Bioinformation by Biomedical Informatics Publishing Group open access
www.bioinformation.net Web Database

A database of annotated tentative orthologs from crop abiotic stress transcripts

Jayashree Balaji[1*], Jonathan H Crouch[2], Prasad VNS Petite[1] and David A Hoisington[3]
[1] Bioinformatics Unit, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India; [2]International Wheat and Maize Improvement Centre (CIMMYT), Apdo. Postal 6-641, 06600 México, D.F., México; [3]Applied Genomics Laboratory, ICRISAT, Patancheru -502324, Andhra Pradesh, India;
Jayashree Balaji*- Email: b.jayashree@CGIAR.ORG;* Corresponding author
received September 28, 2006; accepted October 03, 2006; published online October 07, 2006

Abstract:
A minimal requirement to initiate a comparative genomics study on plant responses to abiotic stresses is a dataset of orthologous sequences. The availability of a large amount of sequence information, including those derived from stress cDNA libraries allow for the identification of stress related genes and orthologs associated with the stress response. Orthologous sequences serve as tools to explore genes and their relationships across species. For this purpose, ESTs from stress cDNA libraries across 16 crop species including 6 important cereal crops and 10 dicots were systematically collated and subjected to bioinformatics analysis such as clustering, grouping of tentative orthologous sets, identification of protein motifs/patterns in the predicted protein sequence, and annotation with stress conditions, tissue/library source and putative function. All data are available to the scientific community at http://intranet.icrisat.org/gt1/tog/homepage.htm. We believe that the availability of annotated plant abiotic stress ortholog sets will be a valuable resource for researchers studying the biology of environmental stresses in plant systems, molecular evolution and genomics.

Keywords: database; orthologs; comparative genomics; abiotic stress transcripts

Background:
Integrated approaches to the study of abiotic stress response in plants are important especially since drought and salinity stress are primary reasons for crop losses worldwide. The study of stress response pathways includes analysis of information from stress related metabolic and physiological changes, comparative genomics, gene expression studies and structural, and functional data of stress proteins. Plants have stress specific adaptive responses as well as responses which protect the plants from more than one environmental stress. Multiple stress perception and signaling pathways exist - some specific; others may cross talk at various steps. **[1, 2]** Identification of genes related to stress is an important aspect in the study of plant response to abiotic stress. A minimal requirement to initiate a comparative genomics study across abiotic stress conditions is a dataset of orthologs. The availability of a large amount of sequence information, especially that derived from cDNA libraries in response to abiotic stress allows for the generation of a putative list of candidate genes using the orthologs approach. Orthologs are genes in different species that have evolved from a common ancestral gene by speciation and generally retain an equivalent or similar function in the course of evolution.

A high degree of sequence conservation across species and the availability of partial gene sequence data led to the development of comprehensive orthologous gene alignment such as the TOGA (tentative orthologous gene alignments from EST datasets) **[3]** and the COG (clusters of orthologous groups of proteins) databases. **[4, 5]** The TOGA database currently contains 25 plant species while fewer plant species are represented in the COG database. We report here the generation and availability of tentative orthologous annotated datasets for 16 economically important crop species that are vulnerable to the abiotic stresses of heat, dehydration, cold and salt; and for which ESTs generated from stress cDNA libraries are available in the public domain. The aim in building the dataset is to provide users with a catalogue of annotated sequences associated with abiotic stress, identify elements common to all conditions from those that differ, identify categories of functions that are affected under stress conditions and provide users with a list of genes that have the highest representation across tentative orthologous sets.

Methodology:
Dataset
Sequences derived from cDNA libraries generated from tissues subject to heat, dehydration, salt and cold stress from sixteen crop species were used to construct the database. The sequences were downloaded from TIGR **[6]**, NCBI **[7]** in 2003 and updated in June 2005.

Bioinformatics analysis
The sequences were assembled into contigs and singletons crop-wise using a parallelized version of cap3 **[8]** on a paracel HPC. To construct tentative ortholog sets, each species-specific dataset consisting of contigs and singletons was Blast searched against every other dataset using Blastn (standalone BLAST version 2.2.6). If a reciprocal best-hit (RBH) relationship between these sequences was revealed, then the reciprocal best hits

formed a tentative ortholog set. An additional constraint was that each set must comprise sequences from at least three crop species. Scripts were written in Visual Basic to search and assemble tentative ortholog sets after the Blast searches. Sequences were searched for microsatellite markers using the tool SSRIT. **[9]** Sequences in each dataset were translated and searched for protein motifs/patterns against the Prosite database of protein families and domains. All datasets were searched against the species specific plant repeats database **[10]** and hits with an e-value < 1e-5 and an alignment of over 30% of length of query sequence were annotated as repeats. Tentative functional descriptions for the remainder of the sequences were retrieved from each of the databases. These annotations were classified under the 28 functional categories described in the MIPS Functional catalogue Funcat. **[11]** Scripts written in Java were used to carry out this classification. Multiple sequence alignments have been built using ClustalW (version 1.83).

Database and GUI

The data is housed in a relational database on the MSSQL server 2000. The database GUI has been developed using Active Server Pages (ASP).

Species	Number of stress libraries	ESTs	Number of clusters (singletons + contigs)	ESTs in orthologous sets	Clusters in orthologous sets
Wheat	28	20130	11037	8394	2806
Maize	19	21439	10194	9292	3032
Rice	10	13784	8128	4890	1939
Barley	8	12414	7315	5976	2403
Sorghum	5	37590	13815	16828	3321
Pearl millet	3	1945	1443	824	464
Rye	2	1351	945	938	594
Arabidopsis	37	18637	10362	3675	984
Common bean	11	412	206	259	97
Tomato	6	901	637	419	275
Soybean	4	18236	10363	5103	1571
Cowpea	3	38	37	14	14
Groundnut	2	860	679	356	266
Potato	2	17	12	7	4
Chickpea	1	358	56	55	19
Medicago	1	8294	5140	2444	976
Total	142	156406	80369	59474	18765

Table 1: Coverage of monocot and dicot stress related sequences

Stress Condition	Number of tentative ortholog sets
Heat + Cold	91
Heat + Dehydration	1171
Heat + Salt	69
Cold + Dehydration	6851
Cold + Salt	348
Dehydration + Salt	3304
Heat + Cold + Dehydration	2105
Heat + Dehydration + Salt	2323
Cold + Dehydration + Salt	10416
Heat + Cold + Salt	371
Heat + Cold + Salt + Dehydration	8390

Table 2: Number of ortholog sets sharing sequences across stress conditions

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Database**

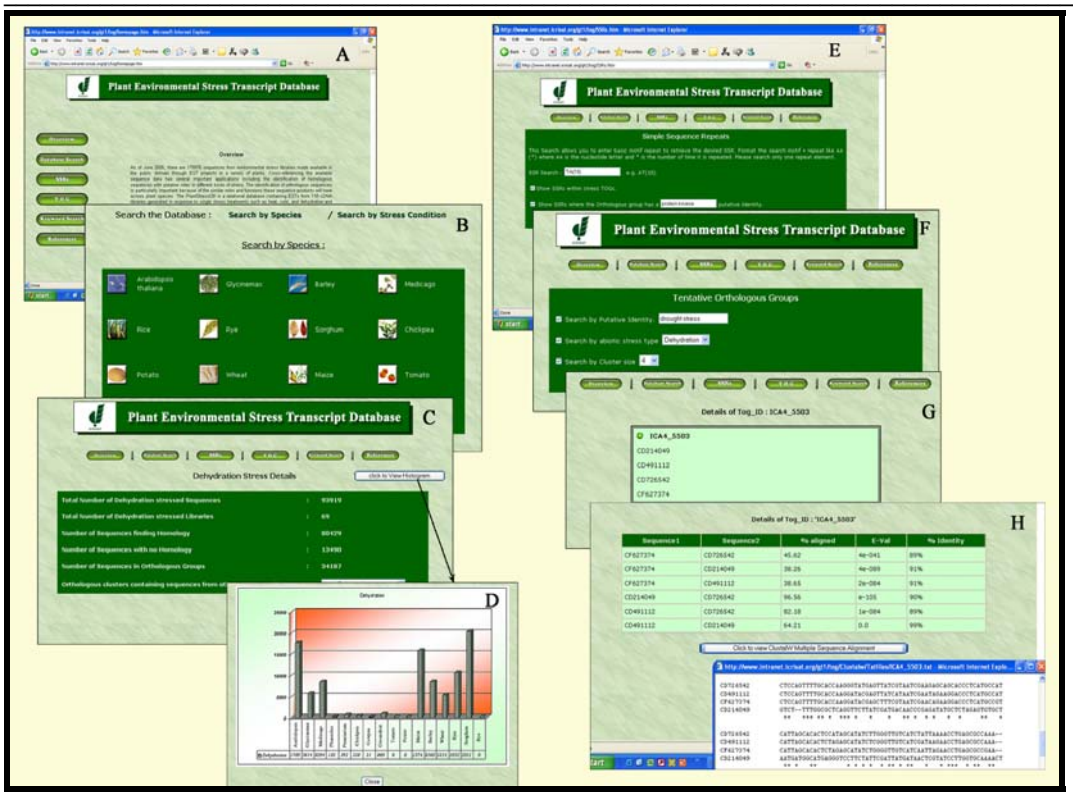

Figure 1: Screen captures of the database GUI. (A) Home page, (B) plant species covered in the current version of the database, (C - H) query pages

Utility:

The database provides a collection of annotated tentative orthologous sequences from sixteen crop species (Table 1) across four abiotic stress conditions (Table 2). The suite of user interfaces (Figure 1) allow the user to browse the database and query for: (a) annotated transcripts that are expressed across stress conditions, (b) transcripts with microsatellites that could be used as conserved functional markers, (c) conserved hypothetical genes that have orthologs in many other species but for which no function has been determined, and (d) ortholog sets with sequence alignment based on annotation, stress conditions or cluster size. The availability of this dataset is a useful resource for researchers studying the biology and genomics of stress response in plants and in the molecular evolution of genes involved in the stress response.

Future development:

We routinely update and expand the database and analyses as additional sequence data becomes available; annotate sequence data with experimental information on candidate genes; and provide users with a reliability score for the ortholog sets constructed along with an analysis of orthologs developed using alternative algorithms.

References:

[01] V. Chinnusamy, *et al*., *J. Exp. Bot.*., 55:225 (2004) [PMID: 14673035]

[02] M. A. Rabbani, *et al*., *Plant Physiol*., 133:1755 (2003) [PMID: 14645724]

[03] Y. Lee, *et al*., *Genome Res*., 12:493 (2002) [PMID: 11875039]

[04] R. L. Tatusov, *et al*., *BMC Bioinformatics*, 4:41 (2003) [PMID: 12969510]

[05] R. L. Tatusov, *et al*., *Nucleic Acids Res*., 28:33 (2001) [PMID: 11125040]

[06] http://www.tigr.org/tdb/tgi/plant.shtml

[07] http://www.ncbi.nlm.nih.gov/

[08] X. Huang, *et al*., *Genome Res.*, 13:2164 (2003)[PMID: 12952883]

[09] S. Temnykh, *et al*., *Genome Res.*, 11:1441 (2001) [PMID: 11483586]

[10] http://www.tigr.org/tdb/e2k1/plant.repeats/

[11] http://mips.gsf.de/projects/funcat

Edited by P. Kangueane

Citation: Balaji *et al*., Bioinformation 1(6): 225-227 (2006)