Quantitative Proteomics of the Archaeon Methanococcus maripaludis Validated by Microarray Analysis and Real Time PCR*§

Qiangwei Xia‡§¶, Erik L. Hendrickson§¶, Yi Zhang‡§¶, Tiansong Wang‡§¶, Fred Taub§, Brian C. Moore§, Iris Porat**, William B. Whitman**, Murray Hackett‡ ‡‡, and John A. Leigh§

For the archaeon Methanococcus maripaludis, a fully sequenced and annotated model species of hydrogenotrophic methanogen, we report validation of quantitative protein level expression ratios on a proteome-wide basis. Using an approach based on quantitative multidimensional capillary HPLC and quadrupole ion trap mass spectrometry, coverage of gene expression approached that currently achievable with transcription microarrays. Comprehensive mass spectrometry-based proteomics and spotted cDNA arrays were used to compare global protein and mRNA levels in a wild-type (S2) and mutant strain (S40) of M. maripaludis. Using linear regression with 652 expression ratios generated by both the proteomic and microarray methods, a product moment correlation coefficient of 0.24 was observed. The correlation improved to 0.61 if only genes showing significant expression changes were included. A novel two-stage method of outlier detection was used for the protein measurements when Dixon’s Q-test by itself failed to give satisfactory results. The log2 transformations of the number of peptides or isotopic peptide pairs associated with each ORF, divided by the predicted molecular weight, were found to have moderately positive correlations with two bioinformatic predictors of gene expression based on codon bias. We detected peptides derived from 939 proteins or 55% of the genome coding capacity. Of these, 60 were overexpressed, and 34 were underexpressed in the mutant. Of the 1722 ORFs encoded in the genome, 1997 or 93% were probed by cDNA arrays. Of these, 50 were more highly expressed, and 45 showed lower expression levels in the mutant relative to the wild type. 15 ORFs were shown to be overexpressed by both methods, and two ORFs were shown to be upregulated by both proteomics and underexpressed by microarray. *Molecular & Cellular Proteomics 5:, 868–881.

The hydrogenotrophic methanogens are a major group of Archaea that conserve energy from the use of molecular hydrogen or formate to reduce carbon dioxide to methane. Methanococcus maripaludis is a model species of hydrogenotrophic methanogen, having favorable laboratory growth behavior, excellent genetic tools (1), and a fully sequenced genome (2). In another publication (3) we describe the molecular biology of a mutant of M. maripaludis (S40) in which the operon encoding the Ehb hydrogenase was disrupted, and functional changes were observed that led to a model for the role of Ehb in energy metabolism. Here we present validation of the multidimensional capillary HPLC/tandem mass spectrometry approach (also known as MudPIT1) to quantitative and qualitative proteomics for M. maripaludis by comparison with mRNA expression. Transcriptome analysis was performed in which the ehb mutant and wild-type strains were compared globally for all known genes under the same culture conditions as for the proteomics. For a subset of genes the proteomic results were further validated by real time quantitative RT-PCR. The relative absence of evidence for posttranscriptional gene regulation and the favorable protein extraction properties of M. maripaludis were suggestive of the hypothesis that, under the experimental conditions reported here, the transcriptome and proteome should both be reasonably complete and show parallel trends for most genes. We also consider codon bias relationships as they relate to the proteomic results, a two-stage algorithm for outlier detection in proteomics, and some of the unique strengths and weaknesses of comprehensive expression ratio calculations based on direct measurement of protein using MudPIT. Our purpose in undertaking these studies was primarily to better understand where the MudPIT approach to global protein measurements fits in to the broader scheme of our armamentarium for

1 The abbreviations used are: MudPIT, multidimensional protein identification technology; CAI, codon adaptation indices; E(g), Karlin index or general expression measure; HEG, highly expressed gene; MAD, median of the absolute deviation; Mmp, M. maripaludis; n, number of unique peptides per ORF; n1, number of 15N/14N peptide pairs per ORF; n2, number of redundant peptides per ORF; POR, pyruvate oxidoreductase; SSC, sodium chloride/sodium citrate buffer; 2D, two-dimensional; dH2O, distilled H2O; NHS, N-hydroxysuccinimide.

*From the Departments of ‡Chemical Engineering and §Microbiology, University of Washington, Seattle, Washington 98195 and **Department of Microbiology, University of Georgia, Athens, Georgia 30602
Received, November 11, 2005, and in revised form, January 12, 2006
Published, MCP Papers in Press, February 17, 2006, DOI 10.1074/mcp.M500369-MCP200

© 868–881 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org
doing systems biology with microbes. More specifically, we address the question is it reasonable to expect whole proteome level expression ratios from this approach to be comparable in terms of coverage and quantitative reliability to what can be achieved for the transcriptome using microarray technology?

MATERIALS AND METHODS

M. maripaludis Strains S2 and S40, Genome Sequence for S2, and Growth Conditions for Qualitative Proteomics of S2—For information regarding the molecular genetics of the M. maripaludis strains used and the construction of the ehb mutant (S40), see Porat et al. (3). Details regarding the genome sequence and annotation of M. maripaludis strain S2 have been published previously (2). The most current annotations are available upon request. Growth conditions for the five chemostat and batch cultures of S2 grown for purely qualitative proteomic analysis are described in the “growth conditions” note to Supplemental Table S3.

Differential $^{15}$N:$^{14}$N Labeling of Proteins—To prepare differentially $^{15}$N-labeled proteins, strains S2 and S40 were grown as described previously (6) in Methanococcus minimal medium plus 10 mM sodium acetate containing either 4.6 mM $^{[15]N}$ammonium sulfate (instead of ammonium chloride) or unlabeled ammonium sulfate in side-arm bottles. The bottles were constructed from 160-ml serum bottles by fusing a 28-ml culture tube to the side. Thus, it was possible to follow the culture absorbance in a spectrophotometer without removing samples from the cultures. Before use, glassware was immersed in 1 M HCl overnight, rinsed with water, autoclaved in water with soap, rinsed in deionized water, and dried. To remove O$_2$ adsorbed to the glassware, the bottles were stored in the anaerobic chamber at least a day before use. Stoppers were autoclaved for 20 min in 0.1 M NaOH, rinsed with deionized water, dried, and stored in the anaerobic chamber at least a day before use. Four bottles, each containing 10 ml of culture, were prepared for each growth condition. Cultures were initially pressurized to 276 kilopascals with H$_2$/CO$_2$ gas (80:20, v/v) initially pressurized to 276 kilopascals with H$_2$/CO$_2$ gas (80:20, v/v) and 30 ml of buffer B (100 mM ammonium bicarbonate (pH 9.0) and 2 mM DTT), the previously (6) in proteins were further reduced by the addition of 5 mM DTT at 37 °C for 1.5–2.0 min until the solution ceased to show obvious viscosity spectrometry. The samples were immediately placed in boiling water for conventional detergents that is less likely to interfere with the mass spectrometry. The samples were rinsed in deionized water, and dried. To remove O$_2$ adsorbed to the Samples were obtained from the two mixtures, two supernatants (CE1 and CE2) and two pellets (PT1 and PT2). For each sample, 200 μl of centrifuged cell lysate containing ~0.5 mg of total protein was brought to a final volume of 370 μl containing 40 μl of DNase/RNase solution (1 mg/ml DNase I, 500 μg/ml RNase A), 0.15% RapiGest (Waters), 5 mM DTT, and 50 mM Tris-HCl, at pH 8.0. The RapiGest is an acid-labile bifunctional molecule designed to serve as a substitute for conventional detergents that is less likely to interfere with the mass spectrometry. The samples were immediately placed in boiling water for 1.5–2.0 min until the solution ceased to show obvious viscosity due to residual DNA. Each sample was then transferred onto ice. The proteins were further reduced by the addition of 5 mM DTT at 37 °C for 30 min and then alkylated with 10 mM iodoacetamide at 30 °C for 30 min in the dark. Each sample was then adjusted to give a solution containing 50 mM NH$_4$CO$_3$ and 5 mM CaCl$_2$. 15 μg of sequencing grade trypsin (Promega, Madison, WI), was added, and the mixture was incubated at 37 °C for 4 h. The sample was acidified with TFA to quench the digestion and concentrated to 200 μl using a vacuum centrifuge (RC10-22, Jouan Inc., Winchester, VA).

Liquid Chromatography and Tandem Mass Spectrometry—Off-line HPLC fractionation, 2D microcapillary HPLC, and LCQ ion trap mass spectrometry were essentially as described previously (7) with minor modifications. Samples were vortexed and centrifuged at 14,000 rpm in a microcentrifuge (Eppendorf 5415C) for 8 min to remove insoluble material. 30 μl of 50% acetonitrile was added followed by another vortex and centrifugation at 14,000 rpm for 4–5 min to remove any remaining insoluble material. Samples were fractionated using a YMC-Pack™ PolymerC18™, S-6, 2.0 × 150-mm reversed-phase HPLC column (Waters). The mobile phase was H$_2$O and acetonitrile with 0.1% TFA. Peptides were eluted with increasing acetonitrile (2–60% in 60 min) at 0.4 ml/min. For samples CE1 and CE2, eluant containing peptides was detected by UV absorption at 214 nm and was fed 1 ml at a time into each of five tubes successively, rotating through all five tubes several times until all peptide-containing eluant was collected, resulting in five combined fractions. For samples PT1 and PT2 only two combined fractions were similarly collected. Each combined fraction was concentrated to 100 μl using a vacuum microcentrifuge, and acetic acid was added to a final concentration of 0.5% (v/v). Approximately 2.5 μl from each combined fraction was analyzed using a 2D microcapillary HPLC system (7, 8) combined with a Thermo Finnigan LCQ Classic or LTQ mass spectrometer in a semiautomated, data-dependent manner as described previously (7). Data acquired for the qualitative analysis of unlabeled strain S2 using the LTQ linear ion trap were similar and differed mainly in the number of usable mass spectra acquired per unit time. In our experiments the LTQ could collect data about 3–5 times faster, consistent with the experience of other laboratories (9).

Proteomic Data Collection and Analysis—Raw mass spectral data were collected, consisting of full scan (MS$^1$) data and, for selected parent ions, CID scan (MS$^2$) data. The SEQUEST algorithm (10) was used to identify MS$^2$ spectra that were candidates for peptides encoded in the M. maripaludis strain S2 ORF database (2) (GenBank™ accession number BX950229). Our M. maripaludis (Mmp) FASTA database (version dated January 21, 2005) was enlarged by adding ~25% of the nrdb human subset to increase the size of the database to about 21 megabytes to avoid statistical problems associated with searching very small databases. Two separate SEQUEST parameter files, one for $^{14}$N-peptides and one for $^{15}$N-peptides, were used (4, 5, 10). DTASelect (11) was then used to filter the results based on the following criteria. Peptides were required to be fully tryptic (beginning and ending at adjacent predicted trypsin digestion sites). The minimal peptide length was four amino acids. ΔCn/Xcorr values for different peptide charge states were 0.08/1.9 for $+1$, 0.08/2.0 for $+2$, and 0.08/3.3 for $+3$. This process generated two DTASelect-filter.txt files, one for $^{14}$N-peptides and one for $^{15}$N-peptides.

A program calculating the protein expression ratios for S40:S2 was written under the FileMaker 7.0 environment (Filemaker, Inc., Santa Clara, CA). Briefly peptide pairs were found for each identified peptide that was a unique peptide (sequence present in only one predicted ORF in the Mmp database) in both DTASelect-filter.txt files as follows. For the parent ion selected for each CID scan, the m/z value for the corresponding isotopic “flip” ion was calculated. The raw MS$^1$ data were then searched for an ion with that m/z value in the full scan from which the original CID scan was derived during the parent ion selection process. If no flip ion was found, then the noise value near the predicted m/z for the flip ion was used instead. A non-0 value for the
missing member of the $^{15}$N/$^{14}$N peptide pair was essential to avoid prematurely rejecting data for potentially highly regulated genes. The raw data were then similarly searched for “heavy” and “light” peptide pairs in the MS1 scans immediately prior to and immediately after the scan used to select the original parent ion. Of the three MS1 scans, data were retained for the scan with the highest signal intensity. This process was repeated for as many heavy $^{15}$N and light $^{14}$N peptide pairs as the program could find for each ORF.

An average ratio was calculated for each protein from all the $^{15}$N/$^{14}$N and $^{14}$N/$^{15}$N peptide pairs associated with that protein. Normalization of the average ratios was done in a manner similar to methods that are used for microarrays (12). Briefly a frequency distribution histogram of the log$_2$-transformed ratios was constructed (data not shown), and then the ratios were normalized by a factor that made the distribution center at 0 (12). The number of $^{15}$N/$^{14}$N peptide pairs per ORF ($n$) was used as a measure of reliability and analytical precision. Differential protein levels were regarded as significant if $n$ was equal to or greater than 3 and the ratio (S40:S2) $\pm$ S.D. was above or below 1.0. These calculations were done separately for each of the four samples, CE1, CE2, PT1, and PT2. Finally data from all four samples were combined and used to make a composite analysis (summary) of the expression pattern for S40:S2 (see Supplemental Table S2). For the composite analysis, peptide ratios determined to be outliers were eliminated. The detection of outliers was done in two steps for each list of peptide ratios associated with an ORF. For the first stage, Dixon’s Q-test (13) was used; in the second stage, a median of the absolute deviation (MAD) modified z-score test (14–16) with a cutoff value of 3.5 was used.

Confirmation by Peptide Synthesis—The following five peptides were synthesized commercially (Anaspec, San Jose, CA) and used for comparison of the original CID mass spectrum (MS$^2$) with the CID spectrum acquired for the synthetic peptide standard. The Mmp number is given for each ORF followed by the sequence tag: Mmp0512, ELQKYAK; Mmp0512, NACKMGAAKFR; Mmp1456, IQIPIFR; Mmp1517, HFGAWSEFR; and Mmp1517, AWYEGLAIFLKL.

Preparation of RNA—Cells were grown as for the differential $^{15}$N/$^{14}$N labeling of proteins except that only $^{14}$N was used. When the average $A_{600}$ reached 0.55, the cultures were rapidly cooled in an ice-water bath. For these experiments, cultures of S2 and S40 reached this optical density after 12 and 36 h, respectively. The cells were collected by centrifugation at 8,000 $\times$ g for 30 min at 4 °C, washed in 1 ml of minimal nitrogen-free growth medium (excluding sodium bicarbonate and stored at $-70^\circ$C. The cells were then resuspended in 200 µl of minimal nitrogen-free growth medium (excluding sulfide (17)) and lysed by the addition of 250 µl of RLT buffer (Qiagen, Valencia, CA). 350 µl of 100% ethanol (Aaper, Shelbyville, KY) was then added, and the RNA was purified on RNeasy columns (Qiagen). Aliquots of RNA were then treated with DNase using DNA-Free (Ambion, Austin, TX). Reactions contained 0.4 units of DNase I/µl of reaction mixture and were terminated with 10 µl of DNase inactivation reagent (Ambion).

The quality of the RNA samples was checked using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Palo Alto, CA). The absence of DNA was confirmed by a PCR using a Biometra T gradient thermocycler (Whatman Biometra, Goettingen, Germany) and primers QR-NAR and QRNAF (Supplemental Table S1) for ORF Mmp06008.

Array Labeling and Hybridization—Arrays constructed by generating PCR products of ~500 base pairs in length for each of the predicted ORFs were purchased from Integrated Genomics (Chicago, IL) and spotted in duplicate on mirrored slides at the University of Washington Center for Expression Arrays. Labeling reactions were conducted with the amino allyl cDNA labeling kit (Ambion) using a modification of the published procedure. For each reaction, 5–10 µg of RNA and 4 µg of random decamers (Integrated DNA Technologies, Coralville, IA) were incubated at 75 °C for 7 min and then briefly centrifuged at room temperature. First strand cDNA synthesis was carried out in 1× RT buffer, 10 units of RNase inhibitor, 0.5 mM dATP, 0.5 mM dCTP, 0.5 mM dGTP, 0.15 mM dTTP, 0.15 mM amino allyl-dUTP, and 400 units of reverse transcriptase for 10 min at room temperature followed by 2 h at 42 °C. The RNA template was then hydrolyzed by adding 4 µl of 1 M NaOH and incubating at 65 °C for 15 min. The reaction was neutralized by the addition of 10 µl of 1 M HEPES; cDNA was precipitated by the addition of 3.4 µl of 3 M sodium acetate, 1 µl of 5 mg/ml glycogen, and 100 µl of 100% ethanol (Aaper), incubation for 15 min at $-20^\circ$C, and centrifugation at maximum speed in a microcentrifuge for 30 min at 4 °C. The samples were then washed with 75% ethanol, and the supernatants were discarded. For NHS-Cy dye coupling, the samples were resuspended in 4.5 µl of coupling buffer (Ambion), 2.5 µl of RNase-free H$_2$O, and 3 µl of either NHS-Cy3 or NHS-Cy5 (Amersham Biosciences) and incubated at room temperature in the dark for 1 h. The reaction was quenched with the addition of 6 µl of 4 M hydroxylamine (Ambion) followed by a 15-min incubation at room temperature in the dark. Unincorporated dye was removed by adding 100 µl of PB buffer (Qiagen) and 750 µl of 100% ethanol (Aaper) and applying the sample to a MinElute reaction cleanup column (Qiagen). Samples were eluted from the column with 60 µl of RNase-free H$_2$O (Ambion). The whole, undiluted sample was analyzed using a spectrophotometer to determine DNA yields and dye concentrations. Only samples with at least 35 pmol of dye/mg of DNA were used for hybridization.

Samples were prepared for hybridization by mixing 100–120 pmol each of the Cy3- and Cy5-labeled cDNA and centrifuging them to dryness in a SpeedVac (Jouan Inc.). The samples were then resuspended in 60 µl of hybridization buffer (Agilent Technologies Inc.), boiled for 3 min, and incubated on ice for 30 s.

Slides were washed in sterilized dH$_2$O and dried with pressurized nitrogen gas. Resuspended samples were then applied along one edge of the slide, and a clean glass 24 × 60-mm coverslip (VWR, West Chester, PA) was lowered slowly over the sample starting with that edge. The slides were then sealed in a plastic container with reservoirs of dH$_2$O to maintain humidity, and the hybridization was carried out at 42 °C for 14–16 h in the dark. All posthybridization washes were performed in 50-ml conical tubes (Greiner, Longwood, FL). A brief wash in 1× SSC + 0.2% SDS was used to remove the coverslip followed by a 10-min wash in 1× SSC + 0.2% SDS and two 10-min washes in 0.1× SSC + 0.2% SDS, all at 54 °C, with rotation in a hybridization oven covered with aluminum foil. Finally slides were subjected to two 1-min washes at room temperature.
M. maripaludis Protein Expression

Table I

Expression of genes involved in energy conversion by proteomic and transcriptional array analyses

| ORF/ function | Descriptiona | Proteomic datab | Array datac |
|---------------|--------------|-----------------|-------------|
|               |              | Average ratio   | n,          | S.D. | Ratio | S.D. | p value   |
| A,A,ATPase    |              |                 |             |      |       |      |           |
| Mmp1038α      | Subunit H    | 0.96            | 42          | 0.25 | 0.65  | 0.14 | 3.43E-04 |
| Mmp1039       | Subunit I    | 1.09            | 85          | 0.33 | 0.65  | 0.18 | 7.96E-04 |
| Mmp1041       | Subunit IE   | 1.17            | 38          | 0.41 | 0.69  | 0.24 | 1.76E-02 |
| Mmp1040α      | Subunit K    | 1.13            | 92          | 0.40 | 0.63  | 0.15 | 1.41E-04 |
| Mmp1042       | Subunit C    | 1.25            | 39          | 0.39 | 0.75  | 0.17 | 1.16E-02 |
| Mmp1043α      | Subunit F    | 1.20            | 150         | 0.38 | 0.74  | 0.13 | 1.04E-03 |
| Mmp1044       | Subunit A    | 1.11            | 103         | 0.39 | 0.71  | 0.11 | 2.00E-03 |
| Mmp1045       | Subunit B    | 1.30            | 55          | 0.56 |       |      |           |
| Mmp1046       | Subunit D    | 1.09            | 209         | 0.46 | 0.67  | 0.10 | 2.57E-03 |
| Mmp1555       | β subunit    | 1.22            | 44          | 0.37 |      |      |           |
| Mmp1556       | Protein D    | 1.31            | 6           | 0.93 | 0.76  | 0.27 | 1.31E-01 |
| Mmp1557       | Protein C    | 1.01            | 194         | 0.33 | 0.74  | 0.12 | 2.57E-03 |
| Mmp1558       | γ subunit    | 0.98            | 188         | 0.35 | 0.81  | 0.08 | 4.70E-03 |
| Mmp1559       | α subunit    | 1.28            | 8           | 0.60 | 0.65  | 0.18 | 7.96E-04 |
| Mmp1560       | Subunit E    | 0.92            | 6           | 0.32 | 0.90  | 0.40 | 1.42E-01 |
| Mmp1562α      | Subunit C    | 1.22            | 44          | 0.37 | 0.71  | 0.15 | 3.43E-04 |
| Mmp1563       | Subunit B    | 1.43            | 17          | 0.29 | 0.70  | 0.18 | 8.33E-03 |
| Mmp1564       | Subunit A    | 0.93            | 29          | 0.33 | 0.71  | 0.14 | 2.62E-02 |
| Mmp1565α      | Subunits A- and F-related proteins | 1.11 | 109 | 0.42 | 0.73  | 0.17 | 5.29E-03 |
| Mmp1566α      | Subunit G    | 0.97            | 3           | 0.47 | 0.34  | 0.15 | 5.51E-05 |
| Mmp1567       | Subunit H    | 0.99            | 22          | 0.40 | 0.36  | 0.16 | 3.56E-06 |
| Flagella      | Flagellin B1 precursor | 1.04 | 2 | 0.24 | 0.39  | 0.16 | 1.70E-06 |
| Mmp1666       | Flagellin B2 | 1.02            | 35          | 0.19 | 0.55  | 0.17 | 1.64E-04 |
| Mmp1667       | Flagellin B3 | 1.09            | 20          | 0.45 | 0.65  | 0.14 | 2.37E-02 |
| Mmp1668       | Flagella accessory protein C | 1.14 | 1 | 0.75 | 0.10 | 4.18E-02 |
| Mmp1669       | Flagella accessory protein D | 0.77 | 0.20 | 1.87E-01 |
| Mmp1670       | Flagella accessory protein E | 0.60 | 0.12 | 2.97E-04 |
| Mmp1671       | Flagella accessory protein F | 0.63 | 0.12 | 2.97E-04 |
| Mmp1672α      | Flagella accessory protein G | 1.14 | 1 | 0.75 | 0.10 | 4.18E-02 |
| Mmp1673       | Flagella accessory protein H | 0.77 | 0.20 | 1.87E-01 |
| Mmp1674α      | Flagella accessory protein I | 0.85 | 0.12 | 1.32E-01 |
| Mmp1675       | Flagella accessory protein J | 0.89 | 0.09 | 2.74E-01 |
| Mmp1676       | Flagella accessory protein K | 2.40 | 3 | 3.23 | 0.89  | 0.09 | 2.74E-01 |

a The ORF description is derived from the genome annotation (2).
b Numerical data for proteomic analysis report average S40:S2 ratios, n, used to derive ratios, and standard deviations for the combination of four samples as described in the text and Supplemental Table S2. Bold indicates significantly higher expression in the mutant; bold italics indicate differential expression. The number of replicates for array analysis was 16.
c Numerical data for array analysis report average S40:S2 ratios, standard deviations, and p values (less than or equal to 0.01 was taken to indicate differential expression). The number of replicates for array analysis was 16.
d mRNA not probed.
e Protein not detected.
f Protein not detected.

Numerical data for proteomic analysis report average S40:S2 ratios, n, used to derive ratios, and standard deviations for the combination of four samples as described in the text and Supplemental Table S2. Bold indicates significantly higher expression in the mutant; bold italics indicate differential expression. The number of replicates for array analysis was 16.

Molecular & Cellular Proteomics 5.5 871
**M. maripaludis** Protein Expression

**Table II**

*Expression of other potentially regulated genes by proteomic and transcriptional array analyses*

Data specifications are as in Table I. See Supplemental Table S2 for the complete list of results for all genes. ABC, ATP-binding cassette; SSU, small subunit.

| ORF/ function | Description | Proteomic data | Array data |
|---------------|-------------|----------------|------------|
|               |             | Average ratio  | n₁ | S.D. | Ratio | S.D. | p value |
| Transport/regulation | | | | | | | |
| Mmp0165      | ABC transporter, family 2 | 1.49 | 2 | 0.93 | 1.77 | 0.67 | 5.94E-03 |
| Mmp0166      | Integral membrane protein | 1.41 | 6 | 1.14 | 2.86 | 1.75 | 3.39E-05 |
| Mmp0167      | ABC transporter ATP-binding protein | 1.30 | 62 | 0.73 | 4.46 | 2.46 | 5.11E-06 |
| Mmp0168      | Transcriptional regulator PadR-like | 0.99 | 16 | 0.23 | 3.32 | 1.71 | 7.29E-06 |
| Mmp1510      | Amidotransferase subunit A | 0.90 | 15 | 0.45 | 1.29 | 0.13 | 3.28E-03 |
| Mmp1511      | Sodium:alanine symporter | 2.31 | 18 | 1.07 | 1.39 | 0.22 | 7.45E-03 |
| Mmp1512      | Alanine racemase | 1.03 | 19 | 0.42 | 1.00 | 0.10 | 9.40E-01 |
| Mmp1513      | Alanine dehydrogenase | 0.98 | 46 | 0.30 | 0.93 | 0.10 | 3.86E-01 |
| Translation | | | | | | | |
| Mmp1319      | SSU ribosomal protein S13 | 1.06 | 22 | 0.74 | 1.25 | 0.17 | 1.75E-01 |
| Mmp1320      | SSU ribosomal protein S4P (S9E) | 1.13 | 71 | 0.40 | 1.49 | 0.49 | 7.04E-02 |
| Mmp1321      | SSU ribosomal protein S11 | 1.04 | 41 | 0.29 | 1.57 | 0.29 | 2.57E-03 |
| Mmp1322      | RNA polymerase, α chain | 1.18 | 18 | 0.46 | 1.52 | 0.22 | 9.11E-04 |
| Mmp1323      | Ribosomal protein L15 | 1.04 | 24 | 0.36 | 1.56 | 0.26 | 1.55E-03 |
| Mmp1324      | Ribosomal protein L13 | 1.14 | 32 | 0.57 | 1.53 | 0.21 | 9.31E-03 |
| Mmp1325      | SSU ribosomal protein S9P | 0.98 | 46 | 0.30 | 1.86 | 0.32 | 2.05E-05 |
| Hypothetical proteins | | | | | | | |
| Mmp0883      | Hypothetical protein | 0.65 | 4 | 0.01 | 0.79 | 0.15 | 5.46E-02 |
| Mmp0884      | Hypothetical protein | 0.84 | 7 | 0.14 | 0.87 | 0.27 | 5.22E-01 |
| Mmp1158      | Conserved hypothetical protein | 1.15 | 8 | 0.67 | 0.62 | 0.11 | 1.04E-04 |
| Mmp1159      | Ferritin | 0.77 | 4 | 0.58 | 0.69 | 0.16 | 1.29E-02 |
| Mmp1160      | Conserved hypothetical protein | 1.70 | 20 | 0.92 | 0.61 | 0.15 | 6.94E-04 |
| Mmp1161      | Conserved hypothetical archaeal protein | 1.14 | 19 | 0.39 | 0.74 | 0.18 | 1.32E-01 |
| Mmp1162      | Flavodoxin:β-lactamase-like protein | 0.71 | 13 | 0.31 | 0.71 | 0.13 | 9.31E-03 |
| Mmp1163      | CutA1 divalent ion tolerance-related protein | 1.03 | 54 | 0.29 | 0.65 | 0.16 | 1.19E-03 |
| Mmp1212      | Undefined product | 1.08 | 30 | 0.49 | 0.70 | 0.16 | 9.31E-03 |
| Mmp1213      | Undefined product | 1.82 | 3 | 1.17 | 0.62 | 0.15 | 1.36E-03 |
| Mmp1714      | Conserved hypothetical protein | 1.68 | 3 | 0.46 | 0.57 | 0.14 | 2.43E-05 |
| Mmp1715      | Putative GTPase | 1.23 | 15 | 0.36 | 0.57 | 0.12 | 3.99E-05 |

* Protein not detected.

formed in triplicate with appropriate controls (no template control and no amplification control), 16 S rRNA was used as a standard (details of TaqMan probes are in Supplemental Table S1).

**Bioinformatics**—The quantitative combined array/proteomic data set (Supplemental Table S2), the qualitative proteomic dataset (Supplemental Table S3), color images of the *M. maripaludis* proteome and transcriptome (Fig. 2), the signal intensity/peptide number calculations (Fig. 4), and the codon bias calculations (Fig. 5) were generated using a Filemaker Pro 7.0 (Filemaker, Inc.) application developed in house.

**RESULTS**

**Summary Statistics for the M. maripaludis Proteome and Transcriptome**—We measured relative gene expression at the proteome and transcriptome levels in the mutant S40 compared with the wild-type S2. Proteomic data were based on mass spectrometric identification and quantitation of peptide pairs as described under “Materials and Methods.” Each peptide pair consisted of a unique amino acid sequence (present in only one predicted Mmp protein) either labeled metabolically with $^{15}$N ($>$98% incorporation) or $^{14}$N at natural abundance. Fig. 1 summarizes the observed expression changes for the entire organism. Table I describes expression changes that were observed for specific genes of interest believed to play a role in energy metabolism. Table II describes other Mmp genes that showed statistically significant expression changes between strains S2 and S40. Proteins encoded by 939 of the 1722 ORFs were detected, representing 55% of the coding capacity of the genome. Of these, 60 were up-regulated in the S40 mutant relative to the wild type, 45 were down-regulated, and the remainder lacked statistical support for regulation. Our spotted arrays contained DNA corresponding to 1597 of the 1722 ORFs encoded in the genome. Of these, 50 were up-regulated in the S40 mutant relative to the wild type, 45 were down-regulated, and the remainder lacked statistical support for regulation. Both mRNA and protein data were recorded for 889 ORFs (probed in arrays and detected in the proteome). Agreement occurred where 15 ORFs were recorded for 889 ORFs (probed in arrays and detected in the proteome).
Fig. 2. Genomic representation of the proteome (A) and transcriptome (B). Each circle represents an ORF in the order it is encoded in the genome (Mmp0001 in upper left, Mmp1722 at end of bottom row). Colored circles indicate results for combined proteome analyses of the sum of four samples (CE1, CE2, PT1, and PT2) and for array analyses. Green, up-regulated in S40 mutant; red, down-regulated in S40 mutant; yellow, data were collected, but there is no statistical evidence for differential expression; white, no data because mRNA was not probed or protein was not detected. Note that the meaning of yellow and white is somewhat different for the proteome and the transcriptome. Because absolute detection is measured in the proteome, yellow circles in A indicate that the protein was present, whereas white circles indicate that no protein was detected at our limit of detection for qualitative identification (an interpretable MS2 daughter ion spectrum was not observed at a signal intensity threshold of 40,000 data system counts for the parent ion when using the LCQ ion trap mass spectrometer). In contrast, yellow in B indicates only that a probe was included in the array, and white indicates that a probe was not included.
M. maripaludis Protein Expression

![Graph showing the correlation of mRNA and protein expression](image)

**Fig. 3.** Scatter plot showing the correlation of mRNA and protein expression. The plot of log$_2$-transformed differential mRNA and protein expression ratios (30) was generated using the proteomic data that contained >3 unique peptides for each ORF. Among a total of 652 ORFs for which both protein and mRNA data were recorded, 15 appeared to be up-regulated in the S40 mutant by both methods (shown in green); two ORFs showed statistically significant opposite trends in regulation between mRNA and protein levels and are marked as foreground in green for protein level (up-regulation) and background in red for mRNA levels (down-regulation). All other data points are in black. Product-moment correlation coefficients (18) were 0.24 for all 652 data points, 0.54 for 121 genes with microarray $p$ values <0.05, 0.61 for 61 genes with $p$ values <0.01, and 0.61 for the 17 data points (shown in color) that showed significant changes in both mRNA and protein levels.

and two were down-regulated in the arrays but up-regulated in the proteome. The complete quantitative datasets for both mRNA and protein level expression ratios, including the statistics used to assign significance with respect to expression changes, are summarized in Supplemental Table S2.

**Corroboration of Quantitative Results**—The results showing differential expression in the proteome and transcriptome were corroborated in several ways. First, as mentioned above 15 ORFs showed similar regulation in the proteome and transcriptome; all were up-regulated in the mutant. Second, in many cases the regulatory pattern was further supported by the observation that the ORFs were components of multicistronic operons in which most of the genes showed the same trends.

This pattern is apparent in Fig. 2 where colored images of the proteome (Fig. 2A) and transcriptome (Fig. 2B) present the ORFs by Mmp number following the physical map of the genome (2). In these plots green indicates statistically supported higher expression in the mutant compared with wild type, red indicates lower expression, yellow indicates no statistical support for differential expression, and white indicates no data for the reasons described in the figure legend. For the sake of simplicity and because the errors associated with the expression ratios are sufficiently large in both the proteome and the transcriptome, we indicate only the direction of change in green and red, not the magnitude of change. This operon-based corroboration was nearly always found in those 15 cases where both the arrays and the proteome showed similar regulation. It was frequently the case also where only the arrays showed regulation: of the 56 ORFs that were regulated in the array only, 29 were in one of eight multicistronic

| Gene name | Mmp number | PCR ratio$^a$ | Array ratio$^b$ | $p$ value$^c$ | Proteomic ratio$^d$ | $n_1$ |
|-----------|------------|---------------|----------------|--------------|---------------------|------|
| 16 S RNA$^e$ | RNA standard | 0.999 ± 0.002 | 1.00 ± 0.09 | 0.73 | 1.2 ± 0.5 | 157 |
| ppsA | Mmp1094 | 1.0 ± 0.5 | 1.02 ± 0.07 | 0.65 | 1.5 ± 0.2 | 10 |
| cbiA | Mmp1478 | 1.0 ± 0.1 | 0.7 ± 0.1 | 1.0 × 10$^{-3}$ | 1.1 ± 0.4 | 103 |
| atpB | Mmp1045 | 0.8 ± 0.0 | 0.6 ± 0.2 | 8.0 × 10$^{-4}$ | 1.3 ± 0.6 | 8 |
| mtrE | Mmp1560 | 0.9 ± 0.3 | 0.7 ± 0.2 | 8.0 × 10$^{-3}$ | 1.4 ± 0.3 | 17 |
| mtrA | Mmp1564 | 0.9 ± 0.2 | 0.6 ± 0.2 | 1.5 × 10$^{-6}$ | 2.0 ± 0.9 | 29 |
| porE$^g$ | Mmp1503 | 4.4 ± 2.1 | 2.8 ± 0.6 | 8.0 × 10$^{-6}$ | 3.5 ± 2.1 | 17 |
| ehlB | Mmp1623 | 4.9 ± 0.2 | 4.4 ± 0.8 | 1.0 × 10$^{-6}$ | 2.6 ± 0.9 | 9 |

$^a$ Except as noted, real time RT PCR assays were done in triplicate from a single RNA sample and normalized to 16 S rRNA as an unregulated standard.

$^b$ Array ratios were calculated from four biological replicates, each with flipped dye slides containing two copies of the array. See Supplemental Table S2 for the complete listing.

$^c$ $p$ values were calculated using a Wilcoxon non-parametric significance test (18) and represent the probability that the null hypothesis (a ratio of 1 indicating no difference between samples) is correct.

$^d$ Proteomic ratios were calculated from the average of $^{15}$N:$^{15}$N or $^{15}$N:$^{14}$N peptide pair ratios of detected peptides. See Supplemental Table S2 for the complete listing.

$^e$ 16 S rRNA served to normalize the assays for RNA concentration and was assayed in triplicate for every set of reactions run. This value represents the average of nine measurements.

$^f$ Real time RT-PCR was performed in triplicate on each of two biological samples. Hence these values represent the average of six measurements.

$^g$ porE encodes one of the subunits of pyruvate oxidoreductase. Pyruvate oxidoreductase activity was measured (3) and had an S40:S2 ratio of 2.6 ± 0.9.
operons with other ORFs that were coregulated. Similarly for the 72 ORFs regulated only in the proteome, 13 ORFs were in one of six operons that were coregulated. Hence there were good reasons to believe that many of the differences in expression observed by only one method were biologically significant. Third, the results from the proteome and the transcriptome were positively correlated overall. This relationship is shown in Fig. 3 where the expression ratios are plotted for the proteome versus the transcriptome for 652 ORFs that passed quantitative quality control criteria for both protein and mRNA measurements in both the mutant and the wild-type strains. Certain ORFs that were probed for but that did not yield signals associated with significant hybridization in the arrays were eliminated from the data used to plot Fig. 3 as were those not detected at the protein level.

Although the product moment correlation coefficient was weak at 0.24, one must keep in mind that these ORFs included most of those for which both proteomic and transcriptomic data were acquired, regardless of standard deviation or p value, implying mostly random noise centered about a net expression change of 0. If only genes showing microarray p values of less than 0.01 were included, the correlation improved to 0.61 (see Fig. 3). Finally the results from proteomics and transcriptome arrays agreed with enzyme activity measurements and real time RT-PCR. Enzymatic assays of pyruvate oxidoreductase (POR) activity showed that S40 contained 2.6 ± 0.9 times the activity found in S2 (3). This value is very close to the average of the expression ratios determined for the ORFs in the operon encoding the POR subunits (Mmp1502–1507) by both proteomics (2.6 ± 0.3) and arrays (2.5 ± 0.4). Furthermore a comparison of real time RT-PCR, transcriptome array, and proteomic results for seven selected ORFs is presented in Table III. ORFs were selected to represent different patterns of regulation in the transcriptome and the proteome. Real time RT-PCR and transcriptome arrays agreed in all cases within their ranges of precision. Real time RT-PCR and proteomics agreed in all cases except one and conflicted in the case of ORF Mmp1478.

**Statistical Evaluation of Protein Level Expression Ratios**—To learn more about the quantitative limits of the Mud-
**M. maripaludis Protein Expression**

**TABLE IV**

Summary of the *M. maripaludis* proteome by Oak Ridge Functional Class

| OFC* | OFC description* | DB* | S40 only† | S2 only‡ | S40 and S2§ | S40, S2 total¶ | Mmp total∥ |
|------|------------------|-----|-----------|---------|------------|-------------|-----------|
| 0    | Hypothetical     | 145 | 6         | 4       | 24         | 34          | 59        |
| 2    | Replication and repair | 36  | 2         | 3       | 14         | 19          | 28        |
| 3    | Energy metabolism | 86  | 4         | 3       | 54         | 61          | 73        |
| 4    | Carbohydrate metabolism | 42  | 2         | 2       | 25         | 29          | 38        |
| 5    | Lipid metabolism  | 8   | 0         | 2       | 3          | 5           | 6         |
| 6    | Transcription     | 41  | 0         | 2       | 25         | 27          | 36        |
| 7    | Translation       | 121 | 1         | 2       | 101        | 104         | 120       |
| 9    | Cellular processes | 33  | 2         | 1       | 21         | 24          | 28        |
| 10   | Amino acid metabolism | 86  | 2         | 7       | 64         | 73          | 84        |
| 12   | Metabolism of cofactors | 68  | 6         | 14      | 24         | 44          | 58        |
| 13   | Conserved hypothetical | 658 | 28        | 50      | 232        | 310         | 461       |
| 14   | Transport         | 82  | 5         | 5       | 17         | 27          | 50        |
| 15   | Signal transduction | 7   | 1         | 0       | 4          | 5           | 7         |
| 16   | Nucleotide metabolism | 43  | 1         | 3       | 31         | 35          | 40        |
| 1, 11 | Other            | 266 | 8         | 22      | 123        | 153         | 201       |
| Total |                  | 1722 | 68        | 120     | 762        | 950         | 1289      |

*Oak Ridge Functional Class Number; see genome.ornl.gov/microbial/mmar.
†Text descriptor for each Oak Ridge Functional Class Number as defined by Oak Ridge National Laboratory.
‡Database of protein-encoding ORFs derived from the genome annotation (2).
§Detected qualitatively only in S40 in the S40/S2 comparison experiment (complete data in Supplemental Table S2).
¶Detected qualitatively only in S2 in the S40/S2 comparison experiment.
∥Detected qualitatively in both S2 and S40 in the S40/S2 comparison experiment.

PIT proteomic approach, we analyzed the effects of increasing peptides measured per ORF for the S40 versus S2 dataset. The total number of redundant peptides involved in the measurement of a given protein correlated well with the number of isotopic 15N/14N peptide pairs as expected (Fig. 4A).

We use the number of “redundant peptides” per ORF (nD) to refer to all peptides that can be associated with a particular ORF, including multiple measurements of the same peptide. The total signal intensity in MS1 summed over all peptides associated with a given protein also correlated well with the total number of redundant peptides (Fig. 4B). Most importantly, both the range of relative standard deviation values (Fig. 4C) and the range of observed ratios (Fig. 4D) were widest at low numbers of peptide pairs per ORF and narrowed markedly at high numbers of peptide pairs. These data support the contention that it is difficult to sort out true regulatory changes based on quantitative proteomic data for low numbers of peptide pairs. On the other hand, true changes in protein expression tended to stand out dramatically as the number of peptide pairs per protein exceeded ~10 as indicated by the ORFs coded green in Fig. 4D for which up-regulation was supported by both proteomics and transcriptome arrays. These ORFs were generally supported by the strongest corroborative evidence for regulation as described above. In contrast, the two ORFs coded red in Fig. 4D did not stand out and may represent false positives in the proteome, the transcriptome, or both.

Proteome Coverage—Mathematical simulation, based on sampling theory and the known data acquisition characteristics of the LCQ mass spectrometer (data not shown), suggest that the true level of measurable protein expression is somewhat higher than the 55% of the annotated protein-encoding ORFs measured for S40:S2, around 70%. This number is based on our estimate of how many parent ions (MS1) that exceed our signal-to-noise threshold can actually be isolated per unit time for purposes of generating a collision spectrum (MS2). For the LCQ spectrometer, only about 3% of the potential parent ions generate daughter ion spectra under our conditions. Indeed proteome coverage was markedly greater in the larger dataset summarized in Supplemental Table S3 in which we have accumulated qualitative data for several samples of *M. maripaludis*. For the cumulative dataset we detected 1289 proteins or 75% of the predicted protein-encoding ORFs. The 1289 proteins detected in the *M. maripaludis* proteome are broken down by Oak Ridge Functional Class in Table IV. Preliminary data (not shown) using the LTQ mass spectrometer suggests that even higher coverage is possible, and we have measured protein expression for more than 80% of the protein-encoding ORFs for strain S2 Δhpt (1) grown under continuous flow (chemostat) conditions.

In addition to the coverage of the proteome in terms of the percentage of encoded proteins detected, our coverage on a per ORF basis was also reasonably good. Thus, in the S40 versus S2 comparison nD ranged from 1 to 325 (see Fig. 4A), although a minimum of three peptide pairs were required to flag the ORF as quantitative expression data (as in Fig. 3 and 4).
M. maripaludis Protein Expression

![Graphs A, B, and C](image)

**Fig. 5. Correlation between gene expression level predicted by codon usage bias and identified peptide numbers per ORF normalized by protein molecular weight.**

- **Graph A** shows $E(g) = 3.2469\ln(n) - 10.275$ with $R^2 = 0.4501$.
- **Graph B** shows $E(g) = 3.8686\ln(n) - 9.721$ with $R^2 = 0.4716$.
- **Graph C** shows $E(g) = 3.2744\ln(n) - 10.438$ with $R^2 = 0.4504$.

This conclusion is based on the data for $n_1$, $n_2$, and number of unique peptides per ORF ($n$) compiled in Supplemental Tables S2 and S3. For all but a few proteins, listed in Supplemental Table S4, qualitative detection was based on multiple $n$ values ranging from 1 to 51 for the cumulative M. maripaludis dataset (see Supplemental Table S3). More details are provided regarding the interrelationships among $n$, $n_1$, and $n_2$ with respect to qualitative and quantitative detection in the notes to the supplemental tables.

We suspect that global coverage for the S40 versus S2 comparison was better on the mRNA side than it was with the proteomics because trends were generally more consistently observed across known operons in the arrays (Fig. 2). However, the difference in coverage may not be as great as might be concluded from a superficial comparison of the numbers of colored (non-white) circles in Fig. 2, A and B. This is because yellow circles in the microarray data (Fig. 2B) do not imply detection of mRNA in any absolute sense, only that the array contained a probe for the gene and that there was no statistically significant difference in mRNA level between S40 and S2. In contrast, yellow circles in Fig. 2A do indicate absolute detection.

**Peptide Number as a Predictor of Protein Abundance—**As has been observed by other laboratories working with different model systems (4, 5), our data tended to contain the highest numbers of peptides for those proteins that are generally known to have high abundance. Thus, the ORFs with the highest numbers of redundant peptides in our cumulative M. maripaludis database (Supplemental Table S3) included the S-layer protein that comprises the cell wall, subunits of the methyl-coenzyme M reductase that catalyzes a step in methanogenesis, the chaperonin Hsp60, translation elongation factors, and ribosomal proteins. These proteins are all known to be abundant in M. maripaludis.

To test the hypothesis that high peptide numbers ($n$, $n_1$, and $n_2$) are a general predictor of protein abundance, we plotted peptide numbers (normalized to the predicted molecular weight of the protein) versus Karlin indices (19), which are predictors of protein expression levels based on codon biases (Fig. 5). For one plot (Fig. 5A) we based our calculations on bioinformatic categories for highly expressed genes, including ribosomal proteins. Ribosomal proteins of M. maripaludis are 100% highly expressed by Karlin’s criteria, similar to bacteria but unlike most Archaea sequenced to date in which about 60% of the ribosomal proteins are highly expressed (20). In two other plots (Fig. 5, B and C), we based our calculations on our own measured peptide numbers. Either way, the plots were essentially the same and showed a moderate correlation between peptide numbers and Karlin indices. To achieve even a moderate correlation it was necessary to divide $n_1$ or $n_2$ by the predicted molecular weight of the protein because the higher the molecular weight, the higher the probability of detection. In general the correlations between codon bias and measured protein expression by peptide numbers are likely to remain moderate as long as mass spectrometer duty cycle, the number of mass spectra that can be acquired per unit time, is a limiting factor.

In addition to Karlin indices, we also calculated codon adaptation indices (CAI (21)), which have also been used to predict protein abundances. Both indices are recorded in Supplemental Table S3. However, unlike Karlin indices, CAI did not correlate well with our peptide numbers. Grocock and Sharp (22) have argued against the predictive value of CAI for...
organisms with strongly biased GC content; the mol % G + C of *M. maripaludis* is quite low at 33% (2).

**DISCUSSION**

**Evaluation of Proteomic Methodology**—To facilitate our discussion of the merits of the MudPIT method, we first present in Fig. 6 an example of mass spectral data used for the protein level expression measurements. Fig. 6 illustrates a quantitation approach based on mass-to-charge peak height in which the isolated parent ion (MS²) still plays the major role with respect to being both a source of fragments for purposes of qualitative identification and the signals used to calculate expression ratios. As suggested by the plots in Figs. 4 and 5, we could have used spectral counting (23), a technique that bases quantitation on the number of peptides recovered, and generated very similar expression ratios, keeping in mind the necessity to correct for molecular weight differences. Spectral counting approaches have been described as potentially having a wider dynamic range (24) relative to those based on ion chromatographic peak detection, and we find this possibility intriguing and worthy of further investigation. Our initial impression of the spectral counting approach as applied to the raw data reported here is that it does in fact extend the dynamic range of the quantitative measurements for many proteins. A detailed comparison of spectral counting and quantitation based on conventional metabolic labeling using stable isotopes is currently being undertaken in our laboratory for two microbes, *M. maripaludis* and *Porphyromonas gingivalis*, and will appear in a future publication.

**Fig. 6.** Representative mass spectral data used to identify the proteins and calculate expression ratios for S40:S2. Here we illustrate the logic flow used by our quantitation software with a single ratio calculation. Details for the chromatographic, data-dependent mass spectral data acquisition, and database searching parameters are given under "Materials and Methods." Briefly after separation of peptides by LC, two kinds of mass spectral scans were obtained. The primary (MS¹) scans contained intact parent ions from peptide mixtures. The CID (MS²) scans contained fragmentation ions derived from individual parent ions of the MS¹ scans. In addition, single ion chromatograms were generated that plot the intensity of a single MS¹ ion versus time. In the data analysis, peptide sequences belonging to predicted ORFs were identified in both heavy-¹⁵N and light-¹⁴N forms. In this example, the identified peptide was KPSTEPYKL whose sequence is unique for ORF Mmp1504 encoding the β subunit of POR (see Table III and Supplemental Table S2). A, CID spectrum (MS² scan number 664) from the doubly charged parent ion 510.4 identifying it as [¹⁴N]KPSTEPYKL. B, CID spectrum (MS² scan number 660) from the doubly charged parent ion 515.3 identifying it as [¹⁵N]KPSTEPYKL. The nomenclature used to designate key peptide sequence ions is that of Biemann (31). Ions labeled y and b indicate sequence-specific ions containing carboxyl and amino termini, respectively, and * indicates a loss of water or ammonia. The CID spectra (MS²) were used to generate a table of identified peptides. Each MS² spectrum was linked by the data system with a specific MS¹ ion, i.e. parent ion. Next single ion chromatograms were checked for each parent ion to determine which MS¹ scan contained the maximum signal intensity. C and D, single ion chromatograms of MS¹ m/z 510.4 and 515.3, respectively, showing that both intensities were maximum at scan number 661 (numbers by peaks indicate scan number). Having identified the MS¹ scan with the maximum observed intensities of the parent ions, this scan was measured for the intensities at m/z 510.4 and 515.3. E, MS¹ scan number 661 in bar graph format showing the two signals used in the calculation. The intensities were 5.14 × 10⁷ and 2.40 × 10⁷ counts, respectively, yielding an S40 heavy:S2 light ratio of 2.14. In total there were 76 ratios calculated (n) from heavy-light signal pairs that were acquired from eight unique peptide sequences for Mmp1504, yielding an average S40:S2 ratio of 2.68. If the average ratio from all measurements ± S.D. did not overlap with a ratio of 1.0, the ratio was judged to indicate a significant difference in expression at the protein level between S40 and S2.
The 2D LC-MS-MS approach to global gene expression differs from two other approaches we use routinely in fundamental ways. Unlike the spotted DNA arrays used here for transcriptome measurements, the proteomics involves detection in an absolute sense before calculation of relative expression levels. Thus, with the proteomics it is possible to ask whether or not a gene is being expressed in a qualitative sense under any given single state in addition to generating the relative quantitation information (e.g. S40 versus S2) typically seen in an array type experiment. In contrast, expression arrays are primarily designed to yield relative quantitation data, although recently certain commercial vendors of transcription microarray technology have made claims for routine absolute detection capability.

Our work agrees with other published accounts that the results with LC-MS-MS are vastly more informative as a read-out of gene expression than 2D PAGE technology (25, 26) at least for microorganisms. Indeed in the current comparison of results with LC-MS-MS are vastly more informative as a read-out of gene expression than 2D PAGE technology (25, 26) at least for microorganisms. In the current comparison of the S40 and S2 strains of M. maripaludis, we identified peptides belonging to 939 proteins, covering about 55% of the predicted protein-encoding ORFs. This is at least an order of magnitude improvement over what we have been able to achieve with 2D gel technology, and the MudPIT type of approach is far less labor-intensive. Although less labor-intensive, the large amount of relatively expensive mass spectrometry time, 2–4 weeks for a complete proteome analysis of a microbe with on the order of 2000 protein-encoding ORFs, limits wider use of the technique at present. Interestingly as we observed with P. gingivalis (8), there was no obvious analytical discrimination based on hydrophobicity or isoelectric point (data not shown). In contrast, 2D PAGE methods are generally limited by those factors. It is recognized that in certain situations where the goal is to get as much information as possible about the mature form(s) of the protein, including all posttranslational modifications and splice variants, spatial resolution of the protein prior to digestion and mass spectral analysis, as in 2D PAGE, can have certain advantages. However, these advantages would apply only to the relatively small numbers of proteins that are amenable to gel-based methods. For a small genome microbial organism, like M. maripaludis, in which there is little splicing and the primary structure of mature protein is generally the product of a single gene, the MudPIT type of approach seems to work well and is superior to the other approaches we have investigated for measuring protein expression.

Our data reported here and from other microbial expression studies suggest that we are discriminating against smaller gene products, proteins below ~7–10 kDa. Because very small ORFs are both difficult to detect in the genome during the annotation process and difficult to probe for in an array experiment, it is also difficult to assess the magnitude of such discrimination at this time. In general, for a method that is limited by the duty cycle of the mass spectrometer, the probability of a protein being detected in such a complex analytical scheme is directly proportional to its molecular weight and the degree to which it is expressed. The inability to reliably detect protein expression by short ORFs is a limitation shared by all global approaches of which we are familiar.

**Differing Trends in Arrays and Proteomics**—Those cases where the results of arrays and proteomics differed deserve further comment. In many cases expression changes in the same direction probably occurred at both the mRNA and proteome levels, but statistical support was found only for one method. In other cases the array or the proteomics data may contain false positives. On the proteomics side false positives become more likely when the number of peptide pairs measured is low as discussed previously. Hence in some cases differences in the arrays and the proteomics may be a statistical issue, and additional experiments are needed when the ORF in question is of particular interest. In other cases post-transcriptional regulation could have occurred, resulting in regulation in the proteome where there was none in the transcriptome. Finally differences could occur due to culture history and the relative stability of proteins. These experiments were performed with batch-grown cultures harvested during the linear growth phase. The linear growth phase is commonly observed during growth of hydrogenotrophic methanogens and represents the point in which the growth rate is limited by the transfer of H2 from the gas phase to the medium (27). In linear growth, the specific activity for methanogenesis is decreasing, and the cellular composition is changing. The cells for transcription analysis and proteomics were harvested at the same point in the growth cycle. However, because proteome analysis measures an analyte that is in general more stable with a much longer half-life in the cell, the proteome will reflect previous mRNA expression. It would include protein expression that occurred at low cell densities prior to harvest. In contrast, transcriptional arrays reflect the levels of mRNA at the time of harvesting. More detailed and useful correlation analysis would require explicit knowledge of the kinetics of biosynthesis and degradation for each mRNA and protein in the dataset.

**Data Compression**—Real time RT-PCR often shows larger -fold differences between conditions than transcriptome arrays (28, 29). The difference between the two techniques seems to correlate with the degree of regulation with larger -fold changes producing greater differences between the array and real time RT-PCR results (29). The results of the M. maripaludis array presented here are consistent with these observations. In general only modest differences were seen between S40 and S2 on the arrays, and these were in general agreement with the results of the real time RT-PCR assays. The consistency of the results is illustrated by the ORF Mmp1503, which encodes PorE, a subunit of pyruvate oxidoreductase (see Table III). For this ORF real time RT-PCR indicated only a slightly higher expression ratio than did arrays, proteomics, and enzyme activity measurements.
The moderate nature of the expression differences between S40 and S2 precluded a full analysis of data compression in the proteome. However, the linear dynamic range of the 2D LC-MS-MS approach in our experiments is only about 2 orders of magnitude, whereas the range of protein expression may be upward of 6–8 orders of magnitude (25, 26). Hence we expect that, similar to arrays, data compression will increase with increasing fold differences.

Outlier Detection and Removal—Low signal-to-noise is inherent in global analyses by quantitative 2D LC-MS-MS, necessitating the measurement of large numbers of 15N/14N peptide pairs to narrow the relative standard deviations of the ratios as shown in Fig. 4C. In some cases we used more than 300 peptide pairs to calculate the mean ratio for a single protein. This situation complicates the detection of outliers. Dixon’s Q-test (13) has been popular in the proteomics community. However, Dixon’s Q-test is not recommended when the number of data points is greater than about 30 (13) and handles groups of randomly clustered outliers characteristic of large datasets poorly, tending to include them as points to be retained. For this reason, we added a second stage of outlier detection based on a MAD modified z-score test (14–16). The MAD modified z-score was better at removing random clusters of multiple outliers for ORFs with large values for n and n1. By Dixon’s Q-test, 300 records were detected as outliers from 21,524 records, or 0.72%. Then by the MAD modified z-score test, 1520 records were detected as outliers from the remaining 21,224 records, or 7.2%. The z-score cutoff of 3.5 used in this study was derived empirically and should be viewed as appropriate for these data, not as a generally useful value.

Statistical Accuracy for Highly Regulated Genes—Our statistical screening appears to have accurately determined many of the proteins that are present at different levels in S40 versus S2 (see ORFs indicated in green in Fig. 4D). The statistical analysis appears to have worked well for most ORFs whose expression ratios using either proteomics or microarrays were within a factor of 4 of a mean normalized value of 1.0 on a linear scale (see Supplemental Table S2). On the proteomics side, ratios outside a factor of four tended to derive from measurements in which true signal, confirmed by a collision spectrum, was measured for one state, whereas noise or very low abundance signals of large standard deviation were measured in the other state. Several of the data points in the upper left portion of Fig. 4, C and D, were typical. Those ORFs were supported by low n1 values and may not reflect real changes in protein expression. However, genes showing an “all on” or “all off” type of regulatory pattern would in principle be especially prone to generating these types of data. This could lead to the paradoxical situation where a gene that may in fact have had a dramatic expression change could yield protein data that would fail to pass our statistical tests, due to highly variable but low signal intensity in the state where little expression was detected, that in turn would influence the standard deviation of the ratio as a whole. Such data would also run a higher risk of being rejected by our outlier detection algorithm. This problem has now been corrected by the use of a weighting scheme4 in which high signal-to-noise data contribute proportionately more to the standard deviation calculation for the ratio as a whole. The p values calculated for the spotted cDNA arrays using the Wilcoxon test were not subject to this problem.

Prospects for M. maripaludis Proteomics—The study presented here illustrates the feasibility of quantitative 2D LC-MS-MS-based proteomics for M. maripaludis. Metabolic labeling allowed the simultaneous comparative analysis of peptides from two conditions, analogous to “two-state” microarray experiments. Coverage was high, and quantitation was sufficiently accurate and precise to detect changes in protein expression that were subsequently confirmed by multiple lines of evidence. The measurement of large n1 values for a given ORF was essential for accurate and precise relative quantitation. We remain skeptical of the idea that biologically meaningful expression ratios can be generated from small numbers of peptide pairs regardless of the labeling or quantitation approach used. The relationship between number of heavy and light peptide pairs and confirmed expression changes shown in color in Fig. 4D is typical of what we see in all of our quantitative proteomic work with microbes and should be understood as a general phenomenon. True, validated, and biologically meaningful expression changes at the protein level in a two-state experiment require a high number of peptide pairs recovered per protein under our conditions. Our thinking about proteomic expression ratios parallels recent thinking in the microarray field where in general larger numbers of replicate measurements are now used to truly differentiate biological regulation from all the various sources of noise and error that can become an issue with such global assessments of gene expression (12). To return to the question asked in the Introduction, can we do as good a job with global measurements of the proteome using MudPIT as we can with the transcriptome using microarrays? The selected data presented in Tables I and II and the complete quantitative dataset presented in Supplemental Table S2 suggest an answer of “no, we are not quite there yet but getting close.” These data were collected with an LCQ instrument that is no longer state-of-the-art. Coverage and quantitation have improved markedly with advances in mass spectrometric technology, and our most recent data3 indicates that by using an LTQ mass spectrometer with a much faster scanning speed we can greatly minimize the duty cycle issue mentioned repeatedly in this report as a fundamental limitation of the MudPIT approach with respect to qualitative coverage and quantitative reliability issues. Thus, for a tractable model or-

4 T. Wang, Q. Xia, M. Meila, and M. Hackett, manuscript in preparation.
ganism with fewer than ~2000 protein-encoding ORFs, such as M. maripaludis, it should now be possible to make global comparisons for all known protein-encoding genes under the assumption that the information will be as reasonably complete for the proteome as for the transcriptome.

Acknowledgments—We thank Kevin Wheeler, Darwin Alonso, and Roger Bumgarner for assistance. We thank Kobi Alfantari for programming, Michrom Bioresources for equipment support, and Dave Tabb and John R. Yates for DTASelect and Contrast software.

* This work was supported by Microbial Cell Project DE-FG03-01ER15252 from the Department of Energy and National Institutes of Health Grant GM60403 (to J. A. L.). Funding for portions of the proteomics software and infrastructure development were provided by National Institutes of Health Grant DE014372 and the University of Washington College of Engineering (to M. H.). 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 data discussed in this publication have been deposited in the NCBI Gene Expression Omnibus (GEO) and are accessible through GEO Series accession numbers GSE2744 for the proteomics and GSE2745 for the spotted cDNA arrays.

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

†† To whom correspondence should be addressed: Dept. of Biological Engineering, Massachusetts Inst. of Technology, 77 Massachusetts Ave., Cambridge, MA 02139-4307.

REFERENCES

1. Moore, B., and Leigh, J. A. (2005) Markerless mutagenesis in Methanococcus maripaludis demonstrates roles for alanine dehydrogenase, alanine racemase, and alanine permease. J. Bacteriol. 187, 972–979

2. Hendrickson, E. L., Kaul, R., Zhou, Y., Bovee, D., Chapman, P., Chung, J., Conway de Macario, E., Dodsworth, J. A., Gillett, W., Graham, D. E., Hackett, M., Haydock, A. K., Kang, A., Land, M. L., Levy, R., Lie, T. J., Major, T. A., Moore, B. C., Porta, I., Palmeiri, A., Rouse, G., Saengphim-machak, C., Soll, D., Van Dien, S., Wang, T., Whitman, W. B., Xia, Q., Zhang, Y., Larimer, F. W., Olson, M. V., and Leigh, J. A. (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen Methanococcus maripaludis. J. Bacteriol. 186, 6856–6869

3. Porat, I., Kim, W., Hendrickson, E. L., Xia, Q., Zhang, Y., Wang, T., Taud, F., Hackett, B. C., Anderson, I. J., Hackett, M., Leigh, J. A., and Whitman, W. B. (2006) Disruption of the operon encoding Ehb hydrogenase limits anaerobic CO2 assimilation in the archaeon Methanococcus maripaludis. J. Bacteriol. 188, 1373–1380

4. Washburn, M. P., Ulaszek, R., Deciu, C., Schietz, D. M., and Yates, J. R., III (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. Anal. Chem. 74, 1650–1657

5. Washburn, M. P., Ulaszek, R. K., and Yates, J. R., III (2003) Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. Anal. Chem. 75, 5054–5061

6. Whitman, W. B., Shieh, J. S., Sohn, S. H., Caras, D. S., and Premachand-ran, U. (1986) Isolation and characterization of 22 mesophilic methano-cocci. Syst. Appl. Microbiol. 7, 235–240

7. Zhang, Y., Wang, T., Chen, W., Yilmaz, O., Park, Y., Jung, I. L., Lamont, R. J., and Hackett, M. (2005) Differential protein expression by Porphyromonas gingivalis in response to secreted epithelial cell components. Proteomics 5, 198–211

8. Wang, T., Zhang, Y., Chen, W., Park, Y., Lamont, R. J., and Hackett, M. (2006) Reconstructed protein arrays from 3D HPLC/tandem mass spectrometry and 2D gels: complementary approaches to Porphyromonas gingivalis protein expression. Analyst 127, 1450–1456

9. Mayya, V., Rezaul, K., Cong, Y. S., and Han, D. (2005) Systematic comparison of a two-dimensional ion trap and a three-dimensional ion trap mass spectrometer in proteomics. Mol. Cell. Proteomics 4, 214–223

10. Eng, J. K., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989

11. Tabb, D. L., McDonald, W. H., and Yates, J. R., III (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. 1, 21–26

12. Quackenbush, J. (2001) Computational analysis of microarray data. Nat. Rev. Genet. 2, 418–427

13. Rorabacher, D. B. (1991) Statistical treatment for rejection of deviant values: critical values of Dixon’s “Q” parameter and related subrange ratios. Anal. Chem. 63, 139–146

14. Muller, J. W. (2000) Possible advantages of a robust evaluation of comparisons. J. Res. Natl. Inst. Stand. Technol. 105, 551–555

15. Hampel, F. R. (1974) The influence curve and its role in robust estimation. J. Am. Stat. Assn. 69, 383–393

16. Burke, S. (2001) Missing values, outliers, robust statistics and non-parametric methods. Stat. Data Anal. LC-GC Eur. Online Suppl. 59, 19–24

17. Blank, C. E., Kessler, P. S., and Leigh, J. A. (1993) Genetics in methanogens: transposon insertion mutagenesis of a Methanococcus maripaludis nifH gene. J. Bacteriol. 177, 5773–5777

18. Sokal, R. R., and Rohlf, F. J. (1981) Biometry: the Principles and Practice of Statistics in Biological Research, 2nd Ed., pp. 448–449, W. H. Freeman, New York

19. Karrin, S., and Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. J. Bacteriol. 182, 5239–5250

20. Karrin, S., Mrazek, J., Ma, J., and Brocchieri, L. (2005) Predicted highly expressed genes in archaeal genomes. Proc. Natl. Acad. Sci. U. S. A. 102, 7303–7308

21. Sharp, P. M., and Li, W. H. (1987) The codon Adaption Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295

22. Grocock, R. J., and Sharp, P. M. (2002) Synonymous codon usage in Pseudomonas aeruginosa PA01. Gene (Amst.) 289, 131–139

23. Gao, J., Opiteck, G. J., Friedrichs, M. S., Dongre, A. R., and Hefta, S. A. (2003) J. Proteome Res. 2, 643–649

24. Liu, H., Sadygov, R. G., and Yates, J. R., III (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201

25. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 196–207

26. Washburn, M. P., and Yates, J. R., III (2000) Analysis of the microbial proteome. Curr. Opin. Microbiol. 3, 292–297

27. Reeve, J. N., Nolling, J., Morgan, R. M., and Smith, D. R. (1997) Methanococcus maripaludis: genes, genomes, and who’s on first? Annu. Rev. Microbiol. 51, 203–243

28. Biemann, K. (1992) Mass spectrometry of peptides and proteins. Molecular & Cellular Proteomics 5.5, 881

29. Wurmbach, E., Yuen, T., Ebersole, B. J., and Sealfon, S. C. (2001) Gona-dotropin-releasing hormone receptor-coupled gene network organization. J. Biol. Chem. 276, 47195–47201

30. Zybalov, B., Coleman, M. K., McDonald, Flenos, L., and Washburn, M. P. (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal. Chem. 77, 6218–6224

31. Biemann, K. (1992) Mass spectrometry of peptides and proteins. Annu. Rev. Biochem. 61, 977–1010