TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation

E. Wingender*, A. E. Kel[1], O. V. Kel[1], H. Karas, T. Heinemeyer, P. Dietze, R. Knüppel, A. G. Romaschenko[1] and N. A. Kolchanov[1]

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany and [1]Institute of Cytology and Genetics SB RAS, pr. Lavrentyeva-10, 630090 Novosibirsk, Russia

Received October 11, 1996; Accepted October 15, 1996

ABSTRACT

Three databases that provide data on transcriptional regulation are described. TRANSFAC is a database on transcription factors and their DNA binding sites. TRRD (Transcription Regulatory Region Database) collects information about complete regulatory regions, their regulation properties and architecture. COMPEL comprises specific information on composite regulatory elements. Here, we describe the present status of these databases and the first steps towards their federation.

INTRODUCTION

Transcriptional control is one of the most important steps that govern the regulation of gene expression. It is mediated mainly by promoter and enhancer regions, but additionally locus control regions (LCR) and scaffold (or matrix) attachment regions (S/MAR) play a specific role in defining the transcriptional activity of a gene (1–3). To characterize the features of the underlying DNA sequences and to enable identification of regulatory signals in newly unravelled genome sequences, we have established and maintain three databases, each of them focussing on distinct aspects of transcriptional control.

TRANSFAC is a database that started with a list of regulatory *cis*-acting DNA elements and a second table giving some information on the binding proteins (*trans*-acting factors, transcription factors) (4). It was subsequently transferred into a simple database format (5), later into network model-based and relational database systems (6–8). TRRD (Transcription Regulatory Region Database) refers to the hierarchy of regulation levels, starting with the regulatory features of a whole gene at the top level and giving information about individual protein binding sites at the bottom level (9). One intermediary level is represented by composite elements which are characterized by a specific combination of two or more single sites which together constitute a novel module of new quality. This information was made available as a separate database, COMPEL (10).

In the following, we describe the present structure and content of each of these three databases individually as well as reporting on our efforts to promote their federation.

TRANSFAC DATABASE

Data structure

Because of some major changes, release 2.5 was followed by release 3.0 (July 1996), r3.1 coming out in January 1997. As a constant feature, the tables SITE and FACTOR and the n:n relation between them represent the main building blocks. The structure of the periphery of SITE (tables METHOD, CELL) remained unchanged. A new field contains a short designation of the functional gene region a certain site belongs to (e. g., promoter, intronic enhancer). For many site-factor interactions, 'qualities' were newly assigned which reflect the reliability of the experimental evidence (8).

The FACTOR entries were re-structured in that the previous field 'cell specificity' was resolved into two new fields describing cell types/tissue in which the corresponding factor is expressed and others where it could not be detected. Position-dependent structural properties are now given in a separate feature table. Global structural properties are still explained as free text. In addition to cross-references to external sequence databases, release 3.1 will include the complete protein sequences of transcription factors. This has been done to facilitate quick access to, e. g., different splice variants or transcription factors encoded by EMBL/GenBank entries that lack the annotation of the coding region and therefore are not in the protein sequence databases.

The assignment of transcription factors to classes (CLASS table) according to the structure of their DNA-binding domain has been re-evaluated according to a more comprehensive hierarchical classification scheme (11).

The matrix table was supplemented by a new field that gives a short description of the corresponding transcription factor in addition to the linked FACTOR table.

Content of TRANSFAC

The database collects information about *cis*-acting elements and *trans*-acting factors from all eukaryotic organisms. The tables

* To whom correspondence should be addressed. Tel: +49 531 6181 427; Fax: +49 531 6181 266; Email: ewi@gbf-braunschweig.de

SITE, FACTOR, and MATRIX contain entries for vertebrates (60, 73 and 72%, respectively), insects (9/9/12%), higher plants (3/5/3%), fungi (10% in all three tables) and viruses (17/2/2%).

The main focus of the releases 3.0 and 3.1 was to reformat and to update the FACTOR table thus increasing the number of its entries by 14% (total of 1763 entries) in r3.0 (Table 1), and by >23% (>1900 entries) in r3.1. Considering the acquisition of additional information to pre-existing entries as well, the whole volume of FACTOR increased by 38% (release 3.0). The content of the SITE table, on the other hand, remained relatively stable (Table 1). However, it has been supplemented by links to a novel GENE table which serves as a link to TRRD (see below). Most important for sequence scanning purposes, the MATRIX table has been considerably enlarged (246 entries in release 3.0).

Table 1. Content of the TRANSFAC tables

Table	Entries
SITES[a]	4299
FACTORS[b]	1763
CLASS	27
MATRIX	246
CELLS	816
METHODS	52
REFERENCES[c]	4112

[a]The total number to SITES entries decreased because some redundant entries were omitted.
[b]Among the FACTOR entries, 800 are assigned to one of the factor classes.
[c]Total number of articles cited in SITE, FACTOR, CLASS and MATRIX, giving rise to more than 10000 citations.

Connected tools

When accessing TRANSFAC through the World Wide Web (WWW), several search and browsing tools are available to retrieve the data of interest. Additionally, visitors can use a PatternSearch routine to scan own sequences with (groups of) sequence elements contained in the SITE table (Karas *et al*., submitted). The user can choose between element groups for vertebrates, insects, plants, fungi, IUPAC consensus strings or can use all elements. Selected weight matrix libraries have been compiled from the MATRIX table and can be used for sequence analysis with the aid of MatInspector (12). They have been grouped into corresponding categories as the sequence elements.

TRRD (TRANSCRIPTION REGULATORY REGION DATABASE)

Data structure

Presently, the data contained in TRRD are recorded in one flat file. The hierarchical organisation of transcription regulatory regions of eukaryotic genomes is put into the database schema (9). The upper level of the hierarchy is represented by genomic regions that may regulate transcription of several genes (LCR, S/MARs). There are corresponding entries in TRRD that describe the regulatory regions of the upper level. DNA regions (5′ and 3′ regions, regulatory regions in introns) that regulate transcription of a single gene constitute the next hierarchical level. This level is represented by gene entries which constitute the main body of entries in the database. The database entries of these two upper levels include information on DNA regulatory sequences of lower hierarchical levels such as promoters, enhancers, silencers, promoter boxes, binding sites of transcription factors. In addition to this structural information, gene entries contain short descriptions of the regulation specificity and other information on functional classification of the gene.

Content of TRRD

TRRD release 3.3 contains 340 entries. Together, they include information on 474 promoters, enhancers and silencers and 1673 binding sites of transcription factors. The information has been extracted from 1365 publications. Genes of the following organisms have been described in the database: human (138 entries); mouse (97 entries); rat (54 entries); chicken (22 entries); viruses (11 entries); frog (6 entries); rabbit (6 entries); other organisms (7 entries).

Four main types of transcriptional regulation have been distinguished in TRRD: (i) gene expression depends on cell cycle stage (43 genes in TRRD); (ii) gene expression depends on the developmental stage (87 genes); (iii) tissue-specific regulation (277 genes; see Table 2); (iv) gene expression depends on the action of external factors such as: heat shock, starvation, chemicals, cytokines**,** hormones, growth factors, vitamins etc. (259 genes).

Table 2. Tissue-specific regulation of genes described in TRRD

Tissue	Number of tissue-specific genes	Number of repressed genes
liver	98	12
kidney	13	11
pancreas	1	5
placenta	10	1
brain	5	12
blood cells	42	5
muscle	23	4
spleen	6	2
carcinoma	5	3
intestine	8	1

COMPEL

The database COMPEL collects information about composite regulatory elements located in the transcription regulatory regions of vertebrate genes. Composite regulatory elements contain two closely situated binding sites for different transcription factors and regulate transcription in a highly specific manner due to specific DNA–protein and protein–protein interactions (see ref. 10 and references cited therein).

Composite elements can be divided into synergistic and antagonistic ones according to the type of interaction between the transcription factors involved. Within synergistic type composite elements two transcription factors simultaneously bind their target sites and non-additively (synergistically) activate transcription. Within antagonistic type composite elements, two transcrip-

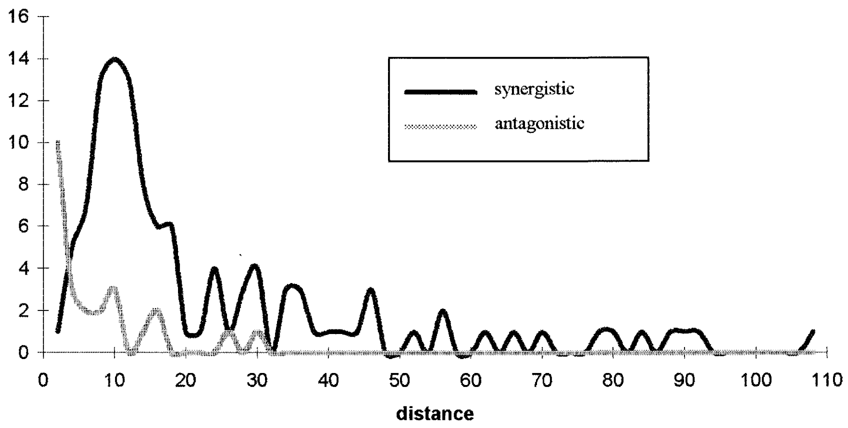

Figure 1. Distribution of distances (base pairs) between two sites in composite elements of synergistic (black curve) and antagonistic type (grey curve).

tion factors influence transcription in opposite directions. The most common mechanism simply involves competitive binding for the same or overlapping binding sites.

Presently, the COMPEL database describes 146 composite elements, among them 118 of synergistic and 28 of antagonistic type. 79 different genes of vertebrates and their viruses are presented.

The distance between binding sites within a composite element can vary significantly. They may be located within each other (13–15), or they may be separated by 80 bp (16). For 80% of composite elements, two binding sites are separated by less then 30 bp (Fig. 1).

Table 3. Functional gene classes in TRANSFAC, TRRD and COMPEL databases

Functional class	Number of genes in GENE table
Structural proteins (histones, ribosomal proteins, actin, myosin, collagen..)	132
Storage and transport proteins (apolipoproteins, ovalbumin, vitellogenin, globins,…)	118
Enzymes	144
Hormones, growth factors, regulatory proteins (insulin, interferon, growth factors, interleukins,…)	145
Proteins related to stress or pathogen defense (immunoglobins, metallothioneins, heatshock proteins,…)	86
Viral genes, transposable elements and retroviruses	71
Unclassified genes	349

FEDERATION OF TRANSFAC, TRRD AND COMPEL

We have developed an integrated GENE table which contains a list of all the genes described in either of the three databases. It contains a functional classification of the genes similar to the system proposed by P. Bucher for the Eukaryotic Promoter Database (EPD; Table 3) (17). Its main function, however, is to provide a basis for the federation of the three databases described here. Thus, it is linked to TRANSFAC SITE entries by a 1:n relation, to TRRD by a 1:1 relation and to COMPEL by a 1:n relation. In future, the GENE table will also provide links to the Eukaryotic Promoter Database (EPD).

Presently, the GENE table contains 1045 genes, among which 215 genes are described in both TRANSFAC and TRRD and 38 gene in all three databases.

The databases are accessible via WWW:
http://transfac.gbf-braunschweig.de or
http://www.bionet.nsc.ru/TRRD
Active links enable the user to shuttle between TRANSFAC SITE, TRRD and COMPEL entries via the GENE table, but COMPEL and TRANSFAC FACTOR are also directly linked. Up to now, the search engines and browsers at either site allow direct access only to the 'domestic' database. Future developments, however, will provide tools for general searches as well.

Users are asked to cite this article when publishing results which have been obtained with the database tools described here.

ACKNOWLEDGEMENTS

We acknowledge the help of R. Appweiler, M. Ashburner, and B. Shomer (EBI, Hinxton) for supporting the mutual database cross-referencing. Part of this work was funded by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (project no. 01 IB 306 A and SX224.6), by the Russian Ministry of Sciences, the Russian National Program 'Human Genome' and the Russian Fundamental Research Foundation (grants N 13241-a and 12757-a) as well as by the North Atlantic Treaty Organization (grant no. 951149).

REFERENCES

1 McKnight, S. L., Yamamoto, K. R. (1992) *Transcriptional Regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
2 Wingender, E. (1993) *Gene Regulation in Eukaryotes*. VCH Weinheim.
3 Bode, J., Schlake, T., Ríos-Ramírez, M., Mielke, C., Stengert, M., Kay, V. and Klehr-Wirth, D. (1995) *Int. Rev. Cytol.* **162A**, 389–454.
4 Wingender, E. (1988) *Nucleic Acids Res.* **16**, 1879–1902.
5 Wingender, E., Heinemeyer, T. and Lincoln, D. (1991) Genome Analysis - From Sequence to Function. In *BioTechForum-Advances in Molecular Genetics* (J. Collins, A. J. Driesel, eds.) **4**, 95–108.
6 Knüppel, R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994) *J. Comput. Biol.* **1**, 191–198.
7 Wingender, E. (1994) *J. Biotechnol.* **35**, 273–280.

8 Wingender, E., Dietze, P., Karas, H. and Knüppel, R. (1996) *Nucleic Acids Res.* **24**, 238–241.

9 Kel O. V., Romachenko A. G., Kel A. E., Naumochkin A. N., Kolchanov N. A. (1995) *Proceedings of the 28th Annual Hawaii International Conference on System Sciences [HICSS]*, Biotechnology Computing, IEE Computer Society Press, Los Alamitos, California, **5**, 42–51.

10 Kel, O. V., Romaschenko, A. G., Kel, A. E., Wingender, E. and Kolchanov, N. A. (1995) *Nucleic Acids Res.* **23**, 4097–4103.

11 Wingender, E., *Molek. Biologiya*, in press.

12 Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) *Nucleic Acids Res.* **23**, 4878–4884.

13 Thanos, D. and Maniatis, T. (1992) *Cell* **71**, 777–789.

14 Lewis, H., Kaszubska, W., DeLamarter, J. F. and Whelan, J. (1994) *Mol. Cell. Biol.* **14**, 5701–5709.

15 Wood, L. D., Farmer, A. A. and Richmond A. (1995) *Nucleic Acids Res.* **23**, 4210–4219.

16 Nerlov, C., De Cesare, D., Pergola, F., Caracciolo, A., Blasi, F., Johnsen, M., and Verde P. (1992) *EMBO J.* **11**, 4573–4582.

17 Bucher, P. and Trifonov, E. N. (1986) *Nucleic Acids Res.* **14**, 10009–10026.