Anal. Chem. **2004,** *76,* 267−275

An Alternative to Tandem Mass Spectrometry: Isoelectric Point and Accurate Mass for the Identification of Peptides

Benjamin J. Cargile and James L. Stephenson, Jr.*

Mass Spectrometry Research Group, Research Triangle Institute, 3040 Cornwallis Road, Research Triangle Park, North Carolina 27709-2194

The traditional approach to the identification of peptides in complex biological samples integrally involves the use of tandem mass spectrometry to generate a unique fragmentation pattern in order to accurately assign its identity to a particular protein. In this article we describe the theoretical basis for a new paradigm for the identification of peptides and proteins. This methodology employs the use of accurate mass and peptide isoelectric point (p*I*) as identification criteria, and represents a change in focus from current tandem mass spectrometry-dominated approaches. A mathematical derivation of the false positive rate associated with accurate mass and p*I* measurements is presented to demonstrate the utility of the technique. The equations for calculation of the experimental false positive rate allow for the determination of the validity of the data. The false positive rate issue examined in detail here is not restricted to accurate mass-based approaches, but also has application to the tandem mass spectrometry community as well. The theoretical proteomes of *Escherichia coli* and *Rattus norvegicus* are used to evaluate the efficacy of this approach. The power of the technique is demonstrated by analyzing a series of peptides with the same monoisotopic masses but with differing isoelectric points. Finally, the speed of algorithm when combined with the experimental peptide analysis has the potential to rapidly accelerate the protein identification process.

The introduction of methods for shotgun proteomics[1−4] has begun to pave the way for high-throughput analysis of the entire protein complement of a cell. Although much of the method development has been purely analytically driven, it is the potential information that such studies could offer about the mechanics of life that is stimulating the growth of this field. It is hoped that such technologies will allow for quantitative comparisons of protein levels,[5−14] analysis of the posttranslational state of proteins such as phosphorylation,[15−18] subcellular protein localization,[19−25] analysis of dynamic turnover,[26−28] and elucidation of the network of protein−protein interactions.[29,30] The possibility exists that more

[*] Corresponding author. E-mail: stephensonjl@rti.org. Phone: (919) 316-3978. Fax: (919) 541-7208.

(1) Wu, S. L.; Amato, H.; Biringer, R.; Choudhary, G.; Shieh, P.; Hancock, W. S. *J Proteome Res.* **2002**, *1*, 459−465.
(2) McDonald, W. H.; Yates, J. R., III. *Dis. Markers* **2002**, *18*, 99−105.
(3) Hancock, W. S.; Wu, S. L.; Shieh, P. *Proteomics* **2002**, *2*, 352−359.
(4) McDonald, W. H.; Yates, J. R., III. *Curr. Opin. Mol. Ther.* **2003**, *5*, 302−309.
(5) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell Proteomics* **2002**, *1*, 376−386.
(6) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.
(7) Kindy, J. M.; Taraszka, J. A.; Regnier, F. E.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 950−958.
(8) Regnier, F. E.; Riggs, L.; Zhang, R.; Xiong, L.; Liu, P.; Chakraborty, A.; Seeley, E.; Sioma, C.; Thompson, R. A. *J. Mass Spectrom.* **2002**, *37*, 133−145.
(9) Yao, X.; Freas, A.; Ramirez, J.; Demirev, P. A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 2836−2842.
(10) Moseley, M. A. *Trends Biotechnol.* **2001**, *19*, S10−S16.
(11) Cagney, G.; Emili, A. *Nat. Biotechnol.* **2002**, *20*, 163−170.
(12) Conrads, T. P.; Alving, K.; Veenstra, T. D.; Belov, M. E.; Anderson, G. A.; Anderson, D. J.; Lipton, M. S.; Pasa-Tolic, L.; Udseth, H. R.; Chrisler, W. B.; Thrall, B. D.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 2132−2139.
(13) Washburn, M. P.; Ulaszek, R.; Deciu, C.; Schieltz, D. M.; Yates, J. R., III. *Anal. Chem.* **2002**, *74*, 1650−1657.
(14) Goshe, M. B.; Smith, R. D. *Curr. Opin. Biotechnol.* **2003**, *14*, 101−109.
(15) Ficarro, S.; Chertihin, O.; Westbrook, V. A.; White, F.; Jayes, F.; Kalab, P.; Marto, J. A.; Shabanowitz, J.; Herr, J. C.; Hunt, D. F.; Visconti, P. E. *J. Biol. Chem.* **2003**, *278*, 11579−11589.
(16) Ficarro, S. B.; McCleland, M. L.; Stukenberg, P. T.; Burke, D. J.; Ross, M. M.; Shabanowitz, J.; Hunt, D. F.; White, F. M. *Nat. Biotechnol.* **2002**, *20*, 301−305.
(17) Zhou, H.; Watts, J. D.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 375−378.
(18) Oda, Y.; Nagasu, T.; Chait, B. T. *Nat. Biotechnol.* **2001**, *19*, 379−382.
(19) Jung, E.; Heller, M.; Sanchez, J. C.; Hochstrasser, D. F. *Electrophoresis* **2000**, *21*, 3369−3377.
(20) Cronshaw, J. M.; Krutchinsky, A. N.; Zhang, W.; Chait, B. T.; Matunis, M. J. *J. Cell. Biol.* **2002**, *158*, 915−927.
(21) Schirmer, E. C.; Gerace, L. *Genome Biol.* **2002**, *3*, 1008.
(22) Dreger, M. *Mass Spectrom. Rev.* **2003**, *22*, 27−56.
(23) Huber, L. A.; Pfaller, K.; Vietor, I. *Circ. Res.* **2003**, *92*, 962−968.
(24) Dreger, M. *Eur. J. Biochem.* **2003**, *270*, 589−599.
(25) Taylor, S. W.; Fahy, E.; Ghosh, S. S. *Trends Biotechnol.* **2003**, *21*, 82−88.
(26) Pratt, J. M.; Petty, J.; Riba-Garcia, I.; Robertson, D. H.; Gaskell, S. J.; Oliver, S. G.; Beynon, R. J. *Mol. Cell Proteomics* **2002**, *1*, 579−591.
(27) Gerner, C.; Vejda, S.; Gelbmann, D.; Bayer, E.; Gotzmann, J.; Schulte-Hermann, R.; Mikulits, W. *Mol. Cell Proteomics* **2002**, *1*, 528−537.
(28) Cargile, B.; Bundy, J.; Grunden, A.; Stephenson, J. L. J. *Anal. Chem.* **2003**, in press.
(29) Ho, Y.; Gruhler, A.; Heilbut, A.; Bader, G. D.; Moore, L.; Adams, S. L.; Millar, A.; Taylor, P.; Bennett, K.; Boutilier, K.; Yang, L.; Wolting, C.; Donaldson, I.; Schandorff, S.; Shewnarane, J.; Vo, M.; Taggart, J.; Goudreault, M.; Muskat, B.; Alfarano, C.; Dewar, D.; Lin, Z.; Michalickova, K.; Willems, A. R.; Sassi, H.; Nielsen, P. A.; Rasmussen, K. J.; Andersen, J. R.; Johansen, L. E.; Hansen, L. H.; Jespersen, H.; Podtelejnikov, A.; Nielsen, E.; Crawford, J.; Poulsen, V.; Sorensen, B. D.; Matthiesen, J.; Hendrickson, R. C.; Gleeson, F.; Pawson, T.; Moran, M. F.; Durocher, D.; Mann, M.; Hogue, C. W.; Figeys, D.; Tyers, M. *Nature* **2002**, *415*, 180−183.

10.1021/ac0352070 CCC: $27.50 © 2004 American Chemical Society
Published on Web 12/09/2003

functionally relevant information will also be acquired, such as enzyme activity,[31] enzyme function,[32] and enzymatic mechanism of reaction,[33] as well as structural information from high-throughput proteomic technologies. The current method of choice for quantitative expression profiling is mRNA array technology because of the relative speed and commercial availability of instruments that require less technical skill than the equivalent mass spectrometry-based approaches. However, it will be difficult, if not impossible, to explore many of the other aspects of proteins mentioned above at the mRNA level, especially for organisms with large genomes. The primary problem with many of these mass spectrometry approaches to proteomic analysis is that the amount of instrument time needed to acquire the data is fairly large, and becomes considerably more so if high sequence coverage of the proteins is required. In fact, because of the time constraint, most high-throughput proteomic studies look at on average between 1 and 5 peptides per protein, with one of the highest ratios reported[34] being ~5 peptides per protein for only ~1500 proteins (which likely took roughly 8 days to complete, assuming 100 fractions and only 2 h of mass spectrometry time per fraction). Thus, some method to significantly shorten the analysis time is needed to make truly high-throughput proteomic studies practical.

The primary time constraint in the current dogma of large-scale shotgun proteomics studies is the need to acquire one tandem mass spectrum for every peptide found. This problem is further compounded by the fact that not every tandem mass spectrum identifies a unique peptide, or any peptide at all, and multiple spectra are often taken of the same peptide to ensure identification. Assuming the absolute minimum requirements (i.e., 100% efficiency of identification), to identify 50 peptides using a commercial tandem mass spectrometry instrument requires 51 scans (1 mass spectrum followed by the 50 subsequent MS/MS spectra to identify the analytes). This is highly improbable in a realistic experimental scenario, as typically some sort of separation is performed on the peptide mixture. Therefore, it would require multiple duty cycles to fully record all of the tandem mass spectra. Looking at the problem in this manner shows the time limiting factor to be the acquisition of tandem mass spectra. The most obvious solution to this problem is to eliminate the need for tandem mass spectra during the identification process.

Recently, such an approach has been introduced in the form of the accurate mass tag strategy for peptide identification,[35] which is based on a modification of the theory behind peptide mapping.[36] In this approach, the mass of a single peptide is measured to such a high degree of mass accuracy that the peptide is unique at that mass in the theoretical "unique" tryptic proteome of an organism. However, a number of challenges exist with this method that could limit its overall application. The first is the question of in what context the term "unique" tryptic digest is defined. Second is the effect of considering missed cleavages for this or any other identification approach. In addition, the introduction of non-predicted peptides or compounds (partially tryptic, modified peptides, and non-peptide compounds) can affect the observed false positive rate as well. Finally, and perhaps most importantly, is the applicability of this method to multicellular organisms with complex genomes. All of these factors are important when considering a single criterion identification (approach employing accurate mass measurement alone).

Because of the potential limitations of the accurate mass approach, other methods to acquire supplementary information on the identity of the peptide have been explored. One method, recently described by Smith and co-workers, has been the introduction of retention time prediction of peptides from reversed-phase columns.[37,38] Even if this added constraint is used in the identification process, a number of issues need to be addressed for this strategy to be practical. The first challenge is the need to apply a generic algorithm[37,38] to "align" the liquid chromatography−mass spectrometry (LC−MS) runs so that the chromatograms can be compared (i.e., use of this unknown algorithm to make the data fit the model). The second is the deviation of a ±10% window to encompass at least 95% of the retained peptides.[38] Although a ±3% window for 50% of the retained peptides is reported, this number is not sufficient to enhance the identification process, given the need to be able to predict where a large number of peptides will be retained for any accurate mass measurement strategy to be effective. Additionally, this model is specific to the chromatographic conditions used, since changing the stationary phase, mobile phase modifiers, and mobile phase solvents is known to affect the relative retention time of peptides. Finally, there exists a strong correlation between mass and retention time, as has been noted before,[39] and this is seen in the high-resolution LC−MS analysis of whole proteome digests.[40] For these reasons, the addition of predicted peptide retention time for a reversed-phase column might be insufficient to compensate for the potential limitations of any single accurate mass approach.

Although accurate mass provides some degree of discrimination for peptide identification, separation techniques other than reversed-phase chromatography can provide information on the amino acid composition of the peptides and thus add a new database constraint. The technique with the highest resolving power and predictability available is that of isoelectric focusing.[39,41] The combination of accurate mass with isoelectric point could provide a powerful identification methodology with a high degree of resolving power. On the basis of our work, the average p*I* of

[30] Gavin, A. C.; Bosche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Hofert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. *Nature* **2002**, *415*, 141−147.

[31] Jessani, N.; Liu, Y.; Humphrey, M.; Cravatt, B. F. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10335−10340.

[32] Adam, G. C.; Sorensen, E. J.; Cravatt, B. F. *Nat. Biotechnol.* **2002**, *20*, 805−809.

[33] Kelleher, N. L.; Nicewonger, R. B.; Begley, T. P.; McLafferty, F. W. *J. Biol. Chem.* **1997**, *272*, 32215−32220.

[34] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43−50.

[35] Conrads, T. P.; Anderson, G. A.; Veenstra, T. D.; Pasa-Tolic, L.; Smith, R. D. *Anal. Chem.* **2000**, *72*, 3349−3354.

[36] Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 5011−5015.

[37] Palmblad, M.; Ramstrom, M.; Markides, K. E.; Hakansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826−5830.

[38] 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. *Anal. Chem.* **2003**, *75*, 1039−1048.

[39] Cargile, B.; Bundy, J.; Freeman, T.; Stephenson, J. L. J. *J. Proteome Res.*, in press.

[40] Shen, Y.; Zhao, R.; Belov, M. E.; Conrads, T. P.; Anderson, G. A.; Tang, K.; Pasa-Tolic, L.; Veenstra, T. D.; Lipton, M. S.; Udseth, H. R.; Smith, R. D. *Anal. Chem.* **2001**, *73*, 1766−1775.

[41] Cargile, B.; Talley, D.; Stephenson, J. L. J. *Electrophoresis*, in press.

the identified peptide aligned well with the predicted p*I* of the immobilized pH gradient (IPG) strip.[41] In addition, a low standard deviation at approximately ±0.25 p*I* unit (2 times the average standard deviation encompassing 95% of the peptides) for the IPG fractions in the area of highest peptide density[41] was observed. In addition, many amino acid-specific trends were observed[41] that should lower the average standard deviation considerably when they are incorporated into the identification algorithm. These features of accurate predictability, combined with the other attributes of the IPG technology, make the use of isoelectric focusing/p*I* prediction a highly favorable match with accurate mass to identify peptides without the need for tandem mass spectrometry.

In this work, the theoretical aspects of using just accurate mass as an identification criterion are explored. The aspects evaluated include the number of missed protease cleavages, organism genome/proteome complexity, introduction of random masses on the specificity of the identification, and the theoretical false positive rate of the method. The complex relationship between the mass of the peptide and the false positive rate is also explored. In addition, limits for the acceptable theoretical and experimental false positive levels are calculated as functions of the number of peptides identified. A new method is described to calculate the false positive rate for experimental data. Next, the contribution of p*I* when combined with accurate mass for the identification process is examined. The levels of peptide uniqueness as well as the specificity of the method are explored at various levels of mass accuracy and p*I* predictability in relation to database size. Finally, the unique benefits of using p*I* rather than attempting higher mass measurement accuracy are discussed.

EXPERIMENTAL SECTION

All calculations were performed with programs written in C with LabWindows/CVI 6.0. The calculation of p*I* used the algorithm described by Bjellqvist et al.[42] All the programs were run on an 800-mHz Pentium personal computer, with the amount of time required varying depending on the search preformed and the database used. Typically, searches took on the order of 30 s to 1 min for the more complex database. The databases used were those from NCBI for *Escherichia coli* K-12 and *Rattus norvegicus* (last updated December 5, 2002). All graphs were made with Microcal Origin version 7.0. For the calculation of peptide uniqueness, all the proteins from the given organisms were digested in silico with trypsin at 100% efficiency (no missed cleavages). The program could then be run to reflect any number of missed cleavages for trypsin. For the calculation of specificity, a number of tryptic peptides of random sequence were generated and compared to the appropriate database of unique tryptic peptides. When looking at the specificity for a given organism, 10 000 random tryptic peptides were generated in the range of 1000−3000 Da, and the false positive rate is simply the percent of matches to unique peptides. When examining the specificity as a function of mass, 1000 peptides in 50-Da windows were generated, and the false positive rate was calculated as the number of unique matches divided by the number of unique matches plus number of unknown masses. This change in the false positive rate calculation stems from the fact that, at low mass, most of the defined mass range has multiple peptides present within a given mass accuracy window.

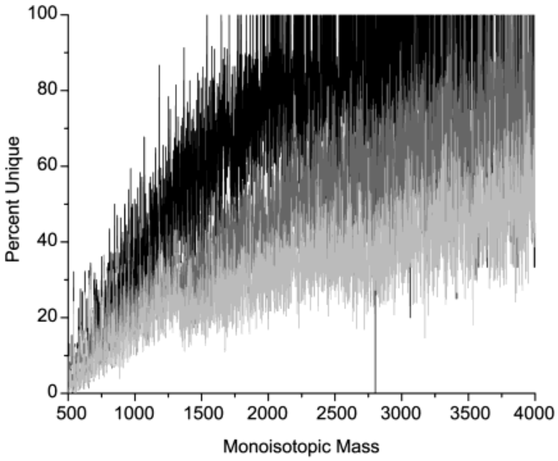

Figure 1. Effect of considering missed cleavages on accurate mass as an identification strategy, shown by comparing multiple missed cleavages (black, zero; medium gray, one; and light gray, two missed cleavages) and examining the percentage of unique peptides identified.

RESULTS AND DISCUSSION

Accurate Mass Alone as a Peptide Identification Criterion. The original introduction of the accurate mass tag (AMT) strategy[35] for identification of peptides by mass alone is highly dependent on the number of missed cleavages observed from a tryptic digest of any given protein sample. For direct comparison with the original publication on employing an accurate mass strategy alone for peptide identification, we define here the term “unique” peptide to mean no missed cleavages as a result of a tryptic digestion. An investigation into the number of unique peptides as a function of the number of missed cleavages, as shown in Figure 1, illustrates that, for the case where there are no missed cleavages, the accurate mass approach is feasible for peptide identification. However, when one or more missed cleavages are considered, as shown in Figure 1, both the total number of peptides increases and the percentage of peptides at higher mass (>1000 Da) increases. This situation limits the accurate mass-only approach because the number of unique peptides goes rapidly toward zero with an increasing number of missed cleavages. A cursory examination of whole proteome digest data from our laboratory and published data from other laboratories[43,34] shows that approximately two-thirds of the identified peptides are from unique tryptic peptides (although this does vary widely, depending on the use of partial tryptics). In contrast, a recent publication, in which stable isotope labeling and Fourier transform ion cyclotron resonance (FTICR) mass spectrometry were used to evaluate[44] the efficiency of tryptic digest, suggested that trypsin works with 94% efficiency. The logic used in the cited paper considered only the number of lysines in the peptide to calculate the number of missed cleavages (thus, a peptide with

[42] Bjellqvist, B.; Hughes, G. J.; Pasquali, C.; Paquet, N.; Ravier, F.; Sanchez, J. C.; Frutiger, S.; Hochstrasser, D. *Electrophoresis* **1993**, *14*, 1023−1031.

[43] Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.

[44] Wierenga, S. K.; Zocher, M. J.; Mirus, M. M.; Conrads, T. P.; Goshe, M. B.; Veenstra, T. D. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1404−1408.

one lysine and one arginine would look like a perfect tryptic peptide, even though there is a missed cleavage). Making this correction, the true tryptic efficiency that was measured was ~80%, which is more in line with the efficiency seen from other laboratories.

Therefore, having a large number of non-"unique" tryptic peptides can significantly affect the accuracy of accurate mass-only-based identifications. Furthermore, even if trypsin possessed 100% efficiency, there would still be masses in the given mixture/spectra from compounds other than unique tryptic peptides such as modified peptides, non-tryptic peptides (generated from cellular proteases), unpredicted/annotated peptides, and other non-peptide sources (chemical noise that possesses an isotopic distribution similar to that of a peptide). This places an additional burden on the accurate mass approach alone as an identification criterion. To accurately distinguish between true unique peptides and this noise, we can define the specificity of the accurate mass-only approach as the percentage of random noise that does not match unique tryptic peptides. Thus, in this case, specificity is equal to 100% minus the false positive rate. First, it would be best to address what is an acceptable amount of specificity or false positive rate. It should be noted here that the amount of specificity does not indicate the experimental false positive rate, but it does give an estimate of what the relative experimental false positive rate would be for a given amount of non-predicted peptides.

For this work there are essentially two false positive rates considered: the false positive rate of the peptide identifications ($\mathrm{FPR}_{\mathrm{peptide}}$) and the false positive rate of the protein identifications ($\mathrm{FPR}_{\mathrm{protein}}$). The theoretical peptide false positive rate will be referred to as specificity, and $\mathrm{FPR}_{\mathrm{peptide}}$ will refer to the experimentally obtained percentage of false positives. The specificity can be measured directly by looking at the ability of the identification approach to distinguish between true peptide hits and the noise or background random matches. To determine what an acceptable specificity is for these types of identification strategies, one must first define the acceptable $\mathrm{FPR}_{\mathrm{protein}}$. The $\mathrm{FPR}_{\mathrm{protein}}$ for protein identifications is slightly more complex and is simply the combined probability that all the peptide identifications for that protein are false for *n* number of peptides of a specific protein:

$$\mathrm{FPR}_{\mathrm{protein}} = \mathrm{FPR}_{\mathrm{peptide}(1)} \times \mathrm{FPR}_{\mathrm{peptide}(2)} \cdots \times \mathrm{FPR}_{\mathrm{peptide}(n)} \qquad (1)$$

Table 1 examines the $\mathrm{FPR}_{\mathrm{protein}}$ identifications as a function of the $\mathrm{FPR}_{\mathrm{peptide}}$ identifications and the number of peptides used to identify a protein (as described in eq 1). For this work, the most important issue is that the $\mathrm{FPR}_{\mathrm{protein}}$ identification should be ~1% or less. Thus, for single peptide identifications the $\mathrm{FPR}_{\mathrm{peptide}}$ should be 1% and for two peptides per protein identified the $\mathrm{FPR}_{\mathrm{peptide}}$ should be ~10% or less. It is quite possible to argue that, for higher peptide-to-protein ratios, the $\mathrm{FPR}_{\mathrm{peptide}}$ could be allowed to fall much lower, but for a truly high-throughput proteomics study that utilizes the shotgun strategy, very low peptide-to-protein ratios should be expected. Analyzing the amount of specificity needed by accurate mass alone as an identification criterion, a peptide false positive rate of 1% and 10% translates into a specificity of 99% and 90% for one- and two-peptide hits, respectively.

Table 1. $\mathrm{FPR}_{\mathrm{protein}}$ Identifications as a Function of the $\mathrm{FPR}_{\mathrm{peptide}}$ Identifications, and the Number of Peptides (from 1 to 5) Used To Identify a Protein (As Described in Eq 1)

		$\mathrm{FP}_{\mathrm{protein}}$ rate (%)				
$\mathrm{FP}_{\mathrm{peptide}}$ rate (%)	probability	1 pept./protein	2 pept./protein	3 pept./protein	4 pept./protein	5 pept./protein
1	0.01	1	0.01	0.0001	0.000001	1×10^{-8}
5	0.05	5	0.25	0.0125	0.00062	3.13×10^{-5}
10	0.10	10	1.00	0.100	0.01	0.001
15	0.15	15	2.25	0.337	0.051	0.0076
20	0.20	20	4.00	0.800	0.16	0.032
30	0.30	30	9.00	2.700	0.81	0.24
50	0.50	50	25.00	12.500	6.25	3.12

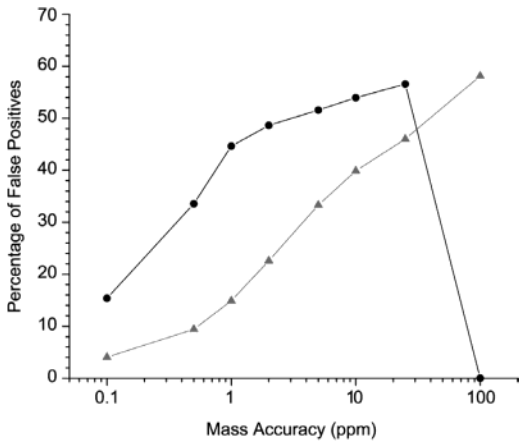

Figure 2. Percentage of false positives from the input of random masses, plotted as a function of the mass accuracy used in the experiment. For *E. coli* (▲), the data show that even at 0.5 ppm for every 10 masses, one hit comes up as a false positive match to a unique peptide. In *R. norvegicus* (●), at 0.5 ppm, one out of three random masses matches a unique peptide. Matches to non-uniques (masses that could be multiple peptides for the given mass accuracy) were discarded. At 100 ppm, all peptide matches are nonunique and a false positive rate cannot be calculated.

To address the issue of actually measuring the specificity of an accurate mass-based approach, a series of random tryptic peptides were generated and searched against a unique tryptic peptide database for various organisms at various levels of mass accuracy (see Figure 2). This approach to measure the specificity of the identification strategy determines how often a mass will appear as a unique match. Analysis of the *E. coli* genome shows that even at 0.5 ppm mass accuracy a specificity of 92% is achieved, and thus the use of less mass accuracy is problematic. Therefore, for any organism that has a proteome of approximately the size of *E. coli*, the use of mass accuracy at less than 0.5 ppm is likely to result in an unacceptable false positive rate because the chance that a random peptide mass will appear to be unique is moderately high. For the *R. norvegicus* genome, it is impossible at any mass accuracy to identify a protein with an acceptable specificity with two or less peptides. With three peptides per protein at a mass accuracy of 0.1 ppm, some *R. norvegicus* proteins could be identified, but to obtain a more reasonable mass accuracy (such as 1−2 ppm), then at least five peptides per protein would be required. The primary problem with requiring such a high number of peptides per protein is that for any measurement made at the peptide level, such as quantitation or synthesis/degradation ratio,[28] it becomes difficult to determine the peptide(s) that are real matches for a given protein since they cannot be separated from the false positives until a very large number of peptides for that protein are observed. Given the levels of mass accuracy needed for even *E. coli* in order to obtain an acceptable false positive rate, the ability of any accurate mass approach to analyze more complex organisms is limited.

Determination of False Positive Rates for Experimental Data. Although the above-mentioned strategy gives an approximate false positive rate through the analysis of specificity, it does not show how to calculate the $\mathrm{FPR}_{\mathrm{peptide}}$ for data obtained experimentally. The aforementioned Monte Carlo-type experiment measures only the average specificity for random matches for data in the mass range of 1000−3000 Da, with a random spread of the "noise" masses over that range. If one examines the specificity as a function of mass at 1 ppm mass accuracy (see Figure 3a), it becomes obvious that the assumptions made above are valid, but a more complex model of the false positive rate is needed for accurate analysis of real data. By looking at the specificity as a function of mass for *E. coli*, a distinct trend can be observed that shows the specificity increasing as the peptide mass increases. This trend is caused by the fact that at higher mass there are fewer unique masses that can be randomly matched with noise. This more specific calculation of the false positive rate becomes important when calculating the $\mathrm{FPR}_{\mathrm{protein}}$ and determining which of the matching peptides for a given protein are real. Here it should be noted that the false positive rate is equal to the number of false uniques divided by the number of unknown masses plus the number of false positives. This means that the random noise that matches non-unique peptides (a single mass that could theoretically have come from two or more peptides because the mass accuracy does not allow them to be distinguished) is not counted in the calculation of specificity. This observation is more important in the examination of the *R. norvegicus* data, because at low mass there are very few masses for tryptic peptides to match, and if it does not match a unique tryptic peptide, then it is not considered a false positive identification in this work. Examination of the specificity in *R. norvegicus* shows that the average number of false positives for any given mass is 50% or more. The primary reason that the false positive rate does not show the same sloping effect as in the *E. coli* data is that at low mass there are very few unique masses. This situation is shown clearly when the percentage of non-unique mass matches from the 1000 random peptide test at each the mass intervals is plotted, as can be seen in Figure 3b.

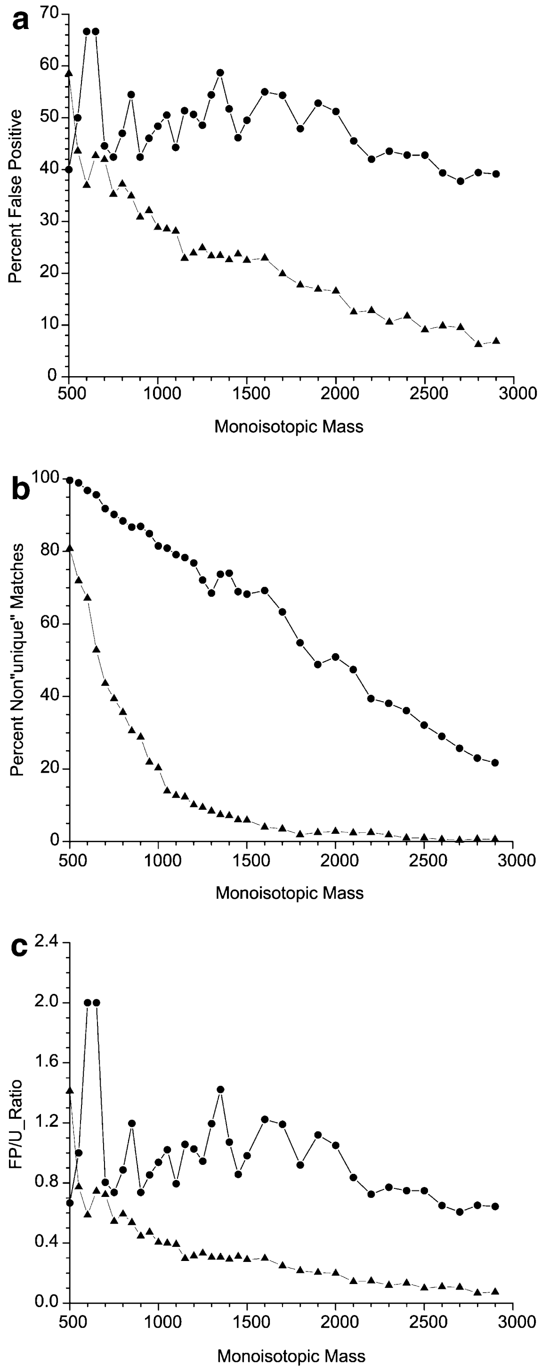

Figure 3. (a) Plot showing that the percent false positive is mass dependent in both organisms. This correlation arises from the fact that there are fewer peptides at high mass as compared to low mass. The difference between the heights of the *E. coli* (▲) and *R. norvegicus* (●) lines is also related to the relative number of tryptic peptides generated from each proteome. (b) Plot showing the percent of matches to non-unique peptides. The higher percentage of matches to non-unique peptides explains the scatter of the data in plot a by showing that few data points could be acquired for unknown or unique matches (*E. coli*, ▲; *R. norvegicus*, ●). (c) Plot showing the ratio of false positive to unknown masses as a function of mass of the peptide. This ratio is important because it will allow a false positive rate for experimental data to be determined (*E. coli*, ▲; *R. norvegicus*, ●). All plots are at a mass accuracy of 1 ppm.

Although the specificity as a function of mass is useful for assessing the ability of an approach like the accurate mass strategy to distinguish real peptides from noise, it does not truly provide a method to accurately assess the $\mathrm{FPR}_{\mathrm{peptide}}$ for real data. For instance, if a sample was composed of only real unique tryptic peptides, the specificity if used directly would suggest that a percentage of these peptides were misidentified, which would not be true for that hypothetically perfect sample. To actually calculate the experimental peptide false positive rate, one needs to determine the number of false positive matches not as a function of the number of identified peptides but as function of some other variable that is independent of the number of unique mass identifications. The reason that the number of identified peptides (IP) cannot be used for this purpose is that this number will be composed of both the real identifications (RI) and false positives (FP):

$$\mathrm{IP} = \mathrm{RI} + \mathrm{FP} \tag{2}$$

The number that can be obtained from the experimental data, which is solely dependent on the noise or random masses, is the number of unknown masses. Like the number of false positive peptides, the number of unknown masses is dependent on the number of random masses entered as a function of the database size and constraints (i.e., mass accuracy) used during the search. Currently there is no way to calculate ab initio the number of unknown mass matches or false positive matches from the knowledge of the database size, database constraints, and number of input random masses. The major limiting factors for such a calculation that would need to be considered include the percentage of lysines and arginines in the proteome under study, the amount of conservation observed between different proteins and their peptides, and the combinations of amino acids that occur at rates above or below what would be expected by chance alone. Because of the aforementioned complications and other more complex issues, such an ab initio calculation will be very difficult to perform. However, since both the number of unknown masses and the number of false positive matches are dependent on the number of random or unpredicted masses, a relationship exists between these two values. This relationship between the number of unknown masses and the number of false positives (see Figure 3c) for a given mass range can be used to determine the expected number of false positives through the ratio of false positive to unknown masses ($\mathrm{FP}/\mathrm{U}_{\mathrm{ratio}}$):

$$\frac{\mathrm{FP}}{\mathrm{U}_{\mathrm{ratio(mass\ range)}}} = \frac{\text{no. of theoretical false positive masses}}{\text{no. of theoretical unknown masses}} \tag{3}$$

This is accomplished by multiplying the number of observed unknown masses for a given mass range as determined experimentally by the $\mathrm{FP}/\mathrm{U}_{\mathrm{Ratio}}$, which has to be determined theoretically with Monte Carlo-type simulations. This value is then the number of expected false positives (E_{FP}):

$$E_{\mathrm{FP}} = \frac{\mathrm{FP}}{\mathrm{U}_{\mathrm{ratio(mass\ range)}}} \times \text{obsd. no. of unknowns}_{\mathrm{(mass\ range)}} \tag{4}$$

This E_{FP} should approximate FP from eq 2 for each mass range; thus, the experimental $\mathrm{FPR}_{\mathrm{peptide}}$ for each mass range can be determined via eq 5:

$$\mathrm{FPR}_{\mathrm{peptide(mass\ range)}} = \frac{E_{\mathrm{FP(mass\ range)}}}{\text{total no. of identified peptides}_{\mathrm{(mass\ range)}}} \tag{5}$$

Using this knowledge, the average $\mathrm{FPR}_{\mathrm{peptide}}$ for the total experimental dataset can be calculated as the sum of the $\mathrm{FPR}_{\mathrm{peptide}}$ for each mass range weighted by the percentage of identified peptides in that mass range:

$$\mathrm{FPR}_{\mathrm{peptide(avg)}} = \sum \frac{\mathrm{FPR}_{\mathrm{peptide(mass\ range)}} \times \text{total no. of identified peptides}_{\mathrm{(mass\ range)}}}{\text{total no. of identified peptides}_{\mathrm{(all)}}} \tag{6}$$

For calculation of the $\mathrm{FPR}_{\mathrm{protein}}$ through eq 1, the expected peptide false positive rate for the mass range of each peptide should be substituted for the $\mathrm{FPR}_{\mathrm{peptide}}$ for that peptide. Thus, two different proteins that both have two peptides per protein will have a different $\mathrm{FPR}_{\mathrm{protein}}$ value if the pairs of peptides did not come from exactly the same mass ranges. In trying to keep false positive protein identifications low, the $\mathrm{FPR}_{\mathrm{protein}}$ value for any given protein should be 1% or less. Although this system is far from perfect, it should provide a crude means for determining the false positive rates for a given dataset and make datasets that use different parameters, such as mass accuracy or databases differences, comparable.

Use of Isoelectric Point with Accurate Mass To Improve the Accuracy of the Identifications. Because of the limitations of accurate mass-based approaches for peptide identifications, other methods to acquire information on the identity of the peptide have been explored. The method encompassing the remainder of this work is that of combining isoelectric focusing and the utilization of p*I* with accurate mass for peptide and protein identification. Initially, the percent of unique tryptic peptides as a function of p*I* and mass accuracy was explored for both the *E. coli* and *R. norvegicus* proteomes (assuming no missed cleavages). For the *E. coli* proteome, the majority of the combinations of p*I* and accurate mass provide high average percentages (>50%) of unique peptides for the peptides in the mass range of 1000−4000 Da (see Figure 4a). Here the exceptions are primarily at the regions of poor p*I* predictability (>1 p*I* unit) and mass accuracy (>10 ppm) levels. At the experimentally relevant levels of 1 ppm and 0.25 p*I* unit, more than 85% of the proteome is unique in the aforementioned mass range. At the 1 ppm mass accuracy level with no p*I* prediction, only 65% of the proteome is unique. This represents a significant increase in the number of possible identifiable peptides by combining the currently achievable p*I* predictability with high levels of mass accuracy. The significance of improving the p*I* prediction algorithm and the power of using p*I* as a database constraint are examined by looking at the two extremes of p*I* prediction at a mass accuracy of 100 ppm (the lowest considered here, and routinely achievable by many commercial instruments). With no p*I* prediction, the percent of unique peptides is 0.15% considering mass accuracy alone. With a 0.01 p*I* unit accuracy, the percent of unique peptides is 59%. Although 0.01 p*I* unit seems like an extremely unrealistic level of isoelectric resolving power, current narrow-range IPG strips come close to, if not equal to, this level of resolution. Thus, the true need for improvement will lie mainly with the prediction algorithms.

Examination of the percent of unique peptides for the *R. norvegicus* data (see Figure 4b) shows trends similar to those observed with the data for *E. coli*. At 1 ppm mass accuracy, with no p*I* prediction, the percentage of unique peptides is only 16% for the 1000−4000 Da mass range. By adding in the p*I* prediction at 0.25 p*I* unit, the percent of unique peptides increases to 52%.

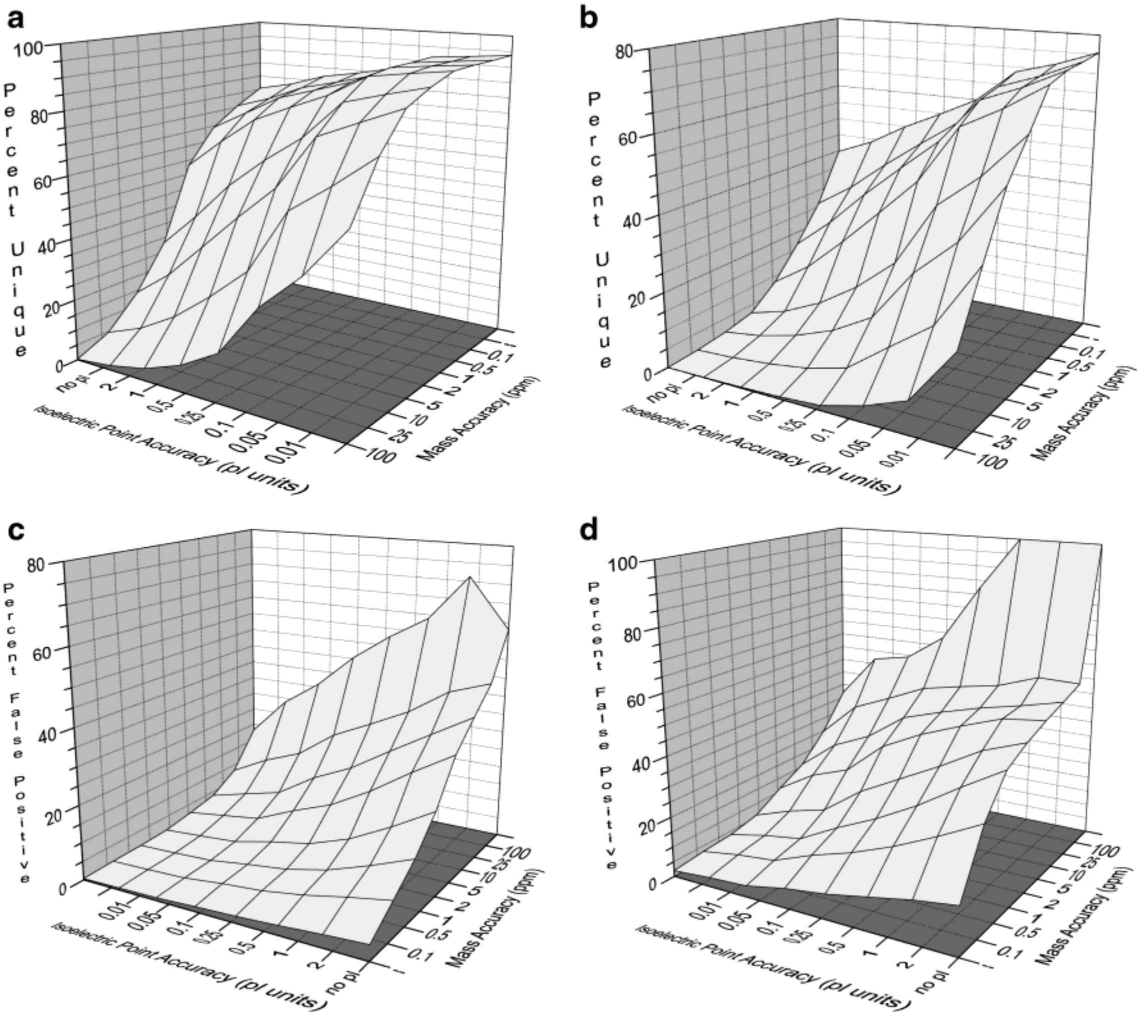

Figure 4. (a) Plot showing the percentage of unique peptides as a function of both the mass accuracy and the predictability of p*I* for *E. coli*. (b) Graph demonstrating the percentage of unique masses in *R. norvegicus* as a function of mass accuracy and the predictability of p*I*. (c) Examination of the false positive rate for random tryptic peptides (in the mass range of 1000−3000 Da) searched against the *E. coli* proteome at various levels of mass accuracy and p*I* . (d) Same search used in panel c, but utilizing the *R. norvegicus* proteome as the database. (Scale on graphs is linear for better view of the 3D surfaces.)

This substantial increase can be attributed to the fact that the peptides from the *R. norvegicus* proteome saturate many of the masses over the majority of the mass range. Again, looking at the effect of using p*I* at 100 ppm mass accuracy shows that, with no p*I* prediction, essentially no peptides can be identified and, with optimal p*I* predictability (0.01 p*I* unit), approximately 23% of the peptides are unique. In the rat, the maximum percentage of unique peptides is only 78% (at 0.01 p*I* unit and 0.1 ppm), and in *E. coli*, the maximum is 94%, which is much closer to being able to identify all the peptide from that organism in the aforementioned mass range.

The true test of each database search routine comes from an examination of the specificity or the theoretical false positive rate. The specificity for both the *E. coli* and *R. norvegicus* proteome at various levels of mass accuracy and p*I* predictability is shown in Figure 4, parts c and d, respectively. For *E. coli*, the specificity at 1 ppm and 0.25 p*I* unit is 97%, which means that the number of false positive protein identifications at this level would be extremely low at 0.09% (for two peptide per protein identifications). Using the cutoff of 90% specificity and a p*I* of 0.25 would allow mass accuracy of greater than 10 ppm to be acceptable for identification of *E. coli* proteins. Decreasing the level of uncertainty in p*I* prediction to 0.1 unit would allow mass accuracy close to 20 ppm to be used. Thus, a wider range of commercial instruments could be used for these analyses, such as many electrospray ionization (ESI) or matrix-assisted laser desorption/ionization (MALDI) time-of-flight instruments rather than the restriction to FT-ICR that is currently in place. This improvement is quite considerable compared to the use of accurate mass alone, which requires about 0.5 ppm mass accuracy to have an acceptable false positive rate. Analysis of the *R. norvegicus* proteome at a specificity of 90% requires a p*I* predictability of 0.1 unit at a mass accuracy of 1 ppm, or a p*I* predictability of 0.25 unit at mass accuracy of 0.5 ppm, to be used. At 0.25 p*I* unit and 1 ppm, the percentage of false positives for peptides is 15%. With three peptides per protein, then the $\text{FPR}_{\text{protein}}$ becomes 0.3%, which while acceptable is far from optimal because of the number of peptides needed for a high level of assurance of protein identification. In contrast, for *R. norvegicus* at a mass accuracy of 0.1 ppm, the specificity is 85%, and at 0.5 ppm it is closer to 66%. Although one could arguably use the 85% specificity with three or more peptide identifications, a specificity of 66% is unacceptable except at very high numbers

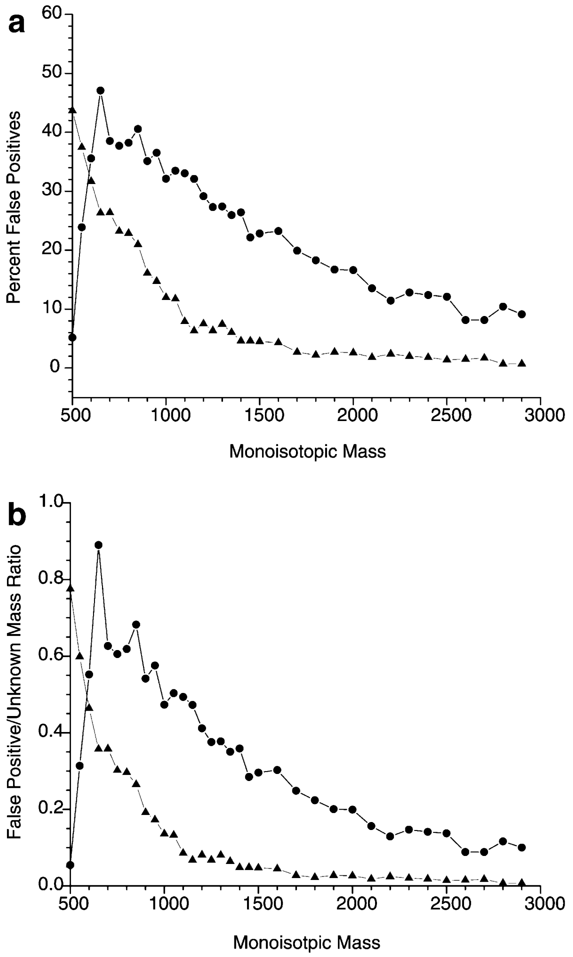

Figure 5. (a) Trend between the mass of a peptide and the false positive rate for random tryptic peptides, shown using a mass accuracy of 1 ppm and ±0.25 p*I* unit. The trends for both the *E. coli* (▲) and *R. norvegicus* (●) proteomes are shown to demonstrate the effect of proteome/genome size. (b) Trend between the ratio of false positives to unknown masses when random tryptic peptides are entered. For *E. coli* (▲), above mass 1000 Da the ratio is large enough that 10−20 unknown masses must be seen before an false positive would be expected. In constrast, the *R. norvegicus* (●) proteome is so large that significant portions of the random tryptic peptide match unique peptides rather than showing up as unknown masses.

Table 2. Series of Isomeric Peptides in the *E. coli* Proteome with Differing Isoelectric Points (p*I*), Demonstrating the Need for p*I* as an Identification Criterion

mono-isotopic mass	elemental composition	p*I*	sequence
503.3141	$C_{23}H_{45}N_5O_5S$	8.75	LLMK
503.3141	$C_{23}H_{45}N_5O_5S$	8.50	MLLK
932.4636	$C_{39}H_{68}N_{10}O_{14}S$	4.37	QDAEVIMK
932.4636	$C_{39}H_{68}N_{10}O_{14}S$	5.66	TAEMTGVPK
920.4966	$C_{42}H_{68}N_{10}O_{13}$	5.84	LAYQLADK
920.4966	$C_{42}H_{68}N_{10}O_{13}$	8.47	SFGAPTITK
1007.4745	$C_{44}H_{69}N_{11}O_{14}S$	4.37	MFDPETLR
1007.4745	$C_{44}H_{69}N_{11}O_{14}S$	6.00	NYNEMLPK
1106.5177	$C_{47}H_{74}N_{14}O_{15}S$	5.83	DGSGIWTICR
1106.5177	$C_{47}H_{74}N_{14}O_{15}S$	6.75	LSFETSMHR
1214.6505	$C_{52}H_{90}N_{14}O_{19}$	4.53	QALLNLEAESK
1214.6505	$C_{52}H_{90}N_{14}O_{19}$	5.84	ISSQTLLGPDGK
1349.6364	$C_{60}H_{87}N_{17}O_{19}$	6.75	YAHIGTGNFNEK
1349.6364	$C_{60}H_{87}N_{17}O_{19}$	4.37	FNGWELDINSR
1421.6786	$C_{60}H_{95}N_{17}O_{23}$	4.37	DSEVSFLTPSGQR
1421.6786	$C_{60}H_{95}N_{17}O_{23}$	4.03	ELNDDSTVNFLR
1574.8336	$C_{67}H_{118}N_{18}O_{23}S$	6.00	IGAVLSEVASGVMNTK
1574.8336	$C_{67}H_{118}N_{18}O_{23}S$	4.53	LASSLSLAECELLAR
1791.9113	$C_{76}H_{125}N_{23}O_{27}$	5.26	TLELHADGTLTTEVHR
1791.9113	$C_{76}H_{125}N_{23}O_{27}$	5.97	VLNTEAATLTSQFNQR
2168.0457	$C_{96}H_{149}N_{23}O_{32}S$	4.17	QCEVFLDPHDPSVIEEALK
2168.0457	$C_{96}H_{149}N_{23}O_{32}S$	3.77	DWIDYLASTDMGIVLVSDR

of peptides per protein, as described above. These values represent only the average specificity of the peptides for peptides in the range of 1000−3000 Da.

As was shown with accurate mass alone, there is a dependency on mass that must be considered to truly establish an accurate false positive rate. The relationship between mass and the specificity is shown for *E. coli* and *R. norvegicus* in Figure 5a. As observed before, there is a distinct trend that as the mass of the peptides increases, the percent of false positives decreases, which is directly related to the density of peptides in each mass range. This trend also holds true when comparing *E. coli* to *R. norvegicus* at any given mass. The percentage of false positives calculated in *E. coli* is lower because there are fewer peptides in the proteome than are present in *R. norvegicus*. The only section of the mass range where this trend does not hold true is at low mass (around 550 Da), and at this mass the statistics to determine this ratio are poor because the majority of the input random masses match to non-unique peptides. As mentioned earlier, the more useful information is the ratio of false positives to unknown masses (see Figure 5b). This ratio is needed to establish an experimental $FPR_{peptide}$. Examination of this plot shows a trend similar to that observed for the the percent false positives for rat. Above a mass of 1200 Da, the average ratio is about 1:3, which means that for every three unknown masses from an experimental spectrum, there should be one false positive peak. Comparing this ratio to an average of 1:1 for all masses below 2000 Da for the accurate mass plot (see Figure 3c) demonstrates the benefit of adding p*I* as an identification criterion. The benefits of using p*I* become more substantial at high mass, as shown by the *R. norvegicus* proteome's ratio of false positives to unknown masses, which becomes 10:1 when using 0.25 p*I* unit but is only 1.5:1 when using mass accuracy alone at a mass of ~2800 Da.

Unique Benefits of Isoelectric Point for Peptide Identification. An additional benefit of using a peptide's isoelectric point rather than a higher level of mass accuracy for peptide identification is the ability, with isoelectric point, to differentiate isomers. Table 2 shows a few isomeric peptides from the *E. coli* proteome that differ significantly in their isoelectric point, but that could never be distinguished by mass measurements only. Many of these peptide pairs have the same elemental composition but different amino acids, which one would expect to allow them to be resolved by their differing chemical properties. For instance, QDAEVIMK and TAEMTGVPK have identical elemental composition ($C_{39}H_{68}N_{10}O_{14}S$) and monoisotopic masses. However, QDAEVIMK has two acidic amino acids, aspartic acid and glutamic acid, and consequently an isoelectric point of ~4.37. The sequence TAEMTGVPK has only one acidic amino acid, glutamic acid, and consequently has an isoelectric point of 5.66. A number of other isomeric peptides that have different amino acid compositions, and thus different isoelectric points, are listed in Table 2. These peptides could never be identified on the basis of accurate mass alone. One of the more surprising features of the isoelectric point calculation that has been seen before[42] and examined in more detail recently[41] is the fact that the order and position of the amino acids in the peptide sequence influence the isoelectric point to some degree. One example shown in Table 2 is the two peptides LLMK and MLLK. These peptides possess the same elemental composition and amino acid composition, with only the ordering of the amino acids being different. The predicted isoelectric points for these peptides are 8.75 and 8.50, respectively. Examination of the p*I* prediction algorithm[42] shows that the algorithm takes into account the effect of the N- and C-terminal amino acids on the N-terminal amine and C-terminal carboxylic acid. Additional data[41] recently found through the use of immobilized pH gradient strips have elucidated other amino acid moieties that cause shifts in the expected isoelectric point both on the termini and for consecutive amino acid pairs. These new data have yet to be incorporated into the p*I* prediction algorithm. This applied correction should reveal additional peptides that have identical amino acid compositions but different amino acid order and p*I*'s, as well as improve the overall accuracy of p*I* prediction.

CONCLUSIONS

Under the current dogma of peptide identification, a tandem mass spectrum of each component in a complex mixture of peptides must be acquired in order to identify that component. With this approach, there is no guarantee that the information in the MS/MS spectrum will be sufficient for peptide identification. Factors such as signal-to-noise of the resultant MS/MS spectrum, lack of sequence coverage for a given peptide, the increased time constraints placed on the analysis, and the various limitations specific to many identification algorithms can make the peptide identification process difficult for complex sample sets. In contrast, the use of accurate mass eliminates these aforementioned problems inherent to tandem mass spectrometry and can, in principle, reduce the minimal number of spectra required to a single mass scan. However, there are limitations inherent to the use of accurate mass alone.

We have examined these issues in detail here for the accurate mass approach and found that, for organisms larger in proteome size than *E. coli*, additional peptide identification criteria would be required to ensure a sufficiently low false positive rate. As an alternative, we have proposed the use of isoelectric point with accurate mass measurements to increase the specificity of peptide identifications. In the examples given, combining the currently achievable level of p*I* prediction with a mass accuracy of 10−25 ppm will allow the analysis of organisms with genome sizes similar to *E. coli* or smaller on commercially available time-of-flight instruments. Furthermore, combining the same p*I* levels with the mass accuracy of FTICR will allow analysis of complex multicellular organisms. The one caveat to these approaches will be the number of unpredicted peaks that arise from non-unique peptides. We are of the opinion that the data presented here compensate for these unknown masses with the analysis of the specificity of the approach. However, if a large number of non-predicted compounds arise in the spectra, then the false positive rate could be so high that the value of the data is significantly reduced. The equations provided for calculation of experimental false positive rate allow for the determination of the validity of the data. The false positive issue examined in detail here is not restricted to accurate mass approaches, but has application to those investigators employing tandem mass spectrometry approaches as well. With consideration of the aforementioned benefits, the combination of isoelectric point with accurate mass measurements can provide for a high-throughput proteomics platform that could rival mRNA array technology in speed.

ACKNOWLEDGMENT

The authors thank the Internal Research and Development Program from the Research Triangle Institute for funding of this research.

Received for review October 13, 2003. Accepted October 22, 2003.

AC0352070