The Amino Acid-Polyamine-Organocation Superfamily

Foon H. Wong Jonathan S. Chen Vamsee Reddy Jonathan L. Day
Maksim A. Shlykov Steven T. Wakabayashi Milton H. Saier, Jr.

Division of Biological Sciences, University of California at San Diego, La Jolla, Calif., USA

Key Words

Transport proteins • Membranes • Evolution •
Amino acid-polyamine-organocation superfamily •
Phylogeny • SuperfamilyTree programs

Abstract

The amino acid-polyamine-organocation (APC) superfamily has been shown to include five recognized families, four of which are specific for amino acids and their derivatives. Recent high-resolution X-ray crystallographic data have shown that four additional transporter families (BCCT, TC No. 2.A.15; SSS, 2.A.21; NSS, 2.A.22; and NCS1, 2.A.39), transporting a wide range of solutes, exhibit sufficiently similar folds to suggest a common evolutionary origin. We have used established statistical methods, based on sequence similarity, to show that these families are, in fact, members of the APC superfamily. We also identify two additional families (NCS2, 2.A.40; SulP, 2.A.53) as being members of this superfamily. Repeat sequences, each having five transmembrane α -helical segments and arising via ancient intragenic duplications, are demonstrated for all of these families, further strengthening the conclusion of homology. The APC superfamily appears to be the second largest superfamily of secondary carriers, the largest being the major facilitator superfamily (MFS). Although the topology of the members of the APC

superfamily differs from that of the MFS, both families appear to have arisen from a common ancestral 2 TMS hairpin structure that underwent intragenic triplication followed by loss of a TMS in the APC family, to give the repeat units that are characteristic of these two superfamilies.

Copyright © 2012 S. Karger AG, Basel

Introduction

Over the years, our research group has identified and expanded superfamilies of transmembrane transporter proteins as recorded in the Transporter Classification Database [TCDB; www.tcdb.org; Saier et al., 2006; Saier et al., 2009]. This database assigns proteins to superfamilies and their respective families or subfamilies while integrating novel bioinformatics software and relevant resources [Reddy and Saier, 2012]. The importance of identification and classification of protein families and superfamilies is emphasized by the fact that structural, functional, and mechanistic data for transporters can be extrapolated from one protein to another if and only if they have been shown to be related by a common descendent [Lam et al., 2011; Saier, 1996, 1999, 2000].

A major focus of our bioinformatics laboratory has been to update and develop novel computer programs to

assist in the identification and expansion of protein superfamilies [Chang et al., 2004; Reddy and Saier, 2012]. These programs have been instrumental in establishing homology amongst proteins sharing common descent. Each year, we become more proficient in reliably detecting and quantitatively evaluating increasingly distant phylogenetic relationships. In parallel, we are perfecting the art of identifying and evaluating the evolutionary pathways that gave rise to these proteins.

The amino acid-polyamine-organocation (APC) superfamily is a large superfamily which on the basis of recent 3D structural analyses [Chan et al., 2010; Fang et al., 2009; Gao et al., 2010; Jeckelmann et al., 2011; Kowalczyk et al., 2011] appears to be much larger than previously thought. Previously we had established superfamily status for the APC superfamily [Chang et al., 2004; Jack et al., 2000; Young et al., 1999], but the five families then recognized for this superfamily included only one of the four families that have been suggested to be related, based on the recent 3D structural data.

The five original families are: (1) the APC (TC No. 2.A.3) family, (2) the amino acid/auxin permease (AAP; TC No. 2.A.18) family, (3) the alanine or glycine:cation symporter (AGCS; TC No. 2.A.25) family, (4) the cation-chloride cotransporter (CCC; TC No. 2.A.30) family, and (5) the hydroxy/aromatic amino acid permease (HAAAP; TC No. 2.A.42) family (table 1). We now present statistical results, based on primary sequence data alone, demonstrating homology of pre-existing APC superfamily members with the following six new families: (1) the betaine/carnitine/choline transporter (BCCT; TC No. 2.A.15) family, (2) the solute:sodium symporter (SSS; TC No. 2.A.21) family, (3) the neurotransmitter:sodium

symporter (NSS; TC No. 2.A.22) family, (4) the nucleobase:cation symporter-1 (NCS1; TC No. 2.A.39) family, (5) the nucleobase:cation symporter-2 (NCS2; TC No. 2.A.40) family, and (6) the sulfate permease (SulP; TC No. 2.A.53) family (table 1). All of these families include members that exhibit the 5 TMS duplicated repeat sequences, although additional transmembrane α -helical segments (TMSs) may exist at their N- and C-termini. We provide evidence that the 5 TMS repeat unit arose by intragenic multiplication where the primordial repeat unit included only 2 TMSs. The pathway for appearance of all APC superfamily members therefore appears to be:

Methods

Sequence Alignment Analyses

FASTA sequences representing members of the APC family (TC No. 2.A.3) and each of the newly added families within the APC superfamily were gathered from the TCDB [www.tcdb.org; Saier et al., 2009]. The Basic Local Alignment Search Tool (BLAST) [Altschul et al., 1990; Altschul et al., 1997], using PSI-BLAST with two iterations, was performed on proteins representing each family to generate seven lists of homologous proteins. Collectively, more than 5,600 proteins were gathered for analysis from the National Center for Biotechnology Information (NCBI) NR Protein database. Over 5,000 proteins were kept for further examination after removing redundancies and proteins that were at least 95% identical to a kept protein using a modified CD-Hit program [Li and Godzik, 2006; Li et al., 2001, 2002; Saier et al., 2009; Yen et al., 2009].

Lists of proteins for each of the seven APC families were compared with one another to identify significant sequence similarities using SSearch [Yen et al., 2009]. SSearch determines protein significance using BLAST. The program identified numerous pairs of proteins that suggested homology, seven of which are displayed in figure 1. The Global Alignment Program (GAP) [Devereux et al., 1984] and the similar GSAT program [Reddy and Saier, 2012] were then used to analyze the sequences, supporting the relationships suggested by SSearch [Reddy and Saier, 2012; Yen et al., 2009]. Both programs randomly shuffle and compare amino acid sequences, correcting for unusual protein composition such as those within integral membrane proteins. The higher the GAP or GSAT scores expressed in standard deviations (SD), the more significant are their sequence similarities. We have

Table 1. Families within the APC superfamily

2.A.3	APC family
2.A.15	BCCT family
2.A.18	AAP family
2.A.21	SSS family
2.A.22	NSS family
2.A.25	AGCS family
2.A.30	CCC family
2.A.39	NCS1 family
2.A.40	NCS2 family
2.A.42	HAAAP family
2.A.53	SulP family

Description, protein members and references of these families can be found in TCDB (www.TCDB.org).

somewhat arbitrarily set 11 SD for a stretch of at least 60 amino acyl residues in comparable positions of the proteins as the cutoff for establishing homology.

Motif Analysis

The MEME program [Bailey and Elkan, 1995] was used to identify common motifs. Default settings were used with the exception that the optimal maximum width was set at 20 residues. Ancestral sequences were generated using the ANCESCON program [Cai et al., 2004] with the 'ancestral sequence reconstruction' option set to 'reconstruct ancestral sequence only for the root (midpoint of the tree) (marginal reconstruction only)' and the 'parameters in ancestral sequence reconstruction' option set to 'use maximum likelihood rate factor (more time and higher precision than AB)'.

Phylogenetic, Hydropathy and Sequence Analyses

Using the TCDB [Saier et al., 2006; Saier et al., 2009], a temporary database file was generated containing proteins that define all members within the APC superfamily. This database file was used to define the criteria for superfamily definition and to determine how that superfamily should be broken down into families or subfamilies. The division of proteins into superfamilies, families, and subfamilies was in general conducted according to assignments in TCDB. In a few cases, SuperfamilyTree 1 (SFT1) detected some errors in TCDB and, following further examination, these errors were corrected.

Multiple Alignment Method

For neighbor-joining methods [Gascuel and Steel, 2006; Saitou and Nei, 1987], the resulting database was used to create a multiple alignment using the ClustalX program. The multiple alignment was used to generate a neighbor-joining ClustalX phylogenetic tree [Thompson et al., 1997]. The resulting file was then viewed as a radial phylogenetic tree using the TreeView program [Zhai et al., 2002].

SFT Methods: SFT1 and SFT2

The temporary database generated from TCDB was used for rapid sequence similarity searches. Using this database, we used PSI-BLAST [Altschul et al., 1990; Altschul et al., 1997] to search the NCBI protein database and matched up potential members for each family. BLAST hits were then classified and sorted into respective families and subfamilies according to assignments in TCDB.

The resulting database files were then used to generate a phylogenetic tree using the SFT1 program by generating tens of thousands of comparative BLAST bit score matrices of the superfamily through 100 repeat shuffles [Chen et al., 2011; Yen et al., 2009; Yen et al., 2010]. The programs, Fitch and Consense [Fitch and Margoliash, 1967; <http://evolution.genetics.washington.edu/phylib/doc/protpars.html>], utilized the matrix information to generate 100 phylogenetic trees and consolidate those trees into a single consensus tree. The resulting SFT1 tree shows the relative phylogenetic positions of all members of the families within the superfamily included in the study. The information from the SFT1 program was then used to combine sequences of selected members into subfamily groupings and each of the constituent families into a single file (SFT2). The same programs and methods are applied to the newly formed database files to generate

Fig. 1. APC superfamily homology established through the use of the GAP and GSAT programs and based on the superfamily principle. Previously established APC superfamily proteins and their homologues [Jack et al., 2000] were used to establish homology between the six other families. The TC No. for each family is listed under the family abbreviation in parentheses. GAP and GSAT scores are expressed in terms of SD.

SFT2 trees. These trees can be viewed as radial phylogenetic trees using the TreeView program. However, branch lengths in trees derived using the SFT programs are not proportional to phylogenetic distance. In previous communications, the reliability of these methods has been evaluated [Chen et al., 2011; Yen et al., 2009; Yen et al., 2010].

Results

Establishing Common Descent

The APC family (TC No. 2.A.3) and the six new families were shown to be related using the superfamily principle which states that if protein A is related to (shares a common origin with) protein B, and protein B is related to protein C, then protein A is related to protein C. The GAP or GSAT comparison scores reported are considered sufficient to establish homology if they give 11 SD or greater [Reddy and Saier, 2012]. This value corresponds to a probability that the degree of sequence similarity occurred by chance of 10^{-29} [Dayhoff et al., 1983].

GAP scores, expressed in SD, for various protein comparisons are recorded in figure 1, and the alignments

(For legend see next page.)

upon which these values were based are shown in online supplementary figures S1–S7 (for all online suppl. material, see www.karger.com/doi/10.1159/000338542). For example, residues 2–327 in Cha1 (TMSs 1–8) of the NSS family aligned with 20–325 in Tpa1 (TMSs 1–8) of the APC family, showing 36.2% similarity and 27.9% identity. This comparison yielded a GAP score of 16.3 SD, which is substantially in excess of what is required to establish homology (online suppl. fig. S1). Representative examples of comparisons that establish homology between all of the families included in this study are presented in online supplementary figures S2–S7, and the resulting comparison scores are summarized in figure 1.

Phylogenetic Analysis

Using all TC entries for families within the newly defined APC superfamily as of February 2011, phylogenetic trees were constructed using the ClustalX and TreeView programs (fig. 2a), as well as the SFT1 and TreeView programs (fig. 2b).

The ClustalX program did not correctly recognize the relationships between members of single families. For example, the APC family (TC No. 2.A.3) is found in seven distinct places within the tree, while the AAP family (TC No. 2.A.18) is found on eight distinct branches and the SulP family (TC No. 2.A.53) is found on seven branches. The NCS1 family (TC No. 2.A.39) is found on five branches, the NCS2 family (TC No. 2.A.40) is found on two branches, the

Fig. 2. Phylogenetic (Fitch) trees for the APC superfamily using the proteins belonging to this superfamily in TCDB as of February 2011. Two different methods of tree construction were used: the ClustalX-based neighbor-joining program showing all APC superfamily members (**a**) and the BLAST-derived SFT1-program showing all APC superfamily members (**b**). In both trees, numbers indicate the protein TC No. (last two digits of the complete

TC No.). In **b**, small numbers adjacent to the branches present the 'bootstrap' values, indicating the reliability of the branching order. See TCDB for protein identification. Comparison of **a** with **b** reveals the superiority of the SFT programs over multiple alignment-based phylogenetic trees when sequences of the constituent proteins are markedly divergent [Chen et al., 2011; Yen et al., 2009; Yen et al., 2010].

SSS family (TC No. 2.A.21) is found on three branches, and the BCCT family (TC No. 2.A.15) is found on two branches. On the other hand, the CCC family (TC No. 2.A.30), the AGCS family (TC No. 2.A.25), the NSS family (TC No. 2.A.22), and the HAAAP family (TC No. 2.A.42) are coherent with all family members clustering together. These results reflect the limitations of the ClustalX program to correctly align sequences when they are very divergent [Chen et al., 2010; Yen et al., 2009; Yen et al., 2010].

By contrast with the results obtained when the ClustalX tree was used, the SFT1 tree shows a high degree of consistency. For example, all members of the APC family cluster together on the upper right hand side of this tree. Members of the CCC, AGCS, and HAAAP families also segregated according to family, and these occur in the upper central part of the tree, sandwiched between the APC and AAAP families. It is interesting to note that all of these families, which had previously been known

to be members of the APC family, comprise the top half of this tree, while all of the families recently added, based on 3D structures and sequence comparison data presented here, occur within the bottom half of the tree. This is of interest since all of the families in the upper half of the tree, except the CCC family, transport amino acids and their derivatives, while the families in the lower half of the tree transport a wide variety of substances including sugars, amino acids, nucleobases, nucleosides, inorganic sulfate, osmoprotectants, and neurotransmitters. The results show that the previously recognized families are more similar to each other than they are to the newly added families. This can be explained since the methods used previously were not as refined and sensitive as the ones we have recently developed [Reddy and Saier, 2012].

All members of each of the eleven families cluster together on the SFT tree as expected with one exception. This one exception is a subfamily within the APC family, the spore germination protein (SGP) subfamily (TC No. 2.A.3.9). Proteins within this subfamily are known to be receptors for amino acids, which trigger germination of *Bacillus* spores [Cooper and Moir, 2011]. It was recognized previously [Jack et al., 2000] that members of this subfamily are much more distant in sequence from other members of the APC family than these other members are from each other. The SGP subfamily members are also shorter with only 10 TMSs. The loss (or lack of addition) of TMSs over evolutionary time correlates with the loss (or absence) of transport activity and conversion of these proteins into simple ligand-binding receptors [Jack et al., 2000; Saier, 2003].

Two additional trees were generated using the SFT2 program, one revealing the relationships of all the subfamilies within the various families of the APC superfamily (fig. 3a) and another revealing only the family relationships with the sole exception of the APC family, where subfamilies are indicated (fig. 3b). With respect to the subfamilies of the APC family, the branching order is essentially the same for the two trees.

Within figure 3a and b, several subfamily and family relationships are worthy of note. For example, in both trees, the relationships are the same without exception. However, when compared to the tree for the APC family reported by Jack et al. [2000], very substantial differences are observed although some similarities such as the tight clustering of the AAT and YAT families with each other were revealed by both methods. In figure 3a and b, we see that the AAAP and the HAAAP families cluster together, the NSS and BCCT families cluster together, and surpris-

ingly the NCS2 and SulP families cluster together with the SSS and NCS1 families, sharing the same common branch but more deeply rooted. It would appear that the SGP receptors are most likely to have arisen from the AGCS transporters in view of (1) their similar substrate specificities and (2) their clustering patterns on the phylogenetic trees. Based on the results presented in figure 3, the SGP family could be considered to comprise its own family within the APC superfamily (see Discussion).

Discussion

It is frequently claimed and generally accepted that sequence divergence occurs more rapidly than 3D fold changes, and that consequently, primary sequence analyses are less sensitive than X-ray crystallographic data for the purpose of defining common protein origin for highly divergent members of a protein superfamily [Abramson and Wright, 2009; Sael et al., 2012; Shenoy and Jayaram, 2010]. While we agree that both methods can provide strong evidence for homology, we are not convinced that this generally accepted notion is correct. The authors of the 3D structures for representative members of the APC superfamily as defined here all claimed that there was insufficient sequence similarity between these proteins to establish homology (see Introduction). Nevertheless, using our standard procedures, we were able to find very significant sequence similarities when members of these different families were compared. In fact, comparison scores were generally so large that there was little doubt about the conclusion of homology.

Our criteria for homology require not only that we identify regions of strong sequence similarity, but also that these regions be in the transmembrane parts of the proteins and that the regions compared are in comparable parts of these proteins [Matias et al., 2010; Saier et al., 2009; Wang et al., 2009; Yen et al., 2009]. Thus, when examining APC superfamily members as defined in table 1, we provided strong evidence that the same TMSs within the two 5 TMS repeat units were being compared. Moreover, we examined all top scores to ensure that they gave agreement. In this way, homology could be established with a much higher confidence level than if these procedures were not followed. It should be noted that failure to find a score that is sufficient to establish homology, or positional disagreement when comparing multiple alignments giving lower scores, never proves a *lack* of homology. Such results can merely reflect extensive sequence

Fig. 3. Phylogenetic (Fitch) trees for the APC superfamily using the proteins belonging to this superfamily in TCDB as of February 2011. Tree construction used the BLAST-derived SFT2 program showing APC superfamily members selectively grouped into TC subfamilies (**a**) and APC superfamily members selectively grouped into TC families (**b**). In **a**, numbers indicate the subfamily TC No. (fourth digits of the complete TC No.). In **b**, fam-

ily abbreviations are presented with TC family numbers in parentheses. Only for the APC family (2.A.3; top) is subfamily TC No. (the fourth digit in the full 5-digit protein TC No.) presented. Small numbers adjacent to the branches present the 'bootstrap' values, indicating the reliability of the branching order. Branch lengths are not proportional to phylogenetic distance. See TCDB for protein identification.

divergence and/or inaccuracies of the programs used to generate the alignments.

The APC superfamily phylogenetic tree, generated with the ClustalX program and based on a multiple alignment, proved highly inaccurate, as individual proteins belonging to single families within the superfamily did not cluster together (fig. 2a). By contrast, our newly developed SFT programs [Chen et al., 2011; Yen et al., 2009; Yen et al., 2010] proved reliable, with members of each family clustering coherently together. These SFT programs base phylogenetic position on tens of thousands of BLAST bit scores, giving much greater degrees of reliability. Thus, although single BLAST scores are known to be inaccurate, by averaging the results of thousands of such scores for large numbers of comparisons, a much higher

degree of accuracy can be obtained than is possible using methods based on multiple alignments. It should be noted, however, that both methods are reliable and in agreement when sequence similarity is sufficient to allow construction of accurate multiple alignments [Chen et al., 2011; Yen et al., 2009; Yen et al., 2010].

Many proteins have the capability of existing in at least two stable states, each one having easily distinguishable 3D structures [Hotze and Tweten, 2012; Littler et al., 2010; Popoff, 2011]. These proteins include secreted pore-forming toxins that act on a cell other than the producer cell (TC subclass 1.C) as well as several α -helical channel-forming proteins that function in the cell that produces them (TC subclasses 1.A and 1.E) [Littler et al., 2010]. These proteins collectively belong to over 200 established

families (see TCDB; www.tcdb.org). The toxins of subclass 1.C are secreted in soluble states and subsequently insert into the membranes of target cells. Well-studied examples include the aerolysin (TC No. 1.C.4) [Popoff, 2011], RTX (1.C.11) [Linhartova et al., 2010], MACPF (1.C.39) [Kondos et al., 2010], and CDC (1.C.12) [Gilbert, 2010] families. Several channel proteins of subclass 1.A have catalytic activities and physiological functions in their soluble cytoplasmic states, distinct from those of their membrane-embedded channel-forming states. Examples include members of the CLIC (1.A.12) [Jalilian et al., 2008], annexin (1.A.31) [McNeil et al., 2006], BASP1 (1.A.71) [Ostroumova et al., 2011], and MICU (1.A.76) [Perocchi et al., 2010] families. These proteins have recently been analyzed from topological standpoints [Cho et al., 2012; manuscript in preparation].

The APC family (TC No. 2.A.3) can be considered the core member of the APC superfamily. It was the first to be recognized as a major family of carrier proteins as indicated by its low TC number. It also appears to have the largest membership of the eleven families that are now recognized as constituents of the APC superfamily (table 1). These facts render appropriate the designation of this greatly expanded superfamily as the APC superfamily.

Phylogenetic analyses of the entire APC superfamily using the SFT programs (fig. 2, 3) [Chen et al., 2011; Yen et al., 2009] indicated that the four families recognized by Chang et al. [2004] as well as the AGCS family (TC No. 2.A.25) are more closely related to each other than they are to other family members. The spore-germination proteins (SGP subfamily; 2.A.3.9) cluster far from other APC family members, closer to the AGCS and AAP families that recognize similar semipolar amino acids. Although the SGP subfamily was originally included in the APC family, it was known to be the most distant member [Jack et al., 2000]. The phylogenetic results depicted in figures 2 and 3 suggest that this subfamily should comprise a family of its own.

Analysis of the SFT trees (fig. 2b and 3a, b) revealed that the NSS and BCCT families cluster together as do the NCS2 and Sulp families in spite of the fact that their substrates differ in character. In agreement with this observation, the NCS1 and NCS2 families, both specific for nucleobases, do not cluster tightly together. It therefore appears that in contrast to the five amino acid transporting families of the APC superfamily, phylogeny does not always reflect the substrate-binding specificities of the carriers.

Some superfamilies, such as the MFS (2.A.1) and DMT (2.A.7) superfamilies, include members that can catalyze

both uptake and efflux of solutes via cation symport and antiport, respectively. However, others consist of members, all of which catalyze only uptake or efflux. The RND superfamily (2.A.6) is an example of one that is exclusively concerned with export, while the APC superfamily is one concerned exclusively with uptake. It will be interesting to understand the evolutionary events required for the interconversion of these two vectorial reactions by members of the various superfamilies of related proteins.

Knowledge of the common origin of members of the APC superfamily implies common structure and mechanism of action as noted above, but to what extent are structure/function extrapolations applicable to this and other families? Is the extent of this structure-function dependency family-specific? Why are some large families highly restrictive in substrates-specificities (e.g. the CaCA family; TC No. 2.A.29), while others are exceptionally broad (e.g. the SSS family; TC No. 2.A.21)? Why can members of some families use multiple mechanisms of energy coupling, while others apparently use only one? Why do some families include members capable of either uptake or export while others include members capable of only one of these two vectorial reactions? We anticipate that structure/function/mechanism relationships will in general prove to be family-specific, but definitive answers to these questions will require extensive research using multiple experimental approaches.

Acknowledgements

This work was supported by NIH grant GM077402. We thank Carl Welliver for assistance with the preparation of the manuscript.

References

- Abramson J, Wright EM: Structure and function of Na(+)-symporters with inverted repeats. *Curr Opin Struct Biol* 2009;19:425–432.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
- Bailey TL, Elkan C: The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* 1995;3:21–29.
- Cai W, Pei J, Grishin NV: Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol* 2004;4:33.

- Chan H, Babayan V, Blyumin E, Gandhi C, Hak K, Harake D, Kumar K, Lee P, Li TT, Liu HY, Lo TC, Meyer CJ, Stanford S, Zamora KS, Saier MH Jr: The p-type ATPase superfamily. *J Mol Microbiol Biotechnol* 2010;19:5–104.
- Chang AB, Lin R, Keith Studley W, Tran CV, Saier MH Jr: Phylogeny as a guide to structure and function of membrane transport proteins. *Mol Membr Biol* 2004;21:171–181.
- Chen JS, Reddy V, Chen JH, Shlykov MA, Zheng WH, Cho J, Yen MR, Saier MH Jr: Phylogenetic characterization of transport protein superfamilies: superiority of Superfamily-Tree programs over those based on multiple alignments. *J Mol Microbiol Biotechnol* 2011;21:83–96.
- Chen LQ, Hou BH, Lalonde S, Takanaga H, Hartung ML, Qu XQ, Guo WJ, Kim JG, Underwood W, Chaudhuri B, Chermak D, Antony G, White FF, Somerville SC, Mudgett MB, Frommer WB: Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* 2010;468:527–532.
- Cooper GR, Moir A: Amino acid residues in the GerAB protein important in the function and assembly of the alanine spore germination receptor of *Bacillus subtilis* 168. *J Bacteriol* 2011;193:2261–2267.
- Dayhoff MO, Barker WC, Hunt LT: Establishing homologies in protein sequences. *Methods Enzymol* 1983;91:524–545.
- Devereux J, Haeblerli P, Smithies O: A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* 1984;12:387–395.
- Fang Y, Jayaram H, Shane T, Kolmakova-Partensky L, Wu F, Williams C, Xiong Y, Miller C: Structure of a prokaryotic virtual proton pump at 3.2 Å resolution. *Nature* 2009;460:1040–1043.
- Fitch WM, Margoliash E: Construction of phylogenetic trees. *Science* 1967;155:279–284.
- Gao X, Zhou L, Jiao X, Lu F, Yan C, Zeng X, Wang J, Shi Y: Mechanism of substrate recognition and transport by an amino acid antiporter. *Nature* 2010;463:828–832.
- Gascuel O, Steel M: Neighbor-joining revealed. *Mol Biol Evol* 2006;23:1997–2000.
- Gilbert RJ: Cholesterol-dependent cytolysins. *Adv Exp Med Biol* 2010;677:56–66.
- Hotze EM, Tweten RK: Membrane assembly of the cholesterol-dependent cytolysin pore complex. *Biochim Biophys Acta* 2012;1818:1028–1038.
- Jack DL, Paulsen IT, Saier MH: The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiol Rev* 2000;146:1797–1814.
- Jalilian C, Gallant EM, Board PG, Dulhunty AF: Redox potential and the response of cardiac ryanodine receptors to CLIC-2, a member of the glutathione S-transferase structural family. *Antioxid Redox Signal* 2008;10:1675–1686.
- Jeckelmann JM, Palacin M, Fotiadis D: A tool for the qualitative comparison of membrane-embedded and detergent-solubilized membrane protein structures in projection. *J Struct Biol* 2011;173:375–381.
- Kondos SC, Hatfaludi T, Voskoboinik I, Trapani JA, Law RH, Whisstock JC, Dunstone MA: The structure and function of mammalian membrane-attack complex/perforin-like proteins. *Tissue Antigens* 2010;76:341–351.
- Kowalczyk L, Ratera M, Paladino A, Bartoccioni P, Errasti-Murugarren E, Valencia E, Portella G, Bial S, Zorzano A, Fita I, Orozco M, Carpena X, Vazquez-Ibar JL, Palacin M: Molecular basis of substrate-induced permeation by an amino acid antiporter. *Proc Natl Acad Sci USA* 2011;108:3935–3940.
- Lam VH, Lee JH, Silverio A, Chan H, Gomolplintant KM, Povolotsky TL, Orlova E, Sun EI, Welliver CH, Saier MH Jr: Pathways of transport protein evolution: recent advances. *Biol Chem* 2011;392:5–12.
- Li W, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;22:1658–1659.
- Li W, Jaroszewski L, Godzik A: Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;17:282–283.
- Li W, Jaroszewski L, Godzik A: Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 2002;18:77–82.
- Linhartova I, Bumba L, Masin J, Basler M, Osicka R, Kamanova J, Prochazkova K, Adkins I, Hejnova-Holubova J, Sadilkova L, Morova J, Sebo P: RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiol Rev* 2010;34:1076–1112.
- Littler DR, Harrop SJ, Goodchild SC, Phang JM, Mynott AV, Jiang L, Valenzuela SM, Mazanti M, Brown LJ, Breit SN, Curmi PM: The enigma of the CLIC proteins: ion channels, redox proteins, enzymes, scaffolding proteins? *FEBS Lett* 2010;584:2093–2101.
- Matias MG, Gomolplintant KM, Tamang DG, Saier MH Jr: Animal Ca^{2+} release-activated Ca^{2+} (CRAC) channels appear to be homologous to and derived from the ubiquitous cation diffusion facilitators. *BMC Res Notes* 2010;3:158.
- McNeil AK, Rescher U, Gerke V, McNeil PL: Requirement for annexin A1 in plasma membrane repair. *J Biol Chem* 2006;281:35202–35207.
- Ostroumova OS, Schagina LV, Mosevitsky MI, Zakharov VV: Ion channel activity of brain abundant protein BASP1 in planar lipid bilayers. *FEBS J* 2011;278:461–469.
- Percocchi F, Gohil VM, Girgis HS, Bao XR, McCombs JE, Palmer AE, Mootha VK: MICU1 encodes a mitochondrial EF hand protein required for Ca^{2+} uptake. *Nature* 2010;467:291–296.
- Popoff MR: Epsilon toxin: a fascinating pore-forming toxin. *FEBS J* 2011;278:4602–4615.
- Reddy V, Saier MH Jr: BioV Suite – a collection of programs for the study of transport protein evolution. *FEBS J* 2012, Mar 30. DOI: [10.1111/j.1742-4658.2012.08590.x](https://doi.org/10.1111/j.1742-4658.2012.08590.x). (E-pub ahead of print).
- Sael L, Chitale M, Kihara D: Structure- and sequence-based function prediction for non-homologous proteins. *J Struct Funct Genomics* 2012, E-pub ahead of print.
- Saier MH: Genome archeology leading to the characterization and classification of transport proteins. *Curr Opin Microbiol* 1999;2:555–561.
- Saier MH Jr: Phylogenetic approaches to the identification and characterization of protein families and superfamilies. *Microb Comp Genomics* 1996;1:129–150.
- Saier MH Jr: A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* 2000;64:354–411.
- Saier MH Jr: Tracing pathways of transport protein evolution. *Mol Microbiol* 2003;48:1145–1156.
- Saier MH Jr, Tran CV, Barabote RD: TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* 2006;34:D181–D186.
- Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C: The Transporter Classification Database: recent advances. *Nucleic Acids Res* 2009;37:D274–D278.
- Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4:406–425.
- Shenoy SR, Jayaram B: Proteins: sequence to structure and function – current status. *Curr Protein Pept Sci* 2010;11:498–514.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–4882.
- Wang B, Dukarevich M, Sun EI, Yen MR, Saier MH Jr: Membrane porters of ATP-binding cassette transport systems are polyphyletic. *J Membr Biol* 2009;231:1–10.
- Yen MR, Chen JS, Marquez JL, Sun EI, Saier MH: Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. *Methods Mol Biol* 2010;637:47–64.
- Yen MR, Choi J, Saier MH Jr: Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. *J Mol Microbiol Biotechnol* 2009;17:163–176.
- Young GB, Jack DL, Smith DW, Saier MH Jr: The amino acid/auxin:proton symport permease family. *Biochim Biophys Acta* 1999;1415:306–322.
- Zhai Y, Tchieu J, Saier MH Jr: A Web-based Tree View (TV) program for the visualization of phylogenetic trees. *J Mol Microbiol Biotechnol* 2002;4:69–70.