Function annotation enrichment assisting function prediction of plant genes

ABSTRACT
There are currently over 100 partially or entirely sequenced plant genomes but despite progress in sequencing, the function of the majority of plant genes remains largely unknown. Experimental researches cannot cover large-scale sequencing projects and this is where function prediction methods can assist. Most of them are based on sequence similarities but such approach is not entirely reliable because similarity does not always reflect homology. This study combines a phylogenomic approach with information from functional annotations, embedded into plant phylogenetic trees, which represent both sequence similarity and homology relationships between genes. Since functional annotations come from both manual and automatic annotations, an in-house semantic network approach was applied to improve their quality. This approach simulates the process of manual annotation and prevents generation of incorrect knowledge about gene function and domain distribution in de novo sequenced genomes. Then, the information from functional annotations, homology relationships and evolutionary distances based on tree topology was summarized for each gene using a scoring system for all potential functions. Functions with the highest score were assigned to uncharacterized genes or used as feedback to address inaccuracies in automatically annotated genes. As a result, genes were assigned into four major groups according to the score and the origin of their predicted function. These groups also reflect the reliability of the prediction. The predicted functional information is being involved in studying evolutionary relationships between genes responsible for the regulation of C3/C4 photosynthesis and can be applied to other major plant phenotypes.

Key words: functional annotation; function prediction; phylogenomics; plants; semantic network

Introduction
Genome analyses of plants have always been a challenge compared to other taxa due to variations in their size, complexity and ploidy. There are more than 100 plant genomes that are partially or entirely sequenced (Figure 1) but the function of the majority of their genes (except in a few model species) remains unknown.

Experimental researches that study gene function are an inseparable part of sequencing projects but cannot compete with neither the speed nor coverage of sequencing and need the assistance of in silico functional prediction methods. Most of these rely on database searches to identify homologous sequences of the uncharacterized gene and assign their function (if known) to it. But sequence similarity does not always reflect functional homology and this approach has its limitations. Such difficulties can be avoided by using information from phylogenetic tree reconstructions. Ancestral states along the trees can be inferred as well as a possible state for any uncharacterized genes in the tree and...
the results generally provide more accurate predictions of gene function than those based solely on sequence similarity. This approach is known as phylogenomic because it integrates evolutionary analysis with genome sequence analysis (Barton et al., 2007).

Even so, an accurate functional annotation of the genes in plant genomes is essential for accurate prediction. Using expert curators to assign functions to sequences is considered the “safest” approach but is far slower than automated annotation. It is a commonly held view that curated annotations are of better quality than automated annotations, however, the error rate of curated annotations can be significant (Jones et al., 2007). Often these errors are caused by the lack of well-structured controlled vocabularies, the inability to detect the context in co-occurrence of specific domains or sometimes they are caused by errors in sequence alignment tools. Manual annotation process, however, can be used as a standard for evaluation of automated annotation.

In our work we propose a function prediction pipeline that uses semantic network approach to improve the quality of the annotations and then apply the enriched and corrected functional annotations to a large set of phylogenetic gene trees. A custom designed scoring system is then used to predict the function of the unannotated genes in the trees according to their homology relationships with the annotated genes.

Materials and Methods

The overall approach of this study (Figure 2) includes obtaining phylogenetic and functional datasets, analyzing the links between them and designing a suitable prediction pathway.

![Figure 2. Function prediction pipeline.](image)

The approach also provides feedback on its initial steps and allows them to be corrected if necessary.

Source data

The phylogenetic data was obtained from the 26th release of the public database Ensembl Plants (Kersey et al., 2013) as flat file dumps containing plant phylogenetic trees in newick format. The data was processed and analyzed by either custom-written Python 2.7 scripts or with the instruments of the ETE Toolkit (A Python Environment for Phylogenetic Tree Exploration; http://etetoolkit.org) which uses the same programming language (Huerta-Cepas et al., 2010).

Functional annotation data was gathered from several databases: the Gene Ontology Consortium; the Plant Ontology Consortium, UniProt; InterPro; Prints and Pfam. A resource description framework (RDF) standard was applied to represent database records and their relationships to controlled vocabularies, families, patterns, motifs and more entities used in the process of annotation.

Semantic network approach

Formal description of entities and relationships between the databases were used to build a semantic network model (Figure 3) that is used to automate large number of checks performed by curators in order to discard false positive annotations.

A rule-based inference engine was designed to describe these formalisms, so that they can be managed to direct decision capabilities of semantic network in order to produce specific knowledge about relationships between sequenced regions and records in public databases (Decker et al. 2000).

Identification of homology relationships

To detect the evolutionary relationships between the annotated genes and uncharacterized genes, a homology
prediction was conducted using ETE’s Species Overlap (SO) algorithm (Huerta-Cepas et al., 2007). It uses the level of overlap in the species connected to two related tree nodes (children) to decide whether their parental node represents duplication (paralogy) or speciation (orthology) event.

Figure 3. A RDF-based semantic model representing the connections (arrows) between the databases (circles) used to obtain functional annotation data.

**Design of function scoring system**

The scoring system used to predict the function of uncharacterized genes included several parameters:

1) **Type of relationship** – Only orthologous relationships were taken into account because paralogs commonly do not share function.

2) **Branch length** – Branch length represents the genetic change (mutations per site) between nodes. The average value of the sums of branch lengths between the query gene/annotated orthologs and their closest common parent indicates how closely related they are (smaller value means closer relationship) and thus it was subtracted from the score.

3) **Accuracy of annotation** – Manual annotations were preferred to automatic, whenever possible.

4) **Frequency of annotation** – As each phylogenetic tree represents genes that are related to each other, their functional annotations are expected to be similar. A gene, however, often has more than one annotation so only annotations that were present in 50%+1 of the other annotated genes were assigned to the query genes.

The scoring system was tested to manually annotated genes from model plant species before being applied to the uncharacterized genes. For each query gene, the functional annotations with the highest score were assigned.

**Results**

After disposal of single-species trees the phylogenetic dataset contained 74 754 trees with 2 283 388 genes from 39 plant and 5 animal species. Since the removal of animal genes requires reconstruction of the entire tree that sometimes leads to losing it, these genes were ignored in the further analyses.

The phylogenetic trees were then linked to the processed functional annotations (Figure 4) and a list of the annotations (exported as Gene Ontology (GO) terms) for the matching genes, was extracted from the databases.

Figure 4. Connections between gene IDs in the phylogenetic dataset and GO terms in the gene ontology tree.

Annotated genes were found in 26% (19 219) of the trees containing 78% (1 790 304) of the genes and the analysis continued with them. To 35% (627 710) of these genes a total of 3 056 433 functional annotation were assigned. The number of annotation is much larger than the genes because most genes have more than one annotation and the list includes all of those repetitions.

The results from the homology prediction algorithm were indicated all possible homology relationships between the genes in a tree, so the output was processed to remove repetitive results. For each unannotated gene in a given tree, a score for each possible functional annotation was assigned.
according to the parameters described in Materials and methods.

According to the obtained scores, the following groups could be distinguished:

1) Genes with predicted function based on manual annotations or automatic annotations.
2) Genes with none, insufficient or contradicting information – the function of such genes could either not be predicted by this method or may be caused by imperfections in the scoring system.
3) Conflicts between already annotated genes
4) Genes with predicted function based on manual annotations or automatic annotations.
5) Genes with predicted function based on manual annotations or automatic annotations.

The accuracy of the proposed function prediction method depends on the accuracy of the annotations but is also influenced by the size of the trees and the correct tuning of the scoring system. The last implies that a threshold needs to be defined that ensures the prediction is correct.

Another major issue is the relatively small amount of functional annotation data available for plant: less than 25% of the sequenced plant genes are annotated compared to almost half of the genes in vertebrate animals. De novo annotations from NGS data may be the key to lower this ratio.

The predicted functional information is being involved in studying evolutionary relationships between genes responsible for the regulation of C3/C4 photosynthesis and can be applied to other major plant phenotypes. Also, the described approach is currently being applied in the development of a web-based tool that could, predict, display, and update functional annotation of user genes.

**Acknowledgement**

This study is supported by the National Science Fund of Bulgaria within the project "Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets" under contract DFNI-102/7, 12th December, 2014.

**References**


RDF Core Working Group. 2014. RDF is a standard model for data interchange on the Web. https://www.w3.org/RDF/