Identification of Proteins in Unsequenced Bacterial Strains Via Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry

Daisy-Malloy Hamburg, Moo-Jin Suh, Patrick A. Limbach, University of Cincinnati
Steven Gregory, Albert Dahlberg, Brown University

This work presents a protocol that has been developed to identify proteins from bacterial strains for which no DNA or protein sequence exists. A particular set of proteins are characterized using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOFMS) to obtain accurate protein molecular weights. The resulting molecular weights are then compared to DNA or protein sequences from reference strains, allowing for straightforward identification of proteins from any non-reference strains. This approach allows for the facile determination of differences in proteins at the amino-acid level as well as the post-translational level. This protocol was developed, validated, and applied to ribosomal proteins from various strains of *Thermus thermophilus*. The protocol allows for rapid identification of proteins and will be useful in phylogenic studies and biomarker identification.

Mass spectrometry (MS) is presently one of most powerful techniques for investigating proteins and their post-translational modifications [1]. The development of proteomics has benefited from the application of “bottom-up” (proteolytic fragment identification) [2] and “top-down” (intact protein mass measurement) [3] strategies coupled to mass spectrometry, allowing for large-scale analysis of proteins present in an organism, tissue, or cell under a given set of physiological conditions [1]. A key component for the successful application of these strategies in proteomics is the presence of DNA or protein databases, which allow experimental data to be characterized based upon the closeness of fit to protein sequences available in these databases. However, there are often investigations requiring knowledge of protein sequences from species or strains, which are not publicly available.

The ribosome, found in all organisms, is the subcellular organelle that performs the activity of protein synthesis. In previous work, we have optimized the MALDI-MS conditions for analyzing *E. coli* ribosomes [4]. The work described here is an extension of that work into a strategy which allows protein identifications to be applied to mixtures of proteins for which no genomic information is available. In particular, bacterial ribosomal proteins from a reference strain serve as the primary database to which MALDI-MS experimental data from ribosomal proteins arising from unsequenced strains is compared. An iteritive matching strategy allows one to readily identify the protein components and determine specific differences arising in ribosomal proteins from the various strains investigated (Figure 1).

This approach was developed and validated using ribosomes from the extreme thermophile *Thermus thermophilus*, chosen in part because of the available crystal structures of ribosomes from this organism [5,6,7] and the two available genome sequences ([8]; GenBank AE017221).

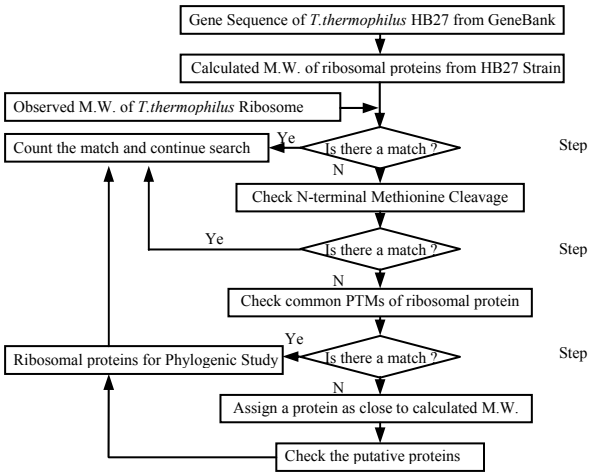

Figure 1 Flowchart outlining the procedure used to assign MALDI-MS data to proteins from strains not present in any protein or DNA database.

Ribosome Preparations. The 30S and 50S ribosomal subunits isolated through a 0-45% sucrose gradient. Sucrose gradients were formed by layering different concentrations of sucrose in dissociation buffer (10mM Tris-HCl at pH 7.6, 1 mM MgCl2, 60 mM NH4Cl, and 4 mM β-mercaptoethanol) followed by diffusion. Approximately 30 A_{260} units of intact 70S ribosomes were applied on the top of the gradient and the gradients were centrifuged in an SW28 rotor in a Beckman ultracentrifuge at 4°C for 17 hours at 19000 rpm. Fractions of 1.1-1.2 mL were collected manually and the absorbance at 260 nm was measured to determine the location of the 0S and the 30S subunits.

MALDI-MS Analysis. All MALDI-TOF MS experiments were done on a Bruker Reflex IV reflectron MALDI-TOF mass spectrometer (Bruker Daltonics, Billerica, MA) equipped with a nitrogen laser. Protein mass spectra were obtained in the positive-ion mode at an acceleration voltage of 20 kV, extraction plate voltage of 17.1 kV, lens voltage of 10.1 kV with 300 laser shots [4].

Data Analysis. The amino acid sequences of *T. thermophilus* HB27 and HB8 ribosomal proteins were obtained from GenBank (http://www.ncbi.nlm.nih.gov/genomes) with accession numbers AE017221 and AP008226 (chromosome). The theoretical molecular weights of these proteins were calculated using the SequenceEditor software provided by the MALDI manufacturer.

The approach developed to readily confirm assigned proteins and to distinguish possible sequence variations is provided in the flowchart of Figure 1. Theoretical molecular weights of ribosomal proteins from *T. thermophilus* HB27 were first calculated based on DNA-derived sequences. Experimental data was obtained from *T. thermophilus* HB8 using the previously described MALDI-MS approach (Figure 2) [4]. The mass spectral data was classified into four categories: 1) Proteins yielding identical molecular weights to DNA-derived sequences: 13 matches-24% of proteins identified, 2) Proteins yielding molecular weights corresponding to N-terminal methionine loss: 35 matches-65% of proteins identified, 3) Proteins yielding molecular weights consistent with conserved post-translational modifications, and 4)Proteins that cannot be assigned directly which may reflect differences at the primary sequence or post-translational level.

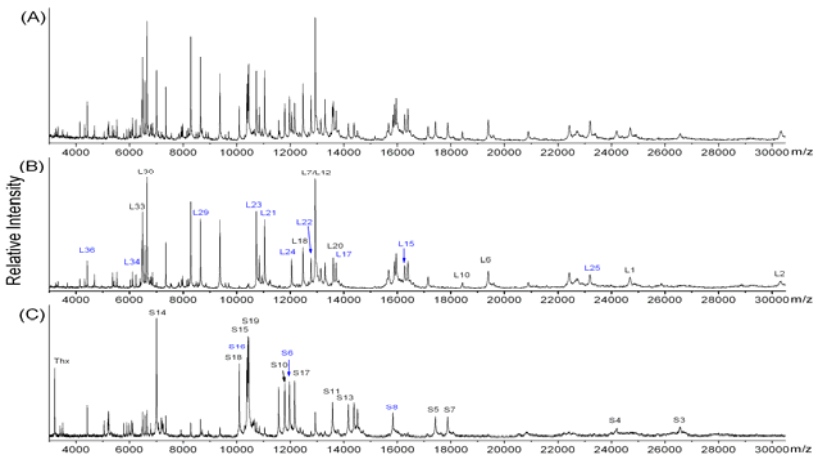

Figure 2 MALDI analysis of *T. thermophilus* HB 8 ribosomal proteins. (A) Intact 70S ribosome, (B) 50S subunit, (C) 30S subunit.

While the results obtained with strains HB8 and HB27 were encouraging, the effectiveness of this strategy is best demonstrated by its extension to ribosomal proteins isolated from a strain of *T. thermophilus* for which no genome (or protein) sequence exists. Thus, ribosomal proteins from the geographically and ecologically distant *T. thermophilus* IB21 strain were isolated and analyzed by MALDI-MS. Figure 3 shows a representative MALDI mass spectrum of ribosomal proteins from *T. thermophilus* IB21. A total of 51 out of 54 ribosomal proteins (94%) could be identified with only ribosomal proteins S3, S2, and S1 not being detected and assigned from MALDI-MS data. This procedure should prove generally applicable to other strains of *T. thermophilus* or other organisms for which proteins from a particular strain have not been sequenced.

This approach also allows one to examine the level of conservation among different strains of the same organism. Figure 4 illustrates the comparison of calculated protein molecular ions from HB27 gene information (Fig. 4a) to the experimental data obtained from HB8 (Fig. 4b) and IB21 (Fig. 4c) strains. As noted in Table 1, 35 ribosomal proteins were found to yield identical molecular weights from HB8, HB27 and IB21 strains.

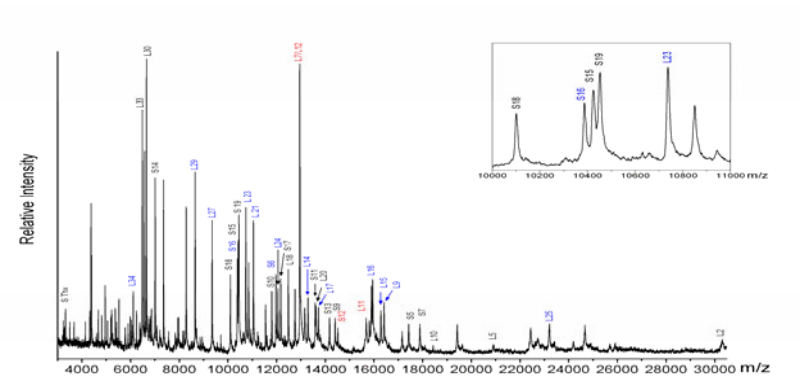

Figure 3 MALDI analysis of *Thermus thermophilus* IB 21 ribosomal proteins from intact 70S ribosomes.

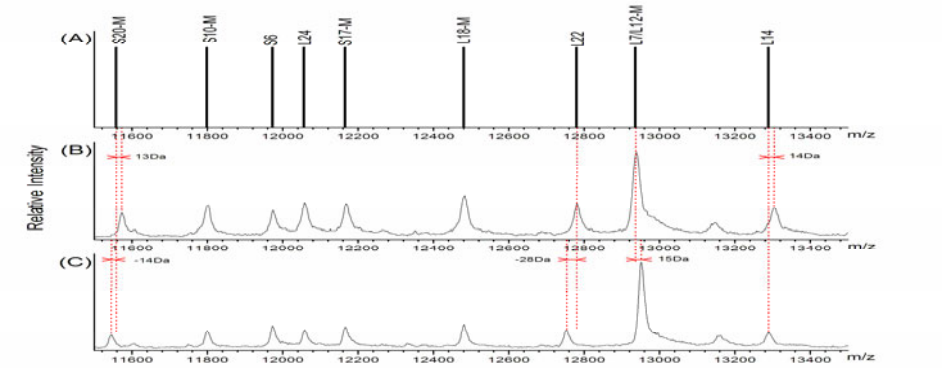

Figure 4 Typical experimental differences in ribosomal protein molecular weights between the *T. thermophilus* reference strain HB27 (calculated from annotated genome, A) and MALDI mass spectra obtained from strains (B) HB8 and (C) IB21.

Table 1 Results obtained from analysis of IB 21 strain using the developed matching strategy to identify ribosomal proteins. 131.2 Da is subtracted from a protein for N-terminal methionine cleavage. Methylation, acetylation, β-methylthiolation are considered as common post-translational modifications of ribosomal proteins from prokaryotic organisms.

	Comparison		
Experimental IB 21 Data Compared to:	**HB 27**	**HB 8**	**HB 27+HB 8**
# of hits in Step 1 (No modification)	**14**	**13**	**17**
# of hits in Step 2 (Consider Loss of Met)	**19**	**20**	**22**
# of hits in Step 3 (Consider PTMs)	**6**	**6**	**6**
# of unmatched proteins	**18**	**18**	**12**
Total	**57***	**57***	**57***

A straightforward procedure for assigning prokaryotic proteins based upon MALDI-MS data and DNA-derived protein sequences has been presented. This procedure allows currently available protein database information specific for a particular strain of an organism to be extended to strains for which no database information exists. As illustrated with *T. thermophilus*, more than 90% of the analyzed proteins can be assigned by this procedure. And while around 60% of the ribosomal proteins for the HB8 and IB21 strains were found to yield molecular weights consistent with the published protein sequences for the HB27 strain, 20% of the investigated proteins yielded molecular weights which differed among the three strains.

Although the MALDI-MS approach provides a rapid and sensitive route for extending protein assignments, the approach as described cannot determine the exact chemical constituents that lead to mass differences between measured and database values. The mass shifts arising from these differences are most likely due to differences in the primary amino acid sequence of the proteins or due to differential post-translational modifications. While the data generated by MALDI-MS cannot distinguish between these two possibilities nor provide information on the specific site of amino acid substitution or post-translational modification, these data do allow one to focus further efforts on only those proteins suspected of differing amongst the various strains. Further, this approach can allow one to readily identify proteins that could serve as potential biomarkers for establishing strain identity from unknown samples.

[1] Aebersold, R. and Mann, M., *Nature* **2003**, *422*, 198-207.
[2] Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M. and Yates, J.R., III, *Nat. Biotechnol.* **1999**, *17*, 676-682.

[3] Kelleher, N.L., Lin, H.Y., Valaskovic, G.A., Aaserud, D.J., Fridriksson, E.K. and McLafferty, F.W., *J. Am. Chem. Soc.* **1999**, *121*, 806-882.
[4] Suh, M.J. and Limbach, P.A., *Eur. J. Mass Spectrom.* **2004**, *10*, 89-99.
[5] Schuenzen, F., Tcolj, A,, Zaravich, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F. and Yonath, A. *Cell* **2000**, *102*, 615-623.
[6] Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Jr., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V., *Nature* **2000**, *407*, 327-339.
[7] Yusupov, M. M.; Yusupova, G. Z.; Baucom, A.; Lieberman, K.; Earnest, T. N.; Cate, J. H. D.; Noller, H. F. *Science* **2001**, *292*, 883-896.
[8] Henne, A.; Brüggemann, H.; Raasch, C.; Wiezer, A.; Hartsch, T.; Liesegang, H.; Johann, A.; Lienard, T.; Gohl, O.; Martinez-Arias, R.; Jacobi, C.; Starkuviene, V.; Schlenczeck, S.; Dencker, S.; Huber, R.; Klenk, H.-P.; Kramer, W.; Merkl, R.; Gottschalk, G.; Fritz, H.-J. *Nature Biotechnology* **2004**, *22*, 547 - 553.
[9] Suh, M.J., Hamburg, D.M., Gregory, S., Dahlberg, A. and Limbach, P.A. *Proteomics* **2005**, in press.