DRTF: a database of rice transcription factors

Ge Gao¹, Yingfu Zhong¹, Anyuan Guo¹, Qihui Zhu¹, Wen Tang¹, Weimou Zheng², Xiaocheng Gu¹, Liping Wei¹,²* and Jingchu Luo¹,²*

¹Center 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 ²The Institute of Theoretical Physics, 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 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 stimuli. Identification, characterization and classification of TFs at the genome scale may provide an important resource for researchers on transcriptional regulation. The DRTF database is available online and can be browsed 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., 2000). 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 families, 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...
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