D274–D278 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 20 November 2008*
doi:10.1093/nar/gkn862

The Transporter Classification Database: recent advances

Milton H. Saier, Jr[1,*], Ming Ren Yen[1], Keith Noto[2], Dorjee G. Tamang[1] and Charles Elkan[2]

[1]Division of Biological Sciences and [2]Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093-0116, USA

Received September 16, 2008; Revised October 14, 2008; Accepted October 16, 2008

ABSTRACT

The Transporter Classification Database (TCDB), freely accessible at http://www.tcdb.org, is a relational database containing sequence, structural, functional and evolutionary information about transport systems from a variety of living organisms, based on the International Union of Biochemistry and Molecular Biology-approved transporter classification (TC) system. It is a curated repository for factual information compiled largely from published references. It uses a functional/phylogenetic system of classification, and currently encompasses about 5000 representative transporters and putative transporters in more than 500 families. We here describe novel software designed to support and extend the usefulness of TCDB. Our recent efforts render it more user friendly, incorporate machine learning to input novel data in a semiautomatic fashion, and allow analyses that are more accurate and less time consuming. The availability of these tools has resulted in recognition of distant phylogenetic relationships and tremendous expansion of the information available to TCDB users.

INTRODUCTION: THE TRANSPORTER CLASSIFICATION DATABASE (TCDB)

The transporter classification (TC) system (1), formally adopted by the International Union of Biochemistry and Molecular Biology (IUBMB) in June 2001, provides a guide to the known types of transport proteins present in living organisms on earth. The development of a classification system for transport proteins has allowed us to comprehensively view transport systems in a coherent and unified fashion from structural, functional and evolutionary standpoints, and to trace pathways taken for their evolutionary appearance (1,2). This development has been strongly influenced by recent progress in computational biology and genome sequencing.

Since our last comprehensive description of TCDB (3), we have expanded the transporter classification system by (i) introducing new classes and families of transporters, (ii) increasing the memberships of pre-existing families, (iii) providing more detailed annotations of these families and proteins, (iv) updating relevant reference citations, (v) creating a more interactive database and (vi) employing machine learning approaches that allow the semiautomated input of published information. The results of our analyses, made possible by these updates, are summarized here. We also describe briefly some of the most important software developed to support TCDB.

More than 500 protein families are currently in the TC system, classified according to transporter class and subclass as presented in Table 1. Affiliation with a family requires satisfying rigorous statistical criteria of homology. Whereas the classes and subclasses distinguish functionally distinct types of transporters, the families and subfamilies provide a phylogenetic basis for classification. The TC system is thus a functional/phylogenetic system. Families sometimes, but rarely, cross class or subclass lines. Hyperlinks have been constructed to identify superfamilies, disease-related transporters, and sources of high resolution 3D structural data. Several types of search tools facilitate protein identification and characterization.

Recognition of a phylogenetic relationship based on sequence similarity allows certain conclusions to be drawn regarding 3D structural features. Any two proteins that can be shown to be homologous (i.e. that exhibit sufficient sequence similarity to establish that they arose from a common evolutionary ancestor) can be expected to exhibit strikingly similar topological features and 3D structures, although a few exceptions have been noted (4). Therefore, extrapolation from one member of a family of known structure to other members becomes justifiable, and the degree of confidence in such an extrapolation is inversely related to the degree of sequence divergence. However, extrapolation of structural data to

*To whom correspondence should be addressed. Tel: +1 858 534 4084; Fax: +1 858 534 7108; Email: msaier@ucsd.edu

Table 1. Classes and subclasses of transport systems included in TCDB (8 August 2008)

1		Channels/pores
	1.A	α-Type channels
	1.B	β-Barrel porins
	1.C	Pore-forming toxins (proteins and peptides)
	1.D	Non-ribosomally synthesized channels
	1.E	Holins
	1.F	Vesicle fusion pores
	1.G	Paracellular channels
2		Electrochemical potential-driven transporters
	2.A	Porters (uniporters, symporters, antiporters)
	2.B	Nonribosomally synthesized porters
	2.C	Ion-gradient-driven energizers
3		Primary active transporters
	3.A	P-P-bond-hydrolysis-driven transporters
	3.B	Decarboxylation-driven transporters
	3.C	Methyltransfer-driven transporters
	3.D	Oxidoreduction-driven transporters
	3.E	Light absorption-driven transporters
4		Group translocators
	4.A	Phosphotransfer-driven group translocators
	4.B	Nicotinamide ribonucleoside uptake transporters
	4.C	Acyl CoA ligase-coupled transporters
5		Transport electron carriers
	5.A	Transmembrane 2-electron transfer carriers
	5.B	Transmembrane 1-electron transfer carriers
8		Accessory factors involved in transport
	8.A	Auxiliary transport proteins
	8.B	Ribosomally synthesized protein/peptide toxins that target channels and carriers
	8.C	Non-ribosomally synthesized toxins that target channels and carriers
9		Incompletely characterized transport systems
	9.A	Recognized transporters of unknown biochemical mechanism
	9.B	Putative transport proteins
	9.C	Functionally characterized transporters lacking identified sequences

other proteins is never justified if homology has not been established.

Similar arguments apply to mechanistic considerations. Thus, the mechanism of solute transport is likely to be similar for all members of a permease family with variations on a specific mechanistic theme being greatest when the sequence divergence is greatest (5,6). By contrast, for members of any two independently evolving permease families, the transport mechanisms may be entirely different. Extensive experimental work has established that phylogenetic data can also be used to predict substrate specificity, polarity of transport and even intracellular localization, depending on the family and degree of sequence divergence observed (1,4).

Since our last description of TCDB (3), this database has expanded with the introduction of six new subclasses (increase of 33%), 143 novel families (increase of 34%), 2009 novel proteins (increase of 67%) and 167 novel subfamilies in the five largest superfamilies (increase of 30%). The number of references cited in TCDB is now 4595, a 40% increase since January 2006. In the last 12 months, the number of visits to TCDB has increased 40%. Thus TCDB is a rapidly growing database which is increasingly useful to the international scientific community.

ESTABLISHING HOMOLOGY BETWEEN PROTEINS

Statistical algorithms are used to establish homology between two proteins, two families of proteins, or two repeat sequences within the proteins of a single family (7,8). In general, these depend on the 'Superfamily Principle' (9,10). This principle simply states that if A is homologous to B, and B is homologous to C, then A is homologous to C (9). Care must be taken, however, that in establishing homology, corresponding domains or regions of the protein are being compared (11,12). Moreover, a reliable program must take into account unusual residue compositions, as, for example, occur with membrane proteins that have a disproportionate percentage of hydrophobic residues, or proteins with multiple short repeat sequences that comprise a substantial fraction of the proteins or protein segments compared (7,12).

An average protein domain is roughly 60 residues long, so we have set the minimal length of sequences to be compared for purposes of establishing homology as 60 residues (10). We use the following rigorous criteria for the purpose of establishing common ancestry. To be homologous, two proteins, when correctly aligned to maximize identities and similarities and minimize gaps, must give a comparison score of 9 SD. This value corresponds to a probability of 10^{-19} that this degree of sequence similarity could have occurred by chance (13). These criteria eliminate the possibility that convergent sequence evolution accounts for the degree of similarity observed (10,14).

The GAP program (15) randomly shuffles the two sequences being compared 100 times and compares the actual aligned sequences with the shuffled sequences. This method eliminates artifacts due to unusual amino acid compositions, but 100 random shuffles are insufficient to give reliable values. We designed a modified program [the InterCompare program (IC); (7) and unpublished modifications], which has several advantages over GAP. First, it automatically conducts five 100-shuffle runs and averages the results, and second, it can take any number of sequences known to be homologous to protein or protein domain A, and compares them to any number of sequences known to be homologous to protein or domain B. If protein/domain C (homologous to A) shows over 9 standard deviations with protein/domain D (homologous to B), then by the Superfamily Principle, A must be homologous to B. The IC program can compare 100 homologues of A with 100 homologues of B to give 10 000 comparison scores. The third advantage is that the program presents the results as specified by the user, most usefully according to the values of the comparison scores. This allows the investigator to quickly identify the best comparisons for further examination (16).

The IC program can take a few hours to compare multiple sequences. Consequently, the number of proteins that can be inputted is limited. If BLAST searches of proteins A and B yield 500 sequences each, this number must be

reduced. This becomes possible due to availability of the CD-Hit program (17). This program eliminates all redundancies and all sequences with a percent identity greater than some specified value. The default setting is 90%. Thus, with this setting, only one protein of all the retrieved sequences with greater than 90% identity will be retained. If too many sequences are still retained, a lower cut-off value can be used. In this way, the desired number of sequences can be fed into the IC program.

A problem with the CD-Hit program is that the retained sequences may be fragments of complete protein sequences rather than the full length sequences. We have therefore modified CD-Hit so that only sequences of 'normal' length are retained. The program works as follows: the script summarizes the sizes of all the proteins obtained in a BLAST search, and a decision is made to exclude presumed fragmentary sequences. This is done by selecting a size range. All smaller sequences are eliminated.

When two sets of proteins are to be compared, two programs can be used: IC and GS (Get Score). The IC program is described above; the GS program functions as follows: The two lists of proteins are compared by (i) BLAST (18) and (ii) SSearch (19). In the latter program, for any binary comparison, the two bit scores are averaged, and based on a standard curve, they are converted to a comparison score expressed in standard deviations. Because SSearch compares the binary alignment with 500 randomly shuffled sequences, this program, like GAP and IC, corrects for abnormal amino acid compositions. An advantage of GS over IC is that it takes only about 1% as much computer time. Using programs to estimate integral membrane protein topologies (WHAT and AveHAS; 20,21), the parts of the proteins compared can be visualized.

ESTABLISHING SUPERFAMILY RELATIONSHIPS BETWEEN DISTANTLY RELATED FAMILIES

A major problem for phylogenetic tree construction arises when the sequences are so divergent that accurate multiple alignments cannot be generated. A novel program is therefore required for quantitating increasingly distant relationships. We have designed such a program and call it 'Supertree'. This program is based on BLAST searches and the resultant bit scores. There are several steps in its use. First, the query protein sequences (one for each family within the superfamily) are BLASTED against the NCBI protein database. Redundancies and sequences of greater than 70% identity (another cut off point can be used) are eliminated using the modified CD-Hit program described above. A small number of sequences (typically five) from each set are randomly selected by the program. All resultant sequences are compared with all other sequences using the Blastall program. The Blastall scores for all comparisons (e.g. if five sequences are selected, this is 5×5, or 25) are averaged, so a mean score is obtained for each family comparison. The resultant matrix is then used to generate a neighbor-joining tree. This process is conducted 100 times, and a consensus tree is generated using the program Consense (source code for both neighbor-joining and Consense are available at http://evolution.genetics.washington.edu/phylip.html). Finally, the tree is drawn using the TreeView program (22).

In some cases, when sequence divergence is not great, the Supertree method can be compared with approaches using traditional programs based on multiple alignments (CLUSTALX, T-COFFEE, etc.). In all such cases, the correlations have proven to be excellent.

DATA INPUT USING MACHINE LEARNING

A major improvement to TCDB has involved the development and implementation of state-of-the-art machine learning techniques. Machine learning is the field of artificial intelligence in computer science that uses computer programs to read specific data and use them to create a generalization called a model. We use proteins and cited articles present in TCDB (and therefore known to be relevant to transport) and create models that can identify *novel* proteins and articles that should be included in TCDB. Since our training sets come from general protein databases such as Swiss-Prot and TrEMBL, or Medline articles available in PubMed, our techniques are general purpose and directly applicable to many of the databases listed in this issue.

In order to keep TCDB constantly up-to-date, we need to identify new data that are relevant to transport but not already in TCDB. We consider two types of data, (i) UniProt protein records and (ii) Medline documents. Our techniques for working with each of these sources are largely similar, so we focus on Medline documents here. For a detailed description of our research involving UniProt records, see (23). The process of updating TCDB with new documents is as follows:

(i) Choose the training set of documents.
(ii) Identify the features of those documents to be used to make classifications.
(iii) Train a model.
(iv) Use the model to identify new documents.
(v) Verify that the predictions are correct.
(vi) Import the data into TCDB.

As noted above, the positive instances in our training set are the documents currently referenced in TCDB. We do not, however, have a corresponding reliably labeled negative set which is typically also provided as part of the input to a learning algorithm. However, we recently showed why we were able to do almost as well by using *unlabeled* Medline documents (24).

The features that we use are words that are associated with each document, either by appearing in the document itself, or by being part of a set of keywords associated with the document. That is, each word is a numerical feature, and its value is proportional to the number of times it appears. This representation is sometimes called a 'bag of words', since the multiplicity of each word is considered, but not the order in which it appears. We also separate out different sources of these words. For instance, we consider the word 'transport' as appearing in the title of an article to be a different feature than if this word were to

appear in the document's abstract. In fact, we do not typically consider words in the body of a document but limit our representation to words in the document's title and abstract. We also weigh author names, affiliations, and keywords associated with the documents.

To accomplish the third and fourth steps (training a model and using it to identify novel documents), we use a classifier model called a support vector machine (25) and a standard associated learning algorithm. We use our derived models to rank a set of candidate articles according to a score that is proportional to the likelihood that the new article is related to membrane transport, given its features (i.e. words). We then examine each article in order, starting with the most likely transport article according to our model. For each, we identify the appropriate proteins and associated information and insert these into TCDB. We continue this process until we determine that the frequencies of relevant articles are insufficient to be of use.

In a period of less than 9 months, we have identified 1255 articles that are related to transport, 742 of which have been added to TCDB (an increase of 21%). The remaining 513 articles were not added to TCDB because they described proteins that are very similar in sequence and function to proteins that were already in TCDB or because they were not important enough for some other reason. For further details about our learning approach and deployment statistics, see (26).

We focus on a set of about 100 journals that are the most cited in TCDB. In a month, we typically have about 6000 articles to consider as potential sources of transport information. We cannot examine them all, but we expect about 2–3% of them to be relevant. The accuracy of our models depends on how many articles we examine. For instance, if we look at the 10 highest-scoring articles, we observe an accuracy of nearly 100%. When we examine the top 100, we observe about 48% accuracy, and if we look at the top 300, we observe about 28% accuracy. This approach gives us a recall comparable to that of a human expert, but the human expert needs only to examine a relatively small number of false positives to find the relevant ones.

CONCLUSIONS AND PERSPECTIVES

We are currently developing a few additional methods to facilitate the introduction of new proteins and families into TCDB. In the next version of the database, we plan to allow users to submit their own sequenced proteins and descriptions for inclusion. We are also experimenting with novel ways to automatically identify proteins that are associated with our document examples. One source is NCBI's curated databases, but these are incomplete and need to be supplemented with named entity recognition (NER) techniques. NER involves automatically parsing text, such as document abstracts, into categories. Once protein names and other identifiers are found, one can look them up in general databases and extract the necessary information.

Associating articles with proteins helps us in two important ways. First, it provides our learning algorithms with key additional features which should increase the accuracy of the models. Second, it will help us identify and organize the data associated with a protein (sequence, protein family, etc.) that go into TCDB. This can help automate the process of importing new data.

The vast amount of protein sequence data now available renders data mining essential for maximizing output. TCDB development often depends on preexisting programs, but we must also design and update software in order to refine and optimize data input concerned with the functions, mechanisms, topologies, structures, phylogenetic relationship and evolutionary origins of transport proteins. TCDB can serve as a model system for the expansion of database technologies useful for many purposes.

FUNDING

This work was supported by National Institutes of Health (grant number GM1077402). Funding for open access charge: National Institute of General Medical Sciences, National Institutes of Health (R01 GM077402).

Conflict of interest statement. None declared.

REFERENCES

1. Busch,W. and Saier,M.H. Jr. (2002) The IUBMB-Endorsed transporter classification system. *Mol. Biotech.*, **27**, 253–262.
2. Saier,M.H. Jr (2003) Tracing pathways of transport protein evolution. *Mol. Microbiol.*, **48**, 1145–1156.
3. Saier,M.H. Jr, Tran,C.V. and Barabote,R.D. (2006) TCDB: The transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.*, **34**, D191–D186.
4. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, **2**, 953–971.
5. Pollock,D.D. (2002) Genomic biodiversity, phylogenetics and coevolution in proteins. *Appl. Bioinformatics*, **1**, 81–92.
6. Yen,M.R., Peabody,C.R., Partovi,S.M., Zhai,Y, Tseng,Y.H. and Saier,M.H. Jr. (2002) Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochim. Biophys. Acta.*, **1562**, 6–31.
7. Zhai,Y. and Saier,M.H. Jr. (2002) A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. *J. Mol. Microbiol. Biotechnol.*, **4**, 375–377.
8. Zhou,X., Yang,N., Tran,C. V., Hvorup,R. and Saier,M.H. Jr. (2003) Web-based programs for the display and analysis of transmembrane α-helices in aligned protein sequences. *J. Mol. Microbiol. Biotech.*, **5**, 1–6.
9. Doolittle, R.F. (1986) *Of URFS and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, CA.
10. Saier,M.H. Jr. (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol. Rev.*, **58**, 71–93.
11. Barabote,R.D., Tamang,D.G., Abeywardena,S.N., Fallah,N.S., Fu,J.Y.C., Lio,J.K., Mirhosseini,P., Pezeshk,R., Podell,S., Salampessy,M.L. *et al.* (2006) Extra domains in secondary transport carriers. *Biochim. Biophys. Acta.*, **1758**, 1557–1579.
12. Serres,M.H. and Riley,M. (2005) Gene fusions and gene duplications: relevance to genomic annotation and functional analysis. *BMC Genomics.*, **6**, 33.
13. Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524–545.

14. Saier,M.H. Jr (2000) A functional-phylogenetic classification system for transmembrane solute trans-porters. *Microbio. Mol. Biol. Rev.*, **64**, 354–411.
15. Devereux,J., Haeberli,P. and Smithies,O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.
16. Chang,A.B., Lin,R., Keith Studley,W., Tran,C.V. and Saier,M.H. Jr. (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Mol. Membrane Biol.*, **21**, 171–181.
17. Li,W. and Godzik,A. (2006) CD-Hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
18. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
19. Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.
20. Zhai,Y. and Saier,M.H. Jr. (2001) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J. Mol. Microbiol. Biotech.*, **3**, 501–502.
21. Zhai,Y. and Saier,M.H. Jr. (2001) A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. *J. Mol. Microbiol. Biotechnol.*, **3**, 285–286.
22. Zhai,Y., Tchieu,J. and Saier,M.H. Jr. (2002) A web-based Tree View (TV) program for the visualization of phylogenetic trees. *J. Mol. Microbiol. Biotechnol.*, **4**, 69–70.
23. Das, S., Saier, M.H. Jr. and Elkan, C. (2007) Finding transport proteins in a general protein database. In *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, *Lecture Notes in Computer Science*, Vol. 4402, Springer, Berlin, Heidelberg, pp. 54–66.
24. Elkan, C. and Noto, K. (2008) Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, ACM, New York, pp. 213–220.
25. Cortes,C. and Vapnik,V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273–297.
26. Noto,K., Saier,M.H. Jr. and Elkan,C. (2008) Learning to find relevant biological articles without negative training examples. In Wobcke,W.R. and Zhang,M. (eds), *AI 2008: Advances in Artificial Intelligence, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg.