318–322 Nucleic Acids Research, 1999, Vol. 27, No. 1

Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms

T. Heinemeyer[1], X. Chen[1,2], H. Karas[3], A. E. Kel[1], O. V. Kel[4], I. Liebich[1], T. Meinhardt[1], I. Reuter[1,3], F. Schacherer and E. Wingender[1,*]

[1]Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany, [2]College of Life Sciences, Peking University, Beijing 100871, China, [3]BIOBASE GmbH, Mascheroder Weg 1B, D-38124 Braunschweig, Germany and [4]Institute of Cytology and Genetics, Novosibirsk, Russia

Received October 1, 1998; Revised October 7, 1998; Accepted October 16, 1998

ABSTRACT

TRANSFAC is a database on transcription factors, their genomic binding sites and DNA-binding profiles. In addition to being updated and extended by new features, it has been complemented now by a series of additional database modules. Among them, modules which provide data about signal transduction pathways (TRANSPATH) or about cell types/organs/developmental stages (CYTOMER) are available as well as an updated version of the previously described COMPEL database. The databases are available on the WWW at http://transfac.gbf.de/

INTRODUCTION

With the amount of available genomic data rapidly increasing, the need for databases and software that help to interprete raw sequence data becomes increasingly urgent. From the interpretation of coding regions, important information on hitherto unknown biological objects can be derived (‘structural genomics’; 1). However, the main characteristic of biological systems is their complexity in terms of information content and information flow. To approach the goal of describing biological systems therefore means that we have to describe functional contexts of biological objects, their interactions, dynamics and emergent properties (‘functional genomics’).

Starting from a genome, one of the most fundamental steps to express the information stored in it is transcription. Starting with a simple compilation of relevant data more than 10 years ago (2), we established the database TRANSFAC on transcription factors, their genomic binding sites and DNA-binding profiles (3–5). TRANSFAC has been tightly linked with the databases TRRD and COMPEL (3,5–7) to allow for the retrieval of additional data on more complex gene regulatory features.

Here, we present the recent developments of the TRANSFAC system as well as our concepts on how to link data about the basal mechanism (transcriptional regulation) with structured representation of higher order mechanisms.

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

TRANSFAC

Contents

The TRANSFAC database provides information about genomic binding sites of eukaryotic transcription factors (TFs) and the binding proteins. It is internally maintained as a relational database management system (RDBMS) which, after a general revision of the system, now comprises 94 different tables. TRANSFAC has been made publicly available and distributed in six ASCII flat files (Table 1). In the last year, TRANSFAC releases 3.3, 3.4 and 3.5 have been fixed in January, May and October, respectively. Due to some major changes in the data structure, the first release in 1999 will be 4.0.

Table 1. Contents of the TRANSFAC database release 3.4

Table	Entries
SITE	4945
GENE[a]	1167
FACTOR[b]	2376
CLASS	31
MATRIX	280
CELLS	938
METHOD	54
REFERENCE[c]	5937

[a]983 entries of which are connected to TRANSFAC, 420 to TRRD and 254 to both TRANSFAC and TRRD.
[b]Among the FACTOR entries, 1264 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 16 000 citations.

Among the six flat files, the SITE table stores data on experimentally proven regulatory sites in eukaryotic genomes (yeast to man; Fig. 1). Mostly, these are individual TF binding sites in specified genes of a given eukaryotic organism. The SITE table also comprises a number of consensus sequences using the 15 letter IUPAC code, as well as a huge number of artifical

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

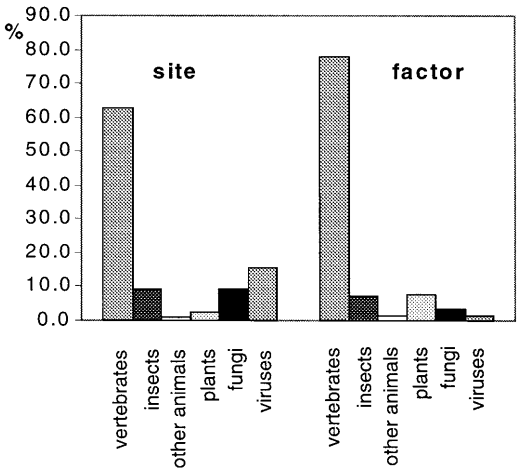

Figure 1. Species distribution of TRANSFAC SITE and FACTOR table entries.

binding sequences. The latter may have been synthesized on the basis of idealized consensi, in order to test certain hypotheses, or come from *in vitro* selection approaches to deduce binding profiles (weight matrices, see below). This part has been significantly increased to nearly 8000 SITE entries to provide the user with sequence sets underlying the matrices stored in the MATRIX table. These attempts also led to a systematic updating of the consensus sequences since we derived one new consensus from any matrix (in total 431 entries). The rules for the use of degenerate codes were similar to those of Cavener (8): (i) one letter codes (A, C, G, T) were used if more than 50% of all sequences in the aligned set were unambigously defined; (ii) double-degenerate codes (S, W, R, Y, K, M) were used if two bases occupy a certain position in at least 75% of all sequences and rule (i) does not apply; (iii) triple-degenerate codes (B, D, H, V) were applied if only one of the bases did not appear at all in the respective position.

The FACTOR table has been systematically updated by data on proteins of the classes fork head/winged helix, POU, runt and zinc fingers of Cys_2His_2 type. Whereas all known proteins of the first three groups appear to act as transcription factors, the mentioned type of zinc finger proteins also comprises members which do not exhibit this function but rather act as, e.g., RNA-binding proteins. Therefore, a careful selection had to be made for including these factors. As for the SITE table, vertebrate entries clearly predominate in the FACTOR table as well (Fig. 1). Transcription factors of plants represent a higher portion in the FACTOR table than plant sites do in the SITE table. This may reflect that a large number of plant TFs have been cloned by homology or have been identified by virtue of a specific genetic effect, but without data about their DNA-binding properties. In contrast, there is a higher proportion of binding sites in viral genes than viral factors, which is easily explained by the fact that viral genomes generally make use of the transcriptional apparatus of the host cell.

As mentioned above, a number of matrices has been newly included, but also old ones had to be revised. For documenting this kind of database maintenance, 'TRANSFAC Reports' have been prepared and made available on the WWW (http://transfac.gbf.de/TRANSFAC/tf_reports/index.html). It turned out that several matrices had to be revised because of author's errors in counting or due to suboptimal alignments of the underlying sequence sets. Thus, nearly all matrices in the database have been double-checked in the meantime. This kind of quality check is important because the MATRIX table is the main source for the matrix library of the MatInspector routine (see below).

For effective maintenance of the database, we have developed a Java-based input client. Having passed the present testing phase, it will be put on the Web to allow interested researchers to directly enter experimental data. These data will be queued into a TRANSFAC-analogous database system until some quality check has been performed and allows final entrance of the data to TRANSFAC.

User interfaces

Several search and browse options are available via WWW as was described previously (5). The 'Extended search' option now enables the user to search for transcription factor names simultaneously in the 'Name' and in the 'Synonyms' fields. This is particularly important since very frequently the same TF has been discovered independently by several laboratories and until its cloning has not been recognized to be identical. Afterwards, it normally takes some time until a commonly accepted nomenclature arises, if any. The output format of the 'Extended search' option has been expanded and includes now some additional core information (e.g., factor name and its biological species).

For the installation on stand-alone PCs we developed 'tinyTRP' (tiny TRANSFAC Retrieval Program). The program is able to handle the structure of the six TRANSFAC flatfiles including their relations which are implemented as hyperlinks. Similar to the WWW version of TRANSFAC, tinyTRP includes a search engine that allows free text search as well as searches within defined fields. Each search result can be saved for later use or for refining selections with AND, OR and NOT operators. Under development are additional features which will enable the user to use tinyTRP not only for TRANSFAC but likewise for all databases distributed in the EMBL flatfile format.

Transcription factor classification

Since release 3.2, a comprehensive transcription factor classification scheme is provided (Table 2; 5). Here, several rearrangements have been done. Thus, cold-shock domains, TFs with a histone-fold and Grainyhead-factors have been assigned to 'superclass 4' (β-scaffold factors/minor groove contacts; 9). Runt factors have been likewise shifted from the so-called 'superclass 0' which assembled all those exhibiting a not yet otherwise assignable TF to superclass 4. Also, we slightly changed the sub-classification of, e.g., the Runt domain factors. The helix–span–helix motif is now in its own class within superclass 1 ('Basic domains') although it comprises only one family (AP-2). The basic principles of the classification scheme have been published previously (9), some recent details will be explained elsewhere (Wingender *et al.*, in preparation).

It should be pointed out that the whole classification scheme has been shifted from a mere list to a fully relational database system. The system has been constructed such that it allows for handling alternative classifications using an elaborate parent–child relations system, as will be described elsewhere (Kel *et al.*, in preparation).

Table 2. Classification of transcription factors (TRANSFAC release 3.4)

Superclass[a]	Class	Number of families	Total number of subfamilies	Number of transcription factors in this class
1. Basic domain (395)	Leucine zipper factors (bZIP)	7	12	152
	Helix-loop-helix factors (bHLH)	9	8	125
	Helix-loop-helix / leucine zipper factors (bHLH-ZIP)	2	8	68
	NF-1	1	-	31
	RF-X	1	-	8
	Helix-span-helix factors (bHSH)	1	-	11
2. Zinc-coordinating DNA-binding domains (207)	Cys4 zinc finger of nuclear receptor type	2	17	111
	Diverse Cys4 zinc fingers	3	2	22
	Cys2His2 zinc finger domain	5	4	60
	Cys6 cysteine-zinc cluster	1	-	12
	Zinc fingers of alternating composition	2	-	2
3. Helix-turn-helix (522)	Homeo domain	4	28	347
	Paired box	2	-	25
	Fork head / winged helix	4	-	48
	Heat shock factors	1	-	16
	Tryptophan clusters	3	2	82
	TEA domain	1	-	4
4. β-Scaffold Factors / Minor Groove Contacts (205)	RHR (Rel homology region)	3	-	39
	p53	1	-	3
	MADS box	3	3	59
	β-Barrel α-helix transcription factors	1	-	2
	TATA-binding proteins	1	-	7
	HMG	6	-	51
	Heteromeric CCAAT factors	1	-	13
	Grainyhead	1	-	5
	Cold-shock domain factors	1	3	7
	Runt	1	5	19
0. Other Transcription Factors (29)	Copper fist proteins	1	-	2
	HMGI(Y)	1	-	6
	STAT	1	-	13
	Pocket domain	1	-	6
	E1A-like factors	1	-	2

[a]With the number of transcription factors assigned to this superclass in parentheses; the sum (1358) is higher than that of all TRANSFAC factors with class assignment (Table 1) since the classification scheme contains some factors which are not yet represented in database entries.

Cross-referencing with external databases

As was reported previously, TRANSFAC is linked with the Transcription Regulatory Region Database (TRRD) through the GENE table which is common to both databases. The presently available GENE table links TRANSFAC 3.5 with TRRD version 3.5. It also connects both these databases with COMPEL (see below).

We extended the cross-referencing with a total of 11 external data sources (Table 3). The mechanisms used for this are quite different: Mutual cross-referencing with EMBL, SwissProt (10) and EPD (11) is done by TRANSFAC and is actively supported by the providers of these databases. References to PIR and PDB are done on our side. *Drosophila* entries are referenced by FlyBase, these links are then included by TRANSFAC.

In contrast to these links which appear in the DR line (Database References) throughout the database, we started to include references to other more specialized data sources as well. In particular, FACTOR entries which deal with developmentally important TF have been linked to 'The Kidney Development Database' by Davies and Brandli [http://www.ana.ed.ac.uk/anatomy/database/kidbase/kidhome.html ; 10 links (12)] and to the 'Gene expression in tooth' page (http://honeybee.helsinki.fi/toothexp/index.htm ; 16 links).

Table 3. Cross-links of TRANSFAC 3.4 to external databases

Database (DB)	↔	TRANSFAC table	Linked TRANSFAC entries	Linked external DB entries	Total number of links
EMBL	↔	SITE	3165	1234	4359
	↔	FACTOR	1437	2155	2340
EPD	↔	SITE	743	164	749
FlyBase	↔	SITE	348	47	351
	↔	FACTOR	138	111	139
SwissProt	↔	FACTOR	1115	1071	1148
PIR	↔	FACTOR	775	982	1064
PROSITE	↔	CLASS	23	20	32
PDB	↔	FACTOR	33	29	63
TRRD	↔	SITE	1603	255	1603
COMPEL	↔	GENE	59	96	116
	↔	FACTOR	79	96	194

External browsing tools

As before, TRANSFAC is still accessible via the Sequence Retrieval System, SRS5.1 (13,14) at http://transfac.gbf.de/srs5/ . Under the subheading 'TransFac/Transcriptional regulation DBs' accesses to the individual TRANSFAC tables (TFSITE, TFFACTOR, TFCELL, TFCLASS, TFMATRIX and TFGENE) are provided. Here, TRRD and COMPEL are also available.

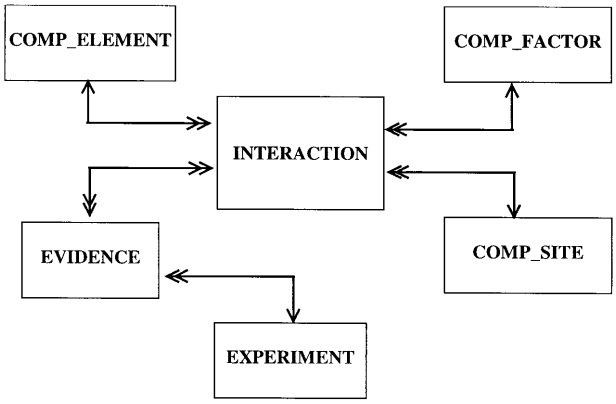

Figure 2. Part of the relational schema of the COMPEL database on composite elements. The central table interaction interconnects information about a certain composite element, its constituent sites, the interacting transcription factors and the experimental evidence by a series of n:m (double arrowheads on either side) or 1:n links (single/double arrowheads).

COMPEL

COMPEL is a database on composite elements (CE; 7). These are combined regulatory genomic sequence elements interacting with two or more distinct transcription factors thereby integrating responses to several signalling pathways with each other or with tissue- or developmental stage-specific transcriptional control mechanisms. The most recent COMPEL version, release 2.1, contains ~180 CE, nearly all of them being linked to the corresponding TRANSFAC FACTOR entries. COMPEL has recently been implemented as a fully relational database system, but only the ASCII flat files are presently accessible over the WWW.

The relational model comprises six main tables reflecting the biological subjects modelled in this database: Comp_element, Comp_site, Comp_factor, Interaction, Evidence and Experiment. The table interaction serves to link each site of the composite element with one or more factors so that each CE is linked with at least two factors (Fig. 2). CEs are usually evidenced by a number of experiments for which different factors may have been used. Therefore each evidence entry presents information on the corresponding experiment type (1:n relation) as well as on a particular interaction (n:m relation).

CONNECTED PROGRAMS

On the TRANSFAC WWW server, two programs can be used for scanning newly unravelled DNA sequences for potential TF binding sites. MatInspector (15) uses a selected library, mainly coming from the TRANSFAC MATRIX table. Publicly available on the WWW is version 2.2 under http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl . Determination of individually optimized score thresholds for each matrix has been described elsewhere (16). The selected matrix library accompanying MatInspector and its basic algorithm are also used by FastM. This tool has been developed by T. Werner and colleagues (17) and searches for specific combinations of potential TF binding sites.

The program PatSearch (http://transfac.gbf.de/cgi-bin/patSearch/patsearch.pl) uses the sequence information contained in the SITE table of TRANSFAC or, optionally, the sequence elements stored in TRRD (18). Generally, it yields long output lists and is recommended mainly for a more detailed characterization of short sequences already delimited by experimental indications or pre-annotated by, e.g., analysis with MatInspector. Its particular strength is to propose aberrant binding sites which are similar to elements which have already been experimentally characterized but are highly divergent so that none of the weight matrices available will detect them.

A similar task is fulfilled by the FindPattern routine of the GCG program package. The sequence information contained in the SITE table is also available in a GCG readable format to be used with this tool.

ADDITIONAL DATABASE MODULES

TRANSPATH

The activity of numerous transcription factors is regulated in response to certain signal transduction pathways. In most cases, these pathways involve phosphorylation cascades triggered by an extracellular signalling molecule (hormone, growth factor, cytokine etc.), but they may also include protein processing steps (as seems to be the case for NF-κB) or direct shuttling of an intercellular messenger molecule to the nucleus of the target cell where it binds to a nuclear receptor. We are modelling this type of information in an object-oriented database system (ODBMS; http://transfac.gbf.de/TRANSPATH/). Presently, nearly all data concern human cells and have been taken from CSNDB (Cell Signaling Networks Database, NIHS, Tokyo, Japan) (19). A publicly accessible interface from TRANSPATH to the TRANSFAC database is presently under development.

CYTOMER

Among other data about transcription factors, the TRANSFAC FACTOR table contains information about TF expression patterns. However, these data are presented in a rather unstructured format as mere text fields and, up to now, not even using a controlled vocabulary. Improving this situation would enable not only the mapping of expression patterns of individual TFs, but will also allow the derivation of expression patterns of their known or suggested target genes. We therefore developed another relational database module (CYTOMER) which lists physiological systems, organs and cell types. It differentiates these objects according to the developmental stage. Presently, this module is restricted to human sources, but can be easily extended to other organisms as well. Cytomer is equipped with an interface to the TRANSFAC database. In addition to the tasks described above, this module can also be used to construct expressions profiles of defined organs and cell types. It will represent a new node for linking TRANSFAC and the other databases.

AVAILABILITY

The TRANSFAC database is and remains publicly available over the WWW (http://transfac.gbf.de/TRANSFAC/) for the academic research domain. There are complete mirror sites at the Institute of Cytology and Genetics, Novosibirsk (http://transfac.bionet.nsc.ru/transfac/), at the Peking University (http://www.cbi.pku.edu.cn/TRANSFAC/), and at the Weizmann Institute of Science, Rehovot (http://bioinfo.weizmann.ac.il/transfac/).

Additional TRANSFAC servers with their own search and browse tools can be found at the University of Pennsylvania, Philadelphia (TESS, http://www.cbil.upenn.edu/tess/) and, integrated into the DBGET system, at the Kyoto University http://www.genome.
ad.jp/htbin/www_bfind?transfac (20). In addition to WWW accesses, the flat files are free for downloading by users from non-profit making organizations after registration (http://transfac.gbf.de/cgi-bin/download/download.pl). A commercial version with additional functionality and an extended data set is available as well but is not described here.

ACKNOWLEDGEMENTS

The authors are indebted to H. Hermjakob (EBI, Hinxton, UK) for his support in establishing TRANSFAC-links to the EMBL data library and to the SwissProt database, and 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, the input given by Wolfgang Fleischmann (EBI) during numerous discussions, and the support of N. A. Kolchanov (Institute of Cytology and Genetics, Novosibirsk, Russia) for the development of the COMPEL database. Finally, we express our gratitude to Mrs A. Bischoff for her technical help in nearly all of the above-mentioned fields. Parts of this work were supported by a grant of the European Commission (BIO4-95-0226), by the German Ministry of Education, Science, Research and Technology (BMBF, grants 0311640 and 01KW9629), by Scientific-Technical cooperation grants of BMBF (X224.6 and CHN-305-97) and the Russian Ministry and by a NATO grant (951149).

REFERENCES

1 Gaasterland,T. (1998) *Nature Biotechnol*., **16**, 625–627.
2 Wingender,E. (1988) *Nucleic Acids Res*., **16**, 1879–1902.
3 Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996) *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) *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) *Nucleic Acids Res*., **26**, 362–367.
6 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*, IEE Computer Society Press, Los Alamitos, CA, Vol. 5, 42–51.
7 Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E. and Kolchanov,N.A. (1995) *Nucleic Acids Res*., **23**, 4097–4103.
8 Cavener,D.R. (1987) *Nucleic Acids Res*., **15**, 1353–1361.
9 Wingender,E. (1997) *Mol. Biol. Engl. Tr*., **31**, 483–497.
10 Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res*., **27**, 49–54
11 Cavin Périer,R., Junier,T., Bonnard,C. and Bucher,P. (1999) *Nucleic Acids Res*., **27**, 307–309.
12 Bard,J.B.L., McConnell,J. and Davies,J.A. (1994) *Mech. Dev*., **48**, 3–11.
13 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol*., **266**, 114–128.
14 Etzold,T. and Verde,G. (1997) *Pacific Symp. Biocomput*., **2**, 134–141.
15 Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) *Nucleic Acids Res*., **23**, 4878–4884.
16 Pickert,L., Reuter,I., Klawonn,F. and Wingender,E. (1998) *Bioinformatics*, **14**, 244–251.
17 Frech,K., Danescu-Mayer,J. and Werner,T. (1997) *J. Mol. Biol*., **270**, 674–687.
18 Wingender,E., Karas,H. and Knüppel,R. (1997) *Pacific Symp. Biocomput*., **2**, 477–485.
19 Igarashi,T. and Kaminuma,T. (1997) *Pacific Symp. Biocomput*., **2**, 187–197.
20 Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pacific Symp. Biocomput*., **3**, 681–692.