C H A P T E R

51

Identification of Proteins from Organisms with Unsequenced Genomes by Tandem Mass Spectrometry and Sequence-Similarity Database Searching Tools

Adam J. Liska and Andrej Shevchenko

I. INTRODUCTION

The analysis of proteomes by mass spectrometric methods that correlate peptide fragments from proteins with database entries *in silico* has been dependent on the sequencing of genomes. Mass spectrometry and database sequences have enabled the analysis of the human, mouse, and *Arabidopsis* proteomes, among others. Due to the high homology between living organisms at the molecular level, it is possible to use the available protein sequences accumulated in databases from a range of organisms as a reference for the identification of proteins from organisms with unsequenced genomes by sequence-similarity database searching. As research continues in organisms such as *Xenopus*, maize, cow, and others with limited database sequence resources, sequence-similarity searching is a powerful method for protein identification. This article focuses on MS BLAST (Shevchenko *et al.*, 2001) and MultiTag (Sunyaev *et al.*, 2003) as bioinformatic methods for the identification of proteins by the interpretation of tandem mass spectra of peptides and sequence-similarity searching.

II. MATERIALS AND INSTRUMENTATION

In analyses where MS BLAST is utilized for protein identification, tandem mass spectra of peptides can be acquired with any ionization source and mass spectrometer that enables *de novo* sequence prediction: nanoelectrospray, LC/MS/MS, o-MALDI (MALDI quadrupole TOF), MALDI TOF-TOF, QqTOF, triple quad, PSD MALDI-TOF, and ion trap. Alternatively, in analyses where MultiTag is applied, tandem mass spectra of peptides can be acquired with any ionization source and tandem mass spectrometer that enables the creation of peptide sequence tags (Mann and Wilm, 1994): nanoelectrospray, LC/MS/MS, QqTOF, triple quad, or other novel system.

Software for *de novo* sequence prediction from tandem mass spectra is often included in the software packages associated with mass spectrometers: BioMultiview, BioAnalyst (both are from MDS Sciex, Canada), BioMassLynx (Micromass Ltd, UK), BioTools (Bruker Daltonics, Germany), and DeNovoX (ThermoFinnigan). The Lutefisk program (Johnson and Taylor, 2000) can be acquired from http://www.

hairyfatguy.com/Sherpa/. BioAnalyst with the Pro-Blast processing script can generate a complete MS BLAST query automatically from multiple-spectra files acquired by nanoelectrospray or LC/MS/MS (Nimkar and Loo, 2002). A web browser such as Internet Explorer or Netscape is also required to gain access to the MS BLAST web interface located at http://dove.embl-heidelberg.de/Blast2/msblast.html. An independent BLAST computer (Paracel BlastMachine system) may also be purchased and installed for rapid and private MS BLAST operation.

Sequence tags that contain a few confidently designated amino acid residues and mass values that lock the short sequence stretch into the length of the peptide can often be generated from tandem mass spectra with the software packages associated with mass spectrometers. Microsoft Excel or an alternative spreadsheet program is required for compiling of search results before submission to MultiTag. BioAnalyst software has an associated processing script that produces a list of database search results from a generated list of sequence tags to accelerate spectra processing with MultiTag.

III. PROCEDURES

A. Identification of Proteins by MS BLAST Database Searching

1. MS BLAST is a specialized BLAST-based tool for the identification of proteins by sequence-similarity searching that utilizes peptide sequences produced by the interpretation of tandem mass spectra (Shevchenko *et al.*, 2001). The algorithm and principles of BLAST sequence-similarity searching are reported in detail elsewhere (Altschul *et al.*, 1997). A useful list of BLAST servers accessible on the web is provided in Gaeta (2000).

2. Peptide sequences are generated from the interpretation of tandem mass spectra from the analysis of a single in-gel or in-solution digest of an unknown protein, edited and assembled into a query list for the MS BLAST search. 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) (Fig. 1). Automated interpretation of tandem mass spectra often requires adjustment of parameters that affect the quality of predicted 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 the standard single-letter symbols for amino acid residues. If the software introduces special symbols for modified amino acid residues, replace them with standard symbols.

3. When interpreting MS/MS spectra manually, 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 or four amino acid sequence stretch deduced from a noise-free high *m/z* segment of the spectrum.

4. Gaps and ambiguities in peptide sequences can occur due to the fragmentary nature of tandem mass spectra of peptides. 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,

DTPS[...]HYNAR, [...] = [S, V] or [D, A]

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

-DTPSSVHYNAR-DTPSVSHYNAR-
DTPSDAHYNAR-DTPSADHYNAR-

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 and therefore it matches weakly any amino acid residue:

-DTPSXXHYNAR-

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

-DTPSXXHYNAR-DTPSXXXHYNAR-

5. Isobaric amino acids need to be altered in the MS BLAST query. L stands for Leu (L) and Ile (I). Z stands for Gln (Q) and Lys (K), if undistinguishable in the spectrum. Use Q or K if the amino acid residue can be determined. The query string needs to be further altered for cleavage site specificity. If the proposed sequence is complete, a putative trypsin cleavage site symbol B is added prior to the peptide sequence:

...-BDTPSVDHYNAR-

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

...-BXXPSVDHYNAR-...

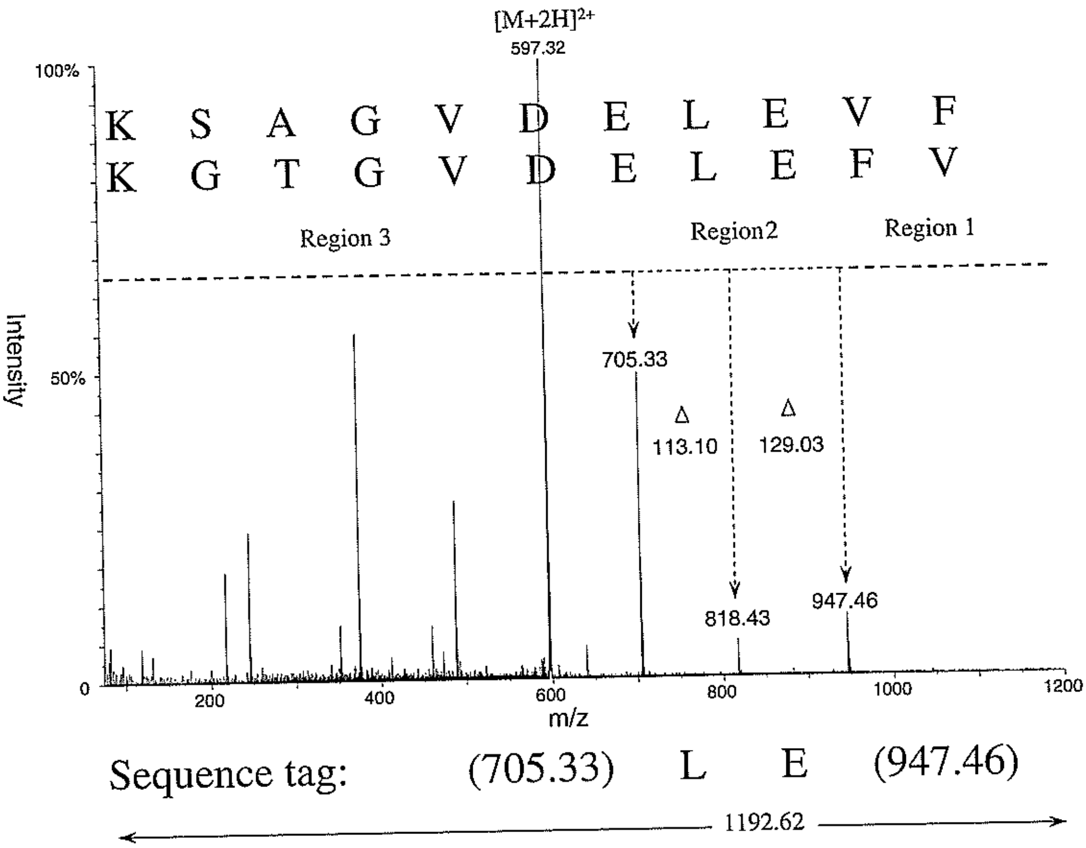


FIGURE 1 A spectrum with *de novo*-predicted amino acid sequences and a manually constructed sequence tag. Multiple candidate peptide sequences can be generated from a single spectrum for MS BLAST analysis, whereas MultiTag requires one sequence tag per spectrum.

MS BLAST will then consider BXX residues in possible sequence alignments.

6. The regular BLAST search must be altered in options and settings for an MS BLAST query:

NOGAP is absolutely essential, it turns off gapped alignment method so that only high-scoring pairs (HSPs) with no internal gaps are reported.

SPAN1 is 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.

HSPMAX 100 limits the total number of reported HSPs to 100. Set it to a higher number (e.g., 200) if a large query is submitted and a complete list of protein hits (including low confidence hits) is required in the output.

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.

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. The EXPECT setting also does not affect the scores of retrieved HSPs.

MATRIX: PAM30MS is a specifically modified scoring matrix. It is not used for conventional BLAST searching.

PROGRAM: blast2p.

DATABASE: nrdb95 are default settings of the MS BLAST interface.

FILTER: Filtering is set to "none" default. However, if the sequence query contains many repeating stretches (as . . . EQEQEQ . . .), filtering should be set to "default."

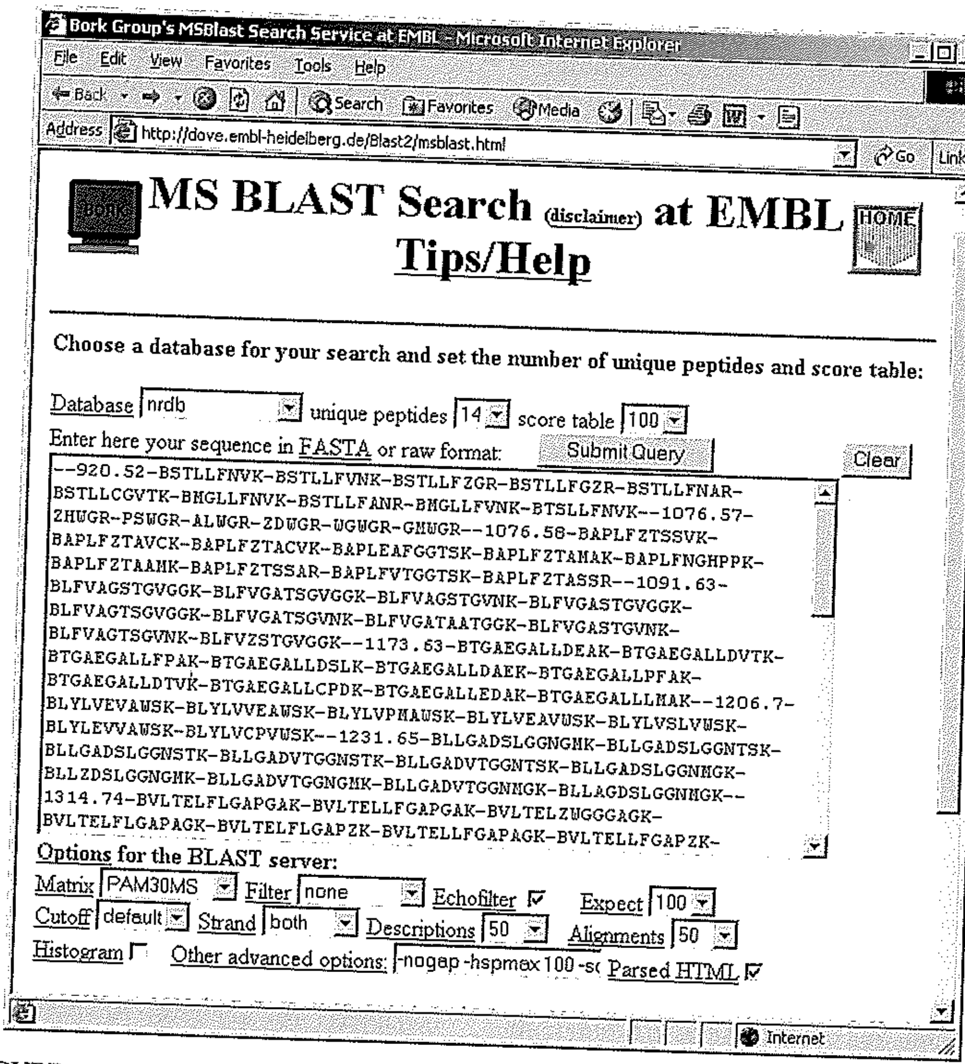


FIGURE 2 The MS BLAST web interface. A generated query is pasted in the input window, the number of tandem mass spectra from which sequences were derived is input in "unique peptides," and all other settings are set automatically according to MS BLAST parameters.

At the EMBL web interface, all parameters are preset and only the number of fragmented peptides and query sequences need to be input.

7. Space all candidate sequence proposals obtained from MS/MS spectra with a "–" (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 (Fig. 2). The query may contain space symbols, hard returns, numbers, and so on, as the server ignores them. For example, it is convenient to keep masses of precursor ions in the query, as it makes retrospective analysis of data much easier. Statistical evaluation is a very important element of MS BLAST protocol, as the query typically comprises many incorrect and partially redundant peptide sequences. Note that the statistics of conventional BLAST searching are not applicable and therefore ignore reported E-values and P-values. Thresholds of statistical significance of MS BLAST hits were estimated in a computational experiment and scoring thresholds were set conditionally on the number of reported HSPs and the size of the database searched. Experimental MS BLAST hits are evaluated based on the number of fragmented precursors (this value is entered in the search parameters), and confident hits appear in red, borderline hits in green, and random matches in black at the web interface cited earlier (Fig. 3).

8. MS BLAST can, in principle, be used to search protein, EST, and genomic databases. The EMBL site only supports protein BLAST searching due to available computational capacity. A script can be written to retrieve specific genomic sequences that lie within the

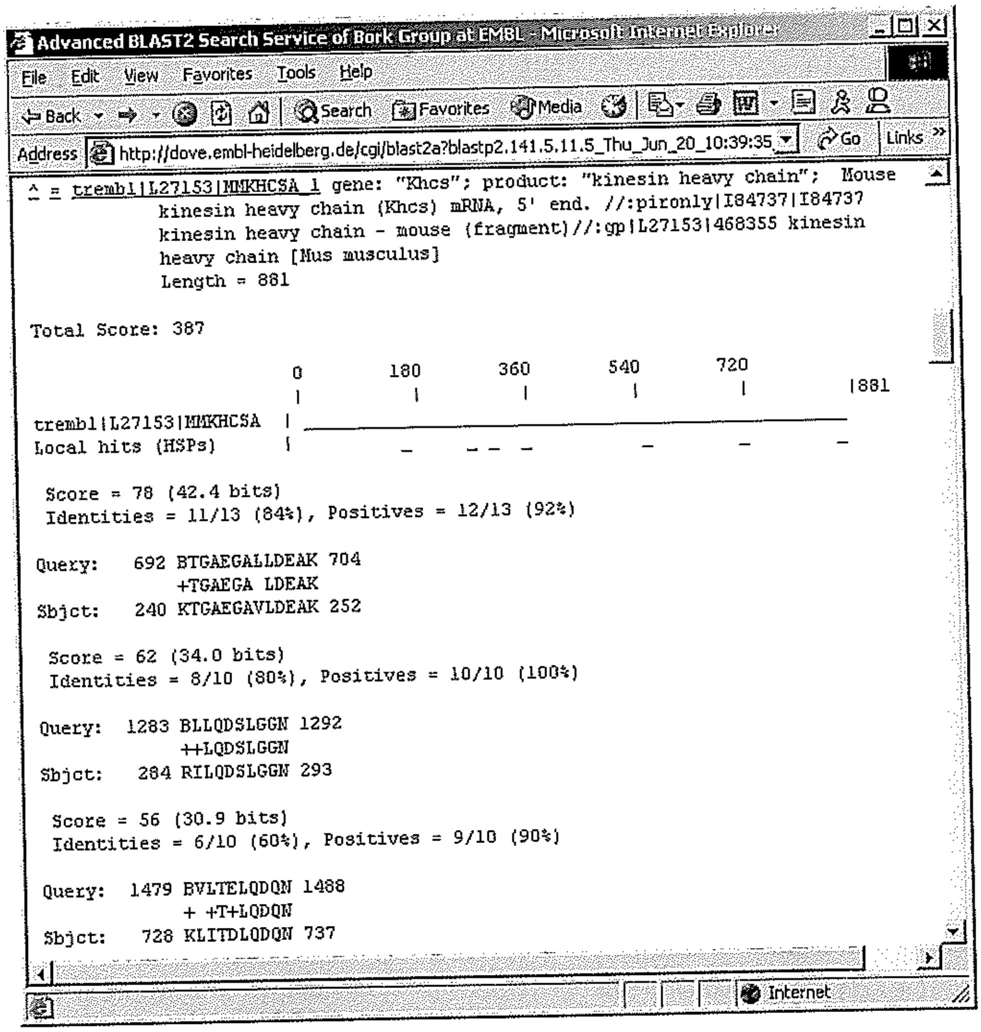


FIGURE 3 A section of the MS BLAST output where three of seven matched peptides to one database entry are shown. A list of matching database entries with the highest significance match at the top of the list is generated. Significant hits are color coded for easy data interpretation.

igned peptides from a *tblastn* search against specific enomic databases. This search enables the use of nannotated genomic sequences and makes it possible) identify novel genes in large genomes. However, in oth EST and genomic searches, different scoring chemes would need to be developed and installed ocally. To set up a local BLAST searching engine, VU-BLAST 2.0 can be acquired from http://blast.wustl.edu/.

3. Identification of Proteins by MultiTag Database Searching

1. MultiTag is a software program that sorts compiled results from database searches with partial and complete sequence tags and calculates the significance of matches that align multiple sequence tags (for a complete description, see Sunyaev *et al.*, 2003). MultiTag is based on error-tolerant searching with multiple partial sequence tags. This technique enables the correlation of search results from multiple searches with sequence tags representative of numerous spectra and gauges the significance of those matches.

2. Sequence tags should be generated manually from tandem mass spectra acquired in the analysis of a single in-gel or in-solution digest of an unknown protein or proteins. Some mass spectrometer software enables the automatic prediction of sequence tags; however, for best results it is advisable to make sequence tags by manual interpretation or gauge the accuracy of the automatic prediction software with a standard protein prior to the analysis of unknown samples. One sequence tag per spectrum should be made using prominent Y-ions, usually larger

than the multiply charged precursor. Sequence tags made with two to four amino acids each from multiple MS/MS spectra should be compiled in a text file list that includes the tag followed by the parent mass.

(360.20)FLL(733.44)918.64

(561.27)LA(745.40)935.48

(866.41)DEA(1181.52)1422.59

3. Each sequence tag is used to search a protein database, and the results from four searches are compiled in a spreadsheet. For the most specific results, mass tolerances should be narrow, taking into consideration the best accuracy of the mass spectrometer employed. The database is first searched using the complete sequence tag:

(866.41)DEA(1181.52)

This search will only find proteins that contain peptides with exactly these amino acids, spaced with exact amino acids residues that give mass combinations to make up the gaps to the peptide's termini. The second search allows for one error within the amino acid representation itself:

(866.41)D?A(1181.52)

The third search allows for errors between the analyzed peptide and the database at the C terminus of the peptide (searching with regions 1 and 2 only):

DEA(1181.52)

The fourth search allows for errors between the analyzed peptide and the database at the N terminus of the peptide (searching with regions 2 and 3 only):

(866.41)DEA

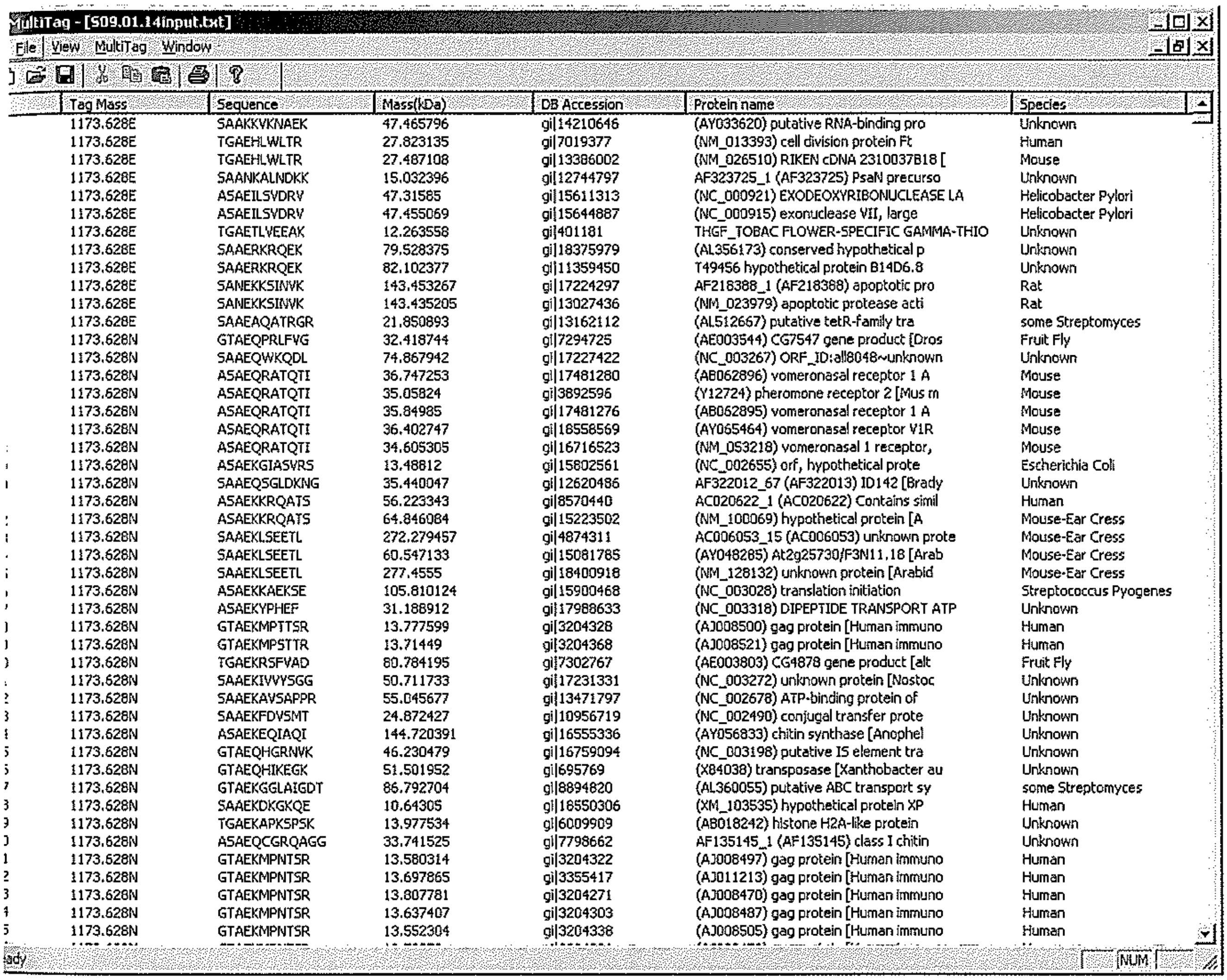
In the first column of the results table, the parent mass should be followed by an "NC", "E", "N", or "C" for sequence tag search results from complete, one-error, regions 1 and 2 (N-terminal match), and regions 2 and 3 (C-terminal match), respectively. The second, third, fourth, fifth, and six columns should include the amino acid sequence of the peptide matched, the molecular weight of the protein the tag matched, the accession number of the entry, the name of the protein, and the species name, respectively (Fig. 4). Compile the search results in a spreadsheet from all searches with the sequence tags generated, and save results as a text file.

4. Submit the search results table to MultiTag. Designate the mass tolerance used for searching (in daltons), input the approximate number of entries in the database searched, input the list of tags used to generate search results, and compute significance. Results will be sorted, with the database entry containing the most unlikely correlation event at the top of the list and probabilities are calculated (Fig. 5). Results can be evaluated based on the number of sequence tags matched, E-values and P-values [count (predicted)]. E-values lower than 1×10^{-3} and P-values lower than 1×10^{-4} can be considered significant matches. Final E-values are highly dependent on the number of tags submitted for database searching; more tags will tend to diminish the significance of the alignment of multiple tags to one database entry. Reported P-values are less affected by the number sequence tags in the query. P-values reflect an approximation of the probability that the tags that are aligned will match randomly to an entry in a database of a specific size and at a specific mass accuracy, while neglecting the query size (number of sequence tags). Low P-values (but with higher E-values) are good indicators of alignments having borderline significance that need further manual evaluation to conclude a confident identification. Three partial sequence tags are normally specific enough to identify one entry in a database of 1,000,000 entries, at a mass accuracy of 0.1 Da.

IV. COMMENTS

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 (Mann and Wilm, 1994; Perkin *et al.*, 1999). Only if the protein is unknown and no convincing cross-species matches can be obtained is it recommended to proceed with *de novo* interpretation of tandem mass spectra and sequence-similarity searching.

The success of MS BLAST and MultiTag identification depends on the size of a query and the corresponding database, the number of peptides aligned, the quality of peptide sequences or sequence tags, and 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 or MultiTag searching to identify the protein by matching to a homologous sequence. With 10 sequences submitted and aligned, MS BLAST can identify 50% of homologues containing 50% sequence similarity, and MultiTag can identify 50% of homologues at 70% sequence similarity. However, because MultiTag can utilize less



<table>
<tr><th>Tag Mass</th><th>Sequence</th><th>Mass(kDa)</th><th>DB Accession</th><th>Protein name</th><th>Species</th></tr>
<tr><td>1173.628E</td><td>SAAKKVKNAEK</td><td>47.465796</td><td>gi|14210646</td><td>(AY033620) putative RNA-binding pro</td><td>Unknown</td></tr>
<tr><td>1173.628E</td><td>TGAEHLWLTR</td><td>27.823135</td><td>gi|7019377</td><td>(NM_013393) cell division protein Ft</td><td>Human</td></tr>
<tr><td>1173.628E</td><td>TGAEHLWLTR</td><td>27.487108</td><td>gi|13386002</td><td>(NM_026510) RIKEN cDNA 2310037B18 [</td><td>Mouse</td></tr>
<tr><td>1173.628E</td><td>SAANKALNDKK</td><td>15.032396</td><td>gi|12744797</td><td>AF323725_1 (AF323725) PsaN precurso</td><td>Unknown</td></tr>
<tr><td>1173.628E</td><td>ASAEILSVDRV</td><td>47.31585</td><td>gi|15611313</td><td>(NC_000921) EXODEOXYRIBONUCLEASE LA</td><td>Helicobacter Pylori</td></tr>
<tr><td>1173.628E</td><td>ASAEILSVDRV</td><td>47.455069</td><td>gi|15644887</td><td>(NC_000915) exonuclease VII, large</td><td>Helicobacter Pylori</td></tr>
<tr><td>1173.628E</td><td>TGAETLVEEAK</td><td>12.263558</td><td>gi|401181</td><td>THGF_TOBAC FLOWER-SPECIFIC GAMMA-THIO</td><td>Unknown</td></tr>
<tr><td>1173.628E</td><td>SAAERKRQEK</td><td>79.528375</td><td>gi|18375979</td><td>(AL356173) conserved hypothetical p</td><td>Unknown</td></tr>
<tr><td>1173.628E</td><td>SAAERKRQEK</td><td>82.102377</td><td>gi|11359450</td><td>T49456 hypothetical protein B14D6.8</td><td>Unknown</td></tr>
<tr><td>1173.628E</td><td>SANEKKSINVK</td><td>143.453267</td><td>gi|17224297</td><td>AF218388_1 (AF218388) apoptotic pro</td><td>Rat</td></tr>
<tr><td>1173.628E</td><td>SANEKKSINVK</td><td>143.435205</td><td>gi|13027436</td><td>(NM_023979) apoptotic protease acti</td><td>Rat</td></tr>
<tr><td>1173.628E</td><td>SAAEAQATRGR</td><td>21.850893</td><td>gi|13162112</td><td>(AL512667) putative tetR-family tra</td><td>some Streptomyces</td></tr>
<tr><td>1173.628N</td><td>GTAEQPRLFVG</td><td>32.418744</td><td>gi|7294725</td><td>(AE003544) CG7547 gene product [Dros</td><td>Fruit Fly</td></tr>
<tr><td>1173.628N</td><td>SAAEQWKQDL</td><td>74.867942</td><td>gi|17227422</td><td>(NC_003267) ORF_ID:all8048~unknown</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>ASAEQRATQTI</td><td>36.747253</td><td>gi|17481280</td><td>(AB062896) vomeronasal receptor 1 A</td><td>Mouse</td></tr>
<tr><td>1173.628N</td><td>ASAEQRATQTI</td><td>35.05824</td><td>gi|3892596</td><td>(Y12724) pheromone receptor 2 [Mus m</td><td>Mouse</td></tr>
<tr><td>1173.628N</td><td>ASAEQRATQTI</td><td>35.84985</td><td>gi|17481276</td><td>(AB062895) vomeronasal receptor 1 A</td><td>Mouse</td></tr>
<tr><td>1173.628N</td><td>ASAEQRATQTI</td><td>36.402747</td><td>gi|18558569</td><td>(AY065464) vomeronasal receptor V1R</td><td>Mouse</td></tr>
<tr><td>1173.628N</td><td>ASAEQRATQTI</td><td>34.605305</td><td>gi|16716523</td><td>(NM_053218) vomeronasal 1 receptor,</td><td>Mouse</td></tr>
<tr><td>1173.628N</td><td>ASAEKGIASVRS</td><td>13.48812</td><td>gi|15802561</td><td>(NC_002655) orf, hypothetical prote</td><td>Escherichia Coli</td></tr>
<tr><td>1173.628N</td><td>SAAEQSGLDKNG</td><td>35.440047</td><td>gi|12620486</td><td>AF322012_67 (AF322013) ID142 [Brady</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>ASAEKKRQATS</td><td>56.223343</td><td>gi|8570440</td><td>AC020622_1 (AC020622) Contains simil</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>ASAEKKRQATS</td><td>64.846084</td><td>gi|15223502</td><td>(NM_100069) hypothetical protein [A</td><td>Mouse-Ear Cress</td></tr>
<tr><td>1173.628N</td><td>SAAEKLSEETL</td><td>272.279457</td><td>gi|4874311</td><td>AC006053_15 (AC006053) unknown prote</td><td>Mouse-Ear Cress</td></tr>
<tr><td>1173.628N</td><td>SAAEKLSEETL</td><td>60.547133</td><td>gi|15081785</td><td>(AY048285) At2g25730/F3N11.18 [Arab</td><td>Mouse-Ear Cress</td></tr>
<tr><td>1173.628N</td><td>SAAEKLSEETL</td><td>277.4555</td><td>gi|18400918</td><td>(NM_128132) unknown protein [Arabid</td><td>Mouse-Ear Cress</td></tr>
<tr><td>1173.628N</td><td>ASAEKKAEKSE</td><td>105.810124</td><td>gi|15900468</td><td>(NC_003028) translation initiation</td><td>Streptococcus Pyogenes</td></tr>
<tr><td>1173.628N</td><td>ASAEKYPHEF</td><td>31.188912</td><td>gi|17988633</td><td>(NC_003318) DIPEPTIDE TRANSPORT ATP</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPTTSR</td><td>13.777599</td><td>gi|3204328</td><td>(AJ008500) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPSTTR</td><td>13.71449</td><td>gi|3204368</td><td>(AJ008521) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>TGAEKRSFVAD</td><td>80.784195</td><td>gi|7302767</td><td>(AE003803) CG4878 gene product [alt</td><td>Fruit Fly</td></tr>
<tr><td>1173.628N</td><td>SAAEKIVVYSGG</td><td>50.711733</td><td>gi|17231331</td><td>(NC_003272) unknown protein [Nostoc</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>SAAEKAVSAPPR</td><td>55.045677</td><td>gi|13471797</td><td>(NC_002678) ATP-binding protein of</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>SAAEKFDVSMT</td><td>24.872427</td><td>gi|10956719</td><td>(NC_002490) conjugal transfer prote</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>ASAEKEQIAQI</td><td>144.720391</td><td>gi|16555336</td><td>(AY056833) chitin synthase [Anophel</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>GTAEQHGRNVK</td><td>46.230479</td><td>gi|16759094</td><td>(NC_003198) putative IS element tra</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>GTAEQHIKEGK</td><td>51.501952</td><td>gi|695769</td><td>(X84038) transposase [Xanthobacter au</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>GTAEKGGLAIGDT</td><td>86.792704</td><td>gi|8894820</td><td>(AL360055) putative ABC transport sy</td><td>some Streptomyces</td></tr>
<tr><td>1173.628N</td><td>SAAEKDKGKQE</td><td>10.64305</td><td>gi|18550306</td><td>(XM_103535) hypothetical protein XP</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>TGAEKAPKSPSK</td><td>13.977534</td><td>gi|6009909</td><td>(AB018242) histone H2A-like protein</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>ASAEQCGRQAGG</td><td>33.741525</td><td>gi|7798662</td><td>AF135145_1 (AF135145) class I chitin</td><td>Unknown</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPNTSR</td><td>13.580314</td><td>gi|3204322</td><td>(AJ008497) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPNTSR</td><td>13.697865</td><td>gi|3355417</td><td>(AJ011213) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPNTSR</td><td>13.807781</td><td>gi|3204271</td><td>(AJ008470) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPNTSR</td><td>13.637407</td><td>gi|3204303</td><td>(AJ008487) gag protein [Human immuno</td><td>Human</td></tr>
<tr><td>1173.628N</td><td>GTAEKMPNTSR</td><td>13.552304</td><td>gi|3204338</td><td>(AJ008505) gag protein [Human immuno</td><td>Human</td></tr>
</table>

FIGURE 4 Compiled and formatted search results from sequence tag searching are input in the MultiTag software via opening a text-formatted results file.

tense and noisy spectra, it can outperform MS BLAST many cases (for a more thorough discussion on omologue identification specificity, see Sunyaev *et al.*, 003).

Both MS BLAST and MultiTag can identify proteins resent in mixtures. Usually two or three components er sample can be identified easily. The sensitivity of oth methods is determined primarily on the quality f the *de novo* sequences or sequence tags.

V. PITFALLS

1. Poor sample preparation can frequently deteriorate the quality of tandem mass spectra of peptides. he digestion of proteins with trypsin or other proeases should be carried out with chemicals of the highest degree of purity available. Plasticware (pipette tips, gloves, dishes, etc.) may acquire a static charge and attract dust, thus leading to contamination of samples with human and sheep (wool) keratin during in-gel or liquid digestion. Any polymeric detergents (Tween, Triton) should not be used for cleaning the laboratory materials.

2. When generating *de novo* sequences or sequence tags, if the software automatically extrapolates the parent mass from the precursor isotope cluster in the MS/MS spectra or the proceeding survey scan in a LC/MS/MS run, it is advisable to manually calculate this value, as software may determine the parent mass incorrectly by designating an incorrect charge state or ^{12}C monoisotopic peak of the parent ion isotope cluster, thus disabling correct *de novo* sequence prediction and sequence tag prediction.

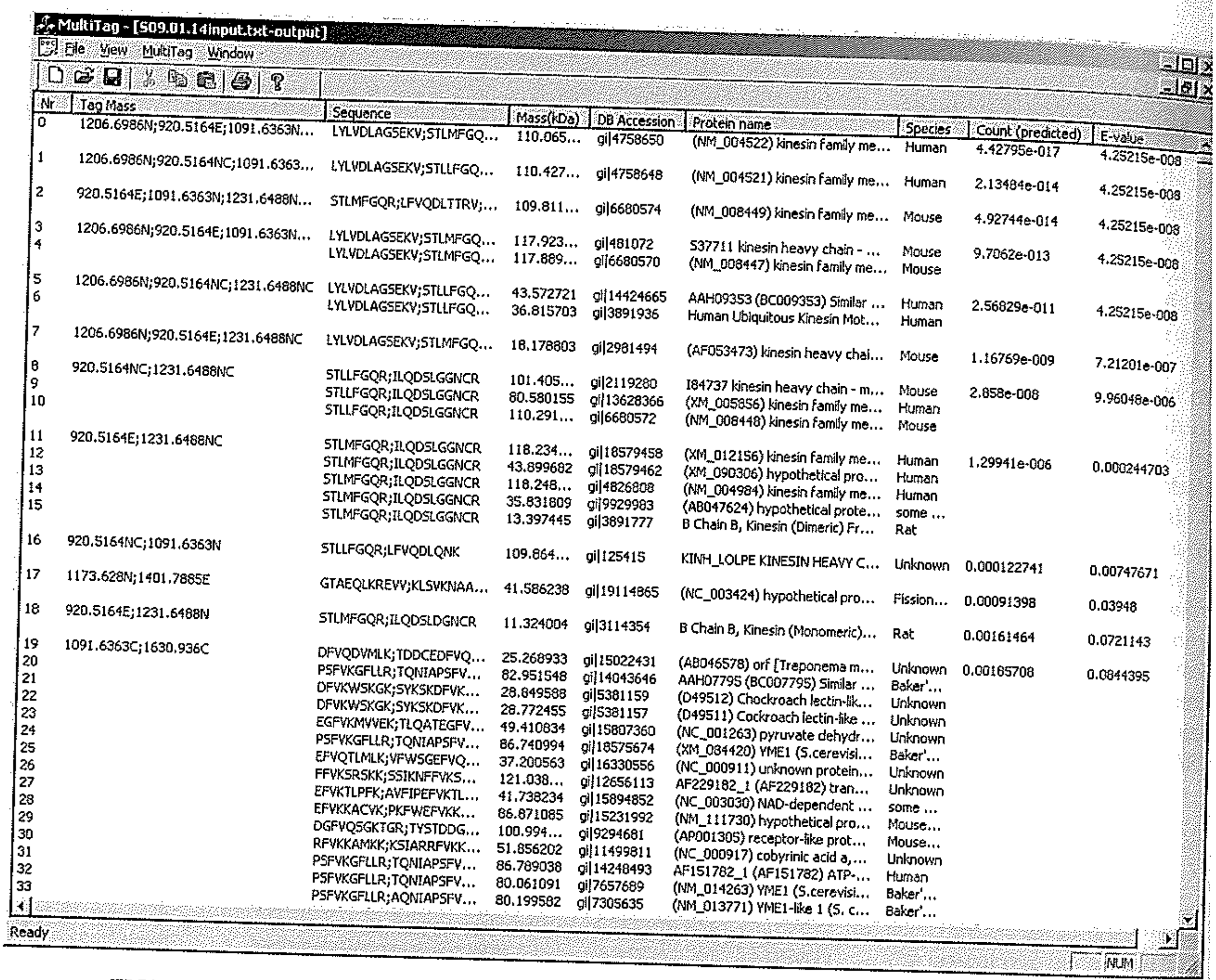


FIGURE 5 A section of MultiTag output. Margins may be adjusted to see the full list of tags matched, as well as full peptide sequences aligned, names, and so on. Results may be saved as text files to be viewed in an appropriate spreadsheet application.

3. MultiTag is laborious. Without scripted sequence tag database searching and processing of search results, manual data processing can demand extended effort; however, in cases where conventional methods fail to identify analyzed proteins, positive identifications are of a high value to cell biological studies.

4. Poor queries tend to obscure protein identification by both MS BLAST and MultiTag. It is best to submit fewer higher quality sequences than numerous lower quality sequences to MS BLAST. MS BLAST is particularly susceptible to low-complexity glycine- and proline-rich sequences generated incorrectly by *de novo* software. These low-complexity sequences tend to mask correct alignments. MultiTag functions best with sequence tags containing multiple (2–4) amino acids that have a low prevalence, such as tryptophan (W) or methionine (M), whereas common amino ac such as leucine (L) in the tag tend to be of less sigr icance and are likely to produce more false positiv Sequence tags generated from larger peptides a have more significance in a database search than th generated from smaller peptides.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Miller, W., and Lipman, D. J. (1997). Gapped BLAST and P BLAST: A new generation of protein database search prograı *Nucleic Acids Res.* 25, 3389–3402.

Gaeta, B. A. (2000). BLAST on the Web. *Biotechniques* 28, 436–440

Johnson, R. S., and Taylor, J. A. (2000). Searching sequence databa: via de novo peptide sequencing by tandem mass spectromet *Methods Mol. Biol.* 146, 41–61.

ınn, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399.

mkar, S., and Loo, J. A. (2002). Orlando FL. Application of a new algorithm for automated database searching of MS sequence data to identify proteins. Abstract 334.

rkins, 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.

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.

Sunyaev, S., Liska, A., Golod, A., Shevchenko, A., and Shevchenko, A. (2003). MultiTag: Multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Submitted for publication.