BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 10 2006, pages 1286–1287
doi:10.1093/bioinformatics/btl107

Databases and ontologies

DRTF: a database of rice transcription factors

Ge Gao[1], Yingfu Zhong[1], Anyuan Guo[1], Qihui Zhu[1], Wen Tang[1], Weimou Zheng[2], Xiaocheng Gu[1], Liping Wei[1,*] and Jingchu Luo[1,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, People's Republic of China and [2]The Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

Received on March 9, 2006; accepted on March 18, 2006
Advance Access publication March 21, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: DRTF contains 2025 putative transcription factors (TFs) in *Oryza sativa* L. ssp. *indica* and 2384 in ssp. *japonica*, distributed in 63 families, identified by computational prediction and manual curation. It includes detailed annotations of each TF including sequence features, functional domains, Gene Ontology assignment, chromosomal localization, EST and microarray expression information, as well as multiple sequence alignment of the DNA-binding domains for each TF family. The database can be browsed and searched with a user-friendly web interface.

Availability: DRTF is available at http://drtf.cbi.pku.edu.cn

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

1 INTRODUCTION

Transcription factors (TFs) play key roles in regulating gene expression at the transcriptional level, controlling or influencing many biological processes such as development, growth, cell division and responses to environmental stimulus. Identification, characterization and classification of TFs at the genome scale may provide an important resource for researchers on transcriptional regulation. The only available online database of rice TFs is RiceTFDB (http://ricetfdb.bio.uni-potsdam.de/) which contains 2856 protein models coded from 2305 loci in 53 TF families for *japonica*. It has limited annotations including DNA-binding domain and InterPro domain hits for each TF and whole length multiple sequence alignment for each family. Xiong *et al*. (2005) identified 1745 putative TF protein models coded from 1611 loci in *japonica*, and provided the list as Supplementary data. A comprehensive, well-annotated resource of TFs in both *indica* and *japonica* can facilitate comparative analysis of TFs between these two rice subspecies and help to explore the distinct morphological differentiations between *indica* and *japonica*.

Combining automated InterPro scans and BLAST searches with careful manual curation, we have identified TFs in both *indica* and *japonica*, and constructed a database of rice TFs named DRTF containing extensive annotations for the TFs and TF families as well as homologous relationship between corresponding *indica*, *japonica* and *Arabidopsis* TFs. The DRTF web server was set up under the Apache/PHP/MySQL environment on a RedHat Linux platform. It can be browsed by TF families or chromosomes, and searched by keywords or sequences. All sequences are available for downloading.

2 IDENTIFICATION OF PUTATIVE TRANSCRIPTION FACTORS

We first compiled and refined a list of sequence signatures for known plant TF families based on the literature (Shiu *et al*., 2005; Xiong *et al*., 2005; Davuluri *et al*., 2003; Riechmann *et al*., 2000) and existing databases (Guo *et al*., 2005, http://datf.cbi.pku.edu.cn/). Most families can be identified by representative HMM profiles for their DNA-binding domains from Pfam (Bateman *et al*., 2004). For the remaining families without DNA-binding domain profiles, either characterized recently or containing few members, we chose representative sequences from the literature and use them as seeds for BLAST. Finally, we collected 63 distinct TF families.

We downloaded 49 710 predicted *indica* proteins from the Beijing Genome Institute (BGI, http://rise.genomics.org.cn/) and 49 472 predicted *japonica* proteins from TIGR (http://rice.tigr.org/). Based on the list of plant TF signatures, we performed HMMER (Eddy, 1998) and BLAST searches against the whole proteomes of *indica* and *japonica*. We choose 0.01 as the default *E*-value cutoff for most TF families in HMMER searches. We manually inspected all alignments of the domains and refined the results carefully. For BLAST searches, we manually inspected the alignments and set the *E*-value cutoff case by case (for details see the DRTF Help page). Finally, we identified 2025 putative TFs from *indica* and 2384 from *japonica*.

3 ANNOTATION OF PUTATIVE TRANSCRIPTION FACTORS

To provide comprehensive information for the putative TFs, we made extensive annotations using a number of bioinformatics tools and databases. In particular, we employed InterProScan (Quevillon *et al*., 2005) to identify protein domains and assign GO terms to the putative TFs; we performed similarity searches against major databases including UniProt (Wu *et al*., 2006), RefSeq (Pruitt *et al*., 2005), EMBL (Cochrane *et al*., 2006) and TRANSFAC (Matys *et al*., 2006) and hyperlinked to them; we made BLASTP searches against the latest PDBselect database (*E*-value <0.01, identity >30%, and overlap ≥50 residues) to find 3D structural relevance; we obtained EST expression information from

*To whom correspondence should be addressed.

UniGene clusters and microarray expression information from the NCBI GEO database using GEO-BLAST; we aligned the TFs to the RIKEN full-length sequences and provided their accession numbers and CloneIDs; lastly, we identified homologs of each TF in the other rice subspecies and *Arabidopsis*. For each TF family, DRTF includes information extracted from the literature, key references, and multiple sequence alignment of the DNA-binding domains.

4 DISCUSSION

The goal of DRTF is to construct a comprehensive resource of rice TFs. Instead of relying on computational prediction completely, we combined automated search and manual curation. Despite the difference in TF numbers of the two rice subspecies, TFs of one subspecies find homologs in the other reciprocally at a rate >97%.

The different TF numbers between some of the co-responding families in DRTF and RiceTFDB could be caused partly by the different HMM profiles used to define certain families. For example, we took CCCH type zinc finger domain (IPR000571) as the defining signature described as 'DNA-binding' in InterPro and 'nucleic acid binding' as the GO term for the C3H family, whereas RiceTFDB used the C3HC4 ring-finger domain (IPR001841) which has no description of DNA-binding function in InterPro, and the GO terms assigned are 'protein-binding' (GO: 0005515) and 'zinc ion binding' (GO: 0008270). The different choice of HMM profiles has resulted in a 6-fold difference in the number of predicted *japonica* TFs of this family, only 90 in DRTF but 541 in RiceTFDB.

The differences between the dataset of putative *japonica* TFs in DRTF and the dataset composed by Xiong *et al*. (2005) are mostly because of the larger number of TF families we classified (63 versus 37), and the newer version (Release 4) of TIGR database which contains 62 827 predicted proteins versus 59 712 in Release 2 of which 409 TFs we identified for DRTF are missed.

DRTF is the first database of TFs for *indica* and the most annotated one for *japonica*. Currently, there is little annotation available for the *indica* genome in the public sequence repository, and DRTF may bridge the gap at least for the TF families. We will maintain and update DRTF regularly as more data and information become available.

ACKNOWLEDGEMENTS

This study was supported by domestic grants: 2003CB715900 (973), 90408015 (NSFC) and the 863 Programme.

Conflict of Interest: none declared.

REFERENCES

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.

Cochrane,G. *et al*. (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*., **34**, D10–D15.

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*, **23**, 4–25.

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

Guo,A. *et al*. (2005) DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics*, **21**, 2568–2469.

Hwang,I. *et al*. (2002) Two-component signal transduction pathways in *Arabidopsis*. *Plant Physiol*., **129**, 500–515.

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

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.

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

Shiu,S.H. *et al*. (2005) Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol*., **139**, 18–26.

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

Xiong,Y. *et al*. (2005) Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. *Plant Mol. Biol*., **59**, 191–203.