Nucleic Acids Research, 1993, Vol. 21, No. 13 3117–3118

Status of the transcription factors database (TFD)

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

ABSTRACT

The Transcription Factors Database is a specialized database focusing on transcription factors and their properties. This report describes the present status of this database and developments during the past year. Within this time, the size of this database has increased by a 2799 total records, and has become accessible through a number of new mechanisms.

Since the last description of the status of this database, there have been relatively few changes to its organization, although it is now accessible through a larger number of mechanisms. In addition to the previously described mechanisms, which include direct downloading of the raw files of the database, it is also now accessible through a variety of Internet gopher servers and electronic mail servers (1). In release 6.4 (April 1993) there are 1941, 1010, 505, 1572, 2060, 38, 2811, 7038, 5427 records in the clones, domains, factors, polypeptides, sites, methods, n__pointers, references, x__pointers tables. The size of the 6.4 release of the database is 10.9 Mbytes in the flat-file representation, and 8.5 Mbytes in the ASN.1 representation. The most recent data dictionary for TFD is presented in Figure 1. This design contains only a minor change, in the SITES table, from previously described designs (2). This change would permit improved searches on the sites.trn__unit field, and allow easier generation of maps of previously reported, experimentally determined, binding sites for a given promoter or transcription unit. In this regard, there now exist tools developed in the Prolog environment, which permit the display and analysis of patterns of computationally determined binding sites on a given genome sequence, based on the previously described PC-Genographics interface (3,4).

The SITES table of this database has increased in size from 1887 entries in release 4.0 to 2060 entries in release 6.0. The information in this table incorporates a variety of published compilations (5–9), and these various compilations account for approximately 30% of the SITES entries. New SITES entries are checked prior to inclusion to avoid exact sequence duplicates, so that an entry from a compilation is not made when the given sequence already has a representation in the SITES table. Although a list of these entries is not provided here due to excessive size, it is available electronically in the printable text file 'SITELIST.TXT', which is derived from the more important SITES fields, from the NCBI repository FTP server and other FTP servers.

Although the SITES table is an essential component of this database, and although transcription factor binding sites are often associated with information involving other aspects of gene regulation, TFD does not attempt to provide a general model for gene regulation events. Although links have been explored between this database and other databases including a promoter database, TFD does have several defined boundaries. Specialized databases of signal transduction, RNA processing, intracellular addressing, translational regulation, and cell lineage development may become at some stage necessary, but this database limits its contents to information about transcription factors. It is also important to point out that by any definition, the term 'transcription factor' does not limit itself to sequence-specific DNA binding proteins. Although one of the most important components of the database is information relating to this category of molecular recognition, there have existed for some time in this database, entries for DNA-binding but non-sequence-specific polypeptides and RNA-binding proteins, and the database also includes entries representing various basal transcription factors and coactivators (10,11) which are currently being studied intensively. The main criterion is that the information represent, or directly pertain to, proteins that participate in some manner in the transcription machinery, which can include such processes as attenuation, elongation, and termination, as well as initiation. The database also includes many entries corresponding to RNA polymerase subunits, which are not considered as transcription factors according to the classical definition (12). The rationale for this inclusion involves reasons of relevance of polymerase polypeptides in the transcription machinery, as well as a potential utility of this information in the testing of object-oriented methodologies to this database problem. These entries might also be useful at some stage for studying the sequence-specific protein-protein interactions that presumably occur within active polymerase complexes. Also it is important to consider that prokaryotic sigma factors, which are required for the formation of active polymerase complexes, have been suggested to possess sequence and potentially structural similarities to eukaryotic transcription factors (13).

Various software tools currently exist for using this database. Commercial sequence analysis packages (e.g. GCG, Intelligenetics GeneWorks, IBI MacVector) contain utilities allowing SITES analysis, and a number of public domain software packages (14,15) are available from FTP servers (16). One tool, 'threedb', allows mixed sequence analysis/SQL-like retrieval queries against amino acid sequence entries in the database, and has been used thus far for one step in the characterization of a sequence motif shared by two different classes of DNA-binding domains (17). Such flexible query methods may be of particular usefulness in performing targeted analyses of sequences involved in molecular recognition phenomena in transcription. General protein sequence analysis queries and text retrieval queries can be performed on this database through the use of the NCBI Blast and Retrieve email servers.

A major current problem of the material for this database relates to the limitations of relational database technology in organizing certain types of information. Object-oriented databases have been

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	Primary reference
clones	ref_n	8	Eight-digit NLMUID
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 (I/C)
domains	aa_start	4	Start position in protein sequence
domains	aa_end	4	Stop position in protein sequence
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	Eight-digit NLMUID
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 (Y/N)
factors	seq_spec	1	Sequence-specific (Y/N)
factors	dna_bindin	1	DNA-binding (Y/N)
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	Alternative names used for this factor
factors	ref_n	8	Eight-digit NLMUID
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

Table	Field	Length	Description
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	Primary reference
polypeptides	ref_n	8	Eight-digit NLMUID
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 NLMUID
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 (I/C)
sites	system	10	Organism system class
sites	trn_unit	30	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	Eight-digit NLMUID
sites	strand	1	Coding or noncoding (C/N)
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 1. Data dictionary for TFD release 7.0. From left to right, the columns indicate table name, field name, length of field, and a brief description of field contents.

suggested to be able to better encode certain relationships such as containment than can relational databases (18). Although object-oriented technologies have a much less-developed theoretical framework than do relational technologies, the information currenly present in TFD may at some stage provide suitable material for the development of an object-oriented biological database. In addition to polymerases, another example of a specific entity which might find better representation in an object-oriented, rather than a relational, database is human TFIID, a multiprotein complex and activity comprised of TBP (TATA-binding protein) and a variety of accessory TAF (TBP-associated factor) polypeptides. In concert with the development of this highly specialized and information-rich subfield of molecular biology, this database can potentially develop into one research tool for the making of knowledge discoveries, in parallel with the making of scientific discoveries (19).

ACKNOWLEDGEMENTS

The mention of commercial products or organizations does not imply endorsement by NCBI or the U.S. Government.

REFERENCES

1. Gilbert, D. (1993) Trends In Bioch Sci **18**, 107–8.
2. Ghosh, D. (1992) Nucleic Acids Res **20S**, 2091–3.
3. Hagstrom, R., Michaels, G.S., Overbeek, R., Price, M., Taylor, R., Yoshida, K., and Zawada, D. (1992) Argonne National Laboratory Technical Report ANL-92/11.
4. Michaels, G.S., Taylor, R., Munson, P., Kazic, T., Hagstrom, R., Price, R., and Overbeek, R. (1993) Comput Chem, in press.
5. Wingender, E. (1988) Nucleic Acids Res **16**, 1879–902.
6. Locker, J., and Buzard, G. (1990) DNA Sequence **1**, 3–11.
7. Turpaev, K.T. and Vasetskii, E.S. (1990) Soviet Genetics **26**, 504–16.
8. Wingender, E. (1990) Adv Mol Genet **4**, 95–108.
9. Faisst, S. and Mayer, S. (1992) Nucleic Acids Res **20**, 3–26.
10. Roeder, R.G. (1991) Trends In Biochemical Sciences **16**, 402–408.
11. Gill, G., and Tjian, R. (1992) Curr Opin Genet Dev **2**, 236–42.
12. Lewin, B. (1992) Gene Expression, Vol. 4, pp. 543–577 (Wiley Intersciences, New York).
13. Jaehning, J.A. (1991) Science **253**, 859.
14. Prestridge, D.S. and Stormo, G. (1993) Comp Appl Biosci **9**, 113–115.
15. Faulkner, D., Smith, T.F., and Ghosh, D. (1993) In preparation.
16. Signalscan is available from the FTP servers: beagle.colorado.edu; molbio.umn.edu. Dynamic is available from the FTP servers: mbcrr.harvard.edu; darwin.bu.edu; phloem.uoregon.edu; reseq.regent.e-technik.tu-muenchen.de; sunbcd.weizmann.ac.il.
17. Toledano, M.B., Ghosh, D., Trinh, F., and Leonard, W.J. (1993) Mol Cell Biol **13**, 852–860.
18. Cattell, R.T.T. (1991) Commun ACM **34**, 31–33.
19. Frawley, W.J., Piatetsky-Shapiro, G., and Matheus, C.J. (1991) In Knowledge Discovery in Databases (ed. Piatetsky-Shapiro and Frawley, AAAI Press, Menlo Park, CA), 1–27.