the FEBS Journal

The transporter–opsin–G protein-coupled receptor (TOG) superfamily

Daniel C. Yee*, Maksim A. Shlykov*,†, Åke Västermark, Vamsee S. Reddy‡, Sumit Arora, Eric I. Sun[1] and Milton H. Saier Jr[1]

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

Keywords
channels; receptors; rhodopsin; secondary carriers; transport proteins

Correspondence
M. H. Saier Jr, Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116, USA
Fax: +1 858 534 7108
Tel: +1 858 534 4084
E-mail: msaier@ucsd.edu

Present address
†University of Michigan Medical School, Ann Arbor, MI, USA

‡University of Calgary, Calgary, Alberta, Canada

*These authors contributed equally to this work

(Received 27 April 2013, revised 6 August 2013, accepted 6 August 2013)

doi:10.1111/febs.12499

Visual rhodopsins 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 a previous paper [Shlykov MA, Zheng WH, Chen JS & Saier MH Jr (2012) *Biochim Biophys Acta* **1818**, 703–717], we characterized the 4-toluene sulfonate uptake permease (TSUP) family of transmembrane proteins, and showed that these 7-transmembrane segment (TMS) or 8-TMS proteins arose by intragenic duplication of a gene encoding a 4-TMS protein, sometimes followed by loss of a terminal TMS. In this study, we show that the TSUP, GPCR and microbial rhodopsin 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, the members of most constituent families exhibit 7-TMS topologies that are well conserved, and these arose by loss of either the N-terminal TMS (more frequent) or the C-terminal TMS (less frequent), depending on the family. Phylogenetic analyses revealed familial relationships within the superfamily and protein relationships within each of the nine families. The results of the statistical analyses leading to the conclusion of homology were confirmed using hidden Markov models, Pfam and 3D superimpositions. 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 evolution of this superfamily.

Introduction

Using functional and phylogenetic information derived from over 10 000 publications on transport systems, members of our laboratory have been able to classify virtually all recognized transport proteins into over 700 families [1,2]. The resulting system of classification is summarized in the IUBMB-approved Transporter Classification (TC) Database (TCDB; http://www.tcdb.org) [3,4]. Our current efforts focus on identification of distant relationships, allowing placement of these families into superfamilies. As 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, i.e. a group of proteins showing a

Abbreviations
GPCR, G protein-coupled receptor; HMM, hidden Markov model; HORC, heteromeric odorant receptor channel; LCT, lysosomal cystine transporters; MR, microbial rhodopsin; NiCoT, Ni^{2+}–Co^{2+} transporters; OST, organic solute transporters; PNaS, phosphate:Na^{+} symporters; PnuC, nicotinamide ribonucleoside uptake permeases; RMSD, root mean square deviation; TCDB, Transporter Classification Database; TOG, transporter/opsin/G protein-coupled; TSUP, 4-toluene sulfonate uptake permease; VR, visual rhodopsin.

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 proteins included in the TOG superfamily derive from a common ancestor via similar pathways. It may therefore be anticipated that the structures of most of these proteins exhibit common features [7,8]. As rhodopsins are the transmembrane proteins with the highest-resolution X-ray structures solved to date [9–11], we are able 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 may vary to accommodate a wide diversity of functions, and may serve as a guide in future studies revealing how sequence divergence may lead to alterations in the scaffold.

Results

All protein families within the TCDB belonging to subclass 2.A comprise electrochemical potential-driven uniporters, symporters and antiporters. Based on preliminary evidence reported by Shlykov *et al.* [12], we used a modified SSearch program [13,14] to compare TSUP homologs with all other secondary carriers, and identified potential superfamily relationships. Subsequently, these analyses were extended to other TC classes. Comparisons to the TC sub-classes 9.A (the GPCRs), 3.E (the MRs), and 4.B (PnuC) proved fruitful.

The MR and LCT families had previously been shown to be related [15]. Analyses involving the MR, LCT, PnuC, PNaS, Sweet and GPCR families provided sufficient evidence to include them in the TOG superfamily. Our results led to formulation of a novel TOG superfamily for which trees were generated using the ClustalX (http://www.clustal.org/) and Superfamily Tree 1/2 (SFT1/SFT2) programs [16–18].

The TOG superfamily consists of nine, possibly ten, currently recognized protein families, with members primarily having six to nine putative transmembrane segments (TMSs) (Table 1). A summary of the comparisons performed is presented in Fig. 1A and Table 2, and a proposed evolutionary pathway for the appearance of various members of the TOG superfamily is presented in Fig. 1B. The TSUP family has been characterized previously [12], and the AveHAS plots, phylogenetic trees and TSUP homologs are described elsewhere [12].

The lysosomal cystine transporter (LCT) and ion-translocating microbial rhodopsin (MR) families (TC# 2.A.43 and 3.E.1, respectively)

The evolutionary pathway of the 7-TMS LCT family has been elucidated [15], and LCT family members were found to be homologous to members of the MR family, including putative fungal chaperone proteins present in the MR family (see Table 1 and TCDB 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 and are generally larger than MR proteins, which have approximately 220–300 residues. Eukaryotic homologs within a single transporter family tend to be approximately 40% larger than their bacterial homologs [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 (Table 1, Tables S1 and S2, and Figs S1A,B and S2A,B).

TMS1–3 in LCT family members duplicated to give rise to TMS5–7, with TMS4 showing insignificant sequence similarity to any of the other six TMSs [15]. The precursor may have been an 8-TMS protein that generated the present-day 7-TMS proteins by loss of TMS1 or TMS8, and strong evidence for this possibility is presented here and previously [12].

The 7-TMS Bba2 protein of the LCT family is homologous to the 7-TMS Aae2 protein of the Sweet family. Alignment of these two proteins using GSAT (http://tcdb.org/progs/?tool=gsat) yielded a comparison score of 12.4 SD (Fig. S1C). These comparisons

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 are shown. Minor topologies, represented in a minority of family members, are listed in parentheses in column 7. Putative topological transitions are described in footnotes a–d.

Family number	Family name	Family abbreviation	TC#	Number of 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 TMSs arose from 8 TMSs by loss of the N-terminal TMS	Archaea, Bacteria, Eukaryota	Bac_rhodopsin
2	Sweet[a]	Sweet	9.A.58	3 or 7	200–290	3 + 1 + 3 (3)	7 TMSs arose from 8 TMSs 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 TMSs arose from 8 TMSs by loss of the N-terminal TMS	Bacteria, Eukaryota	NMN_transporter, AAA_28
4	4-toluene sulfonate uptake permease[b]	TSUP	2.A.102	7–9	250–600	4 + 4 (3 + 1 + 3; 1 + 4 + 4)	8 TMSs arose from internal duplication of 4 TMSs	Archaea, Bacteria, Eukaryota	TauE
5	Ni^{2+}–Co^{2+} transporter[c]	NiCoT	2.A.52	6–8	300–380	4 + 4 (3 + 1 + 3; 2 + 1 + 1 + 2)	8 TMSs 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 TMSs arose from 8 TMSs by loss of the N-terminal TMS	Eukaryota	Solute_trans_a
7	Phosphate:Na^+ symporter[d]	PNaS	2.A.58	8 or 9	500–700	4 + 4 (4 + 4 + 1)	8 TMSs 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 TMSs arose from 8 TMSs 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 TMSs arose from 8 TMSs by loss of the N-terminal TMS	Eukaryota	7tm_1, 7tm_2, 7_tm3

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

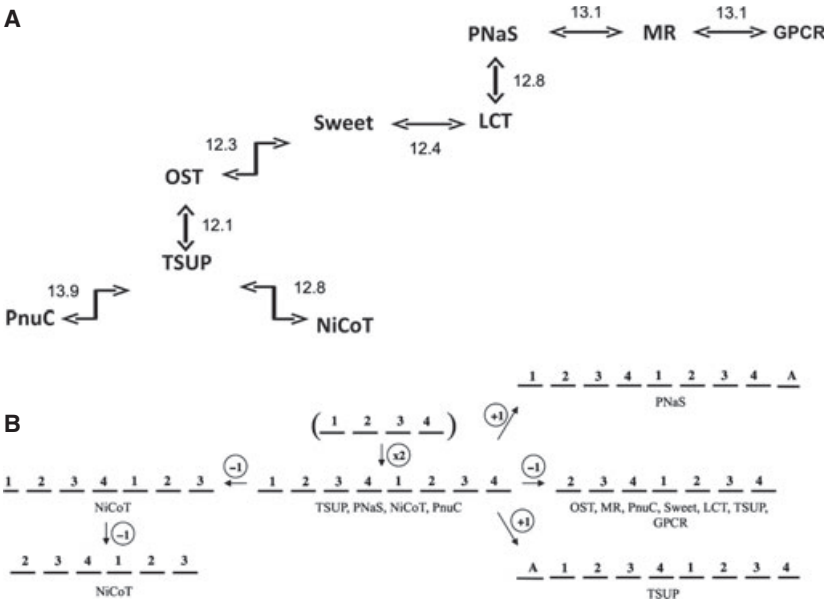

Fig. 1. (A) TOG superfamily homology established through use of GSAT/GAP (http://saier-144-21.ucsd.edu/) and the superfamily principle. TOG superfamily proteins from the TCDB and their homologs were used to establish homology between all members of the nine families. The GSAT/GAP comparison scores are expressed in terms of standard deviations (SD). (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 loss or gain of specific TMSs.

establish homology between the LCT and Sweet families. TMS3–7 of Aae2 aligned with TMS3–7 of Bba2, demonstrating that the two families both arose via the same evolutionary pathway (Fig. S1C). Loss of TMS1 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 TMS2–4 of LCT Ago1 (seven putative TMSs) with TMS6–8 of PNaS Cre1 (11 putative TMSs) yielded a comparison score of 12.8 SD (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 TMS2–4 of Ago1 and TMS6–8 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 probably arose from fusion of a 7-TMS protein with another 4-TMS repeat unit.

The Ni^{2+}–Co^{2+} transporter (NiCoT) family (TC# 2.A.52)

Members of the ubiquitous NiCoT family are typically 300–380 amino acyl residues in size and possess 6–8 putative TMSs [20] (Table S3 and Fig. S3A,B). The NicO family (TC# 2.A.113) includes distant homologs of great size, sequence and topological variation. NiCoT transporters catalyze the uptake of Ni^{2+} and Co^{2+} using a proton motive force-dependent mechanism; however, NicO family members catalyze Ni^{2+} and Co^{2+} export to the external environment [21,22]. NicO family members exhibit 3–8 putative TMSs, but the 6 TMS topology is most common.

Comparing TMS1–3 of TSUP Pla1 (eight putative TMSs) with TMS4–6 of NiCoT Bja1 (six putative TMSs) yielded a comparison score of 12.8 SD (Fig. S3C). This comparison establishes homology between members of these two families, and serves to confirm our proposed evolutionary pathway for appearance of the NiCoT family as a member of the TOG superfamily (Fig. 1A,B). Based on alignments, it is likely that the 6-TMS NiCoT proteins arose by the loss of both TMS1 and TMS8 after the 4-TMS intragenic duplication event took place.

The organic solute transporter (OST) family (TC# 2.A.82)

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 homologs of the α-subunits in plants, fungi and bacteria were retrieved in NCBI searches, but their scores usually bordered in or fell below our threshold cut-off for establishing homology. Furthermore, each well-characterized transporter within this family functions as part of a two-component system utilizing an α-subunit (280–400 amino acyl residues) and β-subunit (180–290 amino acyl residues). The α-subunits generally contain seven TMSs, whereas the β-subunits

Table 2. Highest comparison scores between TOG superfamily members. Protein representatives of families used in these comparisons are listed in Tables S1–S9. Representative alignments for highlighted squares are usually shown 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 SD, which are considered sufficient to establish homology and inter-connect families, are shaded. These values are sufficient to establish homology based on the criteria discussed in Experimental procedures. The mean comparison score and mean number of TMSs in all TOG superfamily alignments are 11.5 SD and 2.3 TMSs. The mean comparison score and mean number of TMSs in alignments used to establish homology and inter-connect all families within the TOG superfamily are 12.8 SD and 2.5 TMSs.

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

contain only one. To date, neither subunit has been found to function without the other (Table S4 and Fig. S4A,B) [23].

Comparing TMS2–3 of TSUP Tsp1 (eight putative TMSs) with TMS1–2 of OST Cre2 (seven putative TMSs) yielded a comparison score of 12.1 SD (Fig. S4C). This comparison demonstrates the loss of TMS1 in OST transporters, and establishes homology between the two families: loss of TMS1 from an 8-TMS precursor generated the 7-TMS topology of the OST family. Another alignment between TMS2–4 of TSUP Gfo1 (seven putative TMSs) and TMS2–4 of OST Dre1 (seven TMSs) also supports homology between the TSUP and OST families and the proposed evolutionary pathway. This comparison yielded a score of 11.3 SD (Fig. S4D).

The Sweet family (TC# 9.A.58)

Eukaryotic Sweet family (Sweet is a collective term for PQ-loop, Saliva, MtN3) channels or carriers catalyze facilitated diffusion (uptake or efflux) of sugars across the ER and plasma membranes of plants and animals [26]. Bacterial pathogens up-regulate specific plant Sweet transporters, allowing them to utilize the sugar efflux function of these proteins to meet their energy needs [27]. Eukaryotic homologs possess seven TMSs in a 3 + 1 + 3 repeat arrangement, and are 200–290 amino acyl residues in size (Table S5 and Fig. S5A,B). Although 7-TMS bacterial homologs exist, most bacterial putative Sweet channels possess three TMSs and are approximately 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 four TMSs per polypeptide chain have been reported, suggesting that the 3-TMS proteins may function as oligomeric channels [28].

Comparing TMS6–7 of Sweet Rco4 (seven putative TMSs) with TMS6–7 of OST Ath8 (seven putative TMSs) yielded a comparison score of 12.3 SD (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 TMS4–6 of Sweet Asu3 (seven putative TMSs) and TMS4–6 of OST Ncr1 (seven putative TMSs) yielded a comparison score of 10.3 SD (Fig. S5D), further confirming the homology between the Sweet and OST families.

The phosphate:Na^+ symporter (PNaS) family (TC# 2.A.58)

Both bacterial and eukaryotic PNaS homologs usually range in size between 500 and 700 residues, but the bacterial homologs may be as small as 350 residues. Most members of this family possess eight or nine TMSs in a 4 + 4 or 4 + 4 + 1 arrangement, as shown previously [28] (Table S6 and Fig. S6A,B). However, some proteins such as NPT2 of *Rattus norvegicus* have as many as 12 TMSs, with the extra ones appearing at the N-terminus [29]. Mammalian PNaS porters may catalyze the electroneutral co-transport of three Na^+ with inorganic phosphate (P_i). Their activities are regulated by parathyroid hormone and dietary P_i.

Comparing TMS4–6 of PNaS Odi8 (eight putative TMSs) with TMS3–5 of MR Hwa1 (seven putative TMSs) yielded a comparison score of 13.1 SD (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.

The nicotinamide ribonucleotide uptake permease (PnuC) family (TC# 4.B.1)

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

Comparing TMS2–3 of PnuC Spr1 (seven putative TMSs) with TMS3–4 of TSUP Cba4 (eight putative TMSs) yielded a comparison score of 12.4 SD (Fig. S7C). This comparison demonstrates homology between the PnuC and TSUP families. An alignment between TMS3–6 of PnuC Sde2 (seven putative TMSs) and TMS4–6 of TSUP Ere1 (eight putative TMSs) provides additional evidence of homology and supports the PnuC evolutionary pathway (Fig. S8D). Our results, and placement of the PnuC family into the TOG superfamily, support the proposal that a 4-TMS precursor duplicated to give 8-TMS proteins, and that the N-terminal TMS was then deleted.

The G protein-coupled receptor (GPCR) family (TC# 9.A.14)

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, the 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,E), 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 Table S8A,B and the TCDB), they are among the most targeted proteins for drugs, and their analysis has tremendous implications for future pharmacological developments [44].

Comparing TMS5–6 of the GPCR Dre1 (seven putative TMSs) with TMS5–6 of Mos1 (seven putative TMSs) of the MR family yielded a comparison score of 13.1 SD (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, probably arose from loss of the N-terminal TMS from the proposed 8-TMS predecessor. TMS1–4 of GPCR Dre1 (seven putative TMSs) also aligned with TMS1–4 of MR Cga1 (seven putative TMSs) and yielded a comparison score of 10.6 SD, further supporting homology (Fig. S7D) between GPCRs and MRs.

The heteromeric odorant receptor channel (HORC) family (TC# 1.A.69)

Olfactory sensory neurons in insects express between one and three members of the channel-forming olfactory receptor gene family, as well as the highly conserved Or83b co-receptor (TC# 1.A.69.1.1). Each functional odorant receptor consists of a heteromeric complex comprising at least one odorant-binding subunit and the aforementioned Or83b co-receptor [45]. Immunocytochemical experiments showed that insect odorant receptors possess a 7-TMS topology, but, in contrast to members of the GPCR family, have a

cytoplasmic N-terminus and an extracellular C-terminus. Several authors [45–47] suggested that heteromeric insect olfactory receptors comprise a new class of ligand-activated non-selective cation channels. We obtained preliminary evidence that insect olfactory receptors and GPCRs are homologous. However, based on our criteria, we could not establish homology because comparison scores were insufficient (10.3 SD). Nevertheless, the intriguing possibility of homology will provide the basis for future investigations.

Controls: the major intrinsic protein (MIP) family (TC# 1.A.8) and the mitochondrial carrier (MC) family (TC# 2.A.29)

Members of the major intrinsic protein (MIP) family are channel proteins that function in transport of water, small carbohydrates, urea, NH_3, CO_2, H_2O_2 and ions 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 involved in transporting keto acids, amino acids, nucleotides, inorganic ions and co-factors across the mitochondrial inner membrane. Proteins of 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 and 10.5 SD, respectively (Table S9). Comparisons of the MC family against the NiCoT and PNaS families yielded a maximal comparison score of 9.5 SD. Comparison of the MIP family against the PNaS family yielded a maximal comparison score of 10.5 SD. The mean score for the best comparisons between TOG superfamily members and the MC family was 8.8 SD, and the mean score for comparisons between TOG superfamily members and the MIP family was 8.9 SD. When compared to each other, the MIP and MC families yielded maximal comparison scores of 9.2 SD. By contrast, the mean score for the best comparisons for the nine TOG superfamily families with each other was 11.5 SD, and the mean score was 12.6 SD between families used to establish homology. Based on these results, we suggest that 12.0 SD (Fig. S9), combined with correct alignment of at least two 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 cut-off 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 edge displaying the highest similarity between MIP and DUF3816), considerably worse than any of the similarities 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.

Integration of topological data

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 (http://saier-144-21.ucsd.edu/baravehas.html) [52] were generated for all families except the GPCR family (Fig. 2), as well as one for all families (Fig. S10). The large GPCR homologs rendered the AveHAS plot too large for easy viewing, but this plot is presented in the supplementary Fig. S10.

The plot reveals seven 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. TMS1 in the 8-TMS homologs is conserved in only a few of the family members. Other less conserved peaks of hydrophobicity are found N- and C-terminally to the seven well-conserved peaks. A closer look revealed that these peaks are primarily due to the larger PNaS homologs. 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 conserved domain database (http://www.ncbi.nlm.nih.gov/cdd/) search identified a member of the death domain superfamily constituting approximately the first 100 residues of the Sko2 N-terminus; death domain proteins 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].

Phylogenetic analyses of the TOG superfamily

Proteins found in the TCDB, representing the various sub-families within each family of the TOG superfamily

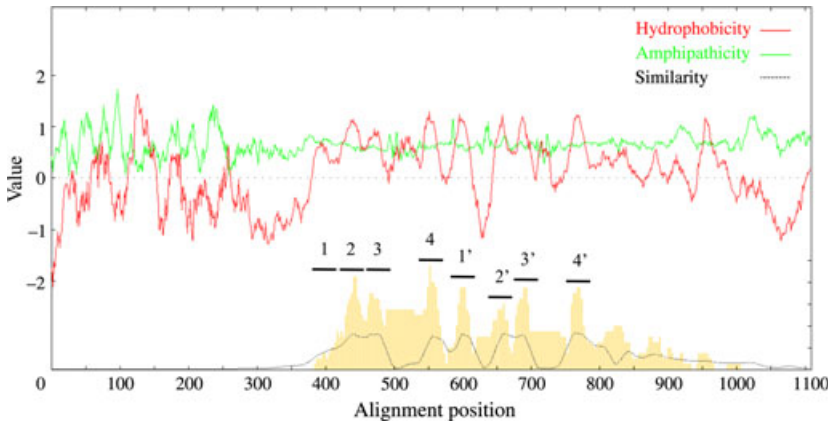

Fig. 2. Average hydropathy, amphipathicity and similarity (AveHAS) plots based on a ClustalX multiple alignment. All members of all TOG superfamily families from the TCDB were included except for the GPCR family (see Fig. S10). The plot reveals eight well-conserved averaged TMSs. However, as many as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

(except the GPCR family), 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) [16–18]. In Fig. S11, the ClustalX/TreeView program revealed the GPCR family (TC# 9.A.14) in five distinct clusters, the TSUP family (TC# 2.A.102) 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 (Fig. 3A), which shows clustering of nearly all protein members coherently according to their respective families, with the exception of the GPCR and NiCoT families, which are found in two closely related clusters. All members of the NiCoT family (TC# 2.A.52.1) form one cluster, and all members within the distantly related NicO family (TC# 2.A.113) form the other. These results reveal the superiority of the SFT1 program over the ClustalX program, an observation that has been noted for many sequence-divergent superfamilies for which multiple alignments are not reliable [16–18]. The SFT2 tree (Fig. 3B) shows the phylogenetic relationships between all nine families within the TOG superfamily. Interestingly, the families that have lost TMS1 cluster together at the bottom of the tree, suggesting, but not proving, that this event may have occurred before these families diverged from each other.

Analyses of internal repeats

Shlykov *et al.* [12] previously reported internal repeats within TSUP family members that corresponded to a 4-TMS α-helical structural precursor [12], and Zhai *et al.* [15] demonstrated that TMS1–3 are homologous to TMS5–7 in the 7-TMS MR proteins. More recently, it has been demonstrated that TMS1–3 are homologous to TMS5–7 in members of the PNaS family (14.6 SD) (E.I.S. and M.H.S., unpublished results). Using the AR and GSAT programs [14], comparing TMS1–4 with TMS5–8 of the 8-TMS TSUP Pas1 protein yielded a comparison score of 15.2 SD (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 [69]. 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).

Non-TOG superfamily proteins previously reported to be related

A recent study [55] described two new Pfam families, one of which (7TMR_DISM) was claimed to be a bacterial family with a domain organization related to mammalian glutamate GPCRs. The other family (7TMR_HD) was reported to be peripherally related to 2.A.102.1.1 of the TSUP family [using hidden Markov models (HMMs) in Pfam]. However, it appears that the comparisons were not performed at the sequence level, and that other 7-TMS transporter families were not included in the screen. Using the TCDB and PSI-BLAST as well as Protocol1 and Protocol2 searches [55a], we could not obtain convincing evidence that membrane domains of these sequences are related to the proteins in the TOG superfamily.

Mapping of TC# 9.A.14 and the GRAFS system

HMMER 3.0 (http://hmmer.janelia.org/) was used to map the GPCR family (TC# 9.A.14) and GRAFS (glutamate, rhodopsin, adhesion, frizzled/taste2, and

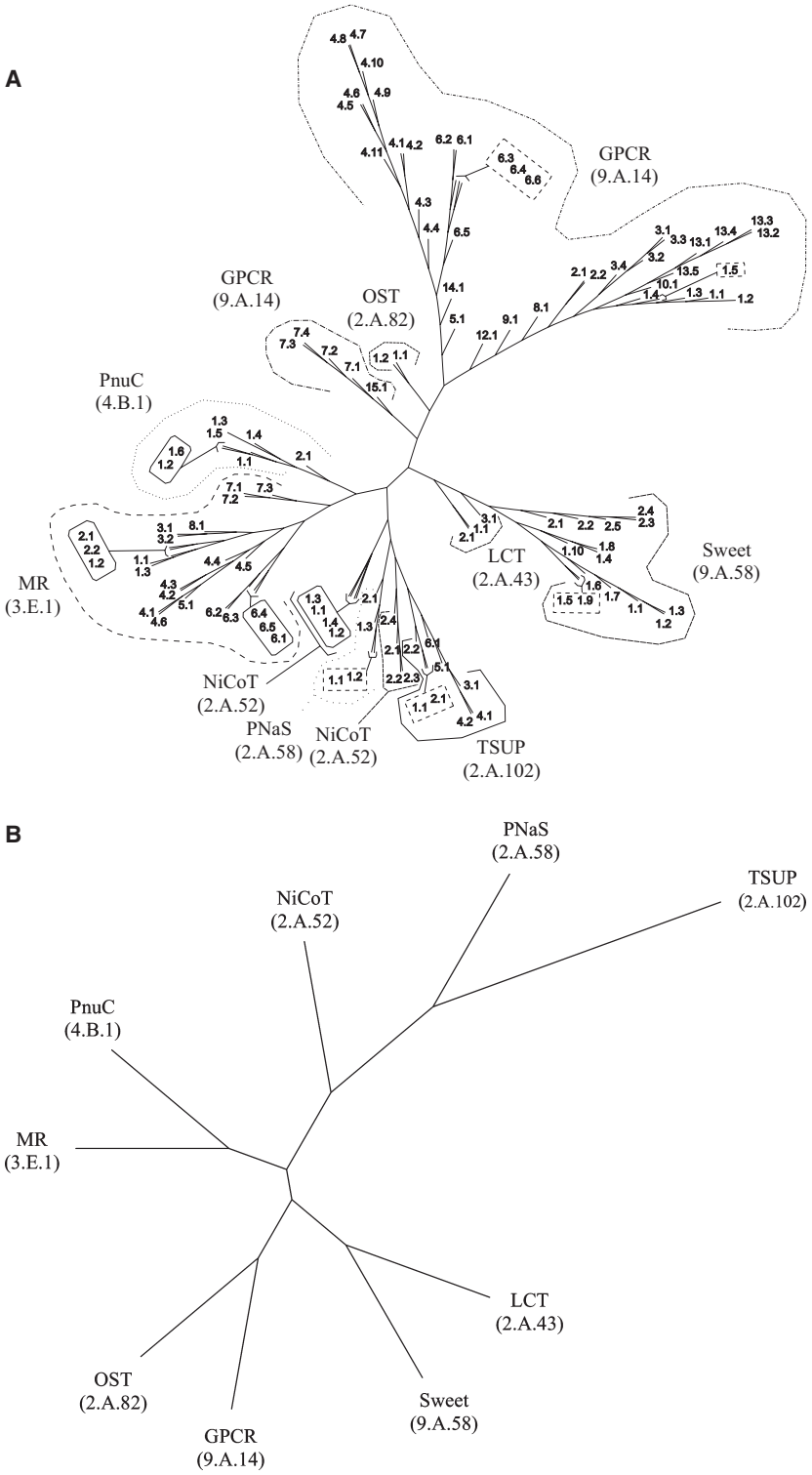

Fig. 3. Phylogenetic (Fitch) trees for the TOG superfamily in the TCDB as of May 2013. Three methods of tree construction were used. (A) The BLAST-derived SFT1 program shows the proteins of families within the TOG superfamily. (B) The SFT2-based tree shows the relationships of the TOG superfamily families to each other. In (A), numbers adjacent to the branches indicate the protein TC# (last two digits of the complete protein TC#), while the family designations and family TC numbers are shown in parentheses. In (B), family abbreviations are shown, with TC family numbers in parentheses (see the TCDB for protein identification). The results for the third method (ClustalX/Tree View) are shown in Fig. S11.

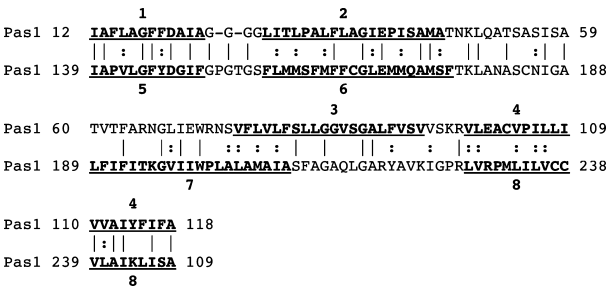

Fig. 4. Demonstration of a 4-TMS repeat unit in the Pas1 protein of the TSUP family. GSAT alignment of TMS1–4 of Pas1 (*Photorhabdus asymbiotica*; gi# 211638062; eight TMSs) with TMS5–8 of the same protein. A comparison score of 15.2 SD was obtained, with 50.5% similarity and 30.3% identity. Identical residues are indicated by vertical lines, close similarities are indicated by colons. GSAT was used at default settings, with a gap creation penalty of 8 and a gap extension penalty of 2, with 500 random shuffles.

secretin) in Cytoscape 2.8.3 (http://www.cytoscape.org/), and a spring-embedded layout was applied. A clear clustering pattern was evident, with the secretin and adhesion receptor families forming a cluster on one side of the rhodopsin cluster. Frizzled/Taste2 had two edges with two sequences, also in the secretin/adhesion cluster.

The glutamate receptors formed a small cluster. Two sequences within the secretin/adhesion receptor family linked these receptors to the rhodopsin cluster (which includes somatostatin receptors, opioid receptors, galanin and the GPR54 binding receptors (somatostatin, opioid, and galanin) and opsin). Five sequences among the glutamate receptors connected this cluster with the rhodopsin cluster. All 12 types of rhodopsins showed good representation with similarity to sequences in the TCDB. This was clearly the largest and most compact cluster. Twenty-six sequences included in TC# 9.A.14 are rhodopsin GPCRs and 28 sequences in TC# 9.A.14 are non-rhodopsin members.

TCDB and Pfam family correspondence

Specific Pfam families corresponded to our TCDB families: TSUP (TC# 2.A.102) equivalent to PF01925 (TauE); LCT (TC# 2.A.43) equivalent to PF04193 (PQ-loop); NiCoT (TC# 2.A.52) equivalent to PF03824 (NicO); PNaS (TC# 2.A.58) equivalent to PF02690 (Na_Pi_cotrans); OST (TC# 2.A.82) equivalent to PF03619 (Solute_trans_a); MR (TC# 3.E.1) equivalent to PF01036 (Bac_rhodopsin); PnuC (TC# 4.B.1) equivalent to PF13521 (AAA_28) and PF04973 (NMN_transporter), and Sweet (TC# 9.A.58) equivalent to PF03083 (MtN3_slv). We also checked the following clans (Pfam 'clans' are superfamilies of similar Pfam families) to determine whether obvious relationships exist between them: TauE and NicO are in the same clan, PQ_LOOP and MtN3_slv are in the same clan, Na_Pi_cotrans is not a member of a clan, Solute_trans_a is not a member of a clan, Bac_rhodopsin is in a large clan called GPCR_A, containing 7TM-7TMR_HD and many GPCRs including the 7tm_1 family, a central node of rhodopsin GPCRs, AAA_28 is in a large clan called P-loop_NTPase, and NMN_transporter is not a member of a clan. Thus, Pfam analyses yielded confirmatory evidence for relatedness among several of the TOG superfamily families.

Statistics of the TC# 9.A.14 network compared with the entire TOG–Pfam network

We used Network Analyzer (a function in cytoscape), treating the network as undirected, to compare TC# 9.A.14 with the entire network. The number of connected components was 4; the network diameter was 16; the network radius was 1; the network centralization was 0.060 (0.175 for TC# 9.A.14); the shortest path was 168 890 (87%); the characteristic path length was 6.229; the mean number of neighbors was 2.832 (5.074 for TC# 9.A.14); the network density was 0.006 (0.096 for TC# 9.A.14), and the network heterogeneity was 1.210 (0.553 for TC# 9.A.14). This shows that the network of TC# 9.A.14 has higher density and lower heterogeneity than the entire TOG–Pfam network.

Location of the GPCRs within the TOG–Pfam network

For the small non-rhodopsin component of TC# 9.A.14, the only edges connecting it with the others were from PgaD and DUF4131. From PgaD, there were onward links to members of the LCT family (TC# 2.A.43.4.1). These were the glutamate GPCRs, some of which display limited similarity to the SOG group of rhodopsin GPCRs, as shown by our studies. For the large non-rhodopsin component of GPCRs (TC# 9.A.14), the number of connections is greater to other families, including the rhodopsin component of the GPCRs (TC# 9.A.14). For example, a direct link between TC# 9.A.14.14.1 (non-rhodopsin GPCR) was shown for 7tm_1, which is the Pfam family that links many rhodopsin GPCRs. The same was true for TC# 9.A.14.6.6 and TC# 2.A.43.2.5 (a PQ-loop repeat-containing protein from *A. thaliana* of the LCT family), which had direct links to Pfam family 7tm_1 (the central node connecting rhodopsin GPCRs in Fig. 5). The similarity was embedded in

motifs, including LxLxV and KxLLxxVxVF. Even the large non-rhodopsin component of the GPCRs (TC# 9.A.14) did not show strong direct links to other TCDB families that are members of the TOG superfamily, showing that the Pfam approach to homology searching is less sensitive than the superfamily principle approach described here. The link from Pfam family 7tm_1 to TC# 2.A.43.2.5 (a plant member of the LCT family) is weaker (0.00021) than the link between the GPCR protein TC# 9.A.14.14.1 and 7tm_1 ($3.2e^{-07}$), as expected, but the former value is still highly significant. The link to the LCT protein TC# 2.A.43.2.5 was much stronger than any link from other GPCRs to other TC families that are members of the TOG superfamily. This means that the closest neighbor to GPCRs in TC# 9.A.14 was connected via the rhodopsin component Pfam family (7tm_1), and that the GPCR family (TC# 9.A.14) itself has a non-rhodopsin component (glutamate GPCRs) that is disconnected from the rest of the GPCRs.

Mapping of GPCRs in the TCDB did not contain disconnected components, and this appeared not to be the result of addition of missing sequences; it instead depended on a change in the database. Nevertheless, members of the glutamate GPCR cluster (e.g. TC# 9.A.14.7.3) were poorly connected to other GPCRs. A significant link between the glutamate cluster and the other GPCRs passed through one Pfam family, PF07077, with e-values of 0.016 and 0.048. Using the Pfam website, PF07077 (DUF1345) and TC# 9.A.14.11.2 (the closest node to the family) did not appear significantly related, even when the threshold was set to 10. The Pfam web service lists the relationship to 7tm_3 (PF00003), one of the core nodes of the glutamate GPCRs, as $3.3e^{-40}$, compared with $1.2e^{-42}$ in our mapping. In summary, these results confirm the conclusion that the GPCR cluster is not held together by high similarity edges, especially glutamate GPCRs. In fact, these edges were weaker than those between rhodopsin GPCRs and the LCT and Sweet families.

HMM:HMM comparisons

The most significant result of our HMM:HMM comparison was 30.6% probability of homology, between the opsin cluster of the human rhodopsin GPCRs (cluster α in the GRAFS classification system) and the MR family (TC# 3.E.1). This is a significant result, indicating homology [56]. In fact, we had 753 match columns, and the e-value was 0.00044. The hit was divided into two halves: the first half showing a 30.6%

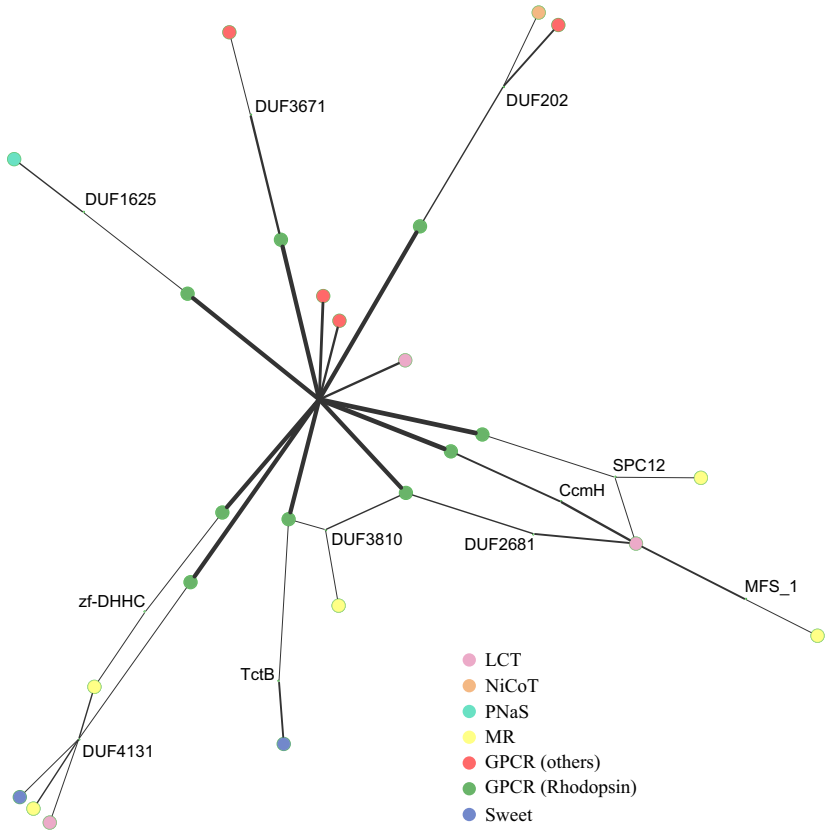

Fig. 5. TOG superfamily constituent interrelationships revealed using Pfam. Cytoscape 2.8.3 visualization (using spring-embedded logic) of our mapping of the proposed TOG superfamily to Pfam using HMMER3 and the default similarity cut-off (10). The Pfam nodes are shown in smaller size, and the edge width represents levels of similarity. The most significant link between any of the GPCR sequences and any of the non-GPCR sequences connects the central node of the rhodopsin GPCR cluster ('7tm_1') to an LCT sequence from *A. thaliana* (colored pink). The LCT sequences are similar to the Sweet sequences (colored violet) and overlap that cluster. The microbial rhodopsins (MR, colored yellow) show connectivity to the rhodopsin GPCR cluster.

probability over 104 columns, and the second showing 14.2% probability over 85 columns. The consensus alignment is shown at the bottom of Table S10.

Using hmmsearch (searching profile(s) against a sequence database) in HMMER 3.0, we confirmed that bovine rhodopsin was a member of the opsin cluster; the score was $1.5e^{-136}$ against the opsin HMM, but only $4.8e^{-37}$ against the amine HMM. The term 'opsin' means rhodopsin without retinal, and both microbial and bovine rhodopsins link retinal to the corresponding lysine amino acyl residue in TMS7. These results may be compared with other significant results between TC families. For example, OST (TC# 2.A.82) and PNaS (TC# 2.A.58) scored a 31% probability of homology in HHsearch (hhsuite-2.0.16; ftp://toolkit.genzentrum.lmu.de/pub/HH-suite/), MR (TC# 3.E.1) and NiCoT (TC# 2.A.52) scored 12%, MR (TC# 3.E.1) and LCT (TC# 2.A.43) scored 26.7%, and Sweet (TC# 9.A.58) and LCT (TC# 2.A.43) scored 97%.

Structural superposition of visual rhodopsin and microbial rhodopsin

Differences in the signaling systems of visual (e.g. squid [57]) and microbial rhodopsins, and the conformational changes of retinal isomerization and helix movements of spectroscopically distinct intermediates have been described elsewhere [58,59]. A superimposition of the bovine visual rhodopsin (VR) structure (PDB ID 1F88, chain A [60]) and a microbial rhodopsin (MR) structure (PDB ID 3NS0) was performed using Chimera 1.7 (http://www.cgl.ucsf.edu/chimera/; Fig. 6 and Movie S1). The individual TM helices matched up in the superimposed configuration, although they were not perfectly aligned, often being tilted at somewhat different angles. Nevertheless, when we oriented the structures to view them through the pore and placed TMS1 at the top, we were able to count the seven TMSs clockwise from one side in a corresponding manner between the structures. It was clear that the N-termini, the C-termini and all the loops between the TMSs corresponded.

When the seven TMSs of the human adenosine receptor (PDB ID 2YDO) (3 Å resolution structure, without the non-membrane helix after position 298) were compared with the seven TMSs in the α-rhodopsin GPCR/MECA [the melanocortin receptors (MCRs), endothelial differentiation G-protein coupled receptors (EDGRs), cannabinoid receptors (CNRs), and adenosine binding receptors (ADORAs)] cluster, and these were compared with VR (PDB ID 1F88, chain A), we found that the adenosine receptor could be oriented so that TMS1 is on top, and the consecutive helices could then be counted clockwise. This superposition was better than that noted above with MR. Here, the root mean square deviation (RMSD) was 4.925 for all α-carbons. A similar result was obtained with the opioid (rhodopsin GPCR) receptor (PDB ID 4EA3, chain A), as the same clockwise arrangement of TMSs was observed, and the RMSD was 4.467 for all α-carbons. These results show that, while the clockwise arrangement of the seven TMSs is shared between VR and MR, the structural superimposition is better between adenosine, opioid and VR, as expected considering their relative phylogenetic distances.

To evaluate how the superimposition is influenced by the various states in which both MR and VR can exist, we compared them in their various states. PDB ID 1F88-A is the dark-adapted conformation of VR, similar to PDB ID 4A4M (constitutively active light-adapted VR). Previously, when we compared 1F88-A (VR) with 3NS0 (ground-state MR), the RMSD was 7.176 for all α-carbons. However, if we used PDB ID 1KGB (another ground state of MR), there was an improvement; the RMSD was 7.019 for all α-carbons. Using the K intermediate of MR (PDB ID 1IXF), the RMSD was 7.639 for all α-carbons, and the RMSD was 8.401 for all α-carbons for the L intermediate of MR (PDB ID 1UCQ). For the early M intermediate (PDB ID 1KG8), we observed an RMSD of 7.118 for all α-carbons, and the RMSD was 7.638 for all α-carbons for the actual M intermediate (PDB ID 1IW9). Hence, the early M intermediate of MR (PDB ID 1KG8) is notably more similar to the dark-inactivated conformation of VR (PDB ID 1F88-A) than the K, L or actual M intermediates. However, when we considered all of the atoms, the best RMSD between VR (PDB ID 1F88-A) and MR was with the ground state of MR, especially 1KGB. Such RMSD values only demonstrate a similar fold, not homology. However, taken together with the other statistical results reported in this study, these results provide confirmation of homology.

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 Fig. 1). No other transport protein superfamily in the 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

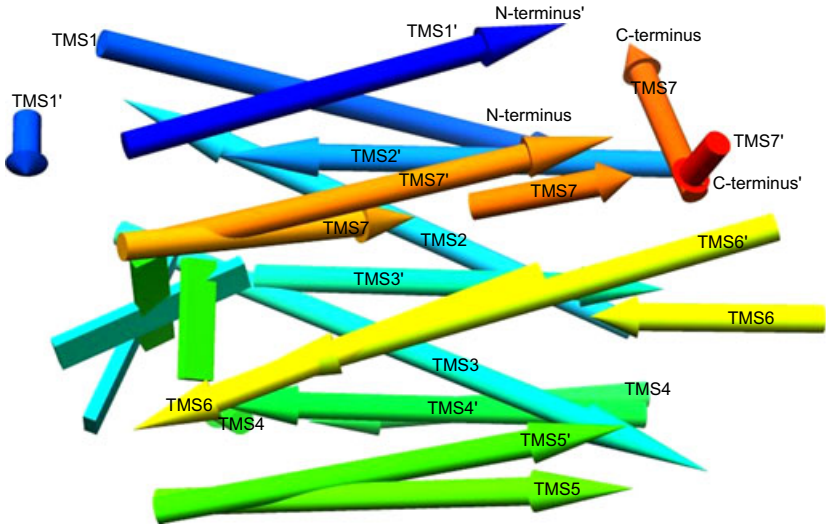

Fig. 6. Structural superimposition of a bovine VR structure (PDB ID 1F88, chain A) and a bacteriorhodopsin (MR) structure (PDB ID 3NS0), both containing seven TMSs, was performed using Chimera 1.7. The prime (′) notation indicates TMS numbering of the MR structure. The overall RMSD is 7.176. While the protein is shown as a set of helices, these may deviate from perfect helices. When such data are used as templates to switch to the 'pipes and planks' (α-helices and β-sheets) representation mode in Chimera, idealized 'pipes' (helices) are placed in a way that best represents the separate stretches of helix-annotated sequences. Several idealized helices are presented, each representing its own discontinuous helix segment. The result is that the idealized TMS arrow is placed in a way that shows a compromise for how that TMS has traversed the membrane. In the figure, all helices point towards the N-termini, except constituent helices of the C-terminal TMSs, which point towards the C-termini. The coloring is a scale starting from the N-termini (blue) and ending in the C-termini (red), going through an intermediate color scale through blue, light blue, turquoise, spring green, yellow and orange to red. The color scale is placed on the sequence considering the entire sequence including the non-TMS sequence, which is not shown in the figure.

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 to diverge into proteins with different modes of action. For example, no member of the major facilitator superfamily functions in transport by a mechanism other than secondary transport, and the only alternative function is that of transmembrane receptor [14,20,61]. Even this alternative functional type is exceedingly rare in nature.

A 2-TMS precursor may 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 with odd-numbered topologies, such as the PnuC family. Topological analyses of the entire superfamily reveal seven well-conserved 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, which is lost in many homologs, is present in enough members of the TOG superfamily to be visualized in the AveHAS plot (Fig. 2).

The greatest topological variation is observed for the Sweet, NiCoT and PNaS families. In the Sweet family, 3-TMS homologs are found in prokaryotes in addition to the 7-TMS proteins found ubiquitously. In the NiCoT family, some members appear to have only six TMSs, due to loss of both TMS1 and TMS8 of the 8-TMS precursor. In the PNaS family, the additional TMSs present in several family members proved to derive from fusions or late duplication events. Thus, the four extra TMSs are homologous to the last four 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 (Fig. 3A,B). 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 the TCDB. ArsB

transporter (TC# 3.A.4) family members of the ion transpoter (IT) superfamily [62], 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) [63]. Members of the phosphotransferase system (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; http://www.tcdb.org/search/result.php?tc=4.B.1) within the TOG superfamily function by group translocation using ATP-dependent nicotinamide ribonucleoside kinase as the energy-coupling enzyme. However, many members of this family are encoded by genomes that lack nicotinamide ribonucleoside kinase, supporting the conclusion that these porters function as secondary carriers [66]. Comparable studies have shown that members of the Na^+−transporting carboxylic acid decarboxylase (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 (cation:proton antiporter) superfamily function as secondary carriers (see the TCDB). A similar situation has been demonstrated for members of the ECF (energy coupling factor) sub-superfamily of the ATP-binding cassette (ABC) superfamily [67]. Some of these porters transport vitamins such as biotin and thiamin either by ATP-dependent primary active transport or by proton motive force-driven secondary active transport [67] (E.I.S. and M.H.S., unpublished results).

It is a common assumption that sequence similarity between visual rhodopsins and microbial rhodopsins is non-existent [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 (a PQ-loop repeat-containing protein from *A. thaliana*), which is a member of the 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 the LCT family being 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 [15]. In summary, the connection between rhodopsin GPCRs and the LCT and Sweet families is stronger than the connections of many GPCRs to the most divergent glutamate GPCRs. Thus, the concept of 7-TMS GPCRs being a closely related group of sequences, while often taken for granted, is not valid, and the conclusion that sequence similarity between visual rhodopsin and bacteriorhodopsin is undetectable is equally invalid.

Using HMM:HMM comparisons (hhsuite-2.0.16) and the GRAFS system for GPCR classification, we detected a 30.6% probability of significant homology between the MR family (TC# 3.E.1) and the opsin cluster of α-rhodopsin GPCRs (Table S10). As a comparison, the glutamate cluster of the GPCRs shows only a 1.2% chance of a distant homologous relationship with the opsin cluster. This observation confirms the conclusion that different GPCRs are more divergent from each other than microbial and visual rhodopsins are from each other.

We are aware that sequence convergence is potentially capable of explaining some degree of sequence similarity when the regions compared are short, but skepticism is appropriate [69]. The need for stable transmembrane segments, along with functional requirements, may dictate sequence convergence in somewhat longer sequences [70–72]. However, we believe that convergence cannot explain a degree of similarity sufficient to give a comparison score of 12–14 SD for a stretch of over 60 amino acyl residues, particularly where two or more α-helical domains are aligned in a manner that makes evolutionary sense and fits a proposed pathway. The results of our control experiments using family members that evolved independently of the TOG superfamily support a current threshold of 12.0 SDs for establishing homology.

The elucidation of superfamily relationships is likely to open up new fields of study by allowing extrapolation of structural data from a well-characterized superfamily homolog to all or most members of the same superfamily. However, when the evolutionary process gives rise to homologs of differing topologies, extrapolation of structural data from one superfamily member to another may not be justified [14]. Future studies are required to reveal the degrees of structural dissimilarity that result from sequence divergence and topological variation within a superfamily.

Experimental procedures

Obtaining homologs and removing redundancies

The query sequences used to identify members of each family were (1) bacteriorhodopsin of *Halobacterium salinarum*

(GenBank accession number: gi# 114811, TC# 3.E.1.1.1), (2) MtN3 of *Medicago truncatula* (GenBank accession number: gi# 75220431, TC# 9.A.58.1.1), (3) PnuC of *Haemophilus influenzae* (GenBank accession number: gi# 81335937, TC# 4.B.1.1.2), (4) Predicted permease of *Pyrococcus abyssi* (GenBank accession number: gi# 74545625, TC# 2.A.102.4.1), (5) YfcA of *Escherichia coli* (GenBank accession number: gi# 82592533, TC# 2.A.102.3.1), (6) Membrane-like protein of *Oryza sativa* (GenBank accession number: gi# 75252893, TC# 2.A.102.5.1), (7) RcnA of *E. coli* (GenBank accession number: gi# 3025266, TC# 2.A.52.2.1), (8) Ostα of *Raja erinacea* (GenBank accession number: gi# 82108802, TC# 2.A.82.1.1), (9) NptA of *Vibrio cholera* (GenBank accession number: gi# 81345622, TC# 2.A.58.1.2), (10) CTNS (cystinosin) of *Homo sapiens* (GenBank accession number: gi# 269849555, TC# 2.A.43.1.1) and (11) ROP (red cone photoreceptor pigment) of *Homo sapiens* (GenBank accession number: gi# 129219, TC# 9.A.14.1.1). Analyses dealing with the HORC family TC# 1.A.69 were performed using Or83b of *Drosophila melanogaster* (GenBank accession number: gi# 14285640, TC# 1.A.69.1.1) as the query. PSI-BLAST searches with two iterations (e^{-4} and e^{-6} cut-offs, respectively) were performed using Protocol1 [14,73] 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 S1–S8). The CD-HIT option of Protocol1 was used to remove redundancies and highly similar sequences [14,18]. 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 mean hydropathy plots and phylogenetic trees. These percentage identity values refer to the values above which redundant sequences were removed. Thus, an 85% cut-off means that no two protein sequences retained for analysis were more than 85% identical. FASTA files from Protocol1 were considered representative of each protein family, although selected proteins that demonstrated homology between families were confirmed using NCBI's Conserved Domain Database [74] and through PSI-BLAST results.

Multiple alignments and topological analyses

ClustalX was used to create 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 generation of a coherent multiple alignment in which all or most sequences are homologous throughout most of their lengths. The results obtained with this program were compared with other programs (see also [75]), and when sequence similarity was sufficient to give reliable multiple alignments, the phylogenetic trees obtained using 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].

Establishing homology between families

Initially, a large screen was performed comparing distantly related TSUP family members [12] against all families of the TC# 2.A, 3.E and 9.A sub-classes. The targeted Smith–Waterman search (TSSearch) feature of Protocol2 [14] was then used in order to compare each family with all other TOG superfamily members [12,13]. TSSearch uses a rapid search algorithm to find distant homologs between two FASTA files that may not readily be apparent in BLAST or PSI-BLAST searches [14]. The most promising comparisons between proteins were automatically analyzed using the Global Sequence Alignment Tool (GSAT) [80] feature of Protocol2 [14]. Comparisons using the GSAT feature of Protocol2 are reported in standard deviations (SD), which refers to the number of SDs that a given score is from the mean raw local bit score of pairwise scores of 200 shuffled residues. Scores were calculated using the Needleman–Wunsch algorithm [80a]. Promising results with a comparison score of 12.0 SD 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 2000 random shuffles; assuming a Gaussian distribution, a comparison score of 12.0 SD corresponds to a probability of 1.77×10^{-33} that the degree of similarity between two proteins arose by chance (see Fig. S9) [81]. Despite this conclusion by Dayhoff *et al.* [81], Gaussian skewing increases the probability of chance similarity for any given standard deviation value [82].

Probabilities for comparison scores were calculated using Mathematica (http://www.wolfram.com/mathematica/; Wolfram Research Inc., Champaign, IL, USA). Comparisons involving at least 60 amino acyl residues, the mean size of a prototypical protein domain, alignment of two or more α-helical domains between compared proteins, and a comparison score of at least 12.0 SD were considered sufficient to provide strong evidence for homology between two proteins or internal repeat units in the studies reported [1,4,18,81]. 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 that we use a minimum of 60 amino acyl residues in defining homology is that, for such a long sequence, convergence to give 12 SD 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 [14] was used to search for internal repeats, and the 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 SDs in the same format as Protocol2.

Controls

A large screen was performed with all members of the TOG superfamily against the major intrinsic protein (MIP; TC# 1.A.8) and 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 performed using the same techniques and programs to establish homology between TOG superfamily members (Protocol1, Protocol2 and GSAT). The best comparison scores were selected using the same criteria as outlined previously; selected comparisons contained at least two 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 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 2000 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 the TCDB, using the default cut-off (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.

Phylogenetic and sequence analyses

The ClustalX program [51] was used to create multiple alignments 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 [16–18] were used to generate SFT1 and SFT2 phylogenetic trees using tens of thousands of BLAST bit scores instead of multiple alignments [18]. The SFT1 phylogenetic tree was generated to visualize relationships between all sub-families within families of the TOG superfamily. The SFT2 tree, drawn using the TreeView program [52], consolidated individual members into their respective families to visualize phylogenetic relationships between families within the TOG superfamily.

Obtaining sequences from the GRAFS system

Our starting point for mapping to the rhodopsin GPCRs was the well-known classification system for human GPCRs, published shortly after completion of the human genome sequence, the so-called GRAFS system [35]. Not all reported sequences are available, but as many as possible were extracted. We obtained a list of IDs from the GRAFS system for secretin (15), adhesion (24), glutamate (15) and frizzled/taste2 (24). Of these, 15 secretin, 24 adhesion, 15 glutamate and 23 Frizzled/Taste2 entries were used; one ID was a duplicate in the original publication. For the adhesion family, eight sequences were nucleotide entries.

The α-group of rhodopsin receptors (89) contains the prostaglandin receptor cluster (15). Of these, we eliminated two nucleotide entries because they did not refer to a single translated sequence and because the actual gene names were not present in the current entry, making them unidentifiable. This left 13 sequences in the prostaglandin receptor cluster. In the amine receptor cluster (40), one entry had been removed at the submitter's request, leaving 39. For the opsin receptor cluster (9), the melatonin receptor cluster (3) and the MECA receptor cluster (22), all sequences were retrievable.

In the β-group of rhodopsin receptors (35), one sequence (NP_004113.2) was rendered obsolete, leaving 34. In the γ-group of rhodopsin receptors (59), all of the sequences in the SOG receptor cluster (15), the MCH (melanin-concentrating hormone) receptor cluster (2) and the chemokine receptor cluster (42) were retrievable except NP_002021.1 in the chemokine receptor cluster, leaving 41 sequences in that cluster.

In the δ-group of rhodopsin receptors (58), one sequence in the MAS oncogene-related (http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=4142) receptor cluster (NP_089843.1), did not exist, leaving seven. All sequences in the glycoprotein receptor cluster (8) were retained. We eliminated one nucleotide entry

(NT_006337.5) in the purine receptor cluster (42), leaving 41. These sequences were saved to FASTA files and parsed so that the header only contained the sequence ID and the sequences were each on single lines in lower-case letters.

Training HMMs on the GRAFS sequences

MAFFT version 7.023b [85] was used to make alignments. The alignments were converted to Stockholm format (http://en.wikipedia.org/wiki/Stockholm_format). Using HMMER 3.0, we built an HMM for each alignment representing the major GRAFS groupings, creating four files for the non-rhodopsin groupings (Adhesion.hmm, Frizzled.hmm, Glutamate.hmm and Secretin.hmm), as well as five files for the first major cluster of rhodopsin sequences (Amine.hmm, MECA.hmm, Melatonin.hmm, Opsin.hmm and Prostaglandin.hmm), one file for the second major cluster of rhodopsin sequences (β.hmm), three files for the third major cluster of rhodopsin sequences (chemokine.hmm, MCH.hmm and SOG.hmm) and three files for the fourth major cluster of rhodopsin sequences (glycoprotein.hmm, MAS.hmm and purin.hmm), resulting in 16 HMMs.

HMMSEARCH was used, using the default similarity threshold (10.0), to search these 16 HMMs against 54 sequences in the GPCR family (TC# 9.A.14). As they are not listed in the GRAFS paper, we ignored the olfactory receptor cluster (estimated at 460), and, the other 7-TMS receptors.

Training HMMs on TOG sequences

On 4 March 2013, we downloaded 8790 proteins from the TCDB (http://www.tcdb.org/public/tcdb). Because some multi-component systems have multiple sequences under a single TC number (e.g. MexA and MexB), there were only 6316 unique IDs. We added a letter (A, B, C…) after the TC# when this was the case to distinguish the sequences. We found 167 sequences in the ten families comprising the TOG superfamily (including the odorant receptors, which are not established members of this superfamily): TC#s 1.A.69, 2.A.43, 2.A.52, 2.A.58, 2.A.82, 2.A.102, 3.E.1, 4.B.1, 9.A.14 and 9.A.58. We used MAFFT version 7.023b 'E-INS-I' (accurate). Each alignment was converted to Stockholm format, and HMMER 3.0 was used to build HMMs for each. Pfam-A was downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/Pfam-A.hmm.gz. Pfam-A was searched against the ten families comprising the TOG superfamily using a relatively stringent value of $1e^{-20}$, as well as the default (10.0) similarity threshold.

Mapping TOG to Pfam

The mapping from the default cut-off was used (parts of which may be seen in Fig. 5). This mapping contained 623 edges, mostly weak links to distantly related Pfam families. A scale of ten edge widths was used to represent each tenth of the distribution of e-values, with thick lines representing higher similarity. Using a spring-embedded layout, the nodes representing Pfam families were decreased in size.

We imported the Pfam mapping of the ten TCDB families to Pfam-A using an e-value threshold of 10 in Cytoscape 2.8.3. In total, 440 nodes and 623 edges were imported. We set the visual style to nested network style and applied the spring-embedded logic. We highlighted sets of nodes, such as TC# 9.A.14, using the Node Attribute Batch Editor. Visual Mapping Bypass was first used to establish a Node Attribute with node size 5 as default and 30 for TCDB nodes. In VizMapper™, a function in Cytoscape, we used a Discrete Mapper for Edge Width based on interaction.

Supplementing the TCDB sequence set with GRAFS sequences

We added three new GPCRs to the TCDB in order that representatives from all classes were present in the GRAFS system. The most recent additions were FZD1 (TC# 9.A.14.16.1), TAS2 (TC# 9.A.14.17.1) and KiSS (TC# 9.A.14.18.1). To ensure that the poor connectedness between TC# 9.A.14 sequences in mapping of the TOG superfamily to Pfam (Fig. 5) was not due to the lack of representation of these families, we used the 8843 sequences in the TCDB on 14 March 2013. Of these, 62 were GPCRs, containing representatives from all branches of the GPCR system. We used default settings in HMMER 3.0, using a threshold e-value of 10 to map these against Pfam-A. As the e-values depend on the database size, the exact e-values are not directly comparable with the other mapping. We applied a spring-embedded logic on 314 edges in Cytoscape 2.8.3, using a pass-through mapper on a 1–10 scale representing edge width bands, sub-dividing our edges.

HMM:HMM comparisons

We downloaded and installed the HHSuite (hhsuite-2.0.16) for HMM:HMM comparisons and used HHMAKE (HHmake version 2.0.15) to retrain HHMs (HMMs from HHmake) in our proposed superfamily. HMMs were not used, as use of the HMMER 3.0 format as input results in severe loss of sensitivity for the nine families (not including TC# 9.A.14, the GPCRs). We used the -M 50 flag for FASTA columns and HHsearch (2.0.16) to compare each of the α, β, γ and δ clusters of rhodopsin GPCRs [35]. In total, 12 HMMs were used from these groups: amine, MECA, melatonin, opsin, prostaglandin, β, chemokine, MCH, SOG, glycoprotein, MAS and purin [35]. These were compared with HMMs representing the nine TC families:

MR (TC# 3.E.1), Sweet (TC# 9.A.58), PnuC (TC# 4.B.1), TSUP (TC# 2.A.102), NiCoT (TC# 2.A.52), OST (TC# 2.A.82), PNaS (TC# 2.A.58), LCT (TC# 2.A.43) and HORC (TC# 1.A.69). For each comparison, we recorded the HHsearch (2.0.16) percentage probability, representing the probability of homology. The relevant TCDB families were also compared internally. The results are presented in Table S10.

Acknowledgements

This work was supported by National Institutes of Health grant number 2 RO1 GM077402-05A1. We thank Carl Welliver for assisting in the preparation of this manuscript, and Benjamin Hass for assisting in statistical interpretations.

References

1 Saier MH Jr (1994) Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol Rev* **58**, 71–93.
2 Saier MH Jr (2000) A functional–phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev* **64**, 354–411.
3 Saier MH Jr, Tran CV & Barabote RD (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181–D186.
4 Saier MH Jr, Yen MR, Noto K, Tamang DG & Elkan C (2009) The Transporter Classification Database: recent advances. *Nucleic Acids Res* **37**, D274–D278.
5 Busch W & Saier MH Jr (2002) The transporter classification (TC) system, 2002. *Crit Rev Biochem Mol Biol* **37**, 287–337.
6 Lam VH, Lee JH, Silverio A, Chan H, Gomolplitinant KM, Povolotsky TL, Orlova E, Sun EI, Welliver CH & Saier MH Jr (2011) Pathways of transport protein evolution: recent advances. *Biol Chem* **392**, 5–12.
7 Chang AB, Lin R, Studley WK, Tran CV & Saier MH Jr (2004) Phylogeny as a guide to structure and function of membrane transport proteins. *Mol Membr Biol* **21**, 171–181.
8 Mansour NM, Sawhney M, Tamang DG, Vogl C & Saier MH Jr (2007) The bile/arsenite/riboflavin transporter (BART) superfamily. *FEBS J* **274**, 612–629.
9 Furutani Y & Kandori H (2002) Internal water molecules of archaeal rhodopsins. *Mol Membr Biol* **19**, 257–265.
10 Hirai T, Subramaniam S & Lanyi JK (2009) Structural snapshots of conformational changes in a seven-helix membrane protein: lessons from bacteriorhodopsin. *Curr Opin Struct Biol* **19**, 433–439.
11 Zhou XE, Melcher K & Xu HE (2012) Structure and activation of rhodopsin. *Acta Pharmacol Sin* **33**, 291–299.
12 Shlykov MA, Zheng WH, Chen JS & Saier MH Jr (2012) Bioinformatic characterization of the 4–Toluene Sulfonate Uptake Permease (TSUP) family of transmembrane proteins. *Biochim Biophys Acta* **1818**, 703–717.
13 Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. *J Mol Biol* **276**, 71–84.
14 Reddy VS & Saier MH Jr (2012) BioV Suite – a collection of programs for the study of transport protein evolution. *FEBS J* **279**, 2036–2046.
15 Zhai Y, Heijne WH, Smith DW & Saier MH 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**, 206–223.
16 Chen JS, Reddy V, Chen JH, Shlykov MA, Zheng WH, Cho J, Yen MR & Saier MH 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.
17 Yen MR, Chen JS, Marquez JL, Sun EI & Saier MH (2010) Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. *Methods Mol Biol* **637**, 47–64.
18 Yen MR, Choi J & Saier MH Jr (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. *J Mol Microbiol Biotechnol* **17**, 163–176.
19 Chung YJ, Krueger C, Metzgar D & Saier MH Jr (2001) Size comparisons among integral membrane transport protein homologues in bacteria, Archaea, and Eucarya. *J Bacteriol* **183**, 1012–1021.
20 Saier MH Jr, Beatty JT, Goffeau A, Harley KT, Heijne WH, Huang SC, Jack DL, Jahn PS, Lew K, Liu J *et al.* (1999) The major facilitator superfamily. *J Mol Microbiol Biotechnol* **1**, 257–279.
21 Iwig JS, Rowe JL & Chivers PT (2006) Nickel homeostasis in *Escherichia coli* – the *rcnR–rcnA* efflux pathway and its linkage to NikR function. *Mol Microbiol* **62**, 252–262.
22 Rodrigue A, Effantin G & Mandrand-Berthelot MA (2005) Identification of *rcnA* (*yohM*), a nickel and cobalt resistance gene in *Escherichia coli*. *J Bacteriol* **187**, 2912–2916.
23 Dawson PA, Hubbert M, Haywood J, Craddock AL, Zerangue N, Christian WV & Ballatori N (2005) The heteromeric organic solute transporter α–β, Ostα–Ostβ, is an ileal basolateral bile acid transporter. *J Biol Chem* **280**, 6960–6968.

24 Seward DJ, Koh AS, Boyer JL & Ballatori N (2003) Functional complementation between a novel mammalian polygenic transport complex and an evolutionarily ancient organic solute transporter, OSTα–OSTβ. *J Biol Chem* **278**, 27473–27482.
25 Wang W, Seward DJ, Li L, Boyer JL & 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 USA* **98**, 9431–9436.
26 Takanaga H & Frommer WB (2010) Facilitative plasma membrane transporters function during ER transit. *FASEB J* **24**, 2849–2858.
27 Chen LQ, Hou BH, Lalonde S, Takanaga H, Hartung ML, Qu XQ, Guo WJ, Kim JG, Underwood W, Chaudhuri B *et al.* (2010) Sugar transporters for intercellular exchange and nutrition of pathogens. *Nature* **468**, 527–532.
28 Saier MH Jr (2003) Tracing pathways of transport protein evolution. *Mol Microbiol* **48**, 1145–1156.
29 Ghezzi C, Murer H & Forster IC (2009) Substrate interactions of the electroneutral Na^+-coupled inorganic phosphate cotransporter (NaPi–IIc). *J Physiol* **587**, 4293–4307.
30 Foster JW, Park YK, Penfound T, Fenger T & Spector MP (1990) Regulation of NAD metabolism in *Salmonella typhimurium*: molecular sequence analysis of the bifunctional nadR regulator and the nadA–pnuC operon. *J Bacteriol* **172**, 4187–4196.
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**, 4410–4420.
32 Penfound T & Foster JW (1999) NAD-dependent DNA-binding activity of the bifunctional NadR regulator of *Salmonella typhimurium*. *J Bacteriol* **181**, 648–655.
33 Kurnasov OV, Polanuyer BM, Ananta S, Sloutsky R, Tam A, Gerdes SY & Osterman AL (2002) Ribosylnicotinamide kinase domain of NadR protein: identification and implications in NAD biosynthesis. *J Bacteriol* **184**, 6906–6917.
34 Singh SK, Kurnasov OV, Chen B, Robinson H, Grishin NV, Osterman AL & 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**, 33291–33299.
35 Fredriksson R, Lagerstrom MC, Lundin LG & Schioth HB (2003) The G–protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256–1272.
36 Lagerstrom MC & Schioth HB (2008) Structural diversity of G protein-coupled receptors and significance for drug discovery. *Nat Rev Drug Discov* **7**, 339–357.
37 Civelli O, Reinscheid RK, Zhang Y, Wang Z, Fredriksson R & Schioth HB (2013) G protein-coupled receptor deorphanizations. *Annu Rev Pharmacol Toxicol* **53**, 127–146.
38 Krishnan A, Almen MS, Fredriksson R & Schioth HB (2012) The origin of GPCRs: identification of mammalian-like *Rhodopsin*, *Adhesion*, *Glutamate* and *Frizzled* GPCRs in fungi. *PLoS One* **7**, e29817.
39 Nordstrom KJ, Sallman Almen M, Edstam MM, Fredriksson R & Schioth HB (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**, 2471–2480.
40 Schioth HB & Fredriksson R (2005) The GRAFS classification system of G–protein coupled receptors in comparative perspective. *Gen Comp Endocrinol* **142**, 94–101.
41 Almen MS, Nordstrom KJ, Fredriksson R & Schioth HB (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 RC (2013) Structure–function of the G protein-coupled receptor superfamily. *Annu Rev Pharmacol Toxicol* **53**, 531–556.
43 Katritch V, Cherezov V & Stevens RC (2012) Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmacol Sci* **33**, 17–27.
44 Tang XL, Wang Y, Li DL, Luo J & Liu MY (2012) Orphan G protein-coupled receptors (GPCRs): biological functions and potential drug targets. *Acta Pharmacol Sin* **33**, 363–371.
45 Sato K, Pellegrino M, Nakagawa T, Vosshall LB & Touhara K (2008) Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452**, 1002–1006.
46 Smart R, Kiely A, Beale M, Vargas E, Carraher C, Kralicek AV, Christie DL, Chen C, Newcomb RD & Warr CG (2008) Drosophila odorant receptors are novel seven transmembrane domain proteins that can signal independently of heterotrimeric G proteins. *Insect Biochem Mol Biol* **38**, 770–780.
47 Touhara K (2009) Insect olfactory receptor complex functions as a ligand-gated ionotropic channel. *Ann NY Acad Sci* **1170**, 177–180.
48 Park JH & Saier MH Jr (1996) Phylogenetic, structural and functional characteristics of the Na–K–Cl cotransporter family. *J Membr Biol* **149**, 161–168.
49 Kuan J & Saier MH Jr (1993) The mitochondrial carrier family of transport proteins: structural,

functional, and evolutionary relationships. *Crit Rev Biochem Mol Biol* **28**, 209–233.
50 Kunji ER & Robinson AJ (2010) Coupling of proton and substrate translocation in the transport cycle of mitochondrial carriers. *Curr Opin Struct Biol* **20**, 440–447.
51 Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948.
52 Zhai Y, Tchieu J & Saier MH Jr (2002) A web-based Tree View (TV) program for the visualization of phylogenetic trees. *J Mol Microbiol Biotechnol* **4**, 69–70.
53 Park HH (2011) Structural analyses of death domains and their interactions. *Apoptosis* **16**, 209–220.
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**, 6284–6287.
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.
55a Reddy VS & Saier MH Jr (2012) BioV Suite – a collection of programs for the study of transport protein evolution. *FEBS J* **279**, 2036–2046.
56 Soding J (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960.
57 Murakami M & Kouyama T (2008) Crystal structure of squid rhodopsin. *Nature* **453**, 363–367.
58 Vinothkumar KR & Henderson R (2010) Structures of membrane proteins. *Q Rev Biophys* **43**, 65–158.
59 Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF & Babu MM (2013) Molecular signatures of G-protein-coupled receptors. *Nature* **494**, 185–194.
60 Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science* **289**, 739–745.
61 Pao SS, Paulsen IT & Saier MH Jr (1998) Major facilitator superfamily. *Microbiol Mol Biol Rev* **62**, 1–34.
62 Prakash S, Cooper G, Singhi S & Saier MH Jr (2003) The ion transporter superfamily. *Biochim Biophys Acta* **1618**, 79–92.
63 Castillo R & Saier MH (2010) Functional promiscuity of homologues of the bacterial ArsA ATPases. *Int J Microbiol* **2010**, 187373.
64 Hvorup RN, Winnen B, Chang AB, Jiang Y, Zhou XF & Saier MH Jr (2003) The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. *Eur J Biochem* **270**, 799–813.
65 Saier MH, Hvorup RN & Barabote RD (2005) Evolution of the bacterial phosphotransferase system: from carriers and enzymes to group translocators. *Biochem Soc Trans* **33**, 220–224.
66 Rodionov DA, Hebbeln P, Eudes A, ter Beek J, Rodionova IA, Erkens GB, Slotboom DJ, Gelfand MS, Osterman AL, Hanson AD *et al.* (2009) A novel class of modular transporters for vitamins in prokaryotes. *J Bacteriol* **191**, 42–51.
67 Hebbeln P, Rodionov DA, Alfandega A & Eitinger T (2007) Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. *Proc Natl Acad Sci USA* **104**, 2909–2914.
68 Josefsson LG (1999) Evidence for kinship between diverse G-protein coupled receptors. *Gene* **239**, 333–340.
69 Doolittle RF (1994) Convergent evolution: the need to be explicit. *Trends Biochem Sci* **19**, 15–18.
70 Baeza-Delgado C, Marti-Renom MA & Mingarro I (2013) Structure-based statistical analysis of transmembrane helices. *Eur Biophys J* **42**, 199–207.
71 Remmert M, Biegert A, Linke D, Lupas AN & Soding J (2010) Evolution of outer membrane β-barrels from an ancestral ββ hairpin. *Mol Biol Evol* **27**, 1348–1358.
72 Ried CL, Kube S, Kirrbach J & Langosch D (2012) Homotypic interaction and amino acid distribution of unilaterally conserved transmembrane helices. *J Mol Biol* **420**, 251–257.
73 Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
74 Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res* **41**, D348–D352.
75 Young GB, Jack DL, Smith DW & Saier MH Jr (1999) The amino acid/auxin:proton symport permease family. *Biochim Biophys Acta* **1415**, 306–322.
76 Tusnady GE & Simon I (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–850.
77 Zhai Y & Saier MH Jr (2001) A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J Mol Microbiol Biotechnol* **3**, 501–502.
78 Bailey TL & Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28–36.
79 Bailey TL & Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54.

80 Devereux J, Haeberli P & Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12**, 387–395.
80a Wallin E, Wettergren C, Hedman F & von Heijne G (1993) Fast Needleman-Wunsch scanning of sequence databanks on a massively parallel computer. *Comput Appl Biosci* **9**, 117–118.
81 Dayhoff MO, Barker WC & Hunt LT (1983) Establishing homologies in protein sequences. *Methods Enzymol* **91**, 524–545.
82 O'Hagan A & Leonard T (1976) Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* **63**, 201–202.
83 Soding J, Remmert M & Biegert A (2006) HHrep: *de novo* protein repeat detection and the origin of TIM barrels. *Nucleic Acids Res* **34**, W137–W142.
84 Park JH & Saier MH Jr (1996) Phylogenetic characterization of the MIP family of transmembrane channel proteins. *J Membr Biol* **153**, 171–180.
85 Katoh K, Misawa K, Kuma K & Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066.
86 Fredriksson R, Lagerström MC, Lundin LG & Schiöth HB (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* **63**, 1256–1272.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site:

Fig. 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 families.
Fig. S2. AveHAS plot and phylogenetic tree for the MR family.
Fig. S3. AveHAS plot, phylogenetic tree for the NiCoT family, and homology between members of the TSUP and NiCoT families.
Fig. 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, and GSAT alignment of TSUP and OST.
Fig. 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, and GSAT alignment of Sweet and OST.
Fig. S6. AveHAS plot, phylogenetic tree for the PNaS family, and homology between members of the Sweet and PNaS families.
Fig. 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, and GSAT alignment of PnuC and TSUP.
Fig. S8. AveHAS plot, phylogenetic tree 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, and GSAT alignment of GPCR and MR.
Fig. S9. Schematic representation of the mean bit score of a given alignment, algorithm input into Mathematica 7.0, and probability of notable Z scores.
Fig. S10. AveHAS plots representing all families within the TOG superfamily.
Fig. S11. Results from the ClustalX-based neighbor-joining program.
Table S1. 135 LCT protein sequences.
Table S2. 104 MR protein sequences.
Table S3. 134 NiCoT protein sequences.
Table S4. 52 OST protein sequences.
Table S5. 198 Sweet protein sequences.
Table S6. 147 PNaS protein sequences.
Table S7. 74 PnuC protein sequences.
Table S8. (A) 41 of the 97 GPCR protein sequences. (B) 56 of the 97 GPCR protein sequences.
Table S9. The highest comparison scores between TOG superfamily and the MC and MIP families.
Table S10. HMM:HMM pairwise comparison scores indicating probabilities of homology as measured by HHsearch in the HHSuite (hhsuite-2.0.16).
Movie S1. Video file of rotation of the superimposition of B-rho (microbial rhodopsin) and M-rho (visual rhodopsin).