Title

Expansion of the APC superfamily of secondary carriers

Journal

Proteins Structure Function and Bioinformatics, 82(10)

ISSN

0887-3585

Authors

Vastermark, Ake
Wollwage, Simon
Houle, Michael E
et al.

Publication Date

2014-10-01

DOI

10.1002/prot.24643

Peer reviewed

Published in final edited form as:
Proteins. 2014 October ; 82(10): 2797–2811. doi:10.1002/prot.24643.

Expansion of the APC superfamily of secondary carriers

Ake Vastermark[1], **Simon Wollwage**[2], **Michael E. Houle**[2], **Rita Rio**[3], and **Milton H. Saier Jr.**[1,*]

[1]Department of Molecular Biology, University of California at San Diego, La Jolla, CA 92093-0116

[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

[3]Department of Biology, West Virginia University, Morgantown, WV 26505

Abstract

The Amino acid-Polyamine-organoCation (APC) superfamily is the second largest superfamily of secondary carriers currently known. In the current study, we establish homology between previously recognized APC superfamily members and proteins of seven new families. These families include the PAAP (Putative Amino Acid Permease), LIVCS (Branched Chain Amino Acid:Cation Symporter), NRAMP (Natural Resistance-Associated Macrophage Protein), CstA (Carbon starvation A protein), KUP (K^+ Uptake Permease), BenE (Benzoate:H^+ Symporter) and AE (Anion Exchanger). The topology of the well-characterized human Anion Exchanger 1 (AE1) conforms to a UraA-like topology of 14 TMSs (12 α-helical TMSs and 2 mixed coil/helical TMSs). All functionally characterized members of the APC superfamily use cation symport for substrate accumulation except for members of the AE family which frequently use anion:anion exchange. We show how the different topologies fit into the framework of the common LeuT-like fold, defined earlier (Proteins. 2014 Feb;82(2):336–46), and determine that some of the new members contain previously undocumented topological variations. All new entries contain the two 5 or 7 TMS APC superfamily repeat units, sometimes with extra TMSs at the ends, the variations being greatest within the CstA family. New, functionally characterized members transport amino acids, peptides, and inorganic anions or cations. Except for anions, these are typical substrates of established APC superfamily members. Active site TMSs are rich in glycyl residues in variable but conserved constellations. This work expands the APC superfamily and our understanding of its topological variations.

Keywords

SuperFamily Tree; Transporter Classification Database; Amino acid-Polyamine-organoCation (APC) superfamily; Anion Exchanger 1; Carbon starvation A protein

INTRODUCTION

The APC Superfamily

Secondary transporters utilize transmembrane molecular gradients created by the transport activities of primary active pumps to accumulate or expel substrates against large

*Corresponding author: Telephone: (858) 534-4084, Fax: (858) 534-7108, msaier@ucsd.edu.

concentration gradients [1]. Each secondary carrier normally consists of a single polypeptide chain and displays a twofold pseudosymmetrical topology, being made up of two homologous units of 4–7 transmembrane helical spanners (TMSs) [2] although in some families these two repeat units are represented by two similar (heterodimeric) or identical (homodimeric) subunits.

The Amino acid-Polyamine-organoCation (APC) superfamily constitutes the second largest superfamily of secondary carriers [3], being smaller only than the Major Facilitator Superfamily (MFS) [4]. Regrettably, re-examination of previously published alignments used as incorrect evidence for proposed homology between the APC and MFS superfamilies [4] revealed that this conclusion was based on pairs of mislabeled sequences, leading to the wrong conclusion. At present, we do not have evidence that the APC and MFS superfamilies are related.

The APC superfamily was initially described by Jack *et al*. [5]. In 2000, it contained several subfamilies within the APC family, as well as the AAAP and HAAAP families. In 2012, Wong *et al*. expanded the superfamily to its current state, having 11 families [3]. Most members exhibit a common 5+5 topology, but two families, SulP and NCS2, display a 7+7 topology (see Table I). The evolutionary origin of this alternative topology from the more common 5+5 topology has been determined and discussed [6].

Compared to other superfamilies in the Transporter Classification Database (TCDB; www.tcdb.org), the APC superfamily is unique in having large numbers of high resolution crystal structures of proteins derived from different families, including the structures of ApcT (APC), LeuT (NSS), Mhp1 (NCS1), BetP (BCCT), CaiT (BCCT) and SglT (SSS), representing a common five TMS repeat unit fold with different constellations of extra TMSs, as well as UraA with the less common 7+7 topology (see Table I) [7]. In a recent paper [6], comparisons were made between two representative structures, AdiC [8] and UraA [7], both containing an easily recognized “spiny” secondary structural element located near the substrate translocation site in equivalent positions of each repeat unit (TMS1; see Discussion). Using this shared feature, we established a new nomenclature to facilitate the identification of a common fold, shared between the 5 and 7 TMS repeat units in all then recognized APC superfamily members [6].

In the present study, we report investigation of seven established families, not previously assigned to the APC superfamily: the (1) PAAP, (2) LIVCS, (3) NRAMP, (4) CstA, (5) KUP, (6) BenE and (7) AE families, individually introduced below and in Table I. After establishing homology, we examined their respective topologies and determined how these topologies fit into the previously established framework of the APC superfamily using innovative bioinformatic methods. As high resolution structural data were not available for a member of any one of these seven families, such efforts had to be carried out solely at the sequence level. We used a novel clustering method to identify conserved constellations of glycyl residues in “spiny” helices, and identified previously unknown variants of such helices in CstA homologues in metagenomic samples which exhibited unexpected topological variation. Phylogenetic trees, generated with the SuperFamilyTree (SFT1 and SFT2) programs, revealed the relationships of the proteins and their families to each other.

These SFT programs have been shown to be superior to those used to generate phylogenetic trees based on multiple alignments when sequence divergence is too great to allow construction of reliable multiple alignments [33–35]. We also demonstrate the superiority of the SFT programs over the SATCHMO-JS program which estimates phylogenetic distances based on hidden Markov models (HMMs).

The PAAP family

The Putative Amino Acid Permease Family (TC# 2.A.120; http://tcdb.org/search/result.php?tc=2.A.120) lacks a functionally characterized member. However, a large majority of the genes encoding these transporters occur in operons or gene clusters encoding enzymes of amino acid metabolism. These incude proteases, sporulation proteins, amino acid aminotransferases, amino acid and oxo acid oxidoreductases and amino acid tRNA synthetases. Members display a uniform 5+5 topology (in a 2+3+2+3 arrangement) with a total of 10 TMSs. PAAP homologs are found in *Bacillus subtilis* where some appear to play roles in sporulation. Two such proteins in this organism are YyaD and YkvI, both of which exhibit the 5+5 TMS topology, a uniform feature of members of this family.

The LIVCS family

The Branched Chain Amino Acid:Cation Symporter Family (TC# 2.A.26) includes members that display 11–12 TMSs. Characterized members of this family transport all three branched-chain aliphatic amino acids, leucine (L), isoleucine (I) and valine (V). They are found in Gram-negative and Gram-positive bacteria, e.g., *Pseudomonas aeruginosa* and *Lactobacillus delbrueckii*, respectively, but apparently not in *Archaea* or *Eukaryota*, and function by a Na^+- or H^+-symport mechanism [9–12].

The Nramp family

Homologues of the Metal Ion (Mn^{2+}-Fe^{2+}) Transporter Family (TC# 2.A.55) are found ubiquitously in various yeast, plants, animals, archaea and bacteria and are termed “Natural resistance-associated macrophage proteins” because some of the animal homologues play roles in resistance to intracellular pathogens such as *Salmonella enterica*, *Leishmania donovani* and *Mycobacterium bovis*. The natural history of this family in vertebrates (also referred to as the SLC11 family) has been discussed [13]. Several human pathologies result from defects in Nramp-dependent Fe^{2+} or Mn^{2+} transporters, including iron overload, neurodegenerative diseases and innate susceptibility to infectious diseases [14]. One such protein, NRAMP2, has been reported to transport many different metal ions in the following order: $Fe^{2+} > Zn^{2+} > Mn^{2+} > Co^{2+} > Ca^{2+} > Cu^{2+} > Ni^{2+} > Pb^{2+}$ [15]. The generalized transport reaction catalyzed by Nramp family proteins is: Me^{2+} (out) + H^+ (out) $\rightleftharpoons$ Me^{2+} (in) + H^+ (in) (symport), where Me^{2+} is a metal ion. See TCDB for a more detailed descriptions of this well-characterized family.

The CstA family

The *cstA* gene of *Escherichia coli*, encoding the CstA protein (TC# 2.A.114.1.1), was first identified as a cyclic AMP-CRP-regulated gene, upregulated during carbon starvation [16]. This protein, the first member of the CstA family in TCDB, contains 18 putative TMSs and

was postulated to be a peptide transporter. The evidence derived from results of an experiment that showed that *cstA* Δ*opp* double mutants had a lower growth rate on peptides than the isogenic Δ*opp* mutant. When the peptides were replaced by amino acids as the carbon source, growth was rescued. These suspicions were recently rigorously confirmed using *Campylobacter jejuni*, an organism that relies almost entirely on amino acids for its growth. Expression of the *C. jejuni cstA* gene is upregulated upon carbon starvation [17,18], and Δ*cstA* mutants have a reduced ability to utilize peptides as nitrogen sources [19]. The ArsA ATPase of *E. coli* has been tentatively implicated in the energization of CstA [20], but this suggestion needs to be confirmed.

The KUP family

Proteins of the K^+ Uptake Permease Family (TC# 2.A.72) include the TrkD protein of *E. coli*. This protein has 12 putative TMSs with a requisite, hydrophilic, C-terminal domain of 182 residues, localized to the cytoplasmic side of the membrane [21]. Uptake is blocked by protonophores such as CCCP (but not arsenate), and evidence for a proton symport mechanism has been presented [22]. The *N. crassa* protein has been shown to be a K^+:H^+ symporter, establishing that the KUP family consists of secondary carriers [23].

The BenE family

The Benzoate:H^+ Symporter Family (TC# 2.A.46) includes two functionally characterized members, the benzoate permeases of *Acinetobacter calcoaceticus* and *E. coli* [24]. These proteins probably span the membrane 12 times and exhibit about 30% identity to each other, functioning by a proton symport mechanism [24]. The generalized transport reaction catalyzed by BenE of *A. calcoaceticus* is: benzoate (out) + H^+ (out) → benzoate (in) + H^+ (in).

The AE family

Protein members of the Anion Exchanger (AE) Family (TC# 2.A.31) found in eukaryotes were among the first transporters to be characterized [25]. The human red blood cell anion exchanger, AE1, is a homodimeric complex with a subunit size of 911 amino acids. Its N-terminal hydrophilic domains interact with cytoskeletal proteins and thereby play a structural role in determining cell shape. AE1 also binds carbonic anhydrase II (CAII), forming a ‘transport metabolon’. CAII binding activates AE1 transport activity 10-fold [26]. AE1 is also activated by interaction with glycophorin which functions to target it to the plasma membrane [27]. The membrane-embedded C-terminal domain of AE1 has been proposed by different workers to span the membrane 13–16 times, but the exact number of TMSs has been controversial as different conflicting predictions have been presented. Current evidence from TOPCONS indicates that AE1 contains 12 TMSs, but this software is not necessarily reliable to predict mixed coil/helical and other “spiny” TMS structures. According to the model of Zhu *et al.* (2003), it spans the membrane 16 times, 13 times as α-helices, and three times (TMSs 10, 11 and 14) as β-strands [28]. The membrane-spanning domain is involved in both anion exchange and cation transport [29]. For detailed information about this well-characterized family, see the family description in TCDB.

MATERIALS AND METHODS

Protocols 1 and 2

To establish homology between a family and a superfamily, a suitable pair of query and target sequences are identified by subjecting a member of the new entrant family to TC-BLAST analysis against TCDB. Protocols 1 and 2 are then used in "tweak mode", allowing retrieval of up to 5,000 homologues, using up to two iterations with a high (99%) cutoff in CD-HIT, eliminating only nearly identical sequences [30]. With these settings, we have the best chance of detecting even remote homologs. The current criterion for homology is a GSAT comparison score of 12 standard deviations (S.D.) (but preferably >13 S.D.; larger values are desired due to the growing sizes of public databases). An alignment must also be longer than 60 amino acids, and contain matching TMSs.

Convergent sequence evolution is not easy to identify as it can also result in convergent motifs and convergent structures. The cutoff of 60 residues is the average size of a protein domain, but it is arbitrary, and longer segments are preferred. We are convinced that convergent evolution cannot be responsible for the observed degree of sequence similarity when the sequence length and the other criteria outlined in the Methods section are met. Results must be confirmed using GSAT with 20,000 random shuffles. We rely on the Superfamily Principle, application of the transitivity rule to homologous relationships [31], to establish homology [32].

SuperFamilyTree

The SFT1 and SFT2 programs use tens of thousands of BLAST bit scores instead of multiple alignments [33–35]. They avoid the pitfalls often encountered when determining the phylogenies of distantly related proteins when multiple alignments are unreliable. While SFT1 constructs trees allowing visualization of individual proteins, SFT2 allows depiction of family/subfamily relationships by pooling homologs designated by the user. We have used SFT1 and SFT2 to create trees, showing (a) all proteins (SFT1) and (b) all families (SFT2) using either one or three subfamily members (available through TCDB) to construct the trees. The SFT programs were run on the Triton Super Computer Cluster (TSCC) of the San Diego Supercomputer Center (SDSC).

SFT results were compared with those generated using SATCHMO-JS, a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction [36]. SATCHMO uses a subtree masking procedure to describe each internal node as an HMM, which can be used to determine the branching order to the root.

SATCHMO creates separate alignments of distant families and then performs HMM:HMM comparisons to detect remote similarities. It relies on comparisons between different sets of sequences that are multiply aligned, a potential weakness as noted above. In this respect, the SuperFamilyTree (SFT1 and 2) programs, which are based on BLAST BIT scores for large numbers of binary comparisons, have been shown in six publications to be superior to other programs for depicting deep phylogenies [3,4,33–35,37]. This is in contrast to programs that construct phylogenetic trees that are based on multiple alignments such as Neighbour-Joining [38], Bayesian integration over alignments [39], and maximum parsimony [40]. When the

proteins are sufficiently similar to generate reliable multiple alignments, the results from SFT agree with those obtained using multiple alignments because similarities can be detected between any set of homologs using variable parts of the sequences (Yen et al. 2009; Yen et al. 2010; Chen et al. 2011). The matrix that is fed into the Fitch algorithm summarizes all similarities detected, rather than an aggregated trend found in a given multiple alignment, which need not be generated.

TMS clustering method

Using TMHMM Server v. 2.0, 44,757 transmembrane segments (TMSs) were predicted in the Transporter Classification Database (TCDB). The amino acid sequence of each TMS was parsed from the data. Using BLASTP with an E-value cutoff of 100, having the query SEG filter off, we subjected the TMSs to a BLAST analysis against themselves, resulting in a matrix of pairwise distances, measured in E-values.

We used the Relevant-Set Correlation (RSC) model for clustering [41], that requires no direct knowledge of the nature or representation of the data. The method can be used to find the relative importance of cluster candidates of various sizes, avoiding problems of bias of methods using fixed neighborhood sizes. Neighborhoods were constructed directly from the 44,757 × 44,757 table, using brute force methods. The minimum threshold on cluster size was set to 3.

We created a look-up table of the clustered objects that shows which TC#s they map to. We went back to the original script that outputted the excised TMSs and numbered them consecutively. Using the look up table, we could reprint only lines in clustered_objects_with_clusters that matched the expanded APC superfamily.

We selected only clusters that had more than 10 members. For these clusters, we recorded the nominal TMS# (could vary depending on the TMHMM predictions; beta sheets were not considered). Selection was partly influenced by family size, since bigger families have more sequences.

454 Sequencing

Termite colonies—Termite samples were collected from a colony located at the West Virginia University Department of Biology Arboretum, Morgantown, WV.

Termite species identity—Genomic termite DNA was isolated [42] from 10 individuals of each colony, and various PCR amplifications were performed for the determination of species identity. The termite species was identified as the Eastern subterranean termite, *Reticulitermes flavipes*, based on cytochrome oxidase subunit II (COII; 99% sequence identity), 16S rRNA (99% sequence identity), 12S rRNA (99% sequence identity), and various morphological markers [43].

Termite dissections—The guts of approximately 200 worker termites were dissected by pulling on abdomens, and P3 hindgut regions were identified (under 40X magnification) and placed in RNAlater tissue buffer (Applied Biosystems).

RNA isolation and Illumina sequencing—DNA was removed from 29 μg of total RNA using the RNAse-Free DNAse set followed by purification on an RNeasy column (Qiagen, Valencia, CA). Bacterial and eukaryotic rRNAs were removed from the total mRNA with the Ribominus Eukaryote and Ribominus Transcriptome isolation kit (Bacteria) (Invitrogen, Carlsbad, CA). Messenger RNA was removed from the rRNA-depleted total RNA with the Oligotex mRNA purification kit (Qiagen, Valencia, CA). The remaining RNA was used to prepare an RNA-seq library using the mRNA Seq Sample Prep kit (Illumina, San Diego, CA). The RNA_seq library was normalized with the Trimmer Direct Kit (Evrogen, Russia) following the manufacturer's protocol. The RNA_seq library was loaded on 4 Illumina lanes (12 pmol/lane), with 80 bp sequenced from each end of the cDNA fragments (paired-end sequencing). Illumina sequencing was performed at the University of Illinois Urbana-Champaign's W.M. Keck Center for Comparative and Functional Genomics. A total of 111 million reads were generated encompassing a yield of 17.8 billion bases.

Bioinformatic analysis of diversity in CstA's spiny helix—Using a script written in the programming language C, paired end reads from the metagenomics data were scanned for the following motif: P-x(7)-CG-x(2)-SG-x-H (http://pfam.xfam.org/family/PF02554#tabview=tab4), which corresponds to the CstA family (Table S2), using degenerate bacterial codons. 25 million paired end reads could be analyzed in less than two minutes. True positives were identified in the output from each run using discontiguous Mega BLAST. As a comparison, we scanned for the following motif, G-x-GN-[ILVM]-x-G-x(7)-GG, in TMS 2 of sodium:alanine symporters (Table S3). The rationale for focusing on the sodium:alanine symporters is that they contain an equally detectable and conserved motif in Pfam (http://pfam.xfam.org/family/PF01235#tabview=tab4).

RESULTS

Inclusion of seven novel families within the APC superfamily

Using Protocol 1 [30], a wrapper of PSI-BLAST and CD-HIT to draw up lists of homologs from the NCBI NR protein database in a standardized way, we obtained a set of homologues using TC# 2.A.120.1.1 (a representative of the PAAP family) as the query sequence, and retrieved 1,190 homologues using two iterations and a 0.99 cutoff as the CD-HIT setting to only filter out nearly identical matches. Using TC-BLAST, we found the closest hit within the APC superfamily to be TC# 2.A.3.8.12 (within the LAT subfamily of the APC family; see Table I). Repeating Protocol 1 on this LAT sequence yielded 5,000 sequences, the maximal number retrievable. Running Protocol 2, a wrapper of the GSAT sequence shuffling and alignment program used to obtain statistical evidence of homology [30], revealed a binary alignment with a GSAT Z-score of 16 standard deviations (S.D.). Confirming this score with 20,000 random shuffles produced a precise score of 15.5 S.D., above the current threshold of 13 S.D. for establishing homology. The alignment was long, containing 8 TMSs, 1 through 8 of both sequences, which had 10 and 12 putative TMSs, respectively (see Fig. S1A). Thus, homology between the PAAP family and the APC superfamily was established. Table II presents the comparison scores for the three pairwise comparisons that established homology, A versus B, B versus C, and C versus D [30]. This

same procedure was applied to establish homology in all new families described below. The next family to be considered is the LIVCS family.

For the next family, LIVCS, Protocols 1 and 2 were used with the same settings. The starting sequences were TC#s 2.A.26.1.1 and 2.A.3.14.1 (APC), which produced a precise Z-score of 14.7 S.D. The alignment was long, including TMSs 1–11 of both homologues (see Figs. S1B and Table II).

For the Nramp family, the sequences TC#s 2.A.55.3.5 and 2.A.26.1.1 (LIVCS) yielded a precise GSAT-Z-score of 16.8 S.D. The alignment was long, containing matching TMSs 2 through 10 (Figs. S1C and Table II).

For the CstA family, using sequences TC#s 2.A.114.1.1 and 2.A.39.1.6 (NCS1, a current member of the APC superfamily), an alignment of homologues scored 14.7 S.D. (Figs. S1D and Table II). The alignment was long, containing 7 aligning TMSs from each homologue (TMSs 7–13 and 3–9, in the two proteins, respectively). Thus, homology was established between an established member of the APC superfamily and the CstA family.

For the KUP family, using sequences TC#s 2.A.72.1.1 and 2.A.3.14.2 (APC), we obtained an alignment of homologues scoring 19 S.D. (Figs. S1E and Table II), producing a long alignment with corresponding TMSs.

For the BenE family, using sequences TC#s 2.A.46.1.1 and 2.A.40.3.2 (NCS2), homologues aligned with a score of 13.1 S.D. (Figs. S1F and Table II). The alignment is long with corresponding TMSs aligned as expected.

For the AE family, using sequences TC#s 2.A.31.1.1 and 2.A.40.1.4 (NCS2), we obtained an alignment of homologues scoring 15.0 S.D. (Figs. S1G and Table II) with matching TMSs. The conclusion of homology applies to both sodium-dependent NBCs and sodium-independent AEs, all members of the AE family in TCDB. These proteins are internally homologous throughout most of their lengths.

Repeat units and topological features of novel APC families

Some of the novel families display interesting and sometimes unique topological features (Table III). PAAP (TC# 2.A.120) has 13 members in TCDB, all of which have the same topology of 5+5 TMSs. The closest protein (outside of 2.A.120; using TC-BLAST) in TCDB is TC# 2.A.3.8.12 (5+5+2 TMSs), and using HHpred, the closest protein appeared to be AdiC (2.A.3; PDB:3L1L). Residues 1–300 (TC# 2.A.120.1.1; TMSs 1–8) aligned with AdiC (PDB:3L1L) (TC# 2.A.3.2.5; TMSs 1–8). PAAP clearly contains just the two repeat units without extra TMSs (Table III).

LIVCS (TC# 2.A.26) has 8 members in TCDB, all of which have 12 TMSs, in a 5 + 5 + 2 arrangement. Using HHpred, the best template proved to be ApcT, (TC# 2.A.3.6.3; PDB: 3GIA; the crystal structures of GadC and ApcT are available); AdiC was in second place. TMSs 1–8 (residues 1–300) of 3L1L and 2.A.26.1.1 aligned, confirming that its two extra TMSs are at its C-terminus (Table III). This pattern is the same as for AdiC [8].

NRAMP (TC# 2.A.55), for which homology to the APC superfamily could easily be established via LIVCS, exhibits greater topological diversity with 11–13 apparent TMSs (Table III). The family has 3 subfamilies; members of subfamilies 1 and 3 display 11 TMSs but subfamily 2 shows greater diversity. The GSAT alignment showed that TMSs 2–10 from both NRAMP and LIVCS homologues aligned (Figs. S1B and C) with a one-to-one correspondence between the TMSs. Using the MEME program, conserved motifs in TMS 1 (DPGN) and TMS 6 (G-x(3)-M-x(5)-L) were identified in the NRAMP homologues. The consistent placement of these motifs suggests that topological variation involves extra TMSs 1–3 at the N-terminal end of the sequences in subfamily 2 of the Nramp family (Table III).

The CstA family has 8 members in TCDB, displaying 13–18 (13, 15, 16 or 18) predicted TMSs, more variation than for other previously established families in the APC superfamily (Fig. 2A). Using MEME, the motif, CG-x(2)-SG, was found in TMS 1 of the second repeat unit. Using Pfam's HMM logo for CstA, this motif could be spotted directly in each sequence in the TMSs that line up with the spiny TMS in the second repeat unit (Fig. S3). AdiC (PDB:3L1L) is the closest relevant homology modeling template for these family members. The sequence diversity of the spiny TMSs is substantial (Table S2). For example, there are conserved motifs towards the end of the spiny TMS in the first repeat unit, and other different but highly conserved motifs in the beginning and end of the spiny helix in the second repeat unit. This revealed that the sequence variation is considerable, even though the helix maintains its spiny appearance in the different crystal structures available for APC superfamily members. Despite the great topological diversity of these sequences, running each sequence through HHpred consistently aligned the CG-x(2)-SG motif-containing TMS with the second spiny TMS (TMS# 6, the first TMS in the second repeat unit). Based on this observation, we oriented the sequences in relation to the "I" of the second repeat unit (i.e. IUV-*I*UV; Fig. 2A). In all cases, the sequences of the CstA family have two extra TMSs after the second repeat unit (e.g. IUV-IUV-V), except in CstA of *E. coli* (TC# 2.A.114.1.1) which has two extra 2 TMS units after the second repeat unit. Furthermore, all sequences have at least one leading (N-terminal) TMS, and sometimes a leading 3 or 4 TMS unit (see Table III).

We used CstA of *E. coli* (TC# 2.A.114.1.1) to run Protocol 1 when establishing homology. This sequence has 18 putative TMSs. The original alignment supports the conclusion that 2.A.114.1.1 has a leading group of 4 TMSs before the first repeat unit. Therefore, the alignment starts at TMS 5 in the CstA sequence which aligns with TMS 1 of AdiC in the IUV-IUV-V configuration. The same proved to be true for a confirmatory alignment, obtained by running Protocols 1 and 2, where the alignment starts with TMS 7 in the CstA homologue and TMS 3 of the homologue in the IUV-IUV-V configuration. Thus, we have not only shown homology of established APC superfamily members to CstA family members, but have also located the repeat units within this highly variable family (Table III and Fig. 2A).

The "K_trans" Pfam family is closely related to the APC family (TC# 2.A.3), and it appears that the KUP family (TC# 2.A.72) shares the common 5+5+2 topology characteristic of AdiC and other APC family members. We conclude that KUP family members have 12 α-helical TMSs.

BenE is most closely related to TC family 2.A.40 (UraA; 7+7 topology). It should be noted that the "spiny" helices of UraA's topology are part β-structure. This makes these spiny structures easy to predict using secondary structure prediction software, such as the one built into HHpred's search engine. We conclude that BenE (2.A.46) has 14 TMSs (12 α-helical TMSs and 2 mixed coil/helical TMSs).

The multifaceted Anion Exchanger 1 (AE1) protein of the human red blood cell (Band 3) was one of the first transporters to be functionally characterized [44]. It functions in chloride:bicarbonate exchange, selenium metabolism, volume regulation, erythrocyte cytoskeletal cell structure and metabolic coordination [45–47]. Its topology has been controversial, varying from 13 to 16 TMSs, depending on the prediction method used. No crystal structure is available for the TMS-containing domain, although multiple crystal structures have been published for the soluble part of this large and versatile protein. Using TOPCONS, AE1 is predicted to have a 12 TMS topology. In Pfam, the family corresponding to TCDB's AE family (HCO3) is a member of the APC clan and is thought to be related to UraA's family (Xan_ur_permease), suggesting a relation to the original 7+7 topology (VIUV-VIUV), of which the spiny helices are partly β-structured in UraA. Using HHpred to model AE1 with UraA as the template, all of AE1 aligned with UraA. The alignment with secondary structure predictions is shown in Fig. S2. It seems clear that TOPCONS mispredicted the spiny TMSs of AE1, probably because these structures are only half length α-helical segments consisting of 50% coil instead of 50% β-structure (Fig. S2). These spiny stretches match up with corresponding sequences in AE1 that, while not predicted by HHsearch's secondary structure prediction to be beta-structured, have a mixed coil-α-structured nature. Thus, we have clarified the long standing issue of AE1's topology. Some of the lower TMS estimates are due to the mixed-coil/helical nature of the spiny helices in AE1. We conclude that AE (2.A.31) members have 14 TMSs (12 α-helical TMSs and 2 mixed coil/helical TMSs) (Table III).

Pfam analysis of families within the APC superfamily: the Pfam APC clan (CL0062)

The Pfam APC clan (CL0062) contains 18 families [48] (Fig. 1). In the relationship diagram (http://pfam.sanger.ac.uk/clan/CL0062#tabview=tab3), it can be seen that these 18 families are organized in three separate clusters. These clusters agree with our SuperFamilyTree (Fig. 3). To understand where established APC superfamily members are placed within the Pfam system, we took a single representative sequence (e.g., 2.A.25.1.1 as a representative of 2.A.25.1) from each subfamily in TCDB and analyzed these proteins using Pfam. Using HMMER 3.1b1, allowing E-values only of less than $1e^{-5}$ against Pfam-A, we scanned TCDB to determine which Pfam families in Pfam's APC clan match TC families within the APC Superfamily. While some families have a clear one-to-one correspondence between the two databases, such as the BCCT (2.A.15) family, which matches the BCCT Pfam family in a one-to-one manner, others have a more complex relationship. For example, the APC (2.A.3) family, which has 15 recognized subfamilies in TCDB, matches both AA_permease and AA_permease_2, but also, to a lesser extent, Spore_permease and AA_permease_C. This last mentioned family partly overlaps the matching pattern of 2.A.30, which also matches AA_permease and AA_permease_2, as well as AA_permease_N. This last Pfam family is

not presently a member of the APC clan but is represented in TCDB as being equivalent to parts of the CCC (2.A.30) family. These results are summarized in Table IV.

TMS clustering and CstA polymorphisms in metagenomics data

Soft clustering yielded 321 clusters, the biggest having 115 members. The average cluster size was ~30. Overall, 9,070 objects were contained in the clusters, constituting ~20% of all objects.

Two files were generated, clustered_objects and clustered_objects_with_clusters. In the clustered_objects file, there were 9,075 lines but only 8,487 unique lines because lines reappear if the same sequence (TMS) is present in multiple clusters (rare but possible). For example, sequence 237 (“plvraglfgfngtlagialpfff”) is found in three clusters: 169, 294 and 295. In the file clustered_objects_with_clusters, there were 9,075 lines and 320 clusters. In principle, we were only interested in clusters containing matches to all 18 families within the APC superfamily.

Examples of glycine-rich, spiny TMSs include TMS 1 of the NSS family (TC# 2.A.22) (which contains G-x(3)-G-x-G), TMS 1 of the SulP family (TC# 2.A.53) (G-x(3)-G), TMS 1 of the Amino Acid/Auxin Permease (AAAP) family (2.A.18) (G-x-G-x(2)-G), TMS 6 of the AAAP family (G-x(7)-G-x-G), TMS 1 of the KUP family (2.A.72) (G-x(2)-G), TMS 1 of the APC family (either GG-x(2)-G-x-G or G-x(3)-G-x-G), and TMS 6 of the APC family (2.A.3) (either GG, G-x(4)-GG or G-x(3)-G) (see Table S1). Several of these longer motifs contain two glycyl residues separated by one helical turn, a pattern noted previously within the drug/metabolite (DMT; 2.A.7) superfamily [49]. Exceptions include glycine-rich helices that are not located at the start of a repeat unit. Examples include TMS 9 of the NSS family (2.A.22) as well as TMSs at the start of a repeat unit that are not glycine-rich, such as TMS 6 in the NSS family.

In order to expand the collection of the CstA polymorphisms, we took advantage of termite gut metagenomics data (Table S2 and Fig. 2B). Two of the polymorphisms were located within the core motif region of CstA in Pfam. These polymorphisms contained previously undocumented variations involving conservative substitutions: FISI instead of FITI, and CGSISG instead of CGAISG. To compare the amount of variability observed for the CstA spiny helix with any other TMS, we scanned for variants of the motif G-x-GN-[ILVM]-x-G-x(7)-GG (glycine rich-but not in a spiny helix) located in TMS 2 of the Pfam sodium:alanine symporter (Na_Ala_Symp) family. The Pfam Na_Ala_Symp family matched the PAAP (2.A.120) and the AGCS (2.A.25) families in TCDB. Surprisingly, 14 of the sequences corresponded to previously recognized sequences in public databases, and only nine of them differed (see Table S3). This showed that the proportion of sequences displaying variations in the spiny helices of CstA homologues are more extensive (~66% of the helical motifs; compare motifs in Tables S2 and S3). The fact that variants of the CstA spiny helix motif (CG-x(2)-SG) could be identified reliably in the termite gut data set indicates a high degree of conservation. This motif in the CstA family, the most topologically diverse confirmed member of the APC superfamily, may be essential for function.

In the clustering results, we can see other cases where we have a spacing of glycyl residues equivalent to that typical of the CstA family. For example, in TMS 1 of 12 of the NSS family (TC# 2.A.22), the clustering approach identified a submotif of G-x-A-x-G. The same glycyl residue spacing is found in the spiny TMS 1 of the SulP family (TC# 2.A.53), TMS 1 of the APC family, and in TMS 6 of the APC family. While serine is a common catalytic site residue that is relatively common in the motifs identified by the clustering approach (Table S1), there appears to be only one case where a conserved serine occurs in another APC family in a constellation with glycyl residues, and that is in spiny TMS 6 of the NSS family.

APC phylogenetic analyses with the SuperFamilyTree and SATCHMO-JS programs

Phylogenetic trees were generated with SuperFamilyTree programs 1 and 2 (SFT1 and 2) as follows: (1) using a single protein from each subfamily within the APC superfamily with SFT1, (2) using up to three proteins from each subfamily with SFT1, and (3) using all the proteins in (2) with the SFT2 program to show only the relationships of the families to each other [33–35]. All of these trees gave excellent agreement with respect to the relative positions of the proteins and the families, with the proteins of each subfamily clustering together to give a coherent pattern, and with all subfamilies of a family clustering together as well. In no case did members of one family intermix with those of another. The SFT2 tree showing the family relationships is presented in Fig. 3, while the SFT1 tree generated using up to three proteins per subfamily is shown in Supplementary Fig. S4. The family clustering patterns are particularly worthy of note. All of the families with members known or postulated to exhibit the 7+7 TMS topology, SulP, NCS2, BenE and AE, appear on the same branch, with NCS2 and BenE being most closely related. This is consistent with the fact that most substrates of these two families are small aromatic compounds of similar structure. The Anion Exchanger (AE) family exhibits substrates similar to those of the SulP family. The association of members of the AE family with SulP and NCS2 family members that have a known 7+7 TMS topology confirms our postulate (see above) that members of the AE family exhibit this topology.

The close association of the functionally uncharacterized Putative Amino Acid Porter (PAAP) family with the branched chain amino acid:cation symporter (LIVCS) family provides phylogenetic evidence that members of the PAAP family do, in fact, take up amino acids by a cation symport mechanism. It is also worthy of note that although many members of the PAAP family are annotated as sodium:alanine symporters, the PAAP and AGCS families are only distantly related. The latter family, also a member of the APC superfamily (see Fig. 3), includes known alanine uptake symporters. Also worthy of note is the loose clustering of the CstA family with the BCCT family and the Nramp family with the NSS family. However the branch point between these last two families occurs so near the center of the tree that this association cannot be considered established.

The SATCHMO tree, based on HMMs, is shown in Fig. S5. Most important is the confirmation that the four established and putative 7+7 topology families cluster together on a single branch. Further, in both trees, the BenE and NCS2 families are more closely related to each other than they are to the SulP and AE families. The close association noted between

the APC and CCC families in the SFT2 tree (Fig. 3) is not apparent in the SATCHMO tree although the CCC family is at the base of the branch bearing the APC family.

Other clustering patterns are completely different in the two trees. For example, in the SATCHMO tree, the PAAP and LIVCS families are distant from each other although they are close in the SFT trees. More disturbing is the fact that while the SFT trees consistently show all members of each family clustering closely together, this is not true of the SATCHMO tree. For example, AAAP family members (TC# 2.A.18) are found in four distant branches of the tree, while APC family members localize to three distant positions as is true of the three PAAP family members included in our study. Further, SSS family members occur on two distant branches. TC-BLAST results showed that these inconsistencies were not reflective of the distances of the sequences but instead reflected inaccuracies in the SATCHMO program in recognizing similarities. In agreement with previously published results [3,4,33–35,37] we conclude that of the available programs, the SFT programs are best at correctly identifying relationships between sequence-divergent proteins within a single superfamily. This should not be considered surprising since multiple alignments are always unreliable when the sequences are highly divergent, and SATCHMO relies only on conserved motifs instead of the entire sequence of the homologues as does SFT.

These trees were compared with the SFT trees for the APC superfamily previously published [3]. In this publication, three different SFT trees were derived and compared with the then available members of the superfamily: (1) an SFT1 program-derived tree containing all APC superfamily members, (2) an SFT2 program-derived tree showing one branch per subfamily for the entire APC family, and (3) a magnification of the SFT2 program-derived output for the APC family. Comparing the major branches of the published SFT trees with those reported in this paper revealed excellent agreement with no significant discrepancies. Figs. 3 and S4 additionally provide the first phylogenetic data for the seven novel families reported here.

DISCUSSION

The APC superfamily, already the second largest superfamily of secondary carriers [3,5], is now much larger than previously recognized. While most members have a 5+5 TMS topology, frequently with extra TMSs at the N- and C-termini, two families, SulP and NCS2, were known to display a 7+7 TMS topology [6]. Thanks to the availability of several crystal structures for different superfamily members, this topological diversity could be understood, mainly by exploiting a shared and easily recognized structural feature, e.g., a “spiny” TMS located near the substrate translocation channel of many but not all superfamily members. Using this feature and others, we established a new nomenclature for structural features of the APC superfamily using the letters I, U and V to describe the common inverted fold. The I corresponds to a single “spiny” TMS (TMS 1), the U to a hairpin structure (TMSs 2–3) connected by a loop of substantial size, and the V (TMSs 4–5) to a hairpin structure, with a loop of smaller size. We postulated that multiple hairpin and domain duplication events were responsible for the evolution of extra TMSs and showed that the basic 5 TMS fold is common to all members of the APC superfamily [6].

Multiple hairpin and domain duplication events evidently created topological diversity within the APC superfamily, and all members share the common 5 TMS unit [6]. In the current study, we establish homology between the APC superfamily and seven new family members: the PAAP, LIVCS, NRAMP, CstA, KUP, BenE and AE families. Their different topologies fit into the framework of the common LeuT-like 5+5 TMS repeat unit fold, although AE family members appear to conform to the 7+7 TMS topology established for UraA [6].

Although large, the PAAP (Putative Amino Acid Permease) family does not have a single functionally characterized member. However, in Pfam, the equivalents of PAAP are annotated as sodium:alanine symporters. Homology could easily be established between the PAAP family and the APC family. The PAAP family does not appear to display topological variation and contains only the repeat units IUV-IUV without extra TMSs.

The LIVCS family has characterized members that transport the aliphatic, hydrophobic amino acids, leucine, isoleucine and valine using a sodium or proton symport mechanism [9–12]. Homology could be established between the LIVCS family and an established APC-1 family of unknown function, conclusively tying them to the APC superfamily. LIVCS family members appear to have an invariant topology, consisting of the repeat unit (5+5) with two extra C-terminal TMSs, the same pattern displayed by AdiC, the structurally characterized argninine:agmatine antiporter [8].

Nramp family proteins transport inorganic divalent metal cations, particularly manganese and iron, using a metal:proton symport mechanism [15]. Homology between NRAMP and the APC superfamily could be established via the LIVCS family using the Superfamily Principle, indicating that NRAMP is more distant from the original group of APC superfamily members. The NRAMP family, which has three subfamilies, displays greater topological diversity than the PAAP and LIVCS families, having 11–13 TMSs, similar to the SSS family. Motif analyses using MEME revealed that this added topological diversity is confined to the ends of the sequences, and that NRAMP family members consistently contain two copies of the repeat unit extending from the N-terminus of each sequence.

CstA homologues are peptide transporters [16–18]. Homology could be established between CstA and the NCS1 family, members of which transport cytosine, hydroxymethyl pyrimidine, pyridoxine, pyridoxal, pyridoxamine, hypoxanthine, adenine, guanine, allantoin, uracil, uridine, hydantoin, thiamine (i.e., mostly pyrimidines and purines). The topological diversity of CstA family members is complex as different members display 13, 15, 16 and 18 putative TMSs. A particularly conserved motif, CG-x(2)-SG, was exploited to identify the common repeat units within and between the sequences. This motif consistently aligned with the spiny TMS in the second repeat unit in other APC superfamily members. While the spiny TMS is easily recognizable in crystal structures, we have not been able to find a clear sequence motif that correlates with this feature throughout the APC superfamily. Taking advantage of the GSAT alignment with the NCS1 homologue used to establish homology, the patterns suspected using the CG-x(2)-SG motif were confirmed, showing that CstA family members display great topological diversity, involving modifications at both ends of the sequences. The variants of the CstA spiny helix identified in the termite metagenomics

data may represent a more ancestral form of a spiny helix motif or a variant of the motif that is resistant to the huge topological variation in CstA, possibly able to function alone or in combination with an ordinary helix.

Proteins of the K^+ Uptake Permease (KUP) family have 12 putative TMSs [21]. Evidence for a K^+:H^+ symport mechanism has been presented [22], establishing that the KUP family consists of secondary carriers [23]. KUP proved to be demonstrably homologous to members of the U-APC1 family.

The Benzoate:H^+ Symporter family contains the benzoate permeases (BenE) [24], which span the membrane 12 times and function by a proton symport mechanism [24]: benzoate (out) + H^+ (out) $\rightarrow$ benzoate (in) + H^+ (in). BenE proved to be demonstrably homologous to NCS2.

Protein members of the Anion Exchanger family were among the first transporters to be characterized [25]. The human red blood cell anion exchanger, AE1, spans the membrane 14 times, 12 times as α-helices and two times (spiny TMSs 3 and 10) as coil/helical-strands. The membrane-spanning domain is involved in both anion exchange and cation transport [29]. AE proved to be demonstrably homologous to NCS2, and shows the greatest similarity to the Xan_ur_permease family in Pfam (UraA's family).

Helix-breaking glycines in spiny helices generally occur in regularly spaced constellations that are conserved through either a family or a subset of a family. The spiny helices do not contain a unifying sequence motif other than being relatively rich in helix-breaking glycines. While the APC superfamily members have a common fold, it is possible that the substrate specificities are determined by the variable amino acyl residues in the spiny helices. It is well known that a double glycine can act as a helix breaker [50], and our results suggest that other spacings of glycines may have the same effect, where a single glycyl residue may be sufficient to cause a discontinuity in a transmembrane α-helix, depending on the context. An example is the motif TQ-x-FFS-x(5)-G in TMS 6 of LeuT.

CstA contains a highly conserved helix that aligns with the second spiny helix of AdiC. CstA homologues show variable topologies, but all known variants contain a highly conserved helix that aligns with the second spiny helix of CstA homologues from multiple species (Fig. 2A). The previously undocumented variation seen in the CG-x(2)-SG helix of CstA from multiple species (Table S2 and Fig. 2B) likely represents functionally crucial adaptations in the substrate translocation channel found in this category of peptide transporter within the rich termite hindgut microbiota. This may be a crucial adaptation required for symbiosis (cross-feeding [51]) within this community (Table S3).

In conclusion, we have expanded the APC superfamily by establishing homology with seven new families not previously recognized as APC superfamily members. The new families exhibit the 5 or 7 TMS APC superfamily repeat units as well as optional extra N- or C-terminal TMSs. Variation proved to be minimal for the PAAP family and maximal for the CstA family. Spiny TMSs are rich in glycyl residues and contain glycines in conserved constellations, in several cases having them spaced by approximately one helical turn. Examination of homologues from termite hindgut samples showed that the spiny helices of CstA homologues display great sequence variation but conservation of key residues. Most

new family members transport amino acids, metal ions and peptides, typical substrates for established APC superfamily members, almost always using a cation symport mechanism. The fact that inorganic anion exchangers of the AE family belong to the APC superfamily represents an expansion in the types of substrates transported. Our results serve to place this exceptionally well-characterized family into an evolutionary perspective by demonstrating a conclusive alignment between UraA and the TMS-containing domain of AE1. Both appear to have the same 7+7 topology. The spiny helices of AE1 are predicted to be 50% helical and 50% coil.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH Grant GM077402A1 (M.H.S.). An USDA grant awarded to R.R. funded the 454 sequencing, USDA NIFA Grant No. 201034158-20857. We thank Philomène Kabran Paul, University of California, San Diego, for providing background information about the YeeE/YedE family and Bryan Lunt, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, Calif., USA, for assisting with the Triton Supercomputer Cluster.

References

1. Forrest LR, Kramer R, Ziegler C. The structural basis of secondary active transport mechanisms. Biochim Biophys Acta. 2011; 1807(2):167–188. [PubMed: 21029721]
2. Forrest LR, Rudnick G. The rocking bundle: a mechanism for ion-coupled solute flux by symmetrical transporters. Physiology (Bethesda). 2009; 24:377–386. [PubMed: 19996368]
3. Wong FH, Chen JS, Reddy V, Day JL, Shlykov MA, Wakabayashi ST, Saier MH Jr. The amino acid-polyamine-organocation superfamily. J Mol Microbiol Biotechnol. 2012; 22(2):105–113. [PubMed: 22627175]
4. Reddy VS, Shlykov MA, Castillo R, Sun EI, Saier MH Jr. The major facilitator superfamily (MFS) revisited. Febs J. 2012; 279(11):2022–2035. [PubMed: 22458847]
5. Jack DL, Paulsen IT, Saier MH. The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. Microbiology. 2000; 146 (Pt 8):1797–1814. [PubMed: 10931886]
6. Vastermark A, Saier MH Jr. Evolutionary relationship between 5+5 and 7+7 inverted repeat folds within the amino acid-polyamine-organocation superfamily. Proteins. 2014; 82(2):336–346. [PubMed: 24038584]
7. Lu F, Li S, Jiang Y, Jiang J, Fan H, Lu G, Deng D, Dang S, Zhang X, Wang J, Yan N. Structure and mechanism of the uracil transporter UraA. Nature. 2011; 472(7342):243–246. [PubMed: 21423164]
8. 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(7282):828–832. [PubMed: 20090677]
9. Braun PR, Al-Younes H, Gussmann J, Klein J, Schneider E, Meyer TF. Competitive inhibition of amino acid uptake suppresses chlamydial growth: involvement of the chlamydial amino acid transporter BrnQ. J Bacteriol. 2008; 190(5):1822–1830. [PubMed: 18024516]
10. Reizer J, Reizer A, Saier MH Jr. A functional superfamily of sodium/solute symporters. Biochim Biophys Acta. 1994; 1197(2):133–166. [PubMed: 8031825]
11. Stucky K, Hagting A, Klein JR, Matern H, Henrich B, Konings WN, Plapp R. Cloning and characterization of brnQ, a gene encoding a low-affinity, branched-chain amino acid carrier in Lactobacillus delbruckii subsp. lactis DSM7290. Mol Gen Genet. 1995; 249(6):682–690. [PubMed: 8544834]

12. Tauch A, Hermann T, Burkovski A, Kramer R, Puhler A, Kalinowski J. Isoleucine uptake in Corynebacterium glutamicum ATCC 13032 is directed by the brnQ gene product. Arch Microbiol. 1998; 169(4):303–312. [PubMed: 9531631]
13. Neves JV, Wilson JM, Kuhl H, Reinhardt R, Castro LF, Rodrigues PN. Natural history of SLC11 genes in vertebrates: tales from the fish world. BMC Evol Biol. 2011; 11:106. [PubMed: 21501491]
14. Cellier MF. Nramp: from sequence to structure and mechanism of divalent metal import. Curr Top Membr. 2012; 69:249–293. [PubMed: 23046654]
15. Nevo Y, Nelson N. The NRAMP family of metal-ion transporters. Biochim Biophys Acta. 2006; 1763(7):609–620. [PubMed: 16908340]
16. Schultz JE, Matin A. Molecular and functional characterization of a carbon starvation gene of Escherichia coli. J Mol Biol. 1991; 218(1):129–140. [PubMed: 1848300]
17. Wright JA, Grant AJ, Hurd D, Harrison M, Guccione EJ, Kelly DJ, Maskell DJ. Metabolite and transcriptome analysis of Campylobacter jejuni in vitro growth reveals a stationary-phase physiological switch. Microbiology. 2009; 155(Pt 1):80–94. [PubMed: 19118349]
18. Cordwell SJ, Len AC, Touma RG, Scott NE, Falconer L, Jones D, Connolly A, Crossett B, Djordjevic SP. Identification of membrane-associated proteins from Campylobacter jejuni strains using complementary proteomics technologies. Proteomics. 2008; 8(1):122–139. [PubMed: 18095373]
19. Rasmussen JJ, Vegge CS, Frokiaer H, Howlett RM, Krogfelt KA, Kelly DJ, Ingmer H. Campylobacter jejuni carbon starvation protein A (CstA) is involved in peptide utilization, motility and agglutination, and has a role in stimulation of dendritic cells. J Med Microbiol. 2013; 62(Pt 8):1135–1143. [PubMed: 23682166]
20. Castillo R, Saier MH. Functional Promiscuity of Homologues of the Bacterial ArsA ATPases. Int J Microbiol. 2010; 2010:187373. [PubMed: 20981284]
21. Bossemeyer D, Schlosser A, Bakker EP. Specific cesium transport via the Escherichia coli Kup (TrkD) K+ uptake system. J Bacteriol. 1989; 171(4):2219–2221. [PubMed: 2649491]
22. Zakharyan E, Trchounian A. K+ influx by Kup in Escherichia coli is accompanied by a decrease in H+ efflux. FEMS Microbiol Lett. 2001; 204(1):61–64. [PubMed: 11682179]
23. Haro R, Sainz L, Rubio F, Rodriguez-Navarro A. Cloning of two genes encoding potassium transporters in Neurospora crassa and expression of the corresponding cDNAs in Saccharomyces cerevisiae. Mol Microbiol. 1999; 31(2):511–520. [PubMed: 10027968]
24. Neidle EL, Hartnett C, Ornston LN, Bairoch A, Rekik M, Harayama S. Nucleotide sequences of the Acinetobacter calcoaceticus benABC genes for benzoate 1,2-dioxygenase reveal evolutionary relationships among multicomponent oxygenases. J Bacteriol. 1991; 173(17):5385–5395. [PubMed: 1885518]
25. Passow H, Fasold H, Gartner EM, Legrum B, Ruffing W, Zaki L. Anion transport across the red blood cell membrane and the conformation of the protein in Band 3. Ann N Y Acad Sci. 1980; 341:361–383. [PubMed: 6772068]
26. Sterling D, Reithmeier RA, Casey JR. A transport metabolon. Functional interaction of carbonic anhydrase II and chloride/bicarbonate exchangers. J Biol Chem. 2001; 276(51):47886–47894. [PubMed: 11606574]
27. Young MT, Tanner MJ. Distinct regions of human glycophorin A enhance human red cell anion exchanger (band 3; AE1) transport function and surface trafficking. J Biol Chem. 2003; 278(35): 32954–32961. [PubMed: 12813056]
28. Zhu Q, Lee DW, Casey JR. Novel topology in C-terminal region of the human plasma membrane anion exchanger, AE1. J Biol Chem. 2003; 278(5):3112–3120. [PubMed: 12446737]
29. Barneaud-Rocca D, Borgese F, Guizouarn H. Dual transport properties of anion exchanger 1: the same transmembrane segment is involved in anion exchange and in a cation leak. J Biol Chem. 2011; 286(11):8909–8916. [PubMed: 21257764]
30. Reddy VS, Saier MH Jr. BioV Suite--a collection of programs for the study of transport protein evolution. Febs J. 2012; 279(11):2036–2046. [PubMed: 22568782]
31. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol. 1997; 273(1):349–354. [PubMed: 9367767]

32. Saier MH Jr, Reddy VS, Tamang DG, Vastermark A. The transporter classification database. Nucleic Acids Res. 2014; 42(Database issue):D251–258. [PubMed: 24225317]
33. 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 SuperfamilyTree programs over those based on multiple alignments. J Mol Microbiol Biotechnol. 2011; 21(3–4): 83–96. [PubMed: 22286036]
34. 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(4):163–176. [PubMed: 19776645]
35. 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. [PubMed: 20419429]
36. Hagopian R, Davidson JR, Datta RS, Samad B, Jarvis GR, Sjolander K. SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction. Nucleic Acids Res. 2010; 38(Web Server issue):W29–34. [PubMed: 20430824]
37. Reddy BL, Saier MH Jr. Topological and phylogenetic analyses of bacterial holin families and superfamilies. Biochim Biophys Acta. 2013; 1828(11):2654–2671. [PubMed: 23856191]
38. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987; 4(4):406–425. [PubMed: 3447015]
39. Huelsenbeck JP, Joyce P, Lakner C, Ronquist F. Bayesian analysis of amino acid substitution models. Philos Trans R Soc Lond B Biol Sci. 2008; 363(1512):3941–3953. [PubMed: 18852098]
40. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 1996; 266:418–427. [PubMed: 8743697]
41. Houle ME. The Relevant-Set Correlation Model for Data Clustering. Statistical Analysis and Data Mining. 2008; 1(3):157–176.
42. Holmes DS, Bonner J. Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry. 1973; 12(12):2330–2338. [PubMed: 4710584]
43. Ye W, Lee CY, Scheffrahn RH, Aleong JM, Su NY, Bennett GW, Scharf ME. Phylogenetic relationships of nearctic Reticulitermes species (Isoptera: Rhinotermitidae) with particular reference to Reticulitermes arenincola Goellner. Mol Phylogenet Evol. 2004; 30(3):815–822. [PubMed: 15012959]
44. Kopito RR. Molecular biology of the anion exchanger gene family. Int Rev Cytol. 1990; 123:177–199. [PubMed: 2289848]
45. Hongoh M, Haratake M, Fuchigami T, Nakayama M. A thiol-mediated active membrane transport of selenium by erythroid anion exchanger 1 protein. Dalton Trans. 2012; 41(24):7340–7349. [PubMed: 22580993]
46. Wu F, Satchwell TJ, Toye AM. Anion exchanger 1 in red blood cells and kidney: Band 3's in a pod. Biochem Cell Biol. 2011; 89(2):106–114. [PubMed: 21455263]
47. van den Akker E, Satchwell TJ, Williamson RC, Toye AM. Band 3 multiprotein complexes in the red cell membrane; of mice and men. Blood Cells Mol Dis. 2010; 45(1):1–8. [PubMed: 20346715]
48. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014; 42(1):D222–230. [PubMed: 24288371]
49. Vastermark A, Almen MS, Simmen MW, Fredriksson R, Schioth HB. Functional specialization in nucleotide sugar transporters occurred through differentiation of the gene cluster EamA (DUF6) before the radiation of Viridiplantae. BMC Evol Biol. 2011; 11:123. [PubMed: 21569384]
50. Chakrabartty A, Schellman JA, Baldwin RL. Large differences in the helix propensities of alanine and glycine. Nature. 1991; 351(6327):586–588. [PubMed: 2046766]
51. Molloy S. Environmental microbiology: disentangling syntrophy. Nat Rev Microbiol. 2014; 12(1): 7. [PubMed: 24336180]
52. Perez C, Ziegler C. Mechanistic aspects of sodium-binding sites in LeuT-like fold symporters. Biol Chem. 2013; 394(5):641–648. [PubMed: 23362203]

53. Hediger MA, Clemencon B, Burrier RE, Bruford EA. The ABCs of membrane transporters in health and disease (SLC series): introduction. Mol Aspects Med. 2013; 34(2–3):95–107. [PubMed: 23506860]

54. 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(9):2261–2267. [PubMed: 21378181]

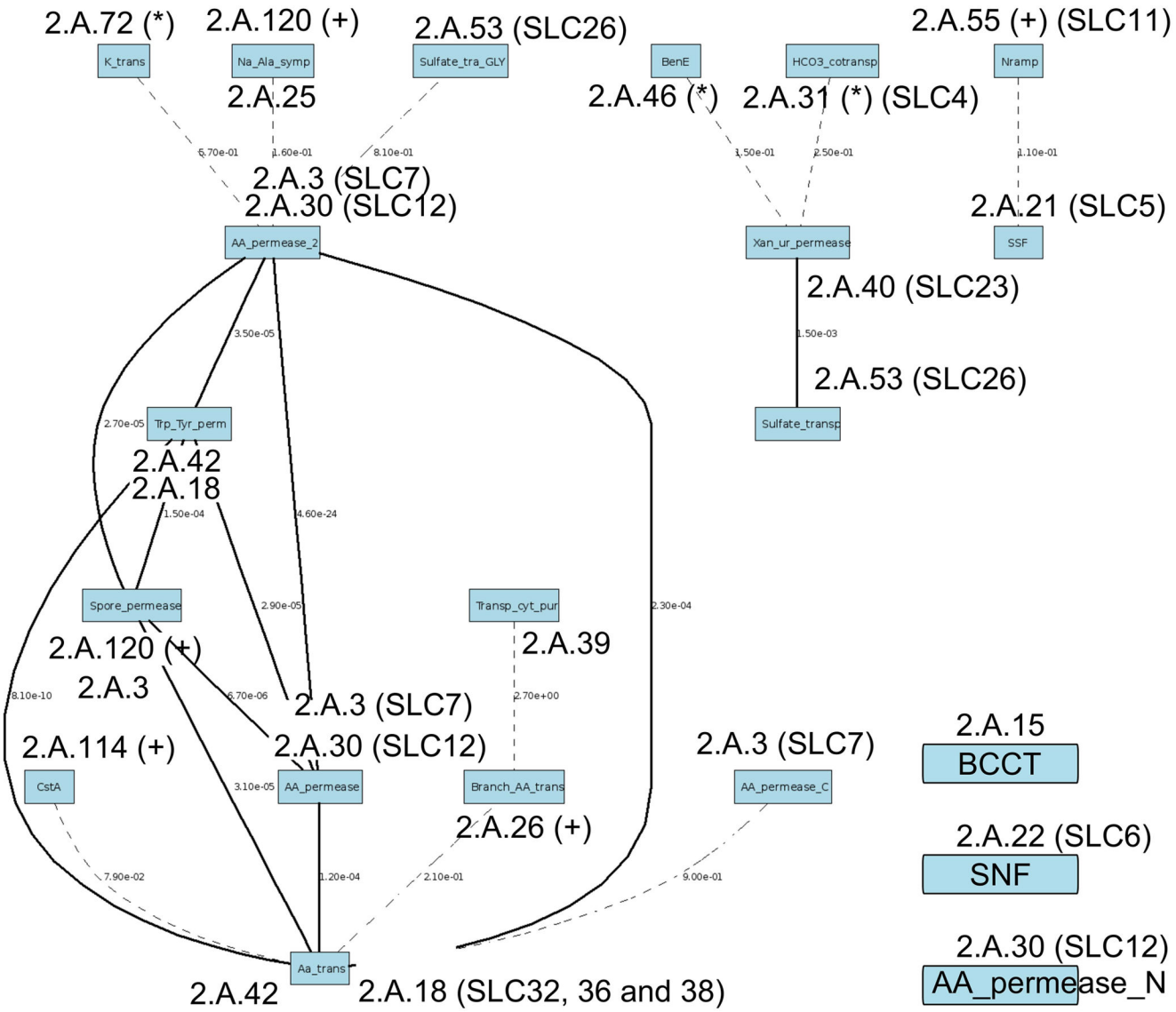

Figure 1.

Diagram (adapted from Pfam) showing family relationships between Pfam families that are members of the APC clan (CL0062). HMMER 3 was used to map the basic correspondences of TC families to these Pfam families. New families, not previously recognized as APC superfamily members, are highlighted by (+). A second set of new families, highlighted by (*), were discovered through this Pfam analysis.

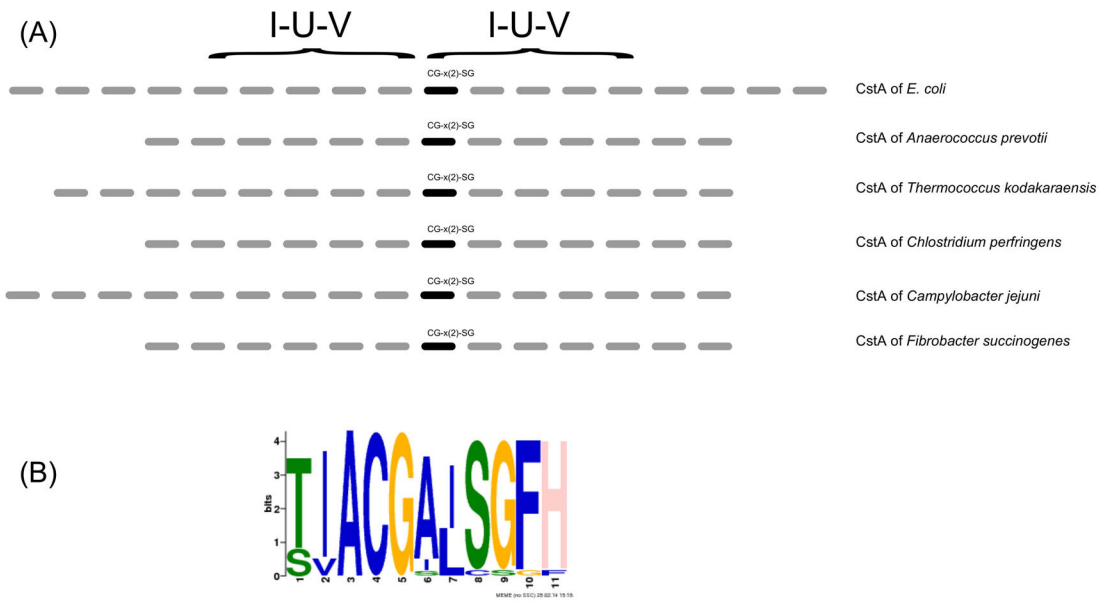

Figure 2.

Figure 2A. Diagram showing location of the CG-x(2)-SG motif in the spiny helix (TMS 1) of the second repeat unit for CstA homologues in six different bacterial species. This degree of topological variation and the high degree of sequence conservation of CG-x(2)-SG are not seen in other APC superfamily members.

Figure 2B. Motif variation in the spiny helix of CstA discovered in the metagenomics data and summarized using MEME. Compared to the HMM logo of CstA in Pfam, the predominance of serine at positions relative to the fully conserved CG is not represented in Pfam for the entire family (where the position is dominated by threonine, T). The position (+1) relative to SG is dominated by phenylalanine (F), whereas in Pfam, this position displays interchangeable non-polar amino acids (FWY). Note also that in the metagenomics sample, in rare cases, there has been a shift of the position of SG relative to CG due to an insertion of one amino acid (C).

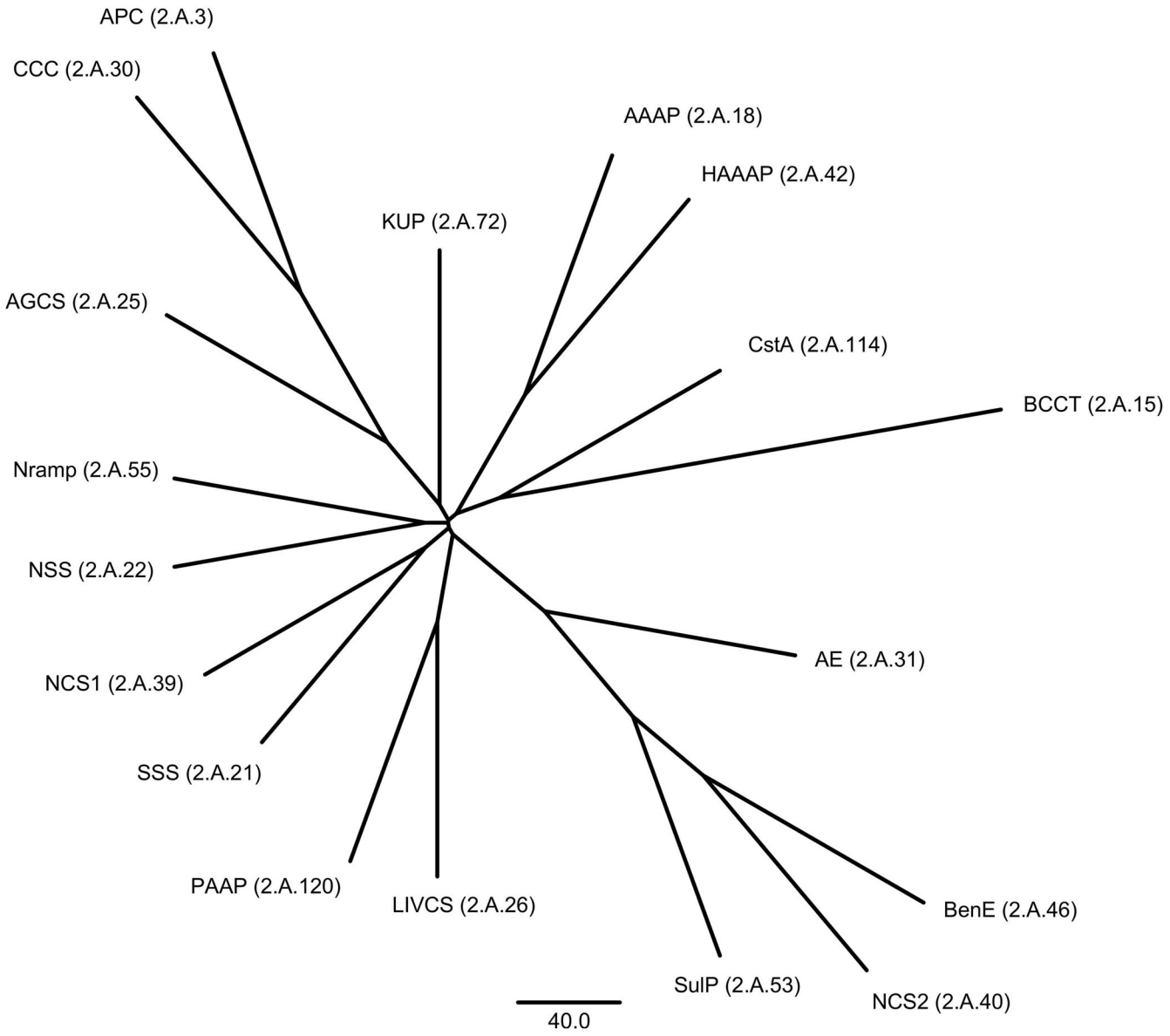

Figure 3.

The SuperFamilyTree-2 (SFT2) tree showing family relationships. As many as three sequences were used to represent each subfamily within each of the families shown when available in TCDB. Note that the SulP, NCS2, BenE and AE families, which are believed to display the 7+7 topology, all cluster together. Note also that (1) CstA and BCCT, (2) AAAP and HAAAP, (3) APC and CCC, (4) Nramp and NSS, (5) SSS and NCS1, and (6) PAAP and LIVCS cluster together, suggesting that these pairs of families are more closely related to each other than to other families in the APC superfamily. In all cases, previously recognized families show relationships that are in full agreement with this tree [3].

Table I

Established and novel families within the APC superfamily.

Abbreviation	Name	TC#	Number of members in TCDB	Most common arrangement of TMSs (Topology)
APC	The Amino Acid-Polyamine-Organocation Family	2.A.3	137	5+5+2 TMSs
BCCT	The Betaine/Carnitine/Choline Transporter (BCCT) Family	2.A.15	14	2+5+5 TMSs
AAAP	The Amino Acid/Auxin Permease (AAAP) Family	2.A.18	55	11
SSS	The Solute:Sodium Symporter (SSS) Family	2.A.21	50	5+5+4 TMSs
NSS	The Neurotransmitter:Sodium Symporter (NSS) Family	2.A.22	48	5+5+2 TMSs
AGCS	The Alanine or Glycine:Cation Symporter (AGCS) Family	2.A.25	8	11
CCC	The Cation-Chloride Cotransporter (CCC) Family	2.A.30	16	12
NCS1	The Nucleobase:Cation Symporter-1 (NCS1) Family	2.A.39	23	5+5+2 TMSs
NCS2	The Nucleobase:Cation Symporter-2 (NCS2) Family	2.A.40	26	7+7 TMSs
HAAAP	The Hydroxy/Aromatic Amino Acid Permease (HAAAP) Family	2.A.42	11	11
SulP	The Sulfate Permease (SulP) Family	2.A.53	47	8–13
PAAP	The Putative Amino Acid Permease (PAAP) Family	2.A.120 (formerly 9.B.160)	13	10 TMSs (5+5)
LIVCS	The Branched Chain Amino Acid:Cation Symporter (LIVCS) Family	2.A.26	8	12 TMSs (5+5+2)
NRAMP	The Metal Ion (Mn2+-iron) Transporter (Nramp) Family	2.A.55	21 in 3 subfamilies	11–13 TMSs (5+5+2)
CstA	The Peptide Transporter Carbon Starvation CstA (CstA) Family	2.A.114	8	13, 15, 16 or 18 TMSs
KUP	The K^+ Uptake Permease (KUP) Family	2.A.72	13	12 TMSs (5+5+2)
BenE	The Benzoate:H^+ Symporter (BenE) Family	2.A.46	8	14 TMSs (7+7)
AE	The Anion Exchanger (AE) Family	2.A.31	22	14 TMSs (7+7)

Table II

Comparison scores to establish homology.

Subject [1)]		Target		Score (S.D.)
PAAP	TC# 2.A.120.1.1	LAT	TC# 2.A.3.8.12	15.5
				A-B 94.5 **C-D 59.7**
LIVCS	2.A.26.1.1	U-APC1	2.A.3.14.1	14.7
				A-B 42.6 **C-D 15.0**
NRAMP [2)]	2.A.55.3.5	LIVCS	2.A.26.1.1	16.8
				A-B 49.4 **C-D 48.7**
CstA [3)]	2.A.114.1.1	NCS1	2.A.39.1.6	14.7
				A-B 94.9 **C-D 500** [4)]
KUP	2.A.72.1.1	U-APC1	2.A.3.14.2	19.0
				A-B 106.8 **C-D 70.3**
BenE	2.A.46.1.1	NCS2	2.A.40.3.2	13.1
				A-B 83.7 **C-D 68.5**
AE	2.A.31.1.1	NCS2	2.A.40.1.4	15.0
				A-B 72.2 **C-D 40.1**

[1)] Two families YeeE/YedE (9.B.102; showing extensive topological diversity, containing two copies of the Sulf_transp motif and a conserved G- and C-based motif) and PiT (2.A.20) showed scores of 13.2 and 13.1 (just over the current threshold), respectively. For alignment of the YeeE/YedE homologues, the TMS numbering was not directly equivalent, and the length was just under the minimal acceptable length. For the PiT family, the same problems applied, the alignment length being even shorter. For these reasons, these potential APC superfamily members were not included in the superfamily or analyzed further.

[2)] Cellier [14] had suggested that, based on homology modeling results, NRAMP displays a LeuT fold, which contains two short extended sequence motifs that interrupt the central parts of TMSs 1 and 6 (i.e. the “spiny” helices), forming an unusual structural pattern that was found in various families of cation-driven transporters.

[3)] CstA in Pfam (PF02554) is a member of the APC clan (CL0062). In the relationship diagram of CL0062 (http://pfam.sanger.ac.uk/clan/CL0062#tabview=tab3), it can be seen that the CstA family displays an E-value of 7.9e-02 to the Aa_trans family (PF01490).

These scores have been amplified by re-running Protocols 1 and 2.

Table III

Topological diversity, predicted using the IUV nomenclature.

Name	TC #	Topology *1)*
PAAP	TC# 2.A.120.1.x	**IUV- IUV**
LIVCS	2.A.26.1.x	**IUV- IUV**-V
NRAMP	2.A.55.1.x	**IUV- IUV**-I
	2.A.55.2.1-3	**IUV- IUV**-VI
	2.A.55.2.4-5, 7–12	**IUV- IUV**-V
	2.A.55.2.6	**IUV- IUV**-I
	2.A.55.3.x	**IUV- IUV**-I
CstA *2)*	2.A.114.1.1	VV-**IUV- IUV**-VV
	2.A.114.1.2	I-**IUV- IUV**-V
	2.A.114.1.3	IV-**IUV- IUV**-V
	2.A.114.1.4	I-**IUV- IUV**-V
	2.A.114.1.5	VV-**IUV- IUV**-V
	2.A.114.1.6	I-**IUV- IUV**-V
KUP	2.A.72	**IUV- IUV**-V
BenE	2.A.46	V**IUV**-V**IUV**
AE	2.A.31 3)	V**IUV**-V**IUV**

1) Without 3-d structures available, it is not possible to determine the physical shapes of the extra TMSs; however, on the basis of homology within repeat units, we assume that they retain the basic IUV (5 TMS repeat unit) or VIUV (7 TMS repeat unit) structure of the LeuT-like fold [52]. The I, U and V labels refer to TMSs 1, 2–3, and 4–5, respectively, in each repeat unit, which physically look like an I, a U and a V, respectively, in the 3-d structures of APC superfamily proteins.

2) Given the topological diversity observed so far, it is likely that further topological variants exist.

Proposed to contain 12 a-helical TMSs and 2 part-α, part-coil TMSs.

Table IV

Pfam family-TC family equivalences.

Pfam family	Pfam clan	TC family	SLC system *1)*
K_trans	APC (CL0062)	KUP (TC# 2.A.72)	
Na_Ala_symp	APC (CL0062)	PAAP (2.A.120); AGCS (2.A.25)	
Sulfate_tra_GLY	APC (CL0062)	SulP (2.A.53)	SLC26
AA_permease_2 *2)*	APC (CL0062)	APC (2.A.3); CCC (2.A.30)	SLC7, 12
Trp_Tyr_perm	APC (CL0062)	HAAAP (2.A.42); AAAP (2.A.18)	
Spore_permease *2)*	APC (CL0062)	PAAP (2.A.120); APC (2.A.3)	
Transp_cyt_pur	APC (CL0062)	NCS1 (2.A.39)	
CstA	APC (CL0062)	CstA (2.A.114)	
AA_permease *2)*	APC (CL0062)	APC (2.A.3); CCC (2.A.30)	SLC7, 12
Branch_AA_trans	APC (CL0062)	LIVCS (2.A.26)	
AA_trans	APC (CL0062)	HAAAP (2.A.42); AAAP (2.A.18)	SLC32, 36, 38
AA_permease_C *2)*	APC (CL0062)	APC (2.A.3)	SLC7
BenE	APC (CL0062)	BenE (2.A.46)	
HCO3_cotransp	APC (CL0062)	AE (2.A.31)	SLC4
Xan_ur_permease	APC (CL0062)	NCS2 (2.A.40)	SLC23
Sulfate_transp	APC (CL0062)	SulP (2.A.53)	SLC26
Nramp	APC (CL0062)	Nramp (2.A.55)	SLC11
SSF	APC (CL0062)	SSS (2.A.21)	SLC5
BCCT *3)*	N/A	BCCT (2.A.15)	
SNF *3)*	N/A	NSS (2.A.22)	SLC6
AA_permease_N *2) 3)*	N/A	CCC (2.A.30)	SLC12

1) Nine APC superfamily families are not found within the SLC system [53] because there is no known mammalian representative.

2) Subfamilies 2.A.3.1-2, 4–8 and 10–15 match both AA_permease and AA_permease_2; subfamily 2.A.3.3 matches AA_permease, AA_permease_2 and AA_permease_C, and the subfamily 2.A.3.9 matches Spore_permease. This last family is incorrectly annotated as these proteins are spore germination receptors, apparently lacking transport activity [54]. Family 2.A.30 has 6 subfamilies. 2.A.30.1 and 3–4 match AA_permease and AA_permease_2, as well as AA_permease_N. However, 2.A.30.2 and 5–6 match only AA_permease and AA_permease_2. These results are attributable to the fact that the subfamily architectures of 2.A.3 and 2.A.30 in TCDB provide higher resolution than the corresponding Pfam system of classification.

3) Some of the TC families occur outside of the APC clan. Thus, Pfam's BCCT, SNF and AA_permease_N are not currently included in Pfam's APC clan. According to our results, these families should be included. We have forwarded these recommendations for clan expansion to the Pfam team.