Identification of The Three Distinct Fold Types of the CPA/AT Superfamily.

Govindarajan Sudha1, Nanjiang Shu1,2 and Arne Elofsson1

1. Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, SE‐171 21 Solna, Sweden.

2. Bioinformatics Short‐term Support and Infrastructure (BILS), Science for Life Laboratory, Solna, Sweden.

[sudha.govindarajan@scilifelab.se](mailto:sudha.govindarajan@scilifelab.se)

[nanjiang.shu@scilifelab.se](mailto:nanjiang.shu@scilifelab.se)

[arne@bioinfo.se](mailto:arne@bioinfo.se)

# Abstract

The CPA/AT transporter superfamily transports a variety of compounds, including ions, amino acids, bile acid and auxin. All known families in this superfamily consists of an inverted repeat structure, in most cases the N and C-terminal subdomains have identical (but inverted) topologies, but in some cases there exist one or more additional helices in one of the subdomains. Here, we show that the evolutionarily related proteins in this superfamily contain unprecedented changes in their topology. This identifies the existence of several novel families with either broken and reentrant helical transporters using an integrated topology annotation method. Using these annotations, we define a hierarchy of topology and evolutionary relatedness; (i) the CPA/AT Superfamily, (ii) Fold-type, (iii) Family and (iv) Subfamily. We identified three evolutionarily distinct fold-types based on the repeat units; the 5 helical broken fold-type, the 6 helical reentrant fold-type and the 7 helical broken fold-type. Each fold type can be made up of families of same/different topologies. Few families can be further made up of subfamilies with different topologies (Pfam families Na\_H\_Exchanger, SBF and Cons\_hypoth698). Topology variations are due to the addition/deletion of helices in the scaffold subdomain. We also identified subfamilies (cons\_hypoth698\_1 and cons\_hypoth698\_2) that have a truncated C terminal core subdomain. Families with a complete change in orientation were also observed both among the broken and reentrant types. Finally, we show that the transition between the broken and reentrant types is accompanied by the gain/loss of helices in the scaffold subdomain.

# Keywords

CPA/AT transporter, Repeats, Topology, Structure classification, Evolution

# Abbreviations

CPA/AT, monovalent cation:proton antiporter /anion transporter , TM , Transmembrane, NhaA, Na+/H+ antiporter from *E coli***,** Na+/H+ antiporter from *Thermus thermophilus*, ASBT, Apical sodium–bile acid transporter, citS, Citrate-sodium symporter,

# 

# Introduction

CPA/AT transporters transport a variety of ions, amino acids etc [[1–4]](https://paperpile.com/c/6Cfpun/xLiA+pAMf+rI7q+DL19), One of the best-studied sodium-proton exchanger, NapA is involved in the exchange of protons for sodium ions to maintain the pH and ion homeostasis. Due to their functional importance, they are ubiquitously present in all the three kingdoms of life [[3,5–7]](https://paperpile.com/c/6Cfpun/rI7q+PGMD+4HL7+FYqL). In humans, malfunctioning of transporters is associated with several diseases, including cardiac diseases, and inflammatory bowel disease. They serve as important drug targets. [[8,9]](https://paperpile.com/c/6Cfpun/FVbj+agTU). In plants, the CPA/AT transporters are involved in the transport of auxin that is crucial for various plant physiological processes [[10]](https://paperpile.com/c/6Cfpun/CEx1).

Structures are available only for NhaA (Pfam family: Na\_H\_antiporter\_1, PF06965)[[11]](https://paperpile.com/c/6Cfpun/kN93)**,** NapA (Pfam family: Na\_H\_Exchanger, PF00999) [[6]](https://paperpile.com/c/6Cfpun/4HL7), ASBT (Pfam family: SBF, PF01758) [[12]](https://paperpile.com/c/6Cfpun/PHIb) and citS (Pfam family: 2HCT, PF03390) [[13]](https://paperpile.com/c/6Cfpun/LVzV). All known structures consist of two inverted symmetric ***repeat units*** consisting of five to six transmembrane segments, see Figure 1a. The repeats are important to enable the different conformational states that are necessary for the transport mechanism [[14–17]](https://paperpile.com/c/6Cfpun/Pj0v+2s38+ch0B+0S1D). Further, the repeat units can be divided into two structurally distinct parts, the ***core*** and the ***scaffold*** ***subdomains,*** see Figure 1a. [*[6]*](https://paperpile.com/c/6Cfpun/4HL7). Structurally, the repeats intertwine in such a way that the two core subdomains come together and the two scaffold subdomains come together, see Figure 1b.

The scaffold subdomain is not directly involved in the transport function but it interacts with the core domain. The SBF transporter has 2TM (Transmembrane helix) scaffold subdomains (Figure 1c) while the other families of known structures (Na-H-exchangers, Na\_H\_antiporter\_1, and 2HCT) all have 3TM scaffold subdomains (Figure 1c). Structural alignment between the scaffold subdomains shows that the two C-terminal helices of the 3TM align to the 2TM subdomain [[6]](https://paperpile.com/c/6Cfpun/4HL7) (Figure 1c). The extra helix of the scaffold subdomain in Na-H-exchangers (PF00999) and 2HCT (PF03390) have been shown to be related to oligomerisation of the scaffold domain which is an important feature for their transport [[6]](https://paperpile.com/c/6Cfpun/4HL7).

The core domains generally consist of 6 transmembrane helices, with the middle helix of the core subdomains being broken, see Figure 1d. The broken helices contain a polar non-helical part in the centre of the membrane region capable of binding/transporting ions [[18]](https://paperpile.com/c/6Cfpun/KCRF). These helices play an important role in the function of the transporter. Three out of the four known structures of CPA/AT transporters contain core domains with broken helices.

Interestingly, the sodium-citrate symporter (2HCT, PF03390) does not contain a broken helix, instead, it contains a helical hairpin, also referred to as a reentrant helix, as the middle helix of both the core subdomain. The reentrant regions functionally resemble the broken helix in ion binding. However, structurally they are different. They do not pass through the membrane, see Figure 1d [[13,19]](https://paperpile.com/c/6Cfpun/cFTr+LVzV). A reentrant helix starts from one side of the membrane and then in the centre of the membrane bends back and forms a hairpin structure to end at the initial side again. Therefore, the topology of the following helices has to differ to accommodate the different location of the C-terminal end of the reentrant/broken helices. However, when the entire domains are superposed, the two reentrant helices form a very similar structure as the two broken helix in the other structures, Figure 1d. Both the reentrant and the broken helices are less hydrophobic than other TM-helices and both are capable of binding ions.

One common feature between the broken and the reentrant transporters are their elevator mode of transport of ions [[20]](https://paperpile.com/c/6Cfpun/qtaK). Here, the core domain traverses the membrane against a rigid immobile and dimerised scaffold domain [[6,20]](https://paperpile.com/c/6Cfpun/4HL7+qtaK). Recent studies have also pointed out the coordinated movement of both the core and scaffold domains within the single protein to be important for the transport activity [[21]](https://paperpile.com/c/6Cfpun/kgrP). Experimental studies have proposed this mechanism for Na\_H\_Exchanger and sodium-citrate symporter families. Dimeric states have also been shown for NhaA (Na\_H\_antiporter\_1) [[22]](https://paperpile.com/c/6Cfpun/5nLG) and human sodium bile acid symporter [[23]](https://paperpile.com/c/6Cfpun/N3FZ). But, It is still not clear if the elevator mechanism of transport is a conserved feature for the entire superfamily. The inverted repeats have been used to generate “Repeat swap models”, describing one conformation using a known structure of the transporter in the complementary conformation as a template [[17,24]](https://paperpile.com/c/6Cfpun/87YC+0S1D).

Several sequence- (Pfam, TCDB) and structure-based classification (OPM, ECOD, CATH) are available for CPA/AT transporter superfamily. In Pfam, the CPA/AT transporter superfamily is named as the “CPA\_AT clan” [[3]](https://paperpile.com/c/6Cfpun/rI7q). The CPA/AT clan (CL0064) (Pfam version 32) is composed of 13 Pfam families. It should be noted that the sodium-citrate symporter (2HCT, PF03390) does not belong to the CPA/AT clan in Pfam. In TCDB, [[25]](https://paperpile.com/c/6Cfpun/lk3U) some members of the Pfam CPA/AT clan are split into two superfamilies, the CPA superfamily and the BART superfamily [[26]](https://paperpile.com/c/6Cfpun/yGZP). Structure-based classification for this superfamily is limited only to families with known structure. OPM database [[27]](https://paperpile.com/c/6Cfpun/3Z1l) classifies all the four structures into the “Monovalent cation-proton antiporter superfamily”. ECOD [[28]](https://paperpile.com/c/6Cfpun/L5NE) classifies the broken and reentrant transporters into two different superfamilies. CATH [[29]](https://paperpile.com/c/6Cfpun/Jd35) classifies the three broken transporter structures into two different superfamilies. There is no classification provided by SCOPe [[30,31]](https://paperpile.com/c/6Cfpun/HSJA+tztR) for CPA/AT transporters. The comparison of different sequence and structure-based classification for CPA/AT superfamily are listed in detail in Table 1. These different categorisations show that there is a need to update and reclassify the relationship between CPA/AT transporters.

The topology of proteins with known structure varies a lot. The transporters are 10 (SBF), 12 (Na\_H\_Antiport\_1) and 13 (Na\_H\_Exchanger and 2HCT) helices long. The repeats can have five or six TM helices. Apart from variations of the core and scaffold domains, there also exist additional TM-helices in a few families that do not belong to the repeat unit [[6]](https://paperpile.com/c/6Cfpun/4HL7). However, a systematic study of the topological diversity of the CPA/AT superfamily is still largely unexplored. Providing accurate topology annotations for the families, without known structure would be useful for accurate modelling of these proteins.

In this study, we are using computational approaches to systematically annotate the topology, including the type of core domain (broken or reentrant), for all the families/subfamilies in the CPA/AT superfamily. We further classify families/subfamilies to have a unique topology. Sequence-based remote homology detection methods [[32]](https://paperpile.com/c/6Cfpun/W2Qg) and structure-based topology annotations lead to the identification of three “Fold types” in the CPA/AT transporter superfamily. Different fold types highlights the structural diversity within the superfamily not previously described. We believe that this study will help to better understand the structure, function and evolution of the CPA/AT superfamily.

# 

# 

# Results:

## 1.CPA/AT superfamily: An improved classification

Even though the Pfam clan (CPA\_AT), is generated using sensitive homology searches, they are not always updated. So, we updated and filtered the Pfam clan and have named it the “CPA/AT superfamily” and includes 12 Pfam families as discussed below:

### a. New members:

First, we identified all Pfam families that are homologous to any family in CPA\_AT clan (CL0064) and renamed this group as the “CPA/AT superfamily”. Several CPA/AT families showed strong relationships with two other Pfam families which are not part of the CPA\_AT clan according to Pfam. Therefore, Abrb (PF05145) and 2HCT (PF03390) families were added as new members to the CPA/AT superfamily.

### b. Lost members:

OAD\_beta, (PF03977) lys\_export (PF03956), sbt\_1 (PF05982) and *LrgB* (PF04172) do not show any significant similarity to other members of the CPA/AT clan. Therefore, they were excluded from the CPA/AT superfamily (see methods section for details). These families are distantly related and failed to pass our strict criteria as families in a CPA/AT superfamily using the remote homology searches. The strict criteria ensured that the topology annotations, alignments with other families are reliable so that evolutionary studies could be carried out.

### c. Broken and the reentrant transporters are related:

The most interesting finding is that our remote homology search showed that the broken and reentrant transporters are evolutionarily related and all belong to the CPA/AT superfamily. Even though 2HCT has been classified to belong to this superfamily using structures in the OPM database, in the current release of Pfam, 2HCT is not part of any Pfam clan. Further AbrB is (wrongly) classified to belong to the clan Membrane\_trans (CL0142).

## 

## 2. Topology annotation identifies new topologies in CPA/AT superfamily:

It is generally assumed that the topology is conserved within a family. However, our analysis shows that proteins in the SBF, Na\_H\_Exchanger and Cons\_hypoth698 families have two distinct topologies. It is quite interesting to note that all are eukaryotic containing families (Figure S1, S2, S3, S5 & S14). Therefore, these families were split into two subfamilies based on their topology, i.e. in total there are 15 families in the CPA/AT superfamily, see Table 1.

The topology has been identified experimentally for some families previously [[33]](https://paperpile.com/c/6Cfpun/vQuZR) [[34]](https://paperpile.com/c/6Cfpun/qm261) [[35]](https://paperpile.com/c/6Cfpun/QBDAd). Nevertheless,the integrated pipeline (Figure 2) helped to identify novel 9, 10, 12 and 14 helical proteins (Table 1, Figure 3, S1-S14) .

(https://github.com/gsudha/CPA-AT-superfamily/tree/master/Homology\_models)

Another interesting finding is the occurrence of new 10 and 12 TM helical proteins in Nin and Nout orientations. We identified 8 families/subfamilies of the broken type transporters and 7 families/subfamilies of the reentrant type of transporters based on helix-type (broken/reentrant) from homologous transporters and mean KR-bias values for broken and reentrant topology models (Table 2) (Figure 4) . The identification of several reentrant transporters in CPA/AT superfamily is novel. The G values clearly show the low hydrophobicity of the broken and reentrant helices in comparison to other transmembrane helices (Figure 3, Figures S1-S14).

The broken type of transporters exists in all three kingdoms of life (Figures S1-S14). The N- and C-core of all the broken transporters are made up of three helices. In contrast, the reentrant transporter families are mostly bacterial, the exception is DUF819 which has some members belonging to Eukaryotes (Viridiplantae). The N- and C-core of all the reentrant transporters are also made up of three helices, except for two of the families (Cons\_hypoth698\_1, Cons\_hypoth698\_2).

Our analysis is also a proof of concept to show that topologies can accurately be predicted, even for proteins with broken or reentrant helices when an integrated approach is used. The classification made by the KR bias calculations provides additional support to the broken/reentrant core classifications, see Figure 4, Table 2.

## 

## Identification of three distinct fold types in the CPA/AT superfamily: A new structural classification scheme:

Based on topology annotation using our methodology, MSA-MSA alignments and hierarchical clustering, we propose a new structure-based classification scheme for the CPA/AT superfamily.This classification is different from the classical Protein fold-superfamily-family hierarchy in SCOP [[31]](https://paperpile.com/c/6Cfpun/tztR) and shows a larger structural diversity than inmost other protein superfamilies (Table 1).

All families are related but clearly evolutionary separated into clusters. The evolutionarily related families in a cluster have a conserved C-repeat helices. Families with such conserved C-repeat units are newly defined as “Fold-type” in this work. The fold-types are namely, (1) 5 helical broken fold-type, (2) 7 helical broken fold-type and (3) 6 helical reentrant fold-type (Figure 5). The numbers (5, 7 and 6) in the fold types denote the number of helices within the C-terminal repeat, as the topology of the N-terminal repeat varies in some families. The multi-scale bootstrapping of the dendrogram shows that the clustering is statistically significant (Figure5, Figure S20).

We would expect the broken transporters (5 helical broken and 7 helical broken ) to be more similar than the reentrant transporters (6 helical reentrant fold-type). Therefore, it is quite intriguing to note that the 7 helical broken fold-type transporters are evolutionarily closer to the 6 helical reentrant fold type transporters than to the 5 helical broken fold-type, see Figure 5.

Cons\_hypoth698\_1 and Cons\_hypoth698\_2 are found to be only distantly related to the other 6 helical reentrant families (Figure 5). The excluded families (Oad\_beta, Sbt\_1, Lys\_export and LrgB) which were distant to the CPA/AT superfamily were also found to be distant to the 7 helical broken and 6 helical reentrant fold-type clusters (Figure 5).

Below we provide the detailed classification of CPA/AT superfamily (Table 1), (Figure 6). The superfamily can be classified based on the decreasing order structural and evolutionary relatedness into (1) Superfamily, (2) Fold-types, (3) Family and (4) Subfamily. Number of helices are denoted as 9H, 10H, 12H, 13H or 14H. Nin and Nout denote the inside and outside orientation of the first helix of the transporter.

### **1. CPA-AT superfamily**

### 1.1 5 helical broken fold-type

***1.1.1.1 SBF\_1 subfamily (PF01758): 10H-Nin topology***

SBF\_1 are Na+/bile acid co-transporters with known structure [[12,36]](https://paperpile.com/c/6Cfpun/F030+PHIb). They are found in both bacteria and eukaryotes. (Figure S5).

***1.1.1.2 SBF\_2 subfamily (PF01758): 9H-Nout topology***

SBF\_2 are Na+/bile acid co-transporters [[37]](https://paperpile.com/c/6Cfpun/Z9VV). Here, the N-terminal repeat has lost one helix in the N-terminal scaffold subdomain. Members are found in both bacteria and eukaryotes (Figure S5).

***1.1.2.1 SBFlike family (PF13593): 10H-Nin topology***

This family is similar to the SBF\_1 and SBF\_2 family and can be found in both bacteria and eukaryotes. see Figure S6 & 5a.

***1.1.3.1 KdgT family* (PF03812): *10H-Nin topology***

It is responsible for the uptake of ketodeoxyuronates [[38]](https://paperpile.com/c/6Cfpun/uDCl). The KdgT family is observed only in bacteria (Figure S8).

***1.1.4.1 Mem\_trans family (PF03547): 10H-Nout topology***

The mem\_trans family is involved in the transport of auxin [[10]](https://paperpile.com/c/6Cfpun/CEx1). This family has a different orientation compared to the other 5-broken transporters (Figure S7 & 5B). Members are observed in both bacteria and eukaryotes.

### 1.2 7 helical broken fold-type

***1.2.1.1 Na\_H\_antiporter family (PF06965): 12H-Nin topology***

These proteins function as a sodium-proton antiporter and have a known structure [[11]](https://paperpile.com/c/6Cfpun/kN93). Here, the N-repeat have lost two helices in the scaffold subdomain (Figure S4). Members are predominantly found in bacteria.

***1.2.2.1 NA\_H\_Exchanger\_1 subfamily (*PF00999): *13H-Nout topology***

This family function as a sodium-proton antiporter and have a known structure [[6]](https://paperpile.com/c/6Cfpun/4HL7). The N-terminal repeat has lost one helix in the scaffold subdomain (Figure S1). Members are seen in bacteria, archaea and eukaryotes. A 12 helical Na\_H\_exchanger (Nhe1) is known previously [[39]](https://paperpile.com/c/6Cfpun/NDAy) but we did not find this topological state to be predominant and therefore did not include it in our analysis.

***1.2.2.2 NA\_H\_Exchanger\_2 subfamily* (PF00999): *14H-Nin topology***

These proteins function as sodium-proton antiporters. These 14-helical proteinsare seen in bacterial and Eukaryotic (metazoan) species (Figure S2 and S3).

### 1.3 6 helical reentrant fold-type

Unless noted these families are unique to bacteria.

***1.3.1.1 Asp\_Al\_Ex family* (PF06826): *12H-Nout topology***

These proteins are aspartate-alanine antiporters [[40]](https://paperpile.com/c/6Cfpun/Gc4C)(Figure S12).

***1.3.2.1 Glt\_symporter family* (PF03616): *12H-Nout topology***

The Glt\_symporter function as sodium-dependent glutamate symporter [[41]](https://paperpile.com/c/6Cfpun/RyXc)(Figure S11).

***1.3.3.1 DUF819 family* (PF05684): *12H-Nout topology***

The function is unknown. It is the only family of reentrant transporters that is found in Eukaryotes, mainly in Viridiplantae (Figure 4).

***1.3.4.1 AbrB family* (PF05145): *12H-Nin topology***

ABrB is involved in the regulation of the protein aidB [[42]](https://paperpile.com/c/6Cfpun/YQRY). (Figure S13 and 5D) This family has an inverted topology compared to the other reentrant families. This change in orientation has been identified for the first time here.

***1.3.5.1 Cons\_hypoth698\_1 subfamily* (PF03601): *10H-Nin topology***

This family is known to be a bacterial inner membrane protein [[43]](https://paperpile.com/c/6Cfpun/NWYL). But the function is not known. Two helices are lost in the C-core subdomain thus showing a truncated 6 helical reentrant fold-type (Figure S14).

**1.3.5.2 Cons\_hypoth698\_2 subfamily (PF03601): *12H-Nin topology***

The function is unknown. Cons2\_hypoth698 has two helices lost in C-core subdomain thus showing a truncated fold-type. Also, the N-terminal repeat has two extra helices in the N-scaffold subdomain (Figure S14).

***1.3.6.1 2HCT family* (PF03390): *13H-Nin topology***

The sodium-dependent citrate symporter has a known structure [[44]](https://paperpile.com/c/6Cfpun/ZsvX). Here, the N-terminal scaffold subdomain has an extra helix.(Figure S9).

## 3. Types of topology variations.

The topology annotation and the MSA-MSA alignments helped us to analyse the variations in topology. Types of topology variations observed between subfamilies, families and fold-types are discussed below. For details see supplemental data,, which provides pairwise alignments between the representative sequence of families.

### Topology variations between subfamilies

**1. Gain/loss of helices in the scaffold subdomain:**

N-terminal scaffold subdomain could have a gain/loss of helices (Figure 7a).

### Topology variations between families

**1. Gain/loss of helices in the scaffold subdomain:**

N-terminal scaffold subdomain could have a gain/loss of helices (Figure 7b).

**2. Gain/loss of helices in the core subdomain:**

These are seen in the subfamilies Cons\_hypoth698\_1 and Cons\_hypoth698\_2. Here, two helices have been lost in the C-terminal core subdomain (Figure 7c). The functional implication of the truncated core domain is not known.

**3. Change in orientation:**

There are cases where the change in orientation have occurred without any addition/loss of helices as confirmed by the KR bias plots (Figure 4b & 4d). To the best of our knowledge, these are identified for the first time here both in the broken and reentrant type of transporters (Figure 7d).

### Topology variations between fold-types:

**1. Gain/loss of helices in the scaffold subdomain:**

Variations are always seen in the C-terminal scaffold subdomain but may or may not have changed in the N-terminal scaffold subdomain (Figure 8a). Structural superimpositions clearly show that the gain of helices is at the periphery of the structure, see Figure 8b.

**2. Broken - reentrant switch of core subdomain:**

One of the most interesting topology variations is the structural transition between the broken and reentrant helices. The switch between the reentrant and broken helices is always accompanied by a change in the scaffold helices. Variations in the helices of C-terminal scaffold subdomain always occur, while variations in the N-scaffold may or may not occur (Figure 8c). Structure alignments between the broken and the reentrant transporters show that the structures are similar in spite of the change in orientation following the broken/reentrant helix. Further, the gain/loss of scaffold helices is necessary during this structural transition between the broken and reentrant type to maintain the inverted nature of the repeat units (Figure 8d).

# Discussion

In this study, we carried out a comprehensive annotation of the CPA/AT superfamily by studying all the available sequences using an integrated pipeline. The presence of broken/reentrant helices makes it difficult to base topology prediction purely on automatic tools. Instead, we found that the best way to identify broken and reentrant helix prediction was by aligning the families to known structures and then use the KR bias plots to confirm the classifications. Broken and reentrant helices (Figure 1d) have been studied in other transmembrane proteins before [[18,45,46]](https://paperpile.com/c/6Cfpun/mOQK+XU0N+KCRF). Although some attempts to identify reentrant regions have been tried [[47]](https://paperpile.com/c/6Cfpun/WqxA), their success is marginal at most, i.e. without structural knowledge it is difficult to identify reentrant regions (and broken helices) resulting in erroneous topology predictions. This is due to their polar nature, marginal hydrophobicity, structural variation and under-representation in nature**.**

Further, the KR bias analysis confirmed the broken or the reentrant classification of transporters and the correctness of the topologies. The KR bias calculations also clearly show the change in orientation in the Mem\_trans and AbrB families. These inverted topologies were also supported by topology predictions. However, they are not evident from the alignment to known structures (Figure S13), i.e. correct topologies can only be obtained by a combination of approaches.

Even though one can predict the topology by just aligning to families of known structure it is still essential to predict the topology by topology programs to identify TM-helices that do not align with helices of known structure (Figure S13). This is especially important when the unknown topology is longer than the known topology or when there are additional helices in the N- or C-terminal ends (Figures S2, S3, S7 & S14).

Using the information from the crystal structure, we have annotated core, scaffold subdomain and inverted repeat units in all the family.

We have identified three families that showed more than one topology. It is appropriate to split these families into subfamilies based on their topology. Presence of eukaryotes in these three families pinpoints that topology variation could be more prevalent in eukaryotes than in bacteria or archaea. Previous work suggests that alternative splicing can result in variations of topology. Functional isoforms exist in the Na\_H\_Exchanger and SBF families [[48–50]](https://paperpile.com/c/6Cfpun/phga+XGDk+UC0V). Therefore, alternative splicing could be one explanation for some of the variations that are observed among these transporters, see Figure S1, S2, S3, S5 & S14.

Considering that these families are sequentially divergent with a lot of gain/loss of helices, multiple sequence alignments containing all the families is not reliable. Therefore, conventional phylogenetic analysis based on multiple sequence alignment is not possible and we resorted to generating dendrogram based on E-values from sensitive HMM-HMM pairwise alignments.

Variations of topology in the scaffold domain contribute to both structural and functional advantages for the transporters. In the case of transporters with three scaffold helices, the first helix helps to dimerise the scaffold domain (Figure 1c). The dimerisation of the scaffold domain is important for the elevator mode of transport [[6,51–53]](https://paperpile.com/c/6Cfpun/2Meq+nBsY+mwCU+4HL7). Variations in the scaffold domain could also provide variations in the scaffold-core interface, that is important for the transport of a wide range of molecules. Crystal structures have shown that families within the same fold-type, e.g. (Na\_H\_exchanger & Na\_H\_antiport\_1 families), show similar core-scaffold interfaces, while families belonging to different fold-types tend to have dissimilar core-scaffold interfaces e.g (Na\_H\_Exchanger & 2HCT families). It is already known that extreme asymmetry in domains results in functional specialisation of elevator type transporters [[20]](https://paperpile.com/c/6Cfpun/qtaK). Evolution of the addition of helices to repeats could also increase asymmetry.

Experiments have also shown that proteins belonging to the (inverted) Mem\_trans family are localised in the plasma membrane for the release of auxin [[54]](https://paperpile.com/c/6Cfpun/I9Sk). The change in orientation help to release the plant hormone auxin out of the cell in the Mem\_trans family. Change in orientation in reentrant transporter AbrB is also newly reported from our analysis. Further, this is the first time a CPA/AT transporter with truncated C-terminal core subdomain has been identified. It would be interesting to know if the transporter is functional and to determine its detailed structure. It is also not clear what are the functional specialities provided by broken and reentrant transporters. Anyhow, the new topological states identified from our annotation clearly show the extensive variations in topology in this superfamily.

# 

# Conclusions

Our study has led to the identification of broken and reentrant transporter families in the CPA/AT superfamily. Further, we show that the families in the CPA/AT superfamily have three evolutionarily distinct fold-types and that within each fold-type families of different families and subfamilies exist. We propose a new structural classification of the CPA/AT superfamily into Superfamily, Fold-types, Family and Subfamily. This classification could potentially be used for other transmembrane transporters with repeat units. We have also analysed the different types of topology variations. These can be categorized into four main types the addition/deletion of helices in (i) scaffold and (ii) core domains as well as (iii) the inversion of the orientation of the entire protein, and (iv) the broken-reentrant transition. These topology variations could be explored further in relation to their structure and function of members in the CPA/AT superfamily.

# Materials and methods

### 1) Annotations of the topology for families/subfamilies in CPA/AT transporters:

Our strategy to annotate topology and reclassify the Pfam CPA/AT clan or superfamily into families/subfamilies involves the following six steps, see Figure 2:

1. Classification of Pfam families with unique topology and assignment of initial topologies.
2. Improved classification of CPA/AT superfamily
3. Identification of broken/reentrant type transporters from proteins with known structures.
4. Obtaining final topology.
5. Identification of core, scaffold subdomains and repeat units from the known structure.
6. Validation of Broken/reentrant type transporters by the positive inside rule [[55]](https://paperpile.com/c/6Cfpun/QqUl).

### i) Classification of Pfam families with unique topology and assignment of initial topologies:

First, the sequences of all the 13 Pfam families (CPA\_AT clan) belonging to reference proteomes from UniProt [[55,56]](https://paperpile.com/c/6Cfpun/QqUl+qURa) were used. Fragment sequences were ignored, and only sequences with ≥ 75% Pfam domain coverage were considered.

Sequence clustering was carried out using blastclust [[57]](https://paperpile.com/c/6Cfpun/8vwK) in two steps. Stage1: Clustering at 90% sequence identity was used to remove identical, highly similar proteins or point mutants. Stage2: Clusters obtained from stage 1 was clustered again at 30% sequence identity. Stage2 clustering was carried out to obtain clusters that could generate multiple sequence alignment with few gaps. For most families, a low number (1-3) of clusters were obtained but for a few families with large sequence diversity, including Na\_H\_exchanger and Mem\_trans, many more clusters were found (Table S1). Finally, all sequences in the clusters were aligned to get multiple sequence alignment using Clustal Omega [[58]](https://paperpile.com/c/6Cfpun/cT4y).

Next, topologies for all the members of the family were predicted using TOPCONS2 [[59]](https://paperpile.com/c/6Cfpun/GKvo). Predicted topologies were mapped onto the multiple sequence alignments to get a “Multiple topology alignment”.

Next, phylogenetic trees were generated from the MSAs using Fasttree [[60]](https://paperpile.com/c/6Cfpun/HBse). The order of the sequences in the multiple topology alignment was changed/reordered according to the branching order of the phylogenetic tree. This is referred to as the “Reordered topology alignment” (Figure 2).

The initial topology for a family was inferred by visually inspecting the reordered topology alignment using the following criteria: A transmembrane helix is assigned for the family when the transmembrane helix from most of the sequences in the cluster are aligned. When a transmembrane helix (TM) is aligned to a loop region then also it is corrected to be a transmembrane helix (TM). This is because, In the case of helices with low hydrophobicity, it can be missed by topology prediction programs. These regions are labelled as loops. When we observe a majority of other sequences annotated as TM in the reordered topology alignment, we assume that the loop is a wrong annotation and relabel it as TM. Cases of TM aligned to the gap region were also noted. Signal peptide was annotated if the dominant majority of the aligned sequences were signal peptide (Signal peptide prediction obtained from TOPCONS2). If the dominant majority is TM, then the signal predictions were considered as an error. Finally, all the MSA clusters were thoroughly checked and one MSA cluster representing each unique topology was selected (Table S1). This was named as a “seed family MSA”.

https://github.com/gsudha/CPA-AT-superfamily/tree/master/Seed\_MSA

In all families there exist a few proteins that differ in predicted topology from the majority. This might be due to sequence errors or just errors in the topology prediction and are therefore ignored. Further, we ignored those clusters which had an additional signal peptide in the Na\_H\_Exchanger, as this does not contribute to a different topology. In the Na\_H\_Exchanger and Mem\_trans families, there were clusters where the TM domain belonging to CPA/AT was fused to another TM domain. These clusters were also ignored since it did not contribute to a different topology within the domain. However, for families with systematic different topologies, the family definitions were reclassified and used for further analysis. These are clearly shown in a reordered topology alignments with different topology in phylogenetically separated clusters, see Figure S5. The MSA is then split based on the topology and the families are split into subfamilies that each have a unique topology (Table S1).

Despite these family classifications, the topology predictions sometimes consistently missed predicting either one or both broken/reentrant helices, see Figure 3. This was examined for those families with known structure/topology, see Figure S9, where the topology prediction fails to predict one or both of the reentrant helices. Therefore, the topologies annotated using the reordered topology alignments were labelled as “Initial topologies”, see Figures 3, S1-S14. To assign correct topology to all families within the CPA/AT superfamily we, therefore, had to use additional methods, as described below. Finally, sequences of all families were assigned into Archaea, Bacteria and Eukaryotes based on the taxonomic lineage to its members.

### ii) Improved classification of CPA/AT superfamily:

Since the definition of the Pfam clan is not completely agreeing with other classification schemes, we used HHsearch version 3.2.0 [[32]](https://paperpile.com/c/6Cfpun/W2Qg) to find possible evolutionary relationships between the family MSA of the CPA/AT clan and other families in Pfam-A\_v32.0 [[3]](https://paperpile.com/c/6Cfpun/rI7q). The evolutionary relationships between new Pfam families and families of the CPA/AT clan were identified. Thereafter, we filtered and renamed the CPA/AT clan as the “CPA/AT superfamily” which only includes Pfam families that have an E-value better than 10-4 with at least one family of the clan (except itself). New families that are added to the CPA/AT superfamily went through the previous step to annotate their topology and check if they have topology variations within the family. A representative sequence and a family MSA was generated for all the families/subfamilies in CPA/AT superfamily as described below:

#### Generation of a representative sequence for each family:

A representative sequence for each family was selected by searching the seed family MSA against the sequence database (UniProtKB) [[61]](https://paperpile.com/c/6Cfpun/mzv9) with an E-value of 10-5 and a single iteration using HMMsearch [[62]](https://paperpile.com/c/6Cfpun/6JWW). The top hit which showed maximum coverage of the TM region with respect to full length and had a single Pfam domain was chosen as the representative sequence. In the families with known structure, the sequences of the structure were considered as the representative sequence. See Table 2 for the UniProt IDs of representative sequences for each family (https://github.com/gsudha/CPA-AT-superfamily/tree/master/representative\_seq).

#### Generation of family MSA:

The representative sequence of each family was searched against Uniclust30 [[56]](https://paperpile.com/c/6Cfpun/qURa) using the HHblits program [[63]](https://paperpile.com/c/6Cfpun/VZko) with default settings. The MSA could be more sensitive when aligning remotely homologous families than the seed family MSA which often has only a few sequences.

(https://github.com/gsudha/CPA-AT-superfamily/Family\_MSA).

### iii) Identification of Broken/reentrant type transporters from known structures:

The next step was to identify the missing broken/reentrant helices if any. The initial topology of the family identified earlier was mapped to the representative sequence of that family. An HMM generated from the family MSA were searched against the PDBmmCIF70\_22\_May database using HHsearch [[32]](https://paperpile.com/c/6Cfpun/W2Qg) to compare the “initial topology” of the family with the topology derived from the crystal structure (Figure 2 & 3) (Supplemental data provides alignments between the representative sequence and best hit from PDB).

We also generated homology models of the representative sequence based on the aligned region and the best structural template using the HHpred server [[64]](https://paperpile.com/c/6Cfpun/uT1V).

(https://github.com/gsudha/CPA-AT-superfamily/tree/master/Homology\_models)

Predicted topology and the topology from the crystal structure were mapped onto the pairwise MSA-MSA alignment to obtain a topology alignment. Transmembrane helices were considered to be aligned to transmembrane region, gaps, loops or signal peptide. Transmembrane helices were considered to be aligned when at least 5 residues of both helices are aligned, as used before [[65]](https://paperpile.com/c/6Cfpun/IYui). If one TM helix shares less than five residues with the corresponding TM helix, it is classified into one of the following types; TM helix aligned to gap regions, TM helix aligned to inside/outside loops, TM helix aligned to signal peptide. The type is chosen based on the dominating composition in the segment of the TM helix that is aligned. Missing helices in the representative sequence was inferred when the TM helix in the topology with known structure was aligned to loops in the representative sequence. The aligned sequence region is then labelled as a TM region. Since we have structural templates for both broken and reentrant type transporter, the type of transporter for the family is classified based on the type of transporter with a known structure which is the best hit (lowest E-value).

### iv) Obtaining the final topology:

Based on the classification of broken/reentrant type, missing helices were added and the orientations were corrected. Now another type of topology prediction error was observed in the cons1\_hypoth698 and cons2\_hypoth698 families. Here, TOPCONS2 predicted a loop region as a transmembrane helix, see Figure S14. This was identified as the TM helix aligned to a loop region of the family with a known structure. We corrected this error in the final topology. 𝞓G values [[66]](https://paperpile.com/c/6Cfpun/YGpu) also indicated that the loop region could be wrongly predicted as a transmembrane helix by TOPCONS2 was not very hydrophobic, see Figure S14. The final topology was then inferred for all families, Figures 3, S1-S14. Here, topology alignments along with the 𝞓G plot summarising the topology, subdomain and repeat annotation were carried out using python scripts. Final topologies of the representative sequence for each family of CPA/AT superfamily is provided.

(https://github.com/gsudha/CPA-AT-superfamily/tree/master/Topology\_representative\_seq)

### v) Identification of core, scaffold subdomains and inverted repeat units from the known structures:

Annotations of scaffold subdomain, core subdomain and additional helices (helices that are not part of the inverted repeat unit) were taken from the literature of the four Pfam families with known structure. Annotation of subdomain and additional helices are transferred from the family with a known structure to the family with an unknown structure based on the definition of aligned TM helices described in previous section (Figure 2 & 3). Here, the N-Core and N-Scaffold subdomain together form the N-repeat and the C-core and C-scaffold subdomains form the C-repeat, see Figure 1a and 1b).

### vi) Validation of Broken/Reentrant type of transporters using the positive inside rule:

Since structures of both broken and reentrant transporters were available, we extrapolated the broken or reentrant type of transporter from the best structural template. However, we wanted to confirm and support the findings of the transporter type (broken/reentrant) also using the positive inside rule via KR bias calculations [[55]](https://paperpile.com/c/6Cfpun/QqUl).

It is well known that the inside loops have an enrichment of positive residues than the outside loops of a transmembrane protein. The positively charged residues Arg (R) and Lys (K) are found to be four times less prevalent in outside or periplasmic loops than the inside or cytosolic loops [[55,67,68]](https://paperpile.com/c/6Cfpun/QqUl+9b44+iads). Therefore, KR bias can be used to identify the orientation of the protein. Here, the number of Lys and Arg residues were counted using the two criteria for the KR bias calculations. (1) Loop lengths of 25 were considered beyond the TM helices on each side. (2) A total of 10 residues inside the TM helix (from the two ends of the TM helix) were also considered as this has been shown to contribute to the positive inside rule [[69]](https://paperpile.com/c/6Cfpun/a6Lp).

KR bias = Σ(K+R)in - Σ(K+R)out

The KR bias is calculated using the family MSA. The final topology of the representative sequence is used to make one broken topology and one reentrant topology. This topology is then extrapolated for all the other sequences of the family, see Figure 2. In the broken type, the N- and C-terminal ends of the helix are on the opposite sides of the membrane (i-H1-o-H2-i-H3-o) or (o-H1-i-H2-o-H3-i) while in the second they are on the same side (i-H1-o-H2-o-H3-i) or (o-H1-i-H2-i-H3-o). If the mean KR bias value is higher for broken, they are labelled as broken and if not they are labelled as reentrant. Classification of broken/reentrant type fails when homologous structures of broken and reentrant type are absent or only one of them are available. In such cases, classification of broken/reentrant type using KR bias can be used solely on the full-length protein.

### 2) Hierarchical clustering of the full-length transporters:

The evolutionary relationships between the family/subfamily MSA of the CPA\_AT clan and other families in Pfam-A\_v32.0 that was carried out in previous steps by HHsearch were obtained. The evolutionary relationship between the full-length transporters was assessed by their E-values. If bi-directional pairs of query-hit are obtained, the pair containing the lowest E-value is obtained. Pairs that were not picked by HHsearch were given an E-value higher than the highest E-value obtained for all pairs compared. Hierarchical clustering of the log10(E-values) was carried out using Hclust in the Heatmap2 program in R using the average cluster method and correlations as the distance. Multiscale Bootstrapping of the hierarchical cluster with 10,000 bootstraps was carried out using the Pvclust program [[70]](https://paperpile.com/c/6Cfpun/5n5nA) in R. AU (Approximately Unbiased) *p*-value and BP (Bootstrap Probability) value are the two types of p-value provided. We considered the AU *p*-value that is computed by multiscale bootstrap resampling as it is a better approximation to unbiased *p*-value than BP value that is computed by normal bootstrap resampling [[70]](https://paperpile.com/c/6Cfpun/5n5nA).

### 3) Generation of topology alignments between families:

MSA-MSA alignment between families obtained in the previous step is converted into pairwise topology alignments. The topology alignment figure was generated using python scripts.

### 4)Structural superposition:

Structure superpositions of the pairs of transporters were carried out separately for the core domain and scaffold domain. Structure superpositions were carried out using TMalign [[71]](https://paperpile.com/c/6Cfpun/0IpL) and visualised using PyMol [[71,72]](https://paperpile.com/c/6Cfpun/0IpL+S9gU).

# Acknowledgements

We thank Claudio Bassot and David Drew for valuable input to this manuscript. We thank the Swedish National Infrastructure for Computing for providing computational resources. This work was supported by a grant VR-NT-2016-03798 from the Swedish National Research Council and a grant from the Kunt and Alice Wallenberg foundation.

# Figure captions

**Figure 1: General structure of CPA/AT transporters:**

The figure shows the sodium bile acid symporter (PDB id:4n7w). (a) They have a 5 transmembrane helical inverted repeat. N-Repeat is followed by C-Repeat (In inverted orientation) in sequence space. Each repeat unit is composed of a scaffold subdomain and a core subdomain. The scaffold subdomains are coloured in brown. The core subdomains are coloured in dark purple respectively. The inside and outside loops are coloured light grey. The arrangements of subdomains in the sequence are shown as a cartoon. The repeats are shaded in grey. (b) In structure space, both the N-scaffold and C-scaffold subdomain come together to form a scaffold domain. Similarly, the N-core and C-core subdomain come together to form the core domain. The lipid bilayer is coloured red and blue to denote outside and inside respectively. (c) Scaffold subdomain can be 2 or 3 helices long. (d) Core domain can be broken type (coloured in dark pink) or reentrant type (Coloured in green). The core subdomains of both types look structurally different. But, they look similar when the entire domain is considered (Highlighted in dark pink and green). The broken helix highlighted as dark pink crosses the transmembrane helix, while the reentrant helix highlighted as dark green bends back to form a hairpin-like structure.

**Figure 2: An Integrated pipeline to annotate the CPA/AT transporters:**

The steps involved in annotation of topology, subdomains repeats, classification of CPA/AT superfamily into families/subfamilies are numbered.

**Figure 3: Topology and subdomain annotation for DUF819 family:**

The topology alignment is reordered according to the phylogenetic tree to obtain a reordered topology alignment. The figure shows a reordered topology alignment. The TM helices (in-out) are coloured dark red or dark grey while TM helices (out-in) are coloured light red or light grey. Reentrant helices (in-in) are coloured yellow and (out-out) are coloured blue. The inside loops are coloured yellow. The outside loops are coloured blue. The vertical bar is coloured based on the taxonomy of the sequences (Bacteria: Purple, Archaea:Dark blue Eukaryotes: Green). The signal peptide is coloured black. The TM helices are numbered. The initial topology is inferred from the reordered topology alignment (See materials and methods section). The representative sequence is aligned with family with known structure/topology (PDB id provided) with the lowest E-value. Based on the alignment between initial topology and known topology, the final topology is corrected for missing helices, errors in topology predictions and correction of orientation. The annotations of core and scaffold subdomain in a family with known topology are extrapolated to the aligned helices. Helices belonging to scaffold subdomains and reentrant core subdomains are shown as brown, and green boxes. Repeats one and two are shown as black trapezoids. ΔG values are obtained for the representative sequence and are plotted to the aligned residues in the representative sequence.

**Figure 4: KR bias calculations to validate the broken or reentrant type of CPA/AT transporters:**

The final topology obtained for the family/subfamily is used to fictitiously generate a broken topology and reentrant topology. Cartoon representation of broken topology is coloured dark pink (7 helical broken fold-type) and dark purple (5 helical broken fold-type) while the reentrant topology (6 helical reentrant fold-type) is coloured green. The in-out, out-in, in-in, out-out helices are coloured dark grey, light grey, light blue, light yellow respectively. KR bias values for broken topology and reentrant topology is shown in x and y-axis of a 2D density scatter plot. KR bias plots and two possible topology cartoon models are provided for the following types: (a) Sbflike: Broken transporter (Nin ) (b) Mem\_trans: Broken (Nout) (C) DUF819: Reentrant (Nout ) (d) AbrB: Reentrant (Nin)

**Figure 5: Heatmap showing the evolutionary relationship between families of CPA/AT superfamily:**

Hierarchical clustering using E-values obtained from MSA-MSA alignment of family MSA versus MSAs in the Pfam database are shown. Families belonging to 5 helical broken fold-type, 6 helical reentrant fold-type and 7 helical broken fold-type are boxed. Cartoon representation of the broken transporter is coloured dark pink (7 helical broken fold-type) and dark purple (5 helical broken fold-type) while the reentrant transporter is coloured green (6 helical reentrant fold-type) for all the figures. The Log10 of E-values were converted to Z scores and scaled column wise in the heatmap. The in-out, out-in, in-in, out-out helices are coloured dark grey, light grey, light blue, light yellow respectively. The clusters with *p*-value *greater than or equal to p=0.95 (\*\*\*)* is represented as asterisks.

**Figure 6: Structure based classification in CPA/AT transporter superfamily:**

The structure based hierarchy in decreasing order of evolutionary relatedness is provided.The 4 levels are (1)Superfamily (2)Fold-types (3) Family and (4) Subfamily. The CPA/AT superfamily has three evolutionarily distinct fold types, 5 helical broken fold-type (Dark purple), 6 helical reentrant fold-type (Green) and 7 helical broken fold-type (Dark pink). Each fold-type can have more than one family and similarly some families can be split into subfamilies due to more than one unique topology. All the family, subfamily name and their topology are listed. The in-out, out-in, in-in, out-out helices are coloured dark grey, light grey, light blue, light yellow respectively.The inside loops are coloured light yellow. The outside loops are coloured blue. The inverted Nand C repeat units are boxed as solid and hashed lines respectively.

**Figure 7: Variation of topology between subfamilies and families:**

Helices belonging to scaffold, broken core and reentrant core subdomains are coloured brown, dark pink (7 helical broken fold-type) or dark purple (5 helical broken fold-type) and green (6 helical reentrant fold-type) respectively. The inside and the outside loops are coloured yellow and blue respectively. (a) Examples of topology variation between subfamilies. (b) (c) & (d) Examples of topology variations between families.

**Figure 8: Variation of topology between fold-types:**

Helices belonging to scaffold, broken core and reentrant core subdomains are coloured brown, dark pink (7 helical broken fold-type), dark purple (5 helical broken fold-type) and green (6 helical reentrant fold-type) respectively. The inside and the outside loops are coloured yellow and blue respectively.(a) Example of topology variations between fold-types (b) Structure superimposition of transporters between 5 helical broken fold-type and 7 helical broken fold-type is shown (PDB id: 4n7w, 4bwz). Gain of helices in one of the broken transporter is highlighted in brown.(c) Example of topology variations between fold-types (d) Structural superimposition of transporters between 7 helical broken fold-type nad 6 helical reentrant fold-type is shown.The broken (coloured dark pink) and reentrant helices (blue and yellow) are highlighted. Also, the gain of a helix in the broken transporter is highlighted in brown (PDB id: 5a1s, 4bwz).

# 

# 

# 

# Tables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Current classification** | | | **Pfam** | **TCDB** | **OPM** | **ECOD** | **CATH** |
| **Family** | | **Subdomain and Repeat annotation (N,C)** |  |  |  |  |  |
| 1.1 5 helical broken Fold-type | | | | | | | |
| *1.1.1.1 SBF\_1 family*  10H-Nin topology  PDB id: 4n7w | | S:1-2, 6-7  C:3-5, 8-10  R:1-5, 6-10 | **(CPA\_AT)**  SBF  (PF01758) | **(BART superfamily)**  The Bile Acid:Na+ Symporter -BASS Family, Arsenical Resistance-3 (ACR3) Family | **(Monovalent cation-proton antiporter)**  Bile Acid:Na+ Symporter (BASS) Family | **(Cation-proton antiporter)**  SBF | **(1.20.1530.20)** |
| *1.1.1.2 SBF\_2 family*  9H-Nout topology | | S:1, 5-6  C:2-4, 7-9  R: 1-4, 5-9 | **(CPA\_AT)**  SBF (PF01758) |  |  |  |  |
| *1.1.2.1 SBFlike family*  10H-Nin topology | | S:1-2, 6-7  C:3-5, 8-10  R:1-5, 6-10 | **(CPA\_AT)**  SBF\_like (PF13593) | **(BART superfamily)**  The Bile Acid:Na+ Symporter -BASS Family |  |  |  |
| *1.1.3.1 KdgT family*  10H-Nin topology | | S:1-2, 6-7  C:3-5, 8-10  R:1-5, 6-10 | **(CPA\_AT)**  KdgT (PF03547) | **(BART superfamily)**  The 2-Keto-3-Deoxygluconate Transporter (KdgT) Family |  |  |  |
| *1.1.4.1 Mem\_trans family*  10H-Nout topology | | S:1-2, 6-7  C:3-5, 8-10  R:1-5, 6-10 | **(CPA\_AT)**  Mem\_trans  (PF03547) | **(BART superfamily)**  The Auxin Efflux Carrier (AEC) Family |  |  |  |
| 1.2 7 helical broken Fold-type | | | | | | | |
| *1.2.1.1 Na\_H\_antiport\_1 family*  12H-Nin topology  PDB: 1zcd | | S:1-2, 6-9  C:3-5, 10-12  R:1-5, 6-12 | **(CPA\_AT)**  Na\_H\_antiport\_1(PF06965) | **()**  The NhaA Na+:H+Antiporter (NhaA) Family | **(Monovalent cation-proton antiporter)**  Sodium-proton antiporter NhaA | **(Cation-proton antiporter)**  Na\_H\_antiport\_1 | **(1.20.1530.10)** |
| *1.2.2.1 NA\_H\_Exchanger\_1 family*  13H-Nout topology  PDB id: 4bwz | | S:1-3,7-10  C:4-6, 11-13  R:1-6, 7-13 | **(CPA\_AT)**  Na\_H\_Exchanger (PF00999) | **(CPA Superfamily)**  The Monovalent Cation:Proton Antiporter-1 (CPA1) Family | **(Monovalent cation-proton antiporter)**  Sodium-proton antiporter\_1 | **(Cation-proton antiporter)**  Na\_H\_Exchanger | **(1.20.1530.20)** |
| *1.2.2.2 NA\_H\_Exchanger\_2 family*  14H-Nin topology | | S:1-4, 8-11  C:5-7, 12-14  R:1-7, 8-14 | **CPA\_AT)**  Na\_H\_Exchanger (PF00999) |  |  |  |  |
| 1.3 6 helical reentrant Fold-type | | | | | | | |
| 1.*3.1.1 Asp\_Al\_Ex family*  12H-Nout topology | | S:1-3, 7-9  C:4-6, 10-12  R:1-6, 7-12 | **(CPA\_AT)**  Asp\_Al\_Exchanger (PF06826) | **(CPA superfamily)**  The Aspartate:Alanine Exchanger (AAEx) Family |  |  |  |
| *1.3.2.1 Glt\_symporter family*  12H-Nout topology | | S:1-3, 7-9  C:4-6, 10-12  R:1-6, 7-12 | **(CPA\_AT)**  Glt\_symporter (PF03616) | **(CPA superfamily)**  The Glutamate:Na+ Symporter (ESS) Family |  |  |  |
| *1.3.3.1 DUF819 family*  12H-Nout topology | | S:1-3, 7-9  C:4-6, 10-12  R:1-6, 7-12 | **(CPA\_AT)**  DUF819  (PF05684) | **(CPA superfamily)**  The Putative Sulfate Exporter (PSE) Family |  |  |  |
| *1.3.4.1 AbrB family*  12H-Nin topology | | S:1-3, 7-9  C:4-6, 10-12  R:1-6, 7-12 | **(Membrane\_trans)**  AbrB (PF05145) |  |  |  |  |
| *1.3.5.1 Cons\_hypoth698\_1 family*  10H-Nin topology | | S:1-3,7-9  C:4-6,10  R:1-6, 7-10 | **(CPA\_AT)**  Cons\_Hypoth698 (PF03601) | **(CPA superfamily)**  The Putative Sulfate Exporter (PSE) Family |  |  |  |
| 1.3.5.2 Cons\_hypoth698\_2 family  12H-Nin topology | | S:1-5,9-11  C:6-8,12  R:1-8, 9-12 | **(CPA\_AT)**  Cons\_Hypoth698 (PF03601) |  |  |  |  |
| *1.3.6.1 2HCT family*  13H-Nin topology  PDB id: 5a1s | | S:1-4, 8-10  C:5-7, 11-13  R:1-7, 8-13 | **()**2HCT  (PF03390) | **()** The 2-Hydroxycarboxylate Transporter (2-HCT) Family | **(Monovalent cation-proton antiporter)**  Hydroxycarboxylate transporter | **(Sodium dependent citrate symport)**  2HCT |  |

**Table 1: Annotation of topology and subdomains for families/subfamilies of CPA/AT superfamily:** The orientation (Nin , Nout) is with respect to the first helix of the protein. Scaffold subdomain, Core subdomain and Repeat annotations are labelled as S,C and R respectively. Comparison of classification of CPA/AT transporter superfamily from different databases are shown.The superfamily are shown in bold . This is followed by the family within the superfamily. It is left blank when the family/superfamilies or both are not known.

# 

|  |  |  |  |
| --- | --- | --- | --- |
| **Family** | **Broken**  **structure template** | **Reentrant structure template** | **Mean KRbias**  **(Broken/ reentrant models)** |
| 1.Na\_H\_antiport\_1:  PF06965  (P13738) | Crystal structure  **1ZCD (2.6E-62)** | 5A1S(0.00046) | **18.47**/-3.10 |
| 2. Na\_H\_Exchanger\_1:  PF009999  (Q9UZ55) | Crystal structure  **4CZ8(2.1E-42)** | 5A1S (5.3E-07) | **-14.73**/3.46 |
| 3. Na\_H\_Exchanger\_2:  PF009999  (M3YCW9) | **4CZB (8.4E-34)** | 5A1S(5.1E-08) | **14.23**/-2.93 |
| 4. SBF\_1:  PF01758  (C4ST46) | Crystal structure  **4N7W(1.1E-44)** | 5A1S(7.1) | **11.38**/1.63 |
| 5. SBF\_2:  PF01758  (H2Y8C2) | **4N7W(3.3E-31)** | 5A1S(5.2) | **-10.72**/0.72 |
| 6. SBFlike:  PF13593  (A0A077PBM8) | **4N7W(2.4E-36)** | 5A1S(0.39) | **14.04**/-0.99 |
| 7. KdgT:  PF03812  (A0A072XF91) | **3ZUY (4.9E-13)** | 5A1S(0.033) | **12.72**/4.03 |
| 8. Mem\_trans:  PF03547  (D4J5G2) | **4N7W (0.17)**  **(3.8E-10)** | 5A1S(5.5) | **-10.71**/-0.31 |
| 9. 2HCT:  PF03390  (G4BX92) | 4CZB (6E-05) | Crystal structure  **5A1S (1.6E-59)** | -0.46/**14.67** |
| 10. DUF819:  PF05684  (A0A0B7I5I0) | 4CZB (0.00068) | **5A1S(1.8E-30)** | 3.94/**-11.21** |
| 11. Glt\_Symporter:  PF03616  (A0A0Q1HN67) | 4CZB(0.0005) | **5A1S( 6.9E-28)** | 3.39/**-13.68** |
| 12. AbrB:  PF05145  (L0MCY2) | 3ZUY(0.027) | **5A1S(5E-6)** | 3.36/**13.99** |
| 13. Asp\_Al\_Ex:  PF06826  (G5HK26) | 4CZB (0.00041) | **5A1S(4.5E-8)** | 8.12/**-14.17** |
| 14. Cons\_hypoth698\_1:  PF03601  (N9X0P8) | 5BZ3(0.15) | **5A1S(5.3E-13)** | 1.78/**13.1** |
| 15. Cons\_hypoth698\_2:  PF03601  (Q251E9) | 4CZB(0.78) | **5A1S(8.6E-8)** | 1.79/**9.76** |

**Table 2: Identification of Broken or reentrant type of transporter:** Family here refers to the Pfam family. Subfamilies are renamed by adding \_1 or \_2 to their Pfam family name. Uniprot ID of the representative sequence for each family is provided in brackets. Structural template with lowest E-value (Shown in bold) is used to identify the broken or reentrant type of transporter. Additionally, Identification of type of transporters based on KR bias calculations is chosen based on the highest mean KR bias value (Shown in bold).

# References

[1] [A.B. Chang, R. Lin, W. Keith Studley, C.V. Tran, M.H. Saier Jr, Phylogeny as a guide to structure and function of membrane transport proteins, Mol. Membr. Biol. 21 (2004) 171–181.](http://paperpile.com/b/6Cfpun/xLiA)

[2] [J.S. Chen, V. Reddy, J.H. Chen, M.A. Shlykov, W.H. Zheng, J. Cho, M.R. Yen, M.H. Saier Jr, Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments, J. Mol. Microbiol. Biotechnol. 21 (2011) 83–96.](http://paperpile.com/b/6Cfpun/pAMf)

[3] [R.D. Finn, J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic Acids Res. 34 (2006) D247–51.](http://paperpile.com/b/6Cfpun/rI7q)

[4] [O. Boudker, G. Verdon, Structural perspectives on secondary active transporters, Trends Pharmacol. Sci. 31 (2010) 418–426.](http://paperpile.com/b/6Cfpun/DL19)

[5] [C.L. Brett, M. Donowitz, R. Rao, Evolutionary origins of eukaryotic sodium/proton exchangers, Am. J. Physiol. Cell Physiol. 288 (2005) C223–39.](http://paperpile.com/b/6Cfpun/PGMD)

[6] [C. Lee, H.J. Kang, C. von Ballmoos, S. Newstead, P. Uzdavinys, D.L. Dotson, S. Iwata, O. Beckstein, A.D. Cameron, D. Drew, A two-domain elevator mechanism for sodium/proton antiport, Nature. 501 (2013) 573–577.](http://paperpile.com/b/6Cfpun/4HL7)

[7] [G. Masrati, M. Dwivedi, A. Rimon, Y. Gluck-Margolin, A. Kessel, H. Ashkenazy, I. Mayrose, E. Padan, N. Ben-Tal, Broad phylogenetic analysis of cation/proton antiporters reveals transport determinants, Nature Communications. 9 (2018). https://doi.org/](http://paperpile.com/b/6Cfpun/FYqL)[10.1038/s41467-018-06770-5.](http://dx.doi.org/10.1038/s41467-018-06770-5.)

[8] [M. Karmazyn, X.T. Gan, R.A. Humphreys, H. Yoshida, K. Kusumoto, The Myocardial Na -H Exchange, Circulation Research. 85 (1999) 777–786. https://doi.org/](http://paperpile.com/b/6Cfpun/FVbj)[10.1161/01.res.85.9.777.](http://dx.doi.org/10.1161/01.res.85.9.777.)

[9] [P.A. Dawson, Role of the intestinal bile acid transporters in bile acid and drug disposition, Handb. Exp. Pharmacol. (2011) 169–203.](http://paperpile.com/b/6Cfpun/agTU)

[10] [K. Palme, L. Gälweiler, PIN-pointing the molecular basis of auxin transport, Curr. Opin. Plant Biol. 2 (1999) 375–381.](http://paperpile.com/b/6Cfpun/CEx1)

[11] [C. Hunte, E. Screpanti, M. Venturi, A. Rimon, E. Padan, H. Michel, Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH, Nature. 435 (2005) 1197–1202.](http://paperpile.com/b/6Cfpun/kN93)

[12] [N.-J. Hu, S. Iwata, A.D. Cameron, D. Drew, Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT, Nature. 478 (2011) 408–411.](http://paperpile.com/b/6Cfpun/PHIb)

[13] [D. Wöhlert, M.J. Grötzinger, W. Kühlbrandt, Ö. Yildiz, Mechanism of Na(+)-dependent citrate transport from the structure of an asymmetrical CitS dimer, Elife. 4 (2015) e09375.](http://paperpile.com/b/6Cfpun/LVzV)

[14] [A.M. Duran, J. Meiler, Inverted topologies in membrane proteins: a mini-review, Comput. Struct. Biotechnol. J. 8 (2013) e201308004.](http://paperpile.com/b/6Cfpun/Pj0v)

[15] [O. Pornillos, G. Chang, Inverted repeat domains in membrane proteins, FEBS Letters. 580 (2006) 358–362. https://doi.org/](http://paperpile.com/b/6Cfpun/2s38)[10.1016/j.febslet.2005.12.054.](http://dx.doi.org/10.1016/j.febslet.2005.12.054.)

[16] [L.R. Forrest, Structural Symmetry in Membrane Proteins, Annual Review of Biophysics. 44 (2015) 311–337. https://doi.org/](http://paperpile.com/b/6Cfpun/ch0B)[10.1146/annurev-biophys-051013-023008.](http://dx.doi.org/10.1146/annurev-biophys-051013-023008.)

[17] [S. Radestock, L.R. Forrest, The alternating-access mechanism of MFS transporters arises from inverted-topology repeats, J. Mol. Biol. 407 (2011) 698–715.](http://paperpile.com/b/6Cfpun/0S1D)

[18] [E. Screpanti, C. Hunte, Discontinuous membrane helices in transport proteins and their correlation with function, J. Struct. Biol. 159 (2007) 261–267.](http://paperpile.com/b/6Cfpun/KCRF)

[19] [N. Reyes, C. Ginter, O. Boudker, Transport mechanism of a bacterial homologue of glutamate transporters, Nature. 462 (2009) 880–885.](http://paperpile.com/b/6Cfpun/cFTr)

[20] [D. Drew, O. Boudker, Shared Molecular Mechanisms of Membrane Transporters, Annual Review of Biochemistry. 85 (2016) 543–572. https://doi.org/](http://paperpile.com/b/6Cfpun/qtaK)[10.1146/annurev-biochem-060815-014520.](http://dx.doi.org/10.1146/annurev-biochem-060815-014520.)

[21] [B. Byrne, It takes two to transport via an elevator, Cell Res. 27 (2017) 965–966.](http://paperpile.com/b/6Cfpun/kgrP)

[22] [M. Appel, D. Hizlan, K.R. Vinothkumar, C. Ziegler, W. Kühlbrandt, Conformations of NhaA, the Na/H exchanger from Escherichia coli, in the pH-activated and ion-translocating states, J. Mol. Biol. 386 (2009) 351–365.](http://paperpile.com/b/6Cfpun/5nLG)

[23] [P.P. Chothe, L.C. Czuba, R.H. Moore, P.W. Swaan, Human bile acid transporter ASBT (SLC10A2) forms functional non-covalent homodimers and higher order oligomers, Biochim. Biophys. Acta Biomembr. 1860 (2018) 645–653.](http://paperpile.com/b/6Cfpun/N3FZ)

[24] [A. Vergara-Jaque, C. Fenollar-Ferrer, D. Kaufmann, L.R. Forrest, Repeat-swap homology modeling of secondary active transporters: updated protocol and prediction of elevator-type mechanisms, Front. Pharmacol. 6 (2015) 183.](http://paperpile.com/b/6Cfpun/87YC)

[25] [M.H. Saier Jr, V.S. Reddy, B.V. Tsu, M.S. Ahmed, C. Li, G. Moreno-Hagelsieb, The Transporter Classification Database (TCDB): recent advances, Nucleic Acids Res. 44 (2016) D372–9.](http://paperpile.com/b/6Cfpun/lk3U)

[26] [N.M. Mansour, M. Sawhney, D.G. Tamang, C. Vogl, M.H. Saier Jr, The bile/arsenite/riboflavin transporter (BART) superfamily, FEBS J. 274 (2007) 612–629.](http://paperpile.com/b/6Cfpun/yGZP)

[27] [M.A. Lomize, I.D. Pogozheva, H. Joo, H.I. Mosberg, A.L. Lomize, OPM database and PPM web server: resources for positioning of proteins in membranes, Nucleic Acids Res. 40 (2012) D370–6.](http://paperpile.com/b/6Cfpun/3Z1l)

[28] [H. Cheng, R.D. Schaeffer, Y. Liao, L.N. Kinch, J. Pei, S. Shi, B.-H. Kim, N.V. Grishin, ECOD: an evolutionary classification of protein domains, PLoS Comput. Biol. 10 (2014) e1003926.](http://paperpile.com/b/6Cfpun/L5NE)

[29] [N.L. Dawson, T.E. Lewis, S. Das, J.G. Lees, D. Lee, P. Ashford, C.A. Orengo, I. Sillitoe, CATH: an expanded resource to predict protein function through structure and sequence, Nucleic Acids Res. 45 (2017) D289–D295.](http://paperpile.com/b/6Cfpun/Jd35)

[30] [N.K. Fox, S.E. Brenner, J.-M. Chandonia, SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures, Nucleic Acids Research. 42 (2014) D304–D309. https://doi.org/](http://paperpile.com/b/6Cfpun/HSJA)[10.1093/nar/gkt1240.](http://dx.doi.org/10.1093/nar/gkt1240.)

[31] [A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.](http://paperpile.com/b/6Cfpun/tztR)

[32] [J. Söding, Protein homology detection by HMM-HMM comparison, Bioinformatics. 21 (2005) 951–960.](http://paperpile.com/b/6Cfpun/W2Qg)

[33] [T. Nodzyński, S. Vanneste, M. Zwiewka, M. Pernisová, J. Hejátko, J. Friml, Enquiry into the Topology of Plasma Membrane-Localized PIN Auxin Transport Components, Mol. Plant. 9 (2016) 1504–1519.](http://paperpile.com/b/6Cfpun/vQuZR)

[34] [K. Nanatani, T. Fujiki, K. Kanou, M. Takeda-Shitaka, H. Umeyama, L. Ye, X. Wang, T. Nakajima, T. Uchida, P.C. Maloney, K. Abe, Topology of AspT, the aspartate:alanine antiporter of Tetragenococcus halophilus, determined by site-directed fluorescence labeling, J. Bacteriol. 189 (2007) 7089–7097.](http://paperpile.com/b/6Cfpun/qm261)

[35] [A. Dobrowolski, I. Sobczak-Elbourne, J.S. Lolkema, Membrane topology prediction by hydropathy profile alignment: membrane topology of the Na(+)-glutamate transporter GltS, Biochemistry. 46 (2007) 2326–2332.](http://paperpile.com/b/6Cfpun/QBDAd)

[36] [X. Zhou, E.J. Levin, M. Quick, Y. Pan, R. Sharma, M. Zhou, Structural Basis of the Alternate-Access Mechanism in a Bile Acid Transporter, Biophysical Journal. 104 (2013) 539a. https://doi.org/](http://paperpile.com/b/6Cfpun/F030)[10.1016/j.bpj.2012.11.2984.](http://dx.doi.org/10.1016/j.bpj.2012.11.2984.)

[37] [B. Hagenbuch, B. Stieger, M. Foguet, H. Lübbert, P.J. Meier, Functional expression cloning and characterization of the hepatocyte Na+/bile acid cotransport system, Proc. Natl. Acad. Sci. U. S. A. 88 (1991) 10629–10633.](http://paperpile.com/b/6Cfpun/Z9VV)

[38] [C. Allen, S. Reverchon, J. Robert-Baudouy, Nucleotide sequence of the Erwinia chrysanthemi gene encoding 2-keto-3-deoxygluconate permease, Gene. 83 (1989) 233–241.](http://paperpile.com/b/6Cfpun/uDCl)

[39] [Y. Liu, A. Basu, X. Li, L. Fliegel, Topological analysis of the Na+/H+ exchanger, Biochim. Biophys. Acta. 1848 (2015) 2385–2393.](http://paperpile.com/b/6Cfpun/NDAy)

[40] [K. Abe, F. Ohnishi, K. Yagi, T. Nakajima, T. Higuchi, M. Sano, M. Machida, R.I. Sarker, P.C. Maloney, Plasmid-encoded asp operon confers a proton motive metabolic cycle catalyzed by an aspartate-alanine exchange reaction, J. Bacteriol. 184 (2002) 2906–2913.](http://paperpile.com/b/6Cfpun/Gc4C)

[41] [M. Kalman, D.R. Gentry, M. Cashel, Characterization of the Escherichia coli K12 gltS glutamate permease gene, Mol. Gen. Genet. 225 (1991) 379–386.](http://paperpile.com/b/6Cfpun/RyXc)

[42] [M.R. Volkert, L.I. Hajec, Z. Matijasevic, F.C. Fang, R. Prince, Induction of the Escherichia coli aidB gene under oxygen-limiting conditions requires a functional rpoS (katF) gene, J. Bacteriol. 176 (1994) 7638–7645.](http://paperpile.com/b/6Cfpun/YQRY)

[43] [D.O. Daley, M. Rapp, E. Granseth, K. Melén, D. Drew, G. von Heijne, Global topology analysis of the Escherichia coli inner membrane proteome, Science. 308 (2005) 1321–1323.](http://paperpile.com/b/6Cfpun/NWYL)

[44] [D. Woehlert, M.J. Groetzinger, W. Kuhlbrandt, O. Yildiz, Crystal structure of the sodium-dependent citrate symporter SeCitS form Salmonella enterica, (2016). https://doi.org/](http://paperpile.com/b/6Cfpun/ZsvX)[10.2210/pdb5a1s/pdb.](http://dx.doi.org/10.2210/pdb5a1s/pdb.)

[45] [H. Viklund, E. Granseth, A. Elofsson, Structural Classification and Prediction of Reentrant Regions in α-Helical Transmembrane Proteins: Application to Complete Genomes, Journal of Molecular Biology. 361 (2006) 591–603. https://doi.org/](http://paperpile.com/b/6Cfpun/mOQK)[10.1016/j.jmb.2006.06.037.](http://dx.doi.org/10.1016/j.jmb.2006.06.037.)

[46] [C. Yan, J. Luo, A Comparison between Transmembrane Helices and Reentrant Loops, 2010 IEEE International Conference on BioInformatics and BioEngineering. (2010). https://doi.org/](http://paperpile.com/b/6Cfpun/XU0N)[10.1109/bibe.2010.54.](http://dx.doi.org/10.1109/bibe.2010.54.)

[47] [H. Viklund, A. Elofsson, OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar, Bioinformatics. 24 (2008) 1662–1668.](http://paperpile.com/b/6Cfpun/WqxA)

[48] [K.N. Lazaridis, P. Tietz, T. Wu, S. Kip, P.A. Dawson, N.F. LaRusso, Alternative splicing of the rat sodium/bile acid transporter changes its cellular localization and transport properties, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 11092–11097.](http://paperpile.com/b/6Cfpun/phga)

[49] [E.R. Slepkov, J.K. Rainey, B.D. Sykes, L. Fliegel, Structural and functional analysis of the Na /H exchanger, Biochemical Journal. 401 (2007) 623–633. https://doi.org/](http://paperpile.com/b/6Cfpun/XGDk)[10.1042/bj20061062.](http://dx.doi.org/10.1042/bj20061062.)

[50] [L.P. Iñiguez, G. Hernández, The Evolutionary Relationship between Alternative Splicing and Gene Duplication, Front. Genet. 8 (2017) 14.](http://paperpile.com/b/6Cfpun/UC0V)

[51] [J.S. Lolkema, D.J. Slotboom, Structure and elevator mechanism of the Na -citrate transporter CitS, Current Opinion in Structural Biology. 45 (2017) 1–9. https://doi.org/](http://paperpile.com/b/6Cfpun/2Meq)[10.1016/j.sbi.2016.10.004.](http://dx.doi.org/10.1016/j.sbi.2016.10.004.)

[52] [M.S. Jin, J.W. Kim, S. Kim, S. Kim, H. Lee, J. -O. Lee, Structural insights into the elevator-like mechanism of the sodium/citrate symporter CitS, (2017). https://doi.org/](http://paperpile.com/b/6Cfpun/nBsY)[10.2210/pdb5x9r/pdb.](http://dx.doi.org/10.2210/pdb5x9r/pdb.)

[53] [D. Wöhlert, M.J. Grötzinger, W. Kühlbrandt, Ö. Yildiz, Mechanism of Na -dependent citrate transport from the structure of an asymmetrical CitS dimer, eLife. 4 (2015). https://doi.org/](http://paperpile.com/b/6Cfpun/mwCU)[10.7554/elife.09375.](http://dx.doi.org/10.7554/elife.09375.)

[54] [C. Luschnig, G. Vert, The dynamics of plant plasma membrane proteins: PINs and beyond, Development. 141 (2014) 2924–2938.](http://paperpile.com/b/6Cfpun/I9Sk)

[55] [G. von Heijne, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225 (1992) 487–494.](http://paperpile.com/b/6Cfpun/QqUl)

[56] [M. Mirdita, L. von den Driesch, C. Galiez, M.J. Martin, J. Söding, M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, Nucleic Acids Res. 45 (2017) D170–D176.](http://paperpile.com/b/6Cfpun/qURa)

[57] [S. Altschul, Basic Local Alignment Search Tool, Journal of Molecular Biology. 215 (1990) 403–410. https://doi.org/](http://paperpile.com/b/6Cfpun/8vwK)[10.1006/jmbi.1990.9999.](http://dx.doi.org/10.1006/jmbi.1990.9999.)

[58] [F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539.](http://paperpile.com/b/6Cfpun/cT4y)

[59] [K.D. Tsirigos, C. Peters, N. Shu, L. Käll, A. Elofsson, The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides, Nucleic Acids Research. 43 (2015) W401–W407. https://doi.org/](http://paperpile.com/b/6Cfpun/GKvo)[10.1093/nar/gkv485.](http://dx.doi.org/10.1093/nar/gkv485.)

[60] [M.N. Price, P.S. Dehal, A.P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix, Mol. Biol. Evol. 26 (2009) 1641–1650.](http://paperpile.com/b/6Cfpun/HBse)

[61] [UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.](http://paperpile.com/b/6Cfpun/mzv9)

[62] [S.R. Eddy, Accelerated Profile HMM Searches, PLoS Comput. Biol. 7 (2011) e1002195.](http://paperpile.com/b/6Cfpun/6JWW)

[63] [M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment, Nat. Methods. 9 (2011) 173–175.](http://paperpile.com/b/6Cfpun/VZko)

[64] [V. Alva, S.-Z. Nam, J. Söding, A.N. Lupas, The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis, Nucleic Acids Res. 44 (2016) W410–5.](http://paperpile.com/b/6Cfpun/uT1V)

[65] [E.L. Sonnhammer, G. von Heijne, A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences, Proc. Int. Conf. Intell. Syst. Mol. Biol. 6 (1998) 175–182.](http://paperpile.com/b/6Cfpun/IYui)

[66] [T. Hessa, N.M. Meindl-Beinker, A. Bernsel, H. Kim, Y. Sato, M. Lerch-Bader, I. Nilsson, S.H. White, G. von Heijne, Molecular code for transmembrane-helix recognition by the Sec61 translocon, Nature. 450 (2007) 1026–1030.](http://paperpile.com/b/6Cfpun/YGpu)

[67] [G. Heijne, The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology, EMBO J. 5 (1986) 3021–3027.](http://paperpile.com/b/6Cfpun/9b44)

[68] [G. von Heijne, Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues, Nature. 341 (1989) 456–458.](http://paperpile.com/b/6Cfpun/iads)

[69] [M.T. Virkki, C. Peters, D. Nilsson, T. Sörensen, S. Cristobal, B. Wallner, A. Elofsson, The positive inside rule is stronger when followed by a transmembrane helix, J. Mol. Biol. 426 (2014) 2982–2991.](http://paperpile.com/b/6Cfpun/a6Lp)

[70] [R. Suzuki, H. Shimodaira, Pvclust: an R package for assessing the uncertainty in hierarchical clustering, Bioinformatics. 22 (2006) 1540–1542. https://doi.org/](http://paperpile.com/b/6Cfpun/5n5nA)[10.1093/bioinformatics/btl117.](http://dx.doi.org/10.1093/bioinformatics/btl117.)

[71] [Y. Zhang, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Research. 33 (2005) 2302–2309. https://doi.org/](http://paperpile.com/b/6Cfpun/0IpL)[10.1093/nar/gki524.](http://dx.doi.org/10.1093/nar/gki524.)

[72] [W.L. DeLano, The PyMOL Molecular Graphics System; DeLano Scientific: San Carlos, CA, 2002, (n.d.).](http://paperpile.com/b/6Cfpun/S9gU)