TRANSFAC®

flat files

BIOBASE

BIOLOGICAL DATABASES

Documentation

2012.1

geneXplain GmbH
Am Exer 10b
D-38302 Wolfenbüttel
Germany

E-mail: info@genexplain.com

URL: http://www.genexplain.com

Content

1. Overview

1.1. Preface

TRANSFAC(r) is a database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors. It covers the whole range from yeast to human. TRANSFAC(r) started 1988 with a printed compilation (Nucleic Acids Res. 16: 1879-1902, 1988) and was transferred into computer-readable format in 1990 (BioTechForum - Advances in Molecular Genetics (J. Collins, A.J. Driesel, eds.) 4:95-108, 1991). The basic structures of Table 1 and 2 of the compilation were taken as the core of the emergent database. The aim of the early compilation as well as of the TRANSFAC(r) database is

- to guide through a meanwhile overwhelming amount of data in a field which is connected to nearly all areas of modern molecular biology;
- to map the regulatory sites in the individual genes and, ultimately, in the genome(s) as a whole;
- to develop a tool for the identification of regulatory elements in newly unravelled genomic sequences;
- to provide a basis for a more comprehensive understanding of how the genome governs transcriptional control.

The TRANSFAC(r) data have been generally extracted from the original literature, occasionally they have been taken from other compilations (Faisst and Meyer, Nucleic Acids Res. 20:3-26, 1992; Dhawale and Lande, Nucleic Acids Res. 21:5537-5546, 1994) which is appropriately indicated. Thus, the main responsibility for the correctness of the data is up to the authors, while we have to assume any responsibility for correctly extracting their publications. In a long term, a direct submission system should be envisaged similar to that established for depositing sequence data in the sequence data libraries.

1.2. TRANSFAC® tables and their relations

This ASCII flat file release comprises the following data files::

FACTOR describes proteins that regulate transcription (by sequence-specific interaction with DNA) and micro RNAs (miRNA) that control stability or translation of messenger RNA (mRNA).

GENE describes a gene where TF-binding sites or ChIP fragments belongs to and/or genes encoding for transcription factors or miRNAs.

SITE gives information on individual (regulatory) protein binding sites (within eukaryotic genes) and mRNA sequence parts that are targets for miRNA interaction.

FRAGMENT contains DNA fragments to which in vivo binding of a transcription factor was shown by ChIP-on-chip, ChIP-Seq, or related experiments.

MATRIX contains nucleotide distribution matrices for the binding sites of transcription factors.

CLASS contains background information about the transcription factor classes.

CELL gives brief information about the cellular source of proteins that have been shown to interact with these sites.

REFERENCE contains references extracted for TRANSFAC(r) (factor, site, etc.) entries together with links to these entries and to PubMed.

1.3. How to cite

Please cite for TRANSFAC(r):

- Matys, V.; Kel-Margoulis, O. V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A. E.; Wingender E. (2006) "TRANSFAC(r) and its module TRANSCompel(r): transcriptional gene regulation in eukaryotes." Nucleic Acids Res. 34(Database issue):D108-110. Please cite for Match(TM):
- Kel, A. E.; Gößling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O. V.; Wingender, E. (2003) "MATCH(TM): A tool for searching transcription factor binding sites in DNA sequences" Nucleic Acids Res. 31, 3576-3579.

1.4. Publications

1. Matys, V.; Kel-Margoulis, O. V.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A. E.; Wingender E. (2006) "TRANSFAC(r) and its module TRANSCompel(r): transcriptional gene regulation in eukaryotes." Nucleic Acids Res. 34(Database issue):D108-110.
2. Kel-Margoulis, O.; Matys, V.; Choi, C.; Reuter, I.; Krull, M.; Potapov, A. P.; Voss, N.; Liebich, I.; Kel, A.; Wingender, E. (2005) "Databases on gene regulation." In Bajic, V.B. and Tan, T.W. (Eds.). Information Processing and Living Systems, Singapore World Scientific Publishing Co. Vol. 2, pp. 709-727.
3. Kel, A. E.; Gößling, E.; Reuter, I.; Cheremushkin, E.; Kel-Margoulis, O. V.; Wingender, E. (2003) "MATCH(TM): A tool for searching transcription factor binding sites in DNA sequences" Nucleic Acids Res. 31, 3576-3579.
4. Matys, V.; Fricke, E.; Geffers, R.; Gößling, E.; Haubrock, M.; Hehl, R.; Hornischer, K.; Karas, D.; Kel, A. E.; Kel-Margoulis, O. V.; Kloos, D. U.; Land, S.; Lewicki-Potapov, B.; Michael, H.; Münch, R.; Reuter, I.; Rotert, S.; Saxel, H.; Scheer, M.; Thiele, S.; Wingender, E. (2003) "TRANSFAC(r): transcriptional regulation, from patterns to profiles" Nucleic Acids Res. 31, 374-378.
5. Kloos, D. U.; Choi, C.; Wingender, E. (2002) "The TGF-beta--Smad network: introducing bioinformatic tools" Trends Genet. 18, 96-103.
6. Wingender, E.; Chen, X.; Fricke, E.; Geffers, R.; Hehl, R.; Liebich, I.; Krull, M.; Matys, V.; Michael, H.; Ohnhäuser, R.; Prüß, M.; Schacherer, F.; Thiele, S.; Urbach, S. (2001) "The TRANSFAC(r) system on gene expression regulation" Nucleic Acids Res. 29, 281-283.

7. Wingender, E.; Chen, X.; Hehl, R.; Karas, H.; Liebich, I.; Matys, V.; Meinhardt, T.; Prüß, M.; Reuter, I.; Schacherer, F. (2000) "TRANSFAC(r): an integrated system for gene expression regulation" Nucleic Acids Res. 28, 316-319.

8. Heinemeyer, T.; Chen, X.; Karas, H.; Kel, A. E.; Kel, O. V.; Liebich, I.; Meinhardt, T.; Reuter, I.; Schacherer, F.; Wingender, E. (1999) "Expanding the TRANSFAC(r) database towards an expert system of regulatory molecular mechanisms" Nucleic Acids Res. 27, 318-322.

9. Knüppel, R.; Dietze, P.; Lehnberg, W.; Frech, K.; Wingender, E. (1994) "TRANSFAC(r) retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins." J. Comput. Biol. 1, 191-198.

10. Wingender, E. (1988) "Compilation of transcription regulating proteins" Nucleic Acids Res. 16, 1879-1902.

For further reading, please see http://www.biobase.de/pages/index.php?id=publications.

2. Table FACTOR

2.1. Contents

The total number of entries in the FACTOR table does not reflect the number of independent transcription factors. First of all, homologous factors from different species, such as human and mouse SRF, are given in different entries since they may differ in some molecular aspects. Moreover, factors which have originally been described by different research groups to bind to different genes may turn out to be identical as soon as they have been cloned. On the other hand, more and more factors are recognized to be representatives of whole transcription factor families, comprising products of distinct but very similar genes or alternative splice products. In many cases, a more general term originally defining just a specific DNA-binding activity such as AP-1 appears as one entry. In most cases, this activity has not been analyzed for its subunit composition by members of the Jun and Fos families. Nevertheless, all fos- and jun-related proteins are included as separate entries.

All factors that are mentioned in the SITE table appear in the FACTOR table as well. However, it includes also polypeptides, which do not bind to DNA by themselves. One well-known example is c-Fos, which is forced to contact DNA only by being complexed with, e.g., c-Jun. Information about non-DNA binding subunits of transcription factor complexes, such as the TAFs, is given by FACTOR as well. There are also proteins that act as inhibitors for a particular DNA-binding activity and which are of regulatory importance. Therefore, proteins such as Id, lkappaB or hsp90 have been included in TRANSFAC® FACTOR.

On the other hand, proteins which carry a putative DNA-binding motif have in general not yet been entered. Thus, there are much more zinc finger proteins known than are included in FACTOR, but for many of them, no data about DNA-binding specificity or about other important gene-regulatory features are available.

In general, a protein is a potential entry for TRANSFAC® if it fits to the following definition: "A transcription factor is a protein that regulates transcription (after nuclear translocation) by sequence-specific interaction with DNA or by stoichiometric interaction with a protein that can be assembled into a sequence-specific DNA-protein complex."

In addition to transcription factors, with release 10.4 we have started to include in the FACTOR table also micro RNAs (miRNA) that control stability or translation of messenger RNA (mRNA) by sequence-specific interaction.

2.2. Fields

It should be noted that in individual entries, some fields may be empty. In this case, these fields are not displayed.

Field	Content and format
AC	Accession number "T" + 5-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

ID Identifier

"T" + 5-digit number (identical with accession number)

DT Created/Updated

date of entry creation; entry author /
date of last entry updating; updater

FA Factor name

(normally the most commonly used) name of the factor (NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

SY Synonyms

alternative names of the transcription factor

OS Species

biological species (in some cases, when the species of the protein used in an experiment was not clearly given in a publication, the species is assigned to a "taxonomic class" (mammalia, vertebrata, ...); complex entries may have more than one species assigned, when the subunits used in the experiment were derived from different species)

OC Taxonomic classification

systematic biological classification of the species

GE Encoding gene

GENE accession no.; short gene term; HGNC: standard gene symbol.

HO Homologs

suggested homologous transcription factors from distant biological species (e. g., yeast HAP2 as homolog to human CP1B)

CL Factor class

assignment of the factor to the comprehensive transcription factor classification (class accession number linked to the CLASS table entry; class identifier; decimal classification number linked to the classification tree.)

TY Type

the type of this factor entry (not yet given in all entries);
possible values are:

family, group entry which summarizes different products of (closely related) paralogous genes

isogroup, group entry which summarizes different products (e.g. alternative splice variants) of the same gene

basic, for specific isoforms (concrete existing monomeric proteins)

complex, factors consisting of more than one non-covalently bound protein/molecule

miRNA, micro RNA

HP Superfamilies

lists generic entries (isogroup or family) to which this factor belongs

HC Subfamilies

lists entries, e.g. splice variants or family members of this isogroup/family entry

SZ Size

length (number of amino acids); calculated molecular mass in kDa (derived from cDNA / genomic clones); experimental molecular mass (or range) in kDa (experimental method, e. g. SDS PAGE, GF/gel filtration)

SQ Sequence

protein sequence of the factor (for miRNAs: RNA sequence)

SC Sequence source

source of the (protein) sequence (e. g., SwissProt, PIR, MIRBASE)

FT Feature table

local features of the factor molecule:

first position last position feature

SF Structural features

global structural features of the factor

CP Cell specificity (positive)

organs / cells in which the factor has been demonstrated to be expressed

CN Cell specificity (negative)

organs / cells in which the factor has been demonstrated NOT to be expressed

EX Expression pattern

organ, cell name, system, developmental stage; relative level of expression (very high, high, medium, low, very low, detectable or none); detection method; molecule type detected, i.e. RNA or protein; [reference]

FF Functional properties

functional properties of the factor including more detailed explanations of its expression pattern and of its regulation

IN Interacting factors

factors which interact physically with the factor of this entry (as the applied methods may also include co-immunoprecipitation from crude cell extracts, or similar, it cannot be excluded that in some cases the binding between the two proteins was mediated by a third protein) (linked accession number; name; biological species)

ST Subunits (Precursors)

subunits of the given factor (complex), or the precursor/unmodified form (for non complexes)

CX Complexes

a list of complexes which contain this factor

MX Matrices

MATRIX table entries providing DNA-binding profiles of the factor (linked accession number; identifier.)

BS Binding sites / Regulated genes

DNA (or RNA) sequences shown to be bound by the factor (linked accession number; identifier; "Quality" of the factor-site interaction on a six level scale;) and for genomic sites: (short gene term; GENE accession no. biological species.)

BR Binding region (ChIP-chip/-Seq)

DNA fragment from ChIP-on-chip or ChIP-Seq experiments (linked accession number; "Quality" of the factor-DNA interaction; biological species.)

DR External database links

Database name (e. g. BKL, EMBL, SwissProt, PIR, Flybase, PDB, DATF, MIRBASE, TRANSCompel, PathoDB, SMARtDB, TRANSPATH): database accession number, identifier (where available).

In case of EMBL cross-links: (r) denotes reference to a RNA/cDNA, (g) to a genomic DNA sequence.

RSNP: accession number; EMBL: accession number; pos: SNP position in EMBL sequence; var: variation introduced by SNP; effect of SNP
(example: RSNP: 97894; EMBL: M61108; pos: 716; var: a,g; amino acid exchange, A47->T);

RN Reference number

[consecutive entry reference number]; reference accession number.

RX PUBMED; link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows:

ä -> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

journal volume:pages (year)

2.3. Explanations

Up to now, no identifiers have been assigned to the FACTOR entries, instead the accession numbers are repeated. The field "Synonyms" covers different spelling (AP-1/AP1) as well as real alternative names (HNF-1: HNF-1alpha, APF, LF-B1). In contrast, the field "Homologs" indicates the names of other proteins, frequently from evolutionary more distant species, which may be functionally and/or structurally related to the factor under consideration.

The field "Factor classification" indicates the major class of DNA-binding domains a factor may be assigned to. It also contains a systematic decimal classification number referring to the proposed transcription factor classification scheme. Note that this is a tentative assignment which may change according to the insights into the structure-function relationships of this large protein category.

The "Size" field shows the number of amino acid residues of a polypeptide and its molecular weight. The method by which this figure has been obtained is indicated in brackets; (cDNA) or (gene) means that is has been calculated after cloning, (SDS) or (sedim.) hints on the corresponding experimental approaches.

The "Sequence" field contains the full amino acid sequence of the transcription factor. It may have been copied from SwissProt or PIR or conceptually translated from an EMBL/GenBank/DDBJ nucleic acid sequence, as is indicated in the "Sequence comment" field. In case that some manual editing has been done, this is also indicated in this line.

The "Feature table" may contain information on:

- ❖ regions that are enriched in some amino acid residues and may therefore represent trans-activating domains; the content is given as (M/N) which means that M out of N residues are of the enriched amino acid;

- positions of the typical DNA-binding/dimerization motifs and the motif structure within the individual molecule; e. g. tryptophan cluster motifs are explained with regard to Trp spacing, and the nature of a leucine zipper is given as well (e.g. L4 means that it consists of four leucine residues spaced by 6-AA-intervals, L2EL2 indicates a motif such as L-X6-L-X6-E-X6-L-X6-L);
- the AA that coordinate the zinc ion(s) in a zinc finger motif (e.g. C2HC for three cysteines and one histidine)
- posttranslational modifications (phosphorylation, glycosylation). For visualization of positional features, please see the online version of TRANSFAC.

The field "Structural features" gives information about global structural features of the factor. Data may be referenced, the source of information used is indicated by a bracketed number that points to the corresponding paper at the end of the entry.

"Cell specificity (positive)" gives predominant occurrence of a factor in certain cell types or tissues. Occasionally, cells from which the factor has been isolated are indicated in brackets; this information does not necessarily point to a true cell specificity. Additionally, "Cell specificity (negative)" lists cells / tissues which have been proven not to express the corresponding factor. For factors from human or mouse the "Cell specificity" fields have been started to be replaced by the better structured "Expression pattern" field. (For a more comprehensive coverage of expression data, please see the online version of TRANSFAC.)

Interactions: Since most transcription factors bind to DNA as dimers, the dimerization partners are indicated in this field. Repeating the factor's name in this field means that it forms homodimers. Also given are inhibitory protein-protein-interactions such as NF-kappaB - IkappaB.

The field "Matrix" gives accession number and identifier of the connected MATRIX table entries.

"External databases" points to corresponding entries within the EMBL, SwissProt, PIR, FlyBase, PDB, RSNP, TRANSCompel®, PathoDB, SMARtDB™ or TRANSPATH® data libraries.

3. Table CLASS

3.1. Contents

This record briefly explains some of the main features of the DNA-binding domains of transcription factor classes. In those cases where an amino acid consensus motif has been identified, the corresponding accession number of the PROSITE database (A. Bairoch) is included.

3.2. Fields

It should be noted that in individual entries, some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

"C" + 4-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

ID Identifier

acronym for the structural class (e.g., bZIP for basic region-leucine zipper domain)

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

CL Class

denomination of the structural class; link to transcription factor classification tree.

CC Description

explanation of the features of the particular class of DNA-binding domains

BF Factors

list of transcription factors assigned to this class (linked factor accession number; factor name; biological species of the factor)

DR External database links

database name (currently always PROSITE): PROSITE accession number.

RN Reference number

[consecutive entry reference number]; reference accession number.

RX PUBMED; link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: „ ä-> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

journal volume:pages (year)

4. Table SITE

4.1. Contents

The SITE table gives information on individual (regulatory) protein binding sites. It contains three different kinds of entries. First, there are sites within eukaryotic genes, the species of which ranging from yeast to human. Second, it comprises artificial sequences which resulted from mutagenesis studies, in vitro selection procedures starting from random oligonucleotide mixtures or from specific theoretical considerations. And finally, SITE includes consensus binding sequences given in the IUPAC code, many of them being taken from the compilation of Faisst and Meyer (Nucleic Acids Res. 20:3-26, 1992). The symbols used in addition to A, C, G, or T for these consensi are:

W = A or T
S = C or G
R = A or G
Y = C or T
K = G or T
M = A or C
B = C, G or T
D = A, G or T
H = A, C or T
V = A, C or G
N = A, C, G or T

A number of consensi has been generated by the TRANSFAC(r) team, generally derived from the profiles stored in the MATRIX table. Here, the use of degenerate codes follows the following rules (adapted from Cavaner, Nucleic Acids Res. 15:1353-1361, 1987):

Rule 1: A single nucleotide is shown if its frequency is at least 50% and at least twice as high as the second most frequent nucleotide.

Rule 2: A double-degenerate code indicates that the corresponding two nucleotides occur in at least 75% of the underlying sequences and rule 1 does not apply.

Rule 3: Usage of triple-degenerate codes is restricted to those positions where one of the nucleotides did not show up at all in the sequence set and none of the afore mentioned rules applies.

Rule 4: All other frequency distributions are represented by the letter "N".

In addition to transcription factor binding sites, with release 10.4 we have started to include also mRNA sequence parts that are targets for miRNA interaction.

4.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field	Content and format
AC	Accession number

"R" + 5-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

ID Identifier

{species acronym}${gene acronym}_{consecutive site number}

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

TY Sequence type

D (DNA) or R (RNA)

DE Description

short gene term (explicit gene name); GENE accession no.

OS Species

biological species

OC Taxonomic classification

systematic biological classification of the species

RE Gene region

functional region of the gene (e.g. promoter, enhancer, intron etc.)

SQ Sequence

Site sequence(s)

EL Element

specific denomination of this site (if available), such as CRE (cAMP-response element)

S1 Reference point

SF Start position

ST End position

the position numbers normally refer to the transcription start site (+1); where this is not the case, the reference point is stated explicitly

BF Binding factors

factors shown to bind to this site (accession number; name; "Quality" of the factor-site interaction on a six level scale; biological species of the factor.

MX Matrices

nucleotide distribution matrices derived from alignment of this and other binding sites of a specific factor (accession number; identifier.

SO Cellular factor source

expression system (tissue, cell line, ...) the factor or binding activity was derived from (linked accession number; short description)

MM Method(s)

methods applied for the identification of this site

CC Comments

any additional comments, e. g. on the functionality of the site or on sequence conflicts with corresponding EMBL entries

DR External database links

name of database (e. g. TRANSPRO, PathoDB, Flybase, EPD): database accession number; identifier (where available).

EMBL: accession number; identifier (first site position:last site position).

RSNP: accession number; EMBL: accession number; pos: SNP position in EMBL sequence; var: variation introduced by SNP.

RN Reference number

[consecutive entry reference number]; reference accession number.

RX PUBMED; link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: „ -> ae, ÷ -> oe, ▯ -> ue; German "s-z" (˜) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

journal volume:pages (year)

4.3. Criteria

The first criterion for a site to be included in TRANSFAC(r) is protein binding, the second is function. Assigned to each site is an unambiguous accession number and an identifier. The latter is composed of an abbreviation for the species (e.g., HS for human), a code for the gene description and a consecutive number for each entry referring to a particular gene. Thus, HS$BAC_02 refers to the second entry for the human gene beta-actin.

The description of a gene is the name of the gene itself or of its product, depending on what the more popular term may be. The positions have preferably been taken from DNase I footprinting studies, if available. The next preference is for chemical modifications, the last for gel retardation assays. In case of different positional information for both DNA strands, the more upstream position has been taken for the 5' border, the more downstream position for the 3' border of the site. If not stated otherwise in the field "Reference point (+1)", the position numbers generally refer to the transcription start site. Occasionally (or normally for yeast genes due to their generally more heterogeneous cap site), they may refer to the translation start codon stated as "Reference point (+1) | ATG". Other reference systems such as defined restriction sites may be indicated as well. If "First position of element" and "Last position of element" are missing, no positions were given by the references cited. If "First position of element" has a negative or positive value, but "Last position of element" is missing, no precise boundaries of the site were given, but it was located "around position" instead.

The sequences depicted have been taken from the literature. Some conflicting data with sequences within the EMBL data library are mentioned in the comment field. In case of diverging site borders on both strands, only the overlapping sequence is given. When the authors emphasized a certain sequence motif within a sequence, it is written in capitals while the rest of the sequence is shown in lowercase letters.

Cross-references to the EMBL data library also give the positions of the TRANSFAC(r) site within the EMBL sequence, negative numbers pointing to the complementary strand.

The factor which binds to this sequence element is given with its TRANSFAC(r) accession number of the FACTOR table and (one of) its name(s) (see FACTOR table for possible synonyms), and a "quality" value ranging from 1 to 6 reflecting the experimental reliability of a certain protein-DNA interaction. These values have the following meaning:

1. functionally confirmed factor binding site
2. binding of pure protein (purified or recombinant)
3. immunologically characterized binding activity of a cellular extract
4. binding activity characterized via a known binding sequence
5. binding of uncharacterized extract protein to a bona fide element
6. no quality assigned

The cellular protein source leading to the identification of a particular site is included in the SITE table as well.

4.4. Methods

Footprinting reactions:

* DNase I footprinting
* genomic in situ DNAse I footprinting
* competitive DNAse I footprint
* methidiumpropyl-EDTA.Fe(II)
* methidiumpropyl-EDTA.Fe(II) in nuclei
* DNase II
* hydroxyl radicals
* photofootprinting in vivo
* copper/phenanthroline footprinting
* copper/phenanthroline footprinting in situ
* neocarzinostatin footprinting
* micrococcal nuclease
* micrococcal nuclease in situ
* nuclease P1 footprinting
* in vivo DMS footprinting

Exonuclease digests:

* exonuclease III
* genomic / in vivo exonuclease III digest
* competition exonuclease III digest
* lambda exonuclease
* T7 gene exonuclease

Gel retardation:

* direct gel shift
* supershift (antibody binding)
* gel shift competition

DNA modification reactions:
* methylation protection
* in vivo / genomic methylation protection
* methylation interference
* Dam methylation
* ethylation protection
* ethylation interference
* DEPC interference
* KMnO4 modification
* genomic demethylation
* uracil interference
* depurination interference
* missing base interference
* carboxymethylation interference
* carboxyethylation interference
* depyrimidination interference
* cytosine and adenine interference

* primer extension footprint

* avidin-biotin complex DNA binding assay

Blotting:
* Southwestern blotting / filter binding
* nitrocellulose filter binding
* southwestern blotting

UV:
* UV-crosslinking
* genomic UV-photofootprint
* UV/primer extension

* immunoprecipitation

* affinity chromatography

* crystallization

* electron microscopy

* chromatin immunoprecipitation procedure

* yeast one hybrid system

* Functional analysis

5. Table CELL

5.1. Contents

This table gives a short explanation of the cellular sources of proteins that interact with the sites listed in the SITE table. Among them may be defined cell lines, tissues / organs, even whole organisms, or recombinant expression systems.

5.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

4-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

SO Cellular factor source

cell, cell line, organ or recombinant expression system used as source for the identification of a certain site; the syntax for recombinant factors is: rec({donor organism} - {acceptor organism}), e.g. rec(human-E.coli); when appropriate, an intermediate organism serving as vehicle is given as well, e.g. rec(rat-vaccina virus-HeLa)

OS Species

biological species

CD Cell description

description of the cellular factor source

BS Binding site

sites bound by a factor derived from this factor source (accession number of entry in site.dat; identifier.)

BR Binding region (ChIP-chip/-Seq)

DNA fragments shown to be bound by a given factor in vivo (ChIP-on-chip / ChIP-Seq experiment) in this cell line. (accession number of entry in fragment.dat)

DR External database links

CLDB: accession number.

TRANSCompel: accession number.

5.3. Terminology

The terminology of the factor sources ("Cellular factor source") corresponds to that used in the SITE table. Cell lines are given with their most common name (e.g., HeLa, 3T3), the most frequently used tissues may be abbreviated (e.g., rl for rat liver), others are explicitly mentioned. Recombinant materials are explained by the denotation rec({source organism} - {expressing cell}). Occasionally, a viral expression system is also indicated: rec(chick-baculovirus-Sf9).

In this table, no taxonomic classification is given since this kind of information is obscure in too many cases.

Where appropriate, links are given to the CLDB

(http://www.biotech.ist.unige.it/cldb/indexes.html).

6. Table FRAGMENT

6.1. Contents

The Fragment table is designed to provide information on in vivo binding DNA fragments. The collected DNA fragments are based on the results of "ChIP-on-chip" experiments that combine in vivo Chromatin Immuno- Precipitation (ChIP) with various microarray technologies (chip) or ChIP- Seq experiments.

The Fragment table is interlinked with other TRANSFAC® tables (Gene, Factor, Cell, Reference) where more details about the corresponding entities can be found.

In addition to binding fragments we provide information about closely located genes. Please note, genes have been associated with the fragments just according to their close location to the in vivo binding DNA fragments. To provide information about the closely located genes we mapped DNA fragments on the respective sequence build and then looked for annotated TSS and exons within a window of 150 kb on both sides of the fragment. On both sides the feature (TSS or exon) with the smallest distance to the fragment was chosen, and the respective gene was assigned to the fragment as "nearest gene". This procedure resulted in one, two, or no genes associated with each fragment.

6.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

"FR" + 7-digit number

AS Accession numbers, secondary

BKL TRANSFAC module accession number of the entry. "FR" + 9-digit number

ID Identifier

"FR" + 7-digit number

DT Created/Updated

Date of entry creation; entry author /

Date of last entry updating; updater

DE Description

GENE accession no.
Genes located within 150 kb on both sides of the fragment. One "nearest gene" on each side of the fragment is shown.

OS Species

Biological species

OC Taxonomic classification

Systematic biological classification of the species

SQ Sequence

Sequence of the ChIP-chip DNA fragment

Capital letters indicate probes (PUBMED: 15082775, PUBMED: 14527995) or binding fragments as authors defined them (PUBMED: 14980218) and small letters indicate flanks added by TRANSFAC team

SC Sequence Source

Sequence build; Chromosome number; Position1-position2 (absolute positions on the chromosome); FORWARD/REVERSE (strand information)

BF Binding factors

Factors shown to bind the DNA fragment in vivo (linked accession number; name; "Quality" of the factor-fragment interaction; biological species of the factor; Cellular source (where in vivo binding was shown): linked accession number; short description).

PS Best supported binding site in the fragment's sequence predicted by Match

Positional weight matrix accession and ID; accession of binding transcription factor; predicted start and end position of the site; Core Similarity Score (CSS) and Matrix Similarity Score (MSS) (calculated by Match algorithm)

MM Method(s)

Methods applied for the identification of the fragment

DR External database links

EMBL: accession number; identifier (first site position:last site position)

RN Reference number

[Consecutive entry reference number]; Reference accession number.

RX PUBMED

link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: ä -> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

Journal Volume:Pages (year)

7. Table GENE

7.1. Contents

The GENE table is a central table for all of our databases, now including links to TRANSFAC(r) and TRANSCompel(r) tables as well as links to TRANSPATH(r), PathoDB(r) and S/MARtDB(TM). It contains all genes for which information is contained in (at least) one of the databases. A GENE entry lists all links to individual sites, within DNA or mRNA, (and to their binding factors) given in TRANSFAC(r) and to the composite elements given in TRANSCompel(r) along with their positions within the gene. Where appropriate, links to encoded FACTORs (transcription factors or miRNAs) in TRANSFAC(r) are given as well. In addition to these internal links, links to a growing number of external databases such as ENTREZ, RefSeq, and OMIM are given. Further, the GENE table contains the promoter classification number as proposed by P. Bucher as well as other information.

7.2. Fields

It should be noted that in individual entries, some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

"G" + 6-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

ID Identifier

{species acronym}${gene acronym}

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

SD Short description

short gene term

DE Description

explicit name of the gene

SY Synonyms

alternative names of the gene

OS Species

biological species

OC Taxonomic classification

systematic biological classification of the species

CH Chromosomal location

chromosome and locus of the gene

HG Host gene

Gene in which the (miRNA-encoding) gene is located.

IG Intronic gene

Gene (encoding for miRNA) which is located within this gene.

BC EPD Promoter classification

decimal classification number according to P. Bucher's promoter classification system

RG Regulation

where and under what conditions the gene is expressed

BS Binding sites / Binding factors

site positions (normally, relative to the transcription start site, but there may be other reference points, which are then stated in the individual site entries) and linked site accession number; site identifier; Binding factors: factor name linked factor accession number

BR Binding region (ChIP-chip/-Seq)

DNA fragment from ChIP-on-chip or ChIP-Seq experiments (linked accession number; upstream/downstream from the gene; biological species.)

CE Composite element

positions and accession number of composite element from TRANSCompel.

FA Encoded factor

linked factor accession number; factor name. (for genes encoding a transcription factor or miRNA)

DR External database links

database name (ENTREZGENE, ENSEMBL, UNIGENE, REFSEQ, TRANSPRO, EMBL, OMIM, BKL, TRANSPATH, SMARtDB, PathoDB): database accession number, identifier (where available).

HGNC, MGI, or RGD: standard gene symbol for human, mouse, or rat gene, respectively.

AFFYMETRIX: chip: probeset. (except for those from chip HuGeneFL, the Affymetrix links are based on those in Ensembl, v.14.31 for human and v.14.30 for mouse)

BRENDA: BRENDA_EC_no.; BRENDA_species_EC_no.

RN Reference number

[consecutive entry reference number]; reference accession number.

RX PUBMED

link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: ä -> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

journal volume:pages (year)

8. Table MATRIX

8.1. Content

The MATRIX table contains nucleotide distribution matrices of aligned binding sequences. These sequences may have been obtained by in vitro selection studies or may be compiled sites of genes. The source is appropriately indicated. The matrix entries have an identifier that indicates one of six groups of biological species (V$, vertebrates; I$, insects; P$, plants; F$, fungi; N$ nematodes; B$, bacteria), followed by an acronym for the factor the matrix refers to, and a consecutive number discriminating between different matrices for the same factor. Thus, V$OCT1_02 indicates the second matrix for vertebral Oct-1 factor. Instead of a consecutive number, the identifier of those matrices which have been generated from TRANSFAC(r) SITE entries, end up with an abbreviation of the least quality of the sites used to construct the matrix. For example, V$CREB_Q2 is a matrix constructed of CREB binding sites of quality 2 or better. This Q-value should not be mixed up with the high/low quality criterion of the matrix, which is given in Match(TM)/Match(TM)Profiler. Finally, a matrix with an identifier like V$AP1_C has been derived from a "consensus description" constructed with the aid of ConsIndex (Frech et al., Nucleic Acids Res. 21:1655-1664, 1993).

The matrix area gives the nucleotide frequencies observed in aligned binding sites of the corresponding transcription factor (or, more general, in aligned sites of the described function); an additional column depicts the IUPAC string consensus derived from the matrix according to the following rules (adapted from Cavener, Nucleic Acids Res. 15:1353-1361, 1987):

Rule 1: A single nucleotide (A,C,G,T) is shown if its frequency is at least 50% and at least twice as high as the second most frequent nucleotide.

Rule 2: A double-degenerate code indicates that the corresponding two nucleotides occur in more than 75% of the underlying sequences and rule 1 does not apply: (W = A or T), (S = C or G), (R = A or G), (Y = C or T), (K = G or T), (M = A or C).

Rule 3: Usage of triple-degenerate codes is restricted to those positions where one of the nucleotides did not show up at all in the sequence set and none of the afore mentioned rules applies:(B = C, G or T), (D = A, G or T), (H = A, C or T), (V = A, C or G).

Rule 4: All other frequency distributions are represented by the letter "N" (= A, C, G or T).

8.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

"M" + 5-digit number

AS Accession numbers, secondary

when two or more entries are merged, the additional accession numbers, separated by commas, are stored in this field

ID Identifier

{species group}${factor}_{discriminating extension}

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

NA Name

designation of the binding transcription factor (or in some cases of the element, e.g. TATA)

DE Factor description

short description of the factor function

TY Type

type of the matrix: family or (factor-)specific

OS Species/Taxon

biological species or higher taxonomic group of the transcription factors, for which the matrix has been compiled

OC Taxonomic classification

systematic biological classification of the species

HP Superfamilies

lists family matrices, to which the factor(s) of the specific matrixbelong

HC Subfamilies

lists factor-specific matrices, that provide a more narrow focus on certain member of the factor family

BF Binding Factors

list of linked entries of the Factor table (factor accession number; factor name; biological species); if a binding site for this factor was used to compile the matrix, this is indicated, otherwise the factor has been linked by its homology to the directly involved factors

P0 Binding Matrix

nucleotide frequency matrix with matrix head (A C G T) and derived IUPAC consensus in the last column underneath: for matrix visualized as sequence logo (adapted from Schneider and Stephens, Nucleic Acids Res. 18:6097-6100, 1990) see separate files

BA Basis

statistical basis of the matrix (and the method that has been applied to generate it)

BS Binding sites

list of aligned sequence segments used for matrix generation (if available) followed by a link to the respective binding site in TRANSFAC (site accession number) or TRANSCompel from which the segment was derived and a description how it was derived (start, length, gaps and orientation of the depicted sequence segment - flanking gaps included - relative to the sequence in the site entry), see below

CC Comments

details on the experimental approach applied to retrieve the sequence set and to compile the matrix

DR External database links

database name (e.g. JASPAR, SwissRegulon, UniPROBE): database accession number

OV Older version

an older version of this matrix, e.g. compiled from a smaller set of sequences

PV Preferred version

the most current version of the matrix

RN Reference number

[consecutive entry reference number]; reference accession number.

RX PUBMED

link to PubMed entry.

RA Reference authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: ä -> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Reference source

journal volume:pages (year)

8.3. Alignment description

BS Sequence segment (including gaps); Site link; Start; Length; Gaps; Orientation.

Examples:

BS TGTTTGTCAAT; R08890; 9; 11;; p. derived from TRANSFAC site R08890

BS TATTTACTTTC; C00278, s00475; 2; 11;; n. derived from TRANSCompel site s00475 in composite element C00278

The positions given refer to the sequence as shown in the site table.

For re-construction of alignment from site sequence

1. introduce gaps at the indicated positions

2. determine matrix start

3. determine matrix length

4. orientation: p (positive), i.e. as in site entry / n (negative) = reverse complement

Examples:

BS AATCCGGAAACGATG; R05296; 1; 17; 1, 2; p

```
                                        0000000001111111111
                                        1234567890123456789
Sequence in site entry:                 AATCCGGAAACGATGCG
1. positions of gaps: 1, 2              --AATCCGGAAACGATGCG
2. matrix start: 1                      --AATCCGGAAACGATGCG
3. matrix length: 17                    --AATCCGGAAACGATG
4. orientation: p
```

BS CCTGGTCA ATGGGTCATG; R11613; 9; 19; 17; p

```
                                        000000000111111111122222222
                                        123456789012345678901234567
Sequence in site entry:                 GGGTCACTCCTGGTCAATGGGTCATG
1. positions of gaps: 17                GGGTCACTCCTGGTCA-ATGGGTCATG
2. matrix start: 9                              CCTGGTCA-ATGGGTCATG
3. matrix length: 19                            CCTGGTCA-ATGGGTCATG
4. orientation: p
```

BS AACTACCGG; R08780; 1; 11; 10, 11; n.

```
                                        00000000011
                                        12345678901
Sequence in site entry:                 CCGGTAGTT
1. positions of gaps: 10, 11            CCGGTAGTT--
2. matrix start: 1                      CCGGTAGTT--
3. matrix length: 11                    CCGGTAGTT--
4. orientation: n                       --AACTACCGG
```

BS CATTACAAAATC; R00097; -1; 12;; n.

```
                                       000000000011
                                      -112345678901
Sequence in site entry:                 ATTTTGTAAT
1. positions of gaps: -                 ATTTTGTAAT
2. matrix start: -1                    GATTTTGTAAT
3. matrix length: 12                   GATTTTGTAATG
4. orientation: n                      CATTACAAAATC
```

(Flanking positions were retrieved from linked EMBL entry.)

9. Table REFERENCES

9.1. Contents

This table contains the references from which the data in TRANSFAC(r), TRANSCompel(r), PathoDB(r) and S/MARtDB(TM) were taken. It gives a link to PubMed and lists all links (accession numbers) to those entries (class, factor, gene, ...) from the respective databases, which are described in this reference.

9.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

"RE" + 7-digit number

RX PUBMED ID

the number in PubMed

RA Authors

(NOTE: accents are omitted, German umlauts are transcribed as follows: ä -> ae, ö -> oe, ü -> ue; German "s-z" (ß) -> ss)

RT Reference Title

(NOTE: Greek letters are expanded to alpha, beta, gamma etc.)

RL Source

journal volume:pages (year)

CL TRANSFAC Class

link to transcription factor class

FA TRANSFAC Factor

link to transcription factor

GE TRANSFAC Gene

link to gene

MX TRANSFAC Matrix

link to binding matrix for certain factors

BS TRANSFAC Site

link to DNA (or RNA) sequence shown to be bound by a factor

FR TRANSFAC DNA fragment

link to DNA fragment shown to be bound by a factor in vivo (e.g. ChIP-on-chip experiment)

EV TRANSCompel Evidence

link to evidence for a composite element

DM PathoDB Diagnosis Method

link to method used to screen for a specific mutation within a factor

or a site.

GT PathoDB Genotype

link to genotype of a mutated factor or mutated site

MF PathoDB Mutated Factor

link to mutated transcription factor correlated with a phenotype

MS PathoDB Mutated Site

link to mutated binding site of a transcription factor correlated with a phenotype

PT PathoDB Phenotype

link to phenotype (disease) provoked by a mutation in a transcription factor or its binding site

SM S/MARtDB S/MAR

link to scaffold/matrix attached region

SB S/MARtDB S/MARBinder

link to nuclear matrix protein that was shown to interact with S/MARs

10. TRANSCompel

10.1. Preface

10.1.1. Contents

TRANSCompel is the only database presenting information on combinatorial gene transcriptional regulation and protein-protein interactions between different transcription factors bound to their cognate promoter elements. TRANSCompel database originates from the COMPEL (Kel-Margoulis et. al., 2000) and collects information about composite regulatory elements (CEs) - pairs of closely situated sites and transcription factors binding to them. We define a composite element as a minimal functional unit within which both protein-DNA and protein-protein interactions contribute to a highly specific pattern of gene transcriptional regulation (Kel O.V. et al, 1995, Nucleic Acids Res., 23: 4097-4103). The factors that cooperate at an individual CE mostly belong to different classes with respect to the structure of protein domains, namely DNA-binding and activation domain. The factors also differ in their functional properties: cell-specificity, inducibility and others. Thus, composite regulatory elements contribute to one of the fundamental principles of genome functioning - combinatorial nature of gene transcriptional regulation.

10.1.2. Status

COMPEL has been developed in a joint effort of the two institutes, Institute of Cytology and Genetics in Novosibirsk, Russia, and GBF, Gesellschaft fuer Biotechnologische Forschung mbH (German Research Centre for Biotechnology GmbH) in Braunschweig, Germany. Development of COMPEL was done mainly with financial support from both the Russian and German Research Ministry, SB RAS, and from NATO until the end of 1997. In 1998, the BIOBASE GmbH granted support to update the database and to develop a relational database management system under MS Access. From this RDBMS, ASCII flat files are regularly generated. The COMPEL database has been registrated in the Federal institute of industrial ownership of the Rospatent (Moscow, Russia) as a joint property of the ICG, Russia and GBF, Braunschweig, Germany. Since Aug. 2000, BIOBASE holds distribution rights for the TRANSCompel and provides further updates of the database as well.

10.2. The concept

Composite regulatory elements contain two closely situated binding sites for distinct transcription factors, and actually are minimal functional units, providing cross-coupling of different regulatory pathways. The term "composite element" was introduced while studying glucocorticoid response element in the mouse proliferin promoter where the glucocorticoid receptor binding site is adjacent to an AP-1 site (Diamond et al., 1990). Further, this term was applied to different pairs of interacting sites and factors (Gutman and Wasylyk, 1990; Jackson et al., 1993; Du et al., 1993; Moulton et al., 1994; Rooney et al., 1995, Brass et al., 1996, Klein- Hessling et al., 1996; Butscher et al., 1998, and others). Based on the known examples, we define a composite element as a minimal functional unit within which both protein-DNA and protein-protein interactions contribute to a highly specific pattern of gene transcriptional regulation (Kel,O.V. et al., 1995b; Kel,O.V. et al., 1997).

Composite elements can be classified based on different criteria:

i. character of interactions between transcription factors involved (synergism or antagonism);

ii. structure of transcription factors, namely structure of DNA-binding domains;

iii. function provided by a composite element (tissue-specificity, inducibility,...).

According to (i), there are two main types of composite elements: synergistic and antagonistic ones.

In synergistic CEs, simultaneous interactions of two factors with closely situated target sites result in a non-additively high level of a transcriptional activation. Highly cooperative binding of factors to DNA and formation of a ternary complex protein-protein-DNA was experimentally shown in many cases (Moreno et al., 1995; Brass et al., 1996; Linhoff et al., 1997; Muhlethaler-Mottet et al., 1998; Butscher et al., 1998, and others). As a result of protein-protein interactions, a new protein surface may be formed which is common for a factor pair. Interaction between two factors may be direct (Brass et al., 1996; Chen et al., 1998, and others), or mediated by a co-activator, for instance by p300/CREB-BP (Butscher et al., 1998).

In some cases, two factors independently binding to DNA still synergistically activate transcription (Zaiman and Lenz, 1996; Ohmori et al., 1997; Cantwell et al., 1998). In this case, synergistic effect may be accounted for by simultaneous interactions of activation domains of two factors with different components of the basal transcription complex, and/or direct factor-factor interactions may elicit conformational changes in activation domains. A number of factors are known to bend DNA and thus permit binding of other factors (Stros et al., 1994; Kerppola and Curran, 1991). Hierarchy of transcription factor loading may occur due to assembling of nucleosome-like structures (Linhoff et al., 1997). Some factors bind primarily to DNA and may serve as gathering centers due to sequence similarity with histone-fold motif as in the case of the subunits of NF-Y factor (Linhoff et al., 1997).

Within an antagonistic CE, two factors interfere with each other. In some cases, competition for overlapping sites leads to a mutually exclusive binding (Casolaro et al., 1995; Klein-Hessling et al., 1996; Takeuchi et al., 1998, and others). In other cases, factors can bind to DNA simultaneously, but binding of a repressing factor possibly "masks" an activation domain of an activator (Diamond et al., 1990). A number of molecular mechanisms are suggested for functioning of both synergistic and antagonistic CEs (Kel,O.V. et al., 1997).

To classify composite elements in terms of a factor's DNA-binding domains (ii.) we applied a previously developed transcription factor classification (Wingender, 1997). The factors interacting at an individual CE mostly belong to different classes. Transcription factors of bZIP, REL and ETS classes play a very important role in composite elements and about 50% of known CEs contain at least one binding site for these proteins. In general, transcription factors of these three classes can be characterized as factors inducible by various extracellular stimuli.

Since functional properties and tissue distribution of factors vary significantly within the same factor class, another criterion (iii.) for classification is suggested based on specific function provided by a CE (Kel,O.V. et al., 1997). Cross-coupling between structurally and functionally different factors on composite regulatory elements seems to be a general regulatory pathway. In fact, it opens up a broad possibility for coding very specific gene expression profiles in the structure of gene regulatory regions.

CEs can be divided into several groups:

1. CEs formed by binding sites for two inducible factors, they provide cross-coupling of signal transduction pathways;

2. CEs formed by binding sites for a tissue-enriched and an inducible factor, they provide tissue-specific responses to inducing signals;

3. CEs formed by binding sites for a tissue-enriched and a constitutive ubiquitous factor, they provide some additional features of the tissue-specific transcriptional regulation;

4. CEs formed by binding sites for an inducible and a constitutive ubiquitous factor, they provide some additional features of the inducible regulation;

5. CEs formed by binding sites for two tissue-enriched factors, they provide some particular aspects of tissue-specific regulation.

References

Brass A. L., Kehrli E., Eisenbeis C. F., Storb U. and Singh H. (1996) "Pip, a lymphoid-restricted IRF, contains a regulatory domain that is important for autoinhibition and ternary complex formation with Ets factor PU.1" Genes Dev. 10, 2335-2347.

Butscher W. G., Powers C., Olive M., Vinson C. and Gardner,K. (1998) "Coordinate transactivation of the interleukin-2 CD28 response element by c-Rel and ATF-1/CREB2" J. Biol. Chem. 273, 552-560.

Cantwell C. A., Sterneck E. and Johnson P. F. (1998) "Interleukin-6- specific activation of the C/EBPdelta gene in hepatocytes is mediated by Stat3 and Sp1" Mol. Cell. Biol. 18, 2108-2117.

Casolaro V., Georas S. N., Song Z., Zubkoff I. D., Abdulkadir S. A., Thanos D. and Ono S. J. (1995) "Inhibition of NF-AT-dependent transcription by NF-kappaB: implications for differential gene expression in T helper cell subsets" Proc. Natl. Acad. Sci. USA 92, 11623-11627.

Chen L., Glover J. N. M., Hogan P. G., Rao A., and Harrison S. C. (1998) "Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA" Nature 392, 42-48.

Diamond M. I., Miner J. N., Yoshinaga S. K. and Yamamoto K. R. (1990) "Transcription factor interactions: selectors of positive or negative regulation from a single DNA element" Science 249, 1266-1272.

Du W., Thanos D., and Maniatis T. (1993) "Mechanism of transcriptional synergism between distinct virus-inducible enhancer elements" Cell 74, 887-898.

Gutman A. and Wasylyk B. (1990) "The collagenase gene promoter contains a TPA and oncogen responsive unit encompassing the PEA3 and AP-1 binding sites" EMBO J. 9, 2241-2246.

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 transcription regulation: TRANSFAC,TRRD and TRANSCompel" Nucleic Acids Res. 26, 362-367.

Jackson D. A., Rowader K. E., Stevens K. Y., Jiang C., Milos P. and Zaret K. S. (1993) "Modulation of liver-specific transcription by interaction between Hepatocyte Nuclear Factor 3 and Nuclear Factor 1 binding DNA in close apposition" Mol.Cell.Biol. 13, 2401-2410.

Karas H., Kel A. E., Kel O. V., Kolchanov N. A., and Wingender E. (1997) "Integrating the knowledge on gene regulation" Mol. Biol. (Mosk). 31, 531-539.

Kel A. E., Kolchanov N. A., Kel O. V., Romashencko A. G., Ananko E. A., Ignatieva E. V., Merkulova T. I., Podkolodnaya O. A., Stepanenko I. L., Kochetov A. V., Kolpakov F. A., Podkolodnyi N. L., and Naumochkin A. N. (1997) "TRRD: Database on Transcription Regulatory Regions of Eukaryotic Genes" Mol. Biol. (Mosk). 31, 626-636.

Kel A., Kel-Margoulis O., Babenko V. and Wingender E. "Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells" J. Mol. Biol. 288, 353-376.

Kel O. V., Romaschenko A. G., Kel A. E., Naumochkin A. N. and Kolchanov N. A. (1995a) "Data representation in the TRRD - a database of transcription regulatory regions of the eukaryotic genomes" Proceedings of the 28th Annual Hawaii International Conference on System Scienses [HICSS]. Biotechnology Computing, IEE Computer Society Press, Los Alamitos, California, 5, 42 -51.

Kel O. V., Romaschenko A. G., Kel A. E., Wingender E. and Kolchanov N. A. (1995b) "A compilation of composite regulatory elements affecting gene transcription in vertebrates" Nucleic Acids Res. 23, 4097-4103.

Kel O. V., Romaschenko A. G., Kel A. E., Wingender E. and Kolchanov N. A. (1997) "Composite regulatory elements: classification and description in the TRANSCompel database" Mol. Biol. (Mosk). 31, 498-512.

Kerppola T. K. and Curran T. (1991) "Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: implications for transcription factor cooperativity" Cell 66, 317-326.

Klein-Hessling S., Schneider G., Heinfling A., Chuvpilo S. and Serfling E. (1996) "HMG I(Y) interferes with the DNA binding of NF-AT factors and the induction of the interleukin 4 promoter in T cells" Proc. Natl. Acad. Sci. USA 93, 15311-15316.

Kolchanov N. A., Ananko E. A., Podkolodnaya O. A., Ignatieva E. V., Stepanenko I. L., Kel-Margoulis O. V., Kel A. E., Merkulova T. I., Goryachkovskaya T. N., Busigina T. N., Kolpakov F. A., Podkolodny N. L., Naumochkin A. N., Romashchenko A. G. (1999) "Transcription regulatory regions database (TRRD): its status in 1999" Nucleic Acids Res. 27, 303-306.

Linhoff M. W., Wright K. L. and Ting J. P.-Y. (1997) "CCAAT-binding factor NF-Y and RFX are required for in vivo assembly of a nucleoprotein complex that spans 250 base pairs: the invariant chain promoter as a model" Mol. Cell. Biol. 17, 4589-4596.

Moreno C. S., Emery P., West J. E., Durand B., Reith W., Mach B. and Boss M. (1995) "Purified X2 binding protein (X2BP) cooperatively binds the class II MHC X box region in the presence of purified RFX, the X box factor deficient in the Bare Lymphocyte Syndrom" J. Immunol. 155, 4313-4321.

Moulton K. S., Semple K., Wu H. and Glass C. K. (1994) "Cell-specific expression of the macrophage scavenger receptor gene is dependent on PU.1 and composite AP-1/ets motif" Mol. Cell. Biol. 14, 4408-4418.

Muhlethaler-Mottet A., Berardino W. D., Otten L. A. and Mach B. (1998) "Activation of the MHC Class II Transactivator CIITA by interferon-gamma requires cooperative interaction between Stat1 and USF-1" Immunity 8, 157-166.

Ohmori Y., Schreiber R. D., Hamilton T. A. (1997) "Synergy between interferon-gamma and tumor necrosis factor alpha in transcriptional activation is mediated by cooperation between Signal Transducer and Activator of Transcription 1 and Nuclear Factor kappaB" J. Biol. Chem. 272, 14899-14907.

Quandt K., Frech K., Karas H., Wingender E. and Werner T. (1995) "New fast and versatile tools for detection of consensus matches in nucleotide sequence data" Nucleic Acids Res. 23, 4878-4884.

Perier R. C., Junier Th. and Bucher P. (1998) "The Eukaryotic Promoter Database EPD" Nucleic Acids Res. 26, 353-357.

Rao A., Luo C. and Hogan P. G. (1997) "Transcription factors of the NFAT family: regulation and function" Annu. Rev. Immunol. 15, 707-747.

Rooney J. W., Sun Y.-L., Glimcher L. H. and Hoey T. (1995) "Novel NFAT sites that mediate activation of the interleukin-2 promoter in response to T-cell receptor stimulation" Mol. Cell. Biol. 15, 6299-6310.

Stros M., Stokrova J. and O'Thomas J. (1994) "DNA looping by the HMG-box domains of HMG1 and modulation of DNA binding by the acidic C-terminal domain" Nucleic Acids Res. 22, 1044-1051.

Takeuchi A., Reddy G. S., Kobayashi T., Okano T., Park J. and Sharma,S. (1998) "Nuclear Factor of Activated T cells (NFAT) as a molecular target for 1alpha,25-dihydroxyvitamin D3-mediated effects" J. Immunol. 160, 209-218.

Wingender E. (1997) "Classification scheme of eukaryotic transcription factors" Mol. Biol. (Mosk). 31, 483-497.

Wingender E., Kel A. E., Kel O. V., Karas H., Heinemeyer T., Dietze P., Kn�ppel R., Romaschenko A. G. and Kolchanov N. A. (1997) "TRANSFAC, TRRD and TRANSCompel: Towards a federated database system on transcriptional regulation" Nucleic Acids Res. 25, 265-268.

Zaiman A. L. and Lenz J. (1996) "Transcriptional activation of a retrovirus enhancer by CBF (AML1) requires a second factor: evidence for cooperativity with c-Myb" J. Virol. 8, 5618-5629.

10.3. Structure of TRANSCompel®

There are two tables in the TRANSCompel database: Compel and Evidence.

The Compel table (for details see compel.txt) contains general information about composite elements including sequence, positions, gene where the CE was experimentally studied, comments on specific transcriptional regulation provided by this CE. Brief information about experimental evidences confirming functional and physical interactions between corresponding transcription factors is given as well. Each entry to this table corresponds to an individual composite element in a particular eukaryotic gene.

The Evidence table (for details see evidence.txt) provides detailed information on experiments that have been undertaken to confirm cooperative functioning of transcription factors: the type of experiment, conclusion, cell type, individual transcription factors involved, and reference.

These two tables are closely interrelated. From the Compel table you can switch to the Evidence table through the evidence accession number. From the Evidence table you can retrieve a composite element entry through the composite element accession number.

10.4. Table COMPEL

10.4.1. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

Accession number of a composite element

ID Identifier

Identifier of a CE. It is the unique characteristic of each composite element. The identifier is composed of brief names of the DNA binding domains of two corresponding transcription factors, separated by the symbol "$", and the consecutive number of elements of this type. For example, a CE contains binding sites for AP-1 (bZIP type of DNA binding domain) and NF-kappaB (REL) transcription factors. In this case the identifier looks like ZIP$ETS_001. (For a list of abbreviations, please see below.)

DT Created/Updated

date of entry creation; entry author / date of last entry updating; updater

CO Copyright information

Copyright (C), BIOBASE GmbH

GE TRANSFAC gene

Gene identifier; brief gene name; species of a gene

SQ Sequence

Sequence of a CE. Sequences of two binding sites (as sites are defined in the corresponding article) are shown in capital letters, the sequence of the spacer between the sites in small letters. When the spacer is too long, one part of it is replaced by points.

ST Reference point for sequence start

Point relative to that the positions are given

PS Position

Positions of a CE. Positions correspond to the first and last nucleotides of the sequence.

DR Link to external databases: EMBL

EMBL accession number and identifier are given. EMBL position corresponds to the first 5' nucleotide of the sequence shown.

NO Number of sites

Number of individual binding sites that constitute a composite element: 2 or 3.

BS Binding site

Position of the DNA binding site; name of the corresponding transcription factor; accession number of this site. This field is repeated two times for CE's containing two sites and three times for CE's containing three sites.

TY Type (synergism/antagonism)

Type of a CE (synergism, antagonism). This field reflects the type of functional cooperation between transcription factors and contains one of two possible meanings: synergism or antagonism. Synergism means that two factors synergistically activate transcription. Antagonism means two factors interfere with each other. Within this particular CE, one of the factors acts as an activator of transcription, another as a repressor.

CL Functional classification

Functional classification of the CE's (for instance, one of the factors is inducible and another is tissue-restricted). Criteria for this classification are suggested based on specific function provided by a CE. (For classification, please see below.)

CG Compel group

Name of a compel group. Synergistic composite elements consisting of two binding sites for TFs are classified into the groups of structural and functional similarity. CEs within one group contain binding sites for the same factor families.

CM Molecular mechanism

Molecular mechanisms of the functioning of a CE.

For synergistic CE's:
cooperative binding of factors to DNA,
interactions with common co-activator,
interactions with basal transcription complex,
possible role of nucleosome, DNA bending by one of the factors.

For antagonistic CE's:
competition for DNA binding site,
competition for a common co-activator.

CC Comment

Comments to this CE: for instance, role of this CE in transcriptional regulation of a given gene in a particular cellular content.

EV Evidence

Accession number of the evidence for this CE. Evidence for a given CE is an experiment of a certain type carried out with two individual interactions within a particular cell type.

EX Experiment type

Type of an experiment

CN Experiment conclusion

Conclusion drawn on the basis of the experiment

CT Cell line

Acc. Number (link to the TRANSFAC Cell Table); short description

RN Reference number

Reference number

RA Author(s)

Authors

RT Title

Reference title

RL Journal

Journal volume: pages (year)

10.4.2. CE Identifier

The identifier of each composite element (CE) contains a short denomination of the DNA-binding domain of the involved factor(s):

C2H2 C2H2 zinc finger type: Sp1, YY1, Egr, RFLAT-1

C4 C4 zinc finger type of nuclear receptors: GR, PR, ER, RAR, T3R, VDR, COUP, HNF-4, SF-1

GATA C4 zinc finger GATA type

ETS HTH, Tryptophan clusters, ETS family

HLH bHLH (E2A, MyoD, myogenin);
bHLH-ZIP (TFE3, USF, SREBP, c-Myc, AhR/Arnt, HIF-1, EBF)

HSH bHSH: AP-2

HMG beta-Scaffold Factors with Minor Groove Contacts, HMGI(Y), Sox9

HOM HTH, homeo domain only (HNF-1, Pbx, Nkx); LIM-homeodomain (lmx-1); paired box (Pax); POU-domain (Oct, Pit)

MADS beta-Scaffold Factors with Minor Groove Contacts, MADS box factors: MEF-2 and SRF families

MYB HTH, Tryptophan clusters, Myb family; HTH, Tryptophan clusters, IRF family

NF1 beta-Scaffold Factors with Minor Groove Contacts, Smad/NF-1, NF-1 factors

NFY	beta-Scaffold Factors with Minor Groove Contacts, heteromeric CCAAT factors: NF-Y family
P53	beta-Scaffold Factors with Minor Groove Contacts, p53 family
REL	beta-Scaffold Factors with Minor Groove Contacts, NF-kappaB and NF-AT families
RUNT	beta-Scaffold Factors with Minor Groove Contacts, AML/PEBP family
SMAD	beta-Scaffold Factors with Minor Groove Contacts, Smad/NF-1, SMAD family
STAT	beta-Scaffold Factors with Minor Groove Contacts, STAT family
TEA	HTH, TEA domain factors, TEF-1
UNCL	DNA binding domain is unclassified (HAF, DPBF)
WH	HTH, winged helix/fork head (HNF-3, E2F, FAST, RFX families)
ZIP	bZIP, AP-1 family, ATF/CREB family, C/EBP family

10.4.3. Functional classification of CEs

Inducible/constitutive CEs

These are CEs that are formed by binding sites for an inducible and a constitutive ubiquitous factor, providing some additional features of the inducible regulation.

Presently known types are:

acute-phase response
cholesterol level response
EGF response in keratinocytes
estrogen response
glucocorticoid response
hypoxia response
IFN-gamma response
IL-1beta response
immune response
keratinocytes
serum growth factor response
steroid hormone response
TGF-beta response
TPA response in keratinocytes
TPA response in smooth muscle cells
viral infection response
convergence of the TGF-beta and Wnt signalling

Inducible/tissue-restricted CEs

CEs that are formed by binding sites for a tissue-enriched and an inducible factor provide tissue-specific responses to inducing signals.

Known CEs of this type functional in:

adipocytes
B-cells
differentiating osteoblasts

hypoxia response, endothelial cells
hypoxia response, liver and kidney cells
kidney cells
liver cells
myeloid cells
pituitary cells
pituitary gonadotropes
pituitary lactotropes
pituitary somatotropes
pituitary thyrotropes
PKA and cAMP response
PTH response in bone cells
Ras/Raf response
T-cells
TGF-beta response in B-cells

Tissue-restricted/tissue-restricted CEs

CEs which are formed by binding sites for a tissue-enriched and a constitutive ubiquitous factor provide some additional features of the tissue-specific transcriptional regulation.

Examples for this kind of CE have been found functional in:

B-cells
intestine, enterocytes
liver cells
muscle cells
myeloid cells
pancreas islet beta-cells (insulin-producing)
pituitary cells
pituitary gonadotropes
T-cells

Tissue-restricted/ubiquitous CEs

These are CEs that are formed by binding sites for a tissue-enriched and a constitutive ubiquitous factor; they provide some additional features of the tissue-specific transcriptional regulation

B-cells
endothelial cells
glia cells
intestine, enterocytes
liver cells
lymphoid cells
macrophages
mammary gland cells
muscle cells
myeloid cells
neuronal cells
ovary and adrenal cortical cells
pancreas islet beta-cells (insulin-producing)
pituitary cells

pituitary lactotropes
pituitary somatotropes
pituitary thyrotropes
placenta cells

10.5. Table EVIDENCE

10.5.1. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

Accession number of a composite element

DT Created/Updated

date of entry creation; entry author /

date of last entry updating; updater

CO Copyright information

Copyright (C), BIOBASE GmbH

CE Composite element

Acc. number and identifier for the corresponding composite element.

EX Experiment type

Type of an experiment (for a list of experiments, please see below, Section 10.5.2)

CN Experiment conclusion

Conclusion drawn on the basis of the experiment

CT Cell line

Acc. Number (link to the TRANSFAC Cell Table); short description

CC Evidence comment

Comments for this evidence. For functional synergism between sites and factors (experiment number 1.1, 3.1, 5.1), quantitative data are provided. For mutational analysis of factors (experiment number 3.3, 4.2, 10.2), protein domains responsible for cooperation are indicated.

IN Interaction

First interaction (accession number of an interaction; positions of the binding site). Interaction corresponds to an individual DNA-protein interaction (binding of an individual transcription factor of a certain origin to its binding site within a CE).

FA Factor

Factor binding to this site (factor name; species; origin)

FD DNA-binding domain

Factor DNA-binding domain

FF Effect on transcription

Effect on the transcription of a particular gene in a given cellular situation: activation or repression.

FC Factor comment

Comments on functional properties of this factor (inducibility, tissue-specificity).

DR Link to external databases: TRANSFAC

Accession number and brief name of the factor in TRANSFAC. This field may be repeated within the same entry.

RN Reference number

Reference number

RA Author(s)

Authors

RT Title

Reference title

RL Journal

Journal volume: pages (year)

10.5.2. Experiments

In the following, experimental approaches are listed that are used to confirm cooperative interactions between transcription factors within composite elements, and the type of conclusions that can be drawn from the particular experiment.

Experiment Acc	Type of experiment	Conclusion drawn on the base of the experiment
1.0		Functional synergism between sites
1.1	Site-directed mutagenesis and study of promoter activity	Functional synergism between sites
1.2	Insertion or deletion of spacer nucleotides between sites and study of promoter activity	Spacing between the sites is important
1.3	This CE is shown to transfer specific properties to a heterologous promoter	Functional synergism between sites
2.1	Site-directed mutagenesis and study of promoter activity	Functional antagonism between sites
3.1	Transient co-transfections	Functional synergism between factors
3.2	Stable co-transfections	Functional synergism between factors
3.3	Co-transfection by plasmids expressing mutated derivatives of the factor(s)	Functional synergism between factors
4.1	Co-transfection experiment	Functional antagonism between factors
4.2	Co-transfection of plasmids expressing mutated derivatives of the factor(s)	Functional antagonism between factors
5.1	Co-transfection experiment combined with site-directed mutagenesis	Both intact sites as well as both factors are essential for synergism
6.0		Co-operative binding of factors to DNA
6.1	Gel mobility shift combined with site-directed mutagenesis	Co-operative binding of factors to DNA
6.2	Methylation interference assay	Co-operative binding of factors to DNA
6.3	DNAse I footprinting analysis	Co-operative binding of factors to DNA
6.4	Gel mobility shift in the presence of oligonucleotide competitors	Co-operative binding of factors to DNA, ternary complex formation
6.41	Gel mobility shift in the presence of oligonucleotide competitors	Co-operative binding of factors to DNA, higher-order complex formation
6.5	Gel mobility shift in the presence of a single factor or both factors together	Co-operative binding of factors to DNA, ternary complex formation
6.51	Gel mobility shift in the presence of a	Co-operative binding of factors to DNA, higher-order

	single factor or both factors together	complex formation
6.6	Cu2+ -phenanthroline footprinting assay	Co-operative binding of factors to DNA, ternary complex formation
6.7	In vivo chromatin immunoprecipitation assay (CHIP)	Co-operative binding of factors to DNA
6.8	Quantitative analysis of DNA binding	Co-operative binding of factors to DNA
6.9	DNA affinity binding assay combined with site-directed mutations	Co-operative binding of factors to DNA
7.0		Competition between TFs for overlapping binding sites
7.1	EMSA	Competition for overlapping binding sites
8.0		Mutually exclusive binding of factors to DNA
8.1	Gel mobility shift	Mutually exclusive binding of factors to DNA
8.2	Methylation interference assay	Mutually exclusive binding of factors to DNA
9.0		Ternary complex formation between two factors and DNA
9.1	Gel mobility shift using increasing amounts of one of the factors	Ternary complex formation between two factors and DNA
9.2	Study of half-dissociation time of the complexes	Ternary complex formation between two factors and DNA
9.21	Study of half-dissociation time of the complexes	High-order complex formation between factors and DNA
9.3	Inspection of protein-DNA complex with antibodies	Ternary complex formation
9.31	Inspection of protein-DNA complex with antibodies	Higher-order complex formation
9.4	Gel mobility shift combined with mutational analysis of the factor(s)	Ternary complex formation
9.41	Gel mobility shift combined with mutational analysis of the factor(s)	High-order complex formation
10.0		Direct factor-factor interactions in the absence of DNA
10.1	In vitro cross-linking between factors	Direct factor-factor interactions in the absence of DNA
10.2	Mutational analysis of factor(s)	Direct factor-factor interactions in the absence of DNA
10.3	In vivo cross-linking between factors	Direct factor-factor interactions in the absence of DNA
11.1	Over-expression of a co-activator restores transcription level	Competition between TFs for a transcriptional co-activator
11.2	In vivo cross-linking between factors and co-activator	Simultaneous interactions of factors with co-activator
12.1	Fluorescence architectural analysis	Cooperative DNA bending
13.1	Crystallization	Detailed structure of a ternary complex

10.6. Citation

How to cite TRANSCompel:

Matys, V.; Kel-Margoulis, O.; Fricke, E.; Liebich, I.; Land, S.; Barre-Dirrie, A.; Reuter, I.; Chekmenev, D.; Krull, M.; Hornischer, K.; Voss, N.; Stegmaier, P.; Lewicki-Potapov, B.; Saxel, H.; Kel, A. and Wingender, E. (2006) "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes" Nucleic Acids Res. 34, D108-D110.

10.7. Publications

Matys V., Kel-Margoulis O., Fricke E., Liebich I., Land S., Barre-Dirrie A., Reuter I., Chekmenev D., Krull M., Hornischer K., Voss N., Stegmaier P., Lewicki-Potapov B., Saxel H., Kel A., Wingender E. (2006) "TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes" Nucleic Acids Res. 34, D108-D110

Kel-Margoulis O., Matys V., Choi C., Reuter I., Krull M., Potapov A. P., Voss N., Liebich I., Kel A., Wingender E. (2005) "Databases on Gene Regulation" In: "Information Processing And Living Systems", eds.: Bajic, V. B. and Wee, T. T., World Scientific Publishing Co.

Kel-Margoulis O., Kel A. E., Reuter I., Deineko I. V., Wingender E. (2002) "TRANSCompel® - a database on composite regulatory elements in eukaryotic genes" Nucleic Acids Res. 30, 332-334

Kel-Margoulis O., Deineko I. V., Reuter I., Wingender E., Kel A. E. (2001) "TRANSCompel® - a professional database on composite regulatory elements in eukaryotic genes" Proceedings of the German Conference on Bioinformatics GCB'01. Braunschweig, Germany, October 7-10, 2001, 185-187

Kel-Margoulis O. V., Romaschenko A. G., Kolchanov N. A., Wingender E. and Kel A. E. (2000) "COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation" Nucleic Acids Res. 28, 311-315

Heinemeyer T., Chen X., Karas H., Kel A. E., Kel O. V., Liebich I., Meinhardt T., Reuter I., Schacherer F., Wingender E. (1999) "Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms" Nucleic Acids Res. 27, 318-322

Kel-Margoulis O. V., Kel A. E., Frisch M., Romaschenko A. G., Kolchanov N. A. and Wingender E. (1998) "COMPEL - a database on composite regulatory elements" Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure, (BGRS'98), ICG, Novosibirsk, Vol.1, 54-57

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 transcription regulation: TRANSFAC, TRRD and COMPEL" Nucleic Acids Res. 26, 362-367

Kel O. V., Kel A. E., Romaschenko A. G., Wingender E., and Kolchanov N. A. (1997) "Composite regulatory elements: classification and description in the COMPEL database" Mol. Biol. (Mosk). 31, 498-512

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

Kel O. V., Romaschenko A. G., Kel A. E., Wingender E. and Kolchanov N. A.(1995) "A compilation of composite regulatory elements affecting gene transcription in vertebrates" Nucleic Acids Res. 23, 4097-4103

11. TRANSPro

11.1. Contents

For a number of analyses promoter sequences are required, those sequences being located just upstream of the TSSs ('Transcription Start Sites') of a primary transcript for a given gene. This includes the development of tools for the prediction of promoters as well as all analyses of gene regulation in gene expression experiments. The extraction of reliable promoter sequences, however, is a time-consuming task, usually including a huge amount of tedious handwork. To meet these needs the TRANSPro™ database was created, a database module within the TRANSFAC® Suite. It contains upstream (5') sequences of human, mouse, rat, dog, chimp, Plasmodium, Arabidopsis and rice genes, together with extensive annotation, the stress obviously lying on elements involved in the process of gene regulation. With the help of BIOBASE tools such as Match™, Patch™, and Catch® the promoter sequences can be searched for potential regulatory elements. The TRANSFAC® Suite now contains everything necessary for the analysis of gene expression data by searching potential regulatory elements in the promoters of co-expressed genes.

For the creation of TRANSPro™, the Genomic Sequence Assemblies created by the international sequencing consortia have been used, extracted from either the EnsEMBL database, or from sequences taken from the NCBI which describe entire chromosomes for those species which are not offered in the EnsEMBL database. Promoter sequences are extracted only for those genes for which an Entrez ID is defined. Genes on mitochondria and chloroplasts are excluded, due to their special modes of transcription.

The calculation of 'virtual TSSs' as reference points for the promoter extraction is based on a collection of TSSs for a given gene. TSSs are taken from EPD, dbTSS, Fantom and EnsEMBL. dbTSS TSSs are assumed to be the first nucleotide of the one-pass mRNA sequences, EnsEMBL TSSs to be the first nucleotide of the most 5' exon of an EnsEMBL mRNA model. Thus, collected TSSs for a given gene are located on a sequence fragment which sometimes spans several thousand nucleotides, in some cases far more than 100 kb. They are frequently not located in tight clusters of only a few dozen nucleotides length, but are often widespread throughout the sequence. In order to define a reasonable number of 'virtual TSSs' for a given gene from this data collection, an algorithm was designed, which applies a set of rules to the data collection in order to find 'clusters' of TSSs. A window of 3000 nt length is slid along the entire sequence fragment. A 'clustering score' is calculated by summing up weighted contributions from each TSS in the window. Each TSS derived from an EnsEMBL mRNA model or an mRNA model extracted from an EMBL/GenBank/DDBJ file (e.g. for Arabidopsis) is scored with 5 evidence points, TSSs derived from a dbTSS one-pass mRNA or from a Fantom sequence are scored with 6 evidence points. We assume EPD TSSs (due to the fact that they are hand- annotated) to have a higher reliability and give 20 evidence points each. The weights of evidence points are additionally multiplied by a distance score: the central position is multiplied by 1, the outer positions are multiplied by 0, and all positions in between by a value taken from a cosine function, according to the distance from the center of the window. The peaks of the resulting clustering score are regarded as potential 'virtual TSSs'. For some of the genes only a handful of evidence points are available, thus resulting in multiple 'virtual TSSs', each consisting of only a few evidence points. Therefore, for all those genes where less than 19 evidence points are available only the most 5' 'virtual TSS' is accepted. For all other genes those peaks are accepted as 'virtual TSSs' for which the respective sequence window contains at least 8% of all evidence points. However, there are genes, for which - although the coverage with data is pretty good - the annotated TSSs are so equally distributed along the sequence, that no prominent peaks occur, and therefore - according to the above mentioned rules - no peak would be accepted. In this case the most prominent peaks are accepted. If there are more than two peaks for which these conditions are true, the most 5' 'virtual TSS' is accepted. The collection of 'virtual TSSs' prepared in this way is the basis for the extraction of the TRANSPro^{TM} promoter sequences. The calculation of 'virtual TSSs' and the subsequent data extraction are fully automated processes; whenever conflicts or inconsistencies occur the respective gene is excluded from the TRANSPro^{TM} database.

A widespread technique in the promoter analysis is the comparison of synthenic promoters of different species. The annotation and/or prediction of 'Transcription Factor Binding Sites' reveals sites which might be conserved in both species or are altered or possibly even lacking in one of them. This gain of knowledge is then the starting point for deeper understanding of the promoter structure and function. To facilitate this kind of analyses a concept was introduced into the TRANSPro^{TM} database which we call the 'Anchor Points'. Prominent points, e.g. TSSs from the EPD or 'virtual TSSs', are defined as 'Primary Anchors'. Fragments are then extracted from the genomic sequence builds, spanning 300 nt upstream and 100 nt downstream of the 'Primary Anchor'. From the EnsEMBL database homology information is extracted to find the genes which are orthologous, resp. paralogous to the one for which the 'Anchor Point' is defined. For those 'Homologous Genes' a sequence fragment is extracted which comprises the entire range containing annotation for exons and TSSs, plus an additional 50,000 nt upstream of the fragment, which is regarded as the gene sequence. Using the best BLAST match of the 'Anchor Point' fragment on the sequence of the 'Homologous Gene' as a starting point, a CLUSTALW alignment is tried for the 'Anchor Point' fragment, not with the entire 'Homologous Gene' sequence, but for the best BLAST fragment, plus an additional 400 nt on both sides. For the alignments two homology values are calculated - the usual one, in which the fraction of matches is calculated, but as well a 'refined homology', in which the contribution of long insertions is regarded to be 4 mismatches only, thus reducing its impact on the entire homology value. Alignments are accepted if the 'refined homology' is better than 50%. On the 'Homologous Gene' sequence the position corresponding to the 'Anchor Point' is calculated and annotated as the 'Secondary Anchor'. During the extraction of the TRANSPro^{TM} sequences pairs of 'Primary Anchors' and their corresponding 'Secondary Anchors' are extracted. The entire sequences of the respective TRANSPro^{TM} entries are aligned to each other using again CLUSTALW and stored in the TRANSPro^{TM} database.

Each entry in the TRANSPro^{TM} database corresponds to a particular promoter of a human, mouse, rat, dog, chimp, Plasmodium, Arabidopsis or rice gene and contains a gene symbol, a gene description, synonyms, and chromosomal location. Sometimes several alternative promoters for one gene are stored in TRANSPro^{TM} in separate entries, each having an individual accession number. For each of the promoter sequences, TRANSPro^{TM} contains 10,000 nt upstream and 1,000 nt downstream relative to the accepted 'virtual TSS'. TRANSPro^{TM} entries contain a wealth of links to external databases, including accession numbers from Entrez Gene, EnsEMBL, the respective nomenclature system (HGNC for human, MGI for mouse, RGNC for rat), as well as links to RefSeq, UniGene, UniProt and Affymetrix. If available, a sequence entry has a link to the respective TRANSFAC® gene. A number of features is annotated if present on the promoter sequence, with their relative positions on the promoter and - if possible - links to the respective data base, including 'Transcription Factor Binding Sites' (from TRANSFAC®), 'Matrix Attachment Regions' (S/MARs) (from the S/MARt™ Database), fragments defined by ChIP-on-Chip or ChIP-Seq experiments with links to the PubMed ID of the paper in which they are described, 'CpG Islands' calculated according to the algorithm of Wang & Leung (PMID:14764558), SNPs from dbSNP, and repeats (e.g. LINE, LTR, etc.). If available the 'Anchor Points' being annotated on a given TRANSPro™ entry are displayed, having a link to a CLUSTALW alignment of the entry with the homologous promoter sequence described through the 'Anchor Point'. In the online version of TRANSPro™, in the BKL Promoter Report, the features being annotated in the TRANSPro™ entries are visualised on the sequence using a coloring schema.

The online version of TRANSPro™ in the BKL interface allows for the selected promoters of an Advanced Search result to retrieve any sequence window around the 'virtual TSS' by indicating the desired relative positions in FASTA format, for analysis of the sequences, e.g. with the Match™ tool, either online or from command line.

11.2. Fields

It should be noted that in individual entries some fields may be empty. In this case, these fields are not displayed.

Field Content and format

AC Accession number

The accession number is unique within TRANSPro(TM) and was created in the form @{species}_@{gene_number}. If two or more promoter sequences per gene exist, an index number, separated by an underscore, is added. The gene number is an internal number with no reference to other databases.

ID Identifier

The identifier is unique within TRANSPro(TM) and was created in the form @{species}_@{gene_number}_@{gene_symbol}. If two or more promoter sequences per gene exist, an index number, separated by an underscore, is added. The gene number is an internal number with no reference to other databases.

GS Gene symbol

Nomenclature gene symbol (HGNC for human, MGI for mouse, RGNC for rat).

DT Date

Date of entry creation.

DE Description

A short description of the gene.

SY Synonyms

Other acronyms or names for the gene.

OS Species

Biological species.

CH Chromosomal location

Chromosome and locus of the gene.

CC Comment

In the current version of TRANSPro(TM) two comment lines are defined for each entry: One of the lines contains a definition of the sequence fragment on the genome assembly, with build info, chromosome, absolute positions, and strand; the other line notes the evidence score points, in absolute numbers as well as the percentage of points accumulated for the TRANSPro(TM) entry.

FT Features

Definitions for those features being located on the genomic sequence fragment displayed in the TRANSPro(TM) entry.

TSSs (from EPD, DBTSS and EnsEMBL): If several TSSs from the same source database are located at the same position, the number of matching TSSs in brackets is added to the accession number.

TRANSFAC(r) SITES and CHIP-ON-CHIP fragments: In the flat file the first position can be higher than the last position, which then means that the site is located on the reverse strand. In the web representation of the data the first position is always smaller than the last position; strand information is given through a '(FORWARD)' or '(REVERSE)' tag at the end of the data line.

Syntax of the features (square brackets '[..]' represent an optional information):

TSS: EPD; @{EPD_accno}; @{position}[(@{number_of_occurrence})].

TSS: dbTSS; @{dbTSS_accno}; @{position}[(@{number_of_occurrence})].

TSS: EnsEMBL; @{ensembl_accno}; @{position}[

(@{number_of_occurrence})].

TSS: Fantom; @{fantom_accno}; @{position}[(@{number_of_occurrence})].

TSS: Build; @{source_accno}; @{position}[(@{number_of_occurrence})].

TRANSFAC SITE: @{transfac_site_accno}; @{first_position}..@{last_position} (@{strand}).

S/MAR: @{smart_accno}; @{first_position}..@{last_position}.

CPG ISLAND: @{first_position}..@{last_position}; (%GC=@{gc_contents}, o/e=@{observed_to_expected_ratio}, #CpGs=@{number_of_cpgs}).

CHIP-ON-CHIP FRAGMENT: @{binding_factor}; @{PubMed_accno}; @{first_position}..@{last_position}.

REPEAT: @{repeat_name}; @{first_pos_of_consensus}..@{last_pos_of_consensus}; @{first_position}..@{last_position}.

SNP: @{dbSNP_accno}; @{variation}; @{position}.

DR External database links

The accession number of the corresponding entry in the respective nomenclature database (HGNC for human, (MGI for mouse, and RGNC for rat), the Entrez Gene ID, the EnsEMBL Gene ID, links to RefSeq, UniGene, UniProt and Affymetrix and the TRANSFAC(r) gene entry.

Syntax:

HGNC: @{hgnc_accno}; @{hgnc_symbol}

MGI: @{mgi_accno}; @{hgnc_symbol}

RGNC: @{rgnc_accno}; @{hgnc_symbol}

ENTREZGENE: @{entrez_accno}

ENSEMBL: @{ensembl_gene_accno}

REFSEQ: @{refseq_accno}

UNIGENE: @{unigene_accno}

UNIPROT: @{uniprot_id}

AFFYMETRIX: @{affy_probe}

TRANSFAC GENE: @{transfac_gene_acc}; @{transfac_gene_symbol}

SQ Sequence

The sequence comprises 10,000 nucleotides upstream and 1,000 nucleotides downstream relative to the calculated virtual TSSs (cf. TRANSPro(TM) documentation).

12. Statistics

12.1. TRANSFAC General

General	Total Number of Entries	
Table	Release 2011.1	Release 2012.1
Factor, total	17,587	18,614
transcription factors	17,182	18,145
miRNA	405	469
Site, total	32,033	35,493
DNA sites	31,129	34,237
RNA sites	904	1,256
ChIP fragment	1,549,846	2,332,432
Gene, total	68,810	71,832
Gene, linked*	59,233	59,517
Matrix	1,455	2,173
Cell	6,662	7,749
Class	57	57
Method	121	129
Reference	25,191	27,331

Genes that are linked to a site, a ChIP-chip fragment and/or a factor

12.2. TRANSFAC Factors

Factor	Number of Entries	
Entries discriminated by taxa*	Release 2011.1	Release 2012.1
Vertebrata	11,833	12,524
Homo sapiens	5,018	5,311
Insecta	398	505
Plants	4,587	4,743
Fungi	686	914
S. cerevisiae	492	503
Viridae	55	61
Nematoda	227	274

The number of factors listed here exceeds the total number of factors because multi-species transcription factor complexes can add to the count in several taxa.

Expression Patterns	Number of Entries	
	Release 2011.1	Release 2012.1
Factors with expression patterns	432	436
Average number of expression patterns per factor	17	17

Cross-links to external databases (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
EMBL	5,588	7,212	7,686	5,644	7,281	7,755
Swiss-Prot	10,899	7,517	11,065	11,499	8,000	11,663
FlyBase	185	153	192	185	153	192
PDB	195	144	438	197	144	440
DATF	1,724	1,434	1,724	1,729	1,434	1,729

Cross-links to BIOBASE databases (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
TRANSCompel	345	425	1,110	346	425	1,111
S/MARt DB	38	38	38	40	40	40
PathoDB	58	10,896	10,896	58	10,896	10,896
TRANSPATH	12,357	12,357	12,357	13,194	13,194	13,194

Factor Superclass	Release 2011.1			Release 2012.1		
	Entries	Linked Sites	Linked Matrices	Entries	Linked Sites	Linked Matrices
I. Basic Domains	1,322	6,462	2,385	1,325	6,987	2,628
II. Zinc-coordinating DNA-binding domains	2,817	10,287	2,045	2,818	11,072	2,320
III. Helix-turn-helix	2,993	8,736	3,047	2,997	9,222	3,424
IV. beta-Scaffold Factors with Minor Groove Contacts	389	2,320	621	390	2,506	723
V. Other Transcription Factors	505	1,887	522	505	2,004	604
Class assignments	8,026	29,692	8,620	8,035	31,791	9,699
Unclassified	9,156	12,456	4,454	10,110	14,561	5,304
miRNA	405	1,030	-	469	1,390	-
Total	17,587	43,178	13,074	18,614	47,742	15,003

The number of linked sites exceeds the total number of site entries, because more than one factor can be linked to one site. Also, orthologous factors can be linked to one matrix, resulting in a higher number of factor-matrix-links.

12.3. TRANSFAC Sites

Site	Number of Entries	
Entries discriminated by taxa	Release 2011.1	Release 2012.1
Vertebrata	15,397	18,157
Homo sapiens	7,983	9,597
Insecta	1,423	1,429
Plants	1,269	1,414
Fungi	1,200	1,271
S. cerevisiae	991	1,033
Viridae	733	750
Other	121	132
Artificial	11,425	11,867
Consensi	465	473

Sequences	Number of Entries	
	Release 2011.1	Release 2012.1
Sequences	31,150	34,443
Nucleotides	753,938	849,805

Cross-links to external databases (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
EMBL	14,374	5,443	15,623	16,338	6,011	17,473
EPD	935	244	1,128	935	244	1,128
FlyBase	381	51	384	381	51	384

Cross-links to TRANSFAC System databases (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
TRANSCompel	72	45	84	72	45	84
PathoDB	7	16	16	7	16	16
TRANSPRO	5,992	2,801	9,415	5,989	2,804	9,409

12.4. TRANSFAC Genes

Gene	Number of Entries	
Entries discriminated by taxa[1]	Release 2011.1	Release 2012.1
Vertebrata	43,396	43,457
Homo sapiens	18,776	18,815
Insecta	221	269
Plants	14,590	14,677
Fungi	853	913
S. cerevisiae	715	728
Viridae	92	98

Cross-links to external databases[2] (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
Affymetrix	43,872	246,101	275,327	43,868	246,101	275,302
EMBL	48,462	335,471	391,518	45,937	293,096	406,104
Ensembl	37,342	37,342	37,342	37,367	37,367	37,367
BRENDA	308	267	320	307	266	319
ENTREZ GENE	66,413	66,413	66,413	69,892	69,892	69,892
HGNC	19,836	19,836	19,836	21,586	21,586	21,586
MGI	22,341	49,054	50,358	22,315	49,024	50,330
OMIM	13,274	15,958	18,388	13,270	15,960	18,430
REFSEQ	45,549	165,889	175,216	45,558	165,974	175,268
RGD	5,203	10,363	10,387	5,250	10,452	10,475
UniGene	47,144	80,201	86,108	47,142	80,204	86,081

Cross-links to BIOBASE databases[2]	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
TRANSPATH	25,576	25,576	25,576	29,196	29,196	29,196
TRANSCompel	284	408	408	284	408	408
TRANSPRO	58,628	114,185	114,185	58,682	114,203	114,203
PathoDB	7	16	16	7	16	16
S/MARt DB	230	363	530	230	363	530

1) only gene entries that are linked to site, ChIP-chip fragment and/or factor were counted

2) all gene entries were counted

12.5. TRANSFAC Matrix

Matrix	Number of Entries	
Entries discriminated by taxa	Release 2011.1	Release 2012.1
Vertebrata	980	1,376
Insecta	69	181
Plants	137	197
Fungi	254	381
Nematoda	7	30
Prokaryota	2	2

Cross-links to external databases[2] (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Links total number	Linked TRANSFAC entries	Linked DB entries	Links total number
JASPAR	-	-	-	850	855	859

12.6. TRANSFAC Cell

Cross-links to external databases (DB)	Release 2011.1			Release 2012.1		
	Linked TRANSFAC entries	Linked DB entries	Linked TRANSFAC entries	Linked DB entries	Linked TRANSFAC entries	Linked DB entries
CLDB	206	260	260	209	263	263

12.7. TRANSFAC Class

Cross-links to external databases (DB)	Release 2010.4			Release 2011.4		
	Linked TRANSFAC entries	Linked DB entries	Linked TRANSFAC entries	Linked DB entries	Linked TRANSFAC entries	Linked DB entries
PROSITE	26	24	35	26	24	35

12.8. TRANSCompel

Entry type	Release 2012.1
Composite elements	428
Genes	283
Evidence codes	1488
References	507

12.9. TRANSPro

Species	Release 2012.1
Promoter sequences (all)	277,337
Human	40,501
Mouse	57,819
Rat	16,274
Dog	17,042
Chimpanzee	16,964
Plasmodium	2,518
Arabidopsis	27,994
Rice	29,456
Soybean	68,769