BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 10 2007, pages 1307–1308
doi:10.1093/bioinformatics/btm113

Databases and ontologies

DPTF: a database of poplar transcription factors

Qi-Hui Zhu[1], An-Yuan Guo[1], Ge Gao[1], Ying-Fu Zhong[1], Meng Xu[2], Minren Huang[2] and Jinchu Luo[1,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering and College of Life Sciences, Peking University, Beijing 100871, P. R. China and [2]The Key Laboratory of Forest Genetics and Gene Engineering, Nanjing Forest University, Nanjing, 210037, China

Received on November 27, 2006; revised on March 12, 2007; accepted on March 15, 2007

Advance Access publication March 28, 2007

Associate Editor: Alex Bateman

ABSTRACT

Summary: The database of poplar transcription factors (DPTF) is a plant transcription factor (TF) database containing 2576 putative poplar TFs distributed in 64 families. These TFs were identified from both computational prediction and manual curation. We have provided extensive annotations including sequence features, functional domains, GO assignment and expression evidence for all TFs. In addition, DPTF contains cross-links to the *Arabidopsis* and rice transcription factor databases making it a unique resource for genome-scale comparative studies of transcriptional regulation in model plants.

Availiability: DPTF is available at http://dptf.cbi.pku.edu.cn

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

1 INTRODUCTION

The genus *Populus* possesses many characteristics that make it ideal for functional genomic studies. *Populus* is a model system for the investigating wood development, crown formation and disease resistance in perennial plants. The availability of the recently sequenced black cottonwood (*Populus trichocarpa*) genome (Tuskan *et al*., 2006), together with two herbaceous plant genomes (*Arabidopsis* and rice), allows investigation of similarities and differences in transcriptional regulators between dicotyledon and monocotyledon (poplar versus rice). These genomes also allow comparative analysis of woody and herbaceous plants (poplar versus *Arabidopsis*).

Transcription factors (TFs) play critical roles in the regulation of gene expression. Genome-wide identification, characterization and annotation of TFs may shed light on understanding the biological function of TFs and exploring the mechanisms of transcriptional regulation. To gain a comprehensive knowledge of TFs in woody plant, we have identified 2576 TFs in *P.trichocarpa* systematically, and constructed a database of poplar TFs (DPTF). DPTF provides extensive annotations of poplar TFs through functional inferences by sequence analysis and expression profiles. In addition, DPTF also contains evolutionary relationships of TF homologs among poplar, *Arabidopsis* and two subspecies of Asian cultivated rice, *indica* and *japonica*.

*To whom correspondence should be addressed.

2 IDENTIFICATION AND ANNOTATION OF PUTATIVE TRANSCRIPTION FACTORS

We obtained 45 555 predicted *P.trichocarpa* proteins from the DOE Joint Genome Institute repository (JGI: http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). In order to provide a complete collection of the poplar TFs and improve the quality of database, we combined automated search and manual curation using the same approach we applied to the construction of the databases of *Arabidopsis* and rice TFs (DATF, Gao *et al*., 2006; DRTF, Guo *et al*., 2005). A value of 0.01 was used as the common E-value cut-off for most TF families in HMMER search. To reduce the false positive and false negative rates, we manually checked and refined the results by inspecting the raw alignments. Some families (e.g. Alfin, HRT, LUG and NZZ) which were either characterized recently or contained only a few members did not have DNA-binding domain profiles. For those families, representative sequences were used as seeds to search for BLAST homologs and the E-value cut-off were manually inspected (for details see the DPTF Help page). As a result of this analysis, a total of 2576 putative TFs were identified and classified into 64 families.

To provide comprehensive information on the putative TFs identified, we made extensive annotations at both family and gene levels. For each TF family, brief introductions and key references were included in the summary pages of the DPTF database website. For each identified TF, DPTF supplied links to various public sequence databases including UniProt (Wu *et al*., 2006), RefSeq (Pruitt *et al*., 2005), EMBL (Cochrane, *et al*., 2006) and TransFac (Matys *et al*., 2006). Furthermore, domain structures were identified and annotated by InterProScan (Quevillon *et al*., 2005), and cross-links to corresponding databases were also provided.

In addition to the annotations based on sequence analysis, we also collected available expression profiling data for all putative TFs. Expression patterns based on EST data were extracted from the NCBI UniGene database. Microarray expression data generated by eQTL analysis of adventitious root development in *Populus* (unpublished data)

were also integrated. The expression information may give valuable insights for further analysis of transcriptional regulations.

A preliminary analysis of the evolutionary relationship between transcription factors in three model plants was conducted. Potential orthologs in *Arabidopsis* and rice (*Oryza sativa* ssp. *indica* and ssp. *japonica*) were detected from best-reciprocal BLAST hits. In addition, neighbor-joining phylogenetic trees were also constructed for each family. These trees can be browsed on the DPTF database website or downloaded as vector graphics format files.

3 IMPLEMENTATION AND USER INTERFACE

DPTF allows users to browse the content taxa by family or chromosome, to query by JGI poplar gene IDs or a combination of keywords. DPTF users can retrieve gene expression information at the genome level. BLAST searches against CDS or protein sequences of poplar, *Arabidopsis* and rice TFs are also available. All the TF CDS and protein sequences can be downloaded through the DPTF website for high-throughput analysis.

4 DISCUSSION

The poplar genome size was estimated to be roughly 485 Mb which is nearly four times of the *Arabidopsis* genome. In spite of this size difference, the predicted number of genes in poplar was estimated at only twice that of *Arabidopsis* (Tuskan *et al*., 2006). Using the same methods and similar cut-off criteria, 2576 putative TFs were identified in poplar, 1922 TFs in *Arabidopsis*, 2025 in *indica* rice and 2384 in *japonica* rice. Some gene families such as MADS have a similar number of members in these two species (111 in poplar and 104 in *Arabidopsis*). A few families such as GeBP contain fewer genes in poplar (7 in poplar and 21 in *Arabidopsis*). Using the best-reciprocal BLAST search approach, 49% potential orthologs in *Arabidopsis*, and 78% in rice were located.

The DPTF database addresses the lack of poplar gene expression data collected in research-accessible public databases. In addition, <40% genes have matched ESTs in UniGene. Expression information from genome-wide *P.deltoides* and *P.euramericana* microarray experiments have been collected, and this data provides expression profiling information for nearly 90% putative TFs. For example, all the 12 BES1 and 13 ARID TFs have been annotated with microarray data, while only three TFs in these two families have EST evidence in public databases. The detailed expression information in DPTF may provide useful information for further functional assessment.

As a large and long-lived forest tree growing in extensive wild populations across continents, poplar has been evolving under selective pressure different from those of annual herbaceous species. Its development involves extensive secondary growth. For example, some TFs in MYB-R2R3 and ZF-HD appear to play an important role in regulating wood formation (Hertzberg *et al*., 2001; Plomion *et al*., 2001). Analysis of data with the same cut-off standard identified 216 MYB and 25 ZF-HD TFs in poplar but 150 MYB and 16 ZF-HD TFs in *Arabidopsis*, and 138 MYB and 15 ZF-HD TFs in rice. These results indicate that some poplar-specific TFs in these two families might play a role in wood formation.

DPTF is the first database of TFs for the perennial woody plant and the first platform to collate the TF databases for poplar, *Arabidopsis* and rice. DPTF will provide comprehensive and valuable resources for research on transcriptional regulation and comparative genomics. DPTF will be maintained and updated regularly as more data and information become available.

ACKNOWLEDGEMENTS

This study was supported by grants: 2003CB715900 (973), 90408015 and 30671705 (NSFC), 20060390012 (MOE), 2006AA02Z334 (863) and the China High-Tech Platform Program.

Conflict of Interest: none declared.

REFERENCES

Cochrane,G. *et al*. (2006) EMBL nucleotide sequence database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.

Gao,G. *et al*. (2006) DRTF: a database of rice transcription factors. *Bioinformatics*, **22**, 1286–1287.

Guo,A. *et al*. (2005) DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics*, **21**, 2568–2569.

Hertzberg,M. *et al*. (2001) A transcriptional roadmap to wood formation. *Proc. Natl Acad. Sci. USA*, **98**, 14732–14737.

Matys,V. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.

Plomion,C. *et al*. (2001) Wood formation in trees. *Plant Physiol*., **127**, 1513–1523.

Pruitt,K.D. *et al*. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.

Quevillon,E. *et al*. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res*., **33**, W116–W120.

Tuskan,G.A. *et al*. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Wu,C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.