Proteome Analyses Using Accurate Mass and Elution Time Peptide Tags with Capillary LC Time-of-Flight Mass Spectrometry

Eric F. Strittmatter, P. Lee Ferguson, Keqi Tang, and Richard D. Smith
Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA

We describe the application of capillary liquid chromatography (LC) time-of-flight (TOF) mass spectrometric instrumentation for the rapid characterization of microbial proteomes. Previously (Lipton et al., *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*, 11049) the peptides from a series of growth conditions of *Deinococcus radiodurans* have been characterized using capillary LC MS/MS and accurate mass measurements which are captured as an accurate mass and time (AMT) tag database. Using this AMT tag database, detected peptides can be assigned using measurements obtained on a TOF due to the additional use of elution time data as a constraint. When peptide matches are obtained using AMT tags (i.e., using both constraints) unique matches of a mass spectral peak occurs 88% of the time. Not only are AMT tag matches unique in most cases, the coverage of the proteome is high; ~3500 unique peptide AMT tags are found on average per capillary LC run. From the results of the AMT tag database search, ~900 ORFs detected using LC-TOFMS, with ~500 ORFs covered by at least two AMT tags. These results indicate that AMT database searches with modest mass and elution time criteria can provide proteomic information for approximately one thousand proteins in a single run of <3 h. The advantage of this method over using MS/MS based techniques is the large number of identifications that occur in a single experiment as well as the basis for improved quantitation. For MS/MS experiments, the number of peptide identifications is severely restricted because of the time required to dissociate the peptides individually. These results demonstrate the utility of the AMT tag approach using capillary LC-TOF MS instruments, and also show that AMT tags developed using other instrumentation can be effectively utilized. (J Am Soc Mass Spectrom 2003, 14, 980–991) © 2003 American Society for Mass Spectrometry

The two dominant separation methods used in proteomics are 2-dimensional gel electrophoresis [1] and, increasingly, liquid chromatography (LC) [2]. The first use of LC to obtain sequence information from proteins was by Fredrick Sanger in the 1940s and LC separation methods have experienced rapidly growing use since online LC-MS analysis based upon electrospray ionization (ESI) was demonstrated in the early 1990s [3–7]. Although most early efforts were used for analysis of peptides from digests of nominally isolated proteins, of most interest currently are methods that can be used to analyze as many as hundreds of thousands of separate species from a single organism or sample with high reproducibility from run-to-run. State-of-the-art technologies for proteomics continually demand higher performance, e.g., high pressure chromatography [8] or use of multiple LC separation dimensions [9]. Our laboratory has focused on an approach using LC separations with highly accurate Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR) mass spectrometry measure to provide more sensitive and comprehensive proteome analyses based on the concept of accurate mass and time (AMT) tags [10].

Because of the complex nature of proteomics, most methodologies use several separation and mass spectrometric steps, including a MS/MS analysis step followed by a algorithmic comparison (via SEQUEST [11] or MASCOT [12]) of this data against a database of proteins derived from genomic sequencing. However, the time required to obtain an adequate MS/MS spectrum in addition to the processing requirements for these spectra has increasingly made this step a bottleneck in proteome studies. Consequently, to harvest the

Published online July 21, 2003

Address reprint requests to Dr. R. D. Smith, Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, MSIN K8-98, P.O. Box 999, Richland, WA 99352, USA. E-mail: rd_smith@pnl.gov

1044-0305/03/$30.00
doi:10.1016/S1044-0305(03)00146-6

Received February 4, 2003
Revised March 3, 2003
Accepted March 4, 2003

maximum amount of information from a capillary LC dataset, we have focused upon greater utilization of accurate mass and retention time information [8, 13].

Approaches using accurate mass measurements of peptide masses rather than for acquisition of MS/MS information are particularly useful for quantitative analysis. Because these methods often involve the measurement of peptides with a natural isotope and a stable isotope enriched tag, accurate mass measurements are useful in identifying characteristic mass differences for these species. Both time-of-flight (TOF) [14] and FTICR [15] methods have been used to obtain protein expression information using stable isotope incorporation or isotope affinity coded tags. With TOF, accuracies of <10 ppm are obtainable given a signal sufficiently large to provide reasonable peak shape [16], and the increasing resolution of this technology ($m/\Delta m$ now exceeds 10,000) should allow for measurements of increasingly complex systems. Key to the increased accuracy of TOF over the last decade are improvements in TOF design including orthogonal ion injection, delayed-extraction [17], ion mirrors [18], and ion collisional cooling [19, 20].

In the AMT tag approach [21] to proteomics involves the combined use of accurate mass measurement and LC elution time information to increase peptide identification confidence [22]. The AMT tag approach is well suited to take advantage of the improving accuracies provided by TOF instruments. In previous work, FTICR measurements have been used to validate identifications from ion-trap MS/MS experiments, effectively improving greatly the confidence of identification without elimination of many "true positives." Under optimal conditions, FTICR measurements can achieve 100 ppb to 1 ppm accuracy for peptides [23–25], although accuracies can fluctuate more under the effect of large ion population variations. With the AMT tag approach, it is clear that improvement of the accuracy of mass measurements enables more comprehensive proteome characterization, especially for more complex eukaryotic organisms [21]. However, as shown by our initial studies, there is no "magic" level of mass accuracy that either allows or disallows the use of AMT tags [10]. However, the better the MMA achieved, the more comprehensive the proteome coverage that will be achieved (i.e., the greater the fraction of peptide masses that will be unique within the context of the proteome being studied).

In this study, we explore using the AMT tag procedure with 10 ppm mass accuracy obtained with ESI-orthogonal TOF instrument and using a set of AMT tags developed using capillary LC-FTICR measurements for *Deinococcus radiodurans* [26, 27], an organism being studied because of its application to bioremediation. A key aspect of this work is the augmentation of accurate mass measurements with the known peptide elution times to improve the overall specificity of the analysis [13].

Experimental

Sample Preparation

The organism *Deinococcus radiodurans* (strain R1) was grown in minimal media (MM) and harvested either at midlog (MLP) or poststationary phase (PSP) using centrifugation. Prior to lysis, the cells were resuspended and washed three times with ammonium carbonate and an EDTA solution. The cells were lysed using bead beating with zirconium beads (0.1 mm) at 5000 rpms. Prior to LC/MS analysis, the protein samples were denatured and reduced with the addition of guandine-HCl, urea, and dithiotheitol and boiled for three min. Digestion of proteins was done by diluting the sample ten-fold with a 50 mM ammonium carbonate and 30 μg of Promega sequencing grade trypsin. A 1.0 μL of $CaCl_2$ (1.0 M) was added to the sample and the digest was carried out overnight at a temperature of 37 °C. The digested sample was ultacentrifuged and dialyzed overnight with a 500 MW cellulose ester membrane. Additional details on the sample handling are given elsewhere [26].

Reverse Phase Chromatography

Packed capillary LC separations were performed at constant pressure using a (50 μm i.d., 360 μm o.d., 82 cm column) packed with 5 μm Jupiter (Torrance, CA) particles (300 Å pore size). The composition of the mobile phase was varied during the LC separation by utilizing a solvent gradient with two solvents **A**, **B**. Solvent **A** is 0.2% acetic acid, 0.1%TFA, and water and Solvent **B** is 90% acetonitrile, 10% water, and 0.1% TFA. The gradient was 100% **A** at the start of the LC run and was mixed using an exponential gradient over 4 h to a final mobile phase composition of 90% **B**. The solvents were delivered at a constant pressure of 9800 psi using two Isco (Lincoln, NE) model 100 DM pumps. The flow rate varied during the run but is approximately 200 nl/min. Under the conditions here, the LC separations were typically 140 to 160 min in duration.

Mass Spectrometry

For most studies, a Micromass (Manchester, UK) Q-TOF Ultima orthogonal time-of-flight mass spectrometer was used. Detection events are acquired at 4 GHz rate. For all measurements, the spectrometer was operated in V mode with typical resolution of $\sim 10^4$. The spectrum integration time was 2.1 s and the interscan time was 0.1 s. The capillary LC eluent from the LC separation was electrosprayed from a 150 μm o.d., 30 μm i.d. fused silica capillary pulled to a narrow tip. All analyses were performed using positive mode ESI using the Micromass nanospray ion source (set at 120 °C). The mass range acquired was between m/z 300 to 2000, resulting in a pusher rate of 16 kHz.

Where additional verification using tandem mass

spectrometry is needed, previously existing MS/MS spectra acquired using an Applied Biosystems (Foster City, CA) Qstar (Q-TOF) mass spectrometer were used for sequence verification. Similar ESI conditions and integration times as for the Micromass Ultima were used. Because of the quality of these MS/MS data, combined with the similar performance characteristics two TOF instruments, it was deemed unnecessary to reanalyze these peptides with MS/MS using the Ultima instrument.

Centroid and Calibration

Data were acquired in centroid mode with the MassLynx software (v 3.5) where the top 80% of the peak distribution was used to determine the centroid. Other parameters used are as follows: centroid threshold was set to 5 counts and the minimum points were set to 3. A deadtime correction employed within MassLynx was used for all analyses to minimize mass measurement errors for high intensity peaks (an Np multiplier setting of 0.7 was used). While these initial processing steps were performed in Mass Lynx, a significant portion of the subsequent data analysis was performed using external tools developed in our lab. Thus, data files containing the mass spectral data for the entire LC run were converted from Micromass to netCDF format, making the data accessible to other data processing platforms.

External calibration was performed using a PEG mixture and a fifth order nonlinear calibration for fitting the known and observed *m/z* values. Although this process gave good mass measurement accuracy performance, temporal drift of the spectrometer often required additional *m/z* calibration. Instead of introducing additional calibrant ions, an additional calibration was performed by examining the mass error for two peptides, GRPQPTPVVHTTTTEPR and TAPGEQGT TLTTR, and adjusting the calibration formula if needed. These peptides were selected arbitrarily and there is no distinguishing feature in using these peptides other than that they are abundant peptides which are routinely detected in *D. radiodurans* lysates. These peptides were also verified with TOF-MS/MS experiments. Using multiple charge states determined for these peptides provides a broad mass range for the calibration. Although the two calibrant peptides only appeared in a small subset of spectra from the entire run, the recalibration was applied globally using the surface calibration expression described elsewhere [28].

Normalized Elution Time

In order to facilitate the comparison of LC retention times for peptides carried out over a series of capillary LC analyses, we have previously described the use of adjusted retention times or normalized elution times (NET) [26]. In order to determine a NET value for a particular monoisotopic mass (M_{mi}) obtained during an LC run, the following eq is used:

$$\mathrm{NET}(M_{mi.}) = slope \times \mathrm{elution\ time}_{Mmi.} + offset \quad (1)$$

The *slope* and *offset* are fit using a least squares procedure with each acquired data set to minimize the error of the NET values of the peptide in the AMT database with the experimentally measured values. The NET value is constrained to be a value from 0 to 1. The normalized elution procedures have been implemented in LaVD2G as part of our AMT tag database. In calculating a single NET value for peptides eluting in a range of spectra, a collective grouping of peptides into universal mass classes (or UMCs) is done in the LaV2DG program. In a UMC, neutral masses of presumed peptide masses are grouped together if they elute in consecutive scans and are within a selected mass accuracy so that adjacent (or with limited gaps) spectra are grouped and attributed to a single eluting species. In cases where it is more convenient to use a single NET value instead of several values obtained over the elution profile of the peptide, UMCs provide a useful method for measuring such values.

AMT Tag Database

The structure and data input of the accurate mass and time tag database previously developed using LC-FTICR instrumentation has been previously described [22, 26]. The AMT tag database operationally utilizes a Microsoft SQL server and contains information describing the peptide, ORF (open reading frame) reference, mass spectral, separation, and other tracking and sample specific data. MS/MS spectra from *D. radiodurans* cultured under a variety of different growth conditions were obtained using a ThermoFinnigan (San Jose, CA) LCQ ion trap mass spectrometer. Both full tryptic, partial and non-trypic peptides were used to build the database of peptides according to the SEQUEST criteria given by Washburn et al. [9], although the final AMT tag data base consisted of mainly tryptic peptides. Because of elution time variability of capillary LC separations, the NET values retained in the database were calculated using a procedure which minimizes the variance obtained for all AMT tag peptides based upon a genetic algorithm. The typical average absolute deviation of the NET value in the database to a measurement from a single LC-MS analysis in the work was 0.027 NET units. Our studies have also indicated that much greater accuracy for peptide NET values can be achieved, primarily by the better control of separation variables (such as temperature and sample matrix) and by the incorporation of elution time internal calibrants. Additionally, experience indicates that low abundance peptides with more ideal peak shapes are preferable for NET calibration. In this work, a tolerance of 0.05 (unless otherwise stated) is used to account for the experimental variance of all AMT tags in our database.

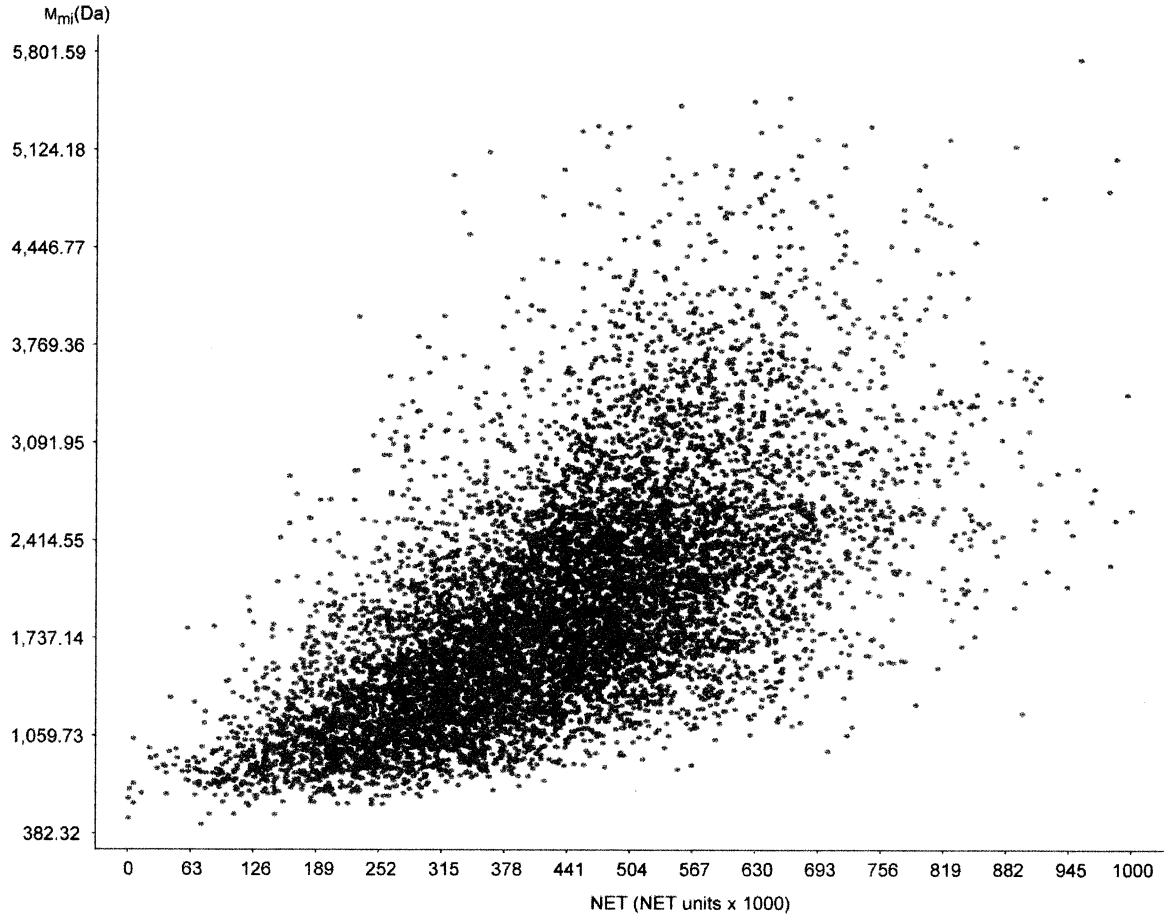

Figure 1. A two-dimensional display depicting the distribution of the 12,600 AMT tag masses and NETs determined from previous LCQ and FTICR experiments.

Results and Discussion

Existing AMT Tags for Deinococcus radiodurans

The genome of *Deinococcus radiodurans* has been completely sequenced, and the proteins expressed under various environmental conditions have been studied. A set of peptide AMT tags has been previously developed, where peptides are identified with both ion trap MS/MS data and accurate mass measurement from FTICR mass measurement [26]. The AMT tag database for *D. radiodurans* consists of peptides that have been previously detected, verified and are distinctive based upon their mass and NET values. Previously we have reported 6997 AMT tags that provide coverage of 1910 ORFs (61% of the predicted ORFs for *D. radiodurans*). Our present database contains ~12,600 AMT tag peptides resulting from 1067 LC-MS/MS analysis and 1084 FT-ICR experiments. This set is approximately twelve times smaller than the 153,000 peptides (by including peptides having one missed tryptic cleavage site) from a tryptic digest of the entire annotated genome.

To facilitate comparison of retention times for peptides identified from different analyses, raw retention times are converted to NETs using eq 1. Using the normalized elution values, it is possible to account for run-to-run variations that occur in the separations. For example, if two LC columns of slightly different length are used among several separations, the NET adjustment will compensate for this difference via the slope in eq 1. In the AMT tag approach, normalized elution time (NET) information, built from a series of LC-MS/MS analyses, is utilized and provides an additional constraint to increase the specificity of peptide identifications from the AMT tag database. The distribution of peptide mass and NET values from the AMT tag database is shown in Figure 1. The NET values range from 0 to 1 and the mass range extends to a maximum of 6000 Da.

Mass Accuracy and Elution Time Information

With TOF mass spectrometry under ideal operating conditions and using internal calibration, accuracies of about 5 ppm are presently achievable. Because an internal mass calibrant was not introduced into each spectrum, recalibration was done using a small number of analyte peptides and applied globally to all the spectra in the LC analysis. Typical mass errors for seven peptides from the abundant *D. radiodurans* protein elongation factor TU were determined and are shown in

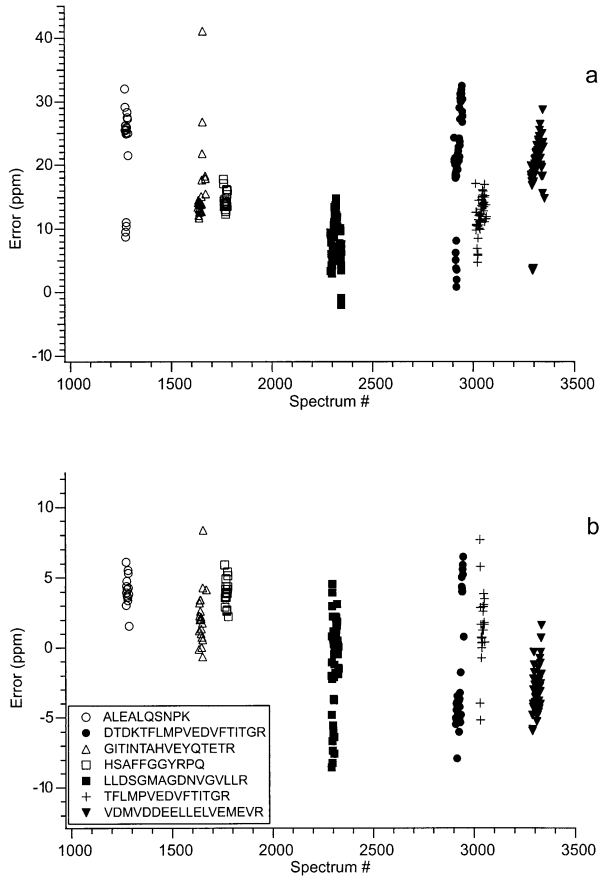

Figure 2. Mass Errors determined for seven peptides from the abundant *D. radiodurans* protein elongation factor TU. Unadjusted (**a**), and adjusted (**b**), *m/z* values were determined by calculating the deviations from the actual mass for these peptides, in ppm. The results in (**a**) represent accuracies obtained from external calibrations.

Figure 2. For these seven peptides, MS/MS data was utilized to obtain additional sequence confirmation. With the abundance of many of these peptides spanning two orders of magnitude, the mass measurement errors would reflect both detector dead-time effects and low peak statistics. From the data in Figure 2 it can be seen that most of the lower molecular weight peptides have spreads in error that are less than 5 ppm. All peptide measurements were within a 10 ppm error tolerance. The errors for the same peptides are significantly larger without the recalibration procedure, where errors otherwise ranged from +30 to −10 ppm.

One question in using capillary LC MS/MS experiments and FT-ICR (or reflectron TOF) spectrometers is whether the accurate mass measurement capability of the mass spectrometer is sufficient for unambiguous identification, or are both accurate mass and retention information needed. There is no single answer to this question; the needed information depends upon the complexity of the system, the specific peptides, and other factors such as sensitivity, the presence of contaminants, etc. In Figure 3, peptides detected from a LC analysis of *D. radiodurans* are shown in a 2-dimensional plot where the monoisotopic ion mass, M_{mi}, is plotted on the vertical axis and NET on the horizontal axis. Two nominally isobaric peptides are highlighted. These peptides have equal mass values within 4 ppm, yet are clearly distinguishable on the separation axis. Also shown are boxes denoting the expected location of any AMT tags using the M_{mi}. and the NET information from the AMT tag database. The width of the box corresponds to a deviation of ±0.027 NET units; smaller than the 0.05 tolerance applied in this work. Because the two peaks at *m/z* 1666.8 are separated well outside the expected error range, they can both be uniquely identified with ease when NET information is considered. This result highlights the desirability of using retention time information with accurate mass identification, where even for a simple prokaryotic organism both pieces of information are needed for unambiguous peptide identification. For eukaryotic organisms, where unique identification at 1 ppm is generally insufficient, the use of retention time information will be even more important.

The utility of the previously developed AMT tag database in making identifications is evident from Figure 4, which shows one spectrum obtained from an LC-MS analysis of a global *D. radiodurans* tryptic digest. Figure 4 shows possible peptide matches to one peak based upon peptides predicted from the annotated genome (i.e., a match is made if the difference between the experimental mass to that calculated from an in silico tryptic digest of *D. radiodurans* proteins is within a certain tolerance). For the intense peak in the spectrum (at *m/z* 699.901) a search with a tolerance of 10 ppm results in a match to 20 possible peptides. Similar results are obtained if most other peaks in the spectrum are searched against the entire set of possible *D. radiodurans* peptides. In contrast, this peak could be assigned to a specific peptide using NET information, as shown in Figure 5. All matches in Figure 5 were identified based on an agreement between the measured mass and predicted mass of 10 ppm and 0.05 between the measured NET (eq 1) for a peptide in the database. For 17 of the peaks one unique peptide can be matched per peak, however two other peaks were each attributable to two peptides. The 17 unique matches can serve as high confidence markers of the parent protein and can also be used for quantitative protein measurements. The two peaks each matching two peptides may still provide useful information (e.g., based upon detection of other peptides form their parent ORFs), but at present this information is not being utilized.

Using Ambiguous Mass Values to Obtain AMT Tag Identifications

Because of the complexity of the LC-MS data and proteome to which it is matched, ambiguities in peptide assignments can occur in many different ways. In Figure 6, a simplified LC-MS analysis is shown, along

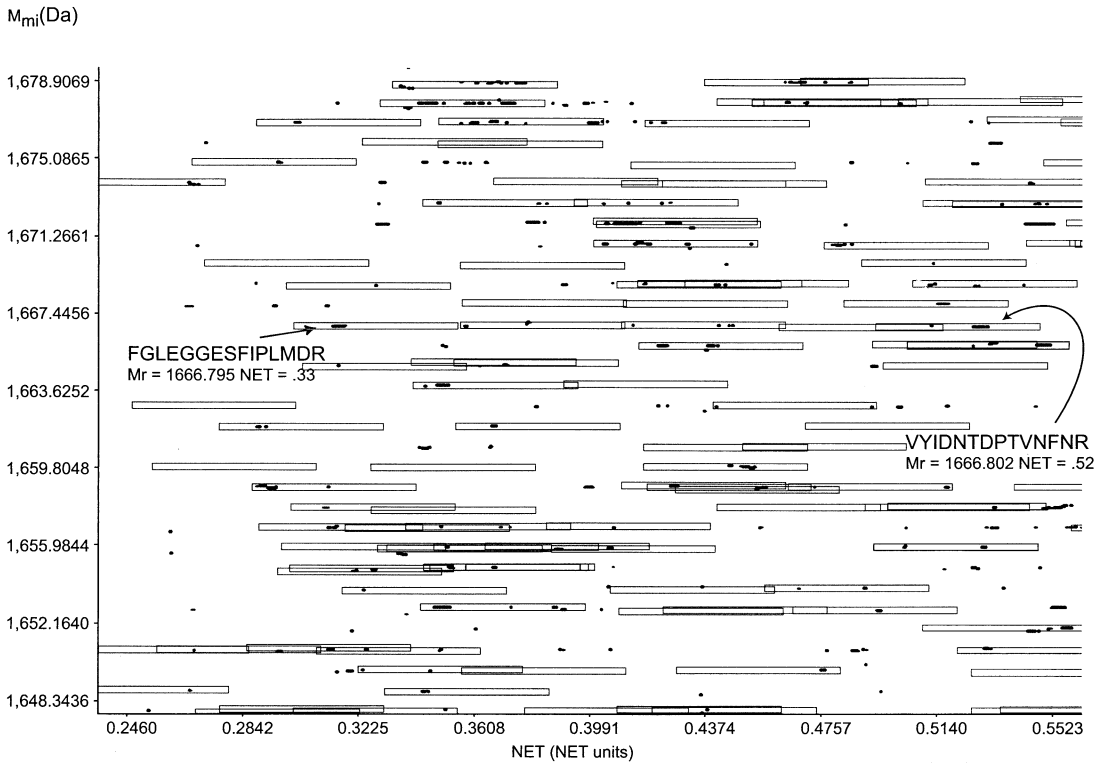

Figure 3. A small region of a two dimensional display obtained from a TOF-MS analysis of *D. radiodurans*. Each spot corresponds to a peptide M_{mi} measurement obtained during the LC separation (0.01 NET is ~54 s). The boxes depict where peptides are expected to elute based on the NET values for peptides in the AMT tag database and a ±0.027 range in this NET value. The height of the boxes, exaggerated to make them visually perceptable, correspond to a ±100 ppm accuracy range.

with information from several stages of data analysis. From an LC analysis, one peptide ID will typically occur in a set of several spectra during the separation. In Figure 6, the peptide NASGEVIALAK occurs in three consecutive spectra and is only counted once. Each peptide has a chromatographic peak width that must be accounted for by the software making the AMT tag identifications. Another level of ambiguity occurs when two peptides can be assigned to one mass value and such a case occurs in Figure 6 for FTQAPELGDAIEPGK and IEDLGDRAVFPGR. Although the ion trap-MS/MS identifications were found uniquely for these latter two peptides when the AMT tag database was constructed, the nearly similar mass and NET values for these peptides makes them difficult to distinguish using the present mass accuracy and elution time normalization routines from LC-MS data only. Thus, in Figure 6, FTQAPELGDAIEPGK and IEDLGDRAVFPGR are non-unique matches, while the peptide NASGEVIALAK is unique (and designated an AMT tag). For a peptide to be an AMT tag in the present context, it must correspond to a single unique peptide and point to a single ORF.

One analysis of the *D. radiodurans* proteome grown under post-stationary phase (PSP) conditions, and two under midlog phase (MLP) conditions were performed (with no difference in sample preparation between MLP-1 and MLP-2). These three LC-TOF MS runs were processed to determine the monoisotopic peptide ion masses for each spectrum. Initially, a set of *m/z* values were obtained from each spectrum and the charge states are determined from their respective ^{13}C-isotopic patterns. Because a theoretical ^{13}C-isotope pattern for a peptide can be adequately predicted using an averaging based model [29], most "non-peptide-like" analytes can be eliminated based on the lack of agreement between the observed and predicted isotopic pattern, given sufficient S/N. Once the list is filtered so as to contain only likely peptide peaks, the monoisotopic ion mass is generated (based on the assumption that the peptide detected is protonated) from the monoisotopic *m/z* and charge state, and ^{13}C enriched (isotopic) peaks are eliminated from further consideration. This process is carried out iteratively for all spectra in the LC analysis.

Two questions to be considered are: (1) What is the ability of LC-TOF-MS to resolve a complex mixture of analytes, and (2) how effectively can unique identifications be made against a database of AMT tag peptide masses and NET values developed using other instrumentation. The total number of monoisotopic peptide masses (M_{mi}) obtained from three LC runs is listed in Table 1. The total number of peptide masses measured from an analysis gives an indication of how well an LC-MS spectrum can resolve a complex mixture. It is difficult to establish a maximum for the number of M_{mi}

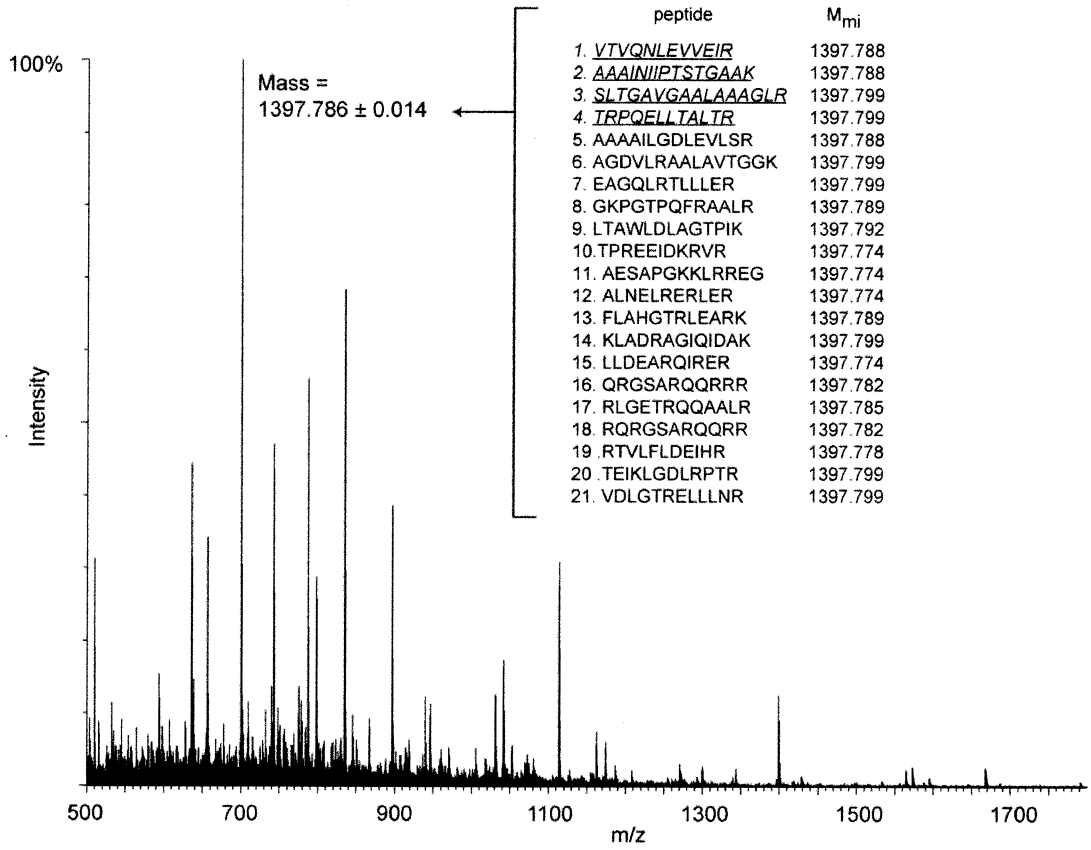

Figure 4. A spectrum from a capillary LC analysis of *D. radiodurans* where one peak is identified using the annotated genome of *D. radiodurans* with the only constraint that a mass agreement be within 10 ppm. Only four of the peptides listed are part of the AMT tag database and these are underlined and italicized.

values that could potentially be determined from a complex mixture of proteins from a cellular lysate, since the concentrations of the proteins vary considerably and the actual number detectable is unknown. However, the number of peptide M_{mi} values can be used in a comparative sense to other methods and, clearly, the larger this value the deeper one can delve into the proteome for an organism.

The number of mass classes, defined as a group of similar peptide masses that are observed in consecutive spectra were also determined. Because many peptides are observed to elute over several spectra (LLDSGMAGDNVGVLLR shown in Figure 1, was over 53 spectra), these "unique mass classes" (UMCs) closely correspond to the number of distinct peptides determined in an LC-MS analysis. The number of UMCs is a better indicator of the complexity of the *D. radiodurans* sample since the number peptide M_{mi} values can be inflated by a small number of highly abundant peptides eluting over an extended period.

All M_{mi} values were searched against the AMT tag database (based upon 10 ppm and 0.05 NET agreement) and matches were obtained for some of them. The total number of M_{mi} that were matched to an AMT tag peptide ranged from 26,000 to 36,000 for the three analyses in Table 1 (i.e., approximately 20% of the total number of M_{mi} values). When grouped by UMCs, these identifications covered approximately 3500 AMT tag peptides, when both unique and non-unique matches are considered. This number is especially large when compared to a single LC-MS/MS run performed on a quadruple ion trap, where 300–500 identifications typically occur. Thus, the resolution and accuracy of TOF MS instrumentation provides the ability to measure an order of magnitude higher peptides than a comparable MS/MS ion trap analysis.

The number of AMT tags uniquely assignable in the present study from the three *D. radiodurans* analysis was also determined (Table 2). From the average of the three samples, unique AMT tag assignments were obtained for ~88% of all peptides that were assigned by the AMT tag database search. The number of UMC peptides matched to two or more peptides are also given in the table. Although these non-unique peptide identifications are of somewhat lesser value than the unique AMT tag identifications, they still provide some information on the identity of the proteins contained in the sample. Dual matches make up the majority of the non-unique identifications.

In this work, the compatibility of accurate TOF mass measurements of peptides using an AMT tag database built up from high field FTICR measurements has been

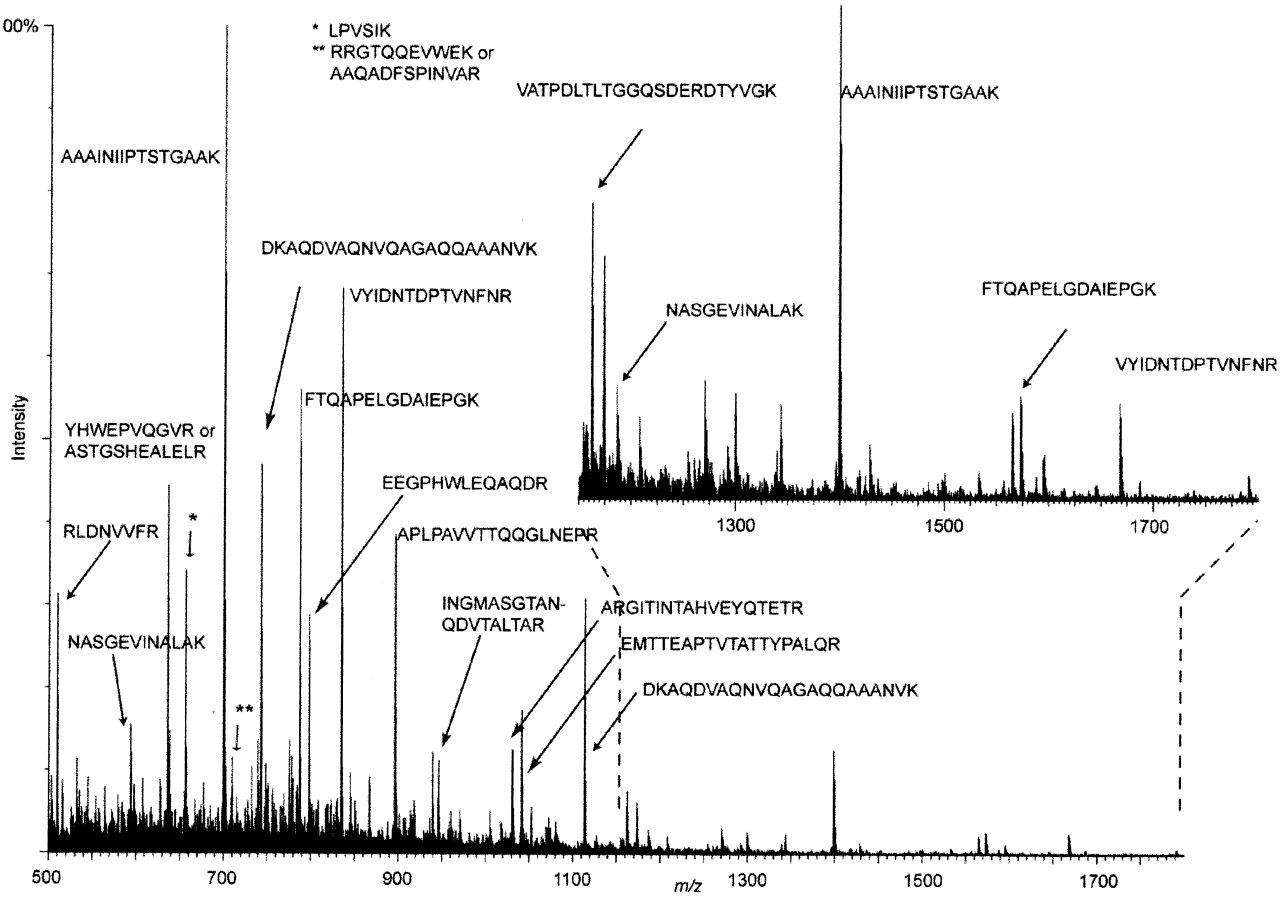

Figure 5. Same mass spectrum as shown in Figure 4, except that identifications are performed using the AMT tag database (within 10 ppm mass and 0.05 NET agreement). In most cases, unique identification of a peak can be made by the search of the AMT tag database. In the case where more than one peak is matched to the database, both matches are shown and separated by an "or", demonstrating that the search cannot distinguish between the two choices.

demonstrated. In these measurements, very similar chromatographic techniques have been applied in both the previous FTICR measurements and the present TOF analyses, and consequently the use of retention time criteria (0.05 NET in this work) should be similar. However, the mass accuracies obtained by TOF differ from those of FTICR measurements, and a systematic study of the impact of different ppm tolerances on the number of peptide identifications was performed. Attention was given to both the crude number of identifications and the fraction of these that are unique and useful AMT tags in LC-TOF analyses. Table 3 shows the number of peptide identifications using a range of mass accuracies from 1 to 12.5 ppm and a NET constraint of ± 0.05. Using 1 ppm accuracy, nearly all identified peptides are effective AMT tags. The number of identified peptides appears to plateau when tolerances of >10 ppm are used and at this level a considerable fraction of the peptides are not unique. The data in Table 3 suggest that an accuracy of between 5 and 10 ppm is optimum in obtaining a large number of matches to the AMT database with limited trade-off in obtaining an excessive number of non-unique matches. This observation is completely consistent with our determination of the errors associated with the mass measurements (i.e., <10 ppm).

A similar study was performed by fixing the mass accuracy criteria at (10 ppm) and varying the NET tolerance criteria from 0.03 to 0.07. In contrast to what was seen for mass accuracy, the number of matches increases linearly with increasing NET value due to the different nature of NET values. However, the number of non-unique identifications become a large fraction of the total number of identification when NET values exceeding 0.05 are used. Similar results were found for the other LC-TOF analyses.

ORF Coverage

The number of predicted proteins (or ORFs) based upon at least one detected peptide is given for the three LC analyses in Table 4. The ~900 ORFs detected based upon at least one AMT tag, corresponds to ~30% of the 3116 proteins predicted for *D. radiodurans*. From these results, it is evident that a large fraction of the proteome of *D. radiodurans* can be analyzed in a single experiment.

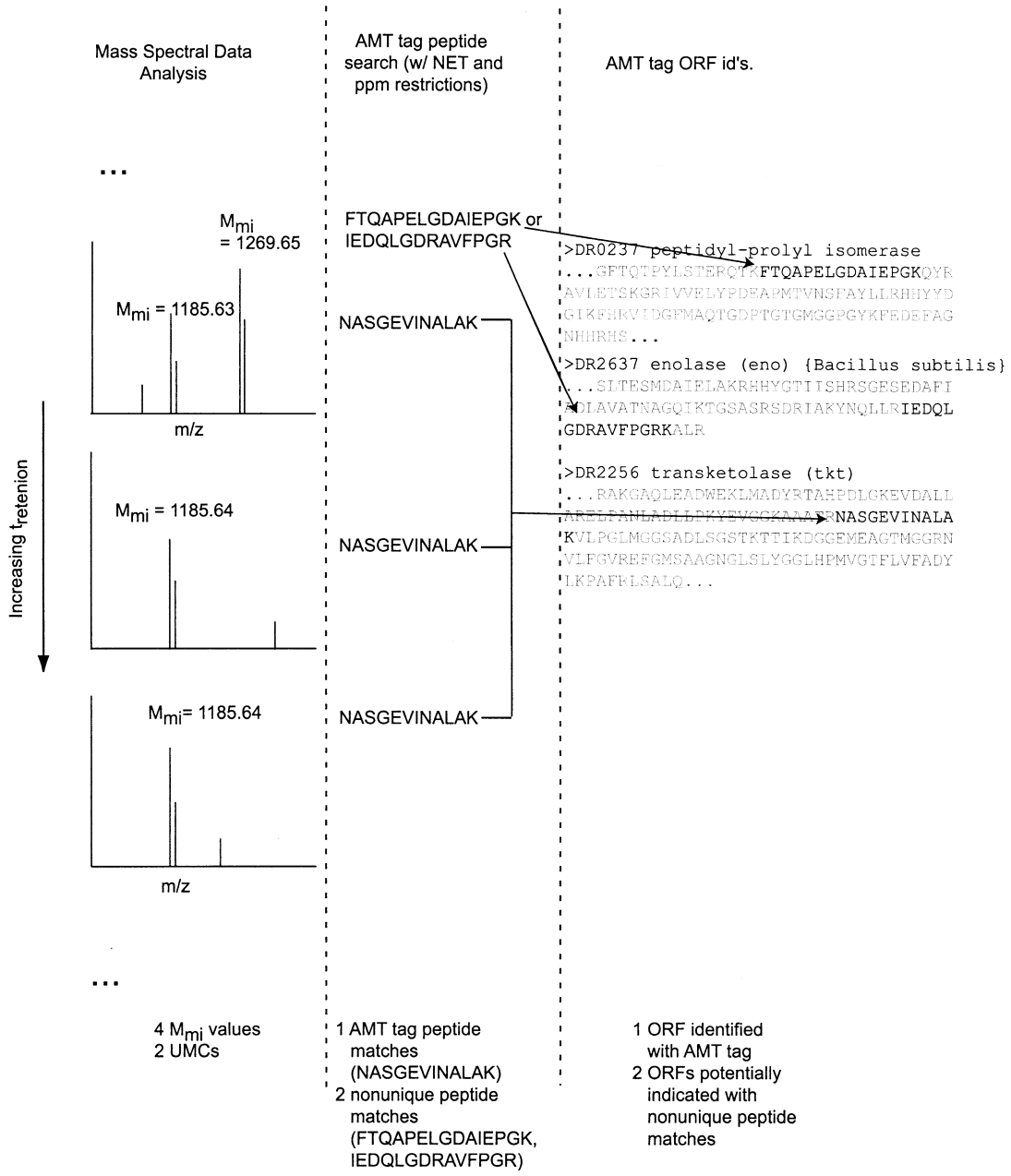

Figure 6. A small set of spectra from a hypothetical LC-MS analysis showing information from the mass spectra and how it refers to protein identifications. Initially, monoisotopic mass values (M_{mi}) are obtained from the thousands of spectra obtained from an LC-MS analysis. With knowledge of the NET values derived from retention times, peptide and AMT tag identifications can then be made. To efficiently perform AMT tag identifications it is important to deal with the multiplicity of M_{mi} values per AMT tag and distinguish non-unique peptide matches from unique AMT tags. Once AMT tags are established they can be used to obtain qualitative or quantitative information on proteins.

As shown in Table 3, changes in both the assumed mass and NET accuracy impact results, it is clear that the ~900 ORFs identified include "false positives". To increase the confidence of identifications one can either increase the specificity of the separation (e.g., NET value) or the mass measurements, or alternatively utilize multiple AMT tags for each identification. To see how such a more restrictive criteria affects the number of identified ORFs, an additional search of data against the AMT tag data set was done with a constraint of having at least two AMT tags for ORF identification. In this case, the number of identified ORFs decreases to ~550, but clearly constitutes a set of proteins identified with much greater confidence. However, we believe that some of the decrease is most likely due to the effect of protein concentration on the number of detectable peptide tags rather than the unreliable ORF identification. Consistent with this, the number of identified

Table 1. Number of peptide neutral masses and AMT tag peptide hits for three *D. radiodurans* whole proteome tryptic digest analyses

Sample[a]	Total peptides[b]	UMCs[c]	Peptide matches
MM, MLP(1)	209,989	62,301	3561
MM, PSP	154,942	48,727	3592
MM, MLP(2)[d]	138,263	44,746	3431

[a]MM = Minimal media, MLP = mid-log phase, PSP = post stationary phase.
[b]The sum of all distinctive peptide masses over all spectra.
[c]Unique mass classes based upon a grouping of all detected masses that correspond to apparent elution of a distinguishable species.
[d]No TDC threshold used for this analysis.

ORFs that are common between all three data sets is 788, indicating that a large fraction of the ORF identifications are occurring reproducibly.

Another test of reproducibility is to tally the number of AMT tags determined for each ORF and compare them among the LC-TOF analyses. In Figure 7, the 70 ORFs with the most AMT tags are shown along with the number of tags identified from each of the three analyses. In general the deviation between the different runs is about 2 AMT tags for the ORFs displayed in Figure 7, while for ORFs that have five or less tags the deviation is in general on the level of a single AMT tag. These results indicate that protein expression (or quantitation experiments) can be performed on a mixture of hundreds up to one thousand proteins in one LC-TOF MS analysis in a reliable fashion.

Matches Against a S. oneidensis Database

A database for the prokaryoitc organism *Shewenella oneidensis* has been constructed using the LCQ and FT-ICR results in a similar way as for *D. Radiodurans*. This database was used to perform AMT peptide searches with *D. radiodurans* lysate LC-TOF analysis to determine the number of false positive matches that could occur with the separation and mass accuracy conditions used in this work. Using the *S. oneidensis* that contains 7760 AMT peptides, 547 AMT tags were determined using *D. radiodurans* lysate LC-TOF analysis MM, MLP (run 1), and a 10 ppm, 0.05 NET tolerence. The AMT tags are contained in 312 ORFs, and yet if a stricter ORF identification criterion of two tags per ORF is required, the number of identified ORFs is reduced to 102. Three AMT tags have the same peptide composition between the two *D. radiodurans* and *S. oneidensis* matches, indicating that sequence similarity between these two organisms does not by itself explain the number of matched ORFs. Although the number of ID ORFs with the false positive database is significantly smaller (by a factor of five and a half for the ORF ID with two AMT tags), these results indicate that using 10 ppm and 0.05 NET tolerence permits overlap between different peptides from different organisms. Improvements in mass calibration and elution time algorithms would aid in reducing false positives even further.

Table 2. Breakdown of the number of instances where one or multiple peptide matches (n = 1, 2, 3, ≥4) to a single UMC are found in three LC analyses

n	1	2	3	≥4
MM, MLP(1)[a]	3037	344	46	4
MM, PSP	3100	439	43	10
MM, MLP(2)[b]	3154	360	45	2
Average	87.8%	10.8%	1.2%	0.2%

[a]MM = Minimal media, MLP = mid-log phase, PSP = post stationary phase.
[b]No TDC threshold used for this analysis.

Table 3. Total number of AMT tag peptide matches, including non unique matches, and AMT tags obtained for MM, MLP (run 1: see Table 2) using different mass accuracies

	Matches	AMT tags
ppm[a]		
1	1357	1286
2.5	2053	1937
5	2769	2553
7.5	3270	2946
10	3561	3154
12.5	3988	3431
NET[b]		
0.03	2561	2342
0.04	3100	2792
0.05	3561	3154
0.06	4216	3667
0.07	4885	4178

[a]using a 0.05 NET criteria.
[b]using a 10 ppm mass accuracy criteria.

Conclusions

The peptide identification approach developed in this work enables high-throughput identification of peptides using time-of-flight mass spectrometry and high performance liquid chromatography. We have demonstrated the ability to use on an existing database built from a series of MS/MS experiments and accurate mass verification obtained using other (i.e., FTICR) instrumentation. However, subsequent measurements for the proteome of this organism can be made using NET information and reasonable mass accuracy requirements. While MS/MS experiments are important for identification, the number of peptides that elute during a typical analysis at almost any time far exceeds the ability of most mass spectrometers to perform MS/MS experiments on them. Once an AMT tag database of

Table 4. ORFs identified from LC-MS analyses using peptide AMT tags for different numbers of AMT tags per ORF (see text)

Sample	≥1 AMT tags	≥2 AMT tags per ORF[a]
MM, MLP(1)	945	558
MM, PSP	935	548
MM, MLP(2)	885	532

[a]Multiple AMT tags provides a higher level of confidence

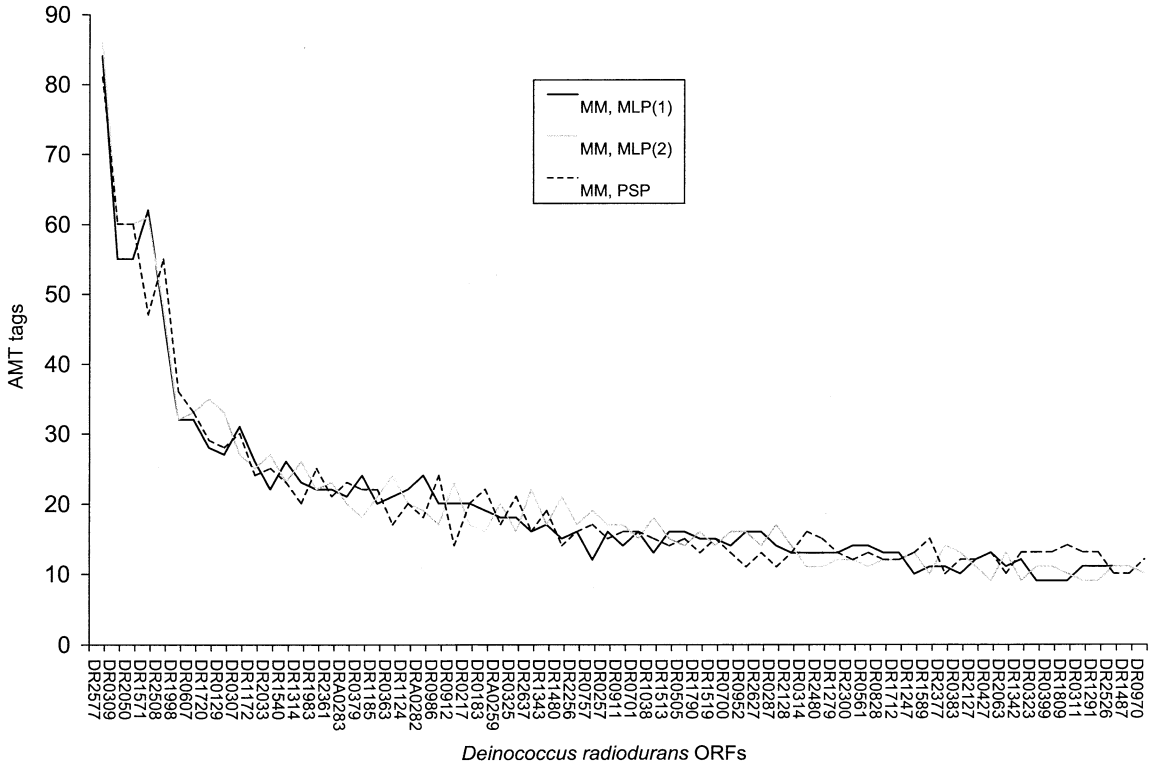

Figure 7. Histogram of 70 ORFs with the most AMT tags found from each of three LC-TOF-MS analyses.

peptides has been developed to contain sufficient number of proteins, i.e., proteome coverage, additional experiments, i.e., quantification of protein levels, can be performed by using the AMT tag identification from a single LC-MS analysis.

The results present here indicate that unique identification of a peptide sequence to monoisotopic mass and elution time in an LC-TOF MS analysis occurs in a majority of cases (~90% of the time). With stricter mass accuracy criteria and improved use of elution time information, this ratio can be further improved. These result were obtained for the organism *D. radiodurans* which has relatively small proteome compared to other sequenced organisms. These results can be extended to more complex organisms (having both larger number of ORFs and post-translationally modified versions of these ORFs) by performing additional separation dimensions to address the greater peptide mixture complexity. The methods demonstrated here will be further improved and extended by anticipated advances in the quality of mass measurements and the accuracy of elution time information.

Acknowledgments

The authors thank Nikola Tolic for his assistance in performing the AMT tag identifications. Portions of research were supported by the U.S. Department of Energy, Office of Biological and Environmental Research. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract no. DE-AC06-76RLO 1830.

References

1. Aebersold, R.; Goodlett, D. R. *Chem. Rev.* **2001,** *101,* 269.
2. Tomer, K. B. *Chem. Rev.* **2001,** *101,* 297.
3. Smith, R. D.; Barinaga, C. J.; Udseth, H. R. *Anal. Chem.* **1988,** *60,* 1948.
4. Huang, E. C.; Henion, J. D. *Anal. Chem.* **1991,** *63,* 732.
5. Deterding, L. J.; Parker, C. E.; Perkins, J. R.; Moseley, M. A.; Jorgenson, J. W.; Tomer, K. B. *J. Chromatogr.* **1991,** *554,* 329.
6. Griffin, P. R.; Coffman, J. A.; Hood, L. E.; Yates, J. R. *Int. J. Mass Spectrom. Ion Processes* **1991,** *111,* 131.
7. Hunt, D. F.; Alexander, J. E.; McCormack, A. L.; Martino, P. A.; Michel, H.; Shabanowitz, J.; Sherman, N.; Moseley, M. A.; Jorgenson, J. W.; Tomer, K. B. Mass Spectrometric Methods for Protein and Peptide Sequence Analysis. *Techniques in Protein Chemistry II;* Villafranca, J. J., Ed.; Academic Press: San Diego, 1991; 441.
8. Shen, Y.; Tolic, N.; Zhao, R.; Pasa-Tolic, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D. *Anal. Chem.* **2001,** *73,* 3011.
9. Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001,** *19,* 242.
10. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002,** *2,* 513.
11. Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976.
12. Perkins, D.; Pappin, D.; Creasy, D.; London, U. *Electrophoresis* **1999,** *20,* 3551.
13. Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.*, in press.
14. Griffin, T. J.; Han, D. K.; Gygi, S. P.; Rist, B.; Lee, H.; Aebersold, R.; Parker, K. C. *J. Am. Soc. Mass Spectrom.* **2001,** *12,* 1238.

15. Pasa-Tolic, L.; Jensen, P. K.; Anderson, G. A.; Lipton, M. S.; Peden, K. K.; Martinovic, S.; Tolic, N.; Bruce, J. E.; Smith, R. D. *J. Am. Chem. Soc.* **1999,** *121,* 7949.
16. Blom, K. *J. Am. Soc. Mass Spectrom.* **1998,** *9,* 789.
17. Wiley, W. C.; McLaren, I. H. *Rev. Sci. Instrum.* **1955,** *26,* 1150.
18. Mamyrin, B. A.; Kavantajev, V. J.; Shmikk, D. V.; Zagulin, V. A. *Sov. Phys. JETP* **1973,** *37,* 45.
19. Douglas, D. J.; French, J. B. *J. Am. Soc. Mass Spectrom.* **1992,** *3,* 398.
20. Krutchinsky, A. N.; Loboda, A. V.; Spicer, V. L.; Dworschak, R.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **1998,** *12,* 508.
21. Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000,** *72,* 3349.
22. Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Masselon, C.; Pasa-Tolic, L.; Shen, Y.; Udseth, H. R. *OMICS* **2002,** *6,* 61.
23. Easterling, M. L.; Mize, T. H.; Amster, I. J. *Anal. Chem.* **1999,** *71,* 624.
24. Gorshkov, M. V.; Masselon, C.; Anderson, G. A.; Udseth, H. R.; Harkewicz, R.; Smith, R. D. *J. Am. Soc. Mass Spectrom.* **2001,** *12,* 1169.
25. Masselon, C.; Tolmachev, A.; Anderson, G.; Harkewicz, R.; Smith, R. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 99.
26. Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R.; Romine, M. F.; Shen, Y.; Strittmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99,* 11.
27. White, O.; Eisen, J. A.; Heidelberg, J. F.; Hickey, E. K.; Peterson, J. D.; Dodson, R. J.; Haft, D. H.; Gwinn, M. L.; Nelson, W. C.; Richardson, D. L.; Moffat, K. S.; Qin, H.; Jiang, L.; Pamphile, W.; Crosby, M.; Shen, M.; Vamathevan, J. J.; Lam, P.; McDonald, L.; Utterback, T.; Zalewski, C.; Makarova, K. S.; Aravind, L.; Daly, M. J.; Minton, K. W.; Fleischmann, R. D.; Ketchum, K. A.; Nelson, K. E.; Salzberg, S.; Smith, H. O.; Venter, J. C.; Fraser, C. M. *Science* **1999,** *286,* 1571.
28. Strittmatter, E.; Rodriguez, N.; Smith, R. D. *Anal. Chem.*, in press.
29. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000,** *11,* 320.