BIOINFORMATICS APPLICATIONS NOTE
Vol. 21 no. 10 2005, pages 2568–2569
doi:10.1093/bioinformatics/bti334

Databases and ontologies

DATF: a database of *Arabidopsis* transcription factors

Anyuan Guo[1,2,3,†], Kun He[1,2,3,†], Di Liu[1,2,3], Shunong Bai[2,3], Xiaocheng Gu[1,2,3], Liping Wei[1,2,3,*] and Jingchu Luo[1,2,3,*]

[1]Center for Bioinformatics, [2]National Laboratory of Protein Engineering and Plant Genetic Engineering and [3]College of Life Sciences, Peking University, Beijing 100871, Peoples Republic of China

Received on September 2, 2004; revised on December 25, 2004; accepted on February 16, 2005
Advance Access publication February 24, 2005

ABSTRACT

Summary: We have probably developed the most comprehensive database of *Arabidopsis* transcription factors (DATF). The DATF contains known and predicted *Arabidopsis* transcription factors (1827 genes in 56 families) with the unique information of 1177 cloned sequences and many other features including 3D structure templates, EST expression information, transcription factor binding sites and nuclear location signals.

Availability: DATF is freely available at http://datf.cbi.pku.edu.cn

Contact: datf@mail.cbi.pku.edu.cn

INTRODUCTION

Transcription factors are the key regulators of gene expression and play critical roles in the life cycle of higher plants (Gong *et al*., 2004). Identification and classification of transcription factors in *Arabidopsis thaliana* are the first step towards understanding its mechanism of gene expression and regulation. A comprehensive and well-annotated database of *Arabidopsis* transcription factors may provide a useful resource for plant molecular biologists.

The Riechmann group (Riechmann *et al*., 2000; Riechmann, 2002), the Sheen lab (http://genetics.mgh.harvard.edu/sheenweb/AraTRs.html), OHIO-ATTFDB (Davuluri *et al*., 2003), RARTF (http://rarge.gsc.riken.jp/rartf/) and TrSDB (Hermoso *et al*., 2004) have compiled transcription factors in *Arabidopsis* and classified them into families. However, they do not classify clearly or provide enough annotation, or have limited browse or search functionality. TRANSFAC (Matys *et al*., 2003) database contains more information on transcription factors than the above three lists do, but the total number of *Arabidopsis* transcription factors it contains is only ~400. Given the importance of *Arabidopsis* transcription factors, there is a strong need for a database that integrates multiple sources of information to give a comprehensive, genome-wide view of transcription factors in *Arabidopsis*. This was the goal for the database of *Arabidopsis* transcription factors (DATF). In particular, DATF provides the unique information of experimental cloned sequences as well as many other annotations of the *Arabidopsis* transcription factors.

COLLECTION OF *ARABIDOPSIS* TRANSCRIPTION FACTORS

We combined automated search and manual curation to generate a collection of *Arabidopsis* transcription factors as complete as possible. First, we used the gene lists and InterPro domains provided by Riechmann *et al*. (2000) to perform BLAST (Altschul *et al*., 1997), HMMER (Eddy, 1998) and Pfscan (Gattiker *et al*., 2002) searches against the protein sequences in the *Arabidopsis* genome. Second, we manually checked the above results and compared them with OHIO-ATTFDB (Davuluri *et al*., 2003) and the list from the Sheen lab. Conflicting cases were individually resolved according to literature and TAIR annotation (http://www.arabidopsis.org/). Third, some families such as bHLH and MADS were updated or added based on recent publications (Bailey *et al*., 2003; Parenicova *et al*., 2003; Riechmann, 2002). Finally, we identified 1789 transcription factors in *Arabidopsis* and classified them into 49 families.

ANALYSIS AND ANNOTATION OF *ARABIDOPSIS* TRANSCRIPTION FACTORS

We aim to provide comprehensive annotations for the transcription factors in DATF. First, DATF includes the unique information as to whether a transcription factor has been cloned by the 'proteomic investigation of the *Arabidopsis* transcription factors' project (Gong *et al*., 2004). Among the 1789 genes, 1177 genes had been cloned, 31 of which were shown to be different from previously reported cDNA or predicted sequences (Gong *et al*., 2004). This information can be browsed and searched in the 'clone information' field in DATF.

Among the 49 families, 4 (AP2/EREBP, GARP, NAC and SBP) have had 3D structures determined for at least one *Arabidopsis* transcription factor. Another 21 families have had 3D structures determined for at least one transcription factor in other species. Most of the known structures correspond not to the complete transcription factors but to the DNA-binding domains only. We performed BLAST (Altschul *et al*., 1997) search of all sequences in DATF against PDBselect (http://homepages.fh-giessen.de/~hg12640/pdbselect/) as well as sequences of new PDB entries from 2004. About half have BLAST hits with E-value <0.01, identity $>30\%$ and length of the hit segment >50 amino acids. Alignment between an *Arabidopsis* transcription factor and its hit segment is shown and ribbon pictures of the hit segment as well as the whole 3D structure of the PDB entry are displayed.

DATF integrates many other features of *Arabidopsis* transcription factors. These include: 62 binding sites for 20 transcription

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

factor families, collected from the literature and TRANSFAC; nuclear location signals in 348 transcription factors, predicted with PredictNLS (Cokol *et al*., 2000, http://cubic.bioc.columbia.edu/predictNLS/); leucine zipper segments in 115 transcription factors, predicted with LZpred (Bornberg-Bauer *et al*., 1998, http://2zip.molgen.mpg.de/); functional domains, predicted with InterProScan (Zdobnov and Apweiler, 2001); gene duplication information, collected from TIGR (http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml); EST expression information, collected from UniGene (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3702). Each entry in DATF is also hyperlinked with several other resources including TAIR, TIGR, MIPS (http://mips.gsf.de/proj/thal/db/), SIGnAL (http://signal.salk.edu/) and PubMed.

IMPLEMENTATION AND USER INTERFACE

All data and information were stored in a MySQL relational database on a Linux server. Queries to the database were implemented in PHP scripts running in an Apache/PHP environment. Graphics were drawn using the PHP module of the GD graphics library. Three-dimensional structure drawings were created using Molscript (Esnouf, 1997, http://www.avatar.se/molscript/).

DATF is accessible online and allows users to browse by family or chromosome. An introduction to each family is available by clicking on the family name. Users can search DATF by AGI locus ID or by using the advanced search page where they can specify a combination of several search criteria or run BLAST searches against the sequences in DATF. Users are encouraged to submit new data to DATF online. Users can download all the raw data through the DATF website.

DISCUSSION

The goal of DATF is to be comprehensive in both the collection of *Arabidopsis* transcription factors and information about each transcription factor. Three families in DATF (C2H2, LIM and PHD) may have members that contain DNA-binding domains but it is uncertain whether they play a direct role in transcription regulation. Users can apply their own judgment regarding these families.

A survey of the existing databases of *Arabidopsis* transcription factors shows that the total number of transcription factors included in these databases varies from 1400 to 2000. There are at least two reasons for this seemingly large difference. First, the databases may define ‘transcription factors’ differently—some include general transcription factors, such as TBP and chromatin-related proteins, such as SWI/SNF family proteins, whereas others do not. Second, the methods and cutoffs used in predicting transcription factors are often different—some are stricter than others. Depending on how the data will be used, either high sensitivity or high specificity may be more desirable. In DATF, we define transcription factors as proteins that show sequence-specific DNA binding and are capable of activating and/or repressing transcription, excluding general transcription factors or chromatin-related proteins. We combined automated search and manual curation instead of relying on any one single method or cutoff, which we believe improve the quality of the database.

The DATF website has been accessed over 1.2 million times between May 2004 when the web site went online and March, 2005 when the manuscript went into print. We have also received user comments and gene submissions from several countries. We will update DATF regularly with new data, new analysis results and user submissions.

ACKNOWLEDGEMENTS

We thank colleagues of NSFC project (30221120261) headed by Drs Xing Wang Deng and Yuxian Zhu for providing us with the Clone Information, Shuqi Zhao and Dr John Gordon Olyarchuk for helpful assistance and comments, and the anonymous reviewers for the valuable suggestions. This work was supported by the China National Key Basic Research Program (973, 2003CB715900), the China National High-tech Program (863, 2004AA231020) and the Natural Science Foundation of China (NSFC, 30170232, 90408015).

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3419–3402.

Bailey,P.C. *et al*. (2003) Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana. Plant Cell*, **15**, 2497–2502.

Bornberg-Bauer,E. *et al*. (1998) Computational approaches to identify leucine zippers. *Nucleic Acids Res*., **26**, 2740–2746.

Cokol,M. *et al*. (2000) Finding nuclear localization signals. *EMBO Rep*., **1**, 411–415.

Davuluri,R.V. *et al*. (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Esnouf,R.M. (1997) An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model*., **15**, 112–113, 132–134.

Gattiker,A. *et al*. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl. Bioinformatics*, **1**, 107–108.

Gong,W. *et al*. (2004) Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes. *Plant Physiol.*, **135**, 773–782.

Hermoso,A. *et al*. (2004) TrSDB: a proteome database of transcription factors. *Nucleic Acids Res.*, **32**, D171–D173.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Parenicova,L. *et al*. (2003) Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: new openings to the MADS world. *Plant Cell*, **15**, 1538–1551.

Riechmann,J.L. *et al*. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Riechmann,J.L. (2002) Transcriptional regulation: a genomic overview. In Somerville,C.R. and Meyerowitz,E.M. (eds), *The* Arabidopsis *Book*. American Society of Plant Biologists, Rockville, MD, pp. 1–46.

Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.