Improving Protein Identification Using Complementary Fragmentation Techniques in Fourier Transform Mass Spectrometry*

Michael L. Nielsen‡, Mikhail M. Savitski, and Roman A. Zubarev

Identification of proteins by MS/MS is performed by matching experimental mass spectra against calculated spectra of all possible peptides in a protein data base. The search engine assigns each spectrum a score indicating how well the experimental data complies with the expected one; a higher score means increased confidence in the identification. One problem is the false-positive identifications, which arise from incomplete data as well as from the presence of misleading ions in experimental mass spectra due to gas-phase reactions, stray ions, contaminants, and electronic noise. We employed a novel technique of reduction of false positives that is based on a combined use of orthogonal fragmentation techniques electron capture dissociation (ECD) and collisionally activated dissociation (CAD). Since ECD and CAD exhibit many complementary properties, their combined use greatly increased the analysis specificity, which was further strengthened by the high mass accuracy (≈1 ppm) afforded by Fourier transform mass spectrometry. The utility of this approach is demonstrated on a whole cell lysate from *Escherichia coli*. Analysis was made using the data-dependent acquisition mode. Extraction of complementary sequence information was performed prior to data base search using in-house written software. Only masses involved in complementary pairs in the MS/MS spectrum from the same or orthogonal fragmentation techniques were submitted to the data base search. ECD/CAD identified twice as many proteins at a fixed statistically significant confidence level with on average a 64% higher Mascot score. The confidence in protein identification was hereby increased by more than 1 order of magnitude. The combined ECD/CAD searches were on average 20% faster than CAD-only searches. A specially developed test with scrambled MS/MS data revealed that the amount of false-positive identifications was dramatically reduced by the combined use of CAD and ECD. ***Molecular & Cellular Proteomics 4:835–845, 2005.***

In the growing field of proteomics, MS has become the primary method for identification of proteins expressed in cells, tissues, and organisms (1, 2). Peptide mass fingerprinting emerged as the primary protein identification technique in which experimentally acquired values of peptide masses are matched against a theoretical digest of the data base in question (3). Although accurate measurement of molecular mass values assures protein identification in many cases, it only allows for reliable identification of isolated proteins of relatively simple mixtures. Furthermore, to minimize the presence of false identifications (false positives) when using peptide mass fingerprinting, a very high mass accuracy is required (4, 5).

To characterize mixtures containing many proteins, the more specific and sensitive identification method referred to as MS/MS is used. Here gaseous peptide cations are collided with inert gas molecules in what is termed collisionally activated dissociation (CAD),[1] which predominantly gives rise to informative N- and C-terminal peptide fragments (so called *b* and *y* ions) as the amide backbone bonds dissociate (6, 7). This approach combined with separation of complex peptide mixtures by LC allows for identification of hundreds of proteins in one run (8–12), but processing this enormous amount of data is not straightforward. Acquired tandem mass spectra are matched against predicted fragments of all peptides that are present in a sequence data base and match the measured molecular masses (13). These predicted peptide fragments are obtained by applying appropriate gas-phase fragmentation rules to the peptides. With the Mascot program (14) or equivalent search engine, every tandem mass spectrum is assigned a list of matching data base peptide sequences accompanied by a score representing the quality of each of these sequence identifications. In Mascot, the score is based upon the probability that the identification is a chance event, and it is calculated by the negative logarithm of this probability. Assignment of peptide scores helps to discriminate between correct and incorrect peptide assignments to spectra and facilitates detection of false-positive identifications, but still the presence of false positives remains today one of the challenging problems in protein identification (15–17). In small data sets, distinguishing between correct peptide assign-

From the Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, S-75123 Uppsala, Sweden

Received, December 16, 2004, and in revised form, March 14, 2005
Published, MCP Papers in Press, March 16, 2005, DOI 10.1074/mcp.T400022-MCP200

[1] The abbreviations used are: CAD, collisionally activated dissociation; AGC, automated gain control; ECD, electron capture dissociation; ID, identification; LTQ, linear ion trap; SIM, selected ion monitoring.

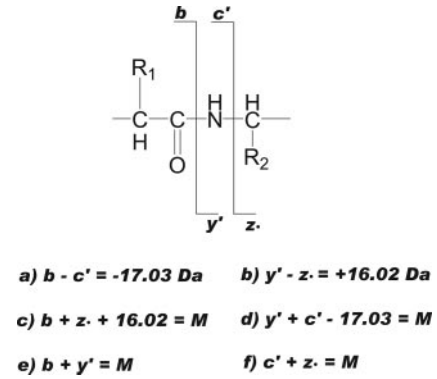

FIG. 1. **Peptide backbone fragmentation nomenclature showing the cleavage of the C–N bond (peptide bond) in CAD yielding *b*- and *y*-type fragment ions and cleavage of the N–C_α bond in ECD yielding *c*- and *z*-type fragment ions.** *a–d*, "golden complementary pair" rules (26). *e* and *f*, regular complementary pairs; these do not give any directional information.

ments and false identifications can sometimes be achieved by investigating spectra manually and verifying the peptide fragment assignment. However, such a time-consuming approach is not possible when the data of interest contains thousands of MS/MS spectra. In these cases, the approach used to eliminate false positives is to apply filtering criteria, *e.g.* to set a high enough cutoff threshold for the score. The problem arising from this approach is that the distribution of search engine scores is usually bimodal (15) with the low score component due to false positives and the higher score distribution due to true identifications. However, since the two score distributions are rarely fully separated, many correct identifications are rejected at high thresholds (false negatives), whereas lowering the threshold increases the risk of false positives. Several statistical models (18–20) have been used in attempts to decrease the amount of false positives and improve the identification score. A common feature of these approaches, however, is that they improve the validity of the assigned score, whereas the statistical nature of the score distributions requires for their separation an improvement in the quality of underlying mass spectrometric data. Recently such improvements have been proposed by several groups. Olsen and Mann (21) used MS^3 to increase the analysis specificity, whereas Gentzel *et al.* (22) used deisotoping and charge state deconvolution of the MS/MS spectra, and Venable *et al.* (23) used data-independent MS/MS to identify low abundance proteins.

Here we propose an alternative solution that is based on a combined use of the novel fragmentation technique electron capture dissociation (ECD) (24, 25) together with the traditional CAD MS/MS. In ECD, peptide fragmentation is induced by recombination of free electrons with multiply protonated peptides. This recombination induces cleavage of the N–C_α backbone bonds, giving rise to *c* and *z*˙ fragment ions (Fig. 1). These fragment ions, when compared with ions derived from CAD spectra of the same peptide, reveal the informative "golden complementary pairs" (26), which not only verify the presence of fragment ions but additionally provide the direction of the sequence by distinguishing between N- and C-terminal fragments. Since electron capture dissociation and collisionally activated dissociation exhibit many complementary properties (*e.g.* ECD does not cleave bonds N-terminal to proline residues, whereas CAD cleaves at such sites at an enhanced rate), their combined use greatly increases the analysis specificity. Regular MS/MS spectra acquired using only one fragmentation technique usually contain intervening peaks that do not necessarily arise from the fragmented peptide. These peaks may be due to contaminations, chemical noise, ion-molecule reactions, stray ions, or electronic noise. The probability that these random peaks will accidentally produce complementary pairs is low, which means that they can be filtered out using complementarity as a selection rule. The use of a complementary pair approach should reduce the complexity of the searched data and increase the specificity because only fragment ions that pass the complementarity test will be submitted to the search engine. The confidence in the protein identification should be further strengthened by the high mass accuracy (≈1 ppm) afforded by FT MS. The current work was undertaken to test these assumptions and to evaluate the benefits and the drawbacks of the complementary pair approach.

EXPERIMENTAL PROCEDURES

One-dimensional SDS-PAGE—200 μg of an *Escherichia coli* cell lysate was kindly donated by Åke Engström from the Institute for Medical Biochemistry and Microbiology (IMBIM) at Uppsala University. 70 μg of this protein mixture was loaded onto a one-dimensional SDS-polyacrylamide gel (~30–200 kDa), and the protein bands were visualized with colloidal Coomassie Blue. 20 equally sized fractions were excised from the lane, and finally the proteins in the gel pieces were reduced, alkylated, and in-gel digested with modified sequence grade trypsin (Promega, Madison, WI) as described previously in the literature (27). Finally the samples were vacuum-centrifuged to remove all organic solvents and reconstituted prior to analysis in 20 μl of HPLC water containing 0.1% TFA (Sigma).

Nanoflow LC-MS/MS—All experiments were performed on a 7-tesla hybrid linear ion trap (LTQ) FT mass spectrometer (Thermo Electron, Bremen, Germany) modified with a nanoelectrospray ion source (Proxeon Biosystems, Odense, Denmark). The high performance liquid chromatography setup used in conjunction with the mass spectrometer consisted of a solvent degasser, nanoflow pump, and thermostated microautosampler (Agilent 1100 nanoflow system). A 15-cm fused silica emitter (75-μm inner diameter, 375-μm outer diameter; Proxeon Biosystems) was used as analytical column. The emitter was packed in-house with a methanol slurry of reverse-phased, fully end-capped Reprosil-Pur C_{18}-AQ 3-μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using a pressurized "packing bomb" operated at 50–60 bars (Proxeon Biosystems). Mobile phases consisted of A (0.5% acetic acid and 99.5% water (v/v)) and B (0.5% acetic acid and 10% water in 89.5% acetonitrile (v/v)). 8 μl of prepared peptide mixture was automatically loaded onto the column and rinsed for 20 min in 4% buffer B at a flow rate of 500 nl/min followed by a 90-min gradient from 4 to 45% buffer B at a constant flow rate of 200 nl/min. Analysis was performed using unattended data-dependent acquisition mode in which the mass spec-

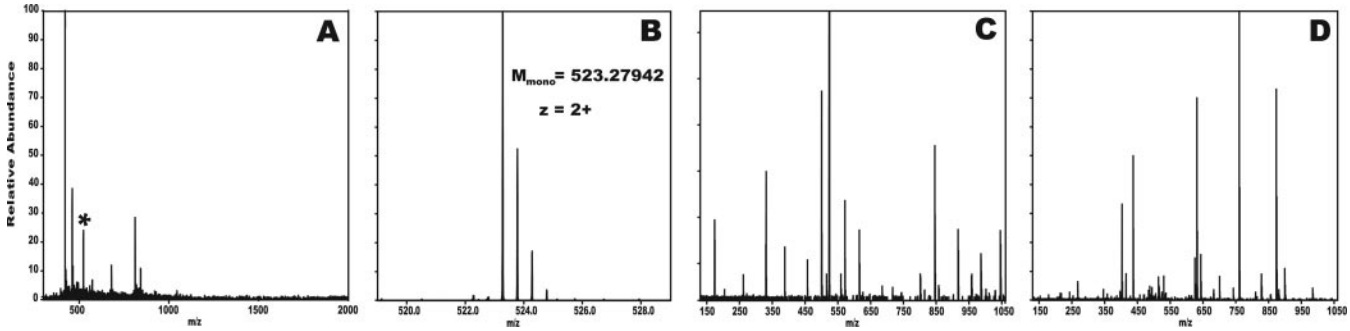

FIG. 2. **Example of the data-dependent MS analysis with LTQ FT.** *A*, survey scan acquired in the LTQ in *m/z* range 300–2000. *B*, SIM scan of the most abundant peak (*) from the survey scan yielding accurate molecular mass and charge state. *C*, the same precursor ion as in the SIM scan is isolated and fragmented by ECD. *D*, the same precursor is once more isolated and then fragmented by CAD. Scans *B–D* were recorded with the FT detector.

trometer automatically switches between a low resolution survey mass spectrum, high resolution "zoom" spectrum, and consecutive ECD and CAD fragmentation of most abundant detected peptides eluting at this moment from the nano-LC column. An illustration of the acquisition cycle is represented in Fig. 2. A survey scan was performed in the LTQ of the instrument with an automated gain control (AGC) value of 3000 (Fig. 2*A*), which is a measure proportional to the number of ions trapped during the acquisition. A zoom scan (Fig. 2*B*) was performed in the FT cell for the most abundant ions from the survey scan. Here a narrow mass window was used (±5 Da at AGC = 50,000 with resolution r = 100,000), and the charge state of the ion and its accurate mass (<2 ppm) were determined. If the ion of interest was multiply charged, consecutive ECD (Fig. 2*C*) and CAD (Fig. 2*D*) MS/MS spectra were acquired with the FT MS detector (AGC = 750,000, r = 25,000). Finally the mass of the fragmented ion was put on the exclusion list for 60 s, which greatly exceeded the typical elution time of 15 s.

As a source of electrons for performing ECD, an indirectly heated dispenser cathode (STD200, HeatWave) was installed on the backside of the FT cell. The electron irradiation event was correlated with the instrument cycle by intercepting and delaying the ion transfer trigger transmitted from the LTQ to the FT part. During this delay, electron irradiation was performed, ending prior to ion excitation and detection in the FT cell. This home-built approach did not allow for acquisition of survey scans in the FT cell, but a commercially ECD setup fixing this has been subsequently installed. The typical irradiation time was 8 ms. The maximum ion accumulation time in LTQ was 1.5 s for MS/MS and 1.0 s for MS scans; the detection time in the FT detector during CAD and ECD was ~250 ms. The longest event in the experimental sequence was the selected ion monitoring (SIM) scan that required 1 s for detection. Inclusion of the ECD event in the sequence led to extension of the sequence cycle duration by approximately 35%. All solvents used for proteolytic digestion and HPLC analysis were purchased from Tamro-Medlab, Mölndal, Sweden unless stated otherwise.

Analysis of Complex Peptide Mixture Using Complementary Fragment Pairs—An in-house written Java program was used for extraction of complementary fragment masses prior to data base searching. Briefly the software compared the peak lists (dta format) from acquired ECD and CAD spectra. First, all peaks were deconvoluted to charge state one for easier comparison followed by removal of all isotope peaks (extraction of the monoisotopic mass). The measured intensity of the fragment masses was not altered at any point during the procedure. Second, to filter out masses that cannot originate from a peptide, we introduced a mass interval (peptide window) in which the fragment masses had to reside to be considered valid. This peptide window was established by dividing the monoisotopic masses of all 20 common amino acids by their nominal mass (Table I). As can be seen from Table I, the isomeric amino acids isoleucine and leucine possess the highest value (1.000743893), and cysteine possesses the lowest (1.000089223). These highest and lowest values were used to establish the peptide window. As an example, a peptide with a nominal mass of 500 Da should have a monoisotopic mass that resides between 500.0446115 (500 × 1.000089223) and 500.3719465 (500 × 1.000743893) Da. Thus the masses 500.030 and 500.400 Da would be rejected as lying outside the peptide window.

Furthermore, whenever it was possible, the software determined the type of a fragment ion for each identified mass according to the formulas in Fig. 1. Performing CAD and ECD on the same peptide of interest will often lead to dissociation between the same pair of residues yielding *b*, *y′* ions from CAD and *c′* and *z*˙ ions from ECD (fragment notation is from Ref. 28). Thus golden complementary pairs (26) can easily be distinguished by comparing masses of the corresponding fragments determined by FT MS with high accuracy (typi-

TABLE I

List of monoisotopic/nominal mass ratios used for establishing a window in which all peptide molecular masses will appear

Highest values reached for isoleucine/leucine and smallest values for cysteine are shown in bold and underlined.

Amino acid	Molecular mass		Ratio monoisotopic/nominal mass
	Monoisotopic	Nominal	
	Da		
Ala	71.03711	71	1.000522676
Arg	156.10111	156	1.000648141
Asn	114.04293	114	1.000376578
Asp	115.02694	115	1.000234260
Cys	103.00919	103	**1.000089223**
Glu	129.04259	129	1.000330155
Gln	128.05858	128	1.000457656
Gly	57.02146	57	1.000376491
His	137.05891	137	1.000430000
Ile	113.08406	113	**1.000743893**
Leu	113.08406	113	**1.000743893**
Lys	128.09496	128	1.000741875
Met	131.04049	131	1.000309083
Phe	147.06841	147	1.000465374
Pro	97.05276	97	1.000543917
Ser	87.03203	87	1.000368160
Thr	101.04768	101	1.000472079
Trp	186.07931	186	1.000426397
Tyr	163.06333	163	1.000388527
Val	99.06841	99	1.000691010

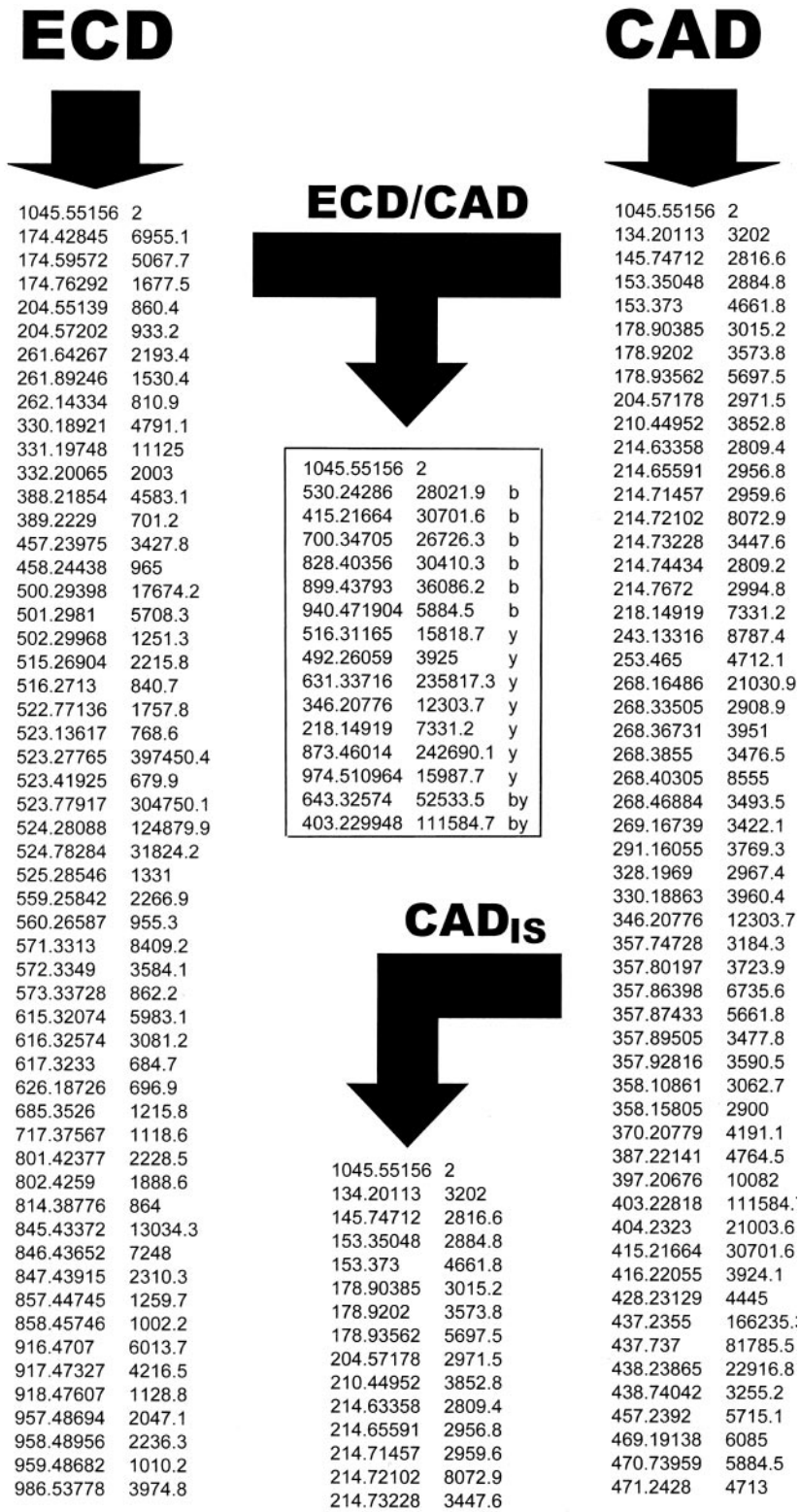

FIG. 3. **An example of data processing before data base search.** First, peak lists are extracted for both ECD and CAD spectra shown in Fig. 2*A*. Then a home-written computer algorithm performs charge and isotope reduction and identifies complementary pairs according to the rules in Fig. 1. Finally a new data file is created including only complementary fragments. If the fragment identity can be determined, it is annotated as "*b*" or "*y*"; otherwise the annotation is "*by*." Mascot does not use these annotations, but they can later be helpful in verification of Mascot fragment assignments. Note that the full length of ECD, CAD, and CAD_{IS} peak lists were not possible to depict.

cally <2 ppm for the molecular mass and <10 ppm for the fragment mass). Possible hydrogen loss from $z^{\bullet}$ ions in ECD was taken into account and addressed. When identifying *y-z* complementary pairs, the first two isotopes of each possible fragment mass were taken into consideration before isotopic removal. For fragment masses <1000 Da, the first isotope was required to have a lower intensity than the second for it to be considered as a hydrogen loss, otherwise it was considered a regular peak. An example of the application of this procedure is given in Fig. 3. Only peaks that were supported by other peaks (*e.g.* complementary pairs) in the MS/MS spectrum from the

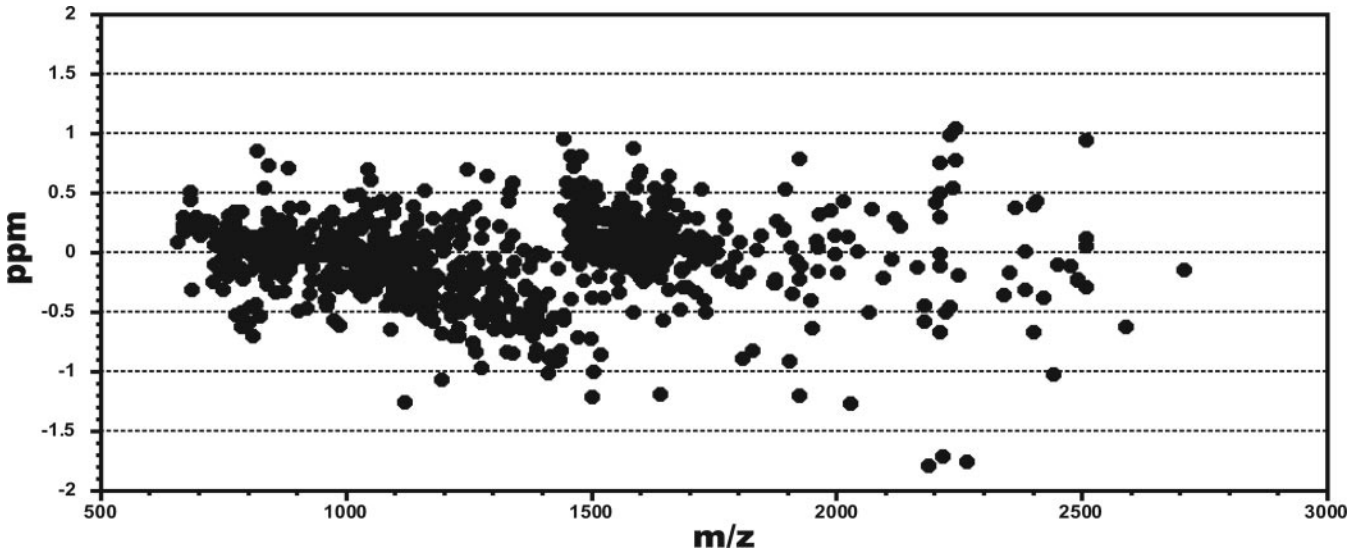

FIG. 4. **Scatter plot showing the mass deviation of 1000 well identified peptides above the statistically significant level derived from the 20 samples analyzed.** All peptides were identified with a mass accuracy of <2 ppm; 98% were identified with a mass accuracy of <1ppm.

same or complementary fragmentation technique and that uphold a monoisotopic mass within the peptide window were submitted to the data base search. The data processing was fast and required less than a second on a typical PC computer for a 60-min LC-MS/MS run containing several hundred MS/MS spectra.

Protein and peptide identification was performed through automated data base searching (14) using the Mascot search engine (Version 2.0.04, Matrix Science, London, UK). All tandem mass spectra were searched against the full NCBI non-redundant data base (www.ncbi.nlm.nih.gov; 2,082,196 sequences; downloaded October 16, 2004). Carbamidomethyl was chosen as the fixed modification, and oxidized methionine was searched as the variable modification. Searches were done with trypsin cleavage specificity allowing 2 missed cleavages; mass tolerance for monoisotopic peptide identification was set to 3 ppm and ±0.02 Da for fragment ions. Finally the instrument setting was "ESI-FTICR," which only permits b, y, $b - NH_3$, and $y - H_2O$ fragment ion types. Analysis of the search results reported by Mascot was performed using the open-source software MSQUANT (msquant.sourceforge.net).

RESULTS

High Mass Accuracy—The advantage of high mass accuracy for peptide and protein characterization is well known (29–31). To assess the performance of the LTQ FT mass spectrometer, a total number of 1000 well identified peptide sequences above the significant level of 34 were extracted, and the deviations of their masses from theoretical values were plotted as a scatter plot (Fig. 4). All peptides turned out to be identified with a mass accuracy below 2 ppm, and >98% of the peptide masses were within 1-ppm accuracy. This assured that the 3-ppm peptide mass tolerance covered all "real" peptide ions, while reducing the probability for false-positive peptide mass identification. Unfortunately Mascot does not currently allow for searching fragment ion data within the ppm range, therefore a more conservative value of ±0.02 Da was used in this study for fragment masses. Still this tolerance is much stricter than the mass tolerance of ±0.8 Da commonly used for searching MS/MS data acquired on ion trap instruments, making the search using the complementary fragment masses even more specific.

Effect of Deisotoping—It has reported that deisotoping and decharging of tandem mass spectra increases the sensitivity and specificity of a data base search (22). To separate the effect of deisotoping from the effect of the complementary pair approach, deisotoped CAD data (referred to as CAD_{IS}) were searched against the peptide data base. The results (Table II) showed that peptides from CAD_{IS} were identified with a somewhat higher average score than CAD but 12.6 Mascot points below the complementary pair technique. The average protein score increased due to deisotoping by 17% compared with regular CAD, a much lower increase than in ECD/CAD (+96%). Although deisotoping increased the number of identified peptides and even proteins compared with unfiltered CAD, it had only a minor effect compared with the complementary pair approach.

Detection Limit—To estimate the sensitivity of the complementary pair approach, we loaded 5 fmol of a BSA digest onto the nano-LC column. At this low femtomole amount, the ECD/CAD technique was able to identify four peptides above the significant level of 34 pointing to the BSA protein. Although the detection limit is higher than that demonstrated using CAD with detection in a low resolution LTQ detector (32), it is quite adequate for the majority of biology-related tasks.

Peptide Identification—Comparison of the results obtained by the fragmentation and data processing techniques used was performed by plotting the distribution of Mascot scores retrieved from the same LC-MS/MS runs (Fig. 5*A*). No peptide score cutoff was applied when performing the search. The CAD (*hatched bars*) and CAD_{IS} (*white bars*) distributions show the typical shape with the majority of peptides identified with lowest

TABLE II

Overall result in peptide identification of 20 samples from E. coli lysate achieved by using complementary fragmentation techniques compared to isotope-deconvoluted CAD and regular CAD

The average (Av.) Mascot peptide score is listed for peptide identifications with no score cutoff as well as the number of peptides identified at three different Mascot score thresholds ($T = 0$, 15, and 34). The relative change in the result is given in parentheses.

Method	Av. score	$T = 0$	$T > 15$	$T > 34$
CAD unfiltered	15.8	5485	2254	615
CAD deisotoped	18.4 (+17%)	5728 (+17%)	2634 (+17%)	848 (+39%)
Complementary pairs	31.0 (+96%)	5290 (−3.6%)	4094 (+82%)	1919 (+212%)

scores and with a rapidly decreasing number of peptides with higher scores. Peptides arising from the combined ECD/CAD search (*black bars*) show a much broader score distribution, which is shifted toward higher Mascot scores. This distribution has a maximum between 20 and 30, whereas both CAD and CAD_{IS} showed maxima between 0 and 10 Mascot score units.

Comparing the average scores (Table II), ECD/CAD gave an increase of 12.6 units compared with deisotoped CAD as already mentioned. Since the Mascot score is calculated as the logarithm ($-10 \times \log(p)$) of the probability (p) that the sequence identification by searching in that particular data base is a chance event, the 12.6 point difference indicates that the combined ECD/CAD on average identifies peptides with more than 1 order of magnitude higher confidence.

Simultaneously with the data base search result, Mascot reports a significance threshold, which is calculated from the size of the data base and depends upon how many precursor masses fall within the tolerated mass window. Searching the full NCBI non-redundant data base with the above listed parameters yielded a significance threshold (individual ions score) of 34 at which 95% of the peptide matches above this threshold are supposed to be correct. Plotting the distribution of Mascot search scores above the significant threshold of 34 clearly showed (Fig. 5*B*) the benefits of the complementary pair approach (see numerical data in Table II). These benefits are due to the highly specific extraction of complementary fragment pairs and the fact that fewer but more reliable fragment masses are submitted for data base searching.

Effect of the Charge State—To reveal the effect of the charge states (z) on the Mascot scores (M), the $M(z)$ relationship was analyzed for all identified peptides (Table III). In all cases, the vast majority of identified peptides (85–91%) were doubly charged and gave the highest average score of all charge states. The 10–12 Mascot points difference between ECD/CAD and CAD_{IS} identifications was charge-independent. A universal tendency toward lowering the Mascot score with the charge increased was probably due to the reduced sequence coverage by MS/MS cleavages for larger peptides (a linear dependence between the mass and the charge state of tryptic peptides has been established previously (33)) as well as the fact that Mascot only matches singly and doubly charged fragment ions. The overwhelming importance of 2+ ions for data base searching is a disadvantage for fragmentation techniques favoring $z \geq 3+$ charge states, such as electron transfer dissociation (34, 35).

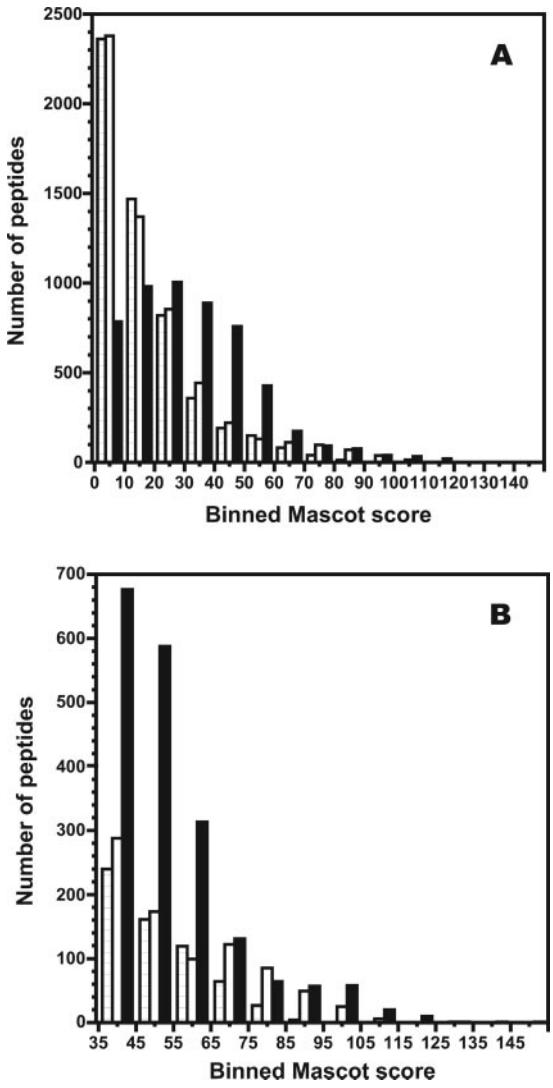

FIG. 5. **Distribution of *peptide* scores from CAD (*hatched*), CAD_{IS} (*white*), and ECD/CAD (*black*).** Peptide scores were binned into 10-Da windows. *A*, all peptides included without any score cutoff. *B*, only peptides with a score above the significant level (>34) are included.

Protein Identification—A protein is considered identified when a certain number of its peptides (usually one or two) are

identified with a significant score. In Table IV, the total number of identified proteins is listed. Table IV reflects both the case when only one peptide was required for positive protein identification and when the protein needed to be identified with a minimum of two peptides. Of course, in the second case fewer proteins were identified but with a much greater certainty. In fact, many proteomic groups have started using a two-peptide threshold as a standard requirement for protein identification by MS/MS (21, 36, 37). Following this approach, only proteins identified by a minimum of two peptides were taken into consideration in the rest of this study.

TABLE III

Result from 20 E. coli samples analyzed regarding the effect of charge state on Mascot score for the ECD/CAD approach (complementary pairs), CAD_{IS}, and CAD

The majority of identified peptides were doubly charged (84–91%), and they yield on average a higher Mascot score.

Charge state	Number of peptides	Percentage	Average Mascot score
Complementary pairs			
2	4795	90.6	31.9
3	476	9.0	22.6
4	17	0.3	15.6
5	2	0.04	10
6	1	0.02	2
CAD_{IS}			
2	4842	84.5	20.1
3	829	14.5	9.8
4	54	0.9	5.1
5	1	0.02	1
6	2	0.03	2.5
CAD			
2	4779	87.1	17.0
3	671	12.2	7.9
4	31	0.6	4.7
5	1	0.02	3
6	3	0.05	1.7

As can be seen from Table IV, complementary fragmentation techniques identified more than twice as many proteins as CAD_{IS}. This increase was due to a larger number of peptides identified at the significant level. Table V shows that the average number of peptides per identified protein also increased from 3.2 (CAD_{IS}) to 5.3 (ECD/CAD). The Mascot protein score is simply the sum of scores of the peptides identifying the same protein. Therefore, an overall increase in average protein score by 51% is not surprising (Table V). Thus, similar to peptide ID, complementary fragmentation techniques provided much higher confidence in protein identification compared with deisotoped CAD.

Analysis of the identity of identified proteins showed that 126 proteins (45%) were identified by both CAD_{IS} and ECD/CAD, and 130 proteins (47%) were identified by ECD/CAD only. CAD_{IS} alone identified 22 (8%) unique proteins.

Furthermore, since the data submitted to the search engine is reduced compared with the CAD fragment list (Fig. 3), the overall time spent for searching was also reduced. As an example, a Mascot search performed on the Pentium IV 3.26-GHz processor was accelerated by 20% (Table V).

Estimation of the False-positive Rate—The usual approach for testing the rate of false positives is to search a reversed data base (16, 38), which then provides a threshold score at which a trade-off between the number of false positives and false negatives can be accepted. Searching the ECD/CAD data in a reversed NCBI data base yielded a total number of 395 peptides above the significant threshold of 34. Of these peptides sequences, 331 were equivalent to real peptides identified in the regular "forward" data base. The large overlap between the forward and "reversed" data bases was probably due to the presence of many palindromic sequences (39, 40) and inversed sequence similarity of proteins (41). This leaves 64 peptides as true "false positives," which corresponds to 3.3% of the total 1919 peptides identified, well within the

TABLE IV

Number of unique identified proteins from CAD, CAD_{IS}, and combined ECD/CAD (complementary pairs)

Two different threshold numbers of identified peptides are listed, 1 and 2. All peptides used for protein identification had a Mascot score above the significant level of 34. A relative increase in protein identifications in the complementary pair and deisotoped CAD approach compared to CAD-only is given in parentheses.

CAD		CAD_{IS}		Complementary pairs	
1 peptide	>1 peptide	1 peptide	>1 peptide	1 peptide	>1 peptide
252	114	309 (+23%)	148 (+30%)	414 (+64%)	256 (+125%)

TABLE V

Overall results in protein identification of data from E. coli lysate achieved by using complementary fragmentation techniques compared to deisotoped CAD and CAD

The relative change in the result is given in parentheses.

	CAD	CAD_{IS}	Complementary pairs
Average protein score	169	183 (+8%)	277 (+64%)
Average number of peptides identifying a protein	3.4	3.2 (−6%)	5.3 (+56%)
Average Mascot search time (s)	459	402 (−12%)	368 (−20%)

estimated 5% incorrect identifications implied by the Mascot significance score of 34.

Using the reversed data base to estimate the rate of false positives has been shown to give good results, but the approach is not without problems. Since trypsin cleaves after both arginine and lysine, the data base reversing will change masses of many tryptic peptides. For instance, reversal of the sequence ATKINSRIGHT will give THGIR, SNIK, and TA peptides instead of ATK, INSR, and IGHT. A quick study performed on a BSA digest shows that almost 50% of all peptides in this protein will end up having different precursor masses (data not shown). Another problem is that reversing sequences does not destroy peptide sequence tags, *i.e.* fragmentation of both THGIR and IGHT peptides can produce the same partial sequence (sequence tag) "THG." Since half of all peptides do not change their mass upon sequence reversal, a combination of the same molecular mass and sequence tag will produce a hit even with the reversed data base. And although this hit will point to the same protein as the regular forward data base, it will be interpreted as a false-positive identification. Even if the search engine does not consider sequence tags, *b* ion series in the forward data base will coincide with [$y - H_2O$] series of the corresponding peptides in the reversed data base. Thus the reversed data base approach significantly overestimates the probability of a false-positive ID and therefore does not provide a satisfactory estimation of the false-positive rate. To avoid the above problems and to estimate the false-positive rate of the complementary pair approach, a new set of MS/MS data files was created that we called "rubbish." These data sets consisted of altered peak lists submitted to Mascot. In the original peak lists derived from ECD and CAD (dta format), 1 Da was randomly added or subtracted from all fragment masses, hereby creating a new data set with the same molecular mass but scrambled fragment masses. This new data set was then searched in the data base.

The search results are depicted in Fig. 6. Since deisotoping gives only minor improvements (see above), we compared raw CAD data with ECD/CAD complementary pairs. For the best evaluation of the false-positive rate, no peptide score cutoff was applied. Fig. 6*A* shows the peptide score distribution of the regular peak lists from the complementary techniques (*black bars*) *versus* the same distribution achieved by searching the rubbish peak lists (*hatched bars*). For comparison, the CAD-only search is presented in Fig. 6*B* (regular, *black bars*; rubbish, *hatched bars*).

The suggestion that the rubbish distribution mostly consists of false-positive identifications was a fair assumption because no hit was found that pointed to the same protein as the regular submission. The overlap between the rubbish and normal peptide distribution is more pronounced in Fig. 6*B* (CAD). To quantify this overlap, we normalized the rubbish distribution by the abundance of the first column of the distribution and calculated the maximum overlap between the former and latter distributions, which estimated the maximum percentage of false positives that the regular distribution may possibly contain. The outcome was that complementary fragmentation techniques gave at most 25% false positives, whereas the same value for CAD was 46%. Furthermore, looking only at peptides above the significant level, the rubbish approach identified only two false peptides, corresponding to a false ID rate of 0.1%. However, similar to the reversed data base, this approach does not estimate the rate of false positives correctly. Amino acids constitute 19 distinct masses and there-

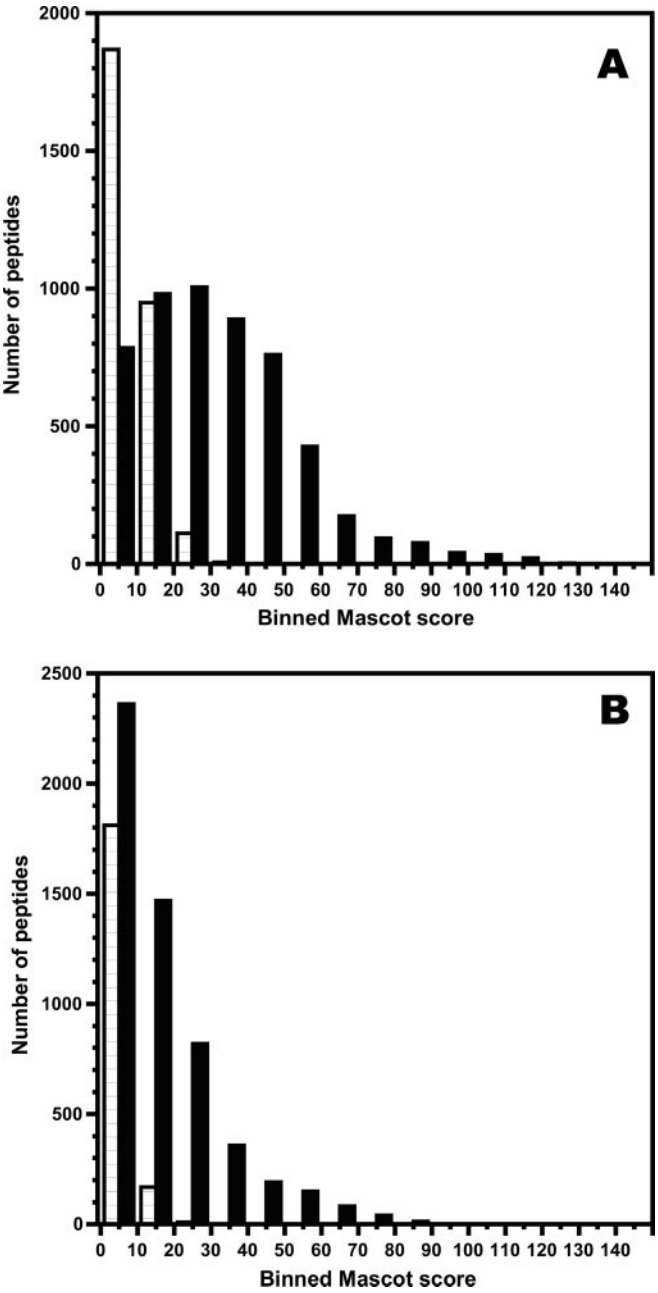

FIG. 6. **Distribution of *peptide* Mascot scores from regular and rubbish (scrambled) data.** *A*, ECD/CAD. *B*, CAD only. No peptide cutoff was applied to either of the searches.

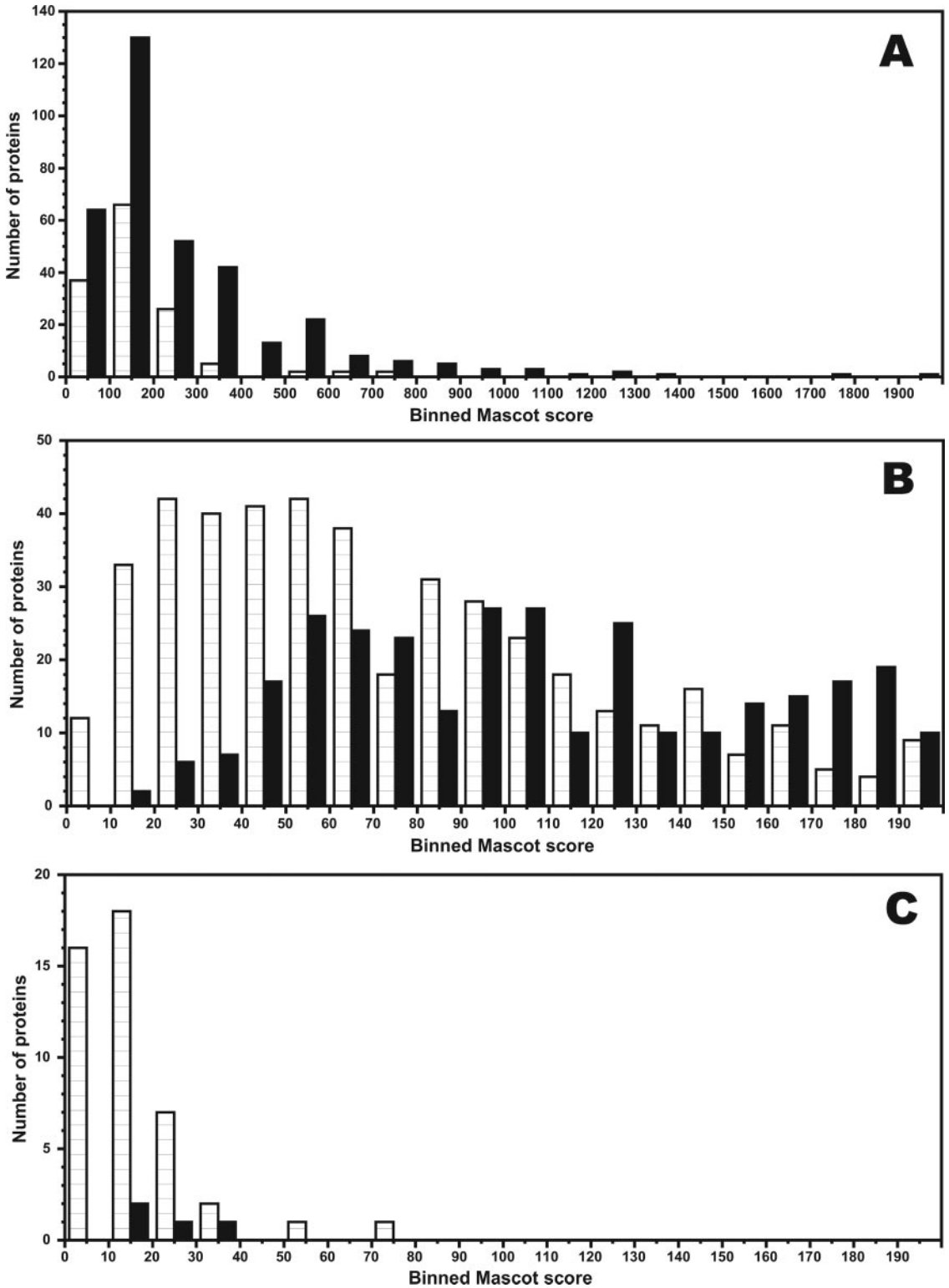

FIG. 7. **Distributions of *protein* Mascot score using the CAD and ECD/CAD approach.** No peptide score cutoff is used, but a minimum of two peptides pointing to a protein is required for the protein to be included. *A*, normal distribution for CAD and ECD/CAD in bins of 100, range 0–2200. *B*, normal CAD and ECD/CAD; data submission range, 0–200; binned into 10-Da windows. *C*, rubbish CAD and ECD/CAD; data submission range, 0–200. A total of 83 proteins was identified with CAD, whereas ECD/CAD only identified three proteins with a score below 30.

fore can only make certain masses when they are combined. By altering fragment masses of ions by ±1 Da, a mass can be created in the peak list that cannot belong to an amino acid combination. Also high scoring false positives often arise from incorrectly matched sequence tags (*e.g. y* ions are matched as *b* ions or vice versa), an effect underestimated by the rubbish approach. The real false ID rate will therefore be somewhere between the results from the two above approaches.

The protein score distribution for the rubbish data was plotted for the case when no peptide score cutoff was applied but when a minimum of two peptides pointing to the same protein was required for positive ID. As can be seen from Fig. 7*C*, the rubbish CAD data (*hatched bars*) still "identified" a total of 45 proteins, whereas the rubbish ECD/CAD data (*black bars*) gave only three protein hits. One protein in the rubbish CAD was identified with a score of $M = 74$, having 12 peptides pointing to it. Many of these peptides were also found in the regular CAD search (with significantly higher scores). The hits were due to the presence of isotopic peaks that gave matches despite ±1 Da scrambling. This result underscores the importance of deisotoping. The four detected proteins in the rubbish ECD/CAD were on the other hand only identified with two peptides each and with the total score below 40. Note that the complementary pair technique only assigned four proteins with a score lower than 34 (Fig. 7*B*, *black bars*) when no peptide score cutoff was used, whereas CAD-only data identified 83 proteins in the same range (*hatched bars*). At least some of these proteins must be due to false-positive identifications because the rubbish CAD search identified many proteins in the same scoring region. At the same time, rubbish ECD/CAD data identified only three proteins with $M \leq 34$. This means that the complementary pair approach did not produce many false-positive protein identifications, perhaps none at all, even when no peptide score threshold was used. Thus it was logical to significantly reduce this threshold when using the complementary pair approach. The number of positive identifications increased to 367 unique proteins with the significance threshold for peptides of 15. A more conservative significance threshold of 25, which lies between the Mascot-suggested value of 34 and the threshold score of 15 suggested by the rubbish approach, would be appropriate. At this threshold the complementary pair approach identifies 307 unique proteins, a 20% increase compared with the threshold of 34. The estimated false-positive rates for the rubbish and reversed data base searches were 1% and 5.6% respectively, still in line with a minimum of 95% correctly assigned peptides.

DISCUSSION

The utility of complementary fragmentation techniques in conjunction with data-dependent LC-MS/MS acquisition on a new hybrid Fourier transform ion cyclotron resonance mass spectrometer for identification of a large number of proteins has been demonstrated for the first time. This type of instrument allows for high mass accuracy, which is very beneficial in protein identification. Yet, despite this benefit, traditional MS/MS produces a measurable risk of false-positive identifications. The fragmentation techniques used in this study, CAD and ECD, yield different fragment ions that increase the specificity of the sequence information. Using this complementary information for protein ID not only improves the confidence in protein identification performed by search engines but for a fixed confidence level also identifies a larger number of peptides and proteins than when only one fragmentation technique is used. Since only complementary fragment masses are submitted, the searches are performed faster. The amount of false positives as determined by both scrambled (rubbish) MS/MS data and a reversed data base search was also reduced. It therefore became logical to decrease the threshold for peptide identification. This further increased the number of positively identified proteins, while the risk of false-positive identifications remained acceptable.

In conclusion, using complementary techniques increases both validity and specificity of protein identifications in complex mixtures. The price to pay, an overall increase of the duration of the experimental cycle by one-third, seems to be affordable compared with the obtained benefits.

Acknowledgments—We thank Thomas Köcher for help with preparing the SDS gel. Christopher M. Adams, Frank Kjeldsen, and Oleg Silivra are acknowledged for fruitful discussions. Thermo Electron (Bremen) is acknowledged for providing technical information on ECD triggering.

* This work was supported by the Wallenberg Consortium North Grant WCN2003-UU/SLU-009 (to R. A. Z.). The purchase of the LTQ FT instrument was supported by the Knut and Alice Wallenberg Foundation grant (to R. A. Z. and Carol Nilsson). 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.

‡ To whom correspondence should be addressed: Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583, S-75123 Uppsala, Sweden. Tel.: 46-18-471-5729; Fax: 46-18-471-5729; Michael.Lund-Nielsen@bmms.uu.se.

REFERENCES

1. Pandey, A., and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405,** 837–846
2. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
3. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 5011–5015
4. Zubarev, R. A., Demirev, P. A., Hakansson, P., and Sundqvist, B. U. (1995) Approaches and limits for accurate mass characterization of large biomolecules. *Anal. Chem.* **67,** 3793–3798
5. Cargile, B. J., and Stephenson, J. L., Jr. (2004) An alternative to tandem mass spectrometry: isoelectric point and accurate mass for the identification of peptides. *Anal. Chem.* **76,** 267–275
6. Hunt, D. F., Yates, J. R., 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
7. Biemann, K., and Scoble, H. A. (1987) Characterization by tandem mass

spectrometry of structural modifications in proteins. *Science* **237,** 992–998
8. Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E., and Engelhard, V. H. (1992) Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255,** 1261–1263
9. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., Yates, and J. R. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
10. Lasonder, E., Ishihama, Y., Andersen, J. S., Vermunt, A. M., Pain, A., Sauerwein, R. W., Eling, W. M., Hall, N., Waters, A. P., Stunnenberg, H. G., and Mann, M. (2002) Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542
11. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature* **419,** 520–526
12. Gygi, S. P., Rist, B., Griffin, T. J., Eng, J., and Aebersold, R. (2002) Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *J. Proteome Res.* **1,** 47–54
13. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
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. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392
16. 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
17. Cargile, B. J., Bundy, J. L., and Stephenson, J. L., Jr. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **3,** 1082–1085
18. MacCoss, M. J., Wu, C. C., Yates, J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74,** 5593–5599
19. Anderson, D. C., Li, W. Q., 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
20. 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
21. Olsen, J. V., and Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 13417–13422
22. Gentzel, M., Kocher, T., Ponnusamy, S., and Wilm, M. (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* **3,** 1597–1610
23. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R., III (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1,** 39–45
24. Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *J. Am. Chem. Soc.* **120,** 3265–3266
25. McLafferty, F. W., Fridriksson, E. K., Horn, D. M., Lewis, M. A., and Zubarev, R. A. (1999) Techview: biochemistry. Biomolecule mass spectrometry. *Science* **284,** 1289–1290
26. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 10313–10317
27. 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
28. Kjeldsen, F., Haselmann, K. F., Budnik, B. A., Jensen, F., and Zubarev, R. A. (2002) Dissociative capture of hot (3–13 eV) electrons by polypeptide polycations: an efficient process accompanied by secondary fragmentation. *Chem. Phys. Lett.* **356,** 201–206
29. Zubarev, R. A., Hakansson, P., and Sundqvist, B. (1996) Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Anal. Chem.* **68,** 4060–4063
30. 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
31. 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
32. Olsen, J. V., Ong, S. E., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3,** 608–614
33. Nielsen, M. L., Savitski, M. M., Kjeldsen, F., and Zubarev, R. A. (2004) Physicochemical properties determining the detection probability of tryptic peptides in Fourier transform mass spectrometry. A correlation study. *Anal. Chem.* **76,** 5872–5877
34. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–9533
35. Pitteri, S. J., Chrisman, P. A., Hogan, J. M., and McLuckey, S. A. (2005) Electron transfer ion/ion reactions in a three-dimensional quadrupole ion trap: reactions of doubly and triply protonated peptides with SO_2^-. *Anal. Chem.* **77,** 1831–1839
36. Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell. Proteomics* **3,** 531–533
37. Nesvizhskii, A. I., and Aebersold, R. (2004) Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* **9,** 173–181
38. 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
39. Hoffman, M., and Rychlewski, J. (1999) Searching for palindromic sequences in primary structure of proteins. *Comp. Meth. Sci. Tech.* **5,** 21–24
40. Giel-Pietraszuk, M., Hoffmann, M., Dolecka, S., Rychlewski, J., and Barciszewski, J. (2003) Palindromes in proteins. *J. Protein Chem.* **22,** 109–113
41. Preissner, R., Goede, A., Michalski, E., and Frommel, C. (1997) Inverse sequence similarity in proteins and its relation to the three-dimensional fold. *FEBS Lett.* **414,** 425–429