316–319 Nucleic Acids Research, 2000, Vol. 28, No. 1

TRANSFAC: an integrated system for gene expression regulation

E. Wingender[1,*], X. Chen[1,2], R. Hehl[3], H. Karas[4], I. Liebich[1], V. Matys[1,4], T. Meinhardt[1], M. Prüß[1,4], I. Reuter[1] and F. Schacherer[1]

[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]Technische Universität Braunschweig, Biozentrum, D-39100 Braunschweig, Germany and [4]BIOBASE GmbH, Mascheroder Weg 1B, D-38124 Braunschweig, Germany

Received September 30, 1999; Accepted October 7, 1999

ABSTRACT

TRANSFAC is a database on transcription factors, their genomic binding sites and DNA-binding profiles (http://transfac.gbf.de/TRANSFAC/). Its content has been enhanced, in particular by information about training sequences used for the construction of nucleotide matrices as well as by data on plant sites and factors. Moreover, TRANSFAC has been extended by two new modules: PathoDB provides data on pathologically relevant mutations in regulatory regions and transcription factor genes, whereas S/MARt DB compiles features of scaffold/matrix attached regions (S/MARs) and the proteins binding to them. Additionally, the databases TRANSPATH, about signal transduction, and CYTOMER, about organs and cell types, have been extended and are increasingly integrated with the TRANSFAC data sources.

INTRODUCTION

Gene regulation is still a major issue of molecular biology, and several new methodologies enable us to globally characterize gene expression patterns and profiles at a phenomenological level. However, the number of organs/cell types, developmental stages and conditional factors is so huge that we cannot expect all combinations of them to be exhaustively characterized by specific experimental setups. Therefore, there is a great need for comprehensive tools for the *in silico* identification of genomic signals that govern gene regulation events, and this requirement increases with the number of genomes that have been sequenced completely and are now to be exploited for biotechnological, pharmaceutical, agricultural or other purposes.

For these reasons, more than a decade ago we began to compile information about *cis*-regulatory DNA sequences and *trans*-acting factors (1). This compilation was transferred into a computer-readable format and released as an EMBL-like ASCII flat-file for public use as TRANSFAC database (2). Since then, it has been reorganized as a hierarchical and, subsequently, a relational database (3,4), increasingly linked with other databases allowing its integration into DBGET (5) and SRS (6,7). A major breakthrough was achieved by linking it with TRRD (Transcription Regulatory Region Database) (8) and by connecting the database COMPEL on composite regulatory elements (9). Additional database modules have been developed more recently, devoted to signal transduction pathways and organs/cell types (10). In the present contribution, the status of these modules along with that of TRANSFAC will be outlined, and two additional TRANSFAC supplements will be described that provide data on pathologically relevant gene regulatory components (PathoDB) and on scaffold/matrix attached regions, S/MAR (S/MARt DB).

THE TRANSFAC DATABASE

Content of TRANSFAC

The underlying relational database system internally comprises two main tables, SITE and FACTOR. Nearly all other tables (a total of 50 tables) are linked directly or indirectly to one or both (such as REFERENCES) of these two. Of particular importance is the MATRIX table which represents DNA binding profiles for individual or groups of transcription factors and, therefore, is properly linked to the FACTOR table. The data structure of the relational system has changed slightly in that an additional table has been introduced linking MATRIX entries with their corresponding training sequences. These sequences have been incorporated into the SITE table (10), and are now explicitly listed in the individual MATRIX entries of the flat-file version as well.

The total number of entries in the individual tables are summarized in Table 1. Special efforts have been undertaken to improve the data content for the plant kingdom. Thus, the number of plant FACTOR entries has been trebled, amounting now to 266 entries (as of September 1999). Comparing this figure to the 136 plant (Viridiplantae) entries of the SWISS-PROT database exhibiting the keyword 'Transcription regulation' may indicate a satisfying coverage rate. However, this comparison should consider that, on the one hand, TRANSFAC may apply slightly more restrictive criteria to include a protein as a 'transcription factor', but that, on the other hand, individual splice variants are represented in TRANSFAC as distinct entries, in contrast to SWISS-PROT; also, multimeric transcription factor

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

complexes are described as own FACTOR entries in TRANSFAC, but of course not in SWISS-PROT.

Table 1. Contents of the TRANSFAC database (status as of September 1999)

Table	Entries
SITE	8390
GENE	1302
FACTOR[a]	2765
CLASS	38
MATRIX	356
CELLS	978
METHOD	67
REFERENCE[b]	6570

[a]Among the FACTOR entries, 1596 are assigned to one of the factor classes.
[b]Total number of articles cited in SITE, FACTOR, CLASS and MATRIX, giving rise to >21 000 citations.

The volume of binding sites for plant transcription factors in the SITE table has been more than doubled, now comprising 104 genomic and 208 artificial binding sites. The latter come mostly from random selection studies to identify the DNA-binding specificity of individual factors and are used to generate binding profiles. Correspondingly, the number of plant transcription factor binding profiles in the MATRIX table has been more than doubled (from 8 to 19 matrices; status as of September 1999). These matrices, together with a total of 18 consensus strings stored in the SITE table, can serve as a basis to identify potential plant transcription factor binding sites in genomic sequences using the search routines provided by the TRANSFAC server (see below). In addition to this more quantitative increase, we paid specific attention to improve the quality of the individual entries since part of this information is used by other databases as well (11).

Cross-links with other databases

The links to the previously listed external databases have been maintained and extended. Links between the CELL table and HyperCLDB (Cell Line Data Base; http://www.biotech.ist.unige.it/cldb/indexes.html), a general collection of information about cell lines comprising data of the American Type Culture Collection (ATCC), the German Collection of Microorganisms and Cell Cultures (DSMZ), and seven additional sources have been newly introduced. Up to now, the URLs of CLDB have been included directly in the commentary/description field of 154 CELL entries. Presently, attempts are being made to include links to GeneCard into the GENE table. To facilitate linking of transcription factor entries to their corresponding genes in GeneCard, all transcription factor genes will be included in the GENE table of TRANSFAC, even if no transcription factor binding sites in their regulatory regions have been described yet.

Classification of transcription factors

Presently (last update May 1999), the overall classification scheme for transcription factors provides links to 1160 factors, embedded in a system of 1321 taxa (see refs 10,12 for a more detailed description). Recent changes comprise the introduction of APETALA2/EREBP-related factors and Dof (DNA-binding with One Finger) factors from plants as new classes.

Connected tools

As has already been described, TRANSFAC data are used by two programs that scan DNA sequences for potential transcription factor binding sites. PatSearch uses the sequence information of the SITE table grouped into six libraries: genomic sites from vertebrates, insects, plants, fungi, consensus strings in the 15-letter IUPAC code, or site information from TRRD, whereas MatInspector (13) works with compiled libraries of binding profiles (for factors from vertebrates, insects, plants, fungi or miscellaneous organisms) which is derived mainly from the TRANSFAC MATRIX table. Likewise, the FastM module developed by Frech *et al*. (14) uses the same matrix libraries while searching for user-defined combinations of transcription factor binding sites. As a new functionality, the output list of MatInspector can be used to suggest expression patterns of the gene analyzed (see below, CYTOMER).

Another program (S_Comp 1.0) designed to specifically detect composite elements of the NFAT type has been made available on the TRANSFAC server and has been described in greater detail elsewhere (http://transfac.gbf.de/dbsearch/funsitep/s_comp.html) (15).

PathoDB

Numerous cases have been published showing that defective transcription factors or transcription factor binding sites lead to pathological defects because normal gene regulation is impaired. It is the aim of the newly developed database PathoDB to collect such data on mutated factors and binding sites. Moreover, the underlying genetic defects and the resulting diseases are considered as well. PathoDB is intended to be an extension of the TRANSFAC database system concerning the pathological aspects of transcriptional regulation.

Structure of the database

PathoDB is a relational database, consisting of 50 linked tables. The database schema was designed with respect to the various interdependencies of entry types. For instance, one mutated gene may encode several mutated proteins due to alternative splicing, and a certain genotype may cause distinct phenotypes depending on ploidity. On the other hand, different genotypes can cause the same pathological phenotype by knocking out the same gene through distinct mechanisms.

The database is composed of 10 main tables and another 40 link tables. The four most important tables contain data about mutated transcription factors (MuFactor), mutated DNA binding sites (MuSite), molecular gene structures (Genotype) and regulatory disorders (Phenotype). In order to access data beyond PathoDB's primary focus, the Phenotype and Genotype entries are connected to important external databases like OMIM (http://www.ncbi.nlm.nih.gov/omim/) (16), MGI (http://www.informatics.jax.org/) (17) and HGMD (http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html?) (18).

Current content and perspective

Presently, the prototype of PathoDB contains detailed information on 80 mutated transcription factors and 20 mutated binding sites (each with a link to the wild-type specified in TRANSFAC), ~100 genotypes and 15 phenotypes of specific diseases. In the current state, mainly developmental defects are considered. In the case of factors, for example, both the mutated pituitary-specific positive transcription factor 1 (Pit-1) and the mutated Prophet of Pit-1 (PROP1) lead to impaired pituitary development (19,20), and the mutated paired-box proteins 3 and 8 (Pax-3, Pax-8) cause early neural tube defects or congenital hypothyroidism, respectively (21,22).

The range of organisms admitted to PathoDB is planned to be nearly as broad as it is in TRANSFAC. However, at the moment only human and murine defects are considered, with special emphasis on the human organism whose genetic diseases might be of greatest interest for medical research. In particular, the wide cancer field will be regarded. The database will soon be accessible via the Internet as a flat-file version.

S/MARt DB

Transcriptional activation of eukaryotic genes is associated with significant changes in chromatin structure. The main change is the transition from a condensed to an 'open' or 'active' structure. It has been suggested that the nuclear matrix may play a role in gene potentiation as well as in genome organization. These functions are believed to be mediated by scaffold or matrix attached regions (S/MARs). Therefore we developed a new database, S/MARt DB (scaffold/matrix attached region transaction database), which is closely linked to TRANSFAC. This database collects information about S/MARs and the nuclear matrix proteins that are supposed to be involved in the interaction of these elements with the nuclear matrix. S/MARt DB is publicly available through the WWW at http://transfac.gbf.de/SMARtDB/index.html . A detailed description of S/MARt DB is the subject of another publication (Liebich,I., Bode,J., Frisch,M., Reuter,I. and Wingender,E., manuscript in preparation).

TRANSPATH

TRANSPATH (http://transfac.gbf.de/TRANSPATH/) focuses on signal transduction networks involved in the regulation of transcription factors and aims at furnishing a collection of data usable for simulation of network dynamics. In our model the signaling network is composed of components like receptors, enzymes, transcription factors and genes, all of which are connected through reactions. While mechanistic reactions represent the minute physical interactions between components needed for simulation, semantic reactions portray the flow of meaning, like 'activation', as usually shown in the literature. Components can be clustered together in families which show similar signaling behavior to reduce the level of redundancy. Reactions can be clustered into pathways. For any signaling component the user can search an interactive graphical tree representation of the connected pathways. The search can take family information into account to broaden the result range. TRANSPATH was developed with a semantic dataset (1514 components and 827 reactions) taken from CSNDB (Cell Signaling Network Database) (23,24). Since we aim at a finer granularity of reactions to make simulation feasible, a second dataset retrieval has been started, which also includes mechanistic reactions (status as of September 1999, 120 components and 80 reactions on top of a set of 10 073 components imported from SWISS-PROT). Nevertheless, the TRANSPATH interface can also be used to view a subset of CSNDB data.

CYTOMER®

The relational database CYTOMER comprises tables for human organs, cell types, physiological systems and developmental stages (10). The organ table is in itself hierarchically structured, since to each anatomical structure listed, the parent (sub-)organ is indicated as an additional attribute. All four tables are linked through a central 'Hub' table that lists the biologically meaningful combinations of these four categories. This table can provide a general framework to map expression patterns (Chen,X., Dress,A. and Wingender,E., manuscript in preparation) which 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. The expression patterns of individual transcription factors can be called from those TRANSFAC FACTOR entries that contain information about expression and non-expression sources within the CP and/or CN line, respectively.

Starting with a list of transcription factors, their expression patterns can be exhibited in a comparative manner to evaluate visually in which organs all or most of them may be expressed. Using this function, an output list such as that generated by MatInspector (see above) will be used to optionally display a table with expression patterns of the potential regulators of the analyzed gene and, thus, suggesting a possible expression pattern of this gene itself.

Using the mouse vocabulary established by Kaufman (25) and implemented with the Gene Expression Database (GXD) (26), the CYTOMER database will be extended to the mouse system and, in the near future, to other intensely studied 'model' organisms as well.

AVAILABILITY

TRANSFAC as well as the other data resources mentioned in this paper are freely available to users from non-profit organizations under http://transfac.gbf.de/TRANSFAC/ and at a number of mirror sites. Users from commercial organizations are requested to license database versions with user interfaces of enhanced functionality and with a data set that has been enlarged mainly by additional data on artificial binding sites and transcription factor DNA-binding profiles. Of course, academic institutions can also license this version.

ACKNOWLEDGEMENTS

The authors are indebted to M. Ashburner (EBI) for regularly providing the links to FlyBase. We also gladly 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 the above-mentioned fields. Parts of this work were supported by the

German Ministry of Education and Research (BMBF, grants no. 0311640 and 01 KW 9629/7), and by a Scientific-Technical cooperation grant of BMBF (CHN-305-97).

REFERENCES

1. Wingender,E. (1988) *Nucleic Acids Res.*, **16**, 1879–1902.
2. Wingender,E., Heinemeyer,T. and Lincoln,D. (1991) In Collins,J. and Driesel,A.J. (eds), *Genome Analysis—From Sequence to Function; BioTechForu—Advances in Molecular Genetics*. Hüthig Buch Verlag, Heidelberg, Volume 4, pp, 95–108.
3. Knüppel,R., Dietze,P., Lehnberg,W., Frech,K. and Wingender,E. (1994) *J. Comput. Biol.*, **1**, 191–198.
4. Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) *Nucleic Acids Res.*, **24**, 238–241.
5. Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pacific Symp. Biocomput.*, **3**, 681–692.
6. Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.
7. Etzold,T. and Verde,G. (1997) *Pacific Symp. Biocomput.*, **2**, 134–141.
8. 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], Biotechnology Computing*. Volume 5, IEE Computer Society Press, Los Alamitos, CA. pp. 42–51.
9. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E. and Kolchanov,N.A. (1995) *Nucleic Acids Res.*, **23**, 4097–4103.
10. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) *Nucleic Acids Res.*, **27**, 318–322.
11. Rombauts,S., Déhais,P., Van Montagu,M. and Rouzé,P. (1999) *Nucleic Acids Res.*, **27**, 295–296.
12. Wingender,E. (1997) *Mol. Biol. Engl. Tr.*, **31**, 483–497.
13. Quandt,K., Frech,K., Karas,H.,Wingender,E. and Werner,T. (1995) *Nucleic Acids Res.*, **23**, 4878–4884.
14. Frech,K., Danescu-Mayer,J. and Werner,T. (1997) *J. Mol. Biol.*, **270**, 674–687.
15. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) *J. Mol. Biol.*, **288**, 353–376.
16. McKusick,V.A. (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*, 12th Edition. Johns Hopkins University Press, Baltimore, MD.
17. Blake,J.A., Richardson,J.E., Davisson,M.T. and Eppig,J.T. (1999) *Nucleic Acids Res.*, **27**, 95–98. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 108–111.
18. Cooper,D.N., Ball,E.V. and Krawczak,M. (1998) *Nucleic Acids Res.* **26**, 285–287.
19. Li,S., Crenshaw,E.B.,III, Rawson,E.J., Simmons,D.M., Swanson,L.W. and Rosenfeld,M.G. (1990) *Nature*, **347**, 528–533.
20. Deladoey,J., Flück,C., Büyükgebiz,A., Kuhlmann,B.V., Eblé,A., Hindmarsh,P.C., Wu,W. and Mullis,P.E. (1999) *J. Clin. Endocrinol. Metab.*, **84**, 1645–1650.
21. Fortin,A.S., Underhill,D.A. and Gros,P. (1997) *Hum. Mol. Genet.*, **6**, 1781–1790.
22. Macchia,P.E., Lapi,P., Krude,H., Pirro,M.T., Missero,C., Chiovato,L., Souabni,A. and Baserga,M. (1998) *Nature Genet.*, **19**, 83–86.
23. Takai-Igarashi,T., Nadaoka,Y. and Kaminuma,T. (1998) *J. Comput. Biol.*, **5**, 747–754.
24. Takai-Igarashi,T. and Kaminuma,T. (1998) *In Silico Biol.*, **1**, 0012.
25. Kaufman,M.H. (1992) *The Atlas of Mouse Development.* Academic Press, London, UK.
26. Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (1999) *Nucleic Acids Res.*, **27**, 106–112. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 115–119.