Nucleic Acids Research, 2007, Vol. 35, Database issue ***D137–D140***
doi:10.1093/nar/gkl1041

TRED: a transcriptional regulatory element database, new entries and other development

C. Jiang, Z. Xuan, F. Zhao and M. Q. Zhang*

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA

Received September 8, 2006; Revised and Accepted November 3, 2006

ABSTRACT

Transcriptional factors (TFs) and many of their target genes are involved in gene regulation at the level of transcription. To decipher gene regulatory networks (GRNs) we require a comprehensive and accurate knowledge of transcriptional regulatory elements. TRED (http://rulai.cshl.edu/TRED) was designed as a resource for gene regulation and function studies. It collects mammalian *cis*- and *trans*-regulatory elements together with experimental evidence. All the regulatory elements were mapped on to the assembled genomes. In this new release, we included a total of 36 TF families involved in cancer. Accordingly, the number of target promoters and genes for TF families has increased dramatically. There are 11 660 target genes (7479 in human, 2691 in mouse and 1490 in rat) and 14 908 target promoters (10 225 in human, 2985 in mouse and 1698 in rat). Additionally, we constructed GRNs for each TF family by connecting the TF–target gene pairs. Such interaction data between TFs and their target genes will assist detailed functional studies and help to obtain a panoramic view of the GRNs for cancer research.

OVERVIEW

TRED was originally designed as a resource for studies on gene function and regulation. It provides *cis*-elements, such as promoters and binding motifs, and *trans*-elements, such as transcriptional factors (TFs). The promoters of target genes came from two sources: experimental determination and sequence-based computational prediction. These two sources complement each other. In TRED, hand curation was applied as a crucial part of the data collection to ensure data accuracy. Based on the reliability of the supporting evidence for each promoter, a quality level was assigned.

One key feature of TRED is the easy access to interaction data between TFs and the promoters of their target genes, including binding motifs reported by the previous studies. Although part of the data was obtained from the existing gene regulation resources, most of the data came from our exhaustive literature curation. The TF-binding motifs were mapped on to the promoter sequences of their target genes. Along with the binding motifs, the experimental evidence and other important pertinent information were also collected. A binding quality level was assigned based on definitiveness of the binding evidence, which was determined by the experimental approaches employed to demonstrate the binding and data interpretation from experts. For example, we assigned 'known' as the binding quality level to a binding that has been proved by gel-shift competition, DNase I footprinting, etc.

In order to provide users with more complete information of the genes, cross-references to other well-known database such as PubMed, GenBank, GeneCards (1) and TRANSFAC (2) were established in TRED. A comprehensive description of the content and the structure of TRED has been published earlier (3). In addition, many on-the-fly tools were implemented for the analysis of sequences retrieved from TRED as well as imported from other resources. The user interface and software functionality were also described in the previous report (3).

RECENT DEVELOPMENTS

New entries

Upon the emergence of high-throughput technologies, a huge amount of large-scale gene expression and regulation profiling data have been made available by microarray and chromatin immunoprecipitation (chip-ChIP) studies. To uncover GRNs among the identified genes would require the knowledge of their promoter sequences. We used our promoter prediction program FirstEF (4) to predict promoters in the genomes of human, mouse and rat. These promoters were then combined with the known promoters extracted from EPD (5), DBTSS (6), GenBank, etc and were deposited in our database CSHLmpd (7). It should be noted that CSHLmpd also contains genes without any promoter.

*To whom correspondence should be addressed. Tel: +1 516 367 8393; Fax: +1 516 367 8461; Email: mzhang@cshl.edu
Present addresses:
C. Jiang, Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, PO Box 980126, Richmond, VA 23298, USA
F. Zhao, Global Business Analytics, Global Pharmaceutical Business, Schering-Plough, Kenilworth, NJ 07033, USA

Table 1. Number of promoters and genes in TRED, with gene numbers in parentheses

Promoter quality	1	2	3	4	5 + 6	Sum
Human	1842 (1767)	13 619 (10 115)	5311 (5150)	7305 (6738)	26 684 (15 222)	54 761 (27 016)
Mouse	172 (156)	8407 (6552)	6551 (6449)	4250 (4041)	23 500 (15 185)	42 880 (25 751)
Rat	91 (82)	996 (680)	3374 (3333)	2917 (2834)	25 681 (16 346)	33 059 (21 440)

Promoter qualities are ranked from high to low: 1, known, curated promoters; 2, known, pipeline collected promoters; 3, predicted promoters with Refseq evidence and putative promoters taking 5′ ends of Refseq as TSS; 4, predicted promoters with mRNA (other than Refseq and EST) evidence; 5, predicted promoters with EST evidence; 6, predicted promoters supported only by gene prediction. Promoters included in a higher ranking are automatically excluded from the lower ranking categories.

In this version, the number of genes with promoter(s) in each genome assembly is lower than that of the previous version after removal of the redundancy. However, the number of known promoters and their related genes are close in both versions. There are more promoters than genes in each species due to alternative transcription start sites in many genes (7). Table 1 gives the statistics of promoters and genes in each quality category.

The human genome codes for ~1850 TFs, which account for 6.0% of its estimated total number of genes (8). It is a daunting task to collect and curate comprehensive and precise interaction data between the TFs and their target genes. Since cancer is one of the greatest threats to human health and has been a field under extensive study, including a broad interest in understanding cell cycle regulatory networks, we started out by focusing on target genes of cancer-related TFs. Previously, TRED contained mainly the target genes and promoters for TF families E2F and Myc (3). In this new release, we expanded it to 34 new TF families that have been implicated in cancer pathways, including p53, AP1, ER and NFκB/Rel. They are involved in many cellular processes, such as proliferation, differentiation, cell motility and apoptosis. There are totally 9308 newly collected target genes (5365 in human, 2526 in mouse and 1417 in rat) and 10 251 target promoters (5956 in human, 2736 in mouse and 1559 in rat) for these TF families. The detailed distribution of the target promoters and the target genes is listed in Table 2.

Although TRANSFAC also provides factor-site interaction data, it contains less information for the TFs and their target genes available in TRED. Its latest version (version 7.0) has collected 1040 factors and 608 genes for human and 765 factors and 417 genes for mouse (2). Therefore, on average each factor has less than one target gene. In contrast, the number of target genes per TF in TRED is much higher. For example, there are >200 target genes on average for each TF family for human in TRED (Table 2). This can provide fairly resolved gene regulatory networks (GRNs) for the 36 TF families involved in cancer pathways. In addition to this, TRED contains relatively complete genome-wide promoter annotation for human, mouse and rat. Moreover, the binding sites in TRED were also mapped on to the assembled chromosomes. These absolute genomic positions make it ready to associate TRED with other genomic data for various studies. However, it should be noted that TRANSFAC also collects factor–site interaction data of species other than human, mouse and rat. Therefore, although TRED and TRANSFAC overlap to certain extent, they complement each other at some aspects.

Table 2. Number of curated target promoters/genes for the 36 TF families

TF	Human	Mouse	Rat
AP1	432/383	217/190	157/143
AP2	338/318	123/123	90/86
AR	69/49	19/19	24/15
ATF	189/173	59/59	26/26
BCL	21/19	15/15	0/0
BRCA	20/20	4/4	0/0
CEBP	335/325	152/134	241/179
CREB	224/220	138/133	95/93
E2F	1593/1329	141/127	11/11
EGR	120/111	67/55	33/26
ELK	47/41	15/13	6/6
ER	169/152	40/39	32/31
ERG	21/21	5/5	0/0
ETS	445/412	207/196	51/51
FLI1	41/41	17/16	0/0
GLI	16/16	8/8	0/0
HIF	119/112	63/60	29/29
HLF	10/10	5/5	2/2
HOX	65/57	93/81	5/5
LEF	40/33	26/23	5/5
MYB	253/239	40/40	6/6
MYC	2676/785	108/38	128/62
NFI	136/127	75/62	73/65
OCT	232/195	123/108	34/34
p53	337/313	135/130	32/30
PAX	52/47	76/61	13/11
PPAR	149/149	125/124	88/84
PR	31/27	14/14	10/10
RAR	233/218	71/71	40/40
REL	445/396	202/181	87/87
SMAD	139/130	76/75	17/17
SP	655/515	296/263	235/220
STAT	245/218	111/106	48/46
TAL1	15/14	9/6	0/0
USF	235/215	94/91	72/62
WT1	78/49	16/16	8/8

Other development

The accurate and comprehensive knowledge of transcriptional regulatory elements in TRED allows one to construct the GRNs for a given TF family by bringing all TF–target gene pairs together. In our initial analysis, we found that some TFs are the target genes of other TFs and often more than one TFs control the expression of the same gene. Furthermore, some target genes affect the expression or stability of TFs by feedback loop. As an example, Figure 1 shows a simplified GRN for TF family GLI (glioma-associated oncogene homolog). For the TF families with hundreds of target genes, such as AP1, CEBP and ETS, they would form more complex networks. There are cross talks between the

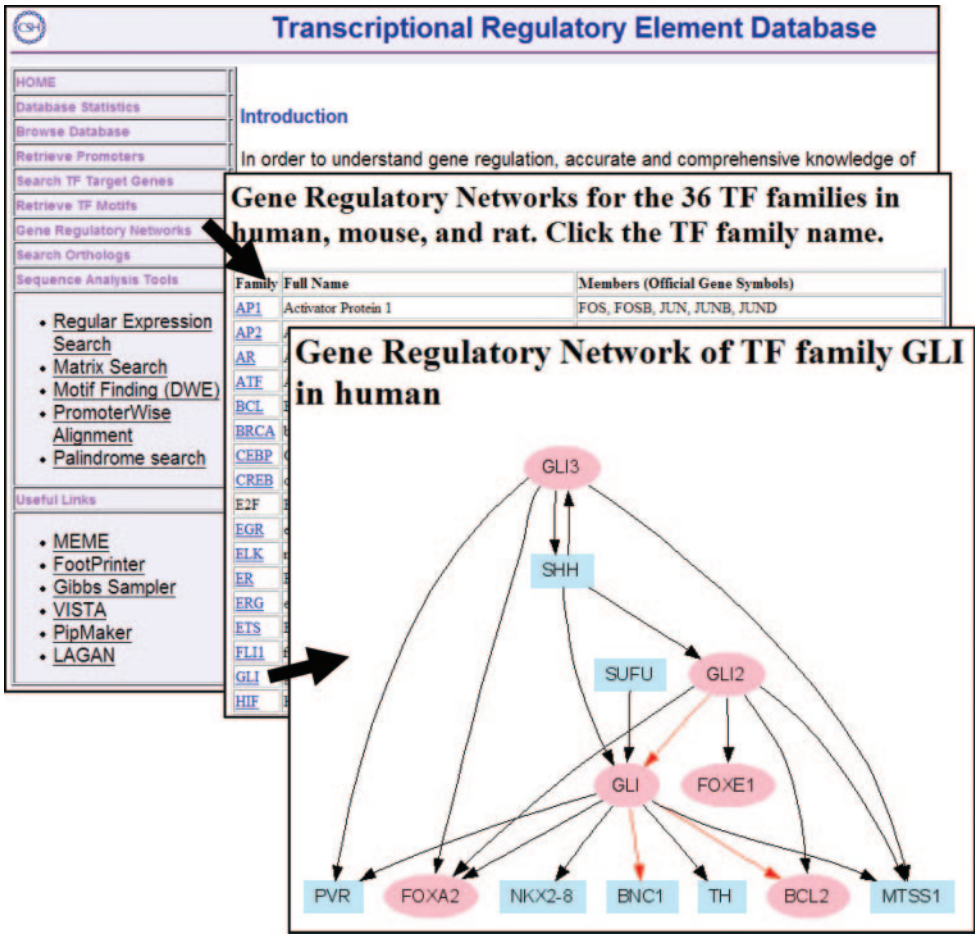

Figure 1. Sample pages showing access to the gene regulatory network of TF family GLI (glioma-associated oncogene homolog) in human. Ellipses, TFs; and squares, genes. Arrows indicate interactions between two genes. Red arrows imply that the binding quality level is known. Only official gene symbols are used in the network.

networks of different TF families through the same target genes or through direct interactions between the members of different TF families. The experimental evidence for each interaction between a TF and its target gene is available through the references provided in TRED. This is an advantage over other networks computationally predicted from expression and/or phylogenetic profiles. In this release, GRNs for the TF families have been generated from the collected interaction data and statically stored in TRED. The dynamic links to GRNs will be provided in the query result in the future. Taken together, TRED can facilitate to decipher the GRNs and help researchers to better understand the gene regulatory mechanisms.

DATA ACCESS

The website (http://rulai.cshl.edu/TRED) offers the following services: easy access to TRED entries through text-based query interface; search for the target genes of a given TF; retrieval of the promoter sequences and the TF-binding motifs; further analysis of the retrieved sequences of promoters and motifs. TRED homepage also provides the access to the GRNs of the TF families in human, mouse and rat, which were constructed from its collection of the interaction data between the TFs and their target genes.

ACKNOWLEDGEMENTS

We thank Dr Thomas Kubarych for the critical reading of the manuscript for English. This work is supported by a contract grant from NCI and partly by a NIH grant HG001696. Funding to pay the Open Access publication charges for this article was provided by NIH/NHGRI.

Conflict of interest statement. None declared.

REFERENCES

1. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
2. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

3. Zhao,F., Xuan,Z., Liu,L. and Zhang,M.Q. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for *in silico* gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.
4. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet*., **29**, 412–417.
5. Schmid,C.D., Perier,R., Praz,V. and Bucher,P. (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res*., **34**, D82–D85.
6. Yamashita,R., Suzuki,Y., Wakaguri,H., Tsuritani,K., Nakai,K. and Sugano,S. (2006) DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res*., **34**, D86–D89.
7. Xuan,Z., Zhao,F., Wang,J., Chen,G. and Zhang,M.Q. (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol*., **6**, R72.
8. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.