Title
Gene Annotation in Rice Genome Using Homology

Author
Athakitkarnka, Pavitra

Publication Date
2013

UNIVERSITY OF CALIFORNIA

Los Angeles

Gene Annotation in Rice Genome
Using Homology

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Biomedical Engineering

by

Pavitra Athakitkarnka

2013

ABSTRACT OF THE THESIS

Gene Annotation in Rice Genome

Using Homology

by

Pavitra Athakitakarnka

Master of Science in Biomedical Engineering

University of California, Los Angeles, 2013

Professor Matteo Pellegrini, Chair

In the functional genomics era, systems-level analysis of genomic and proteomic data quickens the pace of gene-function discovery. I created a program and collected cross-referenced database for the rice genome. This program utilizes sequence similarity as a way to annotate all genes with the gene ontology (GO) terms and performs a functional analysis on genes in rice, an economically important crop with limited functional analysis tools. For demonstration purpose, application of the tool to one gene list created from cytokinin oxidase/ dehydrogenase activity reveals two GO terms that suggest the relevance for cytokinin associated terms as cytokinins was used to generate the gene list. To my knowledge, while there are tools available on the web to interpret the biological function of a gene list, none are set up for use with rice protein sequences. My program overcomes this limitation by permitting the functional analysis of rice amino acid sequences. The example gene list and code are available to reproduce the results.

The thesis of Pavitra Athakitkarnka is approved.

Thomas Graeber

Xinshu Xiao

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2013

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgement

I am deeply grateful for all those who made it possible to complete this thesis, especially my committee chair, Professor Matteo Pellegrini, for his guidance during my research. Professor Thomas Graeber and Professor Xinshu Xiao deserve a special thanks as my thesis committee members. In particular I would like to thank Professor Graeber for his input. I would like to also thank Dr. Dimmer and Dr. Lomin, the authors of "The UniProt-GO Annotation database in 2011" and "Receptor properties and features of cytokinin signaling," and their publishers (Oxford University Press and Park-media Ltd) for permission to reprint their figures.

Background

In the functional genomics era, systems-level analysis speeds the pace of gene-function discovery. The number of genes that have been functionally characterized in rice in the Rice Genome Annotation Project (*Oryza sativa* subspecies *japonica* and *indica*) lags far behind that of *Arabidopsis thaliana*, which is an experimentally tractable model organism.[4] After the rice genome has been sequenced completely, the amino acid sequences of putative rice genes were generated by the Rice Genome Annotation Project. These putative genes are candidates that would benefit from experimental work to further understand their functions. This experimental approach, however, involves labor-intensive techniques used for experimentally validating the putative genes. [4] To facilitate the discovery of novel genes and their biological context, the program that automatically identifies putative functions and biological properties [5] of putative rice genes hold promise for providing clues for experiments most likely to yield successful results [6].

Sequence annotation is one of the computational approaches which overcome the limitations of the experimental approaches and allow for simultaneous characterization of all gene candidates. It takes advantage of sequence and functional conservation among all eukaryotes. Eukaryotes inherit sequences and functions of homologous genes from a common evolutionary ancestor. [7] Knowledge of the relevant features and biological roles of a shared gene in one organism allows us to infer its biological functions in another organism [8]. The shared genes whose functions have been experimentally determined in one organism can be used to infer the functions of related genes in other organisms. We take advantage of genomic DNA sequence known in one organism to predict new genes and their pathways and phenotypes in another organism. The shared genes guide experimental design to enable characterization of DNA sequences encoding proteins with no known function [9]. These predictions have led to the identification of thousands of new genes, some of which have been experimentally validated. [4] This type of sequence annotation represents a way to efficiently fill in the knowledge gap.

One way to predict genes and their functions in rice is automatic protein sequence annotation [10-12]. Protein sequence annotation is more efficient when used to align the genomic DNA segments with proteins rather than with transcribed DNA sequences from other organisms. [13] In specific, this technique of protein sequence annotation helps predict a gene based on sequence similarity between its amino acid sequence and its homologous protein sequences. From its homologous protein sequences, we can infer structure and function of the amino acid sequence into which a DNA gene is translated. Thus homology is used for its structural and functional annotation [10]. For example, a locus (recognized by a locus identifier) in the genomic DNA sequence is predicted to be a putative gene due to sequence similarity between its amino acid sequence, into which this predicted gene is translated, and its homologous protein sequences.

In the context of protein sequence annotation, homology of an amino acid sequence to another protein sequence refers to the existence of a common ancestor between the two sequences as discussed above. Homology is detected by sequence similarity of this amino acid sequence with known proteins. Homology is a transitive relation where given three proteins A, B, and C, proteins A and B as well as proteins B and C have structural similarity due to a common evolutionary ancestor (implied by sequence similarity above a threshold). Hence proteins A and C have structural similarity due to a common evolutionary ancestor. Since homology is evidence for structural and functional annotation of protein sequences, protein sequence annotation is also a transitive relation. In short, sequence similarity between an amino acid sequence and its homologous protein sequences enable the transfer of functional annotation from homologous protein sequences to the amino acid sequence. This technique of protein sequence annotation represents sequence alignment for functional interference. (However, this technique has limitations when dealing with multi-domain proteins, promiscuous domains, and fragmented proteins which would benefit from another mode of functional interference, i.e. the presence of protein domains (domain search).)

A genomic version of this technique of functional annotation includes annotating putative functions of hypothetical amino acid sequences of the putative genes. The hypothetical amino acid sequences are gene products of the putative genes. To investigate the functional and the structural information of the hypothetical amino acid sequences of genes [14], we have used the gene ontology (GO), which is the standard for annotating genes. The GO is the result of an effort to provides a consistent terminology to describe functional information pertaining to gene products and applicable across information repositories [14] when knowledge of what a gene and gene product for which it encodes does is changing rapidly[15]. The GO consists of three ontologies used to describe Molecular Function, Biological Process and Cellular Compartment attributes of gene products. [8] The GO describes the normal biochemical activities and biological goals that a gene product is involved in as well as captures its place in non-disease cells where it is active.[8] What I would like to achieve in this project is to determine the functions associated to the amino acid sequences of putative rice genes at loci (recognized by locus identifiers) in the rice genomic DNA sequence. In doing so I make use of the GO term annotations of rice-specific proteome provided by UniProt (The Universal Protein Resource) [3], which is a central member of the GO consortium and maintains high-quality gene annotation data.

Data and Method

I. The Importance of Annotating Rice Genome

In this project, I created a tool to use the gene ontology (GO) terms to annotate functions of all putative genes in the genome of the two rice subspecies *japonica* and *indica* in the Rice Genome Annotation Project and then perform GO term enrichment from the annotations of these putative genes. Functionally annotating and then charactering rice genes using my program might offer the opportunity for both life science and agricultural research, including improvement of grain yield.[16] *Oryza sativa* is one of the most important grain crops because it is staple food for people especially those in Asia, Latin America and Africa. Additionally, *Oryza sativa* has emerged as a model plant for other grass species including major crop species maize (*Zea mays*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), sorghum (*Sorghum bicolor*), oats (*Avena sativa*) and millet (*Eleusine coracana*) even though it has not been as extensively studied in the lab as *A. thaliana*.[9] Due to the conservation of gene order within linkage groups in these agriculturally important grasses[17], researchers can take advantage of genes known in rice for investigating genes, cellular pathways, phenotypes in other grass species in further analysis.

II. Collecting Database and Cross Referencing

Towards the goal of facilitating our ability to gain insight into a given list of putative genes in rice and other grasses, I wrote a program to access annotations and identify enriched functions of putative rice genes and their amino acid sequences. My program can be divided in to several steps (schematically illustrated in Figure 1). To begin with, I explain how I created the total set of genes from the Rice Genome Annotation Project data in Step 1 (as shown schematically in Figure 2). In addition, I summarize the efforts to create the UniProt protein database. Next, taking advantage of the gene data made available by the Rice Genome Annotation Project, I use the rice genome database to include genes into the total set. Last, I discuss the principles of the basic local alignment search tool (BLAST) implemented by the National Center for Biotechnology Information (NCBI) to find local similarity between two amino acid sequences.

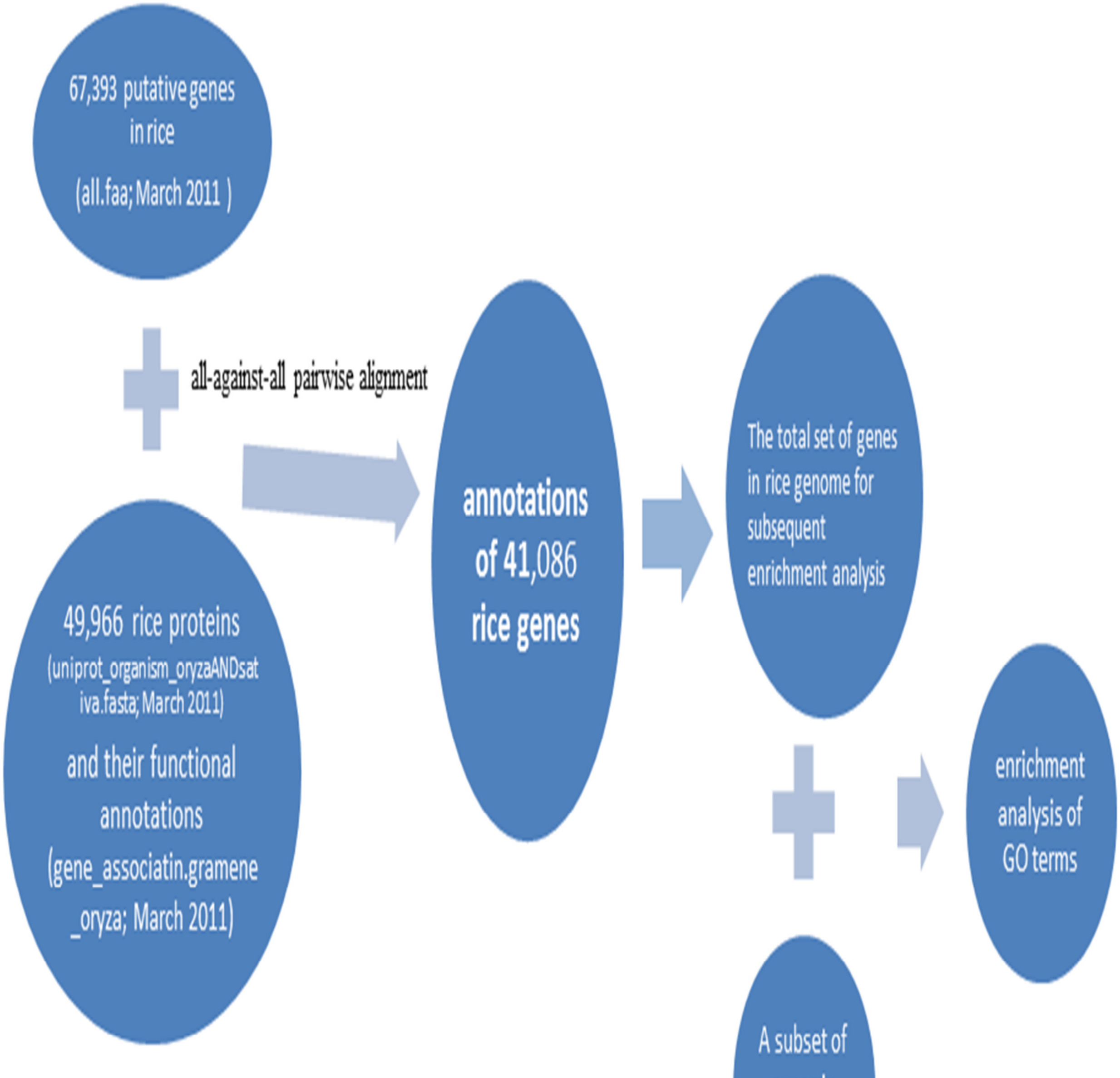

67,393 putative genes in rice
(all.faa; March 2011)
all-against-all pairwise alignment
49,966 rice proteins
(uniprot_organism_oryzaANDsativa.fasta; March 2011)
and their functional annotations
(gene_associatin.gramene_oryza; March 2011)
annotations of 41,086 rice genes
The total set of genes in rice genome for subsequent enrichment analysis
A subset of genes in rice genome
enrichment analysis of GO terms

Figure 1 Schematic showing how the program created in this project functionally annotates the rice reference genome by homology and then functionally categorize a subset of rice genes by GO term enrichment analysis

Step 1: Gene Data Sets and GO Annotations

A reference rice genome was taken from the Rice Genome Annotation Project while all of the rice proteins and their functional annotations were downloaded from UniProt. A full set of the putative genes in the genome of the two rice subspecies *japonica* and *indica,* along with their translated amino acid sequences (all.faa; March 2011; 67,393 amino acid sequences) were taken from the Department of Plant Biology at Michigan State University website (http://rice.plantbiology.msu.edu/data_download.shtml). All of the rice-specific protein sequences (uniprot_organism_oryzaANDsativa.fasta; March 2011; 47,742 proteins) and their functional annotations (gene_associatin.gramene_oryza; March 2011; 49,966 proteins) were downloaded from UniProt (http://www.uniprot.org/taxonomy/39947) and the GO consortium website (http://www.geneontology.org/GO.downloads.annotations.shtml) respectively. These two online data sets (the Rice Genome Annotation Project and UniProt) are publicly available. The data set of functional annotation from UniProt contains 49,966 rice protein sequences each

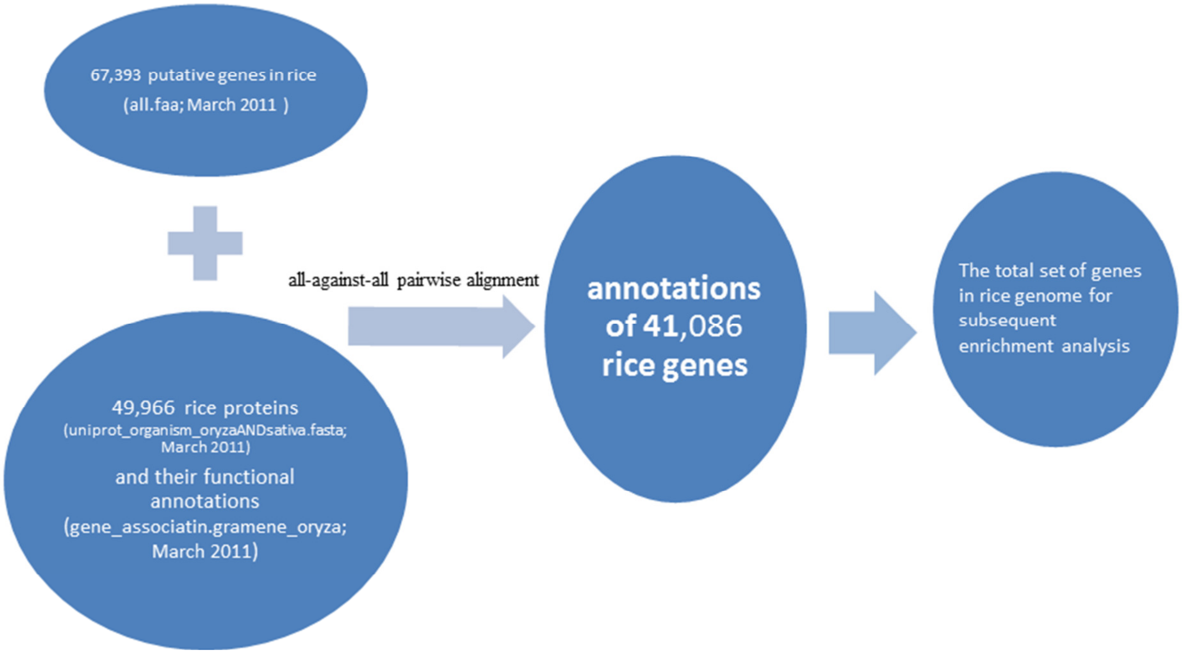

67,393 putative genes in rice
(all.faa; March 2011)
all-against-all pairwise alignment
49,966 rice proteins
(uniprot_organism_oryzaANDsativa.fasta; March 2011)
and their functional annotations
(gene_associatin.gramene_oryza; March 2011)
annotations of 41,086 rice genes
The total set of genes in rice genome for subsequent enrichment analysis

Figure 2 production of the total set of genes in rice genome for GO term enrichment analysis of which is functionally annotated by a GO term. Another dataset from the Rice Genome Annotation Project contains 67,393 amino acid sequences of all the genes in the rice genome. BLAST alignment of the amino acid sequences of all genes in the rice genome against the rice-specific protein sequence database, followed by transferring GO term annotations from the UniProt rice-specific protein sequences to these amino acid sequences in the Rice Genome Annotation Project identified 41,086 close matches, and thus resulted in the production of GO term annotations of 41,086 putative genes in the rice genome. These GO term annotations were, in this place, GO term annotations of 41,086 locus identifiers in the rice genome, and thus became the total set of genes in the rice genome for subsequent enrichment analysis of GO terms. By the total set of genes in the rice genome, I mean the follows: the reference rice genome, or all putative genes in the rice genome, or all putative rice genes is the full set of putative genes in the two rice subspecies *japonica* and *indica*.

As background to the GO annotations being used, UniProt (The Universal Protein Resource) provides a gene annotation file (gene_associatin.gramene_oryza; March 2011) of the rice-specific proteome (uniprot_organism_oryzaANDsativa.fasta; March 2011) [18]. The GO is a structured source of information. I elected to include proteins and their respective annotations that are related to *Oryza sativa* subspecies *japonica* only. To construct the most accurate proteome based on existing evidence, UniProt only includes canonical sequences and isoforms. Each protein is a product encoded by one gene. In addition, annotations of rice proteins which are obtained from the UniProt database were done using GO terms (GO term annotations of rice proteome) as depicted in Figure 3. UniProt proteins have both manually curated GO term annotations and predicted GO term annotations. To help provide high-quality descriptions of the functions of gene products, the GO consortium uses curators who are experts in gene families

[10, 19] to review experimental evidence as well as the automatic predictions by Ensembl, EnsemblGenomes, InterPro and UniProt. These automatic predictions exploit gene orthology, protein sequence signatures, and existing cross-references from external controlled vocabularies. If no accession numbers are available for some proteins, these proteins will not be annotated with GO terms. As a consequence, the UniProt data set of functional annotation (gene_associatin.gramene_oryza; March 2011) contains only 49,966 rice proteins out of all 63,535 proteins in the rice proteome set (uniprot_organism_oryzaANDsativa.fasta; March 2011).

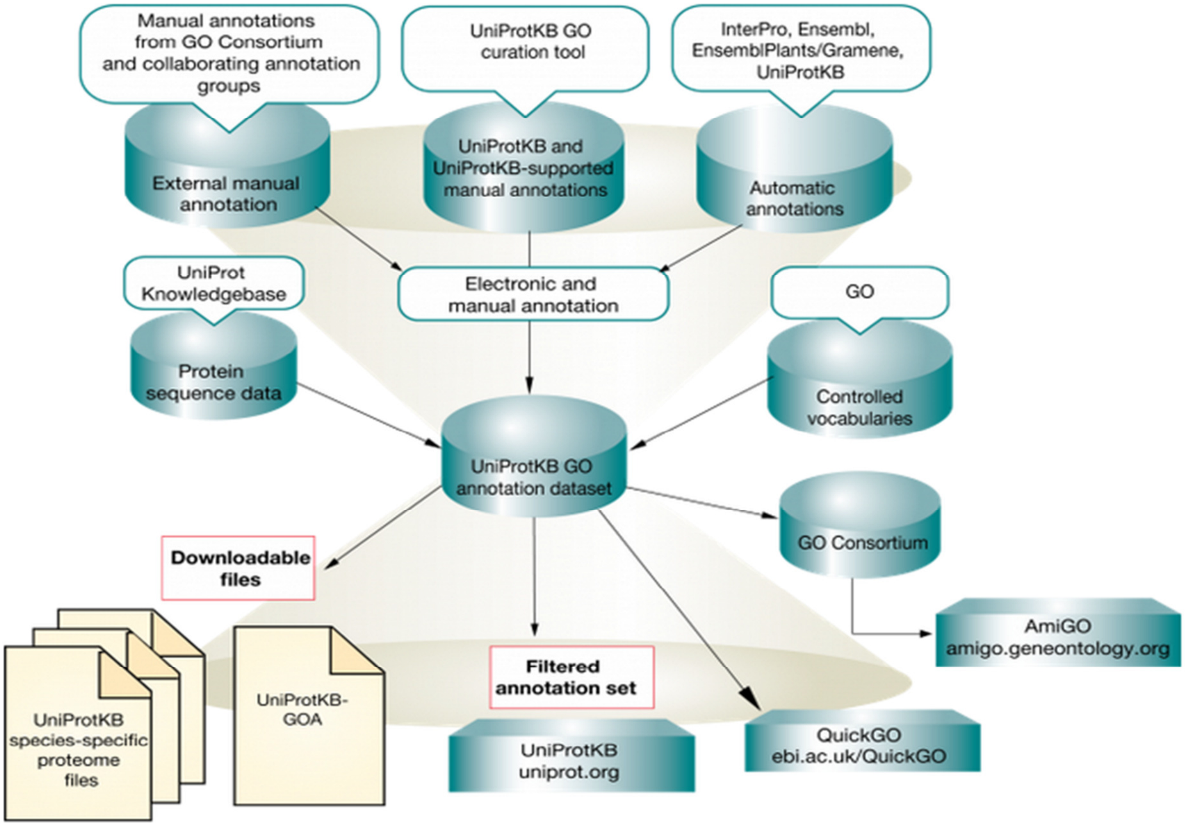

Manual annotations from GO Consortium and collaborating annotation groups
UniProtKB GO curation tool
InterPro, Ensembl, EnsemblPlants/Gramene, UniProtKB
External manual annotation
UniProtKB and UniProtKB-supported manual annotations
Automatic annotations
UniProt Knowledgebase
Electronic and manual annotation
GO
Protein sequence data
Controlled vocabularies
UniProtKB GO annotation dataset
GO Consortium
Downloadable files
AmiGO amigo.geneontology.org
Filtered annotation set
UniProtKB species-specific proteome files
UniProtKB-GOA
UniProtKB uniprot.org
QuickGO ebi.ac.uk/QuickGO

Figure 3 flow of data to create a gene annotation file: Proteins from UniProtKB are annotated using GO terms[1] and AmiGO is a web application that gives access to the resources at the GO consortium. Using GO terms, members of the GO consortium, of which UniProt is a central member, submit annotations to the annotation dataset. [1]

viii

[1] Reprinted with permission from "The UniProt-GO Annotation database in 2011,"by E. C. Dimmer, *et al.*, Jan 2012. *Nucleic Acids Res,* vol. 40, pp. D565-70, Copyright the Author(s) 2011. Published by Oxford University Press

Step 2: Inclusion of Putative Genes

Equally important for creating the total set of all genes in the rice genome for enrichment analysis of GO terms is the inclusion of putative genes. For this purpose, I used the Osa 1 genome annotation database (The Institute for Genomic Research (TIGR) rice genome annotation database) which collects putative genes for *Oryza sativa* subspecies *japonica* and *indica* (all.faa) from the Rice Genome Annotation Project. Each putative rice gene is recognized by a unique locus identifier (LOC_OsXXgXXXX), a model of the gene structure and functional annotations. [10] This attempt to identify all the gene content in the genomes of *japonica* and *indica* was first carried out by the Institute for Genomic Research (TIGR). TIGR references a gene to a gene index, or initially a pseudo-molecule, through the probes and primers used to identify the loci as shown in Figure 4. The resulting rice genome data set contains all putative genes in the rice genome. The size of the current data set of the rice genome is 63,535 putative rice genes. In addition to the results of using gene prediction programs, TIGR gathered evidence for genes in the rice genome from the rice full-length cDNA and assemblies of expressed portions of the genomic DNA sequences (or assemblies of expressed sequence tags (EST))[20]. In 2002, TIGR also integrated protein sequence homology as evidence for genes to support the identification of functional genes in the genomic DNA sequence. In 2007, a tool called EuCAP was built to enable the rice experts (called community curators) to submit functional annotations of the pre-determined genes, and these pre-determined genes are subsequently matched with the genes in the Rice Genome Annotation Project data set.[10] Today's putative rice genes in the current rice genome maintained by the Rice Genome Annotation Project contain evidence from different studies including the results of gene prediction programs, DNA transcription as well as homology to fully characterized proteins.[21] In 2011, Rice-Map was built to incorporate epigenetic modifications, including DNA methylation, to provide additional evidence for the genes in the rice genome of the Rice Genome Annotation Project where no other biological information is available [10]. Despite the effort of the Rice Genome Annotation Project, there is still much room for improvement to identify gene content in the genomes of *Oryza sativa* subspecies *japonica* and *indica*. To my knowledge, no tool has been set up to use GO terms to annotate the amino acid sequences of rice genes in the Rice Genome Annotation Project. Because GO term annotations of amino acid sequences enable us to subsequently perform enrichment analysis, functional characterization works well for any input gene list. This makes the tool more suitable for rice's situation where results of DNA methylation experiments are locus identifers.

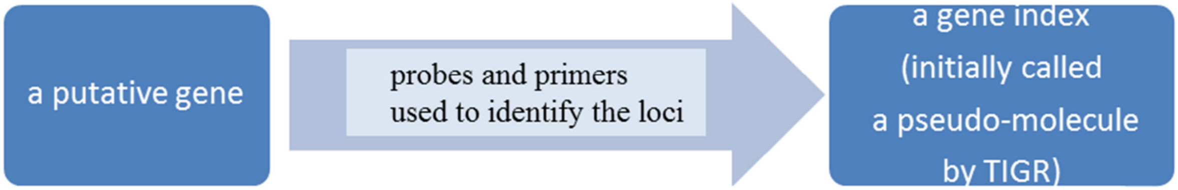


Figure 4 the reference rice genome by mapping putative genes to the gene indices in the genome

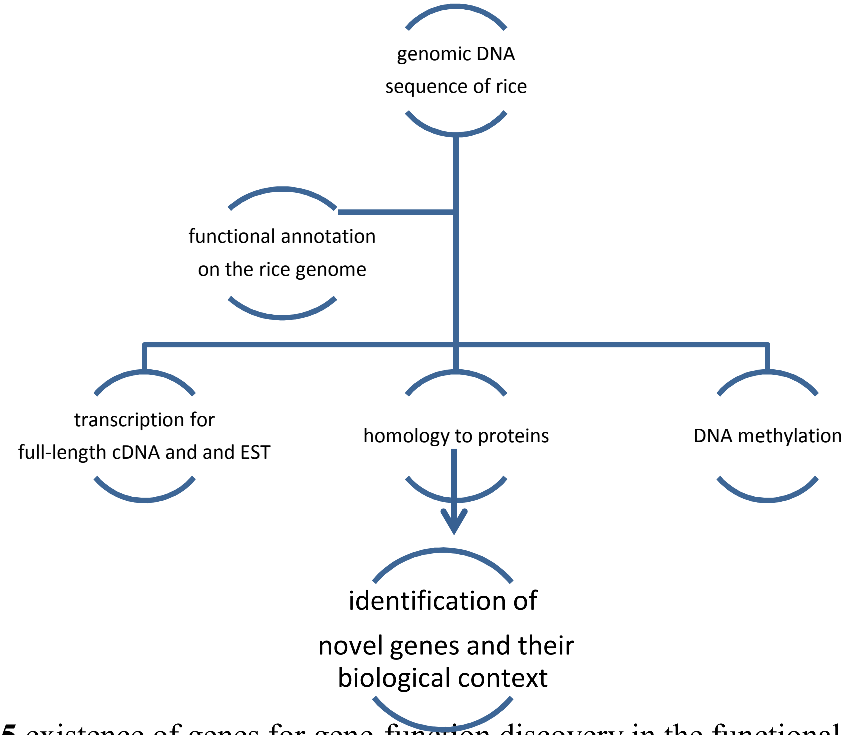


Figure 5 existence of genes for gene-function discovery in the functional genomics era

Step 3: Homology Mappings

I use BLAST as a way of assessing homology, and thus transferring GO term annotations to putative genes. Most importantly, BLAST is an algorithm to search for regions of local similarity by comparing pairs of sequences using a heuristic approach. Generally, of the various tools for comparing similarities between biological sequences, the most widely used is ncbi BLAST. First, BLAST measures sequence similarity by approximate matches in local pairwise alignment. Put simply, sequence similarity is computed between two protein sequences by scoring matches/ mismatches according to some measure of 'sameness'/ 'difference'[22]. A commonly used measure of matches/ mismatches is a matrix for substitutions of amino acids with respect to a pair of sequences, which is calculated from the probability of a biologically meaningful amino acid pair occurring in an alignment [22]. Commonly used substitutions matrices are BLOSUM62 (blocks substitution) and PAM30 (point accepted mutation). It is true that BLAST local alignment is similar to global pairwise alignment with regard to searching for approximate matches. However, insertions and deletions with respect to mutation events are penalized differently in local alignment by comparison to global alignment. Although local alignment and global alignment both penalize for opening a gap (an affine gap penalty) and having an additional gap (an extension penalty)[22], they use different numerical values to achieve different goals. The highest scoring local alignment is achieved by aligning segments of one sequence to segments of the other. On the other hand, the global alignment aims to align two entire sequences from beginning to end. Second, BLAST makes an assumption to decrease the number of comparisons of letters between a pair of sequences being aligned.[22]. Stated briefly, BLAST assumes that the best local alignment will contain at least some words ('hits') that are high-scoring. During the first pass, BLAST finds the high-scoring 'hits' that are not exact matches between the pair of sequences being aligned but are high-scoring alignments nevertheless. Then in the subsequent pass, BLAST ‘extends’ the hits. Prior to BLAST, the highest scoring local alignment of two sequences is exhaustively searched using dynamic programming algorithms, such as Smith-Waterman algorithm to compare every letter to every other. In other words, every single letter of one sequence is compared to every single letter of the other sequence. Certainly, heuristic approaches such as BLAST decrease the search space and therefore increase the speed of finding , but are not guaranteed to find the highest scoring local alignment that Smith-Waterman algorithm is guaranteed to find (but BLAST almost always do).

III. Synthetic Gene List: Proteins Involved in Cytokinin Accumulation and Increased Grain Yield

Previously I discussed how GO term annotations of UniProt protein sequences were transferred to the reference rice genome from the Rice Genome Annotation Project. Now I will discuss below how I sought a list of genes to which genes for known processes are added (as depicted in Figure 6) so that I subsequently applied a GO term enrichment analysis on this input gene list as a demonstration case. In this example gene list described below, I focus on the cytokinin regulating- cytokinin aspect of the OsCKX2 gene. Phytohormone (plant hormone) cytokinin (http://www.plant-hormones.info/cytokinins.htm) plays an essential role in promoting cell division, growth and development. The decrease in OsCKX2 expression has been shown to cause cytokinin accumulation in fluorescence meristems and increase the number of reproductive organs, resulting in increased grain yield. [23] Knowing the biology of OsCKX2 will help engineer the rice that has increased yield. Through querying OsCKX2 on Gramene (http://gramene.org/protein/ ; a comparative mapping resource for grains) for its biological properties, I found that this protein is associated with a single GO term in the Molecular Function ontology, i.e. ‘cytokinin dehydrogenase activity’ whose GO identifier is GO:0019139. Afterwards, this information was used by my program to identify additional proteins. To try to create an example set of genes of interest, I queried by ‘cytokinin dehydrogenase activity’ [24] to search for associated rice-specific protein sequences in UniProt at the GO website using AmiGO version 1.8 (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi; a web app to browse the gene ontology) [25] as shown in Figure 3. After I queried ‘cytokinin dehydrogenase activity’, AmiGO returned a set of 29 orthologous proteins that were annotated to this GO term ‘cytokinin dehydrogenase activity’. Four out of these 29 proteins which share ‘cytokinin dehydrogenase activity’ are in *Oryza sativa*. [24] In fact, these four rice proteins were experimentally annotated with ‘cytokinin dehydrogenase activity’. To pull a list of protein sequences related to cytokinin dehydrogenase activity, these four rice proteins known to be annotated with cytokinin dehydrogenase activity was downloaded in the GO consortium’s gene association file format (idCytokininDehydrogenase.txt).

Subsequently, in order to test my MATLAB tool, I combined these 4 cytokinin dehydrogenase activity proteins and 96 random rice proteins to create an example set of genes for subsequent enrichment analysis of GO terms. Because 4 of the rice proteins in this list were obtained from the GO consortium (through AmiGO) and the remaining 96 proteins were obtained from UniProt, it is true that these 100 proteins (gene_association.cytokinDehydrogenaseList) were annotated with at least one GO term. Nevertheless, only 23 of these 100 proteins, including 4 cytokinin dehydrogenase activity proteins, belong to UniProt rice-specific proteome. (Only 23 of the 100 accession numbers in the original gene list belong to uniprot_organism_oryzaANDsativa.fasta.) This set of 23 proteins (cytokininDehydrogenase.faa for 23 accession numbers, their amino acid sequences, and their annotations using GO terms) serves as an example input to the enrichment analysis (a list of genes of interest for enrichment analysis) of the MATLAB tool written for this project.

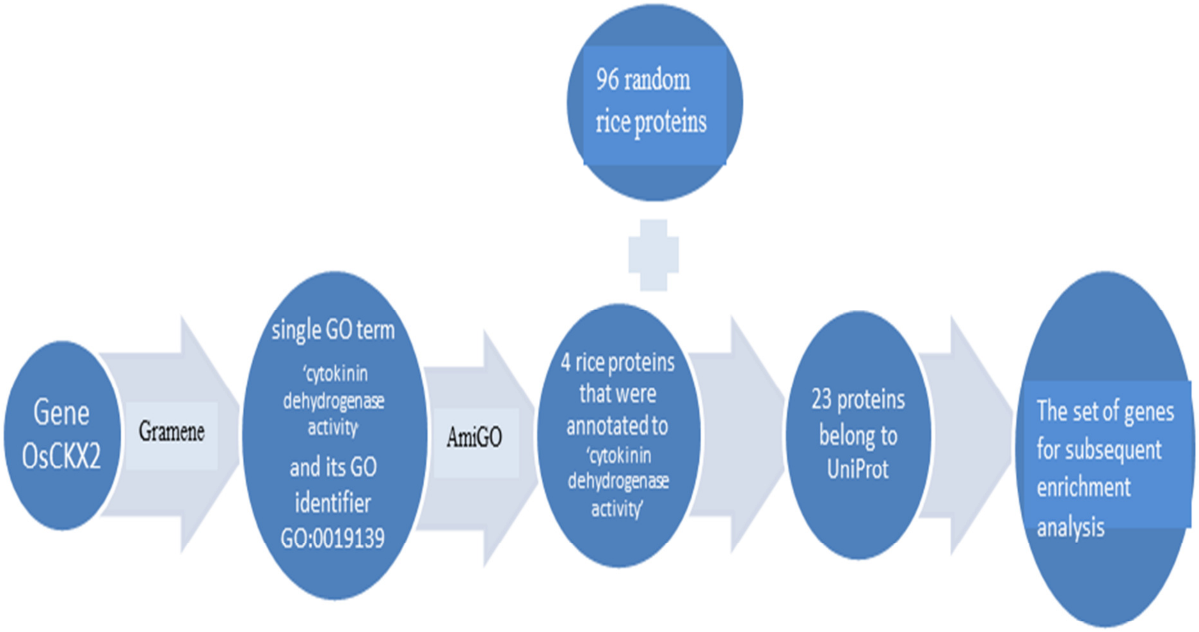

96 random rice proteins
Gene OsCKX2
Gramene
single GO term 'cytokinin dehydrogenase activity' and its GO identifier GO:0019139
AmiGO
4 rice proteins that were annotated to 'cytokinin dehydrogenase activity'
23 proteins belong to UniProt
The set of genes for subsequent enrichment analysis

Figure 6 production of a list of genes for GO term enrichment analysis

Step 4: Enrichment analysis of GO terms

A high-throughput assay usually outputs a list of genes that are biologically related. The comparison between the relative frequency of a GO term in the gene list versus the relative frequency in the reference genome allows us to decide what the biological meaning of the gene list is. The enrichment analysis of GO terms assesses whether each GO term is significant based on its p value [26] that is computed for a gene list of interest relative to the entire genome. In a way the enrichment analysis allows us to functionally categorize any gene list based on enriched

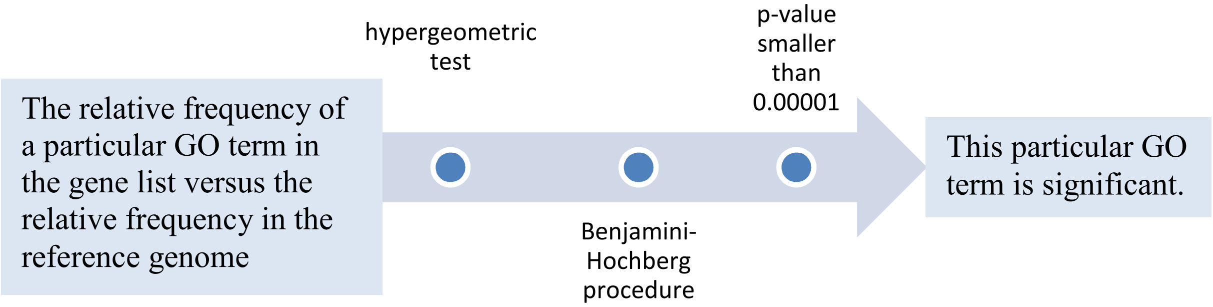
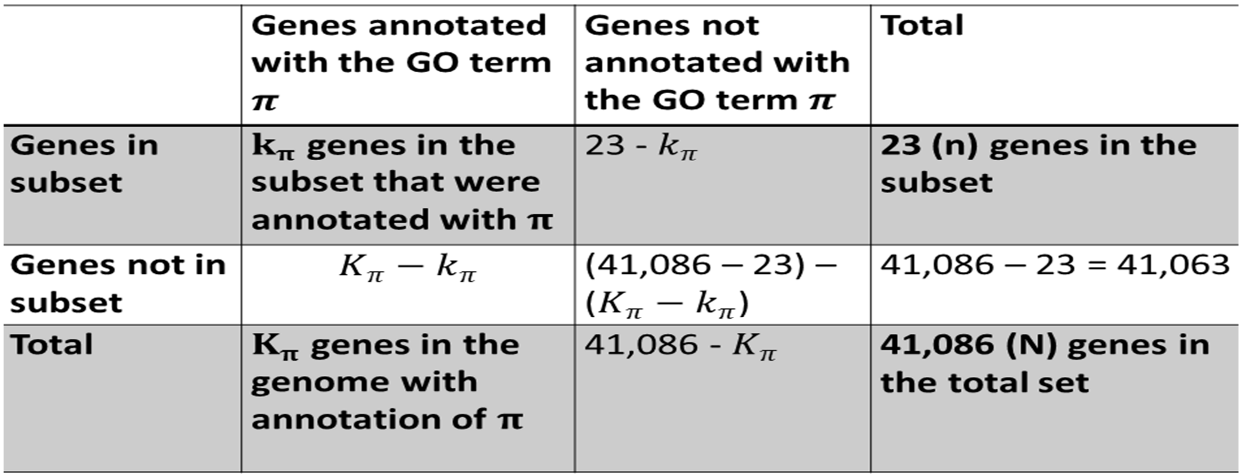

The relative frequency of a particular GO term in the gene list versus the relative frequency in the reference genome
hypergeometric test
Benjamini-Hochberg procedure
p-value smaller than 0.00001
This particular GO term is significant.

Figure 7 p value of a particular GO term is computed by hypergeometric test and then corrected using Benjamini-Hochberg procedure

GO terms. Thus, no doubt the enrichment analysis of GO terms is the most commonly used type of GO analysis. In this section III, I first describe the hypergeometric test and multiple testing correction that are used to compute p value for statistical significance testing of each GO term in this project. Finally, I also explain how I did a proof of principle experiment/ example of using the MATLAB tool in this project to perform enrichment analysis of GO terms.



	Genes annotated with the GO term π	**Genes not annotated with the GO term π**	**Total**
Genes in subset	**k_π genes in the subset that were annotated with π**	23 - k_π	**23 (n) genes in the subset**
Genes not in subset	$K_\pi - k_\pi$	(41,086 – 23) – $(K_\pi - k_\pi)$	41,086 – 23 = 41,063
Total	**K_π genes in the genome with annotation of π**	41,086 - K_π	**41,086 (N) genes in the total set**

Hypergeometric P-value = $$\sum_{i=k_\pi}^{\min(n,K_\pi)} \frac{\binom{K_\pi}{i}\binom{N-K_\pi}{n-i}}{\binom{N}{n}}$$

Figure 8 a 2 by 2 contingency table and hypergeometric test for each GO term π: N – the number of genes in the genome, n - the number of genes in the subset of genes, K_π – the number of genes in the genome with annotation of the GO term π, k_π – the number of genes in the subset of genes that were annotated with this specific GO term π (http://great.stanford.edu/help2/index.php/Statistics#What_is_the_hypergeometric_test_formally.3F)

In the enrichment analysis of GO terms in this manuscript, the exact p value is calculated using hypergeometric tests for k_π number of genes that are annotated with a specific GO term π as shown in Figure 8. In the context of the enrichment analysis, a p-value is the probability of finding k_π or more genes that are annotated with this GO term by chance [3]. The gene list is a subset of the entire genome. Hence, the exact p-value, rather than the approximate p-value, is usually calculated for statistical significance testing of a GO term using a permutation-based test such as a hypergeometric test or a binomial test. In this project, the tool for enrichment analysis of GO terms specifically uses hypergeometric tests to compare GO terms between a gene list and its entire genome by calculating exact p value for a specific GO term because a large number of GO terms are associated with at least one gene and most GO terms are each associated with only a small number of genes[26]. Note that the genes in the entire genome in this project are amino acid sequences from the Rice Genome Annotation Project.

In an enrichment analysis of GO terms, the hypothesis tests are performed for all GO terms and GO terms are not independent. For these two reasons, the need for multiple testing corrections for the p-values becomes even more critical in the enrichment analysis of GO terms. First, the hypothesis tests for 35,119 GO terms (OBO.sgd; October 2011) in each of the two data sets (the total set and the subset) will very likely run into the multiple test problem. If the p-value computed for a GO term is less than a cutoff, which is 0.00001 in this project, this particular GO term in the gene list is then said to be over-represented/ enriched, and so describe the function of the gene list. It is true that the p-value cutoff of 0.00001 in this project is very small. Nevertheless, given that the same hypothesis test (hypergeometric test) was performed 35,119 times to determine enriched GO terms by p-value cutoff of 0.00001, one had already expected some GO terms to be deemed enriched by chance alone. Second, although the Molecular Function, Biological Process and Cellular Compartment attributes of genes are independent, the data structure underlying each of these three ontologies is a hierarchical tree, which is represented by a directed acyclic graph and can represent knowledge to varying levels of completeness[8]. Thus doing hypothesis tests for every GO term violates the independent and identically distributed assumption (iid assumption) of the studied GO terms.[3]

IV. Example: Computing P-value For a Statistically Significant GO Term

The demonstration case described below uses a synthetic gene list. I will explain how I assessed whether a GO term is significantly enriched to identify the function of genes in the subset of 23 genes. In proof of principle experiment, the example set of 23 genes was not known by the program to contain four cytokinin dehydrogenase activity genes. To calculate the enrichment of a GO term, I took two sets of genes, the total set and the subset [27], and then performed a test for statistical significance of the GO term. A total set contains annotations of the reference rice genome was obtained from transferring GO term annotations from UniProt protein sequences to the amino acid sequences of the reference rice genome in the Rice Genome Annotation Project. I previously explained in Step 1: Gene Data Sets and GO Annotations of the section II how the pairwise similarities between the rice-specific proteome from UniProt and amino acid sequences of the reference rice genome from the Rice Genome Annotation Project were detected by BLAST alignments, and then used to infer homology and thus to transfer GO term annotations from UniProt protein sequences to the reference rice genome from the Rice Genome Annotation Project. A subset contains annotations of a subset of genes. I obtained an example subset from query by ‘cytokinin dehydrogenase activity’ (GO:0019139) for protein sequences. Then I used a hypergeometric test (which is equivalent to a one-tailed Fisher’s exact test [3]) to compare the relative frequency of each GO term in the subset of genes of interest versus the frequency in the overall gene set [19] by computing p-value. After the hypergeometric test, a multiple testing correction called Benjamini-Hochberg procedure was also applied to the p-value. If a GO term had p-value less than 0.00001, this GO term was deemed over-represented/ enriched in this subset of rice genes with respect to the total set of rice genes.

V. Implementation

To determine the functions associated to the reference genome from the Rice Genome Annotation Project, I first make use of the protein sequences and their annotating GO terms from the UniProt database to annotate the amino acid sequences which the putative genes in the reference rice genome encode. Next, to demonstrate the feasibility of the tool written to do GO term enrichment for this project, I perform the enrichment analysis of GO terms for a set of cytokinin dehydrogenase activity genes. All the codes were written in MATLAB. In addition to generic statistical tests (hypergeometric test and Benjamini-Hochberg procedure) used to calculate p values, MATLAB also has Bioinformatics toolbox 2.2 that has codes to read proteomic file formats (FASTA), do the gene ontology analysis (GO terms and structure), and visualize a tree display of GO terms (biograph) of gene lists.

As previously discussed in Step 1: Gene Data Sets and GO Annotations of the section II, I wrote MATLAB code to annotate 41,086 putative genes in the rice genome from the Rice Genome Annotation Project to Molecular Function, Biological Process and Cellular Compartment ontologies with GO terms. After downloading all protein sequences of rice from UniProt, their GO term annotations from the GO consortium web page, and rice locus identifiers (LOC_OsXXgXXXX) along with the gene models at those loci (translated amino acid sequences) from the Rice Genome Annotation Project, I performed an all-against-all pairwise alignments for sequence similarities between the amino acid sequences of the reference rice genome (the Rice Genome Annotation Project) and rice-specific protein sequences (UniProt)[7]. This all-against-all pairwise alignment was performed using the command BLASTP in the ncbi BLAST version 2.2.25 package (ncbi-blast-2.2.25) to align UniProt protein sequences to the amino acid sequences of the reference rice genome from the Rice Genome Annotation Project.

To consider only high-scored alignments, I set the E-value to 10^{-10} , a small negative value as a high restrictive threshold. Default values were adopted for the alignment parameters (gap opening = -11, gap extension = -1, substitution matrix = BLOSUM62). [3] Along with alignments between the two rice databases (One contains UniProt proteome and the other contains amino acid sequences of the Rice Genome Annotation Project genome.) which allow for calculating sequence similarity, I annotated putative functions of 41,086 rice genes in the rice genome in the Rice Genome Annotation Project with GO terms. Using the most similar annotated sequences (the top similarity and within 98% of it), I transferred GO annotations from the UniProt rice protein sequences to these amino acid sequences of the reference rice genome in the Rice Genome Annotation Project. At this point, each locus identifier of a rice gene was annotated with GO terms for a total of 41,086 locus identifiers.

Performance Requirement

One difficulty that arises from local sequence alignments is the speed. As fast as personal computers are and as efficient as the local sequence alignments by heuristics using BLAST are, a search of the sequence repository (UniProt rice-specific proteome) in this project on my computer is still too slow. It is clear that all the BLAST sequence alignments need to be carried out in parallel using multiple processors [3]. In this project, I was able to search 67,393 rice amino acid query sequences against the protein sequence database of 49,966 rice-specific protein sequences using the computing power of the Hoffman2 cluster. The Hoffman2 cluster is a grid resource in the UCLA grid and is accessed via the UCLA grid portal, which is operated by the Institute of Digital Research and Education (IDRE) research technology group. The grid infrastructure focuses on resource sharing across institutional borders and among dynamic collections of institutions to enable the collaboration of distributed computing. In the case of the UCLA grid, the Hoffman2 cluster works with 84 UCLA departments, centers and institutes. To meet the high need for computing power and storage capacity, the Hoffman2 cluster provides more than 9,000 processors in three data centers. The grid technologies provide direct access to computers, software, and data. Data center centers, computer nodes and data storage are connected by Ethernet and InfiniBand networks. Globus online (https://www.globusonline.org/), which is a grid middle-ware, enables secure data transfer between my personal computer and the Hoffman2 cluster.

GO terms	GO identifiers	Ontology	Number of genes in the subset	Number of genes in the total set	P value	P value after multiple testing correction
'low-affinity zinc ion transmembrane transporter activity'	6	Molecular Function	23	5982	5.455e-020	3.1929e-016
'riboflavin synthase activity'	3674	Molecular Function	21	2838	8.6515e-023	6.0766e-019
'ribonucleoside-diphosphate reductase activity'	3676	Molecular Function	12	719	8.4156e-016	3.6943e-012
'plasma membrane acetate transport'	5488	Molecular Function	16	1486	1.5063e-018	7.557e-015
'positive regulation of natural killer cell	16491	Molecular Function	4	27	1.2952e-009	5.0541e-006

Table 1 enriched GO terms based on p-values from the analysis for cytokinin dehydrogenase genes

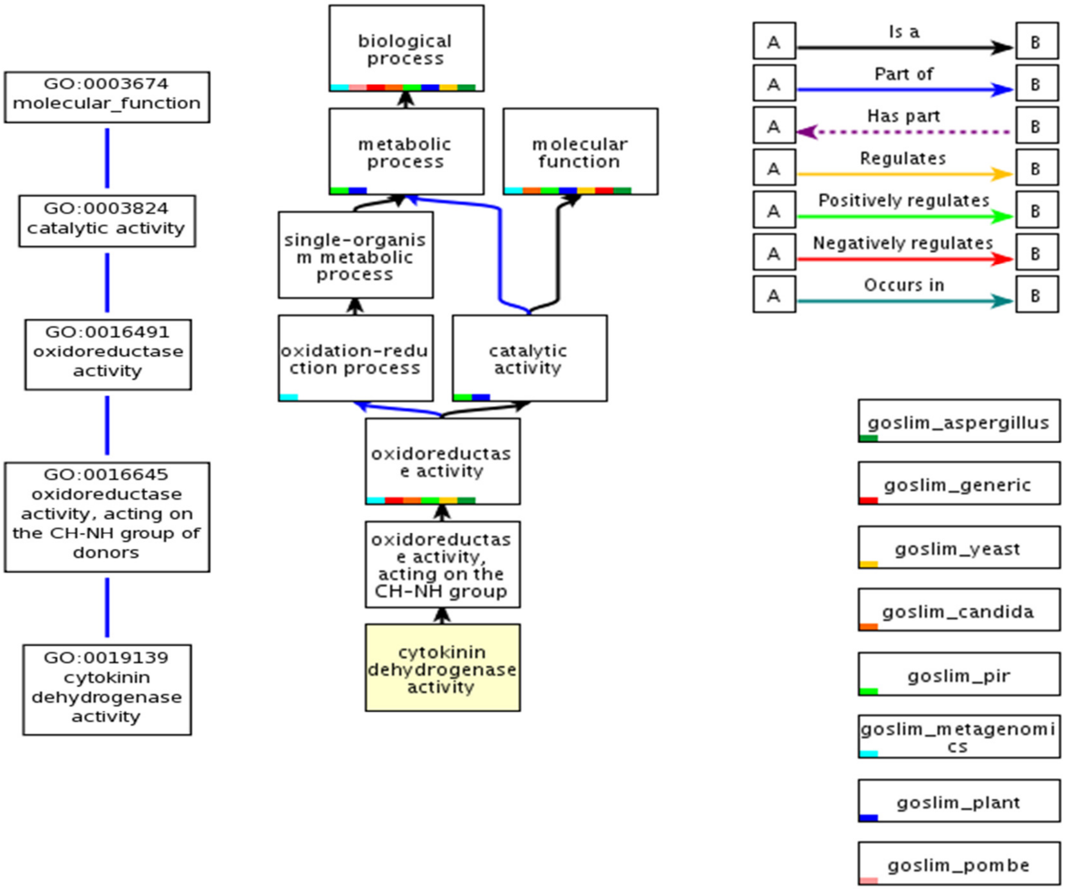

GO:0003674 molecular_function
GO:0003824 catalytic activity
GO:0016491 oxidoreductase activity
GO:0016645 oxidoreductase activity, acting on the CH-NH group of donors
GO:0019139 cytokinin dehydrogenase activity
biological process
metabolic process
molecular function
single-organism metabolic process
oxidation-reduction process
catalytic activity
oxidoreductase activity
oxidoreductase activity, acting on the CH-NH group
cytokinin dehydrogenase activity
A Is a B
A Part of B
A Has part B
A Regulates B
A Positively regulates B
A Negatively regulates B
A Occurs in B
goslim_aspergillus
goslim_generic
goslim_yeast
goslim_candida
goslim_pir
goslim_metagenomics
goslim_plant
goslim_pombe

Figure 9 the gene ontology of 'cytokinin dehydrogenase activity' (GO:0019139) constructed by me on the GO consortium website: The GO term 'cytokinin dehydrogenase activity' is a leaf (that has no children) in the Molecular Function ontology. 'Cytokinin dehydrogenase activity' had been experimentally determined for the four rice proteins.

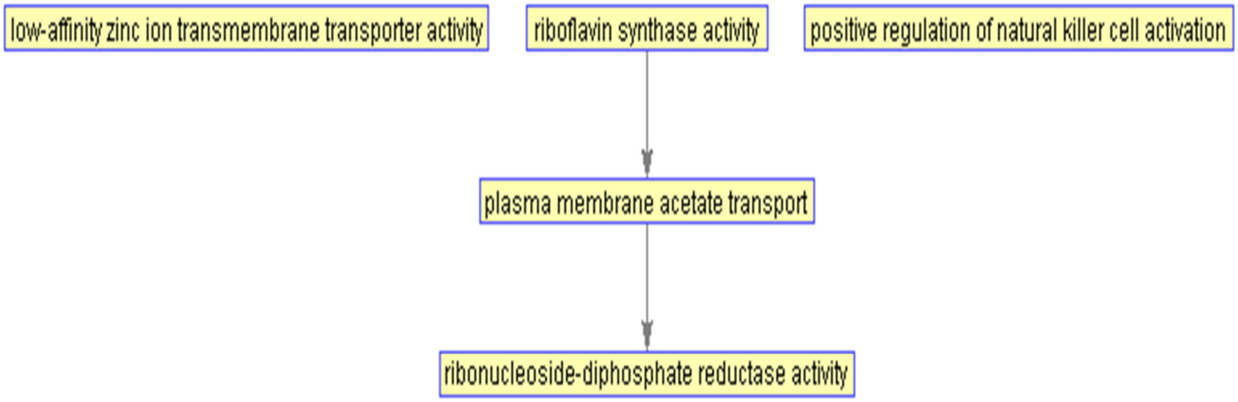

low-affinity zinc ion transmembrane transporter activity
riboflavin synthase activity
positive regulation of natural killer cell activation
plasma membrane acetate transport
ribonucleoside-diphosphate reductase activity

Figure 10 An example output generated by my MATLAB program: to visualize the Molecular Function ontology of enriched terms and the relationships

Discussions of the Enriched GO Terms

The regulation and metabolism of the two types of cytokinin in plants, i.e. trans-zeatin -type cytokinin and isopentenyladenine-type (iP-type) cytokinin, was previously defined by experimental work [2, 28-33]. It is well known that zeatin, which is adenine-type cytokinin, is the most common form of natural cytokinin. Trans-zeatin cytokinin has trans-zeatin-riboside as its side chain while iP-type cytokinin has iP-riboside as its side chain. The enzyme OsCKX2, which is encoded by gene Gn1a, is an enzyme that catalyzes the degradation of cytokinin through cytokinin oxidase/ dehydrogenase (CKX) activity[23] to regulate cytokinin level in plants. Taking advantage of GO term annotations of genes in the example list of 23 genes including four genes related to cytokinin dehydrogenase activity, I applied my tool to perform the enrichment analysis of GO terms.

The p-values of enriched GO terms allow one to predict the biology of genes in the cytokinin dehydrogenase activity gene list for exploratory purpose at the molecular level. Overall I found five enriched GO terms i.e. 'low-affinity zinc ion transmembrane transporter activity', 'riboflavin synthase activity', 'ribonucleoside-diphosphate reductase activity', ‘plasma membrane acetate transport’ and 'positive regulation of natural killer cell activation‘. Table 1 demonstrates a typical result from the enrichment analysis on the cytokinin dehydrogenase activity gene list. All five enriched GO terms were in the Molecular Function ontology. Figure 10 created by my MATLAB tool enables us to observe a Molecular Function ontology graph of the five enriched GO terms in Table 1, which had p-values greater than zero and less than the 0.00001 cutoff.

In particular, applying this MATLAB tool to perform enrichment analysis on the list of 23 genes, out of which 4 were cytokinin dehydrogenase activity proteins, I was able to find two enriched

GO terms that gave clues about the regulation of cytokinin, namely 'ribonucleoside-diphosphate reductase activity' and 'plasma membrane acetate transport'. Nevertheless, the mechanisms of one of the five enriched GO terms, i.e. 'positive regulation of natural killer cell activation’ with respect to the biology of cytokinin is unclear to me while 'low-affinity zinc ion transmembrane transporter activity’ and 'riboflavin synthase activity’ might describe functions and associated processes for many transcription factors and enzymes respectively.

Take 'ribonucleoside-diphosphate reductase activity’ as an example. It is in agreement with the fact that cytokinin biosynthesis happens through the biochemical modification of adenine, which is a riboside. Most important, ribonucleoside is another form of riboside (trans-zeatin-riboside and iP-riboside), and nucleotide of cytokinin can be formed from phosphorylation of nucleoside by adenosine kinase[33]. In addition, the N6-substituted isoprene chain of trans-zeatin type cytokinin or their ribonucleoside was cleaved by cytokinin oxidase, which is also expressed at the base of the flower of mature plants like *Dendrobium* sp [33] and has similar expression to isopentenyladenine transferase gene at the base of siliques. [33, 34]

To illustrate how the enriched GO terms proved of definite help in designing experiments, I will also elucidate how the enrichment of GO term ‘plasma membrane acetate transport‘ is consistent with the fact that endogenous cytokinin was extracted from tissue samples using a 2% acetic acid solution. Going to the original literature conducting biochemical experiments, endogenous cytokinins were extracted from maize using cold methanol:water:acetic acid (70:30:3, v/v) [33]. Furthermore, trans-zeatin-riboside was purified and subsequently separated with linear gradients of a 2% acetic acid solution. Finally, the solution was dried and dissolved in 2% acetic acid[28].

The solubility of cytokinins in acetic acid may account for the enrichment of GO term ‘plasma membrane acetate transport’.

Last but not least, cytokinin can be connected to transport based on two of the enriched terms: ‘low-affinity zinc ion transmembrane transporter activity’ and ‘plasma membrane acetate transport.’ These two concepts annotating multiple genes in this cytokinin related gene set make biological sense. Transport of cytokinin in is known to cause its accumulation which could be explained by the fact that cytokinin regulation involves transportation of this hormone into tissues. Cytokinin is generally found in higher concentrations in meristematic (growing) tissues.

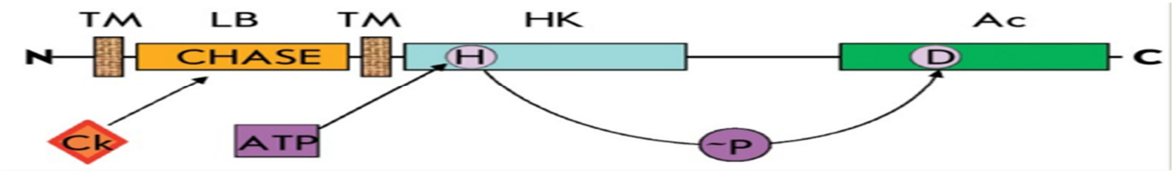


Figure 11 model of cytokinin receptor (exemplified by CRE1/AHK4 from Arabidopsis) TM – transmembrane, HK – histidine kinase, Ck - cytokinin[2] [2]

It is believed that trans-zeatin type cytokinin is transported from root to the shoot [28] and thus accumulation of zeatin-riboside in roots could be attributed partly to the decreased rate of transport of trans-zeatin type cytokinin to the shoots. The decreased iP-type cytokinins in roots may account for the reduced transport of cytokinin in/from source leaves. [28] Furthermore, the finding that the putative signal peptide on the cytokinin oxidase targets it outside of the plasma membrane[33] reiterates the transmembrane property of cytokinin as shown in Figure 11.

Conclusions

I have presented a MATLAB tool that can be used to make sense of the results of high-throughput scanning approaches to increase both the resolution of functional annotations and the

[2] Reprinted with permission from "Receptor properties and features of cytokinin signaling,” by S. N. Lomin, *et al.*, Jul 2012. *Acta Naturae,* vol. 4, pp. 31-45, Copyright 2012 Park-media Ltd. http://www.actanaturae.ru/

inclusion of other yet unknown rice proteins. In this study, the enrichment analysis performed by my tool reveals two GO terms that suggest the relevance for cytokinin associated terms as cytokinins was used to generate the gene list. As already observed in similar enrichment analyses, it is true that some of the GO terms with over-represented numbers of genes from a list of putative genes of interest might be particularly useful for predicting the structure and function of that gene list before confirming these predictions through biochemical experiments. The example gene list (idCytokininDehydrogenase.txt, cytokininDehydrogenase.faa) is available to reproduce the results. My MATLAB tool is particularly valuable to perform functional genomics analysis on rice omics data. As shown above, this tool can annotate lists of rice genes derived from high-throughput profiling approaches, i.e. genes in flowering plants like rice with CG methylations, by performing functional analyses [35].

This tool uses functional inference from sequences to study gene lists. Example future directions include domain searches, excluding annotations that are not experimentally verified, or correcting for multiple testing when calculating the probability of observing a particular enrichment by chance.[36]

Availability and Requirement

MATLAB files
cell2csv.m
cell2float.m
countterm.m
fexact.m
fileread.m
findtag.m
getenrichedterm.m
getgenes.m
gethighscoringpair.m
GOannotate.m
GOannotate2.m
goid2num.m
GOidfrequency.m
main.m
makequeryseqfile.m
opttf.m
printsortblast.m
putgoid.m
quicksort.m
readenrichedterm.m
readtermfrequency.m
savetermcount.m
sortbycol.m
strtok.m
textscantool.m
updateAncester.m
updateGOidfrequency.m
viewsubontology.m
writedatabasefile.m
writequeryseqfile.m

Bibliography

[1] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin*, et al.*, "The UniProt-GO Annotation database in 2011," *Nucleic Acids Res,* vol. 40, pp. D565-70, Jan 2012.

[2] S. N. Lomin, D. M. Krivosheev, M. Y. Steklov, D. I. Osolodkin, and G. A. Romanov, "Receptor properties and features of cytokinin signaling," *Acta Naturae,* vol. 4, pp. 31-45, Jul 2012.

[3] L. Bartoli, L. Montanucci, R. Fronza, P. L. Martelli, P. Fariselli, L. Carota*, et al.*, "The bologna annotation resource: a non hierarchical method for the functional and structural annotation of protein sequences relying on a comparative large-scale genome analysis," *J Proteome Res,* vol. 8, pp. 4362-71, Sep 2009.

[4] K. Jung, G. An, and P. Ronald, "Towards a better bowl of rice: assigning function to tens of thousands of rice genes," *Nature Reviews Genetics,* vol. 9, pp. 91-101, FEB 2008 2008.

[5] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinformatics,* vol. 20, pp. 578-80, Mar 2004.

[6] S. W. Doniger, N. Salomonis, K. D. Dahlquist, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome Biol,* vol. 4, p. R7, 2003.

[7] E. Bolten, A. Schliep, S. Schneckener, D. Schomburg, and R. Schrader, "Clustering protein sequences--structure prediction by transitive homology," *Bioinformatics,* vol. 17, pp. 935-41, Oct 2001.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry*, et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet,* vol. 25, pp. 25-9, May 2000.

[9] Q. Yuan, S. Ouyang, A. Wang, W. Zhu, R. Maiti, H. Lin*, et al.*, "The institute for genomic research Osa1 rice genome annotation database," *Plant Physiol,* vol. 138, pp. 18-26, May 2005.

[10] F. Thibaud-Nissen, M. Campbell, J. P. Hamilton, W. Zhu, and C. R. Buell, "EuCAP, a Eukaryotic Community Annotation Package, and its application to the rice genome," *BMC Genomics,* vol. 8, p. 388, 2007.

[11] X. He and J. Wang, "BGI-RIS V2," *Methods Mol Biol,* vol. 406, pp. 275-99, 2007.

[12] W. Zhao, J. Wang, X. He, X. Huang, Y. Jiao, M. Dai*, et al.*, "BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics," *Nucleic Acids Res,* vol. 32, pp. D377-82, Jan 2004.

[13] A. Louis, E. Ollivier, J. C. Aude, and J. L. Risler, "Massive sequence comparisons as a help in annotating genomic sequences," *Genome Res,* vol. 11, pp. 1296-303, Jul 2001.

[14] P. Mutowo-Meullenet, R. P. Huntley, E. C. Dimmer, Y. Alam-Faruque, T. Sawford, M. Jesus Martin*, et al.*, "Use of Gene Ontology Annotation to understand the peroxisome proteome in humans," *Database (Oxford),* vol. 2013, p. bas062, 2013.

[15] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin*, et al.*, "The Sequence Ontology: a tool for the unification of genome annotations," *Genome Biol,* vol. 6, p. R44, 2005.

[16] J. Wang, L. Kong, S. Zhao, H. Zhang, L. Tang, Z. Li*, et al.*, "Rice-Map: a new-generation rice genome browser," *BMC Genomics,* vol. 12, p. 165, 2011.

[17] C. Liang, P. Jaiswal, C. Hebbard, S. Avraham, E. Buckler, T. Casstevens*, et al.*, "Gramene: a growing plant comparative genomics resource," *Nucleic Acids Research,* vol. 36, pp. D947-D953, JAN 2008 2008.

[18] D. Cozzetto, D. W. Buchan, K. Bryson, and D. T. Jones, "Protein function prediction by massive integration of evolutionary analyses and multiple data sources," *BMC Bioinformatics,* vol. 14 Suppl 3, p. S1, 2013.

[19] F. Thibaud-Nissen, S. Ouyang, and C. R. Buell, "Identification and characterization of pseudogenes in the rice gene complement," *BMC Genomics,* vol. 10, p. 317, 2009.

[20] J. Quackenbush, F. Liang, I. Holt, G. Pertea, and J. Upton, "The TIGR gene indices: reconstruction and representation of expressed gene sequences," *Nucleic Acids Res,* vol. 28, pp. 141-5, Jan 2000.

[21] Y. Lee, R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai*, et al.*, "Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA)," *Genome Res,* vol. 12, pp. 493-502, Mar 2002.

[22] A. Pertsemlidis and J. W. Fondon, "Having a BLAST with bioinformatics (and avoiding BLASTphemy)," *Genome Biol,* vol. 2, p. REVIEWS2002, 2001.

[23] M. Ashikari, H. Sakakibara, S. Lin, T. Yamamoto, T. Takashi, A. Nishimura*, et al.*, "Cytokinin oxidase regulates rice grain production," *Science,* vol. 309, pp. 741-5, Jul 2005.

[24] H. Liu, Z. Z. Hu, and C. H. Wu, "DynGO: a tool for visualizing and mining of Gene Ontology and its associations," *BMC Bioinformatics,* vol. 6, p. 201, 2005.

[25] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, S. Lewis*, et al.*, "AmiGO: online access to ontology and annotation data," *Bioinformatics,* vol. 25, pp. 288-9, Jan 2009.

[26] S. Zhong, C. Li, and W. H. Wong, "ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis," *Nucleic Acids Res,* vol. 31, pp. 3483-6, Jul 2003.

[27] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine*, et al.*, "GoMiner: a resource for biological interpretation of genomic and proteomic data," *Genome Biol,* vol. 4, p. R28, 2003.

[28] T. Kudo, N. Makita, M. Kojima, H. Tokunaga, and H. Sakakibara, "Cytokinin Activity of cis-Zeatin and Phenotypic Alterations Induced by Overexpression of Putative cis-Zeatin-O-glucosyltransferase in Rice," *Plant Physiology,* vol. 160, pp. 319-331, 2012.

[29] D. W. Mok and M. C. Mok, "CYTOKININ METABOLISM AND ACTION," *Annu Rev Plant Physiol Plant Mol Biol,* vol. 52, pp. 89-118, Jun 2001.

[30] D. Mok and M. Mok, "Cytokinin metabolism and action," *Annual Review of Plant Physiology and Plant Molecular Biology,* vol. 52, pp. 89-118, 2001 2001.

[31] S. Li, B. Zhao, D. Yuan, M. Duan, Q. Qian, L. Tang*, et al.*, "Rice zinc finger protein DST enhances grain production through controlling Gn1a/OsCKX2 expression," *Proc Natl Acad Sci U S A,* vol. 110, pp. 3167-72, Feb 2013.

[32] T. Schmülling, T. Werner, M. Riefler, E. Krupková, and I. Bartrina y Manns, "Structure and function of cytokinin oxidase/dehydrogenase genes of maize, rice, Arabidopsis and other species," *J Plant Res,* vol. 116, pp. 241-52, Jun 2003.

[33] N. Brugiere, S. Jiao, S. Hantke, C. Zinselmeier, J. A. Roessler, X. Niu*, et al.*, "Cytokinin oxidase gene expression in maize is localized to the vasculature, and is induced by cytokinins, abscisic acid, and abiotic stress," *Plant Physiol,* vol. 132, pp. 1228-40, Jul 2003.

[34] T. Kakimoto, "Identification of plant cytokinin biosynthetic enzymes as dimethylallyl diphosphate:ATP/ADP isopentenyltransferases," *Plant Cell Physiol,* vol. 42, pp. 677-85, Jul 2001.

[35] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res %M doi:10.1093/nar/gkn923,* vol. 37, pp. 1 - 13, 2009.

[36] S. Y. Rhee, V. Wood, K. Dolinski, and S. Draghici, "Use and misuse of the gene ontology annotations," *Nat Rev Genet,* vol. 9, pp. 509 - 515, 2008.