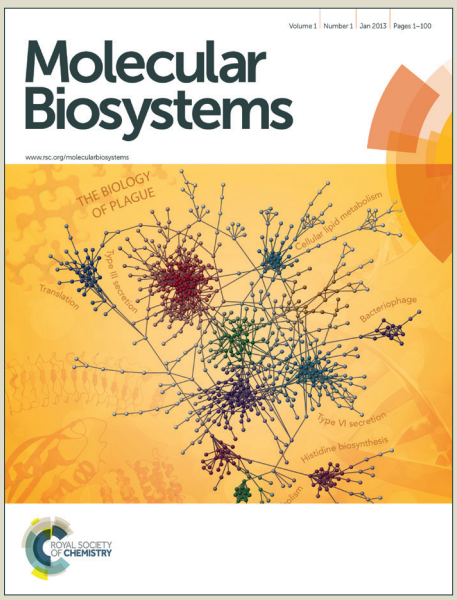
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Proteome-scale identification of Outer Membrane Proteins in *Mycobacterium avium* subspecies *paratuberculosis* using structure based combined hierarchical approach

Aarti Rana[1], Abdur Rub[2] and Yusuf Akhter[1*]

[1]School of Life Sciences, Central University of Himachal Pradesh, Shahpur, District-Kangra, Himachal Pradesh 176206, India.

[2]Infection and Immunity Lab, Department of Biotechnology, Jamia Millia Islamia (A Central University), New Delhi, India-110025.

[*]**Correspondence:** Dr. Yusuf Akhter, Assistant Professor, School of Life Sciences, Central University of Himachal Pradesh, Temporary Academic Block, Shahpur, Distt Kangra-176206, Himachal Pradesh, India. Tel: +91-1892-237288 ext 309, 310, 311. Mobile: +91-8263875426. Fax: +91-1892-237286. Email: yusuf.akhter@gmail.com; yusuf@daad-alumni.de

Abstract

Outer Membrane Proteins (OMPs) in eubacteria have several important roles, which ranges from membrane transport to the host-pathogen interactions. These are directly involved in pathogen attachment, entry and activation of several pathogen-induced signaling cascades in the cell. The cardinal structural features of OMPs include the presence of β-barrel, signal peptide and the absence of the transmembrane helix. This is the first report on proteome-wide identification of OMPs of ruminant pathogen, *Mycobacterium avium* subsp. *paratuberculosis* (MAP). The complete proteome of MAP was analyzed using a pipeline of algorithms, which screens the amino acid sequences and structural features shared by OMPs in other bacteria. Secondary structure of these proteins is also analyzed and scores are calculated for amphiphilic β-strands. From the set of 588 exported proteins, 264 proteins are predicted to be inner membrane proteins while 83 proteins are identified as potential OMPs in MAP. Finally, this study identified 57 proteins as top candidates, on the basis of computed Isoelectric points, as the core set of OMPs. Significantly, the resulting data on OMPs is not only useful in designing novel vaccines but may also open avenues for the development of early serodiagnostic tools for MAP.

Keywords

Outer membrane proteins, β-barrel structure, Inner membrane proteins, Sub-cellular localization, Non-classical secretory proteins.

Introduction

Mycobacterium avium subspecies *paratuberculosis* (MAP) is strikingly slow-growing[1], gram-positive, intracellular[2], acid-fast, mycobactin dependent[3], obligate zoonotic pathogen[4]. It was first observed by Johne & Frothingham in 1895. It causes Johne's disease or Paratuberculosis, an intestinal granulomatous infection among domestic and wild ruminants and has a worldwide distribution[5,6]. Clinical signs of disease include decreased milk production, diarrhoea, weight loss and ultimately death[1]. The disease is of substantial economic importance to livestock industries, especially it is a threat to the dairy production[7]. A major concern with this disease is the unconstrained bacterium outspread. Clinically or sub-clinically infected animal discharges MAP in milk and feces[8,9,10].MAPcontaining milk has been suggested as a potential factor of Crohn's disease in humans[11]. The current challenges include recognition and correlation of the elements that are essential for survival and virulence of the bacterium during infection, particularly those that elicit the immune responses against MAP[12].

Mycobacteria have evolved typically with a complex thick and waxy cell envelope, composed of peptidoglycan-arabinogalactan polymer with covalently bound mycolic acids of substantial size. This also contains a variety of pore-forming proteins and extractable lipids[13]. MspA is a highly stable oligomeric porin present in the outer membrane of *Mycobacterium smegmatis*. It was the first studied Outer Membrane Protein (OMP) from mycobacteria, which provided influencing evidence that OMPs exist in mycobacteria[14]. Mycobacterial cell envelope is highly impermeable[12] and composed of the plasma membrane, the cell wall and the outer membrane with capsule [15]. The capsule is rich in proteins and polysaccharides, and the inner leaflet of outer membrane is composed of mycolic acid residues, while the outer leaflet contains considerable amounts of lipids[16,17]. This asymmetric permeability barrier in mycobacteria confers a distinctive resistance to solutes that includes many therapeutic agents and antibiotics[18, 19]. The cell envelope has been involved in many facets of the pathogenicity of mycobacteria[17,20]. During the interaction between the bacterial cell and its environment, the surface proteins play an essential role[21,22,23]. They are usually involved in attachment, invasion of host cells, sensing physical and chemical conditions of the external environment and transmitting appropriate signals to the cytoplasm, to strengthen the defense against intoxication and host response[12]. At the

same time, number of tasks crucial to the bacterial cells, such as signal transduction, as well as solute and protein translocation is performed by the embedded proteins in the outer membrane. Integral OMPs characteristically fold into antiparallel beta-barrels in contrast to the other integral membrane proteins[24]. To determine the active components and the membrane isolates in their biologically active forms, long established procedures have been developed and utilized to separate the cell envelope by the breakage of the unusual thick wall that encloses mycobacteria into its inner and outer membrane constituents in sucrose gradients[25,26,27,] usually achieved only by harsh conditions, which generally leads to blending of components of both membranes and resulted in a bad yield[28],[29]. This is a big hinderance in protein localization experiments and in turn identification of OMPs in mycobacteria and related species[26]. So far, all known OMPs form β-barrels and are characterized distinguishingly by a pattern of alternating hydrophilic and hydrophobic amino acids in the β-strands[24].On the basis of the current knowledge of OMPs, several computational methods, have been employed for the systematic exploration of potential OMPs in mycobacteria[30]. However, the efficiency of such approaches is narrowed for several reasons like some of the predicted OMPs show sequence homology to well-known periplasmic proteins, cytoplasmic proteins, lipoproteins and non-classical secretory proteins. Unveiled OMPs are currently promising vaccine candidates against a wide variety of bacterial pathogens, and captivating targets for designing of new antibiotics. Therefore, it seems invaluable to perform a β-barrel based and structure knowledge-directed examination with alternating hydrophobic and hydrophilic amino acid residues to identify outer-membrane proteins in mycobacteria[30].

We report the prediction of OMPs of MAP using a proteome-wide computational approach as a first step towards illuminating MAP's physiology and virulence, and furnishing a foundation for the development of the next generation of MAP vaccines and diagnostic markers [1]. Herein, we are predicting 83 OMPs for which no prior experimental work has been reported. Out of which 57 proteins were identified as core set of putative OMPs on the basis of computed Isoelectric points. This work is, to the best of our knowledge, the first comprehensive computational proteomic investigation of the MAP proteins that are potentially OMPs.

Results

1) Exported proteins contain Signal Peptides and respective cleavage sites

Secretory proteins have a specific signal peptide sequence which helps them to cross through the inner membrane. Secretory proteins synthesized as pre-proteins with signal sequences at N-terminus[31]. A signal sequence is referred as signal peptide, cleaved by signal peptidase[3], generally comprises of 18 to 30 residues. The signal peptide consists of positively charged amino acid residues at N-terminus, followed by a non-polar hydrophobic core and polar C-terminal end. The C-terminal domain contains the cleavage site recognized by signal peptidase. Signal peptidase represents membrane-bound protease and removes the signal peptide from the precursor proteins during the translocation reaction. On the other hand, Inner Membrane Proteins (IMPs) generally lack signal peptide, and alternatively have hydrophobic transmembrane helices which acts as an internal signal[32]. Herein, 20 known reference OMPs were scanned by TargetP1.1 server [33] and yielded a minimum value of Signal Peptide (SP) score as 0.532 **(Figure S1).** This minimum value of the SP score (0.532) was set as cutoff value for further screening of 4,356 test proteins in the study of proteome-wide identification of OMPs**.** 588 proteins were predicted to be exported proteins with signal peptidase cleavage site.

2) Identified transmembrane α-helices constitutes Inner Membrane Proteins

In bacteria most of the proteins reach their final destination *via* the General Secretory Pathway (GSP). They are discharged into the periplasm from where they are merged into the inner membrane or are carried over to the outer membrane. Integral IMPs have hydrophobic amino acids forming α-helix. Hydrophobic regions act as membrane anchor signal or stop transfer signal and anchor the proteins in the inner membrane[3]. The hydrophobic α-helices can be identified by an uninterrupted length of 20-30 non-polar residues having predominant aliphatic side chains at the center while both ends are occupied by aromatic residues[34,35]**.** 588 mature protein sequences were examined by HMMTOP (Hidden Markov Model for Topology Prediction) [36] and TMHMM (Transmembrane Hidden Markov Model) [37], which are used to predict transmembrane α-helices. Proteins were not considered as IMPs if both the methods were unable to predict any transmembrane α-helix. Out of 588 predicted exported proteins, 264 proteins were identified with transmembrane α-helices (**Figure 1a**) and were designated as IMPs. Remaining 324 proteins were lacking transmembrane α-helices and were further analyzed. PEDANT (Protein Extraction Description and Analysis Tool) genome database provides exhaustive automatic analysis of genomic sequences by a

large variety of bioinformatic tools. In PEDANT database[35] (http://pedant.gsf.de) 907 IMPs are from MAP. We predicted 264 IMPs, out of which all are present in PEDANT [38] remaining novel 643 IMPs (constitute ~70% of total IMPs from MAP) which does not contain canonical signal peptide (**Figure 1b)**[28].

3) Secondary structure signatures for identification of β-barrel motifs

Various X-ray crystal structures have revealed that most of the OMPs of bacteria have canonical transmembrane β-barrel motifs [35,39]. OMPs with β-barrels are translocated through the inner membrane *via* periplasmic space to get integrated into the outer membrane [40]. *Mycobacterium smegmatis* contains four porins, out of which MspA (PDB ID: 1UUN) consists of two consecutive β-barrels with non-polar outer surfaces forming a homooctameric goblet-like conformation [41]. This indicates that mycobacterial OMPs also have β-barrels. To form a functional β-barrel, there should be at least 2 β-strands in the protein secondary structure. The 30 well characterized reference OMPs were scanned using Jnet algorithm by JPRED3 [68] to predict the secondary structure. Of all the 30 reference OMPs, TolC of *Escherichia coli* scored the minimum β strand score of 0.09 (**Figure 2a**) which was set as threshold value to form β-barrels. This threshold score was crossed by 243 proteins (**Figure 2b**), out of 324 proteins in which transmembrane α-helices were predicted absent.

4) Amphipathic β-barrels present in potential OMPs

In present study, a modified approach to predict the OMPs was used. It included identification of amphipathic β-strands. An algorithm for the identification of OMPs was applied to calculate amphiphilicity of predicted β-strands. To predict amphipathic β-strands having alternating hydrophobic and hydrophilic stretches, residue i was supposed as to lie in the middle of one side of a β-strand extending across the membrane and the mean hydrophobicity of this particular side was calculated as: $H\beta(i) = 1/5 \times (h(i-4)+h(i-2)+h(i)+h(i+2)+h(i+4))$. Zero crossings were calculated using MATLAB (Matrix Laboratory) Proteinprop plot tool (Kyte& Doolittle) [43] for 243 proteins setting window size of 5 residues. Amphiphilicity score close to one shows that the β-strands have a complete pattern of alternating hydrophobic and hydrophilic residues, whereas a score of zero indicates that β -strand residues are completely either hydrophobic or hydrophilic [43]. Of the known 30 reference OMPs, *Escherichia coli* OmpT was set as the lower limit with amphiphilicity score of 0.18 (**Figure 2a).** Single protein (MAP0539) out of 243 proteins was found to have amphiphilicity below 0.18

score and hence removed from the list. Remaining 242 proteins (**Figure 2b**) showed pattern of alternating hydrophobic and hydrophilic residues in their β-strands were fed into further step of the pipeline.

5) Omission of lipoproteins from the list of exported proteins

Most of the lipoproteins in bacteria are located on the periplasmic surface of the outer membrane while some are anchored with the inner membrane. Lipoproteins are not considered as integral membrane proteins but are attached only by an acyl chain in the membrane [44]. A typical feature of all of these lipoproteins is the presence of Cysteine residues at the start of signal peptide at N-terminus [45,46]. PRED-LIPO [45] was used for the prediction of lipoproteins in gram-positive bacteria using Hidden Markov Model. Using this method, 38 proteins were predicted as lipoproteins and were eliminated from the set of 242 amphipathic β–barrel proteins.

6) Removal of cytoplasmic and periplasmic proteins

Proteins are targeted to different sub-cellular locations depending on the presence of the respective localization signal peptides present at N-terminus. To predict the precise localization of proteins, the most accurate bacterial protein Subcellular Localization (SCL) Predictor, PSORTb version 3.0.2 [47] was used and resulted in identification of 58 proteins with similarities to cytoplasmic and periplasmic proteins. Therefore, 58 proteins were eliminated from the set of 204 amphipathic β–barrel proteins and remaining 146 proteins were further analyzed.

7) Exclusion of classically and non-classically secreted proteins

Exported proteins are retained completely or partially inside the cell boundary or on the outer membrane, whereas secretory proteins are transported extracellularly[3]. BLASTP (Basic Local Alignment Search Tool) analysis was carried out for the selected set of 146 protein sequences against the 30 experimentally characterized secretory proteins of eubacteria which were manually retrieved from the literature to check out the similarities. None of the proteins were found homologous to the well known secretory proteins. Literature search shows that there are some proteins that follow signal peptide independent secretion pathway and this phenomenon is termed as non-classical secretion[3]. Proteins secreted by the non-classical secretion pathway can be identified using SecretomeP 2.0

program [48]. 63 Non-Classical Secretory proteins were identified using SecretomeP server. These non-classically secreted proteins were excluded from the selected set of proteins. This reduced the number of putative OMPs to 85.

8) Core set of putative OMPs are rich in acidic amino acid residues

Several other characteristics of OMPs may be used for further narrowing down the list of potential OMPs. Most of the gram-negative bacterial OMPs have a low Isoelectric point (pI) because they are rich in acidic residues. MspA of *Mycobacterium smegmatis*, Rv1698 and Rv1973 of *Mycobacterium tuberculosis* have an acidic Isoelectric point and can be used as worthwhile parameter for the identification of mycobacterial OMPs. Isoelectric points of 30 reference OMP's as computed by ProtParam tool [49] gave theoretical pIs less than 6.5 (**Figure S2a**) leaving out Omp32 of *Delftia acidovorans.* We observed that Omp32 of *Delftia acidovorans* is only one which is rich in basic amino acids and thus has high Isoelectric point [50]. 57 of the 85 predicted OMPs of MAP exhibited a pI score less than 6.5. This represents the core set of putative OMPs (**Figure S2b).**

9) Functional annotation and categorization of potential OMPs

83 identified proteins were classified across the eleven functional groups based on the Pasteur Institute functional classification tree (**Figure 3**)[51]. It is available at http://genolist.pasteur.fr/TubercuList/. Most of the proteins were involved in Cell-wall and cell processes (functional class 3, 27.03%), Conserved hypothetical proteins (functional class 10, 25.00%) and Virulence, detoxification, adaptation (functional class 0, 22.60%). We observed no basis for any of the predicted OMPs that can be categorized in functional category 4 (Stable RNAs), 5 (Insertion Sequences and Phages) and 8 (Proteins of Unknown Function).

10) Gene Ontology terms & Potential O- and N- glycosylation sites

Protein glycosylation is a commonly observed posttranslational modification in bacteria. It involves covalent modifications of proteins by carbohydrates. Glycosylation occurs in all three domains of life and plays important roles during cell–cell recognition, attachment and intracellular sorting [52,53.] Potential 45 N-glycosylation sites and 35 O-glycosylation sites were predicted for the 57 OMPs using GlycoPP program [54]. These set of proteins represent the bacterial glycosylated proteins

and can be involved in bacterial pathogenesis [52,53]. Gene ontology terms were also assigned to the top 57 candidate OMPs and hydrolase activity was observed over-represented (**Supplementary Table 2**).

Discussion

Porin MspA identified as OMP in *Mycobacterium smegmatis* has been shown to share common characteristic features with OMPs of gram-negative bacteria. It provides the evidence establishing the fact that mycobacteria contains OMPs with β-barrels in the outer membrane. A number of computer based programs have been written for identifying β-barrel proteins employing similar amino acid sequence and structural features shared by OMPs. Various programs have been utilized among various studies for the identification of OMPs in both gram positive and gram negative bacteria. E-Komon et al. (2002) used multiple predictors and identified putative OMPs from two available *Pasteurella multocida* genomes: 98 putative OMPs from the avian strain Pm70 and 107 putative OMPs from the porcine non-toxigenic strain 3480 [55]. In 2009, Song et al., was successful in predicting 144 proteins as OMPs of *Mycobacterium tuberculosis* using an algorithm primarily based on physiochemical properties and biological knowledge of OMPs. However, algorithms utilized by Song et al., is limited by a reason: it did not exclude non-classical secretory proteins from the identified set of OMPs. In order to predict integral OMPs of MAP, we employed better computational approach exploiting the similar features shared by bacterial OMPs: existence of signal peptides, absence of hydrophobic transmembrane α-helices and presence of β-barrel structures with alternating hydrophobic and hydrophilic residues. To overcome the limitation of earlier used approach, we eliminated non-classical secretory proteins from predicted OMPs. The overall strategy is explained in **Figure 4** and names of the programs/servers used in this study are summarized in **Table 1**. Combination of multiple predictors will increase the coverage and accuracy of the predicted outer membrane proteome with less false positives. As a control, same pipeline following multiple predictors was first tested on 30 known reference proteins selected across different gram positive and gram negative bacterial species. These proteins were used to define the threshold values for various parameters that were met by 83 proteins of MAP.

Secretory proteins having hydrophobic signal peptides are preferentially targeted by SecB pathway during translocation across cell membranes[32]. 20 known reference OMPs when scanned by

TargetP 1.1 program, resulted in 0.532 (threshold value) as the minimum value obtained of SP score. OmpATb , one of the reference OMP from *Mycobacterium tuberculosis* was not identified as a secretory protein[18],[56] as it has a contrasting molecular mass, and its N-terminal sequence is different from the known porins of mycobacteria [57]. The C- terminal domain of an exported protein contains the cleavage site recognized by signal peptidase. A membrane-bound protease cuts away the signal peptide from the precursor protein during the translocation and forms a mature protein[32].Signal peptide sequences were removed from the protein sequences to obtain mature proteins.

A genetic and biochemical study provided a better understanding of the mechanism of protein translocation and has shown that IMPs are directed to the translocase mediated by Signal Recognition Particle (SRP)[32]. Out of 588 exported proteins of MAP, 264 proteins have hydrophobic transmembrane α-helices as predicted by HMMTOP and TMHMM. These predicted IMPs are present in the list of 907 possible IMPs as provided by PEDANT database. This indicates that greater part of IMPs of MAP do not contain signal peptide. This finding is coherent with the proof that most of the IMPs of many bacteria have signal anchor helix which is non-cleavable in nature and shows high hydrophobicity as compared to signal peptide [58].

Of the 30 reference OMPs, OmpF and MspA structures determined by Raman spectroscopy consist of 50-60% of beta-strand content [59, 60]. When beta-strand contents are predicted for these proteins, OmpF and MspA showed 46% and 32% of beta content respectively. Similar effect was seen in other proteins and reference OMPs scored highest beta-score. This deduces a conclusion that beta-strand content can be used significantly in relationship with amphiphilicity. Both parameters were used altogether and resulted in elimination of 82 exported proteins lacking amphiphilic β-strands.

Exported proteins are translocated across the cytoplasmic membrane *via* the general secretory pathway (GSP). Their final destination can be cytoplasmic membrane or periplasm or outer membrane[3]. *Escherichia coli* contains more than 90 classes of lipoproteins, many of them are located at the periplasmic side of outer membrane, and some are identified in the inner membrane [44].We intended to identify the OMPs only, so we eliminated proteins found similar to known cytoplasmic proteins, periplasmic proteins, lipoproteins and secretory proteins. Two proteins with only one β–

strand were eliminated as they could not form any β–barrel structure. We were successful in short-listing 83 proteins predicted as putative OMPs from MAP **(Supplementary Table 1).**

Nevertheless, despite substantial efforts to limit our search to a certain location, it is almost inevitable that our predicted list of 83 putative OMPs in addition contains false positives representing secreted or periplasmic proteins. Meticulous *in vitro* studies intending subcellular localization of proteins are needed further to validate these results. Many of the surface exposed proteins have low pI values due to the presence of acidic residues like Asp and Glu, it is due to the fact that it is exposed to the aqueous environment. Our data on Isoelectric points of 30 reference OMPs showed theoretical pI values in acidic range (**Figure S2a**). Therefore, we have further screened out more candidate OMPs on the basis of their pIs. Finally, we have come down to the core set of OMPs. The number of predicted 57 OMPs out of 4356 proteins in putative MAP proteome was found consistent with the results of a similar study conducted on putative *Mycobacterium tuberculosis* proteome giving 32 OMPs out of 3991 proteins[28]. However, when these 57 proteins were scanned against experimentally characterized OMPs using BLASTP, no apparent homology and evolutionary conservation was observed.

Glycosylation of bacterial proteins is an important process for bacterial physiology and pathophysiology [61]. Both O- and N-linked glycosylation sites were identified in bacterial glycoproteins. A number of glycosylated proteins have been documented as virulence factors of medically important pathogens. These have been associated with bacterial infection [62,63] and play an important role in pathogenesis [64].

For the veterinary scientific community paratuberculosis still remains an unsettled question despite the large efforts to control it effectively during the past decades [65]. Presently, there is no availability of diagnostic assays and vaccines for paratuberculosis which are adequately effective for early stage diagnosis and infection control. A number of experimental studies have been well cited in literature referencing OMPs as potential candidate molecules for development of vaccines among bacterial pathogens [66]. Comprehending the structure and functions of bacterial OMPs will generate information that is of common interest for the better understanding of its role in bacterial pathogenesis. It will provide a potential repertoire of candidate antigens for immunological

diagnostics and may be used to expedite the vaccine development[22]. Mycopar® from Boehringer Ingelheim is the only vaccine against Johne's disease (JD) available for the limited use in which inactivated bacteria is used for immunization. In spite of the vaccination, MAP shedding continues which spread the disease to new animals [67, 68]. Therefore, better vaccines are required to be developed against JD to rescue the economic loss in the dairy industry [69,70]. The exposed OMPs containing β-barrels may also be the interesting target for the development of better diagnostic tools as the available culture and PCR based diagnostic tools are very sophisticated, time consuming as well as cost effective [71].The OMPs may also be the novel targets for the investigation of the new antibiotics against MAP.

Experimental

Annotated proteome sequence dataset of MAP

The whole genome of MAP, with annotated protein sequences is available at the servers of EMBL/GenBank (European Molecular Biology Laboratory). A FASTA (FAST-ALL) file containing 4,356 protein sequences for the MAP K-10 genome was retrieved from the ftp server at National Centre for Biotechnology Information (NCBI). 30 known reference OMPs as control were selected across the different bacterial species.

Prediction of exported proteins

Total 4,356 protein sequences were subjected to the prediction of their secretory nature. Classically secreted proteins were predicted by detecting the presence of signal sequences using TargetP1.1 server [33]. Threshold values were specified by scanning the 30 reference OMPs of bacteria for the signal sequences and cut-off was set at 0.532. The proteins with predicted scores of 0.532 and above were considered to be exported proteins and sequences corresponding to the potential signal peptides were deleted. Obtained shortened sequences represent the mature proteins and were used for further analysis.

Prediction of transmembrane α-helices

To exclude the α-helix containing transmembrane proteins, sequences of the mature proteins were examined by HMMTOP [36] and TMHMM2.0 [37] programs. MAP proteins were not considered as integral IMPs when the two methods did not predict a transmembrane α-helix. Both these programs were used to predict the presence of transmembrane helices among the protein sequences with standard settings until else noted.

Secondary structure prediction

Proteins without transmembrane α-helices were selected and analyzed further for secondary structure prediction by JPRED3 [42] which is powered by the Jnet algorithm [72]. Among secondary structure prediction algorithms, Jnet with an average accuracy of 76.4% gives the best performance [28,70]. β-strands with a minimum of at least five successive residues were selected in both the strands forming a single β-barrel [43].

Amphiphilicity estimation

To calculate amphiphilicity of the selected β-strands, a specifically developed, Jahnig algorithm [43] was used. The mean hydrophobicity of membrane spanning β-strands Hβ(i) of one side, which comprises of 5 residues, was calculated following Vogel and Jahnig [43] as $H\beta(i) = 1/5 \times (h(i-4)+h(i-2)+h(i)+h(i+2)+h(i+4))$. i is the position of an amino acid residue in a sequence and h(i) represents its hydrophobicity. For amino acid sequence from residue 1 to n, values may be calculated for i = 5, 4,...,n - 4. Hydrophobicity values were taken from Sweet and Eisenberg[71]. Hβm represents the mean values of hydrophobicity along the complete protein sequence. Further, we have calculated the zero crossings of the Hβ – Hβm. Together with these measurements and the secondary structure prediction, the total number of hydrophobicity crossings per residue in protein's β-strand was specified as "amphiphilicity" which was used as discriminating parameter[28].

Prediction of Lipoprotein signal peptides

Potential lipoproteins were predicted among obtained amphiphillic proteins using PRED-LIPO [45]. It uses Hidden Markov Model (HMM) based method for the prediction of lipoprotein signal peptides of Gram-positive bacteria [45]. HMM was trained on a set of 67 experimentally verified lipoproteins. The method performs better than LipoP and other methods, which are based on regular

expression patterns. In different data sets consisting of experimentally characterized lipoproteins, cytoplasmic proteins, secretory proteins and proteins with N-terminal TM segments. This method is highly specific and sensitive in detection of signal peptides .

Subcellular localization

Protein subcellular localization prediction was done by PSORTb version 3.0.2 [47]. PSORTb is the most precise bacterial localization prediction tool available. It uses various analytical modules including the support vector machine (SVM), which is trained to identify the cytoplasmic, inner membrane, cell wall and extracellular proteins. Each localization site predicted by the module was provided with localization score and one of the sites that have a score of 7.5 or above was returned as the final prediction site.

Prediction of non-classically secreted proteins

SecretomeP 2.0 [48] was utilized to predict non-classical secretory proteins from the selected set. This program produces *ab initio* prediction of non-signal peptide targeted proteins using a Neural Network architecture based on sequence-derived protein features. Protein features for each protein sequence were calculated and assigned. Features were encoded by processing them with neural networks, which were trained using three-fold cross validation. Based on correlation coefficient, combination of features yielding best performance was determined. Proteins showing SecP score more than 0.5 were considered to be secreted.

Computation of Isoelectric points of predicted OMPs

Computation of theoretical Isoelectric point (pI) of proteins were carried out using ProtParam tool [49] from ExPaSy. pI of each protein is calculated using pK values of amino acids as described in Bjeequist et al., defined by examining polypeptide migration between pH 4.5 to 7.3 in an immobilized pH gradient gel environment [76].

Functional classification

Selected proteins were further categorized and annotated into 11 functional groups based on the Pasteur Institute functional classification tree [51]. Functional category was postulated for proteins, based on sequence similarities with functionally characterized proteins from related species, *Mycobacterium tuberculosis* H37Rv, classified into eleven functional categories [51]. Proteins without an indication of a function were classified as conserved hypothetical proteins.

Gene Ontology terms & O- and N- glycosylation sites

For prediction of gene ontology terms, sequences of proteins selected as core set of putative OMPs were scanned with GOanna server [77]. GOanna performs a BLAST search against protein sequences that have GO number. The proteins were further analyzed for potential O- and N-glycosylation sites using GlycoPP version 1.0 program [54]. It is a highly accurate O- and N-glycosylation prediction software made available for the analysis of prokaryotic protein sequences. GlycoPP program is trained on an extensive dataset of 116 O-glycosites and 107 N-glycosites extracted from 59 experimentally characterized glycoproteins of prokaryotes. For prediction, SVM (Support Vector Machine) based prediction approaches were used. N- glycosites were predicted using Binary profile of pattern (BPP) method whereas O-glycosites were predicted using PSSM profile of patterns (PPP) strategy [54].

Conclusions

We have achieved a proteome-wide comprehensive account of MAP OMPs, with additional observation of IMPs with signal peptides. 83 distinct candidate OMPs were identified, to our knowledge, this study symbolizes the first *in silico* genome wide analysis of OMPs of MAP. 21 proteins of the identified set of OMPs were conserved hypothetical proteins and 27.3% were predicted to be involved in cell-wall and cell processes. OMPs are engaged in important role in bacterial pathogenicity and host-pathogen interactions. The predicted proteomic profile can be considered for developing strategies to produce vaccine leads and may prove to be a better starting material for the identification of B and T cell epitopes analyzing the potentially antigenic candidates for vaccine development.

Acknowledgments

Authors are thankful to Central University of Himachal Pradesh for providing the research infrastructure. Aarti Rana is supported by UGC Ph.D. fellowship. Research in Dr. A. Rub's lab is supported by grants from INSA, UGC (Govt. of India), and DST (Govt. of India). Dr. S. Gourinath group at Jawaharlal Nehru University, New Delhi is acknowledged for providing software and computational facility.

Conflict of Interest Statement

Authors declare no conflict of interest.

References

1. Li. Lingling, J. P. Bannantine, Q. Zhang, A. Amonsin, B. J. May, D. Alt, N. Banerji, S. Kanjilal, V. Kapur, *PNAS*, 2005; 102(35): 12344–12349.

2. R. B. Gurung, A. C. Purdie, D. J. Begg, R. J. Whittington, *Clin Vaccine Immunol.,* 2012; 19(6):855–64.

3. A. P. Pugsley, *Microbiol Rev.*, 1993; 57(1):50–108.

4. P. S. Rani, L. A. Sechi, N. Ahmed, *BioMed Cent*., 2010: 1–6.

5. R. M. Beard, D. Henderson, M. J. Daniels, A. Pirie, D. Buxton, A. Greig, M. R. Hutchings, I. McKendrick, S. Rhind, K. Stevenson, J. M. Sharp, *Br Vetriniary Rec*., 1999; 145.

6. A. Greig, K. Stevenson, D. Henderson, V. Perez, V. Hughes, I. Pavlik, M. E. Hines, I. McKendrick, J. M. Sharp, *J Clin Microbiol*., 1999; 37(6):1746–51.

7. J. P. Bannantine, C. W. Wu, C. Hsu, S. Zhou, D. C. Schwartz, D. O. Bayles, M. L. Paustian, D. P. Alt , S. Sreevatsan, V. Kapur, A. M. Talaat, *BMC Genomics,* 2012; 13(1):89.

8. I. R. Grant, H. J. Ball, M. T. Rowe, *Appl Environ Microbiol.,* 2002; 68(5):2428–2435.

9. R. N. Streeter, G. F. Hoffsis, S. Bech-Nielsen, W. P. Shulaw, D. M. Rings, *Am J Vet Res*., 1995; 56(10):1322–4.

10. R. W. Sweeney, *Vet Clin North Am Food Anim Pract.,* 1996; 12(2):305–12.

11. R. J. Greenstein, *Lancet Infect Dis*., 2003; 3:507–514.

12. Z. He, J. De Buck, *Proteome Sci*., 2010; 8:21.

13. C. Hoffmann, A. Leis, M. Niederweis, J. M. Plitzko, H. Engelhardt, *Proc Natl Acad Sci USA,* 2008; 105(10):3963–7.

14. C. Stahl, S. Kubetzko, I. Kaps, S. Seeber, H. Engelhardt, M. Niederweis, *Mol Microbiol,* 2001; 40(2):451–64.

15. P. J. Brennan, H. Nikaido, *Annu Rev Biochem,* 1995; 64.

16. P. Brennan, *Tuberculosis,* 2003; 83(1-3):91–97.

17. L. R. Camacho, P. Constant, C. Raynaud, M. A. Laneelle, J. A. Triccas, B . Gicquel, M. Daffe, C. Guilhot, *J Biol Chem.,* 2001; 276(23):19845–54.

18. J. Trias, R. Benz, *Mol Microbiol*, 1994; 14(2):283–90.

19. P. Draper, *Front Biosci.,* 1998; 3:1253–61.

20. M. Daffé, P. Draper, *Adv Microb Physiol.,* 1998; 39:131–203.

21. G. Lindahl, M. Sta, T. Areschoug, *Clin Microbiol Rev.*, 2005; 18(1):102–127.

22. J. Lin, S. Huang, Q. Zhang, *Microbes Infect.,* 2002; 4(3):325–31.

23. H. H. Niemann, W. D. Schubert, D. W. Heinz, *Microbes Infect. ,* 2004; 6(1):101–12.

24. R. Koebnik, K. P. Locher, G. P. Van, *Mol Microbiol.,* 2000; 37(2):239–53.

25. M. R. Loeb, A. L. Zachary, D. H. Smith, *J Bacteriol*., 1981; 145(1).

26. S. Mizushima, H. Yamada, *Biochim Biophys Acta*., 1975; 375(1):44–53.

27. A. P. Pugsley, M. G. Kornacker, A. Ryter, *Mol Microbiol.,* 1990; 4(1):59–72.

28. H. Song, R. Sandie, Y. Wang, M. A. Andrade-Navarro, M. Niederweis, *Tuberculosis*, 2009; 88(6):526–544.

29. M. Rezwan, M. A. Lanéelle, P. Sander, M. Daffé, *J Microbiol Methods*., 2007; 68(1):32–9.

30. R. Pajon, D. Yero, A. Lage, A. Llanes, C. J. Borroto, *Tuberculosis (Edinb).,* 2006; 86(3-4):290–302.

31. G. V. Heijne, *Nature,* 1998; 396:33–37.

32. J. De Keyzer, C. V. D. Does, A. J. M. Driessen, *Cell Mol Life Sci*., 2003; 60(10):2034–52.

33. H. Nielsen, J. Engelbrecht, S. Brunak, G. V. Heijne, *Protein Eng.,* 1997; 10(1):1–6.

34. L. Sipos, G. V. Heijne, *Eur J Biochem*., 1993; 213(3):1333–40.

35. G. E. Schulz, *Biochim Biophys Acta*., 2002; 1565(2):308–17.

36. G. E. Tusnady, I. Simon, *Bioinformatic***s,** 2001; 17(9):849–50.

37. A. Krogh, B. Larsson, G. V. Heijne, E. L. Sonnhammer, *J. Mol. Biol.,* 2001; 305:567–80.

38. M. C. Walter, T. Rattei, R. Arnold, U. Guldener, M. Munsterkotter, K. Nenova, G. Kastenmuller, P. Tischler, A. Wolling, A. Volz, N. Pongratz, R. Jost, H. W. Mewes, D. Frishman, *Nucleic Acids Res.,* 2009; 37:408–11.

39. A. H. Delcour, *J Mol Microbiol Biotechnol*., 2002; 4(1):1–10.

40. W. C .Wimley, *Curr Opin Struct Biol.*, 2003; 13(4):404–411.

41. M. Faller, M. Niederweis, G. E. Schulz, *Science*, 2004; 303(5661):1189–92.

42. C. Cole, J. D. Barber, G. J. Barton, *Nucleic Acids Res.,* 2008; 36:W197–201.

43. H. Vogel, F. Jahnig, *J Mol Biol.,* 1986; 190(2):191–9.

44. S. I. Narita, S. I. Matsuyama, H. Tokuda, *Arch Microbiol.,* 2004; 182(1):1–6.

45. A. S. Juncker, H. Willenbrock, G. V. Heijne, S. O. Brunak, H. Nielsen, A. Krogh, *Protein Sci.,* 2003; (1994):1652–1662.

46. S. Hayashi, H. C. Wu, *J Bioenerg Biomembr.,* 1990; 22(3):451–71.

47. N. Y. Yu, J. R. Wagner, M. R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S. C. Sahinalp, M. Ester, L. J. Foster, F. S. Brinkman, *Bioinformatics,* 2010; 26(13):1608–15.

48. J. D. Bendtsen, L. Kiemer, A. Fausboll, S. Brunak, *BMC Microbiol.* 2005; 5:58.

49. M. R. Wilkins, E. Gasteiger, A. Bairoch, J. C. Sanchez, K. L. Williams, *Methods Mol Biol*., 1999; 112:531–52.

50. U. Zachariae, A. Koumanov, A. Karshikoff, H. Engelhardt, *Protein Sci.,* 2002; (2):1309–1319.

51. J. C. Camus, M. J. Pryor, C. M.digue, S. T. Cole, *Microbiology*, 2002; 148, 2967–2973.

52. M. Zhou, H. Wu, *Microbiology,* 2009;155: 317–327.

53. K. Drickamer, M.E. Taylor*, Trends Biochem Sci.,*1998; 23(9):321-4.

54. J. S. Chauhan, A. H. Bhat, G. P. S. Raghava, A. Rao, *PloS one*, 2012; 7(7): e40155.

55. T. E-komon, R. Burchmore, P. Herzyk, R. Davies, *BMC Bioinformatics,* 2012; 13:63.

56. J. Trias, V. Jarlier, R. Benz, *Science,* 1992; 258(5087):1479–81.

57. S. Mukhopadhyay, D. Basu, P. Chakrabarti, *J Bacteriol*., 1997; 179(19):6205–7.

58. J. Luirink, G. V. Heijne, E. Houben, J. W. D. Gier, *Annu Rev Microbiol*., 2005; 59:329–55.

59. M. Faller, M. Niederweis, G. E. Schulz, *Science*, 2004; 303(5661):1189–92.

60. S. W. Cowan, T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R. A. Pauptit, J. N. Jansonius, J. P. Rosenbusch, *Nature*, 1992; 358(6389):727–33.

61. R. K. Upreti, M. Kumar, V. Shankar, Proteomics, 2003; 3(4):363-79.

62. A. Banerjee, R. Wang, S. L. Supernavage, S. K. Ghosh, J. Parker, N. F. Ganesh, P. G. Wang, S. Gulati, P. A. Rice, *Journal of Experimental Medicine*, 2002;147–162.

63. R. M. Hamadeh, M. M. Estabrook, P. Zhou, G. A. Jarvis, J. M. Griffiss, *Infection and immunity*, 1995; 4900–4906.

64. C. M. Szymanski, B. W. Wren, *Nat Rev Microbiol.,* 2005; 3(3):225-37.

65. F. Bastida, R. A. Juste, *J Immune Based Therapies and Vaccines.*, 2011; 9(1):8.

66. B. S. Ehrt, M. U. Shiloh, J. Ruan, M. Choi, S. Gunzburg, C. Nathan, Q. W. Xie, L. W. Riley, *J. Exp. Med.* 1997; 186(11):1885–1896.

67. C. H. Kalis, J. W. Hesselink, H. W. Barkema, M. T. Collins, *Am J Vet Res.* 2001; 62(2):270–274.

68. J. E. Uzonna, P. Chilton, R. H. Whitlock, P. L. Habecker, P. Scott, R. W. Sweeney, *Vaccine,* 2003; 21(23):3101–3109.

69. R. D. Linnabary, G. L. Meerdink, M. T. Collins, *Counc Agric Sci Technol.,* 2001; 17:1–10.

70. W. C. Losinger, *J Dairy Res*. 2005; 72(4):425–432.

71. F. L. Leite, K. D. Stokes, S. Robbe-Austerman, J. R. Stabel, *J Vet Diagn Invest.*, 2013; 25(1): 27-34.

72. J. A. Cuff, G. J. Barton, *Proteins,* 1999; 34(4):508–19.

73. J. A. Cuff, G. J. Barton, *Proteins*, 2000; 40(3):502–11.

74. R. M. Sweet, D. Eisenberg, *J Mol Biol.,* 1983; 171(4):479–88.

75. P. G. Bagos, K. D. Tsirigos, T. D. Liakopoulos, *J Proteome Res.,* 2008; 7(12):5082–93.

76. B. Bjellqvist, G. J. Hughes, C. Pasquali, N. Paquet, F. Ravier, J. C. Sanchez, S. H. D. Frutiger, *Electrophoresis*, 1993; 14(10):1023–31.

77. F. M. Carthy, N. Wang, G. B. Magee, B. Nanduri1, M. L. Lawrence, E. B. Camon, D. G. Barrell, D. P. Hill, M. E. Dolan, W. P. Williams, D. S. Luthe, S. M. Bridges, S. C. Burgess, *BMC genomics,* 2006; 7:229.

Figure Legends

Figure 1: Inner Membrane Proteins containing transmembrane helices in MAP. TM-helices were predicted for IMPs using TMHMM[37] and HMMTOP[36] programs. **(a)** Out of 588 predicted exported proteins, 55% of the proteins lacked TM-helix. 9% of the 588 proteins showed at least a single TM-helix. 6% were shown to have 2 TM-helices, 7% with 3 TM- helices, 3% with 4 TM-helices. 5% of the proteins were having 5 TM-helices and 2% were having 6 TM-helices. Additionally, 7 or above TM-helices were shown by 13% of the 588 exported proteins; **(b)** Predicted IMPs, when compared with the known 907 IMPs in MAP, provided by PEDANT [38] Database, 264 IMPs were found to have canonical signal peptides and 643 IMPs were without signal peptide (constituting ~70% of total novel IMPs of MAP).

Figure 2: Amphiphilicity and β-strand content. Transmembrane β-strands with minimal length of five residues were considered, **(a)** out of the 29 known reference OMPs (blue coloured diamonds) from eubacteria, Tol C of *Escherichia coli* obtained a minimum β-strand score of 0.09 and OmpT of *Escherichia coli* scored the lowest amphiphilicity score of 0.18; **(b)** Out of the 324 candidate OMPs, 242 proteins (blue coloured diamonds) exhibited a pattern of alternating hydrophobic and hydrophilic residues in their β-strands. These proteins do not have transmembrane helices and have β-score not less than 0.09 and amphiphilicity score of at least 0.18.

Figure 3: **Functional classification of the OMPs of MAP.** Relative distribution, according to Pasteur Institute functional classification tree [51], of the 83 predicted OMPs (grey bars) with *M. tuberculosis* homologs for eleven functional categories. The functional distribution of the *M. tuberculosis* proteome [51] is shown for comparison (black bars). 27.03% of OMPs were involved in

cell-wall and cell processes (functional class 3). Additionally, 25.00% were conserved hypothetical proteins (functional class 10) and 22.60% were involved in virulence, detoxification, adaptation (functional class 0). None of the predicted OMP was categorized as insertion sequences and phages (functional class 5), stable RNAs (functional class 4) and proteins with unknown functions (functional class 8) respectively.

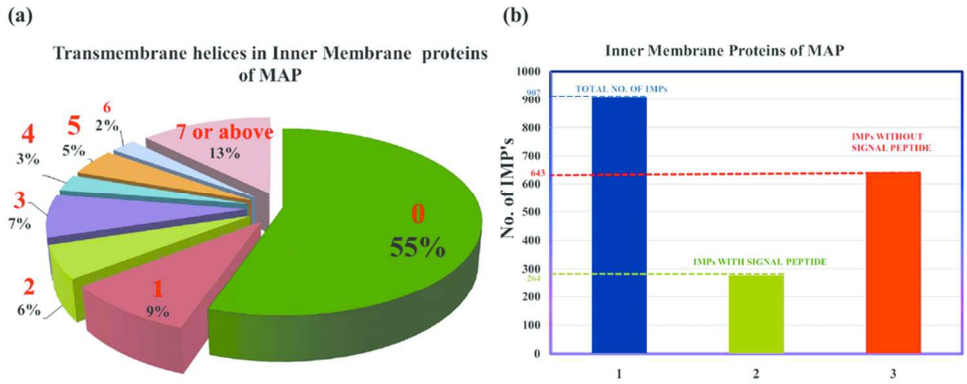
Figure 4: Overall strategy pipeline for computational identification of OMPs. 4, 356 protein sequences of MAP were scanned for the presence of a signal peptide using Target P1.1 program [33]. The 588 mature exported proteins without signal peptide were analyzed using the TMHMM [7]and HMMTOP [36] to predict transmembraneα-helices (Inner Membrane Proteins). β-strand content of 324 proteins without helices were predicted using JPRED3 . 243 proteins with β-strand content above β-score 0.09, which was set as threshold value. The amphiphilicity of the β-strands was calculated using an algorithm of Vogel and Jahnig [43] setting a threshold of amphiphilicity scores 0.18. 242 proteins were predicted to have amphiphilicity score above 0.18. Further, proteins showing similarity to lipoproteins were identified using PRED-LIPO [45] and 38 proteins predicted as lipoproteins were eliminated. Sub-cellular localization of 204 predicted OMPs was carried out using PSORTb [47]. 38 proteins were identified as showing cytoplasmic and periplasmic localization. These 38 proteins were omitted. 63 Non Classical Secretory proteins were identified using SecretomeP server [48]. 2 proteins having only single β-strand were eliminated. 83 putative OMPs were identified out of which 57 top candidate OMPs were shortlisted on the basis of computed Isoelectric points.

TABLE

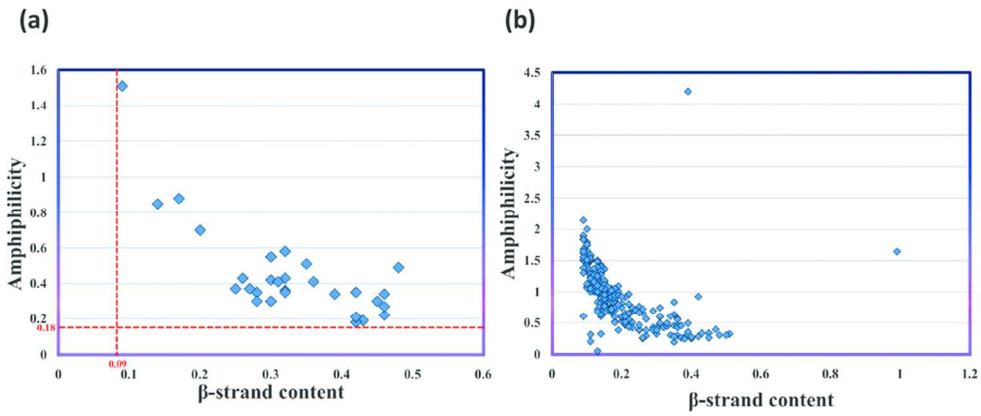
Table 1: Programs used for OMP prediction

SE. NO.	PROGRAMME	ALGORITHM	CUT-OFF USED	PREDICTION
1	TargetP1.1[33]	Neural Network	0.532	Exported Proteins
2	HMMTOP[36] & TMHMM2.0[37]	HMM	-	Transmembrane α-helices

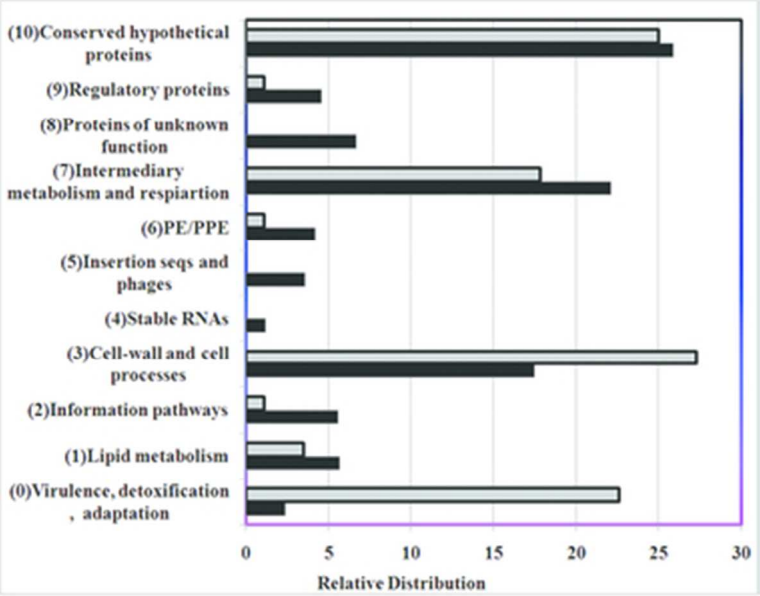
3	JPRED3[42]	Jnet Algorithm	0.09 β-content	Secondary Structure Prediction
4	MATLAB (Proteinpropplot)	Kyte and Doolittle[43]	0.18	Amphiphilicity Score
5	PRED-LIPO[45]	HMM	-	Lipoprotein Prediction
6	PSORTb[47]	SVM	7.5	Subcellular Localization
7	Secretome P2.0[48]	ANN	-	Non-classically Secreted Proteins
8	ProtParam ExPaSy[49]	Edelhoch method	6.5	Isoelectric Point
9	GOanna[77]	BLAST	-	Gene Ontology
10	GlycoPP[54]	SVM	-	O- and N- glycosylation sites



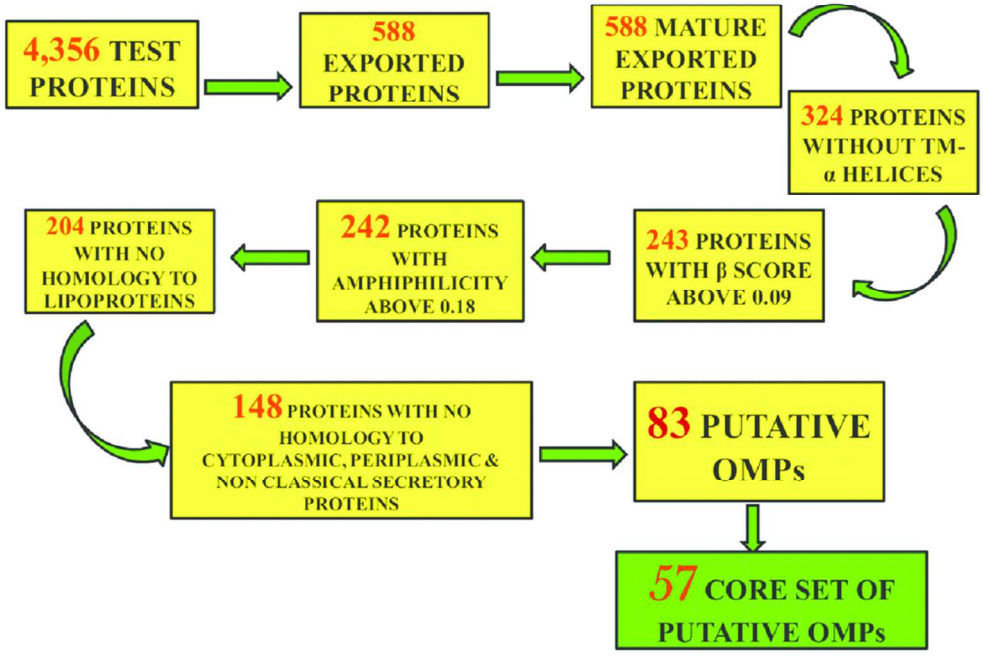
Inner Membrane Proteins containing transmembrane helices in MAP. TM-helices were predicted for IMPs using TMHMM (Ref-37) and HMMTOP (Ref-36) programs. (a) Out of 588 predicted exported proteins, 55% of the proteins lacked TM-helix. 9% of the 588 proteins showed at least a single TM-helix. 6% were shown to have 2 TM-helices, 7% with 3 TM- helices, 3% with 4 TM-helices. 5% of the proteins were having 5 TM-helices and 2% were having 6 TM-helices. Additionally, 7 or above TM-helices were shown by 13% of the 588 exported proteins; (b) Predicted IMPs, when compared with the known 907 IMPs in MAP, provided by PEDANT (Ref-38) Database, 264 IMPs were found to have canonical signal peptides and 643 IMPs were without signal peptide (constituting ~70% of total novel IMPs of MAP).
46x17mm (600 x 600 DPI)



Amphiphilicity and β-strand content. Transmembrane β-strands with minimal length of five residues were considered, (a) out of the 29 known reference OMPs (blue coloured diamonds) from eubacteria, Tol C of Escherichia coli obtained a minimum β-strand score of 0.09 and OmpT of Escherichia coli scored the lowest amphiphilicity score of 0.18; (b) Out of the 324 candidate OMPs, 242 proteins (blue coloured diamonds) exhibited a pattern of alternating hydrophobic and hydrophilic residues in their β-strands. These proteins do not have transmembrane helices and have β-score not less than 0.09 and amphiphilicity score of at least 0.18.
48x20mm (600 x 600 DPI)



Functional classification of the OMPs of MAP. Relative distribution, according to Pasteur Institute functional classification tree (Ref-51), of the 83 predicted OMPs (grey bars) with M. tuberculosis homologs for eleven functional categories. The functional distribution of the M. tuberculosis proteome (Ref-51) is shown for comparison (black bars). 27.03% of OMPs were involved in cell-wall and cell processes (functional class 3). Additionally, 25.00% were conserved hypothetical proteins (functional class 10) and 22.60% were involved in virulence, detoxification, adaptation (functional class 0). None of the predicted OMP was categorized as insertion sequences and phages (functional class 5), stable RNAs (functional class 4) and proteins with unknown functions (functional class 8) respectively.
16x12mm (600 x 600 DPI)



Overall strategy pipeline for computational identification of OMPs. 4, 356 protein sequences of MAP were scanned for the presence of a signal peptide using Target P1.1 program (Ref-33). The 588 mature exported proteins without signal peptide were analyzed using the TMHMM 7and HMMTOP (Ref-36) to predict transmembraneα-helices (Inner Membrane Proteins). β-strand content of 324 proteins without helices were predicted using JPRED3 . 243 proteins with β-strand content above β-score 0.09, which was set as threshold value. The amphiphilicity of the β-strands was calculated using an algorithm of Vogel and Jahnig (Ref-43) setting a threshold of amphiphilicity scores 0.18. 242 proteins were predicted to have amphiphilicity score above 0.18. Further, proteins showing similarity to lipoproteins were identified using PRED-LIPO (Ref-45) and 38 proteins predicted as lipoproteins were eliminated. Sub-cellular localization of 204 predicted OMPs was carried out using PSORTb (Ref-47). 38 proteins were identified as showing cytoplasmic and periplasmic localization. These 38 proteins were omitted. 63 Non Classical Secretory proteins were identified using SecretomeP server (Ref-48). 2 proteins having only single β-strand were eliminated. 83 putative OMPs were identified out of which 57 top candidate OMPs were shortlisted on the basis of computed Isoelectric points.
49x37mm (600 x 600 DPI)