Review

CellPress

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

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

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.

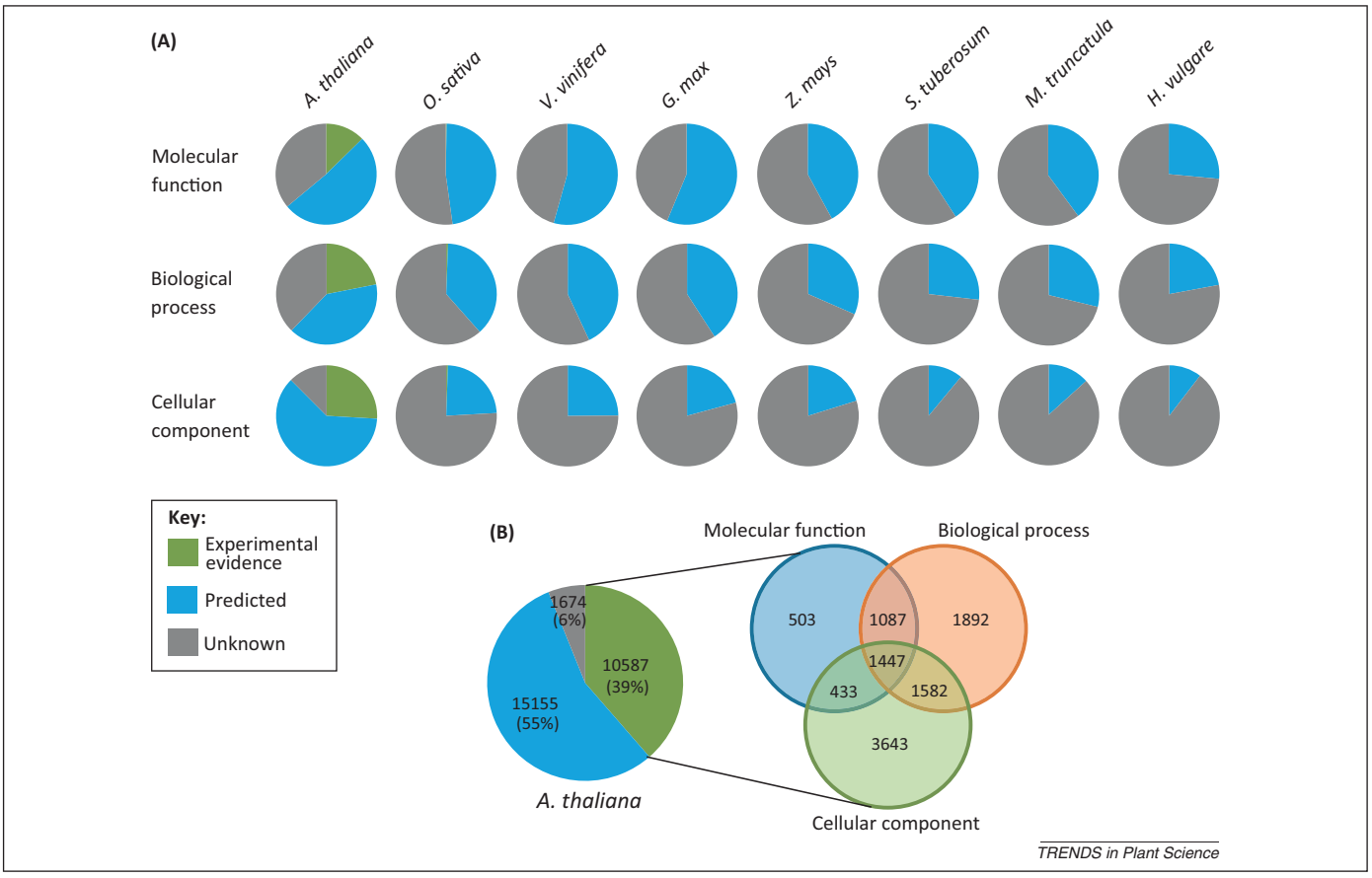

Figure 1. Status of gene function elucidation and annotation in plants: *Arabidopsis thaliana*, rice (*Oryza sativa*), grapevine (*Vitis vinifera*), soybean (*Glycine max*), maize (*Zea mays*), potato (*Solanum tuberosum*), *Medicago truncatula*, and barley (*Hordeum vulgare*). **(A)** Each pie chart 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's 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's 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