Clearing the confusion regarding the putative "membrane coat" proteins in Planctomycetes and Verrucomicrobia

Supplementary Material to "Planctomycetes and eukaryotes: a case of analogy not homology" by James O. McInerney[1]*, William F. Martin[2], Eugene V. Koonin[3], John F. Allen[4], Michael Y. Galperin[3], Nick Lane[5], John M. Archibald[6], and T. Martin Embley[7]

In the paper by Santarella-Mellwig et al. "The compartmentalized bacteria of the Planctomycetes -Verrucomicrobia-Chlamydiae superphylum have membrane coat-like proteins", published in PLoS Biol. 2010, 8(1):e1000281, the authors report finding a family of bacterial proteins (exemplified by *Gemmata obscuriglobus* gp4978*)* with the same domain architecture (β-propeller and α-helical repeat domains) as eukaryotic membrane coat (MC) proteins and propose that MC proteins have evolved from gp4978-like proteins.

We see the following problems with this analysis:

1. Santarella-Mellwig et al. analyzed truncated proteins; domain architecture of full-length gp4978 is more complex than represented in their Fig. 2.
2. Structures of α-helical repeats in gp4978 and its bacterial homologs (HEAT repeats) are different from α-helical repeats (CLTH repeats) in eukaryotic membrane coat proteins.
3. In addition to the members of the PVC superphylum, close homologs of gp4978 are found in multiple representatives of the phylum Bacteroidetes.

1. Santarella-Mellwig et al. analyzed truncated proteins; domain architecture of full-length gp4978 is more complex than represented in their Fig. 2.

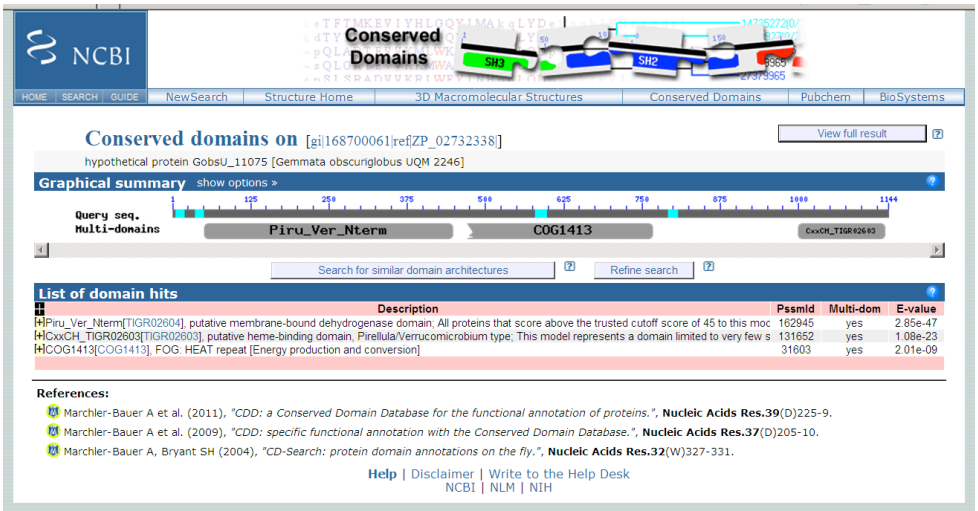
Neither one of the bacterial protein identifiers shown in Fig. 2 of Santarella-Mellwig et al. (gp4978_Gemmata, rb4028_Rhodopirellula, cf3358_Chthoniobacter, La1688_Lentisphaera) corresponds to an entry in any of the public sequence databases (GenBank\ENA\DDBJ, UniProt, UniParc, NCBI protein database; CF3358 is a partial 16S rRNA sequence). The only indication of the sequences used by Santarella-Mellwig et al. is a 362-aa partial sequence of gp4978 given in the Supplementary File Text S1. **A** BLAST search using this sequence as query found a 100% identical hit to a hypothetical protein GobsU_11075 from *Gemmata obscuriglobus* UQM 2246 (gi|168700061|ref|ZP_02732338.1|, length=1144), see below:

>gi|168700061|ref|ZP_02732338.1| hypothetical protein GobsU_11075 [Gemmata obscuriglobus UQM 2246] Length=1144

```
 Score =  757 bits (1954),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 362/405 (89%), Positives = 362/405 (89%), Gaps = 43/405 (11%)

Query  1    KAADKGLKVEVWASAPTMANPVSFCFDEKGKCYVAETTRF--------------------  40
            KAADKGLKVEVWASAPTMANPVSFCFDEKGKCYVAETTRF
Sbjct  46   KAADKGLKVEVWASAPTMANPVSFCFDEKGKCYVAETTRFENGVPDTRGHMKWLDEDLAN  105
```

```
Query  41   ------------------------QVRVVWSSNGRGPADKSEVFSGGYNRPQDGLAAGVLA  77
                                     QVRVVWSSNGRGPADKSEVFSGGYNRPQDGLAAGVLA
Sbjct  106  RSIDDLLKMYNKHNYKGYEKYSDQVRVVWSSNGRGPADKSEVFSGGYNRPQDGLAAGVLA  165

Query  78   RKGSVYFTCIPDLYQLKDTNGDGKADEKKSLFTGFGPTVQFLGHDLHGLRMGPDGKLYFS  137
            RKGSVYFTCIPDLYQLKDTNGDGKADEKKSLFTGFGPTVQFLGHDLHGLRMGPDGKLYFS
Sbjct  166  RKGSVYFTCIPDLYQLKDTNGDGKADEKKSLFTGFGPTVQFLGHDLHGLRMGPDGKLYFS  225

Query  138  VGDRGFMVTTKEGKKLEYPNTGAVLRCDPDGANLEVVHSGLRNPQEIAFDDFGNLFTYDN  197
            VGDRGFMVTTKEGKKLEYPNTGAVLRCDPDGANLEVVHSGLRNPQEIAFDDFGNLFTYDN
Sbjct  226  VGDRGFMVTTKEGKKLEYPNTGAVLRCDPDGANLEVVHSGLRNPQEIAFDDFGNLFTYDN  285

Query  198  NCDSGDRARWVHIVEGGDSGWRGGFQYSTGYHTPEVPQGNRGAWNTEKLWHTQHEGQPAW  257
            NCDSGDRARWVHIVEGGDSGWRGGFQYSTGYHTPEVPQGNRGAWNTEKLWHTQHEGQPAW
Sbjct  286  NCDSGDRARWVHIVEGGDSGWRGGFQYSTGYHTPEVPQGNRGAWNTEKLWHTQHEGQPAW  345

Query  258  IVPPLLHLGNGPAGITHYPGIGLNDKYKDHFFACDFTSSAGSSVIWAVSVKPKGASFEVQ  317
            IVPPLLHLGNGPAGITHYPGIGLNDKYKDHFFACDFTSSAGSSVIWAVSVKPKGASFEVQ
Sbjct  346  IVPPLLHLGNGPAGITHYPGIGLNDKYKDHFFACDFTSSAGSSVIWAVSVKPKGASFEVQ  405

Query  318  KPEPFLRGMVPTDCEFGPDGAFYWSDWVGGWAPQNRGRIFRVTDA  362
            KPEPFLRGMVPTDCEFGPDGAFYWSDWVGGWAPQNRGRIFRVTDA
Sbjct  406  KPEPFLRGMVPTDCEFGPDGAFYWSDWVGGWAPQNRGRIFRVTDA  450
```

It is obvious that the 362-aa ORF gp4978 occupies positions 46-85 and 129-450 of the full-length GobsU_11075 protein. Accordingly, structural predictions by Santarella-Mellwig et al. (Fig. 2 of their paper) relate only to the first 450 amino acid residues of the domain architecture of the GobsU_11075 protein (available, e.g. in the NCBI's Conserved Domain Database, http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?INPUT_TYPE=live&SEQUENCE=168700061), which is shown below. In other words, the combination of a beta-propeller domain and an a-helical repeat domain, which the authors claim to be the hallmark of the PVC gp4978-like family, is observed only in truncated proteins. Full length-proteins of that family contain an additional heme-binding (TIGR2603) domain.



2. Structures of α-helical repeats in gp4978 and its bacterial homologs (HEAT repeats) are different from α-helical repeats in eukaryotic membrane coat proteins

In the absence of statistically significant sequence similarity between eukaryotic MC proteins and any known bacterial proteins, Santarella-Mellwig et al. (2010) make the case for homology (common origin) of the MC proteins and gp4978-like proteins based on the similarity of the structural elements (β-propeller followed by α-helical repeat segments) in both protein families. As explained above, this similarity has been observed primarily in the truncated proteins; full-length members of the gp4978-like protein family contain an additional C-terminal domain. In addition, although Santarella-Mellwig et al. (2010) are correct that both protein families contain α-helical repeat segments, it is important to note that these α-helical repeats are substantially different. The α-helical repeats in gp4978-like proteins represent the HEAT repeat family (Pfam domain PF02985, PDB entry 1b3u),), whereas the structure of nucleoporin C-terminal domain (PF07575, PDB entry 2qx5), although also α-helical, is far less regular. As a result, Schwartz and colleagues, while recognizing common ancestry of the nuclear pore complex and vesicle coats, argued that these structures are not related to the HEAT\ARM repeats (PMID: 17897938, 18974315). In any case, the repetitive nature of this fold makes evolutionary inferences extremely difficult: both families have evolved through multiple cycles of duplication from simple α-helical hairpins. Obviously, such remote structural similarity cannot be used as evidence of homology.

3. In addition to the members of the PVC superphylum, close homologs of gp4978 are found in multiple representatives of the phylum Bacteroidetes.

The following is a list of the best BLAST hits (cut-off E-value, $1x10^{-6}$, >50% query coverage) for the 362-aa gp4978 fragment of the GobsU_11075 protein, modeled by Santarella-Mellwig et al., is attached as an HTML (or PDF) file. Each protein is hyperlinked to its entry in the NCBI protein database, BLAST Link (BLink) output, and its domain architecture according to the CDD database. Also provided, where available, are links to the respective entries in UniProt, Pfam and InterPro databases. This listing clearly shows that, in addition to the representatives of the Planctomysetes, Verrucomicrobia, and Lentisphaerae (members of the PVC superphylum), close homologs of gp4978 are found, often in several copies, in multiple representatives of the phylum Bacteroidetes (see below). This phylum does not appear to be related to the PVC superphylum and Santarella-Mellwig et al. (2010) never suggested it to be a potential ancestor of eukaryotic proteins. These organisms belong to several distinct lineages within the Bacteroidetes, which makes horizontal gene transfer of gp4978-related genes from PVC members extremely unlikely.

In summary, the conclusion of Santarella-Mellwig et al. (2010) that eukaryotic MC proteins have evolved from a single family of PVC-specific gp4978-like proteins contradicts the available sequence, structure, and phylogenetic data.

Phylogenetic distribution of putative "membrane coat-like" proteins (domain architecture as in *Gemmata obscuriglobus* gp4978/GobsU_11075)

Planctomycetes

. Gemmata obscuriglobus UQM 2246	12 hits
. Planctomyces maris DSM 8797	18 hits
. Planctomyces limnophilus DSM 3776	11 hits
. Planctomyces brasiliensis DSM 5305	11 hits
. Blastopirellula marina DSM 3645	17 hits
. Pirellula staleyi DSM 6068	15 hits
. Rhodopirellula baltica WH47	14 hits
. Rhodopirellula baltica SH 1	13 hits
. Isosphaera pallida ATCC 43644	10 hits

Lentisphaerae

. Lentisphaera araneosa HTCC2155	19 hits

Verrucomicrobia

. Verrucomicrobium spinosum DSM 4136	23 hits
. Chthoniobacter flavus Ellin428	15 hits
. Pedosphaera parvula Ellin514	14 hits
. Verrucomicrobiae bacterium DG1235	2 hits

Bacteroidetes

. Dyadobacter fermentans DSM 18053	14 hits
. Algoriphagus sp. PR1	8 hits
. Spirosoma linguale DSM 74	8 hits
. Haliscomenobacter hydrossis DSM 1100	7 hits
. Leadbetterella byssophila DSM 17132	3 hits
. Maribacter sp. HTCC2170	3 hits
. Robiginitalea biformata HTCC2501	2 hits
. Chitinophaga pinensis DSM 2588	1 hit
. Pedobacter heparinus DSM 2366	1 hit