BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 20 2008, pages 2416–2417
doi:10.1093/bioinformatics/btn439

Databases and ontologies

ITFP: an integrated platform of mammalian transcription factors

Guangyong Zheng[1,2,3], Kang Tu[3,4], Qing Yang[2], Yun Xiong[2], Chaochun Wei[5,6], Lu Xie[6], Yangyong Zhu[2,6,*] and Yixue Li[3,6,*]

[1]School of Life Sciences, Fudan University, Shanghai 200433, [2]Department of Computing and Information Technology, Fudan University, Shanghai 200433, [3]Bioinformatics Center, Key Lab of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, [4]Graduate School of the Chinese Academy of Sciences, Beijing 100039, [5]College of Life Sciences and Technology, Shanghai Jiaotong University, Shanghai 200240 and [6]Shanghai Center for Bioinformation Technology, Shanghai 200235, China

Received on May 10, 2008; revised on July 10, 2008; accepted on August 17, 2008

Advance Access publication August 19, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Investigation of transcription factors (TFs) and their downstream regulated genes (targets) is a significant issue in post-genome era, which can provide a brand new vision for some vital biological process. However, information of TFs and their targets in mammalian is far from sufficient. Here, we developed an integrated TF platform (ITFP), which included abundant TFs and their targets of mammalian. In current release, ITFP includes 4105 putative TFs and 69 496 potential TF-target pairs for human, 3134 putative TFs and 37 040 potential TF-target pairs for mouse, and 1114 putative TFs and 18 055 potential TF-target pairs for rat. In short, ITFP will serve as an important resource for the research community of transcription and provide strong support for regulatory network study.

Availability: ITFP can be accessed at http://itfp.biosino.org/itfp

Contact: yyzhu@fudan.edu.cn; yxli@sibs.ac.cn

1 INTRODUCTION

Transcription factors (TFs) play significant roles in various biological processes through binding with *cis*-regulatory elements to control expression levels of downstream genes. Research of TFs and their downstream regulated genes (TF targets) offers a key means for insight into mechanism of transcription regulation. Currently, two databases about TFs and their targets in mammalian have been reported. One is the TRANSFAC repository (Matys *et al*., 2003), which contains some TFs and their targets extracted from published papers. However, it is a commercial system with rigid restriction of data acquirement. The other pioneer work is the TRED database (Zhao *et al*., 2005), which provides 36 cancer-related TF families and targets information in mammalian. Therefore, construction of a free and comprehensive platform including a large number of TFs and their targets for mammalian genomes will offer an important resource for transcription research and facilitate further investigation of regulatory network.

By employing support vector machine (SVM) and ARACNE algorithm (Basso *et al*., 2005; Margolin *et al*., 2006), we have identified TFs and their targets in human, mouse and rat genomes and then developed an informatics platform called ITFP (integrated TF platform). The platform was constructed under the Tomcat/Java/MySQL environment on a unix server. In this platform, users can conveniently browse well-annotated TFs as well as their targets, and query them through text keywords, such as protein symbol, gene name, accession number of Swiss-Prot database and identification number of Entrez-Gene database. Moreover, an online TFs prediction tool named TFMiner is presented in the platform, which can detect whether a protein sequence is a TF or not purely based on the protein sequence. All data presented in the platform can be freely downloaded for non-commercial users.

2 IDENTIFICATION AND VALIDATION OF PUTATIVE TRANSCRIPTION FACTORS

We first downloaded protein sequences of mammalian from the Swiss-Prot database v12.5 to build primal datasets. Three datasets containing 17 658, 14 273 and 6493 proteins were constructed for human, mouse and rat, respectively. Subsequently, based on our previous work (Zheng *et al*., 2008) machine learning model built from SVM algorithm was used to identify TFs from these datasets, and then classification model built from ECOC algorithm was employed to categorize resulting TFs into four subtypes. Finally, we identified 4105 putative TFs for human, 3134 putative TFs for mouse and 1114 putative TFs for rat.

In order to assess accuracy of TFs predicted here, we collected human, mouse and rat TFs from TRANSFAC repository v9.4 and compared with our data. As for human, 934 out of 1032 TFs were detected in our data, the recall rate achieved 90.50%. While for mouse, 693 out of 764 TFs were detected in our data, the recall rate achieved 90.71%. For rat, 224 out of 277 TFs were detected in our data, the recall rate achieved 80.87%. Comparison with the TRANSFAC data demonstrated that result of TFs prediction here was valid to a certain extent.

Extensive efforts were put to annotate the putative TFs. In details, gene message encoding TFs was collected from the Entrez-Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene), pathway information about participated TFs was extracted from the KEGG database (http://www.genome.jp/kegg/kegg2.html), domain and functional-site information of a TF was obtained by the InterProScan software (http://www.ebi.ac.uk/interpro), and

*To whom correspondence should be addressed.

the GO term of a TF was acquired through the EGO web server (http://www.ebi.ac.uk/ego).

3 IDENTIFICATION AND VALIDATION OF POTENTIAL TRANSCRIPTION FACTORS' TARGETS

To detect downstream regulated genes of TFs, we adopted a reverse engineering algorithm named ARACNE (Basso *et al*., 2005; Margolin *et al*., 2006), which used gene expression profile data and TF terms as inputs and produced TF and target pairs as the outcome. We downloaded gene expression profile data from the GEO database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds) and refined them to build primal materials for the algorithm. For human, mouse and rat genome, we collected gene expression profile data from Affymetrix human Genome U133 plus 2.0 Array platform, Affymetrix mouse Genome 430 2.0 Array platform and Affymetrix rat Genome 230 2.0 Array platform, respectively. The refined process was carried out through following steps: (1) selected microarray data from the raw dataset and built a subset, where microarray number of diverse tissues was kept in comparative proportion and (2) normalized microarray data in the subset in order to eliminate bias of different samples. Accordingly, three microarray subsets containing 302 (22 tissues), 310 (34 tissues) and 332 (39 tissues) samples were set up for human, mouse and rat, respectively. Then these subsets and TFs predicted in our work were utilized as inputs for ARACNE method. At last, 69 496 TF-target pairs (related to 1974 TFs), 37 040 TF-target pairs (related to 1340 TFs) and 18 055 TF-target pairs (related to 541 TFs) in human, mouse and rat were gathered with a strict threshold ($P < 1e-4$).

In order to evaluate the correctness of TF-target pairs obtained here, orthologous counterparts comparison among the three species was carried out as following: (1) orthologous mapping information about mouse–human and mouse–rat was collected from the MGI database v4.01 (http://www.informatics.jax.org/); (2) mouse TF-target pairs were mapped to those of human and rat according to the mapping information and (3) relevant numbers of TF-target pairs were calculated for mouse–human and mouse–rat, respectively, so as to do statistic test later. As a result, 3012 and 1009 orthologous counterparts projected from mouse were detected with coverage in human and rat TF-target pairs.

A fisher's exact test was carried out to assess the consistency of performance for our method on different species. The test was operated between mouse and Y (Y stand for human or rat) as below: (1) a 2×2 table is used for the test; (2) the top left cell was the coverage number between orthologous counterparts projected from mouse and pairs of Y; (3) the top right cell was the number of orthologous counterparts projected from mouse but were not predicted in Y; (4) the bottom left cell was the number of orthologous counterparts projected from Y but were not predicted in mouse; (5) the bottom right cell was the number of orthologous pairs between mouse and Y, though they were neither predicted in mouse nor in Y and (6) the odds ratio and *P*-value for the test was calculated. As a result, the odds ratio achieved 101.0018 ($P < 1e-200$) and 134.4441 ($P < 1e-200$) for mouse–human and mouse–rat test, respectively. Result of fisher's exact test indicated that TF-target pairs predicted in different species were fairly consistent, which demonstrated approach used here to predict TF-target pairs were robust and reliable, and information of downstream regulated genes for TFs inferred here was believable to some extent.

4 DISCUSSION

ITFP is a comprehensive information platform. Compared with TRANSFAC and other similar databases, it presents more information of TFs and downstream regulated genes for mammalian. TRANSFAC mainly collects experimentally proven data, so it cannot give any information for new TFs. In our work, plentiful data about TF are offered through computational biology method, which will provide useful reference information for lots of TFs to be investigated. In future, more information about TF [e.g. TF binding sites (TFBS) and binding preference between TF and TFBS] will be added to the platform. We believe that the platform will provide more powerful support for the research of transcription regulation.

Funding: High-Tech Research and Development Program of China (grant no. 2006AA02Z329); National Natural Science Foundation of China (grant no. 60573093); National Basic Research Program of China (grant no. 2006CB910700, 2004CB720103, 2004CB518606, 2003CB715901); Funding of Chinese Academy of Sciences (grant no. KSCX2-YW-R-112); Shanghai Pujiang Program (06PJ14073).

Conflict of Interest: none declared.

REFERENCES

Basso,K. *et al*. (2005) Reverse engineering of regulatory networks in human B cells. *Nat. Genet.*, **37**, 382–390.

Margolin,A.A. *et al*. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7** (Suppl. 1), S7.

Matys,V. *et al*. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Zhao,F. *et al*. (2005) TRED: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Res.*, **33**, D103–D107.

Zheng,G. *et al*. (2008) The combination approach of SVM and ECOC for powerful identification and classification of transcription factor. *BMC Bioinformatics*, **9**, 282.