University of California, San Diego

Bioinformatics Analysis of the MACPF Superfamily

A Thesis submitted in partial satisfaction of the requirements

for the degree Master of Science

in

Biology

by

Bennett Vitug

Committee in charge:

Professor Milton H. Saier Jr., Chair
Professor Nigel Crawford
Professor Maarten Chrispeels

2012

The Thesis of Bennett Vitug is approved and is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2012

Table of Contents

List of Tables

List of Figures

ABSTRACT OF THE THESIS

Bioinformatics Analysis of the MACPF Superfamily

by

Bennett Vitug
Master of Science in Biology
University of California, San Diego, 2012
Professor Milton H. Saier, Jr., Chair

The Membrane Attack Complex/Perforin (MACPF) superfamily consists of a diverse group of proteins from three families involved in eukaryotic immunity, embryonic development, neural migration and bacterial pathogenesis. Characterization of the MACPF family involved recognition of possible orthology and horizontal gene transfer. Phylogenetic analysis of MACPF homologues using bioinformatics methods revealed a remarkably diverse range of proteins spanning both bacterial and eukaryotic kingdoms, with significant variations in the topological, hydrophobic and amphipathic characteristics of their sequences.

The MACPF superfamily was expanded through the addition of the Cholesterol-Dependent Cytolysin (CDC) family. Comparison of the primary and tertiary structures of homologues from these two families revealed sequence similarity in the transmembrane regions of both families. Phylogenetic analysis

demonstrated exclusive clustering of the CDC homologues, thereby identifying it as the second family within the MACPF superfamily.

The third family to be included in the MACPF family was the Pleurotolysin (Pleurotolysin) family. Comparison of Pleurotolysin homologues from TCDB with the homologues obtained from the MACPF and CDC families revealed 15 pairs of proteins with comparison scores greater than 12 S.D. in their respective transmembrane domains. Addition of the pleurotolysin proteins to the phylogenetic tree containing MACPF and CDC homologues showed clustering of the majority of pleurotolysins.

Introduction

Throughout the past two decades, our bioinformatics laboratory has been involved in the identification of over six hundred families of transport proteins while expanding the Transporter Classification Database, TCDB (Saier *et al.* 2006; Saier *et al*. 2009). Although similar to the Enzyme Commission (EC) system for classifying enzymes, the TC system incorporates functional and phylogenetic information which provides the basis for family classification. The classification of transport protein systems is thus based on structural, functional and evolutionary characteristics (Saier *et al.* 2000; Busch & Saier *et al.* 2002).

As discussed in this thesis, the MACPF superfamily consists of pore-forming, cytolytic proteins that are important in both mammalian immunity, embryonic development, neural migration, tumor suppression and prokaryotic toxicity (Anderluh & Lakey, 2008; Estévez-Calvar *et al*. 2011). As shown here, three families compose the MACPF superfamily: the Membrane Attack Complex/Perforin (MACPF) Family (TC# 1.C.39), the Cholesterol-Dependent Cytolysin (CDC) Family (TC# 1.C.12) and the Pleurotolysin Pore-Forming (Pleurotolysin) Family (TC# 1.C.97). Using a common MACPF domain, proteins associated with the membrane attack complex (MAC) and the protein perforin control microbial invasion of the host through pathogen lysis via formation of a C5b-9 pore complex, a process known as C3-mediated opsonization (Wang *et al*, 2000). Other apextrin-like proteins containing the MAC domain are known to play a role in the larval development of eukaryotic organisms, such as the sea urchin,

Heliocidaris erhthrogramma, and the Mediterranean mussel, *Mytilus galloprovincialis* (Haag *et al*. 1999; Estévez-Calvar *et al*. 2011). Furthermore, the MACPF proteins, DBCCR-1 and BRINP-1, are believed to function in both tumor suppression and neural development (Kawano *et al.* 2004; Wright *et al.* 2004).

X-ray structure analysis of the MACPF domain for complement C8α and Plu-MACPF from *Photorhabdus luminescens* showed structural similarity with the bacterial, pore-forming, cholesterol-dependent cytolysins (CDCs) (Hadders *et al*. 2007; Rosado *et al.* 2007). Both families share a common mechanism of membrane insertion as two regions refold into transmembrane ß hairpins to form the lining of the barrel pore (Xu *et al.* 2010). Thus, it has been suggested that lytic MACPF proteins may share a mechanism similar to CDCs in forming pores and disrupting cell membranes (Law *et al.* 2010; Rossi *et al*. 2010). However, the authors of the papers describing the 3-D structures of these proteins claimed that CDC and MACPF show no detectible similarity at the primary sequence level.

Members of the Pleurotolysin Pore-Forming Family have been shown to exhibit cytolytic activity through pore formation in human erythrocytes (Sakurai *et al.* 2004). Pleurotolysins are two-component hemolysins which require the interaction of both non-associated components to exhibit strong cytolytic activity (Shibata *et al.* 2010). Cooperative pore formation causes leakage of potassium ions from cells and subsequent colloid-osmotic hemolysis (Tomita *et al.* 2004). Although the longer Pleurotolysin B protein exhibits similar three-dimensional folds with members of the MACPF superfamily, NCBI BLAST results suggest that

Pleurotolysin A is a member of the Aegerolysin superfamily and may be distantly related to members of the Equitoxin family, TC #1.C.38 (Shogomori *et al.* 2008).

Our study seeks to expand the MACPF family and to demonstrate sequence similarity between the active pore-forming regions of the MACPF, CDC and Pleurotolysin families. The advent of three technological improvements have made it possible to identify increasingly distance homologues using sequence similarity as the primary means. We first found representatives of the major phylogenetic clusters in each family. Second, we identified proteins that may represent ancestral links between these families. Third, we increased the numbers of homologues available for analysis, which allowed us to broaden the scope of sequence diversity due to the availability of ever increasing amounts of genomic sequence data. Fourth, the availability of increasingly sensitive software allowed us to compare more distant homologs of each family. Finally, application of the superfamily principle allowed us to demonstrate homology between each family using "missing link" homologues.

The superfamily principle was originally used to establish homology between distantly related members of extensive superfamilies (Doolittle, 1981). In our study, the superfamily principle was carried out by first establishing sequence similarity throughout the length of proteins or relevant protein domains within a single family. The transmembrane sequences of proteins belonging to different families were statistically compared. If two proteins from two different families showed homology in their transmembrane regions, then it is not

necessary to establish homology for the transmembrane sequences of every protein in the two families.

Although structural studies have shown the MACPF and CDC families to be functionally and structurally similar, sequence similarity between transmembrane regions had never previously been established. The current dogma is that one can detect homology (common ancestry) more reliably using tertiary structure rather than primary structure. We conducted these studies in an attempt to disprove this dogma by showing that while others may not have been able to find sequence similarity, it does in fact exist using the approach detailed above.

It is well known that many proteins can exist in more than one highly dissimilar conformational states. Sometimes these divergent conformations are unrecognizable at the three-dimensional level. For example, prion proteins can typically exist in "native" α-states but can also assume cleaved β-states (Mangé *et al.* 2004). Several soluble proteins with recognized catalytic and structural properties can insert in membranes, forming ion-conducting channels (Anderson & Blaustein, 2008). Toxins are often made in a soluble state, which can then insert within the membranes of target organisms forming pores that result in cytoplasmic leakage and cell death (Czajkowsky *et al.* 2004). In all such cases, massive conformational changes occur. It is therefore clear that reliance on three dimensional (X-ray and NMR) data cannot be considered the preferred approach to establishing homology. Statistical approaches using primary sequence data

may still be the most reliable means to establish the common origin of distantly related macromolecules including proteins and nucleic acids.

Our study establishes homology between the transmembrane regions of these families. It establishes the Pleurotolysin Pore-Forming family to be the third member of the MACPF superfamily. Statistical and phylogenetic analyses, multiple alignments, and hydropathy plots of the three families have revealed the diversity of the MACPF superfamily.

Methods

Representatives of the MACPF superfamily (TCID 1.C.39) were compiled from the Transporter Classification Database (www.tcdb.org). In order to study the distant members of the superfamily, the compilation of MACPF representative proteins was expanded with putative members of the MACPF superfamily proteins by performing Position-Specific Iterated BLAST (PSI-BLAST) searches against NCBI's non-redundant (NR) protein database. Our lab has established that performing a protein PSI-BLAST with a cut-off value of e^{-4} and a subsequent iteration with a cut-off value of e^{-5} consistently retrieves homologues with few false positives. Data from the BLAST searches were organized based on abbreviation of the protein name, description, sequence length, gi number, organismal source and phylum by running the resultant TinySeq XML files through the MakeTable5 program. A file containing the FASTA formatted sequences of all putative MACPF superfamily proteins and 16S/18S rRNA sequences for most genera were also obtained. Additional rRNA sequences were obtained using the NCBI Nucleotide Database. Only full-length proteins were kept, and protein redundancies and close sequences were minimized using the CD-HIT program with a cut-off value of 70%. The proteins included in this study are listed in Tables 1-3 for the MACPF, CDC and Pleurotolysin families, respectively.

Throughout this study, multiple alignments for each family and individual protein clusters were generated using the ClustalX program (Thompson et. al.,

1997). Multiple alignments and their corresponding phylogenetic trees allowed us to elucidate the existence of possible fused domains within exceptionally long protein sequences. By using protein BLAST to analyze the unaligned regions, we were able to determine if additional domains were accountable for the length of these sequences.

Phylogenetic trees corresponding to the multiple alignments for each family or cluster were created using ClustalX and viewed using the TreeView or FigTree program (Zhai & Saier, 2002). Phylogenetic trees allowed us to identify specific clusters in each family, and the subsequent analysis of each cluster revealed the similarities between members of each cluster in terms of organismal source and sequence length. Furthermore, analysis of the phylogenetic tree created using the 16S and 18S rRNA sequences of all genera in Table 1 allowed us to identify possible horizontal gene transfer and orthologs between our MACPF proteins.

The multiple alignments generated with ClustalX were used in the Average Hydropathy, Amphipathicity and Similarity (AveHAS) program to generate an averaged hydropathy plot for multiple related proteins. The Web-based Hydropathy, Amphipathicity and Topology (WHAT) program was used to generate a hydropathy plot for single proteins (Zhai & Saier, 2001). Both programs provide graphical depictions of hydrophobic, hydrophilic and amphipathic regions throughout the length of the protein. Furthermore, both

programs predict any transmembrane segments (TMS) that may be present within the protein.

To determine homology between the three families, the collection of proteins from the MACPF, CDC and Pleurotolysin families was statistically compared to each other using the SSearch program. The SSearch program analyzes two lists of proteins, indicates regions of similarity and provides the corresponding comparison scores expressed in standard deviations (S.D.). Sequences with scores of 7 standard deviations or greater were confirmed and optimized by first isolating the regions of the sequences that were found to be similar by SSearch and subsequently running them on GAP with 500 random shuffles to ensure the reliability of the scores. A value of 10 standard deviations using GAP was considered sufficient for establishing homology.

Sequence similarity between the MACPF and CDC families was further optimized by analyzing the three-dimensional structures of the proteins that exhibited high standard deviation values from SSearch. The homologous sequences were visualized in the program, PyMOL, using representative PDB files from the Protein Data Bank (PDB) to confirm that the homologous sequences were positioned in the transmembrane regions of the respective proteins. ClustalX and GAP were again used to generate alignments of the representative PDB sequences with the sequences of interest from SSearch and its homologues. The CDC protein from each MACPF-CDC pair was compared with the sequence of the PDB protein model, PDB# 1PFO (Rossjohn *et al*. 1997).

The region where the CDC aligned with both the MACPF protein and 1PFO sequence was colored in PyMOL, thereby showing whether the residues compared were included in the transmembrane region. The same method was utilized using the PDB protein model, PDB #2RD7 (Slade *et al*. 2008), for each MACPF protein in each MACPF-CDC pair.

The binary alignment from GAP was also superimposed on the PDB protein model using the program, ConSurf. ConSurf calculates the amino acid conservation scores through the empirical Bayesian or the Maximum Likelihood method along different sites of the protein and visually modifies the original protein model to reflect the varying degrees of conservation (Mayrose *et al*. 2004, Landau *et al*. 2005, Glaser *et al*. 2003).

The SuperfamilyTree program (SFT) was used as the final step in the phylogenetic analysis of the MACPF, CDC, and Pleurotolysin families. Similar to our use of ClustalX and the FigTree program, this program can determine the phylogenetic relationships between families, subfamilies, and individual proteins through BLAST bit-scores and larger protein samplings (Chen *et al.* 2011, Yen *et al.* 2009, Yen *et al.* 2010). This program was used to confirm whether clear segregation of these families occurred as predicted in our phylogenetic trees generated from multiple alignments. Representative proteins from each family in TCDB were used to generate a final MACPF superfamily tree.

Chapter 1: Characterization of the MACPF Family

Extraction of MACPF Homologues

A systematic method was employed for compiling a list of homologues for each MACPF representative in the Transporter Classification Database (www.tcdb.org) (Table 1). The FASTA formatted sequence of a MACPF representative, such as Complement Protein C9 (TC# 1.C.97.1.1) was first obtained from TCDB, and a subsequent protein PSI-BLAST was performed on the NCBI NR protein database. A second iteration was performed for proteins with e-values of less than e^{-4}, and a list of potential homologues was compiled in FASTA format. This process was done for each MACPF representative in TCDB, and the lists of FASTA sequences were combined. Fragmented and redundant sequences were eliminated using the CD-HIT program with a cutoff of 70%. A multiple alignment of the combined list of proteins was then made using the Clustal X program, and a phylogenetic tree was generated (Figure 1).

Phylogenetic Tree Analysis by Cluster

The phylogenetic tree that was generated based on the multiple alignment allowed us to analyze the putative homologues by cluster and expand the MACPF family in TCDB using a representative protein from each cluster. The average sequence length and its standard deviation value was recorded for each cluster without omitting proteins with fused domains and especially long sequences. The phylum and domain of each protein's respective organism was

also recorded. Furthermore, large proteins in each cluster were analyzed in terms of additional protein domains (Table 4).

Clusters were also analyzed using the AveHAS program. Although the program predicted potential transmembrane sequences, the predicted regions usually corresponded with hydrophobic peaks outside of the MACPF domain and sometimes with little conservation for the proteins in an individual cluster. Studies of MACPF and perforin proteins, however, suggested that helical conformations of specific regions in the MACPF domain could insert into the bilayer membrane. The AveHAS plot (Figures 6 to 19) revealed that the MACPF domains for each cluster were highly conserved and significantly more amphipathic than other regions, leading us to believe that the transmembrane region for each cluster is actually present in the MACPF domain.

Cluster 1:

Cluster 1 contained MACPF homologues with an average sequence length of 583 ± 100 residues. All proteins in Cluster 1 belong to metazoans. Although some proteins were either unnamed or predicted, the majority of the proteins were alpha or beta subunits of complement component 8. The proteins in this cluster were shown to be homologous to the MACPF representative with TC# 1.C.39.3.1 in the Transporter Classification Database (www.tcdb.org).

Analysis of the Average Hydropathy, Amphipathicity and Similarity (AveHAS) plot of Cluster 1 revealed relatively higher conservation from positions

275 to 710 and from positions 875 to 1040 of the multiple alignment (Figure 5). Analysis of the plot also revealed that the majority of proteins in Cluster 1 were amphipathic throughout much of the alignment. Hydrophobicity varied throughout the alignment, although two poorly conserved hydrophobic regions were identified from positions 25 to 60 and positions 240 to 298 of the multiple alignment.

Tgu5 (GI# 224058308) was the longest protein in Cluster 1 with a sequence length of 972 amino acids. Analysis of the protein revealed additional domains not found in other members of the cluster. Residues 1 to 71 were shown to be homologous to the conserved domain, Topoisomerase II-associated protein PAT1 (CDD# pfam09770), which is necessary for accurate chromosome transmission during cell division (Wang *et al*, 1996). Residues 719 to 902 were shown to be an adjacent repeat of the MACPF domain (CDD# pfam01823).

Cluster 2:

The MACPF homologues from cluster 2 had an average sequence length of 731 ± 258 amino acids. All proteins in cluster 2 belong to metazoans. The majority of these proteins were either hypothetical proteins or predicted to be similar to complement component 6. BLAST searches against the TCDB database showed that the proteins in this cluster were most similar to the MACPF subfamilies, TC# 1.C.39.1 and 1.C.39.3.

The AveHAS program predicted two well conserved transmembrane regions from positions 5 to 25 and positions 40 to 60 (Figure 6). Most of the proteins in cluster 2 were characteristically amphipathic throughout their sequences, although a well conserved peak of hydrophobicity was found that corresponded to the first predicted TMS, positions 5 to 25. A second hydrophobic peak was found from positions 675 to 705, which corresponded to a third poorly conserved predicted transmembrane sequence for hypothetical proteins that belonged to the organism *Branchiostoma floridae* (GI# 219503573, 219409896, 219443754, 219492604).

Analysis of a protein belonging to the organism, *Branchiostoma floridae*, revealed a possible fused region at the C-termini of Bfl30 (GI# 219431797). A protein BLAST of this region showed it to be homologous to the conserved domain, DNA Polymerase III subunits gamma and tau (CDD# PRK12323) from residues 723 to 1264.

Two large proteins from *Ciona intestinalis*, Cin5 (GI# 198417017) and Cin7 (GI# 198419275), were also found to contain six additional Thrombospondin Type-1 Repeat domains (CDD# smart00209). This domain is known to bind and activate TGF- β (Transforming Growth Factor β), which plays a role cell proliferation and differentiation (Casalena *et al.* 2012). Abnormalities with activation of TGF-β is known to underlie various developmental disorders and pathologies including cancer and autoimmune diseases (Casalena *et al*. 2012).

Cluster 3:

Cluster 3 contains MACPF homologues with an average sequence length of 567 ± 53 amino acids. All proteins contained within this cluster were from metazoans and most similar to complement component 9. BLAST searches were performed against the TCDB database and showed that these proteins were similar to the MACPF subfamilies, TC# 1.C.39.1.

The AveHAS program revealed a sharp peak of hydrophobicity at positions 25 to 50 of the multiple alignment, which was conserved throughout half of the proteins in cluster 3 (Figure 7). This hydrophobic peak corresponds to the only transmembrane region that was predicted by the program.

Cluster 4:

The MACPF homologues in cluster 4 were shown to have an average sequence length of 960 ± 355 amino acids. All proteins in this cluster belong to metazoans and are similar to complement component 6 (TC# 1.C.39.3.2). An exception to this was the hypothetical protein, Oan 1 (GI# 149634247), which appeared to be most similar to the MACPF representative with TC# 1.C.39.1.1 using TC-BLAST.

AveHAS analysis of this cluster revealed substantial hydrophobicity from positions 1 to 25 and 480 to 510 in the multiple alignment (Figure 8). The predicted TMS of the cluster corresponded to the first hydrophobic peak at the N-termini of the proteins.

The significant variation in length of the protein, Clu1 (GI# 73954287), suggested fusion of an extra domain. A protein BLAST search against the NCBI database was performed, and the extra region at the C-terminus of the protein was found to be homologous to the protein isoform hCG1993037: CRA_F of *Homo sapiens* (GI# 119602545).

Cluster 5:

The MACPF homologues in cluster 5 were found to have an average length of 753 ± 5 amino acids. Proteins in this cluster belong to metazoans and resemble complement component 7 (TC# 1.C.39.3.2).

AveHAS analysis revealed the proteins in this cluster to be largely amphipathic throughout most of their lengths (Figure 9). A sharp hydrophobic peak was predicted from residues 60 to 100 of the multiple alignment and corresponded with the putative TMS of the cluster. This predicted TMS, however, was shown to be conserved only amongst half of the proteins in the cluster.

Cluster 6:

Cluster 6 consists of two metazoan proteins with an average length of 559 ± 5 amino acids. Both proteins were described as complement components and were similar to the MACPF representative with TC# 1.C.39.3.1.

AveHAS analysis of the cluster revealed two significant hydrophobic regions at the N- and C-termini of the proteins. The program predicted the TMS

for this cluster to correspond with the hydrophobic peak at the C-terminus of both proteins.

Cluster 7:

Cluster 7 contained MACPF homologues with an average sequence length of 572 ± 133 residues. All proteins in this cluster belong to metazoans and were similar to the lymphocyte pore-forming protein, perforin 1. Proteins in this cluster were found to be similar to the MACPF representative with TC# 1.C.39.2.1 in TCDB.

AveHAS plot analysis of the cluster showed a hydrophobic peak that was highly conserved at the N-terminus of each protein (positions 60 to 85) (Figure 10). This peak corresponded to the predicted TMS for the cluster. A second poorly conserved hydrophobic peak from residue 1 to 25 was found only in Tni9 (GI# 47218949).

The significantly larger length of Tni9 (GI# 47218949) and the gaps in the multiple alignment suggested the presence of additional domains. Following a protein BLAST of the sequence against the NCBI database, the protein was found to contain two conserved tryptophan domains (CDD# cd00201)from positions 8 to 38 and a PPIC-type PPIase rotamase domain from positions 52 to 131 (CDD# pfam00639). The two conserved tryptophan domains are known to bind proline-rich motifs and are important in various cytoplasmic signal transduction proteins and structural proteins (Ermekova *et al.* 1997). Rotamases

encoded by the PPIA gene in humans and are known to accelerate the rate of protein folding by catalyzing cis-trans isomerization (Haendler & Hofer 1990; Holzman *et al*. 1991). Analysis of the unaligned region at the C-terminus of Tni9 with CDD did not reveal additional conserved domains.

Cluster 8:

Cluster 8 consisted of only two MACPF homologues from fungi with no known functions. A protein from *Emericella nidulans,* Eni1 (GI# 168091), was the product of the gene, SpoC-C1C, which has been used for DNA hybridization experiments (Stephens *et al.* 1999). Although the gene had no known function, it was predicted that it may play a role in transcriptional regulation in dormant spores (Stephens *et al.* 1999). The two proteins had an average sequence length of 612 ± 235 residues. A protein BLAST against the TCDB database showed that both proteins belong to the MACPF subfamily, TC# 1.C.39.9.

Analysis of the AveHAS graph revealed a high degree of conservation of the MACPF domain from positions 350 to 750 in the multiple alignment. The program predicted three possible TMSs from position 110 to 190 of the multiple alignment. These predicted TMSs, however, were only present in Ani1 (GI# 67537830).

Cluster 9:

The MACPF homologues in cluster 9 have an average sequence length of 609 ± 5 residues. All proteins in this cluster originated from organisms in the

phylum, Viridiplantae. Protein BLAST searches against the TCDB database showed low similarity scores with the MACPF representatives, TC# 1.C.39.6.1, 1.C.39.1.2, and 1.C.39.10.1. These proteins were thus incorporated into a new subfamily, TC# 1.C.39.11.

The AveHAS plot for cluster 9 revealed multiple hydrophobic peaks, with the most distinct peak occurring between positions 300 and 310 of the multiple alignment (Figure 11). Despite multiple peaks of hydrophobicity, the program did not predict transmembrane regions. Proteins from this cluster showed moderate peaks of amphipathicity throughout the lengths of their sequences with high conservation.

Cluster 10:

Cluster 10 consists of only two bacterial MACPF homologues; a hemopexin-like protein from *Plesiocystis pacific* SIR-1 and a complement-like protein from *Beggiatoa* sp. PS. The average sequence length of the two proteins in this cluster was 521 ± 7 residues. Domain analysis showed that the MACPF domain spans the proteins from positions 160 to 305 while Hemopexin-like repeats occurs from positions 317 to 512. Together, the two proteins compose the TCDB subfamily, 1.C.39.8.

AveHAS plot analysis of the cluster did not reveal putative transmembrane regions. No significant peaks of hydrophobicity were detected, although three moderate peaks were found between positions 200 and 325 of the binary

alignment. The plot shows a sharp peak of amphipathicity at alignment position 250. This peak of amphipathicity corresponds to the putative transmembrane region of the MACPF domain.

Cluster 11:

Cluster 11 contains MACPF homologues with an average sequence length of 1183 ± 417 residues. The protein, Bfl1 (GI# 219460616), is recognizably longer than the other proteins in this cluster with a length of 2433 amino acids. The longer length is attributed to the addition of a C-type lectin domain at the C-terminus (CDD# cd00037), a GCC2 and GCC3 domain (CDD# pfam07699), an eel-Fucolectin Tachylectin-4 Pentaxtrin-1 domain (CDD# smart00607), a scavenger receptor Cys-rich domain (CDD# smart00202), and furin-like repeats (CDD# cd00064). The eel-Fucolectin Tachylectin-4 Pentaxtrin-1 domain binds to cell-surface carbohydrates and are known to play a role in innate immunity (Honda *et al.* 2000). The GCC2_GCC3 domain is found in a variety of extracellular proteins, however, the function is unknown (Araki *et al*. 2011). The scavenger receptor Cys-rich domain is involved with the recognition of low-density lipoproteins, and is usually expressed in membrane-bound secreted proteins of the immune system (Holm *et al.* 2012). The furin-like repeats domain is a part of a family that contains endoproteases and cell-surface receptors (Molloy *et al*. 1999). Furin is a calcium-dependent serine endoprotease that cleaves and catalyzes the maturation of various proprotein subtrates, such as growth factors, receptors and pathogen proteins (Molloy *et al.* 1999). The C-type

lectin domain requires calcium to bind carbohydrates and is involved in cell to cell adhesion, immune response to pathogens and apoptosis (Elgavish & Shaanan, 1997; Holmskov *et al*. 1994). All of the proteins in cluster 11 were derived from metazoans. BLAST searches against the TCDB database showed that these proteins belong to the MACPF subfamily, TC# 1.C.39.5.

Analysis of the AveHAS plot revealed variable degrees of hydrophobicity and amphipathicity (Figure 12). A sharp peak of hydrophobicity that was conserved among half of the cluster 11 proteins occurred from alignment positions 75 to 90. A second better conserved hydrophobic peak occurred from positions 1190 to 1210. Amphipathicity was highly variable, with the sharpest well conserved peaks occurring around positions 175, 340, 850 and 925. The program predicted two clear TMSs, the first at positions 75 to 90, and the second at positions 450 to 485.

Cluster 12:

The MACPF homologues in cluster 12 had an average sequence length of 675 ± 290 residues. Proteins in this cluster originate from protists. TC-BLAST showed that these proteins belong to the MACPF subfamily, TC# 1.C.39.6.

AveHAS analysis of the cluster revealed significant hydrophobic peaks from alignment positions 1 to 25 and positions 780 to 800 (Figure 13). The predicted transmembrane regions correspond to the hydrophobic peaks from

positions 5 to 25 and positions 220 to 255. The MACPF domain, which encompasses residues 124 to 329, is relatively amphipathic.

Tan1 (GI# 85001526) is significantly longer than the other proteins in this cluster. Observation of conserved domains in this protein revealed the presence of three full length MACPF domains that span the protein from residues 172 to 304, 438 to 652 and 990 to 1212.

Cluster 13:

The MACPF homologues in cluster 13 had an average sequence length of 558 ± 77 residues. Proteins in this cluster are from metazoans and are described as being similar to apextrin. TC-BLAST searches showed that these proteins belong to the MACPF subfamily, TC# 1.C.39.7.

AveHAS analysis of the cluster revealed only one significant hydrophobic peak from positions 1 to 25 in the multiple alignment (Figure 14). The program predicted a transmembrane region that corresponded with this hydrophobic peak. Amphipathicity was fairly high for these proteins.

Cluster 14:

Cluster 14 contained MACPF homologues with an average sequence length of 793 ± 538 residues. The majority of the proteins in this cluster are from the protist, Oligohymenophorea, although one protein, Ami1 (GI# 118153966), is from a metazoan. BLAST searches against the TCDB database demonstrated that these proteins belong to the MACPF subfamily, TC# 1.C.39.7.

AveHAS analysis of cluster 14 revealed one significant hydrophobic peak from positions 1 to 25 (Figure 15). The program predicted a transmembrane region that corresponds to this hydrophobic peak. The MACPF domain, which encompasses residues 144 to 328, is more amphipathic than the rest of the protein.

Tth5 (GI# 118371656) is significantly longer than other members of cluster 14. Domain analysis using the Conserved Domain Database revealed the presence of a discontinuous P-Type ATPase-V domain (CDD# TIGR01657) from residue 562 to 1370 and residue 1521 to 1809 on Tth5. The function of this domain is unknown, however, it is found in many eukaryotes and is believed to be involved in cation transport in the endoplasmic reticulum (Axelsen & Palmgren, 1998). Further analysis showed the presence of an E1-E2_ATPase domain (pfam00122) from residues 745 to 1007.

Cluster 15:

Cluster 15 contains two bacterial MACPF homologues from Chlamydiae with an average sequence length of 610 ± 281 residues. A BLAST search against TCDB showed relatively low similarity with 1.C.39.10.1 and 1.C.39.6.1. These proteins were thus incorporated into a new MACPF subfamily, TC# 1.C.39.12. Both proteins were shown to contain a MAC/perforin domain unique to members of the Chlamydiae phylum.

AveHAS analysis of the cluster revealed relatively high peaks of amphipathicity throughout the lengths of both proteins with varying degrees of hydrophobicity. The largest peak of hydrophobicity, shared between the two proteins, occurs at alignment positions 575 to 600. The program did not predict transmembrane regions.

The MACPF domain in both proteins spanned 195 residues. A BLAST search of the longer protein, Cmu1 (GI# 15835049), did not show additional domains. Cpn1 (GI# 15618100), however, showed an additional domain, MIR (CDD# smart00472), near the C -terminus from residues 366 to 409. This domain may function as a ligand transferase, and is present in ryanodine receptors, inositol triphosphate receptors and in protein O-mannosyltransferases (Ponting *et al.* 2000).

Cluster 16:

The bacterial MACPF homologues in cluster 16 had an average sequence length of 480 ± 85 residues. Proteins in this cluster are from Bacteroides. Analysis of the MACPF domain of each protein revealed low similarity scores to the MACPF representatives, TC# 1.C.39.3.2, 1.C.39.4.1, 1.C.39.5.1, and 1.C.39.11.1. These proteins were thus incorporated into a new MACPF subfamily, TC# 1.C.39.13.

AveHAS analysis of the cluster revealed one well-conserved peak of hydrophobicity occurring at alignment positions 20 to 30 (Figure 16).

Amphipathicity varied throughout the lengths of the sequences, although the protein, Bun1 (GI# 160888542) showed a significant peak of hydrophobicity that occurred at positions 640 to 660.

Cluster 17:

Cluster 17 contains MACPF homologues from both eukaryotic and bacterial domains. The cluster consists of proteins from fungi, γ-proteobacteria and actinobacteria. The average sequence length of these proteins is 563 ± 148 residues. A BLAST search against TCDB revealed that bacterial proteins in this cluster are most similar to the MACPF subfamily, TC# 1.C.39.4. The single protein from fungi was incorporated into a new TCDB subfamily, TC# 1.C.39.14.

AveHAS analysis of the cluster revealed no significant peaks of hydrophobicity (Figure 17). A well-conserved peak of amphipathicity occurred at alignment position 660. The transmembrane region was predicted to be from positions 175 to 200, corresponding to a single peak of hydrophobicity and one of amphipathicity.

Cluster 18:

The MACPF homologues from cluster 18 are from a diverse range of organisms from both the bacterial and eukaryotic domains. Proteins in this cluster originate from fungi, mycetozoa, γ-proteobacteria and cyanobacteria. Surprisingly, protein BLAST against TCDB revealed the bacterial proteins, Ter1 and Msp1, to be most similar to the MACPF subfamily, TC# 1.C.39.4, while the

eukaryotic proteins are more similar to members of the Pleurotolysin family, TC# 1.C.97.2.1 and 1.C.97.3.1. Proteins in this cluster were found to have an average sequence length of 623 ± 168 residues.

AveHAS analysis of this cluster revealed two significant peaks of hydrophobicity that were centered at positions 250 and 350 in the multiple alignment (Figure 18). A single well-conserved peak of amphipathicity was found at position 590. No transmembrane region was predicted.

Chapter 1.1: Orthology, Paralogy, and Horizontal Gene Transfer Among MACPF Family Proteins

A tree (Figure 5) was constructed using the complete 16S and 18S rRNA sequences of all genera in our list of MACPF homologues (Table 1). This unrooted tree was produced from a ClustalX multiple alignment using the neighbor-joining method and the FigTree program. Distinct clustering of eukaryotic and bacterial genera is apparent and shows clear segregation of genera based on phylum. The largest cluster consists of 18S rRNA sequences from metazoans. This cluster segregates and forms its own branch opposite the eukaryotic phyla Viridiplantae, Fungi, Oligohymenophorea, Mycetozoa, and Apicomplexa. The eukaryotic genera omitted from the rRNA tree due to the unavailability of complete 18S rRNA sequences include: *Pongo, Macaca, Canis, Felis,Tetraodon, Oryctolagus, Ginglymyostoma, Takifugu, Ctenopharyngodon,* and *Acropora*. Proteins from these genera were not considered in predicting orthology within each of the clusters.

Observation of the smaller bacterial branch of the 16S/18S rRNA tree also shows clustering based on phylum. The phyla represented in this branch consists of Bacteroidetes, Actinobacteria, Chlamydiae, Cyanobacteria, γ-proteobacteria, and δ-proteobacteria. Genera from the γ-proteobacteria phylum compose the largest cluster, which is adjacent to both a δ-proteobacterium and a cluster containing Cyanobacteria, Chlamydiae, and Actinobacteria.

Orthology and evidence of horizontal gene transfer were identified by comparing clustering patterns in the 16S/18S rRNA tree and the MACPF family protein tree. Potential horizontal gene transfer events were more common in clusters containing bacterial proteins, and thus, orthologous relationships were observed less frequently.

Cluster 1 consists of a large and complex group of Metazoan proteins that can be divided into two sub-clusters. Like Cluster 4, Cluster 1 contains orthologs of *Xenopus*: Xla4 (GI# 147901003) from *Xenopus laevis* and Orf3 (GI# 53749700) from *Xenopus (Silurana) tropicalis*. Two MACPF homologs from *Homo sapiens* are present in one of the sub-clusters and are likely paralogs. Three homologs from *Mus musculus* are present in this cluster. Two of these proteins branch closely together in one sub-cluster, consistent with paralogy. The proteins in this sub-cluster correspond to the order of genera in the 16S/18S rRNA tree, consistent with orthology. The third *Mus* protein is positioned in the other sub-cluster, which also contains proteins from genera that corresponds to the Metazoan cluster in the 16S/18S rRNA tree. The genera, *Ginglymostoma* and *Canis*, were excluded from our rRNA tree, and thus, the assumption that these proteins are orthologous cannot be made with certainty.

Cluster 2 can be divided into two sub-clusters. One sub-cluster consists of twelve paralogs from *Ciona intestinalis* while the other contains seven paralogs of the genus, *Branchiostoma*. The majority of proteins in the *Branchiostoma* sub-cluster come from the species, *Branchiostoma floridae*. One protein in the sub-

cluster, however, belongs to *Branchiostoma belcheri*, suggesting orthology between this protein and one of the *B. floridae* proteins in this sub-cluster. Comparison of the two genera in this cluster with the 16S/18S rRNA tree shows that these proteins may be orthologous as they are situated close to each other in both trees.

Cluster 3 contains the proteins Orf1 (GI# 166796971) and Xla2 (GI# 148233806) from *Xenopus (Silurana) tropicalis* and *Xenopus laevis*, respectively. These two proteins cluster closely together in the phylogenetic protein tree and are possibly orthologous. The remaining proteins in this cluster appear to be orthologous as well with the exception of proteins from the genera *Takifugu, Tetraodon, Ctenopharyngodon, Canis,* and *Macaca,* which were excluded from the 16S/18S rRNA tree.

Cluster 4 contains two paralogous proteins from *Xenopus laevis,* which form a branch at a point after divergence from other protein branches. The cluster also contains two non-adjacent paralogs from *Canis lupus*. The proteins in this cluster, with the exception of those from *Canis lupus*, appear to be orthologous since the genera in this cluster correspond with the order that was found in the 16S/18S rRNA tree.

Cluster 5 contains proteins from various Metazoans. Two proteins from *Mus musculus*, two from *Danio rerio*, two from *Tetraodon nigroviridis*, and three from *Rattus norvegicus* cluster closely together and are likely to be paralogs. The proteins Bta2 (GI# 114051808) from *Bos taurus*, Ssc1 (GI# 47523630) from

Sus scrofa, Eca4 (GI# 194223929) from *Equus caballus*, and Hsa2 (GI# 194389200) from *Homo sapiens* were found to cluster together in the protein tree, and their genera corresponded with the order of the 16S/18S rRNA tree, suggesting that these proteins are orthologous. Similarly, the two proteins, Pol2 (GI# 6682831) from *Paralicthys olivaceus* and Omy7 (GI# 185133218) from *Oncorhynchus mykiss*, cluster together in both trees, again suggesting orthology.

Cluster 6 contains only two proteins from Metazoa, Cin4 (GI# 198433282) from *Ciona intestinalis* and Hro1 (GI# 224176461) from *Halocynthia roretzi*. The close clustering of these two proteins in the protein tree and the adjacent branches of their respective genera in the 16S/18S rRNA tree suggest that these two proteins are orthologous.

Cluster 7 is a more complex group of proteins from various Metazoans. Seven proteins from *Danio rerio* are present in this cluster, suggesting paralogy. Five proteins from *Tetraodon nigroviridis* are also present, suggesting paralogy between these proteins as well. The proteins, Xla3 (GI# 148237294) from the genus *Xenopus*, Oan6 (GI# 149472392) from *Ornithorhynchus*, and Gga2 (GI# 118099091) from *Gallus* are located in close proximity to each other, corresponding to their positions on the 16S/18S rRNA tree and suggesting orthology. The proteins, Pol4 (GI# 30519828) from *Paralichthys* and Omy4 (GI# 198442831) from *Oncorhynchus*, may also be orthologous to each other due to their adjacent positions in both the protein and 16S/18S rRNA tree. A final set of potential orthologous proteins, Cja1 (GI# 197112111) from *Callithrix*, Bta1 (GI#

219522060) from *Bos*, Ssc2 (GI# 194042762) from *Sus*, Eca6 (GI# 194205976) from *Equus*, Mmu3 (GI# 200290) from *Mus*, and Rno3 (GI# 149038739) from *Rattus*, are also located in this cluster.

Cluster 8 contains only two proteins from fungi, Ani1 (GI# 67537830) from *Aspergillus nidulans* and Eni1 (GI# 168091) from *Emericella nidulans*. These proteins cluster tightly in our phylogenetic protein tree, corresponding to the branches for *Emericella* and *Aspergillus* in the 16S/18S rRNA tree. Thus, these proteins are likely orthologs.

Cluster 9 consists of proteins from the phylum, Viridiplantae. Two proteins from *Vitis vinifera* are likely to be paralogs. Divergence of the *Vitis* protein, Vvi1 (GI# 157358723) from the *Medicago* protein, Mtr1 (GI# 92870237), occurs after the gene duplication event that resulted in the *Vitis* paralog, Vvi2 (157354261). Thus, Vvi2 is excluded from the orthologous relationship shared Mtr1 and Vvi1. Ptr1 (GI# 224069581) from *Populus*, however, is orthologous to both *Vitis* proteins and Mtr1, since divergence occurs prior to the gene duplication event that gave rise to both *Vitis* paralogs and the species divergence of *Medicago*.

Cluster 10 consists of two proteins from the γ-proteobacteria, *Beggiatoa sp. PS*, and the δ-proteobacteria, *Plesiocystis pacifica SIR-1*. These two proteins cluster tightly together in the phylogenetic protein tree, but are distantly related in the 16S/18S rRNA tree, thereby suggesting that these proteins arose due to horizontal gene transfer.

Cluster 11 contains multiple paralogs from the organism, *Branchiostoma floridae*, and three paralogs from *Nematostella vectensis*. Early divergence of the branches that show the relationships of these paralogs from *Nematostella* and *Branchiostoma* and the distance between these two organisms in our 16S/18S rRNA tree suggest that some of these proteins may have resulted from an early horizontal gene transfer event.

Cluster 12 consists of seven proteins from the phylum, Apicomplexa, and another three proteins from Oligohymenophorea, suggesting trans-phylum horizontal gene transfer. A closer look at the cluster reveals tight clustering of the proteins, Tth1 (GI# 118368397), Tth2 (GI# 118369627), and Tth4 (118366533), from *Tetrahymena*, indicative of paralogy. Four proteins, Pfa1 (GI# 124505319), Pbe1 (GI# 56805561), Pkn1 (GI# 221052646), and Pvi1 (GI# 156094597) from *Plasmodial* species may also be orthologous. These these four proteins divide into their respective branches from a point after divergence from other Apicomplexa proteins and the branching patterns correspond closely to the 16S/18S rRNA tree. This may therefore indicate orthology between the proteins from *Babesia, Theileria,* and *Plasmodium* in this cluster.

Cluster 13 contains a tight cluster of seven apextrin-like proteins from the organism, *Strongylocentrotus purpuratus*. The diversity of these proteins are likely to be a product of late gene duplication events, giving rise to paralogues.

Cluster 14 contains proteins from the genera, *Tetrahymena, Acropora, and Paramecium*. Five proteins from the organism, *Tetrahymena thermophila*

SB210, show tight clustering and are probable paralogs. A single protein from *Paramecium tetraurelia strain d4-2* branches out near the center of the tree and is surprisingly distant from the *Tetrahymena* proteins. The branching pattern suggests closer phylogenetic similarity between this *Paramecium* protein, Pte1 (GI# 145475565), and a protein from a Metazoa, Ami1 (GI# 118153966), which suggests the lack of orthology for these homologues. Furthermore, the presence of a single Metazoan protein among proteins from Oligohymenophorea suggest trans-phylum horizontal gene transfer.

Cluster 15 shows two proteins from the bacterial phylum, Chlamydiae. The genera, *Chlamydia* and *Chlamydophila,* branch closely together in both protein and 16S/18S rRNA trees, thereby suggesting orthology.

Cluster 16 contains eight proteins from the genus, *Bacteroides*. The clustering of three proteins from *Bacteroides fragilis* and two from *Bacteroides cellulosilyticus* suggests paralogy within this cluster. Bfr2 (GI# 53712858) and Bfr3 (GI# 53713977) form a branch prior to Bce1 (GI# 224536709) showing, that these three proteins are not orthologous. The five remaining proteins in this cluster are likely to be orthologs.

Cluster 17 also displays a potential horizontal gene transfer event as a distant branch of the fungal protein, Pma1 (GI# 212532427), is present among proteins from γ-proteobacteria and Actinobacteria.

Cluster 18 consists mostly of distantly related proteins from fungi. The presence of the Cyanobacterial homologue, Ter1 (GI# 113474643), and the γ-proteobacterium, Msp1 (GI# 87122061), in this cluster may be evidence of horizontal gene transfer. Similarly, a protein derivative of Mycetozoa, Ddi1 (GI# 66805335), is present in this cluster and may have resulted from an early horizontal gene transfer event.

Chapter 2: Homology Between the MACPF Family and Cholesterol-Dependent Cytolysin (CDC) Family

Members of the MACPF and CDC families contain structurally similar transmembrane domains in the form of two α-helices with amphipathic character (Rosado *et al*. 2008). Inclusion of the CDC family into the MACPF superfamily was dependent upon showing sequence similarity between the transmembrane domains of both families. A list of CDC (Table 2) and MACPF homologues was screened for similarities with SSearch. Many pairs of MACPF and CDC homologues were found to be similar within their respective domains and were further analyzed with GAP (Figures 22 to 30). GAP comparisons showed these pairs to have comparison scores as high as 14.4 standard deviations, which by our criteria is sufficient to establish homology (Table 5). The three highest comparison scores in our study came from the following MACPF-CDC pairs: Rno6 & Cte1 (123 residues compared with a comparison score of 14.4 S.D.), Clu7 & Cbo5 (272 residues compared with a comparison score of 13.3 S.D.), and Ami1 & Bbr1 (214 residues compared with a comparison score of 12.9 S.D.).

These pairs were then analyzed in terms of sequence similarity within the regions that comprise the transmembrane alpha helices. Structural data were necessary to determine whether the sequences of the MACPF and CDC pairs corresponded to their respective transmembrane domains. The protein structures, PDB# 2RD7 for a MACPF homologue and PDB# 1PFO for a CDC,

were utilized as previous research efforts had revealed the putative transmembrane region for these proteins.

Three-dimensional visualization suggested the MACPF and CDC pairs to be homologous within the sequences of their transmembrane domains. BLAST2 and GAP results for the MACPF homologues versus the sequences of their respective PDB model revealed that four of the ten pairs listed in Table 4 could be further analyzed.

Superimposing and color coding the GAP alignments on the 1PFO and 2RD7 models revealed that all four pairs of MACPF and CDC homologues were similar in regions that either fully or partially encompassed one of two transmembrane helices (Table 7). Comparison of the MACPF protein, Omy3, with the CDC proteins, Cbo2 and Cno1, showed that TMH1 in the MACPF structure, 2RD7, is similar to TMH2 in the CDC protein structure, 1PFO (Figures 31, 32, 35 and 36). The comparison of the MACPF protein, Tth1, with the CDC protein, Cte1, also showed that TMH1 in 2RD7 is similar to TMH2 in 1PFO (Figures 34 and 38). Conversely TMH2 on 2RD7 is also similar to TMH1 in 1PFO in our comparison of the MACPF protein, Spu6, with the CDC protein, Cno2 (Figures 33 and 37).

Use of the ConSurf program further demonstrated the degree of conservation between the MACPF and CDC superfamilies. The program utilized the multiple sequence alignments of the previous pairs of MACPF and CDC homologues to construct a phylogenetic tree. From the phylogenetic tree,

position-specific conservation scores were calculated through the program's empirical Bayesian algorithm. The resultant scores were then visualized in the 1PFO protein model. Moderate to high conservation of the amino acid sequence was observed along the transmembrane helices shared between the MACPF and CDC pairs: Omy3 & Cbo2 and Omy3 & Cno1 (Figures 39 and 40).

Chapter 2.1: Expanding the MACPF Phylogenetic Tree with the Cholesterol-Dependent Cytolysin (CDC) Family

Once sequence similarity between transmembrane regions of CDC and MACPF proteins was established through use of a combination of SSearch, GAP and three-dimensional visualization, we made certain that the CDC proteins formed a specific branch on our phylogenetic tree. Our CDC proteins (Tables 2) were added to the original list of MACPF proteins (Tables 1) and a multiple sequence alignment was obtained. The alignment was then used to formulate a new phylogenetic tree (Figure 2).

The CDC cluster contained proteins with an average sequence length of 532 ± 47 residues. All proteins belong to Gram-positive and Gram-negative bacteria from the phyla Firmicutes, Actinobacteria, Bacteroides, and β-proteobacteria. All proteins in this cluster are exotoxins that require the presence of cholesterol for pore formation. Furthermore, these proteins fall within the 1.C.12.1 sub-family.

An AveHAS plot of the CDC proteins was generated using the multiple alignment (Figure 20). A poorly conserved peak of hydrophobicity was present at alignment position 20 to 50. The multiple alignment showed that this peak was only present in one protein; the human platelet aggregation factor, Smi1 (GI# 84579714). Further analysis of the sequence using NCBI's CDD showed that residues 53 to 178 of Smi1 were similar to the F5/8 Type C domain (pfam00754), which is also known as the discoidin (DS) domain family. This conserved domain

is a coagulation factor that is a part of the FA58C superfamily. The FA58C superfamily consists of cell surface-attached carbohydrate-binding domains that may have been horizontally transferred from eukaryotes to eubacterial genomes (Baumgartner *et al.* 1998).

The highest degree of conservation between the CDC proteins in our phylogenetic tree occurred from alignment position 280 to position 765. Further analysis of this region using CDD showed that the cholesterol-binding thiol-cytolysin (pfam01289) domain was highly conserved throughout these proteins.

Chapter 3: Homology between the MACPF, CDC, and Pleurotolysin Families

Members of the Pleurotolysin family consist of two-component hemolytic proteins that cooperatively assemble into a membrane pore on human erythrocytes (Sakurai *et al*. 2004, Bernheimer & Avigad *et al.* 1979). PSI-BLAST searches of representative Pleurotolysin proteins in TCDB showed that the Pleurotolysin A components belong to the Aegerolysin superfamily. The Pleurotolysin B components and other pleurotolysin-like representative proteins in TCDB were shown to be members of the MACPF superfamily through use of SSearch.

TCDB representative proteins for the Pleurotolysin family were used in the comparison (Table 3). SSearch standard deviation values greater than 12 S.D. in regions with 60 amino acid residues or more that corresponded with the MACPF or CDC domain demonstrated the inclusion of the Pleurotolysin Family in the MACPF superfamily. 68 pairs of MACPF and pleurotolysin proteins were found to have comparison scores greater than 12 S.D. (Table 8). The SSearch comparison scores between CDCs and Pleurotolysins showed that 30 pairs had scores greater than 12 S.D. (Table 9). Furthermore, high identities were observed between Pkn1 (GI# 221052646) of the Pleurotolysin family and members of the MACPF family. Fps1 (GI# 150024210) and Nsp1 (GI# 17228824) also showed high identities with the CDC family, suggesting possible revision of the TCDB representative proteins for the MACPF superfamily.

The MACPF proteins, Nfi1 (GI# 119499704), Afl2 (GI# 220689182), Gze1 (GI# 46126573), Afl1 (GI# 220688529), and Afl3 (GI# 220693297) also showed high identities with the Pleurotolysin proteins, Pos1 (GI# 54312024), Per1 (GI# 261857452), Cgl1 (GI# 116202857), Cli1 (189345610), and Dis1 (GI# 66805335), respectively. This suggested that these MACPF proteins, which were obtained through a PSI-BLAST search of MACPF representative proteins from TCDB against the NCBI protein database, may actually be members of the Pleurotolysin family.

A phylogenetic tree containing the MACPF, CDC and Pleurotolysin families was used to determine whether these three putative protein families actually represented three distinct branches on the tree and to determine whether revision of TCDB representative proteins was necessary based on the Pleurotolysin SSearch data. The resultant tree (Figure 3) showed Pkn1 clustering with the MACPF family's Group 12 homologues, corresponding to the high comparison scores from SSearch. Fps1 and Nsp1 also showed clustering with the CDC family as predicted by the high comparison scores obtained using SSearch. Pkn1 was thus reassigned as the MACPF family, 1.C.39.6.1, while Fps1 and Nsp1 were reassigned as CDC representative proteins (1.C.12.2.1 and 1.C.12.3.1, respectively).

The phylogenetic tree also showed the clustering patterns of the MACPF proteins, Nfi1, Afl2, Gze1, Afl1 and Afl3 to be consistent with the SSearch data. These proteins were shown to form a cluster with the Pleurotolysin family, while

the remaining proteins that formed Group 18 of our MACPF protein list (Table 1) formed clusters with other members of the MACPF family. As a result, only seventeen out of the original eighteen MACPF protein clusters can be observed with the addition of the Pleurotolysin family to our phylogenetic tree.

SSearch comparison scores between the revised lists of MACPF, CDC and Pleurotolysin homologues showed 15 pairs of MACPF and Pleurotolysin homologues with comparison scores greater than 12 S.D. in regions greater than 50 residues that contain the MACPF domain (Table 10). Comparison scores greater than 12 S.D. between the smaller sampling of Pleurotolysin and CDC homologues were not observed.

Phylogenetic analysis of the MACPF superfamily was continued using the SuperfamilyTree program (SFT). Using the revised MACPF, CDC, and Pleurotolysin representative proteins from each subfamily in TCDB, a new phylogenetic tree (Figure 4) was generated based on BLAST comparison scores rather than multiple alignments. The tree showed distinct branching of the MACPF, CDC, and Pleurotolysin proteins, confirming segregation between the three families that constitute the MACPF superfamily. Gze1 was included as a representative of the five MACPF proteins that were found to cluster with other Pleurotolysin proteins in our previous tree to confirm its reassignment as a Pleurotolysin protein. The new tree obtained from SFT showed Gze1 clustering with the pleurotolysin representative protein, 1.C.97.2, thus confirming its inclusion in the Pleurotolysin family.

The pleurotolysin cluster consisted of proteins with an average sequence length of 612 ± 237 residues. The five proteins that formed the pleurotolysin branch in our phylogenetic tree belong to fungi and mycetozoa from the eukaryotic domain and chlorobia from bacteria. Pkn1, which was reassigned as a MACPF family protein, belongs to the phylum, apicomplexa, from the eukaryotic domain while Fps1 and Nsp1, which were reassigned as CDC proteins, are from bacteroidetes and cyanobacteria, respectively, from the bacterial domain. A TC-BLAST of the five pleurotolysin proteins show that they are most similar to the 1.C.97.1, 1.C.97.2, 1.C.97.3, 1.C.97.5, and 1.C.97.6 subfamilies.

An AveHAS plot was generated using a multiple alignment of the five pleurotolysin proteins (Figure 21). The highest degree of similarity between these five proteins occurred from alignment positions 345 to 640, 668 to 698, 738 to 794, and 825 to 863. A peak of both hydrophobicity and amphipathicity occurred from positions 700 to 735. However, this was only present in one protein, Cli1 (GI# 189345610). Another peak of hydrophobicity occurred from positions 50 to 70, which was present only in the protein, Cgl1 (GI# 116202857). A third peak of hydrophobicity occurred from positions 960 to 1000 and was present in both Cli1 and Cgl1. Peaks of amphipathicity occurred from positions 200 to 240, 420 to 425, and 450 to 470.

Discussion

In this paper, we have characterized the MACPF superfamily by analyzing three families, and their sequence similarities with one another have been evaluated. The MACPF family was expanded through the collection of homologues from NCBI, and the diversity of the family was defined through the creation of multiple alignments and phylogenetic trees. Eighteen clusters were analyzed based on the phylogenetic tree that was generated from the multiple alignment of our compiled list of MACPF proteins. As a result, multiple subfamilies were added to the MACPF entry in TCDB based on the data gathered from clustering patterns in our phylogenetic tree and data from our comparison of MACPF transmembrane sequences using SSearch and GAP. Further analysis using a 16S/18S rRNA tree, based on the genera from which our proteins were obtained, showed that horizontal gene transfer was more widespread in the bacterial proteins while orthology was common among the eukaryotic proteins.

The Cholesterol-Dependent Cytolysin (CDC) family was compared with the MACPF family by analyzing sequence similarity through a combination of SSearch and GAP. The comparison scores from SSearch of ten MACPF and CDC protein pairs was optimized using GAP, yielding scores as high as 14.4 standard deviations. This was sufficient in establishing homology between the MACPF and CDC families. The phylogenetic tree that was generated using our original list of MACPF proteins and our list of CDC proteins was analyzed for the

clustering patterns of the CDCs, which confirmed their identity as a separate family. Based on sequence similarity between the TMSs and clustering in the phylogenetic tree, the CDC family was added to the MACPF superfamily entry in TCDB.

Analysis of the CDC family was continued by comparing the primary and tertiary structures of the MACPF and CDC proteins that were analyzed using SSearch and GAP. Many of the pairs with high comparison scores partially contained the MACPF/CDC domain, and it was therefore necessary to confirm that the compared sequences contained their respective transmembrane regions. Four MACPF and CDC protein pairs were analyzed using PyMOL and ConSurf. We found that each pair is similar in one of two transmembrane helices. The comparison of Omy3 with Cbo2, Omy3 with Cno1, and Tth1 with Cte1 showed that TMH1 in the MACPF protein is similar to TMH2 in the CDC protein. Conversely, the comparison of Spu6 with Cno2 showed that TMH2 in the MACPF protein is similar to TMH1 in the CDC protein. Through our study of primary structure, we determined that the MACPF and CDC families share not only structural similarity, but also sequence similarity in their transmembrane regions.

The Pleurotolysins were the final family to be analyzed in our study of the MACPF superfamily. Although the functional Pleurotolysin pore-forming complex consists of two components, only the B component was compared with the CDC and MACPF families. The smaller A component was found to be a part of the

Aerogolysin superfamily through a protein PSI-BLAST search on NCBI. SSearch was again used to compare the Pleurotolysin proteins with our list of MACPF and CDC proteins. We found 68 MACPF/Pleurotolysin pairs and 30 CDC/Pleurotolysin pairs with comparison scores greater than 12 S.D.. The significantly high comparison scores of Pkn1, Fps1, and Nsp1 suggested possible reassignment of these proteins as members of the MACPF or CDC families in TCDB. This was confirmed by generating a phylogenetic tree based on a multiple alignment of all three MACPF families. Pkn1 was found to form a cluster with the MACPF family while Fps1 and Nsp1 formed a cluster with the CDC family. Pkn1, Fps1 and Nsp1 were therefore assigned the TC numbers 1.C.39.6.1, 1.C.12.2.1, and 1.C.12.3.1, respectively. Comparison of the revised list of MACPF, CDC and Pleurotolysin homologues showed 15 pairs of MACPF/Pleurotolysin proteins with comparison scores greater than 12 S.D. in regions that spanned more than 50 residues and contained the MACPF domain.

Phylogenetic analysis of the MACPFs, CDCs and Pleurotolysins was continued in order to confirm their identity as three distinct families. Using a tree generated from the SuperfamilyTree program, we were able to identify distinctive branching and clear clustering of the proteins in each family. Furthermore, we confirmed the reassignment of five proteins (Nfi1, Afl2, Gze1, Afl1, and Afl3) from our original list of MACPF homologues to the Pleurotolysin family by utilizing Gze1 as a representative protein and observing its clustering patterns with 1.C.97.2 of the Pleurotolysin family.

It is interesting to note that the MACPF superfamily is well represented in the bacterial and eukaryotic domains, but not a single member has so far been found in archaea. This fact correlates that pathogenic archaea seem not to exist, or may be extremely rare. The reason for this surprising observation has yet to be clarified.

Appendix

Table 1. All homologues from the MACPF family that were included in our study are listed by the clock-wise order in which they appear on our phylogenetic tree. These homologues were obtained by a PSI-BLAST search with the TCDB representative protein, 1.C.12.1.1 as the query sequence with two iterations. The proteins are organized by cluster, and their abbreviations, protein descriptions, organismal sources, sequence lengths, GenInfo Identifier (GI) numbers, phyla, domains, and TCDB sub-families are provided.

Table 1: MACPF Family Homologues

Abbreviation	Protein Description	Organism	Protein Size	GI Number	Phylum	Domain	TCDB Sub-Family
Group 1							
Oan5	PREDICTED: similar to Complement component 8, alpha polypeptide, partial	Ornithorhynchus anatinus	517	149453920	Metazoa	Eukaryota	1.C.39.3.1
Gci1	complement component 8 alpha polypeptide	Ginglymostoma cirratum	589	157072526	Metazoa	Eukaryota	1.C.39.3.1
Mdo1	PREDICTED: similar to Complement component 8, alpha polypeptide	Monodelphis domestica	775	126306055	Metazoa	Eukaryota	1.C.39.3.1
Mmu2	complement component 8, alpha polypeptide, isoform CRA_c	Mus musculus	584	148698883	Metazoa	Eukaryota	1.C.39.3.1
Rno1	complement component 8, alpha polypeptide	Rattus norvegicus	587	157821013	Metazoa	Eukaryota	1.C.39.3.1
Ocu1	complement component 8, alpha polypeptide	Oryctolagus cuniculus	585	126723297	Metazoa	Eukaryota	1.C.39.3.1
Ssc3	complement component C8A	Sus scrofa	589	147905213	Metazoa	Eukaryota	1.C.39.3.1
Bta4	complement component 8, alpha polypeptide	Bos taurus	589	114053319	Metazoa	Eukaryota	1.C.39.3.1
Clu6	PREDICTED: similar to complement component 8, alpha polypeptide precursor	Canis lupus familiaris	589	73956396	Metazoa	Eukaryota	1.C.39.3.1
Eca5	PREDICTED: similar to Complement component 8, alpha polypeptide	Equus caballus	631	194207404	Metazoa	Eukaryota	1.C.39.3.1
Hsa4	complement component 8, alpha polypeptide, isoform CRA_a	Homo sapiens	586	119627049	Metazoa	Eukaryota	1.C.39.3.1
Mmu7	PREDICTED: similar to complement component 8, alpha polypeptide precursor	Macaca mulatta	584	109005036	Metazoa	Eukaryota	1.C.39.3.1
Hsa3	Chain A, Structure Of C8a-Macpf Reveals Mechanism Of Membrane Attack In Complement Immune Defense	Homo sapiens	334	158430383	Metazoa	Eukaryota	1.C.39.3.1
Tgu3	PREDICTED: similar to MGC80388 protein	Taeniopygia guttata	597	224058304	Metazoa	Eukaryota	1.C.39.3.1
Gga5	PREDICTED: similar to Complement component 8, alpha polypeptide	Gallus gallus	603	118094713	Metazoa	Eukaryota	1.C.39.3.1
Xla4	MGC80388 protein	Xenopus laevis	589	147901003	Metazoa	Eukaryota	1.C.39.3.1
Orf3	complement component 8, alpha polypeptide	Xenopus (Silurana) tropicalis	584	53749700	Metazoa	Eukaryota	1.C.39.3.1
Dre7	novel protein similar to vertebrate complement component 8, alpha polypeptide (C8A, zgc:92465)	Danio rerio	592	157886430	Metazoa	Eukaryota	1.C.39.3.1
Omy1	complement component C8 alpha chain	Oncorhynchus mykiss	615	185133500	Metazoa	Eukaryota	1.C.39.3.1
Tni8	unnamed protein product	Tetraodon nigroviridis	593	47227989	Metazoa	Eukaryota	1.C.39.3.1
Mdo3	PREDICTED: similar to complement component C8 beta subunit	Monodelphis domestica	631	126306053	Metazoa	Eukaryota	1.C.39.3.1
Rno2	C8b protein	Rattus norvegicus	430	59808937	Metazoa	Eukaryota	1.C.39.3.1
Mmu6	complement component 8, beta subunit	Mus musculus	523	123229562	Metazoa	Eukaryota	1.C.39.3.1
Mmu13	C8b protein	Mus musculus	493	18381134	Metazoa	Eukaryota	1.C.39.3.1
Ocu2	complement component 8, beta polypeptide	Oryctolagus cuniculus	590	130501029	Metazoa	Eukaryota	1.C.39.3.1
Ssc4	complement component C8B	Sus scrofa	611	148235610	Metazoa	Eukaryota	1.C.39.3.1
Bta3	complement component 8, beta polypeptide	Bos taurus	590	194097363	Metazoa	Eukaryota	1.C.39.3.1
Mmu14	PREDICTED: complement component 8, beta polypeptide	Macaca mulatta	591	109005039	Metazoa	Eukaryota	1.C.39.3.1
Clu4	PREDICTED: similar to Complement component C8 beta chain precursor (Complement component 8 beta subunit)	Canis lupus familiaris	590	73956394	Metazoa	Eukaryota	1.C.39.3.1
Eca3	PREDICTED: similar to Complement component C8 beta chain precursor (Complement component 8 subunit beta)	Equus caballus	590	149694327	Metazoa	Eukaryota	1.C.39.3.1
Gga3	PREDICTED: similar to Complement component C8 beta chain precursor (Complement component 8 subunit beta)	Gallus gallus	584	118094715	Metazoa	Eukaryota	1.C.39.3.1
Tgu5	PREDICTED: disabled homolog 1	Taeniopygia guttata	972	224058308	Metazoa	Eukaryota	1.C.39.3.1
Dre4	PREDICTED: wu:fi18d09	Danio rerio	562	125805154	Metazoa	Eukaryota	1.C.39.3.1
Omy5	complement component C8 beta	Oncorhynchus mykiss	587	185132590	Metazoa	Eukaryota	1.C.39.3.1
Tni6	unnamed protein product	Tetraodon nigroviridis	590	47227988	Metazoa	Eukaryota	1.C.39.3.1
Pol3	RecName: Full=Complement component C8 beta chain; AltName: Full= Complement component 8 subunit beta; Flags: Precursor	Paralichthys olivaceus	588	20138051	Metazoa	Eukaryota	1.C.39.3.1
Pfl1	complement component C8 beta chain precursor	Perca flavescens	334	156454695	Metazoa	Eukaryota	1.C.39.3.1
Average Sequence Length	583						

Table 1. continued

Standard Deviation of Sequence Length	100						
Group 2							
Cin8	PREDICTED: similar to complement component C6	Ciona intestinalis	562	198420086	Metazoa	Eukaryota	1.C.39.3.1
Cin1	PREDICTED: similar to complement component C6	Ciona intestinalis	566	198431887	Metazoa	Eukaryota	1.C.39.3.1
Cin9	PREDICTED: similar to complement component C6	Ciona intestinalis	563	198431885	Metazoa	Eukaryota	1.C.39.3.1
Cin11	PREDICTED: similar to complement component C6	Ciona intestinalis	569	198421450	Metazoa	Eukaryota	1.C.39.3.1
Cin2	PREDICTED: similar to complement component C6	Ciona intestinalis	564	198421452	Metazoa	Eukaryota	1.C.39.3.2
Cin10	PREDICTED: similar to complement component C6	Ciona intestinalis	567	198421454	Metazoa	Eukaryota	1.C.39.3.1
Cin6	PREDICTED: similar to complement component C6	Ciona intestinalis	421	198419273	Metazoa	Eukaryota	1.C.39.1.1
Cin13	PREDICTED: similar to complement component C6	Ciona intestinalis	569	198419271	Metazoa	Eukaryota	1.C.39.3.1
Cin3	PREDICTED: similar to complement component C6	Ciona intestinalis	575	198419277	Metazoa	Eukaryota	1.C.39.1.1
Cin7	PREDICTED: similar to HyTSR1 protein	Ciona intestinalis	1108	198419275	Metazoa	Eukaryota	1.C.39.3.1
Cin5	PREDICTED: similar to thrombospondin type 1 repeat containing protein	Ciona intestinalis	1167	198417017	Metazoa	Eukaryota	1.C.39.1.1
Cin12	PREDICTED: similar to complement component C6	Ciona intestinalis	1045	198417019	Metazoa	Eukaryota	1.C.39.1.1
Bfl14	hypothetical protein BRAFLDRAFT_81773	Branchiostoma floridae	699	219443754	Metazoa	Eukaryota	1.C.39.3.2
Bfl4	hypothetical protein BRAFLDRAFT_110173	Branchiostoma floridae	722	219503573	Metazoa	Eukaryota	1.C.39.3.2
Bfl28	hypothetical protein BRAFLDRAFT_65208	Branchiostoma floridae	481	219409896	Metazoa	Eukaryota	1.C.39.3.2
Bfl7	hypothetical protein BRAFLDRAFT_105963	Branchiostoma floridae	949	219494025	Metazoa	Eukaryota	1.C.39.3.1
Bbe1	complement component C6	Branchiostoma belcheri	921	13928546	Metazoa	Eukaryota	1.C.39.3.1
Bfl30	hypothetical protein BRAFLDRAFT_76018	Branchiostoma floridae	1264	219431797	Metazoa	Eukaryota	1.C.39.3.2
Bfl35	hypothetical protein BRAFLDRAFT_134616	Branchiostoma floridae	583	219492604	Metazoa	Eukaryota	1.C.39.3.1
Average Sequence Length	731						
Standard Deviation of Sequence Length	258						
Group 3							
Oan2	PREDICTED: similar to complement component 9	Ornithorhynchus anatinus	430	149634257	Metazoa	Eukaryota	1.C.39.1.1
Omy2	complement component C9	Oncorhynchus mykiss	601	185133255	Metazoa	Eukaryota	1.C.39.1.2
Fhe1	complement component C9	Fundulus heteroclitus	577	40457916	Metazoa	Eukaryota	1.C.39.1.2
Pol1	complement component C9	Paralichthys olivaceus	558	6429127	Metazoa	Eukaryota	1.C.39.1.2
Tni7	unnamed protein product	Tetraodon nigroviridis	484	47228394	Metazoa	Eukaryota	1.C.39.1.2
Tru1	RecName: Full=Complement component C9; Flags: Precursor	Takifugu rubripes	586	2499468	Metazoa	Eukaryota	1.C.39.1.2
Cid1	complement component C9	Ctenopharyngodon idella	650	125661173	Metazoa	Eukaryota	1.C.39.1.1
Dre9	complement component 9	Danio rerio	673	220941693	Metazoa	Eukaryota	1.C.39.1.2
Tgu1	PREDICTED: similar to complement protein C9	Taeniopygia guttata	618	224090365	Metazoa	Eukaryota	1.C.39.1.1
Mdo4	PREDICTED: similar to complement protein C9	Monodelphis domestica	523	126321671	Metazoa	Eukaryota	1.C.39.1.1
Clu7	PREDICTED: similar to Complement component C9 precursor	Canis lupus familiaris	589	73954295	Metazoa	Eukaryota	1.C.39.1.1
Ocu3	complement component 9	Oryctolagus cuniculus	557	126723572	Metazoa	Eukaryota	1.C.39.1.1
Eca2	complement protein C9 precursor	Equus caballus	547	126352550	Metazoa	Eukaryota	1.C.39.1.1
Hsa1	complement component 9, isoform CRA_b	Homo sapiens	567	119576392	Metazoa	Eukaryota	1.C.39.1.1
Mmu12	PREDICTED: complement component 9	Macaca mulatta	561	109077053	Metazoa	Eukaryota	1.C.39.1.1
Bta5	complement component 9	Bos taurus	548	78369352	Metazoa	Eukaryota	1.C.39.1.1
Ssc5	complement component C9	Sus scrofa	543	148233690	Metazoa	Eukaryota	1.C.39.1.1
Mmu5	unnamed protein product	Mus musculus	528	755764	Metazoa	Eukaryota	1.C.39.1.1
Rno6	C9 protein	Rattus norvegicus	567	60688421	Metazoa	Eukaryota	1.C.39.1.1
Orf1	C9 protein	Xenopus (Silurana) tropicalis	598	166796971	Metazoa	Eukaryota	1.C.39.1.1
Xla2	hypothetical protein LOC379504	Xenopus laevis	593	148233806	Metazoa	Eukaryota	1.C.39.1.1

Table1, continued

Average Sequence Length	567						
Standard Deviation of Sequence Length	53						
Group 4							
Clu1	PREDICTED: similar to complement component 6	Canis lupus familiaris	2211	73954287	Metazoa	Eukaryota	1.C.39.3.2
Oan1	PREDICTED: hypothetical protein	Ornithorhynchus anatinus	734	149634247	Metazoa	Eukaryota	1.C.39.1.1
Mdo5	PREDICTED: similar to complement component 6	Monodelphis domestica	933	126321661	Metazoa	Eukaryota	1.C.39.3.2
Clu3	PREDICTED: similar to Complement component C6 precursor isoform 2	Canis lupus familiaris	936	73953818	Metazoa	Eukaryota	1.C.39.3.2
Eca1	PREDICTED: similar to complement component C6	Equus caballus	934	149732917	Metazoa	Eukaryota	1.C.39.3.2
Mmu10	PREDICTED: Complement component 6 isoform 1	Macaca mulatta	934	109077080	Metazoa	Eukaryota	1.C.39.3.2
Bta6	complement component 6	Bos taurus	932	114051692	Metazoa	Eukaryota	1.C.39.3.2
Ssc6	complement component C6	Sus scrofa	935	148226535	Metazoa	Eukaryota	1.C.39.3.2
Mmu4	complement component 6	Mus musculus	769	161086891	Metazoa	Eukaryota	1.C.39.3.2
Rno7	complement component 6, isoform CRA_a	Rattus norvegicus	483	149016517	Metazoa	Eukaryota	1.C.39.3.2
Gga4	complement component 6	Gallus gallus	935	221325664	Metazoa	Eukaryota	1.C.39.3.2
Tgu4	PREDICTED: similar to complement component C6	Taeniopygia guttata	929	224090383	Metazoa	Eukaryota	1.C.39.3.2
Xla1	hypothetical protein LOC432346	Xenopus laevis	935	148237974	Metazoa	Eukaryota	1.C.39.3.2
Xla6	similar to complement component 6	Xenopus laevis	934	148225474	Metazoa	Eukaryota	1.C.39.3.2
Dre2	complement component 6	Danio rerio	885	41055345	Metazoa	Eukaryota	1.C.39.3.2
Omy6	complement component C6	Oncorhynchus mykiss	941	185133413	Metazoa	Eukaryota	1.C.39.3.2
Average Sequence Length	960						
Standard Deviation of Sequence Length	355						
Group 5							
Oan3	PREDICTED: similar to Complement component 7, partial	Ornithorhynchus anatinus	646	149419497	Metazoa	Eukaryota	1.C.39.3.2
Xla5	hypothetical protein LOC432189	Xenopus laevis	830	147901594	Metazoa	Eukaryota	1.C.39.3.2
Mmu9	PREDICTED: similar to complement component 7 precursor	Macaca mulatta	552	109077094	Metazoa	Eukaryota	1.C.39.3.2
Clu2	PREDICTED: similar to complement component 7 precursor isoform 1	Canis lupus familiaris	863	73953824	Metazoa	Eukaryota	1.C.39.3.2
Bta2	complement component 7	Bos taurus	843	114051808	Metazoa	Eukaryota	1.C.39.3.2
Ssc1	complement component 7	Sus scrofa	843	47523630	Metazoa	Eukaryota	1.C.39.3.2
Eca4	PREDICTED: complement component 7	Equus caballus	909	194223929	Metazoa	Eukaryota	1.C.39.3.2
Hsa2	unnamed protein product	Homo sapiens	486	194389200	Metazoa	Eukaryota	1.C.39.3.2
Pab1	complement component 7	Pongo abelii	843	197100316	Metazoa	Eukaryota	1.C.39.3.2
Mmu8	mCG114322	Mus musculus	818	148671441	Metazoa	Eukaryota	1.C.39.3.2
Mmu11	PREDICTED: similar to Complement component 7	Mus musculus	708	149266317	Metazoa	Eukaryota	1.C.39.3.2
Rno5	PREDICTED: similar to complement component 7 precursor	Rattus norvegicus	778	109464453	Metazoa	Eukaryota	1.C.39.3.2
Rno4	complement component 7	Rattus norvegicus	540	149016512	Metazoa	Eukaryota	1.C.39.3.2
Rno8	PREDICTED: similar to complement component 7 precursor	Rattus norvegicus	844	109466098	Metazoa	Eukaryota	1.C.39.3.2
Gga1	PREDICTED: similar to complement protein C7	Gallus gallus	443	50761596	Metazoa	Eukaryota	1.C.39.3.2
Tgu2	PREDICTED: complement component 7	Taeniopygia guttata	844	224090381	Metazoa	Eukaryota	1.C.39.3.2
Dre12	PREDICTED: similar to complement protein component C7-1	Danio rerio	820	189533869	Metazoa	Eukaryota	1.C.39.3.2
Omy7	complement protein component C7-1	Oncorhynchus mykiss	808	185133218	Metazoa	Eukaryota	1.C.39.3.2
Pol2	complement component C7	Paralichthys olivaceus	805	6682831	Metazoa	Eukaryota	1.C.39.3.2
Tni4	unnamed protein product	Tetraodon nigroviridis	584	47211138	Metazoa	Eukaryota	1.C.39.3.2
Dre6	PREDICTED: similar to complement component 7	Danio rerio	849	125824340	Metazoa	Eukaryota	1.C.39.3.2
Omy3	complement component C7-2	Oncorhynchus mykiss	845	185132432	Metazoa	Eukaryota	1.C.39.3.2
Tni10	unnamed protein product	Tetraodon nigroviridis	809	47212821	Metazoa	Eukaryota	1.C.39.3.2

Table 1, continued

Average Sequence Length	753						
Standard Deviation of Sequence Length	137						
Group 6							
Cin4	PREDICTED: similar to complement component C6	Ciona intestinalis	592	198433282	Metazoa	Eukaryota	1.C.39.3.1
Hro1	similar to terminal complement component	Halocynthia roretzi	585	224176461	Metazoa	Eukaryota	1.C.39.3.1
Average Sequence Length	589						
Standard Deviation of Sequence Length	5						
Group 7							
Tni9	unnamed protein product PREDICTED: similar to Perforin-1 precursor (P1)	Tetraodon nigroviridis	1206	47218949	Metazoa	Eukaryota	1.C.39.2.1
Gga2	(Lymphocyte pore-forming protein) (PFP) (Cytolysin) PREDICTED: similar to Perforin-1 precursor (P1)	Gallus gallus	491	118099091	Metazoa	Eukaryota	1.C.39.2.1
Oan6	(Lymphocyte pore-forming protein) (PFP) (Cytolysin), partial	Ornithorhynchus anatinus	511	149472392	Metazoa	Eukaryota	1.C.39.2.1
Xla3	hypothetical protein LOC495232	Xenopus laevis	532	148237294	Metazoa	Eukaryota	1.C.39.2.1
Mmu1	PREDICTED: perforin 1 isoform 1	Macaca mulatta	555	109089420	Metazoa	Eukaryota	1.C.39.2.1
Cja1	perforin 1	Callithrix jacchus	555	197112111	Metazoa	Eukaryota	1.C.39.2.1
Bta1	perforin 1 (pore forming protein)	Bos taurus	554	219522060	Metazoa	Eukaryota	1.C.39.2.1
Ssc2	PREDICTED: perforin 1	Sus scrofa	555	194042762	Metazoa	Eukaryota	1.C.39.2.1
Eca6	PREDICTED: similar to perforin 1	Equus caballus	555	194205976	Metazoa	Eukaryota	1.C.39.2.1
Fca1	perforin 1 PREDICTED: similar to Perforin 1 precursor (P1)	Felis catus	555	156071445	Metazoa	Eukaryota	1.C.39.2.1
Clu5	(Lymphocyte pore forming protein) (PFP) (Cytolysin)	Canis lupus familiaris	613	73953408	Metazoa	Eukaryota	1.C.39.2.1
Mmu3	perforin 1	Mus musculus	554	200290	Metazoa	Eukaryota	1.C.39.2.1
Rno3	perforin 1 (pore forming protein), isoform CRA_a PREDICTED: similar to Perforin-1 precursor (P1)	Rattus norvegicus	585	149038739	Metazoa	Eukaryota	1.C.39.2.1
Mdo2	(Lymphocyte pore-forming protein) (PFP) (Cytolysin) PREDICTED: similar to Perforin-1 precursor (P1)	Monodelphis domestica	559	126272286	Metazoa	Eukaryota	1.C.39.2.1
Oan4	(Lymphocyte pore-forming protein) (PFP) (Cytolysin)	Ornithorhynchus anatinus	556	149536560	Metazoa	Eukaryota	1.C.39.2.1
Tni3	unnamed protein product novel protein similar to mouse and rat perforin 1	Tetraodon nigroviridis	517	47213662	Metazoa	Eukaryota	1.C.39.2.1
Dre10	(pore forming protein) (Prf1)	Danio rerio	536	94732622	Metazoa	Eukaryota	1.C.39.2.1
Ssa1	Perforin-1 precursor	Salmo salar	597	209155244	Metazoa	Eukaryota	1.C.39.2.1
Dre1	PREDICTED: similar to perforin 1 (pore forming protein) PREDICTED: similar to Perforin-1 precursor (P1)	Danio rerio	574	125844965	Metazoa	Eukaryota	1.C.39.2.1
Dre11	(Lymphocyte pore-forming protein) (Cytolysin)	Danio rerio	610	189532388	Metazoa	Eukaryota	1.C.39.2.1
Dre3	PREDICTED: similar to perforin PREDICTED: similar to Perforin-1 precursor (P1)	Danio rerio	588	125812563	Metazoa	Eukaryota	1.C.39.2.1
Dre13	(Lymphocyte pore-forming protein) (PFP) (Cytolysin)	Danio rerio	580	125812566	Metazoa	Eukaryota	1.C.39.2.1
Tni2	unnamed protein product	Tetraodon nigroviridis	585	47217798	Metazoa	Eukaryota	1.C.39.2.1
Tni5	unnamed protein product	Tetraodon nigroviridis	325	47200034	Metazoa	Eukaryota	1.C.39.2.1
Pol4	perforin	Paralichthys olivaceus	587	30519828	Metazoa	Eukaryota	1.C.39.2.1
Omy4	perforin	Oncorhynchus mykiss	589	198442831	Metazoa	Eukaryota	1.C.39.2.1
Tni1	unnamed protein product	Tetraodon nigroviridis	523	47217490	Metazoa	Eukaryota	1.C.39.2.1
Dre5	PREDICTED: hypothetical protein LOC559384	Danio rerio	528	189522601	Metazoa	Eukaryota	1.C.39.2.1
Dre8	novel protein (zgc:63021)	Danio rerio	516	169146195	Metazoa	Eukaryota	1.C.39.2.1
Average Sequence Length	572						

Table 1, continued

Standard Deviation of Sequence Length	133						
Group 8							
Ani1	hypothetical protein AN5085.2	Aspergillus nidulans FGSC A4	778	67537830 Fungi	Eukaryota	1.C.39.9.1	
Eni1	SpoC1-C1C	Emericella nidulans	446	168091 Fungi	Eukaryota	1.C.39.9.1	
Average Sequence Length	612						
Standard Deviation of Sequence Length	235						
Group 9							
Ptr1	predicted protein	Populus trichocarpa	615	224069581 Viridiplantae	Eukaryota	1.C.39.11.1	
Vvi1	unnamed protein product	Vitis vinifera	606	157358723 Viridiplantae	Eukaryota	1.C.39.11.1	
Mtr1	Membrane attack complex component/perforin/complement C9	Medicago truncatula	610	92870237 Viridiplantae	Eukaryota	1.C.39.11.1	
Vvi2	unnamed protein product	Vitis vinifera	603	157354261 Viridiplantae	Eukaryota	1.C.39.11.1	
Average Sequence Length	609						
Standard Deviation of Sequence Length	5						
Group 10							
Bsp1	Membrane attack complex component/perforin/complement C9	Beggiatoa sp. PS	526	153871368 γ-proteobacteria	Bacteria	1.C.39.8.2	
Ppa1	hemopexin	Plesiocystis pacifica SIR-1	516	149920404 δ-proteobacteria	Bacteria	1.C.39.8.1	
Average Sequence Length	521						
Standard Deviation of Sequence Length	7						
Group 11							
Bfl25	hypothetical protein BRAFLDRAFT_89470	Branchiostoma floridae	1238	219458626 Metazoa	Eukaryota	1.C.39.5.3	
Bfl1	hypothetical protein BRAFLDRAFT_90586	Branchiostoma floridae	2433	219460616 Metazoa	Eukaryota	1.C.39.5.3	
Bfl6	hypothetical protein BRAFLDRAFT_69406	Branchiostoma floridae	1305	219418090 Metazoa	Eukaryota	1.C.39.5.3	
Bfl10	hypothetical protein BRAFLDRAFT_76619	Branchiostoma floridae	1592	219433035 Metazoa	Eukaryota	1.C.39.5.3	
Bfl18	hypothetical protein BRAFLDRAFT_78467	Branchiostoma floridae	987	219436623 Metazoa	Eukaryota	1.C.39.5.3	
Bfl22	hypothetical protein BRAFLDRAFT_112780	Branchiostoma floridae	415	219510318 Metazoa	Eukaryota	1.C.39.5.3	
Bfl11	hypothetical protein BRAFLDRAFT_123900	Branchiostoma floridae	1217	219449115 Metazoa	Eukaryota	1.C.39.5.3	
Bfl31	hypothetical protein BRAFLDRAFT_77695	Branchiostoma floridae	899	219435072 Metazoa	Eukaryota	1.C.39.5.3	
Nve2	predicted protein	Nematostella vectensis	1170	156389161 Metazoa	Eukaryota	1.C.39.5.2	
Nve1	predicted protein	Nematostella vectensis	967	156408626 Metazoa	Eukaryota	1.C.39.5.2	
Nve3	predicted protein	Nematostella vectensis	1175	156408896 Metazoa	Eukaryota	1.C.39.5.2	
Bfl3	hypothetical protein BRAFLDRAFT_88402	Branchiostoma floridae	1451	219456494 Metazoa	Eukaryota	1.C.39.5.1	
Bfl2	hypothetical protein BRAFLDRAFT_63382	Branchiostoma floridae	992	219406285 Metazoa	Eukaryota	1.C.39.5.1	
Bfl13	hypothetical protein BRAFLDRAFT_63809	Branchiostoma floridae	1503	219406692 Metazoa	Eukaryota	1.C.39.5.1	
Bfl21	hypothetical protein BRAFLDRAFT_83060	Branchiostoma floridae	1324	219446344 Metazoa	Eukaryota	1.C.39.5.1	
Bfl33	hypothetical protein BRAFLDRAFT_97589	Branchiostoma floridae	1672	219475420 Metazoa	Eukaryota	1.C.39.5.1	
Bfl17	hypothetical protein BRAFLDRAFT_86688	Branchiostoma floridae	776	219453206 Metazoa	Eukaryota	1.C.39.5.1	
Bfl36	hypothetical protein BRAFLDRAFT_67061	Branchiostoma floridae	558	219413344 Metazoa	Eukaryota	1.C.39.5.1	
Bfl15	hypothetical protein BRAFLDRAFT_102093	Branchiostoma floridae	981	219485318 Metazoa	Eukaryota	1.C.39.5.1	
Bfl19	hypothetical protein BRAFLDRAFT_86686	Branchiostoma floridae	1455	219453264 Metazoa	Eukaryota	1.C.39.5.1	
Bfl27	hypothetical protein BRAFLDRAFT_111559	Branchiostoma floridae	572	219506840 Metazoa	Eukaryota	1.C.39.5.1	
Bfl23	hypothetical protein BRAFLDRAFT_89006	Branchiostoma floridae	1337	219457635 Metazoa	Eukaryota	1.C.39.5.1	

Table 1, continued

Bfl16	hypothetical protein BRAFLDRAFT_125400	Branchiostoma floridae	1359	219457669	Metazoa	Eukaryota	1.C.39.5.1
Bfl34	hypothetical protein BRAFLDRAFT_88988	Branchiostoma floridae	1353	219457671	Metazoa	Eukaryota	1.C.39.5.1
Bfl20	hypothetical protein BRAFLDRAFT_105135	Branchiostoma floridae	1320	219492152	Metazoa	Eukaryota	1.C.39.5.1
Bfl32	hypothetical protein BRAFLDRAFT_105134	Branchiostoma floridae	1195	219492150	Metazoa	Eukaryota	1.C.39.5.1
Bfl12	hypothetical protein BRAFLDRAFT_112179	Branchiostoma floridae	858	219508445	Metazoa	Eukaryota	1.C.39.5.1
Bfl9	hypothetical protein BRAFLDRAFT_106125	Branchiostoma floridae	1244	219494396	Metazoa	Eukaryota	1.C.39.5.1
Bfl5	hypothetical protein BRAFLDRAFT_82668	Branchiostoma floridae	1731	219445681	Metazoa	Eukaryota	1.C.39.5.1
Bfl26	hypothetical protein BRAFLDRAFT_71536	Branchiostoma floridae	1130	219422389	Metazoa	Eukaryota	1.C.39.5.1
Bfl29	hypothetical protein BRAFLDRAFT_73492	Branchiostoma floridae	573	219426301	Metazoa	Eukaryota	1.C.39.5.1
Bfl8	hypothetical protein BRAFLDRAFT_106127	Branchiostoma floridae	562	219494413	Metazoa	Eukaryota	1.C.39.5.1
Bfl24	hypothetical protein BRAFLDRAFT_101213	Branchiostoma floridae	1694	219483305	Metazoa	Eukaryota	1.C.39.5.1
Average Sequence Length	1183						
Standard Deviation of Sequence Length	417						
Group 12							
Tth1	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	387	118368397	Oligohymenophorea	Eukaryota	1.C.39.6.2
Tth2	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	342	118369627	Oligohymenophorea	Eukaryota	1.C.39.6.2
Tth4	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	681	118366533	Oligohymenophorea	Eukaryota	1.C.39.6.3
Bbo1	mac/perforin domain containing protein	Babesia bovis T2Bo	420	156084486	Apicomplexa	Eukaryota	1.C.39.6.4
Pfa1	hypothetical protein	Plasmodium falciparum 3D7	842	124505319	Apicomplexa	Eukaryota	1.C.39.6.1
Pbe1	sporozoite protein with MACPF related domain	Plasmodium berghei	810	56805561	Apicomplexa	Eukaryota	1.C.39.6.1
Pkn1	Sporozoite protein with MAC/Perforin domain	Plasmodium knowlesi strain H	844	221052646	Apicomplexa	Eukaryota	1.C.39.6.1
Pvi1	hypothetical protein	Plasmodium vivax Sal-1	843	156094597	Apicomplexa	Eukaryota	1.C.39.6.1
Tpa1	hypothetical protein	Theileria parva strain Muguga	357	71026506	Apicomplexa	Eukaryota	1.C.39.6.5
Tan1	perforin-related protein	Theileria annulata strain Ankara	1219	85001526	Apicomplexa	Eukaryota	1.C.39.6.5
Average Sequence Length	675						
Standard Deviation of Sequence Length	290						
Group 13							
Spu7	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	690	115898506	Metazoa	Eukaryota	1.C.39.7.1
Spu3	PREDICTED: similar to apextrin, partial	Strongylocentrotus purpuratus	437	115753697	Metazoa	Eukaryota	1.C.39.7.1
Spu1	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	501	72148293	Metazoa	Eukaryota	1.C.39.7.1
Spu2	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	569	72008499	Metazoa	Eukaryota	1.C.39.7.1
Spu5	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	567	72157597	Metazoa	Eukaryota	1.C.39.7.1
Spu4	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	570	115921094	Metazoa	Eukaryota	1.C.39.7.1
Spu6	PREDICTED: similar to apextrin	Strongylocentrotus purpuratus	570	115955362	Metazoa	Eukaryota	1.C.39.7.1
Average Sequence Length	558						
Standard Deviation of Sequence Length	77						
Group 14							
Tth3	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	520	118396447	Oligohymenophorea	Eukaryota	1.C.39.7.1
Tth5	E1-E2 ATPase family protein	Tetrahymena thermophila SB210	1982	118371656	Oligohymenophorea	Eukaryota	1.C.39.7.1
Tth7	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	518	118371660	Oligohymenophorea	Eukaryota	1.C.39.7.1
Tth6	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	538	118371658	Oligohymenophorea	Eukaryota	1.C.39.7.1
Tth8	MAC/Perforin domain containing protein	Tetrahymena thermophila SB210	536	118396445	Oligohymenophorea	Eukaryota	1.C.39.7.1
Ami1	apextrin	Acropora millepora	854	118153966	Metazoa	Eukaryota	1.C.39.7.1

Table 1, continued

Pte1	hypothetical protein	Paramecium tetraurelia strain d4-2	603	145475565	Oligohymenophorea	Eukaryota	1.C.39.7.1
Average Sequence Length	793						
Standard Deviation of Sequence Length	538						
Group 15							
Cmu1	MAC/perforin family protein	Chlamydia muridarum Nigg	809	15835049	Chlamydiae	Bacteria	1.C.39.12.1
Cpn1	hypothetical protein CPn0176	Chlamydophila pneumoniae CWL029	411	15618100	Chlamydiae	Bacteria	1.C.39.12.2
Average Sequence Length	610						
Standard Deviation of Sequence Length	281						
Group 16							
Bth1	hypothetical protein BT_3120	Bacteroides thetaiotaomicron VPI-5482	470	29348529	Bacteroidetes	Bacteria	1.c.39.13.1
Bin1	hypothetical protein BACINT_03190	Bacteroides intestinalis DSM 17393	474	189466814	Bacteroidetes	Bacteria	1.c.39.13.1
Bce2	hypothetical protein BACCELL_05502	Bacteroides cellulosilyticus DSM 14838	476	224540588	Bacteroidetes	Bacteria	1.c.39.13.2
Bce1	hypothetical protein BACCELL_01585	Bacteroides cellulosilyticus DSM 14838	397	224536709	Bacteroidetes	Bacteria	1.c.39.13.2
Bfr1	hypothetical protein BF1634	Bacteroides fragilis YCH46	372	53712924	Bacteroidetes	Bacteria	1.c.39.13.2
Bun1	hypothetical protein BACUNI_00959	Bacteroides uniformis ATCC 8492	656	160888542	Bacteroidetes	Bacteria	1.c.39.13.2
Bfr2	hypothetical protein BF1566	Bacteroides fragilis YCH46	486	53712858	Bacteroidetes	Bacteria	1.c.39.13.3
Bfr3	hypothetical protein BF2685	Bacteroides fragilis YCH46	506	53713977	Bacteroidetes	Bacteria	1.c.39.13.3
Average Sequence Length	480						
Standard Deviation of Sequence Length	85						
Group 17							
Pma1	hypothetical protein PMAA_069160	Penicillium marneffei ATCC 18224	784	212532427	Fungi	Eukaryota	1.C.39.14.1
Plu1	hypothetical protein plu1415	Photorhabdus luminescens subsp. laumondii TTO1	510	37525369	γ-proteobacteria	Bacteria	1.C.39.4.1
Cmi1	putative perforin	Clavibacter michiganensis subsp. michiganensis NCPPB 382	470	148273566	Actinobacteria	Bacteria	1.C.39.4.1
Spr1	membrane attack complex component/perforin/complement C9	Serratia proteamaculans 568	489	157370628	γ-proteobacteria	Bacteria	1.C.39.4.1
Average Sequence Length	563						
Standard Deviation of Sequence Length	148						
Group 18							
Ter1	membrane attack complex component/perforin/complement C9	Trichodesmium erythraeum IMS101	453	113474643	Cyanobacteria	Bacteria	1.C.39.4.2
Ddi1	hypothetical protein	Dictyostelium discoideum AX4	340	66805335	Mycetozoa	Eukaryota	1.C.97.3.1
Afl3	hypothetical protein AFLA_064630	Aspergillus flavus NRRL3357	664	220693297	Fungi	Eukaryota	1.C.97.2.1
Afl1	conserved hypothetical protein	Aspergillus flavus NRRL3357	736	220688529	Fungi	Eukaryota	1.C.97.2.1
Gze1	hypothetical protein FG07664.1	Gibberella zeae PH-1	785	46126573	Fungi	Eukaryota	1.C.97.2.1
Nfi1	hypothetical protein NFIA_101960	Neosartorya fischeri NRRL 181	764	119499704	Fungi	Eukaryota	1.C.97.2.1
Afl2	conserved hypothetical protein	Aspergillus flavus NRRL3357	618	220689182	Fungi	Eukaryota	1.C.97.2.1
Msp1	hypothetical protein MED121_03928	Marinomonas sp. MED121	588	87122061	γ-proteobacteria	Bacteria	1.C.39.4.3
Average Sequence Length	619						
Standard Deviation of Sequence Length	156						

Table 2. All homologues of the CDC family that were included in our study are listed. These proteins were obtained by a PSI-BLAST search using the TCDB representative protein, 1.C.12.1.1 as the query sequence with two iterations. The proteins are organized by cluster, and the abbreviations, protein descriptions, organismal sources, sequence lengths, GenInfo Identifier (GI) numbers, phyla, domains, and TCDB sub-family of each protein are provided.

Table 2: CDC Family Homologues

Abbreviation	Protein Description	Organism	Protein Size	GI Number	Phylum	Domain	TCDB Sub-Family
Orf1	lytic 1/listeriolysin O fusion protein	synthetic construct	433	190144480	none	Unclassified	1.C.12.1.7
Lmo1	listeriolysin O RecName: Full=Seeligeriolysin; AltName: Full=Thiol-activated cytolysin; Flags: Precursor	Listeria monocytogenes	532	887028	Firmicutes	Bacteria	1.C.12.1.7
Lse1		Listeria seeligeri	530	401156	Firmicutes	Bacteria	1.C.12.1.7
Cno1	tetanolysin O	Clostridium novyi NT	600	118443734	Firmicutes	Bacteria	1.C.12.1.1
Gva1	TAC family cholesterol-binding, thiol-activated cytolysin	Gardnerella vaginalis ATCC 14019	541	227506699	Actinobacteria	Bacteria	1.C.12.1.5
Cno2	tetanolysin O	Clostridium novyi NT	514	118443539	Firmicutes	Bacteria	1.C.12.1.1
Liv1	Hly	Listeria ivanovii subsp. ivanovii	528	40888993	Firmicutes	Bacteria	1.C.12.1.6
Asp1	tetanolysin O	Algoriphagus sp. PR1	513	126648213	Bacteroidetes	Bacteria	1.C.12.1.1
Cbu1	perfringolysin O	Clostridium butyricum 5521	513	182419658	Firmicutes	Bacteria	1.C.12.1.1
Cbo1	tetanolysin O	Clostridium botulinum B str. Eklund 17B	602	187933202	Firmicutes	Bacteria	1.C.12.1.1
Bth1	Alveolysin	Bacillus thuringiensis serovar pakistani str. T13001	512	228961328	Firmicutes	Bacteria	1.C.12.1.3
Cte1	tetanolysin O	Clostridium tetani E88	527	28211522	Firmicutes	Bacteria	1.C.12.1.1
Cbo2	tetanolysin O	Clostridium botulinum C str. Eklund	641	168185437	Firmicutes	Bacteria	1.C.12.1.1
Bce1	Alveolysin	Bacillus cereus Rock4-18	509	229077261	Firmicutes	Bacteria	1.C.12.1.3
Cpe1	perfringolysin O	Clostridium perfringens ATCC 13124	500	110798884	Firmicutes	Bacteria	1.C.12.1.1
Lsp1	cholesterol-dependent cytolysin RecName: Full=Alveolysin; AltName: Full=Thiol-activated cytolysin; Flags: Precursor	Lysinibacillus sphaericus	513	157885779	Firmicutes	Bacteria	1.C.12.1.3
Pal1		Paenibacillus alvei	501	113672	Firmicutes	Bacteria	1.C.12.1.2
Ofo1	pyolysin	Oxalobacter formigenes HOxBLS	505	237746033	β-proteobacteria	Bacteria	1.C.12.1.9
Apy1	pyolysin	Arcanobacterium pyogenes	534	6456474	Actinobacteria	Bacteria	1.C.12.1.9
Bbr1	thiol-activated cytolysin precursor	Brevibacillus brevis NBRC 100599	511	226310317	Firmicutes	Bacteria	1.C.12.1.3
Cbo3	tetanolysin O	Clostridium botulinum D str. 1873	526	253681419	Firmicutes	Bacteria	1.C.12.1.1
Cbo4	tetanolysin O	Clostridium botulinum D str. 1873	547	253681526	Firmicutes	Bacteria	1.C.12.1.1
Ssu1	suilysin	Streptococcus suis	497	90193575	Firmicutes	Bacteria	1.C.12.1.8
Sdy1	streptolysin O precursor	Streptococcus dysgalactiae subsp. equisimilis GGS_124	574	251783417	Firmicutes	Bacteria	1.C.12.1.4
Cbo5	tetanolysin O	Clostridium botulinum C str. Eklund	518	168186337	Firmicutes	Bacteria	1.C.12.1.1
Smi1	human platelet aggregation factor Sm-hPAF	Streptococcus mitis	665	84579714	Firmicutes	Bacteria	1.C.12.1.5
Spn1	pneumolysin Chain A, Crystal Structure Of The Human-Specific	Streptococcus pneumoniae	479	126571356	Firmicutes	Bacteria	1.C.12.1.5
Sin1	Toxin Intermedilysin	Streptococcus intermedius	535	60593463	Firmicutes	Bacteria	1.C.12.1.5
Average Sequence Length	532						
Standard Deviation of Sequence Length	47						

Table 3. All homologues that were initially present in the TCDB entry for the Pleurotolysin family are listed. The proteins are organized by cluster, and the abbreviations, protein descriptions, organismal sources, sequence lengths, GenInfo Identifier (GI) numbers, phyla, domains, and TCDB sub-family of each protein are provided.

Table 3: Pleurotolysin Family Homologues

Abbreviation	Protein Description	Organism	Protein Size	GI Number	Phylum	Domain	TCDB Sub-Family
Pos1	pleurotolysin B	Pleurotus ostreatus	523	54312024	Fungi	Eukaryota	1.C.97.1.1
Per1	erylysin B	Pleurotus eryngii	522	261857452	Fungi	Eukaryota	1.C.97.1.2
Cgl1	hypothetical protein CHGG_09313	Chaetomium globosum CBS 148.51	924	116202857	Fungi	Eukaryota	1.C.97.2.1
Dis1	hypothetical protein DDB_G0289093	Dictyostelium discoideum AX4	340	66805335	Mycetozoa	Eukaryota	1.C.97.3.1
Pkn1	Sporozoite protein with MAC/Perforin domain	Plasmodium knowlesi strain H	844	221052646	Apicomplexa	Eukaryota	1.C.39.6.1
Cli1	hypothetical protein Clim_0052	Chlorobium limicola DSM 245	892	189345610	Chlorobia	Eukaryota	1.C.97.5.1
Fps1	fmo gene product	Flavobacterium psychrophilum JIP02/86	382	150024210	Bacteroidetes	Bacteria	1.C.12.2.1
Nsp1	hypothetical protein	Nostoc sp. PCC 7120	470	17228824	Cyanobacteria	Bacteria	1.C.12.3.1
Average Sequence Length	612						
Standard Deviation of Sequence Length	237						

Table 4. Recognized Conserved Domains of Longer MACPF Proteins. Proteins with significantly longer lengths in each cluster were analyzed for additional conserved domains.

Cluster	Proteins	GI Number	Conserved Domains in Longer Proteins	Description	Residues
1	Tgu5	224058308	pfam09770 pfam01823	Topoisomerase II-associated protein PAT1 Two MAC/Perforin Domain Repeats	1-71 719-902
2	Bfl30 Bfl7	219431797 219494025	PRK12323	DNA Polymerase III subunits gamma and tau	723-1264
2	Cin5 Cin7 Cin12	198417017 198419275 198417019	smart00209 pfam00090	Thrombospondin Type-1 Repeats Thrombospondin Type-1 Domain	309-359 472-520 580-632 633-682 688-735 742-786
4	Clu1	73954287	No conserved domain.	Similar to hCG1993037, isoform CRA_F (*Homo sapiens*) Expect: 3e-4 GI# 119602545	626-2211
7	Tni9	47218949	cd00201 pfam00693	Two Conserved Tryptophan Domains (WWP or rsp5) PPIC-type PPI rotamase	8-38 52-131
11	Bfl1	219460616	smart00607 pfam07699 smart00202 cd00064 cd00037	eel-Fucolectin Tachylectin-4 Pentaxrin-1 Domain GCC2 and GCC3 Scavenger receptor Cys-rich Furin-like repeats. Cysteine rich region. Exact function of the domain is not known. Furin is a serine-kinase dependent proprotein processor. Other members of this family include endoproteases and cell surface receptors. C-type lectin (CTL)/C-type lectin-like (CTLD) domain	920-1058 1106-1153 1351-1398 1451-1551 1554-1655 1658-1758 1830-1911 1909-1190 1988-2065 2063-2140 2270-2230
12	Tan1	85001526	pfam01823	Three MAC/Perforin Domain Repeats	173-304 438-652 990-1212
14	Tth5	118371656	pfam00122 TIGR01657	E1-E2_ATPase domain Discontinuous P-ATPase-V (copper): P-type ATPase of unknown pump specificity	754-1007 562-1370 1521-1809
15	Cpn1	15618100	smart00472	Domain in ryanodine and inositol trisphosphate receptors and protein O-mannosyltransferases	366-409

Table 5. SSearch Comparison Scores Between CDC and MACPF Homologues. Regions which contained their respective CDC or MACPF domain were further analyzed with GAP and listed in Table 5.

MACPF (Residues Compared)	CDC (Residues Compared)	Average Score Expressed in S.D. (SSearch Program)
Bfl1 (922 - 1062)	Smi1 (48 - 186)	25.3
Bfl5 (1239 - 1407)	Smi1 (48 - 217)	28.6
Bfl6 (959 - 1102)	Smi1 (48 - 188)	30.1
Bfl9 (766 - 899)	Smi1 (48 - 181)	24.1
Bfl16 (1080 - 1212)	Smi1 (48 - 181)	39.8
Bfl23 (1184 - 1325)	Smi1 (48 - 181)	23.6
Bfl34 (1196 - 1350)	Smi1 (44 - 201)	31.2
Omy3 (148-283)	Cbo2 (388-522)	6.5
Omy3 (148-336)	Cno1 (347-538)	7.3
Spu6 (199-293)	Cno2 (169-255)	5.5
Tth1 (203-340)	Cte1 (293-425)	5.7
Ami1 (474-686)	Bbr1 (225-439)	6.8
Eca2 (316-409)	Cbo5 (381-476)	6.3
Clu7 (338-425)	Cte1 (386-474)	6.9
Rno6 (307-426)	Cte1 (362-485)	8.0
Clu7 (168-434)	Cbo5 (204-476)	5.2
Rno6 (328-426)	Cbo5 (378-477)	5.5

Table 6. GAP Comparison Scores Between CDC and MACPF Homologues

MACPF (Residues Compared)	CDC (Residues Compared)	Average Score Expressed in S.D. (GAP Program)
Omy3 (148-283)	Cbo2 (388-522)	10.8
Omy3 (148-336)	Cno1 (347-538)	10.9
Spu6 (199-293)	Cno2 (169-255)	12.5
Tth1 (203-340)	Cte1 (293-425)	10.9
Ami1 (474-686)	Bbr1 (225-439)	12.9
Eca2 (316-409)	Cbo5 (381-476)	11.0
Clu7 (338-425)	Cte1 (386-474)	12.4
Rno6 (307-426)	Cte1 (362-485)	14.4
Clu7 (168-434)	Cbo5 (204-476)	13.3
Rno6 (328-426)	Cbo5 (378-477)	10.2

Table 7. Comparison of MACPF and CDC TMHs. The GAP alignments in Table 5 were superimposed on the MACPF structure, PDB# 2RD7, and the CDC structure, PDB# 1PFO, and the TMH included in each alignment was observed.

	Omy3/Cbo2	**Omy3/Cno1**	**Spu6/Cno2**	**Tth1/Cte1**
2RD7	TMH1	TMH1	TMH2	TMH1
1PFO	TMH2	TMH2	TMH1	TMH2

Table 8. Comparison Scores Between Pleurotolysin and MACPF Homologues

MACPF (Residues Compared)	**Pleurotolysin (Residues Compared)**	**Average Score Expressed in S.D. (SSearch Program)**
Cmi1 (1302 - 2058)	Pkn1 (292 - 593)	12.0
Nve1 (1479 - 2108)	Pkn1 (368 - 611)	12.4
Nve3 (1471 - 2056)	Cgl1 (364 - 591)	13.8
Bfl5 (838 - 2102)	Pkn1 (134 - 607)	12.7
Bfl24 (842 - 2102)	Pkn1 (129 - 607)	13.0
Bfl12 (1725 - 2102)	Pkn1 (482 - 607)	12.2
Tth4 (1120 - 2245)	Pkn1 (294 - 624)	18.4
Tth1 (1372 - 2252)	Pkn1 (331 - 632)	16.1
Pkn1 (2424 - 2556)	Pkn1 (480 - 625)	13.2
Pvi1 (2417 - 2556)	Dis1 (474 - 625)	16.3
Tpa1 (1137 - 2255)	Pkn1 (295-622)	23.7
Tan1 (2853 - 3175)	Pkn1 (294 - 622)	64.9
Spu3 (1139 - 2054)	Pkn1 (290 - 565)	16.9
Spu4 (1305 - 2054)	Pkn1 (295 - 581)	12.1
Spu6 (1308 - 2054)	Pkn1 (298 - 581)	12.0
Tth3 (1124 - 2099)	Pkn1 (274 - 604)	30.1
Tth5 (1140 - 2097)	Pkn1 (290 - 602)	22.7
Tth6 (1138 - 2236)	Pkn1 (288 - 629)	30.1
Tth8 (1123 - 2096)	Pkn1 (271 - 601)	29.4
Tth7 (1132 - 2095)	Pkn1 (280 - 600)	19.9
Cmu1 (1725 - 2248)	Pkn1 (482 - 652)	15.1
Cpn1 (1706 - 2247)	Pkn1 (465 - 652)	12.7
Mmu3 (1720 - 2513)	Pkn1 (477 - 874)	13.8
Mdo2 (1720 - 2110)	Pkn1 (476 - 608)	13.2
Tni3 (1696 - 2191)	Pkn1 (447 - 622)	12.4
Tni1 (1728 - 2191)	Pkn1 (485 - 622)	13.3
Dre7 (1721 - 2104)	Pkn1 (478 - 602)	12.7

Table 8, continued

Cin2 (1326 - 2195)	Pkn1 (254 - 626)	22.8
Cin10 (1332 - 2197)	Pkn1 (262 - 628)	27.0
Cin11 (1332 - 2256)	Pkn1 (262 - 677)	17.7
Cin3 (1729 - 2198)	Pkn1 (486 - 629)	14.0
Cin7 (1729 - 2198)	Pkn1 (486 - 629)	18.1
Cin13 (1729 - 2198)	Pkn1 (486 - 629)	15.4
Cin5 (1722 - 2195)	Pkn1 (479 - 626)	12.1
Cin12 (1704 - 2120)	Pkn1 (471 - 618)	12.3
Bfl14 (1491 - 2108)	Dis1 (368 - 609)	12.9
Bbo1 (1395 - 2202)	Pkn1 (294 - 638)	65.4
Ami1 (1426 - 2123)	Pkn1 (291 - 625)	16.6
Pma1 (1390 - 2111)	Pkn1 (261 - 609)	16.2
Ddi1 (1457 - 2108)	Pos1 (382 - 609)	22.2
Ddi1 (1457 - 2108)	Per1 (382 - 609)	19.4
Ddi1 (1420 - 2110)	Cgl1 (400-611)	15.5
Ddi1 (1465 - 2100)	Cli1 (410 - 600)	13.3
Ddi1 (1723 - 2118)	Pkn1 (480 - 617)	12.4
Nfi1 (1748 - 2409)	Pos1 (317 - 609)	13.8
Nfi1 (1752 - 2409)	Per1 (321 - 609)	12.1
Nfi1 (1731 -2438)	Cgl1 (327 - 637)	59.9
Nfi1 (1782 - 2410)	Cli1 (383 - 610)	19.8
Nfi1 (1772 - 2424)	Dis1 (363 -631)	16.3
Afl2 (1468 - 2108)	Pos1 (317 - 609)	17.5
Afl2 (1481 -2108)	Per1 (357 - 609)	16.7
Afl2 (1418 - 2117)	Cgl1 (319 - 618)	61.7
Afl2 (1475 - 2101)	Cli1 (369 - 602)	27.0
Afl2 (1505 - 2230)	Dis1 (374 - 660)	20.5
Gze1 (1474 - 2111)	Cgl1 (349 - 612)	41.0
Gze1 (1662 - 2049)	Cli1 (410 - 594)	16.0
Afl1 (1661 - 2424)	Pos1 (402 - 722)	15.2
Afl1 (1465 - 2193)	Cgl1 (326 - 631)	57.8
Afl1 (1402 - 2109)	Cli1 (296 - 610)	22.5
Afl3 (1477 - 2109)	Pos1 (353 - 616)	27.4
Afl3 (1477 - 2109)	Per1 (353 - 616)	22.5
Afl3 (1390 - 2231)	Cgl1 (281 - 665)	84.5
Afl3 (1372 - 2062)	Cli1 (264 - 597)	28.6
Pkn1 (1098 - 2961)	Pkn1 (1 - 996)	639.5
Pvi1 (1098 - 2961)	Pkn1 (1 - 996)	569.1
Pbe1 (1098 - 2960)	Pkn1 (1 - 995)	458.6

Table 8, continued

Pfa1 (1112 - 2961)	Pkn1 (15 - 996)	369.3
Ddi1 (1400 - 2481)	Dis1 (354 - 863)	298.3

Table 9. Comparison Scores Between Pleurotolysin and CDC Homologues

CDC (Residues Compared)	**Pleurotolysin (Residues Compared)**	**Average Score Expressed in S.D. (SSearch Program)**
Gva1 (313 - 633)	Fps1 (410 - 858)	28.3
Smi1 (282 - 639)	Fps1 (378 - 864)	24.8
Sin1 (339 - 634)	Fps1 (444 - 859)	33.6
Spn1 (320 - 633)	Fps1 (423 - 858)	16.3
Ssu1 (349 - 638)	Fps1 (464 - 863)	33.0
Lmo1 (352 - 633)	Fps1 (467 - 858)	20.0
Lse1 (655 - 743)	Fps1 (422 - 858)	27.3
Liv1 (317 - 633)	Fps1 (422 - 858)	23.1
Orf1 (517 - 633)	Fps1 (663 - 858)	15.9
Cte1 (322 - 649)	Fps1 (419 - 872)	22.3
Cbo4 (349 - 649)	Fps1 (464 - 872)	22.9
Cbo1 (349 - 649)	Fps1 (464 - 872)	21.9
Cno1 (322 - 649)	Fps1 (419 - 872)	23.6
Cbo2 (322 - 649)	Fps1 (419 - 872)	23.6
Cbo3 (349 - 649)	Fps1 (464 - 872)	21.7
Cno2 (349 - 649)	Fps1 (464 - 872)	23.0
Cbo5 (349 - 649)	Fps1 (627 - 879)	24.7
Bth1 (286 - 638)	Fps1 (383 - 863)	18.6
Bce1 (171 - 638)	Fps1 (364 - 863)	20.9
Lsp1 (330 - 649)	Fps1 (437 - 872)	19.5
Bbr1 (330 - 649)	Fps1 (437 - 872)	23.1
Pal1 (139 - 634)	Fps1 (354 - 859)	21.4
Cpe1 (330 - 649)	Fps1 (437 - 872)	24.2
Cbu1 (318 - 649)	Fps1 (420 - 872)	32.3
Sdy1 (245 - 649)	Fps1 (373 - 872)	21.5
Apy1 (347 - 640)	Fps1 (462 - 865)	29.9
Apy1 (655 - 763)	Nsp1 (728 - 899)	16.0
Ofo1 (318 - 641)	Fps1 (422 - 869)	15.3
Ofo1 (655 - 770)	Nsp1 (728 - 906)	25.4
Nsp1 (354 - 648)	Fps1 (469 - 871)	14.3

Table 10: Comparison Scores Between Revised List of Pleurotolysin and MACPF Homologues.

MACPF (Residues Compared)	Pleurotolysin (Residues Compared)	Average Score Expressed in S.D. (SSearch Program)	Location of the MACPF Domain on the MACPF Protein
Nve3 (192-372)	Afl2 (149-354)	13.6	222-417
Nve3 (194-401)	Cgl1 (332-537)	13.6	222-417
Bfl33 (547-725)	Afl2 (149-334)	12.8	585-760
Bfl32 (244-421)	Afl2 (149-334)	13.0	357-466
Bfl24 (859-1019)	Afl2 (166-334)	12.1	885-1054
Pvi1 (428-572)	Cli1 (350-489)	12.1	341-567
Clu6 (246-451)	Cgl1 (344-540)	12.2	292-497
Bfl4 (258-467)	Afl2 (132-333)	12.2	316-531
Ddi1 (22-208)	Pos1 (131-333)	18.7	30-220
Ddi1 (22-208)	Per1 (130-332)	17.9	30-220
Ddi1 (30-200)	Cli1 (288-463)	14.8	30-220
Ddi1 (10-224)	Nfi1 (334-555)	13.7	30-220
Ddi1 (21-239)	Afl2 (176-408)	19.6	30-220
Ddi1 (8-223)	Gze1 (296-540)	12.9	30-220
Ddi1 (21-210)	Cgl1 (357-557)	14.7	30-220

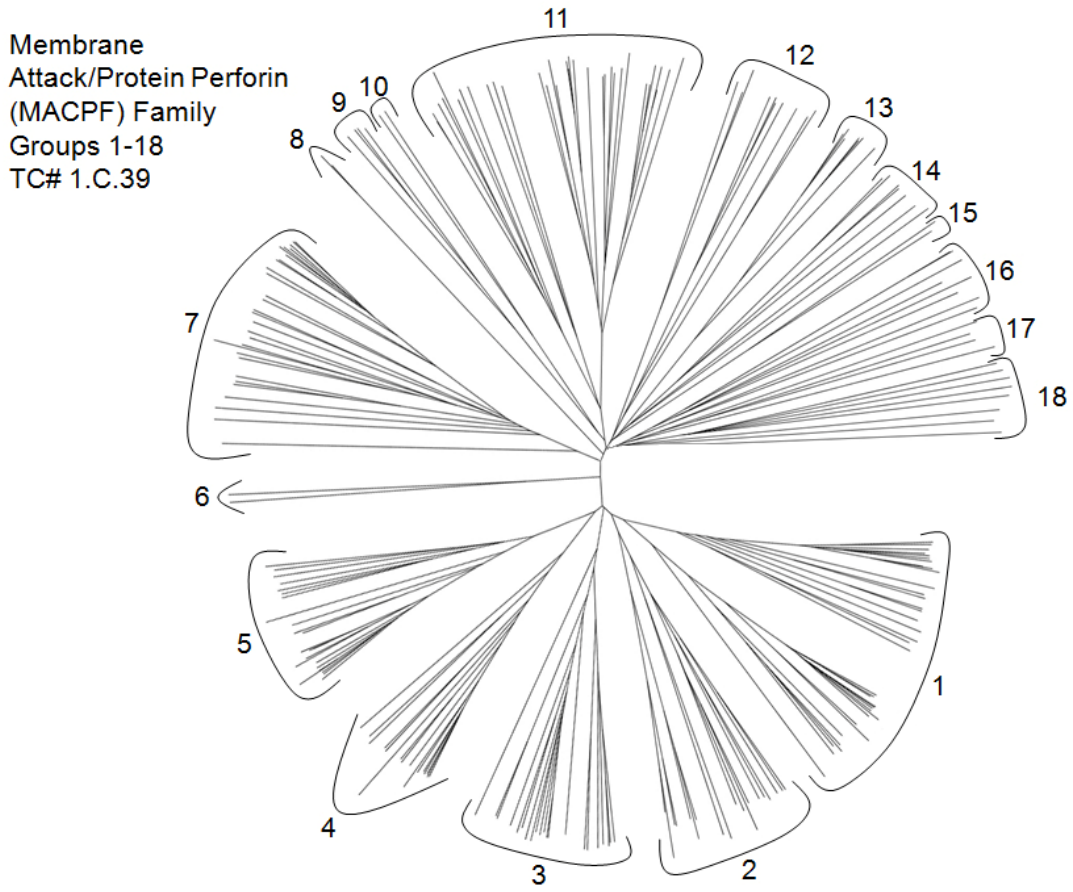


Figure 1. The phylogenetic tree containing all 234 MACPF homologues as listed in Table 1-1. The tree was generated using the ClustalX and FigTree programs and was subdivided into 18 clusters based on branching and clustering patterns as indicated.

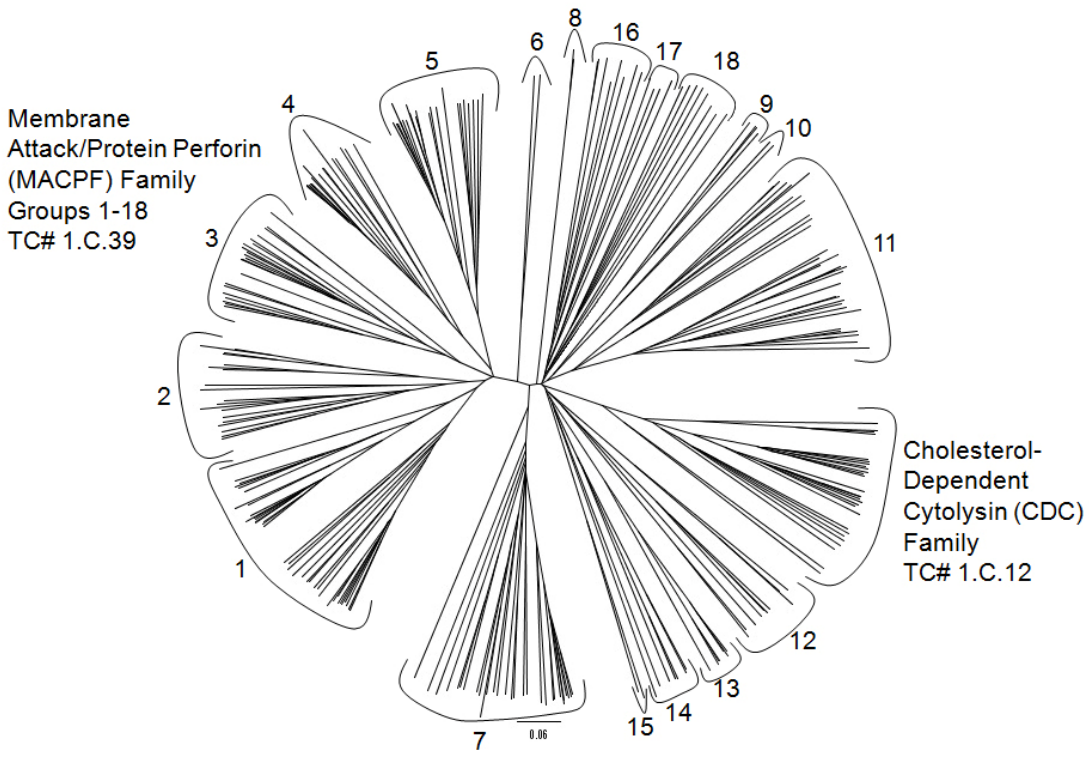


Figure 2. The phylogenetic tree generated by the addition of the Cholesterol-Dependent Cytolysin (CDC) family homologues to the MACPF homologues in Figure 1-1. The tree shows exclusive clustering of all 28 CDC homologues.

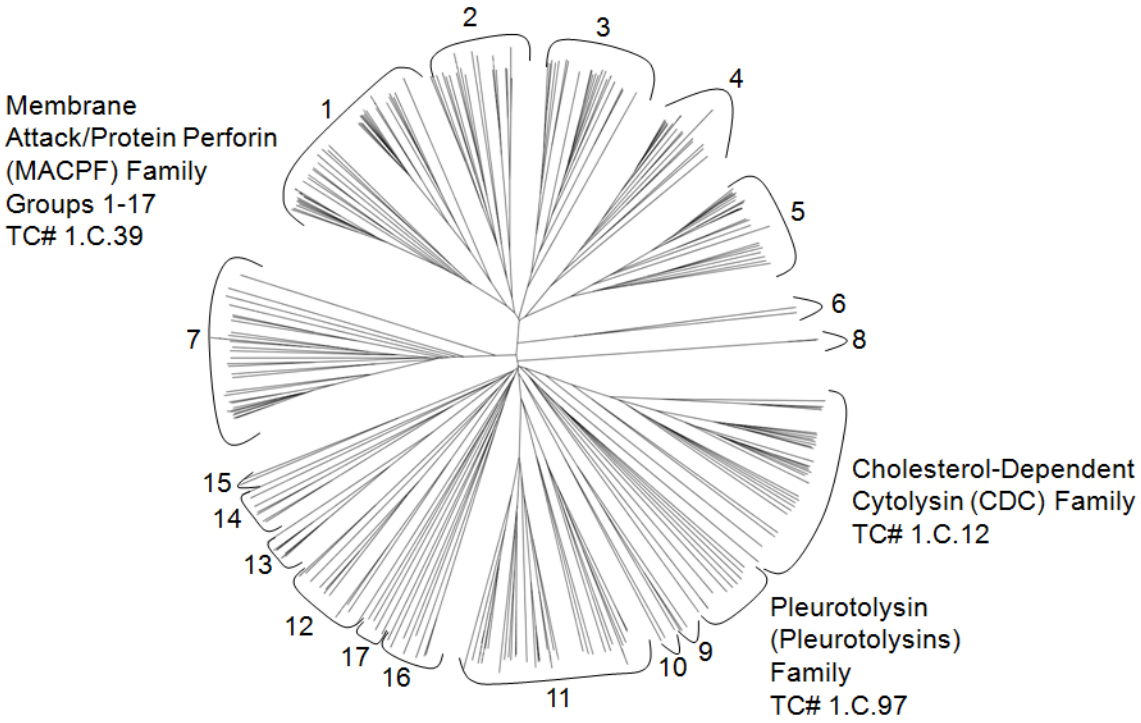


Figure 3. The phylogenetic tree generated by the addition of the 8 Pleurotolysin representatives from TCDB to the phylogenetic tree containing the CDC and MACPF homologues. The Pleurotolysin protein, Pkn1 (GI# 221052646), was shown to cluster with the MACPF family's Group 12 homologues. The Pleurotolysin proteins, Fps1 (GI# 150024210) and Nsp1 (GI# 17228824), were shown to cluster with the CDC homologues. The MACPF proteins, Nfi1 (GI# 119499704), Afl2 (GI# 220689182), Gze1 (GI# 46126573), Afl1 (GI# 220688529), and Afl3 (GI# 220693297) were shown to cluster with the Pleurotolysin proteins. These proteins have been resassigned different TC numbers according to the family with which they associate.

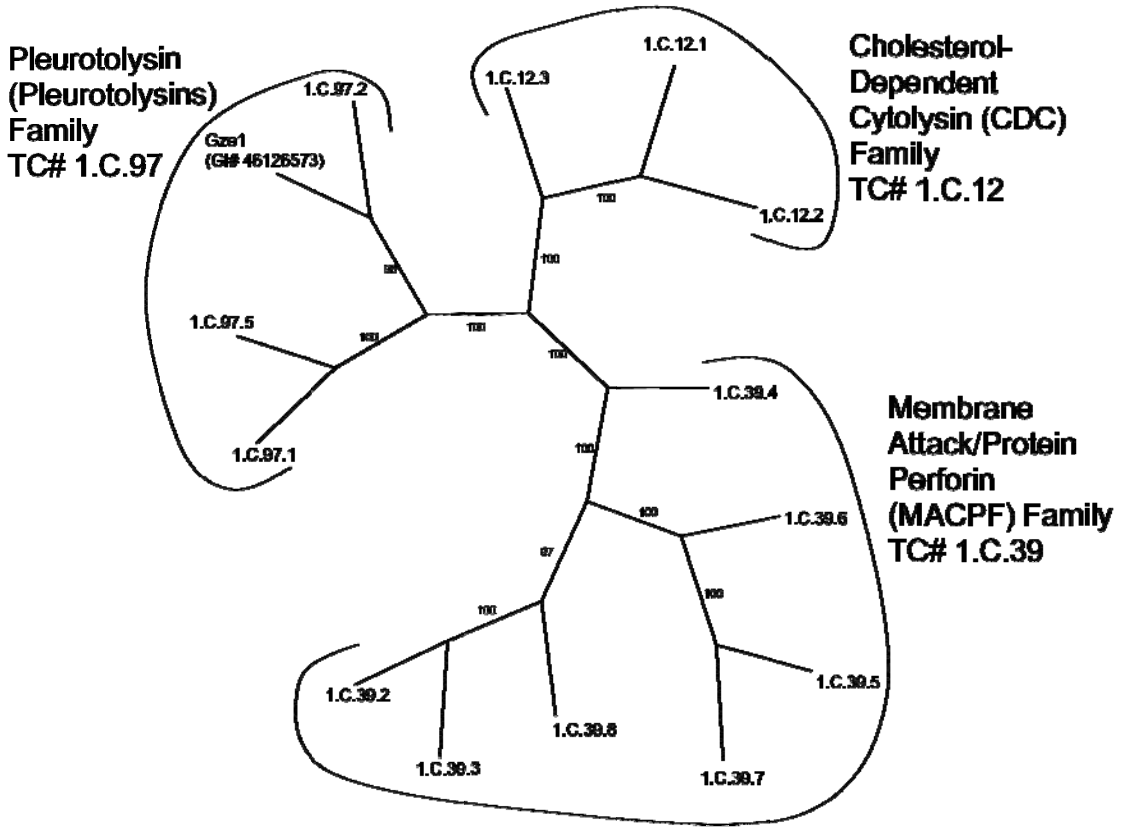


Figure 4. MACPF Superfamily Tree Generated from SFT. Clustering of Gze1 (originally a MACPF homologue collected from NCBI) with 1.C.97.2 confirmed its inclusion in the Pleurotolysin family.

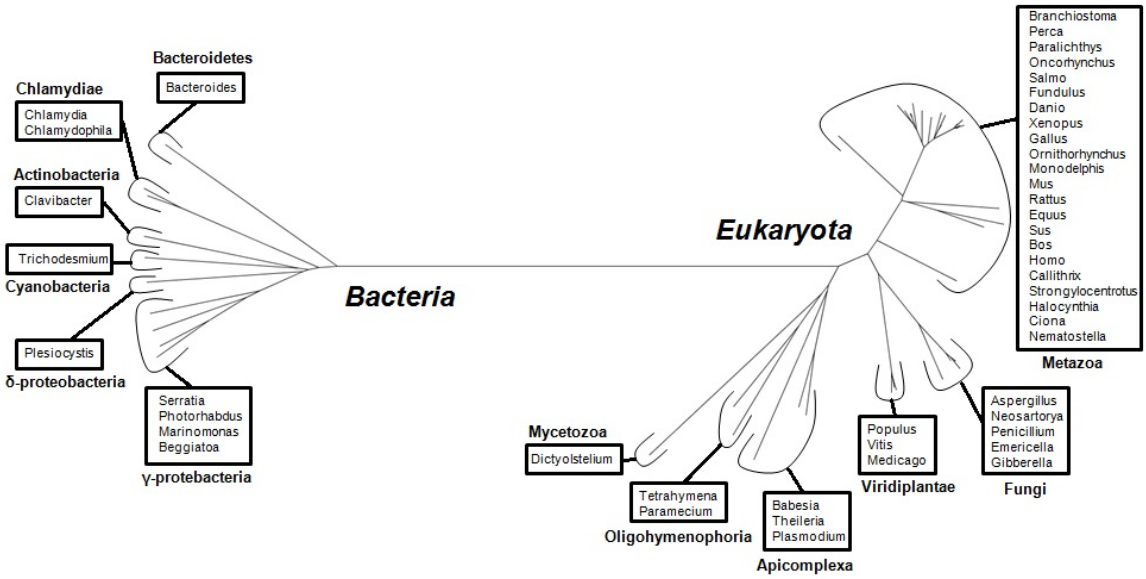


Figure 5. MACPF Family 16S/18S rRNA Gene Tree. Most genera from which our MACPF proteins were derived are included in this phylogenetic tree and are listed in clockwise order. The eukaryotic genera omitted from the rRNA tree due to the unavailability of complete 18S rRNA sequences include: *Pongo, Macaca, Canis, Felis,Tetraodon, Oryctolagus, Ginglymyostoma, Takifugu, Ctenopharyngodon,* and *Acropora.* The right-hand section of the tree shows exclusive clustering of eukaryotic organisms while the left-hand portion shows distinct clustering of bacteria. No archaeal homologues were identified.

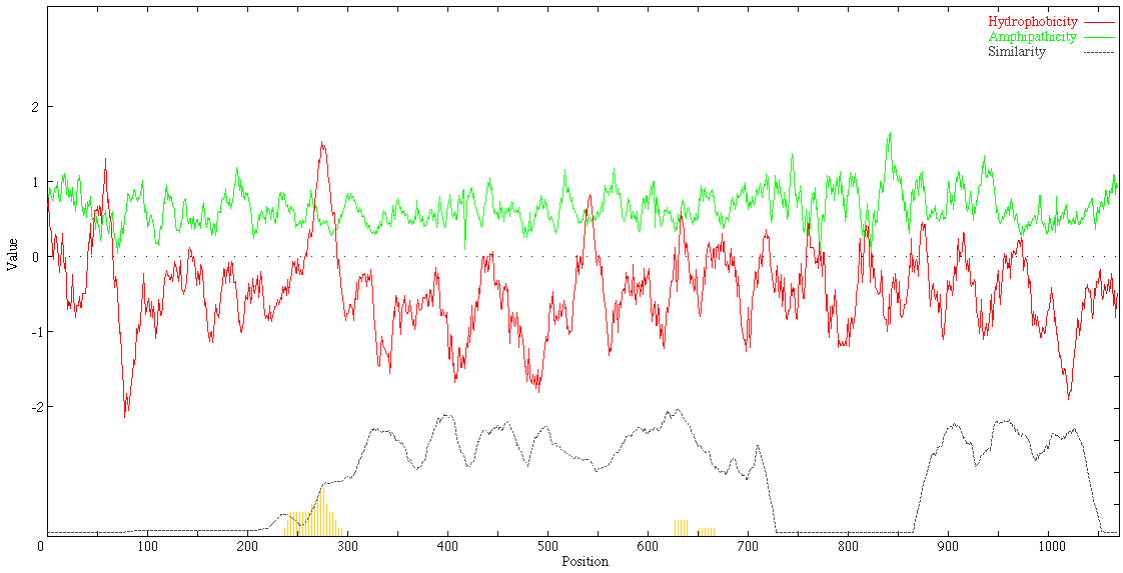


Figure 6. AveHAS plot of MACPF Family Cluster 1

Figure 7. AveHAS Plot of MACPF Family Cluster 2

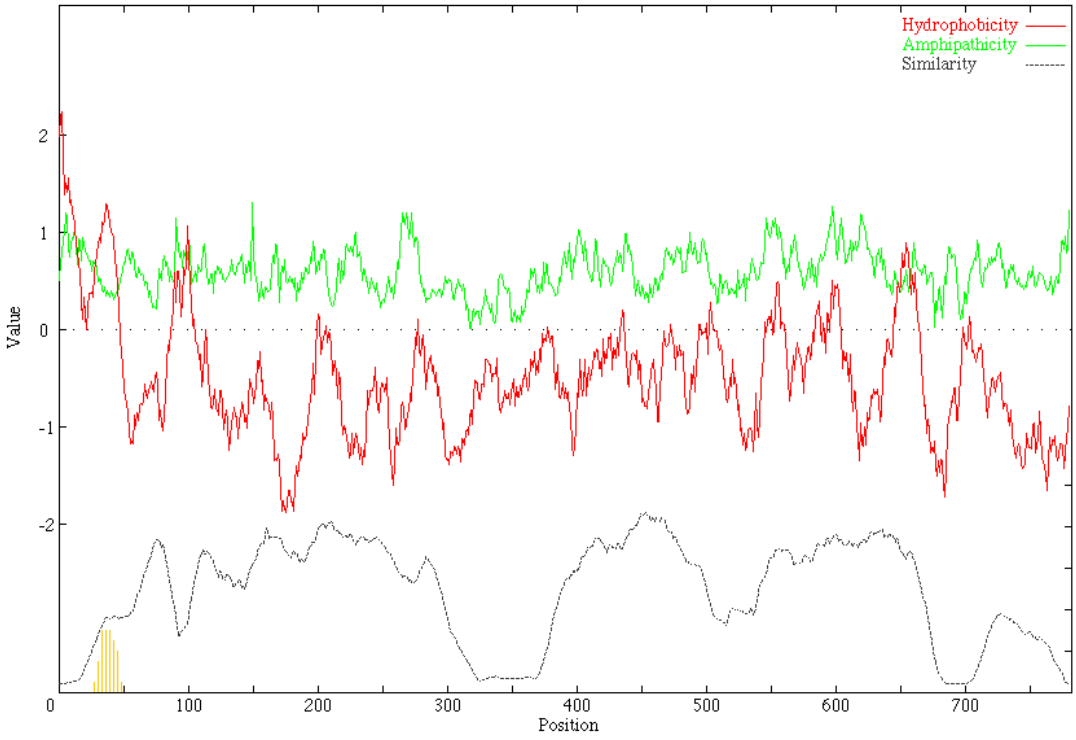


Figure 8. AveHAS Plot of MACPF Family Cluster 3

Figure 9. AveHAS Plot of MACPF Family Cluster 4

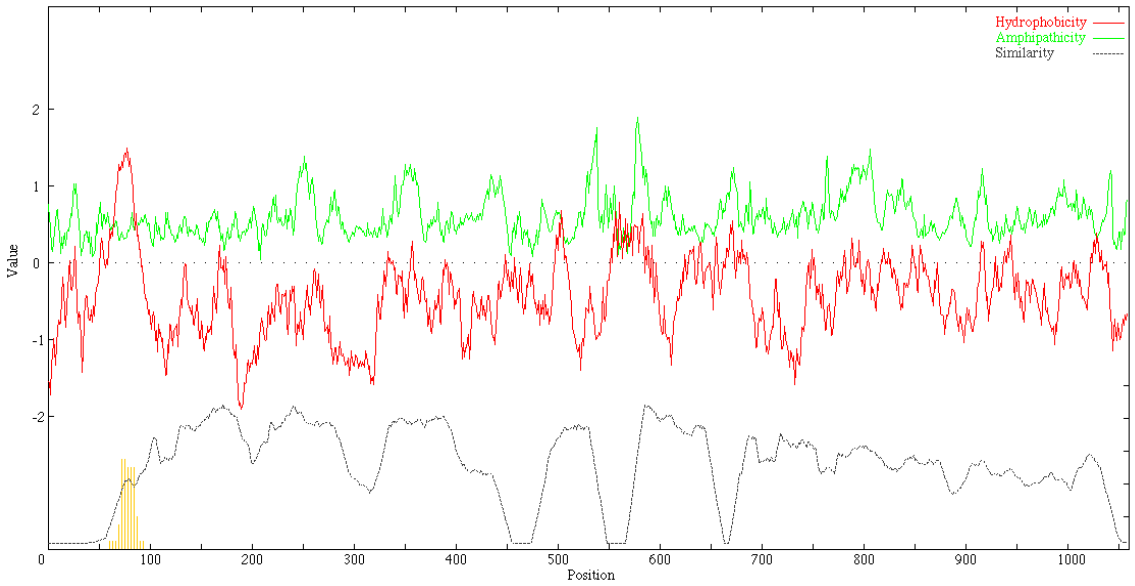


Figure 10. AveHAS Plot of MACPF Family Cluster 5

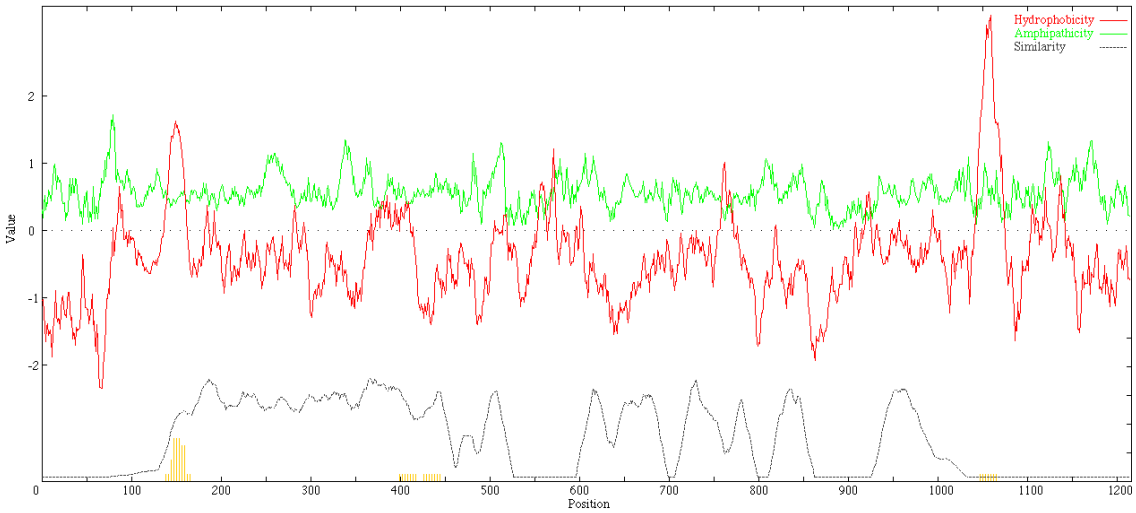


Figure 11. AveHAS Plot of MACPF Family Cluster 7

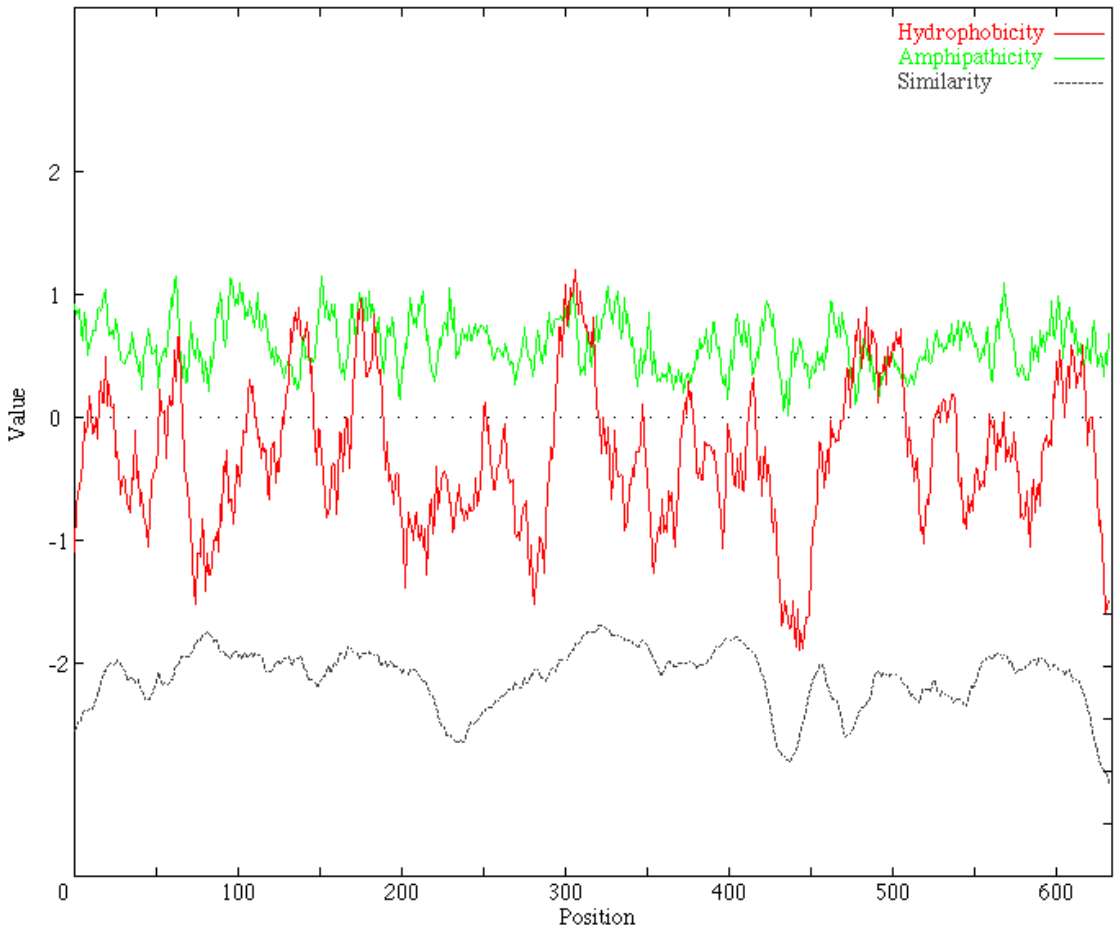
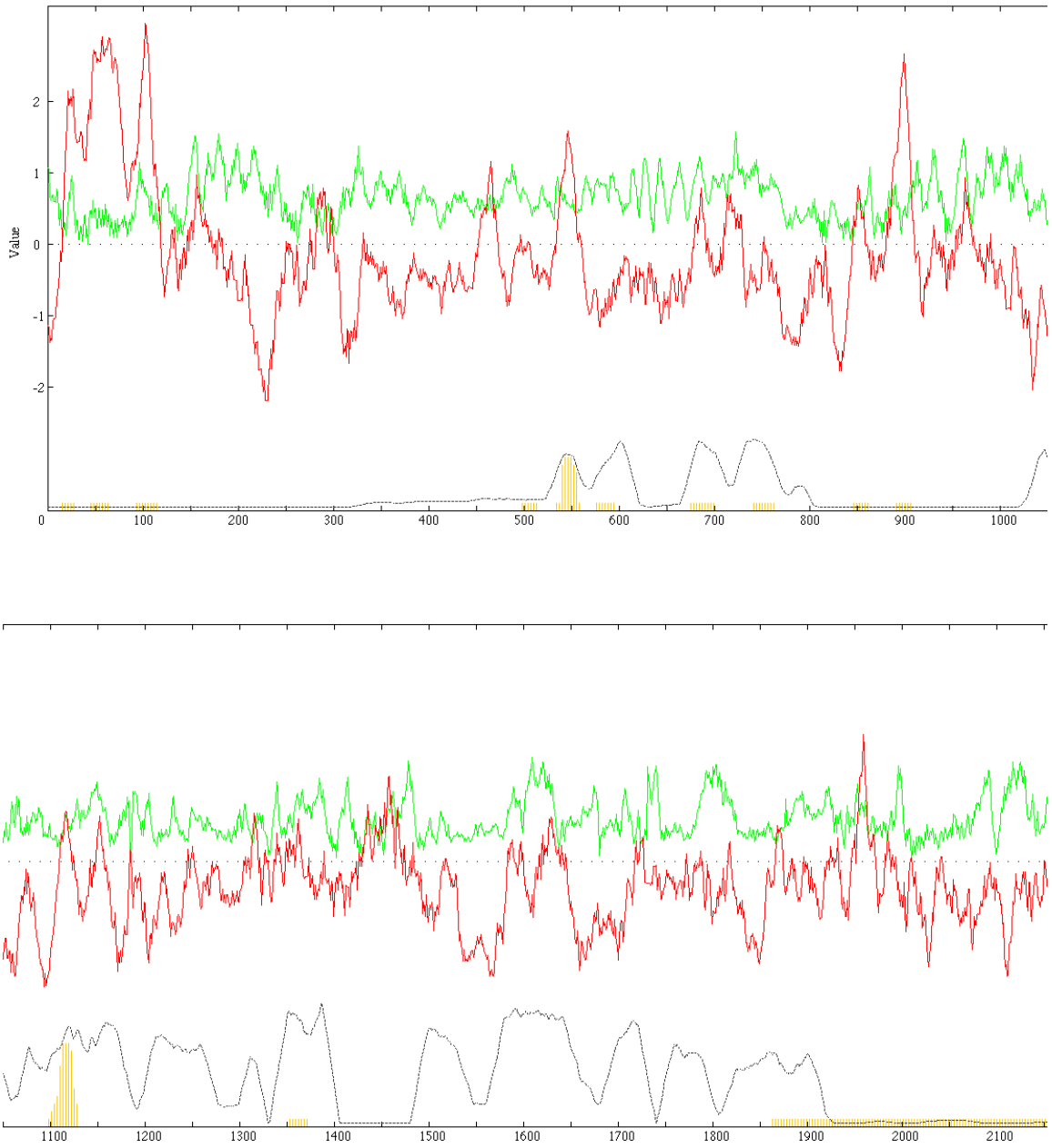


Figure 12. AveHAS Plot of MACPF Family Cluster 9


Value
2
1
0
-1
-2
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100

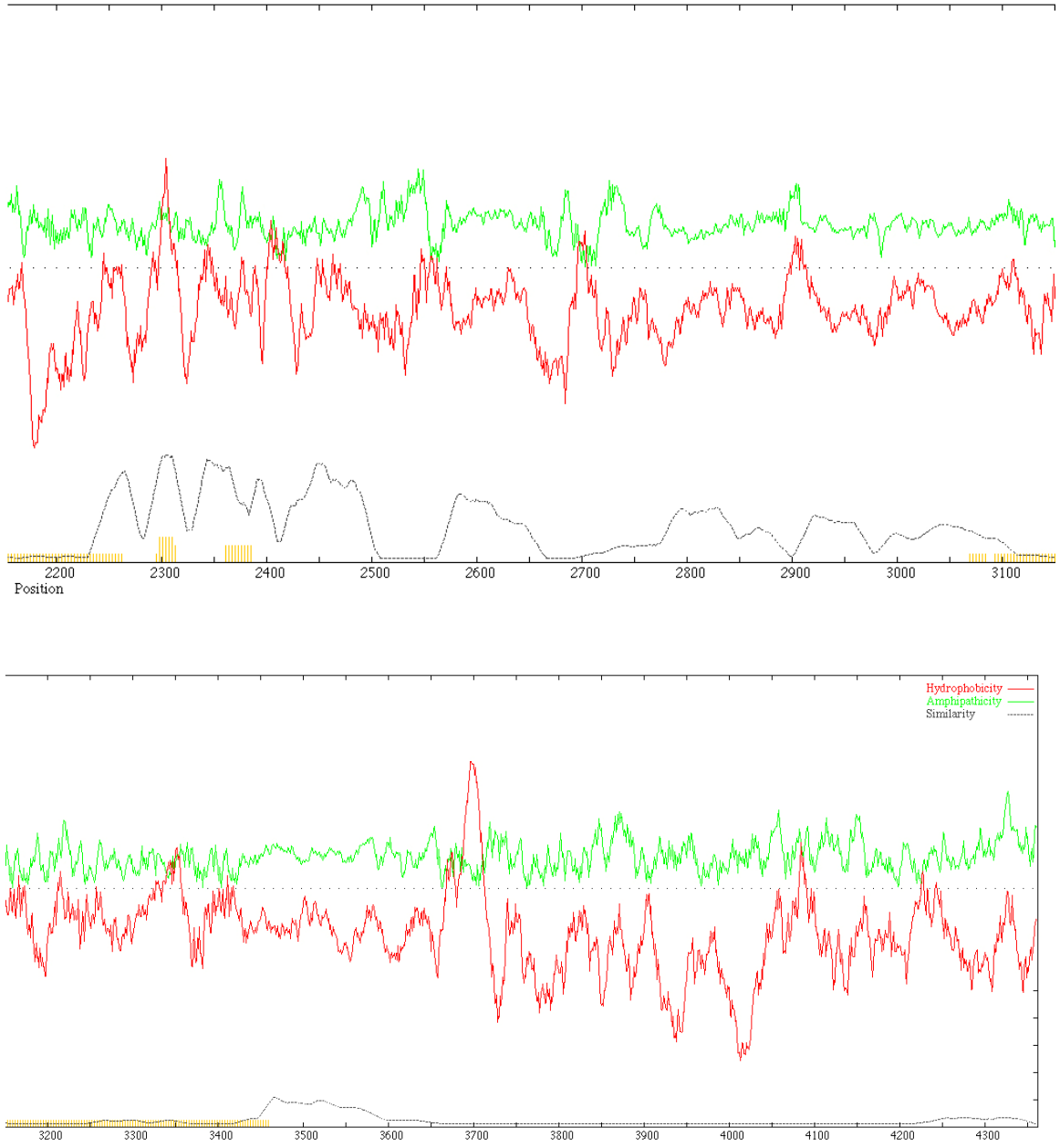


Figure 13. AveHAS Plot of MACPF Family Cluster 11

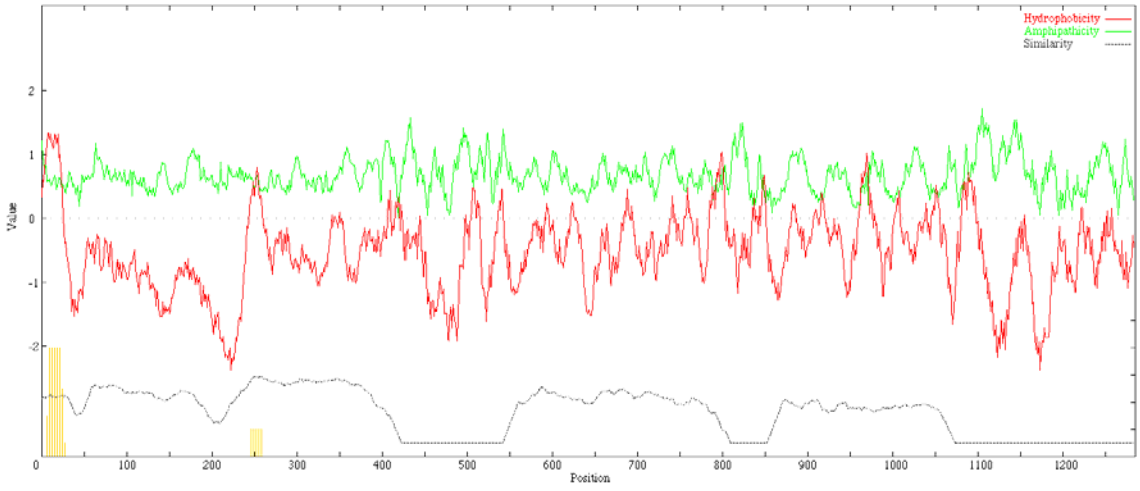


Figure 14. AveHAS Plot of MACPF Family Cluster 12

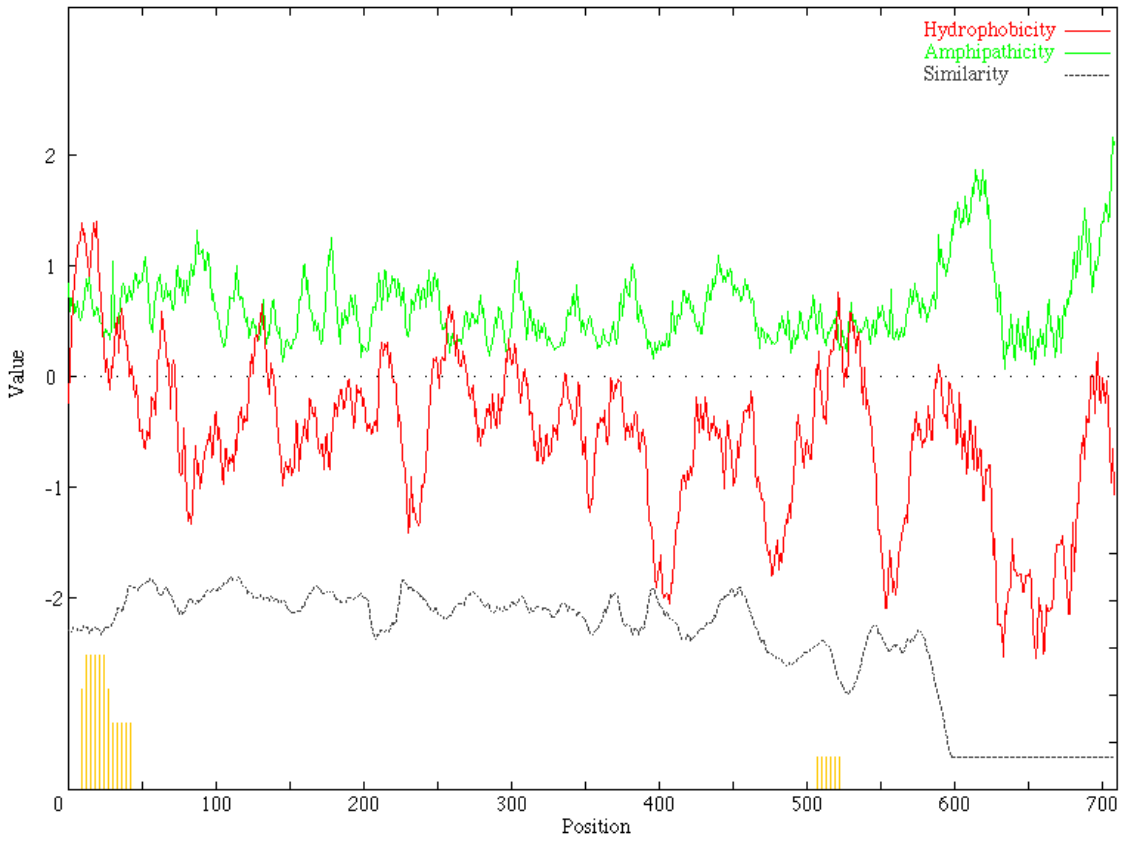


Figure 15. AveHAS Plot of MACPF Family Cluster 13

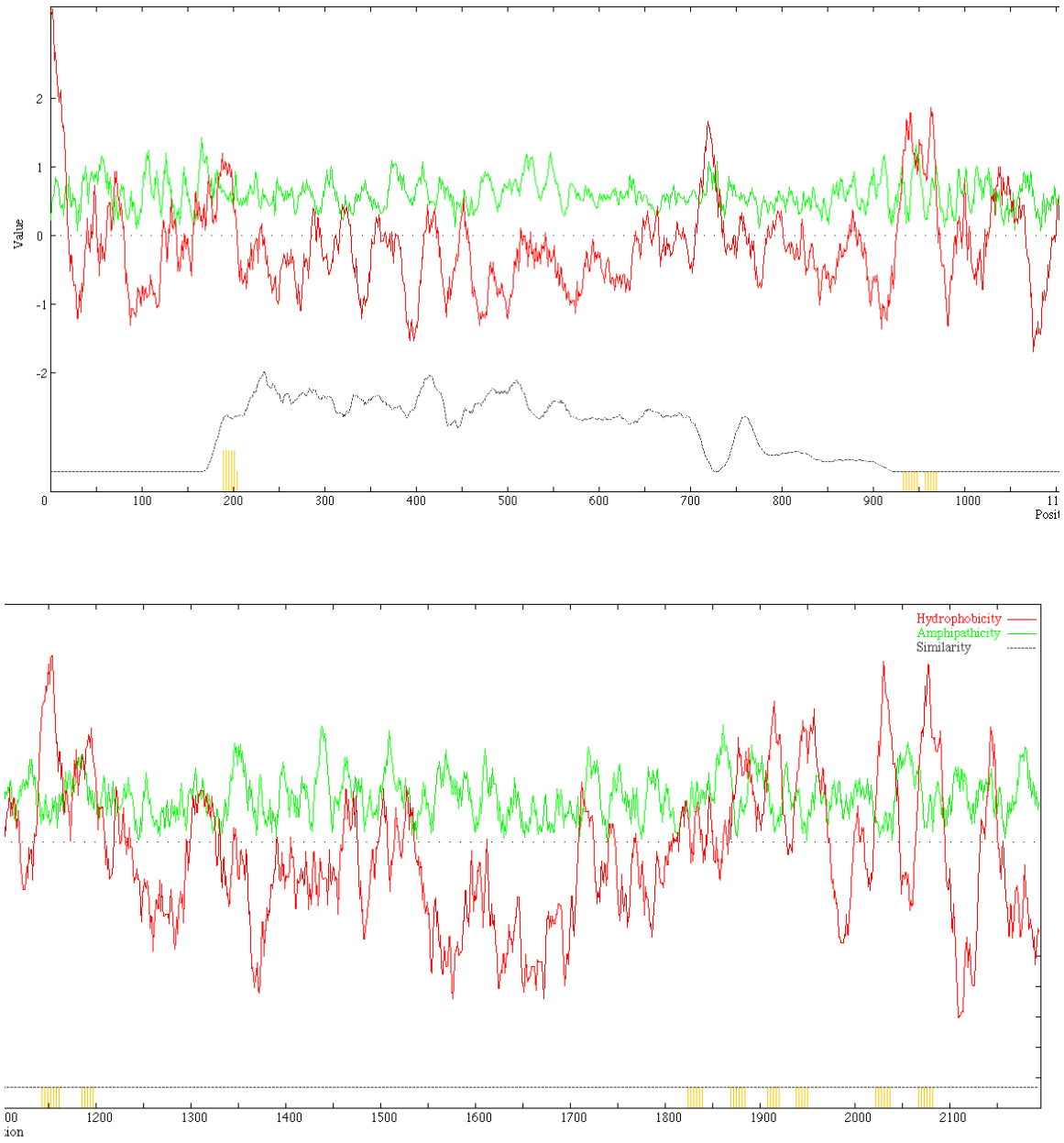


Figure 16. AveHAS Plot of MACPF Family Cluster 14

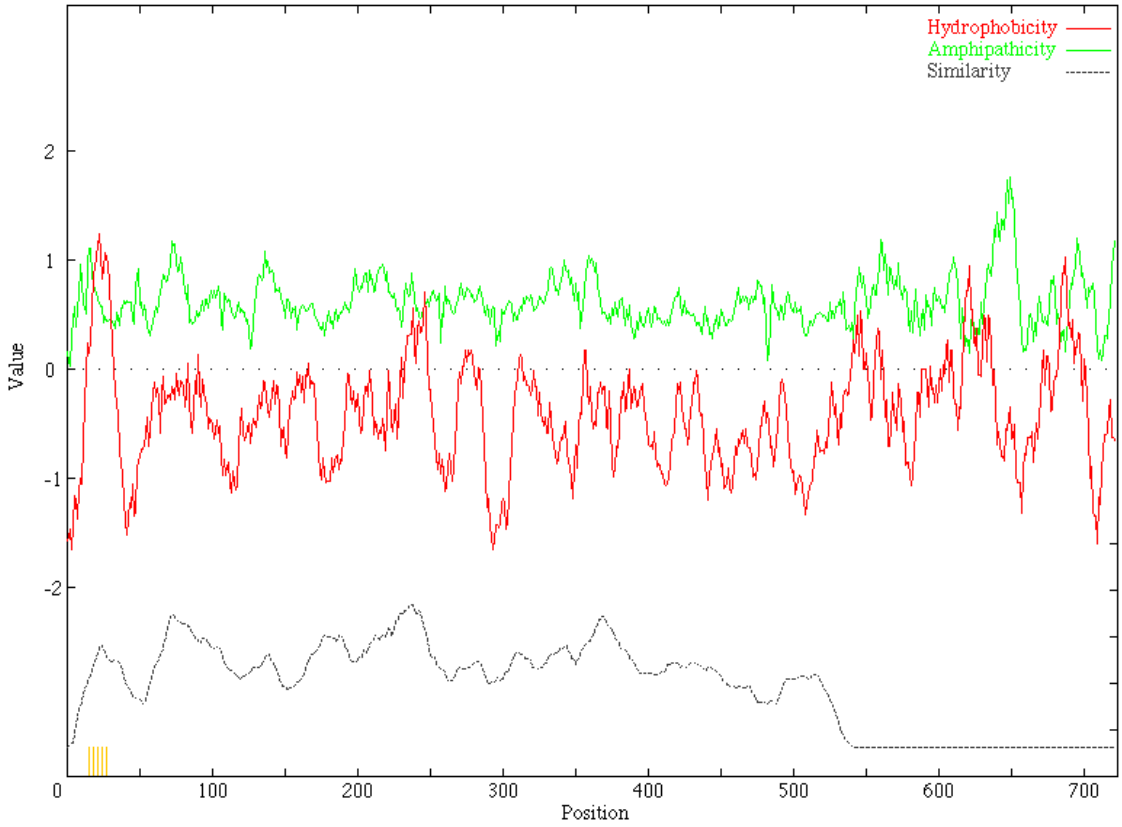


Figure 17. AveHAS Plot of MACPF Family Cluster 16

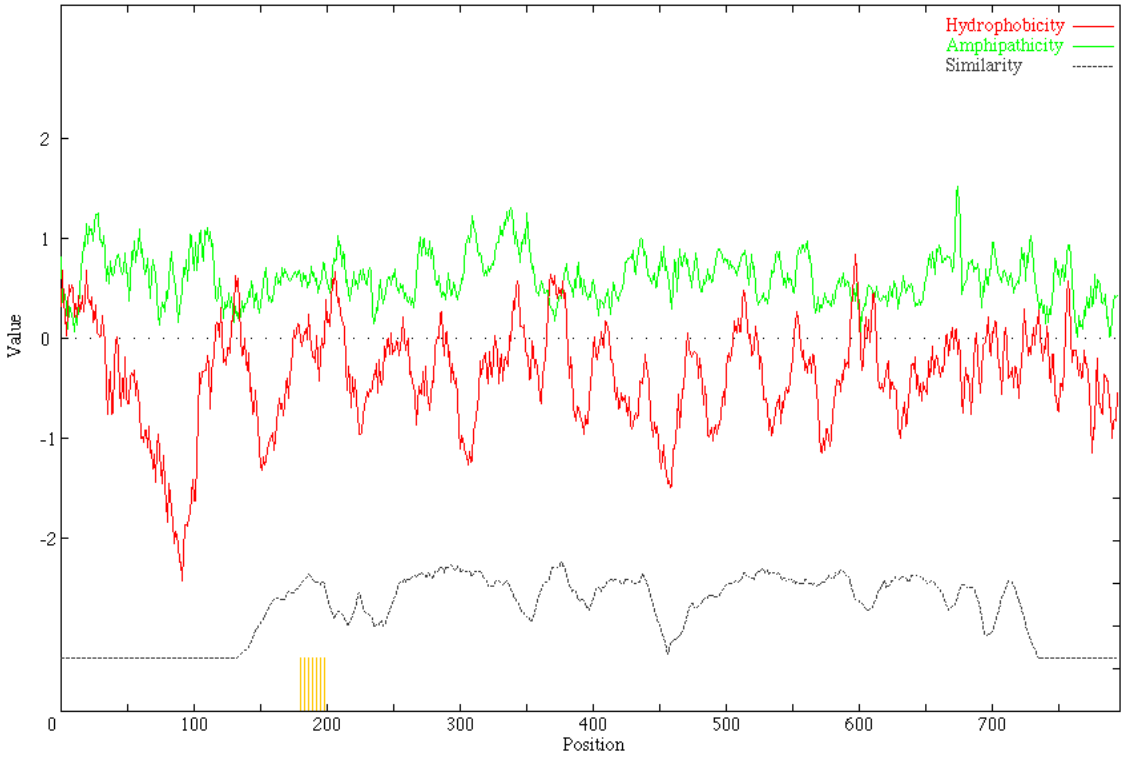


Figure 18. AveHAS Plot of MACPF Family Cluster 17

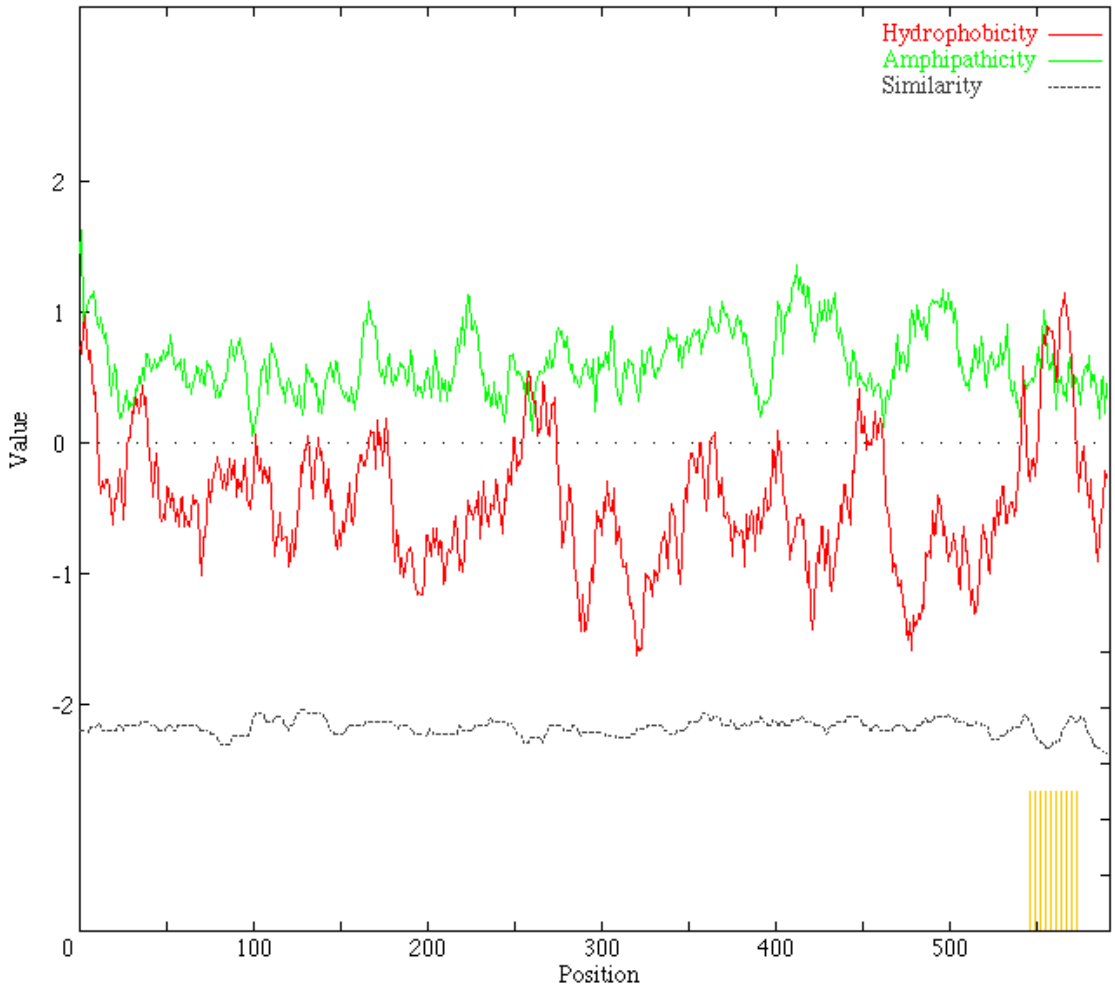


Figure 19. AveHAS Plot of MACPF Family Cluster 18

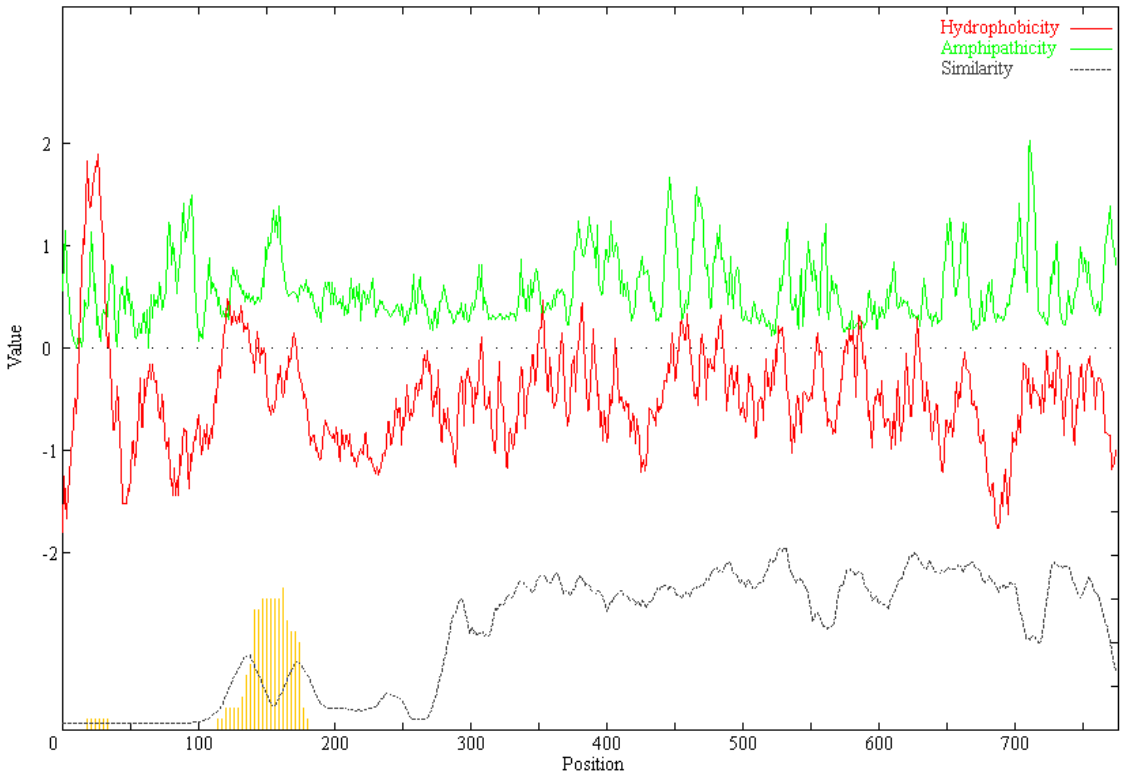


Figure 20. AveHAS Plot of CDC Homologues

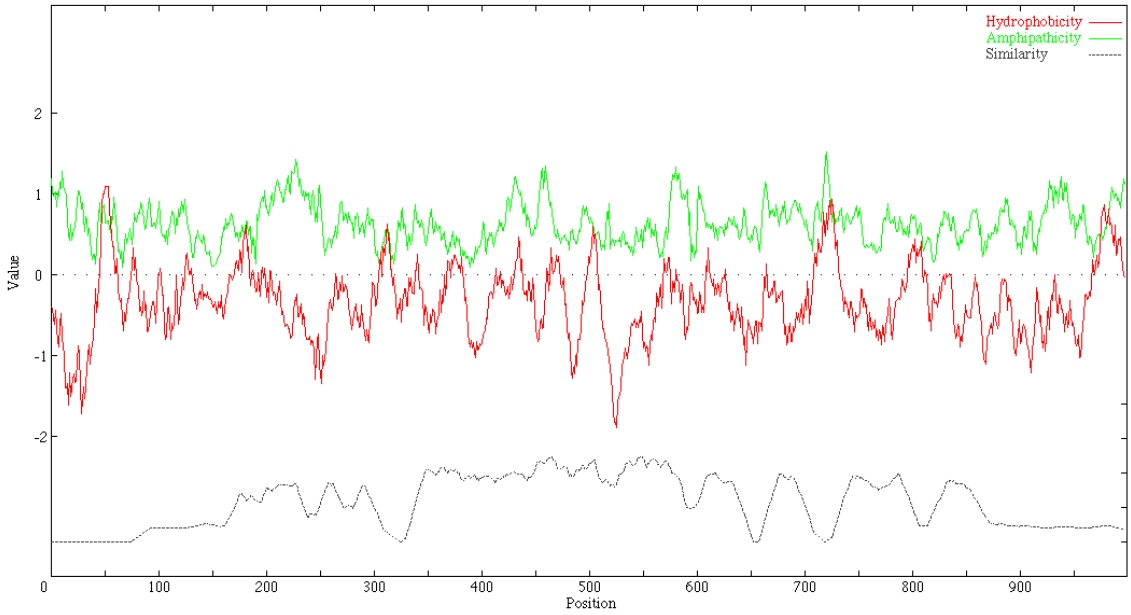


Figure 21. AveHAS Plot of Pleurotolysin Homologues

```
                .         .         .         .         .
Cbo2   1 DSVTLKELKAKGLNKDNPPAYVSNVAYG..RTIYVKLETTSKSLNVKAAF 48
         |.|| |: :   :|  .         |  |  : || :|  ..   | .
Omy3   1 DAVTGKQ.RGSVINTKSYGGQCRTVLSGDNKVIY.RLPQSTLRYNFEVKV 48

                .         .         .         .         .
Cbo2  49 KALIKNQDISGNMEY.KDILNQSSFTATVLGGGAKEHNKVITKNFDEIRE 97
         .   .:  . .  | |||. . . | |  |    : ..   || .
Omy3  49 QNDFSDEFYTSSWSYAKDIVKRETTTGTTTGFNNYDLHQTEEKNRNNHLL 98

                .         .         .         .
Cbo2  98 VIKNN...AEYSPQNPAYPISYTTTFLKDNAVATINSKTDY 135
         |:|||   |::  | | | :| .  | |   .||: .  ||
Omy3  99 VVKNNVEVAQFQNQAPGY.LSLSEEFWK..VLATLPTVYDY 136
```

Figure 22. GAP Optimization Alignment of Omy3 & Cbo2. GAP alignment of the residues compared in SSearch showed a percent identity of 27.0%, 36.2% similarity, and a comparison score of 10.8 S.D.

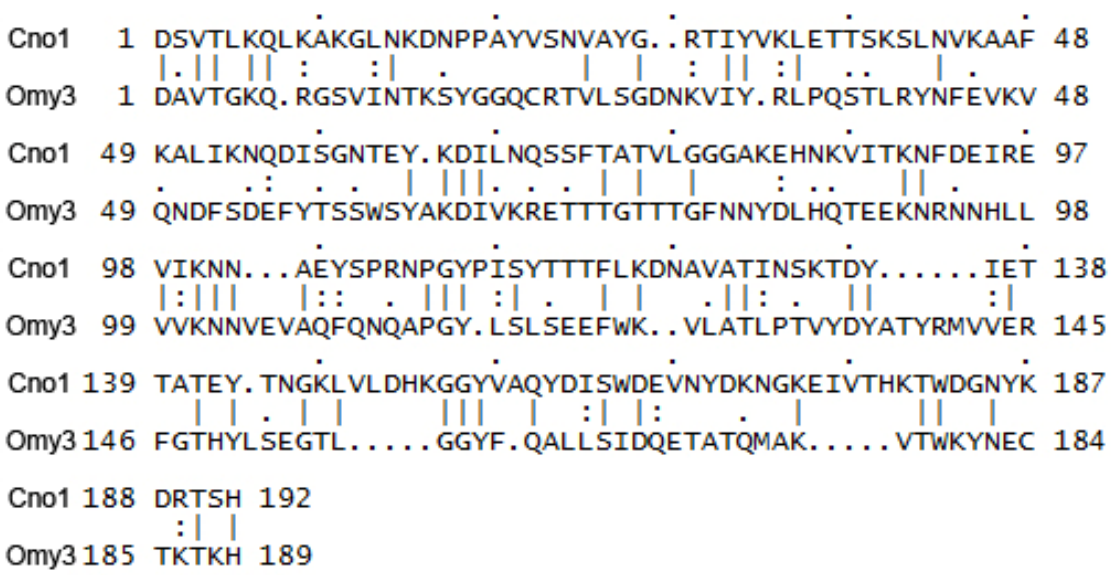


Figure 23. GAP Optimization Alignment of Omy3 & Cno1. GAP alignment of the residues compared in SSearch showed a percent identity of 30.1%, 38.6% similarity, and a comparison score of 10.9 S.D.

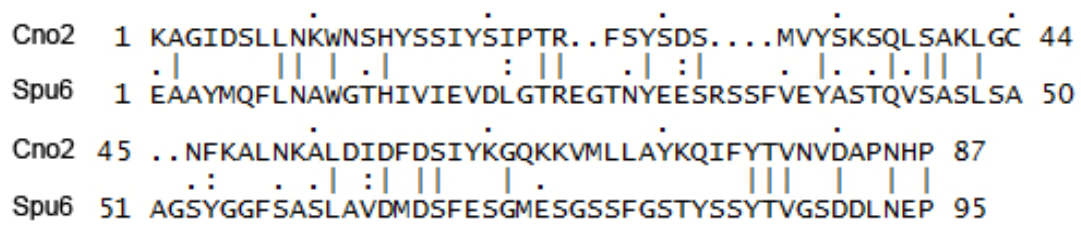


Figure 24. GAP Optimization Alignment of Spu6 & Cno2. GAP alignment of the residues compared in SSearch showed a percent identity of 28.7%, 33.3% similarity, and a comparison score of 12.5 S.D.

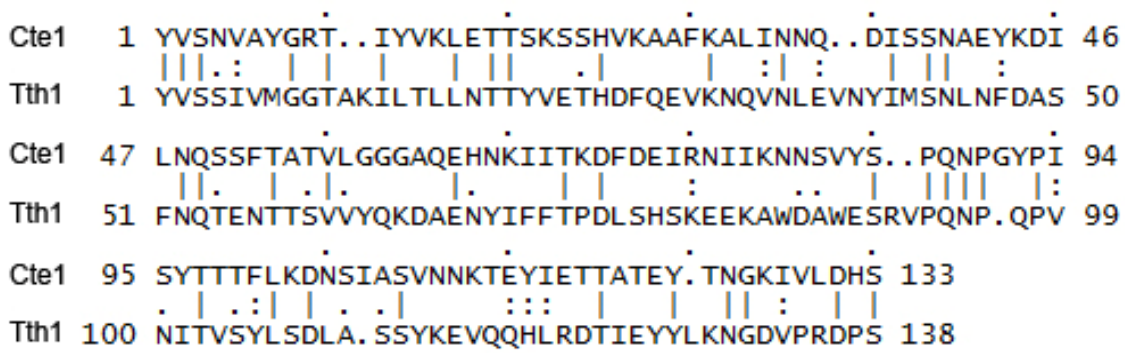


Figure 25. GAP Optimization Alignment of Tth1 & Cte1. GAP alignment of the residues compared in SSearch showed a percent identity of 29.0%, 37.4% similarity, and a comparison score of 10.9 S.D.

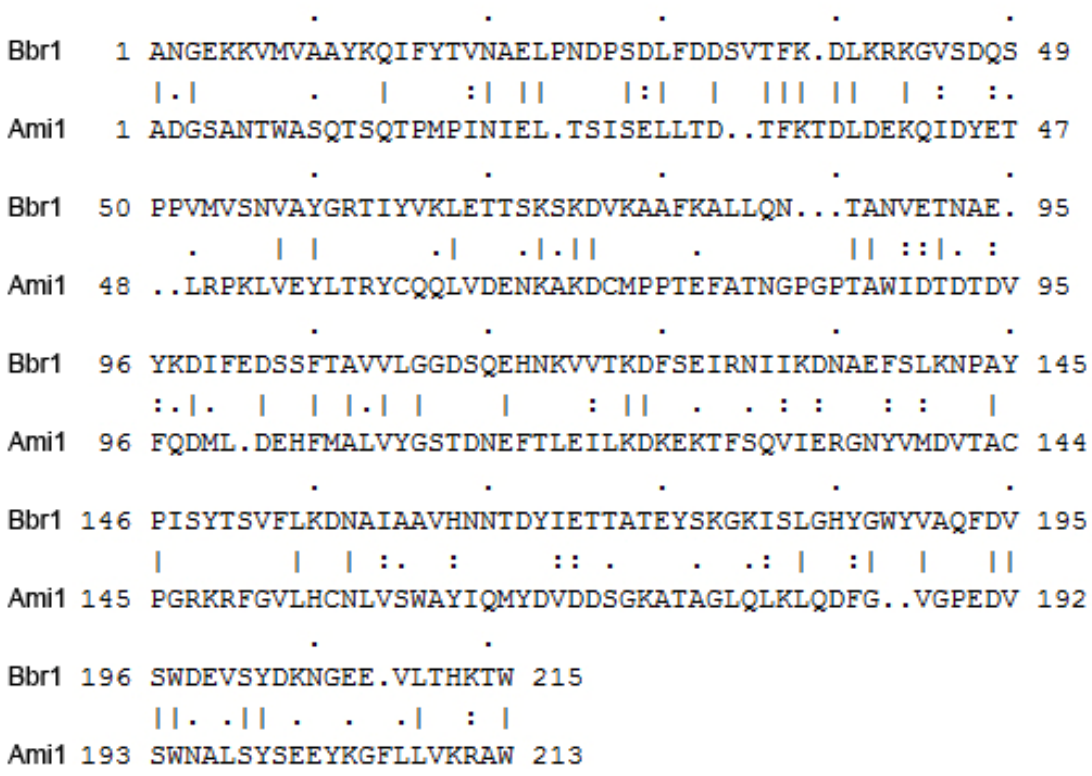


Figure 26. GAP Optimization Alignment of Ami1 & Bbr1. GAP alignment of the residues compared in SSearch showed a percent identity of 23.1%, 32.9% similarity, and a comparison score of 12.9 S.D.

```
                .         .         .         .         .
Cbo5   1 TTTFLKN.NGIATVNNKTDYI...ETTATEY.TNGKLVLDHSGAYVAQFN 45
         ||||| .   : |   | :||   ||  | | ..| |     | |  :
Eca2   1 TTTFLDDIKALPTAYEKGEYIAFLETYGTHYSSSGSL....GGLY..ELI 44
                .         .         .         .         .
Cbo5  46 ITWDEVSYDKKGNEIVEHKAWSGNNRDRTAHFNTEIYLKGNSRNICIKAK 95
            |. | |.|| |: : .   | | | .     |:  |   :| |:|
Eca2  45 YVLDKASMDQKGVELRDIQRCLGFNLDLSLKDKYEVTAK.IDKNDCLKRN 93

Cbo5  96 E 96
         |
Eca2  94 E 94
```

Figure 27. GAP Optimization Alignment of Eca2 & Cbo5. GAP alignment of the residues compared in SSearch showed a percent identity of 36.0%, 44.9% similarity, and a comparison score of 11.0 S.D.

```
                .         .         .         .         .
Cte1   1 ISYTTTFLKD.NSIASVNNKTEY...IETTATEYTNGKIVLDHSGAYVAQ 46
         :  ||||| |  .: |   | ||   :||  | |..   .    | |
Clu7   1 VMLTTTFLDDIKALPSTYEKGEYFAFLETYGTHYSSSGSL....GGYYEL 46
                .         .         .         .
Cte1  47 FQVTWDEVSYDEKGNEIVEHKAWEGNNRDRTAHFNTEIYLKGN 89
           |  |. | :||| |: : .   | | | .     ||  | |
Clu7  47 IYVL.DKASMEEKGVELRDVQRCLGFNLDFSLEAGVEISGKLN 88
```

Figure 28. GAP Optimization Alignment of Clu7 & Cte1. GAP alignment of the residues compared in SSearch showed a percent identity of 35.7%, 42.9% similarity, and a comparison score of 12.4 S.D.

```
                .         .         .         .         .
Clu7   1 VMLTTTFLDDIKALPSTYEKGEYFAFLETYGTHYSSSGSL.....GGYYE 45
         :  ||||| .   : .   | :|   :||  | | ..| |     | |
Cbo5   1 ISYTTTFLKN.NGIATVNNKTDY...IETTATEY.TNGKLVLDHSGAYVA 45
                .         .         .         .         .
Clu7  46 LIYVL.DKASMEEKGVELRDVQRCLGFNLDFSLEAGVEISGKLN.KDDCL 93
            :  |. | :.|| |: : .   | | | .     ||  | | :. |:
Cbo5  46 QFNITWDEVSYDKKGNEIVEHKAWSGNNRDRTAHFNTEIYLKGNSRNICI 95

Clu7  94 KRGE 97
         |  |
Cbo5  96 KAKE 99
```

Figure 29. GAP Optimization Alignment of Clu7 & Cbo5. GAP alignment of the residues compared in SSearch showed a percent identity of 32.6%, 43.5% similarity, and a comparison score of 13.3 S.D.

```
                .         .         .         .         .
Rno6   1 VMLTTTFLDDVKALPVSYEKGEYFGFLETYGTHYSSSGSL....GGLY.. 44
         :  ||||| .   :     | :|   :||  | | ..| |     | |
Cbo5   1 ISYTTTFLKN.NGIATVNNKTDY...IETTATEY.TNGKLVLDHSGAYVA 45
                .         .         .         .         .
Rno6  45 ELIYVLDKASMKEKGVELSDVKRCLGFNLDVSLYTPLQTTLEGPSLTANV 94
         :     |. |  .|| |: : |   | | | . :   :  |.| |    :
Cbo5  46 QFNITWDEVSYDKKGNEIVEHKAWSGNNRDRTAHFNTEIYLKGNSRNICI 95

Rno6  95 NHSDC 99
            :|
Cbo5  96 KAKEC 100
```

Figure 30. GAP Optimization Alignment of Rno6 & Cbo5. GAP alignment of the residues compared in SSearch showed a percent identity of 29.8%, 41.5% similarity, and a comparison score of 10.2 S.D.

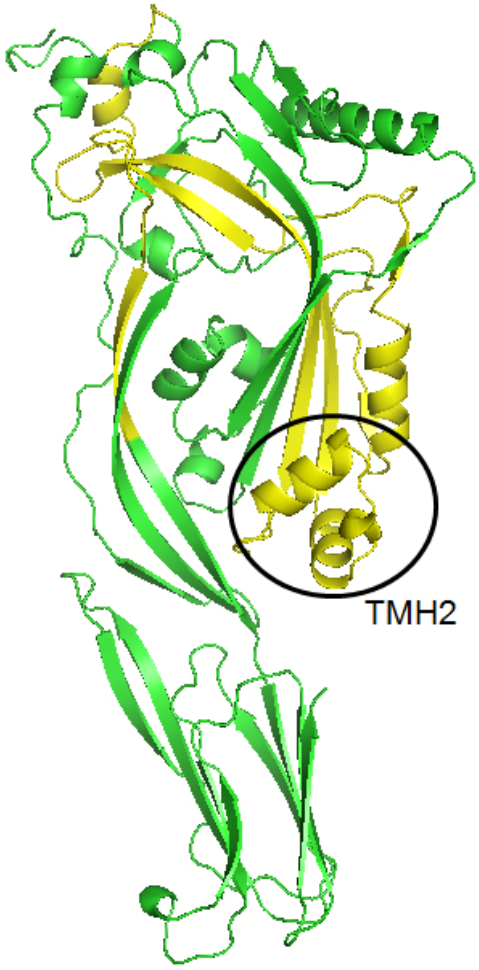


Figure 31. GAP Comparison of Omy3 & Cbo2 Superimposed on 1PFO. Green indicates the Perfringolysin O chain that was not included in the alignment. Yellow indicates where the CDC protein, Cbo2, aligned with both Omy3 and the 1PFO protein using GAP.

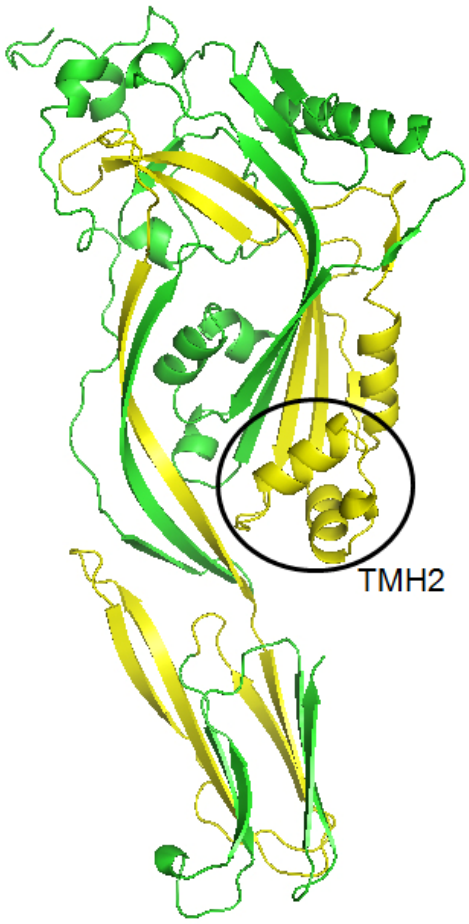


Figure 32. GAP Comparison of Omy3 & Cno1 Superimposed on 1PFO. Green indicates the Perfringolysin O chain that was not included in the alignment. Yellow indicates where the CDC protein, Cno1, aligned with both Omy3 and the 1PFO protein using GAP.

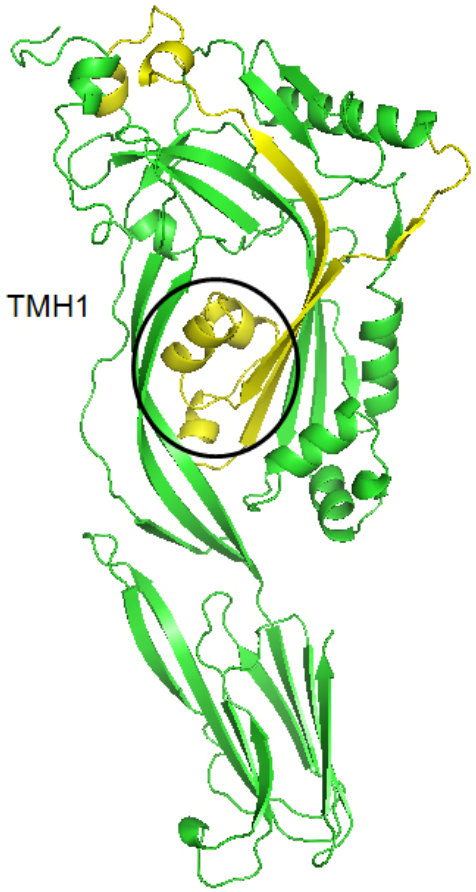


Figure 33. GAP Comparison of Spu6 & Cno2 Superimposed on 1PFO. Green indicates the Perfringolysin O chain that was not included in the alignment. Yellow indicates where the CDC protein, Cno2, aligned with both Spu6 and the 1PFO protein using GAP.

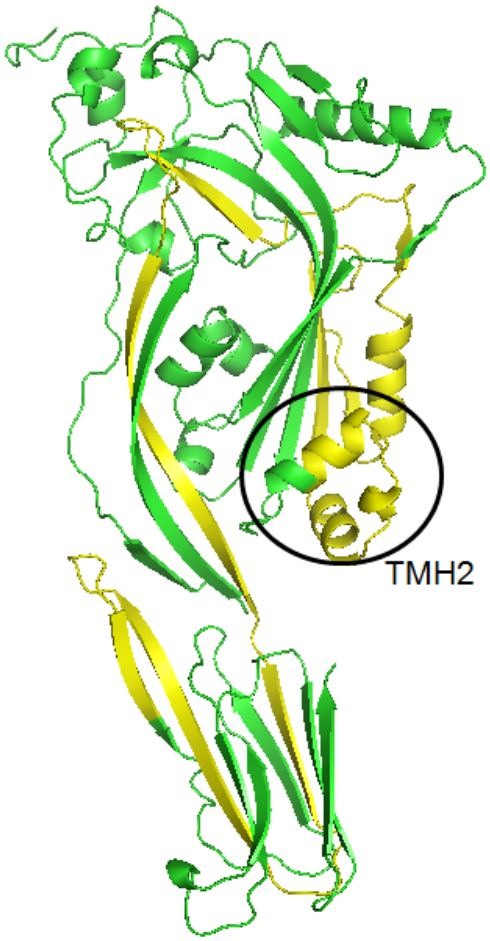


Figure 34. GAP Comparison of Tth1 & Cte1 Superimposed on 1PFO. Green indicates the Perfringolysin O chain that was not included in the alignment. Yellow indicates where the CDC protein, Cte1, aligned with both Tth1 and the 1PFO protein using GAP.

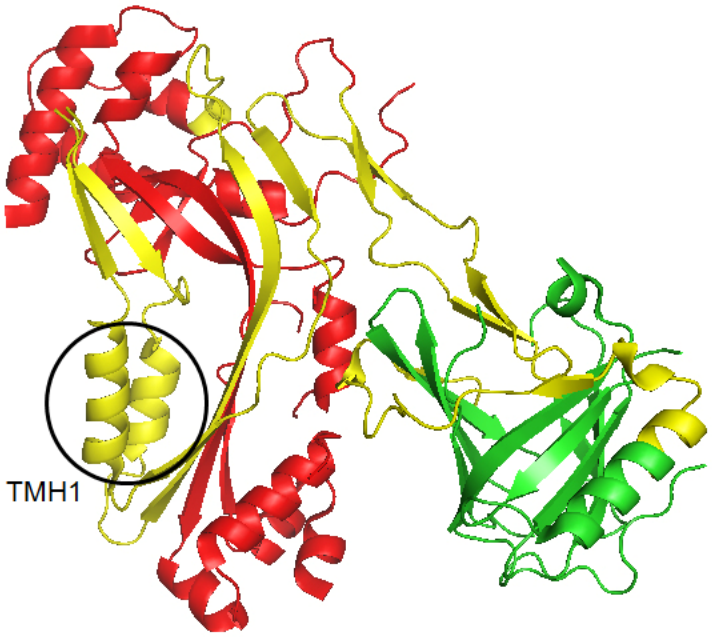


Figure 35. GAP Comparison of Omy3 & Cbo2 Superimposed on 2RD7. Red indicates the complement component C8 alpha chain. Green indicates the complement component C8 gamma chain. Yellow indicates the residues where Omy3 was found to align with both Cbo2 and the 2RD7 protein using GAP.

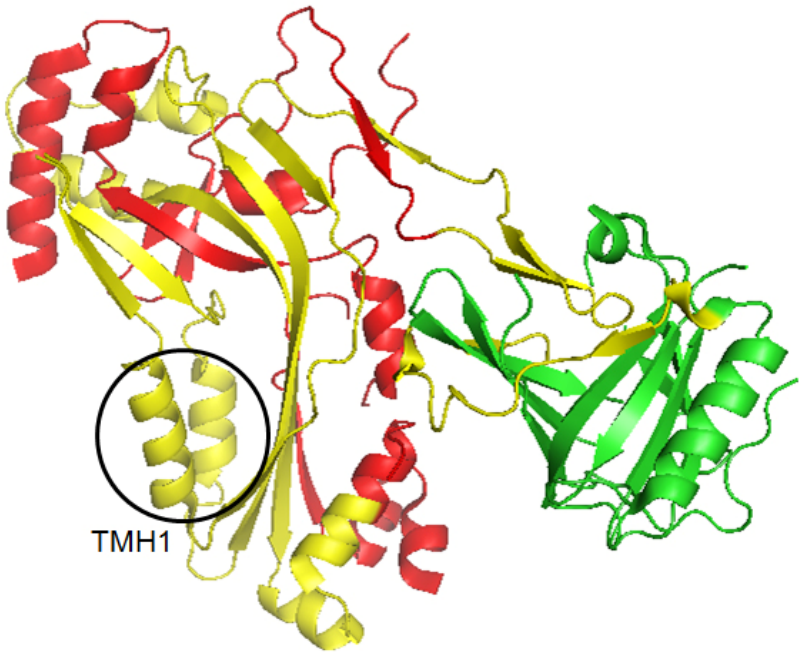


Figure 36. GAP Comparison of Omy3 & Cno1 Superimposed on 2RD7. Red indicates the complement component C8 alpha chain. Green indicates the complement component C8 gamma chain. Yellow indicates the residues where Omy3 was found to align with both Cno1 and the 2RD7 protein using GAP.

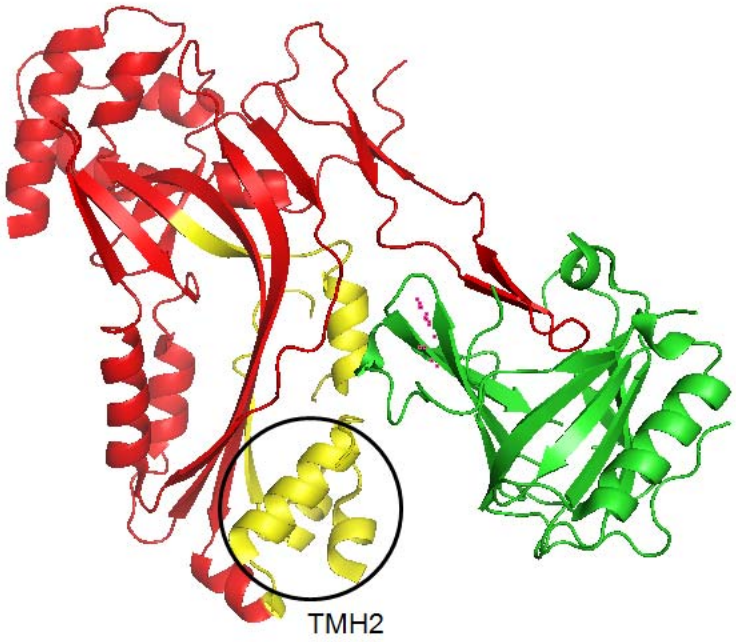


Figure 37. GAP Comparison of Spu6 & Cno2 Superimposed on 2RD7. Red indicates the complement component C8 alpha chain. Green indicates the complement component C8 gamma chain. Yellow indicates the residues where Spu6 was found to align with both Cno2 and the 2RD7 protein using GAP.

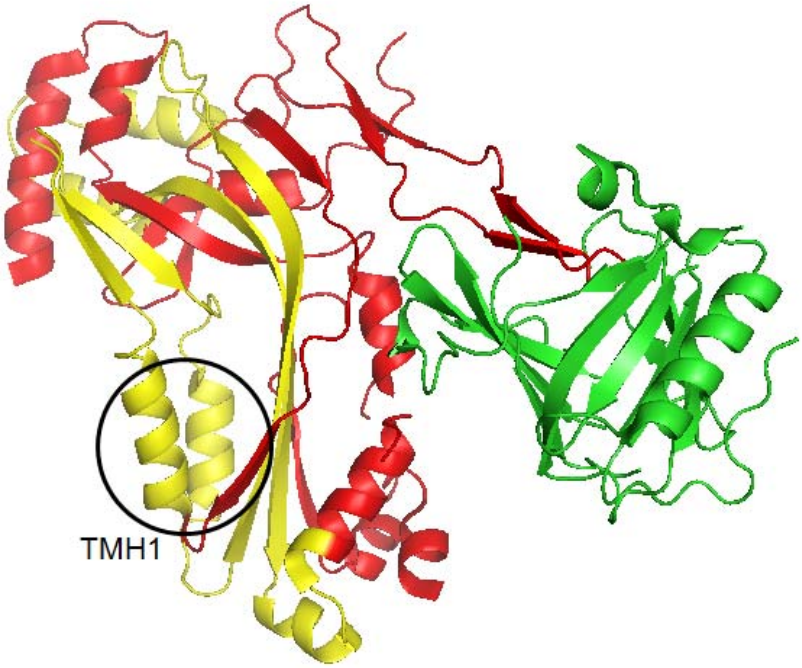


Figure 38. GAP Comparison of Tth1 & Cte1 Superimposed on 2RD7. Red indicates the complement component C8 alpha chain. Green indicates the complement component C8 gamma chain. Yellow indicates the residues where Tth1 was found to align with both Cte1 and the 2RD7 protein using GAP.

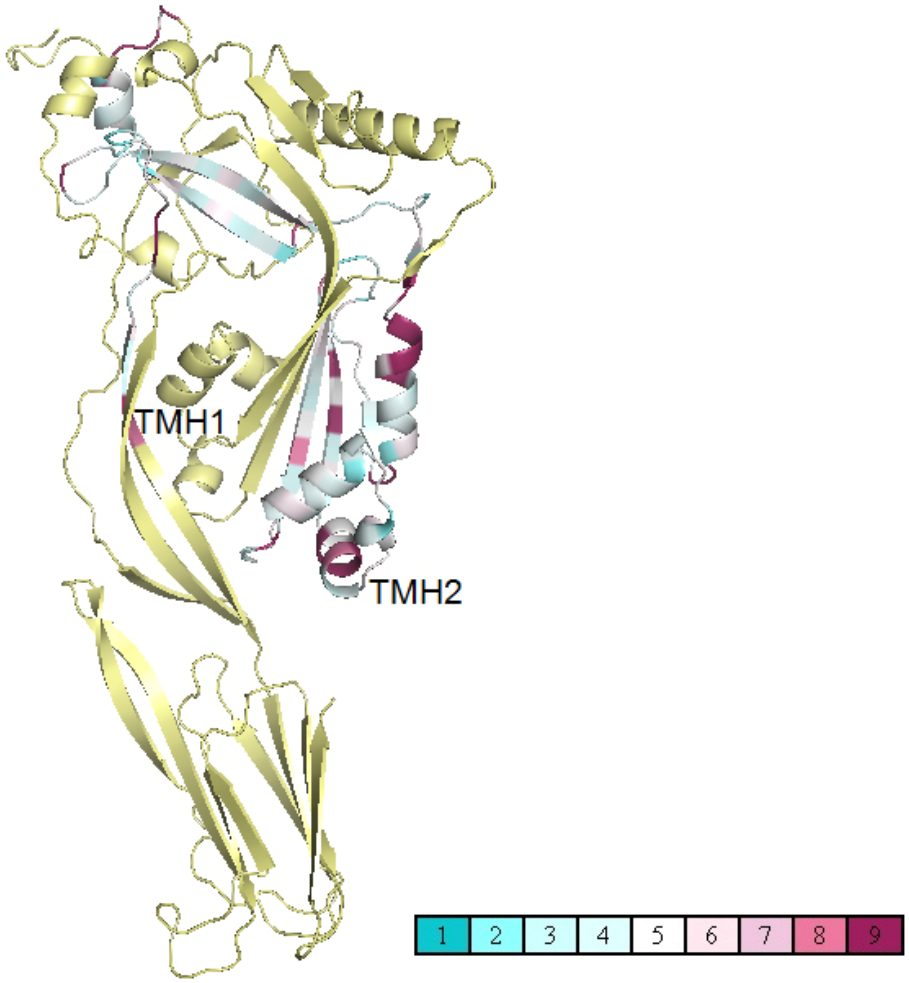


Figure 39. ConSurf Coloring of Omy3 & Cbo2 on 1PFO. Highly conserved residues are indicated with colors closer to 9 in the color key. Poorly conserved residues are indicated with colors closer to 1 in the color key. Light yellow indicates residues that were not aligned.

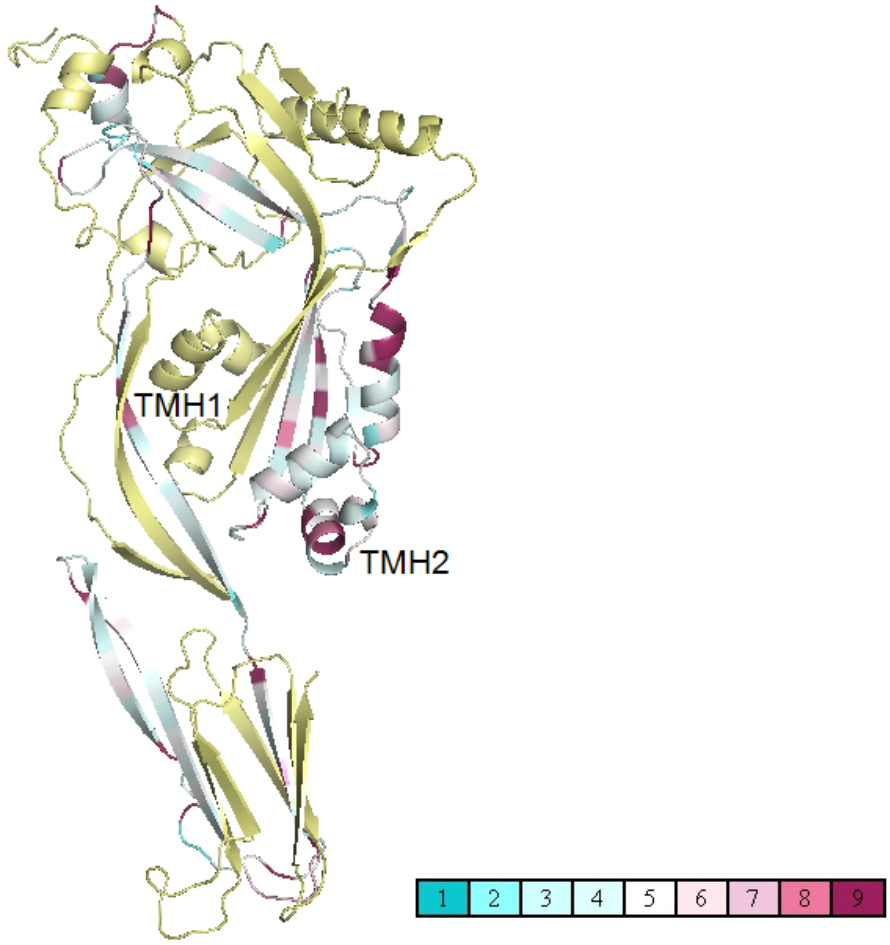


Figure 40. ConSurf Coloring of Omy3 & Cno1 on 1PFO. Highly conserved residues are indicated with colors closer to 9 in the color key. Poorly conserved residues are indicated with colors closer to 1 in the color key. Light yellow indicates residues that were not aligned.

References

Anderluh G. and Lakey J.H. (2008). Disparate proteins use similar architectures to damage membranes. Trends in Biochemical Sciences **33**: 482-490.

Anderson D. S., Blaustein R. O. (2008). Preventing voltage-dependent gating of anthrax toxin channels using engineered disulfides. The Journal of General Physiology **132**: 351–360.

Araki T., Kusakabe M., Nishida E. (2011). A transmembrane protein EIG121L is required for epidermal differentiation during early embryonic development. Journal of Biological Chemistry. **286**(8):6760-8.

Axelsen K. B., Palmgren M. G. (1998). Evolution of substrate specificities in the P-type ATPase superfamily. Journal of Molecular Evolution. **46**(1):84-101.

Baumgartner S., Hofmann K., Chiquet-Ehrismann R., Bucher P. (1998). The discoidin domain family revisited: new members from prokaryotes and a homology-based fold prediction. Protein Science. **7**:1626–1631.

Bernheimer A. W., Avigad L. S. (1979). A cytolytic protein from the edible mushroom, Pleurotus ostreatus. Biochimica et Biophysica Acta. **585**:451–461.

Busch W, Saier M. H. Jr. (2002). The transporter classification (TC) system, 2002. Critical Reviews in Biochemistry and Molecular Biology **37**: 287–337.

Casalena G., Daehn I., Bottinger E. (2012). Transforming Growth Factor-β, Bioenergetics, and Mitochondria in Renal Disease. Seminars in Nephrology. **32**(3):295-303.

Chen J. S., Reddy V., Chen J. H., Shlykov M. A., Zheng W. H., Cho J., Yen M. R. and Saier M. H. Jr. (2011). Phylogenetic Characterization of Transport Protein Superfamilies: Superiority of SFT programs over those based on multiple-alignments. Journal of Molecular Microbiolology and Biotechnology. IN PRESS, 2011.

Czajkowsky D. M., Hotze E. M., Shao Z., Tweten R. K. (2004). Vertical collapse of a cytolysin prepore moves its transmembrane beta-hairpins to the membrane. The EMBO Journal **23**: 3206–3215.

Doolittle, R. F. (1981). "Protein evolution." Science **214** (4525): 1123-1124.

Elgavish S., Shaanan B. (1997). Lectin-carbohydrate interactions: different folds, common recognition principles. Trends in Biochemical Sciences. 22(12):462-7.

Ermekova K. S., Zambrano N., Linn H., Minopoli G., Gertler F., Russo T., Sudol M. The WW domain of neural protein FE65 interacts with proline-rich motifs in Mena, the mammalian homolog of Drosophila enabled. Journal of Biological Chemistry. **272**(52):32869-77.

Estévez-Calvar N., Romero A., Figueras A., Novoa B. (2011). Involvement of pore-forming molecules in immune defense and development of the Mediterranean mussel (Mytilus galloprovincialis). Developmental and Comparative Immunology. **35**(10):1017-31.

Glaser F., Pupko T., Paz I., Bell R.E., Bechor D., Martz E., and Ben-Tal N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics. **19**:163-164.

Haag E. S., Sly B. J., Andrews M. E., Raff R. A. (1999). Apextrin, a novel extracellular protein associated with larval ectoderm evolution in Heliocidaris erythrogramma. Developmental Biology. **211**(1):77-87.

Haendler B., Hofer E. (July). Characterization of the human cyclophilin gene and of related processed pseudogenes. European Journal of Biochemistry. **190**(3): 477–82.

Holm D., Fink D. R., Steffensen M. A., Schlosser A., Nielsen O., Moeller J. B., Holmskov U. (2012). Characterization of a novel human scavenger receptor cysteine-rich molecule SCART1 expressed by lymphocytes. Immunobiology. [Epub ahead of print]

Holmskov U., Malhotra R., Sim R. B., Jensenius J. C. (1994). Collectins: collagenous C-type lectins of the innate immune defense system. Immunology Today. **15**(2):67-74.

Holzman T. F., Egan D. A., Edalji R., Simmer R. L., Helfrich R., Taylor A., Burres N. S. (1991). Preliminary characterization of a cloned neutral isoelectric form of the human peptidyl prolyl isomerase cyclophilin. Journal of Biological Chemistry. **266**(4): 2474–9.

Honda S., Kashiwagi M., Miyamoto K., Takei Y., Hirose S. (2000). Multiplicity, structures, and endocrine and exocrine natures of eel fucose-binding lectins. Journal of Biological Chemistry. **275**(42):33151-7.

Kawano H., Nakatani T., Mori T., Ueno S., Fukaya M., Abe A., Kobayashi M., Toda F., Watanabe M., Matsuoka I. (2004). Identification and characterization of novel developmentally regulated neural-specific proteins, BRINP family. Molecular Brain Research. **125**(1-2):60-75.

Landau M., Mayrose I., Rosenberg Y., Glaser F., Martz E., Pupko T., and Ben-Tal N. (2005). ConSurf 2005: the projection of evolutionary conservation

scores of residues on protein structures. Nucleic Acids Research. **33**:W299-W302.

Law, R.H., Lukoyanova N., Voskoboinik I., Caradoc-Davies T. T., Baran K., Dunstone M. A., D'Angelo M. E., Orlova E. V., Coulibaly F., Verschoor S., Browne K. A., Ciccone A., Kuiper M. J., Bird P. I., Trapani J. A., Saibil H. R., and Whisstock J. C. (2010). The structural basis for membrane binding and pore formation by lymphocyte perforin. Nature **468**: 447-451.

Mangé A., Béranger F., Peoc'h K., Onoder T., Frobert Y., Lehmann S. (2004). Alpha- and beta- cleavages of the amino-terminus of the cellular prion protein. Biology of the Cell **96**: 125–32.

Mayrose I., Graur D., Ben-Tal N. and Pupko T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Molecular Biology and Evolution. **21**: 1781-1791.

Molloy S.S., Anderson E.D., Jean F., Thomas G. (1999). Bi-cycling the furin pathway: from TGN localization to pathogen activation and embryogenesis. Trends in Cell Biology. **9**(1):28-35.

Ponting C. P. (2000). Novel repeats in ryanodine and IP3 receptors and protein O-mannosyltransferases.Trends in Biochemical Sciences. **25**:48–50.

Rosado C. J., Buckle A. M., Law R. H. P., Butcher R.E., Kan W. T., Bird C.H., Ung K., Browne K. A., Baran K., Bashtannyk-Puhalovich T. A., Faux N. G., Wong W., Porter C. J., Pike R. N., Ellisdon A. M., Pearce M. C., Bottomley S. P., Emsley J., Smith A. I., Rossjohn J., Hartland E. L., Voskoboinik I., Trapani J. A., Bird P. I., Dunstone M. A., and Whisstock J. C. (2007). A Common Fold Mediates Vertebrate Defense and Bacterial Attack. Science. **317**(5844):1548-1551.

Rosado C.J., Kondos S., Bull T. E., Kuiper M. J., Law R. H., Buckle A. M., Voskoboinik I., Bird P. I., Trapani J. A., Whisstock J.C., and Dunstone M. A. (2008). The MACPF/CDC family of pore-forming toxins. Cell Microbiology **10**: 1765-1774.

Rossi V., Wang Y., Esser A. F. (2010). Topology of the membrane-bound form of complement protein C9 probed by glycosylation mapping, anti-peptide antibody binding, and disulfide modification. Molecular Immunology. **47**(7-8):1553-60.

Rossjohn, J., Feil, S.C., McKinstry, W.J., Tweten, R.K., Parker, M.W. (1997). Structure of a cholesterol-binding, thiol-activated cytolysin and a model of its membrane form. Cell. **89:** 685-692.

Saier M. H. Jr., Yen M. R., Noto K., Tamang D. G., Elkan C. (2009). The transporter classification database: recent advances. Nucleic Acids Research **37**, 274-278.

Saier M. H. Jr., Tran C. V., Barabote R. D. (2006). TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Research **34**, 181–186.

Saier M. H. Jr., Yen M. R., Noto K., Tamang D. G., Elkan C. (2009), The Transporter Classification Database: recent advances. Nucleic Acids Research **37**: 274-278.

Sakurai N., Kaneko J., Kamio Y., and Tomita T. (2004). Cloning, expression, and pore-forming properties of mature and precursor forms of pleurotolysin, a sphingomyelin-specific two-component cytolysin from the edible mushroom Pleurotus ostreatus. Biochimica et Biophysica Acta **1679**: 65-73.

Shibata T., Kudou M., Hoshi Y., Kudo A., Nanashima N., and Miyairi K. (2010). Isolation and characterization of a novel two-component hemolysin, erylysin A and B, from an edible mushroom, Pleurotus eryngii. Toxicon **56**: 1436-1442.

Shogomori H, Kobayashi T. (2008). Lysenin: a sphingomyelin specific pore-forming toxin. Biochimica et Biophysica Acta **1780**: 612–618.

Slade, D.J., Lovelace, L.L., Chruszcz, M., Minor, W., Lebioda, L., Sodetz, J.M. (2008). Crystal Structure of the MACPF Domain of Human Complement Protein C8alpha in Complex with the C8gama Subunit. Journal of Molecular Biology. **379**: 331-342.

Tomita T., Noguchi K., Mimuro H., Ukaji F., Ito K., Sugawara-Tomita N., Hashimoto Y. (2004). Pleurotolysin, a novel sphingomyelin-specific two-component cytolysin from the edible mushroom *Pleurotus ostreatus*, assembles into a transmembrane pore complex. The Journal of Biological Chemistry. **279**:26975–26982.

Wang X., Watt P. M., Louis E. J., Borts R. H., Hickson I. D. (1996). Pat1: a topoisomerase II-associated protein required for faithful chromosome transmission in Saccharomyces cerevisiae. Nucleic Acids Research. **23**:4791–4797.

Wright K. O., Messing E. M., Reeder J. E. (2004). DBCCR1 mediates death in cultured bladder tumor cells. Oncogene. **23**(1):82-90.

Xu Q., Abdubek P., Astakhova T., Axelrod H. L., Bakolitsa C., et al. (2010). Structure of a membrane-attack complex/perforin (MACPF) family protein from the human gut symbiont Bacteroides thetaiotaomicron. Acta

Crystallographica Section F, Structural Biology and Crystallization Communications. **66**:1297–305.

Yen M. R., Chen J. S., Marquez J. L., Sun E. I., Saier M. H. Jr. (2010). Multidrug resistance: phylogenetic characterization of superfamilies of secondary carriers that include drug exporters. Methods in Molecular Biology **637**: 47-64.

Yen M. R., Choi J. Saier M. H. Jr. (2009). Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. Journal of Molecular Microbiology and Biotechnology **17**(4):163-76. Epub 2009 Sep 18.

Zhai Y., Saier M. H. Jr. (2001). A Web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. Journal of Molecular Microbiology and Biotechnology **3**: 285–286.

Zhai Y., Tchieu J., Saier M. H. Jr. (2002). "A web-based Tree View (TV) program for the visualization of phylogenetic trees." Journal of Molecular Microbiology and Biotechnology **4**: 69-70.