Research Article

Journal of Molecular Microbiology and Biotechnology

J Mol Microbiol Biotechnol 2017;27:252–267
DOI: 10.1159/000481286

Received: January 25, 2017
Accepted: September 5, 2017
Published online: November 17, 2017

The Membrane Attack Complex/Perforin Superfamily

Gabriel Moreno-Hagelsieb[a, b] Bennett Vitug[b] Arturo Medrano-Soto[b]
Milton H. Saier Jr.[b]

[a]Department of Biology, Wilfrid Laurier University, Waterloo, ON, Canada; [b]Department of Molecular Biology, Division of Biological Sciences, University of California at San Diego, La Jolla, CA, USA

Keywords
Cholesterol-dependent cytolysin · Membrane attack complex/perforin · Pleurotolysin · Superfamily · Bioinformatics · Pore formation · Toxin

Abstract
The membrane attack complex/perforin (MACPF) superfamily consists of a diverse group of proteins involved in bacterial pathogenesis and sporulation as well as eukaryotic immunity, embryonic development, neural migration and fruiting body formation. The present work shows that the evolutionary relationships between the members of the superfamily, previously suggested by comparison of their tertiary structures, can also be supported by analyses of their primary structures. The superfamily includes the MACPF family (TC 1.C.39), the cholesterol-dependent cytolysin (CDC) family (TC 1.C.12.1 and 1.C.12.2) and the pleurotolysin pore-forming (pleurotolysin B) family (TC 1.C.97.1), as revealed by expansion of each family by comparison against a large protein database, and by the comparisons of their hidden Markov models. Clustering analyses demonstrated grouping of the CDC homologues separately from the 12 MACPF subfamilies, which also grouped separately from the pleurotolysin B family. Members of the MACPF superfamily revealed a remarkably diverse range of proteins spanning eukaryotic, bacterial, and archaeal taxonomic domains, with notable variations in protein domain architectures. Our strategy should also be helpful in putting together other highly divergent protein families.

Introduction

Over the past 2 decades, the Transporter Classification Database (TCDB) has compiled over 1,000 families of transport proteins [Busch and Saier, 2002; Saier et al., 2006, 2009, 2014, 2016] (http://www.tcdb.org/). Many of these families have been placed into over 60 superfamilies. Although similar to the Enzyme Commission system for classifying enzymes, the TC system incorporates both functional and phylogenetic information as bases for classification. Classification is thus based on structural, functional, and evolutionary characteristics [Busch and Saier, 2002; Saier et al., 2016].

The membrane attack complex/perforin (MACPF) superfamily consists of pore-forming, cytolytic proteins that are important in mammalian immunity, embryonic development, neural migration, tumor suppression, pro-

E-Mail karger@karger.com
www.karger.com/mmb

Milton H. Saier, Jr.
Department of Molecular Biology, Division of Biological Sciences
University of California at San Diego
La Jolla, CA 92093-0116 (USA)
E-Mail msaier@ucsd.edu

karyotic toxicity, and fruiting body formation and sporulation in some fungi and bacteria [Anderluh and Lakey, 2008; Estevez-Calvar et al., 2011]. Moreover, perforins in cytotoxic lymphocytes deliver cationic cargo such as granzyme proteases [Stewart et al., 2014]. As confirmed in the present study (see below), 3 families comprise 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 B pore-forming (pleurotolysin B) family (TC 1.C.97.1) [Rosado et al., 2008]. Using a common MACPF domain for pore formation, proteins associated with the membrane attack complex 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 MACPF domain are known to play a role in larval development in eukaryotes such as the sea urchin, *Heliocidaris erhthrogramma,* and the Mediterranean mussel, *Mytilus galloprovincialis* [Dheilly et al., 2011; Estevez-Calvar et al., 2011; Haag et al., 1999; Kobayashi et al., 2014]. Furthermore, the MACPF proteins DBCCR-1 and BRINP-1 (TC 1.C.39.17) are believed to function in both tumor suppression and neural development [Kawano et al., 2004; Kobayashi et al., 2014; Wright et al., 2004].

X-ray structural analyses of the MACPF domains for complement C8α and Plu-MACPF from *Photorhabdus luminescens* revealed structural similarities with the bacterial, pore-forming CDCs [Hadders et al., 2007; Kondos et al., 2010; Rosado et al., 2008]. Both families share a common mechanism of membrane insertion whereby 2 regions of the soluble proteins 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 that of CDCs in forming pores and disrupting cell membranes [Dunstone and Tweten, 2012; Law et al., 2010; Rossi et al., 2010]. However, the authors of the papers describing the 3-dimensional structures of these proteins claimed that CDCs and MACPFs show no detectable similarity at the primary sequence level.

Members of the pore-forming pleurotolysin B family have been shown to exhibit cytolytic activity through pore formation in human erythrocytes [Sakurai et al., 2004; Schlumberger et al., 2014]. Pleurotolysin proteins are 2-component hemolysins, which require the interaction of both of the 2 nonhomologous components pleurotolysin A and pleurotolysin B to exhibit strong cytolytic activity [Ota et al., 2014; Shibata et al., 2010]. Cooperative pore formation causes leakage of potassium ions from cells and subsequent colloid-osmotic hemolysis [Tomita et al., 2004], and like perforins, pleurotolysins can deliver cationic macromolecules such as granzymes [Stewart et al., 2014]. Although the longer pleurotolysin B proteins exhibit 3-dimensional folds similar to those of MACPF superfamily members, National Center for Biotechnology Information (NCBI) basic local alignment search tool (BLAST) results suggested that pleurotolysin A is a member of the aegerolysin superfamily [Ota et al., 2013, 2014; Shogomori and Kobayashi, 2008].

This study seeks to expand the MACPF family and to demonstrate sequence similarity between the active pore-forming regions of the MACPF, CDC, and pleurotolysin B families. Several developments have made it possible to identify increasingly distant homologues using sequence similarity as the primary means. We first expand the set of representatives of the major phylogenetic clusters in each family. Second, we identify protein sequences that could reveal links between these families. Third, the availability of increasingly sensitive software allows us to compare more distant homologues of each family. In summary, application of the superfamily transitivity principle allows demonstration of homology between each family using "missing link" homologues. We also build and compare hidden Markov models of each protein family to further establish the reliability of our results.

Results

A summary of the overall results showing the relationships revealed by these analyses is presented in Figure 1.

Expansion of Protein Families

Protein families obtained from TCDB were screened against NCBI's nonredundant (NR) protein database using position-specific iterated BLAST (PSIBLAST). Redundant and close sequences were eliminated from the results using CD-HIT (Cluster Database at High Identity with Tolerance) with a percent identity threshold of 90%. This procedure increased the number of MACPF sequences from 68 to 11,650, pleurotolysin B sequences from 9 to 370, and CDC protein sequences from 16 to 1,680. The phylum and taxonomic domain for the organisms from which each protein was derived were also retrieved from NCBI (online suppl. Tables 1–3; see www.karger.com/doi/10.1159/000481286 for all online suppl. material).

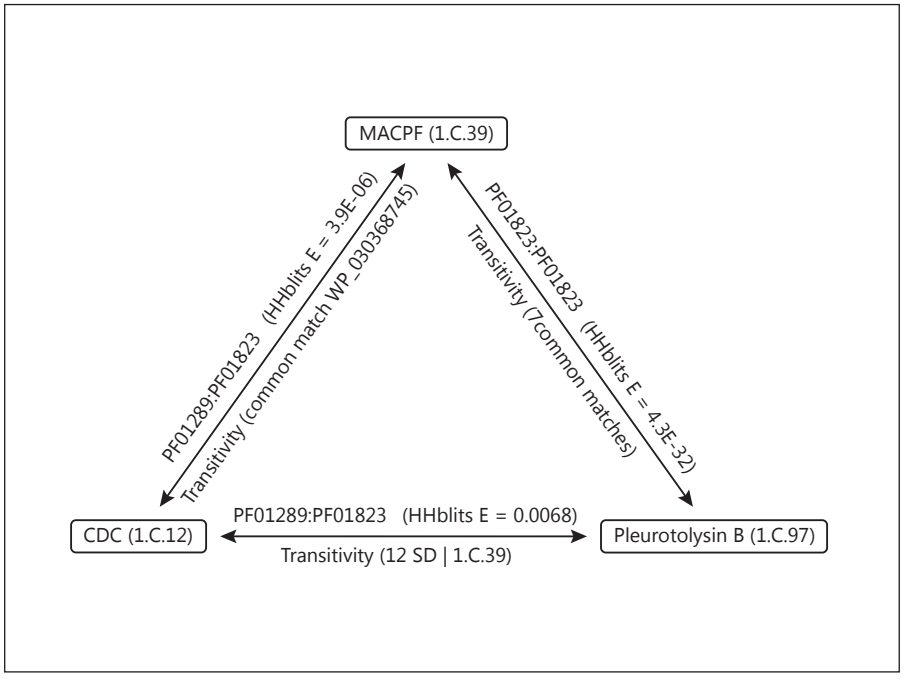

Fig. 1. Summary of sequence-based evidence of homology between MACPF, pleurotolysin B, and CDC protein families. MACPF proteins (1.C.39) were found to be homologous to CDC proteins by finding a homologous protein common to both (WP_030368745), and by the similarity of the hidden Markov models of their respective domains (PF01289:PF01823), compared using HHblits [Remmert et al., 2012]. MACPF proteins were also found to be homologous to pleurotolysin B proteins (1.C.97) because they both match the Pfam model for the MACPF domain (PF01823), and because they both find homologous proteins in common (e.g., KDQ20396). CDC proteins and pleurotolysin B proteins were not found to be homologous using the Protocol2 analysis, but they can be inferred to be homologous from the comparison of the HMM models of their respective domains as compared with HHblits, and because they are both homologous to the MACPF (1.C.39) proteins.

Transitivity Tests

The full table of results from comparing the members of each protein family to NCBI's NR database showed that the CDC family found a total of 70 proteins also found by the members of the MACPF family. Similarly, the MACPF family found 34 proteins in common with those found by members of the pleurotolysin B family (Fig. 1, 2a).

The complete sets of proteins obtained from screening each protein family against NCBI's NR protein database were compared against each other using Protocol2.py from the BioV software suite [Reddy and Saier, 2012]. The CDC set had a top score of 154 SD over randomized sequences when compared against the MACPF family set. The aligned sequences were segments of the same protein retrieved by members of both families (WP_030368745) (Fig. 2b). The top score between the MACPF and pleurotolysin B sets was 113 SD. As with the MACPF and CDC families, the aligned sequences were segments of the same protein matched by members of both families (KDQ20396) (Fig. 2c).

The top score between the CDC and pleurotolysin B sets was only 12 SD, which is within those obtained with the unrelated families examined (negative controls). However, since the MACPF family shows homology to both the CDC and the pleurotolysin B families, we can infer, by the transitivity principle, that the CDC and pleurotolysin B families are also homologous. Homology between the MACPF and CDC transmembrane domains is therefore established based on the analyses of their sequences alone.

MACPF/CDC Protein Domains

When compared against Pfam protein domains [Finn et al., 2014], most proteins within each of the protein families originally present in TCDB-matched domains consistent with their respective group membership (Fig. 1). Most proteins in the MACPF family matched the MACPF domain (PF01823) as did most proteins in the pleurotolysin B family. Most proteins from the CDC family matched the CDC domain (PF01289). Other matches to Pfam domains were specific to fewer members of these protein families (online suppl. Tables 1–3).

Extracting the segments that matched these 2 domains allowed us to compare the MACPF and CDC domains against each other using HHblits [Remmert et al., 2012]. The MACPF domain, as represented by members of the expanded MACPF family, had an alignment E value of 3.9E–06 against the CDC domain, as represented by members of the expanded CDC family (Fig. 3a), while the MACPF domain, as represented in the expanded pleurotolysin B family, had an alignment E value of 0.0068 against the CDC domain (Fig. 3b). These results further suggest that homology between these 3 families can be established from sequence analyses alone.

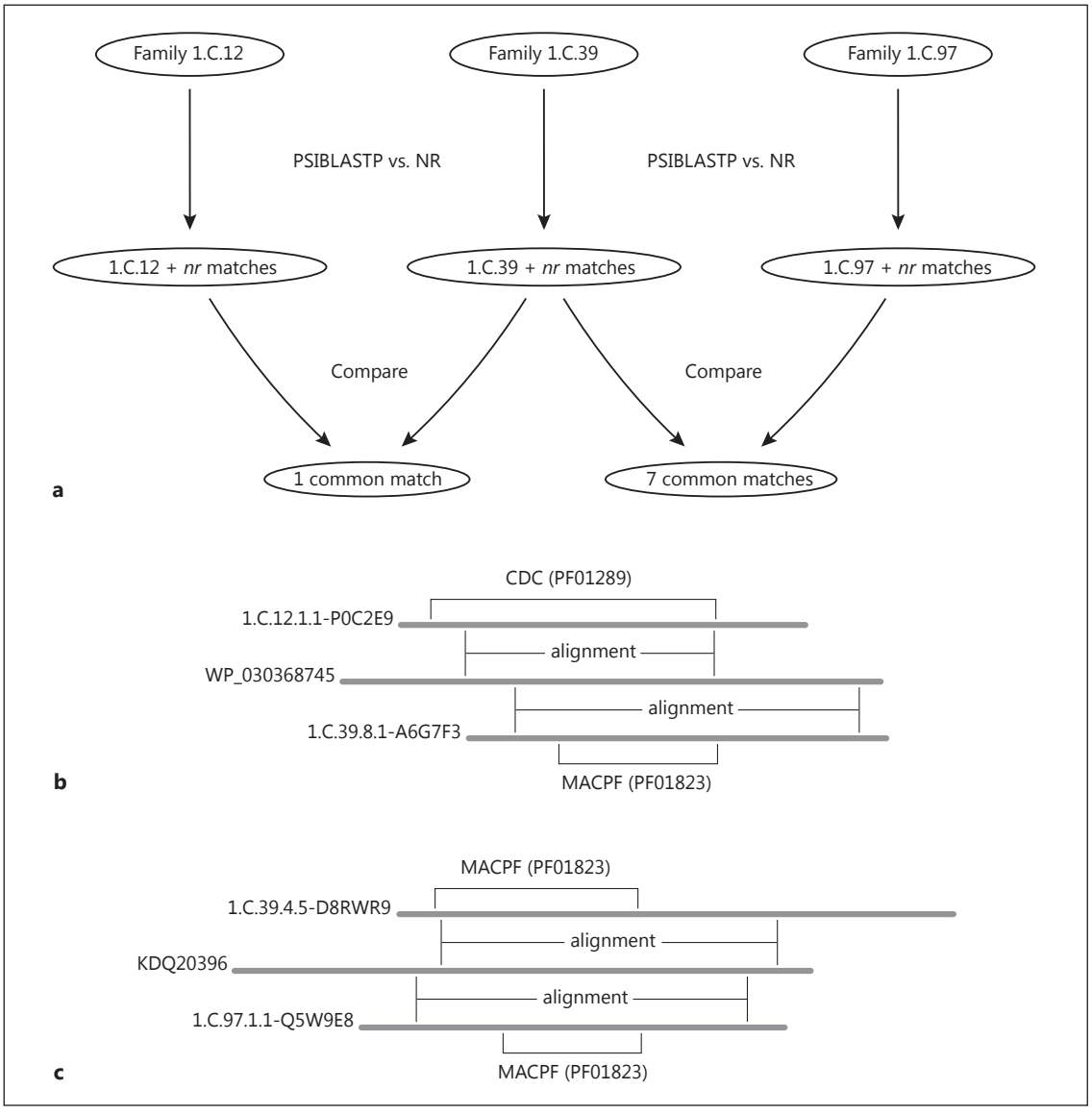

Fig. 2. Homology by transitivity. **a** Strategy and results from the transitivity principle. **b** Alignment between CDC and MACPF domains via their alignment regions with protein WP_030368745. **c** The MACPF domains of pleurotolysin B and MACPF align to each other as shown by their alignment with protein KDQ20396.

Hierarchical Clusters Based on BLAST Bit Scores

After performing hierarchical clustering using the R programming environment [R Core Team, 2016], the CDCs, MACPFs, and pleurotolysin B proteins separated well into 3 larger groups (Fig. 4), with CDCs and pleurotolysins B each falling into their respective groups, while the much larger group of MACPF proteins segregated into 12 clusters for a total of 14 clusters (12 MACPF, plus the CDC and pleurotolysin B groups). The MACPF proteins clustered consistently with their TCDB IDs. We describe the clusters below in the same order as presented in the graph (Fig. 4).

The CDC Cluster

The CDC cluster contains proteins varying in length from 369 to 588 residues. All of these proteins are found in bacteria. They did not match known domains other than the CDC domain in either the Pfam or the CDD database (online suppl. Table 1).

```
Query         PF01289-1.C.12
Match_columns 355
No_of_seqs    132 out of 1471
Neff          7.5
Searched_HMMs 1
Date          Sat May 13 06:51:46 2017
Command       hhalign -aliw 65 -i PF01289-1.C.12.hhm -t PF01823-1.C.39.hhm -o PF01289-1.C.12.PF01823-
1.C.39.hhalign

 No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 PF01823-1.C.39                  87.2 3.9E-06 3.9E-06   44.7   0.0  103  227-335    76-191 (209)

No 1
>PF01823-1.C.39
Probab=87.16  E-value=3.9e-06  Score=44.71  Aligned_cols=103  Identities=28%  Similarity=0.349
Sum_probs=60.4

Q 1M3J_A_consens  227 PPVYVSSVTYGRILYLLIESNES--------SQEVKAALNAAYKGG-----VSVSASLSAEYKSI  278 (355)
Q Consensus       227 ppvYVssV~YGR~~~~~~es~~s----------~~~~~aal~a~~~~~-----~~~~~~~~~~~~k~i  278 (355)
                    -.-||.+|.+|=++.+.+..+.+        ..++++.+.+.+.+.     .+.+.......+..
T Consensus        76 GTH~v~~~~~GG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  140 (209)
T 0000|2QP2_A_co   76 GTHYVTSVTLGGRISYIYTVSSSEFESSEEQKIEIKASASASFGGVSISISGGSSSSSKSSSSSS  140 (209)
Confidence            34699999999777766655433        222445565555233     22222222333344


Q 1M3J_A_consens  279 LSNSSIKVYVIGGSSQGASQVISGNLDELKDFISEGATFSASNPGVPISYTLRYLKD  335 (355)
Q Consensus       279 l~~s~i~v~~~GG~~~~~~~v~~~~~~~l~~~i~~~~~~s~~~g~PIsY~~~~l~d  335 (355)
                    ..+.++++.++||..... .......+.+|++.-   . .+ ..||.+++.-|.+
T Consensus       141 ~~~~~~~~~~GG~~~~~~~~~~~~~~~W~~Sv---~~~-~~~-p~~i~~~l~pI~~  191 (209)
T 0000|2QP2_A_co  141 SQNSSVSVTVIGGDPSSANI-GSDDSEDFSAWAQSV---N-DN-PMPIDIELIPISE  191 (209)
Confidence            577789999999876443 112345566776542   1 13 3888888766644
```

a

```
Query         PF01289-1.C.12
Match_columns 355
No_of_seqs    132 out of 1471
Neff          7.5
Searched_HMMs 1
Date          Sat May 13 06:51:47 2017
Command       hhalign -aliw 65 -i PF01289-1.C.12.hhm -t PF01823-1.C.97.hhm -o PF01289-1.C.12.PF01823-
1.C.97.hhalign

 No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 PF01823-1.C.97                   5.7  0.0068  0.0068   24.6   0.0   70  224-293    72-154 (206)

No 1
>PF01823-1.C.97
Probab=5.65  E-value=0.0068  Score=24.60  Aligned_cols=70  Identities=31%  Similarity=0.297
Sum_probs=39.7

Q 1M3J_A_consens  224 NSNPPVYVSSVTYGRILYLLIESNES--------SQEVKAALNAAYKGG-VSVSASLSAEYK---  276 (355)
Q Consensus       224 ~~~ppvYVssV~YGR~~~~~~es~~s----------~~~~~aal~a~~~~~-~~~~~~~~~~~k---  276 (355)
                    ..+--.+...|.=|=+++-.=+.+.+        ...+++++.+++.++ ++++.+......
T Consensus        72 ~~yG~~f~t~v~LGGrL~st~~~~~~~~~~~~~~~~kaa~~~s~~~~~~~s~~~~s~~~~~~  136 (206)
T 4OEJ_A_consens   72 NKYGHFFPTRVTLGGRLHSTEKSTSSSSSSTEEKKESFKAAASASFASASASASYESSSESSSSN  136 (206)
Confidence            33445777888887777766443322        334777777777444 222222222222


Q 1M3J_A_consens  277 -SILSNSSIKVYVIGGSS  293 (355)
Q Consensus       277 -~~il~~s~i~v~~~GG~~  293 (355)
                     ..-.+.++..-+.||+.
T Consensus       137 s~~~~~~s~~~~a~GGdt  154 (206)
T 4OEJ_A_consens  137 SSSSSSESITWEAIGGDT  154 (206)
Confidence             22344578888999998
```

b

Fig. 3. Alignment of hidden Markov models (hhalign). These alignments show the similarity of hidden Markov models (HMMs) built with protein segments that align to the CDC and MACPF domains from members of each family. We present only the alignments between the CDC HMM and each of the MACPF HMMs (1.C.39 and 1.C.97). As shown, the alignments have similarities beyond what would be expected by chance (note the E values). **a** 1.C.12 (CDC) versus 1.C.39 (MACPF). **b** 1.C.12 (CDC) versus 1.C.97 (pleurotolysin 2-MACPF).

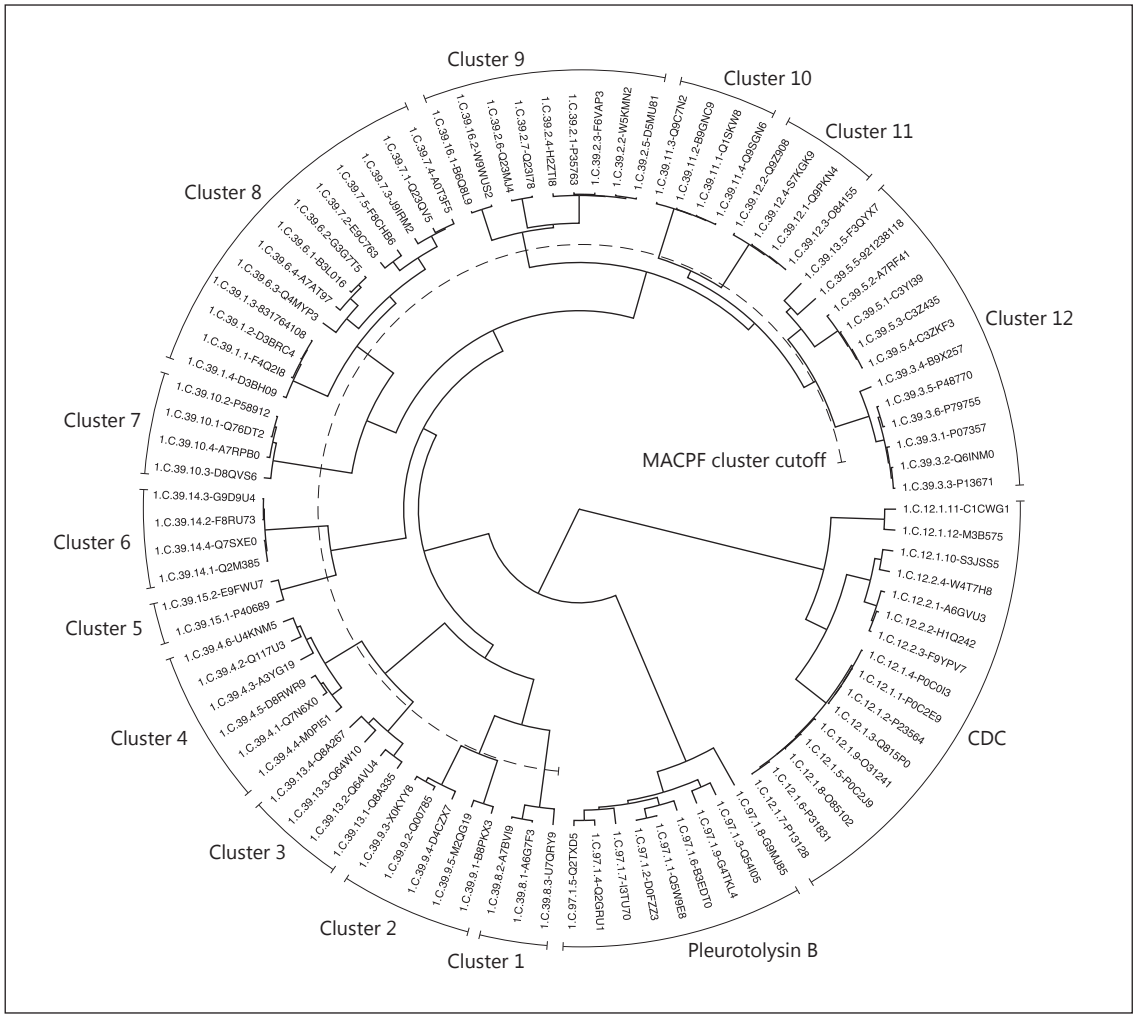

Fig. 4. Similarity clusters of the MACPF, pleurotolysin B, and CDC families. The protein TC number, followed by the UniProt accession number (or occasionally the NCBI number), is provided for each protein included in the tree. Clusters 1–12 are from the MACPF family. The cluster (subfamily or family) designation is provided outside the outer circle.

The Pleurotolysin B Cluster

This cluster contains proteins varying in length from 340 to 1,165 residues. Two of the 7 proteins in TCDB are found in bacteria, while the remaining 5 are from eukaryotes. These proteins did not match known domains other than the MACPF domain in either the Pfam or the CDD database (online suppl. Table 2).

MACPF Cluster 1

The first MACPF cluster contained all 3 proteins in subfamily TC 1.C.39.8 (TC 1.C.39.8.1–3). These proteins are from Proteobacteria. Besides the MACPF domain, they contained from 1–3 copies of the PF00045 domain, described in Pfam as hemopexin. Protein 1.C.39.8.3-U7QRY9 also matched domain PF07472, described as fucose-binding lectin II.

MACPF Cluster 2

This cluster contained all of the proteins in subfamily TC 1.C.39.9 (TC 1.C.39.9.1–5). These proteins did not match Pfam domains other than the MACPF domain (PF01823). All of these proteins are from Fungi.

MACPF Cluster 3

The third cluster contained the proteins in subfamily TC 1.C.39.13 (TC 1.C.39.13.1–4), except for TC 1.C.39.13.5-

F3QYX7, which grouped with other subfamilies in cluster 12 (see below). All of the proteins in cluster 3 are from bacteria, specifically from organisms of the genus *Bacteroides*. No protein in this cluster matched domains other than the MACPF domain (PF01823).

MACPF Cluster 4

The fourth cluster contains most of the proteins in the TC 1.C.39.4 subfamily. Of the proteins in this cluster, TC 1.C.39.4.1–3 and 6 derive from various bacteria, but TC 1.C.39.4.4-M0PI51 is from an archaeon, and TC 1.C.39.4.5-D8RWR9 is from a eukaryote (*Selaginella moellendorffii*). Only the last of these proteins contained more than the MACPF domain (PF01823). It contained 8 copies of the PF09479 domain, described as a *Listeria-Bacteroides* repeat domain involved in host cell invasion [Ebbes et al., 2011].

MACPF Cluster 5

The fifth cluster contained the 2 proteins in subfamily TC 1.C.39.15 (TC 1.C.39.15.1–2). These proteins are eukaryotic (Arthropoda) and match the MACPF domain model from CDD (CDD ID: 214671).

MACPF Cluster 6

The sixth cluster contained all 4 proteins in subfamily TC 1.C.39.14 (TC 1.C.39.14.1–4). All proteins are from eukaryotes, 1 found in humans, 2 in mollusks, and 1 in the zebrafish (*Danio rerio*). These proteins only matched the MACPF domain.

MACPF Cluster 7

The seventh cluster contained all 4 proteins in subfamily TC 1.C.39.10 (TC 1.C.39.10.1–4). All of these proteins are eukaryotic. Three of them are from Cnidaria, and 1 (TC 1.C.39.10.3) is from a plant (*Selaginella*). These proteins only matched the MACPF domain.

MACPF Cluster 8

The eighth cluster can be considered to be a cluster of clusters (Fig. 3). This cluster contained all members of TCDB's subfamilies (TC 1.C.39.1–4), 1.C.39.6 (TC 1.C.39.6.1–4), and 1.C.39.7 (1.C.39.7.1–5). These proteins are from eukaryotes, except for 1.C.39.7.5, which is from a bacterium (*Myxococcus*). Proteins belonging to subfamily TC 1.C.39.1 are all from Dictyosteliida (slime molds), proteins in subfamily TC 1.C.39.6 are from protists of the phylum Apicomplexa, and proteins in TC 1.C.39.7 – except for 1.C.39.7.5-F8CHB6 – are from ciliates (phylum Ciliophora) and Opisthokonta (a sister taxon to Fungi). These proteins only matched the MACPF domain.

MACPF Cluster 9

This cluster is composed of 2 subclusters, a subcluster containing all of the proteins in subfamily 1.C.39.2 (1.C.39.2.1–7) and a subcluster containing both proteins from family 1.C.39.16 (1.C.39.16.1 and 2). All proteins in family 1.C.39.2 are from eukaryotes, the first 5 from vertebrates, and the last 2 from *Tetrahymena*. All of these proteins matched the MACPF domain, with 4 of them also matching domain PF00168, described as a C2 domain, which is a domain that targets proteins to cell membranes [Zhang and Aravind, 2010]. The proteins in family 1.C.39.16 are from Fungi and only matched MACPF domain models.

MACPF Cluster 10

This cluster contains all members of subfamily 1.C.39.11 (1.C.39.11.1–4). These proteins are from plants and matched only MACPF domain models.

MACPF Cluster 11

This cluster contained all 4 proteins in subfamily TC 1.C.39.12 (TC 1.C.39.12.1–4). These proteins are from bacteria (*Chlamydia*). They only matched the MACPF domain.

MACPF Cluster 12

The last MACPF cluster contains all members of TCDB subfamily 1.C.39.5 (1.C.39.5.1–5), as well as 1.C.39.13.5. The first 4 proteins in family 1.C.39.5 are eukaryotic although 1.C.39.5.5 is bacterial (*Pseudomonas*). All of these proteins match the MACPF domain. One of them (1.C.39.5.4) also matches domains PF00754 and PF00530. PF00754 (discoidin domain, F5/8 type C domain) is a cell adhesion domain often found in extracellular and membrane proteins [Couto et al., 1996]. PF00530 (scavenger receptor cysteine-rich domain) is a domain also found in a variety of membrane proteins [Liu et al., 2011]. Protein 1.C.39.13.5 is a bacterial protein (*Paraprevotella*). This protein also matches only the MACPF domain.

Taxonomic Distribution of MACPF/CDC Family Proteins

The total number of homologues found by TCDB's MACPF/CDC proteins in NCBI's NR database was close to 13,700. MACPF/CDC homologues have representative proteins in all 3 domains of life (Fig. 5) (online suppl. Tables 1–3). Of all the proteins found, 70.9% are from eu-

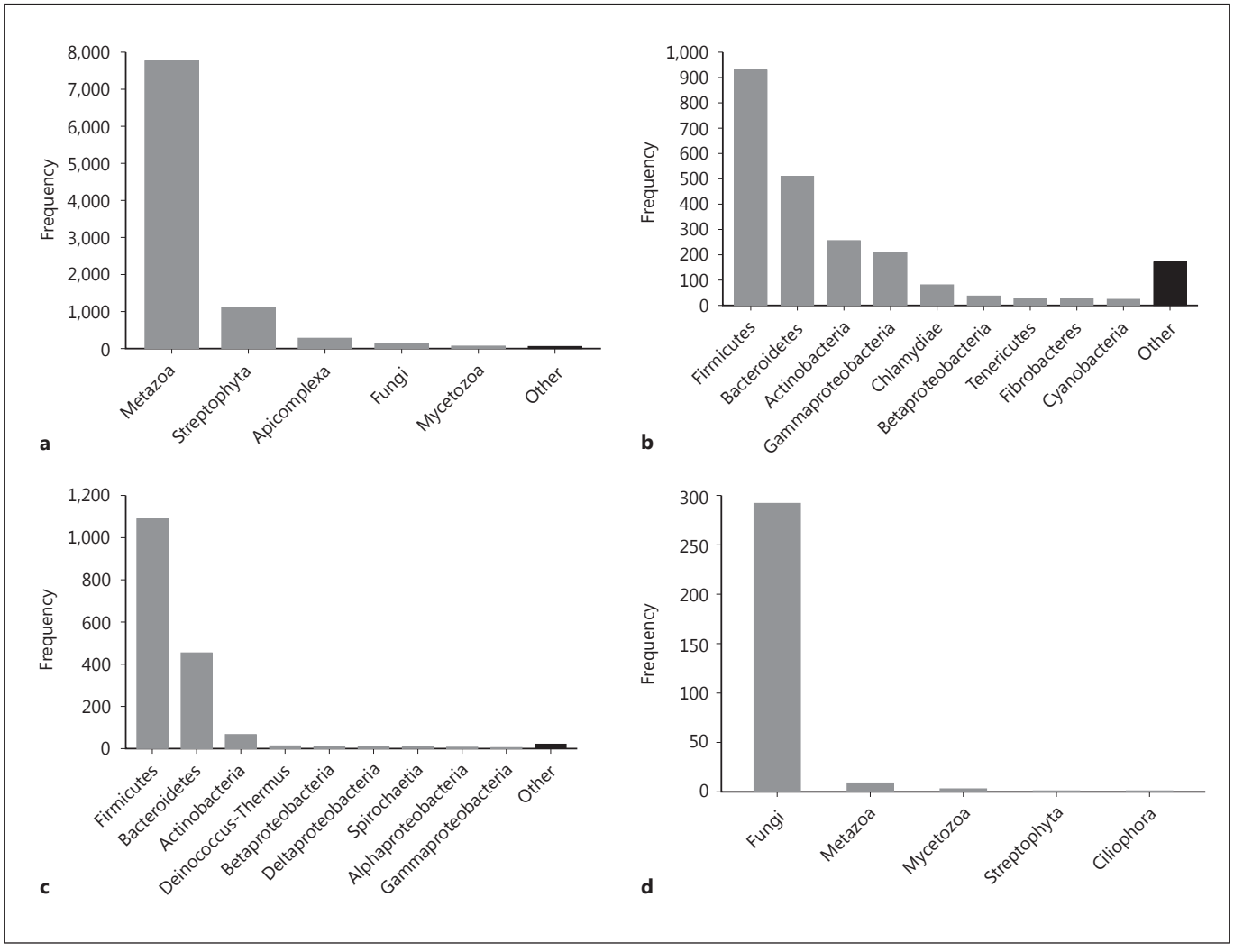

Fig. 5. Most abundant taxa containing MACPF superfamily members. The largest family within the MACPF superfamily is TC 1.C.39. Accordingly, this protein family finds homologues in all 3 domains of life, with dominance of Eukaryotes (**a**), followed by Bacteria (**b**). Family TC 1.C.12 predominates in Bacteria (**c**), while most members of TC 1.C.97 are present in Eukaryotes (**d**). Archaeal proteins are too few to display.

karyotes, 28.6% are from bacteria, and the remaining 0.5% are from archaea (35 proteins) and a few synthetic constructs (15 proteins). As described below, these proteins are widely distributed in nature.

Around 80% of the eukaryotic proteins are from the Metazoa kingdom, followed in proportion by Streptophyta (11%), and Fungi (4%). The remaining 5% of eukaryotic proteins come from organisms belonging to the Apicomplexa, Mycetozoa, Opisthokonta, Ciliophora, Isochrysidales, Dinophyceae, Schizopyrenida, Cercozoa, Bacillariophyta, Florideophyceae, Chromerida, Choanoflagellida, and Chlorophyta.

The bacterial proteins come from representatives of 25 taxonomic phyla. Approximately 51% of these proteins are found in Firmicutes, the next most abundant phyla being Bacteroidetes/Chlorobi (25%), Actinobacteria (8%), and Gammaproteobacteria (6%). The remaining 10% was composed of Chlamydiae, Betaproteobacteria, Alphaproteobacteria, Deltaproteobacteria/Epsilonproteobacteria, Tenericutes, Cyanobacteria/Melainabacteria, Fibrobacteres, Spirochaetia, Deinococcus-Thermus, Verrucomicrobia, Lentisphaerae, Candidatus Hydrogenedentes, Thermotogae, Chloroflexi, Planctomycetes, Fusobacteriia, Chrysiogenetes, Candida-

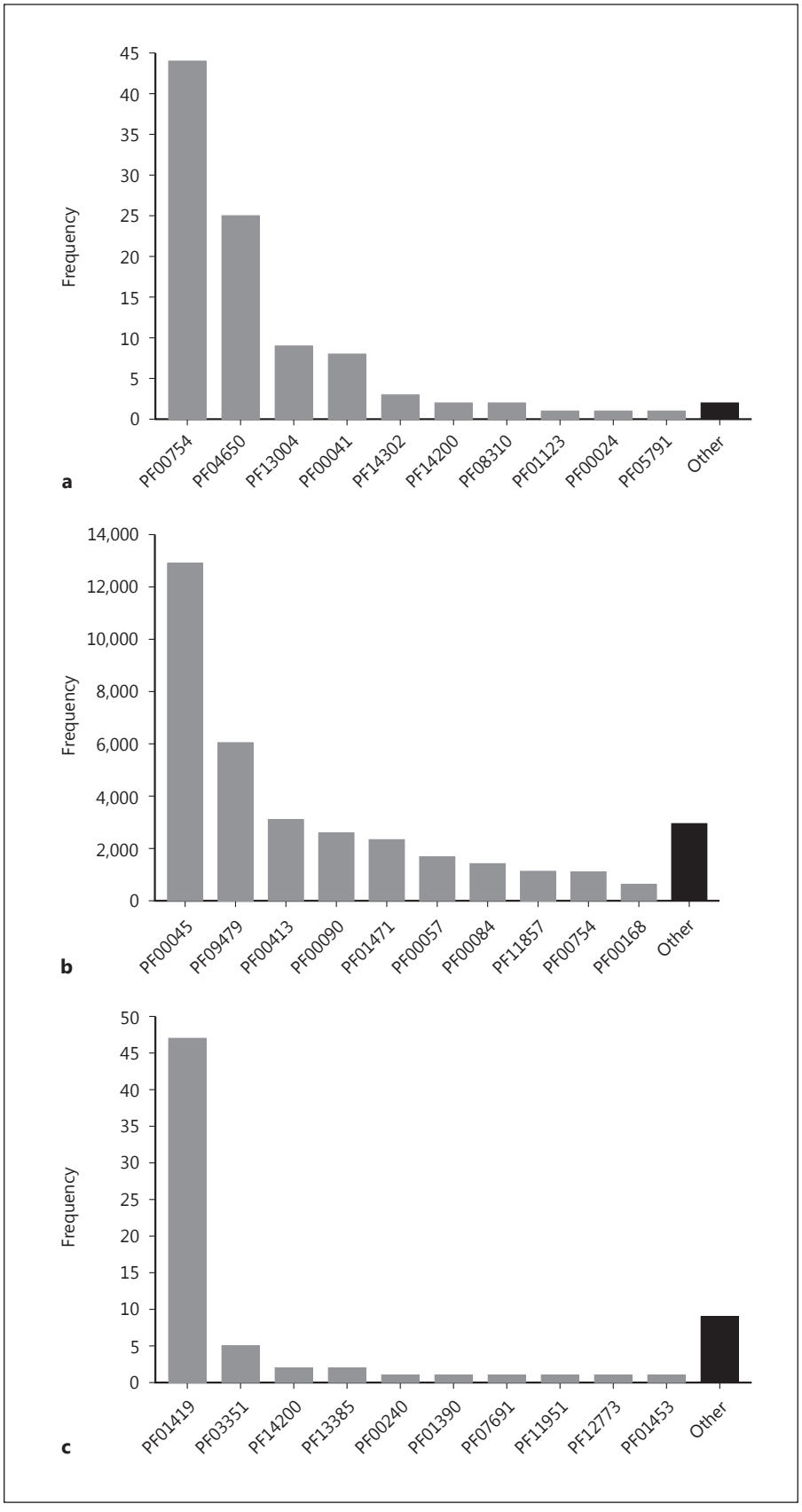

Fig. 6. Most abundant Pfam domains associated with MACPF superfamily members. The most abundant domains, other than the MACPF (PF01823) and CDC (PF01289, PF17440) domains, are shown. **a** 1.C.12. **b** 1.C.39. **c** 1.C.97.

tus Tectomicrobia, Solibacteres, Nitrospira, Candidatus Kapabacteria, and a few unclassified and environmental organisms.

Archaeal proteins come from Thermoplasmata (66%), Methanomicrobia (20%), and Halobacteria (14%).

Domains Associated with the MACPF Superfamily

Examination of the domains associated with the proteins found after running PSIBLAST comparisons of the TCDB proteins against NCBI's NR database showed several domains to be associated with members of the MACPF superfamily (Fig. 6) (online suppl. files 1–3). While most of the original proteins in TCDB matched the expected MACPF and CDC domain models, around 46% of the PSIBLAST retrieved proteins matched the MACPF domain (Pfam: PF01823, NCBI-CDD: 240433, 175998, 307781, and 214671), and around 36% matched the CDC domain (Pfam: PF01289 and PF17440, NCBI-CDD: 307453 and 307781), with 24% matching models for both domains, making a total of approximately 58%. Most of the remaining 42% proteins aligned with the regions of the proteins matching these important domains, suggesting that PSIBLAST found more proteins containing these domains than would be found by the Pfam hidden Markov models (HMMs) in combination with the CDD position-specific matrices. The number of associated domains to the PSIBLAST gathered proteins was 248 Pfam domains and 410 NCBI-CDD domains. The Pfam domains found were PF00045 (hemopexin), already described above as part of the MACPF proteins in cluster 1 (TC 1.C.39.8), and PF09479 (*Listeria-Bacteroides* repeat domain), also described above as part of the D8RWR9 protein (TC 1.C.39.4.5; MACPF cluster 4). Other domains also made sense as domains having functions associated with exported proteins. For example, the thrombospondin domain (PF00090) works in the extracellular matrix [Morris and Kyriakides, 2014]. While a detailed examination of domains associated with MACPF/CDC domains would be interesting, this is outside the scope of the present work.

Discussion

The results obtained by comparing proteins in TCDB associated with the MACPF domain (MACPF: TC 1.C.39 and pleurotolysin B: TC 1.C.97.1), and with the CDC domain (TC 1.C.12.1 and TC 1.C.12.2), showed that homology can be inferred for these protein families by sequence comparison.

Results from PSIBLAST comparisons of each TCDB family against NCBI's NR protein database showed that the MACPF and CDC proteins included shared homologues common to both families (Fig. 1). The alignments from proteins in these 2 families overlap (Fig. 2), and, most importantly, the overlapping regions correspond to the regions that match the Pfam HMM models of the respective domains of each family, namely the MACPF (PF01823) and the CDC domains (PF01289).

Similarly, when compared against NCBI's NR proteins using PSIBLAST, the MACPF and pleurotolysin B proteins found potential homologues in common (Fig. 1), which aligned with overlapping regions of those proteins (Fig. 2). The overlapping alignments also matched the Pfam HMM model of the MACPF domain (PF01823), further confirming homology between these 2 families (Fig. 2).

The superfamily principle (transitivity rule) has been used repeatedly to establish homology between distantly related members of extensive superfamilies [Chang et al., 2004; Doolittle, 1981; Lam et al., 2011; Saier, 2003]. In this study, the superfamily principle was employed by first establishing sequence similarity throughout the lengths of proteins or relevant protein domains within a single family (Fig. 1, 2). It should be noted that homology means sharing a common evolutionary ancestry and does not imply a certain degree of sequence similarity between any 2 homologues.

The results above were confirmed when the segments that match the Pfam domains in each TC expanded family were used to build new HMM models. We compared these models against each other and found them to be significantly similar using the HHsuite software.

Although X-ray and functional analyses had shown the MACPF and CDC family members to be structurally and functionally similar, sequence similarity between these families had not previously been established. The current paradigm is that one can detect homology (common ancestry) more reliably using tertiary structure than primary structure. We conducted these studies in an attempt to show that while others may not have been able to find sequence similarity, it does in fact exist and can be detected using the approaches detailed in this report.

It is well known that many proteins can adopt more than one, highly dissimilar, conformational states. Sometimes these divergent conformations are unrecognizable at the 3-dimensional level. For example, prion proteins typically adopt "native" α-states, but they can also assume cleaved β-states [Duyckaerts, 2013; Mange et al.,

Table 1. Proteins in the CDC/MACPF/pleurotolysinB families present in TCDB

TCDB	Protein ID	Description
Cholesterol-dependent cytolysin (CDC)		
1.C.12.1.1	P0C2E9	Perfringolysin O – *Clostridium perfringens*
1.C.12.1.2	P23564	Alveolysin precursor (thiol-activated cytolysin) – *Paenibacillus alvei* (*Bacillus alvei*)
1.C.12.1.3	Q815P0	Perfringolysin O OS = *Bacillus cereus* (strain ATCC 14579/DSM 31) GN = BC_5101 PE = 4 SV = 1
1.C.12.1.4	P0C0I3	Streptolysin O – *Streptococcus pyogenes* serotype M1
1.C.12.1.5	P0C2J9	Pneumolysin – *Streptococcus pneumoniae*
1.C.12.1.6	P31831	Ivanolysin precursor (thiol-activated cytolysin) – *Listeria ivanovii*
1.C.12.1.7	P13128	Listeriolysin O precursor (thiol-activated cytolysin) – *Listeria monocytogenes*
1.C.12.1.8	O85102	Hemolysin – *Streptococcus suis*
1.C.12.1.9	O31241	Pyolysin – *Arcanobacterium pyogenes*
1.C.12.1.10	S3JSS5	Uncharacterized protein OS = *Treponema medium* ATCC 700293 GN = HMPREF9195_01241 PE = 4 SV = 1
1.C.12.1.11	C1CWG1	Uncharacterized protein OS = *Deinococcus deserti* (strain VCD115/DSM 17065/LMG 22923) GN = Deide_15630 PE = 4 SV = 2
1.C.12.1.12	M3B575	Uncharacterized protein OS = *Streptomyces mobaraensis* NBRC 13819 = DSM 40847 GN = H340_07638 PE = 4 SV = 1
1.C.12.2.1	A6GVU3	Flavomodulin OS = *Flavobacterium psychrophilum* (strain JIP02/86/ATCC 49511) GN = fmo PE = 4 SV = 1
1.C.12.2.2	H1Q242	Putative uncharacterized protein OS = *Prevotella micans* F0438 GN = HMPREF9140_00956 PE = 4 SV = 1
1.C.12.2.3	F9YPV7	Tetanolysin O OS = *Capnocytophaga canimorsus* (strain 5) GN = Ccan_00800 PE = 4 SV = 1
1.C.12.2.4	W4T7H8	Uncharacterized protein OS = *Chryseobacterium indologenes* NBRC 14944 GN = CIN01S_09_02660 PE = 4 SV = 1
Membrane attack complex/perforin (MACPF)		
1.C.39.1.1	F4Q2I8	Putative uncharacterized protein OS = *Dictyostelium fasciculatum* (strain SH3) GN = DFA_07645 PE = 4 SV = 1
1.C.39.1.2	D3BRC4	Uncharacterized protein OS = *Polysphondylium pallidum* GN = PPL_10532 PE = 4 SV = 1
1.C.39.1.3	831764108	Hypothetical protein SAMD00019534_092090 [*Acytostelium subglobosum* LB1]
1.C.39.1.4	D3BH09	Uncharacterized protein OS = *Polysphondylium pallidum* GN = PPL_07811 PE = 4 SV = 1
1.C.39.2.1	P35763	Perforin 1 precursor (P1) (lymphocyte pore-forming protein) (cytolysin) – *Rattus norvegicus* (rat)
1.C.39.2.2	W5KMN2	Uncharacterized protein OS = *Astyanax mexicanus* GN = PRF1 (3 of 7) PE = 4 SV = 1
1.C.39.2.3	F6VAP3	Uncharacterized protein OS = *Xenopus tropicalis* GN = prf1 PE = 4 SV = 1
1.C.39.2.4	H2ZTI8	Uncharacterized protein OS = *Latimeria chalumnae* PE = 4 SV = 1
1.C.39.2.5	D5MU81	Perforin 1 OS = *Carassius auratus langsdorfii* GN = Pfn-1 PE = 2 SV = 1
1.C.39.2.6	Q23MJ4	MAC/perforin domain-containing protein OS = *Tetrahymena thermophila* (strain SB210) GN = TTHERM_01047030 PE = 4 SV = 1
1.C.39.2.7	Q23I78	MAC/perforin domain-containing protein OS = *Tetrahymena thermophila* (strain SB210) GN = TTHERM_01380980 PE = 4 SV = 1
1.C.39.3.1	P07357	Complement component C8 α-chain OS = *Homo sapiens* GN = C8A PE = 1 SV = 2
1.C.39.3.2	Q6INM0	MGC82368 protein OS = *Xenopus laevis* GN = c7 PE = 2 SV = 1
1.C.39.3.3	P13671	Complement component C6 OS = *Homo sapiens* GN = C6 PE = 1 SV = 3
1.C.39.3.4	B9X257	Similar to terminal complement component OS = *Halocynthia roretzi* GN = TCC-like PE=2 SV=1
1.C.39.3.5	P48770	Complement component C9 precursor – *Equus caballus* (horse)
1.C.39.3.6	P79755	Complement component C9 precursor – *Fugu rubripes* (Japanese pufferfish) (*Takifugu rubripes*)
1.C.39.4.1	Q7N6X0	Unknown protein OS = *Photorhabdus luminescens* subsp. *laumondii* GN = plu1415 PE = 4 SV = 1
1.C.39.4.2	Q117U3	Membrane attack complex component/perforin/complement C9 OS = *Trichodesmium erythraeum* (strain IMS101) GN = Tery_0815 PE = 4 SV = 1
1.C.39.4.3	A3YG19	Putative uncharacterized protein OS = *Marinomonas* sp. MED121 GN = MED121_03928 PE = 4 SV = 1
1.C.39.4.4	M0PI51	Putative perforin OS = *Halorubrum kocurii* JCM 14978 GN = C468_01225 PE = 4 SV = 1
1.C.39.4.5	D8RWR9	Putative uncharacterized protein OS = *Selaginella moellendorffii* GN = SELMODRAFT_415635 PE = 4 SV = 1
1.C.39.4.6	U4KNM5	Uncharacterized protein OS = *Acholeplasma palmae* GN = BN85402230 PE = 4 SV = 1
1.C.39.5.1	C3YI39	Putative uncharacterized protein OS = *Branchiostoma floridae* GN = BRAFLDRAFT_71536 PE = 4 SV = 1
1.C.39.5.2	A7RF41	Predicted protein OS = *Nematostella vectensis* GN = v1g196263 PE = 4 SV = 1
1.C.39.5.3	C3Z435	Putative uncharacterized protein OS = *Branchiostoma floridae* GN = BRAFLDRAFT_78467 PE = 4 SV = 1
1.C.39.5.4	C3ZKF3	Putative uncharacterized protein OS = *Branchiostoma floridae* GN = BRAFLDRAFT_69406 PE = 4 SV = 1
1.C.39.5.5	921238118	Hypothetical protein [*Pseudomonas thivervalensis*]
1.C.39.6.1	B3L016	Sporozoite protein with MAC/perforin domain OS = *Plasmodium knowlesi* (strain H) GN = PKH_030600 PE = 4 SV = 1
1.C.39.6.2	G3G7T5	Perforin-like protein 1 OS = *Toxoplasma gondii* GN = PLP1 PE = 2 SV = 1
1.C.39.6.3	Q4MYP3	Putative uncharacterized protein OS = *Theileria parva* GN = TP03_0798 PE = 4 SV = 1
1.C.39.6.4	A7AT97	MACPF domain-containing protein OS = *Babesia bovis* GN = BBOV_II002020 PE = 4 SV = 1
1.C.39.7.1	Q23QV5	MACPF domain-containing protein OS = *Tetrahymena thermophila* SB210 GN = TTHERM_00249780 PE = 4 SV = 1
1.C.39.7.2	E9C763	Uncharacterized protein OS = *Capsaspora owczarzaki* (strain ATCC 30864) GN = CAOG_03644 PE = 4 SV = 1
1.C.39.7.3	J9IRM2	MACPF domain-containing protein OS = *Oxytricha trifallax* GN = OXYTRI_20266 PE = 4 SV = 1
1.C.39.7.4	A0T3F5	Apextrin OS = *Acropora millepora* PE = 2 SV = 2
1.C.39.7.5	F8CHB6	Phospholipase D endonuclease OS = *Myxococcus fulvus* (strain ATCC BAA-855/HW-1) GN = LILAB_21670 PE = 4 SV = 1
1.C.39.8.1	A6G7F3	Hemopexin OS = *Plesiocystis pacifica* SIR-1 GN = PPSIR1_00475 PE = 4 SV = 1
1.C.39.8.2	A7BVI9	Membrane attack complex component/perforin/complement C9 OS = *Beggiatoa* sp. PS GN = BGP_1327 PE = 4 SV = 1

Table 1 (continued)

TCDB	Protein ID	Description
1.C.39.8.3	U7QRY9	Photopexin a/b-like protein OS = *Photorhabdus temperata* J3 GN = O185_23085 PE = 4 SV = 1
1.C.39.9.1	B8PKX3	Predicted protein OS = *Postia placenta* (strain ATCC 44394/Madison 698-R) GN = POSPLDRAFT_98779 PE = 4 SV = 1
1.C.39.9.2	Q00785	SpoC1-C1C protein (fragment) OS = *Emericella nidulans* GN = SpoC1-C1C PE = 4 SV = 1
1.C.39.9.3	X0KYY8	Uncharacterized protein OS = *Fusarium oxysporum* f. sp. *vasinfectum* 25433 GN = FOTG_17677 PE = 4 SV = 1
1.C.39.9.4	D4CZX7	Uncharacterized protein OS = *Trichophyton verrucosum* (strain HKI 0517) GN = TRV_00371 PE = 4 SV = 1
1.C.39.9.5	M2QG19	Uncharacterized protein OS = *Ceriporiopsis subvermispora* (strain B) GN = CERSUDRAFT_115929 PE = 4 SV = 1
1.C.39.10.1	Q76DT2	Toxin AvTX-60A OS = *Actineria villosa* GN = Av60A PE = 1 SV = 1
1.C.39.10.2	P58912	Toxin PsTX-60B OS = *Phyllodiscus semoni* GN = PTX60B PE = 1 SV = 2
1.C.39.10.3	D8QVS6	Putative uncharacterized protein OS = *Selaginella moellendorffii* GN = SELMODRAFT_404579 PE = 4 SV = 1
1.C.39.10.4	A7RPB0	Predicted protein OS = *Nematostella vectensis* GN = v1g200058 PE = 4 SV = 1
1.C.39.11.1	Q1SKW8	Membrane attack complex component/perforin/complement C9 OS = *Medicago truncatula* GN = MtrDRAFT_AC140550g15v2 PE = 4 SV = 1
1.C.39.11.2	B9GNC9	Predicted protein OS = *Populus trichocarpa* GN = POPTRDRAFT_850553 PE = 4 SV = 1
1.C.39.11.3	Q9C7N2	MACPF domain-containing protein CAD1 OS = *Arabidopsis thaliana* GN = CAD1 PE = 2 SV = 1
1.C.39.11.4	Q9SGN6	MACPF domain-containing protein NSL1 OS = *Arabidopsis thaliana* GN = NSL1 PE = 2 SV = 1
1.C.39.12.1	Q9PKN4	MAC/perforin family protein OS = *Chlamydia muridarum* (strain MoPn/Nigg) GN = TC_0431 PE = 4 SV = 1
1.C.39.12.2	Q9Z908	Putative uncharacterized protein OS = *Chlamydia pneumoniae* GN = CPn_0176 PE = 4 SV = 1
1.C.39.12.3	O84155	Uncharacterized protein OS = *Chlamydia trachomatis* (strain D/UW-3/Cx) GN = CT_153 PE = 4 SV = 1
1.C.39.12.4	S7KGK9	MACPF domain protein OS = *Chlamydia psittaci* 10_1398_11 GN = CP10139811_0192 PE = 4 SV = 1
1.C.39.13.1	Q8A335	Putative uncharacterized protein OS = *Bacteroides thetaiotaomicron* (strain ATCC 29148/DSM 2079/NCTC 10582/E50/VPI-5482) GN = BT_3120 PE = 4 SV = 1
1.C.39.13.2	Q64VU4	Putative uncharacterized protein OS = *Bacteroides fragilis* (strain YCH46) GN = BF1634 PE = 4 SV = 1
1.C.39.13.3	Q64W10	Putative uncharacterized protein OS = *Bacteroides fragilis* (strain YCH46) GN = BF1566 PE = 4 SV = 1
1.C.39.13.4	Q8A267	Putative uncharacterized protein OS = *Bacteroides thetaiotaomicron* (strain ATCC 29148/DSM 2079/NCTC 10582/E50/VPI-5482) GN = BT_3439 PE = 1 SV = 1
1.C.39.13.5	F3QYX7	Uncharacterized protein OS = *Paraprevotella xylaniphila* YIT 11841 GN = HMPREF9442_03422 PE = 4 SV = 1
1.C.39.14.1	Q2M385	Macrophage-expressed gene 1 protein OS = *Homo sapiens* GN = MPEG1 PE = 2 SV = 1
1.C.39.14.2	F8RU73	Macrophage-expressed protein OS = *Crassostrea gigas* PE = 2 SV = 1
1.C.39.14.3	G9D9U4	Macrophage-expressed protein OS = Haliotis midae GN = Mpeg1 PE = 4 SV = 1
1.C.39.14.4	Q7SXE0	Macrophage-expressed 1 OS = *Danio rerio* GN = mpeg1.1 PE = 2 SV = 1
1.C.39.15.1	P40689	Torso-like protein OS = *Drosophila melanogaster* GN = tsl PE = 1 SV = 2
1.C.39.15.2	E9FWU7	Putative uncharacterized protein OS = *Daphnia pulex* GN = DAPPUDRAFT_311415 PE = 4 SV = 1
1.C.39.16.1	B6Q8L9	Putative uncharacterized protein OS = *Penicillium marneffei* (strain ATCC 18224/CBS 334.59/QM 7333) GN = PMAA_069160 PE = 4 SV = 1
1.C.39.16.2	W9WUS2	Uncharacterized protein OS = *Cladophialophora psammophila* CBS 110553 GN = A1O5_05526 PE = 4 SV = 1
Pleurotolysin B		
1.C.97.1.1	Q5W9E8	Pleurotolysin B OS = *Pleurotus ostreatus* GN = plyB PE = 2 SV = 1
1.C.97.1.2	D0FZZ3	Erylysin B OS = *Pleurotus eryngii* GN = eryB PE = 2 SV = 1
1.C.97.1.3	Q54I05	Putative uncharacterized protein OS = *Dictyostelium discoideum* GN = DDB_0188257 PE = 4 SV = 1
1.C.97.1.4	Q2GRU1	Putative uncharacterized protein OS = *Chaetomium globosum* GN = CHGG_09313 PE = 4 SV = 1
1.C.97.1.5	Q2TXD5	Predicted protein OS = *Aspergillus oryzae* (strain ATCC 42149/RIB 40) GN = AO090010000188 PE = 4 SV = 1
1.C.97.1.6	B3EDT0	Putative uncharacterized protein OS = *Chlorobium limicola* (strain DSM 245/NBRC 103803) GN = Clim_0052 PE = 4 SV = 1
1.C.97.1.7	I3TU70	Uncharacterized protein OS = *Tistrella mobilis* (strain KA081020-065) GN = TMO_b0300 PE = 4 SV = 1
1.C.97.1.8	G9MJ85	Uncharacterized protein OS = *Hypocrea virens* (strain Gv29-8/FGSC 10586) GN = TRIVIDRAFT_62223 PE = 4 SV = 1
1.C.97.1.9	G4TKL4	Uncharacterized protein OS = *Piriformospora indica* (strain DSM 11827) GN = PIIN_05796 PE = 4 SV = 1

GN, gene name; OS, organism name; TCC, terminal complement component; PE, protein existence; SV, sequence version.

2004]. Several soluble proteins with recognized catalytic and structural properties can insert into membranes, forming ion-conducting channels [Anderson and Blaustein, 2008; Aniya and Imaizumi, 2011]. Toxins are often synthesized and secreted in a soluble state, which can then insert into membranes of target organisms, forming pores that result in cytoplasmic leakage and cell death [Czajkowsky et al., 2004; Feil et al., 2014; Menestrina et al., 2001]. In all such cases, massive conformational changes occur during membrane insertion. It is therefore clear that reliance on 3-dimensional (X-ray and NMR) data alone cannot be considered the preferred approach

Table 2. Protein families used as negative controls

TCDB	Protein ID	Description
Calcineurin-like phosphatase cytolysins		
1.C.12.3.1	Q8YX86	Alr1329 protein OS = *Anabaena* sp. (strain PCC 7120) GN = alr1329 PE = 4 SV = 1
1.C.12.3.2	W9E360	Cytolysin, a secreted calcineurin-like phosphatase OS = *Mesorhizobium loti* R7A GN = MesloDRAFT_1288 PE = 4 SV = 1
1.C.12.3.3	U6B8P7	Cytolysin, a secreted calcineurin-like phosphatase OS = *Candidatus Liberibacter americanus* str. Sao Paulo GN = lam_883 PE = 4 SV = 1
Pleurotolysin A		
1.C.97.1.1	Q8X1M9	Pleurotolysin A OS = *Pleurotus ostreatus* GN = PlyA PE = 2 SV = 1
1.C.97.1.2	D0FZZ2	Erylysin A OS = *Pleurotus eryngii* GN = eryA PE = 2 SV = 1
Aegerolysins		
1.C.97.2.1	F2TLE2	Putative uncharacterized protein OS = *Ajellomyces dermatitidis* (strain ATCC 18188/CBS 674.68) GN = BDDG_07000 PE = 4 SV = 1
1.C.97.2.2	E9D534	Putative uncharacterized protein OS = *Coccidioides posadasii* (strain RMSCC 757/Silveira) GN = CPSG_05302 PE = 4 SV = 1
1.C.97.2.3	A6R8L8	Predicted protein OS = *Ajellomyces capsulatus* (strain NAm1/WU24) GN = HCAG_06659 PE = 4 SV = 1
1.C.97.2.4	W2RUH6	Uncharacterized protein OS = *Cyphellophora europaea* CBS 101466 GN = HMPREF1541_04372 PE = 4 SV = 1
1.C.97.3.1	Q1K511	Hemolysin OS = *Neurospora crassa* (strain ATCC 24698/74-OR23-1A/CBS 708.71/DSM 1257/FGSC 987) GN = NCU03475 PE = 4 SV = 1
1.C.97.3.2	P83467	Ostreolysin (fragment) OS = *Pleurotus ostreatus* PE = 1 SV = 1
1.C.97.3.3	D2QTE8	Aegerolysin OS = *Spirosoma linguale* (strain ATCC 33905/DSM 74/LMG 10896) GN = Slin_6118 PE = 4 SV = 1
1.C.97.3.4	A6UXQ8	Aegerolysin superfamily OS = *Pseudomonas aeruginosa* (strain PA7) GN = PSPA7_0197 PE = 4 SV = 1
1.C.97.3.5	S8AWM9	Uncharacterized protein OS = *Penicillium oxalicum* (strain 114-2/CGMCC 5302) GN = PDE_05655 PE = 4 SV = 1
1.C.97.3.6	C0NGB1	Predicted protein OS = *Ajellomyces capsulatus* (strain G186AR/H82/ATCC MYA-2454/RMSCC 2432) GN = HCBG_02383 PE = 4 SV = 1
1.C.97.3.7	Q06VQ2	Putative uncharacterized protein OS = *Trichoplusia ni ascovirus* 2c PE = 4 SV = 1
1.C.97.3.8	D8T5U4	Putative uncharacterized protein OS = *Selaginella moellendorffii* GN = SELMODRAFT_429335 PE = 4 SV = 1

OS, organism name; GN, gene name; PE, protein existence; SV, sequence version.

for establishing homology. A combination of high-resolution 3-dimensional data with statistical approaches using primary sequence data may be the most reliable means to establish the common origin of distantly related macromolecules. However, it should be noted that the lack of detectable structural sequence similarity cannot be used as evidence supporting the conclusion of independent origin.

Further analysis of the proteins found using PSIBLAST confirmed the presence of these proteins in organisms of the 3 domains of life. We also found that most of the extra domains associated with the MACPF/CDC domain seem to have functions related to transport proteins. These results serve to characterize the MACPF superfamily as defined here for the first time. We showed that, contrary to previous conclusions, primary sequence analyses alone are sufficient to establish homology between the 3 families of the MACPF superfamily. The superfamily appears to be much more widely distributed in many phyla of eukaryotes, bacteria and archaea, than previously thought. These results should inspire further computational analyses and experiments on a wide range of homologues with the expected discovery of novel physiological functions currently unrecognized, such as the roles of domains commonly associated with members of the MACPF superfamily. In bacteria, for example, finding genes coding for members of the superfamily in an operon, associated with genes whose products have a known function, might suggest a role for the MACPF protein. The annotations of the MACPF proteins in TCDB might also give scientists clues as to the functions of the members of the superfamily they might encounter, and thus suggest directions for experimental work.

Besides presenting evidence from primary structure for the MACPF superfamily, this work shows computational strategies by which members of highly divergent protein families might find each other without waiting for 3-dimensional structures to be solved.

Methods

Sequence Data

Sequences of proteins belonging to the membrane attack complex/perforin (MACPF: TC 1.C.39), cholesterol-dependent cytolysin (CDC: TC 1.C.12.1 and TC 1.C.12.2) and pleurotolysin B (TC 1.C.97.1) families were obtained from the TCDB (http://www.

tcdb.org/) [Saier et al., 2016] (Table 1). As negative control data sets (Table 2), we used pleurotolysin A (also under TC 1.C.97.1), aegerolysins (TC 1.C.97.2 and TC 1.C.97.3) and calcineurin-like phosphatase cytolysins (TC 1.C.12.3). The negative controls were chosen because they belong to different protein families and match domains in Pfam other than those in the CDC and MACPF families.

The proteins from these families were used as queries to compare against NCBI's NR protein sequence database using PSI-BLAST with an initial cutoff value of 1E–6 and a subsequent iteration with a cutoff value of 1E–5, requiring minimal alignment coverage of 50% of the query proteins. The aligning segments of subject proteins were retrieved from the NCBI database. Redundant and very similar sequences were filtered out using the CD-HIT program [Li and Godzik, 2006] with a cutoff value of 80%. We automated the procedure using a program, written in PERL, called famXpander.pl, which can be accessed in the TCDBtools repository (https://github.com/SaierLaboratory/TCDBtools). Multiple alignments for each family, for individual protein clusters, and for conserved domains were generated with the Clustal-Omega program [Sievers and Higgins, 2014].

Detecting Homologues with Highly Divergent Sequences

To determine homology by the superfamily principle (transitivity rule), the resulting FASTA files, each containing the proteins from the corresponding family and their matching sequences from NCBI's NR protein database, were compared to each other using "Protocol2.py" from the BioV software suite [Reddy and Saier, 2012] (https://github.com/SaierLaboratory/BioVx). This program runs the ssearch36 implementation of the Smith-Waterman algorithm [Pearson, 2000] and the EMBOSS [Olson, 2002] implementation of the Needleman and Wunsch algorithm [Needleman and Wunsch, 1970] on each pair of protein sequences and assesses the quality of the alignments against 500 randomized sequences. The significance of top scoring sequence alignments was further evaluated with 20,000 randomizations using the "gsat.py" program, also from the BioV suite [Reddy and Saier, 2012].

The proteins for all families were compared against the complete Pfam database [Finn et al., 2014] with HMMscan from the HMMer software suite [Mistry et al., 2013]. A gathering score threshold was set as recommended by the Pfam curators [Finn et al., 2014]. Domains were also found by running RPS-BLAST [Camacho et al., 2009] against the CDD database [Marchler-Bauer et al., 2015] using an E value threshold of 1E–3. Domains were compared against each other using hhmake and hhalign, both from the HHsuite software package [Remmert et al., 2012].

Homologous sequence regions were visualized using the program PyMOL with representative files from the Protein Data Bank (PDB). NCBI's BLASTP [Camacho et al., 2009] was used to generate alignments of the representative sequences in PDB with the sequences of interest obtained from the results above. For example, the CDC protein of each MACPF-CDC pair was compared with the sequence of the PDB protein model, PDB 1PFO [Rossjohn et al., 1997]. The region where a CDC protein aligned with both the MACPF protein and the 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.

Hierarchical Clustering

To obtain clusters of protein families, we compared every protein in the original collection gathered from TCDB against each other using BLASTP with a very permissive E value cutoff (a default cutoff of 10). In accordance with the SuperFamily Tree methods [Chen et al., 2011], we transformed the bit scores into distances using the formula: distance = 100/bit score. Sequences lacking BLASTP results were given a default distance of 10. We used these distances to perform hierarchical clustering as implemented in the R programming environment [R Core Team, 2016]. The clusters thus obtained should reflect phylogenetic relationships, they are easy to obtain, and they have the advantage, unlike common phylogenetic methods, that they use all of the information available when comparing each pair of sequences [Chen et al., 2011]. Traditional phylogenetic methods only use sections in common across all sequences as determined in a multiple alignment.

Acknowledgments

We thank Prof. Michelle Dunstone for valuable discussion, Harry Zhou for assistance with manuscript preparation, and the NIH (GM077402 and GM109895) for financial support.

References

Anderluh G, Lakey JH: Disparate proteins use similar architectures to damage membranes. Trends Biochem Sci 2008;33:482–490.

Anderson DS, Blaustein RO: Preventing voltage-dependent gating of anthrax toxin channels using engineered disulfides. J Gen Physiol 2008;132:351–360.

Aniya Y, Imaizumi N: Mitochondrial glutathione transferases involving a new function for membrane permeability transition pore regulation. Drug Metab Rev 2011;43:292–299.

Busch W, Saier MH Jr: The transporter classification (TC) system, 2002. Crit Rev Biochem Mol Biol 2002;37:287–337.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.

Chang AB, Lin R, Keith Studley W, Tran CV, Saier MH Jr: Phylogeny as a guide to structure and function of membrane transport proteins. Mol Membr Biol 2004;21:171–181.

Chen JS, Reddy V, Chen JH, Shlykov MA, Zheng WH, Cho J, Yen MR, Saier MH Jr: Phylogenetic characterization of transport protein superfamilies: superiority of SuperfamilyTree programs over those based on multiple alignments. J Mol Microbiol Biotechnol 2011;21:83–96.

Couto JR, Taylor MR, Godwin SG, Ceriani RL, Peterson JA: Cloning and sequence analysis of human breast epithelial antigen BA46 reveals an RGD cell adhesion sequence presented on an epidermal growth factor-like domain. DNA Cell Biol 1996;15:281–286.

Czajkowsky DM, Hotze EM, Shao Z, Tweten RK: Vertical collapse of a cytolysin prepore moves its transmembrane beta-hairpins to the membrane. EMBO J 2004;23:3206–3215.
Dheilly NM, Haynes PA, Bove U, Nair SV, Raftos DA: Comparative proteomic analysis of a sea urchin (*Heliocidaris erythrogramma*) antibacterial response revealed the involvement of apextrin and calreticulin. J Invertebr Pathol 2011;106:223–229.
Doolittle RF: Protein evolution. Science 1981;214: 1123–1124.
Dunstone MA, Tweten RK: Packing a punch: the mechanism of pore formation by cholesterol dependent cytolysins and membrane attack complex/perforin-like proteins. Curr Opin Struct Biol 2012;22:342–349.
Duyckaerts C: Neurodegenerative lesions: seeding and spreading. Rev Neurol (Paris) 2013; 169:825–833.
Ebbes M, Bleymuller WM, Cernescu M, Nolker R, Brutschy B, Niemann HH: Fold and function of the InlB B-repeat. J Biol Chem 2011;286: 15496–15506.
Estevez-Calvar N, Romero A, Figueras A, Novoa B: Involvement of pore-forming molecules in immune defense and development of the Mediterranean mussel (*Mytilus galloprovincialis*). Dev Comp Immunol 2011;35:1017–1031.
Feil SC, Ascher DB, Kuiper MJ, Tweten RK, Parker MW: Structural studies of *Streptococcus pyogenes* streptolysin O provide insights into the early steps of membrane penetration. J Mol Biol 2014;426:785–792.
Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M: Pfam: the protein families database. Nucleic Acids Res 2014;42:D222–230.
Haag ES, Sly BJ, Andrews ME, Raff RA: Apextrin, a novel extracellular protein associated with larval ectoderm evolution in *Heliocidaris erythrogramma*. Dev Biol 1999;211:77–87.
Hadders MA, Beringer DX, Gros P: Structure of C8alpha-MACPF reveals mechanism of membrane attack in complement immune defense. Science 2007;317:1552–1554.
Kawano H, Nakatani T, Mori T, Ueno S, Fukaya M, Abe A, Kobayashi M, Toda F, Watanabe M, Matsuoka I: Identification and characterization of novel developmentally regulated neural-specific proteins, BRINP family. Brain Res Mol Brain Res 2004;125:60–75.
Kobayashi M, Nakatani T, Koda T, Matsumoto K, Ozaki R, Mochida N, Takao K, Miyakawa T, Matsuoka I: Absence of BRINP1 in mice causes increase in hippocampal neurogenesis and behavioral alterations relevant to human psychiatric disorders. Mol Brain 2014;7:12.
Kondos SC, Hatfaludi T, Voskoboinik I, Trapani JA, Law RH, Whisstock JC, Dunstone MA: The structure and function of mammalian membrane-attack complex/perforin-like proteins. Tissue Antigens 2010;76:341–351.
Lam VH, Lee JH, Silverio A, Chan H, Gomolplitinant KM, Povolotsky TL, Orlova E, Sun EI, Welliver CH, Saier MH, Jr: Pathways of transport protein evolution: recent advances. Biol Chem 2011;392:5–12.
Law RH, Lukoyanova N, Voskoboinik I, Caradoc-Davies TT, Baran K, Dunstone MA, D'Angelo ME, Orlova EV, Coulibaly F, Verschoor S, Browne KA, Ciccone A, Kuiper MJ, Bird PI, Trapani JA, Saibil HR, Whisstock JC: The structural basis for membrane binding and pore formation by lymphocyte perforin. Nature 2010;468:447–451.
Li W, Godzik A: CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006; 22:1658–1659.
Liu L, Yang J, Qiu L, Wang L, Zhang H, Wang M, Vinu SS, Song L: A novel scavenger receptor-cysteine-rich (SRCR) domain containing scavenger receptor identified from mollusk mediated PAMP recognition and binding. Dev Comp Immunol 2011;35:227–239.
Mange A, Beranger F, Peoc'h K, Onodera T, Frobert Y, Lehmann S: Alpha- and beta- cleavages of the amino-terminus of the cellular prion protein. Biol Cell 2004;96:125–132.
Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH: CDD: NCBI's conserved domain database. Nucleic Acids Res 2015;43:D222–D226.
Menestrina G, Serra MD, Prevost G: Mode of action of beta-barrel pore-forming toxins of the staphylococcal alpha-hemolysin family. Toxicon 2001;39:1661–1672.
Mistry J, Finn RD, Eddy SR, Bateman A, Punta M: Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 2013;41:e121.
Morris AH, Kyriakides TR: Matricellular proteins and biomaterials. Matrix Biol 2014;37:183–191.
Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
Olson SA: EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief Bioinform 2002;3:87–91.
Ota K, Butala M, Viero G, Dalla Serra M, Sepcic K, Macek P: Fungal MACPF-like proteins and aegerolysins: bi-component pore-forming proteins? Subcell Biochem 2014;80:271–291.
Ota K, Leonardi A, Mikelj M, Skocaj M, Wohlschlager T, Kunzler M, Aebi M, Narat M, Krizaj I, Anderluh G, Sepcic K, Macek P: Membrane cholesterol and sphingomyelin, and ostreolysin A are obligatory for pore-formation by a MACPF/CDC-like pore-forming protein, pleurotolysin B. Biochimie 2013;95: 1855–1864.
Pearson WR: Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 2000;132:185–219.
R Core Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2016.
Reddy VS, Saier MH Jr: BioV Suite – a collection of programs for the study of transport protein evolution. FEBS J 2012;279:2036–2046.
Remmert M, Biegert A, Hauser A, Soding J: HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–175.
Rosado CJ, Kondos S, Bull TE, Kuiper MJ, Law RH, Buckle AM, Voskoboinik I, Bird PI, Trapani JA, Whisstock JC, Dunstone MA: The MACPF/CDC family of pore-forming toxins. Cell Microbiol 2008;10:1765–1774.
Rossi V, Wang Y, Esser AF: Topology of the membrane-bound form of complement protein C9 probed by glycosylation mapping, anti-peptide antibody binding, and disulfide modification. Mol Immunol 2010;47:1553–1560.
Rossjohn J, Feil SC, McKinstry WJ, Tweten RK, Parker MW: Structure of a cholesterol-binding, thiol-activated cytolysin and a model of its membrane form. Cell 1997;89:685–692.
Saier MH Jr: Tracing pathways of transport protein evolution. Mol Microbiol 2003;48:1145–1156.
Saier MH Jr, Reddy VS, Tamang DG, Vastermark A: The transporter classification database. Nucleic Acids Res 2014;42:D251–D258.
Saier MH Jr, Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G: The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res 2016;44:D372–D379.
Saier MH Jr, Tran CV, Barabote RD: TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res 2006;34:D181–D186.
Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C: The Transporter Classification Database: recent advances. Nucleic Acids Res 2009; 37:D274–D278.
Sakurai N, Kaneko J, Kamio Y, Tomita T: 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*. Biochim Biophys Acta 2004;1679:65–73.
Schlumberger S, Kristan KC, Ota K, Frangez R, Molgomicron J, Sepcic K, Benoit E, Macek P: Permeability characteristics of cell-membrane pores induced by ostreolysin A/pleurotolysin B, binary pore-forming proteins from the oyster mushroom. FEBS Lett 2014; 588:35–40.
Shibata T, Kudou M, Hoshi Y, Kudo A, Nanashima N, Miyairi K: Isolation and characterization of a novel two-component hemolysin, erylysin A and B, from an edible mushroom, *Pleurotus eryngii*. Toxicon 2010;56:1436–1442.

Shogomori H, Kobayashi T: Lysenin: a sphingomyelin specific pore-forming toxin. Biochim Biophys Acta 2008;1780:612–618.
Sievers F, Higgins DG: Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 2014;1079:105–116.
Slade DJ, Lovelace LL, Chruszcz M, Minor W, Lebioda L, Sodetz JM: Crystal structure of the MACPF domain of human complement protein C8 alpha in complex with the C8 gamma subunit. J Mol Biol 2008;379:331–342.
Stewart SE, Kondos SC, Matthews AY, D'Angelo ME, Dunstone MA, Whisstock JC, Trapani JA, Bird PI: The perforin pore facilitates the delivery of cationic cargos. J Biol Chem 2014; 289:9172–9181.
Tomita T, Noguchi K, Mimuro H, Ukaji F, Ito K, Sugawara-Tomita N, Hashimoto Y: Pleurotolysin, a novel sphingomyelin-specific two-component cytolysin from the edible mushroom *Pleurotus ostreatus*, assembles into a transmembrane pore complex. J Biol Chem 2004;279:26975–26982.
Wang Y, Bjes ES, Esser AF: Molecular aspects of complement-mediated bacterial killing. Periplasmic conversion of C9 from a protoxin to a toxin. J Biol Chem 2000;275:4687–4692.
Wright KO, Messing EM, Reeder JE: DBCCR1 mediates death in cultured bladder tumor cells. Oncogene 2004;23:82–90.
Xu Q, Abdubek P, Astakhova T, Axelrod HL, Bakolitsa C, Cai X, Carlton D, Chen C, Chiu HJ, Clayton T, Das D, Deller MC, Duan L, Ellrott K, Farr CL, Feuerhelm J, Grant JC, Grzechnik A, Han GW, Jaroszewski L, Jin KK, Klock HE, Knuth MW, Kozbial P, Krishna SS, Kumar A, Lam WW, Marciano D, Miller MD, Morse AT, Nigoghossian E, Nopakun A, Okach L, Puckett C, Reyes R, Tien HJ, Trame CB, van den Bedem H, Weekes D, Wooten T, Yeh A, Zhou J, Hodgson KO, Wooley J, Elsliger MA, Deacon AM, Godzik A, Lesley SA, Wilson IA: Structure of a membrane-attack complex/perforin (MACPF) family protein from the human gut symbiont *Bacteroides thetaiotaomicron*. Acta Crystallogr Sect F Struct Biol Cryst Commun 2010;66:1297–1305.
Zhang D, Aravind L: Identification of novel families and classification of the C2 domain superfamily elucidate the origin and evolution of membrane targeting activities in eukaryotes. Gene 2010;469:18–30.