BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 2 2010, pages 290–291
doi:10.1093/bioinformatics/btp645

Databases and ontologies

LegumeTFDB: an integrative database of *Glycine max*, *Lotus japonicus* and *Medicago truncatula* transcription factors

Keiichi Mochida[1,*], Takuhiro Yoshida[1], Tetsuya Sakurai[1], Kazuko Yamaguchi-Shinozaki[2], Kazuo Shinozaki[1] and Lam-Son Phan Tran[1,*]

[1]Plant Science Center, Gene Discovery Research Group, RIKEN, Yokohama 230-0045 and [2]Japan International Center of Agricultural Sciences, Biological Resources Division, Ibaraki 305-8686, Japan

Received on September 15, 2009; revised on November 9, 2009; accepted on November 12, 2009

Advance Access publication November 17, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: We have established a database named LegumeTFDB to provide access to transcription factor (TF) repertoires of three major legume species: soybean (*Glycine max*), *Lotus japonicus* and *Medicago truncatula*. LegumeTFDB integrates unique information for each TF gene and family, including sequence features, gene promoters, domain alignments, gene ontology (GO) assignment and sequence comparison data derived from comparative analysis with TFs found within legumes, in *Arabidopsis*, rice and poplar as well as with proteins in NCBI nr and UniProt. We also analyzed the promoter regions for all of the TFs to identify all types of *cis*-motifs provided by the PLACE database. Additionally, we supply hyperlinks to make available expression data of 2411 soybean TF genes. LegumeTFDB provides an important user-friendly public resource for comparative genomics and understanding of transcriptional regulation in agriculturally important legumes.

Availability: http://legumetfdb.psc.riken.jp/

Contact: tran@psc.riken.jp; mochida@psc.riken.jp

Supplementary information: Supplementary data available at *Bioinformatics* online.

1 INTRODUCTION

Sequence-specific DNA-binding transcription factors (TFs) are major players that control many of the biological processes such as development, growth, cell division and responses to environmental stimuli (Riechmann *et al.*, 2000; Tran *et al.*, 2007). The specific interactions between TFs and *cis*-regulatory sequences play a central role in the regulation of proteins to affect spatial and temporal gene expression (Yamaguchi-Shinozaki and Shinozaki, 2005). Proper characterization of particular TFs often requires study in the biological context of a whole family because functional redundancy is common within the families (Riechmann *et al.*, 2000). With the availability of a number of complete plant genome sequences and the development of high-throughput experimental techniques complementary information describing TF repertoires in many plants, such as *Arabidopsis*, rice, poplar, maize, sorghum, sugarcane and soybean, have been developed and provided (Guo *et al.*, 2008; Iida *et al.*, 2005; Mochida *et al.*, 2009; Riano-Pachon *et al.*, 2006; Riechmann *et al.*, 2000; Yilmaz *et al.*, 2009; Zhu *et al.*, 2007). Recently, the genome sequences of soybean (http://www.phytozome.net/soybean), *Lotus japonicus* (http://www.kazusa.or.jp/lotus) and *Medicago truncatula* (http://www.medicago.org/genome) have been made available to the public. To gain a comprehensive understanding of TF organization in these three important legume species, we computationally analyzed the genome sequences of *L.japonicus* and *M.truncatula* to identify their TF repertoires, which together with the soybean TF repertoire identified previously (Mochida *et al.*, 2009) are housed at LegumeTFDB. We also compared the TF sequences among the legumes and non-legume plants such as *Arabidopsis*, poplar and rice and with those proteins found in NCBI nr and UniProt. LegumeTFDB provides access to relevant annotations of large TF sets of three important legumes as well as tools for comparative genomic analyses.

2 IDENTIFICATION OF PUTATIVE TRANSCIPTION FACTORS IN *L.JAPONICUS* AND *M.TRUNCATULA*

To identify complete sets of TFs in *L.japonicus* and *M.truncatula*, the same strategy and pipeline used to identify soybean TF repertoire were used (Mochida *et al.*, 2009). We started with retrieving the complete sets of predicted proteins from *L.japonicus* and *M.truncatula*, followed by a HMMER search with all hidden Markov models (HMMs) assembled using a predefined threshold of $E < 1e\text{-}5$. We then refined the results by combined automatic and manual inspections of the raw alignments to exclude false-positive hits and determine the true *E*-value for each family (for details see the LegumeTFDB Help page). Finally, 1626 and 1467 TF genes were identified—and grouped into 61 families based on criteria of TF family classification described previously (Zhu *et al.*, 2007)—in *L.japonicus* and *M.truncatula*, respectively. By assessing GenBank and relevant databases, we classified the TFs of *L.japonicus* and *M.truncatula* into four categories of annotation levels according to our confidence in their structure and functionality as TFs (Mochida *et al.*, 2009). Category A genes have sequence identity $\geq 95\%$ and a blastn *E*-value $\leq$1e-100 with their corresponding sequences found in GenBank and reported to have a functional description as TFs. Category B includes TFs which have high homology (blastp *E*-value $\leq$1e-30) to known TFs of *Arabidopsis* and/or rice. Category

*To whom correspondence should be addressed.

C combines possible TFs which show a significant hit with each of the HMM models used for DNA binding domain prediction (Pfam-HMM *E*-value ≤1e-20). Category D contains TFs which have hits with HMM models with Pfam-HMM *E*-values between 1e-20 and 1e-5 (Supplementary Table 1).

3 ANNOTATIONS OF TFs AND CONSTRUCTION OF LEGUMETFDB

We carried out extensive annotations at both gene and family levels to provide comprehensive information on the identified TFs of *L.japonicus* and *M.truncatula*. These data together with detailed characterization and annotations of previously identified soybean TF repertoire were integrated to create LegumeTFDB. The TF search interface for each species provides seven types of search queries for names of TF families, keywords, sequence identifiers, identifiers of domains supported by InterProScan, GO terms and all available *cis*-motifs documented in the PLACE database (http://www.dna.affrc.go.jp/PLACE/) as well as major abiotic stress responsive *cis*-motifs reported in Yamaguchi-Shinozaki and Shinozaki (2005). The search results listed for each TF family with description of each gene based on similarity search with TFs of other legumes, *Arabidopsis*, poplar and rice as well as with sequences found in NCBI nr and UniProt databases are subsequently displayed. Users will be then navigated to the detailed annotation pages to browse the related annotations, including gene structure, cDNA and protein sequences, domain structure predicted by InterProScan, domain alignments, clusters of homologous proteins within families, promoter regions, predicted *cis*-regulatory motifs in −500, −1000 and −3000 bp promoter regions of each TF gene. The result of *cis*-motif sequence pattern search of promoter regions of each TF gene is displayed together with genomic gene structure. GO annotation inferred by comparative analysis with *Arabidopsis* TAIR8 is also accessible. Additionally, hyperlinks linking directly 2411 soybean TFs, whose IDs are highlighted by red colored letters in the search results of TF search page, to their expression patterns documented in Genevestigator (https://www.genevestigator.com/gv/index.jsp) are provided. These expression data together with information of *cis*-motif analyses, GO annotations and sequence similarities inferred from comparative analyses of the legumes can facilitate the systematic functional predictions of identified TFs. LegumeTFDB also supplies links to either species-specific TF databases such as DATF, RARTF, AtTFDB, DRTF or integrative TF database, such as PlnTFDB and Grassius (Guo *et al*., 2008; Iida *et al*., 2005; Riano-Pachon *et al*., 2007; Yilmaz *et al*., 2009) making it a unique resource for genome-wide comparative studies (for details see the LegumeTFDB Help page).

4 DISCUSSION

Since the assemblies of the legume genomes analyzed in this study still need to be improved, part of the repertoires may be affected by future fine-tuning of the annotation. Additionally, our literature analysis depends on the existing literatures about each gene, which will need to be updated as new findings are reported. With the availability of updated HMM libraries or refinements of existing ones and better fine-tuned annotation and continuous search for the newly reported literatures, we will be able to improve the TF prediction accuracy. We will continue to update the web site with new information when it becomes available.

Systematic combinatorial *in silico* analysis of *cis*-motifs and expression patterns has indicated a positive correlation between multi-stimuli response genes and *cis*-element density in upstream regions (Walther *et al*., 2007). Moreover, a great deal of evidence demonstrates that defined *cis*-elements can effectively aid in the genome-wide screening of ABA and abiotic stress-responsive genes (Zhang *et al*., 2005). Integration of expression analysis, comparative sequence analysis and *cis*-motif and GO annotations provided through this study could link many TF genes to their specific biological functions. Our database will serve as an *in silico* analysis-based basic platform for the elucidation of regulatory mechanisms underlying different developmental and physiological processes. Furthermore, with its unique features LegumeTFDB will enable comparative genomics of TF repertoires both within legume species, among legumes, non-legume plants and other organisms.

ACKNOWLEDGEMENTS

The soybean sequence data were produced by the US Department of Energy Joint Genome Institute in collaboration with the scientific user community. The authors thank Kazusa Research Institute and the Medicago Genome Sequence Consortium for *L.japonicus* and *M.truncatula* sequence data, respectively.

Conflict of Interest: none declared.

REFERENCES

Guo,A.Y. *et al.* (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res*., **36**, D966–D969.

Iida,K. *et al.* (2005) RARTF: database and tools for complete sets of *Arabidopsis* transcription factors. *DNA Res*., **12**, 247–256.

Mochida,K. *et al.* (2009) *In silico* analysis of transcription factor repertoire and prediction of stress responsive transcription factors in soybean. *DNA Res.*, **16**, 353–369.

Riaño-Pachón,D.M. *et al.* (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, **8**, 42.

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

Tran,L.-S.P. *et al.* (2007) Plant gene networks in osmotic stress response: from genes to regulatory networks. *Methods Enzymol.*, **428**, 109–128.

Yamaguchi-Shinozaki,K. and Shinozaki,K. (2005) Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends Plant Sci*., **10**, 88–94.

Yilmaz,A. *et al.* (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol*., **149**, 171–180.

Walther,D. *et al.* (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. *PLoS Genet.*, **3**, e11.

Zhang,W. *et al.* (2005) *Cis*-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana*. *Bioinformatics*, **21**, 3074–3081.

Zhu,Q.H. *et al.* (2007) DPTF: a database of poplar transcription factors. *Bioinformatics*, **23**, 1307–1308.