Nucleic Acids Research, Vol. 20, Supplement 2091–2093

TFD: The transcription factors database

David Ghosh
National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA

Since its inception (1), a number of changes (2) have occurred to the Transcription Factors Database, some details of which will be described here. This database is designed to capture and contain information about the properties of transcription factors, their interrelationships, the nucleic acid and amino acid sequences that encode them, and the DNA sequences they recognize. This information has been sequestered into different tables so that, through the use of relational database management systems, the existence of relationships between the various entities represented in the database may be apprehended. Figure 1 presents an ERD (Entity-Relationship Diagram), and Figure 2 the corresponding data dictionary, for release 4.2 of this database, which became available in February 1992. Since release 1.0, five new tables (METHODS, N__POINTERS, POLYPEPTIDES, REFERENCES, and X__POINTERS) have been added, one table (ELEMENTS) has been removed, and the name of one table has been changed (from CDNAS to CLONES). The reference number (ref__n) fields in the various tables have been maintained from the original database definition, but these fields now contain a Medline Unique Identifier (UID), avoiding the need for maintaining a local reference numbering scheme. In release 4.2, the database contains 1658, 929, 434, 1511, 1887, 34, 2552, 5726, and 4872 records in the CLONES, DOMAINS, FACTORS, POLYPEPTIDES, SITES, METHODS, N__POINTERS, REFERENCES, and X__POINTERS tables, respectively.

A question that arises frequently is 'How many transcription factors are there?' Though it is difficult to arrive at a single number that answers this question, the current numbers of entries in this database may provide some handle on this issue. Rat c-Jun, mouse c-Jun, and human c-Jun may be considered as three distinct transcription factors, although in a particular cell line or tissue source, only one of these proteins would be expressed. Furthermore, since Jun only has function within the context of the loosely defined entity termed AP-1, c-Jun could be considered, rather, as a subunit of a transcription factor. A central issue in answering the question of 'How many transcription factors' is whether one defines a transcription factor as an activity or as a molecule, but one answer to this question can be found in the entries of the TFD POLYPEPTIDES table, which currently corresponds to 1447 distinct polypeptide species.

A number of commercial and noncommercial tools now exist for using the sequence information contained in these tables (3). A slightly different class of tools has also been under development. These tools are designed to provide a groundwork for the use of 'visual reasoning' in studying relationships between objects having precomputed positions on large genomes, and are designed to provide graphic workbenches for the study of genome sequences, such as that of the organism *E. coli* (4). The use of the TFD SITES table for this category of analysis (5) may be of some value for the study of the long continuous and contiguous sequences that are now being produced by genome sequencing efforts (6).

Figure 3 presents sample entries in a number of TFD tables. In a typical SITES entry (3A), there exist a number of fields (trn__unit, locat__ref, and n__prob) which did not exist in the 1.0 release. The locat__ref field now serves to distinguish locations that are defined relative to an mRNA start site from those that are defined relative to other numbering schemes such as arbitrary fragment or full genome coordinates. The n__prob field contains a precomputed value that estimates the probability that the SITES entry will occur by chance in a random sequence of identical length, and the values in this field can be used in different ways. For example, sequence analysis datasets are now generated whose contents are restricted to SITES entries whose n__prob values fall below a certain threshold. The n__prob field can also be used, subsequent to a sequence analysis, to compute the statistical significance of a match to a specific SITES entry. In a DOMAINS entry (3B), a distinction now exists between the structural and functional classifications of the respective entries, which are represented by the struc__clas and func__clas fields. In release 4.2, the DOMAINS table contains 412 zinc finger entries, 299 homeodomain entries, 85 helix-turn-helix entries, and 133 entries of other classes.

A principal difference between the 4.2 release and the 1.0 release is in the existence of pointers tables. These tables (N__POINTERS and X__POINTERS) might also be considered

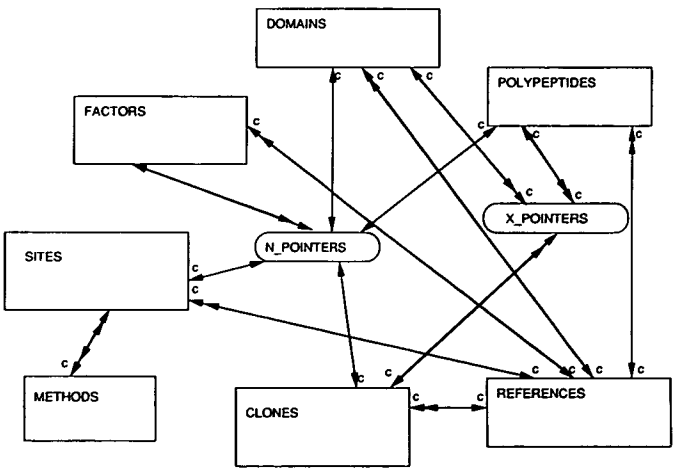

Figure 1; Entity-relationship diagram for eight tables in the 4.0 TFD release, one-to-one relationships are represented by links with single arrows at both ends. Instances of one-to-many relationships are indicated in the cases where one end of the link contains a double arrow. Conditional relationships are indicated by a 'c' character.

Table	Field	Length	Description
clones	clone_id	6	Clones entry identifier
clones	fac_name	20	Name of factor
clones	clone_type	7	Type of clone (cdna or genomic)
clones	source	10	Source of clone
clones	clone_name	10	Name of plasmid
clones	na_seq1	200	Base pairs 1-200
clones	na_seq2	200	Base pairs 201-400
clones	na_seq3	200	Base pairs 401-600
clones	na_seq4	200	Base pairs 601-800
clones	na_seq5	200	Base pairs 801-1000
clones	segment	2	segment identifier
clones	comments	80	Comments
clones	main_ref	60	Literature reference
clones	ref_n	8	Reference number (medline uid)
domains	domain_id	6	Domains entry identifier
domains	fac_name	20	Name of factor
domains	struc_clas	20	Structural or motif classification
domains	domain_num	2	Number of domain relative to amino terminus
domains	seq_type	1	Individual or consensus sequence
domains	aa_start	4	Start position in protein
domains	aa_end	4	Stop position
domains	aa_seq	150	Amino acid sequence entry (one-letter code)
domains	func_clas	15	Functional classification
domains	comments	80	Comments
domains	main_ref	60	Primary reference
domains	ref_n	8	A reference number
domains	organism	20	Organism source
factors	factor_id	6	Factors entry identifier
factors	fac_name	20	Name of factor
factors	distrib	5	Tissue distribution of factor
factors	system	5	System or organism
factors	clone	1	Existing genomic or cdna?
factors	seq_spec	1	Sequence-specific?
factors	dna_bindin	1	DNA-binding?
factors	modifs	15	Modifications
factors	function	25	Function
factors	comments	80	Comments
factors	main_ref	60	Primary reference
factors	source	21	Source for isolation of factor
factors	mw	6	Molecular weight of protein
factors	syns	16	Synonymous or related factors
factors	ref_n	8	A reference number
factors	derivation	15	Description of how identitiy of the factor as a biochemical entity is derived
factors	organism	20	Organism source of the factor
methods	full	50	experimental method
methods	tfdcode	2	two-letter code used by TFD
n_pointers	table1	15	name of first TFD table for this pointer
n_pointers	id_1	7	identifier for TFD entry 1
n_pointers	table2	15	name of second TFD table for this pointer
n_pointers	id_2	7	identifier for TFD entry 2
polypeptides	polypep_id	6	Polypeptides entry identifier
polypeptides	fac_name	20	Name given to factor
polypeptides	subunit	10	Name of this subunit
polypeptides	organism	20	Species name corresponding to biological source
polypeptides	aa_seq1	200	Residues 1-200
polypeptides	aa_seq2	200	Residues 201-400
polypeptides	aa_seq3	200	Residues 401-600
polypeptides	aa_seq4	200	Residues 601-800
polypeptides	aa_seq5	200	Residues 801-1000
polypeptides	segment	2	Which 1000 residue entry of complete sequence
polypeptides	seq_extent	10	Partial or complete ?
polypeptides	comments	80	Comments
polypeptides	main_ref	60	Reference
polypeptides	ref_n	8	A reference number
references	title	200	Title of publication
references	author	200	Authors of publication
references	journal	100	Citation entry corresponding to publication
references	abstract	200	First 200 characters of abstract
references	uid	8	Eight-digit NLM unique identifier
sites	site_id	6	Sites entry identifier
sites	fac_name	25	Name of factor
sites	seq_name	30	Name of sequence or element
sites	na_seq	45	Nucleic acid sequence
sites	seq_type	1	Individual or consensus sequence
sites	system	10	System or organism
sites	genome	1	Viral or cellular genome
sites	trn_unit	20	Name of transcription unit
sites	comments	50	Comments
sites	main_ref	70	Primary reference
sites	fac_source	16	Source of factor used to map site
sites	locat_ref	20	Reference point for coordinates in "location"
sites	location	20	Location relative to mRNA start
sites	method	11	Method used to map site
sites	n_prob	8	Precomputed probability of site occurrence
sites	ref_n	8	A reference number
sites	strand	1	Coding or noncoding
sites	binding	1	Binding or non-binding
x_pointers	tfd_table	15	Name of TFD table referenced by this pointer
x_pointers	tfd_id	6	Identifier of the TFD entry
x_pointers	x_db	10	External database identifier
x_pointers	x_release	4	Release of database referenced in "x_db"
x_pointers	x_ac	10	Accession number of the X_db entry referenced by this pointer
x_pointers	x_entry	10	Entry name of the X_db entry referenced by this pointer

Figure 2. Data dictionary for TFD release 4.0, including (from left to right) table name, field name, length of field, brief field description.

A

SITE_ID	S01874
FAC_NAME	Kruppel
SEQ_NAME	eve-stripe2-kr5
NA_SEQ	TTAATCCGTT
SEQ_TYPE	I
SYSTEM	DROS
GENOME	C
TRN_UNIT	even-skipped
MAIN_REF	Science 254: 1385-7 (1991)
LOCAT_REF	RNA start site
LOCATION	-1562/-1553
METHOD	DF
N_PROB	9.54e-07
REF_N	92073911

B

DOMAIN_ID	D00293
FAC_NAME	c-Myc
STRUC_CLAS	helix-loop-helix
DOMAIN_NUM	1
SEQ_TYPE	I
AA_START	346.00
AA_END	405.00
AA_SEQ	VKRRTHNVLERQRRNELKRSFFALRDQIPELENNEKAPKVVILKKATAYILSVQAEEQKL
FUNC_CLAS	dimerization
COMMENTS	Myc similarity region
MAIN_REF	Genes Dev 4: 167-79 (1990)
REF_N	90249724

C

TFD_TABLE	polypeptides
TFD_ID	P00221
X_DB	PIR
X_RELEASE	27.0
X_AC	A28263
X_ENTRY	TVRTFS

Figure 3. Sample TFD entries in the SITES (3A), DOMAINS (3B), and X__POINTERS (3C) tables.

as 'relation matrices' or 'correlation tables', and can be used as hard links between any of a number of molecular biology databases (Genbank, EMBL, SwissProt, PIR, Genpept, NCBI-Backbone) and TFD. A TFD X__POINTERS entry is presented (Figure 3C) which shows a typical cross-reference between TFD and an external database, which in this case is between a TFD POLYPEPTIDES entry (P00221, rat AP-1/c-Fos) and an entry in the NBRF-PIR database.

Though the ELEMENTS table is no longer contained in TFD, an initial set of cross-references between TFD SITES and EPD (reference 7, the Eukaryotic Promoter Database) has been produced. EPD is linked directly to the EMBL nucleotide sequence database, and it is possible, for those entries which correspond to naturally-occurring promoter sequences, to establish maps between TFD SITES entries and the general purpose nucleotide sequence databases such as Genbank and EMBL. Database retrieval tools that would access these links and cross-references are now under development.

The TFD database is available as a set of flat ASCII text files (each corresponding to a TFD table), or as a simple ASN.1 (Abstract Syntax Notation) file from the NCBI repository, or from the EMBL file server, as well as a number of different CD-ROMS. At NCBI the contents of these files have been loaded into a number of database managers, including Xbase (dBASE, Foxbase, dBXL, etc.), Paradox, Oracle, Sybase, and 4th Dimension, but in principle the flat text files may be imported into any standard relational database manager. SITES datasets as well as amino acid sequence (DOMAINS, POLYEPTIDES) datasets, intended for sequence analysis, are also available from the NCBI server. The internet address for downloading the files by anonymous FTP from the NCBI repository is 'ncbi.nlm.nih.gov' (numerical address 134.14.20.1); the files specific to TFD are located in the repository/TFD directory on this file server.

ACKNOWLEDGEMENTS

I thank Tim Clark and Dennis Benson for their helpful comments on this manuscript. The mention of commercial organizations or products does not imply endorsement by the NCBI or the US Government.

REFERENCES

1. Ghosh, D. (1990) Nucleic Acids Res **18**, 1749–1756.
2. Ghosh, D. (1991) Trends Biochem Sci **16**, 445–447.
3. These tools include, for SITES analysis, Dynamic, Signalscan (public domain software), and GCG, Ig/Suite, and IBI/Macvector (commercial products); for DOMAINS/POLYPEPTIDES analysis, the FASTA and BLAST families of programs.
4. Rudd, K.E., W. Miller, C. Werner, J. Ostell, C. Tolstoshov, and S.G. Satterfield (1991) Nucleic Acids Res **19**, 637–47.
5. Price, M., Hagstrom, R., and Overbeek, R. (1992), In preparation.
6. Pearson, R.L., B. Maidek, M. Chipperfield, R. Robbins (1991) Science **254**, 214–215.
7. Bucher, P. (1992) Eukaryotic Promoter Database of the Weizmann Institute of Science. This database is also available from the NCBI repository and from the EMBL file server.