238–241 *Nucleic Acids Research, 1996, Vol. 24, No. 1*

TRANSFAC: a database on transcription factors and their DNA binding sites

E. Wingender*, P. Dietze, H. Karas and R. Knüppel

Gesellschaft für Biotechnologische Forschung mbH, Department of Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Received September 5, 1995; Accepted October 2, 1995

ABSTRACT

TRANSFAC is a database about eukaryotic transcription regulating DNA sequence elements and the transcription factors binding to and acting through them. This report summarizes the present status of this database and accompanying retrieval tools.

INTRODUCTION

To render raw genomic sequence data into usable biological information requires a lot of experimental work and, thus, depends on additional data. However, there are efforts to circumvent this bottleneck of functional sequence analysis by developing sophisticated computational tools that allow to deduce biological function from mere DNA sequences. The function of genes is to code for specific products, but to know them and to predict their putative biological role is only half of the whole task. The other is to decipher the regulatory code, i.e. to disclose under which conditions this genomic information is expressed. Thus, studying gene expression mechanisms is one of the major tasks of todays molecular biology, giving an enormous and still increasing output of data. This amount of information can only be handled by storing it in an appropriate database system.

TRANSFAC is a database that collects data which are relevant for gene expression at the transcriptional level. Very early, several collections have been published describing transcription factors and the sequences they interact with (1–4). Our own attempts started from a simple compilation of *cis*-acting DNA elements and the proteins (transcription factors) binding to them in two tables (5). From this it was obvious that this kind of data can be optimally managed by a relational model database as was realized for the first time for the Transcription Factor Database (TFD) by D. Ghosh (6,7). For TRANSFAC, this was implemented after an electronically readable ASCII flat file version had been established (8,9). We now present a survey about the progress of the TRANSFAC database, its content as well as the database management systems (DBMSs) which are presently available.

STRUCTURE OF THE DATABASE

The basic mechanism of transcriptional control operates through sequence-specific interactions of a special class of proteins, the transcription factors, with relatively short DNA elements of ~5–25 bp (10). When transferring this knowledge into an appropriate relational data model, information about the two basic constituents appears in two tables, SITES and FACTORS (Fig. 1). They are connected by a many-to-many relation since many sites can interact with several factors, and all known factors bind to more than just one site. The SITES table gives the position of a particular regulatory site, the gene this site belongs to, the biological species this gene has been derived from, and as a free text field some additional, unstructured information such as dissociation constants or inducibility by certain agents is included. The methods by which each regulatory site has been identified are connected since they give a hint on the reliability of this characterization and on the stringency of the sequence displayed. This is also reflected by an assigned 'Quality' parameter (see below). Moreover, published regulatory sites have been identified as being functional in a specific cellular context, therefore information about the cell lines used is given as well. These data are stored in two separate tables, METHODS and CELLS, both of which are linked to SITES (Fig. 1). Similarly, the sequences of the sites are stored in a separate connected table, since some sites may comprise more than one binding sequence, depending on the methods applied. Wherever possible, each sequence is individually linked to the EMBL data library. It has been inserted for users who are interested in the sequence context of a single site. These properties and links are only partially applicable to synthetic sequences and consensus sequences in IUPAC nomenclature which are also part of the SITES table.

The FACTORS table describes the properties of individual transcription factors: the biological species they have been obtained from, any known cell specificities, and data about their size, structural and functional properties. The content of the two latter fields is unstructured information which is stored as plain text. Connected to FACTORS are tables with the synonyms and with interacting factors. The latter is important since most transcription factors act as homo- or heterodimers, the kind of partner frequently determining the biological effect.

We have started to assign 'qualities' to the DNA–protein interactions described in TRANSFAC. They define matrix descriptions according to the stringency of the underlying experimental data. e.g., the least quality of 5 is given to mere *bona fide* elements, while the highest quality of 1 is assigned to protein–DNA contacts for which binding of pure factor has been demonstrated *in vitro* and for which functional importance has been demonstrated, e.g. by

* To whom correspondence should be addressed

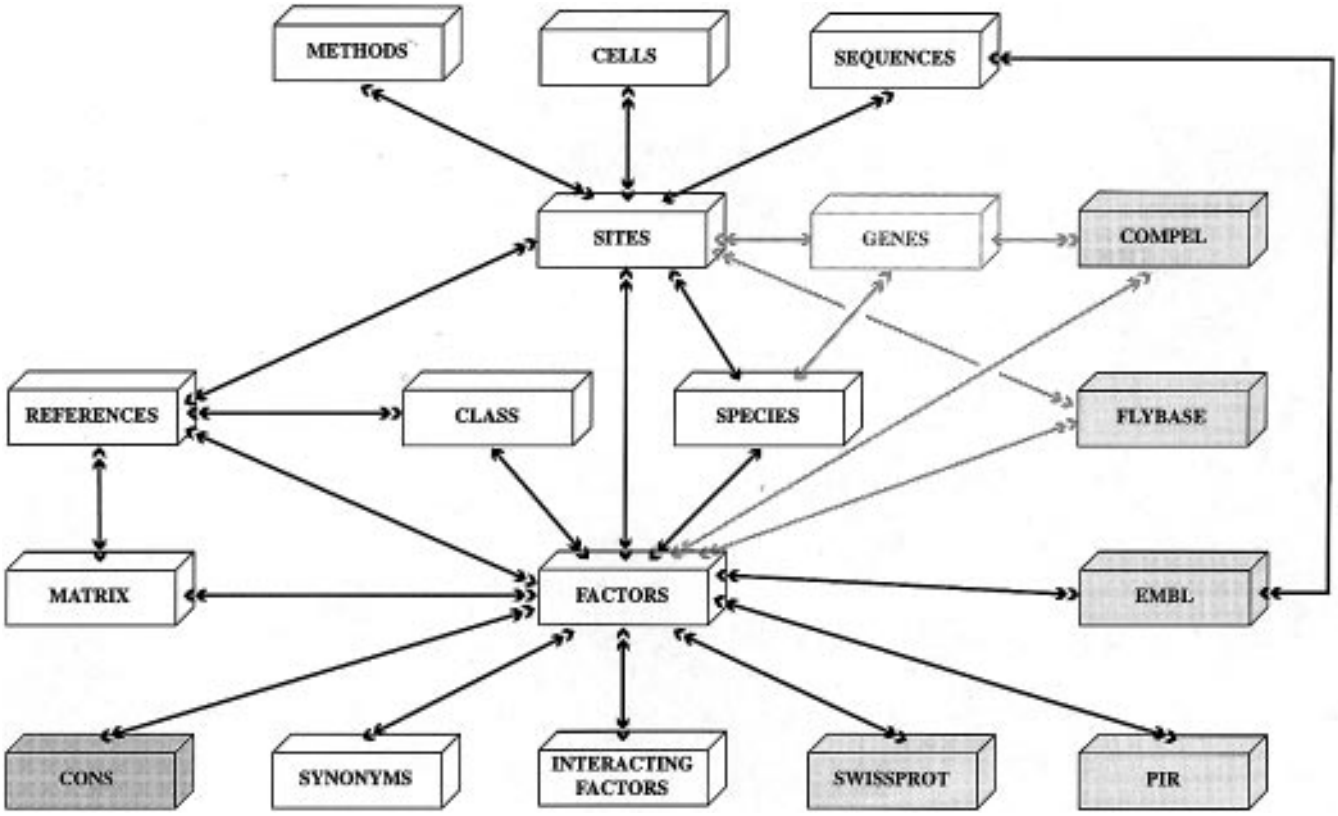

Figure 1. Structure of the TRANSFAC database. Shown is the basic schema of the relational model. Double arrowheads on either side indicate many-to-many relations, one single and one double arrowhead represent one-to-many relations. Black arrows indicate relations which are implemented in release 2.4, those in light grey will appear in release 2.5. The tables indicated by white boxes are genuine TRANSFAC tables, the dark boxes represent tables with externally generated data or external databases.

cotransfection studies of the element in front of a reporter gene and an expression construct of the factor. These 'qualities' are included in the linking table between SITES and FACTORS.

Additionally, cumulating tables have been placed above both SITES and FACTORS. Thus, in most transcription factors known and cloned up to now, distinct DNA-binding motifs such as basic domain-leucine zipper (bZIP) or ets-domains can be detected. Information about the blueprints of these domains is contained within the CLASS table to which many factors entries have now been linked. Each genomic site within the SITES table can be assigned to a gene and therefore has been linked to an entry of the new GENE table (appearing first in release 2.5). In future, it will provide a link to TRRD (Transcription Regulatory Region Database). This database is presently under development at the Institute of Cytology and Genetics, Novosibirsk, Russia, and will be available in the near future. It provides information about the global features of regulatory regions at the different levels of their organization (whole gene, promoters/enhancers, composite elements, individual sites) (11). In fact, the GENES table will be common to both databases.

Also connected to FACTORS are tables with primary as well as secondary data describing the DNA-binding profiles of transcription factors. The MATRIX table comprises nucleotide distribution matrices which may have been obtained by random selection and amplification procedures (primary data) or which have been constructed from compiled sites (secondary data). Some matrices have been derived from aligned TRANSFAC sequences, classified by the (least) quality of the elements used (realized for the first time in release 2.5). These matrices can be used for scanning genomic sequences for potential transcription factor binding sites with higher reliability than simple IUPAC string searches (12). Similarly, the CONS table contains consensus descriptions as they are constructed by the program ConsIndex and can be used by the ConsInspector routine for sequence analysis (13).

In the relational model, the literature references are provided by a separate table which is connected to SITES, FACTORS, CLASS, and MATRIX (Fig. 1). The content of the individual tables is shown in Table 1.

The most commonly used version are the widely distributed ASCII flat files which are organized in a manner similar to the EMBL data library. In contrast to the relational model shown in Figure 1, they comprise only five files harboring the data about sites, factors, class, matrices and cells. For the sake of clarity, the references had to be included into these tables, although this produces considerable redundancy. Moreover, the sequences of the sites and the methods which led to their identification have been integrated into the SITES file, and the cross-references to external databases are displayed in the DR lines of SITES and FACTORS. The cross-links between the different TRANSFAC tables are included in the form of the accession numbers and identifier of the linked entries.

Table 1. Content of the TRANSFAC tables

Table	Entries
SITES[a]	4304
FACTORS[b]	1544
CLASS	27
MATRIX	169
CELLS	816
METHODS	52
REFERENCES[c]	3130

[a]These sites comprise 4042 sequences, among them are 316 artificial (synthetic) sequences and 224 entries with 239 consensus sequences in IUPAC nomenclature; there are 4055 assignments to entries of the EMBL data library.
[b]610 factor entries possess a CLASS assignment; moreover, there are 1014 links to the EMBL dl, 520 to SwissProt, and 514 to PIR.
[c]Only references that are linked to either SITES, FACTORS, CLASS or MATRIX have been counted.

The WWW version has been derived from the ASCII flat files. The main difference is that the references are not included but are connected through active hypertext links. The interconnections between TRANSFAC tables can likewise be activated as is the case for links to entries of the EMBL data library, SwissProt database and PROSITE database and document entries.

DB interconnectivity

TRANSFAC contains cross-references to the EMBL data library and to the SwissProt database (name and accession number). Both databases have inserted pointers to TRANSFAC entries as well. Mutual cross-referencing has also been established with Flybase entries, either from SITES to regulatory regions of genes or from FACTORS to genes encoding them. These references will be included first in TRANSFAC release 2.5. Where appropriate, FACTORS entries are also linked to PIR, and most CLASS entries point to the PROSITE database.

It has been pointed out above that the GENES table provides a link to TRRD. The second level in the regulatory hierarchy displayed by TRRD are composite elements, which are synergistically or antagonistically acting combinations of single elements revealing a novel regulatory quality. Data about these kind of elements are provided by the database COMPEL (11). To enable retrieval of information about the transcription factors binding to these composite elements, COMPEL has been linked to FACTORS.

Browsing tools for the TRANSFAC database

In addition to the Transfac Retrieval Programs (TRPs, see below and ref. 14), we have developed Tiny TRP as a tool that enables the user to browse the various TRANSFAC ASCII flat files under Windows. The links between the tables are modeled as hypertext links. Tiny TRP also allows some very basic searches in the tables using an index system which is similar to that of the EMBL data library. We intend also to implement access to connected entries of the external databases.

The sequence elements compiled by the SITES table are also available in a format to use them with the findpattern program of the GCG program package. Five separate tables are provided that comprise the consensus sequences in IUPAC designation, artificial sequences, sites from arthropoda genes, the plant and fungi elements, or all the rest (mainly vertebrates). To facilitate retrieval of additional information about the TRANSFAC sites which matched, we developed the program ‘tfex’. It enables the user to retrieve the complete SITES entries or a set of selected fields of these entries according to their ID which appear in the findpattern output.

The tools mentioned operate with the ASCII flat files. To provide the user with the full functionality of a relational database, we developed some database management systems (DBMS) which are freely available as well. They have been designated TRANSFAC Retrieval Programs, TRPs. One of them, a network model DBMS (TRPrai), has been described previously (14). It runs under MS DOS and VMS. A relational model DBMS, TRPox, has been developed using Paradox™ (Borland) running under Windows. A third, and maintained, version, TRPro, has been established using FoxPro™ (Microsoft); here, PC Windows and MacIntosh versions are available. They provide full SQL functionality for complex retrieval problems.

Up to now, only the TRPrai version offers connection to the EMBL data library and to the SwissProt database enabling the user to retrieve complete entries of these databases starting from any TRANSFAC entry. Moreover, the sequence analysis program ConsInspector (13) can be invoked. TRPox includes a simple pattern search routine which also uses degenerate IUPAC strings. The results are graphically displayed and listed optionally.

FUTURE PROSPECTS

There are some extensions of the database which have to be done in the future. First of all, enhanced cross-referencing between databases is required to construct a well-working net of information resources for molecular biologists. For instance, pointers to EPD have to be included into TRANSFAC.

Some other changes, however, will require a change in its structure. To improve the access to transcription factor data, we shall abstract relevant information from biological species redundancy in a ‘factors summary table’ where data about obviously identical (or strictly orthologous) factors are accumulated. This will be the first step towards a systematic factor classification we are preparing.

Finally, we intend to link the relational and the WWW version of the database to a more comprehensive sequence analysis program package which includes pattern matching routines (involving genomic sites as well as IUPAC consensus strings), matrix search routines (12) and the ConsInspector program (13).

ACCESS TO THE TRANSFAC DATABASE

The TRANSFAC database is accessible as ASCII flat file. It is distributed in this format on the CD ROM of the EBI along with the EMBL data library or can be downloaded by anonymous ftp from ftp.gbf-braunschweig.de (IP 193.175.244.2) or from ftp.ebi.ac.uk. Moreover, there is access through WWW either directly (http://transfac.gbf-braunschweig.de) or through database access networks such as SRS (http://www.embl-heidelberg.de/srs/srsc) (15,16) or WebDBGET (http://www.genome.ad.jp/htbin/bfind_transfac). The browsing tools can also be obtained at the above-mentioned ftp site.

ACKNOWLEDGEMENTS

The authors wish to thank J. Collins for his continuous support, as well as T. Werner, K. Frech, K. Quandt, N. Kolchanov, A. Kel. and O. Kel for critical discussions and many helpful comments. Moreover, we are indebted to our colleagues at the EBI, in particular B. Shomer, R. Apweiler, and M. Ashburner for their support in mutual cross-referencing the databases. We also thank Borland International, Inc. for making available to us the runtime module of Paradox for Windows 4.5. This work was funded by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (Project No. 01 IB 306 A).

REFERENCES

1 Johnson, P. F., and McKnight, S. L. (1989) *Annu. Rev. Biochem.* **58**, 799–839.
2 Latchman, D. S. (1990) *Biochem. J.* **270**, 281–289.
3 Locker, J., and Buzard, G. (1990) *DNA Sequence* **1**, 3–11.
4 Faisst, S., and Meyer, S. (1992) *Nucleic Acids Res.* **20**, 3–26.
5 Wingender, E. (1988) *Nucleic Acids Res.* **16**, 1879–1902.
6 Ghosh, D. (1990) *Nucleic Acids Res.* **18**, 1749–1756.
7 Ghosh, D. (1993) *Nucleic Acids Res.* **21**, 3117–3118.
8 Wingender, E., Heinemeyer, T., and Lincoln, D. (1991) Genome Analysis—From Sequence to Function in *BioTechForum—Advances in Molecular Genetics* (J. Collins and A.J. Driesel, eds) Vol. 4, 95–108.
9 Wingender, E. (1994) *J. Biotechnol.* **35**, 273–280.
10 Wingender, E. (1993) *Gene Regulation in Eukaryotes.* VCH Weinheim.
11 Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E., and Kolchanov, N.A. (1995) *Nucleic Acids Res.* **23**, 4097–4103.
12 Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) *Nucleic Acids Res.* **23**, 4878–4884.
13 Frech, K., Herrmann, G., and Werner, T. (1993) *Nucleic Acids Res.* **21**, 1655–1664.
14 Knüppel, R., Dietze, P., Lehnberg, W., Frech, K., and Wingender, E. (1994) *J. Comput. Biol.* **1**, 191–198.
15 Etzold, T., and Argos, P. (1993) *Comput. Appl. Biosci.* **9**, 49–57.
16 Etzold, T., and Argos, P. (1993) *Comput. Appl. Biosci.* **9**, 59–64.