Methods in Molecular Biology™

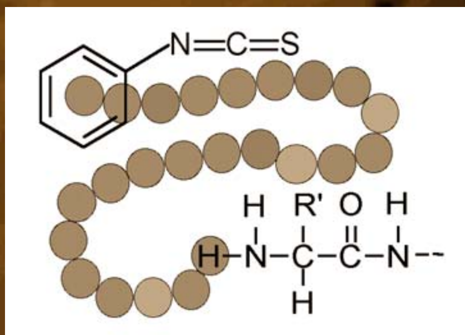
VOLUME 211

Protein Sequencing Protocols

SECOND EDITION

Edited by

Bryan John Smith



19

Nanoelectrospray Tandem Mass Spectrometry and Sequence Similarity Searching for Identification of Proteins from Organisms with Unknown Genomes

Anna Shevchenko, Shamil Sunyaev, Adam Liska, Peer Bork, and Andrej Shevchenko

1. Introduction

Recent developments in technology and instrumentation have made mass spectrometry the method of choice for identification of proteins in a sequence database [reviewed in **refs.** ***(1–4)***]. Regardless of what mass spectrometric technology is applied, it is ultimately required that acquired mass spectra be accurately matched to a protein sequence from the corresponding database entry.

If the sequence of the analyzed protein is not present in a database, the further possible strategy depends on the sequence similarity between the protein of interest and homologous proteins from other species. If the sequence similarity is high and proteins share 5–10 identical tryptic peptides, it would be possible to identify the "unknown" protein of interest by cross-species matching of its peptide mass map to a known sequence of its homolog ***(5)***. If the similarity is low, usually the peptides have to be sequenced *de novo* and the gene of the unknown protein has to be cloned by polymerase chain reaction (PCR) ***(6,7)***.

How is it possible to determine sequences of tryptic peptides? Continuous series of fragment ions containing the C-terminus (Y" ions) ***(8)***, which are usually observed in tandem mass spectra of tryptic peptides, have been successfully recognized for *de novo* sequencing ***(7)***. Partial peptide sequence can be readily obtained by considering precise mass differences between the adjacent Y"- ions in the *m/z* region above the multiply charged precursor ***(9)***. However, in order to determine a complete peptide sequence it is necessary to identify Y"-ions in the low *m/z* region of the spectrum, where ions of other series and ions originating from chemical noise are abundant. Therefore it is necessary to obtain additional

From: *Methods in Molecular Biology, vol. 211: Protein Sequencing Protocols, 2nd ed.*
Edited by: B. J. Smith © Humana Press Inc., Totowa, NJ

evidence that the particular fragment ion indeed belongs to the Y”-series. Masses of the ions of Y” series are shifted upon peptide esterification with methanol ***(7,10)*** and could be identified by comparing tandem mass spectra acquired from esterified and native peptides. Alternatively, C-terminal carboxyl group of the peptide could be selectively labeled with ^{18}O isotope by digesting the protein in a buffer containing 50% of $H_2{}^{16}O$ and 50% $H_2{}^{18}O$ (v/v). Y” ions could be distinguished by a characteristic isotopic pattern: a doublet of peaks, split by 2 mass units ***(11)***. Using any of the approaches, complete and accurate sequences of tryptic peptides can be determined ***(12–16)***. However, *de novo* sequencing remains laborious and time consuming, and requires significantly higher amount of protein compared to conventional protein identification.

Because of the rapid growth of sequence databases, it is becoming increasingly possible to identify unknown proteins via sequence similarity searching. Importantly, it is possible to utilize peptide sequences of much lower quality, compared to the ones required for designing degenerate oligonucleotide primers for PCR ***(17–20)***. Partially redundant peptide sequence candidates determined by automated or rapid manual interpretation of tandem mass spectra are combined (regardless of their completeness, confidence and length) and used in a single search by Mass Spectrometry driven BLAST (MS BLAST) ***(19)***, which is performed over the web on a high computational capacity server. A few thousands candidate sequences can be submitted in a single search, which takes a few minutes to complete. Therefore MS BLAST searching could be coupled with high throughput protein identification methods as MALDI quadrupole TOF ***(21)***, MALDI TOF/TOF ***(22)*** and LC MS/MS ***(23)*** by straightforward script-based automation.

2. Materials

2.1. Chemicals

All chemicals should be of analytical grade or better.

1. Water, acetonitrile methanol, and formic acid high-performance liquid chromatography (HPLC) grade from Merck (Darmstadt, Germany) (*see* **Note 1**).
2. Perfusion sorbent POROS 50 R2 (PerSeptive Biosystems, Framingham, MA) (*see* **Note 2**)
3. Borosilicate glass capillaries GC120F-10 1.2 mm OD × 0.69 mm ID (Harvard Apparatus Ltd, Edenbridge, UK). Needles for nanoelectrospray and columns for micropurification could be purchased from Protana (Odense, Denmark) or manufactured as described in ***(24,25)***.

2.2. Equipment and Software

1. Mass Spectrometers. A QSTAR Pulsar *i* quadrupole time-of-flight mass spectrometer and an API III triple quadrupole mass spectrometer (both from MDS Sciex, Concord, ON, Canada) each equipped with a nanoelectrospray ion source

(24) (Protana, Odense, Denmark and EMBL, Heidelberg, respectively), tuned for peptide sequencing as described in **Note 3**.
2. Benchtop mini-centrifuge (as "PicoFuge," Stratagene, Palo Alto, CA).
3. Micropurification holder purchased from Protana (Odense, Denmark) or manufactured as described in *(25)*.
4. Software for automated interpretation of tandem mass spectra (*see* **Note 4**). Tandem mass spectra acquired on an API III mass spectrometer were interpreted by PredictSequence routine (a part of the BioMultiview software package from MDS Sciex, Concord, Canada). Spectra acquired on a QSTAR Pulsar *i* were interpreted using BioAnalyst v.1.0 supplied by the same company.

   MS BLAST web interface is located at: http://dove.embl-heidelberg.de/Blast2/msblast.html

3. Methods

3.1. Desalting and Concentration of Peptides for NanoES MS/MS Sequencing (see *Note 5)*

1. Pipet ca. 5 µL of POROS R2 slurry (*see* **Note 2**) into the pulled glass capillary (*see* **Subheading 2.1.3.**), here and further down referred as a "column."
2. Spin the beads down and then open the pulled end of the column by gentle touching against a bench top. The amount of POROS R2 resin in the column should be within the tapered region only.
3. Wash the beads with 5 µL of 5% formic acid and make sure the liquid can flow out of the column if gentle centrifuging is applied. Open the column end wider if necessary. Mount the column into the micropurification holder.
4. Dissolve the dried protein digest in 10 µL of 5% formic acid and load onto the column. Pass the sample through the beads layer by centrifuging.
5. Wash adsorbed peptides with another 5 µL of 5% formic acid.
6. Align the column and the nanoelectrospray needle in the micropurification holder and elute peptides directly into the needle with 1 µL of 60% of methanol in 5% formic acid by gentle centrifuging. Add the eluent in three applications of 0.3 µL or two applications of 0.5 µL for best results.
7. Mount the spraying needle with the sample into the nanoelectrospray ion source and acquire mass spectra.

3.2. Sequencing on a Triple Quadrupole Mass Spectrometer (see *Note 6)*

1. After desalting and concentration of peptides (*see* **Subheading 3.2.**), initiate spraying and acquire Q1 spectrum of the peptide mixture. Turn on collision gas and acquire the spectrum in the precursor scan mode with the fragment mass set at *m/z* 86 (*see* **Note 7**).
2. Stop spraying by dropping spraying voltage to zero. Drop the air pressure applied to the spraying capillary to zero. Move the spraying capillary away from the inlet of the mass spectrometer.
3. Examine the acquired spectra and compare them with the spectra acquired from

the control sample (*see* **Note 6**). Select precursor ions for subsequent tandem mass spectrometric sequencing.

4. Add 0.3–0.5 µL of 60% of methanol in 5% formic acid directly to the spraying capillary if the remaining sample volume is less than 0.5 µL. Re-establish spraying and acquire tandem mass spectra from selected precursor ions.

3.3. Sequencing on a Quadrupole Time-of-Flight Mass Spectrometer

Sequencing is performed as described in **Subheading 3.2.**, but no precursor ion scanning is applied and the collision cell is always filled with gas. High resolution of a quadrupole time-of-flight mass spectrometer allows to distinguish genuine multiply charged peptides ion as characteristic sharp isotopic features superimposed on chemical background, which is mostly observed as singly charged broad irregular peaks. Collision energy could be adjusted by monitoring the intensity of the residual precursor ion during acquisition of the tandem mass spectrum ***(26)***.

3.4. Identification of Proteins by MS BLAST Searching (see Note 8)

MS BLAST is a specialized BLAST-based ***(27)*** tool for identification of proteins by sequence similarity searching that utilizes peptide sequences produced by the interpretation of tandem mass spectra ***(19)*** (*see* **Note 9**).

1. Obtain, edit, and assemble peptide sequences.
   a. Automated interpretation of MS/MS spectra: If tandem mass spectra were interpreted by *de novo* sequencing software, disregard relative scores and use the entire list of candidate sequences (or some 50–100 top scoring sequence proposals per fragmented peptide precursor).
   b. Manual interpretation of MS/MS spectra: Try making the longest possible sequence stretches, although their accuracy may be compromised. For example, it is usually difficult to interpret unambiguously fragment ion series at the low *m/z* range because of abundant peaks of chemical noise and numerous fragment ions from other series. In this case, it is better to include many complete (albeit low confidence) sequence proposals into the query rather than using a single (although accurate) three–four amino acid sequence stretch deduced from a noise-free high *m/z* segment of the spectrum.
2. Gaps and ambiguities in peptide sequences. Some *de novo* sequencing programs may suggest a gap in the peptide sequence that can be filled with various isobaric combinations of amino acid residues. For example:

   ASDF[...]FGTR, […] = [L,T] or [D,V]

   If one or two combinations were suggested, include all variants into a searching string:

   -ASDFLTFGTR-ASDFTLFGTR-ASDFDVFGTR-ASDFVDFGTR-

   If more combinations were possible, the symbol X can be used instead to fill the gap. Zero score is assigned to X symbol in PAM30MS scoring matrix (*see* **Subheading 3.4.5.**) and therefore it matches weakly any amino acid residue:
   -ASDFXXFGTR-

Note that MS BLAST is sensitive to the number of amino acid residues that are filling the gap. If the gap could be filled by a combination of two and three amino acid residues, consider both options in the query:

-ASDFXXFGTR-ASDFXXXFGTR-

3. Isobaric amino acids. L stands for Leu and Ile; Z stands for Gln and Lys, if undistinguishable in the spectrum. Use Q or K if the amino acid residue could be determined.
4. Generic trypsin cleavage site. If the proposed sequence is complete, a putative trypsin cleavage site symbol B is pasted prior to the peptide sequence:

   …-BASDFLTFGTR-

   It is often difficult to determine two amino acid residues located at the N-terminus of the peptide. In this case present them as:

   …-BXXDFLTFGTR-…

   MS BLAST will then consider BXX residues in possible sequence alignments.
5. MS BLAST options and settings:
   a. –NOGAP: absolutely essential, it turns off gapped alignment method so that only high scoring pairs (HSPs) with no internal gaps are reported
   b. –SPAN1: absolutely essential, it identifies and fetches the best matching peptide sequence among similar peptide sequences in the query. Therefore the query may contain multiple partially redundant variants of the same peptide sequence without affecting the total score of the protein hit.
   c. –HSPMAX 100 limits the total number of reported HSPs to 100. Set it to higher number (for example, 200) if a large query is submitted and complete list of protein hits including low confidence hits) is required in the output.
   d. SORT_BY_TOTALSCORE places the hits with multiple high scoring pairs to the top of the list. Note that the total score is not displayed, but can be calculated, if necessary, by adding up scores of individual HSPs.
   e. EXPECT: It is usually sufficient to set EXPECT at 100. Searching with higher EXPECT (as, 1000) will report many short low-scoring HSPs thus increasing the sequence coverage by matching more fragmented peptides to the protein sequence. Note that low scoring HSPs do not increase statistical confidence of protein identification. EXPECT setting also does not affect the scores of retrieved HSPs.
   f. MATRIX: PAM30MS is a specifically modified scoring matrix. Do not use it for conventional BLAST searching!
   g. PROGRAM: blast2p; DATABASE: nrdb95 are default settings of MS BLAST interface. In principle, MS BLAST could be also used as a ***tblastn*** program for searching EST or genomic databases.
   h. FILTER: By default, filtering is set to "none." However, if a query contains many low-complexity sequence stretches, (as, …EQEQEQ…), filtering should be set to "default."
6. MS BLAST searching. Space all candidate sequence proposals obtained from all fragmented precursors with "–" (minus) symbol and merge them into a single text string that can be pasted directly into the query window at the MS BLAST

web interface. The query may contain space symbols, hard returns, numbers, and so on since they are ignored by the server. For example, it is convenient to keep masses of precursor ions in the query since it makes retrospective analysis of data much easier.

7. Statistical evaluation of MS BLAST hits. Statistical evaluation is a very important element of MS BLAST protocol since the query typically comprises many incorrect and/or ambiguous peptide sequences. Note, that statistics of conventional BLAST searching is not applicable and therefore ignore reported E-values and P-values.

   Thresholds of statistical significance of MS BLAST hits are set conditionally on a number of reported HSPs and were estimated in a computational experiment (*see* **Table 1**). To evaluate the significance of MS BLAST hits:

   a. Check the number of fragmented precursor ions from which sequences for MS BLAST searching were obtained. Accordingly, pick up appropriate number of expected unique peptides from **Table 1**.
   b. Consider the top hit protein in MS BLAST output and check the list of matching HSPs. Pick up the HSP with the highest score and compare this score with the threshold value reported in the table for a single reported HSP.
   c. If the score of the HSP is higher than the threshold score in **Table 1**, the match is statistically significant. If not, pick up the score of the second ranked HSP, add it to the score of the first ranked HSP and compare the sum with the threshold score reported in Table 1 for two reported HSPs. Again, if the combined score exceeds the threshold, the identification is positive. Otherwise, add the score of the third ranked HSP, compare the sum with the threshold expected for three reported HSPs and so forth.

   Always start the evaluation from the highest ranked HSP reported for the given protein hit!

   If necessary, repeat the procedure for other protein hits in the MS BLAST output.

4. Example Analysis: Identification of a Protein from African Cloned Frog

Currently sequences of less than 3,500 *Xenopus* proteins are available in a public database. Therefore the characterization of proteins isolated from *Xenopus* cells usually requires *de novo* sequencing and cloning of the genes ***(14)***. Direct identification of proteins via sequence similarity searching provides a shortcut to their functional characterization by skipping time-consuming and laborious cloning experiments ***(28)***.

Identification of 120 kDa protein isolated from *Xenopus* oocytes by affinity chromatography and one-dimensional electrophoresis is presented as an example. A Coomassie stained band was excised from the gel, the protein was in-gel digested with trypsin and recovered tryptic peptides were sequenced on a QSTAR Pulsar *i* quadrupole time-of-flight mass spectrometer. Tandem spectra were

Table 1
Threshold Scores for Statistical Evaluation of MS BLAST Hits[a]

Number of reported HSPs	Number of unique peptides in the query		
	10	20	50
1	68	72	75
2	102	106	111
3	143	146	153
4	177	< 208[b]	< 180[b]
5	< 238[b]	n.o.[c]	< 212[b]
6	n.o.	n.o.	< 275[b]
7	n.o.	n.o.	< 285[b]

[a]Threshold scores were estimated in computational experiment ***(19)*** and are provided for guidance rather than as a stringent identification criterion.

[b]The calculated value is statistically unreliable because just a few hits matching with the specified number of HSPs were observed. In those cases the maximal score from the ones observed is presented.

[c]n.o. No random hits with specified number of HSPs were observed.

acquired from 21 multiply charged peptide precursor ions (*see* **Fig. 1**) and were subjected to manual and/or automated interpretation. For example, in the MS/MS spectrum acquired from a doubly charged ion with *m/z* 883.44 three amino acid residues (…DYT…) were confidently called in the *m/z* region above the precursor, which covers the N-terminal part of the peptide sequence (*see* **Fig. 2**). The sequence stretch could be extended a few residues further down to the lower *m/z* region (towards the C-terminus of the peptide). However, the sequence became ambiguous and could not be anchored at C-terminal arginine or lysine residues, expected from the cleavage specificity of trypsin. Automated interpretation produced a few ambiguous candidate sequences that covered C-terminal part of the peptide, but they did not extend far enough towards the N-terminus and did not overlap with the sequence obtained by manual interpretation. All sequences were then edited and merged according to MS BLAST requirements (*see* **Subheading 3.4.**) and combined with sequences obtained by the interpretation of other tandem mass spectra. Altogether, 793 peptide sequences of varying quality, completeness, and length were merged and the text string was submitted to MS BLAST searching at the EMBL server. A few homologous catalytic subunits of DNA polymerase delta from various species were hit with the statistical significance exceeding the threshold scores (*see* **Table 2**) (*see* **Note 10**). Altogether, 10 peptide sequences matched to bovine DNA polymerase delta (although only two tryptic peptides matched its sequence exactly) and additionally two peptides matched to homologous polymerases from other species.

Thus a combination NanoES tandem mass spectrometry and MS BLAST sequence similarity searching allowed rapid and confident identification of the

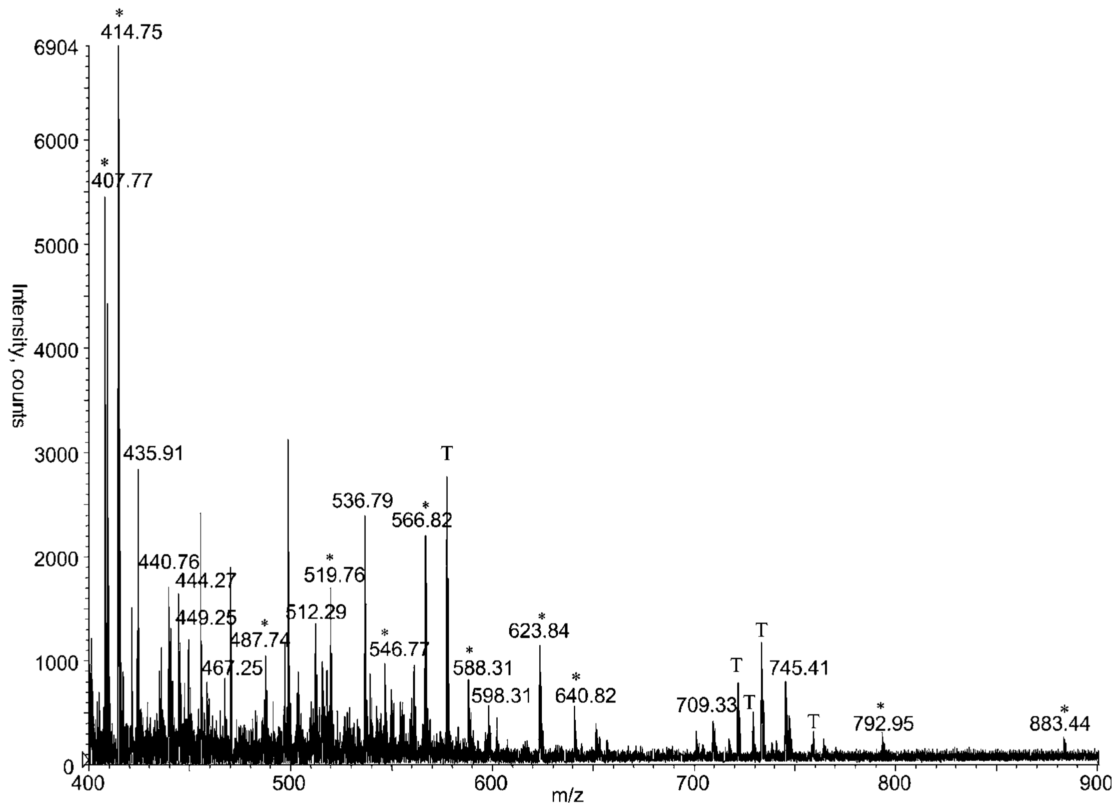


Fig. 1. TOF MS spectrum of the in-gel tryptic digest of 120 kDa *Xenopus* protein. Tandem mass spectra were acquired from peaks designated with *m/z*. Peaks originating from trypsin autolysis products are designated with T. Peaks of the peptides that matched the sequence of bovine DNA polymerase delta are labeled with asterisks.

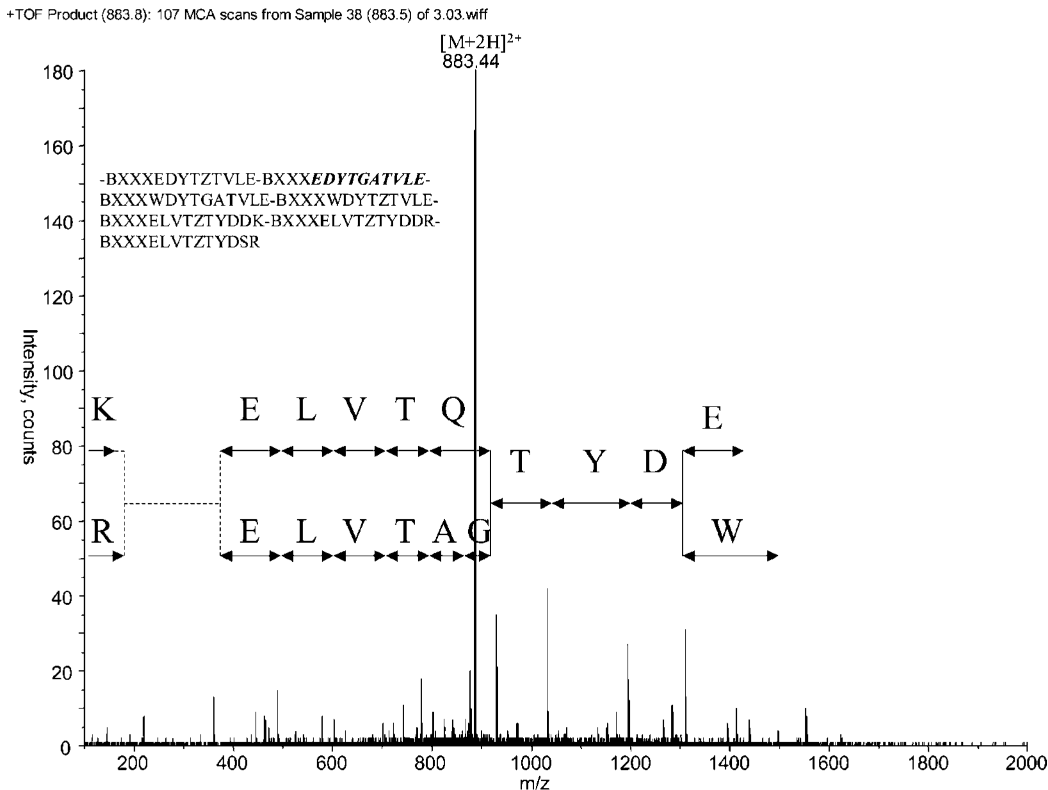


Fig. 2. Tandem mass spectrum, acquired from the doubly charged precursor ion with *m/z* 883.44. Manual interpretation of spectra considered precise mass difference between adjacent Y" ions starting from the m/z segment above the precursor ion (correspondent peaks and amino acid residues are designated by arrows). Automated interpretation resulted in a few partially redundant ambiguous sequences, covering the C-terminus of the peptide (inset).

Table 2
Identification of a *Xenopus* protein by MSBLAST Sequence Similarity Searching

m/z	*z*	Precursor mass	Automated interpretation	Manual interpretation	Bovine	Mouse	Rat	Human	Yeast	Arabidopsis	Score
407.77	2	813.52	/	**BVVSZLLR**	+	+	+	+			47
414.75	2	827.49	/	**BLLEZGLR**	+	+	+	+			46
435.91	3	1304.70	YES	BXXXTAVLZD			+				
440.76	2	879.50	YES	BLAVYD							-
444.27	3	1329.79	YES	BXXAHFNTAVLK,							
				BXXXAHFNTAVLK							-
449.25	3	1344.73	NONE	BXXXXADLL,							
				BXXXXXADLL							-
467.25	2	932.49	NONE	BTPTPT							-
487.74	2	973.46	YES	**BYTLDDGY**K	+	+	+	+			42
512.29	2	1022.57	YES	BVSTFPG							-
519.76	2	1037.52	YES	**BTPTGD**ZV	+	+	+	+			41
536.79	2	1071.56	YES	BLALZDPFLR							-
546.77	2	1091.52	YES	BLZDLSDFZK							-
566.82	2	1131.62	YES	**BLFEPLL**	+	+	+	+			51
588.31	2	1174.60	NONE	**BVLSFDLE**	+	+		+		+	53
598.31	2	1194.62	YES	BYGLNPEDFLK		+	+				
623.84	2	1245.66	YES	BXXSZ**LSALEE**K	+	+	+	+			54
640.82	2	1279.65	/	**BVLSFDLE**E	+	+		+			54
709.33	2	1416.65	NONE	BXXEVPDZ							-
745.41	2	1488.80	YES	BXXXTVAEA,							-
				BXXXXTVAEA							
792.95	2	1583.89	YES	BXADSVYGFT,	+	+	+	+	+	+	58
				BXXADSVYGFT							
883.44	2	1764.86	YES	BXXX**EDYTGATVLE**	+	+	+	+	+		72
				BXXXEDYTZTVLE							

[a]Bovine DNA Polymerase delta was the top hit, scores of matched HSPs are presented. Sequence stretches in bold matched the bovine sequences exactly. Peptides for which the software predicted no sequences are labeled "NONE." Peptides for which the automated sequence predictions were made are labeled "YES." For peptides labeled "/," no automatically predicted sequences were included because high-quality sequences were retrieved directly from Y-ion series in the spectrum.

protein using only rapid and "rough" interpretation of tandem mass spectra (*see* **Note 11**).

5. Notes

1. All chemicals should be of the highest degree of purity available. Plastic ware (pipet tips, gloves, dishes, etc.) may acquire a static charge and attract dust thus leading to massive contamination of samples with human and sheep keratins during in-gel digestion. Any polymeric detergents (Tween, Triton) should not be used for cleaning the laboratory dishes and tools.
2. Methanol (1 mL) is added to *ca.* 30 µL of POROS R2 resin to prepare a slurry. A submicrometer-sized fraction of the resin beads, whose presence increases the resistance to liquid flow, is removed by repetitive sedimentation. Vortex the test tube containing the slurry and then let it stay in a rack until the major part of the resin reaches the bottom of the tube. Aspirate the supernatant with pipet and discard it. Repeat the procedure 3–5 times if necessary.
3. Make sure that the settings controlling resolution of the first quadrupole (Q1) allow good transmission of precursor ions. On the other hand, unnecessary low resolution of Q1 results in the transmission of too many background ions, which may densely populate the low *m/z* region of the MS/MS spectra. The third quadrupole (Q3) should likewise be operated at a low resolution settings in order to improve its transmission and to achieve acceptable ion statistics in the MS/MS spectra. In our experience, resolution of Q3 as low as 250 (FWHM) still allows accurate readout of peptide sequences. The Q1 and Q3 resolution settings can be tuned in tandem mass spectrometric experiment using synthetic peptides.

   Resolution of the first quadrupole (Q1) should be tuned in a similar way as described for a triple quadrupole mass spectrometer. External calibration using masses of [Glu]-fibrinogen peptide fragments allows better than 20 ppm mass accuracy for both TOF MS and product ion operation modes, if calibration and sequencing experiments are performed within *ca.* 2 h.
4. *De novo* sequencing software is often included in the software packages shipped together with mass spectrometers: BioMultiview, BioAnalyst (both are from MDS Sciex, Canada); BioMassLynx (Micromass Ltd, UK), BioTools (Bruker Daltonics, Germany). Lutefisk program ***(20,29)*** can be downloaded from: http://www.immunex.com/researcher/lutefisk/. Automated interpretation of tandem mass spectra often requires adjusting of settings that affect scoring of candidate sequences. It is therefore advisable to test the settings in advance using digests of standard proteins and to adjust them if necessary. Note that the settings may depend on a charge state of the fragmented precursor ion. Use only standard one-letter code for amino acid residues (*see* **Appendix 1**). If the software introduces special symbols for modified amino acid residues, replace them with standard codes.
5. In-gel digestion of proteins with trypsin was described in detail in ***(30–32)***.
6. In-gel digestion of proteins with unmodified trypsin is accompanied by trypsin autolysis. Therefore it is necessary to acquire the spectrum of a control sample

(blank gel pieces processed using the same in-gel digestion method) in advance. Spectra should be acquired both in conventional single MS mode (Q1 scanning on a triple quadrupole or TOF MS mode on a quadrupole TOF machines) and in precursor ion scanning mode (triple quadrupole instruments only).

7. Scanning for precursor ions that produce the characteristic fragment ions with *m/z* 86 (immonium ion of leucine or isoleucine) helps to distinguish genuine peptide ions from chemical noise ***(33)***.
8. It is not known in advance if the sequence of the analyzed protein is already present in a database. Therefore conventional database searching routines based on stringent matching of peptide sequences should be applied first ***(9,34)***. Only if the protein is unknown and no convincing cross-species matches were obtained, it is recommended to proceed with *de novo* interpretation of tandem mass spectra and sequence similarity searching.
9. The algorithm and principles of BLAST homology searching are discussed in detail in ***(27,35)***. A useful list of BLAST servers accessible on the web is provided in ***(36)***.
10. The top peptide matched with the score 72, which as such is not sufficient for positive identification (the threshold score for a single matched peptide is 72, *see* **Table 1**). Taken together with the score of the second ranked peptide matched (58), the total score of two matched peptides is 72 + 58 = 130, which is higher than the threshold score for two matched peptides that is 106 (*see* **Table 1**) and the protein was positively identified. More matching peptide (*see* **Table 2**) additionally validated the hit.
11. The success of MS BLAST identification depends on the number of sequenced peptides, on the quality of peptide sequences and on the sequence similarity between the protein of interest and its homologues available in a database. On average, candidate sequences determined for five tryptic peptides should be submitted to MS BLAST searching to identify the protein by matching to a homolog from the organism within the same kingdom.

References

1. Chalmers, M. J. and Gaskell, S. J. (2000) Advances in mass spectrometry for proteome analysis. *Curr. Opin. Biotechnol.* **11**, 384–390.
2. Lahm, H. W. and Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **21**, 2105–2114.
3. Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405**, 837–846.
4. Mann, M., Hendrickson, R. C., and Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473.
5. Podtelejnikov, A. V., Bachi, A., and Mann, M. (1998) Cross-species identification of proteins by high mass accuracy MALDI peptide mapping, in *Proceedings 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, Abstract no. 212.

6. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., et al. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nanoelectrospray mass spectrometry. *Nature* **379**, 466–469.
7. Shevchenko, A., Wilm, M., and Mann, M. (1997) Peptide sequencing by mass spectrometry for homology searches and cloning of genes. *J. Protein Chem.* **16**, 481–490.
8. Roepstorff, P. and Fohlman, J. (1984) Proposed nomenclature for sequence ions. *Biomed. Mass Spectrom.* **11**, 601.
9. Mann, M. and Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399.
10. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986) Peptide sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **83**, 6233–6237.
11. Shevchenko, A., Chernushevich, I., Ens, W., Standing, K. G., Thomson, B., Wilm, M., et al. (1997) Rapid 'De Novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **11**, 1015–1024.
12. McNagny, K. M., Petterson, I., Rossi, F., Flamme, I., Shevchenko, A., Mann, M., et al. (1997) Thrombomucin, a novel cell surface protein that defines thrombocytes and multipotent hematopoetic progenitors. *J. Cell Biol.* **138**, 1395–1407.
13. Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V., and Cech, T. R. (1997) Reverse transcriptase motifs in the catalytic subunits of telomerase. *Science* **276**, 561–567.
14. Chen, R. H., Shevchenko, A., Mann, M., and Murray, A. W. (1998) Spindle checkpoint protein Xmad1 recruits Xmad2 to unattached kinetochores. *J. Cell Biol.* **143**, 283–295.
15. Muzio, M., Chinnaiyan, A. M., Kischkel, F. C., Rourke, K. O., Shevchenko, A., Jian N., et al. (1996) FLICE, a novel FADD-homologous ICE/CED-3-like protease, is recruited to the CD95 (Fas/APO-1) death-inducing signaling complex. *Cell* **85**, 817–827.
16. Aigner, S., Lingner, J., Goodrich, K. J., Grosshans, C. A., Shevchenko, A., Mann, M., et al. (2000) Euplotes telomerase contains an La motif protein produced by apparent translational frameshifting. *EMBO J.* **19**, 6230–6239.
17. Taylor, J. A . and Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**, 1067–1075.
18. Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A., Wang, C. C., Burlingame, A. A., et al. (2001) Functional assignment of the 20S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* **17**, 28327–28339.
19. Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., et al. (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.
20. Johnson, R. S. and Taylor, J. A. (2000) Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Methods Mol. Biol.* **146**, 41–61.

21. Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., and Standing, K. G. (2000) MALDI Quadrupole Time-of-Flight mass spectrometry: a powerful tool for proteomic research. *Anal. Chem.* **72**, 2132–2141.
22. Medzihradszky, K. F., Campbell, J. M., Baldwin, M. A., Falick, A. M., Juhasz, P., Vestal, M. L., et al. (2000) The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **72**, 552–558.
23. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., et al. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682.
24. Wilm, M. and Mann, M. (1996) Analytical properties of the nano electrospray ion source. *Anal. Chem.* **66**, 1–8.
25. Jensen, O. N., Wilm, M., Shevchenko, A., and Mann, M. (1999) Peptide sequencing of 2-DE gel-isolated proteins by nanoelectrospray tandem mass spectrometry. *Methods Mol. Biol.* **112**, 571–588.
26. Shevchenko, A., Chernushevich, I., and Mann, M. (1998) High sensitivity analysis of gel separated proteins by a quadrupole – TOF tandem mass spectrometer, in *Proceedings 46th ASMS Conference on Mass Spectrometry and Allied Topics*, Orlando, FL, Abstract no. 237.
27. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
28. Tournebize, R., Popov, A., Kinoshita, K., Ashford, A. J., Rybina, S., Pozniakovsky, A., et al. (2000) Control of microtubule dynamics by the antagonistic activities of XMAP215 and XKCM1 in *Xenopus* egg extracts. *Nat. Cell Biol.* **2**, 13–19.
29. Taylor, J. A. and Johnson, R. S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73**, 2594–2604.
30. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858.
31. Shevchenko, A., Chernushevich, I., Wilm, M., and Mann, M. (2000) *Methods Mol. Biol.* **146**, 1–16.
32. Shevchenko, A. (2001) Evaluation of the efficiency of in-gel digestion of proteins by peptide isotopic labeling and MALDI mass spectrometry. *Anal. Biochem.* **296**, 279–283.
33. Wilm, M., Neubauer, G., and Mann, M. (1996) Parent ion scans of unseparated peptide mixtures. *Anal. Chem.* **68**, 527–533.
34. 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.
35. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6**, 119–129.
36. Gaeta, B. A. (2000) BLAST on the Web. *Biotechniques* **28**, 436–440.