BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2493–2495
doi:10.1093/bioinformatics/btq432

Databases and ontologies

Advance Access publication July 27, 2010

proTF: a comprehensive data and phylogenomics resource for prokaryotic transcription factors

Jie Bai[1,2,†], Junrong Wang[3,†], Feng Xue[4,†], Jingsong Li[1], Lijing Bu[1], Junming Hu[4], Gang Xu[1], Qiyu Bao[1], Guoping Zhao[2], Xiaoming Ding[2], Jie Yan[4,*] and Jinyu Wu[1,*]

[1]Institute of Genomic Medicine/Zhejiang Provincial Key Laboratory of Medical Genetics, Wenzhou Medical College, Wenzhou 325035, [2]Department of Microbiology and Microbial Engineering, School of Life Sciences, Fudan University, Shanghai 200433, [3]Maternal and Child Health Hospital of Wenling, Wenling 317500 and [4]Department of Medical Microbiology and Parasitology, College of Medicine, Zhejiang University, Hangzhou 310058, China

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Investigation of transcription factors (TFs) is of extreme significance for gleaning more information about the mechanisms underlying the dynamic transcriptional regulatory network. Herein, proTF is constructed to serve as a comprehensive data resource and phylogenomics analysis platform for prokaryotic TFs. It has many prominent characteristics: (i) detailed annotation information, including basic sequence features, domain organization, sequence homolog and sequence composition, was extensively collected, and then visually displayed for each TF entry in all prokaryotic genomes; (ii) workset was employed as the basic frame to provide an efficient way to organize the retrieved data and save intermediate records; and (iii) a number of elaborated tools for phylogenomics analysis were implemented to investigate the evolutionary roles of specific TFs. In conclusion, proTF dedicates to the prokaryotic TFs with integrated multi-function, which will become a valuable resource for prokaryotic transcriptional regulatory network in the post-genomic era.

Availability: http://centre.bioinformatics.zj.cn/proTF

Contact: med_bp@zju.edu.cn; iamwujy@yahoo.com.cn

Received on January 20, 2010; revised on June 22, 2010; accepted on July 22, 2010

1 INTRODUCTION

Identification, classification and phylogenomics analysis of transcription factors (TFs) within or among specific organisms can help us to highlight how they evolutionarily conserved or diverse in order to fit in the ever-changing environment effectively (Rodionov, 2007). In the past decade, a number of specialized TF databases were established. The TRANSFAC and DBD databases were released as a universal data source covering putative TFs for all completely sequenced genomes and TrSDB contains the TFs derived from nine eukaryotic species (Hermoso *et al*., 2004; Wilson *et al*., 2008; Wingender, 2008). Within the plant kingdom, progressively integrated and comprehensively annotated biological databases had been constructed, such as plantTFDB (Guo *et al*., 2008), DPTF (Zhu *et al*., 2007), RARTF (Iida *et al*., 2005), PlanTAPDB (Richardt *et al*., 2007), TOBFAC (Rushton *et al*., 2008), DATFAP (Fredslund, 2008), GRASSIUS (Yilmaz *et al*., 2009) and PlnTFDB (Perez-Rodriguez *et al*., 2009). Various information of TFs in animal are abundantly explored by databases such as TFcat (Fulton *et al*., 2009), TFdb (Kanamori *et al*., 2004), FlyTF (Pfreundt *et al*., 2009) and ITFP (Zheng *et al*., 2008). In addition, fungal TFs can be accessed from FTFD (Park *et al*., 2008). Up to now, however, no comprehensive platform for computational repository is available to provide access to the large complete sets of prokaryotic TFs. RegTransBase is a database of regulatory interactions in prokaryotes, which captures the knowledge in public scientific literature and contains 1131 TFs in the latest version (Kazakov *et al*., 2007). ooTFD is a database aimed at capturing information regarding the polypeptide interactions, which comprise and define the properties of TFs with limited species and TF entries (Ghosh, 2000). ExtraTrain provides integrated and easily manageable information for 679 816 extragenic regions and for the genes delimiting each of them. For TF, it only contains 16 TF families (Pareja *et al*., 2006). In addition, the currently developed AraC–XylS only holds information about TF family AraC in bacteria and the BacTregulators contains three TF families (TetR, AraC and IclR) in bacteria and archaea (Martinez-Bueno *et al*., 2004; Tobes and Ramos, 2002). DBTBS (Sierro *et al*., 2008), RegulonDB (Huerta *et al*., 1998), TRACTOR_DB (Perez *et al*., 2007), cTFbase (Wu *et al*., 2007) and ArchaeaTF (Wu *et al*., 2008) only aim to focus on the TFs in *Bacillus subtilis*, *Escherichia coli*, gamma-proteobacteria, cyanobacteria and archaea, respectively.

Herein, a new TF database proTF is constructed to provide an integrated useful resource for TFs research and facilitate further investigation of transcriptional regulatory network in prokaryotes. In comparison with other comprehensive TF databases, proTF contains the following prominent characteristics: (i) offered an extensively detailed annotation information of each TF entry in all the completely sequenced prokaryotic genomes; (ii) employed the workset as the basic frame to well organize the retrieved data and save intermediate records; and (iii) implemented a number of phylogenomics analysis tools to investigate the evolutionary roles of specific TFs in or across different prokaryotic organisms. In conclusion, proTF is dedicates to the prokaryotic TFs with multiple integrated phylogenomics function.

*To whom correspondence should be addressed.

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

2 IDENTIFICATION AND ANNOTATION OF TFs

To identify complete putative set of TFs in a given prokaryotic genome, our previously well-established analysis pipelines in cTFbase (Wu *et al*., 2007) and ArchaeaTF (Wu *et al*., 2008) were applied to all the fully sequenced proteomes of prokaryotic species available from KEGG. In brief, we start with the collection of a set of well-characterized/putative TFs from Swiss-Prot/TrEMBL databases (release 15.12) and a number of HMM profiles corresponding to DNA-binding domains (DBDs) from Pfam (version 24.0) and SUPERFAMILY (release 1.73). A combination of BLAST-based and HMM-based search was adopted to obtain significant hits using the BLASTP and hmmpfam program with an *E*-value of 1e-10 and 0.01, respectively. All the identified TFs were classified into different families according to categories of their DBDs.

Once a putative TF was identified and classified, it was extensively annotated using a number of bioinformatics tools and databases. Particularly, the molecular weight and isoelectric point of a given TF was identified using the PepStat program implemented in the EMBOSS package (http://www.ebi.ac.uk/emboss). The InterProScan program (http://www.ebi.ac.uk/Tools/InterProScan/) was used to search its domain architectures. The Gene Ontology terms were obtained from the InterProScan results using InterPro2Go (http://www.geneontology.org/external2go/interpro2go). Sequence similarity alignment was performed using the BLAST program against several major databases, including PDB (http://www.pdb.org/), Uniprot (http://www.uniprot.org/), KEGG (http://www.genome.jp/kegg/genes.html), Swissprot (http://www.expasy.ch/sprot/) and Refseq (http://www.ncbi.nlm.nih.gov/refseq/).

3 WEB INTERFACE

proTF is a relational database hosts on an Apache HTTP server running on Linux operating system. Various separate MySQL database tables are retrieved by the Structure Query Language. PHP is implemented for the connection of database and dynamic production of user-friendly HTML front-end queries. The web interface is organized in an operating system-independent way, which has been tested to work properly in Internet Explorer 7.0, Firefox 2/3 and Opera 10.00 browsers.

proTF presents a user-friendly web interface for researchers to store and interrogate all the putative TFs entries. Users can easily access the data by clicking a specific TF family or the species in the browse page. In the search page, a multi-layered query system is employed for users to retrieve the data based on hierarchized keywords. The search can be performed via locus tag, family or species. Expression in separate fields can be combined with the logical operator AND, OR or NOT. The list of registered species is also arranged hierarchically according to taxonomy to allow users to easily access the TF entry. In BLAST-based search page, the BLAST program (http://indra.mullins.microbiol.washington.edu/blast/viroblast.php) is implemented to enable the identification of the homologs of the query sequences stored in proTF. Either the full-length or the DBD region of the TF sequences can be taken as database to perform BLAST search. In addition, a number of advanced parameters (such as *E*-value, matrix and species) were also provided to allow users to perform more specific BLAST searches. In the result table, basic information of each TF entry matching the query will be listed in a table, in which gene IDs and family IDs are linked to the detailed annotation of the gene and family. By clicking on the entries, detailed annotation information will be displayed, including basic sequence features, Gene Ontology terms, gene domain organization and sequence homolog to other relevant databases.

4 WORKSET AND PHYLOGENOMIC ANALYSIS

Workset, incorporated into proTF, is a significant functionality of having the TF genes and families well organized and conducting a succession of phylogenomics analyses to investigate the evolutionary relationship of specific TF family. Using workset, users can append, remove and configure any retrieved results for further data manipulation, comparative genomics and molecular evolution analysis. Each workset is assigned with a specific ID either generated by the server randomly or named by users themselves. It is available to customize all the data in the workset through appending or deleting the items. Moreover, a saved workset can be loaded again by its corresponding ID to avoid rehandling the retrieved results.

Another prominent feature of proTF is that it can also serve as a comparative genomics and molecular evolution analysis platform for prokaryotic TFs. A number of phylogenomics analysis tools are implemented to allow users to investigate a particular TF within one prokaryotic genome or a bunch of TFs across different ones, as well as the TFs items are stored in the workset. Particularly, the ClustalW (http://www.ebi.ac.uk/clustalw/) and MUSCLE program (http://www.drive5.com/muscle/) were employed to performing multiple sequence alignment for the whole TF sequences or just the DBD sequences of TFs at amino acid or DNA level. Multiple sequence alignment result was graphically displayed using the Jalview program (http://www.jalview.org/). The QuickTree program (http://www.sanger.ac.uk/resources/software/quicktree/) helps users to investigate the evolutionary relationship of TF items stored in the workset. The reliability of the phylogenetic tree can be evaluated by the bootstrap method with replications (at default 100) and the tree is visualized using the ATV program (http://www.phylosoft.org/atv/).

5 PERSPECTIVES

Currently, proTF provided a complete list of centralized putative prokaryotes TFs. It has contained a number of 127 838 TFs from 841 prokaryotic organisms. In future, more prokaryotes TFs from sequenced organisms will be added into the platform to extend its functionality. These existing entries will be updated to keep up with the latest annotation information in linked databases. We believe that the platform will provide a wealth of information and more robust and reliable support for the scientific community to decipher and gain the complete picture of the genetic regulatory networks.

Funding: National Natural Science Foundation of China (30800643); National Science and Technology Key Program for Infectious Diseases of China (2008ZX10004-015).

Conflict of Interest: none declared.

REFERENCES

Fredslund,J. (2008) DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. *BMC Genomics*, **9**, 140.

Fulton,D.L. *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.

Ghosh,D. (2000) Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res*., **28**, 308–310.

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

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

Huerta,A.M. *et al.* (1998) RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*., **26**, 55–59.

Kazakov,A.E. *et al.* (2007) RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res*., **35**, D407–D412.

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

Kanamori,M. *et al.* (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.

Martinez-Bueno,M. *et al.* (2004) BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics*, **20**, 2787–2791.

Pareja,E. *et al.* (2006) ExtraTrain: a database of Extragenic regions and Transcriptional information in prokaryotic organisms. *BMC Microbiol*., **6**, 29.

Park,J. *et al.* (2008) FTFD: an informatics pipeline supporting phylogenomic analysis of fungal transcription factors. *Bioinformatics*, **24**, 1024–1025.

Perez,A.G. *et al.* (2007) Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. *Nucleic Acids Res*., **35**, D132–D136.

Perez-Rodriguez,P. *et al.* (2009) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res*., **38**, D822–D827.

Pfreundt,U. *et al.* (2009) FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res.,* **38**, D443–D447.

Richardt,S. *et al.* (2007) PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. *Plant Physiol.*, **143**, 1452–1466.

Rodionov,D.A. (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.*, **107**, 3467–3497.

Rushton,P.J. *et al.* (2008) TOBFAC: the database of tobacco transcription factors, *BMC Bioinformatics*, **9**, 53.

Sierro,N. *et al.* (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.

Tobes,R. and Ramos,J.L. (2002) AraC-XylS database: a family of positive transcriptional regulators in bacteria. *Nucleic Acids Res.*, **30**, 318–321.

Wilson,D. *et al.* (2008) DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.

Wingender,E. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.*, **9**, 326–332.

Wu,J. *et al.* (2008) ArchaeaTF: an integrated database of putative transcription factors in Archaea. *Genomics*, **91**, 102–107.

Wu,J. *et al.* (2007) cTFbase: a database for comparative genomics of transcription factors in cyanobacteria. *BMC Genomics*, **8**, 104.

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

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