Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues*[S]

Jesper V. Olsen, Shao-En Ong, and Matthias Mann‡

Almost all large-scale projects in mass spectrometry-based proteomics use trypsin to convert protein mixtures into more readily analyzable peptide populations. When searching peptide fragmentation spectra against sequence databases, potentially matching peptide sequences can be required to conform to tryptic specificity, namely, cleavage exclusively C-terminal to arginine or lysine. In many published reports, however, significant numbers of proteins are identified by non-tryptic peptides. Here we use the sub-parts per million mass accuracy of a new ion trap Fourier transform mass spectrometer to achieve more than a 100-fold increased confidence in peptide identification compared with typical ion trap experiments and show that trypsin cleaves solely C-terminal to arginine and lysine. We find that non-tryptic peptides occur only as the C-terminal peptides of proteins and as breakup products of fully tryptic peptides N-terminal to an internal proline. Simulating lower mass accuracy led to a large number of proteins erroneously identified with non-tryptic peptide hits. Our results indicate that such peptide hits in previous studies should be re-examined and that peptide identification should be based on strict trypsin specificity. ***Molecular & Cellular Proteomics 3:608–614, 2004.***

Mass spectrometry (MS)[1]-based proteomics almost invariably involves the enzymatic degradation of proteins to peptides by trypsin (1). This protease has high cleavage specificity, is very aggressive, and is stable under a wide variety of conditions. Most importantly, cleaving C-terminal to arginine or lysine residues leads to peptides in the preferred mass range for effective fragmentation by tandem mass spectrometry (MS/MS) and places the highly basic residues at the C termini of the peptides. This generally leads to informative high mass y-ion series and makes tandem mass spectra more easily interpretable.

From the Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Received, February 10, 2004, and in revised form, March 18, 2004
Published, MCP Papers in Press, March 19, 2004, DOI 10.1074/mcp.T400003-MCP200

[1] The abbreviations used are: MS, mass spectrometry; ppm, parts per million; FTICR, Fourier transform ion cyclotron resonance; LC, liquid chromatography; MS/MS, tandem mass spectrometry; SIM-scan, scan-selected ion monitoring scan; IPI, International Protein Index.

When analyzing peptide mixtures by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS), a large number of fragmentation events occur. The tandem mass spectra are searched against amino acid sequence databases by one of a number of database search algorithms. The identified peptides receive a score and are combined into lists of identified proteins. A critical question in these experiments is what constitutes a reliable peptide and protein hit (2). Some laboratories save raw mass spectrometric data and interpret this raw data in all questionable cases. In some algorithms, the score is itself a probability and can be used to estimate levels of false positives (incorrect hits) and false negatives (missed hits). For other algorithms, this question has been addressed by analyzing defined mixtures of known proteins (3); or by searching in reversed databases that should not yield significant hits (4, 5). On the basis of these findings, a set of parameters for the scores is often defined that will yield a given trade-off of false positives and false negatives. Recently, more sophisticated statistical learning algorithms have been employed to estimate levels of false positives and negatives from parameters including search score, charge state, and length of peptides (6, 7).

A long-standing question in determining the reliability of peptide hits regards the occurrence of non-tryptic peptides: While trypsin is a very specific protease, it is often assumed to also cleave at other residues than arginine or lysine with a certain probability. Thus many research groups allow "non-tryptic" or "half-tryptic" peptides to match in database searches—albeit after requiring a higher identification score—whereas other groups only allow peptides generated by strict tryptic cleavage specificity.

In our group, we require tryptic cleavage specificity of potential sequence matches based on our experience with interpretation and verification of a large number of tandem mass spectra. Specifically, during the last 10 years, we have often used the peptide sequence tag algorithm for peptide identification (8). This algorithm does not require peptides to obey a certain cleavage pattern and is also able to find peptides with modifications and in the presence of sequence errors in databases. From these experiments, we have no clearly documented cases of identification of non-tryptic peptides, other than the C-terminal peptide of the protein and peptides with an N-terminal proline. (These were almost always accompanied by a tryptic peptide of extended N-terminal sequence, which was fully tryptic, indicating that the identified peptide was due to further fragmentation of a proline-

directed y-ion.) However, as peptides that were not fully tryptic were often reported in studies using ion traps instead of the quadrupole time-of-flight mass spectrometers used by our group, it was possible that different fragmentation mechanisms made non-tryptic peptides more readily observable on such instruments.

The recently developed hybrid linear quadrupole ion trap-Fourier transform ion cyclotron resonance (FTICR) (Finnigan LTQ-FT; Thermo Electron, Bremen, Germany) mass spectrometer combines fragmentation in an ion trap instrument with the ability to obtain parent mass accuracies in the low or sub-parts per million (ppm) range in the ICR part (9). This mass accuracy is about a factor 1000 higher than that normally obtained in an ion trap instrument alone and correspondingly allows more than a 100-fold higher discrimination in the identification of peptides. We therefore decided to use this mass accuracy to determine if trypsin does indeed exclusively cleave C-terminal to arginine or lysine. A complex protein mixture was enzymatically degraded, and more than 1000 peptides were identified in a sequence database search. Average absolute mass accuracies of less than 1 ppm were obtained. The only peptides of apparently non-tryptic origin—except the C-terminal peptides of the proteins and peptides seemingly non-tryptic because of database annotation issues—were peptides with an N-terminal proline. As mentioned above, these are well-known breakdown products either of acid conditions in solution or of "nozzle-skimmer" fragmentation. Thus we conclude that trypsin indeed solely cleaves C-terminal to arginine and lysine.

EXPERIMENTAL PROCEDURES

Mouse Liver Protein Preparation—A 3-wk-old male mouse was euthanized by cervical dislocation. The liver was surgically excised and snap frozen in liquid nitrogen. Frozen liver was crushed with a mortar and pestle while chilled with liquid nitrogen and resolubilized in a buffer containing 6 M urea (Invitrogen, Carlsbad CA), 2 M thiourea (Fluka, Buchs, Switzerland), 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS; Invitrogen), and protease inhibitors (Complete Tablet; Roche, Indianapolis, IN). To remove DNA and RNA, a general DNase and RNase, Benzonase (Roche), was added to the mixture on ice, followed by incubation for 30 min. The sample was then centrifuged at 10,000 $\times$ *g* in order to sediment insoluble material.

One-dimensional SDS-PAGE Protein Separation and In-Gel Digest of Mouse Liver Proteins—Protein concentration of the mouse liver fraction was determined by Bradford assay (Bio-Rad, Hercules, CA) and ~90 μg of protein was applied on a 4–12% Bis-Tris gel (Novex; Invitrogen). After staining by colloidal Coomassie (Invitrogen), the entire gel lane was cut into 10 pieces of equal size and subjected to in-gel tryptic digestion essentially as described (10). Briefly, the gel pieces were destained and washed, and, after dithiothreitol reduction and iodoacetamide alkylation, the proteins were digested with porcine trypsin (modified sequencing grade; Promega, Madison, WI) overnight at 37 °C. The resulting tryptic peptides were extracted from the gel pieces with 30% acetonitrile, 0.3% trifluoroacetic acid, evaporated in a vacuum centrifuge to remove organic solvent, then desalted and concentrated on reversed-phase C18 StageTips as previously described (11).

Nanoflow LC-MS/MS and Data Analysis—All nanoflow LC-MS/MS experiments were done on a 7-Tesla Finnigan LTQ-FT mass spectrometer (Thermo Electron) equipped with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark). The liquid chromatography (LC) part of the analytical system consisted of an Agilent Series 1100 nanoflow LC system (Waldbronn, Germany) comprising a solvent degasser, a nanoflow pump, and a thermostated micro-autosampler. Chromatographic separation of the peptides took place in a 20-cm fused silica emitter (75-μm inner diameter; Proxeon Biosystems) packed in-house with methanol slurry of reverse-phase Repro-Sil-Pur C18-AQ 3-μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) at a constant pressure (50 bar) of helium. Then 6 μl of the tryptic peptide mixtures were autosampled onto the packed emitter with a flow of 500 nl/min for 20 min and then eluted with a 90-min gradient from 4–40% acetonitrile (MeCN) in 0.5% acetic acid (AcOH) at a constant flow of 200 nl/min.

The mass spectrometer was operated in the data-dependent mode to automatically switch between MS and MS/MS acquisition. Survey MS spectra (from *m/z* 300–1500) were acquired in the FTICR with r = 25,000 at *m/z* 400 (after accumulation to a target value of 10,000,000). The three most intense ions were sequentially isolated for accurate mass measurements by a FTICR "SIM scan," which consisted of a 10-Da mass range, r = 50,000, and a target accumulation value of 50,000. These were then fragmented in the linear ion trap using collisionally induced dissociation with normalized collision energy of 30% and a target value of 2000. Former target ions selected for MS/MS were dynamically excluded for 30 s. Total cycle time was approximately 3 s. In total, 4755 tandem mass spectra were acquired in the LC experiment.

Proteins were identified via automated database searching (Matrix Science, London, United Kingdom) of all tandem mass spectra against an in-house curated version of the Mouse International Protein Index protein sequence database (IPI, versions 2.18, 40,402 protein sequences; European Bioinformatics Institute, www.ebi.ac.uk/IPI/) containing all mouse protein entries from Swiss-Prot, TrEMBL, RefSeq, and Ensembl as well as frequently observed contaminants (porcine trypsin and human keratins). Carbamidomethyl cysteine was set as fixed modification, and oxidized methionine and protein N-acetylation were searched as variable modifications. Initial mass tolerances for protein identification on MS and MS/MS peaks were 3 ppm and 0.8 Da, respectively. The instrument setting for the Mascot search was specified as "ESI-Trap."

RESULTS

Analysis of a Complex Peptide Mixture by a High Accuracy Ion Trap FTMS Instrument—In order to sequence a large number of peptides from a diverse set of proteins, we extracted proteins from a mouse liver. The protein mixture was run on a one-dimensional gel, excised in 10 equally spaced bands, and the band with the highest apparent molecular mass was prepared for proteomic analysis. Proteins were in-gel digested by trypsin, and the resulting peptide mixture analyzed by LC on-line coupled to electrospray MS/MS. The mass spectrometer used here, the Finnigan LTQ-FT, is a novel combination of a linear ion trap with a FTICR mass spectrometer (FTICR-MS or FTMS). FTMS instruments can have very high mass accuracies in the low or sub-ppm range if the number of ions admitted to the ICR cell is strictly controlled (12, 13). In the LTQ-FT this is achieved with the help of a pre-scan, which estimates the number of ions that will be accumulated in the linear trap and transferred to the FTMS part of the instrument. For peptide sequencing (MS/MS) ex-

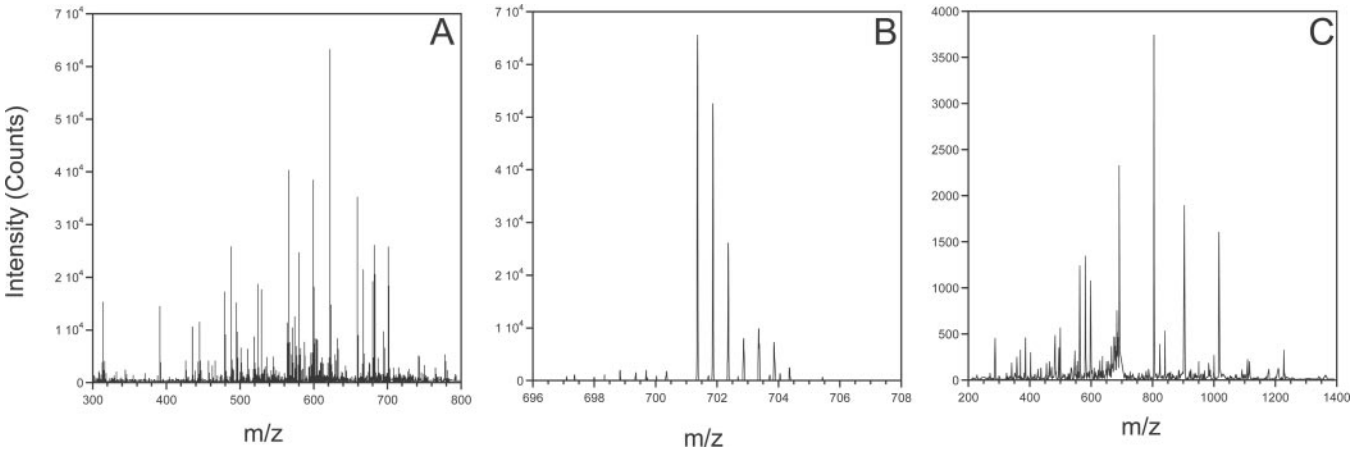

FIG. 1. **Overview of MS analysis in the LTQ-FT.** *A*, MS scan acquired in the FTICR cell for a high dynamic range survey of the total mass range. *B*, a SIM experiment of a precursor from the survey scan to obtain sub-ppm mass accuracy. *C*, the same precursor ion is isolated and fragmented in the linear ion trap to obtain sequence information.

periments, the precursor ion is selectively accumulated in the linear ion trap, fragmented, and the fragments analyzed either in the linear trap or in the FTMS. We chose to analyze the peptide fragments in the linear ion trap because this requires much less accumulation time and thus allows many more peptides to be sequenced. Furthermore, this mode of operation makes our results directly comparable with a large number of proteomic studies published using ion trap instruments. To achieve the highest mass accuracy in MS mode, we included a "zoom scan function" also called "selected ion monitoring (SIM)" scan, with an *m/z* window of 10 mass units around each peptide selected for sequencing, which further limits the number of ions admitted to the FTMS instrument. In the survey scan, a resolution of 25,000 was achieved with 250-ms transients, and in the SIM scan transients of ~440 ms resulted in a resolution of 50,000. Fig. 1 displays the steps in the sequencing cycle and demonstrates the resolution and mass accuracy by way of example.

Determination of Mass Accuracies for Database Searches—Analysis of the peptide mixture led to 4755 fragmentation attempts during the 90-min gradient. We first wanted to determine the precursor mass accuracy suitable for high stringency but inclusive searches. We therefore searched the mouse proteome database (IPI, mouse; European Bioinformatics Institute) with an initial mass accuracy of 10 ppm, which we knew to be conservative from previous experience. To determine the actual mass accuracy achieved, we plotted the mass deviations between measured and calculated masses for the 163 tryptic peptides (104 unique peptide sequences) identifying the top database hit, carbamoyl-phosphate synthase. Fig. 2 shows the binned distribution of identified peptides, and the scatter of mass deviations is shown in the inset. It is apparent from the figure that 95% of peptides are within 1.5 ppm of the correct value and that none of the 163 peptides deviate by more than 2 ppm. We therefore decided to set the precursor mass window to 3 ppm, as a conservative value that should lead to inclusion of all correct peptide hits.

To identify the peptide sequences from their tandem mass spectra, we used the Mascot search engine, a widely used protein identification program, which is based on probability scores for peptide sequence assignment (14, 15). Importantly, Mascot does not take the accuracy of the precursor ion into account when calculating the probability score for the peptide match. Therefore, when using the same cut-off score as in lower mass accuracy experiments, one can use the high mass accuracy as an independent filter for correct matches. In the case of ion trap instruments, parent mass accuracies for database searching are typically set at a few Daltons. A mass accuracy of a few parts per million, therefore, constitutes more than a 100-fold higher confidence in the peptide match given the same probability score. This is because an incorrect peptide hit resulting from a search with a precursor mass window of a few Daltons only has a chance of less than one in 100 to have the correct precursor mass within a few parts per million.

Using the significance score provided by Mascot, we established a level that should lead to 99% confidence in peptide hits even without manual inspection of the spectra. For searches with tryptic peptides, two "missed cleavages" (that is, allowing up to two internal trypsin cleavage sites), and a mass accuracy of 3 ppm, this significance score was 26. Because the Mascot probability score is only an approximation, we independently tested the level of false positives by searching our data in a reverse database (that is, each polypeptide sequence is written in reverse from last residue to first) (4, 5). Only 35 out of the 4755 tandem mass spectra matched a peptide sequence with a score of at least 26. Furthermore, about half of these peptide sequences were equivalent to real peptides that were identified also in the normal, "forward" search. This leaves 17 false positives, which corresponds to 1.5% of the more than 1000 tryptic peptides identified (see below), roughly in line with the pre-

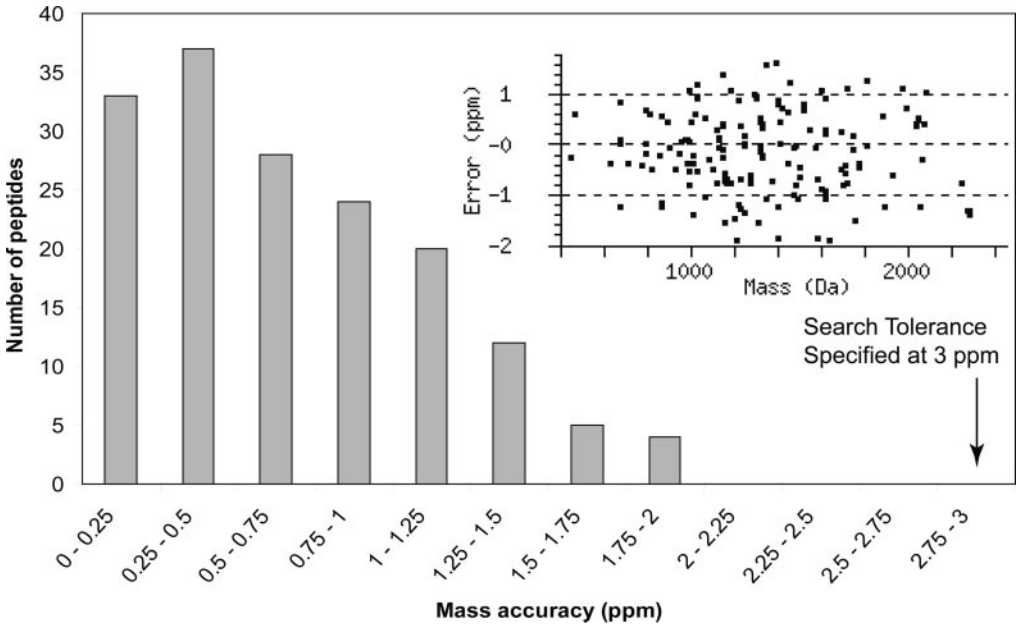

FIG. 2. **Distribution of mass deviations of the 163 peptides identifying the top-ranked protein.** Mass deviations of peptides were binned in 0.25-ppm windows. All peptides were identified with less than 2 ppm mass errors. *Inset*, Scatter plot of peptide masses as determined by the Mascot program.

dicted level of 1% incorrect hits given by the Mascot significance score.

Having established criteria that would lead to 99% correct peptide hits even without manual inspection, we then searched the mouse IPI database with the full dataset under those conditions. Of the 4775 fragmentation spectra, 1131 matched to fully tryptic peptides with at least six amino acids and an identification score of at least 26. The average absolute mass accuracy was 0.7 ppm. Of these peptides, 607 matched the top 50 proteins (Supplemental Table I), which are analyzed in more detail below.

Searches for Not Fully Tryptic Peptides—We next repeated our searches of the dataset with the parameter sets called "SemiTrypsin" or "no enzyme specificity" in the Mascot software. For the SemiTrypsin search, amino acids different from Arg and Lys are allowed at one of the termini of the peptides. This less-restrictive criterion increases the number of candidate sequences to be correlated with each tandem mass spectrum by about a factor 10. Correspondingly, a higher significance score is required. In the case of "No Enzyme" specificity, both termini of the peptide are arbitrary and the peptide can have any number of internal cleavage sites, increasing the number of candidate sequences by more than 100-fold. The significance scores for 99% certain identification for the two searches are 37 and 45, respectively. However, even with the significance score left at 26 only a total of eight and six additional peptides, respectively, matched the top 50 proteins for the two searches. These peptides are listed in Table I along with their peptide sequence, flanking amino acids, and Mascot peptide score. The fact that only 14 not fully tryptic peptides were found compared with 607 tryptic peptides already suggests that using these relaxed search parameters may not yield any great advantage for protein identification. Furthermore, close inspection of the additional peptides revealed that they likely resulted from fully tryptic peptides too: Three of these peptides were in fact fully tryptic but were not identified in the other searches because they contained three internal cleavage sites. Two apparently semi-tryptic peptides were generated from cleavages between aspartic acid (D) and proline (P) in the N terminus (for example, (D)PAKAPNSPDVLEIEFKK(G)). It is well known that the amide bond between D-P residues in peptides is the weakest peptide bond (*e.g.* acid-labile in dilute formic acid) (16) and thereby easily hydrolyzed in solution as well as in gas phase (upon collision-induced dissociation MS/MS (17) or by nozzle-skimmer fragmentation). We believe the former to be the case here as the intact tryptic peptide—(K)TQDPAKAPNSPDVLEIEFKK(G)—was also identified in our LC-MS/MS analysis but at a different retention time in the LC run. Another peptide likewise had an N-terminal proline and probably originated by the same mechanism (Table I).

One tandem mass spectrum matched a semi-tryptic peptide ((N)ALKLQK(G)) with score 42.6, however, it matched a tryptic peptide ((K)ALKLGAK(K)) with only slightly lower score (42.4), which is presumably the correct match. Finally, all except one of the remaining semi-tryptic peptides could be assigned as N-terminal peptides of proteins after removal of the known signal peptide sequences as annotated in the SwissProt-TrEMBL database at ExPaSy; viz. (Y)SEAAADREDDPNFFK(M), (G)YPSSPPVVNTVK(G), and (A)DDEVDVDGTVEEDLGKSR(E) are all adjacent C terminal to signal peptides in their respective protein sequences.

TABLE I
List of additional peptides from "SemiTrypsin" and "No Enzyme" specificity searches

The top 50 protein list from Mascot is listed along with charge states, Mascot scores, and peptide sequences for the additional peptides. Peptides found to be fully tryptic are italicized.

Protein hit no.	Accession and description	SemiTrypsin sequence	Charge	Score	NoEnzyme sequence	Charge	Score
1	Q8C196, carbamoyl-phosphate synthase	(G)ISTGNIITGLAAGAK(S)	2	53			
		(W)PANLDLKK(E)	2	48			
		(D)PNKQNLIAEVSTK(D)	2	73			
2	P16460, argininosuccinate synthase	(D)PAKAPNSPDVLEIEFKK(E)	3	31			
6	P26443, glutamate dehydrogenase	(Y)SEAAADREDDPNFFK(M)	3	28			
11	P54869, 3-hydroxy-3-methyl-glutaryl-coenzyme A synthase				*(R)GLKLEETYTNKDVDKALLK(A)*	3	28
17	Q9ET01, glycogen phosphorylase				*(R)IKKDPKK(F)*	2	34
19	Q91ZV9, carboxylesterase 3	(G)YPSSPPVVNTVK(G)	2	37			
20	P07724, serum albumin				(D)FAEITKL(A)	2	30
					(D)LGEQHFKG(L)		
44	P01027, complement C3 (HSE-MSF)	(N)ALKLQK(G)	2	42			
48	P08113, endoplasmin	(A)DDEVDVDGTVEEDLGKSR(E)	3	27	*(K)EGVKFDESEKTKESR(E)*	3	28
50	Q91YI0, argininosuccinate lyase				(D)LILYGTKEF(S)	2	33

Of the remaining four peptides, only one peptide is above the 99% significance score for these searches (see Table II). It had a score of 53, well above the significance score for semi-tryptic peptides of 37. Because it is the only such peptide out of 607, this single peptide may be a false positive, which would also be consistent with a 1% false positive rate as established with the reversed database above.

Results with Simulated Mass Accuracy Typical of Ion Trap Instruments—The high mass accuracy data obtained above led to virtually exclusive identification of fully tryptic peptides. In order to determine possible causes for not fully tryptic peptides reported in the literature, we then repeated the search using no enzyme specificity and a 2.0-Da mass tolerance typical of published ion trap experiments. Due to the large number of sequences to be compared with every tandem mass spectrum, the Mascot significance score was 65 and only 193 of the spectra met this criterion (Table II). Such strict criteria are rarely applied in practice. If, instead, peptides with a Mascot score of greater than 24 are considered significant, putative peptide matches for almost every tandem mass spectrum (4205 out of 4755 possible MS/MS queries) are obtained. Even when requiring fully tryptic peptides but searching with 2 Da instead of 3 ppm, the search adds 509 (40%) false positives (Table II).

The distribution of peptide hits to correct and incorrect proteins is visualized in Fig. 3, where we plot the incorrect peptides (*red*) from the 2.0-Da search alongside the correct peptides (*green*) from the 3-ppm search. The vast majority of additional peptide matches registers as new peptides to random protein hits throughout the list. It has been noted previously that incorrect peptide hits will tend to distribute to incorrectly identified proteins because the correct peptide hits cluster on proteins correctly identified with several peptides (5).

TABLE II
Peptide statistics for Mascot searches performed with different search parameters

These are all performed with the same mouse liver dataset comprising the 4755 MS/MS queries acquired with the LTQ-FT. The Mascot significance threshold for the respective searches is shown in brackets.

Search parameters	No. of peptides (score >99% Mascot significance threshold)	No. of peptides with score >24
Trypsin 2 MC,[a] 3 ppm	1131 (26)	1258
SemiTrypsin 2 MC, 3 ppm	709 (37)	1579
No Enzyme, 3 ppm	474 (45)	1761
Trypsin 2 MC, 2.0 Da	490 (47)	1767
No Enzyme, 2.0 Da	193 (65)	4205
Reverse db,[b] trypsin 2 MC 3 ppm	17 (26)	68

[a] MC, missed enzyme cleavage site.
[b] Reversed database search: 35 peptides with scores above 99% significance threshold. Of these, 18 are equivalent to peptide sequences identified in the forward search.

DISCUSSION

Our analysis of a complex protein mixture with high mass accuracy nanoflow LC-FTICR-MS provides a high-quality dataset that allows an unbiased and direct evaluation of the specificity of trypsin cleavage. The sub-ppm precursor ion mass accuracy improves confidence in peptide matches by more than a factor of 100 and allowed us to easily separate correct from incorrect hits. Previous studies (5, 6) have also achieved a high degree of certainty in peptide identification,

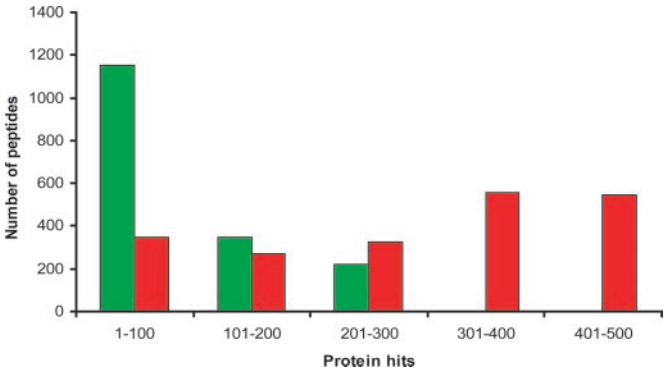

FIG. 3. **Random distribution of incorrect peptide matches.** The distributions of peptides to protein hits (bins of 100 proteins) from the 3-ppm search are shown in *green bars*. The *red bars* indicate additional peptide matches from the 2.0-Da search.

but this has not been used to investigate the existence of not fully tryptic peptides.

Trypsin belongs to the serine protease family and is very similar to chymotrypsin in primary structure. The enzymatic mechanism entails recognition of a target amino acid in a binding pocket and subsequent cleavage of the C-terminal amide bond by a mechanism involving a serine residue on the protease (hence the name). Chymotrypsin has a substrate binding pocket that preferentially recognizes bulky aromatic residues (that is, F,Y,W), whereas trypsin's substrate binding pocket is deep and narrower and has a negatively charged aspartate at the bottom of the binding pocket that binds basic amino acids via an ionic interaction. Thus target amino acids for cleavage need to have long side chains and be positively charged to allow formation of the ionic bond. Only arginine and lysine fulfill these criteria and therefore trypsin might be expected to be a very specific protease on theoretical grounds. Our study identified 607 fully trypic peptides matching to the top 50 proteins in a complex protein mixture. Only 2% half-tryptic or non-tryptic peptides matched to these 50 proteins. Upon further investigation, however, we found that these peptides also originated from specific trypsin cleavage. Some of them only appeared to be non-tryptic peptides but were tryptic with many internal cleavage sites or appeared non-tryptic because of database annotation issues such as the fact that the signal peptide in the database is not automatically removed. Others had been full tryptic at the time of digestion but degraded in-solution or by in-source decay at proline regions. Taken together our data indicate that trypsin exclusively cleaves C terminal to arginine or lysine. Even though the mechanism of trypsin cleavage readily explains such exquisite specificity, to our knowledge this report is the first to establish this fact for any proteolytic enzyme.

It has often in the past been suggested that trypsin preparations may contain some chymotryptic activity. Here we show that trypsin itself has no chymotryptic—or any other unspecific—activity. Such an activity has previously also been attributed to actual chymotryptic contamination of purified trypsin. While we have not observed this contamination here, it is possible that this problem existed several decades ago.

In practical terms, our results suggest that only peptides originating from specific trypsin cleavage be considered in database searches. In order to identify as many peptides as possible, several changes to database search algorithms could, however, be considered. First, peptides with many internal cleavage sites can still be correct hits. Search engines could accommodate this finding by allowing such peptides to match to already identified proteins. Second, peptide sequences with an N-terminal proline can also be correct and could be allowed in a search for fully tryptic peptides. It is well known that protein databases are currently not optimal for proteomic experiments. Clearly, some N-terminal peptides can be retrieved by considering the mature, fully processed form of the protein. Better isoform annotation would additionally allow retrieval of peptides that are not fully tryptic in some but not other isoforms of a protein. However, our results also suggest that all those changes would only add about 2% correctly identified peptides.

It has been known for some time that high mass accuracy can be very beneficial for database searches (18, 19), and this notion is strongly supported by these experiments. The precursor mass accuracy achieved in the experiments reported here allowed us to use database searches with 3 ppm, compared with 200 ppm frequently used for initial searches in quadrupole time-of-flight instruments. (However, the actually achieved mass accuracy in quadruple time-of-flight experiments can be as low as 10 ppm, see for example Ref. 20.) In ion trap experiments, the database search windows are typically several Daltons wide. Thus many more sequences are compared with the tandem mass spectrum, increasing the chance for spurious matches. We have shown here that the LTQ-FT combination allows more than a 100-fold increased confidence in peptide matches, which we have here used to investigate the occurrence of non-tryptic peptides, but which should also be very beneficial in any other proteomic experiment.

This study did not address the causes of reported non-tryptic peptide matches in the literature, except to indicate that low mass accuracy may have contributed. While it is theoretically possible that trypsin was nonspecific under the conditions used by other experimenters, we believe that this is an unlikely cause of these reported peptides. Likewise, non-tryptic peptides are sometimes attributed to proteases in the sample itself. We consider this to be an unlikely possibility based on our experience with a broad range of samples. This notion could be tested by processing samples without adding trypsin.

In several cases, detailed studies in less complex mixtures have been made—such as our own nanoelectrospray studies, combined with peptide sequence tag database studies—and in these cases there was little, if any, evidence for non-tryptic peptides, apart from the N-terminal proline peptides dis-

cussed above. This was also the conclusion of a recent study of 1424 manually interpreted tandem spectra of a single LC-MS/MS run on a quadrupole time-of-flight instrument (21). A more likely explanation for most of the non-tryptic or half-tryptic peptides reported in the literature may be that automated matching of a large number of tandem mass spectra to a similarly large number of possible peptide sequences from the databases does produce a certain number of convincing but spurious hits.

In summary, we have used the LTQ-FT, a state-of-the-art instrument for sub-ppm accuracy analyses of a complex protein mixture to reveal the proteolytic specificity of trypsin in proteomics experiments. Our data clearly demonstrates the extremely high specificity of trypsin and has important implications for proteomics researchers seeking the optimal search parameters to minimize false positives.

Note Added in Proof—To test if in-solution digests may lead to a higher proportion of unspecific digestion products, we analyzed a complex mixture but without gel separation. We found no evidence for nonspecific cleavage by trysin in this experiment, either.

Acknowledgments—We thank other members of our laboratory for help and fruitful discussions. We are grateful to Thermo Electron, especially Drs. S. Horning, R. Pesch, and A. Wieghaus, for help and tips for operation of the Finnigan LTQ-FT. We warmly thank Alexey Nesvizhskii (Institute of Systems Biology, Seattle, WA) for sharing his insights into database identifications of peptides.

* Work in the Center for Experimental BioInformatics (CEBI) is supported by a generous grant by the Danish National Research foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

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

‡ To whom correspondence should be addressed: Center for Experimental BioInformatics (CEBI), Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark. Tel.: 45-6550-2364; Fax: 45-6539-3929; E-mail: mann@bmb.sdu.dk.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
2. Baldwin, M. A. (2004) Protein identification by mass spectrometry: Issues to be considered. *Mol. Cell. Proteomics* **3,** 1–9
3. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6,** 207–212
4. Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926
5. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J., and Gygi, S. P. (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2,** 43–50
6. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
7. Anderson, D. C., Li, W., Payan, D. G., and Noble, W. S. (2003) A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2,** 137–146
8. Mann, M., and Wilm, M. S. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
9. Syka, J. E. P., Marto, J. A., Bai, D. L., Horning, S., Senko, M. W., Schwartz, J. C., Ueberheide, B., Garcia, B., Busby, S., Muratore, T., *et al.* (2004) Novel linear quadrupole ion trap/FT mass spectrometer: Performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J. Proteome Res.*, in press
10. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469
11. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
12. Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **17,** 1–35
13. Belov, M. E., Zhang, R., Strittmatter, E. F., Prior, D. C., Tang, K., and Smith, R. D. (2003) Automated gain control and internal calibration with external ion accumulation capillary liquid chromatography-electrospray ionization Fourier transform ion cyclotron resonance. *Anal. Chem.* **75,** 4195–4205
14. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
15. Creasy, D. M., and Cottrell, J. S. (2002) Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* **2,** 1426–1434
16. Han, K. K., Richard, C., and Biserte, G. (1983) Current developments in chemical cleavage of proteins. *Int. J. Biochem.* **15,** 875–884
17. Hunt, D. F., Yates, J. R., 3rd, Shabanowitz, J., Winston, S., and Hauer, C. R. (1986) Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **83,** 6233–6237
18. Jensen, O. N., Podtelejnikov, A., and Mann, M. (1996) Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Commun. Mass Spectrom.* **10,** 1371–1378
19. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882
20. Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426,** 570–574
21. Parker, K. C., Williamson, B. L., Marchese, J., Juhasz, P., and Martin, S. (2003) Guilty until proven innocent: Protein identifications based on non-tryptic peptides in trypsin peptide-based proteomics experiments are suspect. *51st Annual Conference of the American Society for Mass Spectrometry and Allied Topics*, Montreal, Canada