342–344 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg023

The PlantsP and PlantsT Functional Genomics Databases

Jason H. Tchieu[1], Fariba Fana[1], J. Lynn Fink[1], Jeffrey Harper[3], T. Murlidharan Nair[1], R. Hannes Niedner[1], Douglas W. Smith[2], Kenneth Steube[1], Tobey M. Tam[1], Stella Veretnik[1], Degeng Wang[1] and Michael Gribskov[1,2,*]

[1]San Diego Supercomputer Center and [2]Department of Biology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA and [3]Department of Cell Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037, USA

Received September 11, 2002; Accepted September 20, 2002

ABSTRACT

PlantsP and PlantsT allow users to quickly gain a global understanding of plant phosphoproteins and plant membrane transporters, respectively, from evolutionary relationships to biochemical function as well as a deep understanding of the molecular biology of individual genes and their products. As one database with two functionally different web interfaces, PlantsP and PlantsT are curated plant-specific databases that combine sequence-derived information with experimental functional-genomics data. PlantsP focuses on proteins involved in the phosphorylation process (i.e., kinases and phosphatases), whereas PlantsT focuses on membrane transport proteins. Experimentally, PlantsP provides a resource for information on a collection of T-DNA insertion mutants (knockouts) in each kinase and phosphatase, primarily in *Arabidopsis thaliana*, and PlantsT uniquely combines experimental data regarding mineral composition (derived from inductively coupled plasma atomic emission spectroscopy) of mutant and wild-type strains. Both databases provide extensive information on motifs and domains, detailed information contributed by individual experts in their respective fields, and descriptive information drawn directly from the literature. The databases incorporate a unique user annotation and review feature aimed at acquiring expert annotation directly from the plant biology community. PlantsP is available at http://plantsp.sdsc.edu and PlantsT is available at http://plantst.sdsc.edu.

INTRODUCTION

DNA sequence data from numerous genomic sequencing projects is being rapidly processed by identifying and assigning function to the open reading frames. In spite of this, the biological and biochemical function of a majority of open reading frames in the genomes is still unknown. In order to better understand the genome of the organism, and to turn the genetic 'blueprint' into a functioning organism, attention has shifted from genome mapping and sequencing to the determination of genome function, or functional genomics. Further, scientific advances in sequence analysis, gene expression arrays, yeast two-hybrid experiments, and mass spectrometric identification has made it possible to simultaneously make functional analyses of thousands of proteins.

The databases that we describe in here serve as pioneers in plant functional genomics, that combine knowledge from sequencing efforts and integrate it with sequence analysis, experimental information, and annotation contributed by experts. The PlantsP database, which focuses on protein phosphorylation, and the PlantsT database, which focuses on membrane transport, share a common schema and set of software libraries but present distinct topical views of this information to their users. Both databases aim to provide comprehensive collections of information about specific families of proteins that include all plant species. A key part of the approach is the implementation of a user-contributed annotation and review system. This will eventually allow the scientific community to assume much of the responsibility of curation, leading to increased quality and quantity of functional annotation. Initial versions of the databases necessarily focused on *Arabidopsis*, but a continuing effort is underway to include other plant species, and our specific focus for the next year will be on complete analysis of the rice genome.

DATABASE CONTENT

While the PlantsP and PlantsT databases share a core set of information derived from sequence analyses and literature-based curation, each database comprises some distinct types of experimental information. A summary of the information held in the databases is shown in Table 1.

*To whom correspondence should be addressed. Tel: +1 858 534 8312; Fax: +1 858 822 0873; Email: gribskov@sdsc.edu

Table 1. Database statistics

Database entries	PlantsP	PlantsT
Registered users	281	
Features (sequence motifs and domains)	675 types 11 8587 individual occurrences	
Proteins	3412	2392
RNA	1836	1614
Genomic regions	1872	1297
Families and trees		222/46
Sequence alignments	–	92
Insertional knockouts	3260	–
icp_data	–	518 929 measurements

PlantsP experimental data

The initial experimental focus of the PlantsP database was on the identification of insertional knockouts in tDNA-mutagenized lines produced by the Arabidopsis Functional Genomics Consortium facility (1). Information from insertional knockouts identified in other projects has also been added including information from Syngenta http://www.nadii.com/pages/collaborations/garlic_files/GarlicDescription.html), Cold Spring Harbor Laboratory Arabidposis thaliana insertion database (http://atidb.cshl.org/), Nottingham Arabidopsis Stock Center (http://nasc.nott.ac.uk/insertwatch/insertDatabase.html), and from the Salk Institute Genome Analysis Laboratory (http://signal.salk.edu/). In the future, the experimental focus will be expanded to specifically include rice protein kinases and a broader range of experimental techniques in both *Arabidopsis* and rice.

PlantsT experimental data

A functional approach to identifying mutations involved in ion homeostasis and mineral composition relies on the identification of plants with altered mineral composition. ICP-AES (Inductively Coupled Plasma Atomic Emission Spectroscopy) facilitates rapid screening of the plants; it can measure the concentration of up to 72 minerals in a sample in less than a minute. Each plant thus can be characterized by a 'mineral fingerprint'—a characteristic pattern of concentration of minerals in a plant. Current software supports the download of the experimental data into the database as well as the display of the data (graphically or as text). Although PlantsT focuses primarily on plants, information about yeast membrane transport proteins is also included. Over 2000 samples each consisting of a defined deletion mutant of yeast have been screened. Descriptive information is available for each yeast mutant and aids in understanding the cause of the mutation's effects on organismic mineral composition. In other cases, where little functional data is available about the function of the yeast gene, experimentally determined effects of the mutation on mineral composition can shed some light onto possible functions of the gene.

Classification

Both the PlantsP and PlantsT databases provide systematic classifications of their respective target groups of proteins. The Transporter Classification (TC) System (2), developed by Milton Saier (http://tcdb.ucsd.edu), serves as the starting point for the identification and organization of membrane transporters in *Arabidopsis thaliana* and other plants. Systematic analysis of all available plant protein sequences has been performed resulting in a complete classification for kingdom *viridiplantae*. Briefly, sequences were assigned to groups based on a systematic comparison of all plant protein sequences to panels of sequences that are specific for each group in the TC classification using BLAST (3). This analysis unveiled 65 distinct families of plant membrane transport proteins to which ~1100 members *Arabidopsis* proteins belong. This is roughly 4.3% of the total *Arabidopsis* genes. Of the 4589 total *Arabidopsis* genes that encode membrane proteins (4), about 24% encode for transporters. Multiple alignments and trees have been generated using ClustalX (5) for each of these families. Similar analysis using information from the Yeast Transport Protein Database (6) (YTPdb) against *Arabidopsis* transport proteins (Table 1) identified ~350 proteins in 57 families (roughly 5% of the total yeast genes).

Systematic classifications of both protein kinases (personal communication from M. Gribskov) and protein phosphatase catalytic subunits (7) are available in PlantsP. These classifications are based on sequence comparisons and the pattern of domain conservation between the various protein families.

User annotation

Sequence-derived information is powerful, but is of limited use in predicting the function of proteins. The acquisition of expert annotation is a critical issue for most, if not all, databases. In PlantsP/PlantsT, we have implemented a prototype system for acquiring expert annotation directly from the users of the database. There are several considerations in making such a system work. How do you get people to contribute? How do you assure annotation quality? Our prototype system has several key features.

Users of the databases self-select groups of proteins in which they are interested; we call these proteins 'favorites'. When a user contributes annotation of a particular protein, all of the researchers who have selected that protein as a favorite are notified and can participate in reviewing the submitted annotation. After a period of public review, the submitted annotation is accepted or rejected based on the votes of the reviewers. This adds a peer review step to the curatorial process, which is the key element in ensuring quality in scientific publications.

Encouraging the users of the database to submit annotation is a more complex problem. It is, of course, essential that the annotation submission process be simple to use so that its use is not onerous. Our initial experience has been that the most common user-submitted annotations involve revisions of computationally predicted gene models, or with information that derives from the researcher's own work. We have developed an interface to acquire these kinds of information and plan to enhance the system to handle other types of information as we gain experience with the annotation and review process.

ACCESS

PlantsP and PlantsT are available on the web at http://plantsp.sdsc.edu and http://plantst.sdsc.edu, respectively.

Searches based on keywords, sequence motifs, sequence similarity (BLAST) and classification are available. These approaches give simple but flexible and powerful access to the data and will be expanded in the future. Graphical displays of insertion location, protein sequence motifs and domains, mineral fingerprints, trees and alignments are provided, as well as complete text and XML versions of the information.

FUTURE DEVELOPMENT PLANS

- Complete analysis of the rice genome for protein kinases, protein phosphatases, and membrane transporters.
- Develop approaches for comparison and inference of plant phosphorylation cascades based on those found in animal systems.
- Complete linkage of all plant protein kinases, protein phosphatases, and membrane transporters to their *Arabidopsis* orthologs.
- Extend to other kinds of functional genomic information, especially to protein–protein interactions (e.g., yeast two-hybrid and TAP tagging), localization (e.g., GFP fusions), expression profiling.
- Develop approaches to discover and infer the function of novel membrane protein families not yet incorporated in the TC system.
- Develop techniques to search, analyze and compare ICP-AES data.

ACKNOWLEDGEMENTS

This work is supported by the funding from National Science Foundation Plant Genome Program, DBI-9975808 (PlantsP) and DBI-0077378 (PlantsT). This work is also assisted by the facilities of the National Biomedical Computation Resource at SDSC (NIH P41-RR08605).

REFERENCES

1. Krysan,P.J., Young,J.K. and Sussman,M.R. (1999) T-DNA as an Insertional Mutagen in Arabidopsis. *Plant Cell*, **2**, 2283–2290.
2. Saier,M.H. (1999) A functional-phylogenetic system for the classification of transport proteins. *J. Cell Biochem.*, **32–33** (Suppl.), 84–94.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Ward,J.M. (2001) Identification of novel families of membrane proteins from the model plant *Arabidopsis thaliana. Bioinformatics.*, **17**, 560–563.
5. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
6. Van Belle,D. and Andre,B. (2001) A genomic view of yeast membrane transporters. *Curr. Opin. Cell Biol.*, **13**, 389–398
7. Kerk,D., Bulgrien,J., Smith,D.W., Barsam,B., Veretnik,S. and Gribskov,M. (2002) The complement of protein phosphatase catalytic subunits encoded in the genome of arabidopsis. *Plant Phys.*, **129**, 908–925.