Nucleic Acids Research, 2001, Vol. 29, No. 1 281–283

The TRANSFAC system on gene expression regulation

E. Wingender[1,*], X. Chen[1,2], E. Fricke[3], R. Geffers[3], R. Hehl[4], I. Liebich[1], M. Krull[3], V. Matys[3], H. Michael[1], R. Ohnhäuser[3], M. Prüß[3], F. Schacherer[1], S. Thiele[3] and S. Urbach[3]

[1]Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany, [2]The National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, People's Republic of China, [3]BIOBASE GmbH, Mascheroder Weg 1B, D-38124 Braunschweig, Germany, and [4]Technische Universität Braunschweig, Institut für Genetik–Biozentrum, Spielmannstraße 7, D-38106 Braunschweig, Germany

Received September 28, 2000; Accepted October 12, 2000

ABSTRACT

The TRANSFAC database on transcription factors and their DNA-binding sites and profiles (http://www.gene-regulation.de/) has been quantitatively extended and supplemented by a number of modules. These modules give information about pathologically relevant mutations in regulatory regions and transcription factor genes (PathoDB), scaffold/matrix attached regions (S/MARt DB), signal transduction (TRANSPATH) and gene expression sources (CYTOMER). Altogether, these distinct database modules constitute the TRANSFAC system. They are accompanied by a number of program routines for identifying potential transcription factor binding sites or for localizing individual components in the regulatory network of a cell.

INTRODUCTION

Transcriptional control in eukaryotes is a highly complex subject and, to our present knowledge, is the major step on which the regulation of gene expression is governed. The increasing body of information makes it necessary to provide systematic access to the data about all factors and parameters that influence this process for each individual gene. Starting with our first compilation (1), we provide relevant data about eukaryotic transcription factors, and their genomic binding sites in our TRANSFAC database (2). The database was continuously extended by adding more data sets and by expanding the type of information provided (3–6). The latest step was to complement the TRANSFAC database by additional modules making up now the 'TRANSFAC system' (7). In this contribution, we shall describe the present status of the individual databases of the system as well as the progress of their integration.

THE TRANSFAC DATABASE

Content of TRANSFAC

From the internal relational database system of the TRANSFAC database, which comprises 80 tables, flat file releases are produced where the information is condensed to six files. By far the largest are SITE and FACTOR. They give information about individual transcription factor binding sites in eukaryotic genes and about transcription factors, respectively. Both of these files increased in volume during the last year to different extents (7 and 27%, respectively). The other four files (GENE, CELL, CLASS and MATRIX) provide additional information. Among them, MATRIX is of particular importance since its contents provide the basis for sequence scanning tools such as MatInspector (see below). It represents DNA binding profiles for individual or groups of transcription factors and, therefore, is linked to the FACTOR table.

The total number of entries in the individual tables is given in Table 1. In that table, we also list the distribution of FACTOR entries amongst different species groups. We made particular efforts to update pax (91 entries), homeo domain (457 entries), and forkhead/winged helix factors (92 entries) as well as bHLH-ZIP factors (such as those of the Myc-family; 107 entries). Moreover, yeast transcription factors have been updated and add up to 317 entries now (increase of 66%), and the realm of plant transcription factors has been extended further by 84% to 489 entries.

Quality management

It is essential to standardize the process of data annotation to ensure the consistency and quality of the extracted data. Criteria for the quality of database entries are the absence of errors and redundancy, the completeness, unambigousness and high integration with other data sources and the commentary on existing discrepancies in the literature.

We set up a repository for all quality-relevant information, which comprises the internal database documentation, conventions, syntax rules and format statements. Each step of the annotation process is linked to informatory documents, which set the standards for the implementation of data into TRANSFAC. The checklists contain information for every single field of the entries. The whole internal documentation is being reviewed and updated continuously.

The internal documentation is one part of the quality assurance for TRANSFAC. Besides it, there are several consistency checks, for example automated SQL queries or the adjustment of site sequence data with EMBL.

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

Table 1. Contents of the TRANSFAC database (Release 5.0 September 2000)

Table	Entries
SITE[a]	9009
GENE	1179
FACTOR[b]	3504
Vertebrates	2263
Insects	212
Nematodes	99
Plants	489
Fungi/Yeast	350
Others	91
CLASS	39
MATRIX[c]	374
CELLS	1072
METHOD	69
REFERENCE[d]	7432

[a]Of all SITE entries in release 5.0, the sequences of 5879 sites are accessible without restriction.
[b]Among the FACTOR entries, 2219 are assigned to one of the factor classes.
[c]322 matrices are accessible without restriction.
[d]Total number of articles cited in SITE, FACTOR, CLASS and MATRIX, giving rise to about 24 000 citations.

Connected tools

TRANSFAC has been connected with BLAST to facilitate searching for homologies with transcription factor sequences. The user may apply a protein or a DNA sequence, the latter will be translated into a polypeptide sequence automatically. The tool has been made available as TfBlast and uses BLAST 2.0 from May 26, 2000 (8). In addition to the previously described programs PatSearch (9) and MatInspector (10), we provide the tool AliBaba2 developed by Niels Grabe (11).

PathoDB

A lot of diseases are known to be caused by defects in gene regulation. PathoDB is a new relational TRANSFAC module that provides information about pathologically relevant mutations in transcription factor genes or in their binding sites. Currently, the database stores data about 10 530 transcription factor mutations and about 19 mutated binding sites. Linked to these are tables that hold information about the corresponding phenotypes and diagnostic methods for the underlying genotypes. We have made version 1.0 publicly available. It consists of four flat files (Genotype, MutatedFactor, MutatedSite, Phenotype) that exemplarily provide data about 10 239 p53 mutations, as represented in the p53 database (12), as well as all 19 mutated binding sites.

S/MARt DB

Regulatory domains in a eukaryotic genome are delimited or, at least, functionally influenced by scaffold/matrix attached regions (S/MARs), relatively long sequences in the genomic DNA through which the chromatin is attached to the nuclear matrix structure. The relational S/MARt DB (S/MAR transaction database) presently comprises 313 S/MARs and 61 S/MAR-binding proteins. Thus, the data content of this database module increased by ~38 (S/MARs) or 20% (S/MAR binders), respectively, since its first introduction to the public in October 1999. The number of S/MAR entries that include a DNA sequence has also been enhanced (from 64 to 78% of all S/MARs). Both tables (S/MAR and S/MARBinder) are connected to the TRANSFAC tables GENE and REFERENCES. The data that almost equally refer to animal and plant genomes were extracted mainly from original research reports.

The two flat files that are publicly available provide data about S/MARs and S/MAR binders have the references integrated and can be accessed through a search engine that is similar to that of the TRANSFAC database. SbBlast is a routine connected to S/MARt DB that allows to blast the table of S/MAR binders for similar protein sequences. A detailed description of S/MARt DB and its contents is the subject of another publication (I.Liebich *et al.*, in preparation).

TRANSPATH

Since many transcription factors are regulated in response to extracellular signaling molecules, the signal transduction database TRANSPATH was initiated in 1996. Although it was originally developed as an object-oriented database, it has recently been re-engineered to a relational dabasase management system. Currently, it stores information about 2373 molecules and 2684 reactions taken from 405 references. Thus, its content has been enhanced by a factor of about 20 (molecules) or 30 (reactions) compared to the status reported previously (7).

TRANSPATH is connected with a PathwayBuilder, a routine that exhibits the regulatory cascade aiming at or starting from a user-defined molecule. The depth and cross-connectivity of these pathways can be defined by the user.

CYTOMER®

CYTOMER is a relational database that comprises tables for organs, cell types, physiological systems and developmental stages. The organ table is in itself hierarchically structured, linking anatomical structures to their corresponding sub- and superstructures. The whole organ table comprises presently 5236 entries, 4533 of which are incorporated into the hierarchical system. The other tables list 693 cells, 54 physiological systems and 71 developmental stages. These figures refer mainly to human resources, with a few murine objects attached which are adapted from the mouse vocabulary established by Kaufman (13) and implemented with the Gene Expression Database (GXD; 14). All these terms can be searched, and the transcription factors that are expressed in the corresponding tissues/organs/cells can be displayed. The complete vocabulary of the organ table will be made available to the public.

As a more recent extension, a corresponding compilation of *Caenorhabditis elegans* terms has been initiated. Up to now, it comprises 168 organ entries, all 959 somatic cells including the complete cell lineage, 13 developmental stages and 10 physiological systems. Additionally, we started to establish a plant CYTOMER first for *Arabidopsis thaliana*. Due to the distinct architecture of plants, the structure of this database is

slightly different in having an extra tissue table. In summary, *Arabidopsis* CYTOMER currently comprises entries for 148 organs, 98 tissues, 30 cells, 7 physiological systems and 45 developmental stages.

In each case, a central 'Hub' table is designed to link all biologically existing entries of the other tables. For a given organism, it therefore represents a comprehensive list of all gene expression sources in a spatio-temporal coordinate system and provides a general framework to map expression patterns (15). It is used here to represent the expression patterns of human transcription factors, and to assemble expression profiles for selected organs with regard to the transcription factors they express.

AVAILABILITY

TRANSFAC as well as the other data resources mentioned in this paper are freely available to users from non-profit organizations at http://www.gene-regulation.de/, and some are available at a number of mirror sites. Users from commercial organizations are requested to license database versions either for online use or for inhouse installation. These licensed versions comprise user interfaces of enhanced functionality and data sets that are more frequently updated and extended by proprietary data generated by BIOBASE GmbH. Of course, academic institutions can also license this version.

ACKNOWLEDGEMENTS

The authors are indebted to H. Hermjakob (EBI) for his help in establishing links between TRANSFAC, S/MARt DB and TRANSPATH. We also acknowledge the generous help granted by T. Takai-Igarashi and T. Kaminuma (National Institutes of Health Sciences, Tokyo) in supplying the data set of CSNDB. Finally, we express our gratitude to Mrs A. Bischoff for the technical help in nearly all of the above-mentioned fields. Parts of this work was supported by the German Ministry of Education and Research (BMBF, grant nos 01 KW 9629/7, 01 KW 9906 and 01SF9988/4), by a Scientific-Technical cooperation grant from BMBF (X224.6) and by a grant from the European Commission (QLRI-CT1999-01333).

REFERENCES

1. Wingender,E. (1988) Compilation of transcription regulating proteins. *Nucleic Acids Res.*, **16**, 1879–1902.
2. Wingender,E., Heinemeyer,T. and Lincoln,D. (1991) Regulatory DNA sequences: predictability of their function. In Collins,J and Driesel,A.J. (eds), *Genome Analysis – From Sequence to Function; BioTechForum – Advances in Molecular Genetics*, **4**, 95–108.
3. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
4. Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knüppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. *Nucleic Acids Res.*, **25**, 265–268.
5. Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) Databases on Transcriptional Regulation: TRANSFAC, TRRD, and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.
6. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.*, **27**, 318–322.
7. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Wingender,E., Karas,H. and Knüppel,R. (1996) TRANSFAC Database as a Bridge between Sequence Data Libraries and Biological Function. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing '97 (PSB'97)*. World Scientific, Singapore, pp. 477–485.
10. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector – New fast and sensitive tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
11. Grabe,N. (2000) AliBaba2: Context Specific Identification of Transcription Factor Binding Sites. *In Silico Biol.*, **1**, 0019.
12. Hainaut,P., Hernandez,T., Robinson,A., Rodriguez-Tome,P., Flores,T., Hollstein,M., Harris,C.C., Montesano,R. (1998) IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. *Nucleic Acids Res.*, **26**, 205–213.
13. Kaufman,M.H. (1992) *The Atlas of the Mouse Development.* Academic Press, London.
14. Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (1999) GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. *Nucleic Acids Res.*, **27**, 106–112. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 98–101.
15. Chen,X., Dress,A., Karas,H., Reuter,I. and Wingender,E. (1999) A database framework for mapping expression patterns. In Giegerich,R. Hofestädt,R., Lengauer,T., Mewes,W. Schomburg,D. Vingron,M. and Wingender,E. (eds), *Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB '99)*. GBF-Braunschweig and University of Bielefeld, pp. 174–178.