Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

Structure of a membrane-attack complex/perforin (MACPF) family protein from the human gut symbiont *Bacteroides thetaiotaomicron*

Qingping Xu,[a,b] Polat Abdubek,[b,c] Tamara Astakhova,[b,d] Herbert L. Axelrod,[a,b] Constantina Bakolitsa,[b,e] Xiaohui Cai,[b,d] Dennis Carlton,[b,f] Connie Chen,[b,c] Hsiu-Ju Chiu,[a,b] Thomas Clayton,[b,f] Debanu Das,[a,b] Marc C. Deller,[b,f] Lian Duan,[b,d] Kyle Ellrott,[b,d] Carol L. Farr,[b,f] Julie Feuerhelm,[b,c] Joanna C. Grant,[b,c] Anna Grzechnik,[b,f] Gye Won Han,[b,f] Lukasz Jaroszewski,[b,d,e] Kevin K. Jin,[a,b] Heath E. Klock,[b,c] Mark W. Knuth,[b,c] Piotr Kozbial,[b,e] S. Sri Krishna,[b,d,e] Abhinav Kumar,[a,b] Winnie W. Lam,[a,b] David Marciano,[b,f] Mitchell D. Miller,[a,b] Andrew T. Morse,[b,d] Edward Nigoghossian,[b,c] Amanda Nopakun,[b,f] Linda Okach,[b,c] Christina Puckett,[b,c] Ron Reyes,[a,b] Henry J. Tien,[b,f] Christine B. Trame,[a,b] Henry van den Bedem,[a,b] Dana Weekes,[b,e] Tiffany Wooten,[b,c] Andrew Yeh,[a,b] Jiadong Zhou,[b,c] Keith O. Hodgson,[b,g] John Wooley,[b,d] Marc-André Elsliger,[b,f] Ashley M. Deacon,[a,b] Adam Godzik,[b,d,e] Scott A. Lesley[b,c,f] and Ian A. Wilson[b,f]*

[a]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA, USA, [b]Joint Center for Structural Genomics, http://www.jcsg.org, USA, [c]Protein Sciences Department, Genomics Institute of the Novartis Research Foundation, San Diego, CA, USA, [d]Center for Research in Biological Systems, University of California, San Diego, La Jolla, CA, USA, [e]Program on Bioinformatics and Systems Biology, Sanford–Burnham Medical Research Institute, La Jolla, CA, USA, [f]Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA, USA, and [g]Photon Science, SLAC National Accelerator Laboratory, Menlo Park, CA, USA

Correspondence e-mail: wilson@scripps.edu

Received 26 May 2010
Accepted 15 June 2010

PDB Reference: MACPF family protein, 3kk7.

Membrane-attack complex/perforin (MACPF) proteins are transmembrane pore-forming proteins that are important in both human immunity and the virulence of pathogens. Bacterial MACPFs are found in diverse bacterial species, including most human gut-associated *Bacteroides* species. The crystal structure of a bacterial MACPF-domain-containing protein BT_3439 (Bth-MACPF) from *B. thetaiotaomicron*, a predominant member of the mammalian intestinal microbiota, has been determined. Bth-MACPF contains a membrane-attack complex/perforin (MACPF) domain and two novel C-terminal domains that resemble ribonuclease H and interleukin 8, respectively. The entire protein adopts a flat crescent shape, characteristic of other MACPF proteins, that may be important for oligomerization. This Bth-MACPF structure provides new features and insights not observed in two previous MACPF structures. Genomic context analysis infers that Bth-MACPF may be involved in a novel protein-transport or nutrient-uptake system, suggesting an important role for these MACPF proteins, which were likely to have been inherited from eukaryotes *via* horizontal gene transfer, in the adaptation of commensal bacteria to the host environment.

1. Introduction

Perforin (PF) and components of the membrane-attack complex (MAC; complement proteins C6–C9) are pore-forming proteins of the complement part of the innate immune system. They share a common domain (MACPF) that is also widely distributed in bacteria and protozoa, including many pathogens (Rosado *et al.*, 2008; Voskoboinik *et al.*, 2006). Perforin-like proteins in pathogens play an important role in virulence, for example, by disrupting the plasma membrane and facilitating parasite exit from host cells (Kafsack *et al.*, 2009). The recent structures of two MACPF proteins, a bacterial protein from *Photorhabdus luminescens* (Plu-MACPF) and the human complement membrane-attack complex component C8α, revealed an unexpected structural similarity to the well studied cholesterol-dependent cytolysins (CDCs) of many Gram-positive bacteria, thus suggesting a common mechanism of pore formation (Hadders *et al.*, 2007; Rosado *et al.*, 2007) by CDC and MACPF. CDCs form doughnut-shaped pores by the self-polymerization of 30–50 monomers on target membrane surfaces, followed by a major structural rearrangement and the insertion of two helical regions (Tweten, 2005).

The Gram-negative anaerobic *Bacteroides thetaiotaomicron*, which is a predominant member of the human intestinal tract microbiota, is an important bacterium for the study of the symbiotic relationship between bacteria and humans (Xu *et al.*, 2003; Hooper & Gordon, 2001). Extracellular proteins are crucial for these functions in *B. thetaiotaomicron* and other gut microbes. We initiated a structural genomics project that aims to determine the structures of proteins that are unique to the secretome of human gut microbiota in order to provide broad insights into the molecular mechanisms of bacteria–host symbiosis and pathogenesis. We have selected proteins that do not display significant similarities to proteins of known structure and have determined the structures of more than 60 secreted human gut bacteria proteins thus far. Our structures have revealed that many of these proteins are distant homologs of well known protein families,

Table 1
Summary of crystal parameters, data-collection and refinement statistics for Bth-MACPF (PDB code 3kk7).

Values in parentheses are for the highest resolution shell.

	λ_1 MADSe	λ_2 MADSe	λ_3 MADSe
Space group	$P2_12_12_1$		
Unit-cell parameters (Å)	a = 78.4, b = 127.2, c = 138.3		
Data collection			
Wavelength (Å)	0.9791	0.9184	0.9792
Resolution range (Å)	49.4–2.46 (2.59–2.46)	48.0–2.80 (2.95–2.80)	48.0–2.80 (2.95–2.80)
No. of observations	342564	122453	121355
No. of reflections	49779	34098	33698
Completeness (%)	97.3 (94.2)	98.2 (98.2)	97.0 (99.7)
Mean $I/\sigma(I)$	10.6 (2.3)	7.3 (2.0)	8.6 (2.6)
$R_{\rm merge}$ on I†	0.123 (0.75)	0.153 (0.69)	0.125 (0.52)
Model and refinement statistics			
Resolution range (Å)	49.4–2.46		
No. of reflections (total)	49764		
No. of reflections (test)	2514		
Completeness (%)	97.3		
Data set used in refinement	λ_1 MADSe		
Cutoff criterion	$\lvert F\rvert > 0$		
$R_{\rm cryst}$‡	0.209		
$R_{\rm free}$§	0.252		
Stereochemical parameters			
Restraints (r.m.s.d. observed)			
Bond lengths (Å)	0.014		
Bond angles (°)	1.47		
Average isotropic B value (Å^2)	40.2¶		
ESU†† based on $R_{\rm free}$ (Å)	0.27		
Protein residues/atoms	1001/8046		
Solvent molecules	244		

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$. ‡ $R_{\rm cryst} = \sum_{hkl} \big||F_{\rm obs}| - |F_{\rm calc}|\big| / \sum_{hkl} |F_{\rm obs}|$, where $F_{\rm calc}$ and $F_{\rm obs}$ are the calculated and observed structure-factor amplitudes, respectively. § $R_{\rm free}$ is the same as $R_{\rm cryst}$ but for 5% of the total reflections chosen at random and omitted from refinement. ¶ This value represents the total B that includes TLS and residual B components. †† Estimated standard uncertainty in coordinates (Collaborative Computational Project, Number 4, 1994; Cruickshank, 1999).

which, in many cases, are undetectable based on sequence alone using even the most sensitive fold-detection algorithms. For example, the structure of a putative fimbriae assembly protein BT_1062 from *B. thetaiotaomicron* (PDB code 3gf8) revealed a fold similar to pili components of other bacteria despite no detectable sequential similarity (Xu *et al.*, 2010). Similarly, the structure of BVU_2987 (PDB code 3due) from *B. vulgatus* uncovered an unexpected similarity in fold to the β-lactamase inhibitor protein (BLIP; Das *et al.*, 2010). Therefore, these proteins are also good candidates for exploring the evolution and divergence of protein structures and the underlying sequence–structure relationships. Here, we report the crystal structure of the MACPF protein BT_3439 from *B. thetaiotaomicron* (hereafter referred to as Bth-MACPF) at 2.46 Å resolution, which to our knowledge is the first structure of a potential CDC-like toxin from a gut symbiont.

2. Materials and methods

2.1. Protein production and crystallization

Clones were generated using the Polymerase Incomplete Primer Extension (PIPE) cloning method (Klock *et al.*, 2008). The gene encoding Bth-MACPF (GenBank NP_812351; Swiss-Prot Q8A267) was amplified by polymerase chain reaction (PCR) from *B. thetaiotaomicron* VPI-5482 genomic DNA using *PfuTurbo* DNA polymerase (Stratagene) and I-PIPE (Insert) primers (forward primer, 5′-ctgtacttccagggcAATGAGGAGGAAACTAATAATTATACTC-3′; reverse primer, 5′-ctgtacttccagggcAATGAGGAGGAAACTAATAATTATACTC-3′; target sequence in upper case) that included sequences for the predicted 5′ and 3′ ends. The expression vector pSpeedET, which encodes an amino-terminal tobacco etch virus (TEV) protease-cleavable expression and purification tag (MGSDKIHHHHHHHENLYFQ/G), was PCR-amplified with V-PIPE (Vector) primers (forward primer, 5′-taacgcgacttaattaactcgtttaaacggtctccagc-3′; reverse primer, 5′-gccctggaagtacaggttttcgtgatgatgatgatgatgatg-3′). The V-PIPE and I-PIPE PCR products were mixed to anneal the amplified DNA fragments together. *Escherichia coli* GeneHogs (Invitrogen) competent cells were transformed with the I-PIPE/V-PIPE mixture and dispensed onto selective LB–agar plates. The cloning junctions were confirmed by DNA sequencing. Using the PIPE method, the gene segment encoding residues Met1–Thr18 was excluded from the final construct as it was predicted to encode a signal peptide. Expression was performed in selenomethionine-containing medium at 310 K. Selenomethionine was incorporated *via* inhibition of methionine biosynthesis (Van Duyne *et al.*, 1993), which does not require a methionine-auxotrophic strain. At the end of fermentation, lysozyme was added to the culture to a final concentration of 250 µg ml^{-1} and the cells were harvested and frozen. After one freeze–thaw cycle, the cells were sonicated in lysis buffer [50 m*M* HEPES pH 8.0, 50 m*M* NaCl, 10 m*M* imidazole, 1 m*M* tris(2-carboxyethyl)phosphine–HCl (TCEP)] and the lysate was clarified by centrifugation at 32 500*g* for 30 min. The soluble fraction was passed over nickel-chelating resin (GE Healthcare) pre-equilibrated with lysis buffer, the resin was washed with wash buffer [50 m*M* HEPES pH 8.0, 300 m*M* NaCl, 40 m*M* imidazole, 10%(*v*/*v*) glycerol, 1 m*M* TCEP] and the protein was eluted with elution buffer [20 m*M* HEPES pH 8.0, 300 m*M* imidazole, 10%(*v*/*v*) glycerol, 1 m*M* TCEP]. The eluate was buffer-exchanged with TEV buffer (20 m*M* HEPES pH 8.0, 200 m*M* NaCl, 40 m*M* imidazole, 1 m*M* TCEP) using a PD-10 column (GE Healthcare) and incubated with 1 mg TEV protease per 15 mg of eluted protein. The protease-treated eluate was run over nickel-chelating resin (GE Healthcare) pre-equilibrated with HEPES crystallization buffer (20 m*M* HEPES pH 8.0, 200 m*M* NaCl, 40 m*M* imidazole, 1 m*M* TCEP) and the resin was washed with the same buffer. The flowthrough and wash fractions were combined and concentrated to 19.8 mg ml^{-1} by centrifugal ultrafiltration (Millipore) for crystallization trials. Bth-MACPF was crystallized by mixing 100 nl protein solution with 100 nl crystallization solution in a sitting drop over a 50 µl reservoir volume using the nanodroplet vapor-diffusion method (Santarsiero *et al.*, 2002) with standard JCSG crystallization protocols (Lesley *et al.*, 2002). The crystallization reagent consisted of 5%(*v*/*v*) 2-methyl-2,4-pentanediol, 12%(*v*/*v*) polyethylene glycol 6000, 0.1 *M* HEPES pH 6.7. A cube-shaped crystal of approximate dimensions 40 × 40 × 30 µm was harvested after 42 d at 277 K for data collection. Glycerol was diluted to 10%(*v*/*v*) using the reservoir solution and then added to the drop in a 1:1 ratio as a cryoprotectant prior to mounting. Initial screening for diffraction was carried out using the Stanford Automated Mounting system (SAM; Cohen *et al.*, 2002) at the Stanford Synchrotron Radiation Lightsource (SSRL, Menlo Park, California, USA).

The oligomeric state of Bth-MACPF in solution was determined using a 1 × 30 cm Superdex 200 size-exclusion column (GE Healthcare) coupled with miniDAWN static light-scattering (SEC/SLS) and Optilab differential refractive-index detectors (Wyatt Technology). The mobile phase consisted of 20 m*M* Tris–HCl pH 8.0, 150 m*M* NaCl and 0.02%(*w*/*v*) sodium azide. The molecular weight was calculated using *ASTRA* v.5.1.5 software (Wyatt Technology).

2.2. Data collection, structure solution, refinement and analysis

Multi-wavelength anomalous diffraction (MAD) data were collected on beamline 9-2 at the SSRL at wavelengths corresponding

to the peak (λ_1), high-energy remote (λ_2) and inflection (λ_3) wavelengths of a selenium MAD experiment (see Table 1). The data sets were collected at 100 K using a MAR CCD 325 detector. The MAD data were integrated and reduced using *XDS* and scaled with the program *XSCALE* (Kabsch, 2010). Selenium sites were located with *SHELXD* (Sheldrick, 2008) and refined using *autoSHARP* (mean figure of merit of 0.34 with 22 selenium sites; Bricogne *et al.*, 2003). Density modification was performed by *SOLOMON* (Abrahams & Leslie, 1996) and automatic model building was performed by *Buccaneer* (Cowtan, 2006). Iterative model building and refinement

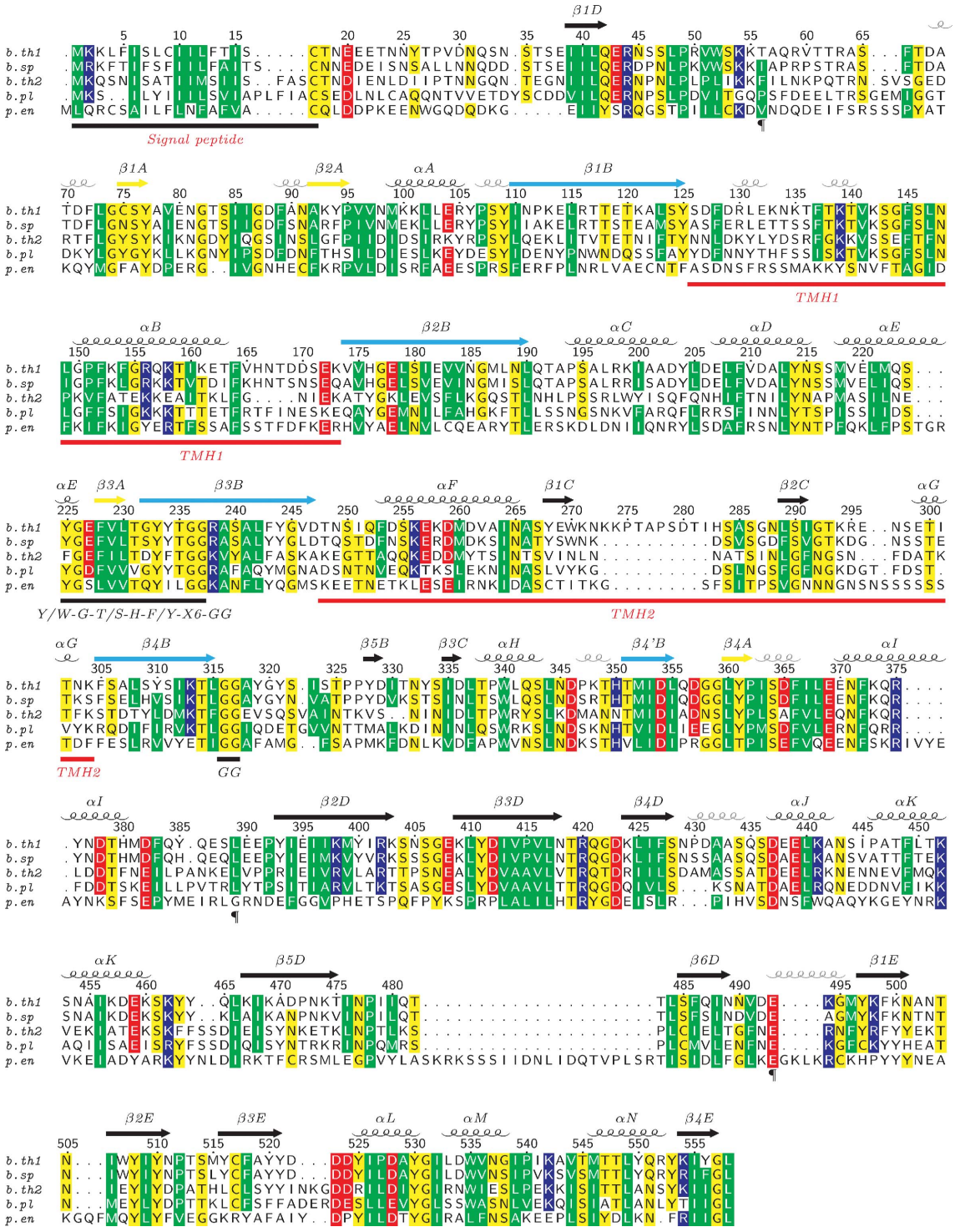

Figure 1
Multiple sequence alignment of Bth-MACPF homologs (sequence identity <90%) with the same domain architecture. The sequence numbering and secondary-structure elements of Bth-MACPF are shown at the top and domain boundaries (¶) and sequence motifs are shown at the bottom. Charged residues are highlighted in red (negative) and blue (positive), hydrophobic residues in green and hydrophilic residues in yellow. The following sequences are shown: *b.th1*, *B. thetaiotaomicron* BT_3439 (Bth-MACPF); *b.sp*, *Bacteroides* sp. 2_2_4 (UniProt accession C3QVE5); *b.th2*, *B. thetaiotaomicron* BT_3437; *b.pl*, *B. plebeius* DSM 17135 (UniProt accession B5CX96); *p.en*, *P. endodontalis* ATCC 35406 (UniProt accession C3J7W9).

were performed with *Coot* (Emsley & Cowtan, 2004) and *REFMAC* (Winn *et al.*, 2003), respectively. The refinement included experimental phase restraints in the form of Hendrickson–Lattman coefficients and TLS refinement with four TLS groups per chain (residues 36–56, 66–389, 390–493 and 494–558). *CCP*4 programs were used for data conversion and other calculations (Collaborative Computational Project, Number 4, 1994). Data-processing and refinement statistics are summarized in Table 1. The quality of the crystal structure was evaluated using *MolProbity* (Chen *et al.*, 2010) and *WHAT IF* (Vriend, 1990). *HHpredict* was used for protein-homology detection and function prediction (Soding *et al.*, 2005). Signal peptides were analyzed using *SignalP* (Emanuelsson *et al.*, 2007) and *LipoP* (Juncker *et al.*, 2003). Oligomers of Bth-MACPF with C_{16} symmetry were predicted using *SymmDOCK* (Schneidman-Duhovny *et al.*, 2005). Molecular graphics were prepared with *PyMOL* (DeLano Scientific). Sequence alignments were rendered using *TEXshade* (Beitz, 2000).

3. Results and discussion

3.1. Bioinformatics analysis

MACPF domains are widely distributed in eukaryotes, but are sporadic in bacteria. Only ∼40 bacterial MACPF proteins are cataloged in the PFAM database (PF01823; Bateman *et al.*, 2004). Chlamydiaceae contain 13 closely related MACPF proteins (Ponting, 1999). *Bacteroides* contain about a third of all bacterial MACPF proteins. The others are found in diverse bacterial species from proteobacteria, actinomycetales and cyanobacteria. It has been suggested that these proteins were acquired from eukaryotes through horizontal gene transfer in order to adapt to the intracellular environment of the host (Ponting, 1999; Wolf *et al.*, 1999). Preliminary phylogenetic analysis (data not shown) suggests that the Bacteroidetes branch is likely to represent an independent horizontal gene-transfer event. Thus, MACPFs in the human gut microbiome may play an important role in the symbiotic relationship, but their specific functions are currently unknown.

Bacterial MACPFs are highly divergent in sequence and domain architecture. Homologs that have significant similarity over the entire sequence of Bth-MACPF are found mostly in other human-related Bacteroidetes, including unclassified *Bacteroides* sp. (strains 2_1_22, 2_2_4 and D1), *B. fragilis* 3_1_12 (Bfra3_17507), *B. plebeius* DSM 17135 (BACPLE_01336), *B. intestinalis* DSM 17393 (BACINT_00423) and *Porphyromonas endodontalis* ATCC 35406 (POREN0001_1212) (Fig. 1), but also in the recently sequenced deep-sea *Zunongwangia profunda* SM-A87 (ZPR_2061). MACPFs from *Bacteroides* are unique as most of them contain lipoprotein signal peptides (Juncker *et al.*, 2003) that are not present in other bacterial MACPFs. *B. thetaiotaomicron* contains two homologous MACPFs (BT_3439 and BT_3437; 33% sequence identity) that are likely to form part of an operon (see more detailed discussion below), as well as a third more distant paralog (BT_3120) that consists of only an MACPF domain. *B. fragilis* YCH46 (BF1566, BF1634 and BF2685) and *B. intestinalis* DSM 17393 (BACINT_00423, BACINT_00829 and BACINT_03190) each contain three MACPFs, with only one protein in each species having the same domain architecture as Bth-MACPF.

Bth-MACPF is located among a cluster of uncharacterized proteins (BT_3442 to BT_3433) that form a putative operon and which are located directly downstream of a well defined operon of cell-division and cell-wall biosynthesis proteins such as FtsZ, FtsA, FtsQ and MurC. This cluster, which appears to contain internal duplications resulting in three homologous pairs (BT_3436/BT_3438, BT_3437/BT_3439 and BT_3433/BT_3440), is rich in potential pore-forming proteins (BT_3433, BT_3434, BT_3437, BT_3439 and BT_3440). Most of the proteins in the cluster also contain similar lipoprotein signal peptides (Fig. 2), suggesting that they are localized to a common area in the cell. BT_3433 and BT_3440 are likely to have a trefoil fold resembling that of hemolytic pore-forming lectins (Mancheno *et al.*, 2005). BT_3434 is likely to be an outer membrane porin, while BT_3435 is a putative inner membrane protein with three transmembrane helices. BT_3441 is a homolog of a hypothetical protein BVU_0276 from *B. vulgatus*, the structure of which has also been determined by the JCSG (PDB code 3d33). It has an immunoglobulin-like fold that is common in cell-surface proteins such as fibronectin and complement C3. BT_3442 is a multi-domain protein containing TPR motifs, which often mediate protein interaction. Therefore, Bth-MACPF is associated with several pore-forming proteins, suggesting a possible role in a cross-membrane transport system. The association of *Bacteroides* MACPFs with lipoproteins and outer membrane porins is also observed in *B. fragilis* YCH46 (Fig. 2).

Bth-MACPF was predicted to be an extracellular protein by *PSORTb* (Gardy *et al.*, 2005) and *SOSUI*$_{GramN}$ (Imai *et al.*, 2008). The N-terminal region of Bth-MACPF (^{1}MKKLFISLCIILFTISC^{17}) matches the lipoprotein signal peptide pattern of Gram-negative bacteria, which usually consists of one or more positive charged residues followed by a stretch of hydrophobic residues and a lipobox motif L(A/S)(G/A)C (Hayashi & Wu, 1990). Similar lipoprotein

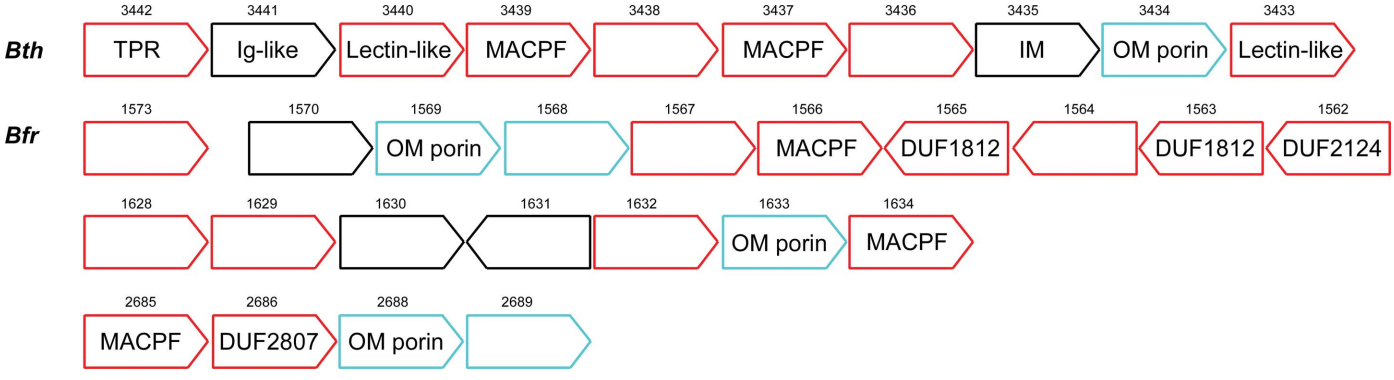

Figure 2
Genomic context of MACPF proteins in two completed genomes of *Bacteroides*: *B. thetaiotaomicron* (Bth) and *B. fragilis* YCH46 (Bfr). Predicted lipoproteins are shown as red boxes. Proteins containing other signal peptides are colored cyan. The locus number of each gene is shown at the top. IM, inner membrane protein; OM, outer membrane protein; DUF, domain of unknown function; TPR, protein containing tetratricopeptide repeats.

signal peptides are also present in structural subunits of the major and minor fimbriae FimA and Mfa1 of *P. gingivalis*, which is a close phylogenetic relative of *B. thetaiotaomicron*, suggesting a common mechanism of translocation across the membrane (Shoji *et al.*, 2004). Lipoproteins are transported across the inner membrane by the general secretion pathway. On the periplasmic face of the inner membrane, the invariant cysteine residue is modified by the diacylglyceryl transferase (Lgt), followed by cleavage of the peptide before the diacylglyceride cysteine by signal peptidase II (LspA) and further modification of the diacylglyceride cysteine by aminoacyl transferase (Lnt; Tokuda, 2009). These proteins are then sorted to their final destinations, but the details of the final steps of translocation of extracellular lipoproteins in *Bacteroides* are currently not clear. The final products could either be tethered to the outer membrane or cleaved and released to the extracellular medium and may be dependent on other residues in close proximity to the cysteine (*e.g.* the conserved acidic residue at position +4; Fig. 1).

3.2. Structural determination

The *BT_3439* gene of *B. thetaiotaomicron* encodes a predicted lipoprotein with a molecular weight of 63 425 Da (residues 1–558) and a calculated isoelectric point of 5.5. We determined the structure using the high-throughput pipeline of the Joint Center for Structural Genomics (JCSG; Lesley *et al.*, 2002) as part of the National Institute of General Medical Sciences' Protein Structure Initiative (PSI; http://www.nigms.nih.gov/Initiatives/PSI/). A selenomethionine derivative of Bth-MACPF was expressed in *E. coli* with an N-terminal TEV-cleavable His tag and was purified by metal-affinity chromatography. To improve the likelihood of obtaining crystals, the predicted N-terminal signal peptide (residues 1–18) was not included in the clone construct. The data were indexed in the orthorhombic space group $P2_12_12_1$ and the structure was determined at 2.46 Å resolution with two molecules per asymmetric unit using the MAD method. The structure was refined to a final R factor of 20.9% and an R_{free} of 25.2%. The model of Bth-MACPF displays good geometry, with an all-atom clash score of 7.8, and the Ramachandran plot produced by *MolProbity* (Chen *et al.*, 2010) shows that all residues are in allowed regions, with 96.7% in favored regions. The final model of Bth-MACPF contains residues *A*/*B*36–558, 239 waters and other solvent molecules that were present in the crystallization or cryoprotection reagents, including one MPD [(4*S*)-2-methyl-2,4-pentanediol] molecule, one chloride ion and three ethylene glycol molecules. The residual residue (Gly0) from the cleaved N-terminal purification tag and segments *A*/*B*19–35, *A*/*B*57–65, *A*277–286, *B*272–286 and *A*482–483 were not included in the model owing to a lack of interpretable electron density. Additionally, side chains for 17 residues were only partially modeled owing to disorder. Data-collection, refinement and model statistics are summarized in Table 1.

3.3. Overall structure

Bth-MACPF (Fig. 3) adopts a flat crescent shape with molecular dimensions of 93 × 58 × 44 Å. The two monomers in the asymmetric unit are nearly identical (with an overall r.m.s.d. of 0.67 Å for 493 C^{α} atoms) with larger deviations located at the two tips, mostly owing to a slight opening of the crescent in molecule *B* compared with molecule *A*. Bth-MACPF consists of three structured domains: an MACPF domain (residues 66–389) and two C-terminal domains, D2

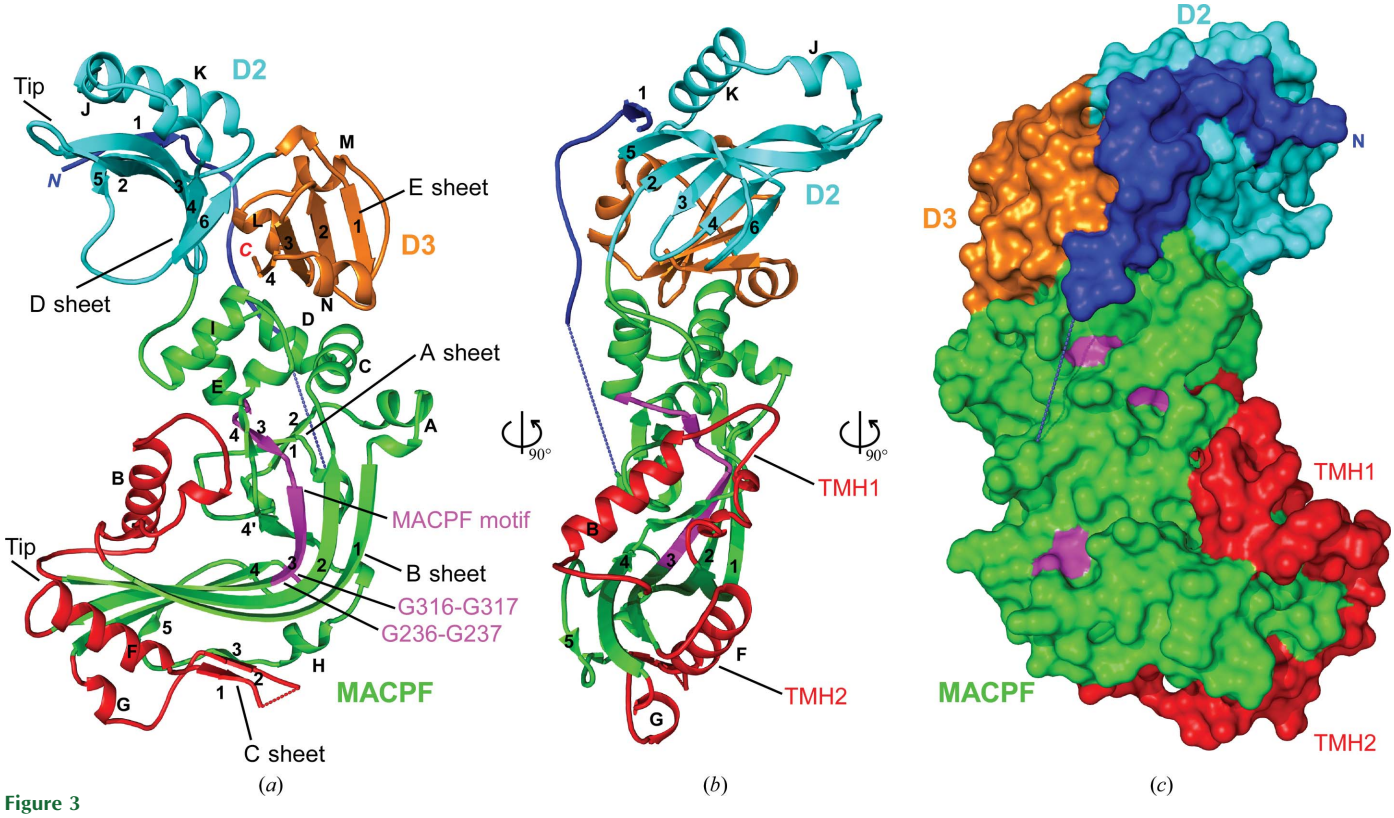

Figure 3
Crystal structure of Bth-MACPF. (*a* and *b*) Ribbon representations of Bth-MACPF in orthogonal views. The color scheme is as follows: the extended N-terminal region is shown in blue, the MACPF domain in shown in green with TMHs in red and the MACPF motif in magenta, domain D2 is shown in cyan and domain D3 is shown in orange. The β-sheets (A–E) and helices (A–N) are labeled alphabetically as in Fig. 1; 3_{10}-helices are not labeled. (*c*) Surface representation of Bth-MACPF color coded by domain as in (*a* and *b*).

(residues 390–492) and D3 (residues 493–558) (Figs. 3*a* and 3*b*). Residues 36–56 of the N-terminus of Bth-MACPF adopt an extended conformation and pack against parts of the MACPF (residues 388–392), D2 (residues 449–474) and D3 (residues 523–531) domains with a total buried surface area of 2029 Å^2 (Fig. 3*b*). The interface contains 32 hydrogen bonds and helps to maintain overall structural integrity. This arrangement places the predicted N-terminal membrane-attachment site (Cys18) away from the MACPF domain. The remaining N-terminal residues that were included in the construct (residues 19–35 and 57–65) were not observed in the electron density and are most likely to be flexible in solution. Furthermore, Bth-MACPF is likely to be a monomer in solution, as supported by crystal-packing analysis and analytical size-exclusion chromatography (data not shown).

The MACPF domain contains two four-stranded β-sheets (A and B) in the central core, which is decorated by several helical insertions. The A sheet with its short strands (strand order 2134) and the B sheet with long strands (strand order 1234) assemble to form a twisted S shape. The B sheet itself is very distorted and bends fairly abruptly in the middle by $\sim$90°. This arrangement of central β-sheets with characteristic geometry is common to both MACPFs and CDCs and allowed the classification of MACPF and CDC into a single family (Rosado *et al.*, 2007; Hadders *et al.*, 2007). The last strand of the B sheet is interrupted (strands 4 and 4′) by an insertion (residues 316–350) at the bend of the sheet. Insertions between β1–β2 and β3–β4 (TMH1 and TMH2, respectively) correspond to the so-called TMH regions of CDCs, which unfold and form transmembrane β-hairpins. TMH1 (residues 126–173) contains one helix (αB) and two short 3_{10}-helices that pack against the inner surface of the B sheet. TMH2 (residues 248–304) contains an antiparallel $\alpha\beta$–$\beta\alpha$ structure that sits on the outer surface of the B sheet. The two strands in TMH2 and another strand from the 4–4′ (B-sheet) insertion forms another β-sheet (C sheet) parallel to the B sheet. The MACPF motif Y/W-G-T/S-H-F/Y-X_6-GG (Ponting, 1999; Rosado *et al.*, 2007) is located on strands 3A and 3B (Fig. 3*a*). The corresponding Bth-MACPF region (^{225}YG**EFV**X_6GG^{237}) is more divergent from the consensus, with nonconserved changes at positions 3–5. Two glycines from the MACPF motif (Gly236 and Gly237) and two additional nearby conserved glycines (Gly316 and Gly317; Figs. 1 and 3) are likely to be essential for structural flexibility in MACPF and CDC (Rosado *et al.*, 2007).

The A sheet is crowned by four helices: αI and a three-helix insertion (αC–αE) between β2B (β2 of the B sheet) and β3A. These helices form the interface between the MACPF and D2/D3 domains. Both D2 and D3 are layered structures with a central β-sheet protected by helices on two sides (see below). The D2 and MACPF interface involves interaction between αD and αI of MACPF and the β3–β4 and β5–β6 loops of D2 and buries a surface area of $\sim$1000 Å^2 (500 Å^2 each). The interface is mostly hydrophilic. In particular, a buried Asp423 in D2 forms a bifurcated hydrogen bond to Arg375 of MACPF. D3 functions as a wedge between D2 and MACPF, with a similar interface area on either side (total $\sim$1400 Å^2 for D3). Leu558 is buried with its C-terminal carboxyl group forming a hydrogen-bond network involving the conserved residues Arg420 and Tyr530. Additionally, the interaction between domains is further stabilized by the N-terminal extended region (residues 36–56) described above. Gap-volume indices between these interacting components are less than 1.7, which is consistent with the expected average (1.8) for intrachain domain–domain interfaces (Jones *et al.*, 2000). Thus, we conclude that the domain arrangement observed in the crystal structure is likely to be representative of the functional protein and not a crystallization artifact.

3.4. D2 and D3 domains

The MACPF domain is usually attached to other auxiliary domains that are expected to regulate the function of MACPF. As discussed earlier, both C-terminal domains of Bth-MACPF are only detected in its closest homologs in sequence-similarity searches (Fig. 1). The D2 and D3 domains show some structural similarity: both have an α/β fold with $\beta\beta\beta\alpha\beta$ topology. However, most structural comparison programs fail to recognize this similarity and also fail to identify significant similarities to other proteins. The $\beta\beta\beta\alpha\beta$ core of D2 and D3 can be partly matched to other structures (Fig. 4), for instance to proteins with the YegP-like fold (SCOP ID 160112), which is characterized by an internal repeat of two domains with a $\beta\beta\beta\alpha\beta$ core. Other examples include the connector domain (residues 321–431; PDB code 1mu2; Ren *et al.*, 2002) of HIV reverse transcriptase (Z = 3.6; r.m.s.d. 3.3 Å for 68 aligned C$^\alpha$ atoms; sequence identity 6%), which is likely to have evolved from the ribonuclease H domain (Malik & Eickbush, 2001; Fig. 4*a*). However, the C-terminal portions of the two structures differ significantly. Domain D3 is similar, for instance, to a viral chemokine (PDB code 1zxt; Luz *et al.*, 2005), with an r.m.s.d. of 2.2 Å (sequence identity 5%) for 44 C$^\alpha$ atoms (Fig. 4*b*). Chemokines adopt a $\beta\beta\beta\alpha$ interleukin 8-like structure stabilized by two conserved disulfide bonds. D3 lacks the long cysteine-containing N-terminal portion observed in chemokines. Instead, it contains an $\alpha\beta$ C-terminal extension and forms a $\beta\beta\beta\alpha\alpha\alpha\beta$ overall structure. The $\beta\beta\beta\alpha\beta$ motif is most likely to represent a repeated structural unit that can be found in nonhomologous proteins with different functions, thus limiting the interpretation of structural similarity in terms of common function.

3.5. Homology of MACPF domains

The MACPF domain in Bth-MACPF is homologous to human MACPFs, as indicated by the significant sequence similarity recognized, for instance, by *FFAS* (Jaroszewski *et al.*, 2005) and *HHpredict* (Hildebrand *et al.*, 2009) and by three-dimensional structural similarity using the *DALI* server (Holm & Sander, 1995). The first two *DALI* hits are the only two previously determined MACPF structures: Plu-MACPF (PDB code 2qp2; Rosado *et al.*, 2007) and the C8α MACPF domain (PDB codes 2qqh and 2rd7; Hadders *et al.*, 2007; Slade *et al.*, 2008). Bth-MACPF is most similar to Plu-MACPF, with a Z score of 17.4, which corresponds to an r.m.s.d. of 3.8 Å and 16% sequence identity for 247 aligned C$^\alpha$ atoms. The second hit, human C8α (PDB code 2qqh), can be superimposed onto Bth-MACPF with 218 aligned C$^\alpha$ atoms, an r.m.s.d. of 5.0 Å and 14% sequence identity (Z = 12.3). More distant similarity is also apparent between Bth-MACPF and CDCs, such as the thiol-activated cytolysin perfringolysin O (PFO; PDB code 1m3i; Rossjohn *et al.*, 1997; Z = 7.2, r.m.s.d. 5.2 Å and 11% sequence identity for 198 aligned C$^\alpha$ atoms). The structural similarity between MACPF domains and the CDC family of toxins has previously been noted, which led to the proposal that MACPF domains use a CDC-like mechanism for pore formation (Rosado *et al.*, 2007; Hadders *et al.*, 2007). In this model, TMH1 and TMH2 undergo conformational changes to form antiparallel hairpins so that the extended β-sheet can oligomerize through the open edges of β1 and β4.

The similarity between the three MACPF domains is even more significant at the topological level (Fig. 5). All contain a common core consisting of sheet A and sheet B. Various insertions occur at specific locations in the conserved strands, most notably between β2A and β1B, β4B and β4B′, β1B and β2B (TMH1), β3B and β4B (TMH2) and β2B and β3A. One common helix within the β4B–β4B′ insertion (αH of Bth-MACPF) is conserved in all known MACPFs and harbors

several highly conserved residues (*e.g.* Trp340) that interact with the region containing the critical glycines that were discussed above. The β4B–β4B′ insertion in Bth-MACPF contains two additional short strands that augment the B sheet and the C sheet, respectively. As a result, this insertion in Bth-MACPF is more similar to PFO. The additional short β-strand in the B sheet of CDCs (β5B in Bth-MACPF) prevents premature oligomerization by blocking access to β4 (Ramachandran *et al.*, 2004). The β-hairpin insertion between β2A and β1B of the C8α MACPF domain and Plu-MACPF are replaced by one helix (αA) and a 3_{10}-helix in Bth-MACPF. This region of C8α is involved in the interaction with the C8γ subunit (Slade *et al.*, 2008). The TMH regions of MACPFs and CDCs are generally not conserved in sequence (Rosado *et al.*, 2007). TMHs of Bth-MACPF contain short stretches of amphipathic regions which might be important for forming transmembrane hairpins (Fig. 1). Both TMHs of Bth-MACPF (48 and 57 amino acids) are longer than the TMHs of CDCs, which generally consist of ∼30 amino acids. Longer TMH regions (∼60 amino acids) are also observed in C8α, C9 and perforin and are likely to be a general feature of MACPF. C8α and Bth-MACPF both contain an $\alpha\beta$–$\beta\alpha$ hairpin, but in different locations (TMH1 in C8α and TMH2 in Bth-MACPF). Interestingly, the two faces of the B sheet in all three MACPFs display amphipathic properties. The interface between the B sheet and TMH1 is mostly polar, whereas the TMH2 interface is more tightly packed and hydrophobic (Fig. 5).

3.6. Functional implications

The helical insertion between β2B and β3A is involved in docking the D2 and D3 domains to the Bth-MACPF domain. These helices are also present in Plu-MACPF and C8α MACPF, but are currently not implicated in protein–protein interactions. Both Plu-MACPF and Bth-MACPF contain additional C-terminal domains. However, the locations of these domains are completely different. The C-terminal β-prism domain of Plu-MACPF is located on the opposite side of the central core (left corner of lower figure of Plu-MACPF in Fig. 5) compared with D2 and D3 (upper left corner) in Bth-MACPF. The arrangements of these auxiliary domains may reflect their different roles. The β-prism domain of Plu-MACPF is similarly located

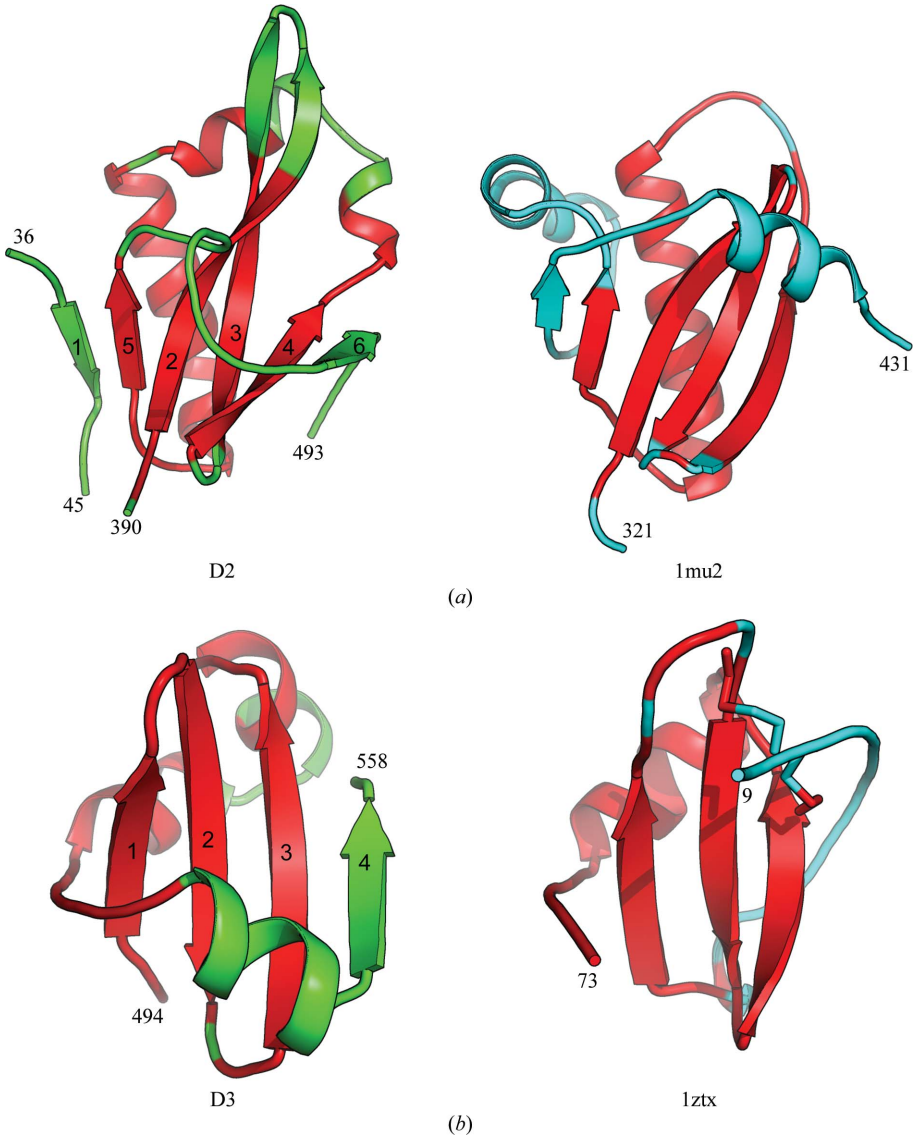

Figure 4
Structural comparisons of the D2 and D3 domains. (*a*) Structural comparison between D2 and the connector domain of HIV reverse transcriptase (PDB code 1mu2). (*b*) Structural comparison between D3 and a viral chemokine (PDB code 1zxt). Equivalent C$^{\alpha}$ atoms are shown in red.

compared with domain 4 of PFO and may interact with the membrane directly (Rosado *et al.*, 2007). In contrast, D2 and D3 of Bth-MACPF, which are distant from the TMH regions, seem more likely to play a role in protein–protein interaction (*e.g.* polymerization or interaction with BT_3442) rather than membrane attachment. The shape of Bth-MACPF appears to be self-complementary, which could facilitate ring-like self-assembly (Hadders *et al.*, 2007) to form pores across membranes. Modeling studies suggest that it is feasible for Bth-MACPF to polymerize *via* the C-terminal auxiliary domains. A model with 16 copies of Bth-MACPF forms a doughnut-shaped molecule with an inner radius of 110 Å, similar in pore size to the the C9 MACPF model (Hadders *et al.*, 2007). The multimer interface involves docking a helical wedge from D2 and D3 (helices K, L and M) into the D2–MACPF interface (D sheet and helix I). The formation of protein complexes involving Bth-MACPF may facilitate structural changes in the MACPF domain which are necessary to form the porin-like transmembrane pore.

MACPFs are well known for killing cells by forming pores and thus are potential virulence factors. Here, we demonstrate the existence of a novel subfamily of secreted MACPF proteins in commensal bacteria. Unfortunately, the physiological functions of these proteins are currently unknown. The properties of the MACPF/CDC fold, such as structural flexibility and membrane penetration, may be utilized for nonlytic purposes (Rosado *et al.*, 2007) and Bth-MACPF may be involved in novel protein-secretion or nutrient-uptake systems. Alternatively, MACPFs may protect the bacteria from host immunity through molecular mimicry (Stebbins & Galan, 2001; Kohm *et al.*, 2003). For example, the presence of these molecules on the cell surface may prevent the assembly of the host MACPF complex. Another possibility is that MACPFs may function as potential toxins, such as bacteriocins against Gram-positive bacteria. Bacteriocins are often produced by nonpathogenic bacteria that colonize the human body and may help to prevent infection by opportunistic pathogenic bacteria. Furthermore, it remains possible

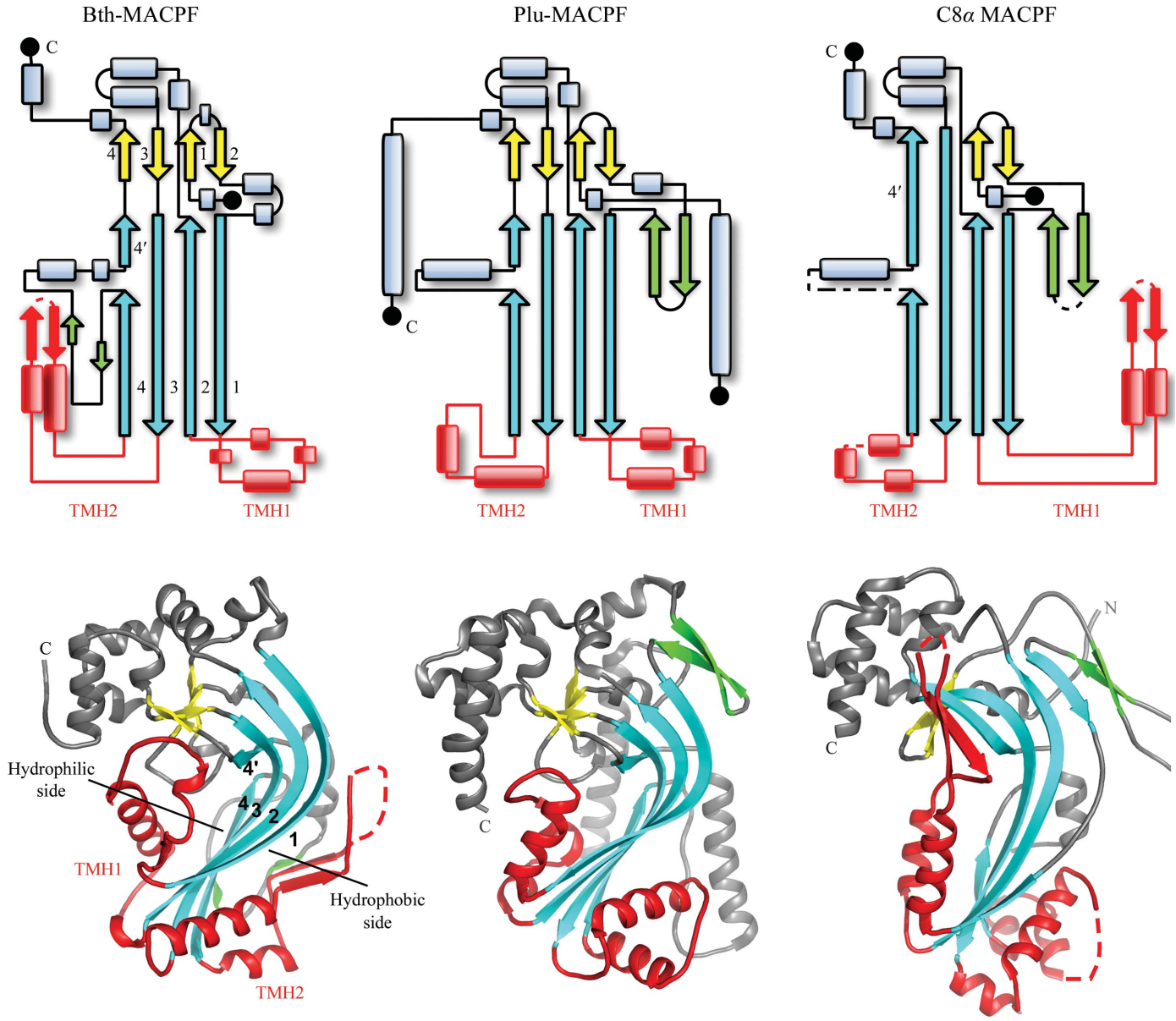

Figure 5
Structural comparison of the MACPF domains in Bth-MACPF, Plu-MACPF and C8α. Top: comparison of the secondary-structure topology diagrams of MACPF domains (sheet A, yellow; sheet B, cyan; TMHs, red). Bottom: ribbon representation of MACPF proteins in the same orientation and color coded as in the topology diagrams.

that these bacterial MACPFs are virulence factors towards the host under certain conditions, as gut symbionts, such as *B. fragilis*, are also opportunistic pathogens. It is well documented that many bacterial virulence-factor genes are located within genomic islands (Juhas *et al.*, 2009). The clustering of potential pore-forming outer-membrane toxins in the *B. thetaiotaomicron* genome suggest that this region could be a pathogenicity island acquired through horizontal gene transfer, as predicted by a genome-wide genomic islands study (Ho Sui *et al.*, 2009).

Although the functions of the MACPFs represented by Bth-MACPF remain to be elucidated, our study provided clues that they are important targets for further exploration of how symbiotic microbes adapt to and influence their host environments. Additional information about the proteins described in this study is available from TOPSAN (Krishna *et al.*, 2010) at http://www.topsan.org/explore?PDBid=3kk7.

This work was supported by the NIH, National Institute of General Medical Sciences, Protein Structure Initiative grant U54 GM074898. Portions of this research were carried out at the Stanford Synchrotron Radiation Lightsource (SSRL). The SSRL is a national user facility operated by Stanford University on behalf of the US DOE, OBES. The SSRL Structural Molecular Biology Program is supported by the DOE, OBER and by NIH (NCRR, BTP and NIGMS). Genomic DNA from *B. thetaiotaomicron* VPI-5482 (ATCC No. 29148D-5) was obtained from the American Type Culture Collection (ATCC). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. & Eddy, S. R. (2004). *Nucleic Acids Res.* **32**, D138–D141.
Beitz, E. (2000). *Bioinformatics*, **16**, 135–139.
Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. & Paciorek, W. (2003). *Acta Cryst.* D**59**, 2023–2030.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *J. Appl. Cryst.* **35**, 720–726.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
Das, D. *et al.* (2010). *Acta Cryst.* F**66**, 1265–1273.
Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). *Nature Protoc.* **2**, 953–971.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M. & Brinkman, F. S. (2005). *Bioinformatics*, **21**, 617–623.
Hadders, M. A., Beringer, D. X. & Gros, P. (2007). *Science*, **317**, 1552–1554.
Hayashi, S. & Wu, H. C. (1990). *J. Bioenerg. Biomembr.* **22**, 451–471.
Hildebrand, A., Remmert, M., Biegert, A. & Soding, J. (2009). *Proteins*, **77**, *Suppl.* 9, 128–132.
Ho Sui, S. J., Fedynak, A., Hsiao, W. W., Langille, M. G. & Brinkman, F. S. (2009). *PLoS One*, **4**, e8094.
Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.
Hooper, L. V. & Gordon, J. I. (2001). *Science*, **292**, 1115–1118.
Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M. & Mitaku, S. (2008). *Bioinformation*, **2**, 417–421.
Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. (2005). *Nucleic Acids Res.* **33**, W284–W288.
Jones, S., Marin, A. & Thornton, J. M. (2000). *Protein Eng.* **13**, 77–82.
Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W. & Crook, D. W. (2009). *FEMS Microbiol. Rev.* **33**, 376–393.
Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. & Krogh, A. (2003). *Protein Sci.* **12**, 1652–1662.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kafsack, B. F., Pena, J. D., Coppens, I., Ravindran, S., Boothroyd, J. C. & Carruthers, V. B. (2009). *Science*, **323**, 530–533.
Klock, H. E., Koesema, E. J., Knuth, M. W. & Lesley, S. A. (2008). *Proteins*, **71**, 982–994.
Kohm, A. P., Fuller, K. G. & Miller, S. D. (2003). *Trends Microbiol.* **11**, 101–105.
Krishna, S. S., Weekes, D., Bakolitsa, C., Elsliger, M.-A., Wilson, I. A., Godzik, A. & Wooley, J. (2010). *Acta Cryst.* F**66**, 1143–1147.
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Luz, J. G., Yu, M., Su, Y., Wu, Z., Zhou, Z., Sun, R. & Wilson, I. A. (2005). *J. Mol. Biol.* **352**, 1019–1028.
Malik, H. S. & Eickbush, T. H. (2001). *Genome Res.* **11**, 1187–1197.
Mancheno, J. M., Tateno, H., Goldstein, I. J., Martinez-Ripoll, M. & Hermoso, J. A. (2005). *J. Biol. Chem.* **280**, 17251–17259.
Ponting, C. P. (1999). *Curr. Biol.* **9**, R911–R913.
Ramachandran, R., Tweten, R. K. & Johnson, A. E. (2004). *Nature Struct. Mol. Biol.* **11**, 697–705.
Ren, J., Bird, L. E., Chamberlain, P. P., Stewart-Jones, G. B., Stuart, D. I. & Stammers, D. K. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 14410–14415.
Rosado, C. J. *et al.* (2007). *Science*, **317**, 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. & Dunstone, M. A. (2008). *Cell. Microbiol.* **10**, 1765–1774.
Rossjohn, J., Feil, S. C., McKinstry, W. J., Tweten, R. K. & Parker, M. W. (1997). *Cell*, **89**, 685–692.
Santarsiero, B. D., Yegian, D. T., Lee, C. C., Spraggon, G., Gu, J., Scheibe, D., Uber, D. C., Cornell, E. W., Nordmeyer, R. A., Kolbe, W. F., Jin, J., Jones, A. L., Jaklevic, J. M., Schultz, P. G. & Stevens, R. C. (2002). *J. Appl. Cryst.* **35**, 278–281.
Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. (2005). *Nucleic Acids Res.* **33**, W363–W367.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Shoji, M., Naito, M., Yukitake, H., Sato, K., Sakai, E., Ohara, N. & Nakayama, K. (2004). *Mol. Microbiol.* **52**, 1513–1525.
Slade, D. J., Lovelace, L. L., Chruszcz, M., Minor, W., Lebioda, L. & Sodetz, J. M. (2008). *J. Mol. Biol.* **379**, 331–342.
Soding, J., Biegert, A. & Lupas, A. N. (2005). *Nucleic Acids Res.* **33**, W244–W248.
Stebbins, C. E. & Galan, J. E. (2001). *Nature (London)*, **412**, 701–705.
Tokuda, H. (2009). *Biosci. Biotechnol. Biochem.* **73**, 465–473.
Tweten, R. K. (2005). *Infect. Immun.* **73**, 6199–6209.
Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1993). *J. Mol. Biol.* **229**, 105–124.
Voskoboinik, I., Smyth, M. J. & Trapani, J. A. (2006). *Nature Rev. Immunol.* **6**, 940–952.
Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.
Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Methods Enzymol.* **374**, 300–321.
Wolf, Y. I., Aravind, L. & Koonin, E. V. (1999). *Trends Genet.* **15**, 173–175.
Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. & Gordon, J. I. (2003). *Science*, **299**, 2074–2076.
Xu, Q. *et al.* (2010). *Acta Cryst.* F**66**, 1281–1286.