DOI 10.1002/pmic.200401201

Regular Article

Towards the proteome of the marine bacterium *Rhodopirellula baltica*: Mapping the soluble proteins

Dörte Gade[1], Dorothea Theiss[2], Daniela Lange[1], Ekaterina Mirgorodskaya[2], Thierry Lombardot[1], Frank Oliver Glöckner[1], Michael Kube[2], Richard Reinhardt[2], Rudolf Amann[1], Hans Lehrach[2], Ralf Rabus[1] and Johan Gobom[2]

[1] Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Max Planck Institute for Molecular Genetics, Berlin, Germany

The marine bacterium *Rhodopirellula baltica*, a member of the phylum *Planctomycetes*, has distinct morphological properties and contributes to remineralization of biomass in the natural environment. On the basis of its recently determined complete genome we investigated its proteome by 2-DE and established a reference 2-DE gel for the soluble protein fraction. Approximately 1000 protein spots were excised from a colloidal Coomassie-stained gel (pH 4–7), analyzed by MALDI-MS and identified by PMF. The non-redundant data set contained 626 distinct protein spots, corresponding to 558 different genes. The identified proteins were classified into role categories according to their predicted functions. The experimentally determined and the theoretically predicted proteomes were compared. Proteins, which were most abundant in 2-DE gels and the coding genes of which were also predicted to be highly expressed, could be linked mainly to housekeeping functions in glycolysis, tricarboxic acid cycle, amino acid biosynthesis, protein quality control and translation. Absence of predictable signal peptides indicated a localization of these proteins in the intracellular compartment, the pirellulosome. Among the identified proteins, 146 contained a predicted signal peptide suggesting their translocation. Some proteins were detected in more than one spot on the gel, indicating post-translational modification. In addition to identifying proteins present in the published sequence database for *R. baltica*, an alternative approach was used, in which the mass spectrometric data was searched against a maximal ORF set, allowing the identification of four previously unpredicted ORFs. The 2-DE reference map presented here will serve as framework for further experiments to study differential gene expression of *R. baltica* in response to external stimuli or cellular development and compartmentalization.

Received: September 1, 2004
Revised: November 16, 2004
Accepted: December 14, 2004

Keywords:
MALDI-MS / PMF / *Rhodopirellula baltica* / 2-DE

1 Introduction

Since the pioneering determination of the *Haemophilus influenzae* [1] and *Mycoplasma pneumoniae* [2] genomes, more than 250 complete genomes from bacteria have been reported (for detailed information see *e.g.*, www.genomesonline.org). Even though a given genome represents the blueprint of life, there is a need for functional analysis on the transcriptional and proteomic level in order to define (i) which of the predicted genes can be expressed in principle, and (ii) the physiological conditions inducing their expression. In contrast to the numerous publicly available genome sequences, only few proteomes (protein maps) have been reported to date. Moreover, only a limited number of proteins is usually identified and annotated. Among the comprehensive protein maps are

Correspondence: Dr. Ralf Rabus, Max Planck Institute for Marine Microbiology, Celsiusstr. 1, D-28359 Bremen, Germany
E-mail: rrabus@mpi-bremen.de
Fax: +49-421-2028-580

those for the very well-studied standard bacteria *Escherichia coli* [3–5] and *Bacillus subtilis* [6, 7], some pathogens like, *e.g.*, *Mycoplasma pneumoniae* [8–10], *Staphylococcus aureus* [11, 12], *Haemophilus influenzae* [13–15] and *Pseudomonas aeruginosa* [16], and some biotechnologically relevant bacteria such as *Corynebacterium glutamicum* [17] and *Streptomyces coelicolor* [18].

Protein maps are most often constructed by applying 2-DE in combination with MS analysis. 2-DE is a well-established technique for high-resolution separation of proteins from complex mixtures [19]. Electrophoretically separated proteins are excised from stained 2-DE gels and cleaved enzymatically (*e.g.*, by trypsin) to defined fragments. The masses of the generated peptides, determined by MS, constitute a PMF of the protein. For protein identification, the PMF is compared to sets of masses calculated for each protein sequence in a database, based on the known cleavage specificity of the protease used [20–22]. MALDI-MS [23] has become the most widely used technique for protein identification. Automation allows a high throughput at the level of spot excision, sample processing and MS analysis.

Over the last one to two decades, the impact of microbial activity on environmental processes has been increasingly recognized. This led recently to the initiation of genome projects on environmentally relevant bacteria. Genomes of such bacteria in conjunction with functional analysis will provide new insights into the molecular basis of microbial activity (and its control) in the natural environment. The first examples are the complete genome sequences of *Synechocystis* sp. [24] (www.kazusa.or.jp), *Caulobacter crescentus* [25] (www.tigr.org) and *Rhodopirellula baltica* [26] (www.regx.de). With 7.145 Mb and 7325 ORF, the genome of *R. baltica* represents one of the largest bacterial genomes sequenced so far. In the case of *C. crescentus*, a protein map with 295 identified proteins has only very recently been reported [27], whereas 57 membrane proteins were identified from *Synechocystis* sp. strain PCC6803 [28]. *R. baltica* is a marine, aerobic bacterium that has been isolated from the Baltic Sea. It belongs to the phylogenetic distinct group of *Planctomycetes* [29], members of which are known to be globally distributed and suggested to be involved in carbon remineralization. Interest in this group of bacteria also comes from their unusual morphological properties. The cells reproduce *via* budding and display a complex life cycle. *R. baltica* cells can occur in two morphotypes, *i.e.*, as single motile cells or attached to each other in aggregates. Peptidoglycans appear to be absent from the proteinaceous cell wall. Individual cells are organized in membrane-defined compartments including a membrane-engulfed nucleoid, termed pirellulosome [30, 31].

To study molecular physiology, cellular development and compartmentalization of this bacterium, we established a 2-DE map of soluble proteins in the pH range of 4 to 7. The master gel contains 626 annotated proteins, which were identified by PMF.

2 Materials and methods

2.1 Growth of cells and preparation of soluble proteins

Cells of *R. baltica* (DSM 10527) were grown in mineral medium with ribose (10 mM), glucose (10 mM) or *N*-acetylglucosamine (10 mM) as sole source of organic carbon [32]. Harvesting of cells was essentially performed as previously described [32]. Cells were harvested in the exponential growth phase by centrifugation (10 000 × *g*, 15 min, 4°C). The pellets were washed with 100 mM Tris/HCl pH 7.5 containing 5 mM $MgCl_2$. Cell pellets were directly frozen in liquid nitrogen and stored at −80°C until cell breakage and 2-DE. Prior to cell breakage, pellets were resuspended in 1 mL lysis buffer (7 M urea, 2 M thiourea, 2% DTT, 2% CHAPS, 0.5% carrier ampholytes; Amersham Biosciences, Freiburg, Germany). Cell breakage was performed with the PlusOne® grinding kit (Amersham Biosciences) following the manufacturer's instructions. Removal of cell debris, DNA and membranes by centrifugation (100 000 × *g*, 1 h, 15°C) yielded the fraction of soluble proteins. The protein content of this fraction was determined using the method described by Bradford [33].

2.2 2-DE, staining, and image acquisition

2-DE was essentially performed as described before [19, 32, 34]. In brief, IEF was performed using the IPGphor™ system and 24 cm long IPG strips (linear pH gradient from 4 to 7; Amersham Biosciences), followed by equilibration of the gels with DTT and iodoacetamide. The second dimension separation was then performed using the Ettan™ Dalt system (Amersham Biosciences) and gels made of 375 mM Tris/HCl, 0.1% SDS and 12.5% Duracryl (Genomic Solutions, Ann Arbor, Michigan, USA). The protein load for preparative gels was 400 µg. Proteins were visualized using colloidal Coomassie (method modified from [35]). For image acquisition the gels were digitalized with the Image Scanner (Amersham Biosciences).

2.3 Gel sample excision and processing

Excision and processing of the gel samples for PMF was performed as described previously [36], with some modifications. Protein spots were sampled from the gel using an automatic excision workstation (Proteineer; Bruker Daltonics, Bremen, Germany). The excision head was equipped with a single needle with a diameter of 2 mm. The excised gel spots were delivered into 96-well polypropylene microtiter plates (MTP) (Costar Thermowell®, Cornis, NY, USA), pre-treated by punching two holes (d < 0.5 mm) in the bottom of each well. This preparation allows removal of the washing solutions and reagents used throughout the digestion procedure by simple flow-through centrifugation, while retaining the gel particles in the wells. To protect the pierced

96-well MTP from environmental contamination, they were placed in a second 96-well MTP and covered by a lid. The second MTP also serves as collector for liquid removed by centrifugation. To ensure that no liquid from the collection MTP reaches or contaminates the pierced MTP, a spacer was placed between these two MTP. Following excision, all liquid was removed from the gel pieces by centrifugation and the sample plates were stored at −80°C prior to further processing.

Prior to digestion the gel particles were washed by incubation for 2 × 30 min in 100 μL 50% ethanol v/v. Following removal of the washing solution by centrifugation, residual water was expelled from the gel particles by incubation for 5 min in 100% ethanol. The sample plates were then placed without lid in a laminar flow-bench for 15 min to allow evaporation of the ethanol. An aliquot of freshly prepared, cooled trypsin (Roche, recombinant porcine) solution (5 μL, 10 ng/μL, 50 mM NH_4HCO_3, pH 7.8) was added to each sample. The sample plates were immediately placed in a refrigerator and incubated at 4°C for 30 min. Thereafter, an aliquot of digestion buffer (50 mM NH_4HCO_3, pH 7.8) was added to each sample, and the MTP were placed in a humidified box and incubated for 4 h at 37°C.

2.4 MALDI-MS

Protein digests were prepared for MALDI using the α-cyano-4-hydroxycinnamic acid affinity sample preparation technique described previously [37]. Mass analysis of positively charged ions was performed on an Ultraflex LIFT and a Reflex III instrument (Bruker Daltonics) operated in the reflector mode and using delayed ion extraction. Positively charged ions in the mass range 700–3 500 Da were analyzed.

2.5 Data processing and protein identification

The success rate and confidence of protein identification by PMF depends to a high degree on the accuracy of the mass measurement. High mass accuracy by MALDI-TOF MS was achieved by using internal reference compounds for spectra calibration. To calibrate the large number of spectra acquired in this study, the following procedure was developed. First, the acquired MALDI-TOF spectra were calibrated externally using a polynomial function according to a previously described procedure [38]. This calibration ensures a maximum error of 500 ppm over the entire MALDI sample support. For a subsequent internal mass correction, each spectrum was searched for signals corresponding to known reference compounds. Three peptides (Angiotensin I, MH^+ 1296.68; Neurotensin 1–13, MH^+ 1 672.9150; ACTH 18–39, MH^+ 2 465.1989; monoisotopic mass values), which were mixed into the MALDI matrix solution, and two abundant signals corresponding to trypsin autoproteolysis (MH^+ 842 510 and 2 211.1045, respectively) were used as internal references. For the spectra in which at least three of these compounds were detected, a linear regression of the relative errors for the reference signals *versus* their calculated m/z values was determined. If the standard deviation of the regression line was below 10 ppm, the regression function was used for correction of the externally calibrated mass values. If a detected calibrant had a relative error >2 SD it was discarded and the linear regression calculated again.

In some cases a sufficient number of reference masses was not detected, and in other cases, an analyte signal with a molecular mass close to the reference compound was erroneously selected as a calibrant. For example, an analyte signal that partially overlapped with the trypsin autoproteolysis signal of m/z 842 510 was erroneously selected as a calibrant. The resulting standard deviation of the linear regression was 17.7 ppm, and the calibration thus discarded. Out of 384 spectra acquired on one MALDI sample support, 190 fulfilled the criteria for internal mass correction. The remaining 194 spectra were calibrated with background signals of unknown identity, as follows: using the internally calibrated spectra, a histogram was constructed of the abundance of signals with mass differences within $\Delta m/z$ 0.05. Mass values within this interval, detected in >25 spectra in the data set were averaged and added to the list of reference masses. Using the new list of internal reference masses, the internal correction procedure was repeated with the remaining 194 spectra, this time with the requirement that at least six signals in each spectrum should match values in the calibrant list. Following this second round of internal correction, all the remaining spectra were successfully calibrated.

The presence of background signals in the spectra decreases the specificity of the database search. Background signals were assigned as described in the previous section, and removed from the data set. In addition, sodium- and potassium-cationized molecular ions, appearing as satellite signals to the protonated peptide molecular ion signal with $\Delta m/z$ 21 982 and 38.090, respectively, were removed.

Database searching was performed using the software MASCOT (Matrix Science, London, UK) [39]. The published ORF set of *R. baltica* (BX119912) was searched using the following settings: mass error tolerance: 50 ppm; fixed modifications: Cys-carbamidomethylation; variable modifications: oxidation; one tolerated missed cleavage. Under these conditions, a probability based MOWSE score >51 was considered significant ($p < 0.05$).

2.6 Generation of theoretical 2-DE gels

The published ORF set of *R. baltica* (Acc. BX119912) was used to create the theoretical 2-DE gels. M_r and pI were calculated for each predicted protein using the program *pepstats* from emboss (www.hgmp.mrc.ac.uk/Software/EMBOSS) [40]. The annotation of the published ORF set was scanned for the keywords "conserved hypothetical" and "hypothetical" in the product key of the description, generating the *conserved hypothetical* and *hypothetical* groups. The remaining proteins were sorted into the group *assigned function*.

2.7 Construction of a maximal ORF set

In order to identify proteins encoded by genes that are not present in the published ORF set of *R. baltica* (BX119912), the following strategy was employed. Based on the genomic sequence of *R. baltica*, a new ORF set was constructed by means of a PERL script according to the following steps. First, the positions of all stop codons in the genome were determined. For each stop codon, all theoretically possible reading lengths with a minimal ORF length of 102 bases were calculated by extending their sequences from the stop codon to all possible start codons detectable until the next stop codon. The resulting ORF list, denoted *Maximum ORF Set* (MOS), comprised 578 949 sequences and represents the maximal coding capacity of the genome. The MOS was translated into amino acid sequences and used as database for protein identification by PMF using data from all three analyzed 2-DE gels, as described in the Section 3.

2.8 Signal peptides and gene expression levels predictions

Signal peptides were predicted by analyzing each theoretical protein encoded by the *R. baltica* genome with the program SignalP 2.0 [41]. From this data set proteins were extracted which corresponded to identified 2-DE-separated proteins by means of a custom PERL script (using the GenDB system) [42]. Proteins with SignalP scores >0.75 were considered as potentially translocated. Expression level prediction based on codon usage optimization was calculated for each gene in the *R. baltica* genome according to the method described by Karlin *et al.* [43]. Highly expressed reference genes including ribosomal proteins, translation factors and chaperonins were extracted from the published annotation of *R. baltica* [26].

3 Results and discussion

3.1 Comparison of theoretical and experimental proteome

Three different theoretical proteome maps of *R. baltica* were created: one for proteins with "assigned function" (Fig. 1A), one for "conserved hypothetical" proteins (Fig. 1B) and one for "hypothetical proteins" (Fig. 1C). Proteins with assigned function are homologous to proteins with known functions. Conserved hypothetical proteins cannot be assigned to any function, however they have homologs in genomes of other organisms. Hypothetical proteins are also of unknown function, but they are to date not known from any organism other than *R. baltica*.

An overlay of these three maps represents the complete theoretical proteome map predicted from the annotated genome sequence. Remarkably, it shows a different isoelectric distribution pattern than those of previously reported bacterial and archaeal proteomes [44, 45]. Typically, prokaryotic theoretical proteome maps display a bimodal distribution with two protein-rich areas in the acidic and alkaline ranges, separated by a pronounced protein-depleted area around pH 7. In contrast, *R. baltica* displays a trimodal distribution with a third area of protein abundance in the neutral range. A protein peak around pH 7 has previously only been described for eukaryotic proteomes [45]. It is assumed that the bi- or trimodality of protein p*I* reflects the subcellular localization of the proteins. While cytoplasmic proteins typically have p*I* values of around 5, integral membrane proteins tend to have p*I* values of around 9. Proteins belonging to these two groups can be found in all genomes in large numbers. The nuclear proteins apparently form the third cluster in eukaryotic proteomes [45]. While a large number of proteins with neutral p*I* are predicted, the analyzed 2-DE gel (Fig. 1D–F, Fig. 2B and D) reveals only a limited number of proteins close to pH 7. Notably, the theoretical proteome of *R. baltica* contains a large number of predicted proteins with p*I* higher than 10, while the alkaline proteins of well-studied bacteria such as *E. coli* center around p*I* 9.

Functions could be assigned to only 32% of the predicted proteins of *R. baltica*. Out of these, the majority is predicted to fall into the acidic region of the theoretical 2-DE gel. Thus, the applied IEF conditions are apparently well suited to study this group of proteins. For the conserved hypothetical proteins (amounting to 14% of the predicted proteins) a similar situation was observed.

More than half (54%) of all predicted proteins belong to the hypothetical proteins, which are unique to *R. baltica*. However, these proteins are apparently under-represented in the set of proteins identified in this study. The theoretical 2-DE gel displayed in Fig. 1C reveals that the majority of hypothetical proteins have theoretical p*I* above 7, in fact, 37% of them have p*I* of greater than 10. Since a pH gradient from 4 to 7 was used in this study, these alkaline proteins could not be detected. Remarkably, many of the predicted alkaline proteins have rather low molecular mass (below 10 kDa), probably hampering their isolation by conventional 2-DE. A contribution of ORF overprediction to the high number of hypothetical proteins cannot be excluded at present. Nevertheless, it is tempting to speculate that *R. baltica* recruits hypothetical proteins for specific functions, *e.g.*, in cellular development or translocation of proteins and solutes across the complex membrane structure.

3.2 Master gel

The soluble protein fraction of *R. baltica* grown under standard conditions was visualized using 2-DE with immobilized pH gradients from 4 to 7. This fraction should represent the major part of the cytosolic proteins. Under these conditions, approximately 2000 proteins of *R. baltica* can be separated and detected, when highly sensitive protein stains such as silver or fluorescent dyes are applied (see accom-

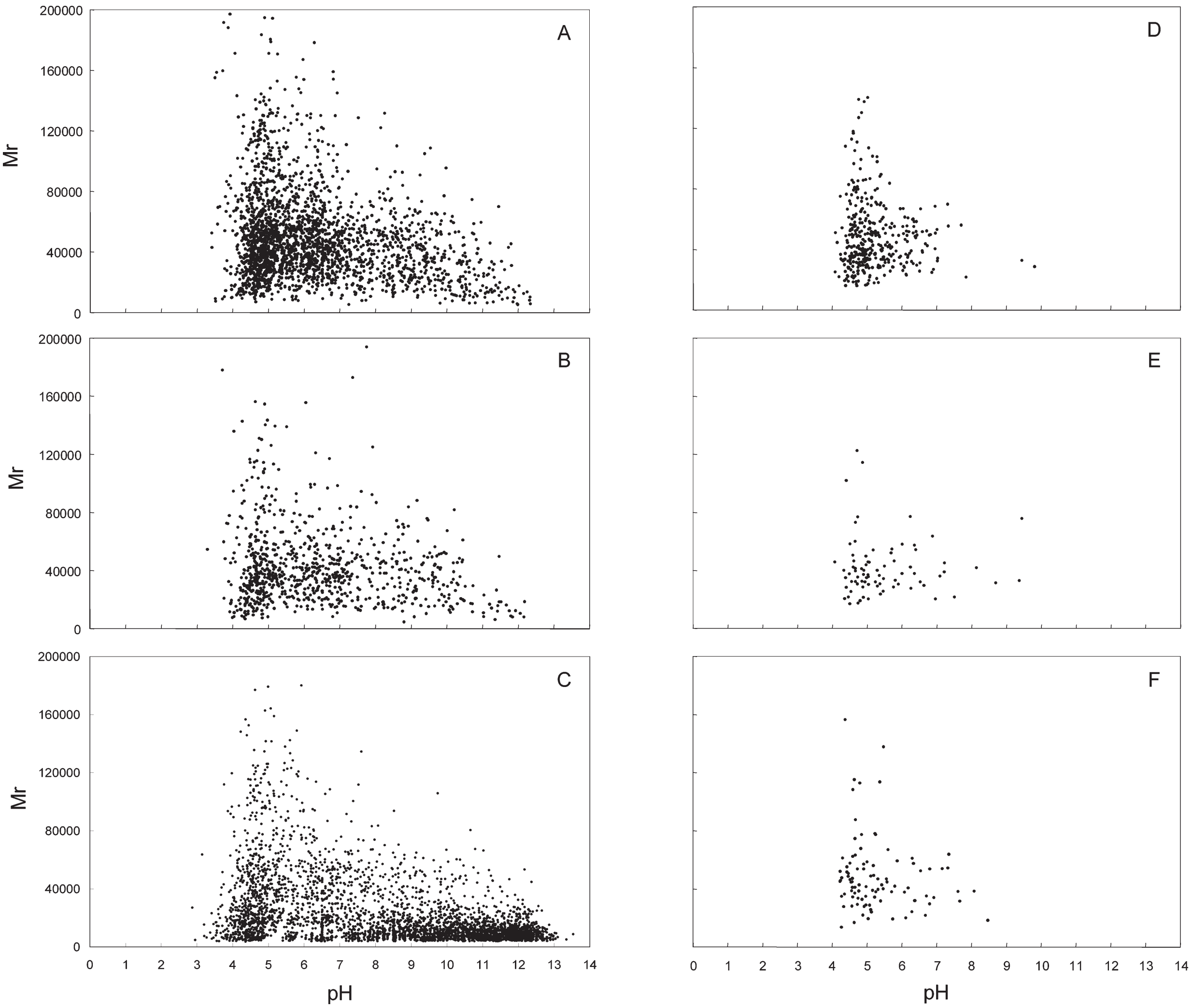

Figure 1. Theoretical 2-DE gels of proteins predicted from the genome of *R. baltica* (A–C) and the subset of proteins experimentally identified in this study (D–F). Isoelectric points and molecular weights were calculated using the "pepstats" program module of emboss. Proteins with functional assignment (A and D), conserved hypothetical proteins (B and E), hypothetical proteins (C and F).

panying publication). Figure 2A-D show the colloidal Coomassie-stained master gel from *R. baltica* cells grown with ribose. From the approximately 1000 excised gel samples, 626 proteins, represented by different spots on the gel, were identified by means of PMF ($p < 0.05$). Since some 30 proteins occurred as at least two spots, the actual number of distinct identified ORFs amounted to 558. The identified proteins were annotated in the master gel sections with the published gene numbers (Fig. 2A–D) and grouped according to functional categories (Table 1). Predicted functions of each identified protein are given in Table 2. To verify the identifications of the master gel, 2-DE and MS analysis of cells grown with glucose and *N*-acetylglucosamine, respectively, were analyzed in parallel (Table 2). Among the 558 identified gene products 301 (54%) were identified from at least two independent 2-DE gels.

Newly developed software was used for processing of the calibrated mass spectrometric data. This included filtering of sodium and potassium adduct signals, filtering of non-peptide-derived masses, filtering of signals derived from known contaminants such as trypsin autoproteolysis products, and statistical filtering of frequently occurring *m/z* values representing unknown gel sample contaminants. This processing improves the quality of the input data for the database search, thereby increasing the number of identified proteins and their respective scores. For example, in a subset of 384 samples prepared on one MALDI target, 205 spectra (53%) resulted in a significant identification score when filtering of the data was not applied. With filtering, the number of significant identification results increased to 262 (68%). Concomitantly, the average MOWSE score of the identified proteins increased from 99 to 119, thereby improving the certainty of the identification results.

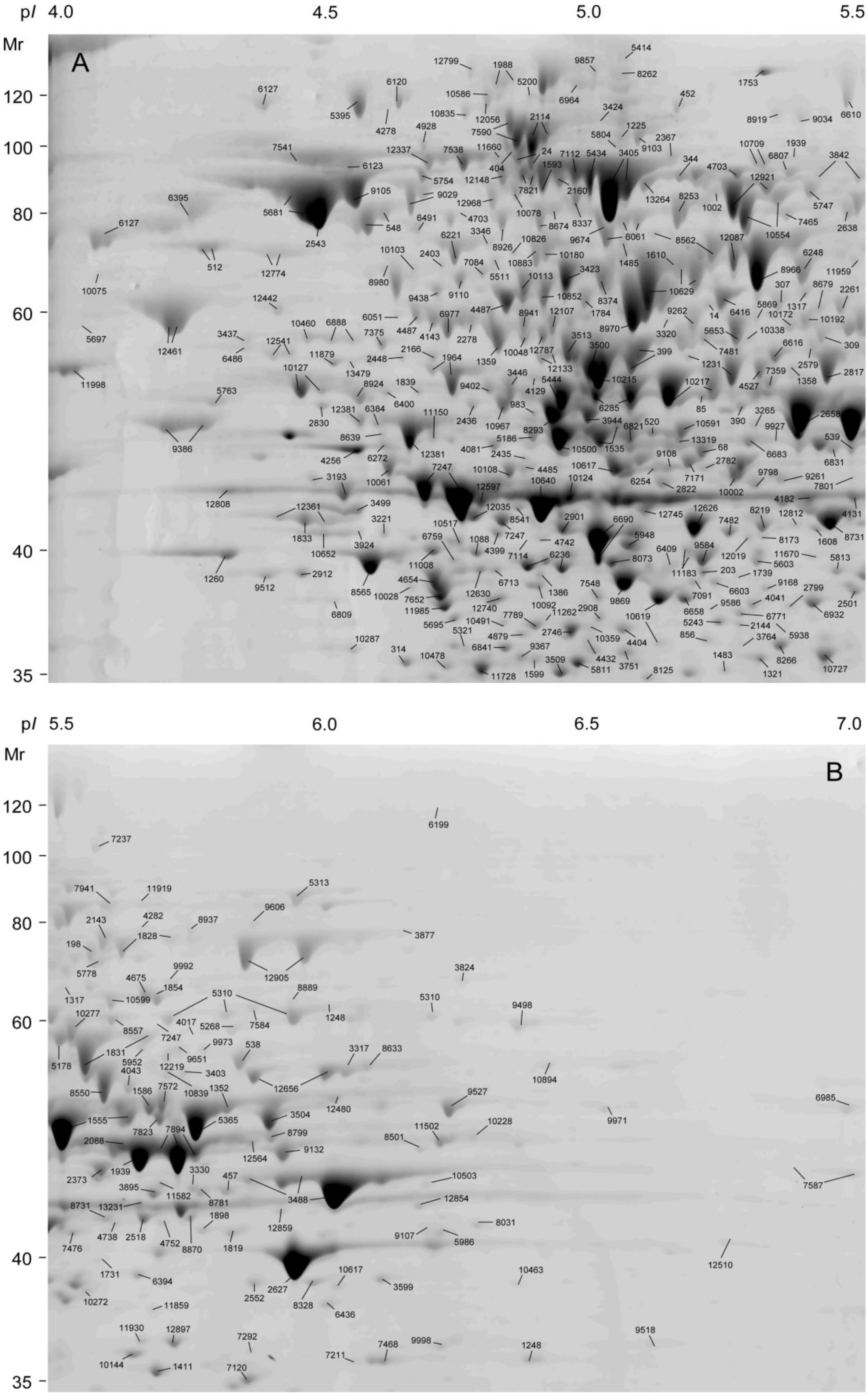
p*I* 4.0
4.5
5.0
5.5
Mr
120
100
80
60
40
35
A
12799
1988
5414
9857
8262
5200
6120
6127
10586
6964
1753
452
3424
8919
9034
6610
5395
10835
12056
2114
4278
4928
7590
1225
5804
2367
7541
12337
7538
11660
24
7112
5434
3405
9103
10709
1939
6807
3842
344
6123
404
1593
4703
12921
5754
12148
7821
2160
9105
9029
13264
8253
1002
5747
6395
5681
12968
4703
10078
7465
6127
548
6491
8674
8337
10554
2638
2543
3346
6061
8562
12087
6221
8926
10826
9674
10103
2403
7084
10883
10180
1485
1610
6248
512
12774
5511
3423
8966
11959
10075
8980
10113
10629
307
8679
2261
9438
9110
10852
8374
5869
12442
6051
6977
4487
8941
12107
1784
9262
14
6416
10172
1317
10192
5697
12461
3437
10460
6888
7375
4487
4143
2278
8970
3320
5653
10338
6486
12541
11879
2166
1359
10048
12787
3513
3500
399
7481
6616
309
2448
1964
12133
1231
2579
10127
13479
11998
3446
10215
7359
2817
8924
1839
9402
5444
4129
10217
4527
1358
5763
6400
983
85
12381
6384
11150
6285
3944
10591
390
3265
2658
2830
2436
10967
8293
6821
520
9927
9386
8639
5186
13319
539
4081
2435
10500
1535
68
6683
4256
6272
12381
9108
2782
6831
7247
10108
4485
10617
3193
10061
10640
10124
6254
7171
9798
9261
12597
2822
10002
7801
12808
12361
3499
12035
4182
3221
10517
8541
2901
6690
12745
12626
7482
8219
12812
4131
1833
6759
1088
7247
4742
5948
9584
8173
1608
8731
10652
3924
4399
7114
6236
6409
12019
11670
5813
1260
2912
11008
8073
11183
203
1739
5603
9512
4654
6713
1386
7548
9168
2799
8565
10028
7652
12630
9869
7091
6603
4041
2501
11985
12740
10092
2908
6658
9586
6932
6809
5695
10491
7789
11262
10619
5243
2144
6771
5321
4879
2746
856
3764
5938
10287
10359
4404
314
6841
9367
10478
4432
3509
3751
1483
8266
11728
1599
5811
8125
1321
10727
p*I* 5.5
6.0
6.5
7.0
Mr
120
100
80
60
40
35
B
6199
7237
5313
7941
11919
9606
2143
4282
8937
3877
1828
198
12905
3824
5778
9992
4675
1854
8889
1317
10599
5310
5310
9498
10277
1248
8557
4017
5268
7584
7247
9973
1831
538
3317
8633
5178
5952
9651
4043
12219
3403
10894
1586
7572
10839
1352
12656
8550
9527
6985
12480
1555
5365
3504
9971
7823
7894
11502
10228
2088
8799
8501
9132
12564
1939
3330
7587
2373
3895
457
10503
11582
3488
8731
13231
8781
12854
1898
8031
7476
4738
2518
4752
8870
1819
12859
9107
5986
12510
1731
6394
10617
3599
10463
2627
10272
2552
8328
6436
11859
11930
12897
9518
7292
9998
7468
1248
7211
10144
1411
7120

Figure 2.

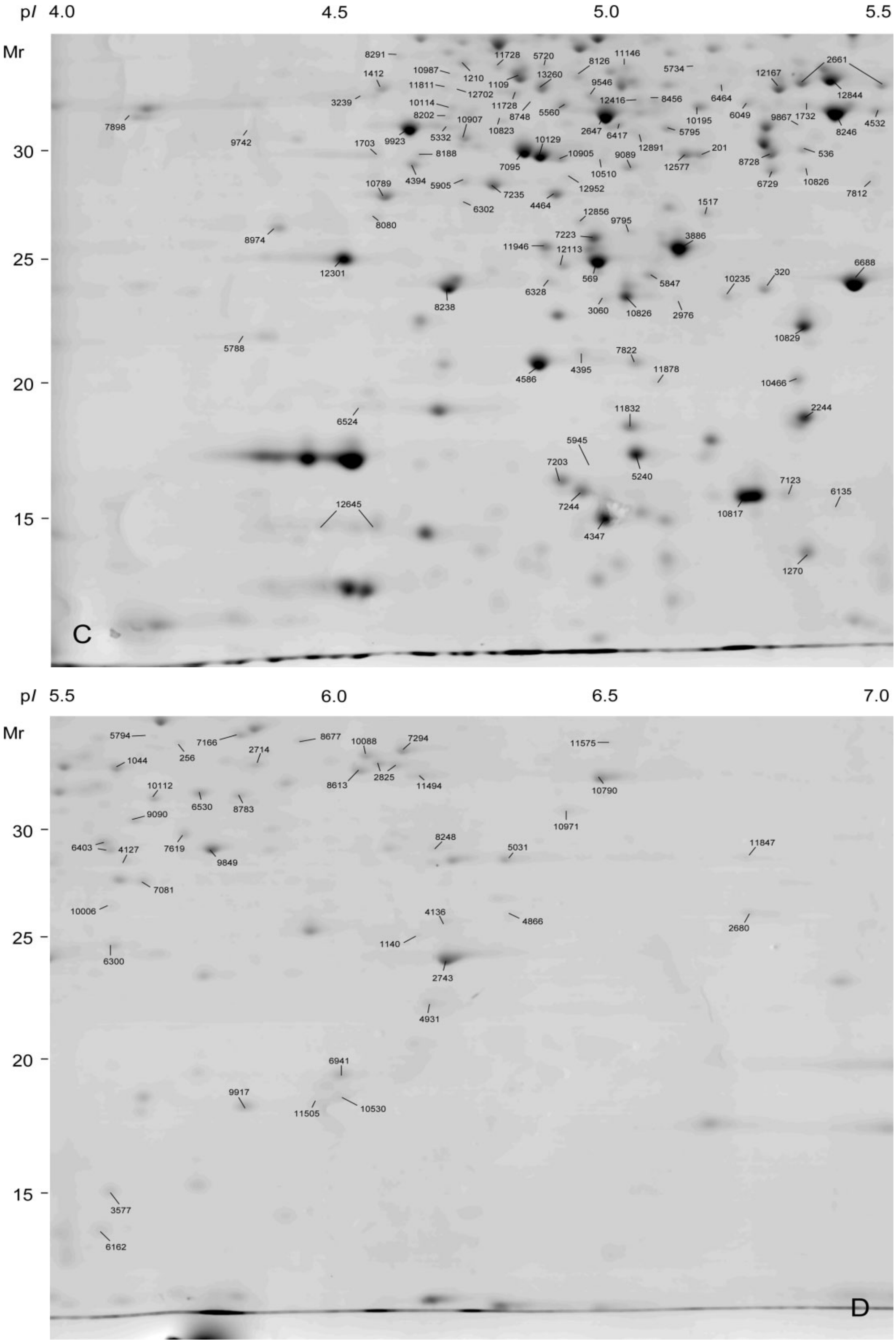

Figure 2. Annotated sections (A–D) of the master gel of soluble proteins of *R. baltica* grown with ribose. Assigned numbers represent genes. Table 2 lists the identified proteins according to functional classes and provides for each protein information on functional prediction, quality/reproducibility of identification, and prediction of signal peptides and expression level.

3.3 Proteins with functional assignment

Functions could be assigned to the majority (366) of the identified proteins. These proteins were classified according to their predicted functions into nine categories (Table 1). The category "metabolism" was further divided into eight sub-categories (Table 2). The sub-category "sulfatases" was used, since the presence of 110 sulfatase-encoding genes was one of the major unexpected findings from the annotation of the *R. baltica* genome [26], 10 of which were identified on the gel.

The pattern of identified proteins displayed in Fig. 2 is typical of exponentially growing cells, with the most abundant proteins involved mainly in housekeeping functions, *e.g.*, GAPDH (RB2627) of glycolysis, malate dehydrogenase (RB7652) of tricarboxic acid cycle, glutamate synthase (RB5653) of amino acid biosynthesis, protease (RB9402) of protein quality control and translation. The fact that several proteins existed as more than a single spot could point to thus far unknown post-translational modifications. In particular, proteins with high molecular weight formed chains with the same M_r but differing p*I*.

Table 1. Distribution of identified proteins among functional groups

Functional groups	Number of identified proteins	Share (%) among identified proteins
Metabolism	250	45
Genetic information processing	52	9
Regulation and signal transduction	24	4
Stress response	13	2
Energy	8	1
Transport	11	2
Conserved hypothetical proteins	94	17
Hypothetical proteins	98	18
Others	8	1

3.4 Hypothetical and conserved hypothetical proteins

About 18% of the identified proteins represent predicted hypothetical proteins that are apparently unique to *R. baltica*. Thus the present study for the first time provides experimental evidence that genes coding for hypothetical proteins are actually expressed under standard growth conditions and consequently have to be considered relevant for the physiology of *R. baltica*. Conserved hypothetical proteins constitute about the same percentage of identified proteins. In both cases no functions could be assigned. Nevertheless, the hypothetical proteins could be of particular interest with respect to the cell cycle and unusual morphological features of *R. baltica*, which may require the activity of thus far unknown proteins.

3.5 Identification of unpredicted proteins

Initial analysis of the *R. baltica* genome sequence with three different ORF prediction programs (Orpheus, Glimmer, and Critica) generated a non-redundant set of 13 331 predicted ORFs. Manual removal of presumably overpredicted ORFs resulted in the published set of 7325 ORF (BX119912) [26]. Thus, the possibility exists that ORFs were initially not predicted or were erroneously removed during manual refinement. This possibility was compounded by the observation that several PMFs with abundant signals did not result in identification of a protein.

As a first attempt to identify proteins encoded by genes that were not present in the predicted ORF set, the PMF data were searched against an amino acid sequence database translated from a *Maximal ORF Set* (MOS); a highly redundant set consisting of 578 949 sequence entries, designed to contain all possible genes and all possible reading lengths thereof. To reduce the number of false positive results, identifications for which the experimental and calculated molecular weight differed by >30% were discarded. This database search retrieved four proteins with scores >51, which are listed in Table 3. All of the newly identified genes code for hypothetical proteins, which are surrounded by further hypothetical or conserved hypothetical proteins in the genomic context. For example, ORF 9191 from MOS was identified with a MASCOT score of 101 and sequence coverage of 62%. The position of the corresponding spot on the 2-DE gel was used as a guide to suggest the ORF length by defining the probable start codon. The product of ORF 9191 is therefore predicted to have a molecular mass of 25 kDa. These results indicate that PMF is not necessarily restricted to identification in protein databases, but can also be used to refine ORF prediction. However, future analysis should include MS/MS to verify the identity of the additional proteins.

3.6 Signal peptides and protein localization

As observed with other described *Planctomycetes*, cells of *R. baltica* contain membrane-separated intracytoplasmic compartments [31]. The internal region is termed pirellulosome and contains the riboplasm with ribosome-like particles and the condensed nucleoid (Fig. 3). The region between the intracytoplasmic and cytoplasmic membranes contains the paryphoplasm that harbors some RNA but no ribosome-like particles. The finding that ribosome-like particles are confined to the riboplasm suggests that protein biosynthesis only takes place in this compartment. Due to the cellular compartmentalization in *R. baltica* an extensive protein translocation can be expected.

Table 2. Predicted functions of proteins annotated in the master gel (see Fig. 2A–D)[a)]

ORF	Putative Function	sp	PHX	Score	*
	Metabolism (250 proteins)				
	C-compound and Carbohydrate				
201	Sugar phosphate isomerase/ epimerase			190	1
307	NAD dependent malic enzyme		+	201	2
344	Xanthan lyase	+		132	
399	Glucose-6-phosphate isomerase			139	1
548	1,4-alpha-glucan branching enzyme			233	2
856	L-Lactate/malate dehydrogenase	+		131	
1210	Hexulose-6-phosphate isomerase		+	105	1
1231	Dihydrolipoamide dehydrogenase		+	135	2
1358	ADP-glucose pyrophosphorylase			198	
1412	Inositol monophosphatase			124	1
1593	Isocitrate dehydrogenase			192	2
1988	Glucose dehydrogenase	+		283	
2114	Aconitate hydratase		+	178	2
2160	Alpha-Amylase		+	229	
2373	Formaldehyde dehydrogenase		+	213	2
2403	D-mannonate oxidoreductase			150	
2518	GDP-mannose 4,6 dehydratase		+	162	2
2627	Glyceraldehyde 3-phosphate dehydrogenase		+	230	2
2638	Glycogen branching enzyme			133	1
2658	Xylose isomerase		+	196	2
2817	6-Phosphogluconate dehydrogenase		+	194	2
3193	Transaldolase		+	233	2
3239	D-tagatose 3-epimerase		+	54	1
3265	Glucose-fructose oxidoreductase			117	
3423	Pyruvate dehydrogenase, E2 component		+	144	2
3424	Pyruvate dehydrogenase, E1 component		+	288	
3488	Sorbitol dehydrogenase			244	
3499	Ribokinase			124	
4131	Alcohol dehydrogenase			108	2
4654	Sugar phosphate isomerase/ epimerase		+	124	2
5200	Alpha-Amylase			129	
5243	Endo-1,4-beta-xylanase B			72	2
5321	Myo-inositol catabolism protein IolH			99	
5948	Alcohol dehydrogenase			157	2
6061	Phosphomannomutase			310	2
6254	Mannose-1-phosphate guanylyltransferase			80	
6394	2-Hydroxy acid dehydrogenase			162	2
6683	Citrate synthase			150	2
6690	Fructose-1,6-bisphosphate aldolase		+	130	2
6729	Deoxyribose-phosphate aldolase		+	55	1
6759	Methenyltetrahydromethanopterin cyclohydrolase			60	
6807	Sialic acid-specific 9-*O*-acetylesterase	+		131	
6841	UDP-*N*-Acetylglucosamine pyrophosphorylase			142	1
6977	UDP-*N*-Acetylhexosamine pyrophosphorylase			158	2
7095	Triosephosphate isomerase		+	109	2
7294	Glucose 1-dehydrogenase			159	1
7572	6-Phosphofructokinase, pyrophosphate-dependent		+	105	1
7652	Malate dehydrogenase		+	138	2
8073	Alpha-L-arabinofuranosidase II	+		151	2
8248	Carboxymethylenebutenolidase			112	
8541	Endoglucanase		+	147	2
8562	Phosphoglycerate mutase			82	
8731	2-keto-3-deoxygluconate kinase		+	91	2
8924	Phosphonopyruvate decarboxylase 1		+	131	2
8941	Ketoglutarate semialdehyde dehydrogenase			67	2
9089	6-Phosphogluconolactonase			117	2
9651	Sialic acidspecific 9-*O*-acetylesterase			80	
10002	Glucose dehydrogenase		+	133	2
10048	Sialic acidspecific 9-*O*-acetylesterase		+	140	
10092	Hydratase, aerobic aromate catabolism			74	2
10124	Polyvinylalcohol dehydrogenase	+		147	
10127	PQQ-dependent glucose dehydrogenase	+		89	2
10144	Endo-1,4-beta-xylanase B	+		121	
10172	Aldehyde dehydrogenase		+	218	2
10277	Pyruvate kinase			138	2
10500	Phosphoglycerate kinase		+	142	2
10554	Succinate dehydrogenase subunit A		+	238	2
10591	PPi-Phosphofructokinase		+	164	2
10617	Succinyl-CoA synthetase beta subunit		+	99	2

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
10619	Succinyl-CoA synthetase alpha subunit		+	103	2
10817	Ribose 5-phosphate epimerase		+	81	
12361	Ribokinase family sugar kinase			175	2
12381	Enolase		+	212	2
12740	Gluconolactonase precursor			101	
12921	Transketolase		+	151	2
13260	Alcohol dehydrogenase			73	
13264	Acetyl-coenzyme A synthetase			220	2
	Amino Acids and Proteins				
1225	Dipeptidyl peptidase IV	+		158	
1317	2-Isopropylmalate synthase			147	
1359	Serine protease			156	2
1411	Dihydrodipicolinate synthase		+	178	2
1732	Beta-Alanine synthetase	+		73	
1898	Dehydroquinate synthase			101	
2261	Carboxypeptidase-related protein		+	187	1
2278	3-Phosphoshikimate 1-carboxyvinyltransferase			131	
2552	*N*-Acetyl-gamma-glutamyl-phosphate reductase		+	172	1
2661	UDP-*N*-Acetyl-enolpyruvoylglucosamine reductase			63	2
2746	Dihydrodipicolinate synthase			100	
3824	L-Aspartate oxidase			176	
3842	Dipeptidyl peptidase IV	+		220	
4282	Matrix metalloproteinase-11			71	
4394	Proteinase		+	65	
4928	Aminopeptidase			244	
5444	S-Adenosylmethionine synthetase		+	234	2
5560	Tryptophan synthase alpha chain			126	2
5653	NADH-Glutamate synthase small chain		+	134	1
5720	Amidohydrolase			74	
5986	Ornithine carbamoyltransferase			151	
6248	Phosphoglycerate dehydrogenase			149	2
6285	Adenosylhomocysteinase		+	152	2
6300	Glutamine amido-transferase			119	
6821	Aspartate aminotransferase		+	88	1
6932	Cysteine synthase			126	2
7359	Gamma-glutamyl phosphate reductase		+	99	1
7375	Aminopeptidase T			115	1
7584	Glycine dehydrogenase (decarboxylating) subunit 2			154	
7587	Aminotransferase-glycine cleavage system T protein			238	
7590	Proteinase			174	2
7823	Transaminase			123	
7941	cysN/cysC bifunctional enzyme			245	2
8080	Phosphoribosylformimino-5-aminoimidazole carboxamide botide isomerase, biosynthesis of histidine			127	
8126	Branched-chain amino acid aminotransferase		+	111	2
8219	Aspartate aminotransferase			135	2
8262	Proline dehydrogenase			147	
8293	Argininosuccinate synthase		+	144	2
8633	Acetylornithine aminotransferase		+	119	
8926	Aspartokinase		+	93	
9029	Metalloproteinase			200	2
9107	Chorismate mutase			150	
9402	Protease		+	170	1
9674	X-Pro dipeptidyl-peptidase	+		134	1
9795	Aspartate-semialdehyde dehydrogenase		+	117	2
9857	5-Methyltetrahydrofolate-homocysteine methyltransferase			117	
9869	Acetohydroxy acid isomeroreductase		+	170	2
10112	Imidazole glycerol phosphate synthase subunit hisF			221	2
10114	Indole-3-glycerol phosphate synthase		+	161	
10180	Peptidase		+	173	2
10272	Pteridine reductase			136	1
10287	Dihydropicolinate synthase		+	166	2
10586	Aminopeptidase			165	1
10826	ATP-dependent clp protease proteolytic subunit			120	2
10829	ATP-dependent clp protease proteolytic subunit		+	88	1
10894	Threonine synthase precursor		+	109	2
11847	Methionine sulfoxide reductase	+		81	
11878	Methionine sulfoxide reductase			56	
11879	Periplasmic serine proteinase		+	91	
11919	Dihydroxy acid dehydratase			59	
11959	Dihydrodipicolinate reductase			114	2
12087	Dihydroxy acid dehydratase		+	94	2
12107	Cytosol aminopeptidase		+	165	2

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
12113	Carbamoyl-phosphate synthase large chain		+	177	2
12133	Succinyl-diaminopimelate desuccinylase		+	178	1
12148	Periplasmic tail-specific proteinase	+	+	287	2
12337	Prolyl endopeptidase	+	+	240	2
12510	Phospho-2-dehydro-3-deoxyheptonate aldolase			143	
12597	3-Isopropylmalate dehydrogenase			131	
12656	3-Isopropylmalate dehydratase large subunit			226	2
12905	Acetolactate synthase III precursor		+	172	2
	Nucleotides				
1964	DNA-directed RNA polymerase alpha chain		+	191	2
256	Formyltetrahydrofolate deformylase			145	
1386	Nucleoside hydrolase	+		128	
1784	UDP-glucose 6-dehydrogenase			195	2
1819	UDP-glucose 4-epimerase			176	2
3751	UDP-glucose 4-epimerase			146	2
4043	Glucose-1-phosphate thymidylyltransferase			89	
4752	Dihydroorotate dehydrogenase			101	
5395	Phosphoribosylformylglycinamidine synthase II		+	367	2
5603	ATP phosphoribosyltransferase			189	2
5695	Beta-alanine synthetase			173	1
5847	Adenine phosphoribosyltransferase			92	2
6135	Phosphoribosylaminoimidazole carboxylase catalytic subunit			67	
6302	ADP-ribose pyrophosphatase			79	
6328	Adenylyl cyclase			107	
6524	Hypoxanthine-guanine phosphoribosyltransferase			79	
6616	Phosphoribosylamine-glycine ligase		+	99	1
7468	Methylentetrahydrofolate cyclohydrolase			59	
8374	GMP synthase		+	265	2
8613	Phosphoribosylformylglycinamidine synthase I			133	2
8748	Dihydroorotate dehydrogenase		+	65	
10113	Bifunctional purine biosynthesis protein purH		+	224	2
10192	Dihydroorotase	+	+	96	2
10510	Cytidylate kinase		+	89	
11832	Nucleoside diphosphate kinase		+	71	2
12745	Phosphoribosylformylglycinamidine cyclo-ligase			109	2
	Lipids, Fatty Acids and Isoprenoids				
314	Malonyl CoA-acyl-carrier-protein transacylase		+	101	2
320	3-Oxoacyl-(acyl-carrier-protein) synthase			173	2
1586	3-Oxoacyl-(acyl-carrier-protein) synthase II			104	2
1839	Thiamine biosynthesis lipoprotein apbE			136	2
2144	Geranylgeranyl pyrophosphate synthetase precursor			78	
2579	Ethanolamine utilization protein EutE		+	67	
2825	Glycerophosphodiester phosphodiesterase			113	1
4527	3-Oxoacyl-(acyl-carrier-protein) synthase		+	177	1
6272	3-Oxoacyl-(acyl-carrier-protein) synthase			97	
6464	Sulfolipid biosynthesis protein			85	
7171	3-Oxoacyl-(acyl-carrier-protein) synthase		+	219	2
7812	Enoyl-CoA hydratase/isomerase			95	
8125	Trans-2-enoyl-(acyl-carrier-protein) reductase		+	75	2
8550	Biotin carboxylase		+	212	2
10466	Probable beta-hydroxyacylACP dehydratase		+	72	1
10790	Enoyl-(acyl-carrier-protein) reductase (NADH)		+	127	2
12812	3-Oxoacyl-(acyl-carrier-protein) synthase III			60	
	Vitamins, Cofactors and Prosthetic Groups				
24	L-sorbosone dehydrogenase			197	2
309	Magnesium protoporphyrin chelatase			95	2
536	Pyridoxal phosphate biosynthetic protein			128	1
2143	1-Deoxy-D-xylulose 5-phosphate synthase			226	

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
6809	Thiamine-monophosphate kinase			69	1
6831	Glutamate-1-semialdehyde 2,1-aminomutase			89	1
6964	L-sorbosone dehydrogenase	+		123	
9090	3-Methyl-2-oxobutanoate hydroxymethyltransferase			64	
10006	Pyridoxamine oxidase			71	
11582	Cysteine desulfurase			86	1
12480	Riboflavin biosynthesis protein RibA		+	91	
	Sulfatases				
198	*N*-acetylgalactosamine-4-sulfatase precursor	+		129	
1610	Arylsulfatase	+		100	
2367	Sulfatase	+	+	121	
3403	*N*-acetylgalactosamine 6-sulfatase	+		65	
3877	Arylsulfate sulphohydrolase	+		137	
4017	Sulfatase			114	
7481	Arylsulphatase A	+	+	162	
9498	Arylsulfatase	+	+	139	2
10599	Sulfatase 1 precursor	+		161	
11502	Alkylsulfatase	+		60	2
	Inorganic Compounds				
5869	Bacterioferritin comigratory protein			79	
6049	Adenylylsulfate kinase			110	
7247	Glutamine synthetase II		+	136	2
7465	Sulfite reductase		+	65	
11670	Ferric enterobactin esterase-related protein			117	
	Others				
203	Oxidoreductase			154	2
1555	NADH-dependent dehydrogenase	+	+	154	2
1608	Esterase	+		125	2
1939	Oxidoreductase		+	212	
2242	Oxidoreductase		+	164	1
3317	NADH-dependent dehydrogenase	+		108	
3330	Dehydrogenase		+	104	1
3405	Hydrolase	+		275	2
4404	Oxidoreductase			124	2
4432	Oxidoreductase	+		67	1
5332	Phosphoesterase			53	
5365	NADH-dependent dehydrogenase	+	+	165	2
6199	Dehydrogenase	+	+	94	2
6985	NADH-dependent dehydrogenase		+	86	
7081	Oxidoreductase			71	
7482	CDP-tyvelose epimerase		+	183	2
7548	Syringomycin biosynthesis enzyme 2			133	2
8679	Oxidoreductase			121	
8728	Oxidoreductase		+	119	2
8781	NADH-dependent oxidoreductase		+	182	
8799	NADH-dependent dehydrogenase			159	2
8937	NADH-dependent dehydrogenase			94	
9168	Nucleotide sugar epimerase			68	2
9584	Oxidoreductase			74	
9586	Oxidoreductase			116	
9971	NADH-dependent dehydrogenase	+	+	116	1
10503	NADH-dependent oxidoreductase		+	142	2
10652	*C*-methyltransferase			103	1
10967	Oxidoreductase		+	91	
10971	Dehydrogenase			96	1
11146	Hydrolase			99	
11859	Hydrolase			167	
12019	Oxidoreductase			74	1
12564	NADH-dependent dehydrogenase	+		139	1
	Stress Response (13 proteins)				
390	Alkylhalidase, dehalogenase			167	
2244	Glutathione peroxidase	+	+	75	
2799	General stress protein 69			107	2
4586	Thiol peroxidase		+	92	2
6384	Thioredoxin related protein	+		116	
6688	Superoxide dismutase, Mn family		+	86	2
7223	Thioredoxin reductase		+	76	2
8238	Peroxiredoxin 2		+	72	2
8674	Thioredoxin	+	+	145	
8870	Multidrug resistance protein		+	127	1
10727	Manganese-containing catalase			73	2
11150	Xenobiotic reductase B			114	1
12541	Thioredoxin	+		214	2
	Transport (11 proteins)				
1248	ATPase component; multidrug transport system		+	117	2
1517	ATP-binding protein, lipoprotein releasing system			75	
4866	ATP-binding protein, lipoprotein releasing system			69	
5795	PTS system, fructose-specific IIABC component			114	2
6236	ATP-binding protein, ABC-transport system			147	2

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
7166	ATP-binding protein, ABC-transport system			89	1
7211	ATP-binding protein, phosphate transport			186	
9998	ATP-binding protein, ABC-transport system		+	67	
10709	Periplasmic dipeptide transport protein precursor	+		203	
11930	ATP-binding protein, ABC-transport system			158	2
12859	ATP-binding protein, oligopeptide transport			124	
	Genetic Information Processing (52 proteins)				
539	Competence-damage inducible protein CinA			93	1
1270	Translation initiation inhibitor		+	65	
1485	DNA polymerase beta family			193	
1964	DNA-directed RNA polymerase alpha chain		+	191	2
2543	30S ribosomal protein S1		+	159	2
3446	Peptidyl-prolyl *cis-trans* isomerase cyp2			100	2
3886	Ribosome recycling factor			106	2
4143	Glutamyl-tRNA amidotransferase subunit A			127	2
4395	Macrophage infectivity potentiator (map) protein		+	83	
4675	Cysteinyl-tRNA synthetase			108	2
5178	Prolyl-tRNA synthetase			218	2
5414	DNA-directed RNA polymerase beta chain		+	53	
5434	Elongation factor G		+	126	1
5681	Trigger factor		+	142	2
5697	Thiol-disulfide interchange protein			52	
5747	Arginyl-tRNA synthetase			205	1
5754	DnaK			97	1
5778	Alkaline phosphatase			234	
5804	Polyribonucleotide nucleotidyltransferase		+	56	
5813	Alkaline phosphatase D	+		105	
6123	Protein disulfide-isomerase		+	201	2
6436	Tryptophan-tRNA synthetase			79	1
7112	Phenylalanyl-tRNA synthetase beta chain			131	
7114	Phenylalanyl-tRNA synthetase alpha chain			129	1
7237	DNA mismatch repair protein MUTS		+	98	
7244	Peptidylprolyl *cis-trans* isomerase			61	1
7821	Elongation factor G	+	+	253	2
7894	Elongation factor Tu		+	187	2
8253	Aspartyl-tRNA synthetase		+	249	2
8328	CMP-binding protein		+	151	
8649	Peptidylprolyl *cis-trans* isomerase	+	+	100	
8889	Alkaline phosphatase D precursor	+		132	
8919	Leucyl-tRNA synthetase		+	172	
8966	60 kDa chaperonin		+	198	2
8970	60 kDa chaperonin		+	148	2
8974	GrpE chaperone		+	57	
9103	ATPases with chaperone activity, ATP-binding subunit			155	1
9105	DnaK		+	168	
9917	Single-strand binding protein			110	
9923	50S ribosomal protein L25		+	117	2
9927	ATP-dependent Clp protease ATP-binding subunit		+	89	
10108	DNA polymerase III, beta chain			212	1
10129	Macrophage infectivity potentiator (map) protein	+	+	76	
10629	GroEL		+	172	2
10640	Elongation factor Ts		+	211	2
10852	Glutamyl-tRNA amidotransferase subunit B		+	114	2
10883	Lysyl-tRNA synthetase	+	+	153	2
12577	Elongation factor P			90	
12626	DNA-directed RNA polymerase alpha chain		+	287	2
12799	DNA polymerase I		+	205	
12854	Methionyl-tRNA formyltransferase			159	2
12856	Peptide deformylase		+	121	1
	Regulation and Signal Transduction (24)				
983	Phosphoprotein kinase	+		189	
1140	Response regulator			101	
1321	Transcription repressor			93	
1483	Sensor histidine kinase/response regulator			58	
2743	Nitrate/nitrite regulatory protein NarP		+	159	2
4081	Regulatory protein			146	2
4136	Regulatory components of sensory transduction system		+	53	
4487	Nitrogen assimilation regulatory protein			168	2
5905	Phosphoprotein phosphatase			60	
6403	Response regulator			119	1
6486	Phosphoprotein kinase	+		165	
6491	RNA polymerase subunit sigma54			134	
6603	MoxR-related protein		+	164	1
7123	Response regulator			56	
7541	Phosphoprotein kinase	+	+	171	2

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
7898	Transcription antiterminator NusG		+	109	1
8173	MoxR-related protein			113	
9108	MoxR-related protein			109	2
9110	Phosphoprotein kinase			162	2
10491	Two-component system regulatory protein			71	
10517	Methanol dehydrogenase regulation homolog YeaC			134	1
10839	Phosphoprotein kinase			101	
11660	Phosphoprotein kinase	+		218	
12952	Two-component system, regulatory protein			69	
	Energy (8 proteins)				
1831	Na+-translocating NADH:ubiquinone oxidoreductase NqrA			220	2
1833	Na+-translocating NADH:ubiquinone oxidoreductase NqrC	+	+	106	
4399	Quinone oxidoreductase			185	1
7084	Pyrophosphatase			165	
10215	H+-transporting ATP synthase alpha chain		+	124	2
10217	H+-transporting ATP synthase beta chain		+	307	2
11946	Thermophilic NAD(P)H-flavin oxidoreductase			125	2
11985	Quinone oxidoreductase			191	1
	Others (8 proteins)				
3895	Internalin	+		93	
4879	Nodulin-26			117	1
10228	Twitching motility protein PilB, biogenesis of pili		+	118	
10338	FlbA protein, biogenesis of flagellae			88	
10463	Ferredoxin-NADP reductase			145	
10905	Phosphoesterase PH1616		+	132	1
10907	Phosphoesterase PH1616			143	2
12774	Type IV fimbrial assembly protein PilB			269	1
	Conserved Hypothetical Proteins (94 proteins)				
85	Conserved hypothetical protein	+		114	
452	Conserved hypothetical protein			215	1
457	Conserved hypothetical protein			183	
520	Conserved hypothetical protein			145	1
538	Conserved hypothetical protein			232	2
569	Conserved hypothetical protein	+	+	74	2
1044	Conserved hypothetical protein			221	1
1109	Conserved hypothetical protein	+	+	70	2
1703	Maf protein			56	
1731	Conserved hypothetical protein			75	
1739	Conserved hypothetical protein			67	
1753	Conserved hypothetical protein		+	76	2
1854	Conserved hypothetical protein			144	
2435	Conserved hypothetical protein	+		157	
2680	Conserved hypothetical protein	+		91	1
2714	Conserved hypothetical protein			122	
2908	Conserved hypothetical protein	+		76	
2912	Conserved hypothetical protein	+		143	2
2976	Conserved hypothetical protein			102	
3221	Conserved hypothetical protein			161	
3509	Conserved hypothetical protein	+		183	
3599	Ring canal kelch protein	+		133	
3924	Conserved hypothetical protein			107	1
3944	Conserved hypothetical protein		+	167	2
4127	Conserved hypothetical protein	+		99	
4129	Conserved hypothetical protein		+	105	2
4278	Conserved hypothetical protein			81	
4347	Conserved hypothetical protein		+	109	
4485	Conserved hypothetical protein			86	2
4532	Conserved hypothetical protein	+		83	
4738	Conserved hypothetical protein			130	2
4742	Conserved hypothetical protein			89	1
5186	Conserved hypothetical protein			55	1
5313	Conserved hypothetical protein		+	157	

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
5511	Conserved hypothetical protein			161	
5788	Conserved hypothetical protein	+	+	58	
5952	Conserved hypothetical protein	+		133	
6120	TolB protein	+		105	2
6395	Conserved hypothetical protein			65	2
6409	Conserved hypothetical protein	+		113	
6416	Conserved hypothetical protein			133	1
6417	Conserved hypothetical protein			81	2
6530	Conserved hypothetical protein			70	
7091	Conserved hypothetical protein			84	
7120	Conserved hypothetical protein	+		102	1
7292	Conserved hypothetical protein			86	
7538	Conserved hypothetical protein		+	158	1
7619	Conserved hypothetical protein			79	2
7789	TolB protein [precursor]			61	
7822	Conserved hypothetical protein			96	2
8031	Conserved hypothetical protein	+		122	1
8188	Conserved hypothetical protein			58	
8202	Conserved hypothetical protein		+	86	
8246	Conserved hypothetical protein	+		184	2
8266	Conserved hypothetical protein		+	60	1
8291	Conserved hypothetical protein			53	
8456	Conserved hypothetical protein			73	
8501	Conserved hypothetical protein	+		158	
8557	Conserved hypothetical protein			130	
8565	Conserved hypothetical protein		+	136	2
8639	Conserved hypothetical protein	+		168	2
8677	Conserved hypothetical protein			152	
8783	Conserved hypothetical protein			102	

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
9132	Conserved hypothetical protein		+	138	2
9261	Conserved hypothetical protein		+	139	
9262	Conserved hypothetical protein		+	113	1
9367	Conserved hypothetical protein	+		64	
9386	FixW protein	+	+	64	2
9438	Conserved hypothetical protein	+		134	2
9546	Conserved hypothetical protein	+		91	1
9606	Conserved hypothetical protein	+		78	
9849	Conserved hypothetical protein	+	+	175	
9992	Conserved hypothetical protein		+	151	
10028	Conserved hypothetical protein			77	1
10061	Conserved hypothetical protein			123	2
10078	Conserved hypothetical protein	+		299	
10088	Conserved hypothetical protein			100	1
10103	Conserved hypothetical protein containing kelch-motif	+	+	102	
10195	Conserved hypothetical protein	+	+	90	2
10235	Conserved hypothetical protein			56	
10359	Conserved hypothetical protein			129	
10478	Conserved hypothetical protein	+		140	2
10789	Conserved hypothetical protein		+	96	
10987	Conserved hypothetical protein			113	
11183	Conserved hypothetical protein			181	2
11262	Conserved hypothetical protein			101	1
11494	Conserved hypothetical protein	+		60	
11505	Conserved hypothetical protein	+	+	68	
11728	Conserved hypothetical protein		+	177	2
11811	Conserved hypothetical protein	+		70	
11998	Conserved hypothetical protein	+	+	156	2

Table 2. Continued

ORF	Putative Function	sp	PHX	Score	*
12056	Conserved hypothetical protein containing TPR domain	+		281	2
12301	Conserved hypothetical protein	+	+	79	
12891	Conserved hypothetical protein			95	
	Hypothetical Proteins (98 proteins)				
14	Hypothetical protein			212	
68	Hypothetical protein			91	1
404	Hypothetical protein			56	
512	Hypothetical protein	+		219	1
1002	Hypothetical protein	+		275	
1088	Hypothetical protein			150	1
1260	Hypothetical protein	+		108	2
1352	Hypothetical protein			132	1
1535	Hypothetical protein		+	152	1
1599	Hypothetical protein	+		95	1
1828	Hypothetical protein	+		210	
2088	Hypothetical protein	+		78	
2166	Hypothetical protein			128	
2436	Hypothetical protein		+	213	2
2448	Hypothetical protein	+		65	
2501	Hypothetical protein			133	2
2647	Hypothetical protein	+	+	113	2
2782	Hypothetical protein			79	
2822	Hypothetical protein	+	+	65	1
2830	Hypothetical protein	+	+	174	2
2901	Hypothetical protein	+		68	2
3060	Hypothetical protein	+		83	
3320	Hypothetical protein	+		259	
3346	Hypothetical protein	+		101	
3437	Hypothetical protein	+		88	
3479	Hypothetical protein			76	
3500	Hypothetical protein	+		71	
3504	Hypothetical protein			184	
3513	Hypothetical protein	+		102	
3577	Hypothetical protein	+		81	
3764	Hypothetical protein	+		67	
4041	Hypothetical protein			146	2
4182	Hypothetical protein	+	+	98	2
4256	Hypothetical protein	+		91	1
4464	Hypothetical protein	+		119	2
4703	Hypothetical protein	+		213	
4931	Hypothetical protein	+	+	76	
5031	Hypothetical protein	+		86	2
5240	Hypothetical protein	+	+	58	1
5268	Hypothetical protein			217	
5310	Hypothetical protein	+		202	
5734	Hypothetical protein			68	
5763	Hypothetical protein	+		143	
5794	Hypothetical protein			110	
5811	Hypothetical protein	+		183	2
5938	Hypothetical protein			95	
5945	Hypothetical protein			53	
6051	Hypothetical protein			90	
6127	Hypothetical protein	+		105	2
6162	Hypothetical protein		+	66	
6221	Hypothetical protein	+		161	2
6400	Hypothetical protein	+		103	1
6610	Hypothetical protein		+	370	
6658	Hypothetical protein	+	+	160	2
6713	Hypothetical protein			164	2
6771	Hypothetical protein	+		220	
6888	Hypothetical protein	+		130	2
6941	Hypothetical protein	+	+	52	
7203	Hypothetical protein		+	157	2
7235	Hypothetical protein	+	+	194	
7476	Hypothetical protein	+		95	
7801	Hypothetical protein			70	
8337	Hypothetical protein			144	
8750	Hypothetical protein	+		112	
8980	Hypothetical protein	+		106	
9034	Hypothetical protein	+		176	
9101	Hypothetical protein	+		108	2
9512	Hypothetical protein	+		82	1
9518	Hypothetical protein	+		65	
9527	Hypothetical protein	+	+	215	1
9742	Hypothetical protein			86	
9798	Hypothetical protein		+	110	
9867	Hypothetical protein			107	
9973	Hypothetical protein			118	
10075	Hypothetical protein			103	
10460	Hypothetical protein	+		89	1
10530	Hypothetical protein			128	
10823	Hypothetical protein	+		143	2
10835	Hypothetical protein			275	1
11008	Hypothetical protein	+		242	2
11575	Hypothetical protein	+		92	1
12035	Hypothetical protein			120	2
12167	Hypothetical protein		+	124	1
12219	Hypothetical protein			91	
12416	Hypothetical protein	+		162	1
12442	Hypothetical protein			165	
12461	Hypothetical protein	+	+	154	2
12489	Hypothetical protein	+		113	
12630	Hypothetical protein	+		90	2
12645	Hypothetical protein		+	114	
12702	Hypothetical protein	+		62	
12787	Hypothetical protein			168	2
12808	Hypothetical protein	+		95	
12844	Hypothetical protein			134	2
12897	Hypothetical protein	+		62	
12968	Hypothetical protein	+		119	1
13231	Hypothetical protein			105	
13319	Hypothetical protein		+	102	

a) The quality of the mass spectrometric protein identification results are characterized by their probability-based MOWSE scores (Score), and the number of gels, in which a protein was identified (*). The presence of a predicted signal peptide (sp) and the predicted level of gene expression (PHX) are provided. Listed are the proteins that received scores >51, corresponding to 95% confidence (for details, see text).

Table 3. Proteins (new ORF) specifically identified from the Maximal ORF Set (MOS)

ORF no.	Start	Stop	Length (aa)	MASCOT score	Predicted function	Genetic context
pir.6532c	1798993	1799565	290	61	Hypothetical	Methionine aminopeptidase, hypotheticals; other strand: ribose-regulated sugar-ADH
pir.8508	2358829	2359248	139	60	Hypothetical	Mostly hypotheticals
pir.9191c	2546400	2546921	173	101	Hypothetical	Mostly hypotheticals, downstream of possible adenylate cyclase
pir.15895	4437587	4438426	279	56	Hypothetical	Mostly hypotheticals; upstream of D-tyrosyl-tRNA(Tyr)-deacylase

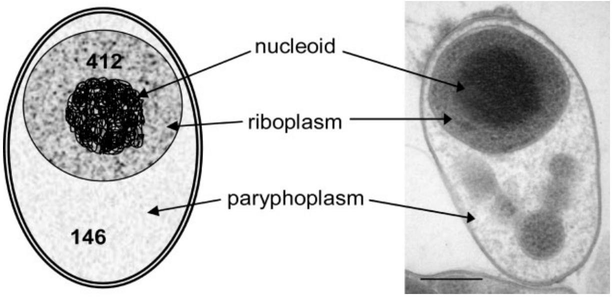

Figure 3. Intracellular compartmentalization of *R. baltica* and possible location of identified proteins (Bar = 0.2 μm).

According to the signal hypothesis [46], the majority of secreted proteins have a signal peptide, which is found in 1160 (16%) of the predicted proteins in *R. baltica*. Out of the 558 identified proteins annotated in the master gel 146 (26%) possess a signal peptide (Table 2). Since the applied methods for cell breakage did not separate riboplasmic from paryphoplasmic proteins one can conclude that the 146 signal peptide containing proteins have potentially been secreted and are actually localized in the paryphoplasm or are cell wall associated.

For 58% (57 proteins) of the hypothetical proteins a signal peptide was predicted. Thirty-six (about 38%) of the 94 conserved hypothetical proteins are secreted according to the signalP prediction. Secreted proteins with functional assignment are mainly dehydrogenases, hydrolases for extracellular macromolecules or involved in signal transduction (phosphoprotein kinases). In contrast, the enzymes performing housekeeping functions seem to be confined to the riboplasm (no signal peptide). Interestingly, nine of the 10 identified sulfatases have a signal peptide prediction. The *R. baltica* genome encodes 110 sulfatases, which are suggested to function in extracellular degradation of sulfated glycopolymers such as, *e.g.*, carrageen [26]. Thus, the identified sulfatases could be in the process of being excreted, since proteins already excreted to the extracellular space would have been lost under the applied conditions of cell harvesting. Expression of sulfatase encoding genes might not require the presence of sulfated substrates, since the studied *R. baltica* cells were grown with ribose as only source of organic carbon. In some cases the correctness of the signal peptide prediction is questionable, *e.g.*, for the elongation factor G and lysyl-tRNA synthase. Both enzymes play an important role in protein synthesis, a process that should exclusively take place in the riboplasm. Thus the presence of a signal peptide alone does not allow defining the exact target region of translocation. Future research on secreted proteins (secretome) will have to differentiate the different compartments present in *R. baltica* cells.

3.7 Predicted highly expressed (PHX) genes

Among the 30 most abundant proteins on the master gel of *R. baltica,* 27 were encoded by genes that were predicted to be highly expressed (PHX) according to codon usage adaptation. Thus, a correlation between experimentally determined protein abundance and codon usage features as it already has been shown for fast-growing bacteria [43] could also be observed for *R. baltica*, a slowly growing environmental bacterium (doubling times between 10–14 h, [32]). However, there are some exceptions where the genes of proteins appearing as highly abundant on 2-DE gels are not PHX; this applies mainly for proteins that were, *e.g.*, specifically induced during growth with ribose (see accompanying publication) or proteins affiliated with lipid metabolism.

4 Concluding remarks

With more than 550 identified gene products, the present study established a solid proteomic framework for further analysis of differential gene expression in *R. baltica*. Considering the nutritional specialization of this bacterium on the utilization of carbohydrates, we will be able to reconstruct the major catabolic routes which are operative in *R. baltica* and to learn about the potential of this bacterium to regulate the expression of catabolic genes in response to the availability of respective growth substrates (see accompanying publication). The master gel will also be beneficial for the identification of proteins involved in cell cycle and development. Such proteins should be related to the two morphotypes (single cells *versus* aggregates) as well as to different growth stages.

We thank Alfred Beck and Thomas Kreitler for bioinformatics support. We are indebted to Friedrich Widdel for continuous support of proteomic work at the MPI in Bremen. This study was supported by the Max Planck Society.

Addendum in proof

A recent proteomic study revealed growth phase dependent regulation of protein composition in *R. baltica* (Gade, D., Stührmann, T., Reinhardt, R., Rabus, R., *Environ. Microbiol.* 2005, *7*, 1074–1084).

5 References

[1] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A. *et al.*, *Science* 1995, *269*, 496–512.

[2] Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E. *et al.*, *Nucleic. Acids Res.* 1996, *24*, 4420–4449.

[3] Tonella, L., Walsh, B. J., Sanchez, J. C., Ou, K., *et al.*, *Electrophoresis* 1998, *19*, 1960–1971.

[4] Tonella, L., Hoogland, C., Binz, P.-A., Appel, R. D. *et al.*, *Proteomics* 2001, *1*, 409–423.

[5] Molloy, M. P., Herbert, B. R., Slade, M. B., Rabilloud, T. *et al.*, *Eur. J. Biochem.* 2000, *276*, 2871–2881.

[6] Ohlmeier, S., Scharf, C., Hecker, M., *Electrophoresis* 2000, *21*, 3701–3709.

[7] Büttner, K., Bernhardt, J., Scharf, C., Schmid, R. *et al.*, *Electrophoresis* 2001, *22*, 2908–2935.

[8] Regula, J. T., Ueberle, B., Boguth, G., Görg, A. *et al.*, *Electrophoresis* 2000, *21*, 3765–3780.

[9] Regula, J. T., Boguth, G., Görg, A., Mayer, F. *et al.*, *Microbiology* 2001, *147*, 1045–1057.

[10] Ueberle, B., Frank, R., Herrmann, R., *Proteomics* 2002, *2*, 754–764.

[11] Cordwell, S. J., Larsen, M. R., Cole, R. T., Walsh, B. J., *Microbiology*, 2002, *148*, 2765–2781.

[12] Hecker, M., Engelmann, S., Cordwell, S. L., *J. Chromatogr. B* 2003, *787*, 179–195.

[13] Fountoulakis, M., Takacs, B., Langen, H., *Electrophoresis* 1998, *19*, 761–766.

[14] Fountoulakis, M., Juranville, J. F., Roder, D., Evers, S. *et al.*, *Electrophoresis* 1998, *19*, 1819–1827.

[15] Langen, H., Takács, B., Evers, S., Berndt, P. *et al.*, *Electrophoresis* 2000, *21*, 411–429.

[16] Nouwens, A. S., Willcox, M. D., Walsh, B. J., Cordwell, S. J., *Proteomics* 2002, *2*, 1325–1346.

[17] Hermann, T., Pfefferle, W., Baumann, C., Busker, E. *et al.*, *Electrophoresis* 2001, *22*, 1712–1723.

[18] Hesketh, A. R., Chandra, G., Shaw, A. D., Rowland, J. J. *et al.*, *Mol. Microbiol.* 2002, *46*, 917–932.

[19] Görg, A., Obermaier, C., Boguth, G., Harder, A. *et al.*, *Electrophoresis* 2000, *21*, 1037–1053.

[20] Pappin, D. J. C., Højrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[21] Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C. *et al. Proc. Natl. Acad. Sci. USA* 1993, *90*, 5011–5015.

[22] Mann, M., Højrup, P., Roepstorff, P., *Biol. Mass Spectrom.* 1993, *22*, 338–345.

[23] Karas, M., Hillenkamp, F., *Anal. Chem.* 1988, *60*, 2301–2303.

[24] Kaneko, T., Sato, S., Kotani, H., Tanaka, A. *et al.*, *DNA Res.* 1996, *3*, 109–136.

[25] Nierman, W. C, Feldblyum, T. V., Laub, M. T., Paulsen, I. T. *et al.*, *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4136–4141.

[26] Glöckner, F. O., Kube, M., Bauer, M., Teeling, H. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 8298–8303.

[27] Vohradsky, J., Janda, I., Grünenfelder, B., Berndt, P. *et al.*, *Proteomics* 2003, *3*, 1874–1882.

[28] Huang, F., Parmryd, I., Nilsson, F., Persson, A. L. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 956–966.

[29] Schlesner, H., Bartels, C., Tindall, B. J., Gade, D. *et al.*, *Int. J. Syst. Evol. Microbiol.* 2004, *54*, 1567–1580.

[30] Lindsay, M., Webb, R. I., Fuerst, J. A., *Microbiology* 1997, *143*, 739–748.

[31] Lindsay, M., Webb, R. I., Strous, M., Jetten, M. S. M. *et al.*, *Arch. Microbiol.* 2001, *175*, 413–429.

[32] Rabus, R., Gade, D., Helbig, R., Bauer, M. *et al.*, *Proteomics* 2002, *2*, 649–655.

[33] Bradford, M. M., *Anal. Biochem.* 1976, *72*, 248–254.

[34] Gade, D., Thiermann, J., Markowsky, D., Rabus, R., *J. Mol. Microbiol. Biotechnol.* 2003, *5*, 240–251.

[35] Dorothy, N. S., Littmann, B. H., Reilly, K., Swindell, A. C., Buss, J. M., Anderson, N. L., *Electrophoresis* 1998, *19*, 355–363.

[36] Nordhoff, E., Egelhofer, V., Giavalisco, P., Eickhoff, H. *et al.*, *Electrophoresis* 2001, *22*, 2844–2855.

[37] Gobom, J., Schuerenberg, M., Mueller, M., Theiss, D. *et al.*, *Analyt. Chem.* 2001, *73*, 434–438.

[38] Gobom, J., Mueller, M., Egelhofer, V., Theiss, D. *et al.*, *Anal. Chem.* 2002, *74*, 3915–3923.

[39] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[40] Rice, P., Longden, I., Bleasby, A., *Trends Genet.* 2000, *16*, 276–277.

[41] Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., *Protein Eng.* 1997, *10*, 1–6.

[42] Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D. *et al.*, *Nucleic Acids Res.* 2003, *31*, 2187–2195.

[43] Karlin, S., Mrázek, J., Campell, A., Kaiser, D., *J. Bacteriol.* 2001, *183*, 5025–5040.

[44] VanBogelen, R. A., Schiller, E. E., Thomas, J. D., Neidhardt, F. C., *Electrophoresis* 1999, *20*, 2149–2159.

[45] Schwartz, R., Ting, C. S., King, J., *Genome Res.* 2001, *11*, 703–709.

[46] Blobel, G., *Chembiochemistry* 2000, *1*, 86–102.