Rapid '*de Novo*' Peptide Sequencing by a Combination of Nanoelectrospray, Isotopic Labeling and a Quadrupole/Time-of-flight Mass Spectrometer

Andrej Shevchenko[1], Igor Chernushevich[2,3], Werner Ens[2], Kenneth G. Standing[2], Bruce Thomson[3], Matthias Wilm[1] and Matthias Mann[1]*

[1]Protein & Peptide Group, European Molecular Biology Laboratory (EMBL), D-69117, Heidelberg, Germany
[2]Department of Physics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada
[3]PE SCIEX, Concord, Ontario, L4K 4V8, Canada

SPONSOR REFEREE: Professor Alain van Dorsselaer, LSMBO, Faculté de Chimie, 1 rue B. Pascal, F-67008, Strasbourg, France

Protein microanalysis usually involves the sequencing of gel-separated proteins available in very small amounts. While mass spectrometry has become the method of choice for identifying proteins in databases, in almost all laboratories 'de novo' protein sequencing is still performed by Edman degradation. Here we show that a combination of the nanoelectrospray ion source, isotopic end labeling of peptides and a quadrupole / time-of-flight instrument allows facile read-out of the sequences of tryptic peptides. Isotopic labeling was performed by enzymatic digestion of proteins in 1:1 $^{16}O/^{18}O$ water, eliminating the need for peptide derivatization. A quadrupole / time-of-flight mass spectrometer was constructed from a triple quadrupole and an electrospray time-of-flight instrument. Tandem mass spectra of peptides were obtained with better than 50 ppm mass accuracy and resolution routinely in excess of 5000. Unique and error tolerant identification of yeast proteins as well as the sequencing of a novel protein illustrate the potential of the approach. The high data quality in tandem mass spectra and the additional information provided by the isotopic end labeling of peptides enabled automated interpretation of the spectra via simple software algorithms. The technique demonstrated here removes one of the last obstacles to routine and high throughput protein sequencing by mass spectrometry.

Received 6 May 1997; Accepted 6 May 1997
Rapid. Commun. Mass Spectrom. **11**, 1015–1024 (1997)
No. of Figures: 6 No. of Tables: 2 No. of Refs: 42

The isolation of biological factors with important functions typically results in minute quantities of proteins, separated on polyacrylamide gels. Often, the protein amount is in the subpicomole range (silver stained gels) and can only be increased by scaling up protein purification to impractical levels. The analysis of small amounts of gel-separated protein thus remains one of the most important analytical problems in molecular biology. In almost all laboratories, this task is still performed by automated Edman degradation (see for example Ref. 1).

Peptides can also be sequenced by tandem mass spectrometry and the development of instrumentation and procedures for mass spectrometric peptide sequencing has long been a focus of biological mass spectrometry.[2–4] Recently, this pursuit has obtained a new direction by the development of efficient methods to identify proteins in the rapidly growing sequence databases (reviewed in Ref. 5). In cases where the full-length sequence of the protein is known, a 'mass fingerprint'[6–10] of peptides obtained after sequence-specific digestion of the protein can now unambiguously identify the protein if these data were obtained with high mass accuracy.[11,12] In complex cases, searching sequence databases by tandem mass spectrometric information provides increased search specificity.[13,14] This even allows identification of proteins whose corresponding genes are only partially contained in sequence databases. Through such databases (i.e. the expressed sequence tag (EST) databases[15]) most human proteins already seem to be amenable to relatively straightforward characterization.[16]

Even the completion of the current genome projects of several model species in the next few years, however, will not remove the need for *de novo* sequencing of the proteins of other species. Previous work in our group has shown that it is now possible to partially sequence gel-separated proteins by mass spectrometry without the aid of databases, even at levels too low for conventional sequencing.[17–19] In our approach, unseparated peptide mixtures obtained after in-gel digestion of proteins are analysed by nanoelectrospray tandem mass spectrometry.[20,21] The experiment is repeated a second time with an esterified portion[3] of the peptide mixture to provide independent information that is needed for high confidence sequence assignment. While a series of proteins has now been partially sequenced and subsequently cloned using that strategy, it was not optimized for high throughput. Sequencing even a single protein can be very time consuming and needs considerable expertise. This problem is addressed in the current study.

Some time ago, we and others began to use isotopic labeling techniques for simplifying the interpretation of

*Correspondence to M. Mann at Protein & Peptide Group, European Molecular Biology Laboratory (EMBL), Meyerhofstr. 1, D-69117 Heidelberg, Germany

CCC 0951–4198/97/091015–10 $17.50

tandem mass spectra. Such techniques have previously been employed for the identification of the C-terminal peptide in a protein and in simplifying the interpretation of tandem mass spectra.[22–28] A label on the C-terminus of a peptide, for example, allows easy identification of the C-terminal or Y" ions[29] by their characteristic isotope pattern. Specifically, protein digestion by trypsin in the presence of ^{18}O water incorporates the ^{18}O atom selectively into the C-terminal carboxyl groups of the peptides. Subsequent fragmentation by MS/MS reveals the Y" ions by a characteristic 2 mass unit shift or split which facilitates 'read-out' of the sequence.

While this technique considerably simplifies spectrum interpretation, the use of a scanning analyser, such as the quadrupole, for read-out of the sequence ions produced by dissociation, limits the sensitivity that can be obtained. It is therefore desirable to evaluate other configurations that allow measurement of sequence ions with higher sensitivity and resolution than the triple quadrupole mass spectrometer.

The quadrupole time-of-flight instrument is such a configuration. It combines the quadrupole one (Q_1) and the quadrupole collision cell (q) of a triple quadrupole with a reflector time-of-flight analyser for the fragment ions (QqTOF). An instrument of this type has already been constructed recently.[30] (The name Q-TOF for quadrupole time-of-flight was suggested but QqTOF more accurately reflects the construction of the device (a separating quadrupole 'Q', an RF only quadrupole 'q' for the dissociation section and a time-of-flight 'TOF' analyser)). High mass resolution and mass accuracy were reported, as expected from a time-of-flight instrument. We have constructed a similar device at the University of Manitoba, using a commercial triple quadrupole (PE Sciex API 300) as the front end and replacing its third quadrupole with the Manitoba electrospray time-of-flight analyser. This analyser has previously been shown to obtain resolution of up to 10 000 full width at half maximum (FWHM).[31,32] Combined with the nanoelectrospray source and ^{18}O labeling, this instrument promises to enable rapid sequencing of small amounts of gel-separated proteins.

EXPERIMENTAL

Materials and reagents

Unless otherwise noted, all chemicals were purchased from Sigma (Sigma Chemicals, St. Louis, MO) and were analytical grade except silver nitrate, which was 'SigmaUltra' grade. MilliQ water (Millipore, Bedford, MA) was used to prepare silver and Coomassie staining solutions. For mass spectrometric analysis and gel spot preparation, HPLC grade water, methanol and acetonitrile (LabScan, Dublin, Ireland) were used. Isotopically labeled water (^{18}O, enrichment greater than 98%) was obtained from Cambridge Isotope Laboratories, Andover, MA.

The yeast proteins were obtained from a collaborating group. They were separated by one-dimensional sodium dodecylsulfate-polyacrylamide gel electrophoresis (SDS-PAGE) on a 12% acrylamide gel of 0.75 mm thickness and visualized by Coomassie Brilliant Blue staining.

The unknown protein was isolated in collaborative work with the group of Dr A. Murray (UCSF, San Francisco, USA). The protein with an apparent mass of 78 kDa was obtained by immunoprecipitation followed by one-dimensional SDS-PAGE.

A stock solution of bovine serum albumin (BSA) was quantified by amino acid analysis. Eighty fmoles of BSA were loaded onto a 12% polyacrylamide gel with a thickness of 0.5 mm. After electrophoresis, the band was visualized by silver staining as described in Ref. 33.

Micro-distillation of $H_2{}^{18}O$

Commercially available $H_2{}^{18}O$ has chemical purity 98% and is thus unsuitable for microsequencing applications. Therefore, 0.5 mL of $H_2{}^{18}O$ was purified by distillation using a sealed glass microdistillation apparatus made in house, and stored as 15 μL aliquots at –20 °C.

Sample preparation

Proteins were prepared for sequencing as previously described.[17,33,34] Briefly, they were excised from polyacrylamide gels, washed with acetonitrile and 0.1 M ammonium bicarbonate, reduced by dithiothreitol (DTT), alkylated by iodoacetamide and in-gel digested with trypsin (Boehringer Mannheim, sequencing grade). To perform ^{18}O isotopic labeling of tryptic peptides, proteins were in-gel digested using the same method except that the digestion buffer contained 50% v/v of $H_2{}^{18}O$ purified by microdistillation.

Tryptic peptides were extracted with 5% formic acid in 50% acetonitrile, extracts pooled and dried down in a vacuum centrifuge. Protein digests were reconstituted in 10 μL 5% formic acid, purified on a pulled capillary containing approximately 50 nL of POROS R2 reversed phase material (Perseptive Biosystems, Framingham, MA), and eluted by 0.5–1.0 μL of 50% methanol in 5% formic acid directly into a nanoelectrospray needle as described.[21,33] The nanoelectrospray ion source was used in all experiments. Needles were pulled and operated as previously described.[21]

Tandem mass spectra interpretation and database searching

Software was developed on the basis of AppleScript™ (Apple, Cupertino, CA) and BioMultiView™ (Sciex, Toronto, Canada). For database identification, the mass fragments with highest m/z in the Y" ion series in the MS/MS spectra of tryptic peptides were joined into a short sequence stretch. Together with the molecular weight information this stretch was assembled into a peptide sequence tag and searched against a protein sequence database using PeptideSearch version 2.9[7,35] (available via www.mann.embl-heidelberg.de). All searches were performed against a non-redundant sequence database (nrdb, updated daily by the group of Dr C. Sander at the European Bioinformatics Institute, Hinxton, UK) presently containing more than 230 000 entries. No restrictions on protein molecular weight, pI or species of origin were applied in the database searches.

The software for complete interpretation of ^{18}O labeled spectra was also developed as a set of AppleScripts. Briefly, the algorithm assembles Y" ion series on the basis of three criteria: the charge state as determined by isotopic spacing must be the same throughout the series; the mass difference between adjacent peaks must correspond to an amino acid mass within 0.06 Da; and the isotopic ratio between ^{16}O and ^{18}O must be consistent with the one observed for the parent ion. If the parent ion does not allow measurement of the isotopic ratio, an abundant fragment ion is used instead. Up to two series from high to low mass and low to high mass are extended and scored. Details of the algorithm will be reported elsewhere.

CONSTRUCTION OF THE QqTOF INSTRUMENT

From our earlier work on electrospray time-of-flight[31,32] as well as from the work of others[30] we expected a dramatic improvement in the quality of tandem mass spectrometric data when using a time-of-flight mass spectrometer as the detector for the fragment ions.

The quadrupole time-of-flight tandem mass spectrometer was constructed by combining the front section of a PE SCIEX API 300 triple quadrupole (from the ion source up to and including the collision cell), with the University of Manitoba ES-TOF spectrometer designed for orthogonal ion injection (Fig. 1). The latter instrument, described previously,[31,32] was modified so that the drift region is now floated at high voltage; this allows operation with a grounded ion storage modulator and with the quadrupoles at low offset voltage.

In brief, the tandem mass spectrometer (Fig. 1) consists of three quadrupoles Q_0, Q_1 and Q_2, two of which (Q_0 and Q_2) are always operated in RF-only mode, followed by the TOF spectrometer. For conventional mass spectrometric analysis,the mass filter Q_1 is operated in the RF-only mode, and mass analysis is performed with the TOF spectrometer. With Q_0, Q_1 and Q_2 in the RF-only mode, ions covering approximately a factor of 20 in *m*/*z* range are transmitted simultaneously into the TOF instrument. If a wider mass range is required, the RF voltage is modulated (stepped between two or more RF levels with a period of approximately 1 second) during spectrum accumulation, providing a larger transmission window averaged over time. Conventional mass spectra can be acquired either with or without collision gas in Q_2. In the former case, the collision energy is kept below 10 eV to avoid fragmentation. In this study, collision gas was used in Q_2 only for tandem mass spectra.

For tandem mass spectrometry, Q_1 is operated in mass resolving mode to select the precursor ion. Q_1 resolution was set to transmit a window approximately three mass units wide in order to transmit all of the isotopic peaks of the precursor ions (some of which were ^{18}O labeled, as described later). The energy (per charge) of ions entering the collision cell in MS/MS mode is typically from 20 to 60 eV depending on the mass and charge of the ion. The pressure in the collision cell (2) is maintained at about 10 mTorr to provide collisional damping of the fragment ions and the remaining precursor ions. As a result, the ions are thermalized, and both spatial and energy spreads are reduced, providing better transmission into and through the TOF spectrometer. After leaving the collision cell, ions are reaccelerated to an energy of approximately 8 eV per charge. A grid and a small DC-quadrupole (3) provide focusing of the ion beam through slit (4) into the ion storage modulator (5) which is initially field free. A pulsed electric field is applied at a frequency of several kilohertz across the

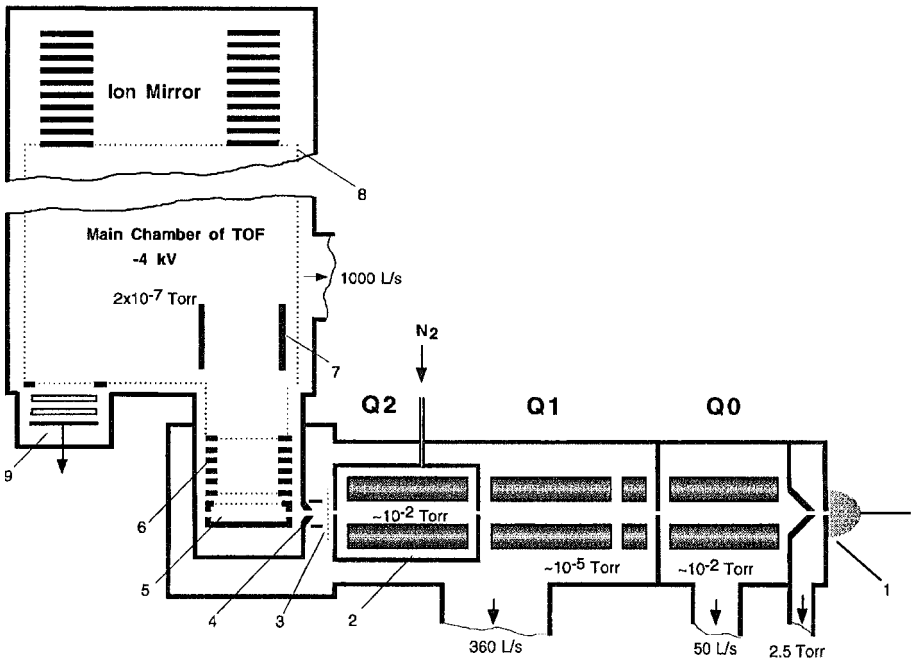

Figure 1. Schematic diagram of the quadrupole-TOF mass spectrometer (QqTOF). 1 — electrospray source, 2 — collision cell, 3 — focusing grid and DC-quadrupole, 4 — slit 2 × 8 mm, 5 — ion storage modulator, 6 — accelerating column, 7 — deflection plates, 8 — linear, 9 — detector.

storage gap, pushing ions in a direction orthogonal to their original trajectory into the accelerating column, where they acquire an energy of 4 keV per charge. The selected ratio of velocities in the two orthogonal directions allows ions to reach the TOF detector (9) naturally, without requiring deflection which could affect the mass resolution.[36,37] A single stage ion mirror provides first order compensation for the initial energy and spatial spread of the ions. A stainless steel liner (8) is floated at –4 kV; this prevents penetration of electric fields into the drift region. All mass spectra (in both MS and MS/MS modes) are recorded on the TOF spectrometer with a time-to-digital converter (Orsay, CTN-M2).

RESULTS AND DISCUSSION

Performance of the quadrupole time-of-flight instrument

First, the resolution and mass accuracy of the instrument were tested. The two mass spectrometers used in the construction of the hybrid had been characterized before and we therefore expected a performance similar to the first and second stage of the Sciex single quadrupole in ion selection and fragmentation and similar to the Manitoba time-of-flight mass spectrometer in fragment mass determination. This turned out to be true in practice. Mass resolution was routinely in the range of 5000 to 7000 (for both conventional MS and MS/MS modes of operation), allowing unambiguous charge state determination for ions at least up to m/z 5000. The mass accuracy obtained for peptides (with external calibration) was typically about 0.03 Da or better. Mass differences could be determined even more accurately and their measurement was mainly limited by the requirement for a number of ions sufficient to define the peak centroids.

This study aimed to investigate the improvements achievable by time-of-flight detection of fragment ions in practical sequencing of gel-separated proteins. To test the performance of the instrument under conditions that are normally used in triple quadrupole measurements at EMBL, a protein standard was run on a polyacrylamide gel, enzymatically degraded and the unseparated peptide mixture analysed by nanoelectrospray tandem mass spectrometry. Fig. 2(a) shows the tandem mass spectrum obtained from 80 fmoles of bovine serum albumin (BSA) loaded onto a polyacrylamide gel followed by preparation as described above and dissociation of the doubly charged peptide ion at m/z 461.7. As is typical of time-of-flight instruments, the spectrum shows no degradation of quality even at these very low levels. The resolution is still above 6000 and the mass accuracy of the eight Y" ions is better than 50 ppm on average.

The amount of protein used to test the quadrupole time-of-flight instrument was close to the limit of detection for our triple quadrupole (API III). For comparison, a BSA sample, prepared in the same way as before, was analysed on that machine. Fig. 2(b) shows the result for the same peptide ion as shown in Fig. 2(a). Note that the resolution and signal-to-noise ratio is much lower compared to the QqTOF instrument. As illustrated in this example, the main advantage of the quadrupole time-of-flight instrument compared to the triple quadrupole was the higher fragment ion current at very low sample levels. Even when not degrading resolution to improve sensitivity (a common practice in low level sequencing of peptides) the peaks start to be less well defined at very low levels. This is because each peak consists of only a few ions, obscuring the peak shape and isotopic pattern. As shown above this problem does not arise with the QqTOF instrument, which concentrates the ion current into a very narrow region (0.1 to 0.2 mass units).

A current limitation of the QqTOF, as well as of the trapping instruments, is their inability to perform parent ion scans. These scans are important to distinguish peptides from chemical noise, which would otherwise determine the limit of detection.[38] We have performed initial experiments with another type of scan, as a means to substitute for the parent ion scan. Quadrupole 1 was scanned over a certain mass range, and selected ions were fragmented in the collision cell. At the same time the detection electronics of the TOF mass analyser was set to accept all ions above the current value of m/z in Q_1. In this way all ions which produce fragments larger than their m/z are recorded, and all singly charged ions are rejected. Since the chemical background is predominantly singly charged and since the ions of interest in peptide sequencing are typically doubly or triply charged, this scheme affords some of the benefits of the parent ion scan technique. These experiments have shown promising results and will be described elsewhere.[39]

Application to protein identification

The higher resolution and mass accuracy obtained by time-of-flight detection of the fragment ions should be advantageous for the database identification of proteins. In the peptide sequence tag algorithm[13] peptides are identified in database searches on the basis of a short stretch of sequence (typically two to three amino acids) combined with the distance, in mass units, to the N- and C-terminus of the peptide.[13,40] The peptide is thus divided into three parts or regions. The mass is known of the first region, from N-terminus of the peptide until the beginning of the sequence. Of region two, in the middle of the peptide, the sequence has been determined. The mass of region three, from the sequence to the C-terminus, is again known. In a database search, either all three regions should match to a sequence in the database (normal search), or only two of the regions (error tolerant search). The mass spectra produced by the QqTOF are easily interpreted, and it is usually trivial to assign a stretch of four amino acids in a tryptic peptide. Furthermore, the search specificity increases proportionately to the square of the obtained mass accuracy because there are two independent mass values that are used in the search. While the high mass accuracy is not necessary for the unambiguous identification by peptide sequence tags, it helps to make the search even more straightforward and 'error tolerant' as described below.

Higher search specificity

To test by experiment these expected improvements in database search specificity, we investigated a number of yeast proteins obtained from ongoing collaborations.

Proteins separated by one dimensional SDS PAGE were prepared as described above and analysed by nanoelectrospray sequencing on the QqTOF. Table 1 lists the resulting identifications. As a typical case, Fig. 3 shows the tandem mass spectrum of one of the peptides from the band at 34 kDa. The figure demonstrates the excellent resolution and distinction between singly and doubly charged ions that can routinely be achieved. As indicated in the figure, the ions at 475.31; 588.39; 735.43; 864.49; and 935.52 can be combined into the search string (475.31)LFEA(935.52) with the monoistopic parent mass of 1674.90 (where L could also be I and F could also be oxidized methionine (M*)). These data were entered into PeptideSearch. With a specified mass accuracy of 50 ppm, the search resulted in a single match, the peptide IKTPETAAEM*INTIK from EIF-2B translation initiation factor. Note that no enzyme specificity, fragment ion type, target protein molecular weight, or any other search restriction was applied. Interestingly, a search of the complete database of expressed sequence tags[41] (dbEST, currently 1 012 567 entries), in which the peptide is not contained, did not retrieve a single match, attesting to the very high search specificity of the tag. Indeed, even a shorter sequence tag, consisting of only two of the four amino acids that could be easily called, already uniquely retrieved the peptide from the comprehensive database. Of course, the subsequent validation by the complete tandem mass spectrum is also more specific due to the higher mass accuracy.

From all four proteins, several sequence tags were obtained that uniquely retrieved the cognate peptide sequence (see Table 1). Unique retrieval, combined with subsequent verification of the complete peptide sequence against the full tandem mass spectrum makes it unnecessary to use probability based methods in the identification of proteins. This is also the case in peptide sequence tag searching with lower mass accuracy; however, the extremely high search specificity of the tags obtainable by the QqTOF instrument allows even more latitude in searching as shown below.

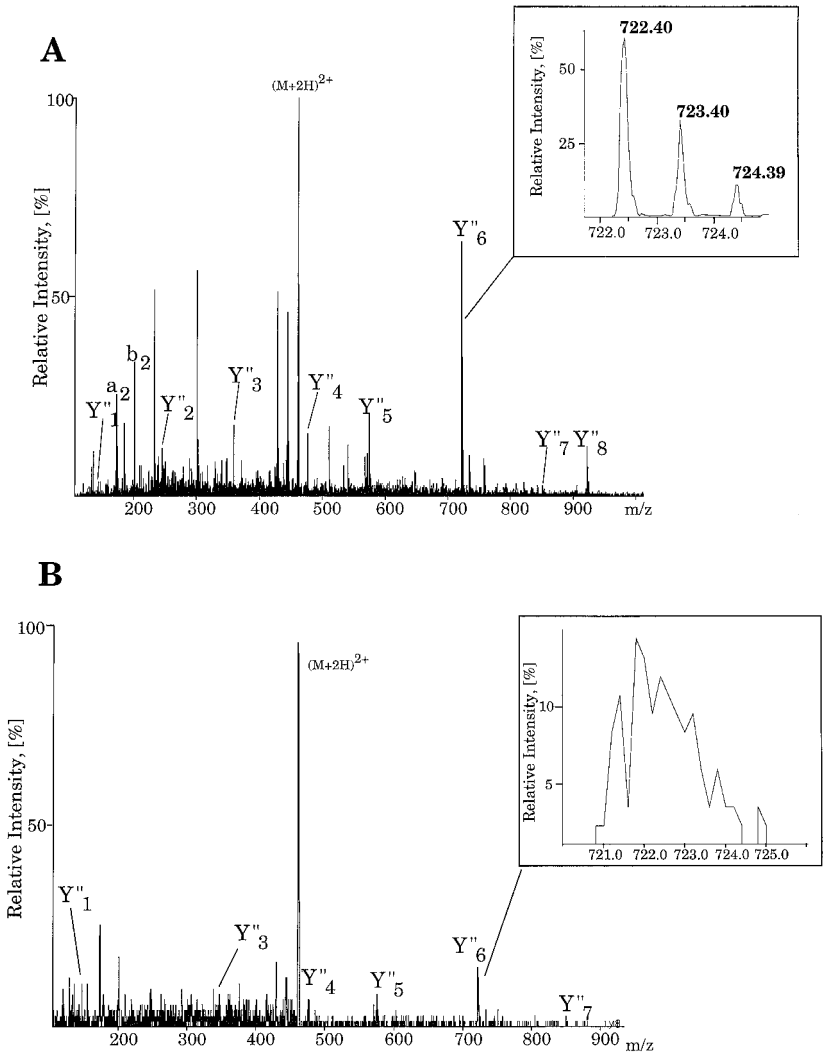

Figure 2. Tandem MS of 80 fmoles BSA loaded on a gel. To demonstrate sensitivity and determine data quality at very low levels two samples containing 80 fmols of BSA on a gel were analysed in parallel on the QqTOF instrument (a) and on an API III triple quadrupole (b). Fragmentation of the doubly charged ion at *m/z* 461.73 is shown. Although the signal-to-noise of the parent ion is poor and it is hardly visible in the chemical noise the mass resolution is still more than 6000 and the mass is within 0.02 mass units of its calculated value.

Table 1. Identification of yeast proteins

Band	Name[a]	Unique tabs[b]
1	IF2G_YEAST TRANSLATIONAL INITIATION FACTOR 2 GAMMA SUBUNIT (58.5 kDa P32481)*	4
2	HS71_YEAST HEAT SHOCK PROTEIN SSA1 (69.7 kDa P10591)	2
3	IF2A_YEAST TRANSLATIONAL INITIATION FACTOR 2 ALPHA SUBUNIT (34.7 kDa P20459)	5
4	E2BA_YEAST TRANSLATION INITIATION FACTOR EIF-2B-ALPHA SUBUNIT (34.4 kDa P1471)	4

[a] With calculated Mr and accession number in SWISSPROT
[b] Number of peptide sequence tags that retrieved only a single peptide entry in the database. Note that there was no attempt to determine all such tags and that positive identification by peptide sequence tag searching can also be achieved when several sequences are retrieved (because of validation of the sequences with the full tandem mass spectrum)

Error tolerant search

The peptide sequence tag algorithm divides a measured peptide into three sections, only two of which have to agree with the database entry in order to yield a match. This feature is normally used when checking against only one or a limited class of sequences (to find mutations, post-translational modifications, etc.) because the search specificity is reduced compared to matching all three regions of the peptide. Using the improved mass accuracy, full database searches can routinely be performed by matching only two of three peptide regions.

As an example, a peptide of one of the above yeast proteins (Fig. 4) easily yielded the following partial sequence …ETLN… which was assembled into the search string (552.30)ETLN(1009.55) with the monoistopic parent mass of 1413.66 Da. A search without any restrictions in the comprehensive database did not identify any sequence. Requiring only 'regions two and three' to match (i.e. the sequence NLTE and the determined mass to the C-terminus of the peptide) produced 18 matches when done at 2 mass units accuracy. A search with 50 ppm mass accuracy, however, resulted in only one match, which moreover was from a yeast protein. (The N-terminal peptide of EIF-2B with sequence MSEFNITETYLR). Alignment of the sequence with the full MS/MS spectrum verified the C-terminal 8 amino acids of the peptide (regions two and three). Inspection of the N-terminal part allowed the Y" ion series to be extended by two more residues, which were identical to the database entry. The remaining mass difference to the N-terminus of the peptide could then be explained by the fact that the methionine from the database sequence was missing in the peptide and that the N-terminal serine was acetylated. This was supported by B-ion fragments which fitted the sequence (N-acetyl)SEFNITETYLR. Straightforward identification of this peptide despite its differences from the corresponding database entry demonstrates the potential of error tolerant search with high mass accuracy as a general tool for identifying post-translational modifications. The high search specificity may furthermore be very useful in searching EST databases, which contain numerous errors.

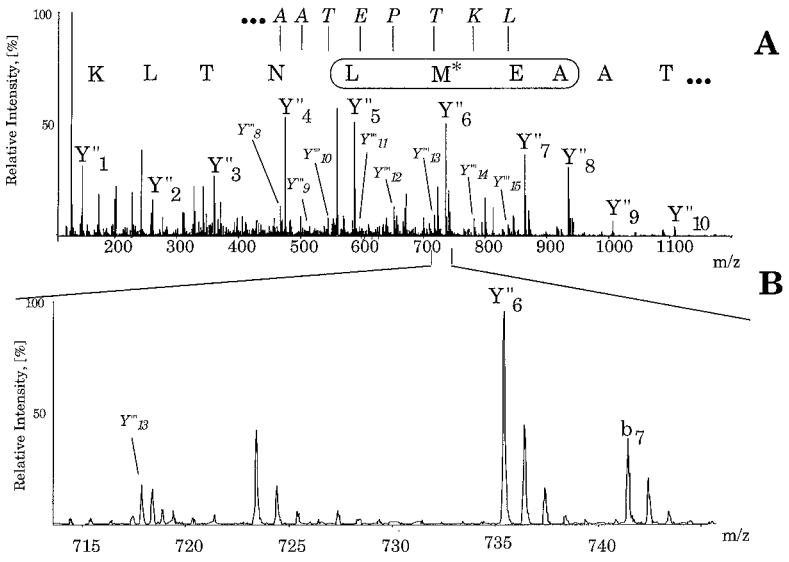

Figure 3. Identification of a yeast protein. The figure shows the tandem mass spectrum of a peptide recovered after in-gel digest of a yeast protein (band 4 of Table 1). The triply charged precursor ion was dissociated, leading to overlapping series of singly and doubly charged fragments. (a) shows the singly charged series …TLK which can be followed in the spectrum. The partial sequence L(F/M*)EA, were F/M* could be either phenylanaline or oxidized methionine, can very easily be assigned (boxed in (a)) and serves as a sequence tag (see text). The series marked by Y''' symbols is doubly charged and gives rise to an extension of the singly charged ion series towards the N-terminus of the peptide, as indicated at the top of (a). Note that singly and doubly charged ions can easily be distinguished by their isotopic spacing (b).

Application to the sequencing of novel proteins

As shown above, the identification of already known proteins profits greatly from the increased mass accuracy and resolution of the quadrupole time-of-flight instrument. These factors are even more important when sequencing novel proteins. While mass spectrometric sequencing of such proteins is routinely performed in our laboratory using a triple quadrupole instrument, this is still very laborious and requires great expertise, at least when performed at low levels (less than two picomoles of gel-separated protein).

To retrieve a sequence from the mass spectrometric data we apply three criteria: (a) the mass difference between two adjacent peaks should precisely fit the mass of an amino acid residue; (b) in the case of tryptic peptides, the lowest mass ion in the C-terminal series should be the $Y"_1$ ion corresponding to Arg or Lys; (c) there should be an independent indication that a particular ion in the spectrum in fact belongs to a given ion series (usually the Y" ion series in the case of tryptic peptides). In our laboratory (EMBL) criterion (c) has generally been provided through a separate experiment, by esterification of the carboxyl groups, including the C-terminus.[3] While construction of partial ion series is much simplified by the high data quality of the QqTOF, we found that criterion (c) must still be provided by an independent source of information if accurate sequence data is desired.

^{18}O labeling of peptide mixture

Because of its high resolution without loss of sensitivity, the QqTOF should be ideal for the application of another labeling of C-terminal ions, the ^{18}O method. Here we discuss application of this technique to one typical case, the partial sequencing of a protein important in cell cycle regulation.

The immunoprecipitated protein, which had an apparent mass of 78 kDa after one dimensional SDS PAGE, was excised and in-gel digested by trypsin in a buffer containing 50% v/v of ^{18}O water. Further sample preparation was done as described previously.[17,33] The nanoelectrospray mass spectrum resulting from the analysis is shown in Fig. 5(a). In the peptide map, labeled peptides could be determined by their characteristic 2 mass unit isotope pattern. The charge state of the peptides could be determined directly from the isotopic spacing. Comparison of the experimentally observed isotopic pattern with the expected one for the same mass showed that most peptide molecules incorporated only one ^{18}O atom. Judging from the isotope ratios, the incorporation of the second ^{18}O atom must be less than 15 percent (see e.g. Fig. 5(c)). Therefore, we did not find it necessary to perform two separate experiments, one in normal and one in ^{18}O water as had been reported by other workers.[28] Rather, the presence of independent information in a single experiment is a great advantage of the ^{18}O method over the esterification technique.

Fig. 5(b) and (c) illustrate how ^{18}O labeling, in conjunction with the high resolution and mass accuracy of the QqTOF instrument, makes sequence read-out very straightforward. The Y" ion series can in most cases be assigned to peaks that differ by the molecular weight of an amino acid residue and also display an intensity ratio between Y" and (Y" + 2) ions consistent with the degree of labeling. These stringent criteria also significantly shorten the interpretation process.

The high quality of the data from the QqTOF reduces ambiguities in a number of other ways as well. Frequently the same ion series can be followed both as a singly charged and as a (partially overlapping) doubly charged series as demonstrated in Fig. 3. In particular, the doubly charged Y" ion series could often be extended all the way to the N-terminus of the peptide

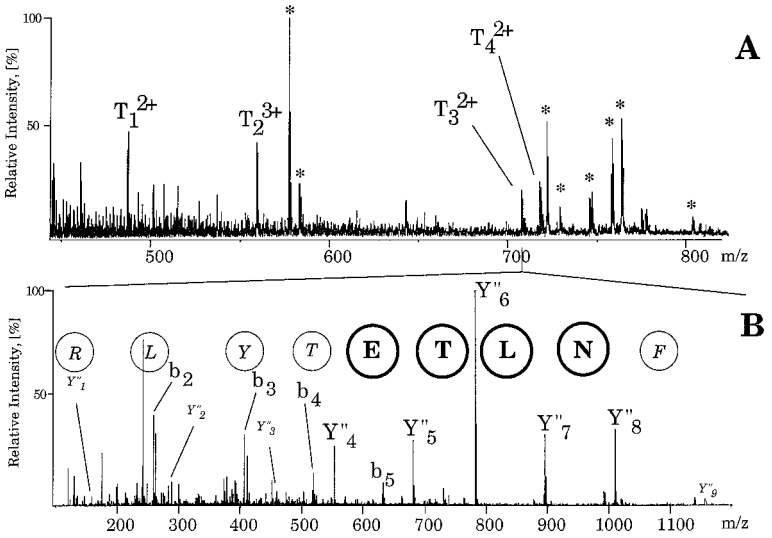

Figure 4. Error tolerant identification of a yeast protein. In-gel digest and nanoelectrospray analysis of a yeast protein (band 4 of Table 1) resulted in the peptide mass spectrum shown in (a). Fragmentation of the peak marked as T_3 yielded the tandem mass spectrum of (b). A series of four amino acids can be assigned obviously and unambiguously (ETLN). The peptide sequence tag at 50 ppm mass accuracy was sufficient to uniquely retrieve the peptide sequence and identify the protein in a comprehensive sequence database in spite of the modifications of the peptide (see text).

allowing unambiguous determination of the sequence order of the two N-terminal amino acids. (This can otherwise be difficult because of the absence of the relevant ions.)

In the example, eight peptides were sequenced, covering 92 amino acid residues (Table 2). All peptides except T_7, a *m/z* 2300 peptide, were sequenced completely. In that peptide fourteen amino acids from the N-terminus and three from the C-terminus were assigned, leaving only two amino acid residues undetermined. Peptide T_8 was found to be N-terminally acetylated, indicating it is the actual N-terminal peptide of the mature protein. After the protein had been cloned using the partial sequences provided by tandem mass spectrometry, two differences between the interpreted sequence and the gene sequence were found (underlined in Table 2). One discrepancy occurred in peptide T_6 where a W was called and a SV was present (SV has the same molecular weight as W). In this case, the presence of SV rather than W could have been deduced from the spectrum and was indeed recognized by our software as described below.

In this and other examples, we have found a number of advantages of the ^{18}O method over the esterification method.

- There is only one experiment, reducing the amount of work and the danger of losing precious material.
- There is no need to determine the *m/z* location of the 'derivatized' peptide.
- It is always possible to analyse the 'derivatized' peptide, even in the case of large, acidic peptides (which are sometimes lost after esterification).
- Incorporation of ^{18}O is only determined by the $^{16}O/^{18}O$ ratio in the buffer and does not depend on the nature of the peptide.
- Even in the low mass region, Y" ions can more readily be identified. (For esterified peptides, with acidic residues near the N-terminus this is not the case).
- The ^{18}O approach lends itself to automatic interpretation (as described below).

Table 2. Sequencing of a novel protein

Peptide[a]	Peptide (MW) mass units	Sequence determined by MS/MS[b] Sequence derived after cloning
T_1	1002.6	SELLTAELK
		SELLTAELK
T_2	1127.6	QQLQEELLK
		QQLQEELLK
T_3	1251.6	LDLTTENAGYR
		IDITTENQYRL
T_4	1271.8	LSAQEQDAALVK
		LSAQEQDAAIVK
T_5	1302.8	FLLEDQAAQLR
		FLLEDQAAQIR
T_6	1331.6	LSESNENLWLK
		LSESNENISVLK
T_7	2358.6	(AL)LDSYDSELTPTEHS(...)LSR
		AILDSYDSELTPTEHSPQLSR
T_8	1753.6	(Ac)M(ox)DDSEDNTTVLSTRLR
		MDDSEDNTTVISTLR

[a] Peptide as shown on Fig. 5
[b] I and L are not differentiated in the MS/MS data

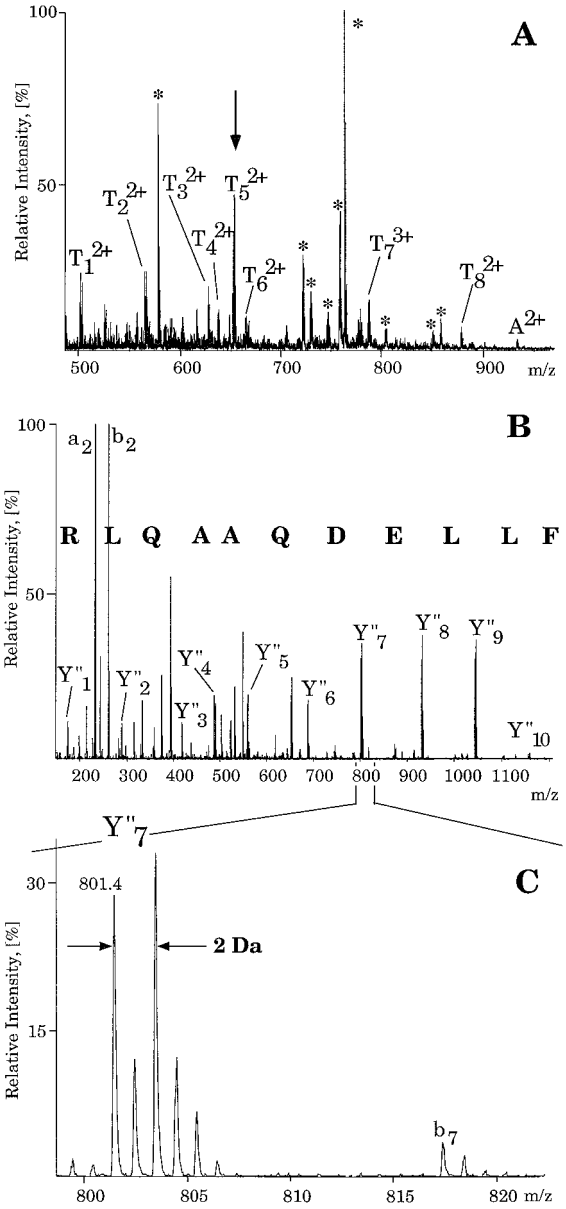

Figure 5. Sequencing of a novel protein. (a) Mass spectrum of the tryptic digest of a novel vertebrate protein. The digest was performed in 1:1 $H_2{}^{16}O/H_2{}^{18}O$. The peaks labeled by T were sequenced. The peaks marked by (a) were also fragmented but belong to the antibody used in the purification and the peaks marked by the asterisk belong to trypsin. (b) Tandem mass spectrum of one of the peaks. As shown by the zoom on peak $Y"_7$ (c), the fragment ion signals are very well resolved, and the ^{18}O labeling allows easy distinction of Y" from B type ions.

Automatic read-out of sequences

Because of the excellent quality of the tandem mass spectra from the QqTOF, we have investigated the possibility of a completely automatic read-out of the sequence. When tryptic peptides are fragmented, the tandem MS spectrum often contains long series of Y" ion because the C-terminal amino acid (Lys or Arg) is basic and therefore retains the charge. The interpretation algorithm identifies Y" ion series by their characteristic $^{16}O/^{18}O$ ratio throughout the spectrum. To assemble a series of Y" ions the spectrum is screened for doublet peaks that are precisely one

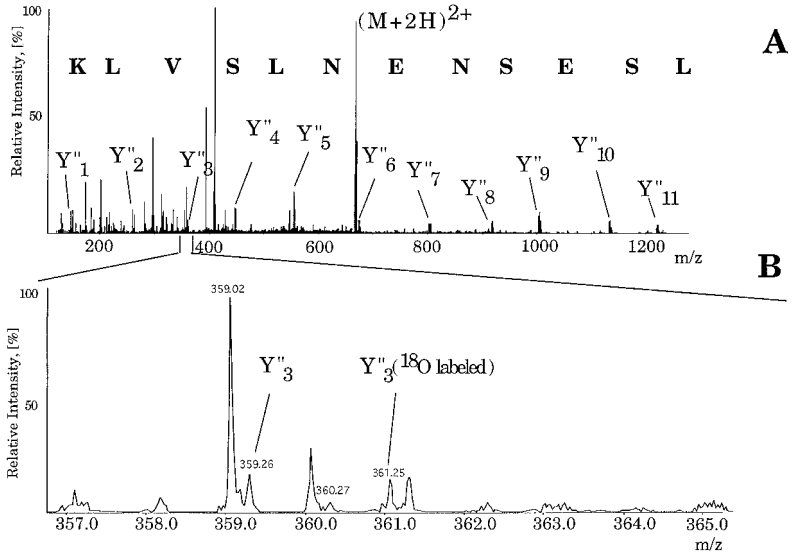

Figure 6. Automated sequence assignment. The MS/MS spectrum in the figure was interpreted by a script developed for the interpretation of tryptic end labeled peptides. The interpretation script switched the order of two amino acids but interestingly it correctly predicted SV instead of W, which had been assigned during manual interpretation (see Table 2). (b) shows the labeled peak, which despite its low intensity was sufficient for automated interpretation.

amino acid residue mass apart. The algorithm is applied to different charge states of fragments since the charge of an ion can easily be determined by its isotopic spacing. Finally, the series produced from different charge states are combined. In cases where more than one sequence is generated the possibilities are scored using the fragment ion intensities taking B-ions into account. The algorithm does not necessarily aim at generating a sequence proposal that corresponds to the complete peptide mass. In some cases the Y''' ion series in the spectrum is not complete although a large part of the peptide sequence is covered. By taking advantage of different charge states, the retrieved sequences can be more than 15 amino acids long.

To test the algorithm, three of the spectra leading to the sequence assignment in Table 2 were analysed automatically. (After manual sequence assignment but before the gene sequence was known to us). In two of the three cases, peptides T_1 and T_5 from Table 2, manual and automatic interpretation gave the same results. In the third case, peptide T_6, there were two differences. In one case, the program had switched the order of two amino acids. Interestingly, the other case the algorithm assigned the sequence correctly, while the manual interpretation was at fault. The error in the manual assignment arose because the distinction between W and SV was not made by the human expert; the intervening ion was very small and furthermore in the 'shadow' of a large but unrelated peak (see Fig. 6). The program, on the other hand, does not weight the results much by the intensity of the fragment ion, and, while the intervening ion was small, it did have the required accompanying ^{18}O labeled satellite peak. The initial results of the automated sequence assignments are very encouraging and we hope that such programs can now finally be useful in the interpretation of tandem mass spectra.

CONCLUSION AND PROSPECTS

A QqTOF instrument was constructed and equipped with the nanoelectrospray ion source. The instrument was applied to low level sequencing experiments as typically performed at EMBL. Significant improvements in the interpretation of the spectra were made possible by the high data quality, even for very low sample amounts. In cases where the protein was included in a database, high mass accuracy sequence tags retrieved unique sequences from the database. Routine identification is now possible even with very limited data and error tolerant searching, e.g. in highly error prone EST sequence databases, is much more straightforward.

To determine the sequence of unknown proteins, it is still necessary to obtain independent information, which in the case of the QqTOF can be very simply done by digesting the protein in the presence of ^{18}O water. The facile read-out of sequences via isotopic labels need not be limited to tryptic peptides; an isotopically labeled N-terminal reagent[42] could be synthesized that would allow identification of the N-terminal ion series (B ions) in the same way as we have shown here for the C-terminal ions of tryptic peptides.

The combination of ^{18}O labeling, nanoelectrospray and QqTOF is here shown to be particularly powerful for the sequencing of novel proteins. Interpretation time is much reduced due to the high quality of the fragment spectra. We have also demonstrated that at least partial automation of sequence interpretation is now feasible. After the success of mass spectrometry in the identification of known proteins, the present results

indicate that novel amino acid sequences for cloning or homology searching may soon be obtainable with similar ease.

Acknowledgements

Tom Covey and Ron Bonner helped in the construction of the hardware and software of the QqTOF instrument. Brian Otter designed the interface between the quadrupole and the time-of-flight parts of the instrument. Work in the Protein & Peptide group is partially supported by generous grants from the German Technology ministry (BMBF) and Glaxo Welcome PLC. Work at Manitoba was partially supported by the NSERC (Canada).

We thank the other members of the Peptide & Protein group for fruitful discussions, Ole N. Jensen for critical reading of the manuscript and Anna Shevchenko at EMBL and Vic Spicer at Manitoba for expert technical support.

REFERENCES

1. P. Tempst, A. Link, L. Riviere, M. Fleming and C. Elicone, *Electrophoresis* **11**, 537–553 (1990).
2. K. Biemann, *Anal. Chem.* **58**, 1289A-1300A (1985).
3. D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston and C. R. Hauer, *Proc. Nat. Acad. Sci., USA* **83**, 6233-6237 (1986).
4. K. Biemann, *Protein Sci.* **4**, 1920-1927 (1995).
5. S. D. Patterson and R. Aebersold, *Electrophoresis* **16**, 791-814 (1995).
6. W. J. Henzel, T. M. Billeci, J. T. Stults and S. C. Wong, *Proc. Nat. Acad. Sci., USA* **90**, 5011-5015 (1993).
7. M. Mann, P. Højrup; and P. Roepstorff, *Biol. Mass Spectrom.* **22**, 338-345 (1993).
8. D. J. C. Pappin, P. Højrup and A. J. Bleasby, *Current Biol.* **3**, 327-332 (1993).
9. P. James, M. Quadroni, E. Carafoli and G. Gonnet, *Biophys. Biochem. Res. Commun.* **195**, 58-64 (1993).
10. J. R. Yates, S. Speicher, P. R. Griffin and T. Hunkapiller, *Anal. Biochem.* **214**, 397-408 (1993).
11. O. N. Jensen, A. Podtelejnikov and M. Mann, *Rapid Commun. Mass Spectrom.* **10**, 1371-1378 (1996).
12. A. Shevchenko, O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie and M. Mann, *Proc. Nat. Acad. Sci., USA* **93**, 14440-14445 (1996).
13. M. Mann and M. S. Wilm, *Anal. Chem.* **66**, 4390-4399 (1994).
14. J. K. Eng, A. L. McCormack and I. J. R. Yates, *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).
15. M. S. Boguski, *Trends Biochem. Sci.* **20**, 295-296 (1995).
16. M. Mann, *Trends Biol. Sci.* **21**, 494-495 (1996).
17. M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis and M. Mann, *Nature* **379**, 466-469 (1996).
18. M. Muzio, A. M. Chinnaiyan, F. C. Kischkel, K. O. Rourke, A. Shevchenko, Jian Ni, C. Scaffidi, J. D. Bretz, M. Zhang, R. Gentz, M. Mann, P. H. Krammer, M. E. Peter and V. M. Dixit, *Cell* **85**, 817-827 (1996).
19. J. Lingner, T. R. Hughes, A. Shevchenko, M. Mann, V. Lundblad, T. R. Cech, *Science* **276**, 561-567 (1997).
20. M. S. Wilm, M. Mann, *Int. J. Mass Spectrom. Ion Processes* **136**, 167-180 (1994).
21. M. Wilm and M. Mann, *Anal. Chem.* **68**, 1-8 (1996).
22. D. Desiderio, M. Kai, *Biomed Mass Spectrom.* **10**, 471-479 (1983).
23. K. Rose, M. Simona, R. Offord, C. Prior, D. Thatcher, *Biochem. J.* **215**, 273-277 (1983).
24. S. J. Gaskell, P. E. Haroldsen, M. H. Reilly, *Biomed. Environm. Mass Spectrom.* **16**, 31-33 (1988).
25. K. Rose, L. Savoy, M. Simona, R. Offord, P. Wingfield, *Biochem. J.* **250**, 253-259 (1991).
26. T. Takao, H,. Hori, K. Okamoto, A. Harada, M. Kamachi, Y. Shimonishi, *Rapid Commun. Mass Spectrom.* **5**, 312-315 (1991).
27. B. Whaley, R. M. Caprioli, *Biol. Mass Spectrom.* **20**, 210-214 (1991).
28. M. Schnolzer, P. Jedrzejewski, W. D. Lehmann, *Electrophoresis* **17**, 945-953 (1996).
29. P. Roepstorff, J. Fohlman, *Biomed. Mass Spectrom.* **11**, 601 (1984).
30. H. R. Morris, T. Paxton, A. Dell, J. Langhorn; M. Berg, R. S. Bordoli, J. Hoyes, R. H. Bateman, *Rapid Commun. Mass Spectrom.* **10**, 889-896 (1996).
31. A. N. Verentchikov, W. Ens, K. G. Standing, *Anal. Chem.* **66**, 126-133 (1994).
32. I. V. Chernushevich, W. Ens and K. G. Standing, In *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation & Applications*, R. Cole (Ed.), Wiley, Chichester 1997.
33. A. Shevchenko, M. Wilm, O. Vorm, M. Mann, *Anal. Chem.* **68**, 850-858 (1996).
34. O. N. Jensen, A. Shevchenko, M. Mann In *Protein Structure — A Practical Approach* 2nd Edition, T. E. Creighton (Ed.), Oxford University Press, Oxford, 1997, pp. 29-57.
35. M. Mann In *Microcharacterization of Proteins*, R. Kellner, F. Lottspeich and H. E. Meyer (Eds.), VCH, Weinheim, 1994, pp. 223-245.
36. A. F. Dodonov, I. V. Chernushevich, V. V. Laiko, In *ACS Symposium Series 549*, R. J. Cotter (Ed.), Washington, DC, 1994, pp 108-123.
37. M. Guilhaus, *J. Am. Soc. Mass Spectrometry*, **5**, 588-595 (1994).
38. M. Wilm, G. Neubauer, M. Mann, *Anal. Chem.* **68**, 527-533 (1996).
39. I. Chernushevich *et al.* (in preparation).
40. R. Bonner, B. Shushan, *Rapid Commun. Mass Spectrom.*, **9**, 1077-1080 (1995).
41. M. S. Boguski, T. M. J. Lowe, C. M. Tolstoshev, *Nature Genetics* **4**, 332-333 (1993).
42. M. Bartlet-Jones, W. Jeffery, H. Hansen, D. Pappin, *Rapid Commun. Mass Spectrom.* **8**, 327-332 (1994).

