UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily /

Permalink

https://escholarship.org/uc/item/6qd7b33c

Author

Yee, Daniel Choi

Publication Date

2014

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily

A Thesis submitted in partial satisfaction of the requirements for the degree Master of Science

in

Biology

by

Daniel Choi Yee

Committee in charge:

Professor Milton H. Saier Jr., Chair
Professor Yunde Zhao
Professor Lin Chao

2014

The Thesis of Daniel Yee is approved and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Chair

University of California, San Diego

2014

DEDICATION

This thesis is dedicated to my parents, my family, and my mentor, Dr. Saier. It is only with their help and perseverance that I have been able to complete it.

TABLE OF CONTENTS

Signature Page iii
Dedication iv
Table of Contents v
List of Abbreviations vi
List of Supplemental Files vii
List of Figures viii
List of Tables ix
Acknowledgements x
Abstract xi
Introduction 1
Methods 3
Results 10
Discussion 23
Supplemental Files 27
Figures 38
Tables 41
References 43

LIST OF ABBREVIATIONS

TOG ..... Transporter-Opsin-G protein-coupled
MR ..... Microbial Rhodopsin
VR ..... Visual Rhodopsin
TSUP ..... 4-Toulene Sulfonate Uptake Permease
PnuC ..... Nicotinamide Ribonucleoside Uptake Permeases
NiCoT ..... Ni^{2+}-Co^{2+} Transporter
OST ..... Organic Solute Transporter
PNaS ..... Phosphate:Na^{+} Symporter
LCT ..... Lysosomal Cystine Transporter
HORC ..... Heteromeric Odorant Receptor Channel

LIST OF SUPPLEMENTAL FILES

Figure S1. AveHAS plot, phylogenetic tree for the LCT family, homology between members of the Sweet and LCT families, and homology between members of the PNaS and LCT 27

Figure S2. AveHAS plot, phylogenetic tree for the MR family 28

Figure S3. AveHAS plot, phylogenetic tree for the NiCoT family, and homology between members of the TSUP and NiCoT families 29

Figure S4. AveHAS plot, phylogenetic tree for the OST family, homology between members of the TSUP and OST families, binary alignment of members of the TSUP and OST families, GSAT alignment of TSUP and OST 30

Figure S5. AveHAS plot, phylogenetic tree for the Sweet family, homology between members of the Sweet and OST families, binary alignment of members of the Sweet and OST families, GSAT alignment of Sweet and OST 31

Figure S6. AveHAS plot, phylogenetic tree for the PNaS family, homology between members of the Sweet and PNaS families 33

Figure S7. AveHAS plot, phylogenetic tree for the PnuC family, homology between members of the PnuC and TSUP families, binary alignment of members of the PnuC and TSUP families, GSAT alignment of PnuC and TSUP 34

Figure S8. AveHAS plot, phylogenetic for the GPCR family, homology between members of the GPCR and the microbial rhodopsin (MR) families, binary alignment of members of the GPCR and MR families, GSAT alignment of GPCR and MR 35

Figure S9. Results from the ClustalX-based neighbor-joining program 37

Table S1. The highest comparison scores between TOG superfamily (left column) and the MC and MIP families 37

LIST OF FIGURES

Figure 1 (A): TOG superfamily homology established through the use of GSAT/GAP and the Superfamily Principle. TOG superfamily proteins from TCDB and their homologues were used to establish homology between all members of the nine families. GSAT/GAP comparison scores, adjacent to the arrows, are expressed in terms of standard deviations (S.D.). **(B)**: Proposed evolutionary pathway for the appearance of nine recognized families within the TOG superfamily. The TOG superfamily is believed to have arisen from a 4 TMS precursor that duplicated to an 8 TMS precursor, common to the superfamily constituents, before diverging in topology via the loss or gain of specific TMSs 38

Figure 2. Average hydropathy, amphipathicity and similarity (AveHAS) plots based on a ClustalX multiple alignment. All members of all TOG Superfamily families from TCDB were included except for the GPCR (see Fig. S10) family. The plot reveals 8 well conserved average TMSs. However, as many as 6 poorly conserved peaks of hydrophobicity can be seen representing additional potential TMSs in the non-homologous regions of some proteins 39

Figure 3. Phylogenetic (Fitch) trees for the TOG superfamily. Phylogenetic (Fitch) trees for the TOG superfamily in TCDB as of May, 2012. The BLAST-derived SFT1 program shows the proteins of families within the TOG superfamily. Numbers adjacent to the branches indicate the protein TC#'s (last two digits of the complete protein TC#) while the family designations and family TC numbers are provided in large print in parentheses. In (B),. The ClustalX-Tree View tree is presented in supplementary figure S11 39

Figure 4. SFT2-based phylogenetic tree. This tree shows the relationships of the TOG superfamily families to each other. Family abbreviations are presented with TC family numbers in parentheses. See TCDB for protein identification..... 40

Figure 5. Demonstration of a 4 TMS repeat unit in the Pas1 protein of the TSUP family. GSAT alignment of TMSs 1-4 of Pas1 (Photorhabdus asymbiotica; gi211638062; 8 TMSs) with TMSs 5-8 of the same protein. A comparison score of 15.2 S.D. was obtained with 50.5% similarity and 30.3% identity..... 40

LIST OF TABLES

Table 1. Summary of nine TOG Superfamily family members. The family number, name, abbreviation, TC number, number of TMSs, normal protein size range in numbers of amino acyl residues, dominant topology, TMS gain or loss, organismal distribution and Pfam id designation are presented. Minor topologies, represented in a minority of family members, are listed in small numbers in parentheses in column 7. Putative topological transitions are described in footnotes 1-4 41

Table 2. The highest comparison scores between TOG superfamily members. Protein representatives of families used in these comparisons are found in tables S1-S9. Representative alignments for highlighted squares are usually provided in figures S1-S7 and S9-S10, but in some cases better scores are reported here than in the figures, based on other alignments. Values above 12.0 S.D., which are considered sufficient to establish homology and interconnect families, are shaded. These values are sufficient to establish homology based on the criteria discussed in the methods section. 42

ACKNOWLEDGEMENTS

I'd like to acknowledge the wonderful Professor Milton Saier for mentoring and teaching me the last two years. He generously provided both guidance and knowledge in my research and other endeavors, and he has been one of the biggest influences in my life. Dr. Saier is an inspiring role model for all of his students, and I am grateful to have worked under him.

I would also like to acknowledge the Saier Lab for their support and company. Their collective knowledge has contributed greatly to my research and they have been an invaluable source of camaraderie and friendship.

This thesis is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC, Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis author was co-first author of this paper.

Lastly, I would like to thank my parents for giving me the opportunity to attend and study at the University of California San Diego. Without their strong support, I would not be where I am today.

ABSTRACT OF THE THESIS

The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily

by

Daniel Choi Yee

Master of Science in Biology

University of California, San Diego, 2014

Professor Milton H. Saier Jr., Chair

Visual Rhodopsins (VR) are recognized members of the large and diverse family of G protein-coupled receptors (GPCRs), but their evolutionary origin and relationships to other proteins, are not known. In an earlier publication (Shlykov *et al.*, 2012), we characterized the 4-Toulene Sulfonate Uptake Permease (TSUP) family of transmembrane proteins, showing that these 7 or 8 TMS proteins arose by intragenic duplication of a 4 TMS-encoding gene, sometimes followed by loss of a terminal TMS. In this study, we show that the TSUP, GPCR and Microbial Rhodopsin (MR) families are related to each other and to six other currently recognized transport protein families. We designate this superfamily the Transporter-Opsin-G protein-coupled

receptor (TOG) Superfamily. Despite their 8 TMS origins, members of most constituent families exhibit 7 TMS topologies that are well conserved, and these arose by loss of either the N-terminal (more frequent) or the C-terminal (less frequent) TMS, depending on the family. Phylogenetic analyses revealed familial relationships within the superfamily and protein relationships within each of the nine families. Proteins functioning by dissimilar mechanisms (channels, primary active transporters, secondary active transporters, group translocators and receptors) are interspersed on a phylogenetic tree of the TOG superfamily, suggesting that changes in the transport and energy-coupling mechanisms occurred multiple times during the evolution of this superfamily.

This chapter is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC, Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis author was co-first author of this paper.

INTRODUCTION

Using functional and phylogenetic information derived from over 10,000 publications on transport systems, our laboratory has been able to classify virtually all recognized transport proteins into over 700 families [1-2]. The resultant system of classification is summarized in the IUBMB-approved Transporter Classification (TC) Database (TCDB; http://www.tcdb.org) [3-4]. Our current efforts serve to identify distant relationships, allowing the placement of these families into superfamilies. Since transport systems play crucial roles in virtually all processes associated with life, their importance cannot be overstated [5-6].

The present study reports the identification of a novel superfamily, a group of proteins showing a common evolutionary origin, that we have designated the Transporter-Opsin-G protein-coupled receptor (TOG) superfamily, based on the best-characterized families of proteins present in this superfamily. In addition to (1) ion-translocating Microbial Rhodopsins (MR; TC# 3.E.1) and (2) G protein-coupled receptors (GPCRs; TC# 9.A.14) including visual rhodopsins (VRs), we show that members of the following families (see Table 1) share a common origin with microbial, invertebrate and vertebrate rhodopsins: (3) Sweet sugar transporters (Sweet; TC# 9.A.58), (4) Nicotinamide Ribonucleoside Uptake Permeases (PnuC; TC# 4.B.1), (5) 4-Toluene Sulfonate Uptake Permeases (TSUP; TC# 2.A.102), (6) Ni^{2+}-Co^{2+} Transporters (NiCoT; TC# 2.A.52), (7) Organic Solute Transporters (OST; TC# 2.A.82), (8) Phosphate:Na^{+} Symporters (PNaS; TC# 2.A.58), and (9) Lysosomal Cystine Transporters (LCT; TC# 2.A.43). Furthermore, our research indicates that the

invertebrate Heteromeric Odorant Receptor Channel (HORC; TC# 1.A.69) family may also share a common origin with members of the TOG superfamily, although this could not be established using our standard statistical criteria.

Our evidence suggests that all of the diverse proteins included in the TOG superfamily derive from a common ancestor via similar pathways. It can therefore be anticipated that the structures of most of these proteins will exhibit common features [7-8]. Since rhodopsins are the transmembrane proteins with the highest resolution X-ray structures solved to date [9-11], we now have the capacity to apply this structural information to the other protein families included within this superfamily. The work reported here provides the groundwork for comparative studies that should lead to a more detailed understanding of how a single structural scaffold can vary to accommodate a wide diversity of functions. It should serve as a guide in future studies revealing how sequence divergence can lead to alterations in this scaffold.

This chapter is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC, Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis author was co-first author of this paper.

METHODS

Query sequences used to identify members of each family were (1) bacteriorhodopsin of Halobacterium salinarum (GenBank:gi# 114811, TC# 3.E.1.1.1), (2) MtN3 of Medicago truncatula (gi# 75220431, TC# 9.A.58.1.1), (3) PnuC of Haemophilus influenzae (gi# 81335937, TC# 4.B.1.1.2), (4) Orf of Pyrococcus abyssi (gi# 74545625, TC# 2.A.102.4.1), (5) YfcA of Escherichia coli (gi# 82592533, TC# 2.A.102.3.1), (6) Orf of Oryza sativa (gi# 75252893, TC# 2.A.102.5.1), (7) RcnA of E. coli (gi# 3025266, TC# 2.A.52.2.1), (8) Ost α of Raja erinacea (gi# 82108802, TC# 2.A.82.1.1), (9) NptA of Vibrio cholera (gi# 81345622, TC# 2.A.58.1.2), (10) CTNS of Homo sapiens (gi# 269849555, TC# 2.A.43.1.1), and (11) ROP of Homo sapiens (gi# 129219, TC# 9.A.14.1.1). Analyses dealing with the HORC family TC# 1.A.69 were performed utilizing OR83b of Drosophila melanogaster (gi# 14285640, TC# 1.A.69.1.1) as the query. NCBI PSI-BLAST searches with two iterations (e^{-4}; e^{-6} cutoffs, respectively) were performed using Protocol1 [13, 73] to identify members of each family. The Protocol1 program compiles homologous sequences from the BLAST searches into a single file in FASTA format, eliminates redundancies and fragmentary sequences, and generates a table of the obtained sequences containing protein abbreviations, sequence descriptions, organismal sources, protein sizes, gi numbers, organismal groups or phyla, and organismal domains (see supplementary tables). Protocol1's CD-HIT option was used to remove redundancies and highly similar sequences [13, 17]. An 85% identity cut-off was used in establishing homology between family members, while a 70% identity cut-off was used to create

more easily viewed average hydropathy plots and phylogenetic trees. These percent identity values thus refer to the values above which redundant sequences were removed. Thus, an 85% cutoff means that no two protein sequences retained for analysis were more than 85% identical. FASTA files from Protocol1 were considered representative of each respective protein family, although selected proteins that demonstrated homology between families were confirmed with NCBI's Conserved Domain Database [74] and PSI-BLAST results.

The ClustalX program was used to create a multiple alignments of homologous proteins, and the few sequences that introduced large gaps into the alignment (usually a reflection of fragmentation, inclusion of introns, or incorrect sequences) were removed. This allowed the generation of a coherent multiple alignment where all or most sequences are homologous throughout most of their lengths. Results obtained with this program have been compared with 5 other programs, and when sequence similarity was sufficient to give reliable multiple alignments, phylogenetic trees obtained with the six programs (Neighbor Joining or Parsimony) were very similar [75]. For topological analyses of single protein sequences, the WHAT, TMHMM 2.0, and HMMTOP programs were used [76-77]. Inputting the multiple alignment files generated by ClustalX into the Average Hydropathy, Amphipathicity and Similarity (AveHAS) program facilitated more accurate topological assessments of multiple proteins or entire families. CDD was also used to analyze protein sequence extensions identified using the AveHAS plot. Motif analyses were performed using the MEME/MAST programs [78-79].

Initially, a large screen was performed comparing distantly related TSUP family members [18] against all families of the TC# 2.A, 3.E and 9.A subclasses. The Targeted Smith-Waterman Search (TSSearch) feature of Protocol2 [13] was then run in order to compare each family to all other TOG superfamily members [12, 18]. TSSearch uses a rapid search algorithm to find distant homologues between two different FASTA files that may not readily be apparent in BLAST or PSI-BLAST searches [13]. The most promising comparisons between proteins were automatically analyzed using the Global Sequence Alignment Tool (GSAT) [80] feature of Protocol2 [13]. Comparisons using the GSAT feature of Protocol2 are reported in standard deviations (S.D.), which refers to the number of S.D.s a given score is from the mean, raw local bit score of pairwise scores of 200 shuffled residues. Scores are calculated with the Needleman-Wunsch algorithm. Promising results with a comparison score of 12.0 S.D. or greater were confirmed and analyzed further using the GSAT and GAP programs set at default settings with a gap creation penalty of 8 and a gap extension penalty of 2 with 2,000 random shuffles; assuming a Gaussian distribution, a comparison score of 12.0 S.D. corresponds to a probability of 1.77×10^{-33} that the degree of similarity between two proteins arose by chance (See supplementary Fig. S9) [81]. In spite of this conclusion by Dayhoff et al. [81], Gaussian skew can increase the probability of chance similarity for any given standard deviation value [82].

Probabilities for comparison scores were calculated using Mathematica (Wolfram Research, Inc., Champaign, IL, USA). Comparisons involving at least 60

amino acyl residues, the average size of a prototypical protein domain, alignment of 2 or more α-helical domains between compared proteins, and a comparison score of at least 12.0 S.D. were considered sufficient to provide strong evidence for homology between two proteins or internal repeat units in the studies reported [1, 4, 17, 81]. Convergent sequence evolution is possible and has been demonstrated for short motifs but never for large segments of proteins such as entire domains. One reason we use a minimum of 60 amino acyl residues in defining homology is that for such a long sequence, convergence to give 12 S.D. is exceedingly unlikely.

Optimization of the GSAT/GAP alignments was performed on sequences by maximizing the number of identities, minimizing gaps, and removing non-aligned sequences at the ends of the alignment, but never in central regions of the alignment. Optimization usually yields a higher comparison score that better represents the level of similarity between two internal sequences.

The Ancient Rep (AR) program [13] was used to look for internal repeats, and results were confirmed using the GSAT/GAP and HHRep programs [83]. The AR program compares potential transmembrane repeat sequences (hydrophobic TMS regions predicted by HMMTOP) within a single protein and between proteins in a FASTA file, giving a comparison score in S.D.s in the same format as Protocol2.

A large screen was performed with all members of the TOG superfamily against the Major Intrinsic Protein (MIP; TC# 1.A.8) and the Mitochondrial Carrier (MC; TC# 2.A.29) families, two large families whose known evolutionary pathways and topologies differ from each other and those of the proposed TOG superfamily [49-

50, 84]. Comparisons between each family were conducted using the same techniques and programs in establishing homology between TOG superfamily members (Protocol1, Protocol2, GSAT). The best comparison scores were selected using the same criteria as outlined previously; selected comparisons contained at least 2 or more aligned α-helical domains and involved at least 60 residues. The evolutionary pathway was not considered in selections. Precise scores of the best alignments fitting these criteria were obtained with GSAT and GAP programs set at default settings with a gap creation penalty of 8 and a gap extension penalty of 2 with 2,000 random shuffles. These scores were then compared against alignments demonstrating homology between members of the TOG superfamily.

As controls, we looked for similarities between members of the MIP family (TC# 1.A.8) and the MC family (TC# 2.A.29) using Pfam-A, a database of HMMs of protein domains. We used HMMER3 to search the current version of TCDB, using the default cutoff (10). We loaded all edges connecting either MC or MIP proteins in Cytoscape 2.8.3 to view the results. Significant similarities were not found.

The ClustalX program [51] was used to create multiple alignment for homologous sequences using default settings, and a neighbor-joining phylogenetic tree for the TOG superfamily was created using the TreeView program [52]. Phylogenetic trees for individual families were also drawn using the FigTree program (http://tree.bio.ed.ac.uk/software/figtree/). To depict phylogenetic relationships more accurately than possible using the multiple alignments provided by the ClustalX program, the SFT programs [15-17] were used to generate SFT1 and SFT2

phylogenetic trees using tens of thousands of BLAST bit scores instead of multiple alignments [17]. The SFT1 phylogenetic tree was generated to visualize relationships between all subfamilies within families of the TOG superfamily. The SFT2 tree, drawn with the TreeView program [52], consolidated individual members into their respective families for visualizing phylogenetic relationships between families within the TOG superfamily.

The ClustalX program [51] was used to create multiple alignment for homologous sequences using default settings, and a neighbor-joining phylogenetic tree for the TOG superfamily was created using the TreeView program [52]. Phylogenetic trees for individual families were also drawn using the FigTree program (http://tree.bio.ed.ac.uk/software/figtree/). To depict phylogenetic relationships more accurately than possible using the multiple alignments provided by the ClustalX program, the SFT programs [15-17] were used to generate SFT1 and SFT2 phylogenetic trees using tens of thousands of BLAST bit scores instead of multiple alignments [17]. The SFT1 phylogenetic tree was generated to visualize relationships between all subfamilies within families of the TOG superfamily. The SFT2 tree, drawn with the TreeView program [52], consolidated individual members into their respective families for visualizing phylogenetic relationships between families within the TOG superfamily.

This chapter is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC, Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis

author was co-first author of this paper.

RESULTS

All protein families within TCDB belonging to subclass 2.A consist of electrochemical potential-driven uniporters, symporters and antiporters. Based on preliminary evidence reported by Shlykov *et al*. (2012), we used a modified SSearch program [12-13] to compare Toluene Sulfonate Uptake Permease (TSUP) homologues with all other secondary carriers and identified potential superfamily relationships. Subsequently, these analyses were extended to other TC classes. Comparisons to the TC subclasses 9.A, 3.E and 4.B proved fruitful.

The MR and LCT families had previously been shown to be related [14]. Analyses dealing with the MR, LCT, PnuC, PNaS, Sweet, and GPCR families provided sufficient evidence to include them in the TOG superfamily. Our results led to the formulation of the novel TOG superfamily for which trees were generated using the ClustalX, Superfamily Tree 1 (SFT1) and SFT2 programs [15-17].

The TOG superfamily consists of nine, and possibly ten, currently recognized protein families with members of primarily 6, 7, 8 and 9 putative TMSs (Table 1). A summary of the comparisons performed is presented in Figure 1A and Table 2, while the proposed evolutionary pathway for the appearance of the different members of the TOG superfamily is presented in Figure 1B. The TSUP family has been characterized previously [18]. Therefore, unlike the other eight families, the AveHAS plots, phylogenetic trees and a table listing TSUP homologues have been excluded from this study and are reported by Shlykov *et al*. [18] . The evolutionary pathway of the 7 TMS LCT family has been elucidated [14], and to members of the MR family, including

putative fungal chaperone proteins present in the Microbial Rhodopsin family (see Table 1 and TC entries under TC# 3.E.1). Most of the known MR transporters are light-driven ion pumps or light-activated ion channels.

LCT family members range in size from 300 to 400 amino acyl residues (aas) and are generally larger than MR proteins, which have ~220 to 300 residues. Eukaryotic homologues within a single transporter family tend to be ~40% larger than their bacterial homologues [19]. Whereas the LCT family is found exclusively in the eukaryotic domain, the MR family is present in all three domains of life (Table 1). Despite these differences, both families possess a 7 TMS topology (Tables 1, S1, S2 and Figs. S1A-B, S2A-B).

TMSs 1-3 in LCT family members duplicated to give rise to TMSs 5-7, with TMS 4 showing insignificant sequence similarity to any one of the other six TMSs [14]. The precursor could have been an 8 TMS protein that generated the present-day 7 TMS proteins by loss of TMS 1 or 8, and strong evidence for this possibility is presented in [18] and here.

The 7 TMS Bba2 protein of the LCT family is homologous to the 7 TMS Aae2 protein of the Sweet family. The alignment between these two proteins using GSAT yielded a comparison score of 12.4 S.D. (Fig. S1C). These comparisons establish homology between the LCT and Sweet families. TMSs 3-7 of Aae2 aligned with TMSs 3-7 of Bba2, demonstrating that the two families both arose via the same evolutionary pathway (Fig. S1C). Loss of TMS 1 in an 8 TMS predecessor yielded the 7 TMS topology found in members of the LCT and Sweet families.

Expansion of the TOG superfamily resulted from comparisons between the LCT and PNaS families. Comparing TMSs 2-4 of LCT Ago1 (7 putative TMSs) with TMSs 6-8 of PNaS Cre1 (11 putative TMSs) yielded a comparison score of 12.8 S.D. (Fig. S1D). This comparison establishes homology between regions of proteins in the LCT and PNaS families and further supports the proposed evolutionary pathway for the LCT family, as TMSs 2-4 of Ago1 and TMSs 6-8 of Cre1 correspond to the last three TMSs in the proposed 4 TMS predecessor. PSI-BLAST searches of Cre1 yielded two separate conserved PNaS domains within the protein. The extended 11 TMS topology in Cre1 likely arose from the fusion of a 7 TMS protein with another 4 TMS repeat unit.

Members of subfamily 1 within the ubiquitous NiCoT family are typically 300 to 380 aas in size and possess 6-8 putative TMSs [20] (Table S3 and Figs. S3A-B). NiCoT subfamily 2 is comprised of distant homologues of great size, sequence and topological variation. NiCoT transporters catalyze the uptake of Ni^{2+} and Co^{2+} using a pmf-dependent mechanism; however, a Ni^{2+} and Co^{2+} resistance protein that is believed to export the two metals to the external environment has been reported [21-22]. Smaller members of subfamily 2 exhibit 6-8 putative TMSs.

Comparing TMSs 1-3 of TSUP Pla1 (8 putative TMSs) with TMSs 4-6 of NiCoT Bja1 (6 putative TMSs) yielded a comparison score of 12.8 S.D. (Fig. S3C). This comparison establishes homology between members of these two families and serves to confirm our proposed evolutionary pathway for the appearance of the NiCoT family as a member of the TOG superfamily (Figs. 1A and 1B). Based on alignments,

it is likely that the 6 TMS NiCoT proteins arose by the loss of both TMSs 1 and 8 after the 4 TMS intragenic duplication event took place.

Members of the OST family are almost exclusive to animals and are known to transport organic anions including estrone-3-sulfate, bile acids, taurocholate, digoxin and prostaglandins [23-25]. Distant homologues of the α-subunits in plants, fungi and bacteria were retrieved in NCBI searches, but their scores usually bordered or fell below our threshold cutoff for establishing homology. Furthermore, each well characterized transporter within this family functions as part of a two-component system utilizing the α- (280-400 aas) and β-subunits (180-290 aas). The α-subunits generally contain seven TMSs, whereas the β-subunits contain only one. To date, neither subunit has been found to function without the other (Table S4 and Figs. S4A-B) [23].

Comparing TMSs 2-3 of TSUP Tsp1 (8 putative TMSs) with TMSs 1-2 of OST Cre2 (7 putative TMSs) yielded a comparison score of 12.1 S.D. (Fig. S4C). This comparison demonstrates the loss of TMS 1 in OST transporters and establishes homology between the two families; the loss of TMS 1 from an 8 TMS precursor generated the 7 TMS topology of the OST family. Another alignment between TMSs 2-4 of TSUP Gfo1 (7 putative TMSs) with TMSs 2-4 of OST Dre1 (7 TMSs) also supports homology between the TSUP and OST families and the proposed evolutionary pathway. It yielded a comparison score of 11.3 S.D. (Fig. S4D).

Eukaryotic Sweet family channels or carriers catalyze facilitated diffusion (uptake or efflux) of sugars across the ER and plasma membranes of plants and

animals [26]. Bacterial pathogens upregulate specific plant Sweet transporters, allowing them to utilize the sugar efflux function of these proteins to meet their energy needs [27]. Eukaryotic homologues possess 7 TMSs in a 3 + 1 + 3 repeat arrangement and are 200-290 aas in size (Table S5 and Figs. S5A-B). Although 7 TMS bacterial homologues exist, most bacterial putative Sweet channels possess 3 TMSs and are about half the size of their eukaryotic and larger bacterial relatives. The 3 TMS proteins show greatest sequence similarity to the first (N-terminal) repeat in the 7 TMS proteins. It is unclear whether the eukaryotic or prokaryotic proteins function as channels or carriers. However, no well-documented examples of carriers with fewer than 4 TMSs per polypeptide chain have been reported, suggesting that the 3 TMS proteins may function as oligomeric channels [28].

Comparing TMSs 6-7 of Sweet Rco4 (7 putative TMSs) with TMSs 6-7 of OST Ath8 (7 putative TMSs) yielded a comparison score of 12.3 S.D. (Fig. S5C). This result indicates that as for the MR and OST families (as well as several other TOG superfamily members), the N-terminal TMS was lost from the 8 TMS topology to generate the 7 TMS Sweet proteins. A second alignment between TMSs 4-6 of Sweet Asu3 (7 putative TMSs) and TMSs 4-6 of OST Ncr1 (7 putative TMSs) also yielded a comparison score of 10.3 S.D. (Fig. S5D), further confirming the establishment of homology between Sweet and OST families.

Both bacterial and eukaryotic PNaS homologues usually range in size between 500 and 700 residues, but the bacterial homologues can be as small as 350 residues. Most members of this family possess 8 or 9 TMSs in a 4 + 4 or a 4 + 4 + 1 TMS

arrangement, which has been demonstrated previously (Saier, 2003; Table S6 and Fig. S6A-B). However, some proteins such as NPT2 of *Rattus norvegicus* can have as many as 12 TMSs, with the extra ones appearing at the N-termini [29]. Mammalian PNaS porters may catalyze the electroneutral cotransport of 3 Na^+ with inorganic phosphate (P_i). Their activities can be regulated by parathyroid hormone and dietary P_i.

Comparing TMSs 4-6 of PNaS Odi8 (8 putative TMSs) with TMSs 3-5 of MR Hwa1 (7 putative TMSs) yielded a comparison score of 13.1 S.D. (Fig. S6C). This comparison demonstrates homology between the MR and PNaS families and further supports the conclusion that TMS loss in PNaS family members occurred at their N-termini.

PnuC family proteins are restricted to bacteria and archaea as well as several bacteriophage. These proteins possess 8 or 7 TMSs in a 4 + 4 or 3 + 1 + 3 repeat arrangement, respectively. The 7 TMS proteins arose by the loss of the N-terminal TMS in the 8 TMS homologues. Some members may be energized by multifunctional NadR homologues, which perform the required step of phosphorylating nicotinamide ribonucleoside (NR), thus allowing its transport in a “group translocation” or “metabolic trapping” process [30-32]. The ribonucleoside kinase domains of NadR homologues are responsible for the transfer of a phosphoryl group from ATP onto NR [33-34]. Therefore, ATP appears to be required for NR accumulation. Proteins of the PnuC family are typically 210 to 270 aas in size (Table S7 and Figs. S7A-B).

Comparing TMSs 2-3 of PnuC Spr1 (7 putative TMSs) with TMSs 3-4 of TSUP Cba4 (8 putative TMSs) yielded a comparison score of 12.4 S.D. (Fig. S7C).

This comparison demonstrates homology between the PnuC and TSUP families. An alignment between TMSs 3-6 of PnuC Sde2 (7 putative TMSs) and TMSs 4-6 of TSUP Ere1 (8 putative TMSs) provides additional evidence of homology and supports the PnuC evolutionary pathway (Fig S8D). Our results and the placement of the PnuC family into the TOG superfamily supports the proposal that a 4 TMS precursor duplicated to give 8 TMS proteins, and that the N-terminal TMS was then deleted.

Members of the GPCR family [35-41] encompass an extremely diverse range of cellular membrane proteins and constitute the largest family of transmembrane proteins found in humans [39, 42]. While all share a general signaling mechanism wherein extracellular signals are transduced into intracellular effectors via ligand binding, members vary tremendously in both ligand type and function. GPCR family members each consists of a 7 TMS α-helical bundle, connected by three extracellular and three intracellular loops. This 7 TMS bundle displays distinctive hydrophobic patterns (Fig. S8A and S8E) and is commonly recognized as the most conserved element of GPCRs [43]. Because the GPCR family includes receptors for a wide variety of hormones, neurotransmitters, chemokines, calcium ions and photons (see Tables S8A, S8B and TCDB), they are among the most targeted therapeutic proteins for drugs, and their analyses have tremendous implications for future pharmacological developments [44].

Comparing TMSs 5-6 of the GPCR, Dre1 (7 putative TMSs) with TMSs 5-6 of Mos1 (7 putative TMSs) of the Microbial Rhodopsin (MR) family yielded a comparison score of 13.1 S.D. (Fig. S8C). This comparison establishes homology

between the GPCR family and the MR family; the topology of members of the GPCR family, like the MR family, likely arose from the loss of the N-terminal TMS from the proposed 8 TMS predecessor. TMSs 1-4 of GPCR Dre 1 (7 putative TMSs) also aligned with TMSs 1-4 of MR Cga1 (7 putative TMSs) and yielded a comparison score of 10.6 S.D., further supporting establishment of homology (Fig S7D) between GPCRs and MRs.

Olfactory sensory neurons in insects express between one to three members of the channel-forming olfactory receptor (OR) gene family as well as the highly conserved Or83b co-receptor TC# 1.A.69.1.1. Functional odorant receptors consist of a heteromeric complex comprised of at least one odorant-binding subunit and the aforementioned Or83b co-receptor [45]. Immunocytochemical experiments demonstrated that insect odorant receptors possess a 7 transmembrane topology, but in contrast to members of the GPCR family, they exhibit a cytoplasmic N-terminus and extracellular C-terminus. Several authors [45-47] suggested that heteromeric insect ORs comprise a new class of ligand-activated non-selective cation channels. We obtained preliminary evidence that insect ORs and GPCRs are homologous. However, based on our criteria, we could not establish homology because comparison scores were insufficient (10.3 S.D.). Nevertheless, the intriguing possibility of homology will provide the basis for future investigations.

Members of the Major Intrinsic Protein (MIP) family are channel proteins that function in water, small carbohydrate, urea, NH_3, CO_2, H_2O_2 and ion transport by energy-independent mechanisms. The observed topology of the MIP family arose

from the intragenic duplication of a 3 TMS predecessor [48]. Members of the Mitochondrial Carrier (MC) family are transporters involved in transporting keto acids, amino acids, nucleotides, inorganic ions and co-factors across the mitochondrial inner membrane. Proteins from the MC family arose from tandem intragenic triplication of a 2 TMS element, giving rise to a 6 TMS topology [49-50]. These two large 6 TMS protein families thus arose via different pathways and are not homologous. They provide an excellent control for homology.

The best comparison scores between the MC and MIP families and TOG superfamily members were 9.5 S.D., and 10.5 S.D., respectively (Table S9). Comparisons of the MC family against the NiCoT and PNaS families yielded a maximal comparison score of 9.5 S.D. A comparison of the MIP family against the PNaS family yielded a maximal comparison score of 10.5 S.D. The average score for all of the best comparisons between TOG superfamily members and the MC family was 8.8 S.D., and the average score for comparisons between TOG superfamily members and the MIP family was 8.9 S.D. When compared to each other, the MIP and MC families yielded maximal comparison scores of 9.2 S.D. By contrast, the average score for all of the best comparisons for the nine TOG superfamily families with each other was 11.5 S.D., and 12.6 S.D. between families used to establish homology. Based on these results, we suggest that 12.0 S.D. (Fig. S9) combined with the proper alignment of at least 2 transmembrane domains that fit a proposed evolutionary pathway, is sufficient to provide strong evidence for homology.

As a negative control, we searched for similarities between the MC (TC# 2.A.29) and MIP (TC# 1.A.8) families using Pfam-A. We found that even considering weak similarities, using the default cutoff of 10, these families showed links only through Pfam family PF12822 (DUF3816), an uncharacterized 5 TMS protein family. The edge linking 1.A.8.1.1 and DUF3816 scored only 2.8 (the best edge between MIP and DUF3816), considerably worse than any of the similarities we have reported to substantiate our conclusions about homology between members of the TOG superfamily. These results further establish the common origin of the family members of the TOG superfamily.

Using a phylogenetic tree that includes members of the nine established families of the TOG superfamily, proteins from each phylogenetic cluster were chosen and combined into a single file. The proteins were then aligned using ClustalX [51], and AveHAS plots [52] were generated for all families except the GPCR family (Fig. 2), as well as one for all families (Fig. S10). The large GPCR homologues rendered the AveHAS plot in Fig. S10 too large for easy viewing.

The plot reveals 7 well-conserved peaks of hydrophobicity with moderate amphipathic nature (peaks 2-8 in Fig. 2), as well as a poorly conserved peak (peak 1). This result is expected, given that the majority of the families consist predominantly of 7 TMS proteins. TMS 1 in the 8 TMS homologues is conserved in only a few of the family members. Other less conserved peaks of hydrophobicity are found N- and C-terminal to the 7 well-conserved peaks. A closer look revealed that these peaks are primarily due to the larger PNaS homologues. The 400 residue extension at the N-

terminal end of the alignment is attributable in part to the Sko2 protein of the PNaS family. A CDD search identified a member of the Death Domain (DD) superfamily constituting approximately the first 100 residues of the Sko2 N-terminus; DDs participate in protein-protein interactions in signaling pathways by recruiting proteins to complexes that sometimes comprise apoptosis pathways [53]. This accessory signaling domain in some PNaS proteins is not unexpected given their roles in phosphate reabsorption in mammalian tissues [54].

Proteins found in TCDB, representing the various subfamilies within each family (except the GPCR family) of the TOG superfamily, were used to generate a tree using the ClustalX neighbor-joining method (Fig. S11). The same proteins were then used to generate a tree using the BLAST-bit score-based SFT1 method (Fig. 3A) [15-17]. In Figure S11, the ClustalX/TreeView program drew the GPCR family (TC# 9.A.14) in five distinct clusters, the TSUP (TC# 2.A.102) family in three clusters, and the PnuC (TC# 4.B.1), LCT (TC# 2.A.43), NiCoT (TC# 2.A.52) and PNaS (TC# 2.A.58) families each in two clusters. Only members of the MR (TC# 3.E.1), Sweet (TC# 9.A.58) and OST (TC# 2.A.82) families clustered coherently within a single cluster according to their respective TC family assignments. This situation contrasts with the SFT1 tree (Figure 3A), which shows clustering of nearly all protein members coherently according to their respective families with the exceptions of the GPCR and NiCoT families, which are found in two closely related clusters. All members of NiCoT subfamily 1 (TC# 2.A.52.1) comprise one cluster, and all members within NiCoT subfamily 2 (TC# 2.A.52.2) comprise the other. These results reveal the

superiority of the SFT1 program over the ClustalX program, an observation now noted for many sequence-divergent superfamilies where multiple alignments are not reliable [15-17]. The SFT2 tree (Fig. 3B) shows the phylogenetic relationships between all nine families within the TOG superfamily. Interestingly, the families that have lost TMS 1 cluster together at the bottom of the tree, suggesting that this event could have occurred before these families diverged from each other.

Shlykov *et al*. [18] previously reported internal repeats within TSUP family members that corresponded to a 4 transmembrane α-helical (TMS) structural precursor [18], and Zhai *et al*. [14] demonstrated that TMSs 1-3 are homologous to TMSs 5-7 in the 7 TMS Microbial Rhodopsins (MR). More recently Sun and Saier (unpublished results) demonstrated that TMSs 1-3 are homologous to TMSs 5-7 in members of the PNaS family (14.6 S.D.). Using the AR and GSAT programs [13], comparing TMSs 1-4 with TMSs 5-8 of the 8 TMS TSUP Pas1 protein yielded a comparison score of 15.2 S.D. (Fig. 4), demonstrating that an intragenic 4 TMS duplication event occurred in TSUP family members. The 4 TMS unit duplicated to yield an 8 TMS protein. By the Superfamily Principle, the internal repeats in the TSUP family are applicable to all families within the TOG superfamily. The evolutionary pathway elucidated for the TSUP and MR families explains the alignment of specific transmembrane domains in the two halves of various families within the TOG superfamily (data not shown).

This chapter is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC,

Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis author was co-first author of this paper.

DISCUSSION

The analyses reported here allowed us to interlink nine integral membrane protein families of diverse mechanistic types to form the novel TOG superfamily (see Table 1 and Figure 1). No other transport protein superfamily in TCDB [3-4] exhibits functional and mechanistic diversity as great as that of the TOG superfamily. This unexpected quality is highlighted by the presence of known and putative secondary carriers, group translocators, light-driven pumps, channels, transmembrane chaperone proteins, photoreceptors and G-protein-coupled receptors. Clearly, this is a case where superfamily assignment is *not* a guide to energy coupling mechanism or mode of transport. Indeed, in contrast to most superfamilies of integral membrane transport proteins [6], several TOG superfamily members are not transporters at all, and there is no correlation between mode of transport or substrate specificity and position in the phylogenetic (SFT) tree (see Fig. 3). The results illustrate the potential of some superfamilies, but not others, to diverge into proteins with different modes of action. For example, no member of the Major Facilitator Superfamily (MFS) functions in transport by a mechanism other than by secondary transport, and the only alternative function is that of transmembrane receptor [13, 20, 61]. Even this alternative functional type is exceedingly rare in nature.

A 2 TMS precursor could have duplicated to yield the 4 TMS unit that gave rise to all nine families in the TOG superfamily, but this has not been demonstrated. More importantly, the comparisons presented resolve some of the uncertainties in the evolutionary pathways of families exhibiting odd numbered topologies, such as the

PnuC family. Topological analyses of the entire superfamily reveal 7 well conserved average TMSs, as is expected given the dominant 7 TMS topology in most families within the superfamily. The N-terminal TMS of the 8 TMS topology was usually lost in the families under study, while loss of the C-terminal TMS occurred with a much lower frequency. Nevertheless, the N-terminal TMS, lost in many homologues, is present in enough members of the TOG superfamily to be visualized in the AveHAS plot (Figure 2).

The greatest topological variation is observed for the Sweet, NiCoT and PNaS families. In the Sweet family, 3 TMS homologues are found in prokaryotes in addition to the 7 TMS proteins found ubiquitously. In the NiCoT family, some members appear to have only 6 TMSs, due to loss of both TMSs 1 and 8 of the 8 TMS precursor. In the PNaS family, additional TMSs present in several family members proved to be due to fusions or late duplication events. Thus, 4 extra TMSs are homologous to the last 4 TMSs in some family members. This fact suggests that at least some members of the PNaS family arose by triplication of the 4 TMS repeat unit.

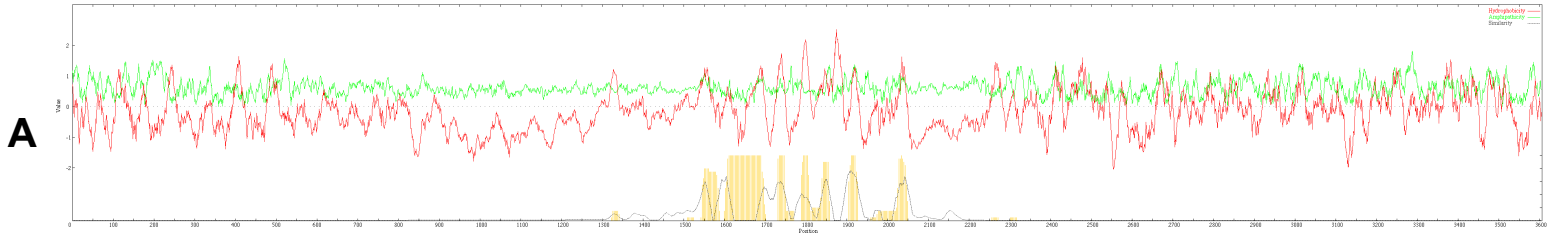
For the most part, the nine families within the TOG superfamily do not cluster according to mechanistic type or substrate specificity. Instead families are interspersed (Figs. 3A and 3B). The diversity within the TOG superfamily is reminiscent of the demonstrated or hypothesized alternative energy coupling mechanisms used by members of certain families found in TCDB. The ArsB transporters (TC# 3.A.4), members of the IT superfamily [62], can function either as secondary carriers or as ATP-driven primary active transporters, depending on the availability of the ArsA

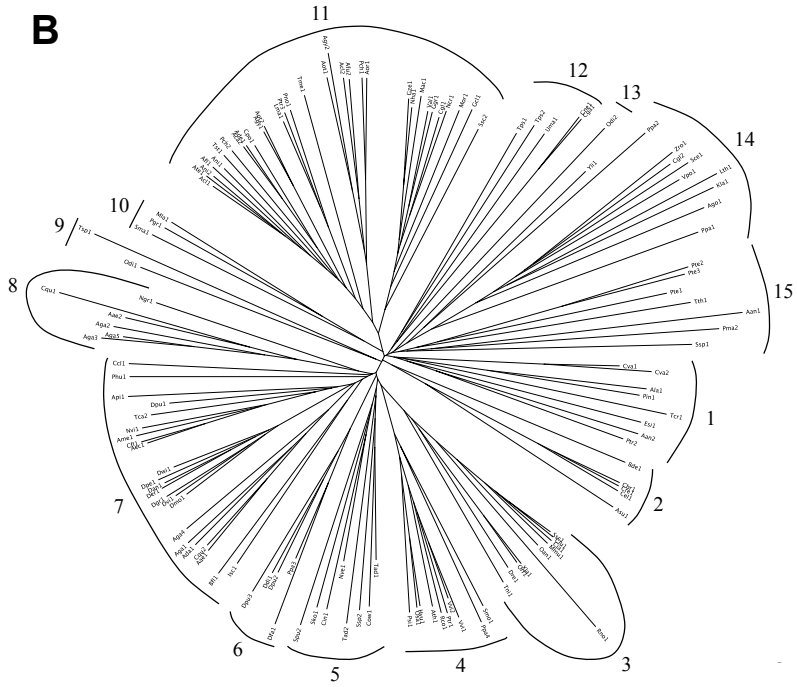
ATPase, and the same may be true of Acr3 porters (TC# 2.A.59) (see Castillo and Saier [63] and references cited therein). Members of the PTS Galactitol (Gat) family (TC# 4.A.5), but not members of the related PTS L-Ascorbate (L-Asc) family (TC# 4.A.7), appear to be capable of functioning either by group translocation involving the PTS energy coupling proteins or by secondary active transport when these proteins are lacking [64-65]. Evidence supports the suggestion that members of the PnuC family TC# 4.B.1 within the TOG superfamily can function by group translocation using ATP-dependent nicotinamide ribonucleoside kinase as the energy-coupling enzyme. However, many members of this family are encoded in genomes that lack this enzyme, supporting the conclusion that these porters function as secondary carriers [66]. Comparable studies have shown that members of the NaT-DC family TC# 3.B.1 catalyze sodium efflux in a process driven by decarboxylation of various carboxylic acids. However, all other members of the CPA superfamily function as secondary carriers (see the CPA superfamily in TCDB). A similar situation has been demonstrated for members of the ECF sub-superfamily of the ABC superfamily [67]. Some of these porters can transport vitamins such as biotin and thiamin either by ATP-dependent primary active transport or by pmf-driven secondary active transport [67] (Sun and Saier, unpublished results).

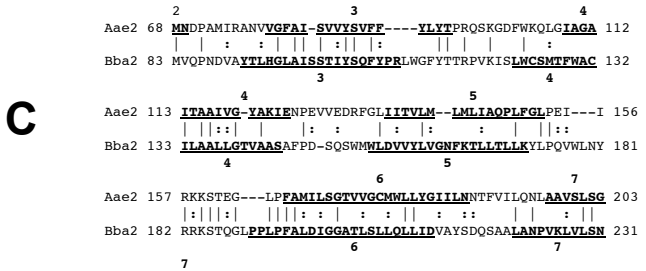
It is a common assumption that sequence similarity between visual rhodopsins and microbial rhodopsins is undetectable [68]. The GPCR family (TC# 9.A.14) sequence set is spread out in three separate network components, one of which, the Glutamate GPCR set, is distantly connected to the others, having only weak links to

other members of the TOG superfamily. One sequence, TC# 2.A.43.2.5 (PQ-loop repeat-containing protein from *A. thaliana*), which is a member of Lysosomal Cystine Transporter (LCT) family and closely related to Sweet, has a similarity of 0.00021 to Pfam family 7tm_1 which is the central node of the Rhodopsin GPCRs. Despite LCTs being larger than MRs, and while the LCT family is found exclusively in the eukaryotic domain, this is the most significant Pfam connection between any member of the GPCRs (TC# 9.A.14) and the rest of the TOG superfamily network. The similarity between LCT and Rhodopsin GPCRs has been noted elsewhere [14]. In summary, the connection between Rhodopsin GPCRs and the Lysosomal Cystine Transporter (LCT) and Sweet families is stronger than the intra-GPCR connections of many GPCRs to the most divergent Glutamate GPCRs. Thus, the concept of 7 TMS GPCRs being a closely related group of sequences, often taken for granted, is not valid, while the conclusion that sequence similarity between visual rhodopsin and bacteriorhodopsin is non-detectable is equally invalid.

This chapter is a reprint from material that appears in The Transporter-Opsin-G protein-coupled receptor (TOG) Superfamily in FEBS Journal in 2013. Yee DC, Shlykov MA, Västermark A, Reddy VS, Arora S, Sun EI, Saier MH Jr. The thesis author was co-first author of this paper.







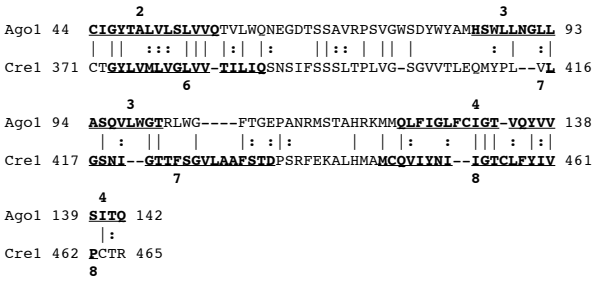


Figure S1: Avehas Plot (A), Phylogenetic Tree (B), and Binary Alignments (C) and (D) for the LCT Family

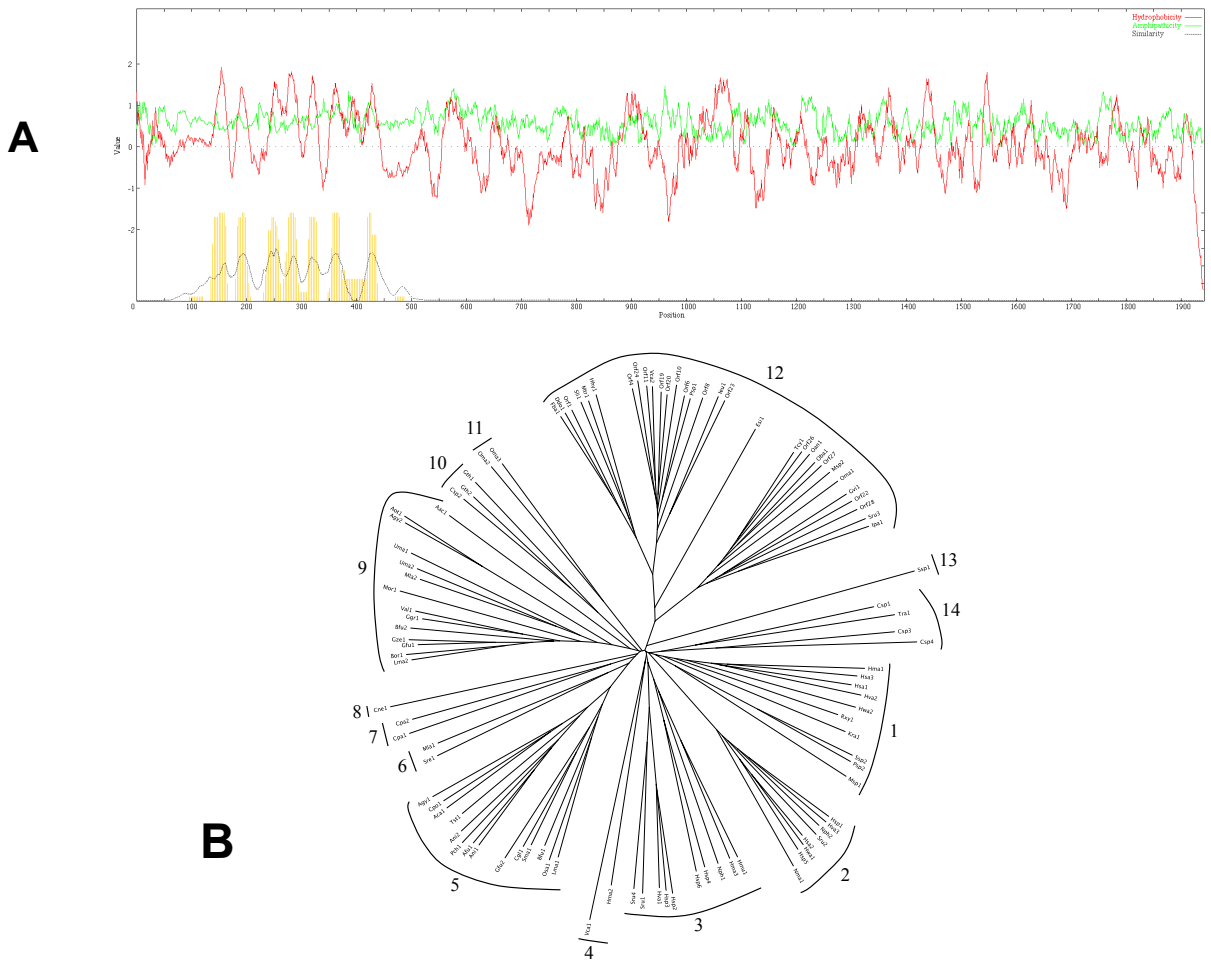
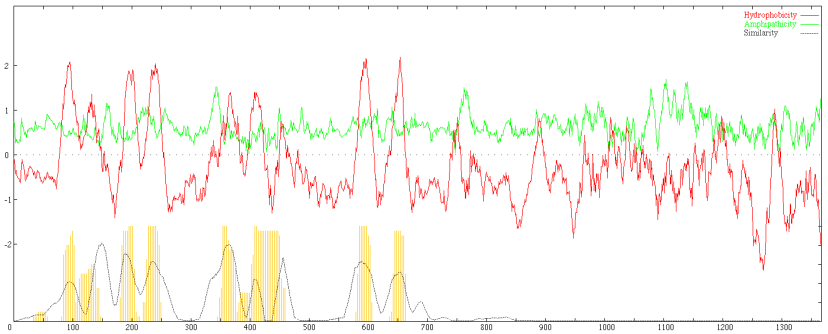
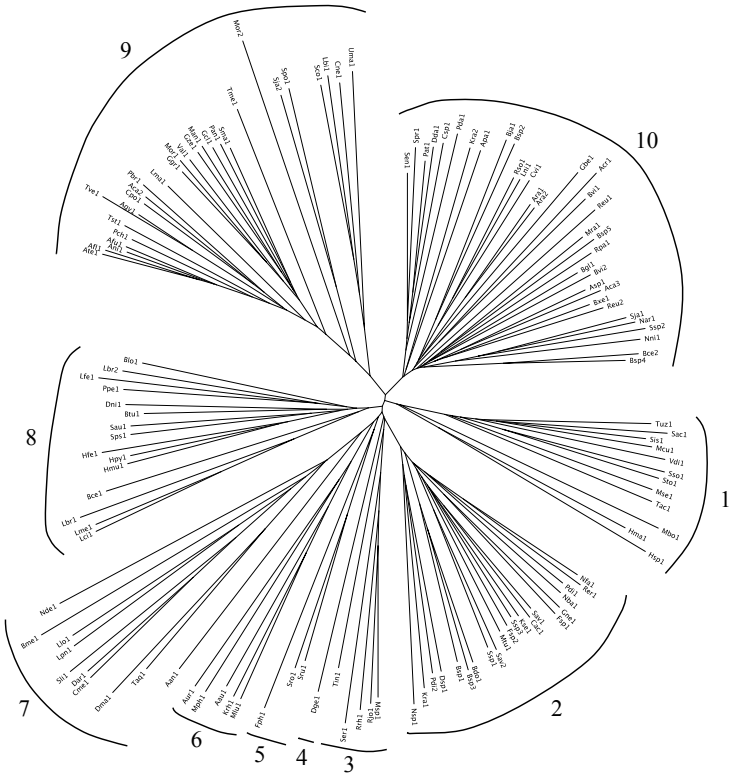


Figure S2: Avehas Plot (A) and Phylogenetic Tree (B) for the MR Family

A



B

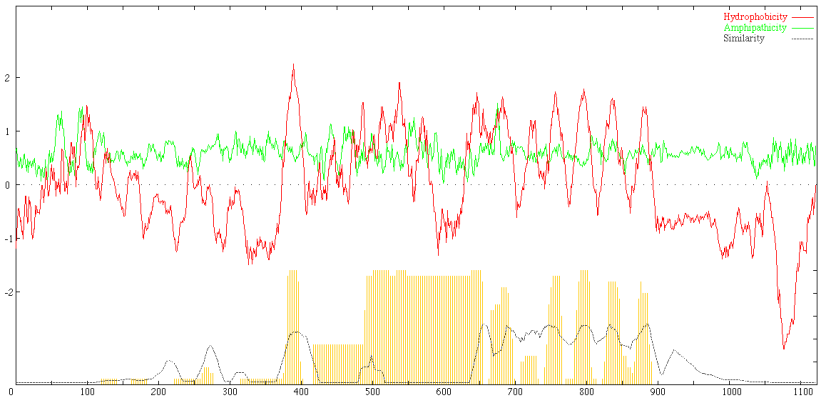


C

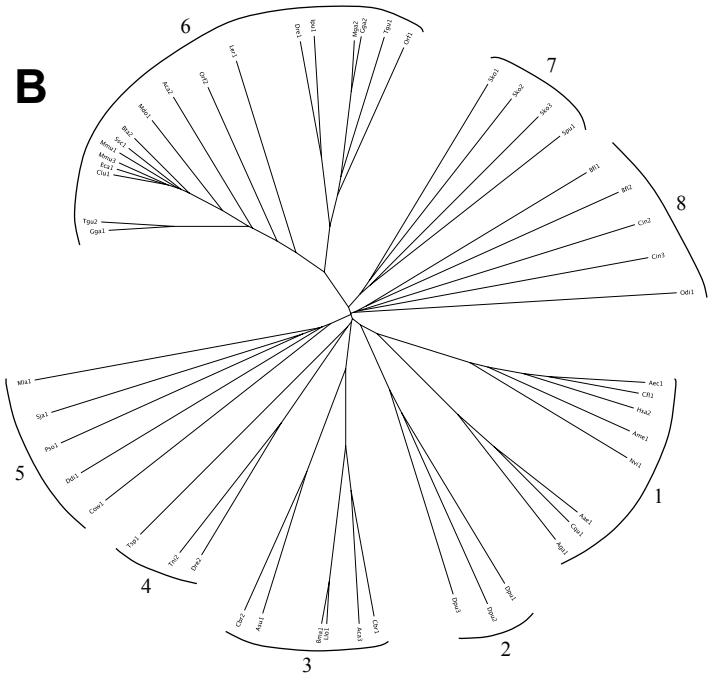
```
Pla1  25 EGAHLHCRNDHAQSHHTPHTHSPRPKFLIRSTANSAIERRGNRSVSPVRG 74
         |  | |   ||   |   | | | |  :       |  |:|      |||
Bja1 179 EHGHVHHDHDH.DTHEHDHAHIPTPADI......RAAKRKG......VRG 215
                                1                  2
Pla1  75 SMQLYLPIADLPVNVFLVLAMGAAVGFVSGMFGIGGGFLMTPLLIFIG.. 122
            : | :   |    :   : |      | | ||    |  :   :|
Bja1 216 MAAMILSVGLRPCTGAILVLLFAV...TQGAFSIG...VMSAIVMSVGTA 259
                     4                                5
                2                                    3
Pla1 123 IT.PAVAVASVASHIAASSFSGAI.SYWRR...RAIDPALASVLLCGGVT 167
         ||  | |   | |   |  | | : | | |   | :  |  ||:|
Bja1 260 ITVSALALMTVFSKRLALRFAGGVDSPWARRVERGLKIAGGSVIL...LF 306
                5                                    6
               3
Pla1 168 GTALGVWTFTQ 178
         |  | |  |||
Bja1 307 GMMLLVASFTQ 317
               6
```

Figure S3: Avehas Plot (A), Phylogenetic Tree (B), and (C) Binary Alignment for the NiCoT Family

A



B



C

```
                    2                                   3
Dre1   1 ILFIVPIYAFDSWLSLLFFTNEEYYVYFDTVRDCYEAFVIYNFLSLCYEY 50
         :| ::|: :  ::||: :| :  ||||  : |:     : :: |::
Gfo1   1 LLMVMPLPSAIAYLSIFYFISTPYYVYKE--RNNIDKLLLKK-LAIS-SL 46
                        2                                3
              3                                          4
Dre1  51 LGGESAIMAEIRGKPIQSSFVYGTCCLWGRTYSIGFLRFCKQATLQFCVV 100
         :|  : |:  |  |||      |   :    : :  ||  |:  ||   :
Gfo1  47 VGVVAGIILLIYAKPILLKKALGVFII---LFVLNSLRPKKE--LQ---L 88
                   3
                 4                                5
Dre1 101 KPLMAIITVILQAFGKYRDGDFNAAGGYLYVMIIYNVSVSLSLFALFLF 149
          | |      |   | :  | | : || |||||: | :  :  |   :|
Gfo1  89 GPKMKYTFGFL---GGFFSGVF-STGGPLYVMIVKNATPDVRTFRATMF 133
                              4                          5
```

D

```
                        2                                3
Tsp1 35 LLGEPIHLAIGTSLTCI.VLSSLSASLTHIRR.GAVLYRVVLLKEVFSVP 82
        :| :|| | : |  | : | |      | |:|     ||   ::: :| ||
Cre2  1 MLRDPITLGLATAATWLSVASAVTQILCHLRNYTEPLYQRYIIRIIFLVP 50
                        1                          2
              3
Tsp1 83 FAVLGAYLSSM 93
        |     :|| |
Cre2 51 FYGVTSWLSIM 61
              2
```

Figure S4: Avehas Plot (A), Phylogenetic Tree (B), and Binary Alignments (C) & (D) for the OST Family

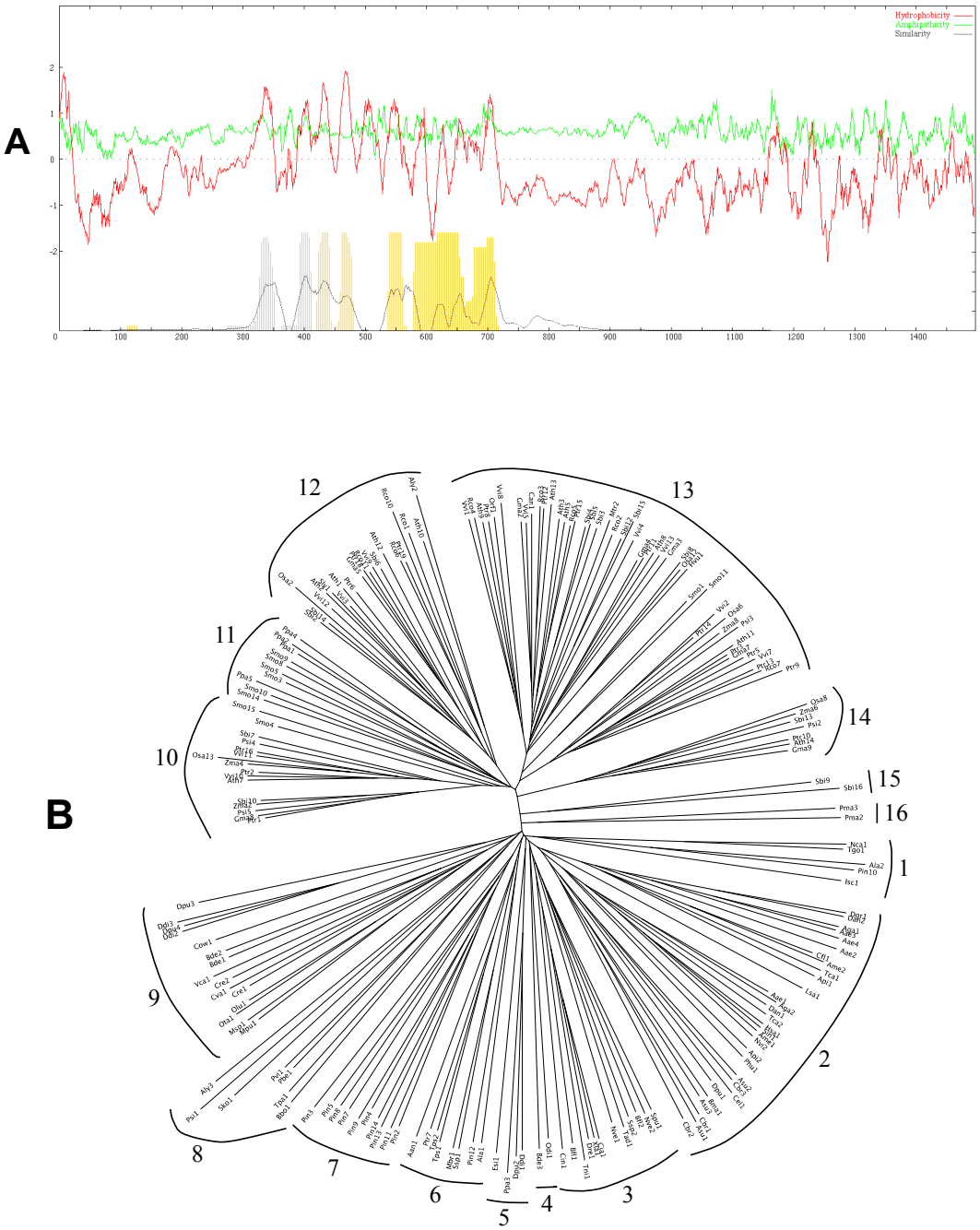


Figure S5: Avehas Plot (A), Phylogenetic Tree (B), and Binary Alignments (C) & (D) for the Sweet Family

C

```
                5                         6                    7
Rco4 145  VAPLAIVRKVIKTKSVEFMPFSLSFFLTLSAVMWFFYGFLKKDLYVAVPN 194
          :||:  : | :  ||: |: :     :     |  | | | |:|   : :
Ath8 249  LAPIKPLAKFLTFKSIVFLTWWQGIIVAFLFSMGLFKGSLAKELKTRIQD 298
                                   6
                    7
Rco4 195 -ILGFMFGVLQMI-LYLIYRNPKKTGDDDQKANELPNQHSIIDVAKLNTR 242
          |:    |:  :: ||:    | | |:   :   : : :: :||
Ath8 299 YIICIEMGIAAVVHLYVFPAAPYKRGERCVRNVAVMSDYASLDVPPDPEE 348
                    7

Rco4 243 VSCCEPNATTVAHSRNDREEQQ 264
         |   |   |   : |:|  |::
Ath8 349 VKDSE-RTTRTRYGRHDEREKR 369
```

D

```
                          4                              5
Asu3 1   MKRRQLRFAAIE-IIFLSLIHLYVEYSQHAKEIILDHLGYIC--VAFNVA 47
         :||  |::| :: :: |: : :    | |  :| |:  ||    : :|::
Ncr1 137 IKRGILQYAWMKPVLALAAVIMKATGSYHEGDIKLNS-GYFWSGIIYNIS 185
                       4                                5
              5                                   6
Asu3 48  -TVAAPLLALGEVIRSKSTENL-PLP--LCLANLLVTSEW--LLYGFLVE 91
          ||:   | |  |  :   :   |:|  ||:  ::  | |     | ||
Ncr1 186 VTVSLYCLGLFWVCMNNDLKPFRPMPKFLCVKAIIFASYWQGFALGILV- 234
               5                                        6

Asu3 92  DFFIKFPN 99
          |   |||
Ncr1 235 -FLGAFPN 241
             6
```

Figure S5, Continued

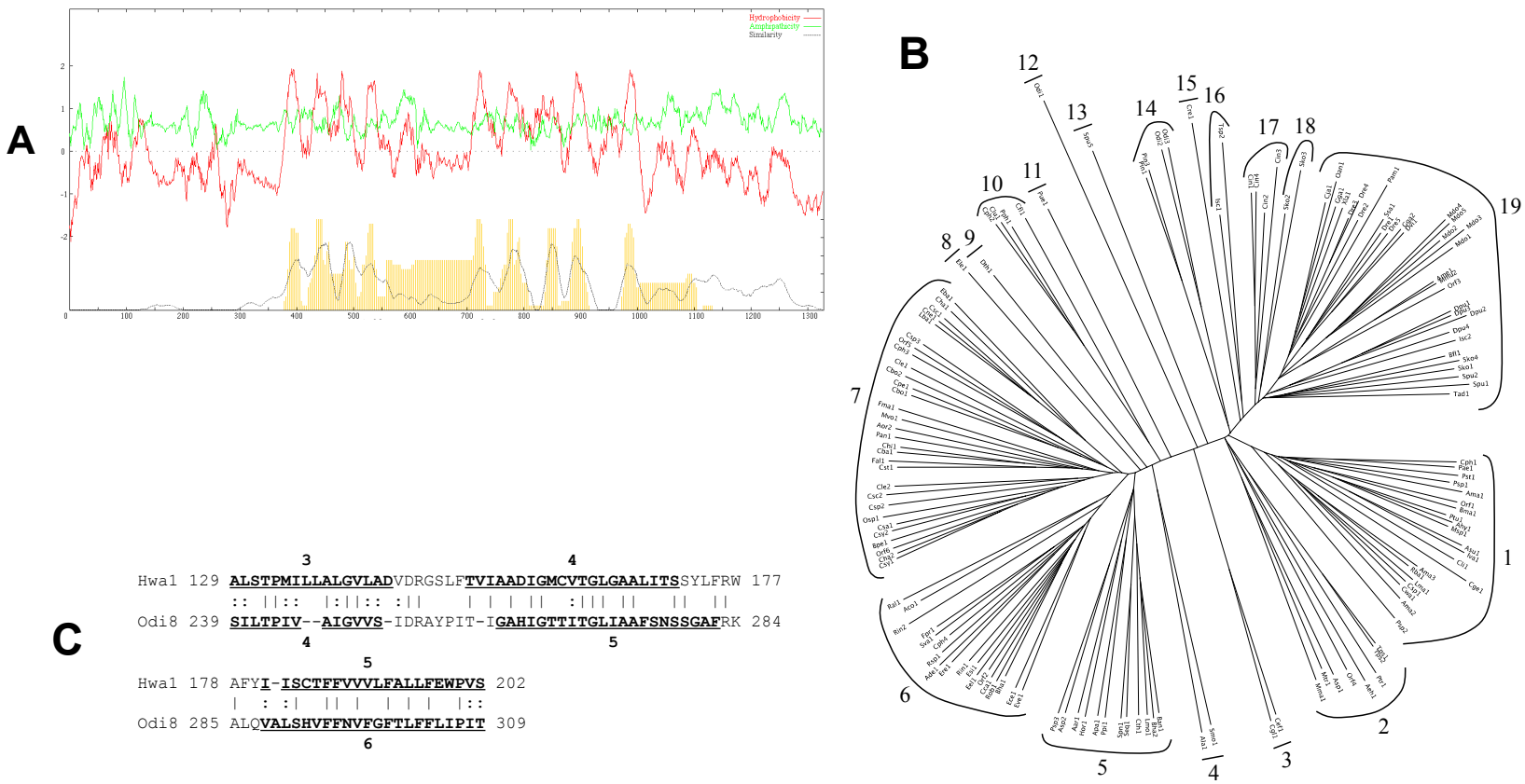


Figure S6: Avehas Plot (A), Phylogenetic Tree (B), and Binary Alignment (C) for the PNaS Family

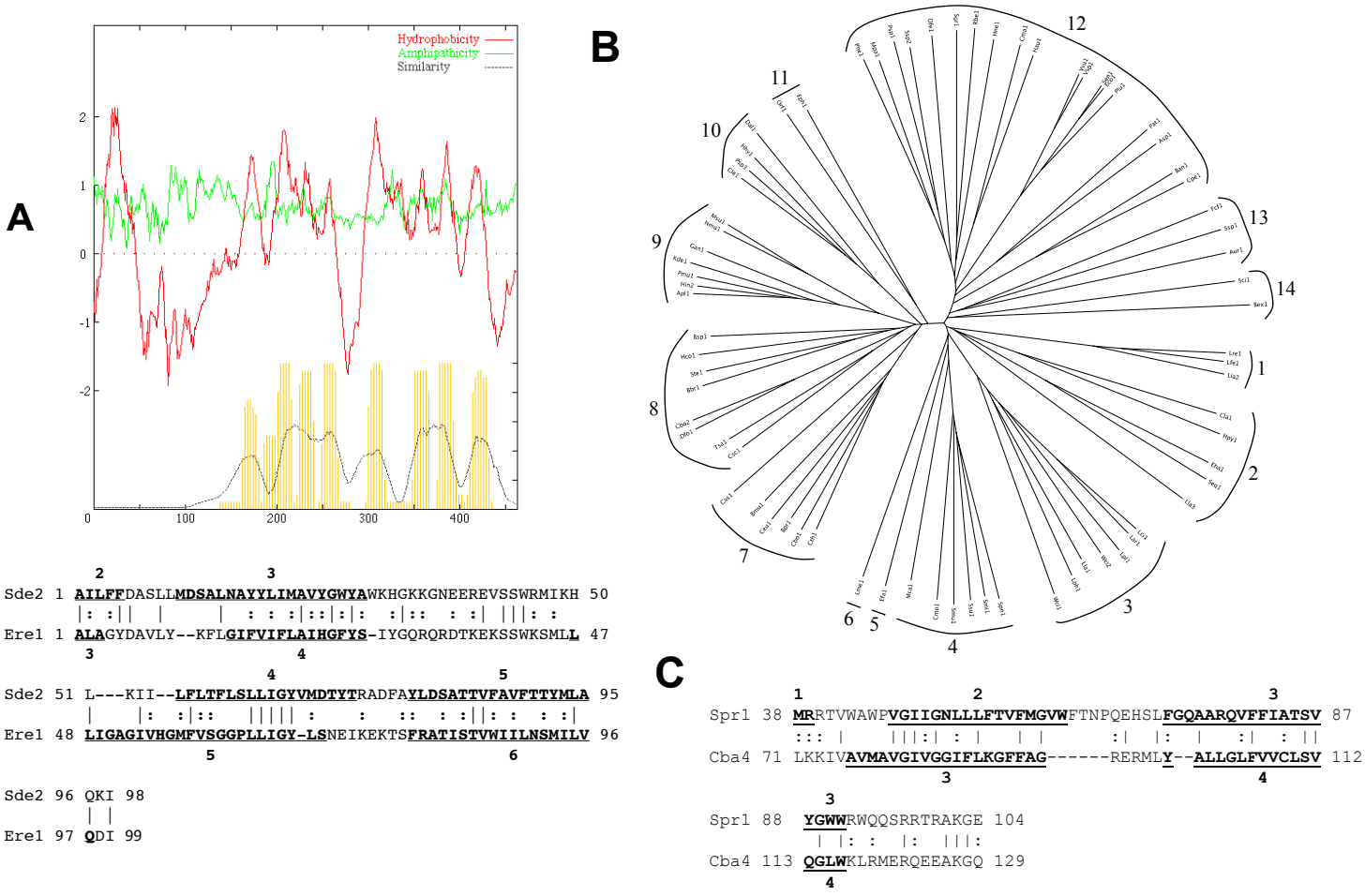


Figure S7: Avehas Plot (A), Phylogenetic Tree (B), and Binary Alignments (C) & (D) for the PnuC Family

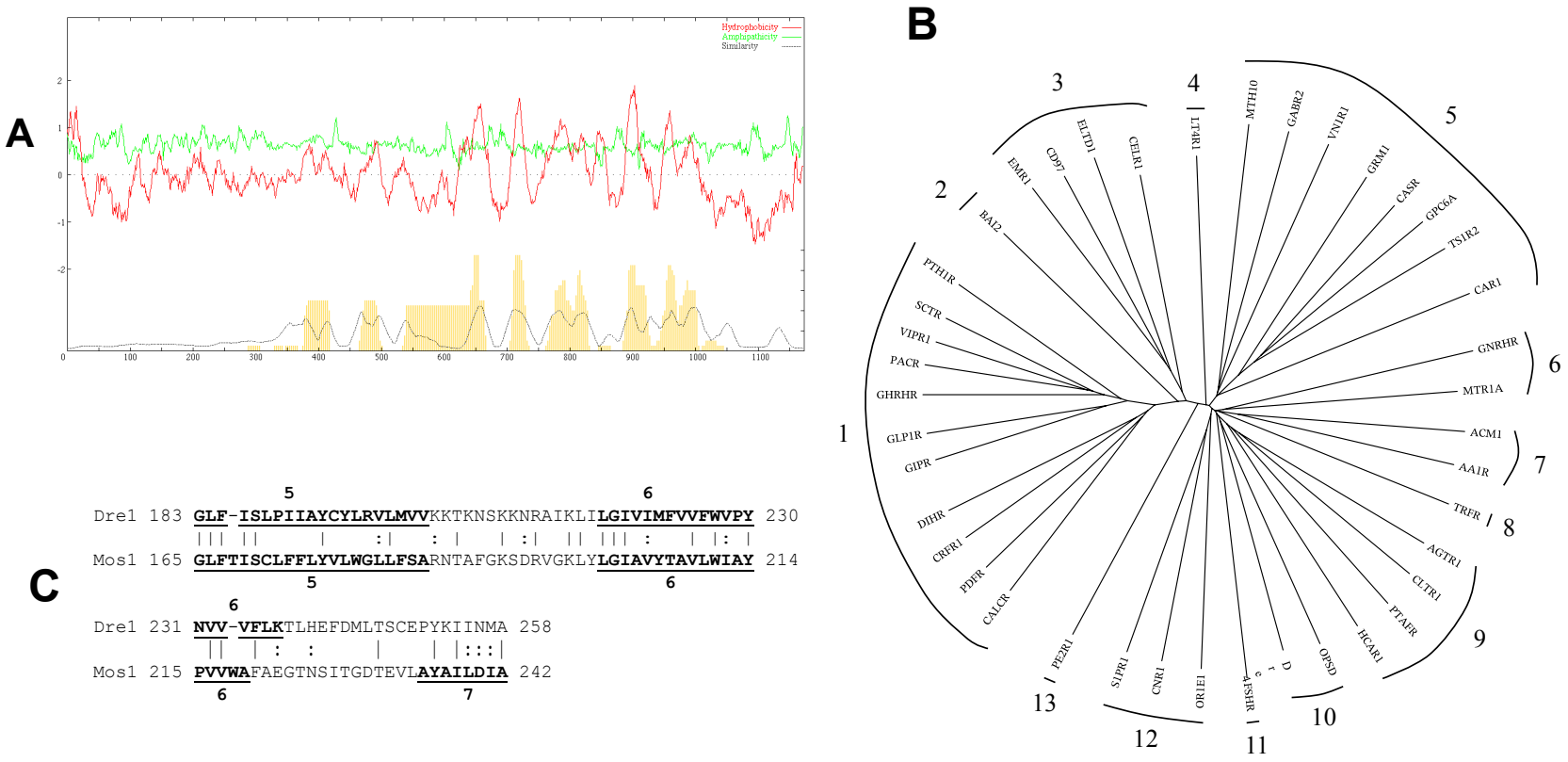
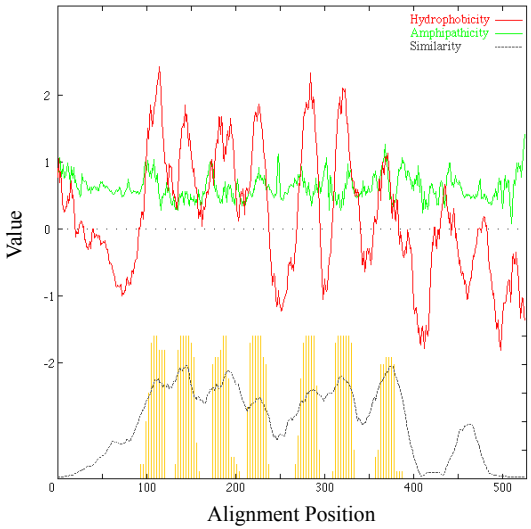


Figure S8: Avehas Plot (A), Phylogenetic Tree (B), Binary Alignments (C) & (D), Secondary Avehas Plot (E), and Secondary Phylogenetic Tree (F) for the GPCR Family

D

```
                                                  1
Dre5  41 GPTHSHHHD-PPHPFPTVDVPDHAHYI-IGSVILIVGITGVIGNALVVYV 88
         || ||||   |  |||:   |   |   :| | : |  |      ||  :
Cga1  27 GPHHSHHLPLPTAPFPSA-TPKFLHATRVGHVSIWV-FTAFFIAGLVAVL 74
                                                  1
                                  2
Dre5  89 FCRSRTLRTAGNMFIVNLAVADFLMSVTQSPVFFAASLHRRWV--FGERP 136
         |  ||| |    :|    ||   : ::|   :  |  :   :|  :|  |
Cga1  75 FLTSRTER-KNRLFHGISAVILTVSALTY--MSLATHIGSTFVPIYGP-P 120
                                  2
                          3
Dre5  137 CE---LYAFCGALFGICSMMTLTAIAADRCLAITQPLALVSRVSRRKAG- 182
               |  |   :| |       ||     :     || ||  :   |
Cga1  121 GHHEPLVHFFRQVFSI--RYIDAAITGPLIILALSRLAGVSPATALSAAL 168
                                  3
                     4
Dre5 183 AVLVVV---WLYSLGWSLPPFFGWSAYVPEGLQTSCSWDYMTFTPSVRAY 229
         | ||||   |  |:|   |    || :
Cga1 169 AELVVVYSAWAGSVGGGWP----WSEH-G
              4
            5
Dre5 230 TILL 233
         |:||
Cga1 214 TVLL 217
```

E



F

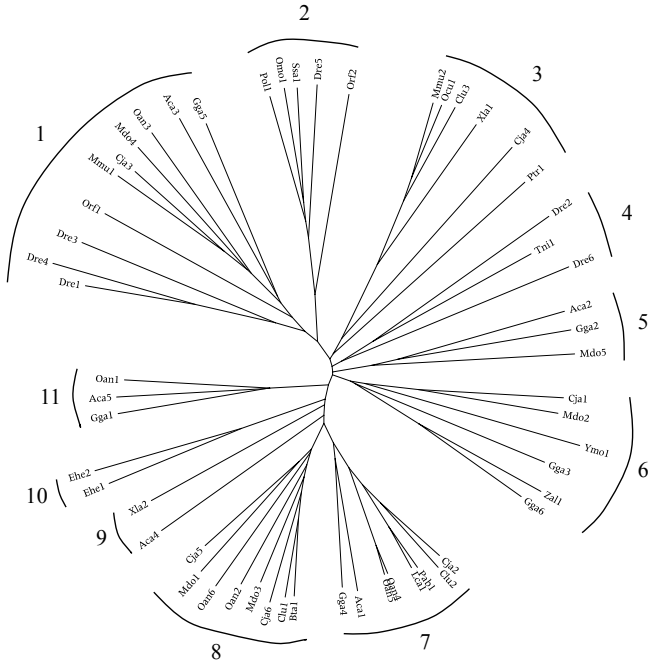


Figure S8, Continued

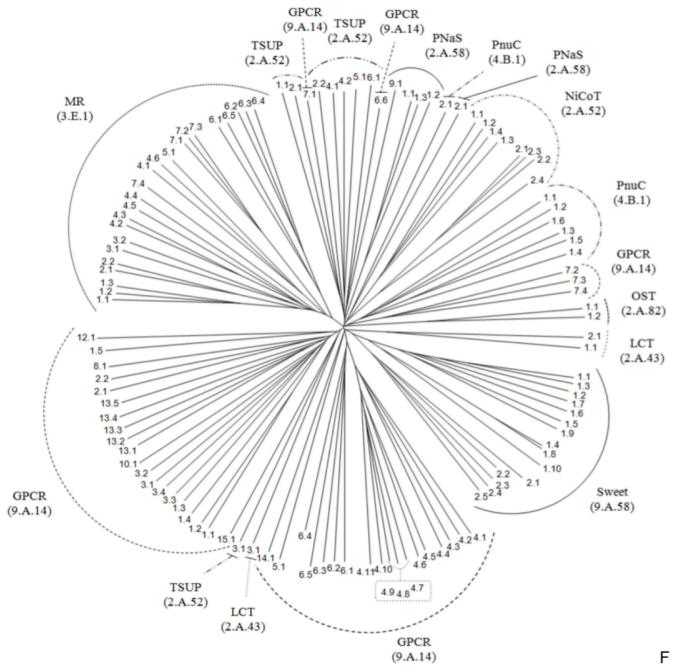


Figure S9: Phylogenetic tree from the ClustalX-based neighbor-joining program

Table S1: The highest comparison scores between TOG superfamily (left column) and the MC and MIP families

	2.A.29 **(MC) 2 TMSs**	2.A.29 **(MC) 3 TMSs**	1.A.8 **(MIP) 2 TMSs**	1.A.8 **(MIP) 3 TMSs**
9.A.14 **(GPCR)**	6.7 S.D.	N/A	5.6 S.D.	2.2 S.D.
2.A.43 **(LCT)**	8.4 S.D.	N/A	9.4 S.D.	8.8 S.D.
3.E.1 **(MR)**	8.7 S.D.	N/A	10.4 S.D.	8.6 S.D.
2.A.52 **(NiCoT)**	9.5 S.D.	9.5 S.D.	8.9 S.D.	7.4 S.D.
2.A.82 **(OST)**	9.1 S.D.	N/A	8.2 S.D.	8.2 S.D.
4.B.1 **(PnuC)**	8.8 S.D.	N/A	7.8 S.D.	7.0 S.D.
2.A.102 **(TSUP)**	9.4 S.D.	9.4 S.D.	10.2 S.D.	9.5 S.D.
2.A.58 **(PNaS)**	9.5 S.D.	N/A	10.5 S.D.	8.5 S.D.
9.A.58 **(Sweet)**	5.9 S.D.	N/A	8.8 S.D.	8.8 S.D.
Average	8.4 S.D.	9.5 S.D.	8.9 S.D.	7.7 S.D.

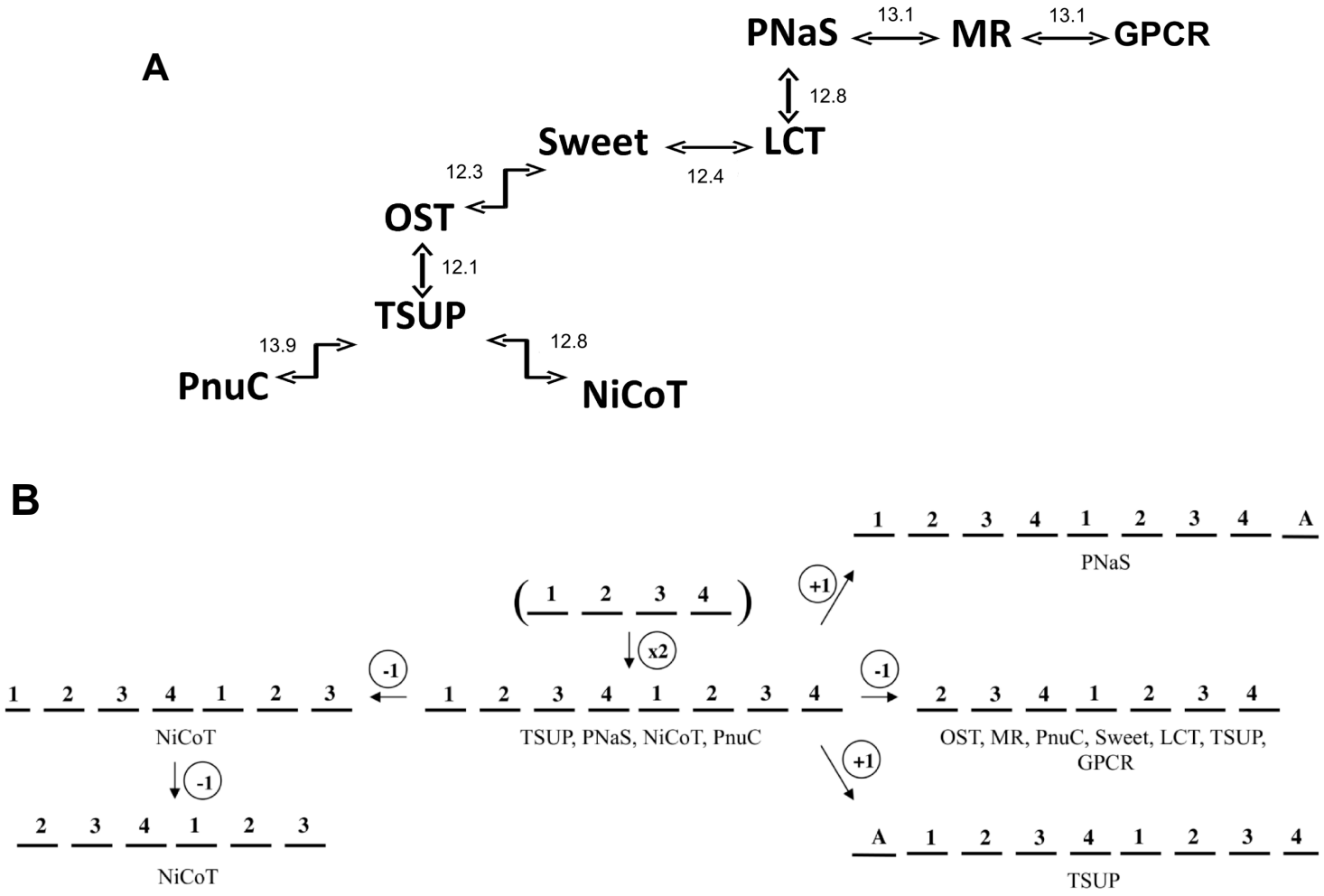


Figure 1: TOG Superfamily Homology (A) and Proposed evolutionary pathway for the appearance of nine reconized families within the TOG superfamily

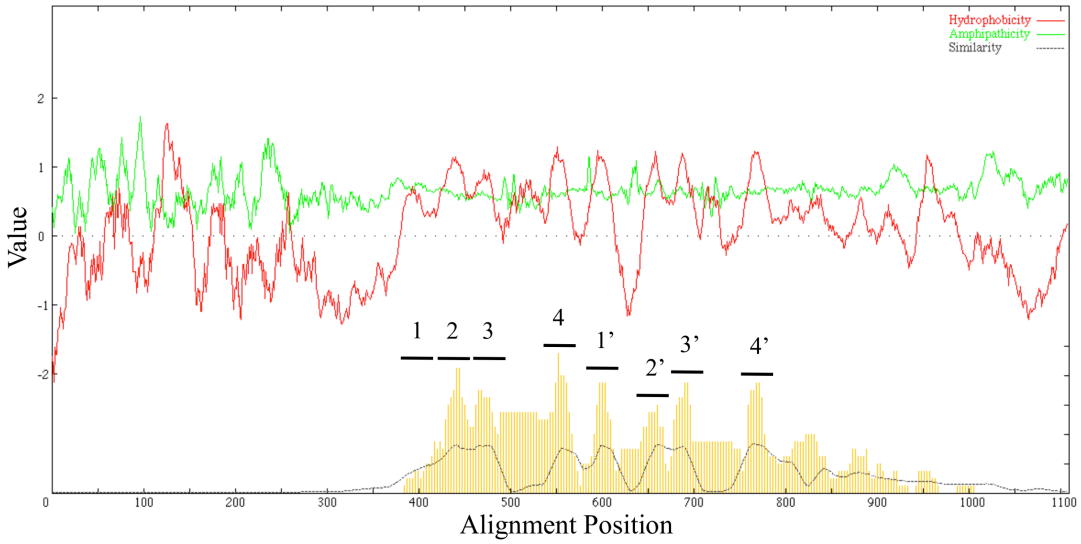


Figure 2: Average hydropathy, amphipathicity, and similarity (AveHAS) plots based on a ClustalX multiple alignment

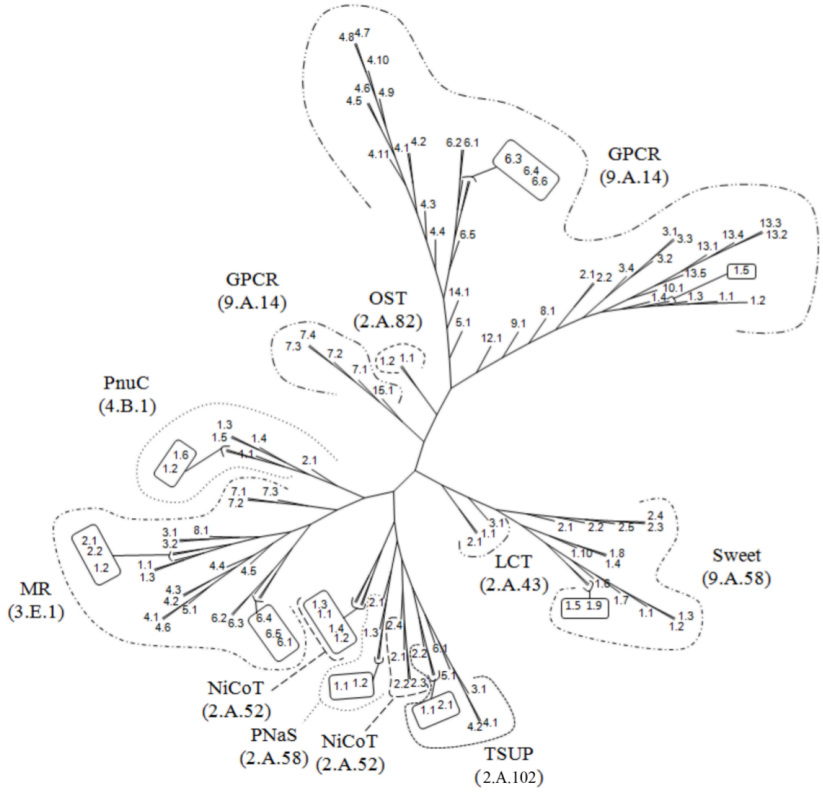


Figure 3: Fitch tree from the BLAST-derived SFT1 program showing the proteins of families within the TOG superfamily

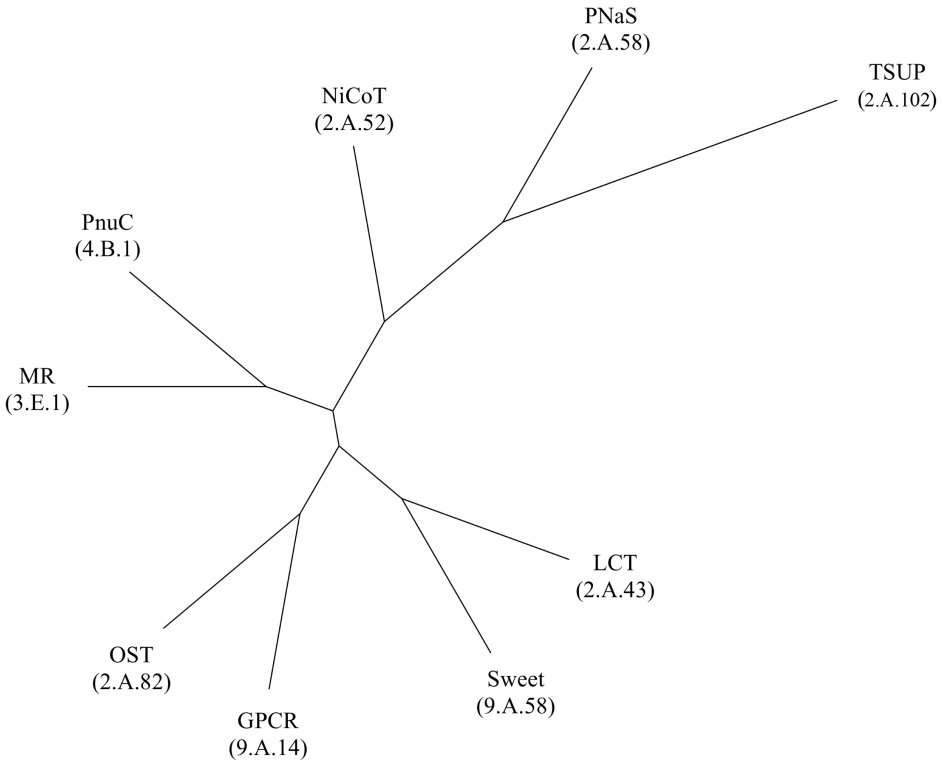


Figure 4: SFT2-based tree showing the relationships of the TOG superfamily families

```
                1                       2
Pas1 12   IAFLAGFFDAIAG-G-GGLITLPALFLAGIEPISAMATNKLQATSASISA 59
          || : ||:| | | | |  : :  :|  |:| : ||:  ||  |  :| |
Pas1 139  IAPVLGFYDGIFGPGTGSFLMMSFMFFCGLEMMQAMSFTKLANASCNIGA 188
               5                       6
                                       3                   4
Pas1 60   TVTFARNGLIEWRNSVFLVLFSLLGGVSGALFVSVVSKRVLEACVPILLI 109
             |   |:| |  :: : : |  |   || :   :  |::   : ::
Pas1 189  LFIFITKGVIIWPLALAMAIASFAGAQLGARYAVKIGPRLVRPMLILVCC 238
                     7                                   8
                  4
Pas1 110  VVAIYFIFA 118
          |:||  | |
Pas1 239  VLAIKLISA 109
               8
```

Figure 5: GSAT alignment demonstrating a 4 TMS repeat within Pas1 of the TSUP family.

Table 1: Summary of nine TOG Superfamily family members. The family number, name, abbreviation, TC number, number of TMSs, normal protein size range, dominant topology, TMS gain or loss, organismal distribution, and Pfam id are presented.

Family #	Family Name	Family Abb'n	TC #	# TMSs	Common Protein Size Range	Topology	TMS Gain or Loss (Primary)	Organismal Distribution	Pfam id
1	Ion-Translocating Microbial Rhodopsin	MR	3.E.1	7	250-350	3 + 1 + 3	7 arose from 8 by loss of the N-terminal TMS.	Archaea, Bacteria, Eukaryota	Bac_rhodopsin
2	Sweet[1]	Sweet	9.A.58	3 or 7	200-290	3 + 1 + 3 (3)	7 arose from 8 by loss of the N-terminal TMS.	Archaea, Bacteria, Eukaryota	MtN3_slv
3	Nicotinamide Ribonucleotide Uptake Permease	PnuC	4.B.1	7 or 8	210-270	3 + 1 + 3 (4 + 4)	7 arose from 8 by loss of the N-terminal TMS.	Bacteria, Eukaryota	NMN_transporter, AAA_28
4	4-Toluene Sulfonate Uptake Permease[2]	TSUP	2.A.102	7-9	250-600	4 + 4 (3 + 1 + 3; 1 + 4 + 4)	8 arose from internal duplication of 4 TMSs	Archaea, Bacteria, Eukaryota	TauE
5	Ni^{2+}-Co^{2+} Transporter[3]	NiCoT	2.A.52	6-8	300-380	4 + 4 (3 + 1 + 3; 2 + 1 + 1 + 2)	8 arose from internal duplication of 4 TMSs	Archaea, Bacteria, Eukaryota	NicO
6	Organic Solute Transporter	OST	2.A.82	7	180-400	3 + 1 + 3	7 arose from 8 by loss of the N-terminal TMS.	Eukaryota	Solute_trans_a
7	Phosphate:Na^+ Symporter[4]	PNaS	2.A.58	8 or 9	500-700	4 + 4 (4 + 4 + 1)	8 arose from internal duplication of 4 TMSs	Bacteria, Eukaryota	Na_Pi_cotrans
8	Lysosomal Cystine Transporter	LCT	2.A.43	7	300-400	3 + 1 + 3	7 arose from 8 by loss of the N-terminal TMS.	Eukaryota	PQ-loop
9	G-protein Coupled Receptor	GPCR	9.A.14	7	300-1200	3 + 1 + 3	7 arose from 8 by loss of the N-terminal TMS 1.	Eukaryota	7tm_1, 7tm_2, 7_tm3

[1] 3 arose from 4 TMSs by loss of one TMS.
[2] 7 arose from 8 by loss of the N-terminal TMS; 9 arose from 8 by gain of an N-terminal TMS.
[3] 7 arose from 8 by loss of the 8th (C-terminal) TMS; 6 arose from 7 by loss of the N-terminal TMS.
[4] 9 arose from 8 by gain of a C-terminal TMS.

Table 2: The highest comparison scores between TOG superfamily members. Values above 12.0 S.D., which are considered sufficient to establish homology and interconnect families, are shaded. These values are sufficient to establish homology.

	9.A.14 **GPCR**	2.A.102 **TSUP**	2.A.82 **OST**	3.E.1 **MR**	9.A.58 **Sweet**	2.A.52 **NiCoT**	2.A.58 **PNaS**	4.B.1 **PnuC**
9.A.14 **GPCR**								
2.A.102 **TSUP**	11.4 S.D. (2 TMSs)							
2.A.82 **OST**	10.8 S.D. (2 TMSs)	12.1 S.D. (2 TMSs)						
3.E.1 **MR**	13.1 S.D. (2 TMSs)	11.0 S.D. (2 TMSs)	12.4 S.D. (3 TMSs)					
9.A.58 **Sweet**	12.3 S.D. (2 TMSs)	10.1 S.D. (3 TMSs)	12.3 S.D. (2 TMSs)	11.2 S.D. (2 TMSs)				
2.A.52 **NiCoT**	10.4 S.D. (2 TMSs)	12.8 S.D. (3 TMSs)	11.9 S.D. (2 TMSs)	10.7 S.D. (2 TMSs)	11.8 S.D. (3 TMSs)			
2.A.58 **PNaS**	11.2 S.D. (2 TMSs)	10.6 S.D. (2 TMSs)	11.5 S.D. (4 TMSs)	13.1 S.D. (3 TMSs)	11.6 S.D. (2 TMSs)	12.2 S.D. (3 TMSs)		
4.B.1 **PnuC**	10.2 S.D. (3 TMSs)	13.9 S.D. (2 TMSs)	11.4 S.D. (2 TMSs)	10.9 S.D. (2 TMSs)	13.1 S.D. (3 TMSs)	11.2 S.D. (2 TMSs)	9.8 S.D. (2 TMSs)	
2.A.43 **LCT**	9.4 S.D. (3 TMSs)	11.3 S.D. (2 TMSs)	11.1 S.D. (2 TMSs)	11.0 S.D. (2 TMSs)	13.6 S.D. (2 TMSs)	11.5 S.D. (2 TMSs)	12.8 S.D. (3 TMSs)	11.3 S.D. (3 TMSs)

[1]Average comparison score and average number of TMSs in all TOG superfamily alignments: 2.3 TMSs; 11.5 S.D.
[2]Average comparison score and average number of TMSs in alignments used to establish homology and interconnect all families within the TOG superfamily: 2.5 TMSs; 12.8

REFERENCES

1. Saier, M. H., Jr. (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution, *Microbiol Rev.* **58**, 71-93.

2. Saier, M. H., Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters, *Microbiol Mol Biol Rev.* **64**, 354-411.

3. 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 Res.* **34**, D181-6.

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

5. Busch, W. & Saier, M. H., Jr. (2002) The transporter classification (TC) system, 2002, *Crit Rev Biochem Mol Biol.* **37**, 287-337.

6. 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, *Biol Chem.* **392**, 5-12.

7. 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, *Mol Membr Biol.* **21**, 171-81.

8. Mansour, N. M., Sawhney, M., Tamang, D. G., Vogl, C. & Saier, M. H., Jr. (2007) The bile/arsenite/riboflavin transporter (BART) superfamily, *Febs J.* **274**, 612-29.

9. Furutani, Y. & Kandori, H. (2002) Internal water molecules of archaeal rhodopsins (Review), *Mol Membr Biol.* **19**, 257-65.

10. Hirai, T., Subramaniam, S. & Lanyi, J. K. (2009) Structural snapshots of conformational changes in a seven-helix membrane protein: lessons from bacteriorhodopsin, *Curr Opin Struct Biol.* **19**, 433-9.

11. Zhou, X. E., Melcher, K. & Xu, H. E. (2012) Structure and activation of rhodopsin, *Acta Pharmacol Sin.* **33**, 291-9.

12. Pearson, W. R. (1998) Empirical statistical estimates for sequence similarity searches, *J Mol Biol.* **276**, 71-84

13. Reddy, V. S. & Saier, M. H., Jr. (2012) BioV Suite--a collection of programs for the study of transport protein evolution, *Febs J.* **279**, 2036-46.

14. Zhai, Y., Heijne, W. H., Smith, D. W. & Saier, M. H., Jr. (2001) Homologues of archaeal rhodopsins in plants, animals and fungi: structural and functional predications for a putative fungal chaperone protein, *Biochim Biophys Acta.* **1511**, 206-23.

15. Chen, J. S., Reddy, V., Chen, J. H., Shlykov, M. A., Zheng, W. H., Cho, J., Yen, M. R. & Saier, M. H., Jr. (2011) Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments, *J Mol Microbiol Biotechnol.* **21**, 83-96.

16. Yen, M. R., Chen, J. S., Marquez, J. L., Sun, E. I. & Saier, M. H. (2010) Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters, *Methods Mol Biol.* **637**, 47-6

17. Yen, M. R., Choi, J. & Saier, M. H., Jr. (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution, *J Mol Microbiol Biotechnol.* **17**, 163-76.

18. Shlykov, M. A., Zheng, W. H., Chen, J. S. & Saier, M. H., Jr. (2012) Bioinformatic characterization of the 4-Toluene Sulfonate Uptake Permease (TSUP) family of transmembrane proteins, *Biochim Biophys Acta.* **1818**, 703-17.

19. Chung, Y. J., Krueger, C., Metzgar, D. & Saier, M. H., Jr. (2001) Size comparisons among integral membrane transport protein homologues in bacteria, Archaea, and Eucarya, *J Bacteriol.* **183**, 1012-21.

20. Saier, M. H., Jr., Beatty, J. T., Goffeau, A., Harley, K. T., Heijne, W. H., Huang, S. C., Jack, D. L., Jahn, P. S., Lew, K., Liu, J., Pao, S. S., Paulsen, I. T., Tseng, T. T. & Virk, P. S. (1999) The major facilitator superfamily, *J Mol Microbiol Biotechnol.* **1**, 257-79.

21. Iwig, J. S., Rowe, J. L. & Chivers, P. T. (2006) Nickel homeostasis in Escherichia coli - the rcnR-rcnA efflux pathway and its linkage to NikR function, *Mol Microbiol.* **62**, 252-62.

22. Rodrigue, A., Effantin, G. & Mandrand-Berthelot, M. A. (2005) Identification of rcnA (yohM), a nickel and cobalt resistance gene in Escherichia coli, *J Bacteriol.* **187**, 2912-6.

23. Dawson, P. A., Hubbert, M., Haywood, J., Craddock, A. L., Zerangue, N., Christian, W. V. & Ballatori, N. (2005) The heteromeric organic solute transporter

alpha-beta, Ostalpha-Ostbeta, is an ileal basolateral bile acid transporter, *J Biol Chem.* **280**, 6960-8.

24. Seward, D. J., Koh, A. S., Boyer, J. L. & Ballatori, N. (2003) Functional complementation between a novel mammalian polygenic transport complex and an evolutionarily ancient organic solute transporter, OSTalpha-OSTbeta, *J Biol Chem.* **278**, 27473-82.

25. Wang, W., Seward, D. J., Li, L., Boyer, J. L. & Ballatori, N. (2001) Expression cloning of two genes that together mediate organic solute and steroid transport in the liver of a marine vertebrate, *Proc Natl Acad Sci U S A.* **98**, 9431-6.

26. Takanaga, H. & Frommer, W. B. (2010) Facilitative plasma membrane transporters function during ER transit, *Faseb* . **24**, 2849-58.

27. Chen, L. Q., Hou, B. H., Lalonde, S., Takanaga, H., Hartung, M. L., Qu, X. Q., Guo, W. J., Kim, J. G., Underwood, W., Chaudhuri, B., Chermak, D., Antony, G., White, F. F., Somerville, S. C., Mudgett, M. B. & Frommer, W. B. (2010) Sugar transporters for intercellular exchange and nutrition of pathogens, *Nature.* **468**, 527-32.

28. Saier, M. H., Jr. (2003) Tracing pathways of transport protein evolution, *Mol Microbiol.* **48**, 1145-56.

29. Ghezzi, C., Murer, H. & Forster, I. C. (2009) Substrate interactions of the electroneutral Na+-coupled inorganic phosphate cotransporter (NaPi-IIc), *J Physiol.* **587**, 4293-307.

30. Foster, J. W., Park, Y. K., Penfound, T., Fenger, T. & Spector, M. P. (1990) Regulation of NAD metabolism in Salmonella typhimurium: molecular sequence analysis of the bifunctional nadR regulator and the nadA-pnuC operon, *J Bacteriol.* **172**, 4187-96.

31. Merdanovic, M., Sauer, E. & Reidl, J. (2005) Coupling of NAD+ biosynthesis and nicotinamide ribosyl transport: characterization of NadR ribonucleotide kinase mutants of Haemophilus influenzae, *J Bacteriol.* **187**, 4410-20.

32. Penfound, T. & Foster, J. W. (1999) NAD-dependent DNA-binding activity of the bifunctional NadR regulator of Salmonella typhimurium, *J Bacteriol.* **181**, 648-55.

33. Kurnasov, O. V., Polanuyer, B. M., Ananta, S., Sloutsky, R., Tam, A., Gerdes, S. Y. & Osterman, A. L. (2002) Ribosylnicotinamide kinase domain of NadR protein: identification and implications in NAD biosynthesis, *J Bacteriol.* **184**, 6906-17.

34. Singh, S. K., Kurnasov, O. V., Chen, B., Robinson, H., Grishin, N. V., Osterman, A. L. & Zhang, H. (2002) Crystal structure of Haemophilus influenzae NadR protein. A bifunctional enzyme endowed with NMN adenyltransferase and ribosylnicotinimide kinase activities, *J Biol Chem.* **277**, 33291-9.

35. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. & Schioth, H. B. (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints, *Mol Pharmacol.* **63**, 1256-72.

36. Lagerstrom, M. C. & Schioth, H. B. (2008) Structural diversity of G protein-coupled receptors and significance for drug discovery, *Nat Rev Drug Discov.* **7**, 339-57.

37. Civelli, O., Reinscheid, R. K., Zhang, Y., Wang, Z., Fredriksson, R. & Schioth, H. B. (2013) G protein-coupled receptor deorphanizations, *Annu Rev Pharmacol Toxicol.* **53**, 127-46.

38. Krishnan, A., Almen, M. S., Fredriksson, R. & Schioth, H. B. (2012) The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi, *PLoS One.* **7**, e29817.

39. Nordstrom, K. J., Sallman Almen, M., Edstam, M. M., Fredriksson, R. & Schioth, H. B. (2011) Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families, *Mol Biol Evol.* **28**, 2471-80.

40. Schioth, H. B. & Fredriksson, R. (2005) The GRAFS classification system of G-protein coupled receptors in comparative perspective, *Gen Comp Endocrinol.* **142**, 94-101.

41. Almen, M. S., Nordstrom, K. J., Fredriksson, R. & Schioth, H. B. (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin, *BMC Biol.* **7**, 50.

42. Katritch, V., Cherezov, V. & Stevens, R. C. (2013) Structure-function of the G protein-coupled receptor superfamily, *Annu Rev Pharmacol Toxicol.* **53**, 531-56.

43. Katritch, V., Cherezov, V. & Stevens, R. C. (2012) Diversity and modularity of G protein-coupled receptor structures, *Trends Pharmacol Sci.* **33**, 17-27.

44. Tang, X. L., Wang, Y., Li, D. L., Luo, J. & Liu, M. Y. (2012) Orphan G protein-coupled receptors (GPCRs): biological functions and potential drug targets, *Acta Pharmacol Sin.* **33**, 363-71.

45. Sato, K., Pellegrino, M., Nakagawa, T., Vosshall, L. B. & Touhara, K. (2008) Insect olfactory receptors are heteromeric ligand-gated ion channels, *Nature.* **452**, 1002-6.

46. Smart, R., Kiely, A., Beale, M., Vargas, E., Carraher, C., Kralicek, A. V., Christie, D. L., Chen, C., Newcomb, R. D. & Warr, C. G. (2008) Drosophila odorant receptors are novel seven transmembrane domain proteins that can signal independently of heterotrimeric G proteins, *Insect Biochem Mol Biol.* **38**, 770-80.

47. Touhara, K. (2009) Insect olfactory receptor complex functions as a ligand-gated ionotropic channel, *Ann N Y Acad Sci.* **1170**, 177-80.

48. Park, J. H. & Saier, M. H., Jr. (1996) Phylogenetic, structural and functional characteristics of the Na-K-Cl cotransporter family, *J Membr Biol.* **149**, 161-8.

49. Kuan, J. & Saier, M. H., Jr. (1993) The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships, *Crit Rev Biochem Mol Biol.* **28**, 209-33.

50. Kunji, E. R. & Robinson, A. J. (2010) Coupling of proton and substrate translocation in the transport cycle of mitochondrial carriers, *Curr Opin Struct Biol.* **20**, 440-7.

51. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007) Clustal W and Clustal X version 2.0, *Bioinformatics.* **23**, 2947-8.

52. Zhai, Y., Tchieu, J. & Saier, M. H., Jr. (2002) A web-based Tree View (TV) program for the visualization of phylogenetic trees, *J Mol Microbiol Biotechnol.* **4**, 69-70.

53. Park, H. H. (2011) Structural analyses of death domains and their interactions, *Apoptosis.* **16**, 209-20.

54. de la Horra, C., Hernando, N., Lambert, G., Forster, I., Biber, J. & Murer, H. (2000) Molecular determinants of pH sensitivity of the type IIa Na/P(i) cotransporter, *J Biol Chem.* **275**, 6284-7.

55. Anantharaman, V. & Aravind, L. (2003) Application of comparative genomics in the identification and analysis of novel families of membrane-associated receptors in bacteria, *BMC Genomics.* **4**, 34.

56. Soding, J. (2005) Protein homology detection by HMM-HMM comparison, *Bioinformatics.* **21**, 951-60.

57. Murakami, M. & Kouyama, T. (2008) Crystal structure of squid rhodopsin, *Nature.* **453**, 363-7.

58. Vinothkumar, K. R. & Henderson, R. (2010) Structures of membrane proteins, *Q Rev Biophys.* **43**, 65-158.

59. Venkatakrishnan, A. J., Deupi, X., Lebon, G., Tate, C. G., Schertler, G. F. & Babu, M. M. (2013) Molecular signatures of G-protein-coupled receptors, *Nature.* **494**, 185-94.

60. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M. & Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor, *Science.* **289**, 739-45.

61. Pao, S. S., Paulsen, I. T. & Saier, M. H., Jr. (1998) Major facilitator superfamily, *Microbiol Mol Biol Rev.* **62**, 1-34.

62. Prakash, S., Cooper, G., Singhi, S. & Saier, M. H., Jr. (2003) The ion transporter superfamily, *Biochim Biophys Acta.* **1618**, 79-92.

63. Castillo, R. & Saier, M. H. (2010) Functional Promiscuity of Homologues of the Bacterial ArsA ATPases, *Int J Microbiol.* **2010**, 187373.

64. 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, *Eur J Biochem.* **270**, 799-813.

65. Saier, M. H., Hvorup, R. N. & Barabote, R. D. (2005) Evolution of the bacterial phosphotransferase system: from carriers and enzymes to group translocators, *Biochem Soc Trans.* **33**, 220-4.

66. Rodionov, D. A., Hebbeln, P., Eudes, A., ter Beek, J., Rodionova, I. A., Erkens, G. B., Slotboom, D. J., Gelfand, M. S., Osterman, A. L., Hanson, A. D. & Eitinger, T. (2009) A novel class of modular transporters for vitamins in prokaryotes, *J Bacteriol.* **191**, 42-51.

67. Hebbeln, P., Rodionov, D. A., Alfandega, A. & Eitinger, T. (2007) Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module, *Proc Natl Acad Sci U S A.* **104**, 2909-14.

68. Josefsson, L. G. (1999) Evidence for kinship between diverse G-protein coupled receptors, *Gene.* **239**, 333-40