REGULAR ARTICLE

Analysis of the cytosolic proteome of *Halobacterium salinarum* and its implication for genome annotation

Andreas Tebbe*, Christian Klein*, Birgit Bisle, Frank Siedler, Beatrix Scheffer, Carolina Garcia-Rizo, Jan Wolfertz, Volker Hickmann, Friedhelm Pfeiffer and Dieter Oesterhelt

Abteilung für Membranbiochemie, Max-Planck Institut für Biochemie, Martinsried, Germany

The halophilic archaeon *Halobacterium salinarum* (strain R1, DSM 671) contains 2784 protein-coding genes as derived from the genome sequence. The cytosolic proteome containing 2042 proteins was separated by two-dimensional gel electrophoresis (2-DE) and systematically analyzed by a semi-automatic procedure. A reference map was established taking into account the narrow isoelectric point (pI) distribution of halophilic proteins between 3.5 and 5.5. Proteins were separated on overlapping gels covering the essential areas of pI and molecular weight. Every silver-stained spot was analyzed resulting in 661 identified proteins out of about 1800 different protein spots using matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) peptide mass fingerprinting (PMF). There were 94 proteins that were found in multiple spots, indicating post-translational modification. An additional 141 soluble proteins were identified on 2-D gels not corresponding to the reference map. Thus about 40% of the cytosolic proteome was identified. In addition to the 2784 protein-coding genes, the *H. salinarum* genome contains more than 6000 spurious open reading frames longer than 100 codons. Proteomic information permitted an improvement in genome annotation by validating and correcting gene assignments. The correlation between theoretical pI and gel position is exceedingly good and was used as a tool to improve start codon assignments. The fraction of identified chromosomal proteins was much higher than that of those encoded on the plasmids. In combination with analysis of the GC content this observation permitted an unambiguous identification of an episomal insert of 60 kbp ("AT-rich island") in the chromosome, as well as a 70 kbp region from the chromosome that has integrated into one of the megaplasmids and carries a series of essential genes. About 63% of the chromosomally encoded proteins larger than 25 kDa were identified, proving the efficacy of 2-DE MALDI-TOF MS PMF technology. The analysis of the integral membrane proteome by tandem mass spectrometric techniques added another 141 identified proteins not identified by the 2-DE approach (see following paper).

**Keywords:**
*Halobacterium salinarum* / Matrix-assisted laser desorption/ionization mass spectrometry / Peptide mass fingerprinting / Reference map / Two-dimensional gel electrophoresis

1 Introduction

*Halobacterium salinarum* is a member of the halophilic archaea and thrives optimally in 4 M sodium chloride in marine salterns and hypersaline lakes. This extremely halo-
available on *H. salinarum*, especially concerning bioenergetics and signal transduction. The organism can grow under aerobic but also under anaerobic conditions where it can use e.g. DMSO as terminal electron acceptor [1, 2], can ferment arginine to ornithine [3–5], or can grow photo-synthetically due to the retinal protein bacteriorhodopsin [6]. It can respond to various different environmental stimuli showing phototactic as well as chemotactic responses [7]. These processes are an attractive subject for a systems biological approach aimed at modeling parts of a cell. A prerequisite for this, however, is the establishment of reliable genomics, transcriptomics and proteomics.

The genome of *H. salinarum* (strain R1, DSM 671; www.halolex.mpg.de) consists of a 2 Mbp chromosome and 4 megaplasmids ranging from 41 to 284 kbp. This genome is very similar to that of the laboratory strain *Halobacterium* sp. NRC-1 [8], especially in its chromosomal genes. A severe ORF overprediction problem was observed in *H. salinarum*. This species contains 9312 ORFs, of which 8649 are longer than 100 codons. After manual annotation this number could be reduced to 2784 protein-coding genes, which cover 90% of the chromosome (88% of the genome). The longest of the remaining 6528 spurious ORFs has more than 1300 codons. In addition, most reading frames stay open in front of the start codon, the longest extension being 653 codons. As a consequence, the start codon assignment is frequently incorrect. Thus, experimental verification of the predicted proteins is necessary, e.g. by using a proteomic approach.

Here we report a systematic proteomic analysis of the cytoplasmic proteins of *H. salinarum*. Our aim was: (i) the experimental identification of existing proteins under a given condition of cell growth that can support and possibly correct ORF prediction; and (ii) to define the limits by a systematic investigation on one of the most widely used mass spectrometric procedures for proteomic analysis, MALDI-TOF MS. PMF [9–12]. 2-DE was combined with automated spot picking, semi-automated protein digestion and automated MALDI-TOF MS in such a way that high-throughput data acquisition was feasible and the analysis reported here became a time manageable enterprise.

2 Materials and methods

2.1 Bacterial strain and growth conditions

*H. salinarum* (strain R1, DSM 671) was grown in complete medium [13] as described previously [14]. Briefly, for preparation of a starter culture, *H. salinarum* was grown aerobically in the dark at 37°C in 1 L of complete medium to the stationary phase. For protein preparation, *H. salinarum* was grown through three successive transfers. For each transfer, 35 mL of fresh medium were inoculated with 1 mL of the previous culture grown to late log phase (30–40 Klett units).

2.2 Sample preparation

Cells were harvested by centrifugation at 5000 × g for 6 min at 4°C and washed once with basal salt (4.3 M NaCl, 81 mM MgSO₄, 27 mM KCl). Cells were lysed as described earlier [15, 16] with minor modifications. In brief, the cells were resuspended in 50% basal salt containing 0.0075% w/v taurodeoxycholate (Sigma-Aldrich, Taukirken, Germany), protease inhibitor mixture (COMPLETE without EDTA; Roche, Basel, Switzerland) and 100 μg/mL DNase I (Sigma-Aldrich). The cell suspension was incubated for 30 min at 4°C and centrifuged at 213 000 × g for 30 min at 4°C to remove cell debris and cell envelope fragments. Proteins of the supernatant (cytosolic fraction) were precipitated by adding a ten-fold excess of ice-cold acetone. After incubation on ice for 20 min the precipitated proteins were centrifuged at 20 000 × g for 15 min at 2°C and the supernatant was carefully removed. The pellet was washed twice with ice-cold 50% acetone to remove salt, dried and stored at –80°C until use. The amount of protein present in the sample was estimated on the basis of cell mass (determined by volume and turbidity of the suspension) used for the preparation [17]. 100 Klett units (Klett, NY, USA) correspond to an O. D. of 1 at 578 nm (Eppendorff photometer; Netheler and Hinz, Hamburg, Germany) and a total protein content of 0.5 mg/mL. Cytosolic proteins are assumed to amount to 2/3 of total protein.

2.3 2-DE

Protein pellets were resuspended in sample buffer (7 M urea, 2 M thiourea, 4% w/v CHAPS, 0.6% v/v Pharmalytes 3–10 (Amersham Biosciences, Uppsala, Sweden), 65 mM DTT), and the suspension was shaken for 30 min. Prior to use the solution was centrifuged at 213 000 × g for 30 min at 18°C. For IEF, 18 cm IPG strips (pH 3.5–4.5, pH 4–5, pH 4.5–5.5; Amersham Biosciences), or 17 cm strips (pH 5–8; Bio-Rad, Hercules, CA, USA) were used. IEF strips were rehydrated in 350 or 300 μL reswelling solution (7 M urea, 2 M thiourea, 4% w/v CHAPS, 0.6% v/v Pharmalytes 3-10, 20 mM DTT). Protein amounts of about 300 μg were applied via application pieces (Amersham Biosciences) at the cathodic end of the IPG strips 3.5–4.5 and 4–5, or at the anodic end of the IPG strips for pH 4.5–5.5 and pH 5–8. Larger amounts of protein (about 800 μg) were applied by in-gel rehydration. IEF was performed with a Multiphor II unit (Amersham Biosciences) employing the following voltage profile: linear increase from 0 to 150 V in 1 min, 150 V for 1h, linear increase from 150 to 300 V in 1 h, 300 V for 1 h, linear increase from 300 to 500 V in 1 h, 500 V for 4 h, linear increase from 500 to 3500 V in 5 h, 3500 V for 11 h. Total focusing time was about 51 500 Vh. The IPG strips were then equilibrated as described earlier [18]. Electrophoresis in the second dimension was carried out either on 9% or 14% T polyacrylamide gels in a Protean II apparatus (Bio-Rad).
Protein spots were visualized using a mass spectrometric compatible silver stain according to Mortz et al. [19] with minor modifications. In brief, gels were washed after fixation with 50% ethanol, half concentrations of formaldehyde were used and the staining was terminated with 5% v/v acetic acid.

2.4 Protein digestion and sample preparation

Protein spots were automatically excised using Bruker’s Proteineer SP spot picking robot (Bruker Daltonics, Bremen, Germany), which transfers the 1.5 mm gel pieces automatically into filter microtiter plates (Multiscreen Durapore MAHV N45; Millipore, Bedford, MA, USA). By using filter microtiter plates all incubation media could be separated from the gel pieces by centrifugation, and the eluate was collected in a microtiter plate placed underneath.

After destaining [20], the gel pieces were washed with water and equilibrated prior to proteolytic digestion. Gel pieces were shrunk with 50% v/v ACN and rehydrated in 50 mM NH₄HCO₃. This treatment was repeated, followed by the addition of 0.1 μg/spot modified trypsin (Promega, Madison, WI, USA). Digestion was carried out overnight at 37°C. The supernatant was collected and combined with the eluates of the following three elution steps: (i) water; (ii) 50% v/v ACN; and (iii) 50% v/v ACN, 0.1% v/v TFA. The combined eluates were frozen in liquid nitrogen and dried in a SpeedVac centrifuge. To remove remaining ammonium bicarbonate the pellet was resuspended in water, frozen and lyophilized. This procedure was repeated once. Each dried sample was finally dissolved in 10 μL 33% v/v ACN, 0.1% v/v TFA. Using Bruker’s MAP pipetting robot, 0.5 μL of each sample was mixed automatically with the same amount of a saturated CHCA in 40% v/v ACN, 0.1% v/v TFA on a ground steel MALDI target.

2.5 Protein identification

MALDI-TOF MS spectra were recorded automatically using a Reflex III spectrometer (Bruker Daltonics) with an acceleration voltage of 20 kV. The spectra were externally calibrated with an in-house optimized mixture of eight peptides ranging from 1046.54 Da to 3494.65 Da. Spectra were automatically annotated using the vendors Xmass program package, which returns monoisotopic masses. After a second recalibration step using an algorithm based on the concept of Zubarev et al. [21] that was implemented and developed further by Bruker Daltonics, an average mass accuracy of 25 ppm was obtained. The search against the database comprising all 9312 ORF’s from H. salinarum was performed with the Biotools software (Bruker Daltonics) integrating the locally installed Mascot search engine (Matrix Science, London, UK) [22]. Search parameters were: 0 missed cleavage sites, +/-100 ppm tolerance (to account for reduced mass accuracy of larger peptides), carbamidomethylation of cysteines as fixed modification and no variable modifications. Results are based on a ‘probability-based MOWSE score algorithm’. For a database of 9312 ORFs, a score of 52 or higher is considered significant at the 5% level according to Mascot. This would mean that only 5% of the ORFs with a score ≥52 are false positives, which turned out to be a gross underestimation under our experimental conditions. This can partially be attributed to casual matches of peptides with one of the trypsin-derived mass peaks. We use this value of 52 as a “significance cut off” in our evaluation, as detailed below.

To minimize false positive identifications, a MOWSE score higher than 72 (20 above significance cut-off) was considered as identification and a MOWSE score higher than 92 (40 above significance cut-off) was considered as a reliable identification. We identified six false positive identifications with scores of 72, 74, 75, 84, 84, and 96. At least nine out of the 66 ORFs (13%) with a score of 62–71 and at least 83 out of the 168 ORFs (49%) with a score of 52–61 are spurious ORFs and thus are considered to be false positives. (These ORFs are spurious because another frame in the same region codes for a real protein. In no case more than one frame was found to be coding when a minimal score of 72 was requested.) It is evident that this rate of false positives is much higher than predicted by Mascot (only 5% false positives expected at a score of 52, 0.5% at 62, 0.05% at 72, and 0.0005% at 92). The Mascot formula is based on ideal mass spectra not considering trypsin-derived or other frequently occurring mass peaks. Due to the high rate of false positives, proteins with a MOWSE score below 72 were not considered as identified in our automatic procedure. However, such proteins may be promoted to the level of “identified” in a manual procedure (status “manual” in Table 1, supplementary material). Criteria considered in this procedure are (a) the score, (b) the number of matching and missing peptides, (c) eventual matches to trypsin derived or other frequently occurring peaks, (d) the occurrence of missed cleavage sites, which are not taken into account during the automatic annotation procedure, (e) the peak intensity, and (f) the correlation between theoretical pI/Mr and gel position.

2.6 The H. salinarum protein sequence database

The H. salinarum protein sequence database was exported from the HaloLex project (www.halolex.mpig.de), and is based on the genome from H. salinarum (strain R1, DSM 671). It contains 9312 ORFs which were obtained by using two gene finders (ORPHEUS [23], Glimmer [24, 25]) and sixframe translation with a minimum length of 100 codons. In total, 8649 ORFs are longer than 100 codons and 663 ORFs are shorter. Currently 2784 ORFs are considered to code for proteins and 6528 ORFs are spurious. This set is continuously updated as required due to new experimental or homology data.

© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim
www.proteomics-journal.de
2.7 Correlation analysis

The correlation between theoretical value and gel position was analyzed for pI and M, values. The theoretical pI was computed using a tool provided by the Expasy server [26]. The molecular weight was computed from the amino acid sequence. Spot positions were recorded using the Bruker SP spot picker device (Bruker Daltonics). Only gels with at least 40 reliable identifications were selected. Identifications with MOWSE scores below the reliability cut-off were not considered in order to exclude adverse effects due to false positive identifications. For pI correlation, a linear line of best-fit was computed from the data using the R statistical program package [27] and used as a calibration curve to read the experimental pH at a given position of the gel strip. The residuals of the data points against the line of best-fit showed a Gaussian distribution. For each gel the SD of the residuals was computed. In an iterative approach, proteins deviating more than 3 × SD were excluded and the analysis was repeated until no strongly deviating proteins were found. Usually, only 1–3 data points had to be excluded. The units of SD to the final line of best-fit were computed for every data point. For molecular weight correlation, a line of best-fit between logM and gel position was computed in the same way.

3 Results

3.1 Analysis of the cytosolic proteome using overlapping gels

The calculation of the theoretical 2-DE map for all 2784 ORFs from *H. salinarum* made obvious that the large majority (82%) of all proteins have pI values between 3.5 and 5.5 (Fig.1). This is especially true for the soluble proteins, of which 90% have pI values between 3.5 and 5.5 and 79% between 4.0 and 5.0. In contrast, other microorganisms show a broader pI distribution on the theoretical 2-DE map, with two pI regions, one on the acidic and one on the alkaline side, where proteins cluster [28–32]. The narrow pI range of soluble halophilic proteins is due to the very high content of acidic residues, which concentrate on the surface and mediate protein solubility in 4 M salt [33–35].

In order to obtain maximal resolution for protein separation in the first dimension, three overlapping narrow range IPG strips (namely 3.5–4.5, 4–5, 4.5–5.5) covering the most highly crowded region, were used (Fig. 2). Proteins with higher pI were analyzed on IPG strips 5–8 (Fig. 3). Proteins with a pI above 8, constituting only 9% of the total proteins and 4% of the cytosolic proteome, and below 3.5 (only nine proteins) were not analyzed in this study. For each pI range gels with two different acrylamide concentrations were used for the second dimension. On the 9% T gels, proteins with a molecular weight down to 19 × 10^3 could be separated, whereas on the 14% T gels the low molecular weight range down to about 6 × 10^3 was resolved. These eight 2-D gels form the reference map series of *H. salinarum* (Fig. 2 and 3).

Exhaustive analysis of the cytosolic proteome of *H. salinarum* covering the overlapping eight gels was done in the following way: every protein spot that could be detected by a mass spectrometric compatible silver stain was excised from at least three different but corresponding gels, with two different protein loads of about 300 and 800 µg. Each protein spot was digested, prepared twice on a MALDI target and analyzed by MALDI-TOF MS PMF. For the about 7500 excised protein spots, a semi-automated protocol was established comprising (a) automated spot excision, (b) in-gel digestion in 96 well filter plates, (c) automated target preparation in duplicate as well as (d) automated acquisition of the about 15000 mass spectra followed by (e) automatic peak annotation, (f) protein identification using the Mascot program and (g) integration of genomic and proteomic data into the HaloLex annotation system. The semi-automatic procedure enabled the processing of this experimental data in 10 weeks to generate a first 2-DE reference map of the cytosolic proteins of *H. salinarum*.

We anticipate 1959 proteins to be located within the cytosolic reference map series. Excluded are (a) ORFs interrupted by ISH elements, (b) duplicated genes coding for identical proteins due to inser-
tion elements and to large-scale duplications in the plasmids, (c) proteins with at least 1 predicted transmembrane domain using the TMHMM program [36] and (d) proteins outside the pI range from 3.5–8.0. The number of 1959 proteins, which we use in the computations below, seems to be a reasonable estimate even if it may be a slight overestimate as e.g. secreted or peripheral membrane proteins have not been excluded.

A total of 661 distinct proteins could be identified (Table 1, supplementary material), representing 33% of the set. Of these only 227 were identified on gels with the higher protein load (800 µg). Most of the protein spots could be matched between corresponding gels. 141 additional proteins were identified on 2-DE gels not fulfilling the specifications of the reference map (gel format, bacterial growth and sample preparation conditions). Only one identified protein on the reference map (OE2015R) contains a predicted transmembrane domain. Six of the identified proteins are encoded by more than one gene, due to the extensive gene duplications observed in H. salinarum. One of the genes (OE6230R/OE7110R) is interrupted by an ISH2 element in both copies, but interestingly the resulting protein of 213 amino acids could be identified. An in-frame stop codon occurs within the ISH2 element 5 codons after the insertion point. No peptide corresponding to the C-terminal 464 amino acids could be identified.

In total about 40% of the soluble proteins of H. salinarum, which can be found theoretically, could be identified using the combination of 2-DE and MALDI-TOF MS PMF (supplementary, Table 1). Two factors turned out to strongly influence the level of protein identification, one being experimental and one of more general importance. The first observation concerns the protein’s size. About 50% of the proteins with a molecular mass >25 kDa but less than 20% of those <20 kDa could be identified (Fig. 4). Theoretical analysis showed that the 14% Tgels in Fig. 2, containing the proteins <20 kDa, should contain one third of all spots. Visual inspection of these gels shows that this “small proteome” is severely underrepresented (Fig. 2, b, d, and f). In addition the small proteins <20 kDa are difficult to identify by PMF due to the reduced number of tryptic peptides. The second factor is the gene localization within the genome. Proteomic analysis showed that mainly chromosomally encoded proteins are expressed at high level. We identified

Figure 2. 2-D reference map of H. salinarum. The cytosolic proteins were separated on overlapping zoom gels, covering the pI range from 3.5–5.5 and stained with silver nitrate. Every pI range was covered with two different acrylamide concentrations (9% and 14% T). The different reference map sectors are marked (a–f). Identified proteins are labelled with yellow triangles. Green triangles indicate bovine DNase I. Corresponding spot patterns in the overlapping regions are highlighted with red circles. Overlapping gel regions are not fully displayed. All gels will be accessible in the HaloLex database (www.halolex.mpg.de).
Figure 3. Reference map of cytosolic proteins from *H. salinarum*. Proteins were separated in a pH gradient 5.0–8.0 and stained with silver nitrate. Additionally gels were run with two different polyacrylamide concentrations (9% and 14% T). The different reference map sectors are marked (g–h). Identified proteins are labelled with yellow triangles. Green triangles indicate bovine DNase I. Corresponding spot patterns in the overlapping region are highlighted with red circles. The overlapping gel region is not fully displayed. All gels will be accessible in the HaloLex database.

Figure 4. Percentage of identified proteins that should be located within the reference map series as a function of size range.

48% of the chromosomally encoded proteins but only 14% of the plasmid encoded proteins. When restricting the analysis to soluble, chromosomally encoded proteins larger than 25 kDa that should be represented on the reference map series, 63% of these were identified. According to our results, only one frame is coding at a given genome position despite the high number of spurious ORFs.

Among the identified proteins 94 were found to have more than one spot. The occurrence of proteins in multiple spots may be due to non-native chemical modifications like deamidation or carbamylation [37] that may occur during sample processing, to partial proteolytic cleavage, or to biochemically relevant post-translational modifications such as phosphorylation. 45 of these proteins were found in spots with different pI-values, 17 in spots with a different apparent molecular mass and 32 proteins shifted in both directions. Supplementary Fig. 9 illustrates a representative overview of proteins found in multiple spots.

On the reference map series about 1800 distinct spots were found. Subtracting the multiple spots a conservative lower estimate is about 1600 unique proteins present on these gels. Theoretical analysis predicts 1959 proteins. This would mean that about 80% of all genes are expressed at a level to reach visibility on a 2-D gel for this organism under the selected growth conditions. This unexpectedly high proportion of expressed genes is also observed in transcriptomic analysis (J. Twellmeyer, unpublished data).

3.2 Correlation analysis

When the theoretical pI of the identified proteins was plotted against their position on the gel an extremely good linear correlation was observed (Fig. 5A). Using the line of best-fit as a linear calibration curve the experimental pH at a given position on the IPG strip can be defined. We thus plotted the theoretical versus the experimental pI value. The large number of proteins with pI values in a narrow pH range permits a detailed statistical analysis of this calibration procedure. We restricted our analysis to 2-D gels with at least 40 reliable identifications (25 gels in total) and computed the line of best-fit in an iterative procedure (see section 2.7). The scattering of the individual data points followed a Gaussian distribution and thus allowed the SD to be computed. The data in Fig. 5 originate from a single 2-D gel, and represent 242 reliable identifications (MOWSE score of at least 92). The theoretical pI values match very closely to the line of best-fit with one SD being 0.034 pH units. Only very few proteins showed deviations above 3 $\times$ SD, i.e. above 0.1 pH units. More than 50% of the proteins deviate less than 0.025 pH units. The data from Fig. 5A are typical for the 25 gels, which have been analyzed accordingly, with standard deviations always being in the range of 0.03–0.06 pH.

The molecular weight data were analyzed in the same way. The theoretical $M_r$ values match to the line of best-fit with one SD being 0.034 pH units. Only very few proteins showed deviations above 3 $\times$ SD, i.e. above 0.1 pH units. More than 50% of the proteins deviate less than 0.025 pH units. The data from Fig. 5A are typical for the 25 gels, which have been analyzed accordingly, with standard deviations always being in the range of 0.03–0.06 pH.

The molecular weight data were analyzed in the same way. The theoretical $M_r$ values match to the line of best-fit with one SD being 0.034 pH units. Only very few proteins showed deviations above 3 $\times$ SD, i.e. above 0.1 pH units. More than 50% of the proteins deviate less than 0.025 pH units. The data from Fig. 5A are typical for the 25 gels, which have been analyzed accordingly, with standard deviations always being in the range of 0.03–0.06 pH.
deviations thereof may indicate an incorrect assignment of the protein. Alternatively, the genome derived protein sequence used to compute the \( p_I \) value may not represent the sequence of an existing protein species. This can be due to post-translational modifications or to an incorrectly annotated start codon. Thus, correlation analysis is turned into a tool to analyze the correctness of annotation and to search for interesting biological phenomena.

As one application of this tool, we checked the \( p_I \) correlation of proteins identified with MOWSE scores above 72 for any indication of false positive identifications by MALDI-TOF MS PMF. We identified 5 false positives with large deviations. For two of them, SD was computed (183 × SD, protein \( p_I \) 11.2, gels pH 3–6, and 26 × SD, protein \( p_I \) 6.0, positioned at \( p_I \) 4.8 on a gel pH 3.5–5.5). For the other three false positives, the SD was not computed but all had a \( p_I \) above 9 and were identified on acidic gels. Only one additional false positive showed a good correlation. For proteins with scores between 52 and 71, rating them as not being identified (see section 2.5), many poor correlations were evident, consistent with the high number of false positives in this score range.

### 3.3 Evaluation and correction of gene predictions by proteome identifications

Gene prediction may result in several types of errors, including overprediction, underprediction, selection of the wrong reading frame, as well as incorrect start codon assignment. For the annotation of the genome of *H. salinarum* all these problems were found in abundance. A solution to these problems by integration of bioinformatic and proteomic analysis is illustrated in the following examples.

On the basis of \( p_I \) correlation analysis, as described above, we checked several proteins, which showed deviations of more than 3 × SD on several gels. This permitted the identification of several incorrect start codon assignments. Criteria for selection of the start codon were (a) the \( p_I \) value of the region between the alternative start codons, (b) consistency to homologous proteins, and (c) the length of the intergenic region to the preceding ORF. The problem in assigning the correct start codon is illustrated by protein OE3065R. The ORPHEUS program predicted a protein 108 residues longer than that computed by the Glimmer program. The identification of 3 peptides in the N-terminal extension clearly showed that the start codon predicted by Glimmer is incorrect. However, protein OE3065R still showed a large deviation from the line of best-fit in the \( p_I \) correlation analysis on several gels (deviations ranging from 2.4 to 5.1 × SD). Therefore, we checked if ORPHEUS also had selected an incorrect start codon. After assigning a start codon that shorted the sequence by 26 amino acids at the N-terminal end, a very good correlation between theoretical \( p_I \) value and gel position was observed. PMF analysis in combination with \( p_I \) correlation analysis thus allowed the correct start codon assignment to be made, disproving start codon selection by two gene finders. A similar case is protein OE1793F, where also ORPHEUS and Glimmer had predicted different start codons, both being incorrect.

The region around genome position 1 490 000 contains the *fad1* gene, coding for a probable enoyl-3-hydroxybutyryl-CoA dehydratase (Fig. 6A). The start codon was predicted to be Met-1 by ORPHEUS and Met-81 by Glimmer. The N-terminal 80 residues are clearly homologous to sequences from...
Figure 6. (A) The fad1 and its preceding gene. The fad1 gene is shown in blue with the incorrectly assigned start codon by Glimmer shown as white triangle. The gene for protein OE3849F, which has been reliably identified from a very weak silver stained spot (B), is shown in green. The spurious ORF OE3848R from the ORF list from NRC-1 is shown in white with a red boundary. An additional seven spurious ORFs, all longer than 100 codons are shown in white with black boundaries. (B) zoom view from a 2-D gel sector d (see Fig. 2). The identified protein spot OE3849F is highlighted (red circle).

other species and thus the Glimmer prediction had to be corrected. A divergently transcribed gene, coding for OE3849F, is located 5’ of fad1. The corresponding protein has been reliably identified from a very weak silver stained spot on the reference map series (Fig. 6B). This protein is not in the protein list for strain NRC-1. Instead, one of the spurious ORFs had been assigned to this gene region (Fig. 6A). It overlaps with the longer and correct version of the fad1 gene but not with the shorter and incorrect one. A total of 7 additional spurious ORFs are encoded in the region covered by these two genes.

Protein identification from a very weak silver stained spot (Fig. 7C) allowed a whole series of gene prediction errors in three distinct regions on two of the megaplasmids to be solved. Near position 272 kbp on pH3, ORPHEUS and Glimmer both predicted the same protein (H. salinarum, code OE5432F, NRC-1 code H1502) (Fig. 7A). No other ORFs had been predicted in that region. In the region around position 53 kbp on pH2, ORPHEUS and Glimmer predicted three proteins encoded on the reverse strand, while ORPHEUS had predicted two additional ORFs on the forward strand (Fig. 7A). One of them (OE6106F) showed 77% protein sequence identity to the C-terminal region of OE5432F, the other (OE6101F) showed 63% protein sequence identity to the N-terminal region of OE5432F.

The high level of sequence identity between OE6101F and OE5432F (77%) and the lack of a homolog for OE6105R favored the forward strand to be coding on pH2 near position 53 kbp. This was disproven by the reliable identification of protein OE6105R, encoded on the reverse strand, making the simultaneous expression of OE6106F highly unlikely. As a consequence, it became unlikely that ORF OE5432F codes for a real protein, which in turn is also true for the two ORFs with sequence homology to its N-terminus: OE6101F and OE6058F. As a result, the region around position 272 kbp on pH3 and the region around position 33 kbp on pH2 no longer contained any candidate genes. However, six-frame translation with a minimal length of 100 codons predicted ORF OE5432B1R with 92% sequence identity to the identified protein OE6105R (Fig. 7B) and OE431A1R with 75% sequence identity to OE6102R. Due to the length cut-off, prediction of the 96 amino acid protein OE5431B1R, a homolog to OE6104R with 82% sequence identity, failed. It was created manually upon visual inspection of the genome region.

As a further complication the newly predicted protein OE5432B1R is interrupted by insertion element ISH3, which has inserted at the codon for Lys-9. The protein created by six-frame translation started with a GTG codon coding for Val-21. The N-terminal region of OE5432F codes for a real protein, which in turn is also true for the two ORFs with sequence homology to its N-terminus: OE6101F and OE6058F. As a result, the region around position 272 kbp on pH3 and the region around position 33 kbp on pH2 no longer contained any candidate genes. However, six-frame translation with a minimal length of 100 codons predicted ORF OE5432B1R with 92% sequence identity to the identified protein OE6105R (Fig. 7B) and OE431A1R with 75% sequence identity to OE6102R. Due to the length cut-off, prediction of the 96 amino acid protein OE5431B1R, a homolog to OE6104R with 82% sequence identity, failed. It was created manually upon visual inspection of the genome region.

As a further complication the newly predicted protein OE5432B1R is interrupted by insertion element ISH3, which has inserted at the codon for Lys-9. The protein created by six-frame translation started with a GTG codon coding for Val-21. The N-terminal residues are encoded on the other side of the ISH3 element (not shown in Fig. 7B). Lys-9 is encoded on both sides of the ISH3 element due to a 5 bp target duplication. When analyzing the region around position 33 kbp on pH2, a pseudogene with an internal stop codon was detected (OE6057C1R). The C-terminal region following the stop codon has 85% protein sequence identity to OE5431A1R. The region preceding the stop codon indicates a breakpoint in the plasmid sequence as the C-terminal part shows 77% protein sequence identity to OE5431A1R, while the N-terminal part shows sequence homology to an unrelated protein. An additional 12 spurious ORFs longer than 100 codons exist in this region (not shown). Thus, on the basis of these results taken together we suggest that the filled colored arrows in Fig. 7B represent real proteins.

3.4 Chromosomes and plasmids in H. salinarum

The genome of H. salinarum, strain R1 (DSM 671, www.hallex.mp.g.de) consists of a 2 Mbp chromosome and 4 megaplasmids ranging from 41 to 284 kbp. The chromosome has a very high GC content of 68%, while the GC-content of the plasmids is considerably lower with 59%. A major difference in the level of protein identification achieved between chro-
Here we document that (a) an episomal, plasmid-derived region of 60 kbp exists in the chromosome and (b) a 70 kbp region of chromosomal DNA, encoding several important and essential genes, has been transferred to pHS3, now defining pHS3 as a second chromosome.

Figure 8 displays the GC content and the identification level for the chromosome and the four plasmids. Data were computed with a window of 45 ORFs using soluble proteins neither encoded on nor interrupted by insertion elements. Both, the average GC content and the identification level differ between chromosome and plasmids. The region between position 11 500 and 71 500 on the chromosome shows a significant drop in both, GC content and identification level, down to values typical for the plasmids (boxed in Fig. 8B). The GC content of each individual ORF is indicated, resulting in a broad band fluctuating around the mean. At position 11 500, the “chromosomal” band suddenly stops and a “plasmid” band starts, which flips back to the chromosomal band at position 71 500. This is evidence for an episomal region in the chromosome, previously described as AT-rich island [38, 39]. Only a few of the proteins in this episomal island have been identified.

The opposite situation is found on plasmid pHS3, where the GC content rises from plasmid level to chromosomal level at position 54 000 and falls back to plasmid level at position 126 000 (Fig. 8E). The identification level also increases up to the chromosomal level, indicating that a part of the chromosome has been transferred to pHS3. On this part of pHS3, several important and essential genes are encoded: the only arginine-tRNA ligase, the only catalase, the two chains of aspartate carbamoyltransferase (pyrB and pyrI), catalyzing the first step in pyrimidine biosynthesis. We conclude that pHS3 should be considered a second chromosome rather than a megaplasmid.

4 Discussion

For a systematic proteomic analysis by 2-DE combined with MALDI-TOF MS PMF, we selected the most general cell state and established a master gel series as a reference for further proteomic experiments. All catalysts and other proteins required for standard metabolism and cell growth should be represented in the proteome under these condi-
Figure 8. Identification level and GC content for the whole chromosome (A) compressed scale, the episomal region in the chromosome (B), and the four megaplasmids, pH51–4; (C–F). The x-axis represents the serial number of the protein on the corresponding replicon. Proteins encoded on or interrupted by insertion elements are excluded as well as proteins with predicted transmembrane domains. The following data are shown. Upper panel (I): (a) the average identification ratio for all proteins encoded on the chromosome (dark blue line, 48.2%), (b) the average identification ratio for all proteins encoded on the plasmids (light blue line, 13.6%) and (c) the average identification ratio computed with a window of 45 ORFs (the ORF itself and 22 ORFs on each side, black curve). Identified proteins are indicated by a green dot on top of the panel (d) and non-identified proteins by a green dot at the bottom (e). Lower panel (II): (f) the average GC content of the chromosome (dark blue line, 68.2%), (g) the average GC content of the plasmids (light blue line, 58.9%); the GC content of each individual gene (red dots), resulting in a band around the average GC content (h) computed with a window of 45 ORFs (the ORF itself and 22 ORFs on each side, black curve). It should be noted that chromosomal genes rarely have a GC content below 65% and that plasmid genes rarely have a GC content above 61%, except for the episomal integration into the chromosome (boxed in B) and for the chromosomal region that has integrated into pH53 (boxed in E).

We set out with our experiments with two aims: (i) to establish the protein inventory of the H. salinarum cytosol, which in addition to the generation of a 2-D reference map would allow the validation and improvement of genome annotation; and (ii) to systematically investigate the efficacy of the 2-DE MALDI-TOF MS PMF standard procedure. The advantage of this method is the possibility of automation and therefore high-throughput analysis.

We were successful in establishing a semi-automatic procedure that could cope with about 7500 gel spots (15 000 MS spectra) in an acceptable time, resulting in the identification of 661 proteins on the reference map and 141 additional proteins identified on other 2-D gels. Altogether the 802 identified proteins listed in Table 1 (supplementary material), represent about 40% of the cytosolic H. salinarum proteome. This is so far the highest number reported for procaryotes using only the combination of 2-DE and MALDI-TOF MS PMF. The identified proteins include many (up to 90%) of the enzymes that can be attributed to the central metabolic path-
ways (gluconeogenesis, citric acid cycle, purine and pyrimidine metabolism, several amino acid and coenzyme biosynthetic pathways). An exception is the respiratory chain, which due to its membrane-bound nature is under-represented in the soluble proteome but is represented in the membrane proteome [41]. Also, ribosomal proteins are not well represented, which can be attributed to their small size (see below for a discussion of the small proteome).

Upon classification into the usual function classes, reliable assignments are not possible for about 50% of the proteins. We introduced the acronym NOF (“no function assigned”) for those hypothetical (HY) or conserved hypothetical (CHY) proteins that have been identified in this study.

When looking at the genomic localization of the identified proteins the identification level differs strikingly for the chromosomal and plasmid encoded proteins. We identified 48% of the chromosomally encoded but only 14% of the plasmid encoded proteins (overall, 40%). We therefore propose that a large fraction of the genes encoded on plasmids, whose function in most cases remains unknown, is not expressed and may only be necessary under special conditions for survival of the cells. While the chromosome of \textit{H. salinarum} is very GC-rich (68% GC), the four megaplasmids of strain R1 with 41–284 kbp have a moderate GC content (59% GC). The chromosome contains a 60 kbp region, which has only a moderate GC content, corresponding to the previously described AT-rich island [38]. Surprisingly, the identification level is greatly reduced in this 60 kbp region, which is interpreted to be a sequence originating from a plasmid. More surprisingly, we identified a region with high GC content and also a high identification level in the largest megaplasmid pH3. We suggest that this is a 70 kbp piece of chromosomal DNA, which has been transferred to the plasmid with preservation of its GC content and protein identification level. The megaplasmid pH3 contains a number of essential and highly important genes (arginine-tRNA ligase, catalase and the two chains of the aspartate carbamoyltransferase) and thus should be regarded as a second chromosome.

Confirmation of chromosomal plasmid DNA exchange and the link between high GC content and high expression rate of proteins is not the only case where proteomics substantially contributed to the genomics of \textit{H. salinarum}. The very high GC content leads to a 3.5-fold ORF overprediction (9,312/2,784) and also results in a notorious misprediction of start codons. In many cases, some of which are documented in the section 3, ORF selection ambiguities could be resolved by proteomic analysis. In addition, proteomics verified the existence of 390 hypothetical proteins. Several proteins with an incorrectly assigned start codon could be recognized by analysis of the correlation between theoretical pI value and gel position of protein spots. This correlation is impaired for proteins with incorrectly assigned start codons as the pI value changes abruptly around the start codon. In general, the coding region is acidic, while theoretical translation of the preceding region results in protein sequences, which are highly alkaline due to the statistical characteristics of halobacterial DNA.

So far, in none of the genomically sequenced organisms has it been possible to identify 100% of all proteins by proteomics. Our inability to identify more than 60% of the cytosolic proteome may be due to: (a) the proteins are not expressed under growth conditions or the expression level is below the level of detection; (b) not all proteins are included in the set, which is subjected to analysis because they are lost during sample preparation; and (c) proteins cannot be identified by 2-DE MALDI-TOF MS PMF. Small proteins turned out to be drastically under-represented in the list of identified proteins. In addition to the potential shortcomings of 2-DE (loss of protein, failure of staining), a major problem is the low number of tryptic peptides in the scanned mass range, which hampers identification by PMF. This problem may be overcome by application of MS/MS technologies like MALDI-TOF/TOF and ESI-MS/MS, which allow protein identification based on a single proteolytic peptide. By using these techniques we have recently identified several small proteins, which escaped identification by PMF. The rapid technical development in proteomic research will further improve sensitivity and throughput. An integrated environment is now being introduced and tested (Proteineer SP, Bruker Daltonics, in combination with ProteinWeaver, Definiens, Munich, Germany). Gels are analyzed automatically by a 2-D gel image analysis software and spots are instantly and automatically detected. After guided selection, picked spots are automatically digested and prepared for MALDI-TOF MS (Bruker ProteinDeeP, Bruker Daltonics) on anchor targets [42] to further improve reproducibility and sensitivity. Alternatively, new sample preparation technologies [43] (GyroLab, Gyros, Sweden) are being tested.

Given the limitations by theory and experimental experience we conclude that beside the 2-DE MALDI-TOF MS approach, nevertheless alternative and additional methods are clearly required. LC-MS/MS methods avoiding gel preparations also added to the identified proteins in the inventory. By using a LC-MS/MS based shotgun technology, Goo et al. [44] identified from the very closely related strain \textit{Halo bacterium sp.} NRC-1 330 proteins without predicted transmembrane domains. Of these 64 proteins are not present in the set of the 802 identified proteins of our study.

We are grateful to Prof. Dr. Michael Hecker for intensive discussion and helpful advice on the 2-DE methodology. Further the authors would like to thank Bruker Daltonics for fruitful cooperation. The technical assistance of Sigrid Bauer was highly effective and is much appreciated.

5 References


© 2005 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

www.proteomics-journal.de