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

N. A. Kolchanov*, E. A. Ananko, O. A. Podkolodnaya, E. V. Ignatieva, I. L. Stepanenko, O. V. Kel-Margoulis, A. E. Kel, T. I. Merkulova, T. N. Goryachkovskaya, T. V. Busygina, F. A. Kolpakov, N. L. Podkolodny, A. N. Naumochkin and A. G. Romashchenko

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

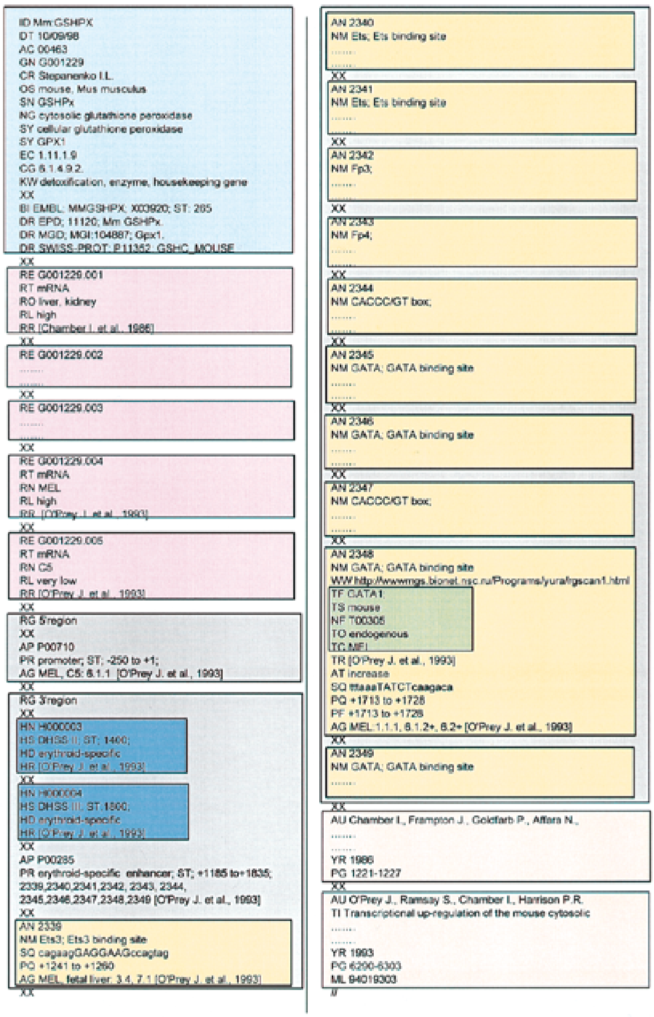
The structure of TRRD entries reflects the modular organization of the gene regulatory regions and the hierarchy of constituent regulatory units. Each line of an entry begins with a two-character line code indicating the type of information contained in the line and denoting some information field in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

General gene description is based on 14 fields: **ID**, identifier; **DT**, date of the last update; **AC**, TRRD gene accession number; **GN**, unified TRRD and TRANSFAC gene accession number (7); **CR**, name of an annotator; **OS**, English and Latin names of species; **SN**, **NG** and **SY**, short, full gene names and synonyms, respectively; **EC,** enzyme classification; **CG**, promoter classification according to EPD database, release 45; **KW**, key words; **BI**, references to EMBL/GenBank and **DR**, references to other databases. Lines of the general description of a gene are marked by blue in Figure 1. Eight lines (**ID**–**NG**) are mandatory for description of any gene. It should be noted that unified gene accession number **GN** was introduced to provide efficient interoperability of TRRD with the TRANSFAC (8) and COMPEL (2,3) databases.

Gene expression patterns are described by 17 lines: **RE**, pattern identifier; **RT**, molecular product used to estimate the expression level (protein or mRNA); **RY**, cell cycle stage (from G0 to M);

*To whom correspondence should be addressed. Tel: +7 3832 333468; Fax: +7 3832 331278; Email: kol@bionet.nsc.ru



```
ID Mm:GSHPX
DT 10/09/98
AC 00463
GN G001229
CR Stepanenko I.L.
OS mouse, Mus musculus
SN GSHPx
NG cytosolic glutathione peroxidase
SY cellular glutathione peroxidase
SY GPX1
EC 1.11.1.9
CG 6.1.4.9.2.
KW detoxification, enzyme, housekeeping gene
XX
BI EMBL: MMGSHPX; X03920; ST: 265
DR EPD; 11120; Mm GSHPx.
DR MGD; MGI:104887; Gpx1.
DR SWISS-PROT: P11352, GSHC_MOUSE
XX
RE G001229.001
RT mRNA
RO liver, kidney
RL high
RR [Chamber I. et al., 1986]
XX
RE G001229.002
.......
.......
XX
RE G001229.003
.......
.......
XX
RE G001229.004
RT mRNA
RN MEL
RL high
RR [O'Prey J. et al., 1993]
XX
RE G001229.005
RT mRNA
RN C5
RL very low
RR [O'Prey J. et al., 1993]
XX
RG 5'region
XX
AP P00710
PR promoter; ST; -250 to +1;
AG MEL, C5: 8.1.1 [O'Prey J. et al., 1993]
XX
RG 3'region
XX
HN H000003
HS DHSS II; ST; 1400;
HD erythroid-specific
HR [O'Prey J. et al., 1993]
XX
HN H000004
HS DHSS III; ST:1800;
HD erythroid-specific
HR [O'Prey J. et al., 1993]
XX
AP P00285
PR erythroid-specific enhancer; ST; +1185 to+1835;
2339,2340,2341,2342, 2343, 2344,
2345,2346,2347,2348,2349 [O'Prey J. et al., 1993]
XX
AN 2339
NM Ets3; Ets3 binding site
SQ cagaagGAGGAAGccagtag
PQ +1241 to +1260
AG MEL, fetal liver: 3.4, 7.1 [O'Prey J. et al., 1993]
XX
AN 2340
NM Ets; Ets binding site
.......
.......
XX
AN 2341
NM Ets; Ets binding site
.......
.......
XX
AN 2342
NM Fp3;
.......
.......
XX
AN 2343
NM Fp4;
.......
.......
XX
AN 2344
NM CACCC/GT box;
.......
.......
XX
AN 2345
NM GATA; GATA binding site
.......
.......
XX
AN 2346
NM GATA; GATA binding site
.......
.......
XX
AN 2347
NM CACCC/GT box;
.......
.......
XX
AN 2348
NM GATA; GATA binding site
WW http://wwwmgs.bionet.nsc.ru/Programs/yura/tgscan1.html
TF GATA1;
TS mouse
NF T00305
TO endogenous
TC MEL
TR [O'Prey J. et al., 1993]
AT increase
SQ tttaaaTATCTcaagaca
PQ +1713 to +1728
PF +1713 to +1728
AG MEL:1.1.1, 8.1.2+, 8.2+ [O'Prey J. et al., 1993]
XX
AN 2349
NM GATA; GATA binding site
.......
.......
XX
AU Chamber I., Frampton J., Goldfarb P., Affara N.,
.......
.......
YR 1986
PG 1221-1227
XX
AU O'Prey J., Ramsay S., Chamber I., Harrison P.R.
TI Transcriptional up-regulation of the mouse cytosolic
.......
.......
YR 1993
PG 6290-6300
ML 94019303
//
```

Figure 1. Hierarchical organization of a TRRD entry. The gene of mouse glutathione peroxidase is taken as an example. Lines describing characteristic features of transcription regulation are united in blocks and marked by various colours.

RD, developmental stage (embryo, fetus, etc.); **RO**, organ (liver, brain, kidney, etc.); **RU**, tissue (muscle, bone marrow, etc.); **RN**, cell type (hepatocyte, myocyte, erythrocyte, etc.); **RX**, sex of an organism; **RL**, gene expression level; **RI**, external signal (heat shock, cAMP, starvation, interleukins, hormones, etc.); **RH**, duration of the signal's effect; **FF**, the effect produced by external signal; **RS** and **RP**, accession numbers of the sites and regulatory units, respectively, that contribute to the expression pattern; **RR**, reference to the paper; **CC**, comments on the general features of gene expression; and **RC**, comments on the particular expression pattern. Gene expression pattern lines are coloured pink in Figure 1. Two fields, **RE** and **RR**, are mandatory.

It is essential that several expression patterns can be revealed experimentally for one and the same gene. For example, five expression patterns of mouse glutathione peroxidase gene designated by pink blocks are described in the entry (Fig. 1). In particular, the pattern **RE** G001229.001 corresponds to the high expression level of this gene in liver and kidney estimated according to mRNA expression level, while the pattern G001229.005 represents a low expression level of the gene in C5 cell line.

In the next item of the entry, the regulatory regions contained in the 5′ and 3′ termini of the gene, in introns and exons are described. In Figure 1, the information describing regulatory regions is presented against a grey background. For instance, two gene regions **RG** are revealed for the gene in question (Fig. 1). One of them, **RG** 5′ region, contains the promoter **PR** with accession number P00710. The other, **RG** 3′ region, has a complex structure. It contains two DNase I hypersensitive sites (accession numbers H000003 and H000004, respectively) denoted by dark blue colour. Four fields are introduced to describe hypersensitive sites: **HN**, accession number of a site in TRRD; **HS**, name of the hypersensitive site and its location; **HD**, its functional characteristics (erythroid-specific, in this case); and **HR**, bibliographic reference. In addition, 3′ region contains an erythroid-specific enhancer **PR** (accession number P00285), housing 11 transcription factor binding sites (accession numbers 2339–2349), given in yellow coloured blocks in Figure 1.

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; **TY**, 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 realized 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 editing 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

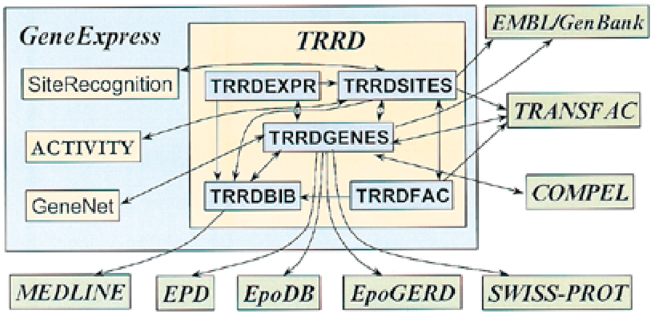


Figure 2. Main tables of TRRD under SRS and the links of the tables to the other databases and computer systems.

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.

TRRD-SRS contains 60 indexed fields providing complicated queries for extraction of information from TRRD.

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%).

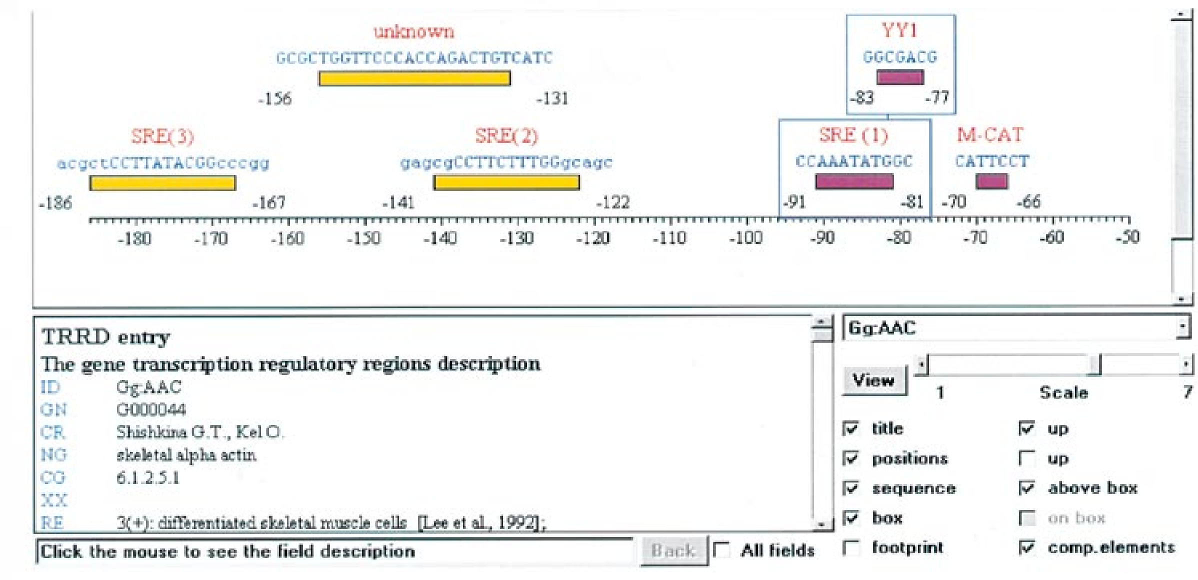
Genes described in TRRD can be subdivided according to their functional role (18–23): interferon-inducible genes (60), erythroid-specific regulated genes (44), genes of lipid metabolism (48), glucocorticoid-controlled genes (35), cell cycle-dependent genes (20), endocrine system genes (41), heat shock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



TRRD entry

The gene transcription regulatory regions description

ID Gg:AAC
GN G000044
CR Shishkina G.T., Kel O.
NG skeletal alpha actin
CG 6.1.2.5.1
XX
RE 3(+): differentiated skeletal muscle cells [Lee et al., 1992];

Click the mouse to see the field description

☐ All fields

Gg:AAC

☑ title ☑ up
☑ positions ☐ up
☑ sequence ☑ above box
☑ box ☐ on box
☐ footprint ☑ comp.elements

Figure 3. An example of the visualization of a gene regulatory map by TRRD Viewer. Boxes represent binding sites for transcription factors. Boxes in rectangles denote composite elements.

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

1 Wingender,E. (1993) *Gene Regulation in Eukariotes.* VCH, Weinheim, Germany.
2 Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E. and Kolchanov,N.A. (1995) *Nucleic Acids Res*., **20**, 4097–4103.
3 Kel,O.V., Kel,A.E., Romashchenko,A.G., Wingender,E. and Kolchanov,N.A. (1997) *Mol. Biol.* (Mosk.), **31**, 498–512.
4 Kel,O.V., Romachenko,A.G., Kel,A.E., Naumochkin,A.N. and Kolchanov,N.A. (1995) *Proceedings of the 28th Annual Hawaii International Conference on System Sciences [HICSS]*, IEE Computer Society Press: Los Alamos, California, pp. 42–51.
5 Kolchanov,N.A. (1997) *Mol. Biol.* (Mosk.), **31**, 581–583.
6 Kel,A.E., Kolchanov,N.A., Kel,O.V., Romaschenko,A.G., Ananko,E.A., Ignatieva,E.V., Merkulova,T.I., Podkolodnaya,O.A., Stepanenko,I.L., Kochetov,A.V. *et al.* (1997) *Mol. Biol.* (Mosk.), **31**, 521–530.
7 Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knuppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) *Nucleic Acids Res*., **25**, 265–268.
8 Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. *et al.* (1998) *Nucleic Acids Res*., **26**, 362–367.
9 Kolchanov,N.A., Ignatieva,E.V., Kel-Margoulis,O.V., Kel,A.E., Ananko,E.A., Podkolodnaya,O.A., Stepanenko,I.L., Merkulova,T.I., Goryachkovsky,T.N., Kolpakov,F.A. *et al.* (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS′98)*, ICG, Novosibirsk, Vol., **1**, pp. 25–28.
10 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,G. (1998) *Nucleic Acids Res*., **26**, 8–15.
11 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
12 Perier,R.C., Junier,T. and Bucher,P. (1998) *Nucleic Acids Res*., **26**, 353–357.
13 Stoeckert,C., Podkolodnaya,O.A., Kel,A.E., Brunk,B., Haas,J., Salas,F., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel,O.V. *et al.* (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure (BGRS′98)*, ICG, Novosibirsk, Vol., **1**, pp. 26–30.
14 Salas,F., Haas,J., Brunk,B., Stoeckert,C.J.,Jr and Overton,G.C. (1998) *Nucleic Acids Res*., **26**, 288–289.
15 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
16 Kolpakov,F.A., Ananko,E.A., Kolesov,G.B. and. Kolchanov,N.A. (1998) *Bioinformatics*, **14**, 529–537.
17 Kolchanov,N.A., Ponomarenko,M.P., Kel A.E., Kondrakhin, Yu. V., Frolov,A.S., Kolpakov,F.A., Kel,O.V., Ananko,E.A., Ignatieva,E.V., Podkolodnaya,O.A. *et al.* (1998*) The Sixth International Conference on Intelligent Systems for Molecular Biology,* 1998 Montreal, Canada, pp. 95–104.
18 Anan’ko,E.A., Bazhan,S.I., Belova,O.E. and Kel’,A.E. (1997) *Mol. Biol.* (Mosk.), **31**, 592–604.
19 Podkolodnaya,O.A. and Stepanenko,I.L. (1997*) Mol. Biol.* (Mosk.), **31**, 562–574.
20 Ignat’eva,E.V., Merkulova,T.I., Vishnevskii,O.V. and Kel’,A.E. (1997*) Mol. Biol.* (Mosk.), **31**, 575–591.
21 Merkulova,T.I., Merkulov,V.M. and Mitina,R.L. (1997) *Mol. Biol.* (Mosk.), **31**, 605–615.
22 Kel’,O.V. and Kel’,A.E. (1997) *Mol. Biol.* (Mosk.), **31**, 548–561.
23 Goryachkovskaya,T.N., Ananko,E.A. and Peltek,S.E. (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure, (BGRS′98)*, ICG, Novosibirsk, Vol., **1**, pp. 50–53.