JOURNAL OF PROTEOMICS 75 (2011) 502–510

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Bacterial proteins with cleaved or uncleaved signal peptides of the general secretory pathway

Gustavo A. de Souza[a,b], Nils A. Leversen[a], Hiwa Målen[a], Harald G. Wiker[a,c,*]

[a]Section for Microbiology and Immunology, the Gade Institute, University of Bergen, N-5021 Bergen, Norway
[b]Proteomic Unit of Bergen, Department of Biomedicine, University of Bergen, N-5009 Bergen, Norway
[c]Department of Microbiology, Haukeland University Hospital, N-5021 Bergen, Norway

ARTICLE INFO

Article history:
Received 31 January 2011
Accepted 18 August 2011
Available online 5 September 2011

Keywords:
Mycobacterium tuberculosis
Mammalian cell entry proteins
Mass spectrometry
Membrane proteins
Secreted proteins
Signal peptides

ABSTRACT

Correct protein compartmentalization is a key step for molecular function and cell viability, and this is especially true for membrane and externalized proteins of bacteria. Recent proteomic reports of *Bacillus subtilis* have shown that many proteins with Sec-like signal peptides and absence of a transmembrane helix domain are still observed in membrane-enriched fractions, but further evidence about signal peptide cleavage or soluble protein contamination is still needed. Here we report a proteomic screening of identified peptides in culture filtrate, membrane fraction and whole cell lysate of *Mycobacterium tuberculosis*. We were able to detect peptide sequencing evidence that shows that the predicted signal peptide was kept uncleaved for several types of proteins such as mammalian cell entry (Mce) proteins and PE or PE-PGRS proteins. Label-free quantitation of all proteins identified in each fraction showed that the majority of these proteins with uncleaved signal peptides are, indeed, enriched in the Triton X-114 lipid phase. Some of these proteins are likely to be located in the inner membrane while others may be outer membrane proteins.

1. Introduction

Membrane and exported proteins are crucial players for maintenance and survival of bacterial organisms, and their contribution to pathogenesis and immunological responses make these proteins relevant targets for medical research [1]. Therefore, sorting and processing of these proteins to the correct compartment is essential for bacterial growth and viability. Overall, the bulk of exported proteins are transported by the general secretory Sec-translocase pathway [2,3]. This is performed by recognition of a signal peptide in the nascent pre-protein, which is subsequently transferred to the machinery that executes its translocation across the membrane.

Despite the detailed knowledge of mechanisms for protein secretion and membrane retention that has been gathered in recent years, major problems are still encountered regarding the distinction between signal peptides of soluble exported proteins, lipoprotein signal peptides, and amino-terminal membrane anchors, based on characteristic features of their amino acid sequence [4]. The amino acid termini of these proteins share highly similar charged and hydrophobic regions, and mainly differ with respect to their processing cleavage site or the absence of it. Intriguingly, a recent proteomic characterization of the *Bacillus subtilis* membrane protein fraction (MPF) identified 31 proteins with the very characteristic Sec-type / signal peptidase I AXA motif. This would characterize them as soluble extracellular proteins, but the study showed they were retained in the membrane [5,6]. Unfortunately, the authors were not able to address if these proteins were true membrane proteins, or if they were residual cytoplasmic soluble proteins still present after the membrane protein extraction. In addition, no peptide upstream of the signal peptidase I cleavage site was characterized.

* *Corresponding author at: Section for Microbiology and Immunology, the Gade Institute, University of Bergen, N-5021 Bergen, Norway*. Tel.: +47 55974650; fax: +47 55974689.
E-mail address: harald.wiker@gades.uib.no (H.G. Wiker).

1874-3919
doi:10.1016/j.jprot.2011.08.016

Proteomic analysis has helped in characterization of secreted proteins with a known start of the mature sequence after cleavage by signal peptidase I. For example, the set of cleaved proteins of *Mycobacterium tuberculosis* was extended to 57 through the use of high-accuracy MS instrumentation and advanced database design [7,8]. This data set was later used to validate signal peptide prediction algorithms [9], and the study found the Hidden Markov model (HMM) of SignalP v3.0 to be the most accurate tool. The HMM correctly predicted the presence or absence of a signal peptide, and the correct cleavage site in a high proportion of the proteins. A majority of the proteins in the validation set had an AXA signal peptidase I cleavage motif, suggesting the importance of alanine in the −1 and −3 position relative to the cleavage site. However, when HMM prediction was applied to all of the annotated proteins in the *M. tuberculosis* genome, many predicted transmembrane proteins were assigned as exported, in the same way as observed with *B. subtilis*[4]. The question is therefore whether predicted signal peptides are cleaved or not. This issue is of particular interest for proteins with several transmembrane helices, but is also important for proteins without transmembrane regions or proteins with untypical signal peptides such as PE or PE-PGRS proteins.

We have recently investigated a Triton X-114 generated MPF of *M. tuberculosis* H37Rv and identified 6741 peptides from 1417 different proteins [10]. Here we searched this database to identify peptides within predicted signal sequences, and identified 40 proteins with predicted signal peptides that were uncleaved. As an additional control, we analyzed whole cell lysates (WCL) of *M. tuberculosis* H37Rv, identifying 15537 peptides from 1874 proteins. We used a label-free quantitation method named emPAI [11,12] to estimate mol% representation for individual proteins in each fraction, and to check if proteins observed in the MPF were truly enriched or not. From the 40 proteins with uncleaved signals, 3 proteins were discarded because they had predicted lipoprotein cleavage prediction closer to the N-terminal than the observed peptide. Seven were found to be more abundant in WCLs indicating that they are soluble proteins. Our data support that uncleaved predicted signal peptides might be essential for correct membrane anchoring for certain membrane proteins. For less hydrophobic proteins such as PE/PE-PGRS proteins the predicted signal peptide may serve as a membrane anchor. Outer membrane proteins, for example the Mce proteins, might fold their alpha helical predicted signal peptide in close proximity to the typical transmembrane outer membrane β-barrell.

2. Materials and methods

2.1. Preparation of whole cell lysate (WCL)

The *M. tuberculosis* H37Rv (ATCC 27294) bacilli were cultured on Middelbrook 7H10 agar plates with OADC enrichment (BD Difco) at 37 °C and 5% CO_2 for 3–4 weeks. Bacterial colonies were harvested by using an extraction buffer consisting of phosphate-buffered saline (PBS), pH 7.4 with freshly added Roche Protease Inhibitor Cocktail (Complete, EDTA-free, Roche Gmbh, Germany). Six hundred microliters of this extraction buffer was added to each agar plate and the mycobacterial colonies were gently scraped off the agar surface using a cell scraper. Aliquots of the resulting pasty bacterial mass were transferred into 1.5 ml Eppendorf tubes and extraction buffer was added to 1.5 ml total volume. The mixture was gently vortexed and bacteria spun down for removal of the supernatant. The washing procedure was repeated 3 times to remove soluble extracellular proteins. Washed bacilli were transferred to 2 ml cryo-tubes with O-rings (Sarstedt, Norway) containing 250 μl of acid washed glass beads (≤106 μm; Sigma-Aldrich, Norway) and an additional 600 μl of extraction buffer, and stored at −80 °C until protein extraction was performed. To prepare WCL, the mycobacteria were disrupted mechanically by bead beating in a Ribolyser (Hybaid, UK) at max speed for 45 s.

2.2. Membrane protein fraction (MPF)

Triton X-114 phase-separation was used to isolate lipophilic proteins following the method of Bordier [13,14]. In brief, 3–4 week old bacilli were lysed by bead beating and centrifuged, initially at 2300 *g* to remove unbroken cells and cell-wall debris. Triton X-114 was added to the supernatant (final detergent concentration 2%, v/v) and the suspension was stirred at 4 °C for 20 min to obtain the protein extract in a single phase. Residual insoluble matter was removed by centrifugation at 15,700 *g* for 10 min, and the phases were separated after 10 min incubation at 37 °C. Upper (aqueous) and lower (detergent) phases were collected and washed four times. Proteins in the pooled aqueous and detergent phases were recovered by acetone precipitation.

2.3. Culture filtrate (CF)

M. tuberculosis H37Rv was cultured as surface pellicle on Sauton medium for 3 weeks without shaking. Bacteria were removed by filtration and the CF was concentrated by 80% ammonium sulfate precipitation. Precipitated proteins were dissolved in buffer, dialyzed against distilled water, lyophilized and dissolved in loading buffer for SDS-PAGE.

2.4. Gel electrophoresis and in-gel digestion of proteins

Fifty micrograms of protein sample was mixed with 5 μl sodium-dodecyl-sulphate (SDS) loading buffer containing 10 mM DTT, and boiled for 5 min before separation on a 4–12% SDS-PAGE (NuPAGE kit, Invitrogen, Carlsbad, CA, U.S.A.). The protein migration was allowed to proceed until the bromophenol dye had migrated to the bottom of the gel. The protein bands were visualized with Coomassie Brilliant Blue R-250 staining kit (Invitrogen). Protein lanes were excised along the visible protein bands ranging from ~3 kDa to ~188 kDa and washed twice with 50% acetonitrile (ACN) at room temperature (RT). The gel pieces were dehydrated by incubating them with 50 μl 100% ACN for 20 min at RT. The proteins were reduced using 10 mM DTT in water at 58 °C for 1 h, and alkylated with 55 mM iodoacetamide in 100 mM NH_4HCO_3 for 45 min at RT. The gel pieces were dehydrated by 100% ACN as described above, and rehydrated in 50 mM NH_4HCO_3 containing 0.125 μg of sequence-grade trypsin (Promega, Madison, U.S.A.) overnight at 37 °C. The trypsin reaction was quenched using 1% trifluoracetic acid. The digested peptides were eluted by incubating the

gel pieces with 50 μl 50% ACN for 20 min at RT two times, plus a final wash with 100% ACN for 10 min. Peptide mixtures were then desalted using STAGE-tips packed with C18 resin (3 M, USA) [15].

2.5. Mass spectrometry

All experiments were performed on a Dionex Ultimate 3000 nano-LC system (Sunnyvale CA, USA) connected to a linear ion trap—Orbitrap (LTQ-Orbitrap) mass spectrometer (ThermoElectron, Bremen, Germany) equipped with a nanoelectrospray ion source. For liquid chromatography separation we used an Acclaim PepMap 100 column (C18, 3 μm, 100 Å) (Dionex, Sunnyvale CA, USA) capillary of 12 cm bed length 100 μm ID self packed with Reprosil_Pur C18-aq (Dr. Maisch Gmbh, Ammerbuch-Entringen, Germany). The flow rate used was 0.3 μL/min for the nano column, and the solvent gradient used was 7% B to 40% B in 87 min, then 40–80% B in 8 min. Solvent A was aqueous 2% ACN in 0.1% formic acid, whereas solvent B was aqueous 90% ACN in 0.1% formic acid.

The mass spectrometer was operated in the data-dependent mode to automatically switch between Orbitrap-MS and LTQ-MS/MS acquisition. Survey full scan MS spectra (from *m/z* 300 to 2000) were acquired in the Orbitrap with resolution R = 60,000 at *m/z* 400 (after accumulation to a target of 1,000,000 charges in the LTQ). The method used allowed sequential isolation of the most intense ions, up to six, depending on signal intensity, for fragmentation on the linear ion trap using collisionally induced dissociation at a target value of 10,000 charges.

For accurate mass measurements the lock mass option was enabled in MS mode and the polydimethylcyclosiloxane ions generated in the electrospray process from ambient air were used for internal recalibration during the analysis [16]. Target ions already selected for MS/MS were dynamically excluded for 60 s. General mass spectrometry conditions were: electrospray voltage, 1.5 kV; no sheath and auxiliary gas flow. Ion selection threshold was 500 counts for MS/MS, and an activation Q-value of 0.25 and activation time of 30 ms were also applied for MS/MS.

2.6. Protein identification

MS/MS peak lists from individual 60 RAW files (10 from CF samples [8], 20 from MPFs [10] and 30 from WCLs) were generated using DTA SuperCharger package, version 1.29, available at the MSQuant validation tool (see below). Protein identification was performed by searching the data separately against *M. tuberculosis* H37Rv protein database available at the Tuberculist website, version R11 (genolist.pasteur.fr/tuberculist/) and CMR-TIGR database. The databases were in-house modified to also contain reversed sequences of all entries as a control of false-positive identifications during analysis. Common contaminants, such as keratins, BSA, trypsin, were also added to the database. We used MASCOT Deamon for multiple searches submission on a local Mascot server v2.1 (Matrix Science). The search parameters used were: Enzyme: Trypsin/P (no proline restriction); Maximum missed cleavages: 3; Carbamidomethyl (C) as fixed modification; N-acetyl (Protein), Oxidation (M), pyro-glu (Q) and pyro-glu (E) as variable modifications; Peptide mass tolerance of ±15 ppm; MS/MS mass tolerance of 0.5 Da. Under these criteria, Mascot indicated a minimal score of 22 for $p \leq 0.01$ and 15 for $p \leq 0.05$. All data had an average mass accuracy of 2.8 ppm. Spectra and protein validation were performed using an open source software called MSQuant (version 1.5a61), largely used for LC–MS/MS data analysis [17]. Proteins were validated statistically, based on the score of their individual peptides. Proteins with at least two tryptic peptides with a minimal score of 22 for each (protein false-positive probability of 0.01%), or those with only 1 peptide but a MS/MS score higher than 38 were accepted (protein false-positive probability lower than 0.25%). Using these criteria, all MS/MS identifications of peptides present in entries with reversed sequences (i.e. false-positive identifications) were not validated, since none of the reversed proteins were identified with 2 peptides with a score higher than 21 each or 1 peptide with a score higher than 38 (the highest Mascot score for a peptide from the reversed database was 32—data not shown). Identifications with only one unique peptide were accepted only after manual validation. Quality criteria for manual validation were the assignment of major peaks, the occurrence of uninterrupted y- or b-ion series of at least 3 consecutive amino acids, the preferred cleavages N-terminal to proline bonds and C-terminal to Asp or Glu bonds, and the possible presence of a2/b2 ion pairs.

2.7. Estimation of protein abundance

Protein abundance expressed as emPAI values was calculated using the number of observable peptides and the number of observed parent ions per identified peptide. The number of observable (or expected) peptides for a protein was calculated through *in silico* trypsin digestion of the *M. tuberculosis* H37Rv database, and the resulting peptide fragments were compared with the scan range of the mass spectrometry. The emPAI values were calculated using a script developed at the Keio University (Japan) (available at http://empai.iab.keio.ac.jp/), using the following parameters: trypsin enzyme, Carbamidomethyl (C) fixed modification, mass range from 300 to 8000 Da, no retention time filtering, Bold red peptides only (i.e., unique peptides in the Mascot result), peptides filtered by peptide Mascot score higher than 22. The Protein Abundance Index (PAI) was obtained by division of the observed parent ions with the number of theoretical observable peptides; emPAI is obtained using the formula $\text{emPAI} = 10^{\text{PAI}} - 1$. To obtain the concentration of a protein in the sample, its emPAI value was divided by the sum of all emPAI values in the sample, and the result multiplied by 100 (resulting in an estimate of the mol percentage of the protein in that sample) [11,12].

2.8. Sequence analysis and prediction

Annotated protein sequences from the *M. tuberculosis* H37Rv genome [18] were submitted to SignalP HMM v3.0 [19], and proteins with signal peptide prediction values above 0.500 were included in this dataset. Transmembrane helix and hydrophobic region predictions were done using the TMHMM v2.0 server from http://www.cbs.dtu.dk[20]. Protein sequence alignment and motif visualization was done using Protein Sequence Logos [21,22].

3. Results

3.1. Proteomic analysis of *M. tuberculosis* H37Rv fractions

We performed a clustered analysis of *M. tuberculosis* H37Rv CF, MPF and WCLs. In total, we report the identification of 2182 proteins present in one or more of the samples. The Supplemental File S1 reports all peptide sequences identified in this work, and the proteins they originated from (both Tuberculist and CMR-TIGR entry names are given). In addition, acquisition data relevant to the quality of the identifications are given, such as observed and theoretical mass measurements, mass accuracy, Mascot scoring values, etc. Peptide sequences in red are unique peptides present only in the Tuberculist database, and those in blue are unique to the CMR-TIGR database, due to differences of translational start site choices as discussed elsewhere [8,23]. Protein entry names similarly shown in red or blue means that a gene was only annotated in the respective database represented by that color. The sheet named 'Protein list' is a simplified representation of the proteins identified in each fraction in a merged list, with the Tuberculist entry name given preference except for CMR-TIGR specific genes.

3.2. Calculation of protein abundance

The data from the peptide list of Supplemental File S1 was submitted to the emPAI calculation tool, and emPAI values of individual proteins identified in each *M. tuberculosis* H37Rv fraction was obtained. Prior to the submission of the data, we manually merged peptides identified for genes Rv1198, Rv1793 and Rv2346c, and renamed this protein group 'ESATx-like'. This was performed because the protein products of these 3 genes (94 amino acids long) share 86 identical amino acids, therefore it is practically impossible to determine which of those contributed more for detected tryptic peptides that are shared among them.

Protein abundance in the fraction was calculated as mol % by dividing individual emPAI values by the sum of all values in the fraction, and multiplied by 100. This also contributes to the normalization of the sample, since differences might be observed due to variation in instrument efficiency over time. Supplemental File S2 reports all emPAI values obtained for all identifications obtained in this article. Mol % values were then compared among the different fractions in order to estimate enrichment or not of a certain protein in MPF or CF. To obtain information about the "purity" of the different fractions, mol % values were summed for selected groups of proteins in each of the protein extracts as shown in Table 1. These results shows that the CF was considerably enriched for secreted proteins with only minimal contamination from the WCL or the MPF and that the latter two had minimal amounts of proteins from the CF. On the other hand, there was considerable overlap in the protein content between the WCL and the MPF. Intracellular proteins from functional group 2, information pathways were enriched by a factor of 2 in the WCL as compared to the MPF. Many membrane proteins were also found in the WCL, but there was an enrichment of membrane proteins estimated to about 6 times in the MPF as compared to the WCL (i.e. proteins with more than 3 predicted transmembrane regions). Supplemental File S2 contains a sheet named 'Membrane Protein Set', where 30 proteins with multiple predicted TMH domains were validated to see if emPAI values correctly showed their probable enrichment in the MPF. All but 2 were found to be membrane enriched. Fig. 1 illustrates examples of the emPAI result: The kinase PknB (Rv0014c), a well characterized membrane protein was found to be enriched in the MPF, while the fructose-biphosphate aldolase Fba (Rv0363c), an intracellular metabolic protein, was more abundant in the WCL. It should be noted that Fba was identified in both the MPF and CF. However, the abundance analysis within all 3 fractions shows that this protein was predominantly found in WCL, suggesting the observations of Fba in the MPF- and CF fractions is due to carry-over contamination of intracellular proteins during sample preparation. If the MPFs were analyzed separately, one might mistakenly assume that Fba is a membrane protein or membrane-associated, due to its presence in a sample prepared in order to target membrane proteins.

3.3. Cleaved and uncleaved signal peptides

Tuberculist protein sequences version R11 were submitted to SignalP prediction using the hidden Markov model, and the proteins with significant signal peptide prediction score values above 0.500 were selected (data not shown). A total of 521 proteins were predicted to have signal peptides, of which we identified 296 in one or more fractions. The identified peptides for these proteins were aligned in the database sequence,

Table 1 – Evaluation of protein composition of the protein extracts. The table shows the sum of mol % values (emPAI) and number (N) of protein observations for different groups of proteins in the CF, MPF and the WCL. The selected groups of proteins were proteins with a predicted signal peptide according to the hidden Markov model of SignalP v3.0; proteins with 1, 1 or 2, >=2 and >=3 predicted transmembrane regions (TMH); and proteins that belong to functional group 2, information pathways. Almost all of the proteins in the latter group are without predicted transmembrane regions.

Protein group	emPAI_CF	N_CF	emPAI_MPF	N_MPF	emPAI_WCL	N_WCL
Predicted signal peptide	82.28	126	18.04	207	8.14	216
Predicted TMH=1	73.93	82	8.64	135	4.53	152
Predicted TMH=1 or 2	73.98	86	14.67	179	6.82	194
Predicted TMH>=2	0.07	9	11.97	217	3.32	142
Predicted TMH>=3	0.02	5	5.94	173	1.03	100
Information pathways	0.14	24	3.39	106	6.42	148

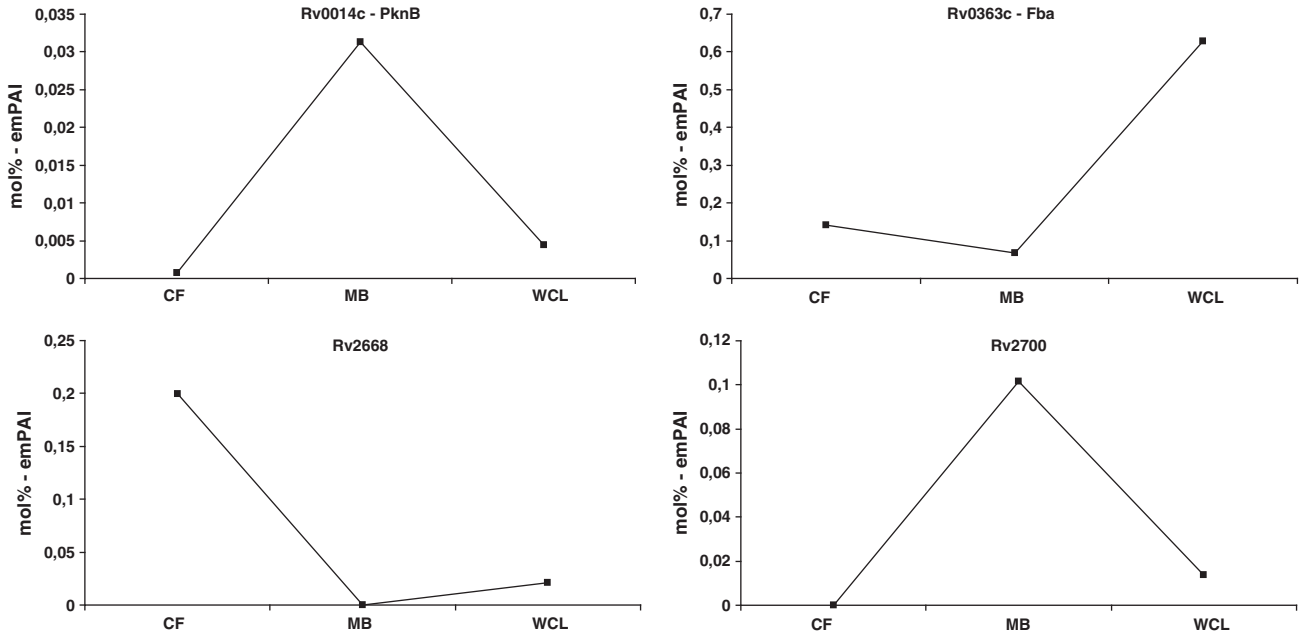

Fig. 1 – emPAI analysis of individual proteins in *M. tuberculosis* H37Rv fractions. Two well characterized proteins: one from the membrane, PknB—transmembrane serine/threonine-protein kinase B (upper left panel), and one from the intracellular space, Fba—probable fructose-bisphosphate aldolase (upper right panel), had emPAI value validations that confirm their presence in the correct, expected fraction. Two examples of proteins with predicted signal peptidase I cleavage sites are shown: Rv2668—possible exported alanine and valine rich protein (lower left panel), in which signal peptidase I processing was detected, is correctly validated in CF fractions; on the other hand, Rv2700—possible conserved secreted alanine rich protein (lower right panel), which presented peptide sequences upstream of the predicted signal peptidase I cleavage site, is observed enriched in the MPF.

as shown in Supplementary file S3. Amino acids marked with a grey background represent the first downstream amino acid relative to the predicted signal peptidase I cleavage site. Sequences listed in bold red are processed N-terminal peptides that have previously been identified by our group [8,9]. Peptides marked with a green background contain amino acids upstream of the predicted cleavage site.

We initially identified tryptic peptides upstream of the signal peptidase I predicted cleavage site for 40 of the 296 identified proteins. The emPAI algorithm was then employed to show that the amino acids originated from truly exported or membrane-enriched proteins, and not from precursors or intracellular proteins wrongly predicted as exported ones. Fig. 1 shows the abundance profile of proteins Rv2668 (Possible exported alanine and valine rich protein) and Rv2700 (Possible conserved secreted alanine rich protein). Rv2668 was more abundant in CF, and indeed its observed N-terminal peptide showed that this protein is cleaved by signal peptidase I. On the other hand, Rv2700 was enriched in the MPF, and its N-terminus was identified as the predicted translational start site, without signal peptidase cleavage.

From the 40 proteins with identified peptides upstream of the predicted cleavage site, 7 were classified as soluble intracellular proteins by the above analysis and were discarded from the group (in red in Supplementary file S4) and 3 were discarded because they had a Signal peptidase II cleavage site upstream of the observed peptide. This file also reports the number of predicted transmembrane helix domains for the proteins with uncleaved signal peptides, plus prediction if the signal peptide might be inside a transmembrane span. All values and graphics for emPAI calculations of all proteins with cleaved or uncleaved signals are given. The final list of proteins having an uncleaved N-terminus is given in Table 2.

3.4. *Sequence alignment of protein N-terminals*

Once we delimited truly exported proteins in our set, we aligned their N-terminal sequences around the predicted cleavage site in order to determine any sequence feature common to those proteins (Fig. 2B). As a positive control we also aligned the proteins with identified cleaved signal peptides (Fig. 2A). The cleavage site is shown in position 24 in both alignments. In the control set, the presence of the AXA motif in positions −1 and −3 relative to the cleavage site is evident, while in the uncleaved signal set, the alanine signature in position −3 seems to be disrupted. In addition, the hydrophobic region of the control set seems to favor alanine more often, while this is not seen in the uncleaved signal set. However, overall the hydrophobic region of the uncleaved signal set seems to keep its hydrophobic characteristics.

4. Discussion

In this work, we took advantage of a descriptive method with high coverage capacity of mass spectrometry-based proteomics to investigate peptide products as evidence that corroborate or

Table 2 – Proteins with uncleaved signal peptide.

	TMhelix	TMH inside sig	Secondary structure prediction
Rv0011c	2	No	Mainly helix
Rv0170	1	Yes	Helix+strand
Rv0171	1	Yes	Helix+strand
Rv0285	0	–	Mainly helix
Rv0426c	2	Yes	All helix
Rv0431	1	Yes	Mixed helix+strand
Rv0514	2	Yes	Helix+strand
Rv0544c	2	Yes	Mainly helix
Rv0592	1	Yes	Helix+strand
Rv0677c	1	Yes	Mainly strand
Rv0732	9	No	Mainly helix
Rv0734	0	–	Helix+strand
Rv0832	0	–	Mainly helix
Rv0906	0	–	Mainly strand
Rv1236	5	Yes	Mainly helix
Rv1339	0	–	Mainly strand
Rv1386	0	–	Mainly helix
Rv1468c	0	–	Mixed helix+strand+coil
Rv1488	1	Yes	Helix+strand
Rv2216	0	–	Helix+strand
Rv2563	4	Yes	Helix+strand
Rv2612c	3	No	All helix
Rv2700	1	Yes	Mixed helix+strand
Rv2959c	0	–	Mixed helix+strand
Rv2982c	0	–	Mixed helix+strand
Rv3069	4	Yes	All helix
Rv3101c	4	Yes	Mainly helix
Rv3587c	1	Yes	Mainly strand
Rv3682	1	Yes	Helix+strand
Rv3851	2	Yes	All helix

not with Sec-type signal peptides in *M. tuberculosis*. Quantitative information about cell compartment enrichment was also assessed to guarantee that the data collected was indeed relevant to exported or membrane proteins, and not from intracellular proteins that remained detectable in the MPFs or CF.

4.1. *Signal peptide structure*

The structure of signal peptides is well known, and many reviews have highlighted the features [24,25]. Overall, the signal is composed of three portions, a positively charged N-terminal, followed by a hydrophobic region and finally, a polar C-terminal part containing the AXA motif in positions −3 and −1 to the cleavage site. While there is no evidence up to date that the signal peptide remains uncleaved for membrane proteins, this feature is expected [24]. But structural differences between Sec-like exported or Sec-like membrane proteins are minimal, mostly differences in size of the N-terminal and the hydrophobic region. One could hypothesize that such differences in peptide length would be a key factor for signal peptidase I accessability, therefore deciding the fate of an exported protein. However, it has been demonstrated that alterations in the length of the N-terminal or the hydrophobic region of an exported protein in *Escherichia coli* and *B. subtilis* did not alter signal peptidase I activity or accuracy significantly [26]. Therefore, we could assume that, while such regions might be key factors for signal peptidase I accessability, they are most probably irrelevant for the export machinery to distinguish between exported or membrane-attached proteins.

4.2. *Sequence characteristics of uncleaved signal peptides*

We might expect that proteins with uncleaved signal peptides observed in *M. tuberculosis* present distinct sequence features that, while predicted as Sec-like exported proteins, are not enough to guarantee export. Thus, the importance of detecting true uncleaved signal peptides in our dataset gave us the opportunity to focus on sequence analysis of the relevant proteins. Alignment of 30 N-terminal regions from proteins with uncleaved signals and 24 with cleaved signal peptides demonstrate that, while the cleaved group seems to be more alanine oriented in the hydrophobic region, the uncleaved group nonetheless have hydrophobic residues in that region overall. As expected, the alanine in the −1 position is highly conserved, since substitution of that amino acid with amino acids having larger side chains will physically block the access of signal peptidase I, and only modification with Glycine or Serine are generally acceptable [27]. While the same is true for the position −3, it is generally considered that this position is more flexible (due to its distance to the cleavage site) and the substitution of Alanine by Serine, Glycine, Valine, Threonine, Leucine or Isoleucine is often observed in precursor exported proteins of gram positive bacteria [25]. However, in our dataset, it is apparent that the cleaved group still favors Alanine as the preferred choice for position −3, while the

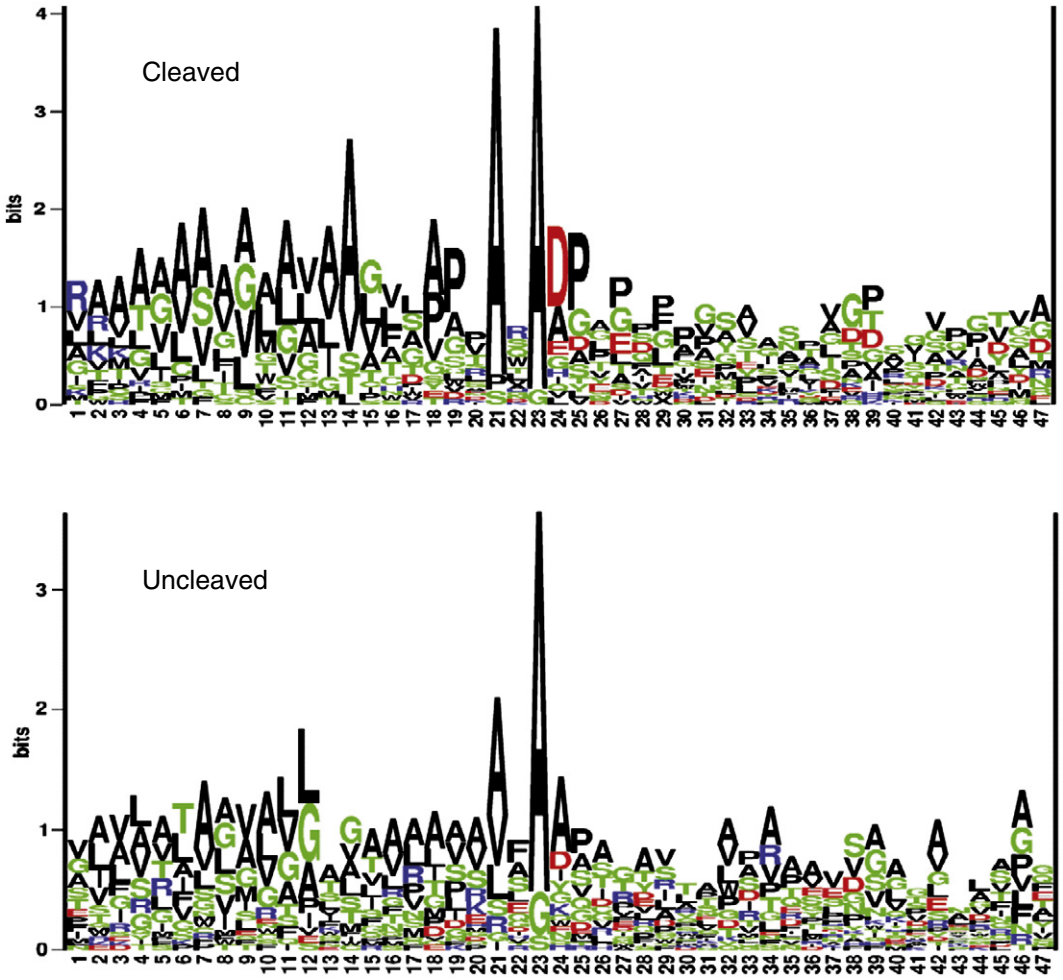

Fig. 2 – Alignment of proteins confirmed to be cleaved or to be uncleaved in the predicted signal peptidase I site. The top row shows alignment of N-terminal sequences of proteins with detected processed sequences [8], where the AXA motif at –3 and –1 position is clearly evident. Bottom row shows the alignment of proteins which were truly enriched in MPFs and with detected sequences upstream to the predicted signal peptidase I site.

uncleaved peptide group shows a greater variation of choices. These data indicate that *M. tuberculosis* signal peptidase I may have more stringent structural rules for the AXA motif that are initially described for *E. coli* or *B. subtilis*.

4.3. *Position of the predicted cleavage site*

Finally, another possibility why predicted signal peptides are not cleaved would be that the cleavage site itself is inside the membrane span, and therefore inaccessible to the signal peptidase I [24]. For our 30 proteins, 18 had predicted transmembrane regions which overlapped with the cleavage site region. However, 15 of our identifications did not fit to this explanation, as there was no overlap between predicted transmembrane region and the cleavage site or the lack of predicted transmembrane region in the signal peptide. Interestingly, 10 of the proteins with uncleaved signal peptides come from proteins with no prediction of transmembrane helix domains, but with overall hydrophobic (grand average of hydropathicity, GRAVY) scores assigned (data not shown). In general, researchers in the field tend to classify such proteins as membrane-associated instead of membrane-attached, or outer membrane proteins if the protein contain predicted β-strands [28].

4.4. *Outer membrane proteins with uncleaved signal peptides*

Interestingly, 10 previously predicted outer membrane proteins [28] were shown to possess uncleaved signal peptides in our data. Curiously, all of those were predicted as outer membrane proteins since they lacked transmembrane helix domains in the mature protein sequence (i.e. after the signal peptide is cleaved). However, it is also important to address that some of the signal peptides might be part of transmembrane helix domains themselves, which implies the site is inaccessible to signal peptidase I as stated above, and as observed for the Mce1B, Mce1C and Mce2D proteins in File S4. Mce proteins were initially characterized as part of an operon containing two integral ABC transporter membrane proteins YrbEA and YrbEB, plus 6 *mce* genes (*mceA*, *mceB*, *mceC*, *mceD*, *mceE* and *mceF*), and 4 copies of this operon is observed in the genome (*Mce*1–4) [18,29]. Mce proteins are important virulence

factors of the organism. The transmembrane regions of outer membrane proteins are typically β-barrel structures. Secondary structure predictions show that all 6 mce proteins in each operon have a very similar structure with a domain of about 100 amino acids with several highly preserved strands in the N-terminal half of the proteins. A structural model of Mce1A [30] also confirms the β-pleated structure in this region. Our view is that even if these proteins have uncleaved signal peptides with an N-terminal transmembrane helix region, they may still be outer membrane proteins.

4.5. *Conclusion*

In conclusion, we report and describe peptides upstream of signal peptidase I cleavage sites in truly membrane-enriched proteins, and we did not observe any peptides from the upstream region for proteins with confirmed signal peptidase I cleavage. Our data shows that for some of these identifications, inaccessibility of the cleavage site or differences in signal peptide structure are not enough to explain why signal peptidase I failed to recognize the signal. It may also indicate that amino acid substitutions in the position -3 of the AXA motif have a bigger penalty for *M. tuberculosis* when compared to other bacterial models. Finally, for similar identifications in proteins with no transmembrane region prediction, the missed cleavage by signal peptidase I might be essential for membrane retention of the protein, and for inner/outer membrane proteins' function and structure.

Supplementary materials related to this article can be found online at doi:10.1016/j.jprot.2011.08.016.

Acknowledgments

This work was supported by grants from the Regional Health Authorities of Western Norway (Projects 911077, 911117 and 911239) and by the National Programme for Research in Functional Genomics in Norway (FUGE) funded by the Norwegian Research Council (Project 175141/S10).

REFERENCES

[1] Daffe M, Etienne G. The capsule of *Mycobacterium tuberculosis* and its implications for pathogenicity. Tuber Lung Dis 1999;79:153–69.
[2] Nouwen N, Berrelkamp G, Driessen AJ. Bacterial sec-translocase unfolds and translocates a class of folded protein domains. J Mol Biol 2007;372:422–33.
[3] Xie K, Dalbey RE. Inserting proteins into the bacterial cytoplasmic membrane using the Sec and YidC translocases. Nat Rev Microbiol 2008;6:234–44.
[4] Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. Microbiol Mol Biol Rev 2000;64:515–47.
[5] Bunai K, Ariga M, Inoue T, Nozaki M, Ogane S, Kakeshita H, et al. Profiling and comprehensive expression analysis of ABC transporter solute-binding proteins of *Bacillus subtilis* membrane based on a proteomic approach. Electrophoresis 2004;25:141–55.
[6] Eymann C, Dreisbach A, Albrecht D, Bernhardt J, Becher D, Gentner S, et al. A comprehensive proteome map of growing *Bacillus subtilis* cells. Proteomics 2004;4:2849–76.
[7] Målen H, Berven FS, Fladmark KE, Wiker HG. Comprehensive analysis of exported proteins from *Mycobacterium tuberculosis* H37Rv. Proteomics 2007;7:1702–18.
[8] de Souza GA, Målen H, Søfteland T, Saelensminde G, Prasad S, Jonassen I, et al. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using *Mycobacterium tuberculosis* as an example. BMC Genomics 2008;9:316.
[9] Leversen NA, de Souza GA, Målen H, Prasad S, Jonassen I, Wiker HG. Evaluation of signal peptide prediction algorithms for identification of mycobacterial signal peptides using sequence data from proteomic methods. Microbiology 2009;155:2375–83.
[10] Målen H, Pathak S, Søfteland T, de Souza GA, Wiker HG. Definition of novel cell envelope associated proteins in Triton X-114 extracts of *Mycobacterium tuberculosis* H37Rv. BMC Microbiol 2010;10:132.
[11] Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol Cell Proteomics 2005;4:1265–72.
[12] Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ, et al. Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 2008;9:102.
[13] Bordier C. Phase separation of integral membrane proteins in Triton X-114 solution. J Biol Chem 1981;256:1604–7.
[14] Målen H, Berven FS, Søfteland T, Arntzen MO, D'Santos CS, de Souza GA, et al. Membrane and membrane-associated proteins in Triton X-114 extracts of *Mycobacterium bovis* BCG identified using a combination of gel-based and gel-free fractionation strategies. Proteomics 2008;8:1859–70.
[15] Rappsilber J, Ishihama Y, Mann M. Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 2003;75:663–70.
[16] Olsen JV, de Godoy LM, Li G, Macek B, Mortensen P, Pesch R, et al. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics 2005;4:2010–21.
[17] Mortensen P, Gouw JW, Olsen JV, Ong SE, Rigbolt KT, Bunkenborg J, et al. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. J Proteome Res 2010;9:393–403.
[18] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 1998;393:537–44.
[19] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 2004;340:783–95.
[20] Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–80.
[21] Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res 1990;18:6097–100.
[22] Gorodkin J, Heyer LJ, Brunak S, Stormo GD. Displaying the information contents of structural RNA alignments: the structure logos. Comput Appl Biosci 1997;13:583–6.
[23] de Souza GA, Søfteland T, Koehler CJ, Thiede B, Wiker HG. Validating divergent ORF annotation of the *Mycobacterium leprae* genome through a full translation data set and peptide identification by tandem mass spectrometry. Proteomics 2009;9:3233–43.

[24] Tjalsma H, van Dijl JM. Proteomics-based consensus prediction of protein retention in a bacterial membrane. Proteomics 2005;5:4472–82.
[25] van Roosmalen ML, Geukens N, Jongbloed JD, Tjalsma H, Dubois JY, Bron S, et al. Type I signal peptidases of Gram-positive bacteria. Biochim Biophys Acta 2004;1694: 279–97.
[26] Carlos JL, Paetzel M, Brubaker G, Karla A, Ashwell CM, Lively MO, et al. The role of the membrane-spanning domain of type I signal peptidases in substrate cleavage site selection. J Biol Chem 2000;275:38813–22.
[27] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997;10: 1–6.
[28] Song H, Sandie R, Wang Y, Andrade-Navarro MA, Niederweis M. Identification of outer membrane proteins of *Mycobacterium tuberculosis*. Tuberculosis (Edinb) 2008;88: 526–44.
[29] Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW. Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. Science 1993;261:1454–7.
[30] Das AK, Mitra D, Harboe M, Nandi B, Harkness RE, Das D, et al. Predicted molecular structure of the mammalian cell entry protein Mce1A of *Mycobacterium tuberculosis*. Biochem Biophys Res Commun 2003;302:442–7.