challenge that is likely to require an innovative combination of empirical and computational methods. For example, machine learning algorithms that use unsupervised or semi-supervised methods [76] combined with genome-wide network data may be more suitable for discovering novel mechanisms and pathways.

Importance of being a hub

Highly connected nodes (hubs) in networks have received much attention for their potential biological roles [77–81]. Although some of the hubs may represent true biological hubs or integrators of multiple processes, many hubs may not have such meaning *in vivo*. For example, many hubs in protein interaction networks are proteins that tend to interact with many proteins *in vitro* such as kinases or trafficking proteins. Similarly, ubiquitous molecules such as water and protons represent hubs in metabolic networks but their inclusion in network analysis may not be meaningful in analyzing network topologies or dynamics [82].

The number of links a gene has in a co-function network can confound the predictability of its function. For example, functions of hubs involved in multiple processes are easier to predict than functions associated with sparsely connected genes. Therefore, when using performance assessment and inference scoring schemes we should consider normalizing the scores against the number of functions a gene has as well as the number of genes to which a function is annotated [83]. Another possible solution is to perform a statistical test against a randomized network (ensuring that the node and edge degree distributions are retained) and consider only the inferences that are statistically significant.

Where do we go from here?

Network-based function inference, despite having some caveats, holds great promise for accelerating gene function discovery in plants. However, there are some bottlenecks that we must overcome to achieve the goal of understanding the function of all genes in a plant genome.

Data matter, perhaps more than algorithms

Although systematic comparisons of different prediction programs are few compared with the plethora of prediction programs, it appears that the quality of underlying data is at least as important as the different algorithms in dictating the performance [63]. Unfortunately, the limited amount of high-quality omics data and experimentally proven gold standard data in plants, particularly for plants other than *Arabidopsis*, is still a bottleneck (Table 1). Besides transcriptome and protein interactome data, we need to identify genome-wide protein subcellular localization and proteome data [64], enzyme active site information [84], post-translational modification data [85], ligand–protein interaction data [86], chromatin and epigenetic marks [87], and transcriptional regulatory information, protein complexes and pathways [88], to name a few.

In addition to expanding the types and depths of omics datasets, gold standard data are critical in assessing and improving function prediction [12,63]. To increase the coverage of gold standard data in public databases, we need more efficient ways of not only experimentally validating the predictions but also curating experimental data from the literature. Text-mining could help automate and triage curation efforts and collaborative projects such as the BioCreative initiative could improve the way gold standard data are extracted from the literature [89]. How much gold standard data do we need to transform the state of knowledge of gene function? A recent study has shown that for an organism where less than 20% of the genes are experimentally annotated (as is the case for most plant species), function prediction can reliably annotate 40% of all genes [56]. According to this study, 80% of reliable function prediction would necessitate 50% of genes with experimental annotation. How much gold standard data from one species can improve the predictability in another species is an open question. A reference set of gold standard data from multiple species is needed to study cross-species predictability of gene function.

Need to benchmark data integration and function prediction

The integration of heterogeneous inference methods is often more effective than single inference analysis [63]. Although variation in the performance of predictive methods poses a problem, it also offers a solution because the heterogeneity of predictions tends to boost true associations and cancel out the limitations of individual methods when they are integrated [12,63]. It is important to compare the performance of inference methods objectively and systematically as the critical assessment of protein function annotation (CAFA) experiment has done [57]. Such performance comparison can help integrate different methods by exploiting method-specific advantages and rationally including, excluding, or weighting different methods.

Need for a community repository of networks, gold standards, tools, protocols, and annotations

The number of networks and inference tools for plants is increasing (Table 2), but they are scattered in various websites and publications. For the entire plant research community to benefit from these tools, it would be useful to have a single, up-to-date portal from which these networks, inference tools, gold standard datasets, and protocols on how to use the tools could be accessed. Efforts such as the iPlant [90] and KnowledgeBase (KBase, http://kbase.science.energy.gov/) projects have been initiated recently to address this problem.

Need for high-throughput experimental testing platforms

Function inference methods typically use internal cross-validation using a proportion of the gold standard data that was partitioned from the training. Performance is rarely tested on external data that have never been seen by the algorithm. Even fewer studies perform experimental validation of predicted functions, and no study has yet performed systematic, large-scale functional validation of predictions in plants. To reach the goal of elucidating the functions of all genes in plants, we need to develop high-throughput experimental validation platforms. Pioneering efforts towards this goal have been initiated, which include high-throughput enzyme assays (BIOLOG, http://www.biolog.com/), plant phenotyping platforms (LemnaTech, http://www.lemnatec.com/), and metabolomics platforms [91].

Concluding remarks

Although network-based gene function prediction has been an active area of research for the past 15 years, its use in plant science has been limited. To exploit this underused technology we need, (i) more data; (ii) better assessment of data and tool quality; (iii) easy access to the data and tools; and (iv) high-throughput experimental validation. Many discoveries in plant science were made without knowing what most of the genes do. It is exciting to ponder what we will discover, those discoveries that are unimaginable now, when we are equipped with the knowledge of all the functions encoded in genomes.

Acknowledgments

Work in S.Y.R.'s laboratory is funded by the US National Science Foundation (grants IOS-1026003, DBI-0640769, and MCB-1052348) and the US Department of Energy (grant BER-65472). M.M. is funded by the Max Planck Institute for Molecular Plant Physiology. We thank Tanya Berardini from TAIR, Rachael Huntley from UniProt, and Marcela Monaco from Gramene for their help with accessing GO annotations, and Lee Chae and Eva Huala for their comments on the manuscript.

References

1 Lamesch, P. *et al.* (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210
2 Youens-Clark, K. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.* 39, D1085–D1094
3 Bard, J.B. and Rhee, S.Y. (2004) Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* 5, 213–222
4 Smith, B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255
5 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29
6 Moss, G.P. (2013) *Enzyme Nomenclature* (http://www.chem.qmul.ac.uk/iubmb/enzyme/)
7 Saier, M.H., Jr *et al.* (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res.* 37, D274–D278
8 Pujar, A. *et al.* (2006) Whole-plant growth stage ontology for angiosperms and its application in plant biology. *Plant Physiol.* 142, 414–428
9 Ilic, K. *et al.* (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol.* 143, 587–599
10 Thimm, O. *et al.* (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939
11 Rhee, S.Y. *et al.* (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515
12 Lee, I. *et al.* (2010) Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol.* 28, 149–156
13 Bassel, G.W. *et al.* (2012) Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *Plant Cell* 24, 3859–3875
14 Csete, M.E. and Doyle, J.C. (2002) Reverse engineering of biological complexity. *Science* 295, 1664–1669
15 Chou, K.C. and Shen, H.B. (2007) Recent progress in protein subcellular location prediction. *Anal. Biochem.* 370, 1–16
16 Tian, W. and Skolnick, J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.* 333, 863–882
17 Vazquez, A. *et al.* (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.* 21, 697–700
18 Pagani, I. *et al.* (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* 40, D571–D579
19 Schatz, M.C. *et al.* (2012) Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* 13, 243
20 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
21 Quevillon, E. *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120
22 Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* 318, 595–608
23 Galperin, M.Y. *et al.* (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* 8, 779–790
24 Nakamura, Y. *et al.* (2007) Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Mol. Biol. Evol.* 24, 110–121
25 Lee, J.M. and Sonnhammer, E.L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13, 875–882
26 Chu, H.Y. *et al.* (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.* 66, 66–79
27 Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96, 4285–4288
28 Gerdes, S. *et al.* (2011) Synergistic use of plant–prokaryote comparative genomics for functional annotations. *BMC Genomics* 12 (Suppl. 1), S2
29 Wu, L.F. *et al.* (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31, 255–265
30 Ryngajllo, M. *et al.* (2011) SLocX: predicting subcellular localization of *Arabidopsis* proteins leveraging gene expression data. *Front. Plant Sci.* 2, 43
31 Persson, S. *et al.* (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proc. Natl. Acad. Sci. U.S.A.* 102, 8633–8638
32 Han, X. *et al.* (2012) Co-expression analysis identifies CRC and AP1 the regulator of *Arabidopsis* fatty acid biosynthesis. *J. Integr. Plant Biol.* 54, 486–499
33 Maeda, H. *et al.* (2011) Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. *Nat. Chem. Biol.* 7, 19–21
34 Ficklin, S.P. and Feltus, F.A. (2011) Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. *Plant Physiol.* 156, 1244–1256
35 Movahedi, S. *et al.* (2011) Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in *Arabidopsis* and rice. *Plant Physiol.* 156, 1316–1330
36 Patel, R.V. *et al.* (2012) BAR expressolog identification: expression profile similarity ranking of homologous genes in plant species. *Plant J.* 71, 1038–1050

37 Giorgi, F.M. *et al.* (2013) Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics* 29, 717–724

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

39 Gavin, A.C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631–636

40 Meier, M. *et al.* (2013) Proteome-wide protein interaction measurements of bacterial proteins of unknown function. *Proc. Natl. Acad. Sci. U.S.A.* 110, 477–482

41 Hishigaki, H. *et al.* (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* 18, 523–531

42 Stark, C. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* 39, D698–D704

43 Huang, H. *et al.* (2007) Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* 3, e214

44 Hart, G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.* 7, 120

45 Yu, H. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104–110

46 Bandyopadhyay, S. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385–1389

47 Bassik, M.C. *et al.* (2013) A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* 152, 909–922

48 Babu, M. *et al.* (2011) Genetic interaction maps in *Escherichia coli* reveal functional crosstalk among cell envelope biogenesis pathways. *PLoS Genet.* 7, e1002377

49 Collins, S.R. *et al.* (2009) From information to knowledge: new technologies for defining gene function. *Nat. Methods* 6, 721–723

50 Brylinski, M. and Skolnick, J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 129–134

51 Roy, A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738

52 Drew, K. *et al.* (2011) The Proteome Folding Project: proteome-scale prediction of structure and function. *Genome Res.* 21, 1981–1994

53 Pavlidis, P. and Gillis, J. (2012) Progress and challenges in the computational prediction of gene function using networks. *F1000 Res.* 1, 1–14

54 De Bodt, S. *et al.* (2010) CORNET: a user-friendly tool for data mining and integration. *Plant Physiol.* 152, 1167–1179

55 Warde-Farley, D. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 38, W214–W220

56 Klie, S. *et al.* (2012) Inferring gene functions through dissection of relevance networks: interleaving the intra- and inter-species views. *Mol. Biosyst.* 8, 2233–2241

57 Radivojac, P. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods* 10, 221–227

58 Franzosa, E. *et al.* (2009) Computational reconstruction of protein–protein interaction networks: algorithms and issues. *Methods Mol. Biol.* 541, 89–100

59 Tian, W. *et al.* (2011) Predicting gene function using omics data: from data preparation to data integration. In *Protein Function Prediction for Omics Era* (Kihara, D., ed.), pp. 215–242, Springer

60 Schietgat, L. *et al.* (2010) Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics* 11, 2

61 Cozzetto, D. *et al.* (2013) Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinformatics* 14 (Suppl. 3), S1

62 Janga, S.C. *et al.* (2011) Network-based function prediction and interactomics: the case for metabolic enzymes. *Metab. Eng.* 13, 1–10

63 Marbach, D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804

64 Tanz, S.K. *et al.* (2013) SUBA3: a database for integrating experimentation and prediction to define the SUBcellular location of proteins in *Arabidopsis*. *Nucleic Acids Res.* 41, D1185–D1191

65 Kaundal, R. *et al.* (2010) Combining machine learning and homology-based approaches to accurately predict subcellular localization in *Arabidopsis*. *Plant Physiol.* 154, 36–54

66 Li, H. *et al.* (2009) TransportTP: a two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinformatics* 10, 418

67 Sharan, R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.* 3, 88

68 Handl, J. *et al.* (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21, 3201–3212

69 Adamcsek, B. *et al.* (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* 22, 1021–1023

70 Ackermann, M. and Strimmer, K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10, 47

71 Pereira-Leal, J.B. *et al.* (2004) Detection of functional modules from protein interaction networks. *Proteins* 54, 49–57

72 Asthana, S. *et al.* (2004) Predicting protein complex membership using probabilistic network reliability. *Genome Res.* 14, 1170–1175

73 Gilchrist, E. and Haughn, G. (2010) Reverse genetics techniques: engineering loss and gain of gene function in plants. *Brief. Funct. Genomics* 9, 103–110

74 Al-Shahib, A. *et al.* (2005) Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl. Bioinformatics* 4, 195–203

75 Graur, D. *et al.* (2013) On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590

76 Yip, K.Y. *et al.* (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol.* 14, 205

77 Ning, K. *et al.* (2010) Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. *BMC Bioinformatics* 11, 505

78 Wu, X. and Qi, X. (2010) Genes encoding hub and bottleneck enzymes of the *Arabidopsis* metabolic network preferentially retain homeologs through whole genome duplication. *BMC Evol. Biol.* 10, 145

79 Manna, B. *et al.* (2009) Evolutionary constraints on hub and non-hub proteins in human protein interaction network: insight from protein connectivity and intrinsic disorder. *Gene* 434, 50–55

80 Han, J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430, 88–93

81 Fox, A.D. *et al.* (2011) Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics* 27, 1135–1142

82 Arita, M. (2004) The metabolic world of *Escherichia coli* is not small. *Proc. Natl. Acad. Sci. U.S.A.* 101, 1543–1547

83 Gillis, J. and Pavlidis, P. (2012) Guilt by association” is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* 8, e1002444

84 Thibert, B. *et al.* (2005) Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics* 6, 213

85 Yao, Q. *et al.* (2012) P(3)DB: an integrated database for plant protein phosphorylation. *Front. Plant Sci.* 3, 206

86 Cheng, F. *et al.* (2012) Prediction of chemical–protein interactions network with weighted network-based inference method. *PLoS ONE* 7, e41064

87 O’Malley, R.C. and Ecker, J.R. (2012) Epiallelic variation in *Arabidopsis thaliana*. *Cold Spring Harb. Symp. Quant. Biol.* 77, 135–145

88 Navlakha, S. *et al.* (2012) A network-based approach for predicting missing pathway interactions. *PLoS Comput. Biol.* 8, e1002640

89 Arighi, C.N. *et al.* (2013) An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database (Oxford)* 2013, bas056

90 Oliver, S.L. *et al.* (2013) Using the iPlant collaborative discovery environment. *Curr. Protoc. Bioinformatics* 42, 1–26

91 Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171

92 Heyndrickx, K.S. and Vandepoele, K. (2012) Systematic identification of functional plant modules through the integration of complementary data sources. *Plant Physiol.* 159, 884–901

93 Obayashi, T. *et al.* (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.* 52, 213–219

94 Kourmpetis, Y.A. *et al.* (2011) Genome-wide computational function prediction of *Arabidopsis* proteins by integration of multiple data sources. *Plant Physiol.* 155, 271–281

95 Ogata, Y. *et al.* (2010) CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* 26, 1267–1268

96 Bradford, J.R. *et al.* (2010) GO-At: in silico prediction of gene function in *Arabidopsis thaliana* by combining heterogeneous data. *Plant J.* 61, 713–721

97 Mutwil, M. *et al.* (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell* 23, 895–910

98 Bassel, G.W. *et al.* (2011) Functional network construction in *Arabidopsis* using rule-based machine learning on large-scale data sets. *Plant Cell* 23, 3101–3116

99 Jupiter, D. *et al.* (2009) STARNET 2: a web-based tool for accelerating discovery of gene regulatory networks using microarray co-expression data. *BMC Bioinformatics* 10, 332

100 Franceschini, A. *et al.* (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815

101 Katari, M.S. *et al.* (2010) VirtualPlant: a software platform to support systems biology research. *Plant Physiol.* 152, 500–515