Title

The Transporter Classification Database (TCDB): recent advances

Journal

Nucleic Acids Research, 44(D1)

ISSN

0305-1048

Authors

Saier, Milton H
Reddy, Vamsee S
Tsu, Brian V
et al.

Publication Date

2016-01-04

DOI

10.1093/nar/gkv1103

Peer reviewed

D372–D379 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 5 November 2015*
doi: 10.1093/nar/gkv1103

The Transporter Classification Database (TCDB): recent advances

Milton H. Saier, Jr[1,*], Vamsee S. Reddy[1,2], Brian V. Tsu[1], Muhammad Saad Ahmed[3], Chun Li[3] and Gabriel Moreno-Hagelsieb[1,4]

[1]Department of Molecular Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA, [2]Department of Medical Sciences, Boston University School of Medicine, 72 E Concord St., Boston, MA 02118, USA, [3]Department of Biological Engineering, School of Life Sciences, Beijing Institute of Technology, Beijing 100081, People's Republic of China and [4]Department of Biology, Wilfrid Laurier University, 75 University Ave W, Waterloo, ON, Canada N2L 3C5

Received September 14, 2015; Revised October 01, 2015; Accepted October 11, 2015

ABSTRACT

The Transporter Classification Database (TCDB; http://www.tcdb.org) is a freely accessible reference database for transport protein research, which provides structural, functional, mechanistic, evolutionary and disease/medical information about transporters from organisms of all types. TCDB is the only transport protein classification database adopted by the International Union of Biochemistry and Molecular Biology (IUBMB). It consists of more than 10 000 non-redundant transport systems with more than 11 000 reference citations, classified into over 1000 transporter families. Transporters in TCDB can be single or multi-component systems, categorized in a functional/phylogenetic hierarchical system of classes, subclasses, families, subfamilies and transport systems. TCDB also includes updated software designed to analyze the distinctive features of transport proteins, extending its usefulness. Here we present a comprehensive update of the database contents and features and summarize recent discoveries recorded in TCDB.

INTRODUCTION

Membrane transporters represent a diverse group of proteins that form intricate networks of channels, carriers, pumps, group translocators and electron flow carriers that determine the molecular compositions and energy states of cells (1). These proteins transfer nutrients, end products of metabolism, toxic substances, macromolecules, signaling molecules, electrons and many other cellular constituents from source to sink, resulting in the cellular uptake and extrusion of compounds (2). Of particular importance to the fields of oncology, microbial pathogenesis and virology, drug efflux pumps play a dominant role in drug resistance (3,4). Thousands of researchers worldwide contribute to the collective understanding of molecular transport across cellular membranes (5).

In June 2001, the International Union of Biochemistry and Molecular Biology (IUBMB) formally adopted the Transporter Classification (TC) system as the only internationally recognized system for the organization of transport protein information derived from a diverse range of organisms (6,7). With the advent of genome sequencing and recent progress made in computational biology, a long-standing interest in membrane transporters has allowed the Saier Laboratory at UCSD to comprehensively design a system in some respects comparable to the Enzyme Commission (EC) system for classifying enzymes, but from evolutionary, structural and functional standpoints (2,8). Earlier versions of the TC database (TCDB) have been described in previous publications in Nucleic Acids Research (7,9,10).

Freely accessible, TCDB (www.tcdb.org), provides structured access to data published in over 11 000 papers. This information is integrated into descriptions and hierarchical structures within TCDB. The database contains more than 10 000 single- or multi-component transport systems from all kinds of living organisms. These systems are classified into more than 1000 transporter families based on their phylogeny and function. The classes and subclasses established in the database are presented in Table 1. TCDB is continually updated as published research regarding transport systems becomes available.

TCDB is not limited to the classification of well-characterized transport proteins, it also attempts to classify and elucidate the roles of many transporters that are poorly understood, even when minimal information is available in the published literature. The TCDB research group maintains an up-to-date list of software for computational analysis of transport proteins, software that has been established as 'best-practices' through time-tested experience. The database also provides the tools to search, view, compare and download relevant information. In this summary

*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 (1 September 2015)

1. Channels/pores
1.A α-Type channels
1.B β-Barrel porins
1.C Pore-forming toxins (protein and peptide)
1.D Non-ribosomally synthesized channels
1.E Holins
1.F Vesicle fusion pores
*1.G Viral fusion pores
*1.H Paracellular channels
*1.I Membrane-bounded channels
*1.J Virion egress pyramidal apertures
*1.K Phage DNA injection channels
*1.L Tunneling nanotubes, TNTs
2. Electrochemical potential-driven transporters
2.A Porters (uniporters, symporters, antiporters)
2.B Nonribosomally synthesized porters
2.C Ion gradient-driven energizers
*2.D Transcompartment carriers
3. Primary active transporters
3.A P-P-bond-hydrolysis-driven transporters
3.B Decarboxylation-driven transporters
3.C Methltransfer-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
*4.D Polysaccharide synthase/exporters
*4.E Vacuolar polyphosphate polymerase-catalyzed group translocators
5. Transmembrane electron carriers
5.A Transmembrane two-electron transfer carriers
5.B Transmembrane one-electron transfer carriers
8. Accessory factors involved in transport
8.A Auxiliary transport proteins
*8.B Ribosomally synthesized protein/peptide toxins/agonists 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

*Added since January 2014.

paper, the contents and web features of TCDB are described with a focus on recent discoveries, improvements and additions.

THE TC SYSTEM

The TC system includes transport proteins, many of which are characterized structurally and functionally. However, efforts have also been made to include constituents of transport protein families with divergent sequences, organismal sources and functions. Each transport system is classified with a five-character TC identifier: N.L.N2.N3.N4 where N is a number and L is a letter. N represents a transporter class, i.e. 1. Channels/pores, 2. Electrochemical potential-driven transporter, 3. Primary active transporters, 4. Group translocators, 5. Transmembrane electron carriers, 8. Auxiliary transport proteins, and 9. (Putative) transport system that are incompletely characterized (Table 1). L represents the transporter subclass of which there are currently 12 channel subclasses (A–L), four carrier subclasses (A–D), five primary active transporter subclasses (A–E), five group translocator subclasses (A–E), two transmembrane electron carrier subclasses (A and B), three auxiliary subclasses (A–C) and three poorly characterized transporter subclasses (A–C) (see Table 1). N2 represents the transporter families of which there are over 1000 in TCDB, while N3 represents subfamilies or phylogenetic clusters within a family. N4 represents the transport system itself with a distinctive range of substrates and/or energy sources. Most families consist of several phylogenetically distinct subfamilies, and over the years, several families have expanded into sequence divergent superfamilies, still maintained under a single (super)familial TC number (N2). However, in most superfamilies, different families are assigned different familial N2 TC numbers. For historical reasons, 'hybrid' superfamilies exist where many families are maintained under a single N2 TC number, but others are listed with dissimilar N2 TC numbers in the TC superfamily hyperlink. The latter are the families that were found to belong to a superfamily after both families became established in TCDB.

EXAMPLES OF FUNCTIONALLY DIVERSE SUPERFAMILIES

The Major Facilitator Superfamily (MFS) is an example where numerous families are listed under the original superfamilial N2 TC identifier (2.A.1), while other families, which were later shown to be distant members of the MFS, are associated with other familial N2 TC numbers (i.e. the POT or PTR family of peptide transporters; TC-ID 2.A.17). While most families of transporters include members that only catalyze transport, the MFS, which apparently arose by two duplication events (3→6→12 TMSs), has long been known to include homologues that function as receptors in addition to or instead of their transport functions (11). More recently, integral membrane proteases that cleave intramembrane peptide bonds in transmembrane proteins (TC-ID 9.B.104) have been shown to be, in part or in full, homologous of MFS carriers. Glycosyl transferases that mediate exopolysaccharide synthesis, possibly with concomitant export of the growing carbohydrate chain (TC-ID 4.D.2), have similarly been found to be members of the MFS, having diverged from a common precursor protein (S. Wang, I. Javadi-Razaz & M. Saier, manuscript in preparation). Interestingly, some Mg^{2+}-import P-type ATPases (TC-ID 3.A.3.4) possess N-terminal domains of unknown function that show extensive sequence similarity with MFS carriers (12). It seems that the ancient MFS repeat structure has evolved with transmembrane functional diversification.

Another superfamily that includes non-transporters is the recently described Transporter-Opsin-G Protein-coupled Receptor (TOG) superfamily (13,14). In addition to eight recognized families of transporters, including secondary carriers, a probable group translocator (PnuC) in a family of secondary carriers (TC-ID 4.B.1), light-driven ion pumps and light-activated ion channels (MR; TC-ID 3.E.1), this superfamily includes the large G-Protein-coupled Receptor (GPCR) family (TC-ID 9.A.14) of diverse seven TMS receptors and opsins, most of which lack recognizable transport functions (15,16).

Table 2. Proposed evolutionary pathways for the appearance of well-characterized families and superfamilies

TC-ID[a]	(Super)family[b]	Pathway[c]	Reference
3.A.1	ABC1	**2**$\overset{x3}{\rightarrow}$**6**	(20)
3.A.1	ABC2	**3**$\overset{x2}{\rightarrow}$**6**	(20,21)
3.A.1	ABC3	**4**$\overset{x2}{\rightarrow}$ { $\rightarrow$ **8**; $\rightarrow$ **10*** }	(20,46)
2.A.3	APC	**5** { $\overset{x2}{\rightarrow}$ **10** $\overset{\pm 1,2 \text{ or } 3}{\rightarrow}$ 9 – 13; $\overset{+2}{\rightarrow}$ 7 $\overset{x2}{\rightarrow}$ 14 }	(47)
2.A.45	ArsB (IT Superfamily)	**6**$\overset{x2}{\rightarrow}$**12**	(48)
2.A.28	BART	**5**$\overset{x2}{\rightarrow}$**10**	(49)
2.A.4	CDF	2 $\overset{x3}{\rightarrow}$ 6 { $\overset{-2}{\rightarrow}$ 4; $\overset{x2}{\rightarrow}$ 12 $\overset{-1}{\rightarrow}$ 11 $\overset{-1}{\rightarrow}$ 10 }	(24)
9.A.55	CopD	**4**$\overset{x2}{\rightarrow}$**8**$\overset{x2}{\rightarrow}$**16**	Unpublished
2.A.7	DMT	**2**$\overset{x2}{\rightarrow}$**4**$\overset{+1}{\rightarrow}$**5**$\overset{x2}{\rightarrow}$**10**	(50)
1.E	Holins	**1**$\overset{+1}{\rightarrow}$**2; 2**$\overset{+1}{\rightarrow}$**3;** 3$\overset{+1}{\rightarrow}$**4; 2**$\overset{x2}{\rightarrow}$**4**	(51)
2.A.75	LysE	**3**$\overset{x2}{\rightarrow}$**6**	(69)
1.A.8	MIP	**3**$\overset{x2}{\rightarrow}$**6**	(52)
2.A.1	MFS	**3**$\overset{x2}{\rightarrow}$**6**$\overset{x2}{\rightarrow}$**12**$\overset{+2}{\rightarrow}$**14****	(53)
2.A.29	MC	**2**$\overset{x3}{\rightarrow}$**6**	(54)
1.A.77	MCU	**2**$\overset{x2}{\rightarrow}$**4**	(55)
1.A.72	Mer	**2**$\overset{x2}{\rightarrow}$**4**	(56)
2.A.66	MOP	**6**$\overset{x2}{\rightarrow}$**12**	(57)
2.A.67	OPT	2 $\overset{x2}{\rightarrow}$ 4 $\overset{x2}{\rightarrow}$ 8 $\overset{x2}{\rightarrow}$ 16($\overset{+1}{\rightarrow}$ 17)***	(58)
4.A.5	PTS-AG	**5**$\overset{x2}{\rightarrow}$**10**	(59)
4.A.1	PTS-GFL	**5+5**$\rightarrow$**10**	(60)
2.A.6	RND	**6**$\overset{x2}{\rightarrow}$**12**	(61)
1.H.1	4JC	**2**$\overset{x2}{\rightarrow}$**4**	(62)
2.A.43	TOG	4 $\overset{x2}{\rightarrow}$ 8 $\overset{-1}{\rightarrow}$ 7	(14)
1.A.28	UT/NQR	**5**$\rightarrow$**10**$\rightarrow$**20**	(63)

[a]Superfamilies do not have a TC-ID and can be found in the superfamily hyperlink in TCDB. The TC-ID provided is for a representative family.
[b]See TCDB for the full Superfamily/Family names and the constituent families found in each superfamily.
[c]The evolutionary pathways are presented in the TC evolutionary pathways link, where the number refers to the numbers of TMSs in a repeat unit or a transmembrane transport protein.
[*]The two extra TMSs in the 10 TMS proteins appear to be centrally located between the two 4TMS repeat units.
[**]Several families include member with 14 instead of 12 TMSs, and the two extra TMSs appear to be located between the two 6 TMS repeat units
[***]Only one subfamily of the OPT family exhibits a 17th TMS.

Another superfamily, the Lysine Exporter (LysE) superfamily, which has recently been expanded to include eleven families, includes DsbD electron carriers (TC-ID 5.A.1) that function in the transmembrane transfer of electron pairs from an intracellular electron donor to an extracellular acceptor (17,18). In this case, the entity exported is not a classical molecular species, but a simple pair of electrons that functions to reduce an extracellular oxidized molecular entity, such as a disulfide bond in a protein (19). Thus, while most families of transport systems contain only transporters, a few have diverged, serving other functions. TCDB, presenting a functional/phylogenetic system of classification, has expanded to include homologues of transporters as well as group translocators that function in both enzyme or electron flow catalysis and molecular transport, even though some of these have lost their original transport function.

The ABC superfamily (3.A.1) provides another interesting but dissimilar case. According to the rules of the TC classification system, family assignment is normally based on the phylogeny of the integral membrane protein(s) that comprise the transmembrane transport pathway (20,21). However, early assignment to the ABC superfamily (3.A.1) by other laboratories was based on the presence of the cytoplasmic ATP hydrolyzing constituents of these systems, the ATPases, because of their greater sequence conservation and consequent ease of recognition (22). Later studies, however, showed that the integral membrane constituents of these systems fall into at least three independently evolving families which appeared via distinctive routes starting with different peptide precursors (see Table 2) (20). Nevertheless, to maintain a static and compatible system of classification, as required by the IUBMB, these multicomponent systems are still maintained under a single superfamilial TC identifier (3.A.1).

Table 3. External databases linked to transport systems in TCDB and their specific roles. Several other external databases are linked to transport systems in TCDB. However, some of these are (partially) redundant with those listed here[a]

External Database (reference)[b]	Role in TCDB
RCSP PDB (26)	Provides high-resolution crystal or NMR structures for transport proteins when available
Interpro (27)	Provides functional analysis of transporters by classifying them according to domains and other important sites or motifs
GENEW (28)	Almost all human transporters can now be found in TCDB and are annotated using the Human Gene Nomenclature DB system
UniProt Disease (29)	Users can identify diseases associated with any transporter class, (super)family, or system
PFAM (30)	Transport systems are cross-referenced with PFAM methods of classification, dependent upon conserved domains
KEGG (31) & BioCyc (32)	These resources allow transport systems to be linked with metabolomics data
EggNOG (33)	Multiple alignments and phylogenetic trees containing a protein of interest can be viewed
RefSeq (34)	Information about a gene source is provided with comments and references
Entrez Gene ID (35)	Information about the protein's gene context can be viewed
DrugBank (36)	Known drug information related to a transporter can be accessed
DIP (37)	Database of Interacting Proteins indicates other proteins that may interact with a transporter
HEGENOM	Phylogenetic trees related to a transporter can be viewed
EcoGene (38)	Gene context with annotations from a curated *E. coli* database can be viewed
GO (39)	Gene Ontology provides functional annotations to transport proteins when available

[a]EchoBase has been deprecated since our last publication, due to its redundant role in our external databases. Several databases serve similar roles but achieve them through different methods (i.e. gene contexts/phylogenetic trees) and consequently have been retained in this table.
[b]Numbers in parentheses provide the most recent reference to the database described.

POORLY CHARACTERIZED FAMILIES

Of the TC classes, only Class 9 (see Table 1) includes transporters of unknown biochemical mechanism and/or mode of transport energization. These systems will be classified elsewhere in the TC system when their functions and mechanisms become elucidated. As noted above, we have recently found that some proteins with receptor or enzyme activities, sometimes lacking transport activities, are present in transport protein families, and these are retained in TCDB, often in class 9. However, if proteins listed in TC class 9 do not prove to be either transporters or transporter homologues, they will be removed from the TC system. Only class 9 is in a continual state of flux.

TRANSPORT PROTEINS THAT CAN FUNCTION BY MORE THAN ONE MECHANISM

The TC system was founded based on the observation that functional class and phylogenetic grouping correlate, almost without exception. However, within a few families, there are examples of transporters that either can employ more than one mode of action or can use an energy-coupling mechanism distinct from that of other members of the family. Thus, the well-defined transporter classes 1–5 usually include completely different, non-overlapping families. However, a limited amount of overlap now exists. Some members of carrier families, for example, are believed to be 'broken' (mutated) and can only function by a more simple channel mechanism, e.g. some members of the CLC family (TC-ID 2.A.49) (23). One superfamily, the CDF Superfamily of six TMS carriers, lost two TMSs, evolving into four TMS channels (see Table 2), a process we have called 'retro-evolution' (24). Additionally, some ATP-hydrolysis-dependent primary active transporters can function without the ATPase, catalyzing solute transport by a channel or a carrier mechanism, e.g. the ArsAB and ArsB families (TC-IDs 3.A.4 and 2.A.45) and the ECF subset of ABC transporters (TC-IDs: 3.A.1 and 2.A.88) (25).

DATABASE CONTENT AND ACCESS

As noted above, TCDB provides free access to structural, functional, physiological, mechanistic and evolutionary information about transporters derived from numerous sources. Further, this resource has been expanded by crosslinking TCDB to several other databases (26–40). These extended databases serve several specific roles in the context of TCDB (Table 3).

Users can access transport systems and their constituent transport proteins as well as transporter classes, subclasses, and families by clicking on 'search' at the top right of the main TCDB page. Single or multiple terms can be used, i.e. UniProt accession numbers (41), PDB ID numbers (26), TC identifiers, protein names, key words, abbreviations, associated diseases, organismal sources, author's names, references, etc. Efforts by the TCDB curators to connect TCDB to other databases can be exemplified by the recent comparison between the contents and philosophies of TCDB and Pfam, which resulted in improved mapping of TCDB protein families and superfamilies to Pfam, and in the addition of new protein families to the upcoming version of Pfam (40).

Using the 'Superfamilies' hyperlink on the home page, researchers can access the TC superfamilies (Table 4) with a brief description of their characteristics, references and a list of all currently recognized family members. In each family entry, one can find a continuously updated family description and access the representative protein members of that family. On the clickable, 'analyze' tab, the user can assess protein properties and determine sequence similarities to other proteins in TCDB. Access to BIO-TOOLS allows the user to determine many features of transport proteins such as topology, conservation, and common motifs.

Table 4. Superfamilies in TCDB

Superfamily abbreviation	Number of families[a]
ABC1, 2 & 3	3 (96)
Aerolysin	7
APC	27
ArsA	4
BB	7
BART	7
CAAX	2
CDF	4
Cecropin	7
CBB	4
CopR	3
CPA	5
Defensin	4
DMT	1 (35)
ENaC/P2X	2
EMPT	3
GT	3
Holin I	1 (3)
Holin II	4
Holin III	7
Holin IV	4
Holin V	2
Holin VI	2
Holin VII	1 (6)
Huwentoxin	8
IT	21
LysE	11
MACPF	3
MIP	2 (18)
Mer	1 (5)
MFS	17 (100)
MC	2 (32)
MOP	1 (12)
Mrp	3
OAPol	2
P-ATPase	1 (20)
Porin I	48
Porin II	2
Porin III	2
Porin IV	2
Porin V	2
PTS-AG	2
PTS-GFL	4
PBB	2
RND	1 (9)
RTX	3
TAMP-I	2
4JC	15
Tim17	3
TOG	9
TRC/TAMP	2
UT/RnfD/NqrB[b]	3
VIC	9 (37)
Env-FP	5

[a]The number of families with distinct familial TC-IDs (e.g. TC-ID N.L.N2) is provided for all entries, while the number in parentheses indicates the total number of families listed under TC-ID N.L.N2.N3 for families that have been assigned superfamily status.
[b]All six constituents of each system within the Rnf family are homologous to the corresponding constituents of each system within the Nqr family.

A left-vertical-frame architecture of the main page gives an option to use either BLAST or *PSI*-BLAST for the query protein against either the non-redundant NCBI (National Center for Biotechnology Information) database or the TCDB protein database. The 'STRUCTURE DATA' option shows 3D structures of transporters when known, along with the source organism, TC-ID, PDB ID and a description of the transport protein.

The new 'HUMAN TRANSPORTERS' link shows the transporters specific for different compounds present in humans, while the revised 'TRANSPORTERS & DISEASES' link shows the transporters associated with specific diseases. Most currently recognized human transporter are now included in TCDB.

DATABASE GROWTH (JANUARY 2014–DECEMBER 2015)

Nearly 100 new families are introduced into TCDB every year, representing the continual effort devoted to expanding database entries. Table 1 indicates the current classes and sub-classes with the newly identified subclasses marked by asterisks, showing close to a 30% increase since January, 2014. These novel subclasses include six specialized types of channels, a newly identified trans-compartment transport carrier, two novel types of group translocators, and two types of transporter-targeting toxins and agonists.

Table 4 presents the TC superfamilies, many of which have been added or expanded during the last two years (2014 and 2015, inclusive). Numerous preexisting superfamilies and families have been expanded and even merged because of the development of more sensitive strategies for detecting distant phylogenetic relationships (42), greater sequence representation and novel published information. This has also allowed us to recognize the existence of super-superfamilies that exhibit distant relationships between preexisting superfamilies. For example, we have evidence that the Voltage-gated Ion Channel (VIC) Superfamily includes the Cation:Proton Antiporter (CPA) and Anoctamin Superfamilies (D.J. McLaughlin, Z.S. Ye and M.H. Saier, Jr., manuscript in preparation). In some cases, the evolutionary pathways giving rise to the appearance of these proteins are predicted as recorded in a new TC link entitled 'Predicted Evolutionary Pathways' (PEP) (see also Table 2). TCDB now includes 820 PDB protein crystal structures, and these are automatically updated on a regular basis as new structures become elucidated.

BIOTOOLS FOR SEQUENCE ANALYSIS

Biotools is a collection of bioinformatics software for the analysis of protein sequences with emphasis on those most useful in transport protein analyses (Table 5). Several of these tools were adopted from outside sources, but all BioV Suite programs (43) and many others were developed in-house. The active user will benefit from the utilization of these analytical programs for confirming or refuting published reports and allowing the generation of novel data.

CONCLUSIONS

In 2006, TCDB contained 3000 transport proteins that were classified into 400 families with only a few recognized superfamilies, but by the end of 2015, these transporters exceeded 10,000 and were classified into more than 1000 families and over 60 superfamilies. The accessibility of TCDB has facilitated major fundamental studies of transporter structure

Table 5. Software available on BioTools for the analysis of transport protein sequences

Biotool Program (reference)[a]	Brief Description
MEMSAT (64)	Predicts the positions of transmembrane segments
Clustal Ω (65)	Creates multiple sequence alignments that can be passed on to TreeView or FigTree to construct phylogenetic trees
MAFFT (66)	Multiple alignment tool for proteins & nucleic acids
GSAT-Pairwise Alignment (43)	A Global Sequence Alignment Tool
TCDB Topology Stats (IH[b])	Statistically analyzes integral membrane proteins for topology in families or superfamilies
HHrepID (67)	Detects protein repeats using a dot-plot for analysis
WHAT (18)	Plots hydropathy and amphipathicity of proteins from N- to C-terminus using an adjustable sliding window
AveHAS (17)	Determines the average hydropathy, amphipathicity & similarity, again as function of protein length, for a set of multiply aligned sequences and predicts topology using two different programs
ChkAll (IH)	Counts amino acids with different properties in aligned sequences
TMS Align (IH)	Aligns multiple protein sequences & labels putative transmembrane segments within each aligned sequence
TMS Split (IH)	Splits protein sequences either between any two TMSs, frequently between equal parts for repeat unit identification
COMPARE (IH)	Compares two FASTA files to determine which sequences are unique and which are shared
Table Extract (IH)	Extracts various information from FASTA files to create a protein table
GI Extract (IH)	Extracts GI numbers from NCBI and removes the redundant GI numbers from NCBI blast results
Sequence Extract (IH)	Extracts protein sequences from NCBI based only on GI number. Can be used in conjunction with REELIMI
REELIMI (IH)	Eliminates protein sequence redundancies based on sequence similarity and other factors
HMMTop (68)	Predicts the positions of transmembrane segments
SortTree (IH)	Sorts any FASTA file or TAB file according to a phylogentic tree
Aln2Fas (IH)	Converts clustal alignment files to FASTA format to select the range of residues desired
BioV - Protocol1 (43)	Automatically generates representative FASTA list and removes similar/redundant sequences from a single input
BioV - Protocol2 (43)	A highly sensitive method for detecting homology between distantly related proteins
BioV - Ancient Rep(AR) (43)	A method used to find very old intergenic transmembrane repeat units, which can be used as criteria for establishing homology in some contexts

[a]References, when available, are provided in parentheses.
[b]IH: in house. Several of the tools listed adapt existing software to TCDB's needs, and/or solve common problems in reformatting and downstream analyses of results from other tools.

and function. Moreover, evolutionary routes for the appearance of these proteins have been elucidated as a result of the research conducted by X-ray crystallographers and the TCDB research group (44). These are now presented in a new link in TCDB entitled 'Predicted Evolutionary Pathways (PEP)', briefly summarized in Table 2. Superfamily identification allows extrapolation of structural, functional and mechanistic data from one or a few homologues to many (45). While our laboratory has been concerned with identifying superfamily relationships, we are aware that this work is still in its infancy. Novel biotools continue to be designed for analyzing the distinctive features of transport proteins and to further extend the capacities and usefulness of TCDB.

ACKNOWLEDGEMENTS

We thank the many professionals and students in the Saier Laboratory who contributed to the advances reported in this communication and Yongxin Hu for assistance with the preparation of this manuscript.

FUNDING

National Institutes of Health [GM 077402]. Funding for open access charge: National Institutes of Health [GM 077402].

Conflict of interest statement. None declared.

REFERENCES

1. Cook,G.M., Greening,C., Hards,K. and Berney,M. (2014) Energetics of pathogenic bacteria and opportunities for drug development. *Adv. Microb. Physiol.*, **65**, 1–62.
2. Busch,W. and Saier,M.H. Jr. (2002) The transporter classification (TC) system, 2002. *Crit. Rev. Biochem. Mol. Biol.*, **37**, 287–337.
3. Putman,M., van Veen,H.W. and Konings,W.N. (2000) Molecular properties of bacterial multidrug transporters. *Microbiol. Mol. Biol. Rev.*, **64**, 672–693.
4. Delmar,J.A., Su,C.C. and Yu,E.W. (2014) Bacterial multidrug efflux transporters. *Annu. Rev. Biophys.*, **43**, 93–117.
5. Yen,M.R., Choi,J. and Saier,M.H. Jr. (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. *J. Mol. Microbiol. Biotechnol.*, **17**, 163–176.
6. Busch,W. and Saier,M.H. Jr. (2003) The IUBMB-endorsed transporter classification system. *Methods Mol. Biol.*, **227**, 21–36.
7. 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**, D181–D186.
8. Saier,M.H. Jr. (2003) Tracing pathways of transport protein evolution. *Mol. Microbiol.*, **48**, 1145–1156.
9. Saier,M.H. Jr, Yen,M.R., Noto,K., Tamang,D.G. and Elkan,C. (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res.*, **37**, D274–D278.
10. Saier,M.H. Jr, Reddy,V.S., Tamang,D.G. and Vastermark,A. (2014) The transporter classification database. *Nucleic Acids Res.*, **42**, D251–D258.

11. Dietvorst,J., Karhumaa,K., Kielland-Brandt,M.C. and Brandt,A. (2010) Amino acid residues involved in ligand preference of the Snf3 transporter-like sensor in Saccharomyces cerevisiae. *Yeast*, **27**, 131–138.
12. Chan,H., Babayan,V., Blyumin,E., Gandhi,C., Hak,K., Harake,D., Kumar,K., Lee,P., Li,T.T., Liu,H.Y. *et al.* (2010) The p-type ATPase superfamily. *J. Mol. Microbiol. Biotechnol.*, **19**, 5–104.
13. Zhai,Y., Heijne,W.H., Smith,D.W. and Saier,M.H. Jr. (2001) Homologues of archaeal rhodopsins in plants, animals and fungi: structural and functional predications for a putative fungal chaperone protein. *Biochim. Biophys. Acta*, **1511**, 206–223.
14. Yee,D.C., Shlykov,M.A., Vastermark,A., Reddy,V.S., Arora,S., Sun,E.I. and Saier,M.H. Jr. (2013) The transporter-opsin-G protein-coupled receptor (TOG) superfamily. *FEBS J.*, **280**, 5780–5800.
15. Krumm,B.E. and Grisshammer,R. (2015) Peptide ligand recognition by G protein-coupled receptors. *Front. Pharmacol.*, **6**, 48.
16. Jaeger,W.C., Armstrong,S.P., Hill,S.J. and Pfleger,K.D. (2014) Biophysical detection of diversity and bias in GPCR function. *Front. Endocrinol. (Lausanne)*, **5**, 26.
17. 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.
18. 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. Biotechnol.*, **3**, 501–502.
19. Kadokura,H. and Beckwith,J. (2010) Mechanisms of oxidative protein folding in the bacterial cell envelope. *Antioxid. Redox Signal.*, **13**, 1231–1246.
20. Wang,B., Dukarevich,M., Sun,E.I., Yen,M.R. and Saier,M.H. Jr. (2009) Membrane porters of ATP-binding cassette transport systems are polyphyletic. *J. Membr. Biol.*, **231**, 1–10.
21. Zheng,W.H., Vastermark,A., Shlykov,M.A., Reddy,V., Sun,E.I. and Saier,M.H. Jr. (2013) Evolutionary relationships of ATP-Binding Cassette (ABC) uptake porters. *BMC Microbiol.*, **13**, 98.
22. Higgins,C.F. (1992) ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.*, **8**, 67–113.
23. Duran,C., Thompson,C.H., Xiao,Q. and Hartzell,H.C. (2010) Chloride channels: often enigmatic, rarely predictable. *Annu. Rev. Physiol.*, **72**, 95–121.
24. Matias,M.G., Gomolplitinant,K.M., Tamang,D.G. and Saier,M.H. Jr. (2010) Animal Ca2 +release-activated Ca2+ (CRAC) channels appear to be homologous to and derived from the ubiquitous cation diffusion facilitators. *BMC Res Notes*, **3**, 158.
25. Rice,A.J., Park,A. and Pinkett,H.W. (2014) Diversity in ABC transporters: type I, II and III importers. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 426–437.
26. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
27. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
28. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32**, D255–D257.
29. Famiglietti,M.L., Estreicher,A., Gos,A., Bolleman,J., Gehant,S., Breuza,L., Bridge,A., Poux,S., Redaschi,N., Bougueleret,L. *et al.* (2014) Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Hum. Mutat.*, **35**, 927–935.
30. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–230.
31. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
32. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.A., Holland,T.A., Keseler,I.M., Kothari,A., Kubo,A. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.
33. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldon,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
34. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
35. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
36. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.
37. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–451.
38. Zhou,J. and Rudd,K.E. (2013) EcoGene 3.0. *Nucleic Acids Res.*, **41**, D613–D624.
39. Gene Ontology C. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
40. Chiang,Z., Vastermark,A., Punta,M., Coggill,P.C., Mistry,J., Finn,R.D. and Saier,M.H. Jr. (2015) The complexity, challenges and benefits of comparing two transporter classification systems in TCDB and Pfam. *Brief Bioinform.* doi:10.1093/bib/bbu053.
41. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
42. Chen,J.S., Reddy,V., Chen,J.H., Shlykov,M.A., Zheng,W.H., Cho,J., Yen,M.R. and Saier,M.H. Jr. (2011) Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments. *J. Mol. Microbiol. Biotechnol.*, **21**, 83–96.
43. Reddy,V.S. and Saier,M.H. Jr. (2012) BioV Suite–a collection of programs for the study of transport protein evolution. *FEBS J.*, **279**, 2036–2046.
44. Vastermark,A., Lunt,B. and Saier,M. (2014) Major facilitator superfamily porters, LacY, FucP and XylE of Escherichia coli appear to have evolved positionally dissimilar catalytic residues without rearrangement of 3-TMS repeat units. *J. Mol. Microbiol. Biotechnol.*, **24**, 82–90.
45. Reddy,V.S., Shlykov,M.A., Castillo,R., Sun,E.I. and Saier,M.H. Jr. (2012) The major facilitator superfamily (MFS) revisited. *FEBS J.*, **279**, 2022–2035.
46. Khwaja,M., Ma,Q. and Saier,M.H. Jr. (2005) Topological analysis of integral membrane constituents of prokaryotic ABC efflux systems. *Res. Microbiol.*, **156**, 270–277.
47. Vastermark,A., Wollwage,S., Houle,M.E., Rio,R. and Saier,M.H. Jr. (2014) Expansion of the APC superfamily of secondary carriers. *Proteins*, **82**, 2797–2811.
48. Prakash,S., Cooper,G., Singhi,S. and Saier,M.H. Jr. (2003) The ion transporter superfamily. *Biochim. Biophys. Acta*, **1618**, 79–92.
49. Mansour,N.M., Sawhney,M., Tamang,D.G., Vogl,C. and Saier,M.H. Jr. (2007) The bile/arsenite/riboflavin transporter (BART) superfamily. *FEBS J.*, **274**, 612–629.
50. Jack,D.L., Yang,N.M. and Saier,M.H. Jr. (2001) The drug/metabolite transporter superfamily. *Eur. J. Biochem.*, **268**, 3620–3639.
51. Reddy,B.L. and Saier,M.H. Jr. (2013) Topological and phylogenetic analyses of bacterial holin families and superfamilies. *Biochim. Biophys. Acta*, **1828**, 2654–2671.
52. Park,J.H. and Saier,M.H. Jr. (1996) Phylogenetic characterization of the MIP family of transmembrane channel proteins. *J. Membr. Biol.*, **153**, 171–180.
53. Pao,S.S., Paulsen,I.T. and Saier,M.H. Jr. (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.*, **62**, 1–34.
54. Kuan,J. and Saier,M.H. Jr. (1993) The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit. Rev. Biochem. Mol. Biol.*, **28**, 209–233.
55. Lee,A., Vastermark,A. and Saier,M.H. Jr. (2014) Establishing homology between mitochondrial calcium uniporters, prokaryotic magnesium channels and chlamydial IncA proteins. *Microbiology*, **160**, 1679–1689.

56. Mok,T., Chen,J.S., Shlykov,M.A. and Saier,M.H. (2012) Bioinformatic analyses of bacterial mercury ion (Hg2+) transporters. *Water Air Soil Pollut.*, **223**, 4443–4457.
57. Hvorup,R.N., Winnen,B., Chang,A.B., Jiang,Y., Zhou,X.F. and Saier,M.H. Jr. (2003) The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. *Eur. J. Biochem.*, **270**, 799–813.
58. Gomolplitinant,K.M. and Saier,M.H. Jr. (2011) Evolution of the oligopeptide transporter family. *J. Membr. Biol.*, **240**, 89–110.
59. Hvorup,R., Chang,A.B. and Saier,M.H. Jr. (2003) Bioinformatic analyses of the bacterial L-ascorbate phosphotransferase system permease family. *J. Mol. Microbiol. Biotechnol.*, **6**, 191–205.
60. Cao,Y., Jin,X., Levin,E.J., Huang,H., Zong,Y., Quick,M., Weng,J., Pan,Y., Love,J., Punta,M. *et al.* (2011) Crystal structure of a phosphorylation-coupled saccharide transporter. *Nature*, **473**, 50–54.
61. Tseng,T.T., Gratwick,K.S., Kollman,J., Park,D., Nies,D.H., Goffeau,A. and Saier,M.H. Jr. (1999) The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J. Mol. Microbiol. Biotechnol.*, **1**, 107–125.
62. Hua,V.B., Chang,A.B., Tchieu,J.H., Kumar,N.M., Nielsen,P.A. and Saier,M.H. Jr. (2003) Sequence and phylogenetic analyses of 4 TMS junctional proteins of animals: connexins, innexins, claudins and occludins. *J. Membr. Biol.*, **194**, 59–76.
63. Minocha,R., Studley,K. and Saier,M.H. Jr. (2003) The urea transporter (UT) family: bioinformatic analyses leading to structural, functional, and evolutionary predictions. *Receptors Channels*, **9**, 345–352.
64. Jones,D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, **23**, 538–544.
65. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
66. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version, 7, improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
67. Biegert,A. and Soding,J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinformatics*, **24**, 807–814.
68. Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.
69. Tsu,B.V. and Saier,M.H. Jr. (2015) The LysE Superfamily of Transport Proteins Involved in Cell Physiology and Pathogenesis. *PLoS One.*, **10**, e0137184.