The Utility of Accurate Mass and LC Elution Time Information in the Analysis of Complex Proteomes

Angela D. Norbeck,* Matthew E. Monroe,* and Joshua N. Adkins
Biological Sciences Division, Computational Sciences and Mathematics Division, Pacific Northwest National Laboratory, Richland, Washington, USA

Kevin K. Anderson and Don S. Daly
Computational Sciences and Mathematics Division, Pacific Northwest National Laboratory, Richland, Washington, USA

Richard D. Smith
Biological Sciences Division, Computational Sciences and Mathematics Division, Pacific Northwest National Laboratory, Richland, Washington, USA

The combination of mass and normalized elution time (NET) of a peptide identified by liquid chromatography-mass spectrometry (LC-MS) measurements can serve as a unique signature for that peptide. However, the specificity of an LC-MS measurement depends upon the complexity of the proteome (i.e., the number of possible peptides) and the accuracy of the LC-MS measurements. In this work, theoretical tryptic digests of all predicted proteins from the genomes of three organisms of varying complexity were evaluated for specificity. Accuracy of the LC-MS measurement of mass-NET pairs (on a 0 to 1.0 NET scale) was described by bivariate normal sampling distributions centered on the peptide signatures. Measurement accuracy (i.e., mass and NET standard deviations of ±0.1, 1, 5, and 10 ppm, and ±0.01 and 0.05, respectively) was varied to evaluate improvements in process quality. The spatially localized confidence score, a conditional probability of peptide uniqueness, formed the basis for the peptide identification. Application of this approach to organisms with comparatively small proteomes, such as *Deinococcus radiodurans*, shows that modest mass and elution time accuracies are generally adequate for confidently identifying most peptides. For more complex proteomes, more accurate measurements are required. However, the study suggests that the majority of proteins for even the human proteome should be identifiable with reasonable confidence by using LC-MS measurements with mass accuracies within ±1 ppm and high efficiency separations having elution time measurements within ±0.01 NET. (J Am Soc Mass Spectrom 2005, 16, 1239–1249) © 2005 American Society for Mass Spectrometry

Genomes for more than 100 organisms have been sequenced, providing potential protein coding sequences that could number in the millions. The introduction of shotgun proteomics, an approach in which Yates and coworkers have made critical contributions [1, 2], has greatly increased the role of mass spectrometry in biological research. The surge in information that has resulted is presenting challenges related to the identification of proteins as well as the desire to increase the throughput of measurements.

Although the use of tandem mass spectrometry can often provide confident identifications, there is increasing interest in higher throughput approaches that exploit highly accurate mass measurements [3–8]. Earlier studies have shown [9, 10] that utilizing accurate mass spectrometric measurements for MS based identification of peptides within ±0.1 ppm uncertainty (tolerance) can allow significant levels of confidence in protein identifications, even from mixtures with the complexity of some smaller eukaryotic systems (e.g., yeast). However, as the genomic, and thus, the proteomic, complexity of an organism increases, the ability to identify proteins (or peptides) on the basis of mass measurements alone decreases. Additional information such as isoelectric point, LC elution time or, most commonly, the analysis of peptide fragment ions must be used to distinguish

Published online June 23, 2005

Address reprint requests to Dr. R. D. Smith, Biological Systems Analysis and Mass Spectrometry, Pacific Northwest National Laboratory, 3335 Q Ave., P.O. Box 999, MSIN: K8-98, Richland, WA 99352, USA. E-mail: rds@pnl.gov

* A.D. Norbeck and M.E. Monroe made equal contributions to this article.

1044-0305/05/$30.00
doi:10.1016/j.jasms.2005.05.009

Received February 16, 2005
Revised May 20, 2005
Accepted May 23, 2005

peptides that have identical or very similar masses [11–14]. Experimental approaches to address this complexity include more extensive protein or peptide separations, or focusing on only those peptides with a specific physical characteristic (e.g., isolation of cysteinyl peptides by chemical labeling or solid phase extraction techniques [15–17], and fractionation techniques to add a second separation dimension, e.g., MudPIT [2, 18–23], in addition to the use of peptide ion fragmentation patterns. While useful, these methods may decrease analysis throughput, result in lower protein coverage, or result in specific protein losses.

When liquid chromatography (LC) separations (e.g., using a micro-capillary C18 column) are combined with high-resolution mass spectrometric measurements, reproducible peptide elution times can be acquired simultaneously with highly accurate mass measurements, producing informative mass and separation time features. The utility of this information increases with the peak capacity of the separations and the reproducibility of peptide elution times [24, 25]. Although the absolute LC elution time of a particular peptide can vary from run to run because of temperature and flow rate, among other factors, these changes can largely be corrected after normalization by using an appropriate algorithm to align multiple analyses [26, 27].

A peptide's expected mass and normalized elution time defines a signature point in the two-dimensional mass-by-normalized-elution-time space. A reference database comprised of these signature points can then be used to identify peptides using high throughput proteomic analyses by comparing distances from the measurements to the signature points. If the database is small, signature points are more likely to be confidently isolated (i.e., assigned), allowing for determination of peptide identity. As the database grows, however, more and more detected species have near neighbors, and an increasing level of ambiguity can apply to the identification. Processing vagaries and measurement errors randomly distribute replicate measurements about the locale of a peptide's true mass and normalized elution time. The probability distribution describing this scattering of measurements is called a peptide's sampling distribution. The sampling distributions of neighboring peptides may overlap. If a measured mass and elution time falls within the overlap of the sampling distributions of two or more peptides, then the identity of the source peptide will be somewhat ambiguous. The level of ambiguity—the amount of sampling distribution overlap—may be quantified for each reference peptide, and for a set of reference peptides, thus quantifying specificity of the peptides in a database to the measurements from an LC-MS analysis.

In this study, bivariate normal sampling distributions of peptide mass and predicted elution time were used to evaluate the effects of proteome complexity, mass measurement accuracy, and LC separation time precision on the confidence of peptide identification. This work exploits the capability previously developed to predict the elution times for peptides using an artificial neural network approach [9]. The present calculations allow for an estimate of the effectiveness of the approach for a range of biological systems and measurement qualities. For this study, elution time information predicted from theoretical tryptic peptide sequences [26] at several levels of mass accuracy was used to determine the likelihood of correctly identifying a peptide by comparing its mass and elution time with that of a peptide in a reference database. The applicability of this method to a variety of peptide reference databases was addressed by comparing four systems of varying complexity.

Methods

Databases

Protein lists for three organisms—*Deinococcus radiodurans*, *Saccharomyces cerevisiae*, and *Homo sapiens*—were obtained from the following protein sequence repositories: *Deinococcus radiodurans* (TIGR, March 21, 2000), *Saccharomyces cerevisiae* (http://www.yeastgenome.org/ provided though Stanford University, January 6, 2003), and *Homo sapiens* (IPI, April 1, 2004). In addition, a fourth system comprised of the combined proteins identified from 436 SEQUEST analyses of LC-MS/MS analyses of human mammary epithelial cells (HMEC) was used to represent an observed subset of human proteins (described in an appendix at the end of this paper). This HMEC dataset represents a set of proteins that have been observed in the proteome for this human cell line [28, 29].

Simulated Processing and Analysis

An *in silico* digestion was performed on the proteins present in each database using Protein Digestion Simulator (PDS), a program written in-house using VB.NET (available online [30]). This program reads a list of protein names and sequences from an input file and performs a virtual tryptic digest on each protein sequence, then uses an improved version of the normalized elution time (NET) prediction program by Petritis et al., to compute the predicted NET values for each sequence [26, 27]. The *in silico* tryptic digestion cleaves each sequence after either lysine or arginine (K or R) sites, but not if the residue is followed by proline. The resultant peptides were permitted to have up to one "missed cleavage" (internal K or R), and were filtered to only include those with a mass between 600 and 4000 Da. The NET prediction portion of PDS is a VB.NET DLL that takes as an input a peptide sequence, its length, and its calculated hydrophobic moment, and computes the NET for the sequence. The predicted NET for a given peptide (on a scale of 0 to 1) is determined

by employing a neural network-based model, developed with the utilization of training data from 20 species and over 200,000 very high quality peptide identifications from LC-MS/MS analyses using strictly controlled separation conditions. Cysteine-only databases were created for each of the four systems by selecting the subset of cysteine-containing peptides from the virtual tryptic digests.

The analysis of individual peptide and overall database specificity for each of the reference databases was achieved by comparing the overlap of peptide sampling distributions using the spatially localized confidence scoring (SLiC) method developed by Anderson et al. [31]. This algorithm estimates the probabilistic distance between a (mass-NET) measurement and each reference peptide (mass-NET), and then computes the SLiC score, the probability (on a 0 to 1 scale) of a match to each reference peptide conditioned on the peptides in the reference set. The measurement is identified with the reference peptide resulting in the largest SLiC score (probability of a match). For an isolated peptide with no neighboring peptides and, hence, no overlapping sampling distributions, identifications have a SLiC score of exactly one. As the number of neighboring peptides or overlap increases, SLiC score decreases.

SLiC Score Calculation

The statistical basis for the SLiC method is derived from estimating the probabilistic distance from a measured mass and time pair, $M_i = (m_i, t_i)$, to the center of a sampling distribution, and then applying Bayes theorem to estimate the likelihood that the point is from that sampling distribution when the results are non-specific [32].

Suppose that the mass and normalized elution time measurement, $M_i = (m_i, t_i)$, for the i^{th} (mass-elution time) observation of the j^{th} peptide is bivariate normally distributed with a mean value, $\mu_j = (\mu_{mj}, \mu_{tj})$ equal to the j^{th} peptide's (mass-NET) signature and with covariance Σ_j. The standardized distance, d_{ij}, from the measurement M_i to the peptide μ_j is computed via

$$d_{ij}^2 = (M_i - \mu_j)' \sum_j^{-1} (M_i - \mu_j) = \frac{(m_i - \mu_{mj})^2}{\sigma_{mj}^2} + \frac{(t_i - \mu_{tj})^2}{\sigma_{tj}^2} \quad (1)$$

under the assumption of independence between the measurements of mass and normalized elution time. If, for every signature point, d_{ij} is greater than 2.43—approximately the 95th percentile of the standard bivariate normal distribution—for all reference peptides j, we will consider the i^{th} measurement as unidentifiable.

If we knew the (a priori) probability π_j that measurements come from the distribution associated with the j^{th} peptide, then, applying Bayes theorem, the conditional probability that M_i comes from the j^{th} peptide, given the measurement M_i, is

$$p_{ij} = \frac{\pi_j \left| \sum_j \right|^{-1/2} \exp(-d_{ij}^2/2)}{\left(\sum_{k=1}^{N} \pi_k \left| \sum_k \right|^{-1/2} \exp(-d_{ik}^2/2) \right)} \quad (2)$$

where N is the number of peptides in the database and the determinant $| \sum_j | = \sigma_{mj}^2 \sigma_{ij}^2$. Because no (a priori) probabilities are available, we will assume they are all equally likely (i.e., $\pi_j = \pi_k$ for all j and k), which yields

$$p_{ij} = \frac{(\sigma_{mj}\sigma_{tj})^{-1} \exp(-d_{ij}^2/2)}{\left(\sum_{k=1}^{N} (\sigma_{mk}\sigma_{ik})^{-1} \exp(-d_{ik}^2/2) \right)} \quad (3)$$

Assuming that the database does not contain the entirety of possible peptides in the analyzed sample, we can admit the possibility that an observed measurement is from an unreferenced source when the d_{ij} values are large for all reference peptides in the database. Eq 3 will always assign probabilities, but if d_{ij} is greater than 2.43, we will consider the i^{th} measurement as unidentifiable.

The SLiC scoring method, developed for the application of the accurate mass and time (AMT) tag approach, uses the measured mass and time information to assist the peptide identification process. The scoring has two steps. The first step computes the standardized distance between the measurement and each peptide signature point and only passes measurements for possible identification that are less than the critical distance, d_C, in the mass-NET space from one or more signature points (i.e., AMT tags). Typically, d_C = 2.4, 3, or 3.7, corresponding to 95, 99, and 99.9% signature probability regions. If a measurement is less than the critical distance from two or more signature points, then the sampling distributions of those signature points—each bounded by the critical distance—overlap. Identifying the peptide source is now an ambiguous endeavor, a situation that is encountered in the real application of the approach, increasingly as the level of complexity increases. To address such ambiguities, the second step of the SLIC scoring method computes a score that aims to estimate the likelihood that the detected species is actually a specific candidate. The detected species can then be assigned a peptide identification for the closest peptide if the quality of the agreement is above a certain (user selected) threshold. By its formulation, the SLiC scoring method accounts for sampling distributions defined by instrumental accuracy and their potential overlap. The issues associated with application of this approach with experimental measurements will be the focus of a future publication.

Comparative Analysis Using Defined Mass and NET Tolerances

To further illustrate the advantages of using a SLiC score for specificity of identification, a parallel analysis (data not shown) was performed on the three smallest databases, *D. radiodurans*, HMEC, and *S. cerevisiae*, us-

Table 1. Description of databases of tryptic peptides, mass range 600 to 4000 Da

Database name	Proteins	Residues (millions)	Tryptic peptides with 0 or 1 missed cleavages
D. radiodurans	3117	0.964	125,640
S. cerevisiae	6360	2.99	416,552
H. sapiens	41,216	19.3	1,683,095
HMEC subset	2759	1.63	234,398

ing defined mass and NET tolerances in a binary function such that a mass-NET pair will either be unique (no other signature mass-NET point within the region of tolerance), or it will be non-unique (another AMT tag on or within the boundary of tolerance). The results showed that signature points considered unique by the binary method typically had SLiC scores of 1.0, but occasionally had scores between 0.98 and 1.0 because of the presence of neighboring points just outside the distribution region. A SLiC score is thus representative of the complexity of the region of mass-NET space for the measurement and provides a value on a scale of uniqueness which allows the selection of a threshold of acceptance (i.e., assignment of a peptide identification).

For this study, a peptide identification was assigned if it had a SLiC score $\geq$0.95. Following *in silico* digestion, each peptide was examined for its uniqueness of mass and elution time compared with that for all other peptides of a particular proteome within mass and NET tolerances of $\pm$0.1, 1, 5, and 10 ppm and $\pm$0.01, and 0.05 NET, or without a NET constraint.

Results and Discussion

Information from the analysis of the tryptic peptide and tryptic cysteinyl peptide databases for each proteome used in this study is provided in Tables 1 and 2, respectively. The HMEC example was included to represent a subset of the human proteome, and is representative of applications where only a fraction of the possible proteins would actually be expressed or experimentally detectable. HMEC is the smallest database, comprised of 2759 proteins, and the *H. sapiens* database is the largest, with 41,216 proteins. That is roughly 13 times the number of proteins in *D. radiodurans* and 6.5 times more proteins than the *S. cerevisiae* database (Table 1). The *H. sapiens* database also has ~13 and 4 times more peptides than the *D. radiodurans* and *S. cerevisiae* databases, respectively. Over 90% of both *H. sapiens* and HMEC proteins contain cysteine residues, whereas only 67% of *D. radiodurans* proteins contain cysteine residues (Table 2). *S. cerevisiae* lies between these two extremes, with 89% cysteine-containing proteins. It should be noted that while the HMEC derived protein list is smaller than the *D. radiodurans* protein list, the HMEC peptide list is more than twice that of *D. radiodurans*, because of differences in average protein size.

Table 2. Description of databases containing cysteine, mass range 600 to 4000 Da

Database name	Cysteinyl peptides	Proteins containing cysteine
D. radiodurans	9%	67%
S. cerevisiae	16%	89%
H. sapiens	26%	95%
HMEC subset	21%	95%

Differences in system complexity are readily observed in Figure 1a, where predicted peptide NET values are plotted against the monoisotopic mass for all tryptic peptides in each of the four systems. Interestingly, it should be noted that the areas of mass-NET space that are sparse in peptides differ somewhat between species. For instance, the sparse region around 0.4 NET and 2500 Da for *D. radiodurans* contrasts to the same region for *S. cerevisiae* that is relatively denser, and even more sharply to the same region for human (i.e., *H. sapiens* and HMEC), thus reflecting the differences in amino acid distributions of prokaryotic and eukaryotic systems. A similar observation applies for regions above 0.4 NET, as is evident for *S. cerevisiae*, and to an even greater extent for *H. sapiens*. Not surprisingly, the data for HMEC appear similar to that for *S. cerevisiae*; however, *S. cerevisiae* peptides are more dense than HMEC in the NET range greater than 0.6, reflecting subtle differences in abundances of lysine and arginine and therefore the frequency of trypsin cleavage sites.

When only cysteinyl peptides are considered (Figure 1b) the complexity of all four systems is substantially reduced compared with the whole proteome (Figure 1a), a significant advantage for approaches that isolate this subset of peptides. Cysteinyl tryptic peptides in Figure 1b are also well distributed across mass and NET values, indicating the viability of this method for obtaining a subset of peptides without biasing the sampling to any particular mass-NET region.

The advantage of experimentally isolating cysteinyl peptides is illustrated in Figure 2. Here, a detailed view is shown for a common "dense" region of the plots in Figure 1 for each of the four systems, illustrating the distribution of peptides by their mass and NET values. The circles show the two levels of mass and NET constraints for one cys-peptide (i.e., a signature point) to illustrate the potential effectiveness for unambiguous peptide identification. The lower precision constraints of $\pm$5 ppm and 0.05 NET, represented by the larger circle, include more peptides than the tighter con-

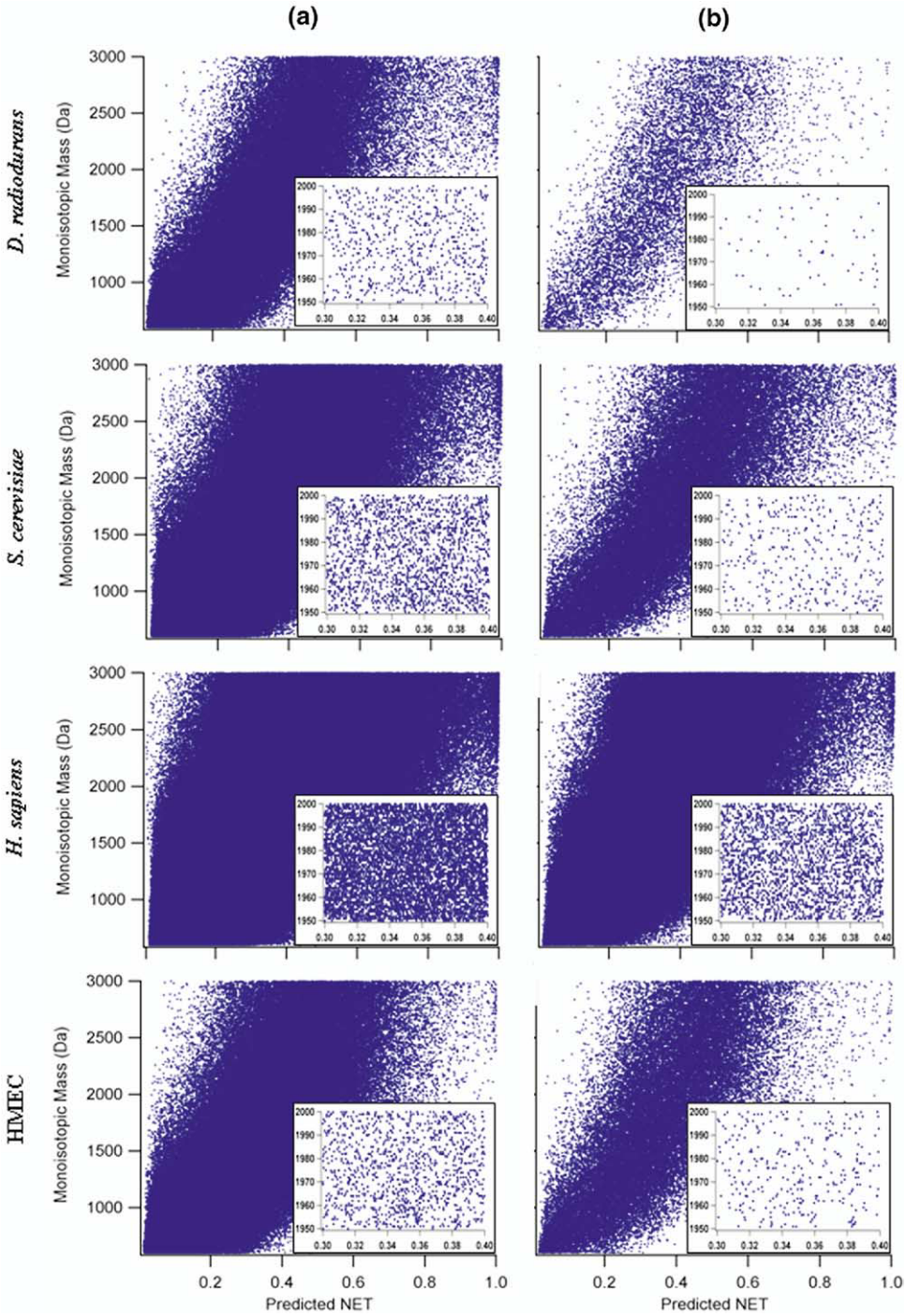

Figure 1. (**a**) Global representation of tryptic digests for all four systems studied. (**b**) Cysteine-containing peptides from tryptic digests for all four systems studied. Predicted Normalized Elution Time (NET) is plotted along the x-axis, and monoisotoptic mass in Daltons is plotted along the y-axis. Inset views are representative of the region contained within 1950–2000 Da and 0.3–0.4 NET.

straints of ±1 ppm and 0.01 NET represented by the smaller circle. Figure 2 shows that if only cysteine-containing peptides are used for peptide identification, the number of unique peptide choices is reduced by more than half, effectively doubling the likelihood a unique peptide will be chosen, illustrated by the disparity of triangles (cysteinyl peptides) compared to dots (non-cysteine containing peptides). In this example, the point chosen in *D. radiodurans* is completely isolated from the other points in the region. The peptide identification, therefore, is unambiguous, regardless of the mass and NET constraints used. But for the *S. cerevisiae* proteome, three peptides fall within the circle corresponding to the more relaxed criteria, which presents

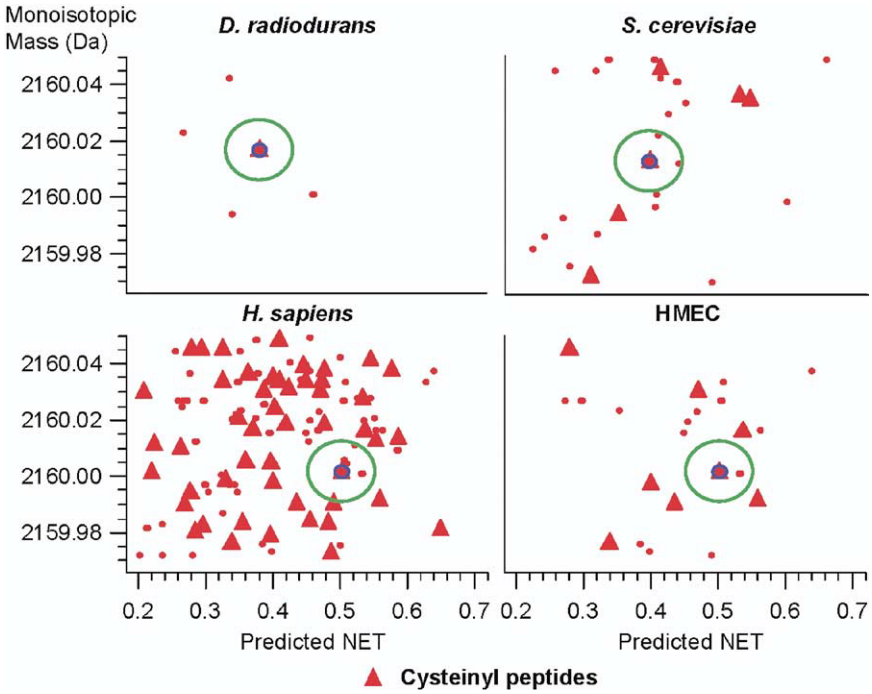

Figure 2. Detailed view of a "dense" region from the plots in Figure 1, where cysteinyl peptides are shown as triangles and non-cysteine containing tryptic peptides as dots. The peptide point chosen for comparison in both the *H. sapiens* and HMEC subset systems corresponds to the same cysteinyl tryptic peptide.

some ambiguity for identification. However, with higher mass accuracy and NET precision (inner circle), the peptide can be distinctly identified. For *H. sapiens*, clearly a ±5 ppm/0.05 NET tolerance for cysteinyl peptides generally is not adequate for uniquely identifying peptides in denser regions of mass-NET space, as indicated by the five peptide species located within the outer boundary region. However, many unique peptide identification assignments can be made if tighter constraints can be applied. Figure 2 also illustrate the benefits of using NET and mass information rather than mass alone. If only mass is used as the criteria for identification, one encounters more than 20 peptides having masses within the selected mass region (2160.002 ± 5 ppm) of the *H. sapiens* plot in Figure 2. This phenomenon applies for the other systems, but because of their smaller proteomes, the number of ambiguous assignments within a given mass range and ±5 ppm accuracy is smaller, and more peptides could be confidently assigned, particularly from less dense mass-NET space.

Figure 3 shows a detailed view of peptides from *H. sapiens*, from a dense region of the plot in Figure 2. The dots represent peptides that do not contain cysteine, and those that are cysteine-containing are represented as triangles. In the example shown, two peptides are located very close to each other, with the target centered on one peptide (A), while another peptide (B) is located just outside the smaller circle. With rigid mass and NET constraints of ±5 ppm and 0.05 NET, defined by the larger region, the four peptides within this region are effectively indistinguishable. However, some assignments in such cases are more likely to be correct, and this information can be useful in several ways (e.g., in establishing the confidence of protein level assignments), and an approach for gauging ambiguity, e.g., using SLiC scores, can serve this purpose.

The effect of the mass-NET distance and the number of neighboring peptides on these scores is illustrated in Figure 3 based upon either ±5 ppm/0.05 NET (outer circle) or ±1 ppm/0.01 NET (inner circle). The underlined SLiC scores for points A–C correspond to the more stringent constraints (inner circle) while the remaining scores correspond to the less stringent constraints (outer circle). The SLiC scores for the ±1 ppm and 0.01 NET tolerances are only shown for three points because the remaining points are too far from the selected peptide and, therefore, have SLiC scores nearly equal to zero. If an LC-MS analysis detected a peptide of mass 2160.002 ± 1 ppm having a NET value of 0.5035 ± 0.01, then the SLiC score for peptide A would be 0.96, while that for peptide B would be 0.04, based upon the more stringent mass and NET constraints. The SLiC score of this peptide is not 1 because of the close proximity of peptide B. Peptide C is farther away, and therefore its presence has less of an effect on the score for peptide A. If the less stringent mass and NET tolerances are used (outer circle), then the presence of all points within that region, as well as the one point

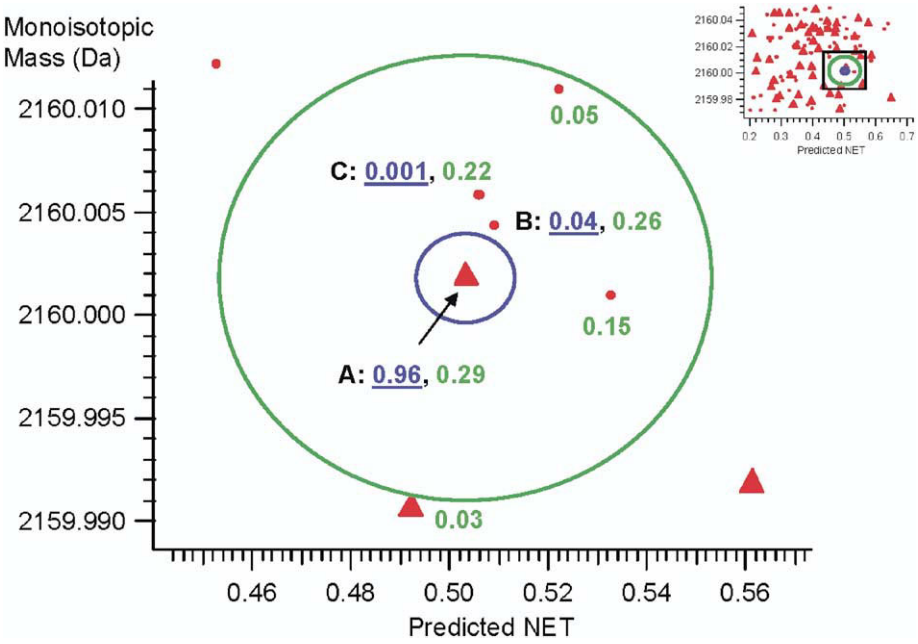

Figure 3. Granular view of dense region from *H. sapiens*. Cys-peptides are represented as triangles. The SLiC scores for ±1 ppm/0.01 NET are underlined, while those for ±5 ppm/0.05 NET are present for each data point. The peptide sequences for points A–C are LVWEEAMSRFCEAEFSVK, FGLLMVENLEEHSEASNIE, and DDLDEQIRHMLFSWAER, respectively.

just outside the region, effectively reduces the SLiC score of peptide A to 0.29. In this way, SLiC is representative of the density of neighboring points in the two-dimensional space, and the closeness of each point to the center. A threshold for SLiC scores can be applied during data analysis, applying a minimum acceptable score given the aims of a study and the manner of use. For example, if high confidence in individual peptide identifications is desired, then one would use a minimum SLiC score of 0.95 to guarantee that all matches are solidly unique. If one wanted to allow borderline matches to pass the filters and use subsequent data processing steps to sort out the matches (e.g., rollup of peptides to proteins), then one might choose a lower SLiC score of 0.75. Thus, the SLiC score can be useful for the assessment of peptide identifications since it assigns a confidence value when ambiguities exist.

Figure 4 shows the uniqueness of tryptic peptides as a function of peptide mass for each system (assuming up to one missed cleavage), at different levels of mass accuracy and NET precision. The SLiC score is used to determine the relative position and number of neighboring peptides for each peptide. If a peptide had a SLiC score ≥0.95, then it was designated as unique, and the number of unique peptides out of the total peptides for each mass range was calculated. The total mass range of 600–4000 Da was divided evenly into 140 binned regions of 25 Da each, and this range of mass was used for the percentage calculations. Figure 4 shows that the general trend of uniqueness is very low at masses less than 1000 Da, but increases as the peptide size increases. Lower accuracy measurements (5 and 10 ppm) are generally not sufficient to maintain greater than 30% uniqueness when no NET constraints are applied for any system. If 50% uniqueness is to be achieved, measurements must be within 0.1 ppm, especially for the larger proteomes and with no NET measurements. With a constraint of ±0.05 NET, greater than 50% of unique identifications can be obtained for peptides greater than 1500 Da for all but the most complex proteome. With the most stringent NET constraints, extremely low mass tolerances are not as important to maintain high levels of peptide distinction. Larger peptides have a higher uniqueness than smaller peptides, and above 2000 Da at ±1 ppm and 0.01 NET, greater than 75% of peptides from all systems are unique.

D. radiodurans

With a mass larger than 2000 Da, greater than 50% uniqueness can be achieved with mass accuracy at or better than 1 ppm, if no NET constraints are used. With lower mass accuracy, greater than 50% uniqueness can only be attained with 5 ppm and at high mass (>3500 Da), and 10 ppm accuracy is not sufficient to reach 50% at any mass. With a NET precision of ±0.05, a mass of at least 2500 Da and mass accuracy of at maximum ±5 ppm or a mass greater than 3500 (10 ppm) is required for ~50% of the peptides to be unique. At ±0.01 NET precision, the percent of unique peptides for all but the lowest mass accuracy and mass greater than 1000 Da increases to ~75%.

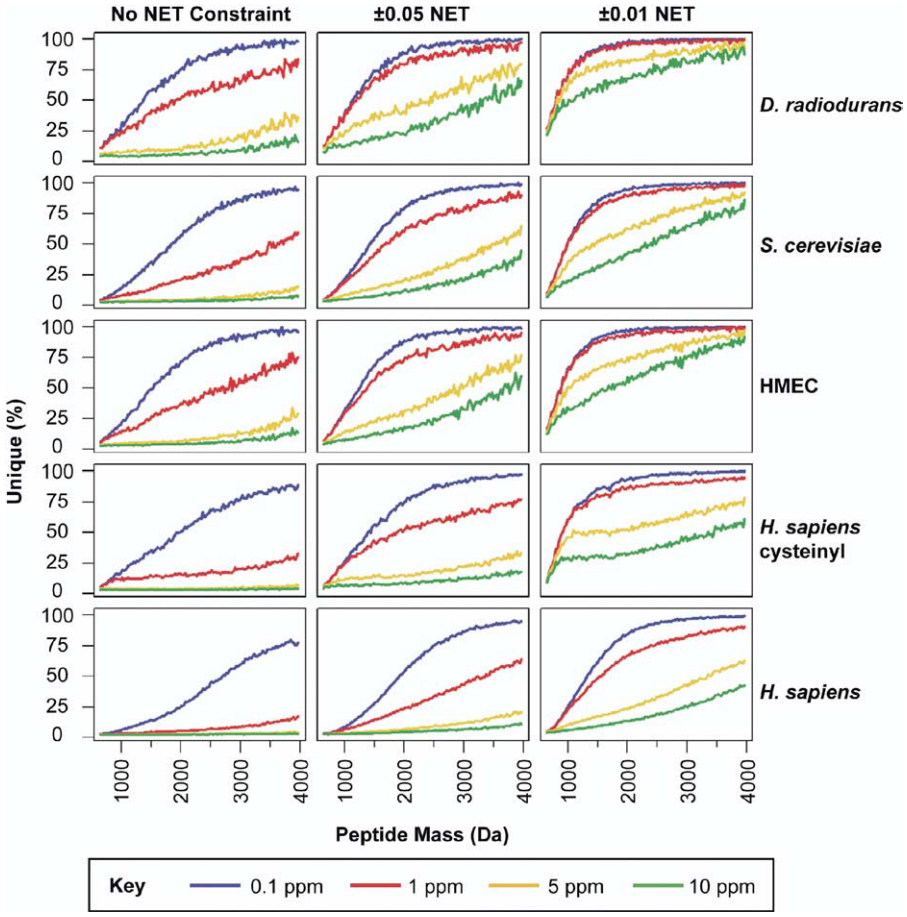

Figure 4. Percent of peptides that are unique versus peptide monoisotoptic mass for the four systems for no NET constraint, and for ±0.05 and 0.01 NET, as well as for different levels of mass accuracy. The peptides plotted have SLiC scores greater than or equal to 0.95. Cysteinyl-only peptides are shown for *H. sapiens* in addition to all peptides for *H. sapiens*.

S. cerevisiae

In comparison with *D. radiodurans*, the *S. cerevisiae* proteome is larger, and thus the level of measurement accuracy required for distinct peptide identifications is increased. If 50% uniqueness is used as the benchmark of acceptance, then a mass measurement accuracy of 0.1 ppm is sufficient for identifications with or without a NET measurement. However, if the mass accuracy is 1 ppm, NET measurements with a precision of at least ±0.05 must be obtained to reach the benchmark. The tightest NET constraints (±0.01 NET) will allow for greater than 50% uniqueness with a mass window of 5 ppm for peptides larger than ~1500 Da.

HMEC

For HMEC with no NET constraints, a minimum of 1 ppm and a mass greater than 2500 Da is required to differentiate 50% of the peptides, and at 5 ppm, the best possible rate is less than 30%, even for large peptides. To achieve the benchmark of acceptance (50%) for peptides around 1500 Da, mass tolerances of 0.1 or 1 ppm must be used, or NET measurements of ±0.01, for mass measurements of 5 ppm.

H. sapiens Using Cysteinyl Peptides

When no NET constraints are employed, the benchmark of uniqueness is only achieved using 0.1 ppm. With 1 ppm, uniqueness is less than 30%, and lower mass accuracies are not sufficient to distinguish peptides. With NET measurements of ±0.05, a mass accuracy of 1 ppm will reach the benchmark at a peptide mass of 2000 Da. Five ppm accuracy measurements are generally useful only with the most stringent NET constraint.

H. sapiens

Not surprisingly, tryptic peptides from *H. sapiens* show a lower average uniqueness in comparison to the other systems, attributable to the greater complexity. For the *H. sapiens* peptides without the use of a NET constraint, less than 10% of peptides can be assigned using any mass tolerance examined other than 0.1 ppm or except

at the largest peptide masses. Even with the most precise NET measurements, 1 ppm is still required to assign 50% peptides at a mass of ~1500 Da. This highlights the advantage of cysteine isolation techniques and high precision elution time measurements.

Conclusions

The use of accurate mass and LC-NET information for peptide identifications in the context of high throughput measurements can be effective for addressing proteomes of high complexity. As the accuracy of these two measurements improves, the extent of protein coverage by confidently identified peptides also increases. Statistically speaking, peptides have the highest possibility of being uniquely identified when both the separation and mass spectrometric measurements are as accurate and reproducible as possible. Now, high accuracy mass spectrometers are able to measure masses within a tolerance of 1 ppm or less, and LC separations have sufficient run-to-run performance to support the use of alignment algorithms that yield corrected (e.g., normalized) elution times within 1%. In spite of inherent uncertainties that apply to accurate mass and NET measurements, these high specificities provide the basis for proteomics approaches that combine accurate mass and NET to identify peptides in a high throughput manner (i.e., without the need to identify every peptide using routine MS/MS). These approaches can be further augmented by using sample preparation techniques that isolate either cysteinyl peptides or a well-defined subset of peptides, making them attractive for studies of the most complex proteomes.

Estimating confidence levels when ambiguities arise can be accomplished by utilizing the SLiC score. This score determines to what extent a peptide is isolated in mass-NET space and can be used to gauge measurement specificity. Such information can aid in determining the more likely identification when uncertainties are present. As illustrated herein, the SLiC score allows the uniqueness of a peptide identification to be assessed by incorporating a degree of certainty that is more useful than fixed acceptance criteria, and allows for an acceptance threshold to be defined by the researcher to best suit the needs of the application and downstream data processing (such as roll-up to the protein level).

The present theoretical approach underestimates the added complexity of protein modifications and "partial" tryptic peptides observed in actual proteomics samples. However, this approach does incorporate many peptides that will not be observed in actual samples due to many proteins not being expressed and/or not being present at detectable levels (e.g., highly hydrophobic peptides that are underrepresented in proteome analyses). While the offset of these two factors will vary from one situation to another, we note that to date, the number of peptides we have experimentally detected in all systems we investigated has been significantly less than the numbers used in this work. The lower numbers are most likely a manifestation of the finite dynamic range of measurements or (similarly) a lack of expression of many proteins.

The present calculations indicate that modest mass accuracies of ± 5 ppm and ± 0.05 NET tolerances will likely be adequate for identifying the majority of peptides for a given system, and particularly for less complex proteomes. For more complex proteomes, the present study provides a basis for estimating the relative practicality of using the combination of highly accurate mass and normalized elution measurements to identify peptides with varying degrees of precision.

Acknowledgments

Portions of this research were supported by the NIH National Center for Research Resources (RR18522) at Pacific Northwest National Laboratory (PNNL), the National Institute of Allergy and Infectious Diseases interagency agreement Y1-AI-4894-01, and the PNNL Biological Sciences Initiative. The authors thank Jon Jacobs, Kostas Petritis, Penny Colton, and Samuel Purvine for helpful discussions and critical reading of the manuscript. This research was performed in the Environmental Molecular Sciences Laboratory (a national scientific user facility sponsored by the U.S. DOE Office of Biological and Environmental Research) located at Pacific Northwest National Laboratory, operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RL0 1830.

Appendix
Generation of the HMEC Database

HMEC Sample Preparation

The whole cell lysates were split into four groups, each representing a different focus for identification. Preparation conditions for each group are summarized below.

Group 1. Global 3D analysis of HMEC. A protein size exclusion separation was performed, with subsequent trypsin digestion (Promega, Madison, WI), half using alkylation by iodoacetamide and half without, followed by strong cation exchange fractionation (SCX) of each of the size exclusion fractions and LC-MS/MS analysis of each SCX fraction This group had 149 samples.

Group 2. Second global analysis. No protein separation was performed before digestion with trypsin. SCX fractionation (plus alkylation) was performed, followed by LC-MS/MS analysis. This group had 67 fractions.

Group 3. Cysteine enrichment global dataset. Same as for Group 2, except that peptides were first treated using quantitative cysteinyl-peptide enrichment technique (QCET) [16] for cysteine enrichment before SCX fractionation. A total of 60 fractions were collected.

Group 4. Secreted protein sample. The media from four different growth treatments of HMEC cell samples were analyzed to target secreted proteins. Each sample was cleaned, protein isolated, and digested (plus alkylation), then SCX fractionated. There were 40 fractions for each of the four samples, for a total of 160 total fractions.

LC-MS Analysis of HMEC Samples

The high-pressure LC (HPLC) system consisted of a pair of Model 100DM 100-mL syringe pumps and Series D controller (Isco, Inc., Lincoln, NE), an in-house manufactured stir-bar style mobile phase mixer (2.5-mL volume), two 4-port, 2-position valves (Valco Instruments Co., Houston, TX) for mobile phase and capillary column selection, and a 6-port, 2-position Valco valve equipped with a 10-μL sample loop for automated injections. The mixer and valves were mounted on an in-house manufactured rack assembly that was custom fit to a PAL autosampler (Leap Technologies, Carrboro, NC) for unattended routine analysis. Reversed-phase capillary HPLC columns were manufactured in-house by slurry packing 5-μm Jupiter C_{18} stationary phase (Phenomenex, Torrence, CA) into a 60-cm length of 360 μm o.d. $\times$ 150 μm i.d. fused silica capillary tubing (Polymicro Technologies Inc., Phoenix, AZ) incorporating a 2-μm retaining screen in a 1/16 inch capillary-bore union (Valco).

The mobile phase consisted of 0.2% acetic acid and 0.05% TFA in water (A) and 0.1% TFA in 90% acetonitrile/10%water (B). Mobile phase was degassed with an in-line Alltech vacuum degasser (Alltech Associates, Inc., Deerfield, IL). The HPLC system was equilibrated at 5000 psi with 100% mobile phase A for initial starting conditions. The mobile phase selection valve was switched from position A to B 20 min after injection, creating an exponential gradient as mobile phase B displaced A in the mixer. An ~5-cm length of 360 i.d. fused silica tubing packed with 5 μm C_{18} was used to split ~25 μL/min of flow before the injection valve. The split flow controls gradient speed under conditions of constant pressure operation. Flow through the capillary HPLC column was ~1.8 μL/min when equilibrated to 100% mobile phase A.

MS analysis was performed using a Finnigan model LCQ Duo or XP ion trap mass spectrometer (ThermoElectron Corp., San Jose, CA) with electrospray ionization (ESI). The HPLC column was coupled to the mass spectrometer using an in-house manufactured interface. No sheath gas or make-up liquid was used. The heated capillary temperature and spray voltage were 200 °C and 2.2 kV, respectively. Samples were analyzed over a mass (m/z) range of 400–2000. For each cycle, the three most abundant ions from MS analysis were selected for MS/MS analysis using a collision energy setting of 45%. Dynamic exclusion was used to discriminate against previously analyzed ions.

HMEC Protein Identifications

A total of 436 LC-MS/MS analyses were performed with the digested lysate samples. The datasets were searched against the *H. sapiens* database using SEQUEST [33]. Proteins with 2 or more peptides of high confidence (XCorr $\geq$ 1.9, 2.2, and 3.75 for 1+, 2+, and 3+ charge states, respectively), a mass between 500 and 4000 Da, and no more than one missed tryptic cleavage were compiled, resulting in a list of 2759 unique proteins. The peptide confidence criteria used here are similar to those used by the Yates and coworkers [2, 33]. This protein list was then treated in the same manner as the protein lists that were downloaded for the other three systems.

References

1. Eng, J. K.; McCormack, A. L.; Yates, J. R. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5*(11), 976–989.
2. Washburn, M. P.; Wolters, D.; Yates, J. R. III. Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. *Nat. Biotechnol.* **2001,** *19*(3),242–247.
3. Brock, A.; Horn, D.; Peters, E.; Shaw, C.; Ericson, C.; Phung, Q.; Salomon, A. An Automated Matrix-Assisted Laser Desorption/Ionization Quadrupole Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for "Bottom-Up" Proteomics. *Anal. Chem.* **2003,** *75*(14),3419–3428.
4. Masselon, C.; Anderson, G.; Harkewicz, R.; Bruce, J.; Pasa-Tolic, L.; Smith, R. Accurate Mass Multiplexed Tandem Mass Spectrometry for High-Throughput Polypeptide Identification from Mixtures. *Anal. Chem.* **2000,** *72*(8),1918–1924.
5. Jacobs, J.; Monroe, M.; Qin, W.; Shen, Y.; Anderson, G.; Smith, R. Ultra-Sensitive, High Throughput, and Quantitative Proteomics Measurements. *Int. J. Mass Spectrom.* **2005,** *240*(3),195–212.
6. Fung, K.; Askovic, S.; Basile, F.; Duncan, M. A Simple and Inexpensive Approach to Interfacing High-Performance Liquid Chromatography and Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry. *Proteomics* **2004,** *4*(10),3121–3127.
7. Liu, T.; Qian, W.; Strittmatter, E.; Camp, D.; Anderson, G.; Thrall, B.; Smith, R. High-Throughput Comparative Proteome Analysis Using a Quantitative Cysteinyl-Peptide Enrichment Technology. *Anal. Chem.* **2004,** *76*(18),5345–5353.
8. Nakamura, T.; Dohmae, N.; Takio, K. Characterization of a Digested Protein Complex with Quantitative Aspects: An Approach Based on Accurate Mass Chromatographic Analysis with Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry. *Proteomics* **2004,** *4*(9),2558–2566.
9. Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolic, L.; Smith, R. D. Utility of Accurate Mass Tags for Proteome-Wide Protein Identification. *Anal. Chem.* **2000,** *72*(14),3349–3354.
10. Zubarev, R.; Hakansson, P.; Sundqvist, B. Accuracy Requirements for Peptide Characterization by Monoisotopic Molecular Mass Measurements. *Anal. Chem.* **1996,** *68*(22),4060–4063.
11. Cargile, B.; Stephenson, J. An Alternative to Tandem Mass Spectrometry: Isoelectric Point and Accurate Mass for the Identification of Peptides. *Anal. Chem.* **2004,** *76*(2),267–275.
12. Palmblad, M.; Ramstrom, M.; Bailey, C.; McCutchen-Maloney, S.; Bergquist, J.; Zeller, L. Protein Identification by Liquid Chromatography-Mass Spectrometry Using Retention Time Prediction. *J. Chromatogr. B* **2004,** *803*(1),131–135.

13. Palmblad, M.; Ramstrom, M.; Markides, K.; Hakansson, P.; Bergquist, J. Prediction of Chromatographic Retention and Protein Identification in Liquid Chromatography/Mass Spectrometry. *Anal. Chem.* **2002,** *74*(22),5826–5830.
14. Spengler, B. De Novo Sequencing, Peptide Composition Analysis, and Composition-Based Sequencing: A New Strategy Employing Accurate Mass Determination by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2004,** *15*(5),703–714.
15. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags. *Nat. Biotechnol.* **1999,** *17*(10),994–999.
16. Liu, T.; Qian, W. J.; Strittmatter, E. F.; Camp, D. G., II; Anderson, G. A.; Thrall, B. D.; Smith, R. D. High-Throughput Comparative Proteome Analysis Using a Quantitative Cysteinyl-Peptide Enrichment Technology. *Anal. Chem.* **2004,** *76*(18),5345–5353.
17. Ihling, C., Sinz, A. Proteome Analysis of *Escherichia coli* Using High-Performance Liquid Chromatography and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Proteomics*, in press.
18. Adkins, J. N.; Varnum, S. M.; Auberry, K. J.; Moore, R. J.; Angell, N. H.; Smith, R. D.; Springer, D. L.; Pounds, J. G. Toward a Human Blood Serum Proteome: Analysis by Multidimensional Separation Coupled with Mass Spectrometry. *Mol. Cell. Proteom.* **2002,** *1*(12),947–955.
19. Durr, E.; Yu, J.; Krasinska, K. M.; Carver, L. A.; Yates, J. R.; Testa, J. E.; Oh, P.; Schnitzer, J. E. Direct Proteomic Mapping of the Lung Microvascular Endothelial Cell Surface In Vivo and in Cell Culture. *Nat. Biotechnol.* **2004,** *22*(8),985–992.
20. Graumann, J.; Dunipace, L. A.; Seol, J. H.; McDonald, W. H.; Yates, J. R., III; Wold, B. J.; Deshaies, R. J. Applicability of Tandem Affinity Purification MudPIT to Pathway Proteomics in Yeast. *Mol. Cell. Proteom.* **2004,** *3*(3),226–237.
21. 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. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. Global Analysis of the *Deinococcus radiodurans* Proteome by Using Accurate Mass Tags. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99*(17),11049–11054.
22. Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome. *J. Proteome Res.* **2003,** *2*(1),43–50.
23. Shen, Y.; Jacobs, J. M.; Camp, D. G., II; Fang, R.; Moore, R. J.; Smith, R. D.; Xiao, W.; Davis, R. W.; Tompkins, R. G. Ultra-High-Efficiency Strong Cation Exchange LC/RPLC/MS/MS for High Dynamic Range Characterization of the Human Plasma Proteome. *Anal. Chem.* **2004,** *6*(4), 1134–1144.
24. Shen, Y.; Tolic, N.; Masselon, C.; Pasa-Tolic, L.; Camp, D. II; Lipton, M.; Anderson, G.; Smith, R. Nanoscale Proteomics. *Anal. Bioanal. Chem.* **2004,** *378*(4),1037–1045.
25. Mohan, D.; Pasa-Tolic, L.; Masselon, C.; Tolic, N.; Bogdanov, B.; Hixson, K.; Smith, R.; Lee, C. Integration of Electrokinetic-Based Multidimensional Separation/Concentration Platform with Electrospray Ionization-Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry for Proteome Analysis of *Shewanella oneidensis*. *Anal. Chem.* **2003,** *75*(17),4432–4440.
26. Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. Use of Artificial Neural Networks for the Accurate Prediction of Peptide Liquid Chromatography Elution Times in Proteome Analyses. *Anal. Chem.* **2003,** *75*(5),1039–1048.
27. Petritis, K. K. L.; Yorn, B.; Strittmatter, E. F.; Camp, D.G. II; Lipton, M.; Xu, Y.; Smith, R. D. Improved Liquid Chromatography peptide Elution Time Prediction by Using Artificial Neural Networks for Improved Proteomics Analysis. *Proceedings of the 52nd ASMS Symposium*; Nashville, TN, May 2004.
28. Jacobs, J. M.; Mottaz, H. M.; Yu, L. R.; Anderson, D. J.; Moore, R. J.; Chen, W. N.; Auberry, K. J.; Strittmatter, E. F.; Monroe, M. E.; Thrall, B. D.; Camp, D. G., II; Smith, R. D. Multidimensional Proteome Analysis of Human Mammary Epithelial Cells. *J. Proteome Res.* **2004,** *3*(1),68–75.
29. Liu, T.; Qian, W. J.; Chen, W. J.; Jacobs, J. M.; Moore, R. J.; Anderson, D. J.; Gritsenko, M. A.; Monroe, M. E.; Thrall, B. D.; Camp, D. G., II; Smith, R. D. Improved Proteome Coverage by Using High Efficiency Cysteinyl-Peptide Enrichment: The Human Mammary Epithelial Cell Proteome. *Proteomics*, in press.
30. Monroe, M. E. *Protein Digestion Simulator*. http://ncrr.pnl.gov/software/ (v2.0.1846.26324, 01/20/05).
31. Anderson, K. K.; Monroe, M. E.; Daly, D. S. Estimating Probabilities of Peptide Assignments to LC-FTICR-MS Observations. *Proceedings of the International Conference METMBS*; Las Vegas, NV, June, 2004, pp 151–156.
32. Groen, F.; Mosleh, A. Foundations of Probabilistic Inference with Uncertain Evidence. *Int. J. Approx. Reason* **2005,** *39*(1),49–83.
33. Ducret, A.; Van Oostveen, I.; Eng, J. K.; Yates, J. R., III; Aebersold, R. High Throughput Protein Characterization by Automated Reverse-Phase Chromatography/Electrospray Tandem Mass Spectrometry. *Protein Sci.* **1998,** *7*(3),706–719.