Title

Characterization of the Tetraspan Junctional Complex (4JC) superfamily

Permalink

https://escholarship.org/uc/item/9h1867r8

Journal

Biochimica et Biophysica Acta (BBA) - Biomembranes, 1859(3)

ISSN

0005-2736

Authors

Chou, Amy
Lee, Andre
Hendargo, Kevin J
et al.

Publication Date

2017-03-01

DOI

10.1016/j.bbamem.2016.11.015

Peer reviewed

Biochimica et Biophysica Acta 1859 (2017) 402–414

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbamem

Characterization of the Tetraspan Junctional Complex (4JC) superfamily

Amy Chou [a,1], Andre Lee [a,1], Kevin J. Hendargo [a], Vamsee S. Reddy [a], Maksim A. Shlykov [b], Harikrishnan Kuppusamykrishnan [a], Arturo Medrano-Soto [a], Milton H. Saier Jr [a,*]

[a] *Department of Molecular Biology, Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116, United States*
[b] *Department of Orthopaedic Surgery, Medical School, University of Michigan, Ann Arbor, MI 48109-5624, United States*

ARTICLE INFO

Article history:
Received 8 October 2016
Received in revised form 22 November 2016
Accepted 29 November 2016
Available online 2 December 2016

Keywords:
Gap junctions
Tight junctions
Connexins
Occludins
Ca^{2+} channels
Superfamily

ABSTRACT

Connexins or innexins form gap junctions, while claudins and occludins form tight junctions. In this study, statistical data, derived using novel software, indicate that these four junctional protein families and eleven other families of channel and channel auxiliary proteins are related by common descent and comprise the Tetraspan (4 TMS) Junctional Complex (4JC) Superfamily. These proteins all share similar 4 transmembrane α-helical (TMS) topologies. Evidence is presented that they arose via an intragenic duplication event, whereby a 2 TMS-encoding genetic element duplicated tandemly to give 4 TMS proteins. In cases where high resolution structural data were available, the conclusion of homology was supported by conducting structural comparisons. Phylogenetic trees reveal the probable relationships of these 15 families to each other. Long homologues containing fusions to other recognizable domains as well as internally duplicated or fused domains are reported. Large "fusion" proteins containing 4JC domains proved to fall predominantly into family-specific patterns as follows: (1) the 4JC domain was N-terminal; (2) the 4JC domain was C-terminal; (3) the 4JC domain was duplicated or occasionally triplicated and (4) mixed fusion types were present. Our observations provide insight into the evolutionary origins and subfunctions of these proteins as well as guides concerning their structural and functional relationships.

1. Introduction

Connexins and innexins are the principal core proteins of gap junctions, while claudins and occludins are tight junctional core proteins [1]. All have the same topology with four α-helical transmembrane segments (TMSs), and all exhibit well-conserved extracytoplasmic cysteines that either are known to, or potentially can, form extracytoplasmic disulfide bridges [2,3].

In metazoan tissues, adjacent cells are often connected by connexin- or innexin-containing gap junctional channels [4] as well as claudin- and occludin-containing tight junctions [2,5–7]. All of these junctional proteins span the two plasma membranes. In the former cases, docking of the two half channels in the plasma membranes of two adjacent cells creates hexameric tori of junctional proteins enclosing an aqueous pore [8]. These densely packed gap junctional channels allow cells to exchange ions and small messenger molecules such as Ca^{2+} and cyclic nucleotides as well as oligonucleotides. They also coordinate electrical activities in excitable tissues [9].

In 2003, our laboratory published sequence, topological and phylogenetic analyses of the proteins that comprise the connexin, innexin, claudin and occludin families [1]. A multiple alignment of the sequences of each family was used to derive average hydropathy and similarity plots as well as a phylogenetic tree. Analyses led to the following conclusions: (1) In all four families, the most conserved regions of the proteins are the four TMSs, although the extracytoplasmic loops between TMSs 1 and 2, and TMSs 3 and 4 are usually well conserved [4]. (2) The phylogenetic trees revealed sets of orthologues except for the innexins where phylogeny primarily reflected the organismal source, probably due to a lack of close organismal sequence data [5]. (3) The two halves of the connexins exhibited similarities suggesting that they were derived from a common origin by an internal gene duplication event, but this possibility could not be demonstrated [6]. (4) Conserved cysteyl residues in the connexins and innexins pointed to a similar extracellular structure involved in hemichannel docking to create intercellular communication channels. Similar roles in homomeric interactions for conserved extracellular residues in the claudins and occludins were suggested. The apparent lack of obvious sequence and motif similarities between the four different families indicated that, if they did evolve from a common ancestral gene, they had diverged substantially to fulfill different functions.

In this work, statistical and other methods provide strong evidence that these four junctional protein families, as well as eleven additional families of ion (most frequently Ca^{2+}) channel and channel-affiliated proteins have, in fact, arisen from a common origin. The fifteen families that comprise the 4JC superfamily are listed with their characteristics in Table 1.

[*] Corresponding author.
E-mail address: msaier@ucsd.edu (M.H. Saier).

[1] These two authors contributed equally to the work reported.

http://dx.doi.org/10.1016/j.bbamem.2016.11.015
0005-2736/© 2016 Elsevier B.V. All rights reserved.

Table 1
Families included in the 4JC superfamily.[a]

Family name	Family Abb'n	TC#	Phyla	Avg seq. length (aas) ± SD	#s of members	# of potential fusion proteins	Pfam designations	References
Connexin	Connexin	1.A.24	Animals	323 ± 107	1690	12	Connexin (PF00029), Connexin43 (PF03508), Connexin50 (PF03509)	[10]
Innexin	Innexin	1.A.25	Animals + dsDNA viruses (3%)	389 ± 138	7971	16	Innexin (PF00876), Pannexin_like (PF12534), LRR_8 (PF13855)	[11]
Intracellular chloride channel	ICC	1.A.36	Animals + dsDNA viruses (1%)	494 ± 120	260	1	MCLC (PF05394)	[12]
Plasmolipin	Plasmolipin	1.A.64	Animals	169 ± 34	2413	1	MARVEL (PF01284)	[13]
The low affinity Ca^{2+} channel	LACC	1.A.81	Fungi	277 ± 31	263	0	Fig1 (PF12351)	[14]
Hair cell mechanotransduction channel	HCMC	1.A.82	Animals	221 ± 34	511	0	L_HMGIC_fpl (PF10242)	[15]
Calcium homeostasis Modulator Ca^{2+} channel	CALHM-C	1.A.84	Animals	351 ± 56	537	4	Ca_hom_mod (PF14798)	[16]
Claudin tight junction	Claudin	1.H.1	Animals	225 ± 35	4770	1	PMP22_Claudin (PF00822), SUR7 (PF06687)	[17]
Invertebrate PMP22-claudin	Claudin2	1.H.2	Animals	243 ± 75	537	2	Clc-like (PF07062), Claudin_2 (PF13903)	[18]
Ca^{2+} channel auxiliary subunit γ1-γ8	CCAγ	8.A.16	Animals	231 ± 60	3839	3	PMP22_Claudin (PF00822), GSG-1 (PF07803), Claudin_2 (PF13903), TMEM37 (PF15108)	[19]
Non-classical protein exporter	NCPE	9.A.27	Fungi	170 ± 13	584	1	MARVEL (PF01284), NCE101 (PF11654)	[20]
Clarin	CLRN	9.A.46	Animals	218 ± 42	407	1	None	[21]
Occludin	Occludin	9.B.41	Animals	531 ± 181	335	4	MARVEL (PF01284), Occludin_ELL (PF07303)	[22]
Tetraspan vesicle membrane protein	TVP	9.B.130	Animals	258 ± 116	812	3	MARVEL (PF01284)	[23]
MscS/DUF475	DUF475	9.B.179	Actinobacteria	298 ± 159	361	0		

[a] Family names and abbreviations are provided in columns 1 and 2, respectively, while family numbers in the Transporter Classification Database (TCDB; www.tcdb.org) [24–26] are provided in column 3. Class 1 indicates a channel function. Class 8 indicates a transporter auxiliary function while class 9 indicates that insufficient information is available to establish the mechanisms of action of these proteins. Phylum representation for each protein family of the 4JC superfamily is provided in column 4. Average sizes of the proteins in each family ± standard deviations (SD) are provided in column 5, while estimates of family sizes, expressed in numbers of proteins retrieved by running Psi-BLAST against the NCBI NR protein database with two iterations and a cutoff of 90% to eliminate redundancies and very similar (>90% identity) sequences can be found in column 6. Potential fusion proteins (column 7) are those that are at least 2× larger than the familial average. Pfam designation(s) for members of a given family, when available, are provided in column 8, and a representative reference is given in column 9. Additional references for each family can be found in TCDB.

Evidence is presented that members of these families arose following a pathway involving duplication of a primordial 2 TMS element to give rise to the current 4 TMS proteins. The gap junctional innexins and connexins proved to be more closely related to each other although the tight junctional occludins and claudins do not appear to be closely related. We suggest that the innexins, present primarily in invertebrates, were the precursors of connexins in vertebrates. Vertebrate pannexins, members of the innexin family [4], may have been obtained by vertebrates from invertebrates via horizontal transfer after vertebrates diverged from invertebrates, giving rise to the current families of connexins and innexins [1].

2. Methods

Representative members the first member of each TC subfamily within the fifteen families included in the 4JC superfamily (Table 1) were obtained from the Transporter Classification Database (TCDB; www.tcdb.org) and expanded using a PSI-BLAST search tool against the NCBI NR protein database within the Protocol1 program with an e-value cut-off of 0.005 and two iterations [27]. Redundant sequences were then removed using the CD-HIT component of Protocol1 with a 0.8 (80%) identity cutoff [28]. Using this approach, all sequences retained for analysis differ from each other by at least 20%.

Comparison scores, expressed in standard deviations (SD), were determined using the GSAT program [28]. GSAT performs a pairwise alignment using the Needleman-Wunsch algorithm, followed by 200 additional alignments using a shuffled sequence in each round. A standard score (z-score) is calculated and returned by the program. High scoring pairs (HSPs) were selected between families using the Protocol2 program [27]. Protocol2 performs a Smith-Waterman search between two FASTA files and selects the HSPs with overlapping TMSs. The HSPs are then analyzed with GSAT using 200 shuffles, and a standard score is determined for each. The greatest HSPs for each family comparison are then selected and again run through GSAT using 2000 shuffles to confirm scores and gain greater accuracy.

The Web-based Hydropathy, Amphipathicity and Topology (WHAT) program was used to determine and plot the hydropathy, amphipathicity, secondary structure and predicted transmembrane topology of individual protein sequences [29]. All TMS predictions for individual proteins were performed using the WHAT program, which predicts integral membrane protein topology using a Hidden Markov Model approach [30,31].

Multiple alignments were created using the ClustalX program [29]. Relative conservation was estimated using the AveHAS program [30], which generates average hydropathy, amphipathicity and similarity plots based on ClustalX multiple alignments, and also predicts topology with greater accuracy than is possible using the WHAT program (plots presented in Figs. 1 and 2).

Phylogenetic superfamily trees were created using the SuperFamilyTree (SFT) programs [32–34]. SFT works by creating 100 distance matrices using tens of thousands of Blast bit scores. The matrices are then built using the Fitch program. The trees are averaged using the Consense program to produce a superfamily tree [32–34]. SFT1 creates a tree showing the individual proteins while SFT2 collapses this tree to show the relationships of the families to each other [32–34].

The Ancient Rep [27], REPRO [35] and HHRepID [36] programs were used to recognize distant transmembrane repeats within a single protein sequence. The former two programs use a variation of the Smith-Waterman local alignment strategy to find non-overlapping top-scoring alignments, but AncientRep also allows screening of multiple homologues for repeats after construction of ClustalX-generated multiple alignments, allowing comparison both within single proteins (horizontal comparisons) and between multiple homologues (vertical comparisons) [27]. TMS repeat units were located using these programs, and their common origin was established using the GSAT program as outlined above.

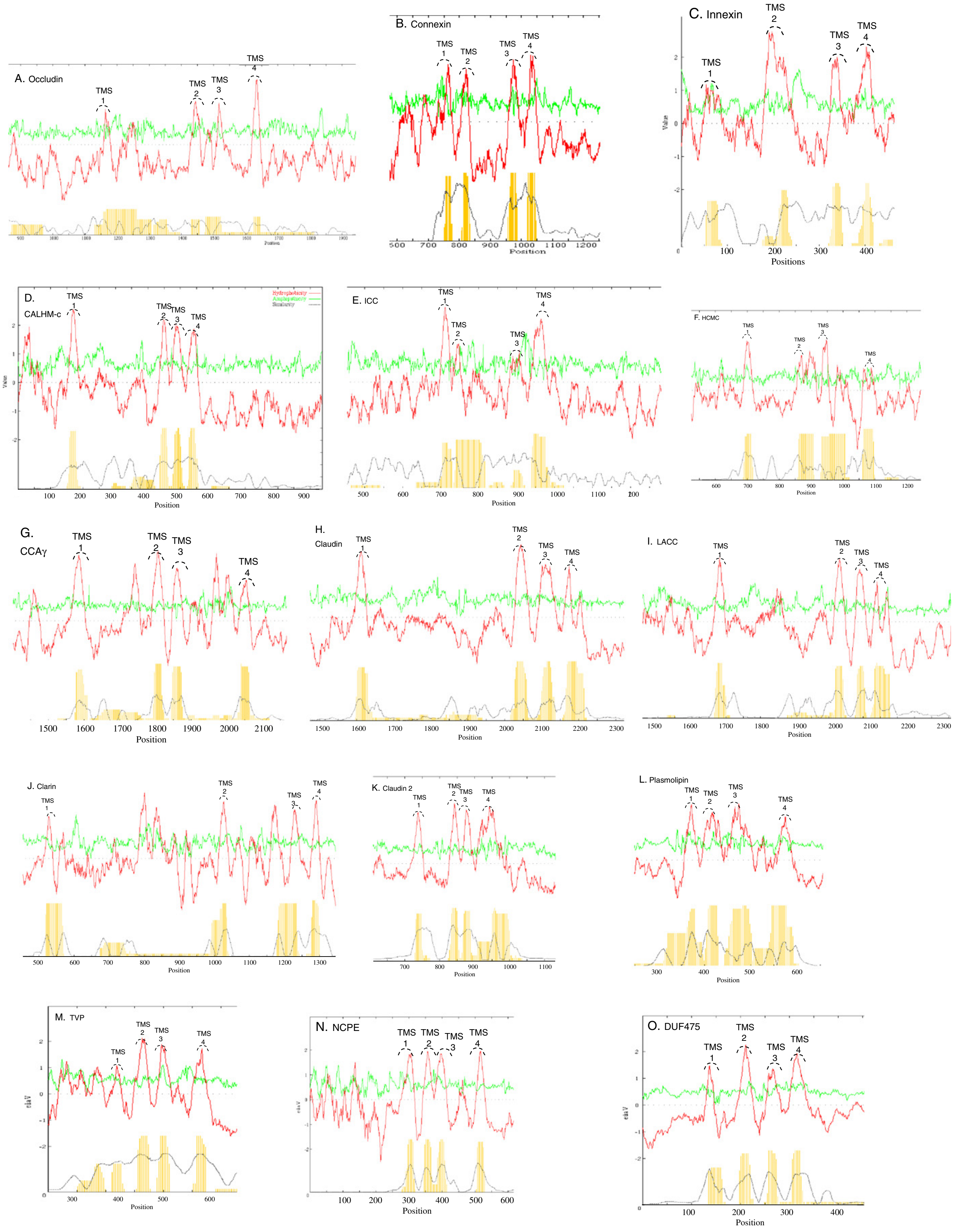


Fig. 1. AveHAS plots for all protein families in the 4JC superfamily. The families are indicated by their familial abbreviations (see Table 1). See Methods and the legend to Fig. 2 for explanation of format.

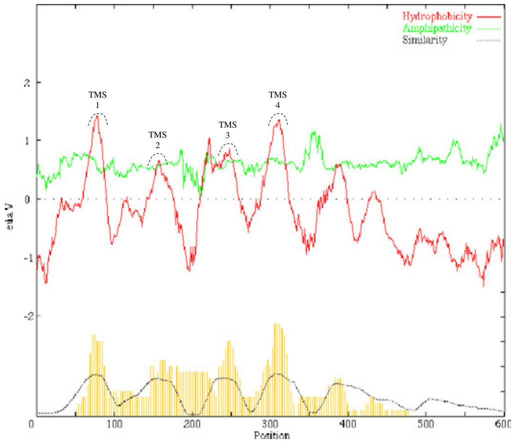


Fig 2. AveHAS plots for the entire 4JC superfamily with sequences aligned using the Clustal X program. See Methods section for procedures and references. The dark red line in the top plot represents average hydropathy, while the light green line represents average amphipathicity. The dotted black line in the lower plot shows the degree of conservation among the proteins at a particular location, while the thin vertical yellow lines indicate probable TMSs using a distinct program.

The phylum composition and average protein size ± S.D. for each of the 15 protein families of the 4JC superfamily were determined by using the Phylum-Size/Topology (PhyST) program [37]. The phyla of origin of the proteins in a family were automatically tabulated and used to quantitatively determine the phylum distribution for each family. The PhyST program was also used to determine the number of family members following use of the CD-hit program with a cut-off of 90% identity, with the CD hit program and to identify large potential fusion proteins for each of the 15 families in the 4JC superfamily. This program was used precisely as describes previously [37].

Three-dimensional structures for members of TCDB families 1.A.24, 1.H.1 and 8.A.16 were obtained from RCSB PDB [38] through sequence and structural similarity searches. First, for family members without structures, we ran the online PDB sequence similarity tool (E-value $\leq 10^{-3}$) to find homologs. Second, for family members with structures, we ran the online PDB structural similarity tool and retrieved structures showing RMSD values of ≤4.0. Finally, we ran HMMTOP [39,40] on all hits obtained and rejected structures with <4 predicted transmembrane segments (TMSs). Representative structural alignments based on pairwise combinations of structures between families were then computed using the Collaborative Computational Project 4 (CCP4) implementation of the Secondary Structure Matching (SSM) algorithm, which superposes structures with an emphasis on matching secondary structural elements as the name suggests, selecting for minimal RMSD values [41–43].

3. Results

3.1. Topological predictions

To predict the common and distinctive topological features of each family found to belong to the 4JC superfamily, average hydropathy, amphipathicity and similarity (AveHAS) plots were generated using homologues obtained using Protocol 1 with a query protein from each family in TCDB, the first member of each subfamily (of the 15 families of the 4JC superfamily) listed in TCDB (Table 1 and Fig. 1) [32–34]. The red lines in the top plots represent hydropathy, while the green lines represent amphipathicity. The dotted black lines below show the degrees of conservation among the proteins at any one location in the alignment while the vertical yellow lines show an independent prediction of TMSs. The AveHAS plots for the fifteen families: Occludins (Fig. 1A), Connexins (Fig. 1B), Innexins (Fig. 1C), CALHM-C (Fig. 1D), ICC (Fig. 1E), HCMC (Fig. 1F), CCAγ (Fig. 1G), Claudin (Fig. 1H), LACC (Fig. 1I), Clarin (Fig. 1J), Claudin 2 (Fig. 1K), Plasmolipin (Fig. 1L), TVP (Fig. 1M), NCPE (Fig. 1N), and DUF475 (Fig. 1O), all demonstrated a conserved 4 TMS topology. All 15 families showed comparable degrees of similarities for the four TMSs with slight differences being observed for a few families. For example, the connexins (Fig. 1B) had TMSs 1 and 2 better conserved than TMSs 3 and 4 while the CALHM-C proteins (Fig. 1D) showed the opposite behavior with TMSs 3 and 4 better conserved than TMSs 1 and 2.

3.2. Topological correspondence among all fifteen families within the 4JC superfamily

In addition to the AveHAS plots for the individual families, the AveHAS plot for the entire 4JC superfamily was generated as shown in Fig. 2. Four clear peaks of hydrophobicity corresponding to four peaks of similarity can be visualized. All four peaks show similar degrees of conservation, but TMSs 3 and 4 may be somewhat better conserved than TMSs 1 and 2. The best conserved TMS appears to be TMS 4. In general, the peaks of similarity are broader than the peaks of hydropathy with similarity preceding peaks 1 and 3 but following peaks 2 and 4. This suggests that the cytoplasmic regions adjacent to the TMSs are better conserved than the remaining parts of the cytoplasmic loops or the corresponding extracellular regions. All four peaks exhibit moderate amphipathicity.

All members of each of the fifteen families of the 4JC superfamily are homologous throughout most of their lengths, although insertions, deletions and fusions, primarily in their hydrophilic regions, have occurred in various protein members during their evolutionary divergence. Proteins used for the initial PSI-BLAST searches were the first member of each sub-family within the 15 families of the 4JC superfamily listed in the Transporter Classification Database (TCDB; www.tcdb.org) [24–26] under their respective families as indicated by abbreviation as summarized in Table 1. The values reported using this expanded dataset yielded scores that suggested homology between all fifteen families (Table 2). The criteria used for establishing homology were comparison scores of 14 standard deviations (SD) or greater, with an alignment of at least 60 amino acyl residues (aas) including corresponding TMSs [26,44].

The phylum representation of each protein family within the 4JC superfamily is provided in Table 1. A majority of the protein families of the 4JC superfamily are from Metazoa. Two families, LACC and NCPE, have proteins derived from Fungi, while the DUF475 family includes proteins only from Actinobacteria. Table 1 also presents the average sizes of the proteins comprising the fifteen 4JC families (column 5) and the relative family sizes (in numbers of proteins recovered as described in the Methods section (column 6)). The numbers of large proteins that could be fusion proteins as determined with the PhyST program were also tabulated. As illustrated in Figs. 1 and 2, they all have at least four conserved TMSs, a unifying characteristic of the 4JC superfamily. If the 4TMS 4JC domain is fully or partially, duplicated, triplicated or fused to another transmembrane domain, there will be more of TMSs, but this occurs rarely (see Table 3). Table 1 also lists Pfam designations for members of the various TC families when available (column 8). Of note is the fact that three families exhibit the Claudin domain while four families exhibit the MARVEL domain. This observation substantiates the conclusion of homology for these families. Table 1 also provides a reference (column 9). Additional references can be found in TCDB for all of the families listed.

3.3. Establishing homology between members of different families

The top comparison scores expressed in SD for each interfamilial comparison were obtained using the GSAT program with 2000 random shuffles. Proteins in TCDB were checked for homology as shown in Table 2 with scores supporting the conclusion of homology. Global sequence alignments for several interfamilial comparisons are presented in Fig. 3.

Table 2
Comparison scores expressed in standard deviations (SD) for the fifteen families in the 4JC superfamily.[a]

Families compared	Proteins compared (UniProt # or GI #)				Comparison score (SD)			
	Protein-1 (A)	Protein-2 (B)	Protein-3 (C)	Protein-4 (D)	A v. B	B v. C	C v. D	A v. D
Connexin v. Innexin	Q8NFK1	Q4SJR0	K8LRA8	Q8IWT6	62.6	17.9	32.3	0.8
ICC v. Innexin	Q96S66	J9JZG4	K1QMI8	Q96QZ0	26.5	14.9	15.3	−0.2
Occludin v. CCAγ	Q16625	H2L4X0	C3ZY32	P54825	154.4	14.2	16.9	2.2
HCMC v. CCAγ	Q8TAF8	291242472	F7BS52	Q06432	22.8	15.0	19.0	0.7
CCAγ v. Claudin	Q9D563	701422218	657540378	P56857	18.6	19.3	20.5	12.5
LACC v. Claudin	I3VPY1	S8ALD9	G8C1J8	P54003	16.3	16.0	95.4	9.5
Clarin v. CCAγ	A7SGP9	C1BSD7	R7TFA9	R7TFA9	21.8	14.3	126.5	9.5
Claudin2 v. Claudin	Q9NGJ7	U1MC19	R4GBS8	P56857	22.4	16.7	35.7	0.8
CCAγ v. Claudin2	Q9NY35	488549030	194750239	F5HJC0	178.5	17.3	125.0	8.1
TVP v. Connexin	P08247	B3RIL2	H3AJZ7	P08050	47.6	15.6	282.5	−1.0
Occludin v. Plasmolipin	Q16625	M7BW70	C3ZW40	P47897	18.5	15.5	23.0	4.2
Occludin v. NCPE	Q16625	F1QIE2	Q6FKV0	Q8NJ01	24.1	14.5	21.7	2.4
Occludin v. TVP	Q16625	432884588	527260494	P08247	152.9	15.7	63.3	1.7
Plasmolipin v. NCPE	P47897	602664033	K3VQ09	Q8NJ01	59.0	14.5	37.7	6.3
TVP v. DUF475	B3RX02	A0A077ZHE5	655407529	Q9KXK6	14.5	14.4	65.6	3.7
TVP v. Plasmolipin	P08247	E9CJG0	C3ZW39	P47897	22.4	14.3	29.4	−1.2
TVP v. NCPE	P08247	641792620	J7SAL5	A5E332	119.4	14.4	14.2	−1.1
CALHM-c v. Claudin	Q8IU99	821384408	E5R4H8	Q06991	46.5	14.0	22.1	−0.4

[a] The Superfamily Principle, which states that if A is related to B, and B is related to C, then A must be related to C (The Transitivity Rule), was used to establish homology. Column 1 gives the family abbreviations (see Table 1). Accession numbers of the four proteins compared (based on Protocol1 and Protocol2 results) are provided in columns 2–5, and the comparison scores for the 3 comparisons (A vs. B, B vs. C, and C vs. D) are given in columns 6–8. Column 9 gives the value obtained when A was directly compared with D. Accession numbers provided are UniProt numbers when available or gi numbers when UniProt numbers were not available.

Three patterns were observed when conducting binary alignments. The first demonstrated all or most of the four TMSs in a subject sequence aligning with their respective counterparts in the target sequence. Alignments of this type can be seen in Fig. 3A, B and D, where (A) a CCAγ homologue is compared with a Claudin homologue, (B) an HCMC homologue is compared with a CCAγ homologue, and (D) a Claudin homologue is compared with a Claudin2 homologue. These three comparisons gave comparison scores of 19.3 SD, 15.0 SD, and 16.7 SD, respectively. The second pattern of binary alignments shown in Fig. 3 always involved comparisons of corresponding TMSs (1 with 1; 2 with 2; 3 with 3; and 4 with 4, respectively) but with only two (Fig. 3G and H) or three (Fig. 3C, E and F) TMSs aligning. 10 SD has been reported to correspond to a probability of 10^{-24} that the observed degree of similarity has arisen by chance [45], but Gaussian skew can substantially increase this probability. Controls showed that 14 SD was 2 SD above the highest value that could be obtained when non-homologous sequences of similar topology were compared at the time these studies were conducted.

The remainder of the comparisons exhibited similar patterns. For example, the best comparison score obtained for the connexins and the innexins was 21 SD. A portion of this alignment is shown in Fig. 3H. The full alignment had 28% identity (I) and 48% similarity (S) for a stretch of 539 residue positions.

The best comparison score for the ICC family compared to the Innexin family (not shown) was 14.9 SD with 24% I and 49% S, spanning 183 residue positions. For the comparison between the HCMC and CCAγ families, a score of 15.0 SD was achieved with 26% I and 46% S, spanning 187 residue positions (Fig. 3B). The comparison of the clarin and CCAγ families gave a score of 14.3 SD with 25% I and 45% S, spanning 193 residue positions (not shown). The comparison between the claudin and claudin2 families (Fig. 3D) gave a score of 16.7 SD with 23% I and 46% S, spanning 192 residue positions. The highest scores for the claudin and LACC families (Figure 3E) were 16.0 SD, with 24% I and 45% S, spanning 150 residues. All other comparisons are listed in Table 2.

3.4. Evidence for 2 TMS repeat units in members of the 4JC superfamily

The third pattern of aligned sequences showed TMSs 1 and 2 of a single 4JC superfamily member aligning with TMSs 3 and 4 of the same protein (or a homologous protein), as shown in Fig. 4. Four alignments for the first and second halves of single proteins are presented, a connexin (A), an innexin (B), an HCMC family member (C), and an occludin (D). Additionally, heterologous comparisons (i.e., the first half of protein A aligned with the second half of Protein B) gave convincing comparison scores. For example, an occludin homologue and a CCAγ homologue gave a comparison score of 14 SD with 25% I and 54% S for a stretch of 67 residue positions. The value of 14 SD was sufficient to establish homology when these studies were conducted [32,37]. The AlignMe program was used to provide further evidence for similarity between the two halves of several of these proteins (See Fig. S1).

3.5. SFT-based phylogenetic trees

Phylogenetic trees for representative members of all 15 families in the 4JC Superfamily are shown in Fig. 5A and B, where A shows the relationships of representative proteins while B shows the integrated family relationships [32–34]. Because the abbreviations for the individual proteins in Fig. 5A may be too small to read, these are reproduced in clockwise order in supplementary Table S1. It will be noted that a very few proteins lie outside of the principle cluster that represents a particular family, but the two trees show good agreement. In both trees, the families cluster into four major groups. Nine families cluster loosely into two groups. Thus, in the first group (Cluster I; top), six families cluster loosely together. These are (from left to right in B): CALHM-C, Connexin, DUF475, LACC, Innexin and ICC. In the second cluster (Cluster II, center left), the Plasmolipin, TVP and NCPE families cluster together. In the lower right hand cluster (Cluster III), the Claudin, Claudin2, CCAγ and HCMC families cluster together with the Clarins at the base of this cluster. Finally, the Occludins (Cluster IV) branch together by themselves from a point near the center of the tree (lower left in A, right in B). These results provide evidence concerning the phylogenetic relationships of the fifteen families within the 4JC superfamily to each other.

3.6. 3D structural comparisons

Of the families believed to be members of the 4JC superfamily, high-resolution 3D structures are available for members of just three of these families. These families are the connexins (TC#1.A.24 [46]), type I claudins (TC#1.H.1, [47,48]), and Ca^{2+} channel γ auxiliary subunits (CCAγ; TC#8.A.16 [49]). The results of comparisons, selecting for minimal RMSD values, are presented in Fig. 6A–C. In each case, the front, back, and top views are shown. The superpositions were sufficient to provide strong

Table 3
Fusion proteins containing 4TC superfamily domains.[a]

Gi #	Sizes (aas)	Domain/order
1.A.25		
Connexins		
431890982	2729	Connexin - PDZ (protein protein interaction domain) - Myosin-XVIIIa (MYSc = myosin motor) - Tropomyosin - Kinetochore (microtubule binding domain) - Opi1 (phosphorylated Tx factor)
537123619	1178	Connexin - Pkc-like cyclin-dependent Ser/Thr kinase
530667615	994	Connexin - SPRY/TRIM domain (regulator of immune system) - olfactory receptor
521027364	760	Connexin - SPRY/TRIM tripartite motif containing domain
190358616	709	Connexin - uncharacterized hydrophilic domain
597731320	755	
736164105	700	
593734441	675	
528770327	839	Connexin - DUF3735 - ABA-GPCR (golgi pH regulator)
465977614	840	4 + 5 + 4 TMS topology; Connexin DUF3735-ABA-GPCR (abscisio acid receptor, G protein)
444706902	930	Connexin - Gtr1_RagA (P-loop NTPase)
47222966	948	LRR-RI-LRR-RI, Leucine-Rich Repeats (11 full repeats; protein-protein interaction domain) - Ribonuclease Inhibitor - Connexin
1.A.25		
Innexins		
669308587	795	Two complete adjacent duplicated innexin domains.
669225467	810	
339248393	813	
541046776	834	
568268171	797	
684378264	969	
669329541	1230	DUF2045 - TAF7 - Innexin
669313956	818	Ndr-Innexin (Ndr may be an α, β-hydrolase (Pfam00561)).
405960508	840	AAT - I (aspartate amino transferease) - Innexin
684379759	844	Innexin fragments - Innexin
674266122	717	(The innexin fragments precede the full length innexin domain)
734560734	836	
684386324	780	
684367491	884	
353231599	1023	
674595321	1006	Innexin - pyruvate kinase
1.A.36		
ICC		
528765010	1089	LisH (microtubule regulation) - W040 (signal transduction) - MCLC (ICC) Cl$^-$ channel
1.A.64		
Plasmolipin		
719732991	339	N-terminal hydrophilic domain – C-terminal MARVEL (plasmolipin domain)
1.A.81 No large proteins		
LACC		
1.A.82 No large proteins		
HCMC		
1.A.84		
CALHM-C		
465989358	1203	Dermatansulfate epimerase - CALHM-C
594679692	659	Duplicated CALHM-C domains. C-terminal hydrophilic domain; The second CALHM-C domain is better conserved.
537213670	668	
676278280	916	
1.H.1		
Claudin		
521024295	908	N-terminal 4 TMS Claudin domain - ARM repeat units (at least 5) (armadillo/β-cateinin repeats; protein-protein interaction domains).
1.H.2		
Claudin 2		
576700697	995	N-terminal 4 TMSs Claudin 2 domain - C-terminal DM10 (3OUF1128) domain; function unknown
322796000	627	1 + 4 + 4 + 3 TMSs. Three (triplicated) Claudin 2 domains
8.A.16		
CCAγ		
641736570	649	Duplicated 4 TMS CCAγ domains. The N-terminal domain resembles 8.A.16.2, while the C-terminal domain more closely resembles 8.A.16.1
351715945	630	
555949535	487	
9.A.27 No large proteins		
NCPE		
9.A.46		
Clarin		
521024245	525	Clarin - EEP (Endonuclease domain) - DUF4205
9.B.41		
Fusions to occludin		
528761243	1094	Two fused Occludin domains. The N-terminal domain is like subfamily 9.B.41.2. The C-terminal domain is more like subfamily 9.B.41.1.
465983309	1280	
528761243	1094	Two full 4 TMS Occludin (MARVEL - Occludin) repeats
465983309	1280	
9.B.130		
TVP family (MARVEL)		
528765018	1135	MARVEL - SCA7 Zn^{2+} binding domain - Cytb$_{561}$
431896453	1088	
432110159	980	MARVEL - Prickle-like protein 3 (PET_Prickle - LIM2-LIM3 Zn^{2+} binding)

[a] The family and its TC# as well as the GenBank ID number (gi#) are provided in column 1. The protein size in number of amino acyl residues (aas) can be found in column 2, and the recognized domains, in order from N- to C-terminus, are presented in column 3.

evidence of homology. Here, comparison of Connexin-26 with Claudin-19 (A) gave an RMSD value of 2.67, that for Connexin-26 with CCAγ (B) gave an RMSD value of 3.35, and that for Claudin-15 and CCAγ (C) gave an RMSD value of 1.90. Other comparisons not shown were as follows: (1) Claudin-19 versus CCAγ, 2.78 over 126 aas (3X29.C versus 3JBR.E), (2) Connexin-26 versus Claudin-15, 2.87 over 114 aas (2ZW3 versus 4P79A), (3) Claudin-19 versus Claudin 15, 1.36 over 151 aas (3X29.A versus 4P79.A). These values are highly suggestive of homology, confirming the conclusions based on primary sequence analyses.

3.7. Potential fusion proteins including 4JC domains

No large homologues were identified for the LACC (1.A.81), HCMC (1.A.82), NCPE (9.A.27) and DUF475 (9.B.179) families, but such proteins were found for all other 4JC families (Table 1).

3.7.1. Connexins (1.A.24)

Twelve connexin homologues from animals proved to be more than two-fold in size relative to the average size of these proteins, and all were examined (Table 3). All but one had full length connexin domains at their N-terminal ends; four of these had long C-terminal hydrophilic domains of unknown function. Two long homologues had an N-terminal connexin domain followed by a 7–9 TMS "Golgi pH Regulatory" domain (TC#1.A.38), which consists of a DUF3735 domain followed by an abscisic acid GPCR receptor (Aba_GPCR) domain. One protein had a C-terminal P-loop NTPase (Gtr1_RagA) domain. Two proteins had an SPRY/TRIM immune system regulatory domain C-terminal to the connexin domain, and one of these also had a C-terminal 7 TMS olfactory receptor domain (TC#9.A.14.8). The largest of these connexins had a size of 2729 aas. Following the connexin domain in this protein was (1) a PDZ protein-protein interaction domain, (2) a myosin-XVIIIa (MYSc myosin motor) domain, (3) a tropomyosin domain, (4) a microtubule binding kinetochore domain, and (5) an Opi1 (phosphorylated transcription factor) domain in this order.

Only one large homologue had the connexin domain at its C-terminus. This protein of 948 aas had an N-terminal multiply repeated leucine-rich domain. The NCBI database also contained shorter 2 TMS proteins corresponding to much of the N-terminal 2 TMS domain or

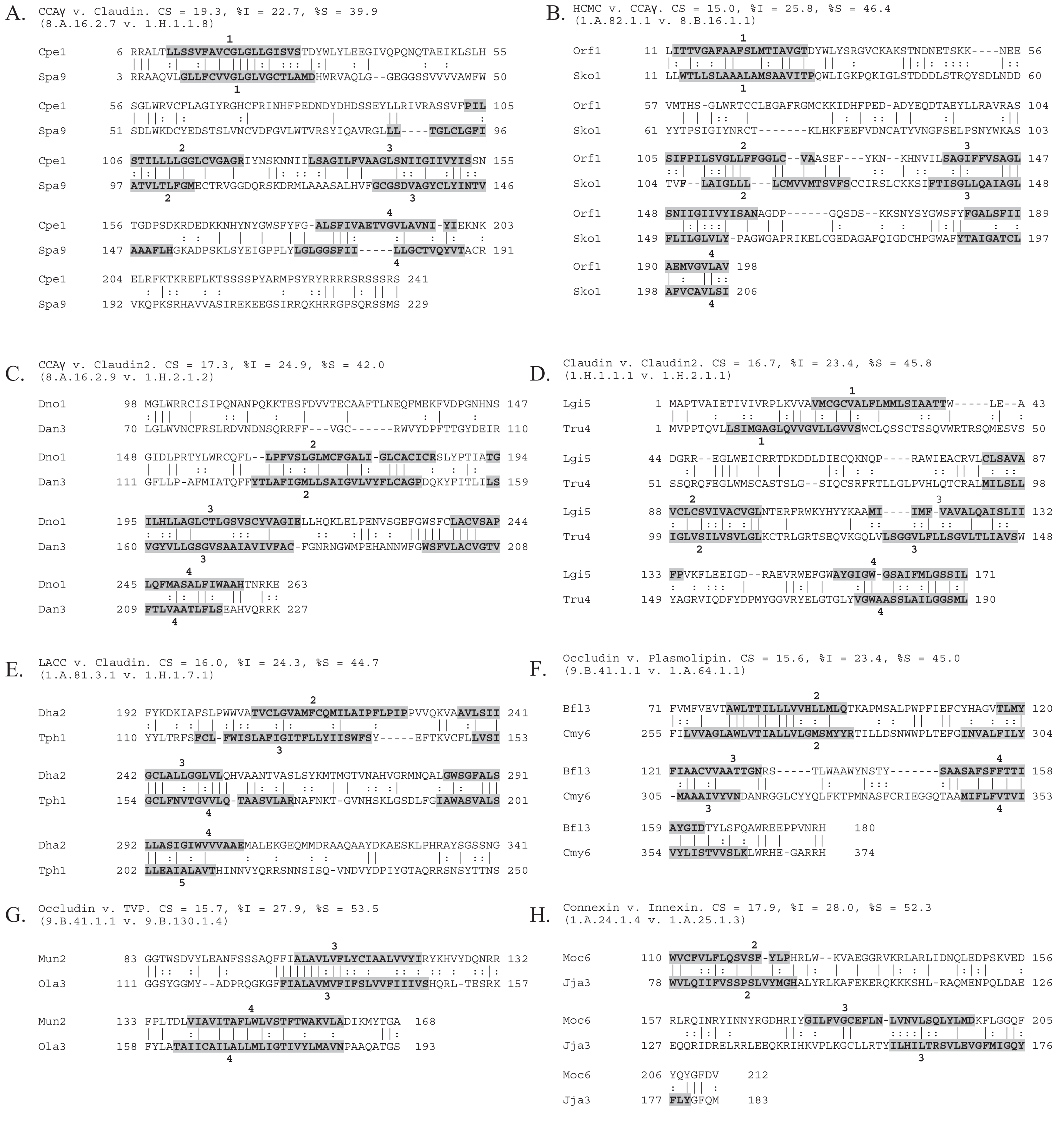

A. CCAγ v. Claudin. CS = 19.3, %I = 22.7, %S = 39.9
(8.A.16.2.7 v. 1.H.1.1.8)
Cpe1
Spa9
B. HCMC v. CCAγ. CS = 15.0, %I = 25.8, %S = 46.4
(1.A.82.1.1 v. 8.B.16.1.1)
Orf1
Sko1
C. CCAγ v. Claudin2. CS = 17.3, %I = 24.9, %S = 42.0
(8.A.16.2.9 v. 1.H.2.1.2)
Dno1
Dan3
D. Claudin v. Claudin2. CS = 16.7, %I = 23.4, %S = 45.8
(1.H.1.1.1 v. 1.H.2.1.1)
Lgi5
Tru4
E. LACC v. Claudin. CS = 16.0, %I = 24.3, %S = 44.7
(1.A.81.3.1 v. 1.H.1.7.1)
Dha2
Tph1
F. Occludin v. Plasmolipin. CS = 15.6, %I = 23.4, %S = 45.0
(9.B.41.1.1 v. 1.A.64.1.1)
Bfl3
Cmy6
G. Occludin v. TVP. CS = 15.7, %I = 27.9, %S = 53.5
(9.B.41.1.1 v. 9.B.130.1.4)
Mun2
Ola3
H. Connexin v. Innexin. CS = 17.9, %I = 28.0, %S = 52.3
(1.A.24.1.4 v. 1.A.25.1.3)
Moc6
Jja3

Fig 3. Global sequence alignments of various families within the 4JC superfamily demonstrating that corresponding TMSs align. Accession numbers for the proteins compared are provided in Table 2 (B vs. C). Residue numbers are provided at the beginning and end of each line. Shaded regions indicate the predicted TMSs which are numbered (1–4). CS, comparison score expressed in standard deviations (SD); %I = percent identity; %S = percent similarity. Vertical lines, identities; colons, similarities. The eight figures (A–H) show eight binary comparisons for the families indicated at the top of each alignments.

the C-terminal 2 TMS domain, each of 200–400 aas in length. The potential existence of such proteins, although functionally uncharacterized, supports the conclusion presented above, that 4JC proteins arose by intragenic duplication of a 2 TMS-encoding element.

3.7.2. Innexins (1.A.25)

Sixteen large Innexin homologues were identified. Six proteins, each from a different invertebrate species, proved to have internal duplications, containing two complete adjacent innexin domains, each with 4 TMSs (Table 3). One such protein was entered into TCDB with TC#1.A.25.1.11. Several other large proteins appeared to contain complete C-terminal innexin domains with N-terminal fragmentary innexin domains of variable sizes. These frequently included partial innexin fragments or sequences that were too distantly related to known proteins to allow their identification, even though some of them represented conserved domains. Recognized N-terminal domains included (1) a DUF2047-TAF7 region and (2) an Ndr domain, thought to be involved in cell differentiation, possibly an α, β-hydrolase (Pfam 00561). One

A) P08050 (Connexin) 1.A.24.1.1

Length = 118, Score = 77.00, Mean score = 0.65

```
              1
 26  LSVLF--IFRILLLGTAVESAW----GDEQSAFRCNTQQPGCENVCYDKS  69
      | ||   |    |       |            |          |
157  ISILFKSVFEVAFL----LIQWYIYGFSLSAVYTCKRDPCPHQVDCFLSR  202
              3

                          2
 70  FPISHVRFWVLQIIFVSVPTLLYLAHVFYVMRKEEKLNKKEEELKVAQTD  119
      |     |         |   |     |||  |  |   |        |
203  -PTEKTIFIIFMLVVSLVSLALNIIELFYVFFKGVKDRVKGRSDPYHATT  251
                    4

120  GVNVEMHLKQIEIKKFKY  137
     |       |     |  |
252  GPLS--PSKDCGSPKYAY  267
```

B) Q19746 (Innexin) 1.A.25.1.1

Length = 103, Score = 78.00, Mean score = 0.76

```
                    1
 26  RLSYVTTATLLAFFSIMVSCKQYVGSAIQCWMPMEFKGGWEQYAEDYC---FIQNTFFIP  82
      | |   |     | |        |     |       |    |  |          |
199  KLMYL--ANVFVQFIIL---NKFLGNETFLW-------GFHTFADLYAGREWQDSGVF--  244
                 3
                                          2
 83  ERSEIPGDVEDRQKAEIGYYQWVPIVLAIQAF--MFYLPSWIW  123
      |     |   |  |    |  |  ||  |   |    ||
245  PRVTL-CDFSVRKLANVHRYT-VQCVLMINMFNEKIYLFIWFW  285
                                             4
```

C) Q8TAF8 (HCMC) 1.A.82.1.1

Length = 94, Score = 46.00, Mean score = 0.49

```
                    1
 29  WGTLTICFSVLVMALFIQPYWIGDSVN---TPQAGYFGLFSYCVGNVLSS  75
     |  |          |     |    |      | |   |   |
132  WMQLAAATGLMIGCLVYPDGWDSSEVRRMCGEQTGKYTL-GHC-------  173
                 3
                                           2
 76 ELICKGGPLDFSSIPSRAFKTAMFFVALGMFLIIGSIICFSLFF  119
                     ||  |         |   |   | |
174 -------------TIRWAFMLAILSIGDALIL---SFLAFVLGY  201
                                        4
```

D) Q16625 (Occludin) 9.B.41.1.1

Length = 116, Score = 95.00, Mean score = 0.82

```
                              1
54   HFYKWTSPPGVIRILSMLIIVMCIA-IFACVASTLAWDRGYGTSLLGGSVGYPYGGSGFG 112
       |   |   |  ||        |  |             ||           |     |
170  RYY--LSVIIVSAILGIMVFIATIVYIMGVNPTAQSSGSLYGSQIYALCNQF-YTPAATG 226
                 3
                                      2
113 SYGSGYGYGYGYGYGYGGYTDPRAAKGFMLA-MAAFCFIAALVIFVTSVIRSEMSR  167
     |   | | |          ||  |    |  |    |  ||  ||     |  | |
227 LYVDQYLYHY-------CVVDPQEAIAIVLGFMIIVAF--ALIIFFAVKTRRKMDR  273
                                   4
```

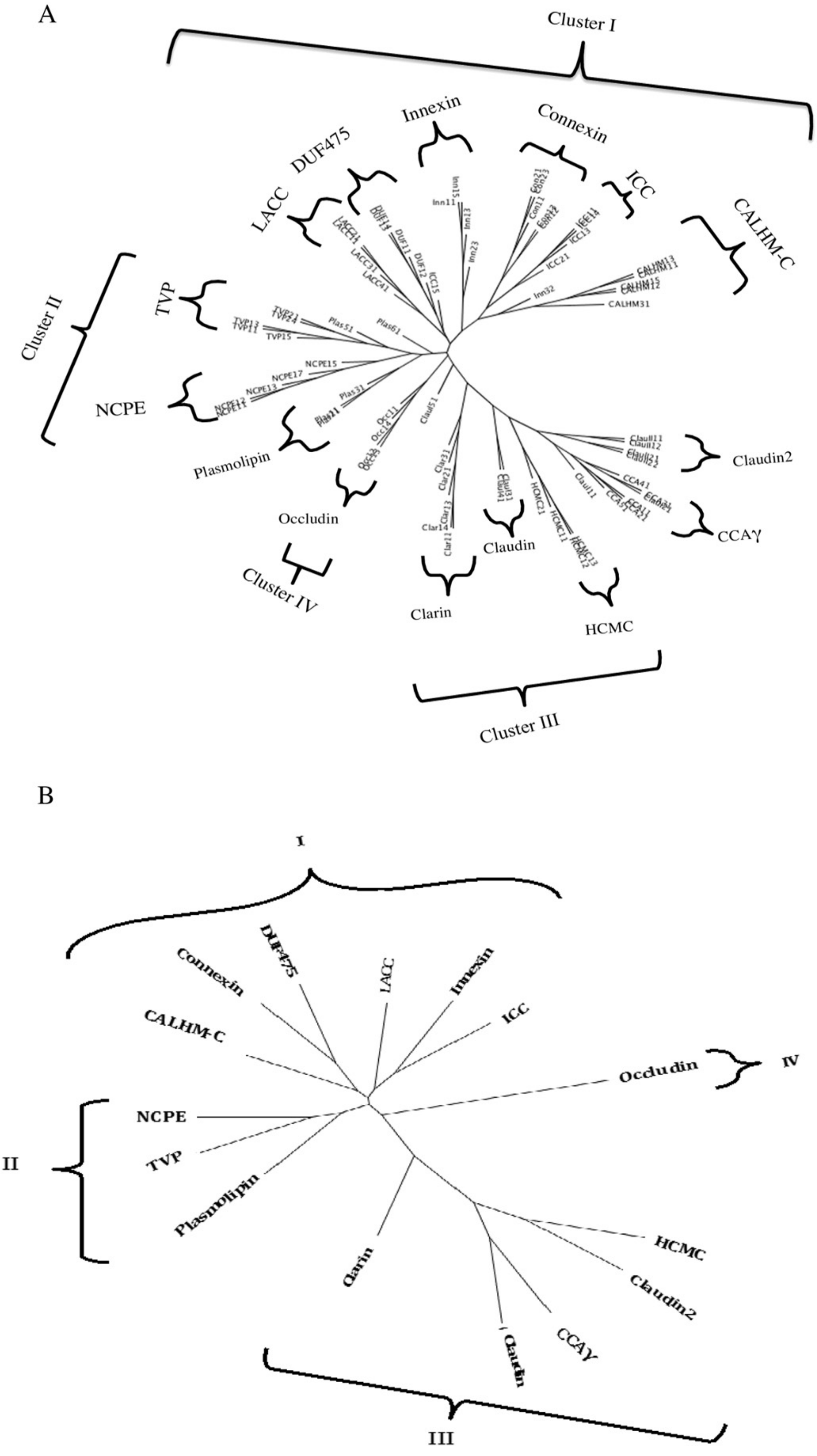
Fig. 4. Alignments of the two halves of several members of the 4JC superfamily. The Repro program was used to identify potential repeats with default settings for gap penalties: 10 for open, 1 for extension, and 50 for N local alignments. A, a connexin, B, an innexin, C, an HMCM protein, D, an occludin. The UniProt accession and TC numbers of the proteins studied are provided above each alignment. TMSs are shaded and numbered.

protein had an N-terminal innexin domain with a C-terminal pyruvate kinase domain.

It is interesting to note that while most connexin fusion proteins have their extra domains fused C-terminal to the connexin domain, the innexin fusion proteins have their extra domains fused N-terminal to the innexin domain. Additionally, while no protein was identified with two full length connexin domains, six homologues with two full length innexin domains were detected, and six more had full length C-terminal innexin domains with what appeared to be N-terminal innexin fragments.

3.7.3. ICC (1.A.36)

Only one large (>2× average) protein was identified for the ICC family. This protein, of 1089 aas, had an N-terminal LisH microtubule regulation domain, a central WD40 signal transduction domain and a C-terminal MCLC chloride channel domain (TC#1.A.36; Table 3).


A
Cluster I
Innexin
Connexin
DUF475
LACC
ICC
CALHM-C
Cluster II
TVP
NCPE
Plasmolipin
Occludin
Cluster IV
Clarin
Claudin
HCMC
Claudin2
CCAγ
Cluster III
Inn15
Inn11
Inn13
Inn23
Inn32
Con21
Con23
Con11
Con13
Con12
ICC11
ICC14
ICC13
ICC21
ICC15
CALHM13
CALHM11
CALHM15
CALHM12
CALHM31
DUF14
DUF13
DUF11
DUF12
LACC21
LACC11
LACC31
LACC41
TVP21
TVP24
TVP13
TVP11
TVP15
Plas51
Plas61
NCPE15
NCPE17
NCPE13
NCPE12
NCPE11
Plas31
Plas21
Plas11
Occ11
Occ14
Occ13
Occ15
Claud51
Clar31
Clar21
Clar13
Clar14
Clar11
Claul31
Claul41
HCMC21
HCMC11
HCMC13
HCMC12
Claul11
ClauII11
ClauII12
ClauII21
ClauII22
CCA41
CCA31
Claul21
CCA51
CCA11
CCA21
B
I
Connexin
DUF475
LACC
Innexin
ICC
CALHM-C
Occludin
IV
II
NCPE
TVP
Plasmolipin
Clarin
HCMC
Claudin2
CCAγ
Claudin
III

3.7.4. Plasmolipin (1.A.64)

One protein of 339 aas was substantially larger than its plasmolipin homologues. This protein had an N-terminal hydrophilic domain, not associated with a conserved domains in CDD or Pfam, followed by a single C-terminal plasmolipin domain.

3.7.5. CALHM-C (1.A.84)

Four large CALHM-C homologues were identified (Table 3). The first had a C-terminal CALHM-C domain fused to a large hydrophilic dermatan sulfate epimerase domain. The other three proteins, each from a different animal species, possessed two internally duplicated, full length, 4 TMS CALHM-C domains. In these cases, the C-terminal domains were better conserved (90–95% identical to its best TC hit) with the N-terminal domain exhibiting only about 30% identity with the same TC CALHM-C homologue. One of these proteins, with gi# 676278280, has a duplicated domain with 5 rather than 4 putative TMSs (peaks of hydrophobicity). Possibly, the domain duplication events occurred late during the evolution of these proteins.

3.7.6. Claudins (1.H.1 and 1.H.2)

Only three large claudin proteins were identified, one for the Claudin (TC#1.H.1) family and two for the Claudin 2 (TC#1.H.2) Family. The large claudin homologue (Table 3) was of 908 aas and had an N-terminal claudin domain of 4 TMSs followed by at least 5 ARM (Armadillo/β-cateinin) repeat units. The ARM domains are probably protein-protein interaction domains. The first of the large Claudin2 homologues, of 995 aas, had at least 3 DM10 (DUF1128) domains of unknown function, C-terminal to the Claudin2 domain. The second large Claudin 2 protein, of 627 aas, had a Claudin2 triplication in a 1 + 4 + 4 + 3 TMS arrangement. Repeat #3 showed 52% identity, repeat #1 showed 32% identity and repeat #2 showed 29% identity with TC#1.H.2.1.1.

3.7.7. CCAγ (8.A.16)

Three large homologues in the CCAγ family were identified, and all shared the same domain patterns. They exhibited duplicated 4 TMS Claudin2-like domains, the first belonging to subfamily 8.A.16.2, and the second belonging to subfamily 8.A.16.1. This fact suggests that they arose by gene fusion rather than by intragenic duplication. Finding multiple homologues with the same domain order increases confidence that these proteins did not result from artifacts of sequencing or exon identification.

3.7.8. Clarins (9.A.46)

Only one large homologue of the Clarins was found. This protein, of 525 aas, had three domains in the order: Clarin - EEP - DUF4205, where the EEP domain is found in endonucleases while the DUF4205 domain has not been characterized.

3.7.9. Occludins (9.B.41)

Two of four large occludin homologues, of 1094–1280 aas, displayed two full length occludin repeats. These proteins appeared to be fusions of two occludins rather than duplications because their N-terminal occludin domains resembled subfamily 9.B.41.2 proteins while the C-terminal domains most closely resembled subfamily 9.B.41.1 occludins. Two such proteins are listed in Table 3. This situation is similar to that observed for the CCAγ family. Two additional proteins also had two full length 4 TMS occludin domains, but their origins could not be ascertained.

3.7.10. Tetraspan vesicle membrane (9.B.130)

The 4 TMS 4JC domain of this family is referred to as MARVEL in CDD. Two proteins had their N-terminal MARVEL domains fused to a long hydrophilic domain with a Zn^{2+} binding SCA7 domain, and one of them had a C-terminal Cytochrome b_{561} domain. One other homologue had the N-terminal MARVEL domain fused to a hydrophilic Prickle-like protein 3 (PET_Prickle-LIM1-LIM2-LIM3) series of domains.

4. Discussion

The statistical analyses presented in this article provide the first evidence that four families of junctional proteins, the innexins, connexins, claudins and occludins, as well as eleven channel, transport auxiliary protein and uncharacterized families (see Tables 1 and 2) all arose from a common ancestor via the same pathway. In all cases, a 2 TMS hairpin structure with its N- and C-termini inside, probably duplicated to give the 4 TMS proteins. Interestingly, additional duplication or fusion events giving 8 TMS proteins with two 4JC domains and even 12 TMS proteins with three 4JC domains (with some variations) were identified (see section entitled “Potential Fusion Proteins” and Table 3). The 4 TMS topology is therefore the basic characteristic of all members of the 4JC superfamily, although in several cases, particularly putative “fusion” proteins, more TMSs were observed (see below). The relatively high frequencies of duplicated or fused 4JC domains, especially among the junctional connexins, innexins, claudins and occludins, is consistent with the conclusion that the proteins of each family form (hetero)oligomeric structures in the intact cell [50–58].

The evidence for homology between the fifteen families of the 4JC superfamily was substantial (all scores at or above the 14 SD cutoff; see Fig. 2 and Table 2), and in the few cases where 3-d structures were available (connexins, claudins, and CCAγ), structural comparisons confirmed this conclusion. Analysis of the phylogenetic trees for the proteins and families of the 4JC superfamily revealed some interesting details (Fig. 5A and B). First, all but two of the known channel proteins (T.C. Class 1.A) occur in cluster I at the tops of the two phylogenetic trees. The two exceptions are the Plasmolipin Family which can be found in Cluster III (middle left), and the HCMC Family, present in Cluster II (bottom). The DUF475 Family, with proteins derived from Actinobacteria, of unknown function, is found in Cluster I, suggesting that these proteins may be ion channels. Second, the plasmolipins cluster with the occludins and two poorly defined families, the TVP and NCPE families. None of the proteins in these families are mechanistically defined. Third, cluster III includes both Claudin families (Claudin and Claudin2) as well as the CCAγ and HCMC families with the Clarin family branching from a position much closer to the center of this radial tree. Fourth, the Occludin family branches from the center of the tree by itself (Branch IV). While Claudins and Occludins are known to be constituents of tight junctions, HCMC family members are mechanosensitive ion channels, while CCAγ proteins are believed to be auxiliary subunits of Ca^{2+} channels. Clarin 1 is a component of the USH complex involved in mechanotransduction [59], responsible, when defective, for deaf-blindness [60]. It is the causative protein which when mutated gives rise to the human Usher syndrome type 3A [61].

Of the 15 families in the 4JC superfamily, a search for proteins at least 2-fold larger than the average size of all members of the family revealed a few proteins that proved to have more than a single domain. Four of these families had no such recognizable fusion proteins, nine families had between 1 and 4 such members, and two, the connexins and innexins, had more, 12 and 15, respectively. These numbers indicate

Fig 5. Phylogenetic trees of representative proteins (A) and families (B) within the 4JC superfamily. The SuperfamilyTree programs (SFT1 and STF2) were used to generate the two trees, respectively (see 2. Methods). In A, the specific proteins examined in each of the 15 families have their protein abbreviations listed in Table S1 in supplementary materials, all listed in clockwise order in the tree. Those proteins that fall outside of their familial cluster are indicated by asterisks. Outside of these branches, indicating the positions of the individual proteins in the tree, the family abbreviation is provided. Finally, the four clusters (I-IV) are shown. B. The integrated tree in which each branch bears a single family. The same four clusters (I–IV) are indicated. Note that TG families corresponding to Pfam claudin/claudin2 families are in Cluster III, while TC families corresnpoding to the Pfam MARVEL family are in clusters II and IV. The family abbreviation used with their full names is presented in Table 1.

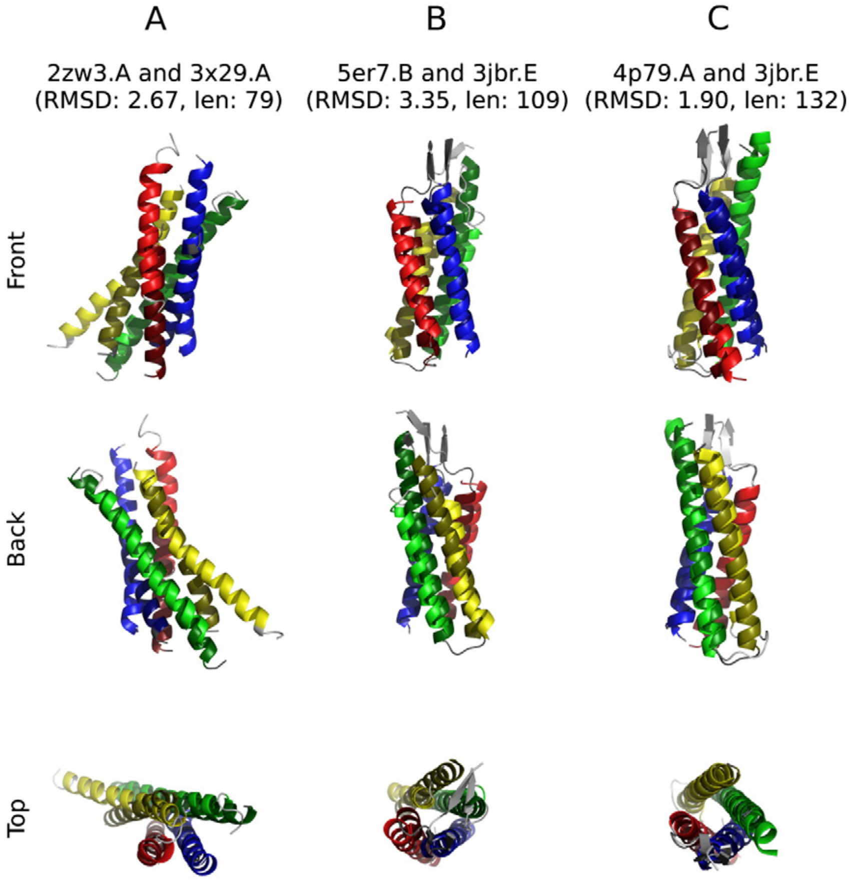


Fig. 6. Representative alignments of available 4JC structures. Left to right: A. Connexin-26 (TCDB: 1.A.24.1.3) and claudin-19 (TCDB: 1.H.1.1.5); B. connexin-26 (TCDB: 1.A.24.1.3) and CCAγ (TCDB: 8.A.16.1.1); C. claudin-15 (TCDB: 1.H.1.1.9) and CCAγ (TCDB: 8.A.16.1.1). Hydrophilic domains and unaligned loops have been excluded for clarity. The color-coding is as follows:

Color	TMS	A	B	C
Light red	1	2ZW3.A	5ER7.B	4P79
Light yellow	2			
Light green	3			
Light blue	4			
Dark red	1	3X29.A	3JBR.E	3JBR.E
Dark yellow	2			
Dark green	3			
Dark blue	4			

that almost all members of the 4JC superfamily consist of single domain proteins; there are only a few exceptions. Some of these large proteins may have resulted from errors in sequencing, incorrect intron/exon assignment or misinterpretation of the sequence data.

The most common occurrence among large putative fusion proteins were internal duplications, having two, or in a few cases, three, 4JC domains. These proteins often have family-specific characteristics. For example, no internally duplicated connexin was identified, although several putative connexin proteins were fused to other domains, almost always with the connexin domain N-terminal. By contrast, six innexin homologues were internally duplicated, and six more displayed C–terminal innexin domains with N-terminal fragments of innexins. In contrast to the connexins, where N-terminal 4JC domains were fused to other domains, the innexin domains were almost always C-terminal.

In each of the ICC and Plasmolipin Families, only a single fusion protein was identified, and in both cases, the 4JC domain was C-terminal. A single CALHM-C protein had a C-terminal 4JC domain with an N-terminal dermatan sulfate epimerase domain, but three other proteins had duplicated 4JC domains, where the C-terminal domains were better conserved. By contrast, Claudins, when fused to other domains, had the 4JC domain N-terminal. A single Claudin2 homologue seemed to have three (triplicated) 4JC domains. Three large CCAγ homologues had the 4JC domains C-terminal with the other domains being N-terminal. The single Clarin fusion protein identified also had its 4JC domain C-terminal.

The large occludins and TVP proteins contained 4JC domains recognized by CDD as MARVEL/occludin domains. Of the occludins, two had two C-terminal 4JC domains with these 4JC domains fused to hydrophilic N-terminal domains. Of the three large TVP homologues, all had the 4JC domain N-terminal to the other domains.

It is apparent, that the occurrence of internally duplicated (or fused) or partially duplicated 4JC domains represented a fairly high percentage of the large proteins and that several of these appeared to have arisen by fusion of two dissimilar 4JC domains, rather than duplication of a single such domain.

Whether these fusions occur with the 4JC domain C-terminal or N-terminal depended on the family to which these proteins belong. One primary function of the fusions could be to anchor a soluble enzyme or structural protein to the membrane with formation of a multiprotein complex. However, it is also possible that these fusions provide cooperative metabolic regulatory functions. Systematic identification of these fusion proteins provides food for thought and future research prospects concerning their evolution and functions.

The observations reported in this communication suggest that the 4 TMS topology, conserved in all members of the 4JC superfamily, is important for a structure and/or function common to all of its members. The two patterns of topological alignment, one showing the same TMSs (1–4) aligning with the corresponding TMSs in members of another family, and the other showing TMSs 1 and 2 aligning with TMSs 3 and 4, substantiate the conclusion of homology and also provide evidence for the conclusion that the proteins of this superfamily arose via intragenic duplication. These suggestions were substantiated using rigorous statistical criteria, and by the identification of 2 TMS elements in some homologues.

In general, we observed that the 4 TMSs in any one family of the 4JC superfamily were conserved to similar degrees. This suggests that these 4 TMSs are of comparable importance, both structurally and functionally, in all 15 families. In some families, differing degrees of conservation were observed, but these were not large differences. In these instances, however, parts of the proteins may serve extra functions not assumed by the others. We suggest that family-specific functions could include subunit:subunit interactions, channel formation, and hemichannel docking. Superfamily-generalized functions could include overall 3-dimensional structural features, subunit stability, and proper biogenesis.

Conflict of interest

This work was supported by NIH grant GM077402. The authors declare no conflict of interest.

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgements

This work was supported by NIH grant GM077402. We thank Joshua Asiaban, Anne Chu and Sabrina Phan for assistance with manuscript preparation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bbamem.2016.11.015.

References

[1] V.B. Hua, A.B. Chang, J.H. Tchieu, N.M. Kumar, P.A. Nielsen, M.H. Saier Jr., Sequence and phylogenetic analyses of 4 TMS junctional proteins of animals: connexins, innexins, claudins and occludins, J. Membr. Biol. 194 (2003) 59–76.

[2] C.M. Morrow, D. Mruk, C.Y. Cheng, R.A. Hess, Claudin and occludin expression and function in the seminiferous epithelium, Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 365 (2010) 1679–1696.

[3] M. Suga, S. Maeda, S. Nakagawa, E. Yamashita, T. Tsukihara, A description of the structural determination procedures of a gap junction channel at 3.5 A resolution, Acta Crystallogr. D Biol. Crystallogr. 65 (2009) 758–766.

[4] M.R. Yen, M.H. Saier Jr., Gap junctional proteins of animals: the innexin/pannexin superfamily, Prog. Biophys. Mol. Biol. 94 (2007) 5–14.

[5] P.M. Cummins, Occludin: one protein, many forms, Mol. Cell. Biol. 32 (2012) 242–250.

[6] M.K. Findley, M. Koval, Regulation and roles for claudin-family tight junction proteins, IUBMB Life 61 (2009) 431–437.

[7] C.E. Overgaard, B.L. Daugherty, L.A. Mitchell, M. Koval, Claudins: control of barrier function and regulation in response to oxidant stress, Antioxid. Redox Signal. 15 (2011) 1179–1193.

[8] A. Oshima, K. Tani, Y. Hiroaki, Y. Fujiyoshi, G.E. Sosinsky, Three-dimensional structure of a human connexin26 gap junction channel reveals a plug in the vestibule, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 10034–10039.

[9] J.C. Herve, P. Phelan, R. Bruzzone, T.W. White, Connexins, innexins and pannexins: bridging the communication gap, Biochim. Biophys. Acta 1719 (2005) 3–5.

[10] K.B. Moore, J. O'Brien, Connexins in neurons and glia: targets for intervention in disease and injury, Neural Regen. Res. 10 (2015) 1013–1017.

[11] G. Dahl, K.J. Muller, Innexin and pannexin channels and their signaling, FEBS Lett. 588 (2014) 1396–1402.

[12] H. Al Khamici, L.J. Brown, K.R. Hossain, A.L. Hudson, A.A. Sinclair-Burton, J.P. Ng, E.L. Daniel, J.E. Hare, B.A. Cornell, P.M. Curmi, M.W. Davey, S.M. Valenzuela, Members of the chloride intracellular ion channel protein family demonstrate glutaredoxin-like enzymatic activity, PLoS One 10 (2015), e115699. .

[13] Y. Yaffe, I. Hugger, I.N. Yassaf, J. Shepshelovitch, E.H. Sklan, Y. Elkabetz, A. Yeheskel, M. Pasmanik-Chor, C. Benzing, A. Macmillan, K. Gaus, Y. Eshed-Eisenbach, E. Peles, K. Hirschberg, The myelin proteolipid plasmolipin forms oligomers and induces liquid-ordered membranes in the Golgi complex, J. Cell Sci. 128 (2015) 2293–2302.

[14] B. Cavinder, F. Trail, Role of Fig1, a component of the low-affinity calcium uptake system, in growth and sexual development of filamentous fungi, Eukaryot. Cell 11 (2012) 978–988.

[15] B. Zhao, Z. Wu, N. Grillet, L. Yan, W. Xiong, S. Harkins-Perry, U. Muller, TMIE is an essential component of the mechanotransduction machinery of cochlear hair cells, Neuron 84 (2014) 954–967.

[16] Z. Ma, A.P. Siebert, K.H. Cheung, R.J. Lee, B. Johnson, A.S. Cohen, V. Vingtdeux, P. Marambaud, J.K. Foskett, Calcium homeostasis modulator 1 (CALHM1) is the pore-forming subunit of an ion channel that mediates extracellular Ca2+ regulation of neuronal excitability, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) E1963–E1971.

[17] C.T. Capaldo, A. Nusrat, Claudin switching: physiological plasticity of the tight junction, Semin. Cell Dev. Biol. 42 (2015) 22–29.

[18] M.H. Jaspers, K. Nolde, M. Behr, S.H. Joo, U. Plessmann, M. Nikolov, H. Urlaub, R. Schuh, The claudin Megatrachea protein complex, J. Biol. Chem. 287 (2012) 36756–36765.

[19] D.M. MacLean, S.S. Ramaswamy, M. Du, J.R. Howe, V. Jayaraman, Stargazin promotes closure of the AMPA receptor ligand-binding domain, J. Gen. Physiol. 144 (2014) 503–512.

[20] P.P. Tovaranonte, S.W. Beasley, K. Maoate, R. Blakelock, A. Skinner, Trends in the use of minimally invasive surgery in children, N. Z. Med. J. 123 (2010) 15–22.

[21] S.R. Gopal, D.H. Chen, S.W. Chou, J. Zang, S.C. Neuhauss, R. Stepanyan, B.M. McDermott Jr., K.N. Alagramam, Zebrafish models for the mechanosensory hair cell dysfunction in usher syndrome 3 reveal that clarin-1 is an essential hair bundle protein, J. Neurosci. 35 (2015) 10188–10201.

[22] S.M. Krug, J.D. Schulzke, M. Fromm, Tight junction, selective permeability, and related diseases, Semin. Cell Dev. Biol. 36 (2014) 166–176.

[23] C.P. Arthur, M.H. Stowell, Structure of synaptophysin: a hexameric MARVEL-domain channel protein, Structure 15 (2007) 707–714.

[24] M.H. Saier Jr., V.S. Reddy, D.G. Tamang, A. Vastermark, The transporter classification database, Nucleic Acids Res. 42 (2014) D251–D258.

[25] M.H. Saier Jr., V.S. Reddy, B.V. Tsu, M.S. Ahmed, C. Li, G. Moreno-Hagelsieb, The transporter classification database (TCDB): recent advances, Nucleic Acids Res. 44 (2016) D372–D379.

[26] M.H. Saier Jr., M.R. Yen, K. Noto, D.G. Tamang, C. Elkan, The transporter classification database: recent advances, Nucleic Acids Res. 37 (2009) D274–D278.

[27] V.S. Reddy, M.H. Saier Jr., BioV Suite–a collection of programs for the study of transport protein evolution, FEBS J. 279 (2012) 2036–2046.

[28] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[29] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.

[30] Y. Zhai, M.H. Saier Jr., A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins, J. Mol. Microbiol. Biotechnol. 3 (2001) 285–286.

[31] Y. Zhai, M.H. Saier Jr., 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 (2001) 501–502.

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

[33] M.R. Yen, J.S. Chen, J.L. Marquez, E.I. Sun, M.H. Saier, Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters, Methods Mol. Biol. 637 (2010) 47–64.

[34] M.R. Yen, J. Choi, M.H. Saier Jr., Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution, J. Mol. Microbiol. Biotechnol. 17 (2009) 163–176.

[35] R.A. George, J. Heringa, The REPRO server: finding protein internal sequence repeats through the Web, Trends Biochem. Sci. 25 (2000) 515–517.

[36] A. Biegert, J. Soding, De novo identification of highly diverged protein repeats by probabilistic consistency, Bioinformatics 24 (2008) 807–814.

[37] H. Kuppusamykrishnan, L.M. Chau, G. Moreno-Hagelsieb, M.H. Saier Jr., Analysis of 58 families of holins using a novel program, PhyST, J. Mol. Microbiol. Biotechnol. 26 (2016) 381–388.

[38] P.W. Rose, A. Prlic, C. Bi, W.F. Bluhm, C.H. Christie, S. Dutta, R.K. Green, D.S. Goodsell, J.D. Westbrook, J. Woo, J. Young, C. Zardecki, H.M. Berman, P.E. Bourne, S.K. Burley, The RCSB Protein Data Bank: views of structural biology for basic and applied research and education, Nucleic Acids Res. 43 (2015) D345–D356.

[39] M. Ikeda, M. Arai, D.M. Lao, T. Shimizu, Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, In Silico Biol. 2 (2002) 19–33.

[40] A. Reddy, J. Cho, S. Ling, V. Reddy, M. Shlykov, M.H. Saier, Reliability of nine programs of topological predictions and their application to integral membrane channel and carrier proteins, J. Mol. Microbiol. Biotechnol. 24 (2014) 161–190.

[41] E. Krissinel, K. Henrick, Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions, Acta Crystallogr. D Biol. Crystallogr. 60 (2004) 2256–2268.

[42] M.D. Winn, C.C. Ballard, K.D. Cowtan, E.J. Dodson, P. Emsley, P.R. Evans, R.M. Keegan, E.B. Krissinel, A.G. Leslie, A. McCoy, S.J. McNicholas, G.N. Murshudov, N.S. Pannu, E.A. Potterton, H.R. Powell, R.J. Read, A. Vagin, K.S. Wilson, Overview of the CCP4 suite and current developments, Acta Crystallogr. D Biol. Crystallogr. 67 (2011) 235–242.

[43] E. Aasum, D.A. Lathrop, T. Henden, R. Sundset, T.S. Larsen, The role of glycolysis in myocardial calcium control, J. Mol. Cell. Cardiol. 30 (1998) 1703–1712.

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

[45] L. Filipovic, M. Hlavka, Our experiences and results in the surgical treatment of metacarpal bone and had phalangeal fractures, Acta Chir. Iugosl. 24 (1977) 485–489.

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

[47] G. Krause, J. Protze, J. Piontek, Assembly and function of claudins: structure-function relationships based on homology models and crystal structures, Semin. Cell Dev. Biol. 42 (2015) 3–12.

[48] H. Suzuki, T. Nishizawa, K. Tani, Y. Yamazaki, A. Tamura, R. Ishitani, N. Dohmae, S. Tsukita, O. Nureki, Y. Fujiyoshi, Crystal structure of a claudin provides insight into the architecture of tight junctions, Science 344 (2014) 304–307.

[49] J. Wu, Z. Yan, Z. Li, C. Yan, S. Lu, M. Dong, N. Yan, Structure of the voltage-gated calcium channel Cav1.1 complex, Science 350 (2015) (aad2395).

[50] W.A. Ayad, D. Locke, I.V. Koreen, A.L. Harris, Heteromeric, but not homomeric, connexin channels are selectively permeable to inositol phosphates, J. Biol. Chem. 281 (2006) 16727–16739.

[51] N. Bonander, M. Jamshad, D. Oberthur, M. Clare, J. Barwell, K. Hu, M.J. Farquhar, Z. Stamataki, H.J. Harris, K. Dierks, T.R. Dafforn, C. Betzel, J.A. McKeating, R.M. Bill, Production, purification and characterization of recombinant, full-length human claudin-1, PLoS One 8 (2013), e64517. .

[52] M.M. Falk, Cell-free synthesis for analyzing the membrane integration, oligomerization, and assembly characteristics of gap junction connexins, Methods 20 (2000) 165–179.

[53] J. Hou, D.A. Goodenough, Claudin-16 and claudin-19 function in the thick ascending limb, Curr. Opin. Nephrol. Hypertens. 19 (2010) 483–488.

[54] G. McCaffrey, W.D. Staatz, C.A. Quigley, N. Nametz, M.J. Seelbach, C.R. Campos, T.A. Brooks, R.D. Egleton, T.P. Davis, Tight junctions contain oligomeric protein assembly critical for maintaining blood-brain barrier integrity in vivo, J. Neurochem. 103 (2007) 2540–2555.

[55] G. McCaffrey, C.L. Willis, W.D. Staatz, N. Nametz, C.A. Quigley, S. Hom, J.J. Lochhead, T.P. Davis, Occludin oligomeric assemblies at tight junctions of the blood-brain barrier are altered by hypoxia and reoxygenation stress, J. Neurochem. 110 (2009) 58–71.

[56] S. Milatz, J. Piontek, J.D. Schulzke, I.E. Blasig, M. Fromm, D. Gunzel, Probing the cis-arrangement of prototype tight junction proteins claudin-1 and claudin-3, Biochem. J. 468 (2015) 449–458.

[57] A. Oshima, T. Matsuzawa, K. Nishikawa, Y. Fujiyoshi, Oligomeric structure and functional characterization of *Caenorhabditis elegans* innexin-6 gap junction protein, J. Biol. Chem. 288 (2013) 10513–10521.

[58] L.A. Stebbings, M.G. Todman, P. Phelan, J.P. Bacon, J.A. Davies, Two Drosophila innexins are expressed in overlapping domains and cooperate to form gap-junction channels, Mol. Biol. Cell 11 (2000) 2459–2470.

[59] O. Ogun, M. Zallocchi, Clarin-1 acts as a modulator of mechanotransduction activity and presynaptic ribbon assembly, J. Cell Biol. 207 (2014) 375–391.

[60] J. Reiners, K. Nagel-Wolfrum, K. Jurgens, T. Marker, U. Wolfrum, Molecular basis of human Usher syndrome: deciphering the meshes of the Usher protein network provides insights into the pathomechanisms of the Usher disease, Exp. Eye Res. 83 (2006) 97–119.

[61] J.B. Phillips, H. Vastinsalo, J. Wegner, A. Clement, E.M. Sankila, M. Westerfield, The cone-dominant retina and the inner ear of zebrafish express the ortholog of CLRN1, the causative gene of human Usher syndrome type 3A, Gene Expr. Patterns 13 (2013) 473–481.