J. Membrane Biol. 194, 59–76 (2003)
DOI: 10.1007/s00232-003-2026-8

The Journal of
Membrane
Biology



# Sequence and Phylogenetic Analyses of 4 TMS Junctional Proteins of Animals: Connexins, Innexins, Claudins and Occludins

**V.B. Hua[1], A.B. Chang[1], J.H. Tchieu[1], N.M. Kumar[2,*], P.A. Nielsen[2], M.H. Saier Jr[1]**
[1]Division of Biology, University of California at San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0116, USA
[2]Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

Received: 21 December 2002/Revised: 3 April 2003

**Abstract.** Connexins and probably innexins are the principal constituents of gap junctions, while claudins and occludins are principal tight junctional constituents. All have similar topologies with four α-helical transmembrane segments (TMSs), and all exhibit well-conserved extracytoplasmic cysteines that either are known to or potentially can form disulfide bridges. We have conducted sequence, topological and phylogenetic analyses of the proteins that comprise the connexin, innexin, claudin and occludin families. A multiple alignment of the sequences of each family was used to derive average hydropathy and similarity plots as well as phylogenetic trees. Analyses of the data generated led to the following evolutionary and functional suggestions: (1) In all four families, the most conserved regions of the proteins from each family are the four TMSs although the extracytoplasmic loops between TMSs 1 and 2, and TMSs 3 and 4 are usually well conserved. (2) The phylogenetic trees revealed sets of orthologues except for the innexins where phylogeny primarily reflects organismal source, probably due to a lack of relevant organismal sequence data. (3) The two halves of the connexins exhibit similarities suggesting that they were derived from a common origin by an internal gene duplication event. (4) Conserved cysteyl residues in the connexins and innexins may point to a similar extracellular structure involved in the docking of hemichannels to create intercellular communication channels. (5) We suggest a similar role in homomeric interactions for conserved extracellular residues in the claudins and occludins. The lack of sequence or motif similarity between the four different families indicates that, if they did evolve from a common ancestral gene, they have diverged considerably to fulfill separate, novel functions. We suggest that internal duplication was a general evolutionary strategy used to generate new families of channels and junctions with unique functions. These findings and suggestions should serve as guides for future studies concerning the structures, functions and evolutionary origins of junctional proteins.

**Key words:** Intercellular communication — Gap junctions — Tight junctions — Connexins — Innexins — Claudins — Occludins — Evolution

## Introduction

Gap junctions, found in the plasma membranes of vertebrate animal cells, consist of clusters of closely packed transmembrane channels, the connexons, in which the principal proteins are referred to as connexins (Beyer et al., 1987; Loewenstein, 1987; Kumar & Gilula, 1996; Harris, 2001; Shibata et al., 2001; Evans & Martin, 2002a; Hand et al., 2002). Topologically related putative gap junctional proteins found in both invertebrates and vertebrates exhibiting little or no significant sequence similarity to the connexins are called innexins (White & Paul, 1999; Phelan & Starich, 2001; Potenza et al., 2002). Connexins and innexins comprise two distinct protein families whose structures and functions have been suggested to be overlapping (Curtin et al., 1999; Ganfornina et al., 1999; Landesman et al., 1999; White & Paul, 1999; Stebbings et al., 2000).

Gap junctional complexes provide direct electrical coupling and metabolic communication by allowing the free flow of ions and other small molecules

*Correspondence to:* M.H. Saier Jr; email: msaier@ucsd.edu

**Present address for N.M.K.:* Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, 1905 West Taylor Street, Chicago, IL 60612-7243

between neighboring cells (Bevans et al., 1998; Kim et al., 1999; Landesman et al., 1999). They play important roles in a variety of pathological conditions such as congenital deafness (Kitamura et al., 2000; D'Andrea et al., 2002), convulsive seizures (Jahromi et al., 2002), congenital cataracts (Mackay et al., 1999), erythrokeratodermia variabilis (Richard et al., 1998), and Charcot-Marie tooth disease (Omori et al., 1996). Their dynamic assembly (Lopez et al., 2001; Evans & Martin, 2002b) and regulation by ATP and protein kinases (Ghosh et al., 2002) and by $Ca^{2+}$ and calmodulin (Sotkis et al., 2001) are complex. Vertebrate connexons consist of homo- and heterohexameric arrays of connexins, and the connexon in one plasma membrane docks end to end with a connexon in the membrane of a closely opposed cell (Yeager et al., 1998; Unger et al., 1999; Delmar, 2002). Although invertebrate innexins have been much less studied, both *Drosophila* and *C. elegans* innexins have multiple paralogues, some of which have been studied with respect to their capacity to form intercellular channels (Starich et al., 2002; Stebbings et al., 2002). Recently, innexins have been proposed to have orthologues in vertebrates based on sequence similarity (Panchin et al., 2000) although this has not been confirmed by functional studies.

Tight junctions, also found in the plasma membranes of animal cells, form charge-selective paracellular diffusion barriers that regulate the diffusion of small molecules across epithelial and endothelial cell sheets and serve as major cell adhesion molecules (Balda et al., 2000; Tsukita & Furuse, 2000; Blaschuk et al., 2002; Colegio et al., 2002; D'Atri & Citi, 2002). They also prevent the intermixing of apical and basolateral proteins, especially in the extracytoplasmic leaflet of these membranes (Tsukita & Furuse, 2002). Protein constituents of the tight junction include the claudins and the occludins (Tsukita & Furuse, 2000; Heiskala et al., 2001; Kollmar et al., 2001; D'Atri & Citi, 2002; Langbein et al., 2002). These oligomeric transmembrane proteins are regulated by phosphorylation (Cordenonsi et al., 1999). Like connexins, but unlike innexins in this regard, occludins are found in vertebrate animals. Claudins may be found in both vertebrates and invertebrates (Ando-Akatsuka et al., 1996; *see* below). Evidence suggests that claudins and occludins cooperate in the regulation of paracellular permeability (Balda et al., 2000; Morcos et al., 2001). As is well established for the connexins, claudins are differentially synthesized in various tissue and cell types (Kiuchi-Saishin et al., 2002). Interestingly, some of the claudins have been shown to secondarily serve as receptors for *Clostridium perfringens* enterotoxin (McClane, 2000; Long et al., 2001). Occludin isoforms of altered structure are synthesized in variable amounts, depending on conditions, and these isoforms may contribute to the regulation of occludin function (Ghassemifar et al., 2002).

Connexins, innexins, claudins and occludins share certain structural features but also exhibit distinctive characteristics. All four of these protein types exhibit four putative transmembrane α-helical spanners (TMS). They vary in size between about 20 kDa and 60 kDa with overlapping size variation within each of these four protein families (*see* below). Three-dimensional structural data are available for connexon membrane channels (Unger et al., 1999). Electron density analyses of the dodecameric channels, formed by end-to-end docking of two hexamers with a total of 48 TMSs, are consistent with an α-helical configuration for all four TMSs of each connexin subunit (Unger et al., 1999). The extracellular vestibule forms a tight seal to prevent the exchange of substances with the extracellular milieu.

We have identified all currently available homologues of the connexins, innexins, occludins and claudins in the publicly available databases using BLAST search tools. These searches were initially conducted in January, 2002, but the tabulations have been updated. However, the analyses reported were conducted with the family members available when the analyses were conducted. The sequences of the proteins in these four families were multiply aligned, and the alignments were used to generate average hydropathy, amphipathicity and similarity plots. Phylogenetic trees were constructed allowing definition of the sequence relatedness of proteins within each of these four families. The reported results not only define the current members of these four families of (putative) junctional proteins, they also allow predictions regarding the evolutionary origins of some of them. Thus, we can predict (1) which proteins are orthologues (having arisen in different species exclusively by speciation), (2) which proteins are recent versus early diverging paralogues (homologues that arose by gene duplication in a single organism), and (3) what the relative rates of sequence divergence were for different orthologous sets. We suggest that although these protein families do not exhibit significant sequence or motif similarity, the evolutionary precursor of the connexins and the innexins might have been the same. The same is possible for the claudins and occludins. We consider the possibility that at least some of these junctional proteins arose by an internal gene duplication event in which one or more 2-TMS-encoding genetic element(s) gave rise to the present-day 4-TMS-encoding gene. This hypothesis presupposes that this duplication event occurred more than once during the evolution of these protein families. Internal duplication may be a general evolutionary strategy that has been used to generate new families of channels and junctions with unique functions (Saier, 2000, 2001).

**Table 1.** Sequenced proteins of the connexin family[1]

| Abbreviation (based on gene symbol) | alternative Abbreviation (based on protein size) | Organism | Size[2] | Accession # |
|---|---|---|---|---|
| BT-α1 | Cx43 | *Bos taurus* (cow) | 383 | P18246 |
| BT-α3 | Cx44 | *Bos taurus* | 402 | P41987 |
| BT-β1 | Cx32 | *Bos taurus* | 284 | O18968 |
| CF-α5 | Cx40 | *Canis familiaris* (dog) | 357 | P33725 |
| CF-α7 | Cx32 | *Canis familiaris* | 396 | P28228 |
| CM-β1 | Cx31.5 | *Chrysophrys major* (red sea bream) | 275 | BAA90669 |
| DA-α1 | Cx43 | *Devario aequipinnatus* (fish) | 382 | AAC19098 |
| DR-α1 | Cx43 | *Danio rerio* (zebrafish) | 381 | O57474 |
| DR-α7 | Cx43.4 | *Danio rerio* | 380 | Q92052 |
| DR-44.2 | Cx44.2 | *Danio rerio* | 391 | AAD42022 |
| GG-α1 | Cx43 | *Gallus gallus* (chicken) | 381 | P14154 |
| GG-α3 | Cx56 | *Gallus gallus* | 510 | P29415 |
| GG-α5 | Cx42 | *Gallus gallus* | 369 | P18860 |
| GG-α7 | Cx45 | *Gallus gallus* | 394 | P18861 |
| GG-α8 | Cx45.6 | *Gallus gallus* | 400 | P36381 |
| GG-β2 | Cx31 | *Gallus gallus* | 263 | AAC64043 |
| HS-α1 | Cx43 | *Homo sapiens* (human) | 382 | NP_000156 |
| HS-α3 | Cx46 | *Homo sapiens* | 435 | AAD42925 |
| HS-α4 | Cx37 | *Homo sapiens* | 333 | NP_002051 |
| HS-α5 | Cx40 | *Homo sapiens* | 358 | NP_005257 |
| HS-α7 | Cx45 | *Homo sapiens* | 396 | NP_005488 |
| HS-α8 | Cx50 | *Homo sapiens* | 433 | AAF32309 |
| HS-α9 | Cx36 | *Homo sapiens* | 321 | AAD54234 |
| HS-α10 | Cx59 | *Homo sapiens* | 515 | AAG09406 |
| HS-α11 | Cx31.9 | *Homo sapiens* | 294 | AAM53649 |
| HS-α12 | Cx47 | *Homo sapiens* | 431 | AAB94511 |
| HS-α13 | Cx62 | *Homo sapiens* | 543 | AAK51676 |
| HS-β1 | Cx32 | *Homo sapiens* | 283 | NP_000157 |
| HS-β2 | Cx26 | *Homo sapiens* | 226 | NP_003995 |
| HS-β3 | Cx31 | *Homo sapiens* | 270 | O75712 |
| HS-β4 | Cx30.3 | *Homo sapiens* | 266 | CAB90270 |
| HS-β5 | Cx31.1 | *Homo sapiens* | 273 | AAD18005 |
| HS-β6 | Cx30 | *Homo sapiens* | 261 | NP_006774 |
| HS-β7 (HH-25) | Cx25 | *Homo sapiens* | 223 | CAC93845 |
| HS-ε1 | Cx31.3 | *Homo sapiens* | 279 | AAM21145 |
| HS-25 | Cx25 | *Homo sapiens* | 223 | CAC93845 |
| HS-37[3] | Cx37 | *Homo sapiens* | 293 | AAD56533 |
| HS-40.1 | Cx40.1 | *Homo sapiens* | 370 | CAC93846 |
| MA-α9 | Cx35 | *Morone americana* (white perch) | 304 | AAC31884 |
| MA-a9′ | Cx34.7 | *Morone americana* | 306 | AAC31885 |
| MM-α1 | Cx43 | *Mus musculus* (mouse) | 382 | AAA53027 |
| MM-α3 | Cx46 | *Mus musculus* | 417 | Q64448 |
| MM-α4 | Cx37 | *Mus musculus* | 333 | NP_032146 |
| MM-α5 | Cx40 | *Mus musculus* | 358 | NP_032147 |
| MM-α6 | Cx33 | *Mus musculus* | 283 | XP_284759 |
| MM-α7 | Cx45 | *Mus musculus* | 396 | NP_032148 |
| MM-α8 | Cx50 | *Mus musculus* | 440 | NP_032149 |
| MM-α9 | Cx36 | *Mus musculus* | 321 | NP_034420 |
| MM-α11 | Cx30.2 | *Mus musculus* | 278 | AAN65188 |
| MM-α12 | Cx47 | *Mus musculus* | 437 | CAC19434 |
| MM-α13 | Cx57 | *Mus musculus* | 505 | NP_034419 |
| MM-β1 | Cx32 | *Mus musculus* | 283 | P28230 |
| MM-β2 | Cx26 | *Mus musculus* | 226 | NP_032151 |
| MM-β3 | Cx31 | *Mus musculus* | 270 | NP_032152 |
| MM-β4 | Cx30.3 | *Mus musculus* | 266 | NP_032153 |
| MM-β5 | Cx31.1 | *Mus musculus* | 271 | NP_034421 |
| MM-β6 | Cx30 | *Mus musculus* | 261 | NP_032154 |
| MM-ε1 | Cx29 | *Mus musculus* | 258 | CAC29245 |
| MU-α3 | Cx32.2 | *Micropogonias undulatus* (Atlantic croaker) | 285 | P51915 |
| MU-α3′ | Cx32.7 | *Micropogonias undulatus* | 283 | P51916 |
| OA-α3 | Cx44 | *Ovis aries* (sheep) | 413 | AAD56220 |
| OA-α8 | Cx49 | *Ovis aries* | 440 | AAF01367 |

*continued on next page*

Table 1. Continued

| Abbreviation (based on gene symbol) | alternative Abbreviation (based on protein size) | Organism | Size[2] | Accession # |
|---|---|---|---|---|
| OA-β2 | Cx26 | *Ovis aries* | 226 | P46691 |
| RN-α1 | Cx43 | *Rattus norvegicus* (rat) | 382 | NP_036699 |
| RN-α3 | Cx46 | *Rattus norvegicus* | 416 | P29414 |
| RN-α4 | Cx37 | *Rattus norvegicus* | 333 | Q03190 |
| RN-α5 | Cx40 | *Rattus norvegicus* | 356 | P28234 |
| RN-α9 | Cx36 | *Rattus norvegicus* | 321 | CAA76528 |
| RN-β1 | Cx32 | *Rattus norvegicus* | 283 | P08033 |
| RN-β2 | Cx26 | *Rattus norvegicus* | 226 | P21994 |
| RN-β3 | Cx31 | *Rattus norvegicus* | 270 | P25305 |
| RN-β4 | Cx30.3 | *Rattus norvegicus* | 265 | P36380 |
| RN-β5 | Cx31.1 | *Rattus norvegicus* | 271 | P28232 |
| RN-β6 | Cx30 | *Rattus norvegicus* | 261 | AAD50911 |
| RN-33 | Cx33 | *Rattus norvegicus* | 286 | P28233 |
| RO-α9 | Cx35 | *Raja ocellata* (skate) | 302 | Q92107 |
| XL-α1 | Cx43 | *Xenopus laevis* (frog) | 379 | P16863 |
| XL-α2 | Cx38 | *Xenopus laevis* | 334 | P16864 |
| XL-α4 | Cx41 | *Xenopus laevis* | 352 | P51914 |
| XL-β1 | Cx30 | *Xenopus laevis* | 264 | P08983 |

[1]Since the completion of this work, several new connexins have been discovered. Several of these have been included in the table although they are not included in our analyses.
[2]Size of the proteins is expressed in numbers of amino-acyl residues (# aas) in this and subsequent tables in this paper.
[3]HS-37 is a polymorphic α4 variant.

## Results

### Connexins

Table 1 presents the sequenced connexin homologues we have identified from publicly available databases. All contain four transmembrane regions and are derived exclusively from vertebrates including mammals, birds, fish and amphibians.

Several organisms exhibit multiple paralogues. For example, six chicken paralogues, 12 rat paralogues, 14 mouse paralogues and 21 human paralogues are listed in Table 1. Because these proteins often do not exhibit sequence relationships suggestive of orthology with proteins from other organisms (*see* below), mammals, and possibly birds, may have as many as 22–24 connexin paralogues. However, one or more of these may be pseudogenes. Recently, the human genome was reported to contain 20 connexin paralogues as determined from genomic databases from Celera and NIH (Eiberger et al., 2001; Willecke et al., 2002). These are the same as the 20 sequence-divergent full-length human paralogues we report here.

Connexins tabulated in Table 1 are reported to be maximally 542 and minimally 223 amino-acyl residues (aas) in length. Because several of the largest and smallest proteins are found with comparable sizes, connexins probably exhibit just slightly greater than a 2× size variation.

The proteins listed in Table 1 were aligned using the CLUSTAL X program (Thompson et al., 1997). The complete multiple alignment (available on our ALIGN website; www-biology.ucsd.edu/~msaier/transport/)[1] revealed that most of the size variation observed for these proteins occurred in their C-terminal regions and the single cytoplasmic loop between the second and third TMSs. The 4-TMS topology, originally deduced using site-directed antibody localization approaches (Milks et al., 1988; Yeager et al., 1998), and confirmed and extended by electron density analyses (Unger et al., 1999) is now well established. Both of the variable regions cited above are located intracellularly. Thus, residue positions 1–110 are well conserved; positions 121–200 are poorly conserved; positions 201–300 are well conserved, and the remaining residue positions of the alignment are poorly conserved. In the first well-conserved region (alignment positions 56–80), the following consensus motif was identified:

$$\overset{*}{C}\ N\ T\ X\ \overset{*}{Q}\ \overset{*}{P}\ \overset{*}{G}\ \overset{*}{C}\ X\ N\ V\ \overset{*}{C}\ Y\ \overset{*}{D}\ X_2\ F$$

$$\overset{*}{P}\ I\ \overset{*}{S}\ H\ (I/V)\ \overset{*}{R}\ (F/Y/L)\ \overset{*}{W}$$

[X = any residue; alternative residues at a single alignment position are indicated in parentheses; *: a fully conserved position]

[1]Figures on the website: (www-biology.ucsd.edu/~msaier/transport/); Fig. S1 Multiple alignment of all connexins; Fig. S2 Multiple alignment of the 22 human connexins; Fig. S3 Phylogenetic tree of the 22 human connexins; Fig. S4 Multiple alignment of all innexins; Fig. S5 Multiple alignment of all claudins; Fig. S6 Multiple alignment of all occludins.

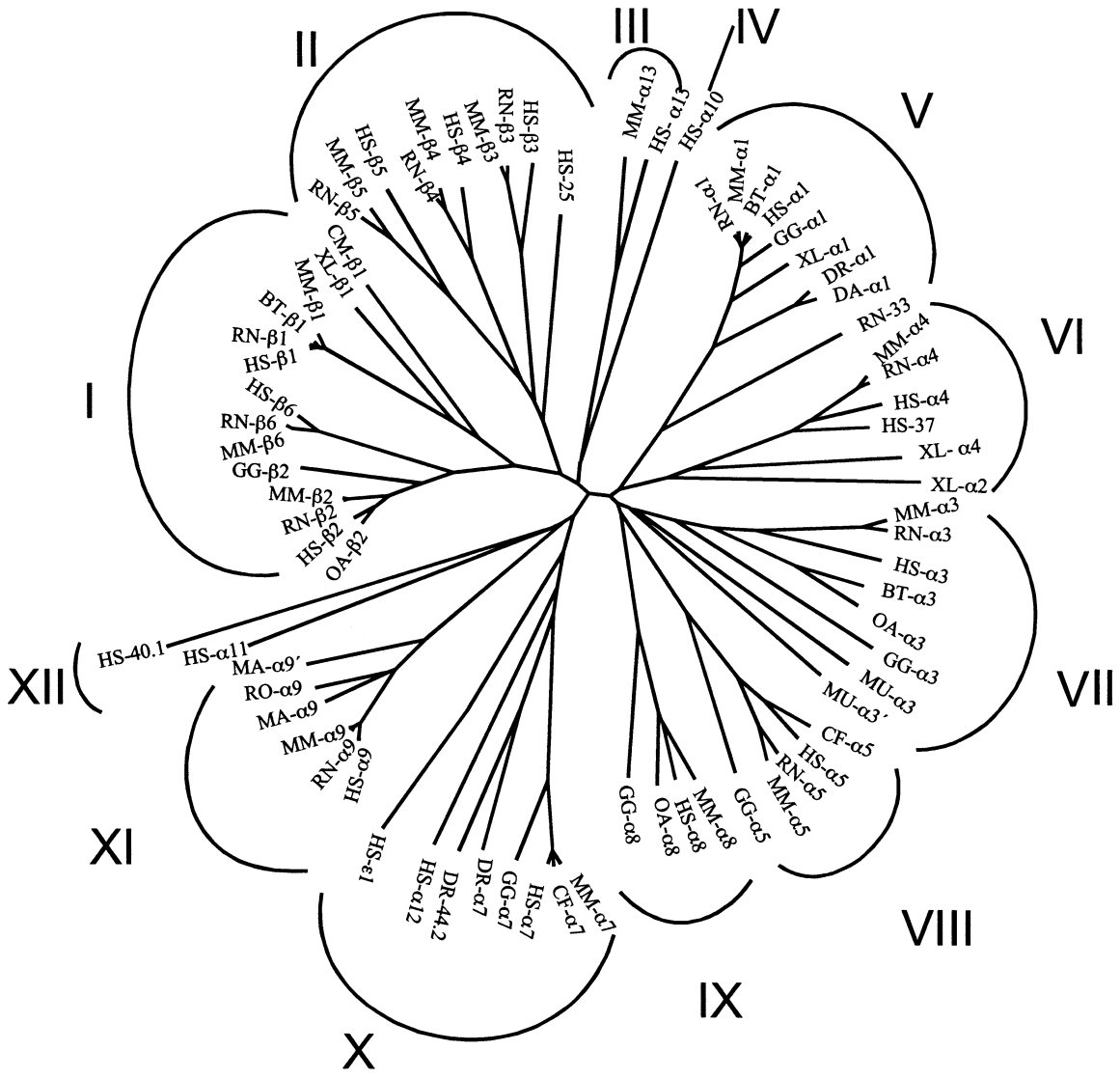


**Fig. 1.** Phylogenetic trees for the complete connexin family. Protein abbreviations are as indicated in Table 1. The Clustal X program (Thompson et al., 1997) was used to derive this tree and all other trees presented here and on our website. *See* text for explanation of the clustering patterns. The multiple alignment for all connexins is shown on our website (www-biology.ucsd.edu/~msaier/transport/; Fig. S1). That for the 22 human proteins is shown in website Fig. S2, and the tree for the human proteins is shown in Fig. S3.

All of these residues are in the extracellular loop between TMSs 1 and 2.

In a second well-conserved region, a less well-conserved cysteine-rich motif was identified. This motif occurs at alignment positions 246–269 as follows:

$$C - X_3\ P\ \overset{*}{C}\ P\ X\ X\ V\ D\ \overset{*}{C}\ F\ V\ S\ R$$

$$\overset{*}{P}\ T\ \overset{*}{E}\ \overset{*}{K}\ T\ I/V\ F$$

[–: a one-residue gap in the alignment of most proteins.]

Three orthologous connexins, connexin β3 of the mouse, rat and human, display an additional residue at alignment position 247 corresponding to the gap (–). The best signature sequence for the connexin family (alignment positions 56–80) corresponding to the first conserved motif (*see* above) is:

$$C\ X\ (T\ S)\ X\ Q\ P\ G\ C\ X_3\ C\ (Y\ F)\ D\ X_3$$
$$P(L\ I\ V)\ S\ X\ (L\ I\ V\ Y)\ R\ (F\ Y\ L)\ W$$

The connexin phylogenetic tree, based on the complete multiple alignment presented on our website (Fig. S1), is shown in Fig. 1, and the corresponding tree for the human proteins, based on the alignment shown in Fig. S2, is shown on our website

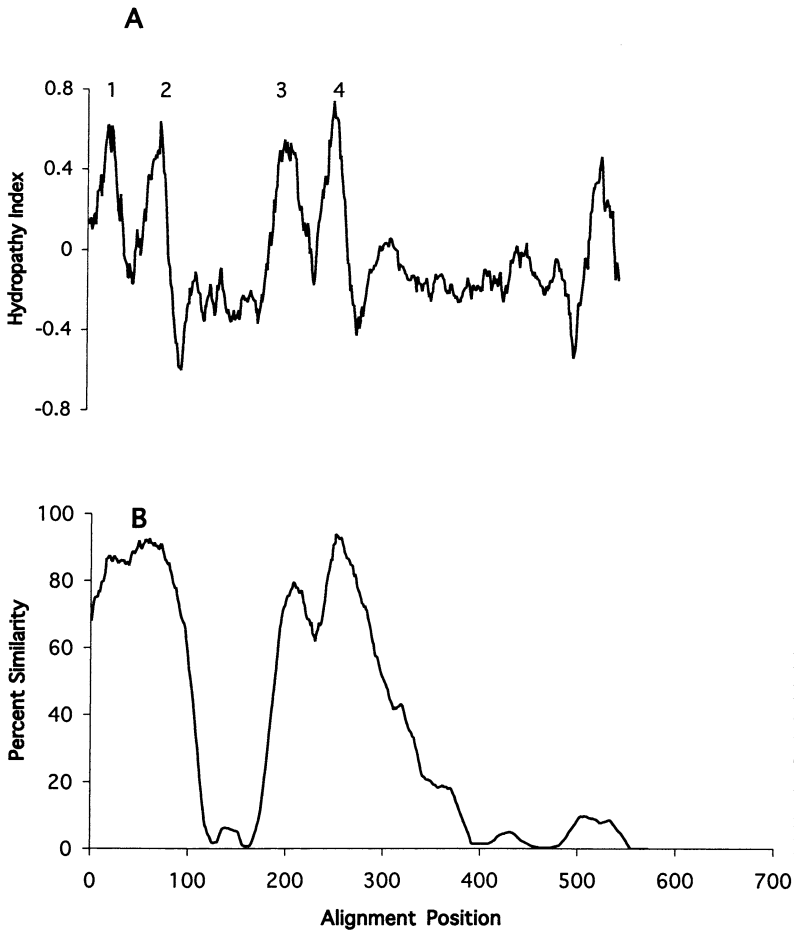


**Fig. 2.** Average hydropathy (*A*) and similarity (*B*) plots for the connexins. Proteins used for this study are the 19 sequence-divergent proteins included in the two partial multiple alignments shown in Fig. 3. The AveHAS program (Zhai & Saier, 2001) was used for both plots with a sliding window of 21 residues. Hydropathy values were those used by Kyte and Doolittle (1982).

in Fig. S3. The proteins fall into 12 clusters that branch from points near the center of the unrooted tree as indicated by the roman numerals (I–XII). Human proteins are found in all 12 of these clusters, and four of the clusters include only mammalian proteins. Sequences from birds (the chicken) appear in six clusters; those from fish are found in five clusters, and those from amphibians are found in two clusters. The absence in these organisms of several of the connexin paralogues found in mammals may reflect a deficiency of sequence data. The configuration of the tree leads to the suggestion that most (but not all) of the sequence divergence observed for the connexins arose due to fairly early gene duplication events prior to divergence of most of the vertebrate species represented.

The six clusters that include both mammalian and avian proteins reveal that in each cluster, the avian protein is more distant from the mammalian proteins than the latter are from each other. In all six cases it can be concluded that the chicken protein is orthologous to the mammalian proteins. Similarly, in the clusters including both mammalian and fish or amphibian proteins, the fish or amphibian proteins are always more distant from the mammalian and avian proteins than the latter are from each other. These observations provide evidence regarding potential orthologous relationships. They reveal that while the major clusters arose by fairly early gene duplication events, several late gene duplication events gave rise to similar sequence paralogues that cluster together. Thus, sets of orthologues as well as non-orthologous proteins can be visualized.

In almost all cases, a single human connexin is present in each set of mammalian orthologues. Cluster I includes three sets of probable orthologues (β1, β2 and β6), and of these, an avian protein is associated with one of them, while both fish and amphibian proteins are associated with another. Cluster II includes four sets of mammalian orthologues (β3, β4, β5 and HS-25). Clusters III and IV include exclusively mammalian proteins, and the two deep-rooted branches each bears only a single human protein. Cluster V includes one human protein (α1) and potential orthologues from other mammals, the chicken, the frog and fish, but surprisingly, one distant rat homologue (RN-33) that has no recognized human counterpart is found in this cluster. Cluster VI consists of one mammalian cluster (α4) with two human homologues (α4 and HS-37) and two associ-


**Fig. 3.** Alignments of the two well-conserved regions of 19 sequence-divergent connexins. Residues comprising the two putative TMSs in each alignment are presented in bold print, as are the three fully conserved cysteyl residues in each of the two inter-TMS loop regions. Fully conserved residues are indicated by a line adjacent to the lower right of the one-letter abbreviation of the amino acid. To be noted are the facts that the TMSs and two of the three fully conserved cysteyl residues align in the top and the bottom figures. The asterisk between the fully conserved Y and the largely conserved G in the lower alignment is the site of single amino-acyl residue insertions in three of these proteins.

ated distant frog proteins (XL-α4 and XL-α2). Based on the phylogenetic tree, at least one of these frog proteins (α2) is not likely to have a mammalian counterpart, possibly due to a unique function in *Xenopus* oocytes. Cluster VII consists of a single mammalian/avian cluster (α3) with two loosely associated fish proteins, both from the Atlantic croaker. As for the two frog proteins in cluster VI, at least one of these fish proteins probably lacks a mammalian counterpart. Clusters VIII (α5) and IX (α8) both include mammalian and avian proteins, but cluster X consists of a single mammalian/avian cluster (α7) with two distantly related human paralogues and two loosely associated fish proteins. Cluster XI consists of a single mammalian cluster (α9) with three related fish homologues, two of which are from the white perch. Finally, cluster XII consists of two distantly related human proteins with orthologues from the mouse that were revealed after this work was completed (*see* Footnote 1 to Table 1).

Further analysis of the tree shown in Fig. 1 revealed that some of the clusters of mammalian/avian orthologues have undergone very little sequence divergence, while others have undergone much more. For example, the α1 orthologues in cluster V exhibit minimal sequence divergence, while the α3 orthologues in cluster VII exhibit maximal divergence. The proteins in other probable orthologous clusters have diverged at intermediate rates. The results clearly suggest that all of the chicken homologues are orthologous to proteins in mammals, but that some of the fish and frog proteins lack mammalian orthologues. The human paralogues exhibit the phylogenetic relationships shown in Fig. S3 (*see* our ALIGN website). All relationships are in accord with those presented in Fig. 1.

Average hydropathy, average similarity and average amphipathicity (angle set at 100° for an α-helix) plots were derived using a sliding window of 21 residues (Kyte & Doolittle, 1982; Le et al., 1999; Zhai & Saier, 2001). The former two plots are presented in Fig. 2*A* and *B*, respectively. Four clear peaks of hydrophobicity are apparent, the first pair separated from the second pair by a poorly conserved hydro-

**Table 2.** Sequenced proteins of the innexin family

| Abbreviation | Database name or description | Organism | Size[1] | Accession or GI#[2] |
|---|---|---|---|---|
| CE-Unc | unc-7 protein | *Caenorhabditis elegans* (worm) | 522 | Q03412 |
| CE-T | Transmembrane protein | *Caenorhabditis elegans* | 428 | AAB09671 |
| CE-Unc2 | unc-9 | *Caenorhabditis elegans* | 386 | AAB51534 |
| CE-Unc3 | Similar to *C. elegans* unc-7 and *Drosophila* passover gene | *Caenorhabditis elegans* | 385 | AAB95049 |
| CE-Unc4 | Similar to *C. elegans* UNC-7 | *Caenorhabditis elegans* | 408 | AAB93310 |
| CE-Em | Embryonic membrane protein | *Caenorhabditis elegans* | 420 | AAB09670 |
| CE-Unc5 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 420 | AAC17640 |
| CE-O1 | Similar to ogre | *Caenorhabditis elegans* | 465 | CAA99940 |
| CE-Unc6 | Similar to *C. elegans* unc-7 and *Drosophila* ogre and shaking-b | *Caenorhabditis elegans* | 559 | AAA83332 |
| CE-Unc7 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 404 | AAC17026 |
| CE-Unc8 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 522 | CAA92633 |
| CE-Eat | eat-5 | *Caenorhabditis elegans* | 423 | AAB09669 |
| CE-Pfam1 | Similar to Pfam domain | *Caenorhabditis elegans* | 475 | AAC69093 |
| CE-P1 | Similar to the *Drosophila* passover gene | *Caenorhabditis elegans* | 378 | AAC16426 |
| CE-F08G12.10 | F08G12.10 | *Caenorhabditis elegans* | 447 | CAB54206 |
| CE-Unc9 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 317 | CAA92634 |
| CE-Pfam2 | Similar to Pfam domain | *Caenorhabditis elegans* | 556 | AAA83313 |
| CE-Unc10 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 362 | AAC17025 |
| CE-Unc11 | Similar to *C. elegans* unc-7 | *Caenorhabditis elegans* | 389 | CAB60997 |
| CE-P2 | Similar to *Drosophila* passover and ogre | *Caenorhabditis elegans* | 382 | AAC17030 |
| CE-O2 | Similar to ogre | *Caenorhabditis elegans* | 434 | CAA96621 |
| CE-O3 | Similar to ogre | *Caenorhabditis elegans* | 392 | CAB05813 |
| CE-Unc9-1 | Similar to *C. elegans* unc-9 | *Caenorhabditis elegans* | 508 | AAF60675 |
| CE-Unc9-2 | Similar to *C. elepans* unc-9 | *Caenorhabditis elegans* | 526 | AAF60654 |
| CE-HP1 | Hypothetical protein F26D11.10 | *Caenorhabditis elegans* | 554 | T33294 |
| CE-HP2 | Hypothetical protein R12H7.1 | *Caenorhabditis elegans* | 409 | T24203 |
| CL-gjp | Putative gap junction protein pannexin | *Clione limacina* (naked sea butterfly) | 426 | AAF75839 |
| DM-OLP | Ogre locus | *Drosophila melanogaster* (fly) | 362 | P27716 |
| DM-ShakB | shak-b (lethal) protein | *Drosophila melanogaster* | 372 | AAB34769 |
| DM-P | Passover gene | *Drosophila melanogaster* | 361 | 1095426 |
| DM-GJP1 | Gap junction protein prp33 | *Drosophila melanogaster* | 367 | AAD50378 |
| DM-GJP2 | Gap junction protein prp7 | *Drosophila melanogaster* | 438 | AAD50379 |
| DM-GP1 | Cg1448 gene product | *Drosophila melanogaster* | 395 | AAF56822 |
| DM-GP2 | Cg10125 gene product | *Drosophila melanogaster* | 367 | AAF50655 |
| DM-GP3 | inx7 gene product | *Drosophila melanogaster* | 481 | AAF50922 |
| DM-GP4 | Cg7537 gene product | *Drosophila melanogaster* | 428 | AAF48923 |
| GT-gjp | Putative gap junction protein pannexin | *Girardia tigrina* (flatworm) | 439 | AAF75840 |
| HS-Pan1 | Pannexin 1; mrs 1 protein | *Homo sapiens* (humans) | 357 | 14794511 |
| MM-Pan1 | Pannexin 1 | *Mus musculus* (mouse) | 448 | 9506951 |
| SA-GJP1 | Invertebrate gap junction protein | *Schistocerca americana* (grasshopper) | 361 | AAD29305 |
| SA-GJP2 | Invertebrate gap junction protein | *Schistocerca americana* | 359 | AAD29306 |

[1]Size of the proteins is expressed in numbers of amino-acyl residues (# aas) in this and subsequent tables in this paper.
[2]GI #, Genbank index number.

philic region of variable length (residue positions 100–190). A second variable hydrophilic region follows the fourth putative TMS (residue positions 300–550). As seen in the average similarity plot (Fig. 2*B*), not only the four TMSs, but also the extracellular loops connecting TMSs 1 and 2, and TMSs 3 and 4 are well conserved. All cytoplasmically localized hydrophilic regions are poorly conserved. Interestingly, TMSs 1 and 2 and the intervening extracytoplasmic loop are much better conserved than TMSs 3 and 4 and the intervening loop. This fact clearly suggests that while TMSs 1 and 2 serve an important and universal functional role, TMSs 3 and 4 are either less important or provide functions that differ for different protein members of the family, e.g., such as forming the lining of the channel pore. The average amphipathicity plot was uninformative and is therefore not presented.

For further similarity analyses, 19 sequence divergent proteins from all of the 12 clusters shown in Fig. 1 were selected for construction of a multiple alignment using the TREE program (Feng & Doo-

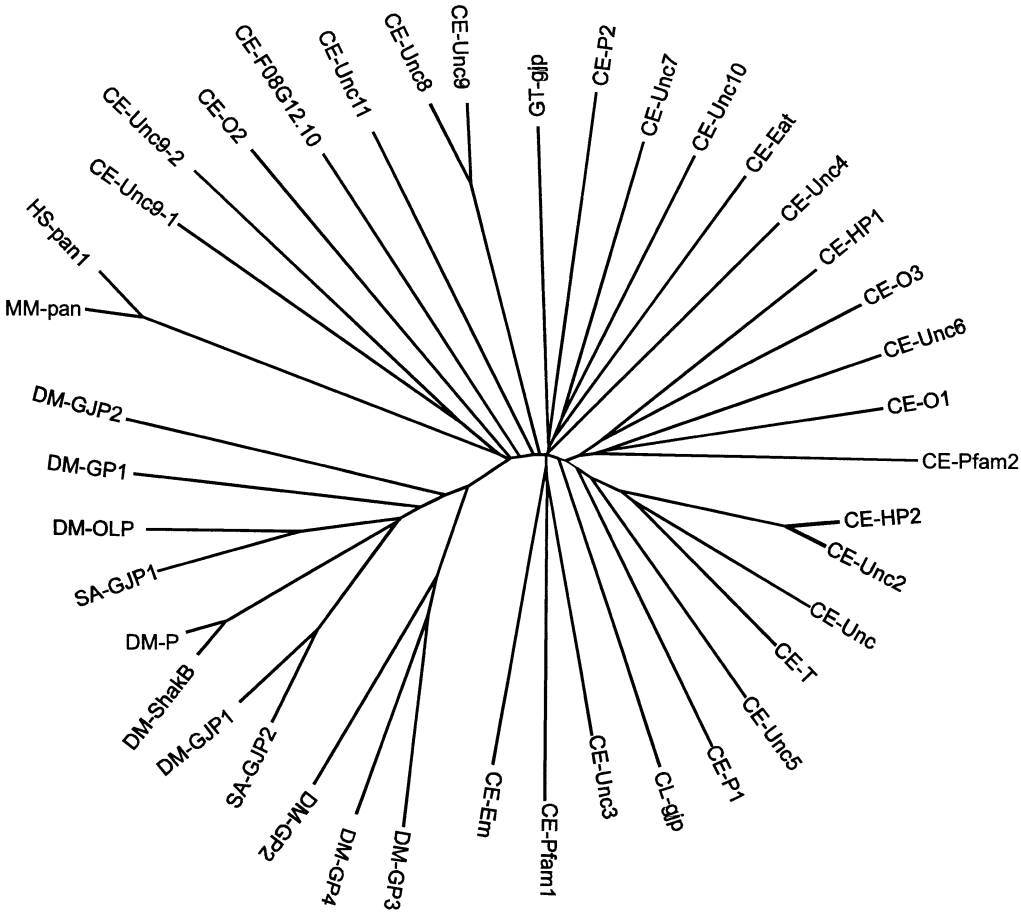


**Fig. 4.** Phylogenetic tree for the innexin protein family. Abbreviations of the proteins are as indicated in Table 2. Format of presentation and the program used were the same as described in the legend to Figure 1. The multiple alignment upon which the tree was based is shown on our website (Fig. S4).

little, 1990). As seen in Fig. 3*A* and *B*, the first two TMSs are separated from each other by exactly the same number of residues as are the second two TMSs, showing that the two extracellular loops in these connexins are of the same length. The only exceptions are three of the aligned proteins, which have a single amino-acid insertion in this region (*see* legend to Fig. 3). Additionally, two of the three fully conserved cysteyl residues in the inter-TMS loops are conserved in position in the two alignments. Although there is little further residue conservation between these two protein segments, we suggest that the positional similarities of the TMSs and cysteyl residues argue that the connexins arose by an internal gene duplication event. The primordial protein presumably was half sized and exhibited just 2 TMSs. The proposed intragenic duplication event doubled the size of and number of TMSs.

## Innexins

Table 2 presents the innexin homologues retrieved from the databases as of January 2002. Forty-two sequences were identified. Of these, twenty-six are from *Caenorhabditis elegans* (Starich et al., 2001) and nine are from *Drosophila melanogaster* (Stebbings et al., 2002). Both the *C. elegans* and *D. melanogaster* genomes had been fully sequenced when these studies were conducted, so these numbers presumably correspond to the total numbers encoded. It is surprising that the worm encodes three times as many innexin paralogues as does the fly. In addition to the worm and fly, only a few organisms, *Schistocerca americana* (grasshopper) and three closely related vertebrates are represented (Panchin et al., 2000). The vertebrate proteins have been suggested to be innexins based on sequence similarity with the invertebrate innexins, but it is not known whether they are able to form functional gap junction channels. After the completion of the work reported here, an innexin gene was cloned from the Annilida polychaete worm *Chaetopterus variopedatus* (Potenza et al., 2002).

As can be seen from the data summarized in Table 2, innexins fall roughly into the same size range as do the connexins (317–554 amino-acyl residues). However, excluding the single *C. elegans* unc9

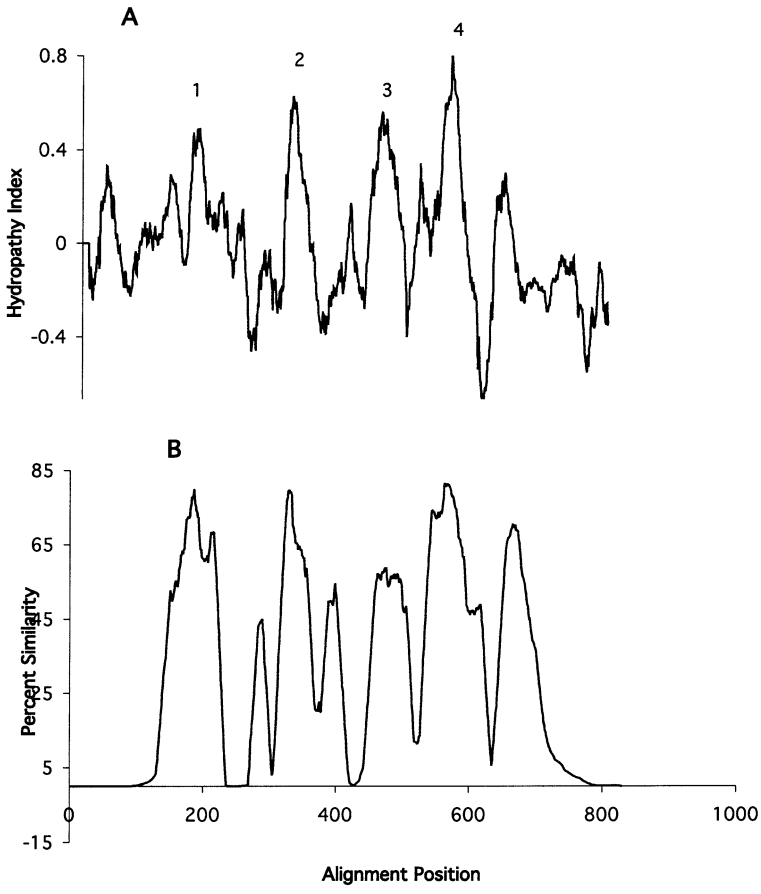


**Fig. 5.** Average hydropathy (*A*) and similarity (*B*) plots for the innexins. The format of presentation and the programs used were the same as for Fig. 2. The innexin family multiple alignment, from which these plots were derived using the AveHAS program (Zhai & Saier, 2001), is shown in Fig. S4 (*see* our ALIGN website).

homologue, the smallest protein is of 359 residues. Assuming that unc9 is an incomplete sequence, the size range of the innexins (359–554 residues) is narrower than that for the connexins (223–543 residues).

The complete multiple alignment of the innexin family proved to be much more divergent than that of the connexins in spite of their more narrow size range. Only seven fully conserved residues were identified (G189, C194, C214, P325, W329, F501 and K542; numbers refer to the alignment positions; *see* Fig. S4 in our ALIGN website). These were scattered throughout the alignment, as indicated. Only two of these seven residues proved to be cysteines. The alignment also revealed an increased proportion of gaps between putative transmembrane segments compared with the connexins (*see* below). As invertebrates evolved over a much greater time period than did the vertebrates, and the innexin family includes both invertebrate and vertebrate proteins, the degree of divergence is in accordance with expectation. The gaps and sequence divergence observed for the innexin alignment precluded derivation of a reliable signature sequence characteristic of this family.

The innexin family tree, shown in Fig. 4, differs greatly from the connexin tree shown in Fig. 1. All of the *Drosophila* and grasshopper proteins cluster separately from the C. *elegans* proteins, and the three mammalian proteins comprise a tight cluster that branches from a point between the worm and insect proteins. Moreover, there are far greater numbers of branches stemming from points near the center of the tree and far fewer large clusters than observed for the connexin tree. This latter fact reflects (1) the lack of more than a few sequence-similar paralogues in both *C. elegans* and *D. melanogaster*, and (2) the lack of close orthologues to any but a few of the innexins. The former fact contrasts with the situation for connexins in mammals, where relatively close paralogues have evolved as a result of more recent gene duplication events. The lack of close orthologues may reflect a deficiency of invertebrate sequence data. Thus, very scant sequence data are available for invertebrate organisms other than *C. elegans* and *D. melanogaster*. The absence of close paralogues between these two organisms represents a fundamental difference between vertebrate connexins and invertebrate innexins.

Fig. 5 shows the average hydropathy (*A*) and average similarity (*B*) plots for the innexin family. Both plots show four clear peaks of hydropathy (1–4 in *A*) corresponding to the four putative TMSs. The inter-TMS loops between TMSs 1 and 2 and TMSs 3 and 4 are poorly conserved. This fact contrasts with the situation for the connexins where both loops were well conserved. Not all of the inter-TMS loop regions

are poorly conserved, however. Comparison of Fig. 5*A* with Fig. 5*B* shows that relatively well-conserved regions occur to the left of TMSs 1 and 3 and to the right of TMSs 2 and 4. These facts also become apparent when the width of the peaks in Fig. 5*B* (average similarity) are compared with those in Fig. 5*A* (average hydropathy). The latter are much sharper than the former. The plots shown in Fig. 5 also reveal that most of the size variation observed for the innexins occurs in the N-terminal region preceding TMS1, and to a lesser extent, in the C-terminal region following TMS4. Since none of these regions is well conserved, they presumably either do not serve an important functional role or their functions are not common to many innexins. This observation correlates with the great phylogenetic distance separating most of these proteins.

Partial multiple alignments of putative TMSs 1 and 2 as compared with TMSs 3 and 4 revealed that the TMSs align approximately with each other, although there are many inter-TMS gaps. In contrast to the alignment of the connexin sequences, the cysteyl residues in the two segments do not align. This is not surprising in view of the fact that so many gaps are present in the alignment. If the innexins arose by an internal gene duplication event, many insertions and deletions must have been introduced during the evolution these proteins.

## Claudins

Table 3 tabulates the current members of the claudin family. Fifty-six sequences were identified, and of these, 17 are from humans, 22 are from the mouse, and 6 are from the rat. In addition to mammalian proteins, bird (chicken), fish (zebrafish), amphibian (frog) and chordate (ascidian) proteins are represented. These proteins are generally smaller than the connexins and innexins, the size range being 191–305 residues. Excluding the two largest and two smallest homologues, the size range is 207–264. Claudins have evidently undergone little size divergence during their evolution.

According to the database entries provided, one claudin homologue is a senescence-associated epithelial protein, while another is found in brain endothelial cells, and a third is associated with oligodendrocytes. Dysentery-inducing bacteria such as *Shigella* spp. can regulate tight junction function both by regulating claudin-1 association and by influencing occludin phosphorylation (Sakaguchi et al., 2002). Claudin 4 can secondarily serve as a receptor for the *Clostridium perfringens* enterotoxin (*see* Introduction). Examination of the claudin family multiple alignment revealed that only three residues, two cysteines at alignment positions 122 and 136 and a glycyl residue at position 272 were fully conserved.

Tepass et al. (2001) notes that *D. melanogaster* encodes two possible claudin-like proteins (CG3770 and CG6982). Both of these invertebrate proteins are about 210 residues long and have four predicted transmembrane domains with a single large inter-TMS loop between putative TMSs 1 and 2. They show a low degree of sequence similarity with claudins and much more with mammalian lens fiber intrinsic membrane proteins and p53 apoptosis effectors. Sequences from *C. elegans* have also been suggested to be claudin-like. These include NP_509257, NP_508583, NP_509800 and NP_509847). Although some similarity is observed, the sequence similarity of these proteins with claudins is insufficient to establish homology, and no functional data suggest a role in tight junction formation. They were therefore not included in our study.

The claudin family tree, based on the multiple alignment shown in Fig. S5, is shown in Fig. 6. No two mammalian paralogues from the human, mouse or rat are closely related to each other, showing that the gene duplication events that gave rise to these paralogues occurred relatively early. This suggestion is substantiated by the observation that close mammalian orthologues occur frequently. Moreover, the two chicken proteins represented are probably orthologues of the mammalian CLD3 and CLD5 claudins. By contrast, none of the fish, frog or ascidian proteins cluster closely with any mammalian protein. Orthologous relationships of these proteins can therefore not be assigned.

Average hydropathy and similarity plots for the claudin family are shown in Fig. 7*A* and *B*, respectively. The four peaks of hydropathy are clearly displayed. In contrast to the connexins and innexins, the claudins show comparable degrees of similarity in the loop regions between TMSs 1 and 2, and between TMSs 2 and 3, with substantially less similarity in the loop between TMSs 3 and 4. The N- and C-termini are poorly conserved. These facts suggest that the first extracellular loop as well as the central cytoplasmic loop may be more important for functions conserved among the proteins than the terminal extracellular loop.

## Occludins

Only 7 tight-junctional occludins were identified following database searches (Table 4). These proteins are derived from mammals (4), the chicken (1), the kangaroo rat (1) and the frog (1). They are large proteins (489 to 522 residues) of fairly uniform size.

The occludin multiple alignment, including all seven sequenced members of the family, revealed considerable sequence conservation throughout the alignment (*see* Fig. S6 on our ALIGN website). The average hydropathy and average similarity plots for the occludins are shown in Figure 8. Like the connexins, the extracellular loops of the occludins are well conserved while the central cytoplasmic loop is

**Table 3.** Sequenced protein of the claudin family

| Abbreviation | Database name or description | Organism | Size | GI # |
|---|---|---|---|---|
| BT-CLD16 | paracellin-1 | *Bos taurus* (bovine) | 235 | 6469051 |
| CA-CLD4 | claudin-4 (*C. perfringens* enterotoxin receptor) | *Cercopithecus aethiops* (vervet monkey) | 209 | 6685274 |
| CF-CLD2 | claudin-2 | *Canis familiaris* (dog) | 230 | 13991613 |
| CF-CLD3 | claudin-3 | *Canis familiaris* | 218 | 13991615 |
| DR-CLDX | claudin 7 | *Danio rerio* (zebrafish) | 215 | 6685322 |
| DR-ORF1 | claudin-like protein | *Danio rerio* | 208 | 6685321 |
| DR-ORF2 | claudin-like protein | *Danio rerio* | 209 | 6685320 |
| GG-CLD3 | claudin-3 | *Gallus gallus* (chicken) | 214 | 13377867 |
| GG-CLD5 | claudin-5 | *Gallus gallus* | 216 | 13377869 |
| HR-ORF1 | putative claudin | *Halocynthia roretzi* (ascidian) | 224 | 8919611 |
| HS-CLD1 | claudin-1 (senescence-associated epithelial membrane protein) | *Homo sapiens* (human) | 211 | 6685283 |
| HS-CLD2 | claudin-2 | *Homo sapiens* | 230 | 9966781 |
| HS-CLD3 | claudin-3 | *Homo sapiens* | 220 | 4502875 |
| HS-CLD4 | claudin-4 (*C. perfringens* enterotoxin receptor) | *Homo sapiens* | 209 | 4502877 |
| HS-CLD6 | claudin-6 | *Homo sapiens* | 220 | 11141863 |
| HS-CLD7 | claudin-7 | *Homo sapiens* | 211 | 12654455 |
| HS-CLD8 | claudin-8 | *Homo sapiens* | 225 | 6912318 |
| HS-CLD9 | claudin-9 | *Homo sapiens* | 217 | 11141861 |
| HS-CLD10 | claudin-10 | *Homo sapiens* | 228 | 5921465 |
| HS-CLD11 | claudin-11 | *Homo sapiens* | 207 | 10938016 |
| HS-CLD12 | claudin-12 | *Homo sapiens* | 244 | 6912312 |
| HS-CLD14 | claudin-14 | *Homo sapiens* | 239 | 6912314 |
| HS-CLD15 | claudin-15 | *Homo sapiens* | 228 | 7656981 |
| HS-CLD16 | paracellin-1 | *Homo sapiens* | 305 | 5729970 |
| HS-CLD17 | claudin-17 | *Homo sapiens* | 224 | 6912316 |
| HS-CLD18 | claudin-18 | *Homo sapiens* | 261 | 7705961 |
| HS-CLD20 | claudin-20 | *Homo sapiens* | 219 | 7387580 |
| MM-CLD1 | claudin-1 | *Mus musculus* (mouse) | 211 | 7710002 |
| MM-CLD2 | claudin-2 | *Mus musculus* | 230 | 7710004 |
| MM-CLD3 | claudin-3 | *Mus musculus* | 219 | 6753438 |
| MM-CLD4 | claudin-4 | *Mus musculus* | 210 | 6753440 |
| MM-CLD5 | claudin-5 (brain endothelial cell clone 1) | *Mus musculus* | 218 | 6685276 |
| MM-CLD6 | claudin-6 | *Mus musculus* | 219 | 9055190 |
| MM-CLD7 | claudin-7 | *Mus musculus* | 211 | 8393144 |
| MM-CLD8 | claudin-8 | *Mus musculus* | 225 | 9055192 |
| MM-CLD9 | claudin-9 | *Mus musculus* | 217 | 9938018 |
| MM-CLD10 | claudin-10 | *Mus musculus* | 231 | 10946728 |
| MM-CLD11 | claudin-11 (oligodendrocyte transmembrane protein) | *Mus musculus* | 207 | 6679186 |
| MM-CLD12 | claudin-12 | *Mus musculus* | 228 | 9799020 |
| MM-CLD13 | claudin-13 | *Mus musculus* | 211 | 10048432 |
| MM-CLD14 | claudin-14 | *Mus musculus* | 239 | 9506495 |
| MM-CLD15 | claudin-15 | *Mus musculus* | 227 | 14149748 |
| MM-CLD16 | paracellin-1 | *Mus musculus* | 235 | 13926043 |
| MM-CLD18 | claudin-18 | *Mus musculus* | 264 | 9790075 |
| MM-CLD19 | claudin-19 | *Mus musculus* | 193 | 9789476 |
| MM-ORF2 | putative protein | *Mus musculus* | 229 | 12860621 |
| MM-ORF3 | putative protein | *Mus musculus* | 220 | 12843248 |
| MM-ORF5 | putative protein | *Mus musculus* | 296 | 12844063 |
| MM-ORF6 | putative protein | *Mus musculus* | 209 | 12839895 |
| RN-CLD1 | claudin-1 | *Rattus norvegicus* (rat) | 211 | 13928976 |
| RN-CLD3 | claudin-3 | *Rattus norvegicus* | 219 | 6685268 |
| RN-CLD5 | claudin-5 | *Rattus norvegicus* | 206 | 13124033 |
| RN-CLD7 | claudin-7 | *Rattus norvegicus* | 191 | 6685270 |
| RN-CLD11 | claudin-11 | *Rattus norvegicus* | 207 | 12276167 |
| RN-CLD16 | paracellin-1 | *Rattus norvegicus* | 235 | 14028680 |
| XL-ORF1 | tight junction protein claudin | *Xenopus laevis* (frog) | 214 | 12004995 |

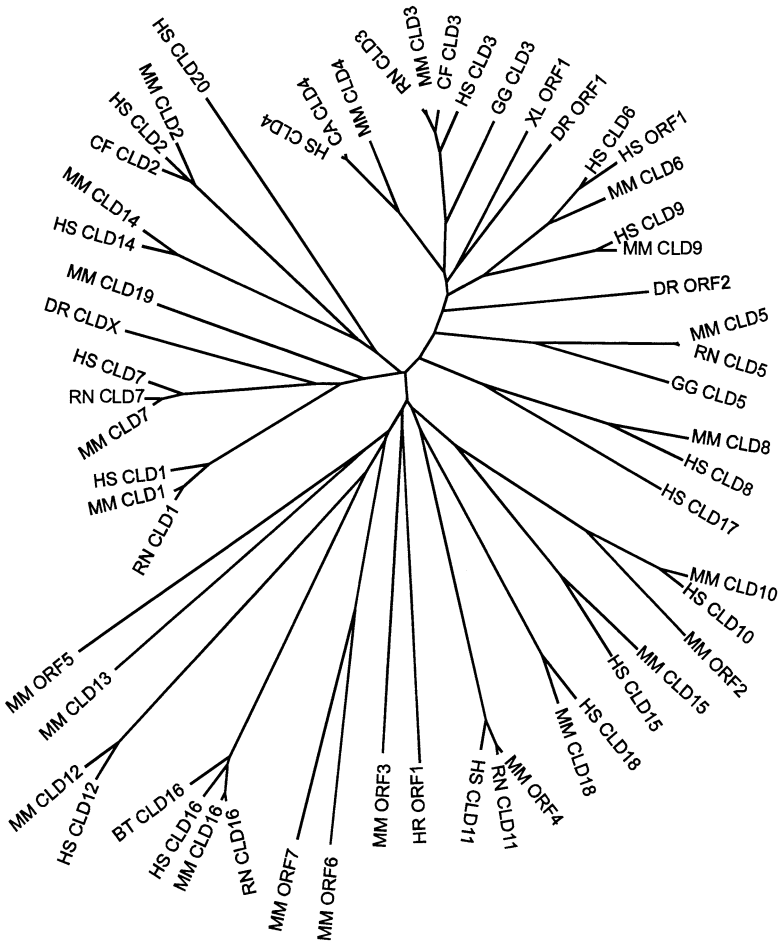


**Fig. 6.** Phylogenetic tree for the claudin protein family. Protein abbreviations are as indicated in Table 3. The claudin family multiple alignment is shown in Fig. S5 on our ALIGN website.

not. Several extended well-conserved motifs including four fully conserved cysteyl residues (*underlined*) were present as follows:

(1) F Y X W X S P P G $X_{12}$ C $X_3$ F X C V A S T L X W
(2) R X A X G F $X_2$ A M $X_3$ C F $X_3$ L
(3) Q Y X Y H Y C X V X P Q E A

The tree for the occludins is shown in Fig. 9. All mammalian proteins cluster tightly together, and the shape of the mammalian cluster suggests that these proteins are orthologous in agreement with the fact that only one occludin is found per organism. The kangaroo rat protein clusters loosely with the chicken protein, far from the frog homologue. However, in contrast to the connexins, large segments of the N- and particularly the C-terminal hydrophilic domains are well conserved. This fact suggests an important unified function for these large domains.

## Perspectives and Conclusions

In this article we have analyzed the sequences of integral membrane 4 TMS proteins implicated in junction formation in animals. Four protein families were analyzed: the connexins, innexins, claudins and occludins. The uniform structural features of these proteins are illustrated in Fig. 10. The multiple sequence alignments for these 4 protein families revealed a higher degree of sequence similarity for the connexins than for the innexins, in agreement with the facts that invertebrates have evolved over a much greater period of time than have the vertebrates, and that innexin homologues, but not connexins, are shared by invertebrates and vertebrates. One might propose that the connexins arose from a primordial innexin precursor, but the similarities between the two halves of the connexins suggest that the gene duplication event that gave rise to these proteins occurred long after any duplication event or events that might have given rise to the innexins. If any two of these four families of junctional proteins are related, there is no compelling evidence. However, extensive sequence divergence could have obscured such an event. Multiple duplication events have been documented during the evolution of other protein superfamilies (Nies et al, 1998; Pao et al., 1998; Tseng et al., 1999; Saier, 2000, 2001).

Connexins exhibit uniform topological features as well as the presence of conserved cysteyl residues in

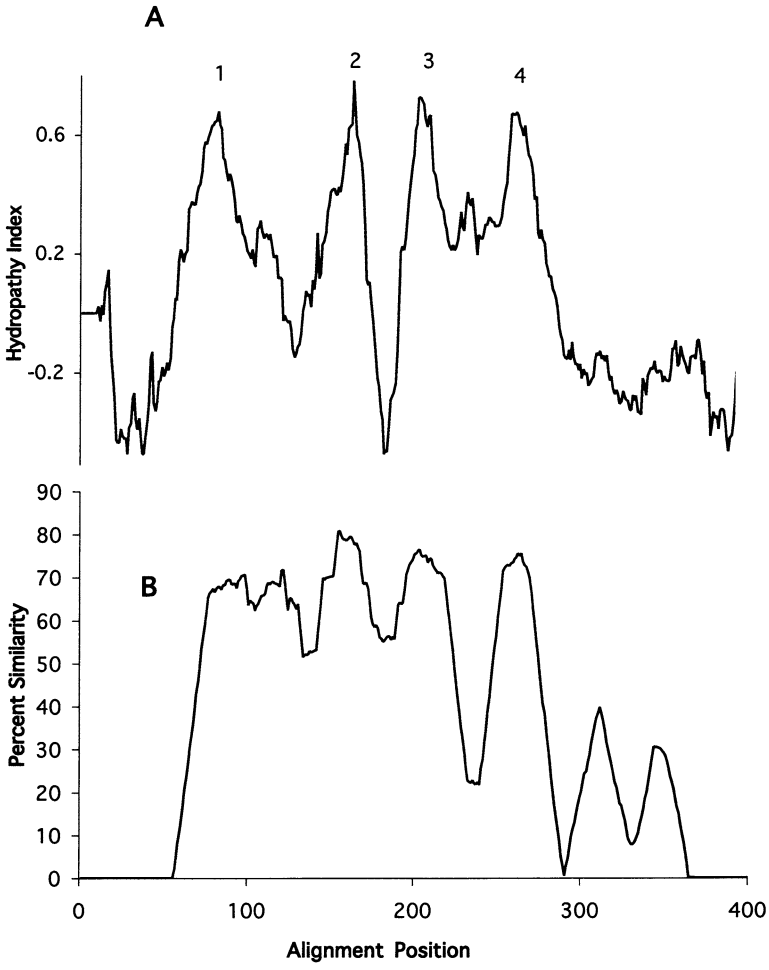


**Fig. 7.** Average hydropathy (*A*) and similarity (*B*) plots for the claudins. The format of presentation and the programs used were the same as for Fig. 2.

**Table 4.** Sequenced proteins of the occludin family

| Abbreviation | Database name or description | Organism | Size | GI# |
|---|---|---|---|---|
| CF | Tight junction structural protein | *Canis familiaris* (dog) | 521 | 7407642 |
| GG | Integral membrane protein localizing at tight junction | *Gallus gallus* (chicken) | 504 | 539507 |
| HS | Tight junction protein | *Homo sapiens* (humans) | 522 | 3914196 |
| MM | Tight junction protein | *Mus musculus* (mouse) | 521 | 3914209 |
| PT | Integral membrane protein localized at tight junction | *Potorous tridactylus* (kangaroo rat) | 489 | 1276981 |
| RN | Tight junction protein | *Rattus norvegicus* (rat) | 522 | 4126664 |
| XL | Tight junction protein | *Xenopus laevis* (frog) | 492 | 5833878 |

the loops between TMSs 1 and 2, and TMSs 3 and 4. Except for the vertebrate innexins, this family similarly exhibits well-conserved cysteyl residues. Other residues are fully or well conserved within each of these families, but not between the two families. Thus, when the complete multiple alignment of the innexins was derived, several residues proved to be largely conserved, and these residues occur exclusively in the extra-cytoplasmic loops and in the even-numbered TMSs. The conserved residues include four cysteyl residues, two between TMSs 1 and 2, and two between TMSs 3 and 4. The two cysteyl residues in each extracytoplasmic loop are separated by 16 or 17 residues. Fully conserved residues in the first halves of the innexins are G, C, C, Y, W, P, and W while in the second halves they are F, C, C, N, K, and W. These fully conserved residues are generally not conserved in nature or position between the two halves. Assuming that these fully conserved residues are of structural or functional significance, we conclude that the two halves of these proteins serve dissimilar functions. The same argument can be made

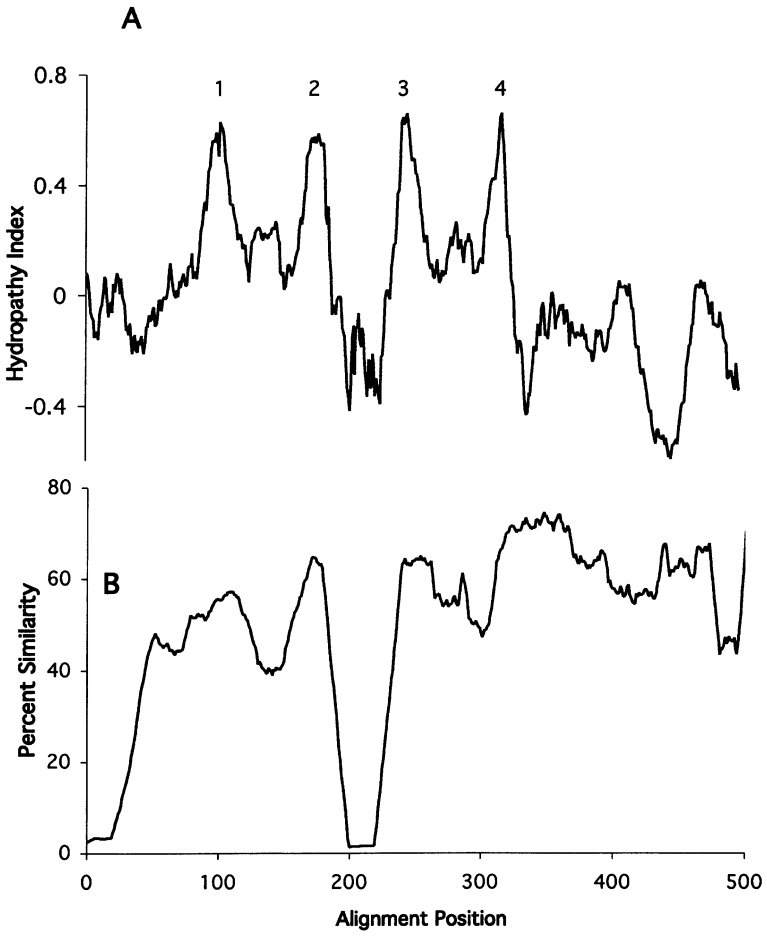


**Fig. 8.** Average hydropathy (*A*) and similarity (*B*) plots for the occludins. The format of presentation and the programs used were the same as for Fig. 2.

for the connexins, where except for the cysteyl residues, the fully conserved residues in the first extracellular loop differ in both nature and position from those in the second extracellular loop.

Multiple paralogues were identified for the connexin, innexin and claudin families but not for the occludins. Thus, 22 paralogous connexin homologues are present in humans, 26 and 9 paralogues of innexins were found in *C. elegans* and *D. melanogaster*, respectively, and 22 paralogous mouse claudins were identified. Many of these paralogues are likely to serve cell type or tissue-specific functions. However, the presence of over 200 cell types in a mammal clearly suggests that many cell types share the same junctional proteins.

Analyses of the data reported in this article led to the following evolutionary and functional suggestions: (1) In all four families, the most conserved regions of the proteins are the four TMSs. However, the loops between TMSs 1 and 2, and TMSs 3 and 4 are well conserved in the connexins and innexins (although less well conserved in the innexins). The loops between TMSs 1 and 2, and TMSs 3 and 4 are also well conserved in the claudins, and all loops plus flanking hydrophilic cytoplasm domains are well conserved in the occludins. This last fact may reflect the small number of occludins and the total lack of paralogues. (2) The phylogenetic trees for these four families allowed us to propose the existence of sets of orthologous proteins in all families except the innexins where phylogeny reflects the organismal source. Whether this is due to a lack of sequence information for other organisms or is a biological property of the innexin family remains to be determined. In this context, it is interesting to note that, unlike many vertebrate cells, gap junctional communication between cells from different insect orders could not be detected (Epstein & Gilula, 1977). (3) In the case of the connexins, evidence was presented to suggest that the two halves of the proteins derived from a common origin by internal gene duplication. Only the cysteyl residues that form disulfide bridges in the connexins and innexins on the external surfaces of the two adjacent cells are positionally well conserved both between the two halves of these proteins and between these two families (Kumar & Gilula, 1996; Yeager et al., 1998). This fact suggests an essential function, possibly as a receptor for specific protein-protein interactions, for the disulfide bridges that they form and leads to the

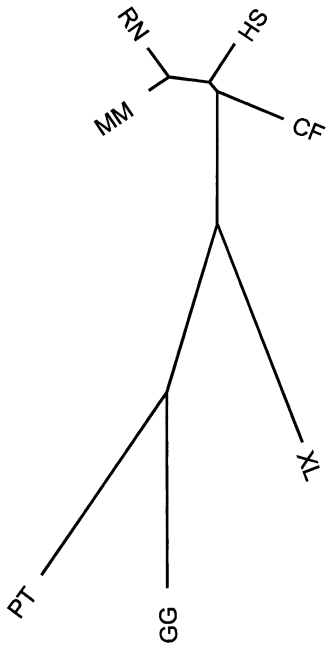


**Fig. 9.** Phylogenetic tree for the occludin protein family. Protein abbreviations are as indicated in Table 4. The occludin family multiple alignment is shown in Fig. S6 on our ALIGN website.

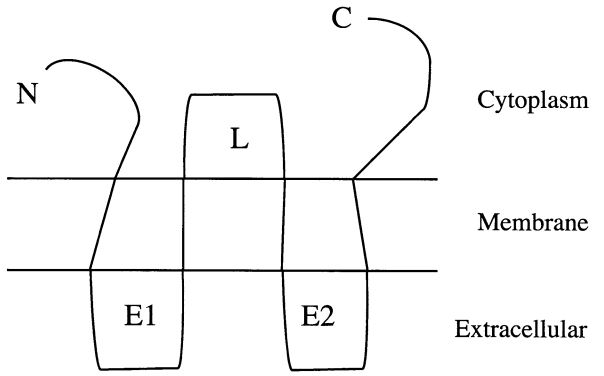


**Fig. 10.** Schematic representation of the transmembrane topologies of all four types of junctional proteins examined in this report. *N* and *C* correspond to the N- and C-termini of the proteins—*E1* and *E2* are the two extracytoplasmic loops, while *L* is the single cytoplasmic loop.

very tenuous suggestion that connexins and innexins share a common origin. (4) No evidence for a common origin of claudins and occludins, or for an origin resulting from intragenic duplication was obtained. Thus, if they do share a common 2TMS precursor with each other or with the gap junctional proteins, they have diverged in sequence from the precursor peptide beyond recognition. Perhaps 3-dimensional structural evidence will provide evidence for or against such a proposal. We suggest a similar role for conserved extracellular residues in the claudins and occludins. These findings and suggestions should serve as guides for future studies concerning the functions and origins of junctional proteins.

We thank Mary Beth Miller for assistance in the preparation of this manuscript. This work was supported by NIH grants GM55434 and GM64368 from the National Institute of General Medical Sciences (to MHS), an NEI grant EY13605 (to NMK), an RPB grant of unrestricted funds from Research to Prevent Blindness (to the UIC), and a grant from the Danish Research Council (to PAN).

## References

Ando-Akatsuka, Y., Saitou, M., Hirase, T., Kishi, M., Sakakibara, A., Itoh, M., Yonemura, S., Furuse, M., Tsukita, S. 1996. Interspecies diversity of the occludin sequence: cDNA cloning of human, mouse, dog, and rat-kangaroo homologues. *J. Cell Biol.* **133:**43–47

Balda, M.S., Flores-Maldonado, C., Cereijido, M., Matter, K. 2000. Multiple domains of occludin are involved in the regulation of paracellular permeability. *J. Cell. Biochem.* **78:**85–96

Bevans, C.G., Kordel, M., Rhee, S.K., Harris, A.L. 1998. Gating connexin 43 channels reconstituted in lipid vesicles by mitogen-activated protein kinase phosphorylation. *J. Biol. Chem.* **274:**5581–5587

Beyer, E.G., Paul, D.L., Goodenough, D.A. 1987. Connexin 43: A protein from rat heart homologous to a gap junction protein from liver. *J. Cell Biol.* **105:**2621–2629

Blaschuk, O.W., Oshima, T., Gour, B.J., Symonds, J.M., Park, J.H., Kevil, C.G., Trocha, S.D., Michaud, S., Okayama, N., Elrod, J.W., Alexander, J.S. 2002. Identification of an occludin cell adhesion recognition sequence. *Inflammation* **26:**193–198

Colegio, O.R., Van Itallie, C.M., McCrea, H.J., Rahner, C., Anderson, J.M. 2002. Claudins create charge-selective channels in the paracellular pathway between epithelial cells. *Am. J. Physiol.* **283**:C142–C147

Cordenonsi, M., Turco, F., D'atri, F., Hammar, E., Martinucci, G, Meggio, F., Citi, S. 1999. *Xenopus laevis* occludin. Identification of *in vitro* phosphorylation sites by protein kinase CK2 and association with cingulin. *Eur. J. Biochem.* **264:**374–384

Curtin, K.D., Zhang, Z., Wyman, R.J. 1999. *Drosophila* has several genes for gap junction proteins. *Gene* **232:**191–201

D'Andrea, P., Veronesi, V., Bicego, M., Melchionda, S., Zelante, L., Di Iorio, E., Bruzzone, R., Gasparini, P. 2002. Hearing loss: frequency and functional studies of the most common connexin26 alleles. *Biochem. Biophys. Res. Commun.* **296:**685–691

D'Atri, P., Citi, S. 2002. Molecular complexity of vertebrate tight junctions. *Mol. Membrane Biol.* **19:**103–112

Delmar, M. 2002. Connexin diversity: discriminating the message. *Circ. Res.* **91:**85–86

Eiberger, J., Degen, J., Romualdi, A., Deutsch, U., Willecke, K., Sohl, G. 2001. Connexin genes in the mouse and human genome. *Cell Adhes. Commun.* **8:**163–165

Epstein, M.L., Gilula, N.B. 1977. A study of communication specificity between cells in culture. *J. Cell Biol.* **75:**769–787

Evans, W.H., Martin, P.E.M. 2002a. Gap junctions: structure and function. *Mol. Membrane Biol.* **19:**121–136

Evans, W.H., Martin, P.E. 2002b. Lighting up gap junction channels in a flash. *Bioessays* **24:**876–880

Feng, D.-F., Doolittle, R.F. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183:**375–387

Ganfornina, M.D., Sanchez, D., Herrera, M., Bastiani, M.J. 1999. Developmental expression and molecular characterization of two gap junction channel proteins during embryogenesis in the grasshopper *Schistocerca americana*. *Dev. Genet.* **24:**137–150

Ghassemifar, M.R., Sheth, B., Papenbrock, T., Leese, H.J., Houghton, F.D., Fleming, T.P. 2002. Occludin TM4$^-$: an isoform of the tight junction protein present in primates lacking the fourth transmembrane domain. *J. Cell Sci.* **115:**3171–3180

Ghosh, P., Ghosh, S., Das, S. 2002. Self-regulation of rat liver GAP junction by phosphorylation. *Biochim. Biophys. Acta* **1564:**500–504

Hand, G.M., Muller, D.J., Nicholson, B.J., Engel, A., Sosinsky, G.E. 2002. Isolation and characterization of gap junctions from tissue culture cells. *J. Mol. Biol.* **315:**587–600

Harris, A.L. 2001. Emerging issues of connexin channels: biophysics fills the gap. *Q. Rev. Biophys.* **34:**325–472

Heiskala, M., Peterson, P.A., Yang, Y. 2001. The roles of claudin superfamily proteins in paracellular transport. *Traffic* **2:**93–98

Jahromi, S.S., Wentlandt, K., Piran, S., Carlen, P.L. 2002. Anticonvulsant actions of gap junctional blockers in an *in vitro* seizure model. *J. Neurophysiol.* **88:**1893–1902

Kim, D.Y., Kam, Y., Koo, S.K., Joe, C.O. 1999. Gating connexin 43 channels reconstituted in lipid vesicles by mitogen activated protein kinase phosphorylation. *J. Biol. Chem.* **274:**5581

Kitamura, K., Takahashi, K., Tamagawa, Y., Noguchi, Y., Kuroishikawa, Y., Ishikawa, K., Hagiwara, H. 2000. Deafness genes. *J. Med. Dent. Sci.* **47:**1–11

Kiuchi-Saishin, Y., Gotoh, S., Furuse, M., Takasuga, A., Tano, Y., Tsukita, S. 2002. Differential expression patterns of claudins, tight junction membrane proteins, in mouse nephron segments. *J. Am. Soc. Nephrol.* **13:**875–886

Kollmar, R., Nakamura, S.K., Kappler, J.A., Hudspeth, A.J. 2001. Expression and phylogeny of claudins in vertebrate primordia. *Proc. Natl. Acad. Sci. USA* **98:**10196–10201

Kumar, N.M., Gilula, N.B. 1996. The gap junction communication channel. *Cell* **84:**381–388

Kyte, J., Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157:**105–132

Landesman, Y., White, T.W., Starich, T.A., Shaw, I.E., Goodenough, D.A., Paul, D.L. 1999. Innexin-3 forms connexin-like intercellular channels. *J. Cell Sci.* **112:**2391–2396

Langbein, L., Grund, C., Kuhn, C., Praetzel, S., Kartenbeck, J., Brandner, J.M., Moll, I., Franke, W.W. 2002. Tight junctions and compositionally related junctional structures in mammalian stratified epithelia and cell cultures derived therefrom. *Eur. J. Cell Biol.* **81:**419–435

Le, T., Tseng, T.T., Saier, M.H., Jr. 1999. Flexible programs for the prediction of average amphipathicity of multiply aligned homologous proteins: Application to integral membrane transport proteins. *Mol. Membr. Biol.* **16:**173–179

Loewenstein, W.R. 1987. The cell-to-cell channel of gap junctions. *Cell* **48:**725–726

Long, H., Crean, C.D., Lee, W.H., Cummings, O.W., Gabig, T.G. 2001. Expression of *Clostridium perfringens* enterotoxin receptors claudin-3 and claudin-4 in prostate cancer epithelium. *Cancer Res.* **61:**7878–7881

Lopez, P., Balicki, D., Buehler, L.K., Falk, M.M., Chen, S.C. 2001. Distribution and dynamics of gap junction channels revealed in living cells. *Cell Adhes. Commun.* **8:**237–242

Mackay, D., Ionides, A., Kibar, Z., Rouleau, G., Berry, V., Moore, A., Shiels, A., Bhattacharya, S. 1999. Connexin46 mutations in autosomal dominant congenital cataract. *Am. J. Hum. Genet.* **64:**1357–1364

McClane, B.A. 2000. *Clostridium perfringens* enterotoxin and intestinal tight junctions. *Trends Microbiol.* **8:**145–146

Milks, L.C., Kumar, N.M., Houghten, R., Unwin, N., Gilula, N.B. 1988. Topology of the 32-kd liver gap junction protein determined by site-directed antibody localizations. *EMBO J.* **7:**2967–2975

Morcos, Y., Hosie, M.J., Bauer, H.C., Chan-Ling, T. 2001. Immunolocalization of occludin and claudin-1 to tight junctions in intact CNS vessels of mammalian retina. *J. Neurocytol.* **30:**107–123

Nies, D.H., Koch, S., Wachi, S., Peitzsch, N., Saier, M.H., Jr. 1998. CHR, a novel family of prokaryotic proton motive force-driven transporters probably containing chromate/sulfate antiporters. *J. Bacteriol.* **180:**5799–5802

Omori, Y., Mesnil, M., Yamasaki, H. 1996. Connexin 32 mutations from X-linked Charcot-Marie tooth disease patients: functional defects and dominant negative effects. *Mol. Biol. Cell* **7:**907–916

Panchin, Y., Kelmanson, I., Matz, M., Lukyanov, K., Usman, N., Lukyanov, S. 2000. A ubiquitous family of putative gap junction molecules. *Curr. Biol.* **10:**R473–R474

Pao, S.S., Paulsen, I.T., Saier, M.H., Jr. 1998. The major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* **62:**1–32

Phelan, P., Stanch, T.A. 2001. Innexins get into the gap. *Bioessays* **23:**388–396

Potenza, N., del Gaudio, R., Rivieccio, L., Russo, G.M., Geraci, G. 2002. Cloning and molecular characterization of the first innexin of the phylum annelida—expression of the gene during development. *J. Mol. Evol.* **54:**312–321

Richard, G., Smith, L.E., Bailey, R.A., Itin, P., Hohl, D., Epstein, E.H., Jr., DiGiovanna, J.J., Compton, J.G., Bale, S.J. 1998. Mutations in the human connexin gene GJB3 cause erythrokeratodermia variabilis. *Nature Genet.* **20:**366–369

Saier, M.H., Jr. 2000. Vectorial metabolism and the evolution of transport systems. *J. Bacteriol.* **182:**5029–5035

Saier, M.H., Jr. 2001. Evolution of transport proteins. *In:* Genetic Engineering. Principles and Methods, Vol. 23. J.K. Setlow, editor, pp. 1–9. Kluwer Academic/Plenum Publishers, New York

Sakaguchi, T., Kohler, H., Gu, X., McCormick, B.A., Reinecker, H.C. 2002. *Shigella flexneri* regulates tight junction-associated proteins in human intestinal epithelial cells. *Cell Microbiol.* **4:**367–381

Shibata, Y., Kumai, M., Nishii, K., Nakamura, K. 2001. Diversity and molecular anatomy of gap junctions. *Med. Electron Microsc.* **34:**153–159

Sotkis, A., Wang, X.G., Yasumura, T., Peracchia, L.L., Persechini, A., Rash, J.E., Peracchia, C. 2001. Calmodulin colocalizes with connexins and plays a direct role in gap junction channel gating. *Cell Adhes. Commun.* **8:**277–281

Starich, T., Sheehan, M., Jadrich, J., Shaw, J. 2001. Innexins in *C. elegans*. *Cell Adhes. Commun.* **8:**311–314

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

Stebbings, L.A., Todman, M.G., Phillips, R., Greer, C.E., Tam, J., Phelan, P., Jacobs, K., Bacon, J.P., Davies, J.A. 2002. Gap junctions in *Drosophila*: developmental expression of the entire innexin gene family. *Mech. Dev.* **113:**197–205

Tepass, U., Tanentzapf, G., Ward, R., Fehon, R. 2001. Epithelial cell polarity and cell junctions in *Drosophila*. *Annu. Rev. Genet.* **35:**747–784

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G. 1997. The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:**4876–4882

Tseng, T.-T., Gratwick, K.S., Kollman, J., Park, D., Nies, D.H., Goffeau, A., Saier, M.H., Jr. 1999. The RND permease superfamily: An ancient, ubiquitous and diverse family that includes human disease and development proteins. *J. Mol. Microbiol. Biotechnol.* **1:**107–125

Tsukita, S., Furuse, M. 2000. The structure and function of claudins, cell adhesion molecules at tight junctions. *Ann. N.Y. Acad. Sci.* **915:**129–135

Tsukita, S., Furuse, M. 2002. Claudin-based barrier in simple and stratified cellular sheets. *Curr. Opin. Cell. Biol.* **14:**531

Unger, V.M., Kumar, N.M., Gilula, N.B., Yeager, M. 1999. Three-dimensional structure of a recombinant gap junction membrane channel. *Sci. Mag.* **283:**1176–1180

White, T.W., Paul, D.L. 1999. Genetic diseases and gene knockouts reveal diverse connexin functions. *Annu. Rev. Physiol.* **61:**283–310

Wiliecke, K., Eiberger, J., Degen, J., Eckardt, D., Romualdi, A., Guldenagel, M., Deutsch, U., Sohl, G. 2002. Structural and functional diversity of connexin genes in the mouse and human genome. *Biol. Chem.* **383:**725–737

Yeager, M., Unger, V.M., Falk, M.M. 1998. Synthesis, assembly and structure of gap junction intercellular channels. *Curr. Opin. Struct. Biol.* **8:**517–524

Zhai, Y., Saier, M.H., Jr. 2001. The AveHAS program for the determination of average hydrophobicity, amphipathicity, and similarity. *J. Mol. Microbiol. Biotechnol.* **3:**285–286