Identifying bacterial and archaeal homologs of pentameric ligand-gated ion channel (pLGIC) family using domain-based and alignment-based approaches

Gloria Rendon,1,* Miriam R. Kantorovitz,1,2 Jeffrey L. Tilson3 and Eric Jakobsson1,4,*

1National Center for Supercomputing Applications; 2Department of Mathematics; 3Beckman Institute for Advanced Science and Technology; Department of Molecular and Integrative Physiology University of Illinois at Urbana-Champaign; Champaign, IL USA; 4Renaissance Computing Institute; Chapel Hill, NC USA

Key words: homology, prokaryotic channels, ART-LGIC, channel phylogeny, channel evolution, channel functional domains

Identification of bacterial and archaeal counterparts to eukaryotic ion channels has greatly facilitated studies of structural biophysics of the channels. Often, searches based only on sequence alignment tools are inadequate for discovering such distant bacterial and archaeal counterparts. We address the discovery of bacterial and archaeal members of the Pentameric Ligand-Gated Ion Channel (pLGIC) family by a combination of four computational methods. One domain-based method involves retrieval of proteins with pLGIC-relevant domains by matching those domains to previously established domain templates in the InterPro family of databases. The second domain-based method involves searches using ungapped de novo motifs discovered by MEME which were trained with well-characterized members of the pLGIC family. The third and fourth methods involve the use of two sequence alignment search algorithms BLASTp and psIBLAST respectively. The sequences returned from all methods were screened by having the correct topology for pLGIC’s, and by returning an annotated member of this family as one of the first ten hits using BLASTp against a comprehensive database of eukaryotic proteins. We found the domain-based searches to have high specificity but low sensitivity, while the sequence alignment methods have higher sensitivity but lower specificity. The four methods together discovered 69 putative bacterial and archaeal members of the pLGIC family. We ranked and divide the 69 proteins into groups according to the similarity of their domain compositions with known eukaryotic pLGIC’s. One especially notable group is more closely related to eukaryotic pLGIC’s than to any other known protein family, and has the overall topology of pLGICs, but the functional domains they contain are sufficiently different from those found in known pLGIC’s that they do not score very well against the pLGIC domain templates. We conclude that multiple methods used in a coordinated fashion outperform any single method for identifying likely distant bacterial and archaeal proteins that may provide useful models for important eukaryotic channel function. We note also that the methods used here are largely standard and readily accessible. The novelty is in the effectiveness of a strategy that combines these methods for identifying bacterial and archae relatives of this family. Therefore the paper may serve as a template for a broad group of workers to reliably identify bacterial and archaeal counterparts to eukaryotic proteins.

Introduction

The cys-loop ligand-gated ion channels are referred to by a variety of acronyms; in this paper we will use the terminology pLGIC’s for Pentameric Ligand-Gated Ion Channels. All members of this family exhibit homology and topological commonality to the archetype of the group, the nicotinic acetylcholine receptor channel. Included in those attributes is a distinctive feature of the extracellular domain called the cys-loop and a channel structure that is a pentamer in which all five subunits contribute to the formation of the pore. Members of this family have proliferated widely among metazoa where they are found as post- synaptic channels in both neuronal synapses and neuromuscular junctions. They have evolved sensitivity to a variety of ligands, a wide range of conductances, and selectivities for both cations and anions.

Until 2005, it was thought that this family existed only in metazoa, because BLAST, the common bioinformatics tool used for homology search, revealed no homologous sequences in any other biological classes. However Tasneem et al. discovered bacterial and archaeal homologs by using different known domains of the eukaryotic channels as independent probes and thereby found sequences that shared the same domain architecture as the eukaryotes. These sequences had escaped detection by BLAST because of the variability that had developed in those parts of the full protein sequence not contained in the pLGIC signature...
domains during the intervening time since the eukaryotic lineage diverged from the bacteria and archaea.

In general the utility of searching for bacterial and archaeal counterparts to eukaryotic membrane proteins is that detailed structural information, especially by x-ray diffraction, is more readily obtained from bacterial and archaeal than eukaryotic counterparts to eukaryotic membrane proteins. A prime recognition of this fact was the 2003 Nobel Prize awarded to Rod MacKinnon for structure determination of bacterial homologues of voltage-gated channels, and the demonstration that this work was critically important in understanding the biophysics of the corresponding channels in excitable membranes in eukaryotes. Indeed the Tasneem et al. discovery was followed by the functional characterization as a channel of one of the predicted pLGIC’s and by x-ray structure determination.3,4

In the present paper we return to the issue of finding bacterial and archaeal pLGIC’s in order to enlarge the field of bacterial and archaeal sequences that may be useful to study in the larger databases that have developed in the ensuing time, to evaluate automated methods that may provide the ability to do such searches much more rapidly, and to increase our knowledge of the evolution of this family of channels. In particular we implement and evaluate a combination of several search methods: (1) search for domains based on a comprehensive compilation of domain information from all protein sequences, such as in contained in the InterPro integration of domain databases,5-7 (2) take a de-novo approach to constructing domains from the eukaryotic sequences and using them to search/discover the bacterial and archaeal sequences using the programs MEME and MAST,8-13 and (3) return to the BLAST approach, using BLASTp and psi-BLAST6,12 to ensure that all relevant sequences will be captured. As validation we perform reverse BLASTp and topology analysis against a database of eukaryotic pLGIC’s in order to filter out the false positives that the methods inevitably retrieve.

Results

As a starting point, we used the 25 eukaryotic pLGIC proteins of the TCDB 1.A.9 family,3 shown in Table 1. We identified a group of 69 likely bacterial and archaeal homologs of the eukaryotic members of the pLGIC family (Table 2), using a combination of computational methods and databases. All of these 69 proteins passed our validation test, namely, they have the topology characteristic of pLGIC’s and, when BLASTed against all eukaryotic sequences in UniProt, returned annotated pLGIC’s among

Table 1. The eukaryotic members of the TCDB 1.A.9 family

| TCDB_id          | Accession | Protein names                                      | Organism               |
|------------------|-----------|----------------------------------------------------|------------------------|
| 1.A.9.1.1.ACH7   | P36544    | Neuronal acetylcholine receptor subunit alpha7     | Homo sapiens           |
| 1.A.9.2.1.ACR2   | P48182    | Acetylcholine receptor subunit beta-type acr-2     | Caenorhabditis elegans |
| 1.A.9.1.2.ACH5   | Q23022    | Acetylcholine receptor subunit-alpha type unc-38   | Caenorhabditis elegans |
| 1.A.9.2.3.ACR3   | Q93149    | Acetylcholine receptor subunit beta-type acr-3     | Caenorhabditis elegans |
| 1.A.9.2.1.SHT3   | P46098    | 5-hydroxytryptamine receptor 3A                    | Homo sapiens           |
| 1.A.9.2.2.1.SHT3R| P35563    | 5-hydroxytryptamine receptor 3A                    | Rattus norvegicus      |
| 1.A.9.2.2.2.SHT3B| Q9JJ16    | 5-hydroxytryptamine receptor 3B                    | Rattus norvegicus      |
| 1.A.9.3.1.1.GL3  | O17793    | Glycine receptor subunit alpha3                    | Homo sapiens           |
| 1.A.9.3.1.2.GRA1 | P23415    | Glycine receptor subunit alpha1                    | Homo sapiens           |
| 1.A.9.3.1.3.GLRA2| P23416    | Glycine receptor subunit alpha2                    | Homo sapiens           |
| 1.A.9.3.1.4.GLRB | P48167    | Glycine receptor subunit beta                      | Homo sapiens           |
| 1.A.9.3.2.HcIA   | A1KYB4    | Histamine-gated chloride channel subunit           | Drosophila melanogaster|
| 1.A.9.3.3.HcIB   | Q8WS32    | Histamine-gated chloride channel subunit B         | Drosophila melanogaster|
| 1.A.9.4.1.GLUC  | Q94000    | Glutamate-gated chloride channel (DrosGlCl)        | Drosophila melanogaster|
| 1.A.9.4.2.1.GluCl_1| O17793   | Avermectin-sensitive glutamate-gated chloride channel GluCl alph) | Caenorhabditis elegans |
| 1.A.9.4.2.2.GluCl_1-avr_14| Q8IFY7| Altered avermectin sensitivity protein 14, isoform c | Caenorhabditis elegans|
| 1.A.9.4.2.3.GluCl_1-avr_15| Q9TW41| Protein RT1G10.1.a, confirmed by transcript evidence | Caenorhabditis elegans|
| 1.A.9.5.1.GAB1   | P15431    | Gamma-aminobutyric acid receptor subunit beta1     | Homo sapiens           |
| 1.A.9.5.2.1.GBRAP| O9S166    | Gamma-aminobutyric acid receptor-assoc protein     | Homo sapiens           |
| 1.A.9.5.2.2.GBR2 | P47869    | Gamma-aminobutyric acid receptor subunit alpha2    | Homo sapiens           |
| 1.A.9.6.1.MOD_1  | Q9GGQ00   | Modulation of locomotion defective protein 1, isoform a | Caenorhabditis elegans |
| 1.A.9.6.2.LGC-53 | Q2PJ95    | Putative uncharacterized protein                    | Caenorhabditis elegans |
| 1.A.9.6.3.LGC-55 | Q9TV17    | Protein Y113G7A.5, partially confirmed by transcript evidence | Caenorhabditis elegans|
| 1.A.9.6.4.LGC-40 | Q22741    | Putative uncharacterized protein                    | Caenorhabditis elegans |
| 1.A.9.7.1.EXP_1  | Q9TZ15    | Excitatory GABA receptor EXP-1B                     | Caenorhabditis elegans |

Date of access May 1st, 2010.
Table 2. The 69 bacterial and archaeal putative members of the pLGIC family

| RefSeq_id | GenBank_id | UniProt_id | Similarity score | Cluster | NEW_ID        |
|-----------|------------|------------|------------------|---------|---------------|
| YP_898128.1 | A0Q559_FRATN | 15.8010 | G1               | G1_A0Q559_FRATN |
| ZP_01621420.1 | A01YQ9_9CYAN | 15.8010 | G1               | G1_A01YQ9_9CYAN  |
| ZP_01901882.1 | A6FMG2_9RHOB | 15.8010 | G1               | G1_A6FMG2_9RHOB  |
| ZP_04989386.1 | A7JKN7_FRANO | 15.8010 | G1               | G1_A7JKN7_FRANO  |
| YP_001677092.1 | B0TZ2N_FRAP2 | 15.8010 | G1               | G1_B0TZ2N_FRAP2  |
| YP_001891319.1 | B25FA9_FRATM | 15.8010 | G1               | G1_B25FA9_FRATM  |
| EDY19711.1 | B4AG47_FRANO | 15.8010 | G1               | G1_B4AG47_FRANO  |
| YP_002484938.1 | B8HXN2_CYAP4 | 15.8010 | G1               | G1_B8HXN2_CYAP4  |
| ZP_04956879.1 | B8KRE1_9GAMM | 15.8010 | G1               | G1_B8KRE1_9GAMM  |
| ZP_05248412.1 | C6YS78_9GAMM | 15.8010 | G1               | G1_C6YS78_9GAMM  |
| NP_927143.1 | Q7NDN8_GLOVI | 15.8010 | G1               | G1_Q7NDN8_GLOVI  |
| ZP_03246216 | 15.8010 | G1               | G1_ZP_03246216 |
| ZP_04754727 | 15.8010 | G1               | G1_ZP_04754727 |
| ZP_01729418.1 | A3IQT1_9CHRO | 13.9280 | G2               | G2_A3IQT1_9CHRO  |
| ZP_01905378.1 | A6FXF8_9DEL | 13.9280 | G2               | G2_A6FXF8_9DEL   |
| YP_001805637.1 | B1WSJ2_CYAA5 | 13.9280 | G2               | G2_B1WSJ2_CYAA5  |
| YP_002374163.1 | B7KSV4_CYAP8 | 13.9280 | G2               | G2_B7KSV4_CYAP8  |
| YP_002985991.1 | C6C8Z4_DICDC | 13.9280 | G2               | G2_C6C8Z4_DICDC  |
| YP_003002644.1 | C6CS3_DICZE | 13.9280 | G2               | G2_C6CS3_DICZE   |
| YP_003139739.1 | C7Q5R1_CYAP0 | 13.9280 | G2               | G2_C7Q5R1_CYAP0  |
| YP_003335322.1 | D2BY24_DICDA | 13.9280 | G2               | G2_D2BY24_DICDA  |
| 2VLJ | ELIC_ERWCH | 13.9280 | G2               | G2_ELIC_ERWCH   |
| ZP_00518157.1 | Q4BXY7_CROWT | 13.9280 | G2               | G2_Q4BXY7_CROWT  |
| YP_00112767.1 | A4IX99_FRATW | 12.6880 | G3               | G3_A4IX99_FRATW  |
| EDN34916.1 | A7IDX5_FRATT | 12.6880 | G3               | G3_A7IDX5_FRATT  |
| ZP_04987931.1 | A7IG12_FRANO | 12.6880 | G3               | G3_A7IG12_FRANO  |
| YP_001427905.1 | A7NAM6_FRATF | 12.6880 | G3               | G3_A7NAM6_FRATF  |
| ZP_05247929.1 | C6YQUS_FRATT | 12.6880 | G3               | G3_C6YQUS_FRATT  |
| ADA79409.1 | D2AK90_FRATT | 12.6880 | G3               | G3_D2AK90_FRATT  |
| YP_763066.1 | Q0BNAS_FRATO | 12.6880 | G3               | G3_Q0BNAS_FRATO  |
| YP_667426.1 | Q14GQ1_FRAT1 | 12.6880 | G3               | G3_Q14GQ1_FRAT1  |
| YP_513219.1 | Q2A4Y3_FRATH | 12.6880 | G3               | G3_Q2A4Y3_FRATH  |
| YP_170294.1 | Q5NF98_FRATT | 12.6880 | G3               | G3_Q5NF98_FRATT  |
| ZP_02275012.1 | 12.6880 | G3               | G3_ZP_02275012.1 |
| ZP_01124763.1 | A4CW13_SYNVP | 10.3100 | G4               | G4_A4CW13_SYNVP  |
| CAM7462.1 | A4TGVS_9PROT | 10.3100 | G4               | G4_A4TGVS_9PROT  |
| YP_001992176.1 | B3Q6J2_RHOPT | 10.3100 | G4               | G4_B3Q6J2_RHOPT  |
| EER62016.1 | C5TAJ8_ACIDE | 10.3100 | G4               | G4_C5TAJ8_ACIDE  |
| ZP_05786088.1 | D0CTE8_9RHOB | 10.3100 | G4               | G4_D0CTE8_9RHOB  |
| ZP_06359755 | D2MC58_RHOPA | 10.3100 | G4               | G4_D2MC58_RHOPA  |
| YP_781543.1 | Q07NC1_RHOPS | 10.3100 | G4               | G4_Q07NC1_RHOPS  |
| YP_569928.1 | Q13EQ2_RHOPS | 10.3100 | G4               | G4_Q13EQ2_RHOPS  |
| YP_532370.1 | Q214YS_RHOPB | 10.3100 | G4               | G4_Q214YS_RHOPB  |
| YP_486367.1 | Q2IWFW4_RHOP2 | 10.3100 | G4               | G4_Q2IWFW4_RHOP2  |

These are the members found by the various methods employed in this paper. Columns 1 and 2 contain their identifiers. Note that some of the sequences are found in the UniProt database only and others in the NCBI RefSeq database only hence the missing identifiers. The third column in the table is the net similarity score to eukaryotic LICS based on the commonality of InterPro domains. High score is most similar. The clustering scheme (G1-G7) in the fourth column indicates the degree of InterPro domain commonality with the proteins in the TCDB 1.A.9 family. G1 indicates the highest degree of commonality, while G7 indicates no domain commonality at all.
Table 2. The 69 bacterial and archaeal putative members of the pLGIC family (continued)

| Accession | Gene symbol | InterPro score | Domain | InterPro score | Domain |
|-----------|-------------|----------------|--------|----------------|--------|
| NP_948199.1 | Q6NSW2_RHOPA | 10.3100 | G4 | G4_Q6NSW2_RHOPA |
| YP_001226021.1 | A5GP99_SYNPW | 5.7630 | G5 | G5_A5GP99_SYNPW |
| ZP_020004689.1 | A7BVC3_9GAMM | 5.0320 | G5 | G5_A7BVC3_9GAMM |
| EDN69487.1 | A7BVE6_9GAMM | 4.8300 | G5 | G5_A7BVE6_9GAMM |
| YP_001404166.1 | A7I712_METB6 | 5.7060 | G5 | G5_A7I712_METB6 |
| YP_002466490.1 | B8G1I2_METPE | 5.7060 | G5 | G5_B8G1I2_METPE |
| ZP_01472912.1 | Q05QM7_9SYNE | 5.7060 | G5 | G5_Q05QM7_9SYNE |
| YP_678409.1 | Q11U47_CYTH3 | 5.7060 | G5 | G5_Q11U47_CYTH3 |
| YP_419787.1 | Q2WAG9_MAGSA | 4.9770 | G5 | G5_Q2WAG9_MAGSA |
| ZP_00207957.1 |  | 4.9770 | G5 | G5_ZP_00207957.1 |
| YP_003073700.1 | CSBJM6_TERTT | 3.5000 | G6 | G6_CSBJM6_TERTT |
| NP_766720.1 | Q89Y75_BRAJA | 3.5000 | G6 | G6_Q89Y75_BRAJA |
| YP_001018133.1 | A2CBK8_PROM3 | 1.0000 | G7 | G7_A2CBK8_PROM3 |
| ZP_01124751.1 | A4CWH1_SYNVP | 1.0000 | G7 | G7_A4CWH1_SYNVP |
| ZP_01124757.1 | A4CWH7_SYNVP | 1.0000 | G7 | G7_A4CWH7_SYNVP |
| YP_001224323.1 | A5GJB1_SYNPW | 1.0000 | G7 | G7_A5GJB1_SYNPW |
|  | B0P1B9_9CLOT | 1.0000 | G7 | G7_B0P1B9_9CLOT |
| YP_00863983.1 | B2J4R8_NOSP7 | 1.0000 | G7 | G7_B2J4R8_NOSP7 |
| ZP_05069790 | B6BT74_9RICK | 1.0000 | G7 | G7_B6BT74_9RICK |
| YP_731820.1 | Q016V2_SYN53 | 1.0000 | G7 | G7_Q016V2_SYN53 |
| YP_729853.1 | Q01CG9_SYN53 | 1.0000 | G7 | G7_Q01CG9_SYN53 |
| YP_676919.1 | Q11YD7_CYTH3 | 1.0000 | G7 | G7_Q11YD7_CYTH3 |
| YP_376599.1 | Q3AZC2_SYN59 | 1.0000 | G7 | G7_Q3AZC2_SYN59 |
| YP_113255.1 | Q60AU2_METCA | 1.0000 | G7 | G7_Q60AU2_METCA |

These are the members found by the various methods employed in this paper. Columns 1 and 2 contain their identifiers. Note that some of the sequences are found in the UniProt database only and others in the NCBI RefSeq database only hence the missing identifiers. The third column in the table is the net similarity score to eukaryotic LIC’s based on the commonality of InterPro domains. High score is most similar. The clustering scheme (G1-G7) in the fourth column indicates the degree of InterPro domain commonality with the proteins in the TCDB 1.A.9 family. G1 indicates the highest degree of commonality, while G7 indicates no domain commonality at all.

The top ten hits (see Methods section). Our findings integrate results from the UniProt and NCBI RefSeq databases. Two of the proteins are in archaea and the remaining 67 proteins are in bacteria. We found four of these proteins missing from the UniProt microbial databases and eight of these proteins missing from the NCBI RefSeq microbial database. Thirty-seven of the proteins were picked up by all the methods we used for discovery of the bacterial and archaeal pLGICs. One protein in the group is annotated as a glycine receptor, eight proteins are annotated as neurotransmitter-gated ion channel binding proteins, three are annotated as LGIC related, seven are annotated as hypothetical membrane proteins, three proteins are annotated as ABC transporters, one protein is annotated as blr0080 protein, one protein is annotated as FhipP2 protein and the rest of the proteins, over 40, are annotated as putative uncharacterized proteins.

The 25 eukaryotic pLGIC proteins in the TCDB 1.A.9 family and the 69 bacterial and archaeal putative pLGICs of Table 2 were used to reconstruct a likely phylogeny of the pLGIC family, as shown in Figure 1 using ClustalW. It is noteworthy that not all eukaryotic proteins are together in the same subtree; rather, the nicotinic acetylcholine (TCDB 1.A.9.1) and serotonin (TCDB 1.A.9.2) subfamilies are in one branch of the tree and the other six eukaryotic pLGIC subfamilies (TCDB 1.A.9.3 through 1.A.9.7) are in a different branch of the tree. The branch with subfamilies TCDB 1.A.9.1 and TCDB 1.A.9.2 also contains one bacterial protein (UniProt id: B0P1B9_9CLOT), albeit stemming from a very deep root. The two archaeal proteins appear in the same branch of the tree as each other, but do not form their own branch.

The InterPro domains search method. For each protein in the eukaryotic pLGIC probe set of Table 1, we identified its functional domain composition using InterProScan. We found that the repertoire of InterPro domains in the eukaryotic pLGIC proteins consisted of 18 entries (including overlapping domains); see Table 3. Only five of these 18 InterPro domains were also found in prokaryotic proteins: IPR006201, IPR006202, IPR006028, IPR006029 and IPR018000. The domain IPR006201 is the ‘generic’ pLGIC domain, of over 400 residues in length, its HMM profile was constructed with over 500 proteins among which there were GABA-A, nicotinic, glycine, glutamate and 5HT3 receptors. IPR006022 corresponds roughly to the extracellular ligand-binding motif, IPR006029 corresponds roughly...
We used similarity metrics based on InterPro domain composition to compare, cluster and rank the bacterial and archaeal proteins we identified as pLGIC homologs against members of the TCDB 1.A.9 family (see Methods). Seven non-eukaryotic pLGIC cluster groups were observed that denote ranking and are named G1 through G7; G1 (most similar) and G7 (least similar). The first cluster, denoted G1, aligns well with GLIC; the second cluster, denoted G2, contains 10 proteins and it aligns well with ELIC; the G3 cluster contains 11 proteins and it aligns well with the novel pLGIC proteins found in C. Elegans, subfamilies 1.A.9.6 and 1.A.9.7. The remaining clusters G4-G7 do not align particularly well with the eukaryotic pLGICs in the TCDB 1.A.9 family except for the above-mentioned B0P1B9_9CLOT to the entire transmembrane motif (composed of four helices); IPR006028 is labeled as GABAA receptor domain but it also picks up the corresponding conserved domain in nicotinic acetylcholine and glycine receptors. IPR018000 corresponds to the distinctive "cys-loop". We note that all of the prokaryotes that were found to contain this cys-loop domain turned out to be false positives; i.e., they failed the validation tests of topology and/or reverse BLAST. It is possible that some of these false positives are soluble ligand-binding proteins. None of the 69 prokaryotic proteins in Table 2, including GLOVI and ELIC, contained the InterPro cys-loop domain. The InterPro domain architecture of these 69 likely bacterial and archaeal pLGIC proteins is shown in Figure 2A.

Figure 1. Overall retrieved in this study. The shaded delineations around the edge of the tree show the major clusters into which the tree is divided. Composition of the clusters is discussed in the text. Phylogenetic tree of the eukaryotic probes and their non-eukaryotic correlates.
that aligns well with the TCDB’s 1.A.9.1 and 1.A.9.2 subfamilies. The G7 cluster group does not appear to contain domain patterns in common with those in the TCDB 1.A.9 family. In fact, the proteins in the G7 cluster have no InterPro domains at all. However they have the topology characteristic of pLGIC’s and, when BLASTed against all eukaryotic sequences in UniProt, returned annotated pLGIC’s among the top ten hits.

Ab initio motifs approach. We used the MEME motif discovery program to find conserved motifs in the pLGIC family. As a training set for MEME, we took the 25 eukaryotic pLGIC sequences in the TCDB 1.A.9 family. Table 4 shows the consensus sequences for the top 15 motifs found by MEME (their Hidden Markov Model logos are provided in Sup. material). Figure 2B shows where these motifs occur in the sequences of the training set (with p-value less than 0.0001). We observed that motif 1 contains the cys-loop. Motif 4 corresponds to the “beta1-beta2 loop”, a stretch of amino acids that lie in between the extracellular neurotransmitter-binding portion of the receptor and the transmembrane pore. Motif 5 corresponds to what is known as “loop A”, a stretch of amino acids involved in neurotransmitter binding. Motif 2 corresponds to the first TM. In the Glycine, GABA and glutamate subfamilies, motifs 7, 3 and 6 are the second, third and fourth TMs respectively. In the ACh and 5HT3 subfamilies, motifs 12 and 15 are the second and fourth TMs, respectively. Note that the long loop between TM3 and TM4 in the eukaryotic pLGIC sequences is clearly seen in Figure 2B.

Searching the databases (UniProt and RefSeq) with the MAST program for proteins with occurrence of the pLGIC motif signature (see Methods) resulted in 52 non-eukaryotic sequences. All of these 52 sequences were true positives by our validation procedure.

Figure 2C shows the occurrence of the MEME motifs in the 69 sequences found by the combined methods. Note that, in contrast to the InterPro cys-loop motif, the MEME cys-loop motif (motif 1) was detected as present in most of these sequences.

Another observation from Figure 2B and C is that the motif diagrams in many of the prokaryotic sequences are highly similar to the motif diagrams in the eukaryotic pLGIC sequences. For example, the prokaryotic sequence C6CH53_DICHE has 7 out of the 10 motifs in the Human GABA sequence P47870, where motifs occurred in the same order in both sequences.

Note that the pLGIC motifs 11, 12 and 15 are found in the ACh and 5HT3 subfamilies and not in the GABA, glutamate and Glycine subfamilies, while motifs 5, 6, 7, 8 and 9 are associated with the GABA, glutamate and Glycine subfamilies and not the ACh and 5HT3. Most of the prokaryotic sequences are more similar to the GABA, glutamate and Glycine subfamilies, in the sense that they contain motifs from these subfamilies. A few sequences have motif diagrams that are more similar to the ACh subfamily. For example, the motifs in the transmembrane region of the bacterial sequence A7BVE6_9GAMM have the same pattern as the TM motifs in the eukaryotic ACh sequence Q23022.

Interestingly, a few prokaryotic sequences contain correctly positioned motifs that come from both GABA and ACh subfamilies, such as B6BT74_9RICK, which contains both, motif 11 and 6 and Q05QM7_9SYNE, which contains both motif 7 and 15. There is an interesting tension between the facts that the prokaryotes, which contained the InterPro cys-loop signature domain, were ultimately found to be false positives, while the true positives contained the MEME-derived motif corresponding to the cys-loop. Even though the InterPro domain IPR018000 and the MEME-derived motif 1 describe the same segment of proteins from the same family, the algorithms underlying the

Table 3. InterPro domains of the pLGIC Family

| InterPro domain | Hits in Eukaryotes | Hits in Bacteria | Hits in Archaea |
|-----------------|--------------------|-----------------|-----------------|
| IPR001390       | 138                | 0               | 0               |
| IPR002289       | 151                | 0               | 0               |
| IPR002394       | 812                | 0               | 0               |
| IPR005432       | 11                 | 0               | 0               |
| IPR006028       | 913                | 21              | 0               |
| IPR006029       | 2130               | 24              | 0               |
| IPR006201       | 2622               | 43              | 0               |
| IPR006202       | 2265               | 44              | 2               |
| IPR008060       | 19                 | 0               | 0               |
| IPR008127       | 48                 | 0               | 0               |
| IPR008128       | 12                 | 0               | 0               |
| IPR008129       | 8                  | 0               | 0               |
| IPR008130       | 5                  | 0               | 0               |
| IPR008132       | 34                 | 0               | 0               |
| IPR008133       | 25                 | 0               | 0               |
| IPR008134       | 4                  | 0               | 0               |
| IPR015680       | 156                | 0               | 0               |
| IPR018000       | 1833               | 7               | 0               |
Figure 2A. InterPro domain architecture of the non-eukaryotic pLGIC sequences discovered in the study.
creation of the motifs in the two cases were sufficiently different that different pattern matching tools (InterProScan and MAST) did not agree on the presence or absence of the motif. An inspection of the two patterns to be matched reveals the reason for the discrepancy. IPR01800 is defined by the PROSITE pattern C-x-[LIVMFQ]-x-[LIVMF]-x(2)-[FY]-P-x-D-x(3)-C. Note that there are two cysteines, in addition to a P and a D, that are absolutely mandatory. Motif 1, on the other hand, is defined by a hidden Markov model with the consensus sequence LKS SCP MDL EWF PMD QVN CSL QFE SWG YTT. This sequence contains the two C’s, the P, and the D, but they are not mandatory, so that sequences that differ in one or more of those four positions may still be judged to contain the motif if they are sufficiently probable based on the hidden Markov model (Fig. 5A). There is no way to judge which approach is correct in general. In some situations a motif requires absolute conservation of particular residues for functional or structural reasons. In other situations, more flexibility in conservation patterns is permitted or even favored. Lacking a priori knowledge of the precise meaning and significance of particular pattern matching, it is clear that there is an advantage to protocols involving redundant tools and algorithms, and methods for achieving consensus among the results of the redundant procedures.

The structural positions of some of the MEME-derived motifs are shown in Figure 3A–C. The motifs are superposed on structural pictures of Torpedo (Fig. 3A and PDB ID 2BG9), GLIC (Fig. 3B and PDB ID 3EAM), and ELIC (Fig. 3C and PDB ID 2VL0). Notable features of the figures are that four motifs are common to all three structures, namely motifs 1 and 4, in the extracellular domains, and motifs 2 and 3, in the transmembrane region. Motif 6 and motif 15 are alternative versions of the fourth transmembrane helix. The bacteria ELIC and GLIC have motif 6 and the Torpedo has motif 15. Figure 4A and B shows further details of the sequence logos and alignment for Motif 6 and Motif 15 respectively. Inspection of the sequence logos in Figure 4A, shows the common patterns of conservation for the prokaryotic and eukaryotic motifs. Comparison of the TCDB ID’s in the alignment in Figure 4A, with the IDs on the tips of the phylogenetic tree in Figure 1 shows that presence of motif 6 is characteristic of the eukaryotic sequences in branch S2, which contains the TCDB subfamilies 1.A.9.3, 1.A.9.4, 1.A.9.5, 1.A.9.6 and 1.A.9.7. Likewise, comparison of the TCDB IDs in the alignment of Figure 4B, with the IDs on the tips of the tree in Figure 1 shows that the other eukaryotic subfamilies (1.A.9.1 and 1.A.9.2) located in branch S1 contain a different motif in the fourth transmembrane helix, represented by motif 15. It is clear that motif 15 has little or no relationship to motif 6. It was presumably acquired from another eukaryotic source at the time that the Ach receptors/cation channels (1.a.9.1 and 1.a.9.2 in the TCDB) diverged from Gaba/glycine members of the pGLIC family (TCDB subfamilies 1.a.9.3–7).

The sequence logo representations for the motifs that are common to all of the structures in Figure 3 (motifs 1, 2, 3 and 4) are shown in Figure 5A–D. Motif 1 (Fig. 5A) contains the noted cys-loop, with C’s completely conserved in positions 5 and 19 in the eukaryotic pLGICs. In the prokaryotic pLGICs, however, there are no cysteines at either positions 5 or 19 even though the overall residue conservation patterns of the entire motif clearly identify the region as being homologous to the eukaryotic motif 1. On the other hand, there is a proline at position 13 that is completely conserved across prokaryotes and eukaryotes, and very strong conservation at positions 12, 15 and 17. From a perspective that encompasses both prokaryotes and eukaryotes, the “cys-loop” might have been called a “P-loop”. Motifs 2, 3 and 4 show similar degrees of correspondence and conservation across the prokaryotic-eukaryotic divide (sequence logos in Fig. 5B–D,星空.
the 25 pLGIC proteins in the TCDB 1.A.9 family and the 69 likely non-eukaryotic pLGICs. Figure 6A–D visually capture such information. Figure 6A and B show each of the InterPro domain sets in eukaryotes and prokaryotes respectively. Figure 6C and D show the corresponding information for the MEME-derived motifs. The connecting tie lines indicate that the connected domain/motif sets have at least one common element. Examination of Figure 6 respectively). The percent identity of the consensus sequences between prokaryotes and eukaryotes for the four motifs, based on most common residue at each position, are: Motif 1, 50% (15/30), motif 2, 40% (12/30), motif 3, 23% (7/30) and motif 4, 23% (7/30).

To inspect protein-domain co-occurrence patterns we examined the same set of 94 proteins we used in the phylogeny study;
leads to the following inferences: (A) By measures of the numbers of motif/domains in individual sets, the eukaryotic proteins are on the average more complex than the prokaryotic proteins. (B) It appears that the MEME-derived motif analysis reveals relationships that are not revealed in the InterPro domain analysis. Some of the prokaryotic family members do not have any of the InterPro domains contained in the eukaryotic set, whereas all of the prokaryotic members have at least two of the motifs derived by MEME from the eukaryotic set. We presume this is because the InterPro domain definitions are generally tied to functional assignment, and we know much less about the functional characteristics of the prokaryotic members of this family than of the eukaryotic ones. What we do know suggests that some functional characteristics are quite different. For example, one bacterial member of the family is gated by protons, quite a different sensitivity from the eukaryotic members. Since the MEME-derived motifs are based solely on homology, relationships of motifs of so-far-unknown functional significance are readily revealed.

In summary we used a combination of methods and databases to conduct the search for the likely non-eukaryotic homologs of the eukaryotic members of the pLGIC family. With respect to methods, we used BLASTp and psiBLAST for high sensitivity-low specificity sequence-alignment based search, MEME and MAST for search by de-novo domain discovery from a well-characterized training set, high throughput InterProScan implemented in MotifNetwork[22] for search by previously characterized domains, and reverse BLASTp and TMHMM for screening results from all four methods. Of the 69 sequences identified here, five sequences were missing from the UniProt database and eight sequences were not in the NCBI refSeq database. Just over 50% of the sequences (37) were identified by ALL methods. The summary of different methods and databases used and their performance is presented in Table 5, and they are described in more detail in the Methods section.

Methods’ performance. Table 5 shows how well the various methods performed on the different databases in terms of sensitivity and selectivity. We see that BLASTp has good sensitivity in UniProt (Archaea 2 tp 0 fn, Bacteria 53 tp, 12 fn) but only acceptable sensitivity in RefSeq (40 tp, 21 fn). Whereas the case is reversed with regards to specificity; BLASTp yielded good specificity in RefSeq (40 tp, 1 fp) and poor specificity in UniProt (Archaea 2 tp 144 fp, Bacteria 53 tp 137 fp).

The sensitivity of psiBLAST showed significant improvement over that of BLASTp on both databases. However, the specificity of psiBLAST was much worse than that of BLASTp on both databases. The MEME/MAST method had the highest specificity (no false positives) and sensitivity similar to that of the InterPro domains method.

Discussion

The work described in this paper is significant on the following counts:

1. We have discovered new non-eukaryotic counterparts to the pLGIC family of ion channels, including some that by some measures are more eukaryote-like than the non-eukaryotic counterparts previously identified. Thus we have significantly enlarged the pool of candidate bacterial and archaeal sequences that may serve to illuminate the structure and biophysics of these channels, and their function in prokaryotes.

2. We have shown that for a family of proteins well characterized by entries in the InterPro domain database, de novo construction of domains and searching by de novo domains (MEME/MAST) is essentially as powerful in discovery of bacterial and archaeal counterparts as searching by the known InterPro domains. This provides validation for future domain-based analysis on families where the InterPro database is not as complete as it is for the pLGIC’s.

3. We have developed a protocol that uses a combination of methods for discovery of bacterial and archaeal counterparts to eukaryotic proteins. Such a protocol is especially significant for structural and biophysical studies of membrane proteins, since eukaryotic membrane proteins are generally more difficult to crystallize than their non-eukaryotic counterparts. It is also instructive in understanding the evolution of proteins and their functions. Each method in the protocol as implemented in this paper begins with the same set of reference sequences, in this case the eukaryotic pLGIC sequences in the Transport Classification Data Base (TCDB 1.A.9). We show that the methods in combination are more powerful than the methods separately. Specifically the combination of the four methods retrieved 69 bacterial and archaeal counterparts to the eukaryotic pLGIC’s, as validated by the dual criteria of topology and reverse blast. Out of those 69, the four methods separately retrieved 48, 49, 55 and 64 of the 69 respectively, for InterProScan search, de novo domain search, BLASTp and psiBlast respectively, each followed by a validation step. As seen in Table 5, the more sensitive methods initially retrieved many false hits that were screened out by topology matching and reverse BLAST. The false positives retrieved by the InterProScan are a special and definable case. They appear to be soluble proteins with homology to ligand-binding domains but are not connected directly to a channel structure.

Table 4. Consensus sequences for MEME motifs

| Motif | Consensus sequence |
|-------|-------------------|
| 1     | LKS SCP MDL EWF PMD VQO CSL QFE SWG YTT |
| 2     | LRR QMG YYV IQL YIP CCL IVI LSW VSF WIN |
| 3     | SLP KVS YVX AID IWM GGC MVF VFC ALL EYA |
| 4     | SVD EKN MDY TTN IWF RQK WND HRL QWN EYP |
| 5     | DPT MAD QIW MPD TFF TFE ANE KGA HFH EIT TPN |
| 6     | AKR IDK ISR IFL PMA FLI FNI FYW CYY |
| 7     | MDA APA RVT LGI TTV LTM TTQ |
| 8     | RTK GYD ARI RPN FKG PPV NIVG CNI FIN SFQ |
| 9     | RIH KNG TVL YSI RIT |
| 10    | ENK ETG YCT KHY NTG KFT CLE |
| 11    | YDG IKK LHI PTD QIW VPD IVI NNN ADG NY |
| 12    | SGE KVS FKI SIL LGY TVF LLI VSE ILP ATA |
| 13    | NDL IFE WQE HDP VQV KEG LHL |
| 14    | EDE HRL YHH LRR NYN KHV RPV RNW GKP TTV |
| 15    | DWK YVA CVL DRL LFI IFY AAT LAG TIG I |
(4) The compute intensity of the methods was in direct relationship to the number of sequences of interest retrieved. The greater the compute intensity, the more non-eukaryotic pLGIC’s were retrieved, but at a successively high computational cost. InterProScan is the most compute efficient, since it consists only of searching a precomputed database for protein ID’s that match known InterPro domains. The MEME/MAST de novo domain searching method is the next most efficient because, once the domains are defined by the training set, the domains that are being scanned across the database being searched are significantly shorter than the complete proteins. BLASTp is more compute intensive than the domain methods because it seeks to match complete protein sequences and also because it recovers a significant number of false hits that must be removed from the hit list by a combination of topology analysis and reverse BLAST. Finally psiBlast is the most compute intensive, firstly because it involves several rounds of BLAST with successively modified substitution matrices tuned to the alignment of the particular protein family, and secondly because it generates a very high number of false hits and consequently requires very compute intensive screening by reverse BLAST.

(5) We find that most of the 69 bacterial and archaeal pLGIC’s we ultimately retrieved were incompletely annotated in the sequence databases. Since all the individual methods we used in this study are readily available, and since the overall strategy we deployed could be coupled into a high throughput workflow, we suggest incorporation of these strategies into annotation pipelines.

(6) The work has revealed the following main points about the evolution of this particular family of channels:
• The building blocks for prokaryotic and eukaryotic proteins of this family are almost the same, as evidenced by the number of
motifs that are common to both prokaryotes and eukaryotes, and the structure of those motifs, which show strong conservation across the prokaryotic/eukaryotic divide at multiple locations. A corollary to this point is that evolution within prokaryocytes and eukaryocytes consists as much of rearranging the memes as it does of point substitutions.

• We infer from the data that there was likely a “primeval scaffold” on which the subsequent members of the family were built, that consists of the few memes that are almost completely conserved across prokaryotic and eukaryotic members.

• The “cys-loop” is the most universal signature meme, although it does not always have cys’s. (If we were starting over with nomenclature, we might call it the “p-loop”).

• In general, the extracellular ligand-binding region is more strongly conserved with respect to meme composition than the rest of the sequence, with the exception of the primeval fourth transmembrane domain.

Methods

Selection of the training set or probe set. As representative eukaryotic members of the pLGIC family, we used the 25 proteins provided by the Transport Classification DataBase (TCDB 1.A.9), date of access May 1st, 2010 (Table 1). These eukaryotic proteins represent seven out of the nine pLGIC subfamilies identified by TCDB; they are: 1.A.9.1 nicotinic acetylcholine-activated cation-selective channels, 1.A.9.2 serotonin (5-hydroxytryptamine)-activated cation-selective channels, 1.A.9.3 adult glycine-inhibited chloride (anion selective) heteropentameric channels, Ort and histamine-gated chloride channels, 1.A.9.4 glutamate-inhibited chloride (anion-selective) channels and glutamate gated chloride channels, 1.A.9.5 GABA-inhibited chloride channels, 1.A.9.6 MOD-1, high affinity dopamine or tyramine receptor chloride channels and low-affinity serotonin receptor channel, and 1.A.9.7 EXP-1. We excluded from this list P60517, a GABA associated protein that does not itself appear to be a channel of this family, as well as the two bacterial proteins ELIC and GLIC that form the TCDB pLGIC subfamilies 1.A.9.8 and 1.A.9.9 respectively.

Selection of the search space: the databases. We chose two databases of protein sequences to search with the blast tools, NCBI RefSeq microbial (release 39) and UniProt bacteria+archaea (release 14); they are vast but not completely comprehensive and to the best of our knowledge they overlap in a significant yet not complete way; in fact, of the nearly 70 sequences identified in this study as likely non-eukaryotic homologs of the pLGIC family, four sequences were missing from UniProt [ZP_00207957.1, ZP_02275012.1, ZP_03246216, ZP_04754727] and eight sequences were not in NCBI refSeq [A4TVG5, A7BVE6, B0P1B9, B4AQ47, C5T0J8, D2AK90, P0C7B7, Q0QM43]. Note that we did not combine these two databases into a single search space; instead, we ran the BLAST searches and the MAST motif discovery tool in each database as a separate search space. To search with the InterProScan search engine, we used release 25 of the InterPro database.

Validation procedure. Two computational tests were used to validate the results of a particular tool, namely (A) use the results of TMHMM to check if the number of predicted transmembranes in the protein was consistent with those of an pLGIC protein which contain between three and five transmembranes and (B) run reverse BLAST on the UniProt Eukaryotes database to see if among the top ten closest homologs in that database, there were any hits annotated as pLGIC proteins.

Measuring the performance of the tools. The quality of the tools was assessed in terms of sensitivity and specificity of the results they gave (Table 5).

\[
\text{Sensitivity} = \frac{tp}{tp + fn} \quad (1)
\]

\[
\text{Specificity} = \frac{tp}{tp + fp} \quad (2)
\]

where \( tp \) is number of true positives, non-eukaryotic proteins that were predicted by the specific tool to be homologs of the pLGIC family and were confirmed by the validation procedure. \( fn \) is number of false negatives, non-eukaryotic proteins that were not predicted by the specific tool to be homologs of the pLGICs but were predicted by one of the other tools and confirmed by the validation procedure. \( fp \) is the amount of noise in the signal, non-eukaryotic proteins that were predicted by the
tool to be homologs of the pLGIC family but did not pass the validation procedure.

Ranking the results based on InterPro similarity. In this section we describe how the 69 bacterial and archaeal proteins were ranked and then grouped by similarity in domain composition to the eukaryotic LIC’s in the TCDB database. There are multiple possible similarity metrics.\textsuperscript{19-21,25} We choose to use the arithmetic mean of four similarity metrics, two unweighted (cumulative association coefficient, or CACF and Jaccard index) and two weighted metrics (inversed document frequency, or IDF, and distinct partners, or DP), as described below.

\begin{align}
\text{Jaccard}(p_1, p_2) &= \frac{n_{1,2}}{(n_1 + n_2) - n_{1,2}} \quad (3) \\
\text{CACF}(p_1, p_2) &= \frac{2 \times n_{1,2}}{n_1 + n_2}; \quad (4) \\
\text{IDF}(S_1, S_2) &= 2 \times \frac{\text{WIDF}(S_1, S_2)}{\text{WIDF}(S_1) + \text{WIDF}(S_2)} \quad (5)
\end{align}

**Figure 4A.** MEME motifs that include the last transmembrane helix. Motif 6. The first diagram shows the sequence logo of all eukaryotic pLGIC probes that contain this motif, the second diagram shows the sequence logo of all non-eukaryotic proteins that contain this motif, the third diagram shows the alignment of the motif lifted from the eukaryotic pLGIC probes.
Let us illustrate with an example. Let us compare A0Q559_FRATN and ACHA7_HUMAN. The domain content of A0Q559_FRATN is IPR006028 IPR006029 IPR006201 IPR006202 IPR006203. The domain content of ACHA7_HUMAN is IPR006202 IPR006029 IPR006204 IPR018000 IPR002394. So, if we denote \( P_1 \) as A0Q559_FRATN; then the value of \( n_1 \) is 4 and the value of \( s_1 \) is [IPR006028 IPR006029 IPR006201 IPR006202]. \( P_2 \) denotes ACHA7_HUMAN; the value of \( n_2 \) is 5 and the value of \( s_2 \) is [IPR006202 IPR006201 IPR006029 IPR018000 IPR002394]. \( S_{1,2} \) is [IPR006029 IPR006202 IPR018000] and \( n_{1,2} \) is 3. In addition, we have these database statistics. Size of the database \( p = 7513312 \) proteins.

Using formula (3) to calculate the Jaccard similarity measure we obtain:

\[
\text{Jaccard}(P_1, P_2) = \frac{3}{((4 + 5) - 3)} = 0.5
\]

Using formula (4) to calculate CACF similarity measure we obtain:

\[
\text{CACF}(P_1, P_2) = \frac{(2 * 3)}{(4 + 5)} = 0.66667
\]

Using formulas (5) and (6) to calculate IDF similarity measure we obtain:

\[
\begin{align*}
\text{WIDF}(S_i) &= \sum \text{for each domain } d \text{ in } S_i \left( \log_2 \left( \frac{P}{n_{pd}} \right) \right) \\
\text{DP}(S_1, S_2) &= 2 \times \frac{\text{WDP}(S_{1,2})}{(\text{WDP}(S_1) + \text{WDP}(S_2))} \\
\text{WDP}(Si) &= \sum \text{for each domain } d \text{ in } S_i \left( \frac{1}{\text{partners}(d)} \right)
\end{align*}
\]

where \( P_1 \) and \( P_2 \) are proteins; \( n_1, n_2 \) is the number of domains shared by both proteins \( P_1 \) and \( P_2 \); \( n_1, n_2 \) is the number of domains of \( P_1 \); \( S_1 \) is set of domains in protein \( P_1 \); \( S_2 \), the set of domains in common between proteins \( P_1 \) and \( P_2 \); \( P \) is the size of the protein InterPro v25 database expressed as the number of sequences; \( n_{pd} \) is the number of proteins that contain domain \( d \) in the InterPro v25 database. Partners(d) is the number of domain signatures that contain domain \( d \).

Let us illustrate with an example. Let us compare A0Q559_FRATN and ACHA7_HUMAN. The domain content of A0Q559_FRATN is IPR006028 IPR006029 IPR006201 IPR006202 IPR006203. The domain content of ACHA7_HUMAN is IPR006202 IPR006029 IPR006204 IPR018000 IPR002394. So, if we denote \( P_1 \) as A0Q559_FRATN; then the value of \( n_1 \) is 4 and the value of \( s_1 \) is [IPR006028 IPR006029 IPR006201 IPR006202]. \( P_2 \) denotes ACHA7_HUMAN; the value of \( n_2 \) is 5 and the value of \( s_2 \) is [IPR006202 IPR006201 IPR006029 IPR018000 IPR002394]. \( S_{1,2} \) is [IPR006029 IPR006202 IPR018000] and \( n_{1,2} \) is 3. In addition, we have these database statistics. Size of the database \( p = 7513312 \) proteins.

Using formula (3) to calculate the Jaccard similarity measure we obtain:

\[
\text{Jaccard}(P_1, P_2) = \frac{3}{((4 + 5) - 3)} = 0.5
\]

Using formula (4) to calculate CACF similarity measure we obtain:

\[
\text{CACF}(P_1, P_2) = \frac{(2 * 3)}{(4 + 5)} = 0.66667
\]

Using formulas (5) and (6) to calculate IDF similarity measure we obtain:

\[
\begin{align*}
\text{WIDF}(S_i) &= \sum \text{for each domain } d \text{ in } S_i \left( \log_2 \left( \frac{P}{n_{pd}} \right) \right) \\
\text{DP}(S_1, S_2) &= 2 \times \frac{\text{WDP}(S_{1,2})}{(\text{WDP}(S_1) + \text{WDP}(S_2))} \\
\text{WDP}(Si) &= \sum \text{for each domain } d \text{ in } S_i \left( \frac{1}{\text{partners}(d)} \right)
\end{align*}
\]
The combined similarity score was normalized thus \[
\frac{(\text{Jaccard} + \text{CACF} + \text{IDF} + \text{DP})}{4}
\]. The results of the ranking obtained for each bacterial and archaeal protein against each one of the seven eukaryotic pLGIC subfamilies are shown in Table 2. Those rankings helped us group the bacterial and archaeal proteins into seven clusters. They go from G1 for the highest scoring proteins.

Using formulas (7) and (8) to calculate DP similarity measure we obtain:

\[
\begin{align*}
WDP(S_1) &= \frac{1}{59} + \frac{1}{113} + \frac{1}{114} + \frac{1}{129} = 0.042323 \\
WDP(S_2) &= 0.152614 \\
WDP(S_1, S_2) &= 0.25573
\end{align*}
\]

\[
\text{DP}(S_1, S_2) = 0.260326
\]

**Figure 5A and B.** Sequence logos of the MEME motifs common to Torpedo, ELIC and GLIC (A) motif 1 (B) motif 2.
The search methods. **BLASTp.** BLASTp is an alignment-based search tool for protein sequences. We used the 25 eukaryotic sequences of Table 1 as probes to search both NCBI RefSeq microbial and UniProt bacteria + archaea databases. The versions of BLASTp used were 2.2.18 for RefSeq and 2.2.22 for UniProt; the input parameters were the default values except for cutoff e-value of 1.0. All hits reported by the 25 probes were collected and the number of probes each hit appeared in was also calculated.

**psiBLAST.** psiBLAST is another sequence alignment-based search tool, like BLASTp, but it adjusts its weight matrix with each iteration it performs on the database. We used the same 25 eukaryotic probes that were used with BLASTp and the RefSeq-Microbial database (v 39). The input parameters were left unchanged with their default values except for cutoff e-value = 1 and number of iterations = 7. All hits reported by the 25 probes

![Sequence logos of the MEME motifs common to Torpedo, ELIC and GLIC (C) motif 3 (D) motif 4.](image-url)
set of domains it contains. Table 3 shows the list of the unique functional domains found in the probe set of the pLGIC family.

**The MEME-MAST ab-initio motifs.** We used the MEME motif discovery program (v 4)\(^9,10,12\) to find conserved motifs in the pLGIC family. As a training set for the MEME, we took the 25 eukaryotic pLGIC sequences in the TCDB 1.A.9 family (Table 1). We ran MEME with the following parameters: minimum length of motif set to 6, maximum length of motif set to 30 and maximum number of motifs set to 15. All other parameters were at the default setting.

To determine a motif-diagram signature by which to define the bacterial and archaeal members of the pLGIC family, we first looked for a short sequence of motifs (order preserving) that is common to most of the training sequences. These candidate motif-diagrams were then tested for specificity as follows. Using MAST (with the default setting for all parameters), we searched NCBI non-redundant protein database with the 15 MEME motifs. We then checked that the eukaryotic annotated sequences with the candidate motif-diagram signature belonged to the pLGIC family. We chose two motif-diagram signatures that also contained the strongest motifs. The motif-diagram signatures were: [4]-[1]-[2] and [1]-[2]-[3], where [n] stands for motif n. We then searched with MAST (using the original 15 motifs) the bacterial and archaeal databases and collected all sequences that had one of the two strongest motif-diagram signatures as a sub-diagram.

Figure 6A and B. (A) Domain network diagram of the InterPro domain sets represented among the eukaryotic pLGICs used as probes in this study. The 25 eukaryotic proteins comprise 15 different domain sets. In the domain network diagram, a connecting line is drawn between each pair of domain sets that have at least one domain in common. Domain sets are related to each other according to the number of domains they have in common, as described in the text. (B) Domain network diagram of the InterPro domain sets represented among the 69 non-eukaryotic pLGICs identified in this study.

were collected and the number of probes each hit appeared in was also calculated.

**The InterPro domains.** For each protein in the probe set of Table 1, we identified its functional domain composition using InterProScan and the InterPro database (release 25). Overlapping domains were included. We then represented each protein by the
Figure 6C and D. (C) Domain network diagram of the MEME motifs in the 25 eukaryotic pLGIC proteins. (D) Domain network diagram of the MEME motifs of the 69 non-eukaryotic pLGIC proteins.
Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
This work was supported by grant NSF 0835718 to Eric Jakobsson and Jeffrey L. Tilson. \textquoteleftHierarchical Modularity in Evolution and Function\textquoteright. We thank Professor Claudio Grosman for useful discussions that led to new insights in the revised version of the paper. We also thank the anonymous reviewers for useful suggestions.

Authors\' Contributions
Gloria Rendon participated in the design of the study, performed the searches with InterPro domains method, pBLAST and psi-BLAST, designed the scoring function, performed the validation tests, performed the structural and phylogenetic studies, evaluated the performance of the tools, prepared figures and tables and wrote parts of the manuscript.

Miriam R. Kantorovitz designed the study involving the ab initio motifs and performed the tests and statistical analysis, participated in the design of the methods and wrote parts of the manuscript. Jeffrey L. Tilson designed a high performance workflow for utilizing the InterPro domain analysis. Eric Jakobsson conceived of the study, participated in its design and wrote parts of the manuscript.

Table 5. Performance of the tools used to discover bacterial and archaeal pLGIC proteins

| Tool                      | Fp-true positives | Fp-signal noise | Fp-false negatives | Total hits | Sensitivity (tp/tp+fn) | Specificity (fp/fp+tp) | Size of database (total number of sequences) | Size of database (total number of letters) |
|---------------------------|-------------------|-----------------|--------------------|------------|------------------------|------------------------|---------------------------------------------|------------------------------------------|
| InterPro v25              | 49                | 19              | 20                 | 58         | 0.7101                 | 0.84482759            | 12440305                                    | n/a                                      |
| BLASTp - UniProt archaea v15 | 2                | 144             | 20                 | 146        | 1                      | 0.01369863            | 211367                                      | 59414885                                  |
| BLASTp - UniProt bacteria v15 | 53               | 137             | 14                 | 190        | 0.7910                 | 0.27894737            | 6635810                                     | 2025601752                                |
| BLASTp - Refseq microbial v39 | 40               | 1               | 29                 | 41         | 0.5797                 | 0.97560976            | 4775066                                     | 1481264742                                |
| psiBLAST - Refseq microbial v39 | 54               | 414             | 15                 | 464        | 0.7826                 | 0.11538462            | 4775066                                     | 1481264742                                |
| psiBLAST - UniProt archaea v15 | 2                | 31              | 0                  | 33         | 1                      | 0.06060606            | 211367                                      | 59414885                                  |
| psiBLAST - UniProt bacteria v15 | 62               | 118             | 5                  | 1180       | 0.9254                 | 0.05254237            | 6635810                                     | 2025601752                                |
| MEME-MAST - Refseq microbial v39 | 43               | 0               | 26                 | 43         | 0.6323                 | 1                     | 4775066                                     | 1481264742                                |
| MEME-MAST - UniProt microbial v15 | 48               | 0               | 21                 | 50         | 0.6957                 | 1                     | 6847177                                     | 2085016637                                |

References
1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997; 25:3389-402.
2. Taneen A, Iyer I, Jakobsson E, Aravind L. Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. Genome Biology 2004; 6:4.
3. Boquoi N, Prado de Carvalho L, Carraud J, Neyton J, Le Poucon C, Tal Y, et al. A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. Nature 2007; 445:116-9.
4. Hill RJCC, Dzurlar R. X-ray structure of a prokaryotic pentameric ligand-gated ion channel. Nature 2008; 452:375-9.
5. Hunter S, Aspweiler R, Artwood TK, Bairoch A, Baretan A, Binnis D, et al. InterPro: the integrative protein signature database. Nucl Acids Res 2009; 37:213-5.
6. Mulder N, Aspweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 2007; 396:59-79.
7. Mulder NJ, Aspweiler R, Artwood TK, Bairoch A, Baretan A, Binnis D, et al. New developments in the InterPro database. Nucl Acids Res 2007; 35:224-8.
8. Bailey TL. Discovering sequence motifs. Bioinformatics 2008;231-51.
9. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clements L, et al. MEME SUITE: tools for motif discovery and searching. Nucl Acids Res 2009; 37:202-8.
10. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning 1995; 21:51-80.
11. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics 1998; 14:48-54.
12. Bailey TL. The value of prior knowledge in discovering motifs with MEME. Proceedings of the Third International Conference on Intelligent Systems for Molecular biology: AAAI Press 1995; 21-9.
13. Bailey TL. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: AAAI, Ed. Proc Int Conf Intell Syst Mol Biol AAAI Publishing, Menlo Park, CA USA 1994; 8.
14. Schaffer AA, Aravind L, Madden TL, Shavrin S, Spouge JL, Wolf YI, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-specific gap penalties and weight matrix choice. Nucl Acids Res 1994; 22:4673-80.
15. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 1977; 74:5088-90.
16. Lin K, Zhu L, Zhang D Y. An initial strategy for comparing proteins at the domain architecture level. BMC Bioinformatics 2006; 22:2081-6.
17. Song N, Sedgewick RD, Darand D. Domain architecture comparison for multidomain homology identification. J Computational Biol 2007; 14:496-516.
18. Tilson JL, Rendon G, Gier MF, Jakobsson E. MotifNetwork: A grid-enabled workflow for high-throughput domain analysis of biological sequences: implications for annotation and study of phylogeny, protein interactions and intraspecies variation. BIBE 2007 Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, Boston, MA: IEEE 2007; 620-7.
19. Bailey TL, Yen MR, Noto K, Tamang DG, Elkan C. The transporter classification database: recent advances. Nucl Acids Res 2009; 37:274-8.

Supplemental materials can be found at: www.landesbioscience.com/journals/channels/article/16822
24. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 2001; 305:567-80.

25. Lee A. Measures of Distributional Similarity. Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics. University of Maryland, College Park, MD USA: Association of Computational Linguistics 1999; 25-32.