TargetDB: a database of peptides targeting proteins to subcellular locations

Tao Wei¹ and Mary O’Connell²

¹Research Computing Center, Harvard Medical School, 220 Longwood Avenue, Goldenson 259, Boston, MA 02115 and ²Department of Agronomy and Horticulture, MSC 3Q, New Mexico State University, PO Box 30003, Las Cruces, NM 88003, USA

Received on November 16, 1998; revised on March 18, 1999; accepted on April 13, 1999

Abstract

Summary: TargetDB is a relational database designed to represent data on protein targeting sequences, mutant signals, subcellular targets and source organisms.

Availability: TargetDB is accessible at http://molbio.nmsu.edu:81. The web interface supports both direct data authoring and database query functions.

Contact: moconnel@nmsu.edu, tao_wei@hms.harvard.edu

Eukaryotic cells are characterized by subcellular compartmentalization or membrane-bounded organelles, each with unique functions. For example, genetic material is stored as chromatin and transcribed inside the nucleus. Mitochondria are responsible for generating metabolic energy and reducing power for other cellular processes. All of these functions require organelle-specific proteins. Proteins are either translated on ribosomes in the cytosol and transported post-translationally into the respective subcellular compartments or translated on ribosomes attached to the rough endoplasmic reticulum and are co-translationally transported into the Golgi–endoplasmic reticulum–lysosome system (Alberts et al., 1994). These processes are called protein targeting, sorting or trafficking. Proper targeting requires collaboration between the nascent protein and some other molecular apparatus. Information for protein localization is encoded by short sequences in the primary translation product. For example, usually, an amino terminus ‘signal peptide’ directs co-translational transport, while an amino terminus ‘transit peptide’ directs post-translational transport, and one or more internal amino acid sequences directs proteins to the nucleus or other organellar destinations (Alberts et al., 1994). A database representing these signal or target sequences in proteins would be useful in efforts to understand the requirements for proper processing of proteins and to engineer organelle-specific accumulation of proteins.

Swiss-Prot (http://www.expasy.ch/) annotates targeting site and signal peptide sequence (Bairoch and Apweiler, 1998), and PSORT (http://psort.nibb.ac.jp) predicts a targeting site given a primary sequence. TargetDB intends to represent detailed information about amino acid sequences within proteins that regulate the subcellular sites of accumulation. TargetDB provides information not represented in Swiss-Prot, specifically information about variants and mutants in signal and target peptides, the experimental basis for the identification of the peptide as a targeting sequence is identified clearly, and the nature of the protein complex involved in the processing event is reported. Further, TargetDB is a relational database and, therefore, can support complex queries and representations of the data.

Data for the population of TargetDB were retrieved from the primary literature on experimentally determined targeting sequences and from Swiss-Prot 35 (Bairoch and Apweiler, 1998) for sequences predicted to target proteins to subcellular volumes. Peptide sequences experimentally demonstrated to target a protein to a particular cellular location were recorded in the database along with the literature citation. Sequences predicted to target proteins were identified by searching the features field of Swiss-Prot for ‘transit’ or ‘signal’.

Currently, the database contains records of peptides that target proteins to nuclei, mitochondria, endoplasmic reticulum (ER), ER lumen, ER membrane, chloroplast, chloroplast stroma, chloroplast thylakoid, plasma membrane/extracellular secretion, vacuolar membrane, microbody, cyanelle and hydrogenosome. Most of the data describe peptides in eukaryotic proteins; there are a few entries from prokaryotes for protein secretion. Currently, there are 5160 peptide sequences for 5095 different proteins in the database. The majority of these sequences are predicted; only 103 sequences in the database are based on biochemical experimental data. The data represent proteins from a wide taxonomic range of organisms: 601 species from 176 families. We expect to enter more experimental data into the database from the literature.

TargetDB is implemented as a relational database managed by SQLAnywhere server software (Sybase). TargetDB has eight related tables, each of which captures characteristics of one data entity (Figure 1). For example, the protein table records sequence, name, symbol, EC number, tissue and source organism. The signal table records properties of a protein targeting peptide. Together, these tables represent sources of a signal peptide, properties including sequence,
mutations, protein complexes and references. Definitions of the data fields are available at the TargetDB web site. [The word signal in TargetDB data fields means any amino acid sequence that directs a protein to a subcellular volume.]

TargetDB is designed to focus on data of both targeting peptide and variations in these peptides generated either artificially or naturally. Allelic variations in targeting peptide sequences are recorded in the signal variant table relative to usually the first reported version of the peptide sequence. This use of the signal variant table extends to variations not formally demonstrated to be allelic variation, but also variation observed within a species for the same protein. The database also represents the relationship between the target peptide and the processing complex. For example, the signal peptide of bilirubin UDP-glucuronosyltransferase can interact with signal recognition particles and be processed by a signal peptidase (Seppen et al., 1996). These protein complexes can also interact with or process signal peptides of many other proteins. Such a many-to-many relationship is implemented with the signal complex linker table.

TargetDB has a client server design. The database is managed by Sybase SQLAnywhere 5.0 software. The front end uses web-based forms managed by WebBase (ExperTelligence, Inc.) providing data authoring and query functions. Data authoring requires a userid and password access. TargetDB is also available in an MS Access97 format through ftp. Complex database queries can then be made locally via SQL scripting.

**Acknowledgements**

We thank Tim Jones for his help with WebBase. This work was supported in part by US ARO under grant number DAAH04-96-1-0415, USDA CSREES 98-34387-5897 and NM Agricultural Experiment Station.

**References**

