Outer membrane 1 pore protein prediction in mycobacteria using genomic comparison

Nancy Mah, Carolina Perez-Iratxeta, Miguel A. Andrade-Navarro

Published in final edited form as:
Microbiology. 2010 Aug ; 156 (8): 2506-2515 | doi: 10.1099/mic.0.040089-0
The Society for General Microbiology

Outer membrane pore protein prediction in mycobacteria using genomic comparison

Nancy Mah [1], Carolina Perez-Iratxeta [2], Miguel A. Andrade-Navarro [1]

[1]Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany

[2]Ottawa Health Research Institute, 501 Smyth Road, Ottawa, Ontario K1H 8L6, Canada

Correspondence: Nancy Mah, Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany; Telephone: +49 30 9406-2831; Fax: +49 30 9406-4240; E-mail: nancy.mah@mdc-berlin.de

Running title: outer membrane pore protein prediction in mycobacteria

Contents category: Research Paper (Microbial Pathogenicity)

Abbreviations: OMP, outer membrane protein; Mt, *Mycobacterium tuberculosis*; Mb, *Mycobacterium bovis*; Ml, *Mycobacterium leprae*; Mu, *Mycobacterium ulcerans*; Ms, *Mycobacterium smegmatis*, Ma, *Mycobacterium avium*; Mm, *Mycobacterium marinum*, aa, amino acid; TM, transmembrane helix

Summary

Proteins responsible for outer membrane transport across the unique membrane structure of *Mycobacterium* spp. are attractive drug targets in the treatment of human diseases caused by the mycobacterial pathogens, *M. tuberculosis*, *M. bovis*, *M. leprae* and *M. ulcerans*. In contrast to *E. coli*, relatively few outer membrane proteins (OMPs) have been identified in *Mycobacterium* spp., largely due to the difficulties in isolating mycobacterial membrane proteins and our incomplete understanding of secretion mechanisms and cell wall structure in these organisms. To further expand our knowledge of these elusive proteins in Mycobacterium, we have improved upon our previous method of OMP prediction in mycobacteria by taking advantage of genomic data from seven mycobacteria species. Our improved algorithm suggests 4333 sequences as putative OMPs in these seven species with varying degrees of confidence. The most virulent pathogenic mycobacterial species are slightly enriched in these selected sequences. We present examples of predicted OMPs involved in horizontal transfer and paralogy expansion. Analysis of local secondary structure content allowed identifying small domains predicted to perform as OMPs; some examples show their involvement in events of tandem duplication and domain rearrangements. We discuss the taxonomic distribution of these discovered families and architectures, often specific to mycobacteria or the wider taxonomic class of Actinobacteria. Our results suggest that OMP functionality in mycobacteria is richer than expected and provide a resource to guide future research of these understudied proteins.

Introduction

Mycobacteria are responsible for some of the most terrible human diseases including leprosy and tuberculosis (Cosma *et al.*, 2003). However, not all mycobacteria are pathogenic to humans despite their considerable genomic similarity. Part of their variable properties in infective ability and specificity are likely related to their variable cell wall (Brennan & Nikaido, 1995). Outer membrane proteins (OMPs) are an important component of the mycobacterial cell wall, yet they are poorly studied in mycobacteria (Niederweis *et al.*, 2010). OMPs are transmembrane proteins that form a beta-barrel structure consisting of amphipathic beta strands and are secreted into the periplasmic space and inserted into the outer membrane to act as channels (Faller *et al.*, 2004). This type of protein is therefore an important target for antibacterial therapy and an object of study to the elucidation of the mechanisms of pathogenicity (Niederweis, 2008). However, currently there is evidence of just some mycobacterial OMPs, in large degree for MspA (Stahl *et al.*, 2001), and less for another two Mt proteins: Rv1973 (Song *et al.*, 2008) and Rv1698 (Siroy *et al.*, 2008; Song *et al.*, 2008). This is not only due to the difficulties of culturing mycobacteria, but also to the difficulty of identifying these proteins.

Computational methods of OMP detection have been developed and applied to *Mycobacterium tuberculosis* with relative success (Pajon *et al.*, 2006; Song *et al.*, 2008), but there is room for improvement, especially when it comes to prioritizing targets for research. The increasing number of related mycobacterial genomes offers a unique opportunity to support the predictions by addressing their coherence across orthologs and to pinpoint OMP families specific to pathogenic mycobacteria.

In this work we accomplished parameter optimization of a previous method that predicts OMPs based solely on their potential to be secreted and to form an amphiphilic beta-barrel (Song *et al.*, 2008). We explored the predictive power of a set of OMP-related properties by contrasting the robustness of the results on the complete proteomes for seven mycobacteria: three obligate pathogens *M. tuberculosis*, *M. bovis* and *M. leprae*, two facultative pathogens *M. marinum* and *M. ulcerans*, one opportunistic pathogen *M. avium*, and the non-pathogenic *M. smegmatis*.

The relatedness between these species and their pathogenic properties are heterogeneous. The closest genomes by far are those of *M. tuberculosis* and *M. bovis*, but their host ranges are different (*M. bovis* can cause tuberculosis in several mammals, whereas the natural hosts of *M. tuberculosis* are humans). Therefore, both genomes were included in the analysis since we considered that a comparison of OMPs between these two species can lead to insights that help to explain the differences in ability to infect different hosts.

The predictions on these seven genomes directed us to a number of OMP domains present in mycobacteria, most of them exclusive to actinobacteria and without homologs in eukaryotes.

Methods

Calculation of parameters for outer membrane protein prediction in mycobacteria

Protein sequences were obtained for seven mycobacterial genomes (Table 1) [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/], including: *Mycobacterium avium* 104, *Mycobacterium bovis* AF2122/97, *Mycobacterium leprae* TN, *Mycobacterium marinum* M, *Mycobacterium smegmatis* str. MC2 155, *Mycobacterium tuberculosis* H37Rv, and *Mycobacterium ulcerans* Agy99. All proteins were predicted to be OMPs based on the two main properties: 1) the ability to be secreted to the outer membrane; and 2) the ability to form beta-barrel structures.

Secreted proteins were predicted according to both classical secretion mechanisms (SignalP-v3.0) and non-classical secretion such as twin arginine translocation (TatP-v1.0) or leaderless secretion (SecretomeP-v1.0) (Bendtsen *et al.*, 2004a; Bendtsen *et al.*, 2004b; Bendtsen *et al.*, 2005). Prediction of classically secreted proteins is well-studied and most secreted bacterial proteins are exported in this manner (Malen *et al.*, 2007); therefore predictions determined by SignalP-v3.0 algorithm or the presence of a single predicted transmembrane alpha helix in the first 70 aa of protein (TMHMM; (Krogh *et al.*, 2001)) were considered to be superior to other prediction methods. TatP prediction is specific to bacteria and was considered to be next most reliable. SecretomeP prediction was the least specific of all methods, but was still considered to be useful as we are aiming for high recall.

To demonstrate the efficacy of the secretion prediction methods, the algorithms were run on positive and negative validation sets. A set of 53 experimentally verified, classically secreted mycobacterial proteins was obtained from Leversen and co-workers (Leversen *et al.*, 2009). A set of non-classically secreted mycobacterial proteins, including proteins secreted by Tat or SecA2 systems, was assembled by literature search. Negative controls for secretion were represented by 1725 reviewed cytoplasmic proteins from Mycobacterium sp. (UniProtKB release 15.3).

In the second step of the OMP prediction, the propensity of the proteins to form beta-barrel structures was determined using various beta-strand properties. Secondary structure for all proteins was predicted using Jnet v1.0 (Cuff & Barton, 1999; Cuff & Barton, 2000). Beta-strands of 5 or more residues (B5 strands) were evaluated for amphiphilicity (FracB5) as previously described (Song *et al.*, 2008). As additional measures of 'betaness', the overall proportion of residues in beta strands (PercentBeta), the number of B5 strands (numB5) and the total number of residues in B5 strands (numB5res) were recorded. Positive and negative control sets for beta-barrels were used to demonstrate the utility of these parameters to predict beta-barrel structure. Positive validation of beta-barrel structure was represented by 428 bacterial and eukaryotic sequences taken from proteins containing Pfam motifs or bacterial sequences with solved 3D structures annotated as forming a beta-barrel. Negative validation of beta sheet prediction consisted of 90 actinobacterial proteins with solved 3D structure of low beta content. Protein structures were obtained from the Protein Data Bank (PDB, www.pdb.org).

Additional parameters, including the number of cysteine residues (numcys) and the isoelectric point (pI) (BioPerl pICalculator, www.bioperl.org) were also evaluated. Initial OMP prediction (Method 1) was carried out with similar parameters and thresholds as previously used by Song et al. (Song *et al.*, 2008), namely: FracB5 ≥ 0.19, PercentBeta ≥ 0.10, Smean (from the signal peptide predictor SignalP) ≥ 0.50 or numpredhel (number of predicted transmembrane helices) = 1 and firsthel (position in amino acids of the most N-terminal predicted transmembrane helix) ≤ 70 (Supplementary Table S1). Here, the OMP prediction method was further refined by using sequence homology information and optimization of the algorithm as described below.

Clustering of Mycobacterial sequences

Clustering of the seven mycobacterium genomes listed in Table 1 was carried out using a very strict sequence similarity criterion that enforces all proteins in a cluster to be homologous along their full lengths, ensuring their domain content and architectures are equivalent (Perez-Iratxeta *et al.*, 2007). About 1% of the sequences were not used because they were too short (< 50 aa). The remaining sequences (30,605) were distributed into 11,633 clusters. Pfam motif information (Finn *et al.*, 2008) of representative members of the clusters was retrieved to further characterize them.

Optimization of OMP prediction

To optimize the OMP prediction, two training sets of clusters (with at least five sequences each) were defined: 1) OMP-rich clusters which contained ≥ 80% OMP-predicted sequences (1151 sequences in 168 clusters); and 2) OMP-poor clusters, which contained ≤ 20% OMP-predicted sequences (981 sequences in 96 clusters). First, optimization was carried out by applying varying thresholds for OMP prediction on parameters that were not included in Method 1, namely pI, sequence length, number of cysteines, numB5, numB5res, Tat-secretion score, and leaderless secretion score. Next, the Method 1 thresholds on parameters including frac, PerBeta, and Smean were optimized. The optimal cutoff values were defined to be those that most reduced the fraction of predicted OMPs in OMP-poor clusters while retaining over 90% of the predicted OMPs in OMP-rich clusters. The new set of optimized criteria was called Method 2 and used for the remainder of the analysis.

A scoring system (producing a score ranging from zero to 16) was used to indicate the confidence of the OMP prediction. Propensity for secretion and beta-barrel formation were given equal weight (maximum of 8 points each). Proteins predicted to be secreted by general secretion or Tat mechanisms were awarded 8 points. In the absence of these two predictions, proteins were awarded 3 points if the SecretomeP score ≥ 0.574. Beta-sheet related parameters were assessed on: 1) the entire protein length; and 2) within a 300 aa sliding window of protein sequence to detect local regions of high beta content. A window size of 300 aa was chosen because this is the size of the beta barrel domain in some known OMP structures (Faller *et al.*, 2004; Song *et al.*, 1996); however, such domains can be formed by association of beta strands from multiple protein monomers, and it can happen that an OMP protein or region is much smaller than 300 aa. One point was awarded for each of the following criteria satisfied as frac ≥ 0.28, PerBeta ≥ 0.11, numB5 ≥ 3, numB5res ≥ 17, for a maximum of 8 points.

Results

Optimization of OMP prediction

To optimize our predictions in seven mycobacterial genomes, we clustered their protein sequences and evaluated the fraction of proteins in each cluster predicted as OMP (see Methods). Contrasting the parameters of OMP-rich and OMP-poor clusters enabled us to further refine the OMP criteria, based on the assumption that all proteins in a cluster should be predicted either as being OMPs or not; and therefore, that OMP-poor clusters indirectly reflected false positives.

In the first instance, we tested nine parameters that could play a minor role in the initial prediction of OMPs but were not previously used. The fraction of OMP predicted sequences in OMP-rich and OMP-poor groups was monitored when applying increasingly restrictive thresholds in each of these nine parameters (Supplementary Fig. S1). Four parameters (pI, number of cysteines, first helix, dvalue score from TatP) showed little improvement in reducing the number of potential false negatives in the OMP-poor group of clusters. Potential improvements could be made for the remaining five parameters: sequence length, number of residues in B5 sheets, number of B5 sheets, number of predicted TM helices, and nnscore (prediction of leaderless score from SecretomeP).

Sequence length was rejected as a constraining factor, as it was undesirable to eliminate short (~100 aa) predicted OMPs potentially composing homologous multimeric structures. Requiring the number of B5 sheets and the number of residues in B5 sheets to be a minimum of 3 and 17, respectively, was successful in reducing the number of OMPs in the OMP-poor clusters by >10%, while keeping 96% of the OMPs in the OMP-rich clusters (Supplementary Fig. S1). A rather stringent threshold was used for prediction of leaderless secretion. At nnscore >= 0.71, 56% of sequences from the OMP-rich clusters were retained, while rejecting 74% of sequences from the OMP-poor clusters as targets for leaderless secretion. This was not considered to be overly stringent, since proteins with signal sequences, (which would be identified by SignalP-3.0 or TatP- 1.0) were likely to have a high nnscore anyway (Bendtsen *et al.*, 2004a).

In the second instance, the five parameters initially used to predict OMPs were varied and the fraction of predicted OMPs was recorded. This analysis suggested fine adjustments in the cutoff values for amphiphilicity (frac), proportion of residues in beta strands (perbeta), and general secretion score (Smean) (Supplementary Fig. S2). The optimized criteria were applied to the dataset, expanding the number of predicted OMPs, compared to the original method (Tables 1 and Supplementary Table S1). Up to this point, the optimization had been carried out by varying parameters on an individual basis. A scoring system was implemented to summarize the effect of all the optimized parameters (with 8 points for beta-barrel formation and 8 points for secretion, see Methods). Assuming that most of the sequences in OMP-rich clusters should actually be OMPs, and that sequences in OMP-poor clusters should not be OMPs, a threshold of OMP score = 12 to accept an OMP prediction was found to be optimal (Fig. 1; Supplementary Fig. S3(a)). At this threshold, 94% of sequences from OMP-rich clusters are classified as OMPs, while 89% of the sequences in the OMP-poor clusters are rejected as OMPs.

Validation of signal sequence and beta-barrel prediction in mycobacteria

None of the secretion prediction programs were specifically designed to predict signal sequences in *Mycobacterium* spp., although the SignalP neural net predictions were based on Gram-positive bacteria, and the TatP server was trained on both Gram-negative and Gram-positive sequences. Mycobacteria are classified as Gram-positive, despite the fact that the mycobacterial outer membrane has distinct properties not found in either functionally classified Gram-negative nor Gram-positive bacteria (Hett & Rubin, 2008). Therefore, it was important to test these algorithms for their ability to detect signal sequences in mycobacteria.

Using known cytoplasmic and known mycobacterial proteins secreted by the general secretory pathway, it could be shown that the optimized cutoff (Smean = 0.54) was sufficient to correctly predict secretion in 93% of the known GSP proteins and reject 98% of the cytoplasmic proteins (Supplementary Fig. S3(b)). Secretion by the non-classical Tat system could be predicted at a cutoff of dvalue = 0.36 in 79% of the known Tat-secreted mycobacterial sequences, while rejecting 98% of the cytoplasmic proteins for secretion (Supplementary Fig. S3(c)).

Prediction of leaderless secretion in mycobacteria at the chosen cutoff nnscore ≥ 0.71 correctly identified 50% (6/12) known leaderless secreted proteins (Supplementary Fig. S3(d)), which included secreted proteins by the recently described bacterial export system ESX-1 and the SecA2 (Sec-independent) system. 81% of the known cytoplasmic proteins were predicted as being secreted in this instance, making the leaderless secretion prediction the least precise of all three secretion prediction methods. As a result, less emphasis was placed on leaderless secretion scores in the OMP prediction.

Prediction of beta-barrel structures was based on beta-sheet content and the amphiphilicity of predicted beta strands (computed as in (Song *et al.*, 2008)). As a measure of the protein's propensity to form beta-barrel structures, beta-barrel scores were calculated globally (over whole protein) and locally (sliding window) for a maximum score of 8 (see Methods). For a beta-barrel score ≥ 6, a total of 97% of known bacterial OMP and 90% of annotated beta-barrel proteins were correctly identified as containing beta-barrels, whereas non beta-barrel structures were predicted in 74% of solved sequences lacking certain beta-barrel structure (Fig. 2).

After the optimization stage, Method 2 was able to correctly identify 90% (27/30) of known bacterial OMPs with high scores (score ≥14; Supplementary Fig. S4) corresponding to strong predictions. Among the three OMPs missed by Method 2, there were two from *Rhodobacter* spp. Although one of them (PORI_RHOBL) contained sufficient beta structure for a beta-barrel, it was predicted to be secreted by leaderless secretion resulting in a weak OMP prediction (OMP score = 11). OmpG from *E. coli* was as well not identified as OMP by this method, due to a lack of beta-strand prediction from Jnet v1.0.

The selection criteria of both the optimized Method 2 and the previous method show substantial differences between the results (Supplemental Table S1). When applying them to seven mycobacterial genomes (see Methods) 3340 proteins are predicted to be OMPs by both methods. A total of 993 proteins are newly predicted by Method 2 whereas 406 proteins predicted to be OMPs by Method 1, are now rejected as false positives.

Identification of OMPs

Table 1 specifies the number of sequences and predicted OMPs. The seven genomes analysed have a genomic size in the 4000-5000 range except for the small Ml genome (1605 genes, an extreme case of genome downsizing (Cole *et al.*, 2001)) and the larger Ms (6,716 genes). When considering the percentage of predicted OMPs it is interesting to note that the three obligate pathogenic organisms (Mt, Mb, and Ml) have the largest percentage (15.1 -15.8%) while the opportunistic pathogen Ma and the non-pathogenic species Ms have the lowest values (12.5 -12.6%).

We showcase the results of our method with some examples of newly OMP-predicted mycobacterial proteins. Because the taxonomic distribution of a protein can give an indication of its functional relevance, we have categorized the examples by this property. Our selection of examples was facilitated by the clustering used for the optimization of the method, e.g. when searching for OMPs present in all seven mycobacteria. The complete results are available in Supplementary Table S2.

OMPs present in Mt but not in Mb

Though Mt and Mb are very closely related (they both belong to the Mycobacterium tuberculosis complex – Mt complex) and share many 100% identical proteins, they have obvious differences regarding pathogenicity. It is therefore interesting to find OMPs in Mt that have no equivalent in Mb. Two outstanding examples are Mt proteins mce3c (Rv1968; 410 aa) and mce3e (Rv1970; 377 aa) encoded by two of six genes from the putative mce3 operon Mce3A-F. Both Mce3C and Mce3E proteins were found to react with antibodies from serum of TB patients (Ahmad *et al.*, 2004) and have one predicted MCE domain each (at positions 38-114 and 36-112, respectively).

The presence of the MCE domain in these sequences is relevant because many of the genes with this domain have been shown to be expressed during natural infection of Mt and it is thought that they are related to mycobacterial pathogenicity (Ahmad *et al.*, 1999). Mt has a total of 24 genes with this domain arranged in four mce operons, which contain two integral membrane proteins followed by six genes with the MCE domain (Cole *et al.*, 1998). In our previous work (Song *et al.*, 2008) we predicted that 23/24 of Mt MCE genes were OMPs. The present method predicts all 24 MCE genes in Mt H37Rv as OMPs with a score of 16 (the maximum possible). We speculate that the MCE domain is actually a beta-barrel characteristic of OMPs, which is coherent with their proposed role at the mycobacterial cell surface (Flesselles *et al.*, 1999).

Present in the Mt complex but not in all mycobacteria

We observed many OMP clusters with members in Mt/Mb but missing in all or some of the five species outside the Mt complex. This can be either due to genes being invented (or horizontally transferred) within the mycobacteria lineage, or to selective gene loss (as in the massive pseudogenization that occurred in the Ml genome (Cole *et al.*, 2001)). Here we show examples of each of those.

An OMP unique to Mt/Mb

Rv1351 is a Mt 109 aa protein that we predict to be an OMP. The Mb ortholog is 100% identical, and there are no homologous sequences in the other five mycobacteria analysed in this work, or outside mycobacteria (no PSIBLAST hits with E-value below 8.3). We also predict that the gene next to it, Rv1352 (encoding a 123 aa protein), is also a small OMP. According to the STRING database (Jensen *et al.*, 2009), these two predicted OMPs are in an operon conserved between Mt/Mb, which includes upstream genes Rv1348 and Rv1349 (two uncharacterized ABC transporter ATP-binding proteins) and fabG2/Rv1350 (predicted as 3-ketoacyl-(acyl-carrier-protein) reductase). Therefore, these two predicted OMPs seem to form part of an Mt/Mb specific operon and although they are rather small they could form a barrel by multimerization, which would explain the need to co-express them in an operon. Such operon could carry out a function inherent to the Mt complex. The three genes, Rv1348, Rv1349 and Rv1350, are essential genes for growth of Mt as determined by Sassetti et al. (Sassetti *et al.*, 2003).

A mycobacterial OMP with horizontal transfer

Mt Rv1914c (135 aa) is predicted as an OMP with orthologs in Mb/Mu/Mm but without apparent equivalents in Ml/Ma/Ms. Curiously, the only match in the database outside mycobacteria is a very clear hit (>50% identity) on a distant bacteria, Proteobacteria *Geobacter uraniireducens* (sequence GI:148265072, 135 aa). This suggests an event of horizontal transfer of this gene between mycobacteria and geobacteria. One could speculate that the function of this OMP would not be associated to pathogenicity given its presence both in pathogenic and non-pathogenic mycobacteria (and in *G. uraniireducens*). Incidentally, Rv1914c was one of 224 genes found to be deleted in one or more clinical isolates of a H37Rv strain from San Francisco (Tsolaki *et al.*, 2004).

C4: a novel putative OMP domain that occurs as a tandem repeat.

Mt Rv2270 (175 aa) defines a family with orthologs in five of seven mycobacteria tested (missing in Ms/Ml) and corynebacteria. This implies that the gene was invented in Corynebacterineae and that there was a selective loss of this gene within some members of the mycobacteria lineage indicating that it is not essential for them.

Sequence analysis indicated that the family contains a C-terminal 120 aa domain (that we termed C4 for its conserved four cysteines, see Supplementary Fig. S5), which is present in other two protein families, one where the domain is tandemly repeated (with orthologs in all seven mycobacteria, e.g. Mt Rv3835), and another where it is combined with an N-terminal Ser/Thr Kinase C domain (present exclusively in a series of Actinomycetales species, e.g. *Stackebrandtia nassauensis* GI:229864975, 577 aa; see Supplementary Fig. S5).

The prediction of Mt Rv2270 as containing a lipoprotein anchor signal may invalidate the OMP function, but the predicted C4 domain has high beta content and high amphiphilicity; its involvement in variable domain architectures suggests that it can be used as a biological module.

Present in all seven mycobacterial genomes

We found a total of 588 clusters with sequences from all seven mycobacteria, and 61 of these were predicted as OMP families. These families are likely to represent proteins important for all mycobacteria but possibly not for pathogenicity since they are present both in pathogenic and in non-pathogenic organisms. We present two interesting cases below.

ACT: an actinobacteria OMP domain greatly expanded in Corynebacterineae

Rv0431 is an Mt predicted OMP with orthologs in all seven mycobacteria. Sequence analysis indicated that the family contains a C-terminal domain of about 100 aa (that we name ACT for the names given to the proteins where it is present: Alanine rich, CpsA, Tuberculin related) present in five Mt sequences that define five families (see Supplementary Fig. S6). In three of the five families the ACT domain is preceded by a predicted domain of around 170 aa of unknown function (PFAM LytR_cpsA_psr).

The ACT domain is present in some genera outside but close to mycobacteria, chiefly Nocardiodes and Corynebacterium, but not all species have the five sequences and Ma has an extra copy of one of the five. These results suggest that the ACT domain was invented before the divergence of mycobacteria, corynebacteria and nocardiodes. Its high level of duplication and a number of gene losses and duplications in mycobacteria suggest that it confers some kind of low-specific functional advantage.

An OMP essential for Mycobacteria growth

Rv0227c is another predicted OMP in a cluster with orthologs in all seven mycobacterial genomes. The proteins in this cluster have no known function, and closer analysis by PSI-BLAST revealed that there are distant homologs in nocardia and corynebacteria. The protein itself is characterized by a by a signal peptide with a predicted cleavage point before the first TM helix and a 300 aa beta-domain surrounded by two TMs. Mutagenesis and comparative genomic analyses have identified Rv0227c as being a 'core' mycobacterial gene, required for optimal growth (Marmiesse *et al.*, 2004; Sassetti *et al.*, 2003).

Example OMPs identified by new criteria

Method 2 includes two predictive features that have not been used before: export signals other than those reported by SignalP and a window analysis of secondary structure. Those allowed the identification of many extra OMPs respect to our previous work. Here we present two examples of OMPs detected by each of these new criteria.

An actinobacteria-specific protein with low global beta content

One of our clusters represents a family with members in five of the mycobacteria tested, four of which have OMP scores of 15 (Rv2345, MAV2041, Mb2374, MMAR_3652; ~660 aa) and one with an OMP score of 13 (MSMEG_4484). Notably absent are sequence homologs in Ml and Mu (confirmed using PSI-BLAST under default parameters), but we found orthologs of this protein in a wide range of Actinobacteria. These proteins contain a predicted Pfam domain of unknown function (DUF477), followed by a predicted TM, and a very variable glycine-rich region at the end. The percentage of beta structure of the whole sequence is well below the threshold of 0.11 that we use for selection. However, the window analysis shows that the DUF447 domain has a high percentage of beta content and high amphiphilicity, potentially characterizing an OMP function (Fig. 3).

Similarity searches uncovered a much shorter second homolog in Ma (MAV_2102), also present in *M. intracellulare*, which keeps the N-terminal domain, the predicted TM following it, and a C-terminal domain, but lacks the middle region and the Glycine-rich region (Fig. 3). We predict that both the long and the short families are OMPs.

Mycobacteria-specific OMPs secreted by the Tat system

An example found using the predicted Tat-system secretion that would not have been detected using SignalP was Rv2577 from Mt. This is an OMP predicted protein with orthologs in Mm and Ma (all of them with maximum OMP score = 16), apparently absent from Mu and Ml. The C-terminal end contains a predicted metallophosphoesterase domain (similarity to COG1409 Predicted phosphohydrolases according to database annotations) with clear homologs to other species outside actinobacteria.

In Mb AF2122/97, the syntenic gene of Rv2577 (529 aa) is separated into two open reading frames (Mb2607 and Mb2608) due to a base transversion (G-A), which introduces a stop codon. Mb2607 (83 aa) contains the signal sequence and sufficient beta strand structure for an OMP prediction of perfect score. Mb2608 (434 aa) matches Rv2577 from position 96 on, so that just 12 amino acids of the Mt protein are not represented in any of the two Mb proteins. The complete predicted phosphoesterase domain is intact. The N-terminal region has high content of potential amphiphilic beta-strand but this extends further to the region of homology to Mb2608. In the absence of sequences with homology to Mb2607 but not to Mb2608, we cannot support that Mb2607 can form an independent domain, although the gene split suggests this possibility.

Both Mb2607/Mb2608 transcripts have been shown to be up-regulated in a virulent strain of *M. bovis* during bacterial replication in macrophages (Blanco *et al.*, 2009). The G-A transversion is absent in avirulent *M. bovis* strains used for human vaccine development (*Mycobacterium bovis* BCG str. Tokyo 172, *Mycobacterium bovis* BCG Pasteur 1173P2). The splitting of this gene may extend host-specific modular functions of this protein in *M. bovis* AF2122/97, which is pathogenic to cattle (Garnier *et al.*, 2003).

Discussion

Outer membrane proteins (OMPs) act as gatekeepers to the external environment. They are exposed as quorum sensors or acting in response to its environment, and are likely to be essential for general survival of the cell. In pathogenic species, the function of the OMPs may play important roles in host-cell interactions that enable the persistence of mycobacterial infection. As such, OMPs are logical drug targets - not only in tuberculosis and leprosy, but also in opportunistic infections in immunocompromised patients, which in total kill millions of people world-wide every year and are complicated by new problems like co-infection with HIV and resistance to drugs (Meya & McAdam, 2007).

Existing computational methods to predict beta-barrel outer membrane proteins primarily focus on known OMPs in Gram-negative bacteria (Berven *et al.*, 2004; Bigelow *et al.*, 2004; Casadio *et al.*, 2003; Remmert *et al.*, 2009; Zhai & Saier, 2002). Unfortunately, OMPs of the beta-barrel type are almost totally uncharacterized in mycobacteria. Models for Gram-positive and Gram-negative OMPs can only partially extend to mycobacteria due to its unique cell wall construction that eludes clear functional classification as Gram-positive or Gram-negative. For example, the beta-barrels of the OMPs of mycobacteria have to be longer than those typically known from Gram-negative bacteria, in agreement with the greater thickness of the mycobacterial wall (Alahari *et al.*, 2007; Hoffmann *et al.*, 2008; Zuber *et al.*, 2008); this is the case for MspA (Faller *et al.*, 2004). As a result, the study of mycobacterial OMPs has to rely on tools specific to it.

Computational methods can be used to predict OMPs based on two properties: OMPs must form an amphipathic beta-barrel and be secreted. However, in view of the current small number of mycobacterial OMPs with which to benchmark such methods, and aware of the fact that mycobacterial OMPs are expected to be very different to those known outside Actinobacteria, we resourced to benchmark our method according to its ability to produce coherent predictions across proteins with high similarity to Mt proteins.

Accordingly, the method we presented here uses clusters of homologous sequences from seven mycobacterial genomes to optimize OMP prediction, based on the assumption that cross-genomic sequences within the same cluster should share similar properties, and therefore if the majority of sequences within a cluster were predicted to be OMPs, then those sequences that escaped prediction were also likely to be OMP sequences. In this manner, we were able to examine the effect of changing the thresholds of different parameters on the number of predicted OMPs to set final thresholds that reduced the number of spurious OMP predictions in clusters with low OMP content while maintaining OMP predictions for clusters with initially high OMP content. Moreover, we performed a sliding window analysis on all proteins to identify local regions of beta-content within larger proteins with low overall propensity to form beta-barrels. This method predicts practically all known mycobacterial OMPs with close to maximum scores.

We computed a set of 4300 potential OMPs in seven genomes (+600 alone in Mt). It is unlikely that all of them will be OMPs as current estimations of OMPs in Mt are in the order of 100s (Niederweis *et al.*, 2010). We do not think that with the current information on mycobacterial OMPs we can devise a more sensitive scoring system. In any case we note that this dataset includes a higher proportion of sequences from obligate pathogenic mycobacteria compared to opportunistic or non-pathogenic mycobacteria (15% versus 13%) suggesting that the set is enriched in genes with a function related to pathogenicity. Many of these proteins, as we have shown, define families specific to Mycobacteria or Actinobacteria that remain yet to be functionally characterized.

Our work proposes a number of putative OMP domains. Some of them are reused in multiple domain architectures and duplicated in paralogs (e.g. the ACT domain) or inside genes (e.g. the tandemly repeated C4 domain in Rv3835). Some of these domains or even some of the entire OMP predicted proteins are probably too small to form a beta-barrel by themselves (<150 aa). However, the many cases where such proteins appear together in putative operons (e.g. the mce operons) suggest that they may associate to form a barrel. OMP formation by oligomerization is already suspected in the predicted OMP Rv1698. Rv1698 has been observed to dimerize and the observation that channel complexes containing Rv1698 have variable conductance states suggest that Rv1698 might form oligomers (Siroy *et al.*, 2008). The formation of self-associations is also a possibility that has been reported. For example, both Mt MspA and the alpha-hemolysin porin of *Staphylococcus aureus* (from different phylum firmicutes) form a beta-barrel with each monomer contributing just a 50 amino acids loop to the beta-barrel associating as homo-octamer (Faller *et al.*, 2004) or homo-heptamer (Song *et al.*, 1996), respectively.

Conclusions

In summary, our results suggest that potential OMPs are a large contributor to the protein baggage of mycobacteria, possibly of Actinobacteria. Should a large fraction of our predictions be demonstrated experimentally to be OMPs, this would point to this function as an important factor for shaping the evolution, variability, and adaptability of these organisms. Using genomic information we have been able to tune an OMP prediction algorithm and produced a set of OMP predictions for more than 4300 mycobacterial proteins. Their profiles of taxonomic conservation can be used to hypothesize the functional importance and pathogenicity relevance.

We note that while this manuscript was under review, one of our predicted OMPs, Rv0899, has been the focus of an experimental effort to characterize it as an OMP (Teriete et al., 2010). Although the result was negative, this indicates that our method produces targets that align well with those that the researchers in the field choose using their intuition and knowledge. As new experimental evidence accumulates, we will be able to refine our algorithm. In addition, the expected sequencing of novel mycobacterial genomes will allow us to further complete the picture of the evolutionary history of OMPs and to pinpoint their association to pathogenicity, hopefully leading to new strategies to combat a number of terrible diseases.

Acknowledgements

This work was supported by funding from the Bundesministerium für Bildung und Forschung and from the Helmholtz Association. We thank M. Niederweis (University of Alabama at Birmingham) for helpful discussion.

References

Ahmad, S., Akbar, P. K., Wiker, H. G., Harboe, M. & Mustafa, A. S. (1999). Cloning, expression and immunological reactivity of two mammalian cell entry proteins encoded by the mce1 operon of Mycobacterium tuberculosis. *Scand J Immunol* **50**, 510-518.

Ahmad, S., El-Shazly, S., Mustafa, A. S. & Al-Attiyah, R. (2004). Mammalian cell-entry proteins encoded by the mce3 operon of Mycobacterium tuberculosis are expressed during natural infection in humans. *Scand J Immunol* **60**, 382-391.

Alahari, A., Saint, N., Campagna, S., Molle, V., Molle, G. & Kremer, L. (2007). The N-terminal domain of OmpATb is required for membrane translocation and pore-forming activity in mycobacteria. *J Bacteriol* **189**, 6351-6358.

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G. & Brunak, S. (2004a). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel* **17**, 349-356.

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004b). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795.

Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. (2005). Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**, 167.

Berven, F. S., Flikka, K., Jensen, H. B. & Eidhammer, I. (2004). BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res* **32**, W394-399.

Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D. & Rost, B. (2004). Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res* **32**, 2566-2577.

Blanco, F. C., Nunez-Garcia, J., Garcia-Pelayo, C. & other authors (2009). Differential transcriptome profiles of attenuated and hypervirulent strains of Mycobacterium bovis. *Microbes Infect* **11**, 956-963.

Brennan, P. J. & Nikaido, H. (1995). The envelope of mycobacteria. *Annu Rev Biochem* **64**, 29-63.

Casadio, R., Fariselli, P., Finocchiaro, G. & Martelli, P. L. (2003). Fishing new proteins in the twilight zone of genomes: the test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria. *Protein Sci* **12**, 1158-1168.

Cole, S. T., Brosch, R., Parkhill, J. & other authors (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-544.

Cole, S. T., Eiglmeier, K., Parkhill, J. & other authors (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007-1011.

Cosma, C. L., Sherman, D. R. & Ramakrishnan, L. (2003). The secret lives of the pathogenic mycobacteria. *Annu Rev Microbiol* **57**, 641-676.

Cuff, J. A. & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**, 508-519.

Cuff, J. A. & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**, 502-511.

Faller, M., Niederweis, M. & Schulz, G. E. (2004). The structure of a mycobacterial outer-membrane channel. *Science* **303**, 1189-1192.

Finn, R. D., Tate, J., Mistry, J. & other authors (2008). The Pfam protein families database. *Nucleic Acids Res* **36**, D281-288.

Flesselles, B., Anand, N. N., Remani, J., Loosmore, S. M. & Klein, M. H. (1999). Disruption of the mycobacterial cell entry gene of Mycobacterium bovis BCG results in a mutant that exhibits a reduced invasiveness for epithelial cells. *FEMS Microbiol Lett* **177**, 237-242.

Garnier, T., Eiglmeier, K., Camus, J. C. & other authors (2003). The complete genome sequence of Mycobacterium bovis. *Proc Natl Acad Sci U S A* **100**, 7877-7882.

Hett, E. C. & Rubin, E. J. (2008). Bacterial growth and cell division: a mycobacterial perspective. *Microbiol Mol Biol Rev* **72**, 126-156, table of contents.

Hoffmann, C., Leis, A., Niederweis, M., Plitzko, J. M. & Engelhardt, H. (2008). Disclosure of the mycobacterial outer membrane: cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. Proc Natl Acad Sci U S A 105, 3963-3967.

Jensen, L. J., Kuhn, M., Stark, M. & other authors (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37**, D412-416.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580.

Leversen, N. A., de Souza, G. A., Malen, H., Prasad, S., Jonassen, I. & Wiker, H. G. (2009). Evaluation of signal peptide prediction algorithms for identification of mycobacterial signal peptides using sequence data from proteomic methods. *Microbiology* **155**, 2375-2383.

Malen, H., Berven, F. S., Fladmark, K. E. & Wiker, H. G. (2007). Comprehensive analysis of exported proteins from Mycobacterium tuberculosis H37Rv. *Proteomics* **7**, 1702-1718.

Marmiesse, M., Brodin, P., Buchrieser, C., Gutierrez, C., Simoes, N., Vincent, V., Glaser, P., Cole, S. T. & Brosch, R. (2004). Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the Mycobacterium tuberculosis complex. *Microbiology* **150**, 483-496.

Meya, D. B. & McAdam, K. P. (2007). The TB pandemic: an old problem seeking new solutions. *J Intern Med* **261**, 309-329.

Niederweis, M. (2008). Nutrient acquisition by mycobacteria. *Microbiology* **154**, 679-692.

Niederweis, M., Danilchanka, O., Huff, J., Hoffmann, C. & Engelhardt, H. (2010). Mycobacterial outer membranes: in search of proteins. *Trends Microbiol*.

Pajon, R., Yero, D., Lage, A., Llanes, A. & Borroto, C. J. (2006). Computational identification of beta-barrel outer-membrane proteins in Mycobacterium tuberculosis predicted proteomes as putative vaccine candidates. *Tuberculosis (Edinb)* **86**, 290-302.

Perez-Iratxeta, C., Palidwor, G. & Andrade-Navarro, M. A. (2007). Towards completion of the Earth's proteome. *EMBO Rep* **8**, 1135-1141.

Remmert, M., Linke, D., Lupas, A. N. & Soding, J. (2009). HHomp--prediction and classification of outer membrane proteins. *Nucleic Acids Res* **37**, W446-451.

Sassetti, C. M., Boyd, D. H. & Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* **48**, 77-84.

Senaratne, R. H., Mobasheri, H., Papavinasasundaram, K. G., Jenner, P., Lea, E. J. & Draper, P. (1998). Expression of a gene for a porin-like protein of the OmpA family from Mycobacterium tuberculosis H37Rv. *J Bacteriol* **180**, 3541-3547.

Siroy, A., Mailaender, C., Harder, D., Koerber, S., Wolschendorf, F., Danilchanka, O., Wang, Y., Heinz, C. & Niederweis, M. (2008). Rv1698 of Mycobacterium tuberculosis represents a new class of channel-forming outer membrane proteins. *J Biol Chem* **283**, 17827-17837.

Song, H., Sandie, R., Wang, Y., Andrade-Navarro, M. A. & Niederweis, M. (2008). Identification of outer membrane proteins of Mycobacterium tuberculosis. *Tuberculosis (Edinb)* **88**, 526-544.

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. & Gouaux, J. E. (1996). Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science* **274**, 1859-1866.

Stahl, C., Kubetzko, S., Kaps, I., Seeber, S., Engelhardt, H. & Niederweis, M. (2001). MspA provides the main hydrophilic pathway through the cell wall of Mycobacterium smegmatis. *Mol Microbiol* **40**, 451-464.

Teriete, P., Yao, Y., Kolodzik, A., Yu, J., Song, H., Niederweis, M. & Marassi, F. M. (2010). Mycobacterium tuberculosis Rv0899 adopts a mixed alpha/beta-structure and does not form a transmembrane beta-barrel. *Biochemistry* **49**, 2768-2777.

Tsolaki, A. G., Hirsh, A. E., DeRiemer, K. & other authors (2004). Functional and evolutionary genomics of Mycobacterium tuberculosis: insights from genomic deletions in 100 strains. *Proc Natl Acad Sci U S A* **101**, 4865-4870.

Zhai, Y. & Saier, M. H., Jr. (2002). The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci* **11**, 2196-2207.

Zuber, B., Chami, M., Houssin, C., Dubochet, J., Griffiths, G. & Daffe, M. (2008). Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state. J Bacteriol 190, 5672-5680.

Tables

Table 1 - Predictions for seven genomes.

Genome (NCBI Taxon ID)	Habitat	Annotated proteins	Suggested OMP proteins with score >=12 (% of total)	Suggested OMPs unique to a genome*
Mycobacterium tuberculosis H37Rv (83332)	obligate pathogen	3991	629 (15.8 %)	35
Mycobacterium bovis AF2122/97 (233413)	obligate pathogen	3920	617(15.7%)	39
Mycobacterium leprae TN (272631)	obligate pathogen	1605	242 (15.1%)	77
Mycobacterium marinum M (216594)	environmental, facultative pathogen	5462	799 (14.6%)	184
Mycobacterium ulcerans Agy99 (362242)	environmental, facultative pathogen	4160	561 (13.5%)	111
Mycobacterium smegmatis str. MC2 155 (246196)	environmental, not pathogenic	6716	844 (12.6 %)	459
Mycobacterium avium 104 (243243)	environmental, facultative opportunistic pathogen	5120	641 (12.5%)	251
Total		30974	4333 (14%)	

* numbers of predicted OMPs in within clusters in other genomic patterns: present in all genomes: 189; present only in obligate pathogens (Mt, Mb, Ml): 48; present only in facultative pathogens (Mm, Mu, Ma): 66; all other combinations: 2874

Figure legends

Fig. 1 - OMP scores for sequences in OMP-rich and OMP-poor clusters.

The minimum score for OMP prediction was set to >= 12 (dotted vertical blue line). At this threshold, 94% of sequences from OMP-rich clusters (black line) are classified as OMPs, while 85% of the sequences in the OMP-poor clusters (red line) are rejected as OMPs.

Fig. 2 - Beta-barrel prediction scores for control sequences.

Positive controls for beta-barrels include 428 bacterial or eukaryotic proteins from Pfam or PDB with annotated beta-barrel structures and solved structure information. Negative controls include 90 actinobacterial sequence fragments with low beta content, as determined from solved structures in PDB. At beta-barrel score ≥ 6, 97% and 90% of known bacterial OMPs and annotated beta-barrels, respectively, are predicted to be beta-barrels, and 74% of low beta sheet content sequences are predicted to be without beta-barrel structure.

Fig. 3 - Frac and PerBeta in a sliding window for Rv2345.

Rv2345 defines a family conserved in Actinobacteria and present in the mycobacteria tested with the exception of Ml and Mu. Top: average on a 300 aa window of percentage of beta sheet (PercentBeta) and amphiphilicity of beta strands (FracB5) for Rv2345. The horizontal lines indicate the thresholds used for these two parameters. The plot suggests that the N-terminal of Rv2345 contains a highly amphiphilic beta structure. Bottom: the N-terminal end of Rv2345 and orthologs contains a predicted Pfam domain of unknown function (DUF477). Ma protein MAV_102 represents a different architecture but is potentially a shorter OMP as it keeps the N-terminal domain. Other predicted sequence features for these proteins are: transmembrane helices (blue boxes), a 300 aa domain (blue hexagon), a C-terminal domain (yellow oval), and a G-rich amino acid biased region (orange bar).

Supplementary material legends

Supplementary Fig. S1– Fraction of OMPs remaining with increased restriction of non-Method 1 parameters.

This figure shows the fractions of OMPs remaining (y-axis) from the OMP-rich (S1, solid line) and OMP-poor (S2, dotted line) groups of clusters, as the parameter thresholds become increasingly restrictive (x-axis).

Supplementary Fig. S2- Effect of changing original parameters on OMP prediction in OMP-rich (S1) and OMP-poor (S2) clusters.

Cutoff criteria for the parameters frac, percent beta, and general secretion score (Smean) were varied and the fraction of predicted OMPs relative to the Method 1 prediction, was recorded. Optimal cutoffs (shown in red) eliminated 5-25% of the OMPs in S2 clusters while maintaining at least 94% of the OMPs in S1 clusters.

Supplementary Fig. S3– Validation of OMP prediction and signal sequence prediction in mycobacteria.

(a) Recall-precision curve for predicted OMPs. This figure shows the recall and precision curve, based on the assumptions that OMP-rich clusters (S1) contained true-positives and OMP-poor clusters (S2) contained true negatives. An OMP score of 12 was chosen as the threshold for an OMP prediction. **(b)** General secretion scores for mycobacterial proteins. This figure shows the Smean scores, as determined from SignalP-v3.0, for 1723 cytoplasmic and 58 proteins known to be secreted by the general secretion pathway. Cytoplasmic proteins were taken from annotated, reviewed proteins in UniProtKB. Experimentally verified GSP secreted proteins were taken from the literature. Proteins with Smean≥0.54 (vertical dotted blue line) were considered to be secreted. At this cutoff, 93% of known GSP-secreted proteins are correctly predicted to be secreted, while 98% of the cytoplasmic proteins are not predicted to be secreted. **(c)** Twin arginine translocation scores for mycobacterial proteins. In this figure, experimentally verified Mycobacterium proteins secreted by

Tat system (19) were found by literature search. At the selected cutoff (TatP dvalue=0.36, vertical dotted blue line), the TatP algorithm correctly predicts 79% of mycobacterial positive validation proteins to be Tat-secreted. 98% of the cytoplasmic proteins were not predicted to be Tat-secreted. **(d)** Leaderless secretion prediction for mycobacterial sequences. In this figure, 50% of known leaderless secreted proteins (12) are correctly predicted to be secreted at nnscore ≥0.71 (SecretomeP-v1.0). This includes 5/6 ESX-1 secreted proteins and 1/5 SecA2 secreted proteins. A single protein (GLNA1_MYCTU), whose secretion mechanism is unknown, was not predicted to be secreted. 81% of known cytoplasmic proteins were correctly predicted to not be secreted.

Supplementary Fig. S4– OMP scores for known bacterial OMPs.

In this figure, each horizontal bar summarizes the points awarded to each bacterial OMP (indicated by UniProtKB accession). Points were awarded for prediction of secretion by signal sequence prediction (general or Tat secretion, 8 points; black bar) or leaderless secretion (3 points; red bar). For the beta-barrel structure, one point was awarded for 4 parameters (frac, perbeta, numB5, resB5) over the whole sequence (green bar) or for a sliding window of 300 aa (blue bar), for a maximum of 8 beta-barrel points.

Supplementary Fig. S5- C4: a tandemly repeated OMP domain

We found a domain that occurs in Mt proteins Rv2770 and Rv3835, tandemly repeated in the latter. In some Actinomycetales the domain occurs with a Ser/Thr protein kinase catalytic domain at the N-terminal and a predicted TM helix in between.

Supplementary Fig. S6- ACT. A domain duplicated and lost many times.

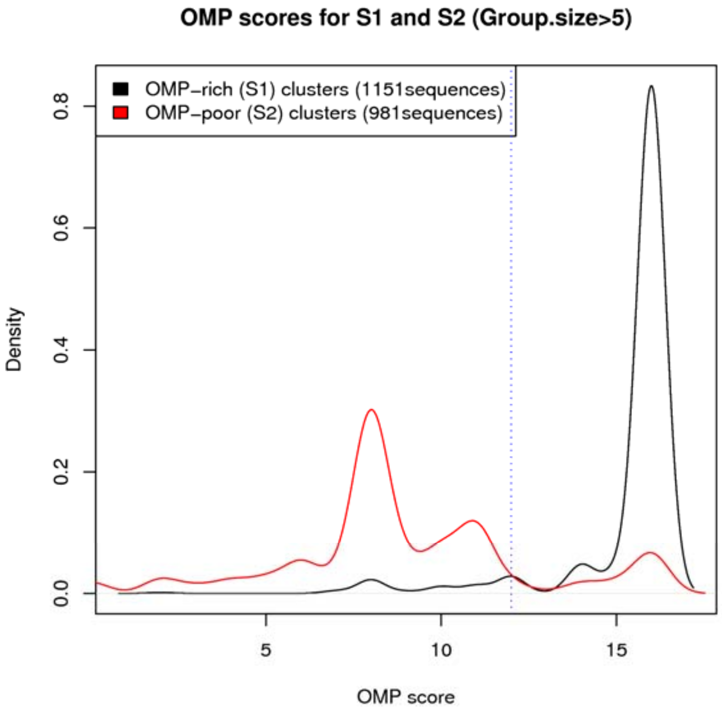
We identified a novel domain (ACT) as a candidate OMP domain occurring C-terminal in Mt proteins Rv0431, Rv2700, Rv0822c, Rv3267 and Rv3484. In three of them it is combined with an N-terminal extracellular domain of unknown function (LytR) found in a number of putative membrane-bound proteins. Left: phylogenetic tree from an alignment of instances of the domain in the seven mycobacterial species analyzed and in Corynebacterineae species: *Corynebacterium amycolatum* SK46 (Ca), *Rhodococcus opacus* B4 (Ro) and *Nocardia farcinica* (Nf). Right: sequence features of the five Mt sequences. Trans-membrane alpha helix (TM, blue) and signal peptide (SP, red) were predicted using TMHMM and SignalP-v3.0, respectively. The TM in Rv0822c was under the default cut-off of TMHMM and was not predicted but the 18 aa region reported displays a maximum of probability of being a TM (with scores above 0.6).

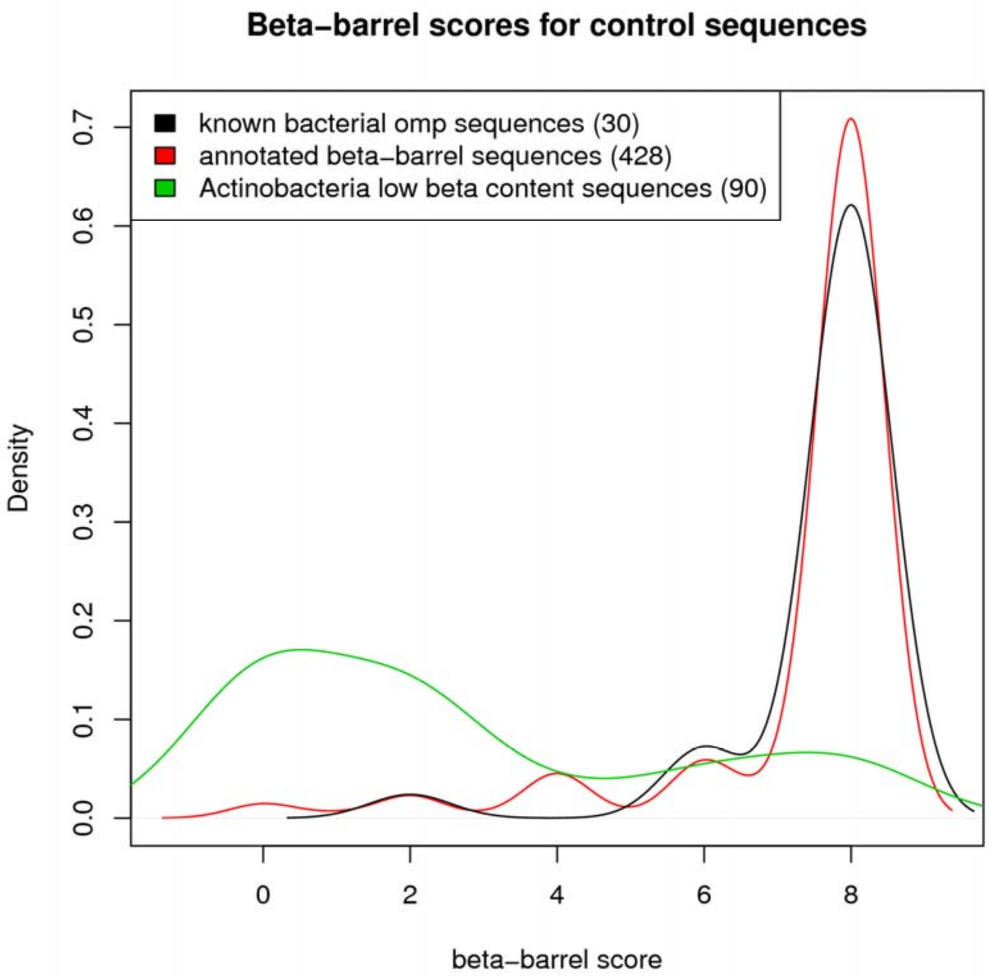
Supplementary Table S1– Criteria for OMP prediction.

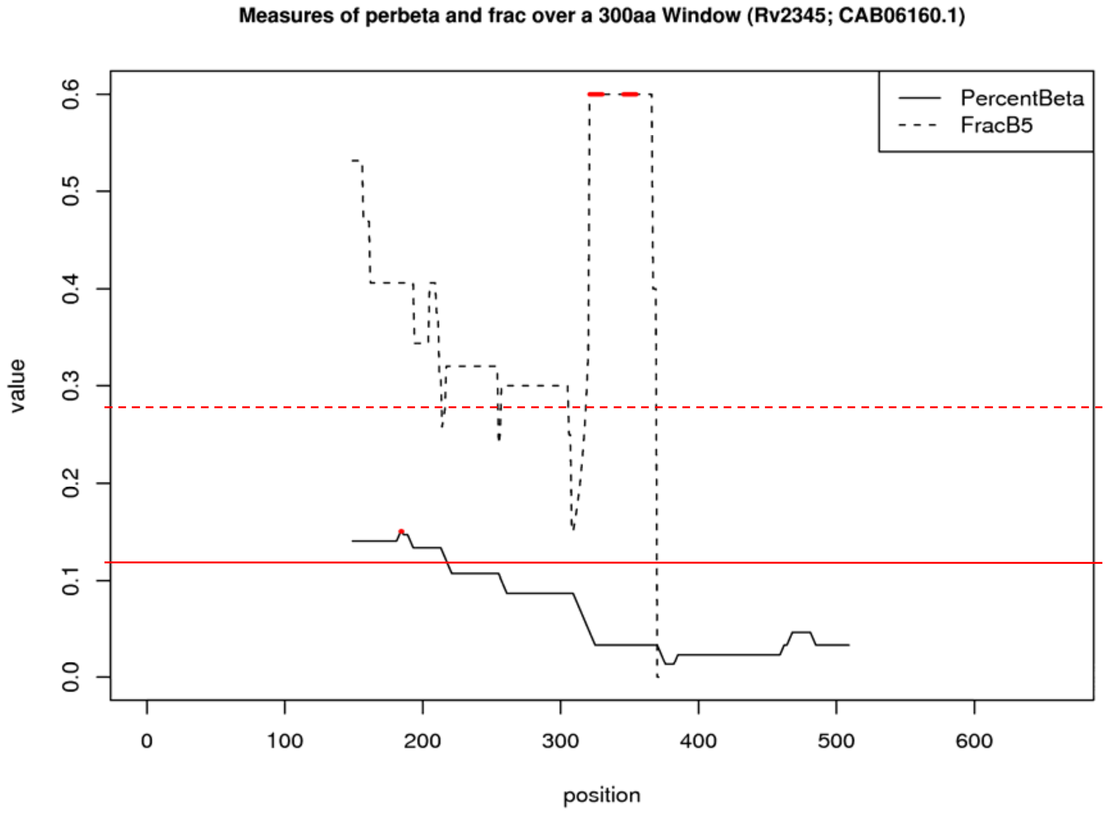
This table shows the criteria used to predict OMPs in Method 1 and Method 2 (this study). The parameter values for 30 known OMPs are included for comparison.

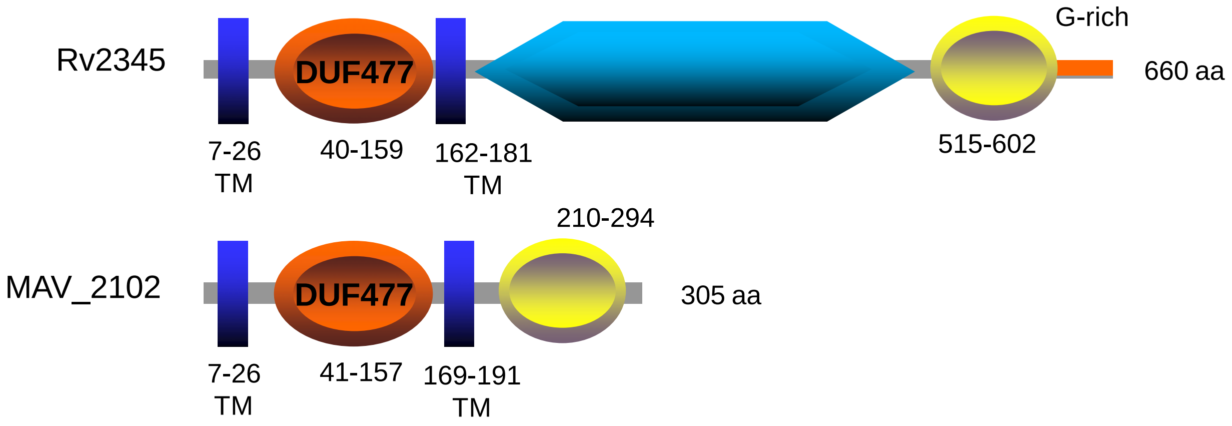
Supplementary Table S2– OMP score for mycobacterial sequences.

This file provides the results of the different scores and tests for protein sequences of seven mycobacteria used in this manuscript (30,605 sequences of length 50 amino acids or more). Each row represents a sequence. The columns indicate (1) species, (2) gene identifier, (3) OMP score, (4) Frac, (5) PerBeta, (6) Smean, (7) Dval (from tat), (8) number of predicted beta strands of five residues or longer and (9) number of residues in those, (10) number of predicted transmembrane helices, (11) position of the first transmembrane helix, (12) length of the sequence, (13) computed pI, (14) number of cysteines. Columns 15-20 regard the properties found on a 300 amino acid window whose position was selected as indicated in Methods: (15) window left start position, (16) number of beta strands of length five residues or more, (17) residues on those, (18) percentage of beta structure predicted and (19) Frac inside the 300 amino acid window, and (20) window beta score.


OMP scores for S1 and S2 (Group.size>5)
■ OMP−rich (S1) clusters (1151sequences)
🟥 OMP−poor (S2) clusters (981sequences)
Density
0.0
0.2
0.4
0.6
0.8
5
10
15
OMP score


Beta−barrel scores for control sequences
known bacterial omp sequences (30)
annotated beta−barrel sequences (428)
Actinobacteria low beta content sequences (90)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
beta−barrel score


Measures of perbeta and frac over a 300aa Window (Rv2345; CAB06160.1)



Rv2345
DUF477
G-rich
660 aa
7-26
TM
40-159
162-181
TM
515-602
210-294
MAV_2102
DUF477
305 aa
7-26
TM
41-157
169-191
TM