Proteome Analysis of *Halobacterium* sp. NRC-1 Facilitated by the Biomodule Analysis Tool BMSorter*§§

Rueichi R. Gan‡, Eugene C. Yi§¶, Yuling Chiu‡, Hookeun Lee‖, Yu-chieh P. Kao***, Timothy H. Wu**, Ruedi Aebersold§, and Wailap Victor Ng**§§¶¶

To better understand the extremely halophilic archaeon *Halobacterium* species NRC-1, we analyzed its soluble proteome by two-dimensional liquid chromatography coupled to electrospray ionization tandem mass spectrometry. A total of 888 unique proteins were identified with a ProteinProphet probability (P) between 0.9 and 1.0. To evaluate the biochemical activities of the organism, the proteomic data were subjected to a biological network analysis using our BMSorter software. This allowed us to examine the proteins expressed in different biomodules and study the interactions between pertinent biomodules. Interestingly an integrated analysis of the enzymes in the amino acid metabolism and citrate cycle networks suggested that up to eight amino acids may be converted to oxaloacetate, fumarate, or oxoglutarate in the citrate cycle for energy production. In addition, glutamate and aspartate may be interconverted from other amino acids or synthesized from citrate cycle intermediates to meet the high demand for the acidic amino acids that are required to build the highly acidic proteome of the organism. Thus this study demonstrated that proteome analysis can provide useful information and help systems analyses of organisms. *Molecular & Cellular Proteomics* 5:987–997, 2006.

*Halobacterium* species NRC-1 is an extremely halophilic archaeon containing a highly acidic proteome with a median pI of 4.9, a property that is essential to the maintenance of the solubility and function of the proteins in a high salinity environment of about 5 M salts (1, 2). Genome sequence analysis has revealed 2,630 putative protein-coding genes in the 2,571,010-bp genome (2). Among the predicted proteins, 1,658 can be matched to sequences in public databases. Of the matches, 1,067 are proteins of known or predicted function, and 591 are proteins of unknown function. The possession of a relatively small and completely sequenced genome, the availability of a full arsenal of genetic manipulation tools, and the relative ease of culture make *Halobacterium* sp. NRC-1 an attractive systems biology model organism of the domain Archaea (3–6).

The genomes of *Halobacterium* species are extremely unstable (7–9). Early studies of *Halobacterium* sp. NRC-1 (also known as *Halobacterium halobium*) and the closely related *Halobacterium salinarium* discovered unusually high spontaneous mutation frequencies of 0.01% for the production of bacteriorhodopsin- or bacterioruberin-deficient phenotypes and a more striking 1% for partial or total gas vesicle-deficient phenotypes. The species is also noteworthy for the large number of insertion sequence elements that are harbored in this unstable genome. Molecular genetic analysis of the bacteriorhodopsin- and gas vesicle-deficient mutants established a relationship between transposable insertion sequence-mediated insertional inactivation or deletions of structural or regulatory genes and the high mutant rates (10–16). Upon the completion of the genome sequence, DNA analysis revealed the presence of 91 copies of insertion sequence elements belonging to 12 families in the *Halobacterium* sp. NRC-1 genome (2).

Mass spectrometry is a powerful technology for protein identification in the postgenomic era. Our previous shotgun peptide sequencing analysis of the *Halobacterium* sp. NRC-1 membrane and soluble proteomes identified a total of 426 unique proteins representing approximately one-fifth of the predicted proteome (2, 17). Among these, 232 were identified predominantly in the soluble fraction, 165 were in the membrane fraction, and 29 were in both fractions. Metabolic reconstruction found 103 of the identified proteins could be matched to enzymes in 52 metabolic pathways found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) data-

---

1 The abbreviations used are: KEGG, Kyoto Encyclopedia of Genes and Genomes; p, PeptideProphet probability; P, ProteinProphet probability; CGI, common gateway interface; SIF, simple interaction format; TBP, TATA box-binding protein; TFB, transcription factor B; aa, amino acids.


**MATERIALS AND METHODS**

**Protein Preparation**—One liter of *Halobacterium* sp. NRC-1 (ATCC 700922) cells was grown in *Halobacterium* medium to early log phase (\(A_{600} = 1\)) at 37 °C (17, 18) and harvested by centrifugation at 5,500 \(\times g\) for 10 min at 4 °C. The cell pellet was resuspended in a total of 25 ml of basal salt solution containing 1 mM PMSF and 0.5 mg each of DNase I and RNase A. The cells were transferred to a dialysis tube (Spectrapor\(^{2}\); membrane molecular weight cutoff: 3,500; Spectrum Laboratories, Inc.) and lysed by osmotic shock by dialysis against four changes of 4 liters of deionized water at 4 °C over a total of 2 days. Cell debris were removed by centrifugation at 10,000 \(\times g\) for 30 min at 4 °C. Membrane and insoluble materials were sedimented by ultracentrifugation at 53,000 \(\times g\) for 16 h at 4 °C. The supernatant containing the soluble proteins was aliquoted into 1.5-ml microcentrifuge tubes and stored at −20 °C.

**Fractionation of Peptides**—An aliquot of 3 mg of soluble proteins was digested with 60 \(\mu\)g of sequencing grade modified trypsin (Promega, Madison, WI) in a total volume of 3 ml of a solution containing 50 mM ammonium bicarbonate (pH 8.3) at 37 °C overnight. The resulting peptides were fractionated by strong cation exchange chromatography as described previously (19). The fractionated peptides were desalted using 96-well format spin columns containing silica C\(_{18}\) matrix (Nestec Group, Southborough, MA) according to the following procedure. First the matrix was washed twice by filling each well with 200 \(\mu\)l of 0.4% acetic acid and centrifuged at 770 \(\times g\) for 2 min. After loading the peptide samples (−200 \(\mu\)l/well), the plate was incubated at room temperature for 30 min and then centrifuged as above to remove the buffer. The reversed phase C\(_{18}\)-bound peptides were washed three times with 200 \(\mu\)l of 0.4% acetic acid and eluted with 200 \(\mu\)l of a solution containing 3 volumes of acetonitrile and 1 volume of 0.4% acetic acid. The eluents were vacuum-dried, and each peptide samples was resuspended in 10 \(\mu\)l of 0.4% acetic acid before analysis by LC-MS/MS.

**Mass Spectrometric Analysis of Peptides**—The desalted tryptic peptides were analyzed using a LCQ-DECA ion trap tandem mass spectrometer (Thermo Finnegan, San Jose, CA) coupled with a C\(_{18}\) trap ESI-emitter/micro-liquid chromatography column as described previously (20).

**Computation Analysis of Mass Spectra**—Tandem mass spectra were analyzed using the SEQUEST (21) algorithm to search against the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) *Halobacterium* sp. NRC-1 proteome database (2). The searches were performed with a peptide mass tolerance of 3 daltons without enzyme specification. The other parameters were left as default. The SEQUEST outputs were consolidated to a single hypertext file and further analyzed using the proteomic data analysis pipeline (Institute for Systems Biology; www.systemsbiology.org). As part of this pipeline, the programs PeptideProphet and ProteinProphet (22, 23) assign probability values to each peptide and protein identification, respectively, that indicate the likelihood that the respective analyte has been identified correctly.

**Biomedical Analysis of Proteomic Data**—BMSorter is a bioinformatic tool developed to facilitate systems analysis of proteins detected by tandem mass spectrometry. It is a set of common gateway interface (CGI) scripts written in Perl language to provide an interface to access the results of analyses via web browser. The scripts require an environment with an Apache web server and Perl module: Bioperl1.4, CGI-3.04, and GD-2.16. BMSorter constructed a list of expressed proteins with interactions between the proteins. It acquired the protein list with identification probabilities from the ProteinProphet output file as well as the pathway information in the form of a template of interactions from the KEGG database. Cross-reference between pathway maps from KEGG PATHWAYS database and enzyme/gene information from KEGG LIGAND database implemented the integration between proteins and interactions. The tool set organized the proteomic data and KEGG pathway information in a summary table. Each row of the table is the summary of a pathway (or biomodule) containing the total number of reactions in the reference pathway, number of *Halobacterium* sp. NRC-1-specific reactions, percentage of NRC-1 reactions, number of identified proteins with ProteinProphet probabilities \((P) \geq 0.9\), and percentage of identified proteins plus the hyperlinks to the pathway protein list tables and pathway maps with the ProteinProphet identification probabilities indicated.

**Plotting the Metabolic Network**—The publicly available Cytoscape tool (www.cytoscape.org) was used to display the protein identification information on amino acid metabolisms and the citrate cycle pathways in terms of the enzyme-metabolite interaction networks (24). The Java programs HalKGMLtoSIF.java and HalnoaFactory.java were developed to process the KEGG pathways KEGG markup language (KGML) files and proteomic data, respectively. The HalKGMLtoSIF.java program was used to parse the extensible markup language (xml) structure of the KGML files to generate the Cytoscape simple interaction format (SIF) and node attribute files. The HalnoaFactory.java was used to define the node properties (protein names and ProteinProphet probability values) in node attribute files.

**RESULTS**

**Peptide Fractionation and Mass Spectrometric Analysis**—Three milligrams of *Halobacterium* sp. NRC-1 trypsin-digested soluble protein were separated into 100 fractions by strong cation exchange HPLC. The peptides contained in most fractions were desalted in a 96-well format C\(_{18}\) plate. The strong cation exchange fractions 11–85 were either analyzed as single fractions or pools of up to three consecutive fractions by LC-MS/MS. A total of 46 samples were thus analyzed of which 16 were analyzed twice, resulting in 62 LC-MS/MS datasets.

**Protein Identification**—The SEQUEST output data were analyzed with the PeptideProphet and ProteinProphet software to evaluate and assign, respectively, the peptide and protein identification probabilities. A total of 888 proteins were identified with \(P\) between 0.9 and 1.0. Among these were 562 proteins with annotated functions, 189 conserved hypothetical proteins, and 137 hypothetical proteins of unknown functions (Supplemental Tables I and II). The distributions of probabilities and functional categories of identified proteins are summarized in Fig. 1. A total of 681 and 147 proteins were

---

2 The software BMSorter, HalKGMLtoSIF.java, and HalnoaFactory.java are available upon request to wvng@ym.edu.tw.
identified with a probability of 1.0 and 0.99, respectively. The estimated number of false positive identifications of $N$ proteins with a probability value of $P$ is equal to $N \times (1.0 - P)$. The total number of false positive protein identifications was ~5.

**Systematic Analysis of Identified Proteins**—The identified proteins were assigned to appropriate biomodules by the BMSorter tool (Fig. 2). From the proteomic data and the information on enzymes, metabolites, and chemical reactions available from the KEGG database, BMSorter generated a table containing the number of reactions in the reference pathways and their corresponding *Halobacterium sp. NRC-1* pathways, the number of predicted and mass spectrometric identified proteins in each NRC-1 pathway, the percentages of NRC-1 reactions, and the percentages of identified NRC-1 proteins together with hyperlinks to the pathway maps and lists of enzymes in each biomodule (Table I). A total of 297 proteins ($P > 0.9$) were matched to 76 biomodules in the KEGG database. The proteins identified in representative biomodules or functional categories are discussed in detail below.

**DNA Replication**—Genome analysis has revealed five genes encoding three DNA polymerase types in *Halobacterium* sp. NRC-1 including two family B polymerases (*polB1* on the chromosome and *polB2* on the minichromosome pNRC200), a bacteriophage-like family A polymerase (*polC*), and the heterodimeric family D polymerase (*polA1* and *polA2*) (2). Among these only PolA2 and PolB1 were detected in this study. In addition, DNA polymerase sliding clamp (Pcn) protein, clamp loader (RfcA, RfcB, and RfcC subunits), DNA topoisomerase VI subunits A (Top6A) and B (Top6B), DNA topoisomerase I (TopA), DNA gyrase subunits A (GyrA) and B (GyrB), DNA helicase (Hel), and replication protein A (Rpa) involved in single strand DNA binding were also identified (Supplemental Table I).

**Transcription**—The transcription of *Halobacterium* sp.
NRC-1 genes is driven by a eukaryotic RNA polymerase II-like system. It consists of 12 subunits (A, B’, B”, C, D, E’, E”, H, K, L, M, and N) encoded by genes located at six loci. Except for the H and K subunits, all the subunits were identified by tandem mass spectrometry with \( P > 0.9 \).

Transcription initiation in Archaea requires the binding of the TATA box-binding protein (TBP) and transcription factor B (TFB) to the promoter region (25–27). Genome analysis identified the presence of six tbp and seven tfp genes (2, 28, 29). Only two TATA-box binding proteins and one Tfp protein, namely the chromosomal TbpE and TfbG and the minichromosomal TbpB, were identified by mass spectrometric analysis (Supplemental Table I). Two peptides of TbpE (\( P = 1 \)) were found for a total of 14 times. The TbpB protein (\( P = 1 \)), encoded by both pNRC100 and pNRC200, had a peptide identified once. The only detectable Tfb, TfbG (\( P = 1 \)), had one peptide identified 10 times. The transcription initiation factor IIE \( \alpha \) subunit (TieA) had three peptides identified a total of 17 times. In addition, one peptide each from the termination-antitermination factors NusA and NusG was also identified.

Among the 27 transcription regulators predicted from the genome sequence (2), 12 were identified in this study (Supplemental Table I). These included the putative regulator ArcR expressed from a gene in the arcRACB gene cluster on pNRC200 that also codes the enzymes for arginine deiminase (ArcA), carbamate kinase (ArcC), and catabolic ornithine transcarbamylase (ArcB), all of which are required for fermentative growth using the arginine deiminase pathway (30). The phosphate transport system regulatory protein PhoU and phosphate regulatory protein homolog Prp1 encoded by genes in the phosphate transporter phoU-pstB2A2C2-phoX gene cluster and the juxtaposed downstream gene \( prp1 \) were identified. The putative transcription regulators ArsR, CinR, SirR, Trh1, Trh4, Trh5, and Hox-like transcription regulators (Hlx1 and Hlx2) were also detected by mass spectrometry.

**Translation** — A total of 117 proteins have been predicted to be associated with the translation in *Halobacterium* sp. NRC-1 (2). Among these 92 (78.6%) were identified in this study (Supplemental Table I). These included the ribosomal structural proteins and proteins involved in aminoacyl-tRNA synthesis, translation initiation, elongation, and release of synthesized polypeptides.

Except for the asparaginyl- and glutaminyl-tRNA synthetases, all other synthetase-coding genes have been found in the *Halobacterium* sp. NRC-1 genome, and these included the duplicated tryptophanyl-tRNA synthetase genes (\( trpS1 \) and \( trpS2 \)) (2). Proteome analysis indicated the expression of almost all synthetases with the exception of the TrpS2 protein. In addition, the identification of amidotransferase subunits GatA, GatB1, GatB2, and GatC suggested that *Halobacterium* sp. NRC-1 could synthesize all necessary aminoacyl-tRNAs for protein synthesis.

All of the nine eukaryotic-like translation initiation factors eIF-1A (Eif1a1 and Eif1a2); eIF-2 subunits \( \alpha \) (Eif2a), \( \beta \) (Eif2b), and \( \gamma \) (Eif2g); eIF-2B subunits \( \alpha \) (Eif2ba) and \( \delta \) (Eif2bd); ATP-dependent RNA helicase homologs eIF-4A (Eif4a) and eIF-5A (Eif5a); and bacterial-like initiation factor IF-2 (InfB) were identified. Also identified were translation elongation factors eEF1 subunits \( \alpha \) (Eef1a1) and \( \beta \) (Eef1b) and eEF-2 (Eef2) and peptide chain release factor eRF1 (Erf1).

The *Halobacterium* sp. NRC-1 genome encodes 32 large and 25 small ribosomal subunit protein-coding genes (2). Proteome analysis indicated the expression of 27 (84.4%) large subunit and 23 (92%) small subunit proteins. Ribosomal subunit proteins that were not identified were Rpl10e, Rpl24e, Rpl39e, Rpl40e, Rpl44e, Rps14p, and Rps27e.

**Energy and Amino Acid Metabolism** — Proteome analysis indicated the expression of almost a complete set of all predicted citric acid cycle enzymes and 48 of the 89 amino acid metabolic enzymes (Supplemental Table I). These included arginine deiminase (ArcA), ornithine transcarbamylase (ArcB), and carbamylase kinase (ArcC) presumably involved in the fermentative arginine degradation via the arginine deiminase pathway (30). There are 16 amino acid metabolic/degradation pathways listed in the KEGG database. Eleven pathways had three to nine proteins identified, whereas the alanine-aspartate metabolism and valine-leucine-isoleucine degradation pathways had 13 proteins, and the arginine-proline, glutamate, and glycerine-serine-threonine metabolic pathways had 16 proteins identified by mass spectrometry analysis (Table I).

The amino acids that may be used in energy production were evaluated by systems analysis of the proteomic data on the amino acid metabolism pathways and citrate cycle (tricarboxylic acid cycle) networks (Fig. 3). The results suggest that up to eight amino acids may be involved in energy production. The amino acids aspartate, asparagine, proline, and arginine may enter the tricarboxylic acid cycle via oxaloacetate, fumarate, or oxoglutarate. Glutamate and glutamine, and glycine and serine may enter the cycle via oxoglutarate and oxaloac-

---

**Fig. 2. Flow chart of BMSorter analysis of proteomic data.** The protein names, putative functions, identification probabilities, and gene numbers from the ProteinProphet output file (bottom right) and the protein names, reactions, EC numbers, and gene numbers from the KGML file were parsed out using Perl scripts. The unique gene numbers were used to match the proteomic data to KEGG information and generate the summary table (please also see Table I). “ProtList” is a hyperlink to a protein table containing a list of the proteins in each pathway, and “PathwayMap” is a hyperlink to the modified KEGG pathway map containing the abbreviated gene names in large rectangles and identification probabilities in small rectangles on a gray to red color scale representing the identification probabilities between 0 and 1. Large rectangles with yellow color background indicate that more than one protein catalyzes the reactions, green color background represents NRC-1 proteins without MS identification, and white color background represents proteins absent in NRC-1. The identified proteins on protein tables and pathway maps are hyperlinked to the ProteinProphet output file for reexamination of MS data on a web browser.
etate, respectively. Arginine may be used to produce ATP by the arginine deiminase pathway. Although a complete set of histidine utilization enzyme-coding genes (hutG, -H, -I, and -U) is present in the genome, none of the corresponding proteins were identified suggesting that histidine was not used to produce energy via the citrate cycle. The absence of detectable tryptophan biosynthesis enzymes TrpA, -B, -C, -D1, -E1, -F, and -G1 suggested that tryptophan was not being synthesized when the cells were harvested for proteome analysis. In addition, the interaction network suggested that the synthesis of the acidic amino acids aspartate and glutamate may occur to meet the demand for acidic amino acids that are needed as building blocks for the highly acidic proteome.

**Nucleic Acid Metabolism**—Analysis of the *Halobacterium* sp. NRC-1 nucleotide metabolic pathways in the KEGG database found 45 proteins that may be involved in catalyzing 38 reactions associated with purine metabolism and 46 proteins that may be involved in 36 reactions associated with pyridine metabolism (Table I and Supplemental Tables III and IV). Thirty-five proteins of 22 purine metabolic enzymes and 31 proteins of 16 pyrimidine metabolic enzymes were detected by mass spectrometry ($P < 0.01$) (Fig. 4). These included 15 proteins from four enzymes (DNA polymerase, DNA-directed RNA polymerase, ribonucleoside-diphosphate reductase, and nucleoside-diphosphate kinase) that are present in both pathways.

**DISCUSSION**

The application of a two-dimensional separation approach greatly increased the total number of protein identified in this

---

**TABLE I**

Summary of the percentage of mass spectrometry-identified *Halobacterium* sp. NRC-1 proteins, percentage of NRC-1-specific reactions, number of MS-identified proteins, numbers of NRC-1 predicted proteins and reactions, and number of reactions in the reference map in each KEGG pathway/protein complex/cellular function.

Biomodules with less than 20% NRC-1-specific reactions are shaded. The column containing the hyperlinks to the protein table and pathway map of each module is not shown in this table (see Fig. 2). NA, not available.
study. In a total of 62 mass spectrometric analyses of the strong cation exchange fractions of the *Halobacterium* sp. NRC-1 soluble proteome, 888 proteins were identified with \( P \geq 0.9 \). This can be compared with our previous study (17) that identified 255 proteins (\( P > 0.9 \)) in the soluble proteome by only reversed phase C\(_{18}\) fractionation and gas phase fractionation of precursor ions at multiple m/z windows (20) over a total of 53 MS analyses; our new approach gave an \( \sim 3.5 \) fold increase in protein identification in this analysis.

Proteome analysis revealed the expression of more than one-third of the genes predicted from the *Halobacterium* sp. NRC-1 genome, which consists of 2,413 unique protein-coding genes. Of the 888 proteins identified at high confidence level (\( P > 0.9 \)), 562 were expressed from genes with annotated functions, and 326 were expressed from genes of unknown function. Among the predicted proteins with defined functions, a significant fraction of those detected by mass spectrometry were involved in vital life processes, namely 48 (53.9%) of 89 amino acid metabolism; 38 (66.7%) of 57 nucleotide metabolism; 31 (56.4%) of 55 cofactor metabolism; 28 (25.0%) of 57 carbohydrate metabolism; and 27 (30.0%) of 91 lipid metabolism.

### Table I—continued

| Table I: *Halobacterium* sp. NRC-1 Proteome Analysis |

| Arginine and proline metabolism | 83.8% | 16.2% | 16 | 18 | 14 | 86 |
| Cysteine metabolism | 60.0% | 38.4% | 6 | 10 | 10 | 26 |
| Glutamate metabolism | 80.0% | 38.8% | 16 | 20 | 14 | 36 |
| Glycine, serine and threonine metabolism | 69.5% | 31.7% | 16 | 23 | 20 | 63 |
| Histidine metabolism | 52.9% | 36.8% | 9 | 17 | 14 | 38 |
| Lysine biosynthesis | 100% | 16.2% | 6 | 6 | 6 | 37 |
| Lysine degradation | 87.5% | 10.0% | 7 | 8 | 5 | 50 |
| Methionine metabolism | 71.4% | 37.0% | 5 | 7 | 10 | 27 |
| Phenylalanine metabolism | 60.0% | 2.0% | 3 | 5 | 1 | 50 |
| Phenylalanine and tryptophan biosynthesis | 38.0% | 48.5% | 8 | 21 | 17 | 35 |
| Tryptophan metabolism | 75.0% | 11.2% | 9 | 12 | 9 | 80 |
| Tyrosine metabolism | 63.6% | 30.3% | 7 | 11 | 11 | 106 |
| Valine, leucine and isoleucine biosynthesis | 75.0% | 30.4% | 6 | 8 | 7 | 23 |
| Valine, leucine and isoleucine degradation | 81.2% | 30.2% | 13 | 16 | 13 | 43 |
| Urea cycle & metabolism of amino groups | 100% | 11.4% | 6 | 6 | 4 | 35 |
| Metabolism of Cofactors and Vitamins |
| Thiamine metabolism | 66.6% | 20.0% | 2 | 3 | 3 | 13 |
| Riboflavin metabolism | 75.0% | 28.5% | 3 | 4 | 4 | 14 |
| Nicotinate and nicotinamide metabolism | 33.3% | 20.0% | 2 | 6 | 7 | 35 |
| Pantothenate and CoA biosynthesis | 75.0% | 10.7% | 3 | 4 | 3 | 28 |
| Folate biosynthesis | 66.6% | 27.7% | 4 | 6 | 10 | 36 |
| One carbon pool by folate | 71.4% | 26.6% | 5 | 7 | 8 | 30 |
| Porphyrin and chlorophyll metabolism | 63.6% | 26.4% | 14 | 22 | 23 | 87 |
| Ubiquinone biosynthesis | 25.0% | 29.6% | 5 | 20 | 8 | 27 |
| Energy Metabolism |
| Oxidative phosphorylation | 44.1% | 85.7% | 19 | 43 | 6 | 7 |
| Carbon fixation | 70.0% | 27.5% | 7 | 10 | 8 | 29 |
| Lipid Metabolism |
| Fatty acid biosynthesis (path 1) | 100% | 17.5% | 2 | 2 | 7 | 40 |
| Fatty acid biosynthesis (path 2) | 85.7% | 51.5% | 6 | 7 | 17 | 33 |
| Fatty acid metabolism | 78.5% | 53.1% | 11 | 14 | 25 | 47 |
| Biosynthesis of steroids | 60.0% | 11.2% | 6 | 10 | 8 | 71 |
| Bile acid biosynthesis | 92.3% | 17.3% | 12 | 13 | 8 | 46 |
| Glycolipid metabolism | 56.2% | 12.3% | 9 | 16 | 12 | 97 |
| Metabolism of Other Amino Acids |
| Beta-Alanine metabolism | 71.4% | 11.7% | 5 | 7 | 4 | 34 |
| Taurine and hypotaurine metabolism | 0.0% | 14.2% | 0 | 1 | 2 | 14 |
| Aminophosphonate metabolism | 50.0% | 13.3% | 1 | 2 | 2 | 15 |
| Selenoamino acid metabolism | 87.5% | 29.1% | 7 | 8 | 7 | 24 |
| Cyanamido acid metabolism | 80.0% | 13.6% | 4 | 5 | 3 | 22 |
| Biosynthesis of 2° Metabolites |
| Terpenoid biosynthesis | 50.0% | 37.5% | 4 | 11 | 6 | 34 |
| Limonene and pinene degradation | 85.7% | 7.2% | 6 | 7 | 4 | 55 |
| Stilbene, coumarine and lignin biosynthesis | 0.0% | 5.0% | 0 | 2 | 1 | 20 |
| Alkaloid biosynthesis | 50.0% | 4.3% | 2 | 4 | 2 | 40 |
| Streptomyccin biosynthesis | 28.5% | 10.5% | 2 | 7 | 2 | 19 |
| Novobiocin biosynthesis | 60.0% | 2.8% | 3 | 5 | 1 | 35 |
| Biodegradation of Xenobiotics |
| γ-Hexachlorocyclo-hexene degradation | 100% | 4.3% | 2 | 2 | 1 | 23 |
| Fluorene degradation | 0.0% | 7.1% | 0 | 2 | 2 | 28 |
| Benzene degradation via CO ligation | 72.7% | 17.7% | 8 | 11 | 2 | 45 |
| Benzene degradation via hydroxylation | 71.4% | 8.7% | 5 | 7 | 5 | 35 |
| Glycan Biosynthesis and Metabolism |
| Peptidoglycan biosynthesis | 100% | 11.1% | 1 | 1 | 3 | 27 |
Halobacterium sp. NRC-1 Proteome Analysis

74 (66.7%) of 111 energy metabolism; 34 (57.6%) of 59 cell envelop components; 35 (29.2%) of 120 transport; 71 (52.2%) of 136 cellular processes; 28 (42.4%) of 66 DNA replication, repair, and recombination; 18 (51.4%) of 35 transcription; 12 (42.9%) of 28 regulation; and 92 (78.6%) of 117 translation-related proteins or enzymes were identified.

Halobacterium sp. NRC-1 is an aerobic chemoorganotroph usually cultured in a complex medium in the laboratory (3, 18). A minimal medium described for Halobacterium species contains all but five (aspartate, asparagine, glutamine, histidine, and tryptophan) of the 20 amino acids (31). Genome analysis suggested that several amino acids, including arginine and aspartate, may enter the citric acid cycle via the metabolic intermediates 2-oxoglutarate and oxaloacetate, respectively.

Figure 3. Interaction network of the enzymes, amino acids, and intermediates in 17 amino acid metabolic pathways and citric acid cycle. Large red triangles indicate the five known non-essential amino acids aspartate, asparagine, glutamine, histidine, and tryptophan, and large blue triangles represent the essential amino acids. Circles and ellipses indicate the enzymes and pathway names (metabolism (META), degradation (DEGR), and biosynthesis (BIOSYN), respectively. The colors of the circumference of the circles indicate ProteinProphet probability (black, P = 1; red to dark red, 0.9 ≤ P ≤ 0.99; and red to green, 0.9 ≥ P > 0), and the color scale is shown in the lower right corner. Enzymes involved in more than one pathway are shown in yellow color. Small triangles indicate the metabolic intermediates. Thick lines and arrows indicate the reactions related to energy production, intermediate lines and arrows represent the other enzymatic reactions, and thin lines link the enzymes to corresponding pathways. All pathways, except for the arginine deiminase pathway, were derived from the KEGG database. The network was displayed using the Cytoscape software. TCA, tricarboxylic acid.
The purine (A) and pyrimidine (B) metabolic pathway maps were generated by the BMSorter, which incorporated the proteomic data to the KEGG pathway maps. The small circles indicate the metabolic intermediates, and the large rectangles indicate the enzymes with abbreviated gene names or enzyme catalog numbers. The rectangles are color-coded: white rectangles, no protein predicted in Halobacterium sp. NRC-1; green rectangles, proteins are predicted in NRC-1 but were not identified by MS; yellow rectangles, contain more than one predicted protein and at least one was identified by MS; gray to red rectangles, contain one MS-identified protein with a ProteinProphet probability as shown in the color scale. The identification probabilities are also indicated in the small rectangles.
as a source of energy (2, 30). An integrative analysis of the mass spectrometry-identified proteins in the amino acid metabolisms and citrate cycle network suggested that up to eight amino acids, arginine, aspartate, asparagines, glutamate, glutamine, glycine, proline, and serine, but not histidine, may be involved in energy production via the citrate cycle under laboratory culture conditions. Furthermore it also suggested that aspartate and glutamate can be synthesized to meet the high demand for acidic amino acids during protein synthesis. Although network analysis suggested the presence of a complete tryptophan synthesis pathway (Fig. 3), the lack of evidence of expression of seven *trp* genes, *aroC*, *psc*, *VNG1244C*, and *VNG1245C* suggested that tryptophan was not synthesized under our culture conditions.

*Halobacterium* sp. NRC-1 lives in harsh environments, which may expose the cells to extended UV irradiation, periodic changes of temperature, fluctuation of nutrient supplies, etc. Proteome analysis revealed the basal level expression of the bacterial-type excision repair system proteins UvrA, UvrB, and UvrD; eukaryote-type excision-repair proteins Rad2, Rad3b, and RadA1; and the photolyase Phr1; these should allow the cells to repair UV damage and tolerate a high level of sunlight exposure (2, 32, 33). The expression of thermosome subunits A (CctA) and B (CctB) and heat shock proteins Hsp1, Hsp5, DnaJ, DnaK, GrpE, and Lon were supported by the mass spectrometry analysis results providing support for a lifestyle in an extreme environment. In addition, proteome analysis also demonstrated the presence of chemotaxis proteins CheA, CheB, CheC1, CheC2, CheR, CheW1, CheW2, and CheY and 11 signal transducers, Htr1 to Htr6, Htr10, and Htr12 to Htr15. The expression of these proteins suggests that the cells are equipped with the essential biomolecules for prompt adaptation to the dynamic conditions found in such extreme environmental conditions (Supplemental Table I).

An examination of this and three other proteome analysis results suggests that at least 40% of predicted proteins are expressed in *Halobacterium* sp. NRC-1 (2, 17, 34, 35). A total of 972 (*P* > 0.9) unique proteins have been identified in this and our previous (17) proteome analyses. Using a Protein-Prophecy probability of 0.9 as the cutoff, most of the proteins (89 of 162 membrane, 218 of 226 soluble, and 26 of 29 membrane/soluble fraction proteins) found in our previous analysis were also identified in this study. The proteome analysis of a closely related strain, *H. salinarium*, resulted in the identification of 802 proteins in the soluble proteome (34) and 141 proteins in purified membrane (35). The *H. salinarium* proteome analyses detected ~680 unique proteins classified as reliable or trusted identifications. Among these, 459 proteins were found in both strains at a high confidence level suggesting that more proteins may be expressed in *Halobacterium* sp. NRC-1 under standard laboratory culture conditions.

In addition to the expression level, at least two other factors, the size and the gene copy numbers, may affect protein identification in proteome analysis. Generally speaking, the number of tryptic peptides generated from each protein is proportional to its size. The larger the protein, the higher the number of tryptic peptides, and thus the higher the probability the protein will be identified by mass spectrometry. Maybe for this reason the relatively small housekeeping large ribosome subunit proteins L24E (62 aa), L38E (50 aa), L40E (47 aa), and L44E (61 aa); small ribosome subunit proteins S14P (52 aa) and S27E (57 aa); and DNA-directed RNA polymerase sub-units N (66 aa) and K (60 aa) proteins were not identified by mass spectrometry. On the other hand, it may not be necessary to express more than one copy of the duplicated genes to carry out a certain biochemical function. This may be why only one of the histidinol-phosphate transaminase and tryptophanyl-tRNA synthetase gene products, HisC1 and TrpS1, respectively, but not HisC2 and TrpS2, were detected by mass spectrometry.

Similar to many other organisms with completed genome sequences the functions of a significant fraction of predicted genes have yet to be identified. In this study, we documented the identification of 888 proteins among which 326 were translated from genes of no defined function. Tandem mass spectrometry is a very useful tool for detecting expressed proteins. The reliability of most identified proteins can be confirmed through examination of mass spectra. In our proteome studies, a suggested list of expressed proteins that contains more than 40% of the predicted proteome was established to facilitate the design of future experiments such as the systematic functional analysis of all predicted genes in the genome.

* This work was supported in part by Grant NSC9323314-B010014 from the National Science Council and an Aim for the Top University grant from the Ministry of Education in Taiwan (to W. V. N.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

To whom correspondence should be addressed: Inst. of Biotechnology in Medicine, National Yang Ming University, 155 Li Nong St., Section 2, Taipei, Taiwan 112, Republic of China. Tel.: 886-2-2826-7321; Fax: 886-2-2826-4092; E-mail: wvng@ym.edu.tw.

REFERENCES

1. Kennedy, S. P., Ng W. V., Salzberg, S. L., Hood, L., and DasSarma, S. (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.*, **11**, 1641–1650.

2. Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Strobaugh, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Wett, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., Danson, M. J., Hough, D. W., Maddocks, D. G., Jablonski, P. E., Krebs, M. P., Angervine, C. M., Dale, H., Isenbarger, T. A., Peck, R. F., Pohlschroeder, M., Spudich, J. L., Jung, K. W., Alam, M., Freitas, T., Hou, S., Daniels, C. J., Dennis, P. P., Omer, A. D., Ehrhardt, H., Lowe, T. M., Liang, P., Riley, M., Hood, L., and DasSarma, S. (2000) Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 12176–12181.

3. DasSarma, S., Robb, F. T., Place, A. R., Sowers, K. R., and Schreier, H. J.
(1995) Archaea: a Laboratory Manual—Halophiles, pp. 197–208, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
4. Peck, R. F., DasSarma, S., and Krebs, M. P. (2000) Homologous gene knockouts in the archaeon Halobacterium salinarum with ura3 as a counterselectable marker. Mol. Microbiol. 35, 667–676
5. Cline, S. W., Lam, W. L., Charlebois, R. L., Schalkwyk, L. C., and Doolittle, W. F. (1989) Transformation methods for halophilic archaeabacteria. Can. J. Microbiol. 35, 148–152
6. Cline, S. W., and Doolittle, W. F. (1987) Efficient transfection of the archaeabacterium Halobacterium halobium. J. Bacteriol. 169, 1341–1344
7. Sapienza, C., and Doolittle, W. F. (1982) Unusual physical organization of the bacteriorhodopsin gene. Nature 295, 384–389
8. Pfeifer, F., Friedmann, J., Boyer, H. W., and Betlach, M. (1984) Characterization of insertions affecting the expression of the bacteriorhodopsin gene in Halobacterium halobium. Nucleic Acids Res. 12, 2489–2497
9. Pfeifer, F., and Blaseio, U. (1989) Analysis of insertion mutants reveals two new genes in the pNRC100 gas vesicle gene cluster of Halobacterium halobium. Nucleic Acids Res. 17, 7785–7793
10. Ng, W.-L., and DasSarma, S. (1989) Analysis of insertion mutants reveals two new genes in the pNRC100 gas vesicle gene cluster of Halobacterium halobium. Nucleic Acids Res. 17, 7785–7793
11. DasSarma, S. (1989) Mechanisms of genetic variability in Halobacterium halobium: the purple membrane and gas vesicle mutations. Can. J. Microbiol. 35, 65–72
12. Jones, J. G., Hackett, N. R., Halladay, J. T., Scothorn, D. J., Yang, C. F., Ng, W. L., and DasSarma, S. (1988) Analysis of insertion mutants reveals two new genes in the pNRC100 gas vesicle gene cluster of Halobacterium halobium. Nucleic Acids Res. 17, 7785–7793
13. Ng, W.-L., Arora, P., and DasSarma, S. (1994) Large deletion in class III gas vesicle-deficient mutants of Halobacterium halobium. Syst. Appl. Microbiol. 16, 506–568
14. Pfeifer, F., Friedman, J., Boyer, H. W., and Betlach, M. (1984) Characterization of insertions affecting the expression of the bacterio-opsin gene in Halobacterium salinarum. Mol. Microbiol. 16, 506–568
15. Keseler, I. M., Huerta-Cerda, G., Casadaban, M. J., and Davidson, N. (2001) Transformation methods for halophilic archaebacteria. J. Bacteriol. 183, 2201–2205
16. Soppa, J., and Link, T. A. (1997) The TATA-box-binding protein (TBP) of Halobacterium salinarum. Cloning of the tcp gene, heterologous production of TBP and folding of TBP into a native conformation. Eur. J. Biochem. 249, 318–324
17. Baliga, N. S., Goo, Y. A., Ng, W. V., Hood, L., Daniels, C. J., and DasSarma, S. (2000) Gene expression in Halobacterium NRC-1 regulated by multiple TBP and TFB transcription factors. Mol. Microbiol. 36, 1184–1185
18. Ng, W. V., Ciufo, S. A., Smith, T. M., Bumgarner, R. E., Baskin, D., Faust, J., Hall, B., Loretz, C., Seto, J., Slagel, J., Hood, L., and DasSarma, S. (1998) Construction of a large dynamic replica in a halophilic archaeon: megaplasmid or minichromosome? Genome Res. 8, 1131–1141
19. Rupp, A., and Soppa, J. (1996) Fermentative arginine degradation in Halobacterium salinarum (formerly Halobacterium halobium): genes, gene products, and transcripts of the arcRACB gene cluster. J. Bacteriol. 178, 4942–4947
20. Gray, V. L., and Fit, P. S. (1976) An improved synthetic growth medium for Halobacterium cutirubrum. Can. J. Microbiol. 22, 440–442
21. McCready, S., Muller, J. A., Boubriak, I., Berquist, B. R., Ng, W. L., and DasSarma, S. (2005) UV irradiation induces homologous recombination genes in the model archaeon, Halobacterium sp. NRC-1. Salmon. Systems 1, 3
22. McCready, S., and Marcello, L. (2003) Repair of UV damage in Halobacterium salinarum. Biochem. Soc. Trans. 31, 694–698
23. Tebbe, A., Klein, C., Bisle, B., Siedler, F., Scheffer, B., Garcia-Rizo, C., Wolfertz, J., Hickmann, V., Pfeiffer, F., and Oesterhelt, D. (2005) Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome annotation. Proteomics 5, 168–179
24. Klein, C., Garcia-Rizo, C., Bisle, B., Scheffer, B., Zischka, H., Pfeiffer, F., Siedler, F., and Oesterhelt, D. (2003) The membrane proteome of Halobacterium salinarum. Proteomics 5, 180–197