Title
Topological Predictions for Integral Membrane Channel and Carrier Proteins /

Permalink
https://escholarship.org/uc/item/3fx815q7

Author
Reddy, Abhinay Boddu

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Topological Predictions for Integral Membrane Channel and Carrier Proteins

A thesis submitted in partial satisfaction of the requirements

For the degree Master of Science

in

Biology

by

Abhinay Boddu Reddy

Committee in charge:

Professor Milton H. Saier, Chair
Professor James W. Golden
Professor Hector Viadiu-Ilarraza

2013

The thesis of Abhinay Boddu Reddy is approved, and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank and acknowledge Dr. Milton Saier for all of his help, support, guidance and encouragement as I worked on this project. I want to highlight the chance he gave me when allowing me to join the lab, subsequently helping me develop as a student and as a biologist. Most importantly, I am truly indebted for his willingness to support me when I needed to gain admission into the BS/MS program at UCSD. He deserves all of my sincerest gratitude for working on every draft and revision of my thesis, and helping me interpret my results when I could not make sense of them. I would also like to acknowledge post-doctoral scholar Ake Vastermark for his ideas on expanding my study.

I would like to thank and acknowledge Dr. Randy Hampton and all the members of his lab for all the great times I shared with them, and their support and help as I grew as a student, a teacher and a biologist. In particular, I would like to thank Dr. Randy Hampton for his efforts to help admit me into the BS/MS program, and Mengxiao Ma for being a source of motivation and inspiration, along with providing much needed scientific advice.

I would like to thank my father, Jagjeevan Reddy, my good friend Michael L. Matthews, and the lab programmers, Vamsee Reddy and Bryan Lunt, for all of their technical assistance and efforts in setting up my data collection programs. Without their help, this project would not have been possible. I would also like to acknowledge Carl Welliver for his assistance in getting the manuscript formatted for publication.

I would like to recognize Andy Lukosus, the BS/MS coordinator in the Division of Biological Sciences for always being available to assist me with the administrative and procedural aspects of being part of the graduate program, and I wish him all the best at his new job at ASU.

Finally, I wish to thank the members of my thesis committee, Dr. James Golden and Dr. Hector Viadiu-ilarraza for taking time out of their busy schedules to be a part of my thesis committee.

The Results chapter of this thesis is being prepared for publication. Reddy Abhinay; Cho Jaehoon; Ling Sam; Reddy Vamsee. The thesis author was the primary investigator and author of this material. This work was supported by NIH grant GM077402 from the National Institute of General Medical Sciences.

ABSTRACT OF THE THESIS

Topological Predictions for Integral Membrane Channel and Carrier Proteins

by

Abhinay Boddu Reddy

Master of Science in Biology

University of California, San Diego, 2013

Professor Milton H. Saier, Chair

We evaluated topological predictions for nine different programs, HMMTOP, TMHMM, SVMTOP, DAS, SOSUI, TOPCONS, PHOBIUS, MEMSAT, and SPOCTOPUS. These programs were first evaluated using four large topologically well-defined families of secondary transporters, and the three best programs were further evaluated using topologically more diverse families of channels and carriers. In the initial studies, the order of accuracy was:

SPOCTOPUS>MEMSAT>HMMTOP>TOPCONS>PHOBIUS>TMHMM>SVMTOP>DAS>SOSUI. Some families, such as the Sugar Porter family (2.A.1.1) of the Major Facilitator Superfamily (MFS; TC# 2.A.1) and the Amino acid/Polyamine/Organocation (APC) Family (TC# 2.A.3), were correctly predicted with high accuracy while others, such as the Mitochondrial Carrier (MC) (TC# 2.A.29) and the K^+ transporter (Trk) families (TC# 2.A.38), were predicted with much lower accuracy. For small, topologically homogeneous families, SPOCTOPUS and MEMSAT were generally most reliable, while with large, more diverse superfamilies, HMMTOP often proved to have the greatest prediction accuracy. We next developed a novel program, TM-STATS, that tabulates HMMTOP, SPOCTOPUS or MEMSAT-based topological predictions for any subdivision (class, subclass, superfamily, family, subfamily, or any combination of these) of the Transporter Classification Database (TCDB; www.tcdb.org) and examined the following subclasses: α-type channel proteins (TC subclasses 1.A and 1.E), secreted pore-forming toxins (TC subclass 1.C) and secondary carriers (subclass 2.A). Histograms were generated for each of these subclasses, and the results were analyzed according to subclass, family and protein. The results provide an update of topological predictions for integral membrane transport proteins as well as guides for the development of more reliable topological prediction programs, taking family-specific characteristics into account.

Introduction

Transport proteins function by multiple mechanisms, allowing hydrophilic molecules to cross biological membranes [1]. The simplest of these proteins form pores or channels which allow the free diffusion of molecules from one side of the membrane to the other [2]. Some of these proteins are small peptides that contain only one or two transmembrane segments (TMSs). In order to form transmembrane pores, these peptides form oligomeric structures with TMSs approximately perpendicular to the plane of the membrane. Others contain many more TMSs, often having arisen by multiplication of a small basic peptide unit with just a few TMSs. The larger number of repeat units minimizes the need for a greater number of subunits necessary to form the pore [3; 4].

Various physical and/or chemical agents often gate the larger proteins but usually not the smaller ones. We have postulated that the simplest of these channel proteins were the primordial systems that gave rise to more complicated channels via intragenic duplication [5]. Two types of channel proteins can be distinguished, one of which exerts its action in the cell that produces it, and the other which targets a cell other than the one that makes it [6]. The latter proteins are toxins that form pores in the membranes of a target organism, releasing nutrients for the predatory organism while killing the target [7; 8]. Toxins can similarly exist in small and large forms, where simple peptide toxins usually have no more than 1 or 2 transmembrane segments. Although the larger protein toxins may have more, this is not always the case. This is because protein toxins often include protein domains that serve any of a variety of functions, such as subcellular targeting and functional regulation [3].

From larger channel proteins, we have postulated that carriers, capable of recognizing their substrates and shuttling them across the membrane, arose in part as a result of point mutations [5]. In contrast to channel proteins, very few carriers have been documented that exhibit fewer than 4 TMSs. Even in the two or three examples where fewer than four TMSs for hypothesized carriers have been suggested, the mode of transport is uncertain. Thus it appears that in order to form a carrier, a larger, more constrained, less oligomeric structure may be required.

In previous studies, we noted the presence of repeat sequences in many secondary carrier proteins [5]. It was this observation that led to the proposed pathway described above. Repeat sequences were initially detected using computer programs that allowed prediction of numbers of transmembrane segments. Several such programs are available. Among these are: HMMTOP [9], SVMTOP [10], TMHMM [11], DAS [12] SOSUI [13], TOPCONS [14], PHOBIUS [15], MEMSAT-SVM (hereafter called MEMSAT) [16] and SPOCTOPUS [17]. The authors describing each of these programs have claimed a high degree of accuracy, usually over 90%, with the exceptions of SVMTOP, which has a reported accuracy of over 70%, and HMMTOP, with a reported accuracy of 88.5%. However, seldom have independent research groups confirmed the observations reported by these investigators.

In the present study, we have compared the nine programs mentioned above using several independently evolving families of transport proteins. Initially we examined members of four different well-characterized families, all of known topology, to evaluate the relative accuracies of these nine programs. Using these data sets, we could establish that HMMTOP, SPOCTOPUS and MEMSAT were the top performers for all 4 families. Consequently, these programs were used to design a novel program (TMStats) that provides statistical analyses of integral membrane transport protein topologies. Upon application of TMStats, we found that SPOCTOPUS and MEMSAT commonly performed most accurately when presented with small, topologically homogenous groups of proteins, but that HMMTOP is the most accurate topological prediction program when certain larger, diverse superfamilies of transport proteins are analyzed.

The Transporter Classification Database (TCDB: www.tcdb.org) categorizes all transport systems according to class, subclass, family, subfamily, and protein [18; 19]. In addition, a hyperlink exists that delineates superfamily relationships among these families. Altogether, TCDB includes over 700 families, many of them included within superfamilies. The novel TMStats program can examine any of these categories, or combinations of these categories simultaneously to make topological predictions.

After first evaluating the nine above-mentioned programs with four selected families, we conducted studies with whole classes of transport proteins. We examined first, α-helical type channels (TC # 1.A), second, small α-helical holin-type channel-forming proteins (TC # 1.E) frequently involved in bacteriophage lysis or bacterial programmed cell death, third, pore-forming toxins that insert into membranes of a target organism other than the one that produces it (TC # 1.C), and fourth, secondary carriers that shuttle substrates across the membrane in a process that involves major conformational changes coupled to the transport cycle (TC # 2.A).

Channels and carriers can be distinguished because the former have turnover rates roughly 1000-fold higher than those of the latter. While channels are often diffusion limiting, carriers never are. Only the latter generally exhibit high stereospecificity for their substrates. In this paper we analyze these transport proteins by subclass, family, and protein. Using large datasets, we confirm previous results concerning average numbers of TMSs of the different classes of proteins. We also notice characteristics that distinguish families or superfamilies. The results should allow refinement of evolutionary predictions and guides to mechanistic details mediated by these proteins.

Methods

Topological Analyses

Several programs were used for comparative purposes to predict integral membrane transport protein topologies. The nine programs examined were HMMTOP, SVMTOP, TMHMM, DAS, SOSUI, TOPCONS, PHOBIUS, MEMSAT and SPOCTOPUS. HMMTOP (http://www.enzim.hu/hmmtop/html/document.html) is a topology prediction program developed by G.E. Tusnady at the Institute of Enzymology at the University of Hungary; it uses a hidden Markov model to predict the number of transmembrane segments [9]. SVMTOP, developed at the Institution of Information Science, Academia Sinica, Taiwan, (http://biocluster.iis.sinica.edu.tw/~bioapp/SVMtop/about.php) is a program that predicts transmembrane helices using a “support vector machine” method that hierarchically classifies transmembrane segments based on inside versus outside loops. [10]. TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/), developed at the Center for Biological Sequence Analysis in Denmark, uses the hidden Markov model to predict transmembrane helices [11]. DAS (http://mendel.imp.ac.at/sat/DAS/DAS.html) is a dissimilar topology prediction program developed at the Biological Research Center at the Institute of Enzymology, Hungarian Academy of Sciences; it uses a “dense alignment surface” algorithm that creates a hydrophobicity profile for the query by comparing it to a predetermined library and scoring matrix (http://mendel.imp.ac.at/sat/DAS/abstract.html) [12]. SOSUI (http://bp.nuap.nagoya-u.ac.jp/sosui/)

is a topology prediction program developed by the Mitaku Group in the Department of Applied Physics at Nagoya University. The batch version of this program was used [13]. TOPCONS (http://topcons.cbr.su.se/) was developed at Stockholm University, and uses multiple topology prediction algorithms to generate a consensus prediction [14]. PHOBIUS (http://phobius.sbc.su.se/) was created at the Center for Genomics and Bioinformatics in the Karolinska Institute in Stockholm, Sweden [15]. MEMSAT (http://bioinf.cs.ucl.ac.uk/psipred/) uses an improved support vector machine model relative to SVMTOP that was developed by the Department of Computer Science: Bioinformatics Group at the University College London [16]. Finally, SPOCTOPUS (http://octopus.cbr.su.se/) was developed at Stockholm University and uses the OCTOPUS algorithm to detect transmembrane segments and a signal peptide prediction algorithm to detect signal peptides; OCTOPUS uses a combination of hidden Markov models and neural networks along with a BLAST search to generate a sequence profile that is annotated with transmembrane properties [17].

In order to set a standard for accuracy, Average Hydropathy, Amphipathicity and Similarity (AveHAS) plots were generated. Because the input file for the AveHAS program is a multiple alignment file produced by the ClustalX program, the results for many proteins are averaged, giving plots that provide much greater predictive accuracy than is possible with individual sequences [20]. The multiple alignments used in the generation of these plots can be found as supplementary materials on our website (http://www.biology.ucsd.edu/~msaier/supmat/TMStats/index.html).

Another program, called WHAT (Web-based Hydropathy and Amphipathicity) allows the generation of hydrophobicity plots for single sequences and provides TMS predictions using HMMTOP; the input for this program is the query sequence in FASTA form. Usage of the WHAT program allows topological verification by counting the number of hydrophobic peaks that represent potential TMSs [21].

Results

Comparison of nine topological prediction programs using proteins from four different superfamilies

In our initial studies, we chose to compare frequently used methods of integral membrane topological prediction using four moderately sized families of transport systems for which the topologies have been established experimentally. These families are (1) the Sugar Porter Family (TC# 2.A.1.1) of the Major Facilitator Superfamily (MFS), members of which have 12 experimentally established TMSs [22], (2) the Amino Acid-Polyamine-organoCation (APC) Family (TC# 2.A.3) within the APC Superfamily, members of which have 10, 12, 14 or 15 established TMSs [23], (3) the Mitochondrial Carrier (MC) Family (TC# 2.A.29) within the MC Superfamily, members of which have 6 established TMSs [24] and (4) the Potassium Transporter (Trk) Family (TC# 2.A.38) within the VIC Superfamily, members of which have 8 established TMSs [25; 26]. The data in this section was obtained on 3/11/2012.

The nine programs examined are listed in Table 1. Eighty-four proteins, derived from TCDB, were the test set used for the Sugar Porter Family within the MFS. These proteins were multiply aligned (ClustalX), and the average topology for these 84 proteins, based on the AveHAS, program is shown in Figure 1A. By averaging the results for these proteins, the prediction of topology becomes clear. The peaks of hydropathy corresponding to TMSs are labeled 1 through 12. As is established for the MFS [27], these proteins consist of two halves, each of which contains six transmembrane segments [22]. The picture obtained by averaging the hydropathy predictions for these sequences is much more clear than when the individual protein hydropathy plots, using the WHAT program, were displayed. The peaks of hydropathy shown in the top panel correspond to the peaks represented by vertical lines in the bottom panel, and also correspond to peaks of similarity as shown by the dashed line in the bottom panel. This plot agrees with our general observation that the TMSs in integral membrane transport proteins are better conserved than the hydrophilic loop regions between them. It also confirms experimental data, including X-ray crystallographic studies, showing that these proteins have 12 TMSs [27].

The data presented in Table 1A reveals the number of proteins predicted to have anywhere between 6 and 13 TMSs. For the MFS, the HMMTOP program predicted 92% of the proteins (77 proteins) having 12 TMSs; the remaining 8% (7 proteins) were predicted to have either 10 or 11 TMSs. Examination of the plots for these seven proteins revealed that HMMTOP missed one or two of the TMSs for each protein. The second program, listed in Table 1, SVMTOP, proved to be much less reliable, with only 49 proteins predicted to have 12 TMSs. The others were predicted to have 10, 11 or 13 TMSs. The third program, TMHMM, predicted 54 proteins to have 12 TMSs, and the exceptions had anywhere from 8 to 11 TMSs. The DAS program predicted only 36 proteins to have 12 TMSs, with large numbers of proteins predicted to have 7 through 11 and 13 TMSs. The SOSUI program predicted only 29 proteins to have 12 TMSs, with the others having anywhere from 6 to 13 TMSs. TOPCONS predicted 75 proteins to have 12 TMSs, with the remainder of the proteins having between 9 and 11 TMSs. PHOBIUS performed most poorly, predicting 0 proteins to have 12 TMSs, and predicted the majority of proteins to have either 9 or 10 TMSs. MEMSAT predicted 78 proteins to have 12 TMSs, with a total of 6 proteins having either 9, 10, or 11 TMSs. Finally SPOCTOPUS was the top performer, predicting 80 proteins to have 12 TMSs, 1 protein to have 9 TMSs, and 3 proteins to have 11 TMSs. Thus, the top performers were SPOCTOPUS, MEMSAT and HMMTOP, predicting 77-80 of 84 proteins correctly.

The average hydropathy plot for the APC family (91 proteins derived from TCDB) is shown in Figure 1B. Of the 91 proteins included in this study, 83 are believed to have 12 TMSs, four have 10 TMSs, six have 14 and one has 15 TMSs (see section on the APC Superfamily below). The average hydropathy plot revealed 12 well conserved peaks of hydropathy as expected for the dominant members of this family.

Examination of Table 1B reveals that TOPCONS predicts the largest number of proteins to have 12 TMSs, but HMMTOP appears to have the best overall prediction accuracy as it found sixty-nine 12 TMS proteins, and correctly predicted the four 10 TMS proteins as well as the six 14 and one 15 TMS proteins. With regard to the APC family, the order of correct predictions was HMMTOP > TOPCONS > PHOBIUS > MEMSAT > SPOCTOPUS > TMHMM > SVMTOP > DAS > SOSUI.

Members of the MC superfamily are known to have six TMSs with no reported topological variations. Examination of the hydropathy plots for the 88 proteins derived from TCDB and included in this study revealed that all could be interpreted as having six TMSs. However, these plots were usually ambiguous in contrast to the MFS and APC superfamilies. SPOCTOPUS was the strongest predictor, with 79 out of 88 proteins (89%) predicted correctly. MEMSAT produced the next best results, predicting the correct topology for 68/88 proteins (77%). HMMTOP predicted 23 proteins to have 6 TMSs, with large numbers predicted to have fewer than 6 TMSs. Only two were predicted to have 7 TMSs. Thus, only 26% of these proteins were correctly predicted. By contrast, very few proteins were predicted to have 6 TMSs by any of the other six programs used (Table 1C). In all such cases, fewer than 6 TMSs were predicted. It is clear that while SPOCTOPUS and MEMSAT predicted the correct topology for these proteins well, the other programs did extremely poorly.

While hydropathy plots for the individual proteins in the MC Superfamily were often confusing, the use of the AveHAS program to generate average hydropathy and similarity plots for the 88 proteins gave clear results as shown in Figure 1C. Here, one can see that 6 TMSs are predicted, where TMSs 1, 3 and 5 are high sharp peaks, while 2, 4 and 6 are lower and broader. This pattern reflects the presence of three 2 TMS repeat units present in all mitochondrial carriers. Each of these 6 peaks is well conserved. These results again illustrate the advantage of using the AveHAS program for topological predictions.

Potassium transporters of the Trk family (20 proteins from TCDB included in this study) were predicted less accurately than the mitochondrial carriers. These proteins are known to have four repeat units derived from the channel-forming element of members of the voltage-gated ion channel superfamily [10; 25; 26]. In fact, the Trk family is a constituent member of the VIC Superfamily (see TCDB Superfamilies: http://tcdb.org/superfamily.php). These channels consist of two TMSs with a central semi-hydrophobic P-loop that dips into the membrane but does not traverse it. This topology was not readily apparent when individual proteins were examined with the WHAT program, but the situation was much clearer when the average hydropathy plot was displayed. This plot revealed four quadrants, each with two hydrophobic peaks separated by a small semi-polar peak. Odd numbered peaks, as indicated in Figure 1D, (the first TMS in each repeat unit) are sharp and high, while even numbered peaks (the second TMS in each repeat unit) are broader and lower. The P-loop is apparent in all four quadrants. This plot again reveals the greater predictive capabilities observed when many proteins are averaged to give a hydropathy plot.

The predictions obtained for the Trk family using the nine different programs are summarized in Table 1D. The majority of the programs predicted fewer than 10 of the 20 proteins to have 8 TMSs, with SPOCTOPUS being the only exception, predicting 10 proteins to have 8 TMSs; MEMSAT was a close second, predicting seven proteins correctly. In contrast to the MC family discussed above, virtually all mispredictions were overpredictions (except in the cases of SPOCTOPUS and MEMSAT). These overpredictions resulted because some or all of the P-loops were counted as TMSs. The correct number of predictions was therefore at best only 50%, or less for all remaining programs. In fact, all programs predicted the average total number of TMSs to be between 9.7 and 10.8. Thus, in this case, none of the programs proved to provide accurate predictions. This can be explained by the fact that all of them predict at least some of the P-loops to be transmembrane.

Summarizing, for the MFS and APC superfamilies, SPOCTOPUS, MEMSAT and HMMTOP are the most reliable programs for topological predictions, while TOPCONS follows. The other 5 programs are less reliable. However, reliability with any program depends upon the family of proteins being analyzed. Some families, such as the MC family were poorly predicted by most programs, and all programs poorly predicted the Trk family. It is imperative that improved prediction methods be developed.

As noted above, the results presented in this section reveal that in general, SPOCTOPUS, MEMSAT and HMMTOP are the most reliable programs available for predicting the topologies of integral membrane transport proteins. For that reason, we chose to use these programs for the quantitative evaluation of predictions for various TC classes, subclasses, superfamilies, families, subfamilies, or any combination of these using the integrated TMStats program (see Methods section). While using three programs theoretically should afford the most comprehensive prediction coverage, we show that SPOCTOPUS and MEMSAT mostly produce the more accurate results when certain small, topologically similar families of proteins are considered, but HMMTOP produces the most accurate predictions when considering certain large families including several superfamilies.

Transmembrane α-helical Channels (Subclass 1.A)

Subclass 1.A includes channel-forming proteins that consist primarily of α-helical TMSs. The entire subclass was analyzed collectively for topological types using TMStats with the three topology prediction programs, HMMTOP, MEMSAT and SPOCTOPUS, both with and without auxiliary proteins. Including all auxiliary proteins, a total of 916 proteins were analyzed, and without these auxiliary proteins, there were 820 proteins (5/30/2013). The average numbers of TMSs, including auxiliary proteins, were 5.5 +/- 4.9 S.D. using HMMTOP, 5.4 +/- 5.3 S.D. using MEMSAT, and 5.1 +/- 5.0 S.D. using SPOCTOPUS; without the auxiliary proteins, the averages were 5.9 +/- 5.0 S.D., 5.7 +/- 5.4 S.D. and 5.4 +/- 5.1 S.D. respectively. In the analyses reported below, only the results obtained when auxiliary proteins were excluded are reported.

A plot of topological types with frequencies of occurrence on the Y-axis and the numbers of TMSs on the X-axis, revealed the distributions using all three programs (Figure 2). There were more 2 than 1 TMS proteins, more 4 than 3 TMS proteins, and more 6 than 5 TMS proteins, showing that even numbered channel-forming proteins are favored. This was true regardless of which of the three programs were used (Figure 2). This observation reinforces the conclusion of an earlier publication using a much smaller data set [5]. The prevalence of even numbered proteins can be explained by the fact that duplication of any number of TMSs gives rise to proteins with even numbers of TMSs, and most transport proteins have arisen via pathways involving intragenic duplication. Surprisingly, however, there are substantially more 11 than 10 TMS proteins, regardless of the program used. Of the proteins predicted to have 11 TMSs, the majority proved to belong to the Amt channel family (1.A.11) regardless of the program used. In fact, the two Amt channel proteins for which high-resolution X-ray structures are available display 11 established TMSs, and most members of this family are predicted to have 11 TMSs. Of the remaining 11 TMS proteins, HMMTOP and MEMSAT predict the cholesterol/dsRNA uptake (CUP) family (1.A.79) to include the most such proteins. This is in accordance with one of the two models proposed for the topology of these proteins. Examining the larger proteins, those predicted to have 16 to 25 TMSs, even numbered proteins are more numerous than odd numbered proteins, almost without exception.

The absolute numbers of proteins having various topologies in subclass 1.A are also indicated in Figure 2. HMMTOP predicted 31 proteins to have 0 TMSs, 48 proteins to have 1 TMS, 97 with 2 TMSs, 75 with 3 TMSs, 127 with 4 TMSs, 60 with 5 TMSs, 142 with 6 TMSs, and 78 with 7 TMSs. MEMSAT predicted only 5 proteins to have 0 TMSs, 76 proteins predicted to have 1 TMS, 132 to have 2, 67 to have 3, 131 to have 4, 101 to have 5, 112 to have 6, and 36 to have 7 TMSs. SPOCTOPUS predicted 39 proteins to have 0 TMSs, 41 proteins to have 1, 117 proteins to have 2, 86 proteins to have 3, 144 proteins to have 4, 104 proteins to have 5, 134 proteins to have 6, and 27 to have 7 TMSs. From these numbers it is clear that HMMTOP and SPOCTOPUS reasonably agree, while MEMSAST greatly underpredicts 0 TMS proteins while overpredicting 1 and 2 TMS proteins. This error when using MEMSAT will be discussed below.

Only about 20% of the proteins had more than 7 TMSs regardless of the prediction program used. These observations confirm our earlier conclusion that a majority of channel-forming proteins are small with few TMSs, while carriers and primary active transporters are larger with more TMSs [5] (see below).

0 TMS channels

We first analyzed the proteins predicted to have 0 TMSs. With HMMTOP, fourteen families were represented among the 31 proteins in this category. Of these, three families had 4-5 members each predicted to have 0 TMSs: the Intracellular Chloride Channel (CLIC) Family (1.A.12), the Annexin (Annexin) Family (1.A.31), and the Nucleotide-sensitive Anion-selective Channel (ICln) Family (1.A.47). Three members of the Epithelial Chloride Channel (E-ClC) family and the Poliovirus 2B Viroporin (2B Viroporin) family were represented. Two members of the Brain Acid-soluble Protein Channel (BASP1 Channel) Family (1.A.71), and the Mitochondrial EF Hand Ca^{2+} Uptake Porter/Regulator (MICU) family (1.A.76) were also in this category. Each of the remaining proteins predicted to have 0 TMSs was the only member of its respective family included in TCDB at the time of these studies. Protein families known to be bifunctional, with one function associated with a soluble form and the other function associated with the membrane-integrated channel-forming form, are listed in Table 2. Using SPOCTOPUS, the CLIC family had

9 proteins predicted to have 0 TMSs, while four members each from the E-ClC, Annexin, and ICln families were represented. Three proteins each from the Cation Channel-forming Heat Shock Protein-70 (HSP-70) family and the MICU family lacked predicted TMSs, and 2 proteins from the 2B Viroporin family were represented. Members of the Cation-selective Channel-forming Heat Shock Protein-70 (Hsp70) Family (1.A.33) are predicted to have 0 or 1 TMSs per polypeptide chain. These proteins are normally present as soluble chaperone proteins, but they apparently can insert into the membranes of eukaryotes to form cation-selective channels [28]. This is another example of proteins that can exist in either soluble or membrane-integrated forms (Table 2). Each of the remaining proteins predicted to have 0 TMSs was the only member of its respective family included in TCDB at the time of these studies.

The CLIC family includes proteins that have dual functions; first, they are soluble glutathione-S-transferases, and second, they have the capacity to insert into the membrane to form channels. One of these proteins (1.A.12.3.1) is the bacterial CLIC homologue, stringent starvation protein A, SspA of *E. coli*. Five out of eight CLIC family members were tabulated as being 0 TMS proteins, by HMMTOP, while the other three are predicted to have 1 TMS. SPOCTOPUS predicted all 8 to lack TMSs while MEMSAT predicted all to have 1 TMS. Although the topology of the membrane- integrated form is not yet known, the ambiguous nature of these proteins presumable reflects their ability to exist both in soluble and membrane-integrated forms.

Annexins similarly have 0 putative TMSs according to HMMTOP and SPOCTOPUS, but 1 TMS by MEMSAT. Annexins, structurally conserved, mediate reversible Ca^{2+}-dependent intracellular membrane/phospholipid binding. Like CLIC family members, these proteins can exist in both soluble and membrane-associated forms. Membrane association is critical for their proposed functions that include vesicle trafficking, membrane repair, membrane fusion and ion channel formation [29].

A fourth family with members exhibiting 0 TMSs by HMMTOP and SPOCTOPUS, but not MEMSAT, is the ICln (1.A.47) family. ICln proteins are multifunctional proteins in animals, being essential for cell volume regulation. They are found in the cytosol but are also associated with the cell membrane. They regulate cell volume by activating a swelling-induced Cl^- conductance pathway. ICln reconstituted in artificial bilayers forms ion channels [30]. Cell swelling causes ICln to redistribute from the cytosol to the cell membrane. The coexistence of these proteins as both soluble and membrane- integrated forms again explains the prediction that they exhibit no putative TMSs.

Like members of the three families described above, members of the BASP1 family (1.A.71), lack observed hydrophobic peaks in hydropathy plots, and again, while HMMTOP and SPOCTOPUS predicted 0 TMSs, MEMSAT predicted 1. These proteins become membrane-associated by virtue of myristoylation and show cation-selective ion channel activity in artificial membranes [31]. Thus, the majority of the BASP1 proteins predicted to have 0 TMSs by HMMTOP and SPOCTOPUS exist as soluble proteins that can insert into membranes as a result of lipid derivatization. In summary, and in accordance with other results, HMMTOP and SPOCTOPUS more reliably predict topologies with -0 – 3 TMSs compared to MEMSAT. Several soluble proteins can insert into membranes to form channels. However, the configuration of the polypeptide chains in association with membranes, in general, is not known.

1 TMS Channels

HMMTOP predicted 48 proteins to have 1 TMS, MEMSAT predicted 76 proteins to have 1 TMS, and SPOCTOPUS predicted 41 proteins to have 1 TMS. We found that 0 TMS proteins were consistently predicted to have 1 TMS by MEMSAT, but not by HMMTOP or SPOCTOPUS (see above). All six members of the Phospholemman (PLM) Family (1.A.27) in TCDB were predicted to have a single TMS using HMMTOP, but using SPOCTOPUS, 4 members were predicted to have 1 TMS, and 2 were overpredicted to have 2 TMSs. These proteins are known to have a single TMS [32; 33; 34] and function in a variety of capacities, both as regulators of Na^+, K^+-ATPases and as anion-selective channels Table 3; [35; 36].

Bcl-2 proteins (1.A.21), involved in both necrosis and apoptosis, play both death and anti-death roles in higher eukaryotes [37]. These proteins may have a single C-terminal TMS that serves to anchor them to the membrane, but all three programs predicted more 2 TMS proteins than 1 TMS proteins. Like phospholemmans and their homologues, all members appear to have very similar topologies (Table 3).

The Colicin Lysis Protein (CLP) Family (1.A.73) [38; 39] consists of three members, all of which have a single N-terminal TMS using HMMTOP and MEMSAT; with the SPOCTOPUS algorithm, one member is underpredicted to have 0 TMSs, while the remaining two are predicted to have 1 TMS. In the case of all channel-forming proteins having a single TMS, one can predict that formation of the channel depends upon the formation of oligomeric structures, either homo- or heterooligomers. As noted above, single TMS channel-forming peptides are common, especially within TC subclasses 1.C (pore-forming toxins) and 1.E (holins). Many pore-forming families consist of members that are single peptides of less than 100 residues with a single TMS. They are from viruses and a wide variety of organisms from bacteria to man.

2 TMS Channels

97 proteins were predicted to have 2 TMSs using HMMTOP, 132 proteins with MEMSAT and 117 using SPOCTOPUS. Families with multiple 2 TMS members will be discussed. The first of these is the Voltage-gated Ion Channel (VIC; 1.A.1) family within the VIC superfamily. The channel is formed by tetramers of 2 TMS subunits, each separated by a well-conserved P-loop. The 2 TMS members of the VIC superfamily retrieved in this search were all of this type. Several members of the VIC family and the Inward Rectifier K^+ Channel (IRK-C) family (1.A.2) are predicted to have 3 TMSs. When 3 TMSs are predicted, the moderately hydrophobic P-loop is predicted to be transmembrane, thus explaining the erroneous prediction. Out of 17

proteins in the IRK-C family, HMMTOP predicts only 1 to have 2 TMSs, with 12 proteins predicted to have 3 TMSs; in contrast, both MEMSAT and SPOCTOPUS correctly predicted all 17 proteins to have 2 TMSs. Ion channels of both families can be homo- or heterooligomeric tetrameric structures.

The Epithelial Na^+ Channel (ENaC) Family (1.A.6) and the ATP-gated P2X Receptor Cation Channel (P2X Receptor) Family (1.A.7) are members of a single superfamily, and all members of both families have 2 TMSs separated by a large hydrophilic extra-cytoplasmic domain. They are involved in Na^+ and Ca^{2+} transport. These channels generally exhibit heterotetrameric architecture. Protein members of this superfamily all exhibit the same apparent topology, each with N- and C-termini on the inside of the cell and two amphipathic transmembrane spanning segments, M1 and M2 [40].

The Mer Superfamily (1.A.72) can be split into five families including MerC, MerE, MerH, MerP, and MerT. All five families show sequence similarity within TMSs 1 and 2, but TMSs 3 and 4, when present, are either non-homologous or arose by an intragenic duplication event [41; 42]. These channels all catalyze uptake of Hg^{2+} into bacterial cells in preparation for reduction by mercuric reductase, MerA.

Additional families that exhibit 2 putative TMSs are the Non-selective Cation Channel-2 (NSCC2) Family (1.A.15), the Chloroplast Envelope Anion Channel-forming Tic110 (Tic110) Family (1.A.18), the Bcl-2 (Bcl-2) Family (1.A.21) and the CorA Metal Ion Transporter (MIT) Family (1.A.35). The 2 TMSs in most of these families are in close proximity to one another. An x-ray structure for the *E. coli* CorA protein has established the 2 TMS topology. The Membrane Mg^{2+} Transporter (MMgT) Family (1.A.67) includes members that all have 2 TMSs.

3 TMS Channels

75 proteins in TCDB were predicted by HMMTOP to have 3 TMSs, 67 by MEMSAT, and 86 by SPOCTOPUS. Within the Bacterial Flagellar Motor/Outer Membrane Transport Energizer (MotAB-ExbBD) Superfamily (1.A.30), three were predicted to have 3 TMSs while five proteins were predicted to have 4. In fact, MotA members of the MotAB family have 4 established TMSs while the homologous ExbB and TolQ proteins have 3 TMSs. In the latter proteins, the 3 TMSs correspond to TMSs 2-4 in the former proteins [43].

The Ctr family of copper channels (1.A.56) probably exhibits a uniform topology, which is however, difficult to predict. The hydropathy plot reveals two hydrophobic peaks, the second of which is broad. This peak is predicted to include 1 or 2 TMSs, depending on the protein, but the 3 TMS topology is favored with two TMSs predicted near the C-termini. These eukaryotic proteins can trimerize and harbor a putative copper-binding M-XC-XM-XM motif near their N-termini that is essential for function [44; 45; 46].

4 TMS Channels

Some members of the VIC family contain 4 TMSs per polypeptide chain, and two such proteins form homodimeric channels with four channel-forming units and a total of 8 TMSs per channel. Insufficient sequence similarities make recognition of these P-loops difficult. As in other members of the VIC Superfamily, these P-loops play important roles in ion-selectivity and ion flux control.

14 out of 23 proteins in the Neurotransmitter Receptor, Cys loop, Ligand-gated Ion Channel (LIC) Family (1.A.9) display a correctly predicted 4 TMS topology using HMMTOP, 22/23 using MEMSAT, and 11/23 using SPOCTOPUS. The hydropathy plots reveal four narrow peaks, two of them close to each other and one lone TMS at the N-termini, and another lone TMS at the C-termini. Members of this family have a ligand-binding domain with a number of key residues that are conserved [47]. The five subunits are arranged in a ring with their 'M2' transmembrane helical spanners lining the central channel. They come together in the middle of the membrane to form the channel gate, and the gate opens upon binding acetylcholine or another ligand [48].

Another family that presents a 4 TMS topology is the gap junction-forming Connexin Family (1.A.24). The hydropathy plot suggests a 2 TMS duplication, creating the 4 TMS display. The channels consist of clusters of closely packed pairs of connexins through which small molecules diffuse between neighboring cells. Connexins consist of homo- or heterohexameric arrays of connexins, and the connexin in one plasma membrane docks end-to-end with another connexin in the membrane of a closely opposed cell [49]. The connexin 4 TMS topology is well established.

Similar to members of the Connexin family, gap junction-forming Innexin Family (1.A.25) members are predicted and are known to have 4 TMSs. These proteins form intercellular gap junctional channels primarily in invertebrates that allow electrical coupling and free flow of small molecules between cells. As for the connexins, a 2 TMS duplication probably gave rise to the 4 TMS proteins [50, submitted]. HMMTOP and MEMSAT sometimes erroneously predict a fifth TMS, with HMMTOP the most inaccurate with 8 proteins predicted to have an extra TMS; SPOCTOPUS correctly predicts a 4 TMS topology for all members of this family.

The H^+- or Na^+-translocating Bacterial Flagellar Motor (Mot) Family (1.A.30.1) includes 5 out of 6 TC entries with an established 4 TMS topology correctly predicted by HMMTOP and MEMSAT. SPOCTOPUS correctly predicted all 6 members of the subfamily to have a 4 TMS topology. The hydropathy plot revealed two broad TMSs at both ends of these proteins with a loop in between. These flagellar motor proteins contain clusters of charged residues at both termini, promoting non-covalent interactions between the two components of these motors, MotA and MotB.

Members of the Ca^{2+} Release-activated Ca^{2+} (CRAC) Channel Family (1.A.52) also exhibit a 4 TMS topology. Hydropathy plots predict 4 TMS proteins with large loops between TMSs 3 and 4. When antigens stimulate the immune cells, they trigger Ca^{2+} entry through these tetrameric channels that stimulate the immune response to pathogens. CRAC channel proteins exhibit a teardrop-shape each with a long, tapered cytoplasmic domain. These channels consist of tetramers formed upon Stim-induced dimerization of the Orai subunit [51].

Proteins in the Synaptic Vesicle-associated Ca^{2+} Channel “Flower” Family (1.A.55) were predicted to have 3 or 4 TMSs. Synaptic vesicles promote neurotransmission in presynaptic terminals, regulated by Ca^{2+} [52]. The hydropathy plots for these proteins show two major broad peaks. The first of these peaks is always predicted to consist of 2 TMSs, but the second peak is sometimes predicted to be 1 and sometimes 2 TMSs. One of the family members (Flower) has been shown to have 4 TMSs [52].

5 TMS Channels

60 proteins were predicted to have 5 TMSs using HMMTOP, 101 using MEMSAT, and 104 using SPOCTOPUS. The family with multiple proteins predicted to have 5 TMSs will be discussed in this section.

The “Tweety” Anion Channel Family (1.A.48) is a recently identified family of channel proteins found in animals and plants. Three out of the five TC entries in the family appear to have a 5 TMS topology with HMMTOP, and all 5 members of the family are predicted to have 5 TMSs with both SPOCTOPUS and MEMSAT. These proteins contain 5 (or 6) TMSs in a probable arrangement: 2 + 2 + 1, with an extra N-terminal TMS present in some plant homologues. They produce large conductance chloride currents [53].

6 TMS Channels

142 proteins were predicted by HMMTOP to have 6 TMSs, 112 using MEMSAT and 134 using SPOCTOPUS, making this the largest topological type in class 1.A. Among the proteins predicted to have 6 TMSs, two families predominate: the VIC Family and the Major Intrinsic Protein (MIP) (1.A.8) Family. Using HMMTOP, the VIC superfamily includes 32 TC entries predicted to have 6 TMSs and 15 to have 5. Most or all of the latter were incorrectly predicted and actually have 6. Most of them are K^+ channels, and they usually consist of homotetrameric structures. Many voltage-sensitive K^+ channels function with subunits that modify K^+ channel gating. Non-integral subunits can be homologous to oxidoreductases that co-assemble with the tetrameric channel-forming subunits [54].

Ryanodine-Inositol 1,4,5-triphosphate Receptor Ca^{2+} Channel (RIR-CaC) Family (1.A.3) members have either a 6 or an 8 TMS predicted topology. They are usually homotetrameric complexes. Pore-forming P-loop sequences occur between the fifth and sixth TMSs as for 6 TMS members of the VIC family. The ryanodine channels function in the release of Ca^{2+} from intracellular storage sites in animal cells, thereby regulating various Ca^{2+}-dependent physiological processes. They are members of the VIC superfamily [55; 56].

Seven proteins from the Transient Receptor Potential Ca^{2+} Channel (TRP-CC) Family (1.A.4) present a topology with 6 putative TMSs using HMMTOP, 14 members with MEMSAT and 21 members using SPOCTOPUS. The topological prediction varies with the most common being 5 TMSs. Nevertheless, they all probably have 6 TMSs. This family can be divided into 7 subfamilies that all share a common Ca^{2+} (cation) channel function. As cellular sensors, TRP channels are activated by a variety of different stimuli and function as signal integrators [57].

The VIC family is the dominant family predicted to have 7 TMSs with 31 proteins of the 78 proteins in the category using HMMTOP; SPOCTOPUS and MEMSAT predict 1 and 2 members respectively to have 7 TMSs. However, almost all these predictions are overpredictions. The P-loop between TMSs 5 and 6 is counted as a TMS, erroneously predicting 7 instead of 6 TMSs, especially by HMMTOP.

The Major Intrinsic Protein (MIP) Family of aquaporins and glycerol facilitators (1.A.8) has 55 of 68 members correctly predicted to have 6 TMSs in a 3 + 3 arrangement due to a single intragenic duplication event [58] using HMMTOP, and 64/68 using MEMSAT and SPOCTOPUS. Two proteins were predicted to have 5 TMSs, in one case because TMS 1 was missed, and in the other because TMS 3 was missed; ten proteins were predicted to have 7 TMSs using HMMTOP. One of these proteins, the Major Intrinsic Protein (MIP), makes up about 60% of the proteins in the lens of the eye. During lens development, MIP becomes proteolytically truncated. These truncated tetramers form intercellular adhesive junctions, yielding a crystalline array that mediates lens formation [59].

The Glutamate-gated Ion Channel (GIC) Family of Neurotransmitter Receptors (1.A.10), members of the VIC superfamily, have a topology with one TMS at the N-termini and the remaining 5 TMSs near their C-termini. The extracellular amino terminal domain, S1, and the large extracytoplasmic loop domain between

TMSs 2 and 3, bind the neurotransmitter, which regulates channel formation and ion selectivity [60]. There are three types of GIC receptors [61]. HMMTOP predicted 4 proteins to have 6 TMSs, while MEMSAT and SPOCTOPUS predicted 0 proteins to have 6 TMSs. When analyzing mispredictions, HMMTOP predicted 4 proteins to have 5 TMSs, MEMSAT predicted 7 proteins to have 4 TMSs, and SPOCTOPUS predicted 13 proteins to have 3 TMSs. In the cases of the erroneous predictions, the programs frequently counted the P-loop and another minor peaks that may or may not be TMSs. Four narrow TMSs in the middle of the sequence and a lone TMS at the N-terminus are displayed using the WHAT program for several of these proteins. One protein has the opposite arrangement: the lone TMS is at the C-terminus, while the 4 other putative TMSs are located at a position similar to that of the other proteins. In this case, HMMTOP proved most reliable, followed by MEMSAT and SPOCTOPUS in that order.

Members of the Small Conductance Mechanosensitive Ion Channel (MscS) Family (1.A.23) comprise a group of topologically diverse proteins with a well-characterized function: osmotic adaptation. These proteins are predicted to have 2, 4, 5, 8, 10, 11, 12 and 13 TMSs. The X-ray crystal structure of an *E. coli* MscS allowed prediction of the types of motions these proteins undergo [62; 63]. The structure also provides a framework to address the mechanism of tension sensing that is defined by channel-lipid interactions.

The Urea/Amide Channel (UAC) Family (1.A.29) has 4 members with 6 putative TMSs, using HMMTOP and SPOCTOPUS, and 2 members with 6 putative

TMSs using MEMSAT. These proteins exhibit 6 broad peaks of hydrophobicity corresponding to the six predicted TMSs. These proteins are encoded within operons that also encode ureases or amidases in bacteria.

7 TMS Channels

The Transient Receptor Potential Ca^{2+} Channel (TRP-CC) Family (1.A.4) within the VIC superfamily includes 11 proteins predicted to have 7 TMSs near the C-termini using HMMTOP, 9 using MEMSAT, and 4 using SPOCTOPUS. TRP channels comprise distinct categories of cation channels that are either highly permeable to Ca^{2+}, nonselective, or Ca^{2+} impermeable. Of these proteins, all probably have 6 TMSs; the 7 TMS prediction results because the P-loop is often considered trans-membrane.

The Polycystin Cation Channel (PCC) Family (1.A.5), still another member of the VIC superfamily, has many predicted topologies, but a 6 or 7 TMS topology is probably correct. Polycystin 1 contains 16 polycystic kidney disease l (PKD) domains, one LDL-receptor class A domain and one C-type lectin family domain [64]. These proteins exhibit 1 TMS at their N-termini with the rest at the C-termini.

The Homotrimeric Cation Channel (TRIC) Family (1.A.62) mediates efficient Ca^{2+} mobilization from intracellular stores through Ca^{2+} release channels. They present a topology with an intragenic duplication of a three-TMS polypeptide-encoding genetic element followed by a seventh TMS at their C termini [65].

9 TMS Channels

Members of the Calcium-dependent Chloride Channel (Ca-ClC) Family (1.A.17) are important for the survival of animals. These channels are required for normal electrolyte and fluid secretion, olfactory perception, and neuronal and smooth muscle excitability in animals [66]. They generally have 9 TMSs; an 8 TMS prediction is probably incorrect.

All members of the Presenilin Endoplasmic Reticular Ca^{2+} Leak Channel (Presenilin) Family (1.A.54), accountable for about 40% of familial Alzheimer's disease cases, are predicted to have 9 TMSs [67] using HMMTOP. MEMSAT predicts 5/7 members to have 9 TMSs, and SPOCTOPUS predicts none of the members correctly. All of these proteins have a 9 TMS topology [68]. They resulted from a 3 TMS triplication. A hydrophilic domain follows the first 6 TMSs, and then 3 more TMSs follow. A distant member, Signal peptide peptidase-2A, was predicted to have 8 TMSs, but the HMMTOP program missed an N-terminal TMS. The order of accuracy of the three programs was therefore HMMTOP > MEMSAT > SPOCTOPUS.

11 and 12 TMS Channels

Ammonia Channel Transporter (Amt) Family (1.A.11) members have dual functions, transporting NH_3 or NH_4^+ and regulating nitrogen metabolism by directly interacting with regulatory proteins such as the *E. coli* PII protein and its homologue, GlnK. They are sometimes thought of as gas channels with two structurally similar halves that span the membrane with opposite polarity [69].

HMMTOP predicts 17/28 members to have 11 TMSs, and 11/28 members to have 12 TMSs. MEMSAT predicts 21/28 members to have 11 TMSs and 7/28 members to have 12 TMSs. Finally, SPOCTOPUS predicts 20/28 proteins to have 11 TMSs and 8/28 proteins to have 12 TMSs. The 11 TMSs (M1-M11) of the *E. coli* and archaeal AmtB proteins for which X-ray crystal structures are available form a right-handed helical bundle surrounding each channel [69; 70]. Probably all members of this family have 11 TMSs.

18-25 TMS Channels

The only protein outside of the VIC Superfamily that was predicted to have 20 TMSs is the Kidney Vasopressin Regulated Urea Transporter (1.A.28.1.1) in the Urea Transporter (UT) Family (1.A.28). Most of the UT proteins vary in size from 380-400 residues and exhibit 10 putative TMSs, but mammalian urea transporters such as UT-A1 of the rat are 920-930 residues long and exhibit an internal duplication yielding a total of 20 TMSs.

Many Ca^{2+} and Na^{+} channels of the VIC Superfamily (1.A.1) have 24 TMSs due to quadruplication of a 6 TMS unit, but a few Ca^{2+} channels have 12 TMSs due to duplication. The HMMTOP program mispredicted many of these protein topologies [71]. MEMSAT afforded the best accuracy with these proteins, correctly predicting 8 to have 24 TMSs. SPOCTOPUS predicted only 2 proteins to have 24 TMSs. Errors are generally due to overpredictions; for instance, SPOCTOPUS predicted 13 proteins to have > 25 TMSs, and MEMSAT predicted a total of 10 proteins to have more than 25 TMSs.

>25 TMS Channels

Only one family of channel proteins included members with >25 TMSs. This is the Mechanical Nociceptor Piezo Family (1.A.75). These proteins are believed to be cation-selective channels that mediate responses to noxious mechanical stimuli [72; 73]. The proteins were predicted to have 30, 37, 39, 41 and 43 TMSs with HMMTOP, 26, 35 37, 38, 39, and 40 TMSs using MEMSAT, and 21, 30, 31, 33 and 37 TMSs using SPOCTOPUS. Examination of the largest of these (1.A.75.1.6) revealed that this protein consists of several domains, possibly an internally repeated sequence, each exhibiting about 7 putative TMSs. At the C-termini of these proteins, DUF3595 domains were identified. These proteins can be found in a wide range of eukaryotes including plants, animals, protozoans, slime molds and ciliates, but not in prokaryotes. It is likely that all the programs were poorly predictive for members of this family, but HMMTOP may have predicted most accurately.

Holins (TC Subclass 1.E)

Subclass 1.E includes 53 families of putative Holin proteins (Table 4). This subclass was analyzed collectively for topological types with the TMStats program without auxiliary proteins on 5/29/2013. A total of 323 proteins were analyzed. The average topology as determined by HMMTOP was 2.2 +/- 1.0 S.D. while the average topologies calculated by MEMSAT and SPOCTOPUS were 1.9 +/- 1.0 S.D. and 1.9 +/- 1.1 S.D., respectively.

The distribution was revealed in a plot of predicted topological types with the frequency of occurrence on the Y-axis and the number of TMSs on the X-axis (Figure 3). Interestingly, MEMSAT predicted more proteins with 1 and 2 TMSs, and fewer proteins with 3 and 4 TMSs compared to the other two programs. More proteins were predicted to have 1 TMS by MEMSAT and SPOCTOPUS, but more proteins appeared to have 2 TMSs when HMMTOP was used. Overall the order of topological types was 1 TMS > 2 TMSs > 3 TMSs > 4 TMSs. Members of 4 or 5 families (depending on the prediction algorithm used) had 4 TMSs (Table 4). There are probably no holins with 5 or more TMSs.

1 TMS Holins

The proteins with 1 TMS were analyzed first. Several families were predicted to contain proteins with 1 TMS (Table 4). The first of them is the T4 Holin Family (1.E.8). T4 holin is hydrophilic with 49 acidic and basic residues that promote its function as a holin-endolysin system for cell lysis. The lone TMS resides near the N-terminus. The T7 Holin Family (1.E.6) similarly exhibits a 1 TMS topology, as does the φAdh Holin Family (1.E.12).

The BlyA Holin Family (1.E.17) also exhibits a 1 TMS topology. BlyA and the BlyB soluble accessory protein are encoded on the conserved cp32 plasmid of *Borrelia burgdorferi.* BlyA can promote endolysin-dependent lysis of an induced lambda lysogen that is defective for the lambda holin S gene [74]. The *Pseudomonas aeruginosa* Hol Holin Family (1.E.20) has 1-2 TMSs. Hol by itself, in a broad host-range expression vector under IPTG control, exhibits strong lytic activity, but expression of both Hol and Lys together induces lysis under conditions where neither one alone is effective [75].

2 TMS Holins

The P21 Holin S Family (1.E.1) has two TMSs with both the N- and C-termini on the cytoplasmic side of the inner membrane of *E. coli*. It functions in the export of an endolysin, but the holin channel also allows release of small ions and metabolites, thereby promoting cell death. The HP1 Holin Family (1.E.7) includes members that aid in the release of lysozymes to the peptidoglycan wall. They have 2 broad hydrophobic peaks and a positively charged C-terminus within the short sequence.

The T4 Immunity Holin Family (1.E.9) is best known for its function in blocking DNA entry into the bacterial cytoplasm [76]. Although T4 Holins usually have 2 TMSs, one family member has 3. TMSs 1 and 2 are homologous to the two TMSs of other family members. Members of the *Bacillus subtilis* φ29 Holin Family (1.E.10) with 2 broad hydrophobic peaks aid in cell lysis. φ11 Holin Family (1.E.11) members are hydrophobic peptides with 2 TMSs that similarly exhibit inner membrane disruptive activity. The 2 narrow peaks of hydrophobicity, corresponding to TMSs, are found near the N-termini.

The *Lactococcus lactis* Phage r1t Holin (1.E.18) has 2 TMSs separated by a short β-turn region. The r1t genome includes two adjacent genes, Orf48 and Orf49, encoding a holin and a lysin. The Bacterophase Dp-1 Holin (1.E.24) is encoded with a lytic phage enzyme that shows an operon organization similar to those of *Streptococcus pneumonia* and its bacteriophage [77]. The φU53 Holin (1.E.13) and The ArpQ Holin (1.E.15) both exhibit 2 TMS topologies.

3 TMS Holins

Phage Lambda Holin S (1.E.2) has a 3 TMS topology with the N-terminus in the periplasm and the C-terminus in the cytoplasm. Two products of the same gene have opposite functions: pore formation (S105), and blockage of pore formation (S107). They have 3 evenly spaced TMSs [78], and the single pore formed has a large diameter [79]. The ratio of these two gene products determines the timing of cell lysis. Holin S is expressed at a specific time after phage infection terminates. The 3 TMSs present are evenly spaced.

Another 3 TMS family includes the PRD1 Phage P35 Holin (1.E.5), an element of the holin-endolysin system that lyses the host bacterial cell. The P35 holin has three TMSs with charged residues in the loop regions [80]. Members of the *Listeria* Phage A118 Holin Family (1.E.21) exhibit a 3 TMS topology with broad peaks of hydropathy evenly spaced. Hol118 appears in the cytoplasmic membrane shortly after infection. A second shorter translation product, like the Lambda phage

S105 protein, has a different translational start rate at position 40, lacks the first TMS, and inhibits pore formation [81].

The *Bacillus* Spore Morphogenesis and Germination Holin Family (1.E.23) also has members exhibiting a 3 TMS topology. Involved with spore morphogenesis and germination, its absence results in spores lacking the usual striatal pattern, and the outer coat fails to attach to the underlying inner coat [82]. Other families that include members exhibiting a 3 TMS topology are the P2 Holin TM Family (1.E.3), the LydA Holin Family (1.E.4), and the Cph1 Holin (Cph1 Holin) Family (1.E.16).

4 TMS Holins

Four or five families have holins that appear to contain 4 TMSs according to HMMTOP, MEMSAT and SPOCTOPUS. The most prevalent of these is the LrgA Holin Family (1.E.14). LrgA is a murine hydrolase exporter, and homologues are present in large numbers of bacteria (both Gram-positive and Gram-negative) as well as archaea. These proteins function in programmed cell death that is analogous to apoptosis in eukaryotes [83]. The 4 TMSs arose by duplication of a 2 TMS precursor.

The *Clostridium difficile* TcdE Holin (1.E.19) is a 4 TMS protein. This organism produces two large toxins, both encoded within a pathogenicity locus; the *tcdE* gene is sandwiched in between the two toxin genes [84]. Both toxins may be released via *tcdE*. This action can lead to death to *E. coli* cells.

Pore-forming Toxins (TC subclass 1.C)

411 proteins in TCDB were listed as pore-forming toxins in subclass 1.C as of 5/29/2013. The average number of putative TMSs for these proteins using HMMTOP is 0.68 +/- 0.82, is 1.0 +/- 0.56 using MEMSAT, and is 0.63 +/- 0.68 using SPOCTOPUS. With HMMTOP, 205 proteins were predicted to have 0 TMSs, 151 were predicted to have 1 TMS, 42 to have 2, 9 to have 3, and 4 to have 4. Using MEMSAT, 17 were predicted to have 0 TMSs, 357 were predicted to have 1 TMS, 25 were predicted to have 2 TMSs and 9 were predicted to have 3 TMSs. Finally, with SPOCTOPUS, 193 proteins were predicted to have 0 TMSs, 180 proteins were predicted to have 1 TMS, 29 proteins were predicted to have 2 TMSs, and 6 proteins were predicted to have 3 TMSs (see Figure 4). The proteins predicted to have 4 TMSs are in the two component Bacterial Type III-target Cell Pore (III TPC) Family (1.C.36). As noted above, MEMSAT does not give reliable results for 0, 1 and 2 TMS proteins.

0 TMS Toxins

Examination of the proteins predicted to have 0 TMSs, revealed that many of their hydropathy profiles displayed substantial peaks of hydrophobicity. For example, all members of the Channel-forming δ-Endotoxin Insecticidal Crystal Protein (ICP) Family (1.C.2) exhibit two striking peaks of hydrophobicity near their N-termini while the remainder of these proteins are hydrophilic. Members of the α-Hemolysin Channel-forming Toxin (αHL) Family (1.C.3) exhibit a single N-terminal peak of hydrophobicity, possibly representing the signal sequence for export via the general secretory pathway (3.A.5). Members of the Aerolysin Channel-forming Toxin Family (1.C.4) lack hydrophobic peaks of sufficient magnitude to pass through the membrane as α-helices. Members of the Botulinum and Tetanus Toxin (BTT) Family (1.C.8) exhibit only one hydrophobic peak centrally located in these polypeptide chains. Members of the Pore-forming RTX Toxin Family (1.C.11) are predicted to have zero, one or two TMSs based on hydropathy plots. However, all members of this family exhibit three hydrophobic peaks in their central domains, the first being the smallest and the last one being the largest.

RTX Toxins exhibit tremendously varied sizes, ranging from three hundred residues to about three thousand residues. The same was observed for the Clostridial Cytotoxin (CCT) Family (1.C.57), which is also a member of the RTX-superfamily. Members of the small peptide Magainin (Magainin) Family (1.C.16) were predicted to have either one or zero TMSs, but all of these small proteins exhibit an N-terminal signal sequence, specifying export via the general secretory pathway (3.A.5). These examples confirm that the TMStats program, based on HMMTOP, MEMSAT and SPOCTOPUS, provides approximate values in predicting TMSs but cannot be considered accurate. Every family must be considered separately, as some of these programs are more reliable for some families when others are more reliable for other families. Using the AveHAS program, predictions can be much more accurately verified.

1 TMS Toxins

Most members of the Channel-forming Colicin Family (1.C.1) are predicted to have a single TMS. These proteins exhibit a single broad hydrophobic peak at their extreme C-termini, but in some cases these peaks split into two predicted TMSs. Most members of the Channel-forming ε-toxin Family (1.C.5) exhibit a single N-terminal hydrophobic peak, undoubtedly corresponding to the export signal sequence. Similarly, all members of the Thiol-activated Cholesterol-dependent Cytolysin (CDC) Family (1.C.12) exhibit a single N-terminal signal TMSs. Again, when members of the Membrane Attack Complex/Perforin (MACPF) Family (1.C.39) were examined, a single N-terminal peak of hydrophobicity was observed. These two families belong to a single superfamily and are therefore homologous. Although they are from prokaryotes and eukaryotes, respective, it would appear that both are secreted via the general secretory (Sec) pathway [85]. Further examination of proteins predicted to exhibit a single TMS showed that the majority of these occur at the extreme N-termini of the proteins. Most of these toxins are secreted to the external medium whereupon they undergo massive conformational changes when they insert into the membranes of their target cells.

2 TMS Toxins

Among the proteins that were predicted to have two TMSs were members of the Pore-forming Haemolysin E (HlyE) Family (1.C.10). These family members are about three hundred residues in length. In this family we find proteins with two peaks of hydrophobicity separated by about one hundred residues. Although, all members of this family exhibit two peaks of hydrophobicity, the programs in use do not always predict them to be transmembrane.

The Cecropin (Cecropin) Family (1.C.17) and the Melittin (Melittin) Family (1.C.18) contain members that are predicted to have one or two TMSs. However, all members show two hydrophobic peaks, the first being the targeting signal sequence, and the second being the single TMS in the mature protein that comprises the oligomeric channel [86]. The Pediocin Family (1.C.24), the Lactacin X Family (1.C.26), the Divergicin A Family (1.C.27) and the Bacteriocin AS-48 Cyclic Polypeptide Family (1.C.28), all members being bacteriocins, exhibit similar characteristics with two putative TMSs. The Cecropin and Melittin superfamilies have recently been shown to include proteins that are homologous to each other (A.J. Le and M.H. Saier, unpublished results).

Toxins with >2 TMS

The Bacterial Type III-Target Cell Pore (IIITCP) Family (1.C.36) includes members that exhibit from zero to four predicted TMSs. These systems consist of two nonhomologous proteins, one predicted to have 0 or 1 TMS, while the other is predicted to have 2-4 TMSs. These proteins insert into the membrane of the target animal or plant cell to facilitate injection of bacterial proteins into the eukaryotic cells via a Type III protein secretion system (injectisome). Most of the larger proteins exhibit two striking centrally localized peaks of hydrophobicity, the first broad, probably encompassing two TMSs, and the second sharp, almost always predicted to be a single TMS. We suggest that these toxins to have three TMSs. While the IpAB protein (1.C.36.3.1), is likely to have 3 TMSs, the other, BopB (1.C.36.4.1) is homologous to other members of this family except that it has two additional hydrophobic peaks C-terminal to the usual 3 TMSs, common to all members of this family.

Porters (uniporters, symporters, and antiporters; TC subclass 2.A)

A histogram of predicted topologies generated using the HMMTOP prediction algorithm for all proteins included in TC subclass 2.A revealed that of the 2582 proteins, the average size was 10.5 +/- 2.9 S.D. With the MEMSAT and SPOCTOPUS algorithms, the means and standard deviations of predicted topologies were 10.5 +/- 2.7 S.D. and 10.0 +/- 2.8 S.D. respectively. The largest numbers of proteins, 953, 968, and 765 for the three programs respectively, contain 12 putative TMSs; however, proteins exhibiting 10 to 14 TMSs were prevalent (Figure 5) regardless of the prediction algorithm used. Of the proteins of smaller sizes, there is a peak of proteins exhibiting 6 TMSs, but substantial numbers of proteins display 7 through 9 TMSs with smaller numbers having 4 and 5 putative TMSs when SPOCTOPUS and MEMSAT were used; HMMTOP exhibited the opposite behavior, with greater numbers of proteins displaying 4 and 5 TMSs, and smaller numbers of proteins displaying 7 through 9 TMSs. These proteins were analyzed further.

1 or 2 TMS Porters

Most members of the Mitochondrial Inner Membrane K^+/H^+ and Ca^{2+}/H^+ Exchanger (LetM1) Family (2.A.97) are predicted to have two TMSs by HMMTOP; however, MEMSAT and SPOCTOPUS predict all 4 members of LetM1 to have 1 TMS. These proteins exhibit hydrophobic peaks near their N-termini, but in cases of members of human origin, they only display 1 TMS. These topological features are typical of channels, and this family is the only family of carriers reported to have fewer than 3 TMSs [87]. In our opinions, the claim that these proteins function as carriers should be further investigated.

3 TMS Porters

Putative carriers predicted to have three TMSs by HMMTOP include members of the Mitochondrial tRNA Import Complex (M-RIC) Family (2.A.91; [88], the Bilirubin Transporter (BRT) Family (2.A.65 [89], and the Mitochondrial Pyruvate Carrier (MPC) Family (2.A.105) [90]. MEMSAT and SPOCTOPUS both predict members of all of these families to have 0 and 2 TMSs, respectively. The M-RIC Family contains a protein with over six hundred residues that displays three probable TMSs at its N-terminus. The BRT family consists of a single functionally characterized protein in TCDB, the bilitranslocase, which exhibits 1 N-terminal TMS and two C-terminal TMSs. These proteins should also be reexamined for their channel versus carrier properties, as there are very few putative carriers that have been reported to have just 3 TMSs.

The Mitochondrial Pyruvate Carrier (MPC) Family (2.A.105) contains 7 proteins that each includes 3 putative TMSs [90]. HMMTOP correctly predicts six 3 TMS proteins and one 4 TMS protein, while SPOCTOPUS and MEMSAT predict 0 and 1 proteins to have 3 TMSs, respectively. SPOCTOPUS predicted two 0 TMS proteins, four 1 TMS proteins and one 2 TMS protein while MEMSAT predicted one 1 TMS protein, five 2 TMS proteins, and one 3 TMS protein. Since these proteins are known to have 3 TMSs, this is an example where HMMTOP proves most reliable of the three programs.

4 or 5 TMS Porters

In addition to the superfamilies described in more detail below, several families include members that were predicted to have 4 and 5 TMSs. The Cytochrome Oxidase Biogenesis (Oxa1) Family (2.A.9) consists of 9 proteins in TCDB. One member was predicted to have 3 TMSs, three members were predicted to have 4 TMSs using HMMTOP, three members were predicted to have 5 TMSs, as has been established experimentally for representative members [91], and members were predicted to have 6 TMSs. MEMSAT predicts 1 protein in this family to have 5 TMSs, 6 to have 6 TMSs, and 2 to have 7 TMSs. SPOCTOPUS predicts 1 protein to have 2 TMSs, and 2 proteins to each have 3, 4, 5, and 6 TMSs. Thus, HMMTOP performed best with this family, although no program proved particularly reliable for all members. The 4 TMS Multidrug Endosomal Transporter (MET) Family (2.A.74) includes 5 proteins in TCDB, one predicted to have 3 TMSs, two predicted to have 4 TMSs, and two predicted to have 5 TMSs using HMMTOP. MEMSAT predicts 3 proteins to have 4 TMSs and 2 proteins to have 5 TMSs. SPOCTOPUS correctly predicts 4 proteins to have 4 TMSs, and only 1 protein to have 5 TMSs. The Threonine/Serine Exporter (ThrE) Family (2.A.79) includes a protein (2.A.79.2.1) that is predicted to have 4 TMSs by all three programs, but examination of the WHAT plot, based on HMMTOP, suggests 5. These homologs can exist as full-length 10 TMS proteins or half-sized 5 TMS proteins. The Vitamin Uptake Transporter (VUT or ECF) Family (2.A.88) includes four members predicted to have 4 TMSs by HMMTOP, 0 by MEMSAT, and 3 members by SPOCTOPUS. Most members of this family exhibit 5 or 6 TMSs as predicted by all three programs.

The Cation Diffusion Facilitator (CDF) Family (2.A.4) consists of 34 proteins in TCDB. Fifteen proteins of this family were predicted to have 5 TMSs by HMMTOP, 2 by MEMSAT and 9 by SPOCTOPUS. However, many of these proteins are known to have 6 TMSs with three 2-TMS repeats [51]. Analysis of proteins predicted to have 5 TMSs revealed that every one of these proteins exhibits six hydrophobic peaks, one of which was missed by the various programs. HMMTOP correctly predicted 12 proteins to have 6 TMSs, MEMSAT correctly predicted 29 proteins to have 6 TMSs, and SPOCTOPUS correctly predicted 20 proteins to have 6 TMSs. A single member of the Tellurite-resistance/Dicarboxylate Transporter (TDT) Family (2.A.16) exhibits 5 TMSs with a long hydrophilic C-terminus by HMMTOP. 11 of the remaining members of the family have 10 TMSs with two 5-TMS repeats as displayed by HMMTOP. In comparison, MEMSAT predicts 9 proteins to have 10 TMSs, and

SPOCTOPUS predicts only 3 proteins to have 10 TMSs. The ATP-dependent subtelomeric helicase, RecQ of *Schizosaccharomyces pombe* (2.A.16.2.2; 2100 amino acids), has a 5 TMS N-terminal domain (residues 43-210). It is 94% identical to 2.A.16.2.1, a malate transporter of the same species. It is not known whether RecQ catalyzes transport, but this close similarity certainly suggests a transport function. The Aromatic Acid Exporter (ArAE) Family (2.A.85) includes members of varying predicted topology, the most prevalent of which being 5 and 10 TMSs as predicted by HMMTOP, 5 and 10 TMSs by MEMSAT, and 6 and 8 TMSs by SPOCTOPUS. Other families in which 5 TMS members were identified were in agreement with previously known or predicted topologies.

6 or 7 TMS Porters

Many proteins were found to exhibit 6 TMSs, and 6 TMSs is a common topology for transporters. Most of the proteins predicted to have 6 TMSs belong to families known to consist of 6 TMS members, but a few consist of 7 TMS proteins where one TMS was missed by the prediction programs. A few proteins were identified that belong to families with most members exhibiting more TMSs, but in these cases, the TC entries were shown to be truncated sequences, and the erroneous TC entries were replaced in TCDB by the correct sequences. The same considerations proved to be true for the 7 TMS proteins. As for the 6 TMS proteins, several of the proteins predicted to have 7 TMSs belong to families that have been discussed above.

8 TMS Porters

Proteins with 8 TMSs were much less numerous than 6 TMS proteins, but there were more 8 than 7 TMS proteins regardless of the prediction program used, indicating that the 8 TMSs topology is common among transporters. One such family is the K^+ Transporter (Trk) Family (2.A.38) where an 8 TMS topology has been established for three members of the family [25; 26] as discussed above.

Porters with ≥ 10 TMSs

The following superfamilies with members having multiple TMSs were studied in greater detail:

-The Major Facilitator (MFS) Superfamily (2.A.1 plus others; see "Superfamily" link in TCDB)

-The Amino Acid-Polyamine-Organocation (APC) Superfamily (2.A.3 plus others)

-The Resistance-Nodulation-Cell Division (RND) Superfamily (2.A.6)

-The Drug/Metabolite Transporter (DMT) Superfamily (2.A.7)

-The Mitochondrial Carrier (MC) Superfamily (2.A.29)

-The Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase Superfamily (2.A.66).

The Major Facilitator Superfamily (MFS; 2.A.1)

TMStats seldom showed anomalous predictions for the Major Facilitator Superfamily (MFS; 2.A.1), as noted for the sugar porter family discussed above, regardless of the program used. The MFS includes 652 proteins in TCDB as of 5/29/2013, and according to the TM distribution histogram, it showed an average topology of 12.2 +/- 1.1 TMSs using HMMTOP, 12.0 +/- 1.2 TMSs with MEMSAT and 11.4 +/- 1.7 TMSs with SPOCTOPUS.

474 (72%) MFS carriers were predicted to have 12 TMSs while 63 (10%) were predicted to have 11 and another 57 (9%) were predicted to have 14 TMSs (Figure 5) using HMMTOP. MEMSAT predicted 447 (69%) to have 12 TMSs, 76 (12%) to have 11 TMSs, and 69 (11%) to have 14 TMSs. SPOCTOPUS predicted only 330 (51%) to have 12 TMSs, 81 (13%) to have 11 TMSs and 54 (8%) to have 14 TMSs. Examination revealed that most proteins predicted to have 11 TMSs actually have 12, but the programs missed one of them. Proteins predicted to have 10 TMSs also appeared to have 12 TMSs where the programs missed 2. Those predicted to have 13 TMSs probably had either 12 or 13 TMSs. Most of the MFS permeases predicted to have 14 proved to have 2 extra TMSs separating the two six TMS repeat units [92]. Only one protein each with 8, 17, 18 and 24 TMSs was detected with HMMTOP, and one protein each with 8, 18, 18, and 24 TMSs detected with MEMSAT. SPOCTOPUS detected 20 proteins with 8 TMSs, 1 protein with 17 TMSs and one protein with 24 TMSs. Very few proteins were predicted to have 5, 6 or 7 TMSs by any of the programs, suggesting that few, if any, half sized (6 TMS) MFS proteins are

included in TCDB. However, a family of lysyl tRNA synthetases (9.B.111) includes 5 or 6 TMS N-terminal sequences that are clearly related to the second halves of MFS carriers of the DHA2 (2.A.1.3) family of the MFS. The proteins predicted to have 8 TMSs could be interpreted as 12 TMS proteins. In these proteins, 2 TMSs within each of four hairpin structures were so close together that the programs predicted the structure to be a single TMS rather than two TMSs in each hairpin.

Three proteins were predicted to have 16 TMSs by HMMTOP, one by MEMSAT, and zero by SPOCTOPUS; the proteins predicted to have 16 TMSs by HMMTOP are located within the DHA2 family (2.A.1.3), most members of which are known to have 14 TMSs. Two small peaks of hydrophobicity in the C-terminal region were predicted to be TMSs in these proteins but not in other members of this family. Their actual topology is most likely to be 14 TMSs, but this must be determined experimentally. The single protein predicted to have 14 TMSs by MEMSAT is actually a protein that belongs to the Unidentified Major Facilitator-16 (UMF-16) Family (2.A.1.67) and has 24 TMSs. This protein is known to consist of two fused MFS permeases exhibiting two distinct but related functions. The first is a nitrate:proton symporter, and the second is a nitrate:nitrite antiporter [93]. Proteins predicted to have 17 and 18 TMSs prove to be fusion proteins where in one case, the fusion was to a 5 TMS sensor kinase domain, and in the other case, it was fused to a 6 TMS YedZ domain [94].

The Amino Acid-Polyamine-Organocation (APC) Family (2.A.3)

The Amino Acid-Polyamine-Organocation (APC) family within the APC Superfamily (2.A.3) was studied in some detail because of the anomalous behavior exhibited by some of its members. The superfamily includes 134 proteins in TCDB with an average number of TMSs equal to 12.2 +/- 1.0 TMSs according HMMTOP, 12.3 +/- 1.2 TMSs according to MEMSAT and 11.9 +/- 1.2 TMSs according to SPOCTOPUS. 95 (71%) of the proteins were predicted to have 12 TMSs with HMMTOP, 87 (65%) with MEMSAT and 95 (71%) with SPOCTOPUS. Six (4.5%) were predicted to have 10 TMSs with HMMTOP, 1 (0.8%) with MEMSAT and seven (5.2%) with SPOCTOPUS; these proteins proved to be homologous to the proteins containing 12 TMSs throughout most of their lengths. However, they differ from the proteins with 12 TMSs in that they lack approximately 100 residues containing the two C-terminal TMSs. Out of the proteins predicted to have 10 TMSs, HMMTOP correctly predicted 4/6 to come from the Spore Germination Protein (SGP) Family (2.A.3.9), MEMSAT incorrectly predicted 0/1 SGP proteins, and SPOCTOPUS only predicted 1 SGP protein correctly. Possibly because of the loss of these two TMSs, these proteins have lost their transport function and have become receptors. In this case, HMMTOP proved superior to MEMSAT and SPOCTOPUS.

Seven proteins (5.2%) were predicted to have 11 TMSs with HMMTOP, 9 (6.7%) with MEMSAT and 6 (4.5%) with SPOCTOPUS. Examination revealed that they actually possess 12 hydrophobic peaks, corresponding to predicted TMSs. In each of these putative 11 TMS proteins, all three programs missed a single hydrophobic peak, most frequently, the TMS at their extreme N-termini. The 9 (6.7%) proteins predicted to have 13 TMSs by HMMTOP, 17 proteins (13%) by MEMSAT and 4 proteins (3.0%) by SPOCTOPUS actually have 12 hydrophobic peaks, but the programs predicted an extra TMS in each case. The 14 proteins predicted to have 14 TMSs and one protein predicted to have 15 TMSs by HMMTOP, 2 proteins with 14 TMSs and 12 proteins with 15 TMSs by MEMSAT and 15 proteins with 14 TMSs by SPOCTOPUS are homologs of the proteins displaying 12 TMSs with extensions at their C-termini that contain the extra 2 (and in a few cases 3) putative TMSs. In this case, HMMTOP provided the most accurate predictions.

Assuming a 12 TMS topology with these few exceptions, only 39 out of 134 proteins (30%) were mispredicted by HMMTOP, 47 out of 134 by MEMSAT (35%), and 39 out of 134 by SPOCTOPUS (30%). Thus, while HMMTOP and SPOCTOPUS correctly predicted 84% of the proteins in the MFS, they predicted 70% of APC family members correctly; MEMSAT predicted 65% correctly. However, the situation is more complex. The APC superfamily topological analyses are of particular interest because the high-resolution x-ray structures of several members of this superfamily have been solved. In all cases, these proteins consist of two repeat units of 5 TMSs with two extra TMSs most frequently at the C-termini of these porters. In contrast to the MFS, the two halves of these proteins have an odd number of TMSs and consequent opposite orientations in the membrane. Interestingly, we have been able to provide statistical evidence that the MFS and APC superfamilies are homologous and therefore share a common ancestry [50]. We propose that the 6 TMS repeat unit of the MFS lost a single TMS before the intragenic duplication occurred in APC family members, and subsequently, two additional TMSs arose, possibly by a second duplication of a hairpin structure, either TMSs 7 and 8 or TMSs 9 and 10. These possibilities are under investigation.

The Resistance-Nodulation-Cell Division (RND) Superfamily (2.A.6)

The Resistance-Nodulation-Cell Division (RND) Superfamily (2.A.6) includes 97 proteins in TCDB with an average prediction of 11.6 +/- 1.5 TMSs according to HMMTOP, 11.8 +/- 1.3 TMSs according to MEMSAT, and 11.8 +/- 1.3 TMSs according to SPOCTOPUS. Sixty-eight (70%) were predicted to have 12 TMSs with a clear repeat unit having a 1 + 5 arrangement using HMMTOP, 76 (78%) were predicted to have 12 TMSs by MEMSAT and 77 (79%) were predicted to have 12 TMSs by SPOCTOPUS. Five (5%) homologues were predicted to have 13 TMSs by HMMTOP, and SPOCTOPUs, and 4 (4.1%) homologues were predicted to have 13 TMSs by MEMSAT. They proved to be most similar to the Niemann-Pick C type (NPC) proteins from cellular and acellular slime molds. One of the NPC proteins (2.A.6.6.7) possesses N-terminal domains of about 400 amino acyl residues with 4 extra putative TMSs in a 1+3 arrangement to make a total of 16 TMSs. Ten (10%) were predicted to have 11 TMSs by HMMTOP, and MEMSAT, and 8 (8.2%) were predicted to have 11 TMSs by SPOCTOPUS. These appeared to have the usual 12 TMSs topology. The prediction algorithms missed one TMS in a region where 5 TMSs cluster tightly together.

There were 2 proteins with 6 putative TMSs with HMMTOP, one with MEMSAT and 3 with SPOCTOPUS; they displayed similar placements of hydrophobic regions as that of its counterparts with 12 TMSs. Two of these proteins, the two correctly predicted by HMMTOP, are the SecD and SecF half sized bacterial proteins of *E.coli*, which together comprise a full-length transporter of ill-defined function, but involved in protein secretion via the general secretory (Sec) pathway (TC #3.A.5), possible acting as a membrane-integrated chaperone powered by the proton motive force to facilitate a step-of ATP-independent protein translocation [95]. Another two proteins predicted to have 7 TMSs by HMMTOP and MEMSAT, actually display 12 TMSs, but the two programs missed the last 5 TMSs at their C-termini; SPOCTOPUS did not predict any proteins to have 7 TMSs. Six proteins were predicted to have 9 TMSs with HMMTOP and one protein was predicted to have 9 TMSs with both MEMSAT and SPOCTOPUS. Two proteins displayed 14 TMSs with HMMTOP, while the MEMSAT and SPOCTOPUS programs only predicted one protein to have 14 TMSs. The three programs failed to pick up the second, seventh, and eight TMSs for the proteins predicted to have 9 TMSs. The two proteins displaying 14 putative TMSs with HMMTOP had 2 moderately hydrophobic peaks near their C-termini.

The Drug/Metabolite Transporter (DMT) Superfamily (2.A.7)

Drug/Metabolite Transporter (DMT) Superfamily (2.A.7) members exhibit variable topologies. The superfamily includes 199 proteins with an average size of

8.9 +/- 2.0 TMSs according to HMMTOP, 8.7 +/- 2.2 TMSs with MEMSAT and 8.4 +/- 1.9 TMSs with SPOCTOPUS. Recent bioinformatic data have led to the conclusion that a 2 TMS-encoding genetic element duplicated to 4 TMSs, added one TMS at the N-terminus to give 5 TMSs, and then duplicated to give 10 TMS proteins [96]. The pathway was thus: 2 -> 4 -> 5 ->10 TMSs. In fact, all of these topological types are found among current DMT family members. Of the 199 DMT proteins, 120 (60%) were predicted to have 10 TMSs with HMMTOP, 104 (52%) were predicted to have 10 TMSs by MEMSAT and only 71 (36%) were predicted to have 10 TMSs by SPOCTOPUS. Six (3%) appeared to have 5 TMSs, 17 (8.5%) may have 4, and 2 (1%) have 2 TMSs according to HMMTOP. MEMSAT predicted one to have 5 TMSs (0.5%), 13 (6.5%) to have 4 TMSs and 2 (1%) to have 2 TMSs. SPOCTOPUS predicted 5 to have 5 TMSs (2.5%), 18 to have 4 TMSs (9%) and 2 (1%) to have 2 TMSs. The functions of the 2 TMS proteins are not known, but many bacteria have them, so they are not likely to be artifactual.

Nine proteins were predicted to have 3 TMSs with MEMSAT (none were predicted by HMMTOP or SPOCTOPUS), but careful examination revealed that these proteins probably have 4 TMSs; the N-terminal TMSs were repeatedly missed by MEMSAT. Based on WHAT program analyzes, the proteins predicted to have 6, 7, 8 or 9 TMSs, also appear to have 10 TMSs. The programs missed certain TMSs throughout these proteins. The proteins predicted to have 11 TMSs probably have 10 TMSs as well. However, one protein (2.A.7.11.2) predicted to have 12 TMSs proved to be homologous to the 10 TMS proteins except for an N-terminal extension that introduced two extra TMSs. The observations reported revealed that the order of correct predictions for the DMT superfamily was HMMTOP > SPOCTOPUS > MEMSAT. Thus, while a particular topological prediction program is relatively reliable for some families of transport proteins, it can be less reliable for others, and unreliable for still others.

The Mitochondrial Carrier (MC) Superfamily (2.A.29)

The Mitochondrial Carrier Family (MC; 2.A.29) gave anomalous results when examined with HMMTOP but to a lesser degree when MEMSAT or SPOCTOPUS was used. This family included 129 proteins in TCDB, and according to the TM distribution histogram using HMMTOP, it showed an average of 4.3 +/- 1.9 TMSs. MEMSAT predicted an average topology of 5.9 +/- 0.52 TMSs, and SPOCTOPUS predicted an average topology of 5.74 +/- 1.00 TMSs. Of HMMTOP's results, 8 were predicted to have 0 TMSs; six, 1 TMS; eleven, 2; eight, 3; thirty-four, 4; seventeen, 5; forty, 6 and five, 7. MEMSAT predicted one 2 TMS protein, one three TMS protein, two 4 TMS proteins, one hundred and twenty-three 6 TMS proteins, and two 7 TMS proteins. SPOCTOPUS predicted one 0 TMS protein, one 1 TMS protein, four 2 TMS proteins, one 3 TMS protein, four 5 TMS proteins, and one hundred and eighteen 6 TMS proteins. These proteins are known to consist of 3 repeat units, each having 2 TMSs [97]. We were unable to come up with evidence for exceptions. Only 40 of the 129 proteins (31%) were correctly predicted to have 6 TMSs by the HMMTOP program; in comparison MEMSAT correctly predicted 123/129 (95%) proteins and

SPOCTOPUS correctly predicted 118/129 (91.4%) proteins. These proteins were therefore examined in greater detail to understand the reasons for this tremendous discrepancy.

All seven members of the family predicted to have 0 TMSs by HMMTOP displayed 6 peaks of hydrophobicity with the WHAT program. However, the degrees of hydrophobicity of these peaks were frequently below the threshold for identification of a TMS by HMMTOP, which missed the N-terminal TMS most frequently and the third TMS least frequently. The statistics in terms of percent missed are TMS1: 66%, TMS2: 48%, TMS3: 31%, TMS4: 43%, TMS5: 41%, and TMS6: 41%. Overall, HMMTOP missed 45% of the TMSs for the MC Family.

The Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase Superfamily (2.A.66)

At the time when this study was conducted, the Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Superfamily (2.A.66) included 79 proteins in TCDB with an average size of 12.4 +/- 1.1 TMSs according HMMTOP, 12.0 +/- 1.2 TMSs according to MEMSAT and 11.3 +/- 1.7 TMSs according to SPOCTOPUS. Forty-one (52%) of the proteins were predicted to have 12 TMSs by HMMTOP, thirty-seven (47%) of the proteins were predicted to have 12 TMSs by MEMSAT, and thirty-four (43%) of the proteins were predicted to have 12 TMSs by SPOCTOPUS. These proteins exhibit two duplicated halves of 6 TMSs based on bioinformatic and x-ray crystallography studies [98; 99]. Careful examination of the proteins predicted to have

9, 10, or 11 TMSs, revealed that all appear to have 12 TMSs; the prediction algorithms missed one or more hydrophobic peaks. The proteins predicted to have 13 or 14 TMSs are homologous to the proteins predicted to have 12 TMSs with extensions at either the N- or C-terminal ends of the sequences with either one or two putative TMSs.

Chapter acknowledgement:

The Results chapter of this thesis is being prepared for publication. Reddy Abhinay; Cho Jaehoon; Ling Sam; Reddy Vamsee. The thesis author was the primary investigator and author of this material. This work was supported by NIH grant GM077402 from the National Institute of General Medical Sciences.

Discussion

We have compared nine programs to determine their topological prediction accuracies. Initially, we used four representative families where the protein topologies have been well established. While two of those families were predicted with reasonably high levels of confidence, the other two were not. For three of the four families, the order of accuracy was the same: SPOCTOPUS>MEMSAT>HMMTOP>TOPCONS>PHOBIUS>TMHMM>SVMTOP>DAS>SOSUI. The results indicated that a combination of SPOCTOPUS, MEMSAT and HMMTOP were the best performers, and these were used in subsequent studies. All three prediction algorithms were incorporated into the novel TMStats program. Interestingly, as shown in Table 5, the order of accuracy observed for these three programs is not the same when analyzing different families or superfamilies in subclass 2.A (carriers). The results show that in many cases, the HMMTOP algorithm outperforms its counterparts in prediction accuracy. Thus, when the MFS (2.A.1), APC (2.A.3), DMT (2.A.7), MOP (2.A.66) and MPC (2.A.105) families were examined, HMMTOP outperformed MEMSAT, and MEMSAT usually outperformed SPOCTOPUS. In fact, while HMMTOP was superior in 5 out of the 9 cases tabulated, SPOCTOPUS was superior in 3 and MEMSAT was superior in only 1. When small families of uniform topology were examined, the reverse trend was sometimes observed. Using the sum totals of the results generated in this study, we propose that MEMSAT and SPOCTOPUS often predict small families of proteins with greatest accuracy, as shown when the Sugar Porter, Mitochondrial Carrier (MC) and Potassium

Transporter (Trk) families were examined, but that HMMTOP often excels at predicting larger, topologically heterogeneous superfamilies of proteins. MEMSAT was particularly unreliable in predicting proteins with 0, 1 or 2 TMSs.

Two families that consistently resulted in poor predictions with most of the nine programs initially examined were the Mitochondrial Carrier (MC) and the Potassium Transporter (Trk) families. In the case of the MC family, erroneous predictions resulted from low hydropathy values for the individual TMSs, but even for the fairly hydrophobic peaks, 7 of the 9 programs often grossly underpredicted these TMSs. Only SPOCTOPUS and MEMSAT predicted these proteins with 90% and 80% accuracy, respectively (Table 1 and Table 5). In the case of the Trk family, the errors resulted from the prediction that P-loops, that dip into the membrane on one side primarily and exit on the same side, were often predicted to be transmembrane. Even SPOCTOPUS and MEMSAT had low prediction accuracy, predicting only 50% and 35% of proteins correctly, respectively. Thus, different causes for the errors were observed for these two families. However, seven of the nine programs examined consistently made these same errors. These observations suggest a systematic problem shared by many topology prediction programs. They should provide impetus for the design of novel improved programs that can more accurately predict transmembrane protein topologies; improvements can be made to different prediction models by training new HMMs or SVMs to better understand and predict transmembrane segments. It may also be beneficial to introduce family-specific programs for structural biologists and bioinformaticians to use, although it would be desirable to incorporate the new knowledge gained into a single, generalized program that could determine the best prediction algorithm to use with a given protein.

Table 6 summarizes the distribution of topological types among three types of channels (TC Subclasses 1.A, 1.C and 1.E) as well as carriers (2.A); predictions based on the best three programs are reported. TC subclass 1.A channels include all alpha-type channels, except the small prokaryotic holins of subclass 1.E. Subclass 1.C includes secreted channel-forming toxins, and subclass 2.A includes all secondary carriers. Topological distributions for the same subclasses are shown in Figures 2-5. For subclass 1.A, 2 TMS channels outnumber putative 1 TMS channels, 4 TMS channels outnumber putative 3 TMS channels, and 6 TMS channels outnumber 5 or 7 TMS channels. However, 8 and 9 TMS channels are present in about equal numbers, while putative 11 TMS channels exceed the numbers of 10 TMS channels. Many proteins that are predicted to contain 11 TMSs are members of the Amt (1.A.11) family. The Amt family has been experimentally determined to have either 11 or 12 TMS proteins, with most members of this family containing 11. HMMTOP predicts 12 proteins to have 12 TMSs, which is more than either MEMSAT or SPOCTOPUS, which predict 7 and 8 proteins to have 12 TMSs, respectively. With all three programs, however, the most commonly predicted topology is 11 TMSs reflecting established experimental data, with SPOCTOPUS and MEMSAT predicting 20 and 21 proteins to have 11 TMSs, respectively, and HMMTOP predicting 17 proteins to have 11 TMSs. For the remainder of the graph, the situation where channels of even numbers of TMSs generally outnumber those of odd numbered TMSs reoccurs. Thus, there are more 16 TMS channels than 17 TMS channels, more 18 TMS proteins than 19, and more 20 TMS proteins than 21 (see Table 6 and Figure 2). This observation, in agreement with previously published data [5], confirming the postulate that channels with larger numbers of TMSs arose as a result of intragenic duplication events, also confirming the presence of repeat sequences in many of these channel proteins.

The situation for channel-forming toxins is strikingly different. Table 6 depicts the average percentages and numbers of topological predictions obtained with the top three programs. Upon first examination, it is shown that MEMSAT has predicted a significantly larger number of 1 TMS toxins than either of its counterparts; 357 compared to 151 for HMMTOP and 180 for SPOCTOPUS. Conversely, a substantially lower number of 0 TMS predictions were made by MEMSAT, predicting only 17 out of 408 proteins. This is in stark contrast to HMMTOP and SPOCTOPUS, which predicted 205 and 193 proteins, respectively, to have 0 TMSs. Upon examination of these results, it became clear that MEMSAT consistently overpredicted the number of 1 TMS proteins, and severely underpredicted the number of 0 TMS proteins. Analysis of subclass 1.C proteins using solely SPOCTOPUS and HMMTOP showed that a large percentage of toxins, 49%, were predicted to lack α-helical TMSs; 40% were predicted to have 1 TMS; 8%, 2 TMSs and 2%, 3 TMSs. (see also Table 6 and Figure 4). No toxin was predicted to exhibit more than 4 TMSs. Recalling that these proteins can exist in both soluble and membrane integrated forms, it is not surprising that half of them lack observable TMSs. The structures of most of the membrane-integrated forms are unknown, but some of them integrate as transmembrane β-structured proteins [100]. Most subclass 1.C channels are non-specific, transporting ions and small metabolites as well as proteins in some cases.

Holins exhibit a very distinctive topological pattern. About 4% of proteins were predicted to have 0 TMSs, roughly 35% exhibit 1 TMS, about 30% display 2 apparent TMSs, about 20% have 3 TMSs, and the remainder were predicted to have 4 TMSs (See Table 6 and Figure 3). No holin or putative holin was identified with more than 4 putative TMSs. These results indicate that, in contrast to subclass 1.A channels, there has been little intragenic duplication of transmembrane regions for members of subclass 1.E, although one case of a holin family with members having 2 or 4 TMSs, depending on the protein, proved to be a consequence of a 2 TMS duplication. These observations can be explained since most holins form oligomeric channels of low specificity that evolved to export autolysins. Some holins have been shown to form gigantic pores, where virtually all of the subunits in the cell form the borders [79; 101], but others, called pinholins, form well-defined small pores [102].

Secondary carriers (TC subclass 2.A) exhibit a very different pattern than noted for any of these three subclasses of channel proteins. There are very few proteins predicted to have 0, 1, 2 or 3 TMSs. Of these proteins, 62% proved to be members of the Mitochondrial Carrier Family (2.A.29) and are clearly mispredictions. It can be seen in Figure 5 that the orders of prevalence of proteins with different predicted topologies (different numbers of putative TMSs) are 1 < 2 < 3 < 4 < 5 < 6, and 7 < 8 < 9 < 10 < 11 < 12. Furthermore, there are far fewer proteins predicted to have 1, 2 or 3 TMSs than 4, 5, or 6 TMSs, and there are far fewer proteins predicted to have 7, 8 or 9 TMSs than 10, 11 and 12 TMSs. We believe that the considerable number of proteins predicted to have odd numbers of TMSs is in part, artifactual due to program mispredictions. It is notable that 6 and 12 TMS proteins are the most prevalent types of carriers. Additionally, the order of prevalence of proteins predicted to have 12 or more TMSs is 12 > 13 > 14 > 15 > 16 > 17. Once again, the surprisingly large numbers of proteins with odd numbers of TMSs is at least partially due to program errors, but it is worthy of note that a number of carriers have been confirmed to have odd numbers of TMSs, with 5, 7, 9, 11 and 13 TMSs [16; 23; 103].

These results, taken together, are consistent with the model that simple channels with 1, 2 or 3 TMSs were the precursors of larger channel and carrier proteins that arose by intragenic multiplication events. The predominance of proteins with even numbers of TMSs is in agreement with this model as duplication and quadruplication events occurred more frequently than triplication events, and the most frequent basic repeat unit in many carriers appear to be a single 2 or 3 TMS element [50; 51]. While we and other laboratories have identified the internal repeats in many of these proteins, further research will be required to extend, quantitate and confirm the results of these studies (Figure 6).

We have found that several families of channel-forming proteins have 0 or 1 predicted TMS and function in at least two capacities. One may be a soluble enzymatic catalytic or chaperone function while the other is a membrane-integrated channel-forming function. In other dual function proteins, the secondary function is regulatory, in addition to their primary channel-forming functions. It seems likely that in the former cases, the soluble function evolved as the primary function, while the channel function was secondarily acquired, but in the latter case, channel-formation may have arisen first, and a receptor, signal-transduction or other regulatory function may have evolved secondarily. Several examples of transporters that have evolved receptor or regulatory functions are known, but it is surprising how seldom this functional transition has occurred.

Summarizing, almost all large α-type channels and carriers appear to have evolved from small channel-forming peptides by intragenic multiplication processes; the membrane insertion of soluble proteins to form channels is only occasionally observed, and this pathway for the evolution of carriers seems never to have been taken. Moreover, "once a transporter, always a transporter" seems to be the general rule, violated by only occasional exceptions such as those noted above. Thus, transporters in general evolved as a distinct class of proteins, evolving independently of the other protein classes such as enzymes, structural proteins and most regulatory proteins. The restraints imposed on the evolutionary processes are only now emerging, although the molecular bases for these constraints are not understood [104; 105]. This is a fertile area for future studies.

References

1. Lee, A. G. (2011). Biological membranes: the importance of molecular detail. *Trends in biochemical sciences* **36**, 493-500.

2. Elinder, F., Nilsson, J. & Arhem, P. (2007). On the opening of voltage-gated ion channels. *Physiol Behav* **92**, 1-7.

3. Barabote, R. D., Tamang, D. G., Abeywardena, S. N., Fallah, N. S., Fu, J. Y., Lio, J. K., Mirhosseini, P., Pezeshk, R., Podell, S., Salampessy, M. L., Thever, M. D. & Saier, M. H., Jr. (2006). Extra domains in secondary transport carriers and channel proteins. *Biochimica et biophysica acta* **1758**, 1557-79.

4. Pivetti, C. D., Yen, M. R., Miller, S., Busch, W., Tseng, Y. H., Booth, I. R. & Saier, M. H., Jr. (2003). Two families of mechanosensitive channel proteins. *Microbiology and molecular biology reviews : MMBR* **67**, 66-85, table of contents.

5. Saier, M. H., Jr. (2003). Tracing pathways of transport protein evolution. *Molecular microbiology* **48**, 1145-56.

6. Saier, M. H., Jr. (2000). Families of proteins forming transmembrane channels. *The Journal of membrane biology* **175**, 165-80.

7. Fischer, W. B., Wang, Y. T., Schindler, C. & Chen, C. P. (2012). Mechanism of function of viral channel proteins and implications for drug development. *Int Rev Cell Mol Biol* **294**, 259-321.

8. Saris, N. E., Andersson, M. A., Mikkola, R., Andersson, L. C., Teplova, V. V., Grigoriev, P. A. & Salkinoja-Salonen, M. S. (2009). Microbial toxin's effect on mitochondrial survival by increasing K+ uptake. *Toxicol Ind Health* **25**, 441-6.

9. Tusnady, G. E. & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849-50.

10. Lo, A., Chiu, H. S., Sung, T. Y., Lyu, P. C. & Hsu, W. L. (2008). Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. *J Proteome Res* **7**, 487-96.

11. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology* **305**, 567-80.

12. Cserzo, M., Eisenhaber, F., Eisenhaber, B. & Simon, I. (2002). On filtering false positive transmembrane protein predictions. *Protein Eng* **15**, 745-52.

13. Hirokawa, T., Boon-Chieng, S. & Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378-9.

14. Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. (2009). TOPCONS: consensus prediction of membrane protein topology. *Nucleic acids research* **37**, W465-8.

15. Kall, L., Krogh, A. & Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* **338**, 1027-36.

16. Nugent, T. & Jones, D. T. (2010). Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLoS Comput Biol* **6**, e1000714.

17. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928-9.

18. Saier, M. H., Jr., Tran, C. V. & Barabote, R. D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic acids research* **34**, D181-6.

19. Saier, M. H., Jr., Yen, M. R., Noto, K., Tamang, D. G. & Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic acids research* **37**, D274-8.

20. Zhai, Y. & Saier, M. H., Jr. (2001). A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of

multiply aligned homologous proteins. *Journal of molecular microbiology and biotechnology* **3**, 285-6.

21. Zhai, Y. & Saier, M. H., Jr. (2001). A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *Journal of molecular microbiology and biotechnology* **3**, 501-2.

22. Pao, S. S., Paulsen, I. T. & Saier, M. H., Jr. (1998). Major facilitator superfamily. *Microbiology and molecular biology reviews : MMBR* **62**, 1-34.

23. Jack, D. L., Paulsen, I. T. & Saier, M. H. (2000). The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology* **146 (Pt 8)**, 1797-814.

24. Palmieri, F. (2013). The mitochondrial transporter family SLC25: identification, properties and physiopathology. *Mol Aspects Med* **34**, 465-84.

25. Kato, Y., Sakaguchi, M., Mori, Y., Saito, K., Nakamura, T., Bakker, E. P., Sato, Y., Goshima, S. & Uozumi, N. (2001). Evidence in support of a four transmembrane-pore-transmembrane topology model for the Arabidopsis thaliana Na+/K+ translocating AtHKT1 protein, a member of the superfamily of K+ transporters. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 6488-93.

26. Zeng, G. F., Pypaert, M. & Slayman, C. L. (2004). Epitope tagging of the yeast K(+) carrier Trk2p demonstrates folding that is consistent with a channel-like structure. *The Journal of biological chemistry* **279**, 3003-13.

27. Guan, L., Smirnova, I. N., Verner, G., Nagamori, S. & Kaback, H. R. (2006). Manipulating phospholipids for crystallization of a membrane transport protein. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1723-6.

28. Arispe, N. & De Maio, A. (2000). ATP and ADP modulate a cation channel formed by Hsc70 in acidic phospholipid membranes. *The Journal of biological chemistry* **275**, 30839-43.

29. McNeil, A. K., Rescher, U., Gerke, V. & McNeil, P. L. (2006). Requirement for annexin A1 in plasma membrane repair. *The Journal of biological chemistry* **281**, 35202-7.

30. Ritter, M., Ravasio, A., Jakab, M., Chwatal, S., Furst, J., Laich, A., Gschwentner, M., Signorelli, S., Burtscher, C., Eichmuller, S. & Paulmichl, M. (2003). Cell swelling stimulates cytosol to membrane transposition of ICln. *The Journal of biological chemistry* **278**, 50163-74.

31. Ostroumova, O. S., Schagina, L. V., Mosevitsky, M. I. & Zakharov, V. V. (2011). Ion channel activity of brain abundant protein BASP1 in planar lipid bilayers. *The FEBS journal* **278**, 461-9.

32. Moorman, J. R., Ackerman, S. J., Kowdley, G. C., Griffin, M. P., Mounsey, J. P., Chen, Z., Cala, S. E., O'Brian, J. J., Szabo, G. & Jones, L. R. (1995). Unitary anion currents through phospholemman channel molecules. *Nature* **377**, 737-40.

33. Kowdley, G. C., Ackerman, S. J., Chen, Z., Szabo, G., Jones, L. R. & Moorman, J. R. (1997). Anion, cation, and zwitterion selectivity of phospholemman channel molecules. *Biophysical journal* **72**, 141-5.

34. Cheung, J. Y., Zhang, X. Q., Song, J., Gao, E., Chan, T. O., Rabinowitz, J. E., Koch, W. J., Feldman, A. M. & Wang, J. (2013). Coordinated regulation of cardiac Na(+)/Ca (2+) exchanger and Na (+)-K (+)-ATPase by phospholemman (FXYD1). *Advances in experimental medicine and biology* **961**, 175-90.

35. Geering, K. (2006). FXYD proteins: new regulators of Na-K-ATPase. *American journal of physiology. Renal physiology* **290**, F241-50.

36. Nilius, B., Eggermont, J., Voets, T. & Droogmans, G. (1996). Volume-activated Cl- channels. *General pharmacology* **27**, 1131-40.

37. Arbel, N. & Shoshan-Barmatz, V. (2010). Voltage-dependent anion channel 1-based peptides interact with Bcl-2 to prevent antiapoptotic activity. *The Journal of biological chemistry* **285**, 6053-62.

38. Cavard, D. (2002). Assembly of colicin A in the outer membrane of producing Escherichia coli cells requires both phospholipase A and one porin, but

phospholipase A is sufficient for secretion. *Journal of bacteriology* **184**, 3723-33.

39. Chen, Y. R., Yang, T. Y., Lei, G. S., Lin, L. J. & Chak, K. F. (2011). Delineation of the translocation of colicin E7 across the inner membrane of Escherichia coli. *Archives of microbiology* **193**, 419-28.

40. Gonzales, E. B., Kawate, T. & Gouaux, E. (2009). Pore architecture and ion sites in acid-sensing ion channels and P2X receptors. *Nature* **460**, 599-604.

41. Yamaguchi, A., Tamang, D. G. & Saier, M. H. (2007). Mercury Transport in Bacteria. *Water, air, and soil pollution* **182**, 219-234.

42. Mok, T., Chen, J., Shlykov, M. & Saier, M., Jr. (2012). Bioinformatic Analyses of Bacterial Mercury Ion (Hg2+) Transporters. *Water, Air, & Soil Pollution* **223**, 4443-4457.

43. Yonekura, K., Maki-Yonekura, S. & Homma, M. (2011). Structure of the flagellar motor protein complex PomAB: implications for the torque-generating conformation. *Journal of bacteriology* **193**, 3863-70.

44. Banci, L., Bertini, I., Cantini, F. & Ciofi-Baffoni, S. (2010). Cellular copper distribution: a mechanistic systems biology approach. *Cellular and molecular life sciences : CMLS* **67**, 2563-89.

45. Dumay, Q. C., Debut, A. J., Mansour, N. M. & Saier, M. H., Jr. (2006). The copper transporter (Ctr) family of Cu+ uptake systems. *Journal of molecular microbiology and biotechnology* **11**, 10-9.

46. Petris, M. J. (2004). The SLC31 (Ctr) copper transporter family. *Pflugers Archiv : European journal of physiology* **447**, 752-5.

47. Connolly, C. N. (2008). Trafficking of 5-HT(3) and GABA(A) receptors (Review). *Molecular membrane biology* **25**, 293-301.

48. Thompson, A. J. & Williamson, R. (2010). Protocol for quantitative proteomics of cellular membranes and membrane rafts. *Methods in molecular biology* **658**, 235-53.

49. Maeda, S., Nakagawa, S., Suga, M., Yamashita, E., Oshima, A., Fujiyoshi, Y. & Tsukihara, T. (2009). Structure of the connexin 26 gap junction channel at 3.5 A resolution. *Nature* **458**, 597-602.

50. Reddy, V., Shlykov, M. A., Castillo, R., Sun, E. I. & Saier, M. H., Jr. (2012). The Major Facilitator Superfamily (MFS) Revisited. *The FEBS journal* **IN PRESS**.

51. Matias, M. G., Gomolplitinant, K. M., Tamang, D. G. & Saier, M. H., Jr. (2010). Animal Ca2+ release-activated Ca2+ (CRAC) channels appear to be homologous to and derived from the ubiquitous cation diffusion facilitators. *BMC research notes* **3**, 158.

52. Yao, C. K., Lin, Y. Q., Ly, C. V., Ohyama, T., Haueter, C. M., Moiseenkova-Bell, V. Y., Wensel, T. G. & Bellen, H. J. (2009). A synaptic vesicle-associated Ca2+ channel promotes endocytosis and couples exocytosis to endocytosis. *Cell* **138**, 947-60.

53. He, Y., Ramsay, A. J., Hunt, M. L., Whitbread, A. K., Myers, S. A. & Hooper, J. D. (2008). N-glycosylation analysis of the human Tweety family of putative chloride ion channels supports a penta-spanning membrane arrangement: impact of N-glycosylation on cellular processing of Tweety homologue 2 (TTYH2). *The Biochemical journal* **412**, 45-55.

54. Norris, A. J., Foeger, N. C. & Nerbonne, J. M. (2010). Neuronal voltage-gated K+ (Kv) channels function in macromolecular complexes. *Neuroscience letters* **486**, 73-7.

55. Chang, A. B., Lin, R., Keith Studley, W., Tran, C. V. & Saier, M. H., Jr. (2004). Phylogeny as a guide to structure and function of membrane transport proteins. *Molecular membrane biology* **21**, 171-81.

56. Du, G. G., Sandhu, B., Khanna, V. K., Guo, X. H. & MacLennan, D. H. (2002). Topology of the Ca2+ release channel of skeletal muscle sarcoplasmic reticulum (RyR1). *Proceedings of the National Academy of Sciences of the United States of America* **99**, 16725-30.

57. Latorre, R., Zaelzer, C. & Brauchi, S. (2009). Structure-functional intimacies of transient receptor potential channels. *Quarterly reviews of biophysics* **42**, 201-46.

58. Park, J. H. & Saier, M. H., Jr. (1996). Phylogenetic characterization of the MIP family of transmembrane channel proteins. *The Journal of membrane biology* **153**, 171-80.

59. Gonen, T. & Walz, T. (2006). The structure of aquaporins. *Quarterly reviews of biophysics* **39**, 361-96.

60. Gouaux, E. (2004). Structure and function of AMPA receptors. *The Journal of physiology* **554**, 249-53.

61. Mayer, M. L. (2006). Glutamate receptors at atomic resolution. *Nature* **440**, 456-62.

62. Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. (2002). Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. *Science* **298**, 1582-7.

63. Wang, W., Black, S. S., Edwards, M. D., Miller, S., Morrison, E. L., Bartlett, W., Dong, C., Naismith, J. H. & Booth, I. R. (2008). The structure of an open form of an E. coli mechanosensitive channel at 3.45 A resolution. *Science* **321**, 1179-83.

64. Gallagher, A. R., Germino, G. G. & Somlo, S. (2010). Molecular advances in autosomal dominant polycystic kidney disease. *Advances in chronic kidney disease* **17**, 118-30.

65. Silverio, A. L. & Saier, M. H., Jr. (2011). Bioinformatic characterization of the trimeric intracellular cation-specific channel protein family. *The Journal of membrane biology* **241**, 77-101.

66. Yang, Y. D., Cho, H., Koo, J. Y., Tak, M. H., Cho, Y., Shim, W. S., Park, S. P., Lee, J., Lee, B., Kim, B. M., Raouf, R., Shin, Y. K. & Oh, U. (2008). TMEM16A confers receptor-activated calcium-dependent chloride conductance. *Nature* **455**, 1210-5.

67. Tu, H., Nelson, O., Bezprozvanny, A., Wang, Z., Lee, S. F., Hao, Y. H., Serneels, L., De Strooper, B., Yu, G. & Bezprozvanny, I. (2006). Presenilins form ER Ca2+ leak channels, a function disrupted by familial Alzheimer's disease-linked mutations. *Cell* **126**, 981-93.

68. Laudon, H., Hansson, E. M., Melen, K., Bergman, A., Farmery, M. R., Winblad, B., Lendahl, U., von Heijne, G. & Naslund, J. (2005). A nine-transmembrane domain topology for presenilin 1. *The Journal of biological chemistry* **280**, 35352-60.

69. Khademi, S., O'Connell, J., 3rd, Remis, J., Robles-Colmenares, Y., Miercke, L. J. & Stroud, R. M. (2004). Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 A. *Science* **305**, 1587-94.

70. Andrade, S. L., Dickmanns, A., Ficner, R. & Einsle, O. (2005). Crystal structure of the archaeal ammonium transporter Amt-1 from Archaeoglobus fulgidus. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14994-9.

71. Nelson, R. D., Kuan, G., Saier, M. H., Jr. & Montal, M. (1999). Modular assembly of voltage-gated channel proteins: a sequence analysis and phylogenetic study. *Journal of molecular microbiology and biotechnology* **1**, 281-7.

72. Coste, B., Mathur, J., Schmidt, M., Earley, T. J., Ranade, S., Petrus, M. J., Dubin, A. E. & Patapoutian, A. (2010). Piezo1 and Piezo2 are essential components of distinct mechanically activated cation channels. *Science* **330**, 55-60.

73. Kim, S. E., Coste, B., Chadha, A., Cook, B. & Patapoutian, A. (2012). The role of Drosophila Piezo in mechanical nociception. *Nature* **483**, 209-12.

74. Damman, C. J., Eggers, C. H., Samuels, D. S. & Oliver, D. B. (2000). Characterization of Borrelia burgdorferi BlyA and BlyB proteins: a prophage-encoded holin-like system. *Journal of bacteriology* **182**, 6791-7.

75. Nakayama, K., Takashima, K., Ishihara, H., Shinomiya, T., Kageyama, M., Kanaya, S., Ohnishi, M., Murata, T., Mori, H. & Hayashi, T. (2000). The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. *Molecular microbiology* **38**, 213-31.

76. Labrie, S. J., Samson, J. E. & Moineau, S. (2010). Bacteriophage resistance mechanisms. *Nature reviews. Microbiology* **8**, 317-27.

77. Sheehan, M. M., Garcia, J. L., Lopez, R. & Garcia, P. (1997). The lytic enzyme of the pneumococcal phage Dp-1: a chimeric lysin of intergeneric origin. *Molecular microbiology* **25**, 717-25.

78. Graschopf, A. & Blasi, U. (1999). Molecular function of the dual-start motif in the lambda S holin. *Molecular microbiology* **33**, 569-82.

79. Savva, C. G., Dewey, J. S., Deaton, J., White, R. L., Struck, D. K., Holzenburg, A. & Young, R. (2008). The holin of bacteriophage lambda forms rings with large diameter. *Molecular microbiology* **69**, 784-793.

80. Rydman, P. S. & Bamford, D. H. (2003). Identification and mutational analysis of bacteriophage PRD1 holin protein P35. *Journal of bacteriology* **185**, 3795-803.

81. Vukov, N., Moll, I., Blasi, U., Scherer, S. & Loessner, M. J. (2003). Functional regulation of the Listeria monocytogenes bacteriophage A118 holin by an intragenic inhibitor lacking the first transmembrane domain. *Molecular microbiology* **48**, 173-86.

82. Real, G., Pinto, S. M., Schyns, G., Costa, T., Henriques, A. O. & Moran, C. P., Jr. (2005). A gene encoding a holin-like protein involved in spore morphogenesis and spore germination in Bacillus subtilis. *Journal of bacteriology* **187**, 6443-53.

83. Bayles, K. W. (2003). Are the molecular strategies that control apoptosis conserved in bacteria? *Trends in microbiology* **11**, 306-11.

84. Tan, K. S., Wee, B. Y. & Song, K. P. (2001). Evidence for holin function of tcdE gene in the pathogenicity of Clostridium difficile. *Journal of medical microbiology* **50**, 613-9.

85. Saier, M. H., Ma, C. H., Rodgers, L., Tamang, D. G. & Yen, M. R. (2008). Protein secretion and membrane insertion systems in bacteria and eukaryotic organelles. *Advances in applied microbiology* **65**, 141-97.

86. Bechinger, B. (1997). Structure and functions of channel-forming peptides: magainins, cecropins, melittin and alamethicin. *The Journal of membrane biology* **156**, 197-211.

87. Jiang, D., Zhao, L. & Clapham, D. E. (2009). Genome-wide RNAi screen identifies Letm1 as a mitochondrial Ca2+/H+ antiporter. *Science* **326**, 144-7.

88. Basu, S., Mukherjee, S. & Adhya, S. (2008). Proton-guided movements of tRNA within the Leishmania mitochondrial RNA import complex. *Nucleic acids research* **36**, 1599-609.

89. Passamonti, S., Terdoslavich, M., Margon, A., Cocolo, A., Medic, N., Micali, F., Decorti, G. & Franko, M. (2005). Uptake of bilirubin into HepG2 cells assayed by thermal lens spectroscopy. Function of bilitranslocase. *The FEBS journal* **272**, 5522-35.

90. Herzig, S., Raemy, E., Montessuit, S., Veuthey, J. L., Zamboni, N., Westermann, B., Kunji, E. R. & Martinou, J. C. (2012). Identification and functional expression of the mitochondrial pyruvate carrier. *Science* **337**, 93-6.

91. Sato, T. & Mihara, K. (2009). Topogenesis of mammalian Oxa1, a component of the mitochondrial inner membrane protein export machinery. *The Journal of biological chemistry* **284**, 14819-27.

92. Paulsen, I. T., Brown, M. H., Littlejohn, T. G., Mitchell, B. A. & Skurray, R. A. (1996). Multidrug resistance proteins QacA and QacB from Staphylococcus aureus: membrane topology and identification of residues involved in substrate specificity. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 3630-5.

93. Goddard, A. D., Moir, J. W., Richardson, D. J. & Ferguson, S. J. (2008). Interdependence of two NarK domains in a fused nitrate/nitrite transporter. *Molecular microbiology* **70**, 667-81.

94. von Rozycki, T., Schultzel, M. A. & Saier, M. H., Jr. (2004). Sequence analyses of cyanobacterial bicarbonate transporters and their homologues. *Journal of molecular microbiology and biotechnology* **7**, 102-8.

95. Tsukazaki, T., Mori, H., Echizen, Y., Ishitani, R., Fukai, S., Tanaka, T., Perederina, A., Vassylyev, D. G., Kohno, T., Maturana, A. D., Ito, K. & Nureki, O. (2011). Structure and function of a membrane component SecDF that enhances protein export. *Nature* **474**, 235-8.

96. Lam, V. H., Lee, J. H., Silverio, A., Chan, H., Gomolplitinant, K. M., Povolotsky, T. L., Orlova, E., Sun, E. I., Welliver, C. H. & Saier, M. H., Jr. (2011). Pathways of transport protein evolution: recent advances. *Biological chemistry* **392**, 5-12.

97. Kuan, J. & Saier, M. H., Jr. (1993). The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Critical reviews in biochemistry and molecular biology* **28**, 209-33.

98. He, X., Szewczyk, P., Karyakin, A., Evin, M., Hong, W. X., Zhang, Q. & Chang, G. (2010). Structure of a cation-bound multidrug and toxic compound extrusion transporter. *Nature* **467**, 991-4.

99. Hvorup, R. N., Winnen, B., Chang, A. B., Jiang, Y., Zhou, X. F. & Saier, M. H., Jr. (2003). The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. *European journal of biochemistry / FEBS* **270**, 799-813.

100. Berne, S., Sepcic, K., Anderluh, G., Turk, T., Macek, P. & Poklar Ulrih, N. (2005). Effect of pH on the pore forming activity and conformational stability of ostreolysin, a lipid raft-binding protein from the edible mushroom Pleurotus ostreatus. *Biochemistry* **44**, 11137-47.

101. Dewey, J. S., Savva, C. G., White, R. L., Vitha, S., Holzenburg, A. & Young, R. (2010). Micron-scale holes terminate the phage infection cycle. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 2219-23.

102. Pang, T., Savva, C. G., Fleming, K. G., Struck, D. K. & Young, R. (2009). Structure of the lethal phage pinhole. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18966-71.

103. Young, G. B., Jack, D. L., Smith, D. W. & Saier, M. H., Jr. (1999). The amino acid/auxin:proton symport permease family. *Biochimica et biophysica acta* **1415**, 306-22.

104. Norris, V., den Blaauwen, T., Cabin-Flaman, A., Doi, R. H., Harshey, R., Janniere, L., Jimenez-Sanchez, A., Jin, D. J., Levin, P. A., Mileykovskaya, E., Minsky, A., Saier, M., Jr. & Skarstad, K. (2007). Functional taxonomy of bacterial hyperstructures. *Microbiology and molecular biology reviews : MMBR* **71**, 230-53.

105. Norris, V., den Blaauwen, T., Doi, R. H., Harshey, R. M., Janniere, L., Jimenez-Sanchez, A., Jin, D. J., Levin, P. A., Mileykovskaya, E., Minsky, A., Misevic, G., Ripoll, C., Saier, M., Jr., Skarstad, K. & Thellier, M. (2007). Toward a hyperstructure taxonomy. *Annual review of microbiology* **61**, 309-29.

Appendix

Table 1. Comparison of 9 topology prediction algorithms to evaluate prediction accuracy.

A. Major Facilitator (MFS): Sugar Porter Family (84 Proteins)

Proteins predicted to have the specified #s of TMSs

TC# 2.A.1.1 (Actual: 12 TMSs)	**# TMSs:**	**6**	**N/A**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
HMMTOP		0	0	0	0	0	4	3	77	0
SVMTOP		0	0	0	0	0	5	27	49	3
TMHMM		0	0	0	2	4	14	10	54	0
DAS		0	0	1	1	4	6	28	36	8
SOSUI		2	0	1	3	4	13	31	29	1
TOPCONS		0	0	0	0	1	7	3	75	0
PHOBIUS		1	0	1	3	20	61	0	0	0
MEMSAT		0	0	0	0	2	2	2	78	0
SPOCTUPUS		0	0	0	0	1	0	3	80	0

B. Amino Acid-Polyamine-Organocation (APC) (91 Proteins)

TC # 2.A.3 (Actual: 12 TMSs[1])	**# TMSs:**	**8**	**N/A**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
HMMTOP		0	0	0	4	5	69	6	6	1
SVMTOP		1	0	0	5	10	52	16	6	1
TMHMM		0	0	1	9	17	55	3	5	1
DAS		0	0	0	5	13	44	22	6	1
SOSUI		0	0	3	6	28	39	10	5	0
TOPCONS		0	0	0	6	1	74	3	7	0
PHOBIUS		0	0	1	9	6	62	5	7	1
MEMSAT		1	0	1	4	4	70	4	2	5
SPOCTUPUS		0	0	1	6	4	61	4	15	0

C. Mitochondrial Carrier (MC) (88 Proteins)

TC # 2.A.29 (Actual: 6 TMSs)	**# TMSs:**	**0**	**N/A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**
HMMTOP		9	0	2	7	8	25	12	23	2	0
SVMTOP		55	0	0	2	6	9	8	6	1	1
TMHMM		50	0	7	13	11	5	2	0	0	0
DAS		22	0	23	16	18	6	2	1	0	0
SOSUI		73	0	1	8	3	3	0	0	0	0
TOPCONS		66	0	0	3	8	8	2	1	0	0
PHOBIUS		46	0	11	15	10	4	1	1	0	0
MEMSAT		0	0	1	2	0	5	12	68	0	0
SPOCTUPUS		0	0	3	1	1	0	4	79	0	0

D. Potassium Transporter (Trk) (20 Proteins)

TC # 2.A.38 (Actual: 8 TMSs[2])	**# TMSs:**	**3**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
HMMTOP		0	0	0	3	3	6	4	4
SVMTOP		1	0	0	0	0	6	7	6
TMHMM		0	0	0	3	5	8	3	1
DAS		0	0	0	0	2	6	7	5
SOSUI		0	0	1	1	2	10	3	3
TOPCONS		0	0	0	2	3	8	5	0
PHOBIUS		0	0	0	1	3	6	6	4
MEMSAT		0	3	3	7	1	5	1	0
SPOCTUPUS		0	0	4	10	2	6	0	0

[1] **Four members of the APC superfamily in TCDB are believe to have 10 TMSs, 6 have 14 TMSs, and 1 has 15 TMSs (See APC family description).**

[2] **One member of the Trk family in TCDB (2.A.38.1.5) is believed to have 10 TMSs, with 2 extra N-terminal TMSs preceding the four repeats.**

Table 2. Families of bi- or multi- functional proteins that exist in both soluble and membrane-integrated channel-forming states

					Function(s)	Function(s)	
TC #	**Family Name**	**Family Abbreviation**	**Average Size (# of aa)**	**Predicted # TMSs**	**Soluble Form**	**Membrane Form**	**Organismal Types**
1.A.12	Intracellular Chloride Channels	CLIC	437	0-2 TMSs	Glutathione S-transferase	Intracellular membrane voltage-sensitive Cl- channels	Ubiquitous channel-formation in animals
1.A.31	Annexin	Annexin	333 or 666	0 TMSs	Vesicle trafficking	Ca^{2+} preferring animal ion channels	Eukaryotes
1.A.33	Cation Channel-forming Heat Shock Protein-70	HSP70	640	0-1 TMS	Chaperones	Cation-selective channels	Ubiquitous; channel activity demonstrated in animals
1.A.47	Nucleotide-sensitive Anion-selective Channel	ICln	228	0 TMSs	Methylosome subunit, inhibits snRNP formation	Cation-selective channels	Eukaryotes
1.A.71	Brain Acid-soluble Protein Channel	BASP1	233	0 TMSs	Lipid raft composition control	Cation-selective channels	Eukaryotes
1.A.76	Mitochondrial EF Hand Ca^{2+} Uptake Porter	MICU	442	0-1 TMSs	Glutamine Amido-transferase/ anthanylate phosphoribo-syl transferase	Ca^{2+} uptake porter	Eukaryotes

Table 3. Bifunctional proteins that can form integral membrane channels.

TC #	Family Name	Family Abbreviation	Average Size (# of aa)	Predicted # TMSs	Function 1 Non-Channel	Function 2 Channel	Organism Type
1.A.15	Non-Selective Cation Channel-2	NSCC	~340	2 TMSs	Component of the general secretory pathway, Sec 62.	Cation-specific channels	Eukaryotes
1.A.21	Bcl-2	Bcl-2	~230	1 TMS	Cell death	Cell anti-death	Mammalian cells
1.A.27	Phospholemman	PLM	178	1 TMS	Na^{+}, K^{+}-ATPase regulator	Intracellular organic and inorganic anion channels	Animals
1.A.37	CD20 Ca^{2+} Channel	CD20	~300	4 TMSs	IgE receptor,β-subunit	Ca^{2+} channels	Animals
1.A.48	Anion Channel-forming Bestrophin	Bestrophin	572	4 TMSs	Regulates L-type Ca^{2+} Channels	Anion channels	Animals
1.A.50	Phospholamban	PLB	52	1 TMS	Regulates Ca^{2+} ATPase	Ca^{2+} channels	Animals
1.A.54	Presenilin E.R. Ca2^{+} Leak Channel	Presenilin	444	9 TMSs	Protease (produces amyloid peptides)	Ca^{2+} channels	Ubiquitous
1.A.64	Plasmolipin	Plasmolipin	180	4 - 8 TMSs	Regulates NKCC2 (TC # 2.A.30.1.1); stabilizes kidney apical membranes; facilitates protein sorting	Voltage-dependent K^{+} channels	Animals

Table 4. Topological Distribution of Holins in TCDB According to Family

TC #	Family Name	Predicted Topologies (HMMTOP)	Predicted Topologies (MEMSAT)	Predicted Topologies (SPOCTOPUS)	Likely Topology
1.E.1	P21 Holin S Famiy	1 - 2	1	1	1
1.E.2	λ Holin S Family	1 - 3	1 - 3	0 - 3	2 - 3
1.E.3	P2 Holin TM (P2 Holin) Family	2 - 3	2 - 3	3	3
1.E.4	LydA Holin (LydA Holin) Famil	2 - 3	1 - 2	2	2 - 3
1.E.5	PRD1 Phage P35 Holin (P35 Holin) Family	2 - 3	1 - 3	1 - 3	3
1.E.6	T7 Holin (T7 Holin) Family	1 - 2	1	1	1
1.E.7	HP1 Holin (HP1 Holin) Famil	1	1	1	1
1.E.8	T4 Holin (T4 Holin) Family	1 - 2	1	1	1
1.E.9	T4 Immunity (T4 Imm) Family	1 - 3	1 - 2	1 - 3	2
1.E.10	Bacillus subtilis φ29 Holin (φ29 Holin) Family	2 - 3	3	3	3
1.E.11	φ11 Holin (φ11 Holin) Family	1 - 2	1 - 2	1 - 2	2
1.E.12	φAdh Holin (φAdh Holin) Family	0 - 1	1	0 - 1	1
1.E.13	Firmicute phage φU53 Holin (φU53 Holin) Family	3	2	3	2
1.E.14	LrgA Holin (LrgA Holin) Family	3 - 4	3 - 4	4	4
1.E.15	ArpQ Holin (ArpQ Holin) Family	2	2	2	2
1.E.16	Cph1 Holin (Cph1 Holin) Family	3	2 -3	2 - 3	3
1.E.17	BlyA Holin (BlyA Holin) Family	1	1	1	1
1.E.18	Lactococcus lactis Phage r1t Holin (r1t Holin) Family	0 - 2	1 - 2	1 - 2	2
1.E.19	Clostridium difficile TcdE Holin (TcdE Holin) Family	1 - 4	3 - 4	2 - 4	4
1.E.20	Pseudomonas aeruginosa Hol Holin (Hol Holin) Family	3	1 - 2	1 - 2	1 - 2
1.E.21	Listeria Phage A118 Holin (Hol118) Family	3	2 - 3	2 - 3	3
1.E.22	Neisserial Phage-associated Holin (NP-Holin) Family	1	1	1	1
1.E.23	Bacillus Spore Morphogenesis and Germination Holin (BSH) Famil	3	2	2 - 3	3
1.E.24	Bacterophase Dp-1 Holin (Dp-1 Holin) Family	2	1 - 2	1 - 2	2
1.E.25	Pseudomonas phage F116 Holin (F116 Holin) Family	1 - 2	1 - 2	1 - 2	2
1.E.26	Holin LLH (Holin LLH) Family	1 - 2	1	0 - 1	1

Table 4. Continued.

TC #	Family Name	Predicted Topologies (HMMTOP)	Predicted Topologies (MEMSAT)	Predicted Topologies (SPOCTOPUS)	Likely Topology
1.E.27	BlhA Holin (BlhA Holin) Family	1	1	1	1
1.E.28	Streptomyces aureofaciens Phage Mu1/6 Holin (Mu1/6 Holin) Fan	2	1 - 2	0 - 2	2
1.E.29	Holin Hol44 (Hol44) Family	2 - 3	1 - 2	1 - 3	2
1.E.30	Vibrio Holin (Vibrio Holin) Family	1	1	0	1
1.E.31	SPP1 Holin (SPP1 Holin) Family	1 - 2	1 - 2	0 - 2	2
1.E.32	Actinobacterial 1 TMS Holin (A-1 Holin) Family	3 - 1	1 - 2	1 - 2	1 - 2
1.E.33	2 or 3 TMS Putative Holin (2/3 Holin) Family	2 - 3	1 - 3	1 - 3	2 - 3
1.E.34	Putative Actinobacterial Holin-X (Hol-X) Family	2	2	2	2
1.E.35	Mycobacterial 1 TMS Phage Holin (M1 Hol) Family	0 - 1	1	0	1
1.E.36	Mycobacterial 2 TMS Phage Holin (M2 Hol) Family	1 - 5	1 , 2, or 4	0 - 4	2
1.E.37	Phage T1 Holin (T1 Holin) Family	1	1	1	1
1.E.38	Staphylococcus phage P68 Putative Holin (P68 Hol) Family	2	2	1	2
1.E.39	Mycobacterial Phage PBI1 Gp36 Holin (Gp36 Hol) Family	2	2	2	2
1.E.40	The Mycobacterial 4 TMS Phage Holin (MP4 Holin) Family	4	2 - 4	2 - 5	4
1.E.41	Deinococcus/Thermus Holin (D/T-Hol) Family	3	2	3	3
1.E.42	Putative Holin-like Toxin (Hol-Tox) Family	1	1	1	1
1.E.43	Putative Transglycosylase-associated Holin (T-A Hol) Family	3 - 4	2 - 3	3 - 4	3
1.E.44	The Putative Lactococcus lactis Holin (LLHol) Family	1 - 2	1 - 2	1 - 2	1 - 2
1.E.45	Xanthomonas Phage Holin (XanPHol) Family	2	1	1	1
1.E.46	Prophage Hp1 Holin (Hp1Hol) Family	1	2	1	1
1.E.47	Caulobacter Phage Holin (CauHol) Family	2	2	2	2
1.E.48	Enterobacterial Holin (EBHol) Family	1	1	1	1
1.E.49	Putative Treponema 4 TMS Holin (Tre4Hol) Family	0, 1, or 4	1, 3, or 4	1, or 4	4
1.E.50	Beta-Proteobacterial Holin (BP-Hol) Family	1	1	1	1
1.E.51	Putative Listeria Phage Holin (LP-Hol) Family	1	1	1 - 2	1
1.E.52	Flp/Fap Pilin Putative Holin (FFPP-Hol) Family	1 - 2	1	1	1
1.E.53	Toxic Hok/Gef Protein (Hok/Gef) Family	0 - 1	1 - 2	1	1

Table 5. Prediction algorithm accuracy rank for selected superfamilies in subclass 2.A

Family/Superfamily	TC ID	Program #1 Program #2 Program #3	Notes
MFS Superfamily	2.A.1	HMMTOP > MEMSAT > SPOCTOPUS	HMMTOP correctly predicted 72% of proteins
Sugar Porter Family	2.A.1.1	SPOCTOPUS > MEMSAT > HMMTOP	SPOCTOPUS correctly predicted 95% of proteins
APC Superfamily	2.A.3	HMMTOP = SPOCTOPUS > MEMSAT	HMMTOP and SPOCTOPUS both correctly predicted 70% of proteins
RND Superfamily	2.A.6	SPOCTOPUS > MEMSAT > HMMTOP	SPOCTOPUS correctly predicted 79% of proteins
DMT Superfamily	2.A.7	HMMTOP > MEMSAT > SPOCTOPUS	HMMTOP correctly predicted 52% of proteins
Trk Family	2.A.38	SPOCTOPUS > MEMSAT > HMMTOP	SPOCTOPUS correctly predicted 50% of proteins
MC Superfamily	2.A.29	MEMSAT > SPOCTOPUS > HMMTOP	MEMSAT correctly predicted 95% of proteins
MOP Superfamily	2.A.66	HMMTOP > MEMSAT > SPOCTOPUS	HMMTOP correctly predicted 52% of proteins
MPC Family	2.A.105	HMMTOP > MEMSAT > SPOCTOPUS	HMMTOP correctly predicted 100% of proteins

Table 6. Distribution of topological types among analyzed subclasses from all three prediction programs.

# TMSs	1.A				1.C				1.E				2.A			
	HMMTOP	MEMSAT	SPOCTOPUS	Average %	HMMTOP	MEMSAT	SPOCTOPUS	Average %	HMMTOP	MEMSAT	SPOCTOPUS	Average %	HMMTOP	MEMSAT	SPOCTOPUS	Average %
0	31	5	39	4	205	17	193	49	7	0	16	4	8	1	4	0
1	48	76	41	5	151	357	180	40	92	147	134	35	8	7	12	0
2	97	132	117	14	43	25	29	8	100	103	85	30	17	12	14	1
3	75	67	86	9	9	9	6	2	88	47	54	20	23	12	21	1
4	127	131	144	16	4	0	0	0	35	26	33	10	79	47	55	2
5	60	101	104	11					1	0	1	0	98	58	88	3
6	142	112	134	16									112	236	230	7
7	78	36	27	6									88	69	101	3
8	33	32	23	4									66	96	158	4
9	29	33	19	3									103	157	211	6
10	15	8	5	1									314	257	330	12
11	32	36	32	4									325	312	329	13
12	17	7	8	1									953	968	765	35
13	1	3	2	0									172	164	93	6
14	0	0	0	0									150	131	112	5
15	1	0	1	0									38	26	33	1
16	3	0	1	0									10	7	3	0
17	2	2	0	0									6	0	2	0
18	7	1	0	0									4	1	0	0
19	2	0	0	0									3	0	0	0
20	5	1	1	0									1	0	0	0
21	2	1	2	0									2	3	3	0
22	5	1	3	0									0	0	1	0
23	3	10	10	1									0	0	0	0
24	2	8	2	0									1	1	0	0
25	2	4	10	1									1	1	1	0
26	1	6	0	0												
27	0	1	3	0												
28	0	0	0	0												
29	0	0	0	0												
30	1	0	1	0												
31	0	0	3	0												
32	0	0	0	0												
33	0	0	1	0												
34	0	0	0	0												
35	0	1	0	0												
36	0	0	0	0												
37	1	1	1	0												
38	0	2	0	0												
39	2	1	0	0												
40	0	1	0	0												
41	1	0	0	0												
42	0	0	0	0												
43	1	0	0	0												

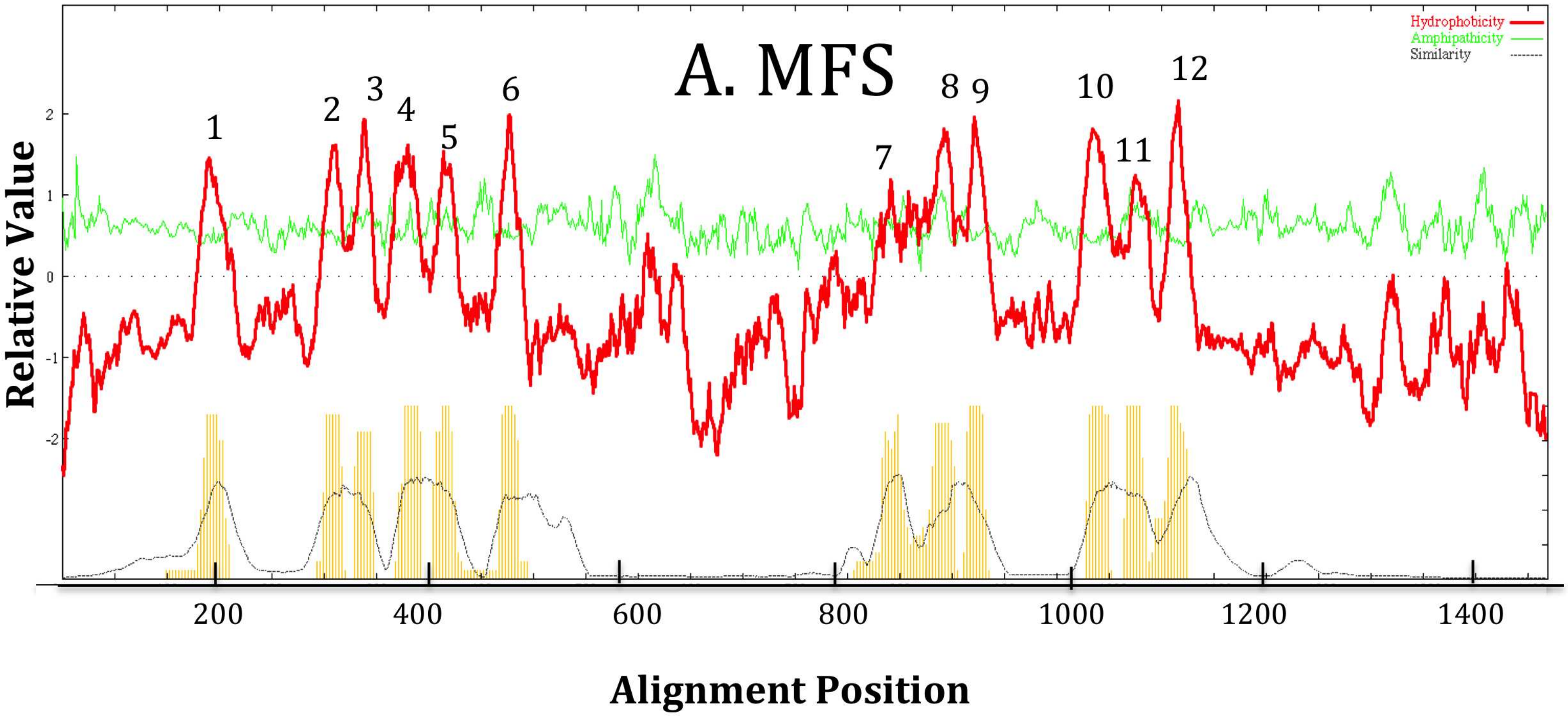


Figure 1A: Average hydropathy, amphipathicity, and similarity plots obtained using the AveHAS program for the Sugar Porter Family in the MFS. The upper panels show the average hydropathy plots (black lines) and the average amphipathicity plots (grey lines) while the lower panels show independently predicted hydropathy plots (vertical lines) and average similarity plots (dotted lines).

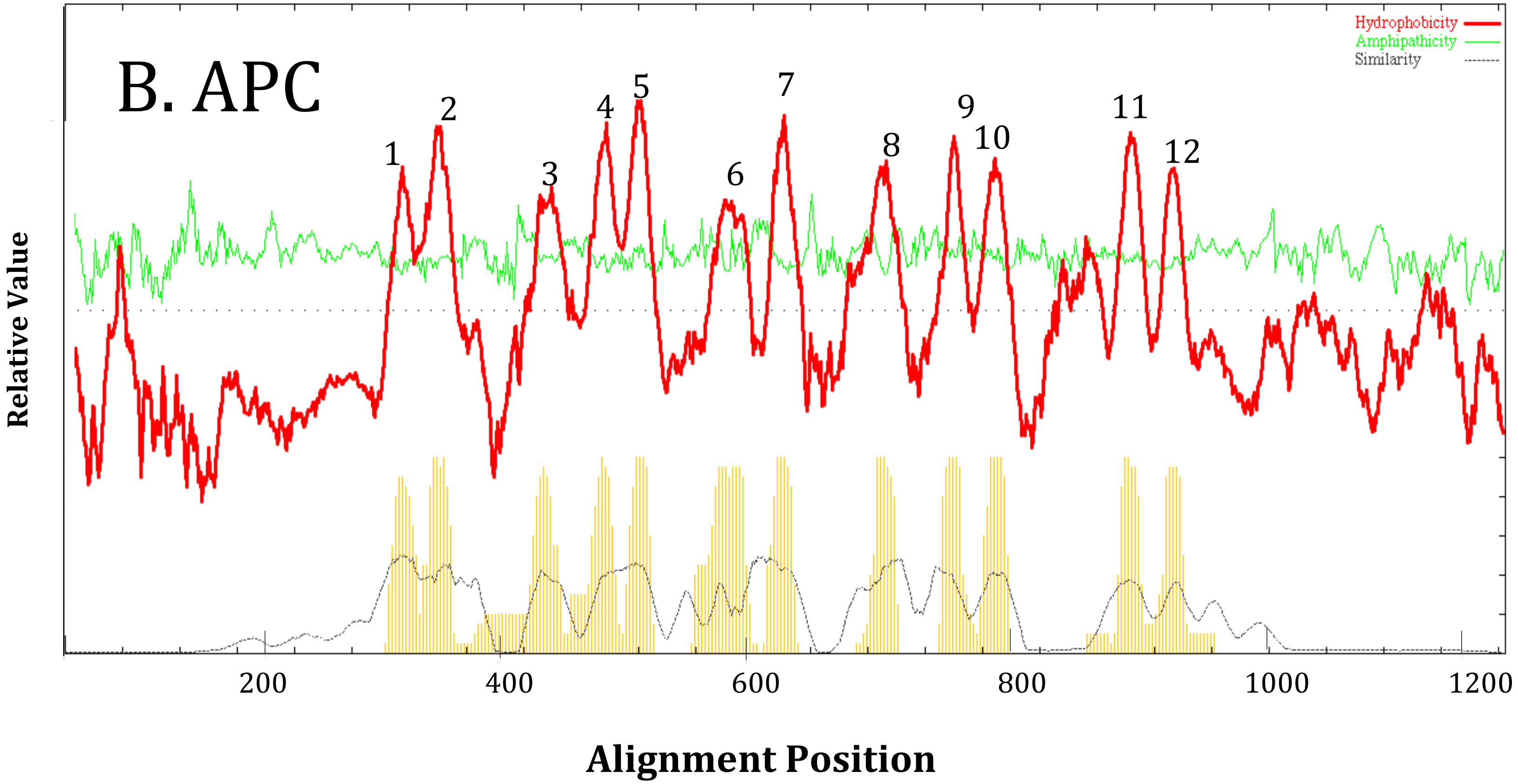


Figure 1B: Average hydropathy, amphipathicity, and similarity plots obtained using the AveHAS program for the APC Family within the APC Superfamily. The upper panels show the average hydropathy plots (black lines) and the average amphipathicity plots (grey lines) while the lower panels show independently predicted hydropathy plots (vertical lines) and average similarity plots (dotted lines).

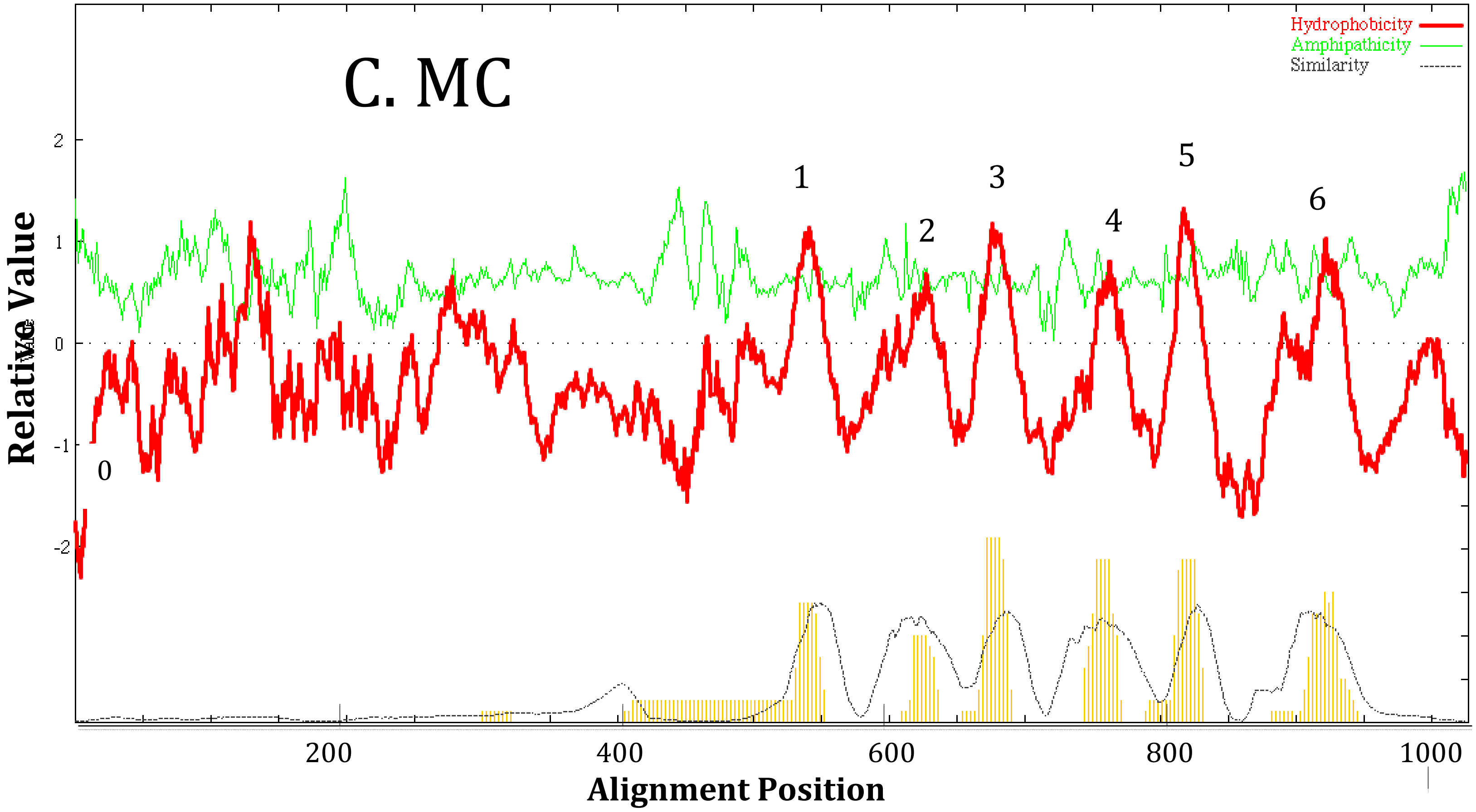


Figure 1C: Average hydropathy, amphipathicity, and similarity plots obtained using the AveHAS program for the MC Family within the MC Superfamily. The upper panels show the average hydropathy plots (black lines) and the average amphipathicity plots (grey lines) while the lower panels show independently predicted hydropathy plots (vertical lines) and average similarity plots (dotted lines). Short vertical lines of uniform length as shown between residue positions 420 and 530 in Figure C and residue positions 200 and 780 in Figure D show regions of poor conservation.

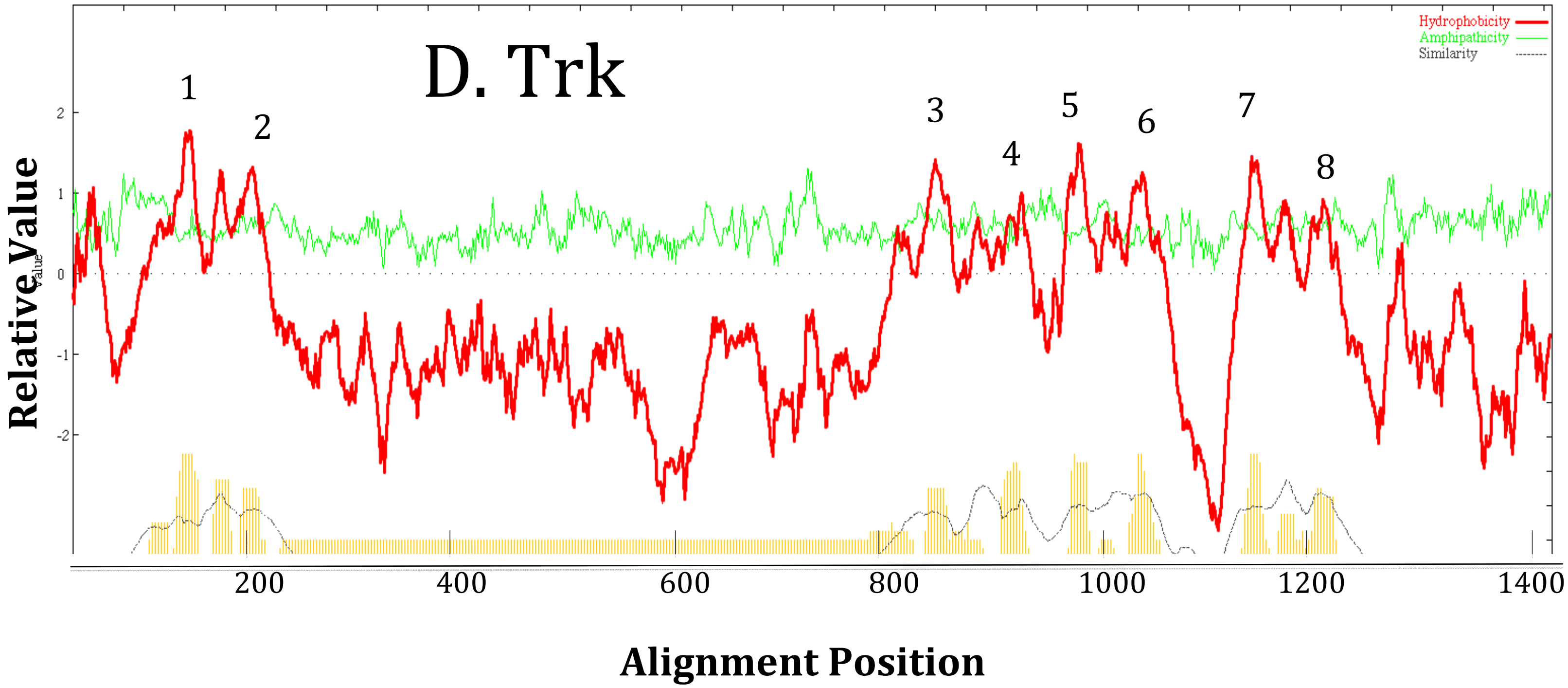


Figure 1D: Average hydropathy, amphipathicity, and similarity plots obtained using the AveHAS program for the Trk Family within the VIC Superfamily. The upper panels show the average hydropathy plots (black lines) and the average amphipathicity plots (grey lines) while the lower panels show independently predicted hydropathy plots (vertical lines) and average similarity plots (dotted lines). Short vertical lines of uniform length as shown between residue positions 420 and 530 in Figure C and residue positions 200 and 780 in Figure D show regions of poor conservation.

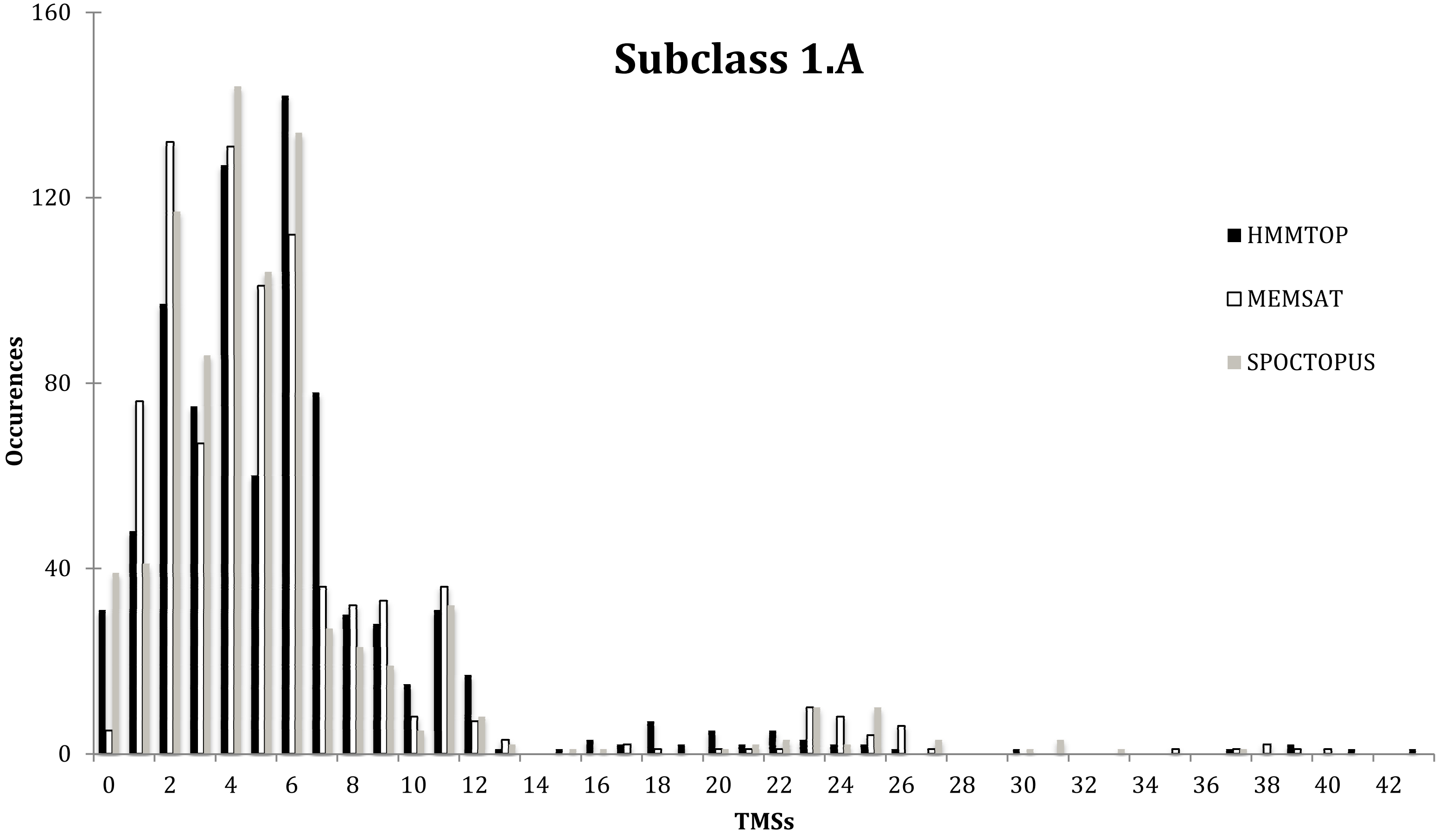


Figure 2: Comparative distribution of topological types predicted using the TMStats system for HMMTOP in black, MEMSAT in white and SPOCTOPUS in grey, for the proteins included in subclass 1.A of TCDB as of 5/29/2013.

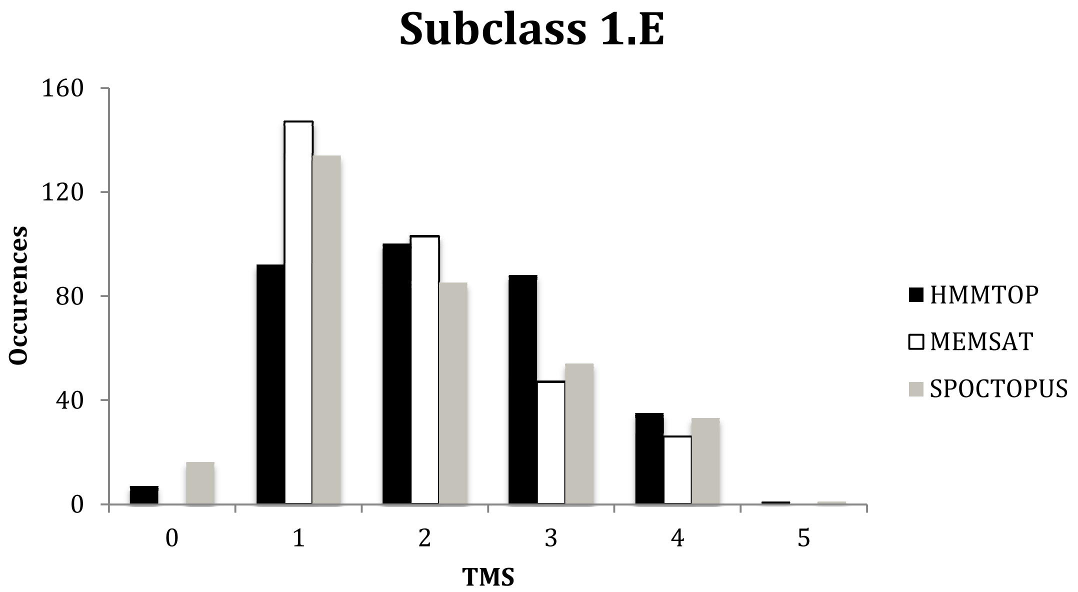
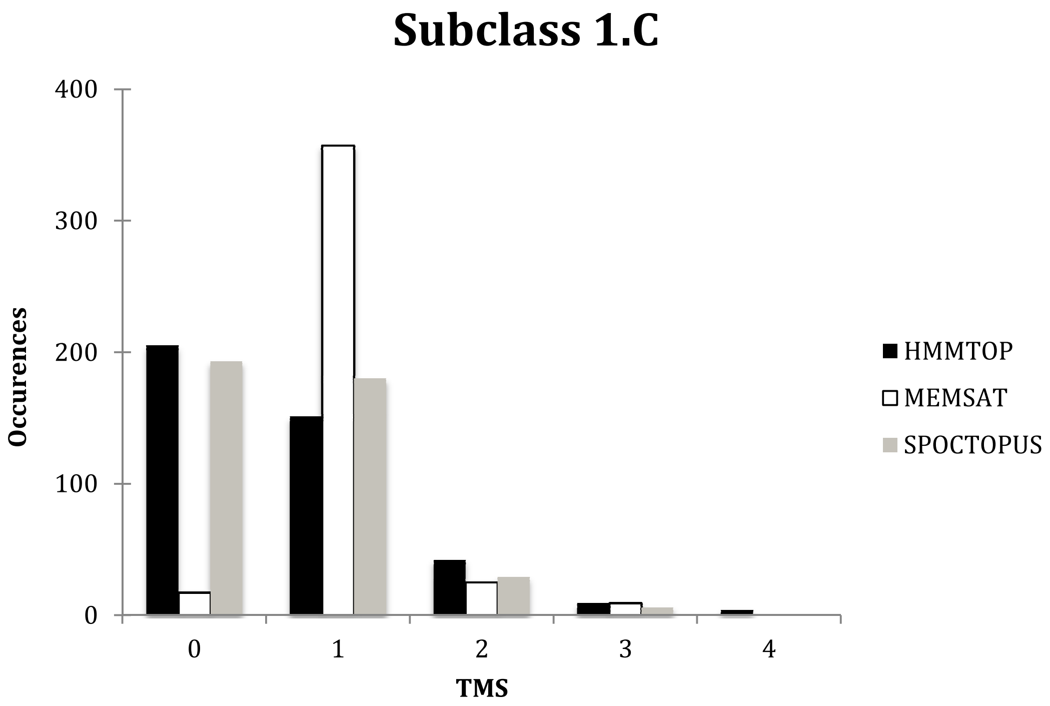


Figure 3: Comparative distribution of topological types predicted using the TMStats system for HMMTOP in black, MEMSAT in white and SPOCTOPUS in grey, for the proteins included in subclass 1.E of TCDB as of 5/29/2013.



F**igure 4:** Comparative distribution of topological types predicted using the TMStats system for HMMTOP in black, MEMSAT in white and SPOCTOPUS in grey, for the proteins included in subclass 1.C of TCDB as of 5/29/2013.

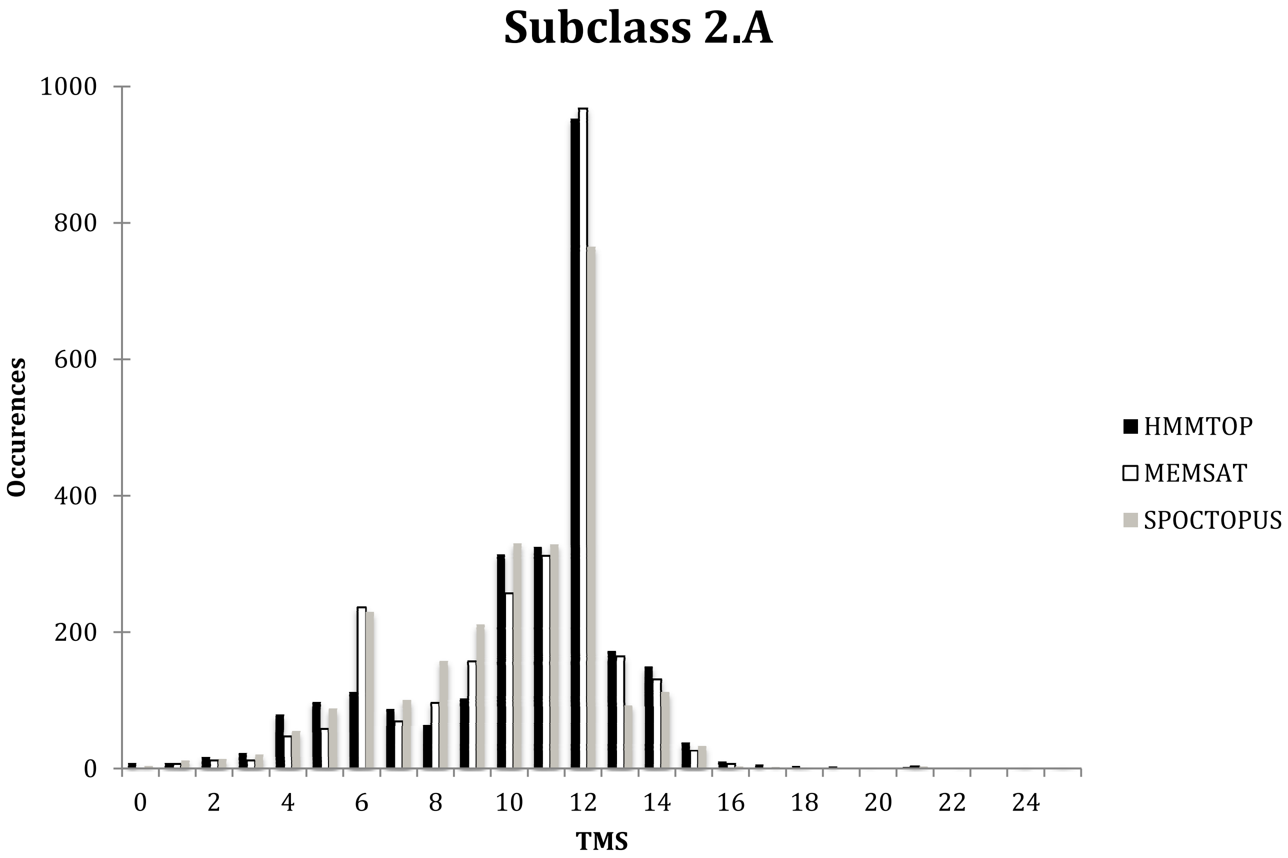


Figure 5: Comparative distribution of topological types predicted using the TMStats system for HMMTOP in black, MEMSAT in white and SPOCTOPUS in grey, for the proteins included in subclass 2.A of TCDB as of 5/29/2013.

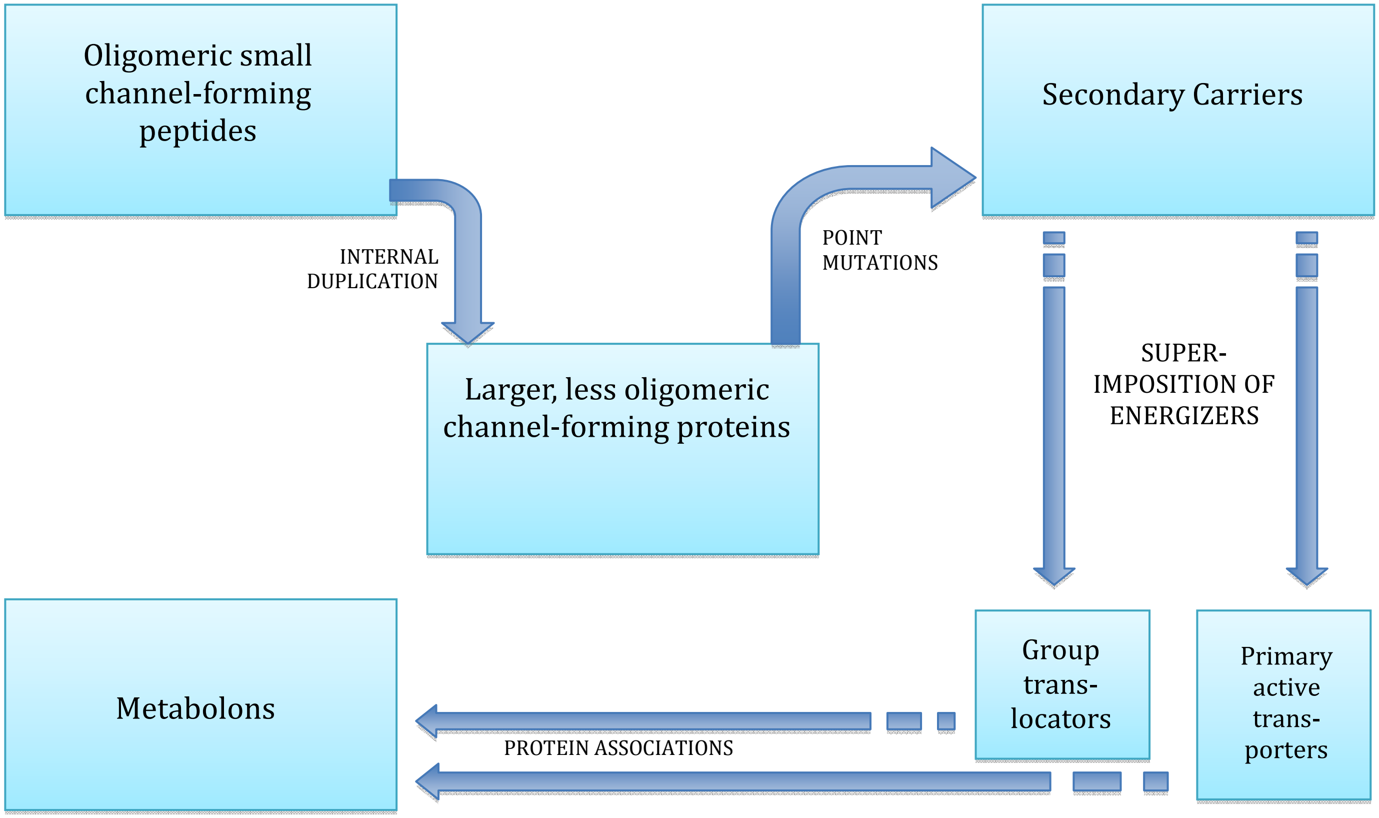

Oligomeric small channel-forming peptides
INTERNAL DUPLICATION
Larger, less oligomeric channel-forming proteins
POINT MUTATIONS
Secondary Carriers
SUPER-IMPOSITION OF ENERGIZERS
Group trans-locators
Primary active trans-porters
PROTEIN ASSOCIATIONS
Metabolons

Figure 6: Schematic depiction of the proposed pathway for the evolution of transport proteins including different types of channel-forming proteins and secondary carriers. We further propose that primary active transport carriers and group translocators arose by the superimposition of energy coupling enzymes such as ATPases. Finally, the integration of these systems into metabolic pathways resulted in the physical construction of complex but coordinated metabolons. See Norris et al., 2007 and Saier, 2003 for earlier considerations regarding this proposal.