Object-oriented Transcription Factors Database (ooTFD)

David Ghosh*

Institute for Transcriptional Informatics, PO Box 2556, Pittsburgh, PA 15230, USA

Received October 1, 1999; Accepted October 4, 1999

ABSTRACT

ooTFD (object-oriented Transcription Factors Database) is an object-oriented successor to TFD. This database is aimed at capturing information regarding the polypeptide interactions which comprise and define the properties of transcription factors. ooTFD contains information about transcription factor binding sites, as well as composite relationships within transcription factors, which frequently occur as multisubunit proteins that form a complex interface to cellular processes outside the transcription machinery through protein–protein interactions. In the past year, a few additions and changes were made to this database and associated tools, which are accessible through the IFTI-MIRAGE web site at http://www.ifti.org/

ORGANIZATION AND CONTENTS OF THE DATABASE

This database is based on data structures derived from the tables of the original database (1,2), now referred to as rTFD (relational Transcription Factors Database). Data structures as described initially (3) for containment and composite relationships in the object-oriented version of the database are being further developed. The contents of main data structures are presented in Table 1, and the flow of information between this database, related data resources, and query and analysis tools is presented in Figure 1. The Mood (Materials object oriented database) and ozone (a pure java object database) systems have been used to maintain the object-oriented version of this database in the past year, although a number of other object database systems are now also being investigated. Entries in the Sites component of this database, which is used in conjunction with transcription factor binding sites sequence analysis tools, are also accessible through network-based interfaces using mysql, a standard relational database system.

Table 1. Primary ooTFD data structures

Data structure	No. entries
Factors	540
Polypeptides	3179
Sites	6037
Domains	1030
References	20 812
Names	1137
siteProfiles	457
domainProfiles	15
Interactions	93

DATA RETRIEVAL TOOLS

The two categories of retrieval tools and services, namely sequence analysis queries and database queries, are summarized in Table 2 and presented graphically in Figure 1. Protein sequence analyses can be performed against the Polypeptides/Domains dataset (tfdaa) through a standard BLASTP analysis. In cases where BLAST matches are to a polypeptide with known composite relationships in ooTFD, associated links indicating the composite information are provided. Sites analyses can be performed with Tfsitescan, a service which constructs an imagemap in association with sequence analysis results which is linked to individual Sites entries. Object database queries are currently possible through an interface to the ozone implementation of ooTFD, and these generally provide links to the individual database entries. A simple ‘knowledge based’ web tool, TF-Advisor, is available to a new user for selecting the most suitable approach to performing a particular query.

Table 2. ooTFD-associated analysis and retrieval tools

Associated tool	URL
Tfsitescan	http://www.ifti.org/cgi-bin/ifti/Tfsitescan.pl
Tfdaa/BLASTP	http://www.ifti.org/cgi-bin/ifti/TfdaaAnalysis.pl
TF-Advisor	http://nbcr1.sdsc.edu:8001/tfadvisor/tfadvisor.html
ooTfd query tools	http://www.ifti.org/cgi-bin/ifti/ootfd.pl
Structure viewer	http://www.ifti.org/cgi-bin/ifti/Structures.pl

EXTERNAL DATA INTERFACES AND FUTURE DIRECTIONS

The original relational database contained a table, X_POINTERS, which was used to map TFD entries to general sequence databases, including GenBank, EMBL, SWISS-PROT and PIR (2). Mainly, this table was used to refer to entries in these other databases which represented the same biological entity. Similar data structures are now being developed to allow the recording of a different type of database relationship, one that is representative of a molecular relationship. These types of pointers, in contrast to the hard links which indicate identical

*Tel/Fax: +1 412 682 0550; Toll-free Tel/Fax (USA only): +1 800 894 4770; Email: dghosh@ifti.org

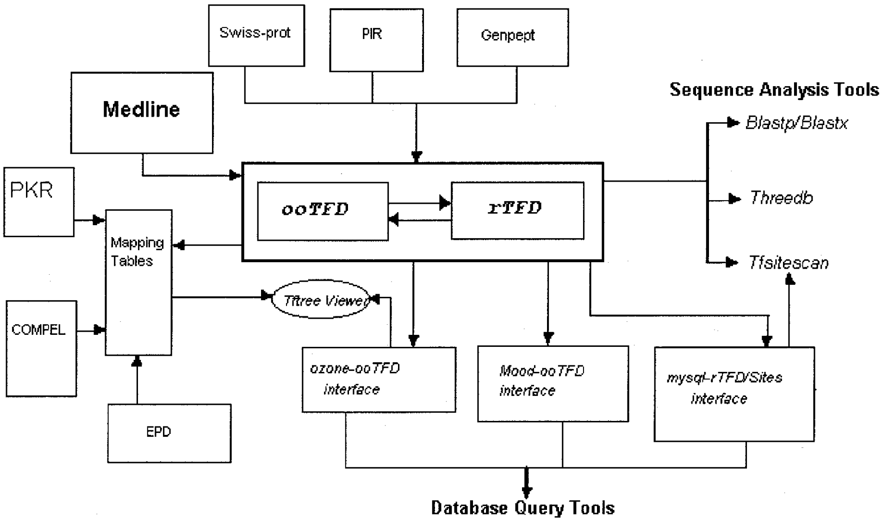

Figure 1. Organization and flow of information in ooTFD, rTFD and data access tools.

information in multiple databases, represent information not present in either database. For example, one of the ways in which transcription factors transmit cellular and extracellular information to the gene expression machinery is through phosphorylation and dephosphorylation events (4,5). The PKR (Protein Kinase Resource) is a database which contains a variety of information on protein kinases and their biologically relevant sites in their respective protein sequences (6). An interface currently in development links ooTFD entries to PKR entries, allowing end-users to explore the known protein kinase–transcription factor interactions that are a part of the interactions of the cellular machinery with the transcriptional machinery at a particular transcription factor binding site. A java applet in development similar to one presented in an earlier publication (3) would allow the viewing of these interactions with transcription factor polypeptides as a result that is returned from an ooTFD object query.

Another example of inter-database links representing information not present in either database involves low *P*-value sequence matches between ooTFD/Sites sequence entries and sequence entries in COMPEL, a database of composite regulatory elements (7). These mappings thus include sequence relationships other than, and in addition to, those that are represented in the COMPEL annotation, which is based on experimentally observed data for transcription factor interactions at a composite transcriptional regulatory element. Such links could suggest possible protein–protein interactions between two sequence-specific transcription factors other than those observed by the investigators who originally reported these sites; in these cases, associated retrieval tools would indicate that the suggested composite interaction is one that is predicted from computation.

Another precomputed dataset that has been developed as an accessory to ooTFD involves low *P*-value sequence matches of TFD/Sites sequences to sequences in EPD, Eukaryotic Promoters Database (8). Query tools for retrieving entries from this precomputed dataset are in development, and further versions of the promoter sequence analysis tool available at the IFTI-MIRAGE web site, Tfsitescan, would provide results which automatically link to this precomputed dataset.

The current consensus in the bioinformatics field is that transcription factor binding sites, especially when represented as matrices, do not contain sufficient information for predictive value for a new binding site from sequence analysis (9). Although there is some usefulness in this view, simple methods for organizing sequence analysis results based on *P*-value, as are currently performed by Tfsitescan, can serve as one approach to screen out spurious and statistically insignificant matches. As in the cases of homeodomain proteins, some transcription factors may simply be less specific for their target sequences than others, and the primary determinants for transcription factor access to DNA may in these cases be modulated at the level of chromatin organization. Although maintaining and updating information in an object database system is less straightforward than maintaining information as a relational database, the increasing complexity of the information being generated regarding gene regulation suggests a need for approaches to data management beyond that which is afforded by relational database management systems. The available dataset of co-regulated genes in complete genomes, their common upstream sequence elements, and reports regarding the role of combinatorial transcription factor interactions in gene regulation, is increasing. As it becomes available, automated systems to link such publicly available information to information derived from transcription factor data retrieval, would be developed in the context of this database resource.

ACKNOWLEDGEMENTS

This work was supported in part by a gift from Procter & Gamble Co. to the University of California at San Diego, by the facilities of the National Biomedical Computation Resource (NIH RR-08605) at the San Diego Supercomputer Center, and by the facilities of the Biomolecular Engineering Research Center at Boston University.

REFERENCES

1. Ghosh,D. (1990) *Nucleic Acids Res.*, **18**, 1749–1756.
2. Ghosh,D. (1991) *Trends Biochem. Sci.*, **16**, 445–447.
3. Ghosh,D. (1998) *Nucleic Acids Res.*, **26**, 360–361.
4. Hunter,T. and Karin,M. (1992) *Cell*, **70**, 375–387.
5. Karin,M. (1994) *Curr. Opin. Cell. Biol.*, **6**, 415–424.
6. Smith,C.M., Shindyalov,I.N., Veretnik,S., Gribskov,M., Taylor,S.S., TenEyck,L.F., Bourne,P.E. (1997) *Trends Biochem. Sci.*, **22**, 444–446.
7. Kel-Margoulis,O.V., Kel,A.E., Frisch,M., Romaschenko,A.G., Kolchanov,N.A. and Wingender,E. (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure, (BGRS '98)*. ICG, Novosibirsk, Vol. 1, 54–57.
8. Perier,R.C., Junier,T., Bonnard,C. and Bucher,P. (1999) *Nucleic Acids Res.*, **27**, 307–309. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 302–303.
9. Bucher,P. (1999) *Curr. Opin. Struct. Biol.*, **9**, 400–407.