Transcription Regulatory Regions Database (TRRD): its status in 1999


Institute of Cytology and Genetics (Siberian Branch of the Russian Academy of Sciences), Lavrentieva 10, Novosibirsk 630090, Russia

Received September 24, 1998; Accepted September 28, 1998

ABSTRACT

The Transcription Regulatory Regions Database (TRRD) is a curated database designed for accumulation of experimental data on extended regulatory regions of eukaryotic genes, the regulatory elements they contain, i.e., transcription factor binding sites, promoters, enhancers, silencers, etc., and expression patterns of the genes. Release 4.1 of TRRD offers a number of significant improvements, in particular, a more detailed description of transcription factor binding sites, transcription factors per se, and gene expression patterns in a computer-readable format. In addition, the new TRRD release provides considerably more references to other molecular biological databases. TRRD 4.1 is installed under SRS and is available through the WWW at http://www.bionet.nsc.ru/trrd/

INTRODUCTION

The Transcription Regulatory Regions Database (TRRD) is designed for accumulation of experimental data on extended regulatory regions of eukaryotic genes. The TRRD format allows to describe the modular structure of transcription regulatory regions and the hierarchy of theirs constituent regulatory units. The following regulatory units are considered: (i) cis-elements providing the interaction of transcription factors with DNA (1); (ii) composite elements supporting DNA–protein and protein–protein interactions between the neighbouring sites, and the corresponding factors causing either synergistic or antagonistic regulatory effects (2,3); (iii) promoters providing formation of the basal transcription complexes; (iv) enhancers and silencers modulating transcription level; (v) extended transcription regulatory regions located in 5′- and 3′-flanking regions of the genes and introns (4,5); and (vi) the system of integral regulation of gene transcription, which comprises all these regulatory elements and provides transcription regulation depending on the cell cycle stage, the stage of development, tissue-specificity, environmental conditions, the effect of inducers, etc. A database entry corresponds to a single gene.

The TRRD database is being developed at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, since 1993. The previous TRRD releases were described earlier (4,6–9). During the last year, TRRD was considerably improved by introducing a new format of data representation. The TRRD release 4.1 provides a more complete description of gene expression regulation patterns and the structural peculiarities of regulatory regions so that the maximum information is presented in a computer-readable form. In addition, new functional groups of genes have been added. TRRD 4.1 is installed under SRS and is available through the WWW at http://www.bionet.nsc.ru/trrd/

STRUCTURE OF A DATABASE ENTRY

The structure of TRRD entries reflects the modular organization of the gene regulatory regions and the hierarchy of constituent regulatory units. Each line of an entry begins with a two-character code indicating the type of information contained in the line and denoting some information field in TRRD.

At the moment, TRRD contains 82 various line types. As an example, the entry containing the information on transcription regulation of mouse glutathione peroxidase gene is shown in Figure 1.

The following regulatory units are considered: (i) cis-elements providing the interaction of transcription factors with DNA (1); (ii) composite elements supporting DNA–protein and protein–protein interactions between the neighbouring sites, and the corresponding factors causing either synergistic or antagonistic regulatory effects (2,3); (iii) promoters providing formation of the basal transcription complexes; (iv) enhancers and silencers modulating transcription level; (v) extended transcription regulatory regions located in 5′- and 3′-flanking regions of the genes and introns (4,5); and (vi) the system of integral regulation of gene transcription, which comprises all these regulatory elements and provides transcription regulation depending on the cell cycle stage, the stage of development, tissue-specificity, environmental conditions, the effect of inducers, etc. A database entry corresponds to a single gene.

*To whom correspondence should be addressed. Tel: +7 3832 333468; Fax: +7 3832 331278; Email: kol@bionet.nsc.ru
The description of a site is presented in the following lines: AN, site accession number; NM, site name; AT, alteration of transcription level caused by binding of the factor to the site; SQ, site sequence; PQ and PF, the positions of site and footprint relative to the referring point; BF, position of the first nucleotide of the site according to EMBL/GenBank; and AG, codes of an experiment. In the example considered, the site Ets3 (AN 2339) is located between positions +1241 and +1260 and has a core sequence GAGGAAG. It was estimated in MEL cells and fetal liver cell extracts by applying the method 3.4 (cross-competition gel-shift assay). In case the Internet-accessible program for recognition of the site is available, the reference to this program is provided in the line WW (as is indicated for GATA AN 2348).

Information on the factors binding to the sites is described in 11 lines including: TF, the abbreviated and full name of the factor; SY, synonymous names; TS, species; NF, accession number of the factor in TRANSFAC database; TO, type of the factor (recombinant, endogenous, etc.); TG, TU, TC, organs, tissues, and cells used for isolation of the factor; TD, cell inductor name; TR, bibliographic reference; CC, comments. For example, in the case of GATA site AN 2348, information about interacting endogenous transcription factor GATA1, NF T00305, isolated from MEL cells, is given in five lines of green coloured block.

To provide the standardization of information in TRRD, the program TRRD-INPUT was developed. This program realizes in Visual FoxPro 5.0 using OLE technology and ActiveX elements produces an entry corresponding to an individual gene in a flat file format as described above. This program enables both edition and creating of new entries. Individual lines of an entry are checked for their compliance with the vocabularies supported within the TRRD database. In total, TRRD contains 21 vocabularies comprising over 3500 words: five vocabularies are stable, the rest are being expanded.

### INSTALLATION OF TRRD UNDER SRS

The TRRD 4.1 release is installed under SRS (Sequence Retrieval System) (http://www.bionet.nsc.ru/trrd/), to provide easy information retrieval and integration with the other molecular biological databases and software for data analysis. Installation under SRS distributes the contents of the flat file corresponding
to an entry between five interconnected tables (Fig. 2): TRRDGENES, TRRDEXP, TRRDSITES, TRRDFAC and TRRDBIB, described below.

(i) **TRRDGENES** table contains general information about all the genes described in TRRD and their regulatory units. It is the central table of the database providing linkage with all the other tables. The TRRDGENES table is also supplied with the references to the databases TRANSFAC (8), COMPEL (3,4), EMBL/GenBank (10,11), EPD (12), EpoGERD (13), EpoDB (14), SwissProt (15) and GeneNet (16). The automated MEDLINE search is available according to the key words included in the TRRDGENES table. In addition, the TRRDGENES table connects the database with the graphic interface TRRD-Viewer.

(ii) **TRRDSITES** table contains the detailed information on transcription factor binding sites. Site description is provided by the references to TRRDGENES, TRRDFAC and TRRDBIB tables of TRRD. In addition, the TRRDSITES table is connected to TRANSFAC and EMBL/GenBank databases, as well as to ACTIVITY and SiteRecognition software of the GeneExpress System (17).

(iii) **TRRDFAC** table contains the detailed description of the transcription factors binding to the sites stored in TRRD. Factors contained in TRRDFAC table are supplied by the references to their descriptions in TRANSFAC database (8) and the references to TRRDGENES, TRRDSITES and TRRDBIB tables.

(iv) **TRRDEXP** table contains the description of the gene expression patterns. It is supplied with references to TRRDGENES and TRRDBIB tables as well as to TRRDSITES table offering information on the sites providing this particular expression pattern.

(v) **TRRDBIB** table contains the complete bibliographic references of all the papers annotated in TRRD and is linked with TRRDGENES table and MEDLINE.

**GRAPHIC USER INTERFACE**

The Java applet TRRD-Viewer provides a visualization of the data on transcription factor binding site location in a map form (Fig. 3) and the textual description of genes and sites. Applying the applet, a user selects a gene identifier from the list, and the graphical visualization of information on transcription factor binding sites and composite elements appears on the screen.

**OPTIONS ALLOWING DIFFERENT SITE REPRESENTATIONS ARE PROVIDED.** Besides, in the text window, a description of the gene extracted from TRRDGENES table appears. If the user clicks the site image, site description from TRRDSITES table is displayed in the text window, where the references from TRRDBIB table can also be seen.

**CONTENT OF THE CURRENT RELEASE**

The current release, TRRD 4.1, comprises the description of 514 genes, 717 regulatory units including 432 promoters, 139 enhancers, 34 silencers, 74 composite elements and 2472 transcription factor binding sites. More than 1700 scientific publications were annotated.

The genes contained in TRRD could be classified into groups according to species specificity, type of protein encoded by the gene, and the functional role of a gene.

The genes described in TRRD refer to different eukaryotic species: human (39.7%), mouse (25.1%), rat (16%), chicken (6%), hamster (1.2%) and others (4.3%). This release also contains information on transcription regulatory regions in plants (7.7%) that was lacking in the previous releases.

**TRRD contains the information on genes encoding proteins with a wide variety of functions.** According to EPD database classification (12), TRRD is subdivided into several groups: genes encoding structural proteins (16%), storage and transport proteins (20%), enzymes (19%), regulatory proteins, including hormones, growth factors, etc. (20%), proteins related to stress or pathogen defence reactions (10%) and others (15%).

Genes described in TRRD can be subdivided according to their functional role (18–23): interferon-inducible genes (60), erythroid-specific regulated genes (44), genes of lipid metabolism (48), glucocorticoid-controlled genes (35), cell cycle-dependent genes (20), endocrine system genes (41), heat shock genes (26) and plant genes (40), available via http://www.bionet.nsc.ru/trrd/

**FUTURE PROSPECTS**

The following extensions of TRRD are planned in future. First, the format of experimental data presentation in TRRD will be improved. In particular, the novel formats will be developed for description of interaction of transcription factors to basal transcription machinery, the influence of nucleosomal chromatin organization, methylation, and mutations on transcription regulation. Second, we intend to start the integration of TRRD with a variety of other molecular biological databases available via Internet using CORBA technique. We plan to continue integration of TRRD with various software for analysis and recognition of regulatory genomic sequences within the framework of GeneExpress system (17). Finally, we plan to continue the expansion of TRRD with the new experimental data on transcription regulation, with special emphasis on the genes controlling hematopoiesis, stress response, and functioning of nervous, endocrine and immune systems.

**AVAILABILITY**

TRRD 4.1 is available through the WWW at http://www.bionet.nsc.ru/trrd/. TRRD flat files are available on a collaborative basis.

No inclusion of TRRD into other databases without explicit permission of the authors. All rights reserved. The administrator of TRRD, Nikolay A. Kolchanov, can be contacted by Email at
kol@bionet.nsc.ru. Please send comments, corrections and requests for additional information to us by Email or Fax (+7 3832 331278). Users are asked to refer to this paper and ref. 4 in reporting results based on the use of TRRD.

ACKNOWLEDGEMENTS

The authors are grateful to I. V. Lochova for assistance in bibliographic search, to G. B. Chirikova for translation of the paper into English, and to G. V. Orlova for helpful comments. The work is supported by the Russian Foundation for Basic Research (grants Nos. 96-04-50006, 97-04-49740, 98-04-49479, 98-07-91078), Integration Program of Siberian Branch of Russian Academy of Sciences (IGSBRAS-97*13), Russian Human Genome Project and Russian Ministry of Sciences.

REFERENCES