

TRRD: A DATABASE ON EXPERIMENTALLY IDENTIFIED TRANSCRIPTION REGULATORY REGIONS AND TRANSCRIPTION FACTOR BINDING SITES

Kolchanov N.A. , Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Khlebodarova T.M., Merkulov V.M., Merkulova T.I., Podkolodny N.L., Romashenko A.G.*

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia

* Corresponding author: e-mail: kol@bionet.nsc.ru

Key words: transcription regulation, regulatory region, transcription factor binding site, gene expression

SUMMARY

Motivation: The main goal of TRRD (Transcription Regulatory Regions Database) development is the most complete and adequate description of the structural and functional organization of transcription regulatory gene regions in eukaryotes based on the data obtained experimentally.

Results: The overall information contained in the current TRRD release is represented as eight libraries: TRRDGENES, TRRDUNITS, TRRDEXP, TRRDSITES, TRRDFACTORS, TRRDLCR, TRRDSTARTS, and TRRDBIB. TRRD compiles the data on 2344 genes, 14 407 patterns of their expression, 3490 regulatory units, and 10 135 transcription factor binding sites associated with them. This database contains only experimentally confirmed information. TRRD is filled in by manual annotation of scientific publications. The data incorporated into TRRD is a result of annotation of 7609 scientific papers. The main tool for searching TRRD and navigation in it is SRS. A large number of indexed fields in the SRS version of TRRD allow the user to generate complex queries both within individual libraries and involving several libraries. TRRD has thesauruses that provide additional options for data access. The number of databases linked to TRRD has been increased.

Availability: <http://www.bionet.nsc.ru/trrd/>.

INTRODUCTION

The structure–function organization of regulatory regions in the genes transcribed by RNA polymerase II is typically very intricate. The presence of alternative promoters and remote regulatory regions localized to both the 5'- and 3'-gene–flanking regions as well as to introns and exons are typical of the numerous genes studied so far. Active contributors to combinatorial gene regulation are the structural elements of core promoters (Smale, Kadonaga, 2003). Transcription factor binding sites within a regulatory unit (promoter, enhancer, or silencer) may be organized in functional modules that determine one or another expression pattern of a gene. One more functionally important characteristic is the multiple transcription starts. This particular information may be very important, as individual transcription starts of one promoter are frequently used for producing

transcripts in different tissues or under different conditions (under the action of inducers, at various ontogenetic stages, etc.).

All these facts clearly indicate that the description of an integrated system of transcription regulation requires the comprehensive information about the regulatory elements of the gene. Creation of collections of experimentally discovered data on the regulatory elements acting at all levels is absolutely necessary for both forming the concepts of what is the nature of regulation of individual genes and developing the computer methods for prediction of regulatory elements, construction of gene networks, and functional genome annotation. TRRD, which we are presenting here, is a unique information resource that is developed aiming to provide an integrated description of transcription regulation of the eukaryotic genes transcribed by RNA pol II. The database is being constantly supplemented with new information, and the TRRD format is being permanently developed. Based on the information contained in TRRD, tools for analyzing regulatory regions of the genes transcribed by RNA pol II were developed.

Structure of the TRRD database and data source

All the information contained in TRRD² is distributed between eight interconnected libraries. The TRRDGENES is the central, integrating library, which compiles the information identifying the gene, internal references to other TRRD libraries, and references to external databases and resources as well as hierarchically organized representation of the regulatory elements of all levels. The rest information tables of TRRD are TRRDSITES collating the information about transcription factor binding sites; TRRDUNITS describing regulatory units (promoters, enhancers, and silencers); TRRDLCR containing the structure–function characteristics the locus control regions (LCR); TRRDSTARTS containing the data on transcription initiation starts; TRRDEXP compiling the description of the qualitative specific features of gene expression; and TRRDBIB containing the bibliographic information. The description of information fields was given in detail previously (Kolchanov *et al.*, 2000, 2002). TRRD is filled in by manual annotation of scientific papers. The database contains only experimentally confirmed information obtained in experiments of various types (<http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+FieldInfo+-lib+TRRDSITES4+-bf+ExperimentCodes>). The data input is standardized via the system of controlled vocabularies.

RECENT DEVELOPMENTS

Development of the TRRD 7.0 format

The format of TRRD is being constantly developed to enhance the search of the database and simplify the data access. In TRRD release 7.0, the TRRDGENES library contains a considerably larger number of links to external databases: in addition to the previously available references to SWISS-PROT and EMBL/GenBank, note the links of the current release to Entrez Gene, GeneCards, MGI, RGD, FlyBase, and MaizeDB (overall, more than 20 databases).

A new library, TRRDSTARTS, was developed. This library compiles the data on the experimentally determined transcription start sites of genes. TRRDSTARTS contains the absolute genome coordinates of the major and minor transcription starts of genes (with indication of the chromosome and the release of genomic database).

The format of TRRDSITES library was extended. A new field, PreferredName (NP), was added; this field contains the standard (preferred) site name. The field PreferredName

² In the public version of TRRD, a number of information fields are not available, in particular, the sequences of transcription factor binding sites (TFBS) and regulatory regions, their localization in the corresponding entries of EMBL or GenBank database of nucleotide sequences, and the TFBS localization relative to a particular reference point within the gene.

is filled in automatically based on the data from the field TF of the block FACTOR, connected to this site. In this process, a specialized vocabulary of transcription factors is used, where the relations (the first order hierarchy and synonymy) between the names are fixed too. Thus, a query to the field PreferredName gives the possibility to obtain the entire information contained in TRRD that is related to the transcription factor binding sites of the user-specified type independently of what synonymic factor names were used in the query.

The TRRDFACTORS library was revised. The vocabularies of transcription factors were unified to assign a unique identifier to each factor. Description of the subunit composition for multimeric factors is provided.

The extension of TRRD content

TRRD is filled up constantly with the new information. The number of entries in TRRD release 6.0 (Kolchanov *et al.*, 2002) and the current release 7.0 (as of September 01, 2005) are listed in Table 1.

Table 1. The information content of TRRD

Library name	Number of entries in release 6.0	Number of entries in the current release 7.0	Including the species (%)			
			Human	Mouse	Rat	Others
TRRDGENES	1167	2344	32	22	15	31
TRRDUNITS	1714	3490	36	19	14	31
TRRDEXP	5335	14 407	37	37	18	18
TRRDSITES	5537	10 135	36	18	14	32
TRRDBIB	3898	7609	37	21	16	26

The sections on a number of subjects are being developed in TRRD that include genes united according to various functional characteristics. Each of the sections contains a group of genes expressed under certain conditions or involved in a certain process. Overall, nine sections were described earlier (Kolchanov *et al.*, 2002). At present, TRRD contains 18 sections of this type (<http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/sections1.shtml>). These sections are also a tool for quick access to the information contained in TRRD.

New possibilities for access to TRRD data

SRS (Sequence Retrieval System) version 6.1.3.11, which provides searching for information over 132 indexed fields, is the main tool for accessing TRRD. In addition, several specialized search systems were developed. These systems are based on the controlled vocabularies of tissues, cells, organs, developmental stages, external stimuli, and transcription factors, on the one hand, and thesauruses on organs and tissues in mammals (<http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/>), on the other. During operation of these searching systems, the relations of the types “general–particular”, “part–whole”, “synonymy”, etc., are realized. The queries to the SRS version of TRRD are realized not only according to a specified term, but also by all the related terms (daughter with reference to the query term) in the corresponding vocabulary as well as by all the synonyms simultaneously. These searching systems (<http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/search.html> and http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/search_hidden.html) provide (1) search for the genes induced (or repressed) by an external stimulus; (2) search for the genes expressed in a specified organ, tissue, cell type, or stage of organism development; (3) a combined search for the genes expressed in a specified tissue, organ, or cell type when induced by a specified external stimulus (simultaneously); and (4) search for the genes or sites regulated by a specified transcription factor.

New tools for analysis of DNA sequences using TRRD

Three new tools for prediction of transcription factor binding sites and promoters were developed based on the information collected in TRRD: (1) SITECON (Oshchepkov *et al.*, 2004) and SiteGA (Levitsky *et al.*, 2006) for site recognition and (2) ARGO

(Vishnevsky, Kolchanov, 2005) for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters.

ACKNOWLEDGEMENTS

The authors are grateful to I.V. Lokhova for bibliographical support and to G.B. Chirikova for translation of the paper into English. The work was supported by RFBR (grants Nos 05-07-98012 and 05-04-49111), Siberian Branch of the Russian Academy of Sciences (integration project No. 119), the government contract with the Federal Agency for Science and Technology “Identification of potential targets for novel medicinal drugs based on reconstructed gene networks”, the priority direction “Living systems”, innovation project of Federal Agency of Science and Innovation IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)”, NATO (grant No. PDD(CP)-(LST.CLG 979815), and INTAS (project No. 2382).

REFERENCES

- Kolchanov N.A. *et al.* (2000) Transcription regulatory regions database (TRRD): its status in 2000. *Nucl. Acids Res.*, **28**, 298–301.
- Kolchanov N.A. *et al.* (2002). Transcription regulatory regions database (TRRD): its status in 2002. *Nucl. Acids Res.*, **30**, 312–317.
- Levitsky V.G. *et al.* (2006) Method SiteGA for recognition of transcription factor binding sites. *Biofizika* (in Russ.) (in press).
- Smale S.T., Kadonaga J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
- Oshchepkov D.Y. *et al.* (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res.*, **32**, W208–W212.
- Vishnevsky O.V., Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucl. Acids Res.*, **33**, W417–W422.