

TC-DB: An Architecture for Membrane Transport Protein Analysis

Can V. Tran, Nelson M. Yang, Milton H. Saier, Jr.

Division of Biological Sciences, University of California at San Diego, La Jolla, CA USA
can@ucsd.edu, nyang@ucsd.edu, msaiier@ucsd.edu

Abstract

TC-DB is a comprehensive relational database containing structural, functional and evolutionary information about transmembrane transport proteins. The database contains factual material extracted from more than 9000 references, covering approximately 3000 representative proteins, classified into over 400 families. TC-DB is the primary resource allowing access to the data upon which the TC system is based. The TC system has been adopted by the International Union of Biochemistry and Molecular Biology (IUBMB). The functional ontology developed for TC-DB provides an infrastructure to develop powerful queries that yield biological insight. The TC-DB website offers a number of software tools that facilitate the analysis of membrane transporters. Multiple avenues of access are supported by the web interface including classification drill-down, parametric searching, and full-text searching. TC-DB has been used for the annotation of transport proteins in newly sequenced genomes and provides tools to trace evolutionary pathways. The database as well as web services are accessible free of charge at <http://tcd.db.ucsd.edu>.

1. Introduction

Membrane transport proteins (MTPs) provide the key means other than via passive diffusion by which molecules enter and exit cells and organelles. MTPs facilitate the transport of critical substrates by several different mechanisms. Some systems directly couple ATP hydrolysis to the transport reaction, while others take advantage of existing electrical or chemical gradients to power the transport process. Proper function of MTPs is critical for the survival of any organism. In humans, defects in MTPs may result in diseases such as cystic fibrosis, Graves' disease, and several neurodegenerative disorders.

Analysis of MTPs has shown that each type of protein typically uses just one mechanism to facilitate transport. It has been shown that members of any given family share functional and structural characteristics. Substrate specificity of an MTP can be altered by point mutations, but their new specificity will reflect the original class of substrates transported.

Taking these facts into account, we have developed a comprehensive classification system for MTPs named the Transporter Classification (TC) system [1, 2]. The TC

system is analogous to the EC enzyme classification system since both feature a similar numbering scheme. However the TC system uses five digits rather than the 4 digits of the EC system, and incorporates phylogenetic data [3].

2. Architecture

The TC-DB web-application is based upon a three-tier architecture characteristic of many other web-database applications. The underlying tier of the system is the open-source database MySQL. Forming the middle tier is the Apache-PHP application server, which retrieves tuples from the database and returns populated HTML data to the web browser client. This architecture resides upon two dual processor computers, one running Solaris for Sparc and the other running OSX for PowerPC.

The relational schema upon which TC-DB is built revolves around the table that represents the TC taxonomy. We have chosen to represent this taxonomy as an adjacency list within the table since it is a sparse directed graph. Since the TC system is not a polyhierarchical system, no two parent nodes have edges to the same child node. For every node in the adjacency list there is only one edge that points to the parent node in the taxonomy. However, as we discover distant relationships between different families, the adjacency list will take on new dimensions, as each node will have multiple edges leading not only to the parent node but also to the related cousin nodes. Revolving around the main TC system fact table are the dimension tables that provide the details on each node within the taxonomy. These details include references, descriptions, proteins, domains, motifs, and other biologically relevant data.

3. User Interface

TC-DB's web-interface allows the user to drill-down and roll-up within the hierarchical taxonomy of the TC system. Included within the interface are a variety of search tools including keyword and parametric search. In addition, the protein sequences within the taxonomy may be searched for sequence similarity with protein sequences entered by the user using the Smith-Waterman algorithm or the BLAST algorithm [4, 5]. The web-interface also supports other data mining capabilities that allow researchers to carry out a thorough analysis of MTPs using software catering to the analysis of this distinct class of proteins.

4. Applications

With the ever-increasing number of genomes being sequenced today, we are faced with a deluge of data to be manipulated and comprehended. One of the primary means of annotating new genomes utilizes a sequence similarity search against a protein repository such as GenBank or SWISS-PROT/TrEMBL [6, 7]. We have hand selected a comprehensive set of functionally characterized proteins to be included within the TC system, and knowledge of these proteins is maintained up-to date by continued literature evaluation. This set of proteins excludes closely related homologues of the same substrate specificity but displays the rich sequence diversity found within the families. This enables researchers to perform high-throughput genome screens for transporters without significantly taxing the computational nodes. By removing closely related homologues, thus keeping the sequence number relatively small, we can search genomes for transporters using the Smith-Waterman algorithm rather than a heuristic such as BLAST or FASTA [8, 9].

5. Conclusion

TC-DB provides a comprehensive, yet manageable way to visualize the state of our knowledge concerning MTPs, as well as tools that enable researchers to conceptualize the great functional, structural, and sequence diversity of MTPs. TC-DB will ultimately incorporate additional information such as relevant disease states, regulatory pathways, three-dimensional structural data, and pharmacological information. As new data are incorporated, more complete analyses will be possible. Although curation of TC-DB is a slow laborious process, the benefits are well worth the time and effort. Current methods to automatically extract data from the primary literature have not provided the requisite degree of reliability. Such techniques will become more feasible once methods are available for tagging documents with meta-data that encapsulate not only keywords but also biological intuition.

6. Acknowledgement

We thank Rolf Apweiler, Amos Bairoch, Andre Goffeau, Arnost Kotyk, Andre Lupas, and Ian Paulsen for valuable discussions. This work was supported by NIH grants GM55434 and GM64368 from the National Institute of General Medical Sciences.

7. References

- [1] W. Busch and M.H. Saier Jr., "The transporter classification (TC) system, 2002.", *Crit. Rev. Biochem. Mol. Biol.*, CRC Press, Boca Raton, Florida, 2002 Oct 1, pp. 287-337.
- [2] M.H. Saier Jr., "A functional-phylogenetic classification system for transmembrane solute transporters.", *Microbiol. Mol. Biol. Rev.*, ASM Press, Washington DC, 2000 Jun, pp. 354-411.
- [3] Webb, E.C. and International Union of Biochemistry and Molecular Biology, *Enzyme Nomenclature*, Academic Press, San Diego, California, 1992.
- [4] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences.", *J. Mol. Biol.*, Academic Press, Orlando, Florida, 1981 Mar 25, pp. 195-197.
- [5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic local alignment search tool.", *J. Mol. Biol.*, Academic Press, Orlando, Florida, 1990 Oct 5, pp. 403-410.
- [6] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, "GenBank.", *Nucleic Acids Res.*, Oxford University Press, Oxford, UK, 2003 Jan 1, pp. 23-27.
- [7] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.", *Nucleic Acids Res.*, Oxford University Press, Oxford, UK, 2003 Jan 1, pp. 365-370.
- [8] W.R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms.", *Genomics*, Academic Press, Orlando, Florida, 1991 Nov, pp. 635-650.
- [9] W.R. Pearson, "Flexible sequence similarity searching with the FASTA3 program package.", *Methods Mol. Biol.*, Wiley, Hoboken, New Jersey, 2000, pp. 185-219.