BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2860–2862
doi:10.1093/bioinformatics/bth300

TargetDB: a target registration database for structural genomics projects

Li Chen, Rose Oughtred, Helen M. Berman* and John Westbrook

Research Collaboratory for Structural Bioinformatics, Protein Data Bank (RCSB PDB), Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854, USA

Received on March 9, 2004; revised on April 12, 2004, accepted on April 21, 2004
Advance Access publication May 6, 2004

ABSTRACT

Summary: TargetDB is a centralized target registration database that includes protein target data from the NIH structural genomics centers and a number of international sites. TargetDB, which is hosted by the Protein Data Bank (RCSB PDB), provides status information on target sequences and tracks their progress through the various stages of protein production and structure determination. A simple search form permits queries based on contributing site, target ID, protein name, sequence, status and other data. The progress of individual targets or entire structural genomics projects may be tracked over time, and target data from all contributing centers may also be downloaded in the XML format.

Availability: TagetDB is available at http://targetdb.pdb.org/

Contact: berman@rcsb.rutgers.edu

The dramatically increasing number of new protein sequences resulting from both genomics and proteomics is driving the development of tools and methods for the rapid and reliable determination of structure and molecular function. The structural genomics projects aim to produce a complete inventory of all three-dimensional protein folds, thereby providing valuable structural information that may be combined with sequence information in order to predict the function of proteins. A number of structural genomics centers have been established worldwide with the common goal of large-scale, high-throughput structure determination using X-ray crystallography and nuclear magnetic resonance (NMR). Efficient structure determination on such a large-scale requires the open exchange of information that can be effectively achieved only through a central registry, which lists target sequences and the status of the work in progress at each site. To this end, the Protein Data Bank (RCSB PDB) (Berman *et al*., 2000) has created a centralized target registration database, named TargetDB (http://targetdb.pdb.org/), which contains sequences from essentially all of the worldwide structural genomics projects.

*To whom correspondence should be addressed.

TargetDB was launched in July 2001 and builds on the work of an earlier target database named PRESAGE (Brenner *et al*., 1999). TargetDB serves as the target registration database for structural genomics projects worldwide. It consolidates target data from the 9 NIH Protein Structure Initiative (PSI) centers and 10 other international structural genomics sites in North America, Europe and Asia (see http://www.rcsb.org/pdb/strucgen.html). Links are also provided to each individual website from the TargetDB home page. TargetDB functions as the primary target registration database from which other target databases derive their data. These databases include SPINE (Structural Proteomics in the Northeast) (Bertone *et al*. 2001; Goh *et al*. 2003, http://spine.nesg.org/sum.pl) and MSDtargets (http://www.ebi.ac.uk/msd-srv/msdtarget/).

In the following sections, we describe the contents and query capabilities of TargetDB and provide examples of how it can be used to track information on protein targets and structural genomics projects.

TARGET DATA

TargetDB is a target registration database funded by the NIH that was developed to provide registration and tracking information for the NIH PSI structural genomics projects. A number of other worldwide structural genomics centers have also contributed data to TargetDB on a voluntary basis. The target data and status information is collected weekly, in XML format, from the 19 contributing structural genomics sites and loaded into the TargetDB database. The organization of the XML document follows the recommendations of the International Task Force on Target Tracking (2001, http://www.nigms.nih.gov/news/meetings/airlie.html).

The XML target specifications include data items on target selection, protein sequence, cloning, expression, purification and structure determination via X-ray crystallography or NMR. The document type definition for the target data file is available at http://targetdb.pdb.org/apps/target.dtd. Currently, TargetDB stores information on ~55 000 targets from

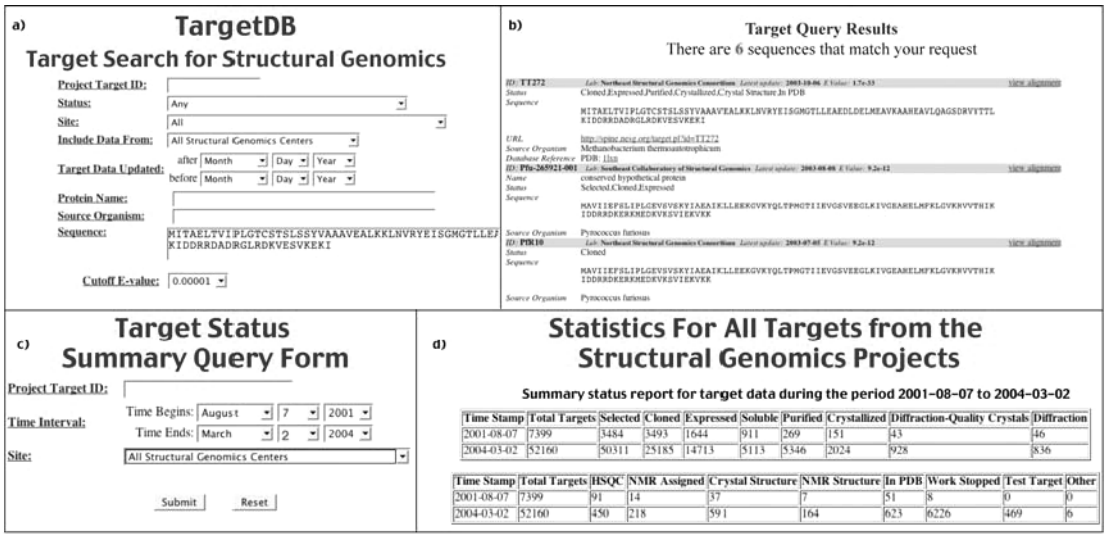

Fig. 1. Screenshot of a TargetDB session. (**a**) FASTA sequence searches may be performed using specific cut-off *E*-values. (**b**) Matching target sequences are returned with their corresponding *E*-values and links to sequence alignments. Query results contain a description of each target, its current status and links to relevant sites. Only three of six results are shown for this query. (**c**) Summary reports may be obtained on the status of individual targets, a single structural genomics center or on all the centers combined. A sample query is shown for all of the contributing sites between two time points. (**d**) The resulting summary report gives the status and statistics for all of the targets from all sites between the specified dates.

the contributing structural genomics sites and all the targets can be downloaded from TargetDB as a single XML document.

QUERY CAPABILITIES

The query capabilities of TargetDB include sequence searches using the FASTA method (Pearson and Lipman, 1988) (Fig. 1a and b). Sequence searches may be performed using specific cut-off *E*-values and all matching target sequences are returned with their corresponding *E*-values and sequence alignments. Searches may be carried out on target sequences from the NIH centers, from all the structural genomics sites, or from all the sites and inclusive of the PDB archive. Currently, TargetDB stores ∼55 000 sequences from the contributing sites and ∼58 000 sequences from experimentally determined structures in the PDB, including sequences that PDB depositors have approved for pre-release. Thus, TargetDB provides wide coverage of the relevant sequence space which allows for more comprehensive sequence searches to be carried out using a single query.

A simple search form also permits queries based on project site, target ID, protein name, source organism, date of last modification and the current status of the target (Fig. 1a). The status search category includes details on target preparation (selected, cloned, expressed, soluble, purified), crystallization (crystallized, diffraction-quality crystals, diffraction, crystal structure), NMR structure determination [Heteronuclear Single Quantum Coherence (HSQC), NMR assigned, NMR structure] and deposition status (in PDB). The option is also provided to indicate if work has been stopped on a particular target.

Query results are summarized as a list of target reports. Each report includes the target's project ID, protein sequence and current status (Fig. 1b). Some contributing sites also include links to their project center and other databases. If a target structure has been deposited in the PDB, then an active link is provided to its corresponding Structure Summary report on the PDB website. The search results may be viewed either as HTML reports, FASTA data files or in the XML format.

The status search information for each target has been preserved over time which facilitates searching on the progress of individual targets or entire structural genomics projects. This information is being used to generate two types of summary reports. One gives the progress for an individual target (or list of targets) according to its change in status over time. The second describes the aggregate status information of each structural genomics center, or of all the centers combined, between two time points (Fig. 1c and d), thereby allowing the progress of the entire structural genomics initiative to be tracked over time.

FUTURE DEVELOPMENTS

The current scope of TargetDB is to provide, in a timely manner, status and tracking in formation on the production and

structure determination of protein targets. Future developments will include extending the contents of TargetDB via an additional database named PEPCdb (Protein Expression, Purification and Crystallization database), which will make available the protocols for target cloning, expression, purification, crystallization and structure solution. Together, TargetDB and PEPCdb will facilitate the open exchange of information on targets and protocols, thereby reducing duplication of effort and contributing to the overall success of the structural genomics initiative.

Acknowledgements

The RCSB PDB is operated by Rutgers, The State University of NewJersey; The San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology—three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The work reported in this paper has been supported by grants from the NSF, the DOE, and six units of the NIH: the NIGMS, NLM, NCI, NCRR, NIBIB and the NINDS.

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bertone,P., Kluger,Y., Lan,N., Zheng,D., Christendat,D., Yee,A., Edwards,A.M., Arrowsmith,C.H., Montelione,G.T. and Gerstein,M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898.

Brenner,S.E., Barken,D. and Levitt,M. (1999) The PRESAGE database for structural genomics. *Nucleic Acids Res.*, **27**, 251–253.

Goh,C.S., Lan,N., Echols,N., Douglas,S.M., Milburn,D., Bertone,P., Xiao,R., Ma,L.C., Zheng,D. *et al*. (2003) SPINE 2: a system for collaborative structural proteomics within a federated database framework. *Nucleic Acids Res.*, **31**, 2833–2838.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci., USA*, **85**, 2444–2448.

Task Force on Target Tracking (2001) Task Force Reports from the Second International Structural Genomics Meeting, Airlie, VA.