MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry

Shamil Sunyaev,*[,†,‡] Adam J. Liska,[§,⊥] Alexander Golod,[†,⊥] Anna Shevchenko,[§] and Andrej Shevchenko*[,§]

European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany, Genetics Division, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, Massachusetts 02115, and Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany

The characterization of proteomes by mass spectrometry is largely limited to organisms with sequenced genomes. To identify proteins from organisms with unsequenced genomes, database sequences from related species must be employed for sequence-similarity protein identifications. Peptide sequence tags (Mann, 1994) have been used successfully for the identification of proteins in sequence databases using partially interpreted tandem mass spectra of tryptic peptides. We have extended the ability of sequence tag searching to the identification of proteins whose sequences are yet unknown but are homologous to known database entries. The MultiTag method presented here assigns statistical significance to matches of multiple error-tolerant sequence tags to a database entry and ranks alignments by their significance. The MultiTag approach has the distinct advantage over other sequence-similarity approaches of being able to perform sequence-similarity identifications using only very short (2−4) amino acid residue stretches of peptide sequences, rather than complete peptide sequences deduced by de novo interpretation of tandem mass spectra. This feature facilitates the identification of low abundance proteins, since noisy and low-intensity tandem mass spectra can be utilized.

Developments in genomic sequencing and mass spectrometry have shifted the focus of biochemical research from the characterization of individual proteins to the large-scale analysis of the proteome, the protein complement of the genome (reviewed in ref 1). Proteins are typically resolved by one-dimensional or two-dimensional gel electrophoresis, followed by enzymatic digestion and identification by peptide mass fingerprinting (PMF) or tandem mass spectrometry (MS/MS). An alternative approach is to digest a complex protein mixture in solution and to identify proteins via two-dimensional LC−MS/MS (reviewed in refs 2 and 3). Masses of intact tryptic peptides (in PMF) and masses of fragment ions (in MS/MS) are submitted for database searching using specialized software (reviewed in ref 4). Regardless of differences in database searching algorithms and mass spectrometry platforms, the software correlates the observed masses with theoretically predicted masses derived from peptide sequences produced by in silico digestion of protein database entries and calculates the statistical significance of hits. Importantly, the software does not require a full representation of the fragment ions in the tandem mass spectrum and can positively identify the peptide even if only some of the fragment ions are matched. The significance of hits increases if more fragment ions are detected and if more than one peptide sequence originating from the same database entry is recognized. Thus, conventional database mining software is inherently biased toward exact matching of spectra to catalogued sequences and in practice is mostly applied to the identification of proteins already residing in available databases. It is, therefore, not surprising that proteomics is largely limited to organisms with sequenced genomes, despite the fact that phylogenetically related organisms share significant molecular homology and that extensive protein sequence information may be available from related species.

The proteomes of organisms with unsequenced genomes can be analyzed effectively with the aid of methods for the correlation of peptides with database sequences by sequence similarity (reviewed in ref 5). Recently, methods have been developed for protein identification by modified FASTA[6] (FASTS)[7] and BLAST[8] (MS BLAST)[9] database searches, which allow the mass spectro-

* Correspondence may be addressed to either author. E-mails: shevchenko@mpi-cbg.de; ssunyaev@rics.bwh.harvard.edu.

[⊥] Authors contributed equally to this work.

[†] EMBL.

[‡] Brigham & Women's Hospital and Harvard Medical School.

[§] Max Planck Institute of Molecular Cell Biology and Genetics.

(1) Anderson, N. L.; Matheson, A. D.; Steiner, S. *Curr. Opin. Biotechnol.* **2000**, *11*, 408−412.

(2) Mann, M.; Hendrickson, R. C.; Pandey, A. *Annu. Rev. Biochem.* **2001**, *70*, 437−473.

(3) Peng, J. M.; Gygi, S. P. *J. Mass Spectrom.* **2001**, *36*, 1083−1091.

(4) Fenyo, D. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.

(5) Liska, A.; Shevchenko, A. *Proteomics* **2003**, *3*, 19−28.

(6) Pearson, W. R. *Genomics* **1991**, *11*, 635−650.

(7) Mackey, A. J.; Haystead, T. A.; Pearson, W. R. *Mol. Cell. Proteomics* **2002**, *1*, 139−147.

(8) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(9) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem* **2001**, *73*, 1917−1926.

10.1021/ac026199a CCC: $25.00
Published on Web 00/00/0000

metric identification of a sizable proportion of proteins sharing more than 50% of the sequence identity to their closest neighbors in a database.[7,10] The MS BLAST method may utilize redundant, degenerate, and inaccurate sequences produced by automated interpretation of tandem mass spectra[9] and can be linked to high-throughput protein characterization techniques, such as LC−MS/MS, through a simple scripting interface.[11] Importantly, both MS BLAST and FASTS methods provide independent means of evaluating the statistical significance of hits, and therefore, it is not necessary to compare retrospectively the matched peptide sequences with actual tandem mass spectra to rule out false positive hits.

A major limitation to sequence-similarity identification rests with the quality of de novo interpretation of raw tandem mass spectra, rather than in database searching. Tandem mass spectra are inherently deficient; peptide sequence fragments are often underrepresented. At the same time, spectra often display ions that originate from fragmentation of side chains of amino acid residues and are not accounted for by typical scoring schemes. It is common in femtomole sequencing that low peptide content and high chemical noise allows only a few informative fragment ions to be detected in MS/MS spectra, from which software-assisted interpretation cannot produce credible peptide sequence proposals and sequence-similarity identification will likely yield a false negative result.

The peptide sequence tag approach for error-tolerant database searching developed by Mann and Wilm in 1994 helps to overcome those limitations.[12] The sequence tag utilizes a short (2−4 amino acid residues) sequence stretch, which can be easily determined from a low-energy CID spectrum acquired from a multiply charged precursor and a pair of masses that lock the determined stretch in the full length peptide sequence, namely, the combined mass of all amino acid residues between the N-terminus of the tryptic peptide and the sequenced region and the mass of all amino acid residues between the sequenced region and the tryptic peptide's C-terminus. In stringent database searches, both masses and the sequence are required to match. Currently, sequence tags are employed in protein, EST[13] (expressed sequence tag), and genomic sequence[14] database searching. However, no evaluation of the significance of matches is provided in these searches. Therefore, even if a single hit is retrieved upon database searching, the match between corresponding peptide sequence from a database entry and the tandem mass spectrum has to be verified retrospectively by manual inspection.

Sequence tags can be used for error-tolerant searching allowing for one of the regions of the sequence tag (and, consequently, the intact mass) to mismatch. The approach enables cross-species identifications in protein sequence databases.[15] However, loose matching requirements result in a dramatic loss of search specificity so that many hundreds of hits are typically retrieved, and manual inspection of all of them is tedious.

In the present paper, we have extended the capability of the sequence tag search with the implementation of a statistical evaluation for the matching of multiple partial sequence tags in the identification of proteins from organisms with unsequenced genomes. We demonstrate that the MultiTag approach enables identification of distantly related proteins by sequence-similarity searching using only very short stretches of peptide sequence derived from tandem mass spectra, and is, therefore, a vastly simplified and sensitive method of exploring the proteomes of organisms with unsequenced genomes.

MATERIALS AND METHODS

Software. MultiTag is a stand-alone application on the MS Windows platform. MultiTag code was written using C++ language with Microsoft Visual C++ and Microsoft Foundation Classes (Microsoft Inc. CA). Sorting and statistical evaluation of ∼5000 hits takes ∼1 s on a Pentium IV workstation.

Mass Spectrometry Analysis. Proteins in a partially purified extract from *Xenopus laevis* oocytes were separated on a one-dimensional polyacrylamide gel and visualized by staining with Coomassie. Protein bands were excised and in-gel digested with trypsin as previously described.[16] Extracted peptides were first analyzed by PMF on a Reflex IV (Bruker Daltonik, Bremen, Germany) matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometer, and obtained peptide mass fingerprints were submitted for database searching by Mascot (Matrix Science Ltd, U.K.) software.[17] None of the samples were positively identified. Samples were further analyzed by nanoelectrospray tandem mass spectrometry on a QSTAR Pulsar *i* quadrupole time-of-flight (QqTOF) instrument (MDS Sciex, Canada).

Interpretation of Tandem Mass Spectra and Database Searching. Sets of uninterpreted tandem mass spectra were used to search databases first with Mascot, and when no positive identifications were achieved, the spectra were interpreted manually. Sequence tags were determined by the interpretation of tandem mass spectra using BioAnalyst QS software (Applied Biosystems, CA). Database searching was performed using the PepSea program (a part of the BioAnalyst QS package) against a comprehensive nonredundant protein sequence database downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). No constraints on the protein molecular weight or species of origin were imposed. The mass tolerance was set to 0.05 Da for fragment ions and 0.1 Da for precursor ions. Hits from error-tolerant searches were pooled in a spreadsheet (Microsoft Excel) and were encoded by a peptide precursor mass and a letter code for the matched regions of the corresponding sequence tag in order to facilitate subsequent processing by the MultiTag program. The entire pool of hits was downloaded to the MultiTag program for sorting and statistical evaluation. MS/MS spectra were further analyzed by MS BLAST sequence-similarity database searches at http://dove.embl-heidelberg.de/Blast2/msblast.html against the nrdb protein database.

(10) Habermann, B.; Sunyaev, S.; Shevchenko, A. Unpublished data.
(11) Nimkar, S.; Loo, J. A. Proc. 50th ASMS Conference on Mass Spectrometry and Allied Topics, Orlando FL 2002; Abstract 334.
(12) Mann, M.; Wilm, M. *Anal. Chem.* **1994**, *66*, 4390−4399.
(13) Mann, M. *Trends Biochem. Sci.* **1996**, *21*, 494−495.
(14) Kuster, B.; Mortensen, P.; Andersen, J. S.; Mann, M. *Proteomics* **2001**, *1*, 641−650.
(15) Shevchenko, A.; Keller, P.; Scheiffele, P.; Mann, M.; Simons, K. *Electrophoresis* **1997**, *18*, 2591−2600.
(16) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.
(17) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

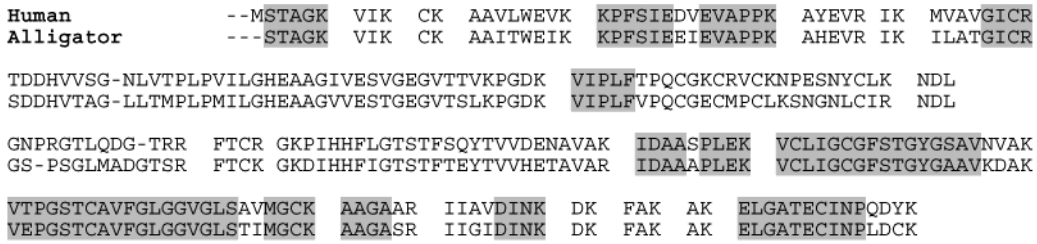

Figure 1. Partial amino acid sequences for human and alligator alcohol dehydrogenases (75% identity) are aligned. Regions alignable by error-tolerant sequence tags between the two sequences are highlighted in gray. These regions are theoretical tryptic peptides over six amino acids in length with more than three conserved amino acids from the N-terminus or more than four conserved amino acids from the C-terminus. Tryptic cleavage sites designated above are shared between both sequences. Tryptic cleavage sites not at the same point on the sequences are not designated by spaces; cleavage sites do not occur in the gray regions. Accession numbers: human, P00325; alligator, AAB28120. The sequences were aligned using the Clustal X program.

RESULTS AND DISCUSSION

MultiTag Protein Identification Strategy. MultiTag is a sequence-similarity searching approach for identifying unknown proteins via their homology to known proteins available in a sequence database. Comparison of homologous sequences of proteins from different species often shows that varying amino acid residues are distributed randomly along a polypeptide backbone, with some regions having more conservation than others. Although a single tryptic peptide may not be completely identical between the two protein sequences, their partial identity frequently occurs (Figure 1). We propose that error-tolerant searching with sequence tags can reveal regions of partial identity without determining complete peptide sequences. Although those regions are rather short to claim positive identification of a protein homologue, typically, many peptides are sequenced from a protein digest. The MultiTag software reveals proteins to which multiple fragmented peptides are matched in an error-tolerant fashion and computes the statistical significance of the hits to discriminate true hits from false positives.

The first step of the analysis is the construction of peptide sequence tags based on raw mass spectra (Figure 2). Sequence tags are typically called from the high m/z region of tandem mass spectra of tryptic peptides, which are dominated by abundant y-ions and partial interpretation of the spectrum is straightforward. Sequence tags were assembled for as many fragmented tryptic peptides as possible and were used for searching a database in a stringent fashion (matching regions 1, 2, and 3) and error-tolerant fashion: a search tolerating a mismatch of the C-terminal mass (matching regions 1 and 2), a search tolerating a mismatch of the N-terminal mass (matching regions 2 and 3), and searches tolerating one mismatch in the amino acid sequence (matching regions 1 and 3). The hits were additionally encoded by the mass of the precursor ion and by the abbreviated matching region (NC, N, C, or E, respectively) in the sequence tag. Importantly, matches of retrieved sequences to corresponding tandem mass spectra were not further inspected, and the redundant hits (matching the same peptide sequence in another database entry, or in another search) were not removed. The full list of hits was then imported to the MultiTag program. The software identified multiple hits originating from the same protein entry, eliminated redundant hits to the same peptide in the same entry and assigned the significance to all matches by computing an estimate of the probability that such a combination of tags may hit a protein entry at random.

Calculation of *E* Values. The major problem of database searching with multiple sequence tags is the need to identify hits corresponding to truly homologous proteins in the large pool of randomly matching proteins produced in multiple degenerate searches, and therefore, the evaluation of statistical significance of hits is ultimately required. The classic way to interpret the results of a database search in the statistical framework is to assign an E value to each hit resulting from the search. E values represent the expected number of better or equally good matches found in a database at random. In the case of the MultiTag search, a database search hit is a protein sequence, which matches some sequence tags in a degenerate or a nondegenerate manner. E values here give the expected number of sequences from a random database that would match the same combination of tags with the same level of degeneracy or even a more specific (less likely) combination of tags. The combination of tags can be more specific as a result of either a higher number of tags matched or a lower degeneracy of the matches. To compute E values, we first have to determine the probability that a given tag with a given type of degeneracy would match a random amino acid sequence. The probability that a given combination of tags would match a random sequence can then be computed as a product of the probabilities corresponding to individual matches. Further, we will need to find the probability that any possible combination of tags more specific (less likely) than a given combination would match a random sequence. Finally, the E value would be given by multiplication of the latter probability to the total number of database sequences. Below, we present the detailed consideration of each of those steps.

Let us consider a sequence tag, which is represented by an N-terminal mass m_N, three amino acids, a_1, a_2, and a_3, and C-terminal mass m_C. The probability that a random tryptic peptide would match this tag in a nondegenerate manner would be given as a product of the three following probabilities: First, the probability that the random tryptic peptide has an N-terminal fragment of any length, whose mass lies in the interval $(m_N - \Delta m, m_N + \Delta m)$, where Δm is mass tolerance of the instrument. Second, the probability that this fragment of random peptide has amino acids a_1, a_2, and a_3. This is simply given by the product $f(a_1)f(a_2)f(a_3)$, where $f(a_i)$ denotes frequency of amino acid a_i. And third, the probability that the mass of the random peptide fragment between these amino acids and the C-terminus would be between $m_C - \Delta m$ and $m_C + \Delta m$.

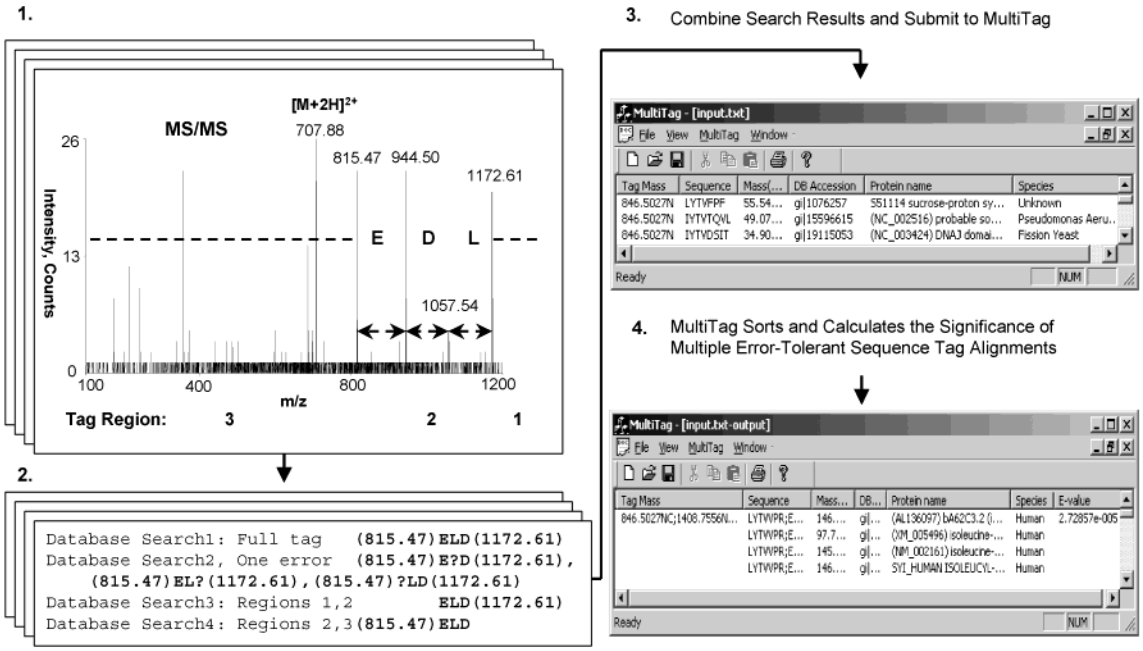

Figure 2. The MultiTag approach consists of constructing sequence tags from peptide tandem mass spectra, error-tolerant database searches, and sorting and calculation of the significance of multiple error-tolerant sequence tag alignments by the MultiTag software. Panel 1 shows a tandem mass spectrum of a low-abundance peptide with an overlaid sequence tag. Panel 2 shows one complete and three error-tolerant sequence tag database searches, which is done for each MS/MS spectrum and corresponding sequence tag. Panel 3 shows the combined list of search results (most of the 8000 entry list is not shown) from all spectra and all searches in the analysis of a single sample. Tag Mass column indicates the tag's parent mass followed by an NC for search results with complete tags, an N for searches with tag regions 1 and 2, an E for searches with tags with one amino acid error, or a C for searches with tag regions 2 and 3; Sequence column is the retrieved sequence found from the database search; Mass column indicates the protein's total mass in kDa from which the peptide originated; DB Accession, the proteins accession number; Protein name; Species. Panel 4 shows the MultiTag output. Tag Mass column lists the tag-search code for the tags aligned; Sequence lists all of the full peptide sequences error-tolerantly aligned; Mass−Species, same as Panel 3; *E* values, for the probability of the alignment of the group of sorted sequence tags. Additional column Predicted Counts reflecting the number of expected random matches of a given combination of tags is not provided here for the sake of presentation clarity.

To derive probabilities corresponding to m_N and m_C, we would regard the mass of a random tryptic peptide as being a result of a random process. We imagine that the sequence of the random tryptic peptide was constructed by a random generator, which consequently generates amino acids, one at a time, and the probability that the amino acid next in the sequence will be a_i is given by its frequency $f(a_i)$. Obviously, at each moment of time, the generator can produce a trypsin cleavage site (K or R residue) with the probability $q = f(K) + f(R)$ and thus stops the process. The mass, M, of the random tryptic peptide can be regarded as an accumulated sum of masses of randomly generated amino acids.

$$M = M_1 + M_2 + M_3 + \ldots \quad (1)$$

In probability theory, random values represented as successive sums of positive identically distributed variables as in eq 1 are called a renewal process.[18] Obviously, masses of randomly generated amino acids obey the probability distribution determined by amino acid frequencies so that the probability $p(m)$ that the random mass would be exactly m is given by combined frequency of amino acids of mass m. Then, the distribution of the mass accumulated after $n + 1$ step, that is, the probability that the peptide fragment of length $n + 1$ would have its mass smaller than t can be computed via successive convolutions.

$$F_{n+1}(t) = (1 - q)\sum_i F_n(t - m_i)p(m_i) \quad (2)$$

Summation here is carried over all values of amino acid masses. Multiplication to $(1 - q)$ is needed to take into account that the process survived the $(n + 1)$-th step, that is, the tryptic peptide, has more amino acids than $n + 1$.

The distribution of the total mass of the tryptic peptide, that is, the probability that the peptide's total mass would not exceed t is given by allowing for all possible lengths of the peptide,

$$F(t) = q\sum_{n=0}^{\infty} F_n(t) \quad (3)$$

which implies that the probability that the peptide's mass would be in the interval from $m - \Delta m$ to $m + \Delta m$ is

$$P(m, \Delta m) = F(m + \Delta m) - F(m - \Delta m) \quad (4)$$

Although not intuitively obvious, this formula holds both for the whole mass of the peptide and for any of its fragments between a fixed amino acid position and the cleavage site. Indeed, if we

(18) Feller, W. *An Introduction to Probability Theory and Its Applications*; John Wiley & Sons: New York, London, Sydney, 1966.

further consider our analogy with the renewal process, it will retain its properties regardless of the point we consider the process started (the process has no memory). Therefore, after the position of the sequence tag on the peptide sequence is fixed through matching one mass and the short sequence stretch, probability that the second mass would also match is given by eq 4.

Now consider matching of the sequence tag as a sequence of three consecutive independent events, namely, match of the first mass, match of the short sequence stretch, and the following match of the second mass. Although the consideration is obviously symmetric with regard to N- and C-termini of the peptide, without loss of generality, we would assume that the N-terminal mass is the first mass to match. The probability that the mass of any N-terminal fragment of the peptide would be in the interval ($m_N - \Delta m$, $m_N + \Delta m$) is given by

$$Q(m_N, \Delta m) = \frac{1}{q}[F(m_N + \Delta m) - F(m_N - \Delta m)]$$

or

$$Q(m_N, \Delta m) = \frac{1}{q}P(m_N, \Delta m) \quad (5)$$

We introduced a multiplier $1/q$ because, in this case, the process survives the step with this mass; i.e., all peptides with arbitrary lengths with N-terminal parts matching the mass would satisfy the condition. We note that eq 5 holds only if the mass tolerance of the instrument is lower than any of the amino acid masses. Otherwise, it corresponds to the expectation and not to the probability.

Since the probability of the nondegenerate match of the sequence tag would be a product of probabilities of the N-terminal mass match (which importantly fixes the position of the tag along the peptide), sequence stretch match, and the C-terminal mass match, it will be expressed as

$$P_{\text{nondegenerate}} = \frac{1}{q(1-q)}P(m_N, \Delta m)f(a_1)f(a_2)f(a_3)P(m_C, \Delta m) \quad (6)$$

Additional multiplier $1/(1 - q)$ simply reflects the fact that we do not consider zero-length tryptic peptides, allowed by the model if the cleavage site comes at the first step. Therefore, we work only with $1/(1 - q)$ fraction of realistic peptides.

Examples of probabilities for degenerate matches are given by

$$P_{\text{N-terminal}} = \frac{1}{q(1-q)}[1 - P(m_N, \Delta m)]f(a_1)f(a_2)f(a_3)P(m_C, \Delta m)$$

and

$$P_{\text{second residue}} = \frac{1}{q(1-q)}P(m_N, \Delta m)f(a_1)[1 - f(a_2)]f(a_3)P(m_C, \Delta m) \quad (7)$$

As the next step, we compute the probability that a random protein sequence containing K tryptic peptides would match multiple sequence tags, taking into account the tags being matched and the type of degeneracy of the match. For instance, if we had three sequence tags and the random sequence matched simultaneously sequence tag 1 with an error in the C-terminal mass, sequence tag 2 with an error in the N-terminal mass and sequence tag 3 with a mismatch at the second identified amino acid, the probability of the event would be given by eq 8.

$$P = (1 - e^{-K\cdot P_1(m_{1N})\cdot f(a_{11})\cdot f(a_{12})\cdot f(a_{13})\cdot(1-P_1(m_{1C}))})(1 - e^{-K(1-P_2(m_{2N}))\cdot f(a_{21})\cdot f(a_{22})\cdot f(a_{23})\cdot P_2(m_{2C})})(1 - e^{-K\cdot P_3(m_{3N})\cdot f(a_{31})\cdot(1-f(a_{32}))f(a_{33})\cdot P_3(m_{3C})})$$

This example shows how to compute the probability that a random amino acid protein sequence would match an arbitrary combination of sequence tags.

To calculate *E* values, we should first compute the probability that any combination of tags would match a random amino acid sequence, which is equally or more specific than the combination observed. In other words, we will need to sum up probabilities (eq 8) of all possible matches that do not exceed the probability of the actual database hit. It is definitely too demanding computationally to directly enumerate all less likely combinations of tags. However, it appears to be much easier to enumerate all combinations that are, in contrast, more likely to happen because they mostly involve matches with a very few tags. Therefore, we compute the probability that a random sequence would produce a less specific match than the actual hit (taking care of possible statistical dependence of various combinations of tags) and subtract the result from 1. The *E* value is then computed by multiplying the result to the database size.

Series of computational simulations have been carried out to validate the computation of *E* values described above (data not shown).

Software implementation of MultiTag uses precomputed distribution function $F(t)$. The software imports sequence tags in the conventional format $(m_c)a_1 \ldots a_n(m_n)$[12] and peptide mass, and computes probabilities for each tag to match a random tryptic peptide. Further, the software imports a full list of hits produced by multiple degenerate and nondegenerate sequence tag searches and identifies hits corresponding to the same protein. For each hit, MultiTag first computes the probability of the match (similarly to eq 8), then it identifies all tag combinations giving the same or higher probability, and on the basis of this information, assigns an *E* value to the hit. At the final step, MultiTag sorts all hits according to *E* values.

Specificity, Performance, and Limitations of Error-Tolerant MultiTag Searching. MultiTag aligns multiple partial and complete sequence tags to increase the coverage of a database sequence from available tandem mass spectrometry data to raise the significance and lower the *E* value of identifications. We used sequence tags from the identification of DNA polymerase (Table 1) to perform alignments with MultiTag to demonstrate factors that contribute to final *E* values (Table 2). High *E* values are given for "poor" quality tags that have short mass lengths for tag regions 1 and 3 and designate common amino acids with a high frequency

Table 1. Identification of *Xenopus* Proteins by MultiTag[a]

MultiTag protein identifications	tags submitted	mass	matching tags	MS BLAST identifications	MS BLAST alignments
isoleucyl-tRNA synthetase	(371.24)VTY(734.42)	846.5	(371.24)VTY(734.42)	no identification	none
human P41252	(637.34)VL(849.49)	935.51			
E value: $\mathbf{2.73 \times 10^{-5}}$	(587.36)VSQ(901.52)	1047.55	(587.36)VSQ		
	(488.34)LEL(843.55)	1099.56			
E value of top	(602.42)EQ(859.53)	1134.64			
false positive: 0.15	(979.62)DVS(1280.74)	1408.76	DVS(1280.74)		
	(1166.60)DLL(1507.80)	1619.81			
	(1363.59)VV(1561.73)	1918.98			
	(985.50)NT(1200.59)	2329.25			
glutamyl-prolyl-tRNA synthetase	(492.31)TY(756.42)	902.49	(492.31)TY	no identification	none
human XP_001958	(559.31)QAS(845.44)	942.03			
E value: $\mathbf{7.14 \times 10^{-7}}$	(559.33)QVS(873.49)	971.57			
	(545.27)LWT(945.48)	1058.58			
E value of top	(705.35)LE(947.47)	1192.62			
false positive: 0.52	(626.40)LLDE(1096.64)	1357.67			
	(1177.60)LLA(1474.81)	1544.87	(1177.60)LLA(1474.81)		
	(926.54)FSLTDT(1590.84)	1688.94			
	(561.34)AVEP(957.54)	1711.92	(561.34)AVEP		
DNA polymerase delta	(401.30)PVP(694.48)	750.5		DNA polymerase delta	
human S35455	(486.38)SE(702.45)	772.49		human P28340	
E value: $\mathbf{8.14 \times 10^{-7}}$	(441.21)PF(685.33)	797.42			
	(456.31)FT(704.42)	816.48	(456.31)?T(704.42)		LTFALPR
E value of top	(345.25)QEL(715.43)	827.49	(345.25)QEL(715.43)		
false positive: 0.35	(385.28)LY(661.42)	846.49	(385.28)LY(661.42)		DAYLPLR
	(421.29)EAW(807.44)	877.48			
	(583.32)LGG(810.44)	908.5	LGG(810.44)		VGGLFAFAK
	(543.30)LNL(883.51)	995.6			
	(470.29)FVL(829.51)	1071.58			
	(533.32)LPE(872.50)	1131.65	LPE(872.50)		
	(889.45)QS(1104.54)	1190.56			
Hsp70/Hsp90 organizing protein	(494.26)DSLL(922.48)	992.53		stress-induced	
Chinese hamster AAB94760	(408.23)FQLA(867.48)	995.51		phosphoprotein STI1	
E value: $\mathbf{7.82 \times 10^{-9}}$	(550.27)ELL(905.48)	1017.56	(550.27)E?L(905.48)	*Xenopus* AAM77586	
	(585.40)GVDF(1003.59)	1115.67	(585.40)G?DF(1003.59)		LFDVGLLALR
E value of top	(674.38)NGAS(1003.52)	1186.65	NGAS(1003.52)		ALSAGNLD
false positive: 0.59	(416.24)ELL(771.45)	1350.72	(416.24)ELL		
	(856.51)NLYA(1317.73)	1415.8			VAYLNPD
heat shock protein 90-beta	(385.26)FLL(758.50)	828.53	(385.26)FLL(758.50)	heat shock protein 90-beta	ALLFLPR
zebrafish O57521	(567.28)Y(730.35)	876.43	(567.28)Y(730.35)	salmon AF135117 (nucleotide)	FYDGFTK
E value: $\mathbf{2.60 \times 10^{-9}}$	(401.29)ES(617.37)	729.45	(401.29)ES(617.37)		LSELLR
	(708.37)NAV(992.52)	1234.64	(708.37)N?V(992.52)		
E value of top	(716.34)NLL(1056.55)	1241.69	NLL(1056.55)		
false positive: 0.2					
					LTPDQPVV

[a] For each sample, sequence tags were constructed from multiple MS/MS spectra from the analysis of a single in-gel digest (tags submitted) and error-tolerantly searched against a protein database (resulting list of entries not shown). The mass column contains the mass of the intact peptide for each sequence tag. Results were sorted by MultiTag. Groups of matching partial sequence tags resulted (matching tags). *E* values for the group of partial sequence tags were calculated by the MultiTag software (first column in bold). The final MultiTag report gave a list of database entries with diminishing *E* values (data not shown). *E* values are cited (column 1, top false positive) for the first database entry in the list to not correspond by annotated function (i.e., HSP 90) to the most significant hit. Protein identifications made by MS BLAST are found in the column MS BLAST identifications, and peptide sequences aligned are in MS BLAST alignments.

in proteins, that is, leucine. Lower *E* values are given for uncommon amino acids such as tryptophan, for tags with more amino acids in the sequence stretch and for tags covering a longer amino acid sequence by the "lock" masses in regions 1 and 3. The probability that a combination of partial sequence tags will match a single database entry is lower than if an individual tag is matched, with an increasing significance as more partial sequence tags are aligned. Two partial sequence tags might not be significant enough for a confident identification at any practical mass accuracy, depending on the character of the tags. However, the alignment of three or more partial sequence tags lowers *E* values to the range of 1×10^{-6} to 1×10^{-9}, even at 0.1 Da mass accuracy, enabling confident protein identification. Sequence tags assembled with narrower mass tolerance increase the specificity of database searching and lower the *E* values of hits. As in the case of conventional database sequence-similarity searches, specificity of the MultiTag search decreases with the growing size of the database.

In the case of analysis of protein mixtures, many sequence tags from a MultiTag query are not expected to match any single protein sequence in a database, even if it is a true homologue of one of the proteins present in the mixture. Therefore, we estimated the robustness of MultiTag identification with respect to the number of irrelevant tags in the query. Simple simulation on a database with random sequences (data not shown) suggested that the dependence of *E* values on the full number of tags in the query

Table 2. *E* Values Are Dependent on Amino Acids in the Tag, Number of Tags, Mass Accuracy, and Database Size[a]

	mass[c]	sequence tags in the identification of DNA polymerase	*E* values			PredCount	*E* values[b]	
			1.0 Da[d]	0.5 Da[d]	0.1 Da[d]	0.1 Da[d]	1 600 000 DB entries[e]	200 000 DB entries[e]
1	816.48	(456.31)?T(704.42)	6.06×10^{3}	2.73×10^{3}	2.56×10^{3}	1.76×10^{2}	5.11×10^{3}	6.39×10^{2}
2	827.49	(345.25)QEL(715.43)	1.96×10^{2}	8.83×10^{1}	8.34×10^{1}	1.23	1.67×10^{2}	2.09×10^{1}
3	846.49	(385.28)LY(661.42)	2.61×10^{2}	1.30×10^{2}	1.29×10^{2}	2.66	2.58×10^{2}	3.22×10^{1}
4	908.50	LGG(810.44)	8.36×10^{3}	6.53×10^{3}	6.92×10^{3}	9.81×10^{2}	1.38×10^{4}	1.73×10^{3}
5	1131.65	LPE(872.50)	1.73×10^{3}	1.01×10^{3}	9.52×10^{2}	7.06×10^{1}	1.90×10^{3}	2.38×10^{2}
6		LGG(810.44) + LPE(872.50)	5.08×10^{1}	2.71×10^{1}	2.51×10^{1}	9.23×10^{-2}	5.16×10^{1}	6.45
7		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42)	1.24×10^{-1}	3.09×10^{-2}	2.86×10^{-2}	2.17×10^{-5}	5.72×10^{-2}	7.15×10^{-3}
8		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42)	1.97×10^{-5}	3.23×10^{-6}	3.01×10^{-6}	7.68×10^{-11}	6.02×10^{-6}	7.53×10^{-7}
9		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42) + (345.25)QEL(715.43)	1.49×10^{-6}	8.60×10^{-7}	**8.14×10^{-7}**	1.26×10^{-16}	1.63×10^{-6}	2.03×10^{-7}
10		LGG(810.44) + LPE(872.50) + (456.31)?T(704.42) + (385.28)LY(661.42) + (345.25)QEL(715.43)	3.64×10^{-9}	4.17×10^{-9}	3.55×10^{-9}	1.31×10^{-16}	7.11×10^{-9}	8.88×10^{-10}

[a] The *E* value in bold is shown in Table 1. In the calculation of *E* values for row 1−9, all tags submitted were included from Table 1. In row 10, only the tags that matched the database entry were included in the list of tags submitted for MultiTag calculations. Da corresponds to mass accuracy (in Da) used in database searches and input into MultiTag for calculations of *E* values. [b] 800 000 and 200 000 database entries correspond approximately to the NCBI nonredundant (nrdb) and SwissProt protein databases, respectively. [c] Mass of the intact peptide corresponding to the sequence tag in column 3. [d] 800 000 database entries. [e] Mass accuracy of 0.1 Da.

is nonlinear, and its strength is dependent on the number of full and partial tags matched to the protein sequence.

We, therefore, considered a highly simplified scenario with a query of *N* tags that are having identical specificity (only one type of match is allowed) and *n* tags out of total *N* tags match the protein in a database. Let us estimate an increment of the *E* value if the query size would be increased up to *MN* tags without any new matching tags added. A reasonable low limit estimate could be presented by a ratio of binomial coefficients C^n_{MN}/C^n_N For example, with 5 tags matched, expanding the query from 5 to 10 tags would increase the *E* value by 252 times, which is in good agreement with example calculations in Table 2. However, increasing the number of tags under the same scenario from 20 to 40 would increase the *E* value by only 42 times.

For large queries, $C^n_{MN}/C^n_N \sim M^n$. Let us consider the data from Table 1 as a practical example to evaluate the performance of MultiTag in "shotgun-type" identification of unknown proteins from unseparated mixtures Although these proteins were identified from Coomassie stained individual bands, let us consider the case in which the same analysis would have been performed from mixtures of, say, 10 unknown proteins, each producing 10 peptides upon their tryptic digestion, so that altogether 100 tryptic peptides would have been sequenced in each case, with sequence tags determined and submitted to MultiTag search. *E* values of the hits in Table 1 are in the range of 1×10^{-5} to 1×10^{-9} for about 10 tags submitted, with 3−5 tags matched. If the same tags would be matched to the same proteins in the same way, but from the queries comprising yet another 90 nonmatching tags, we would observe their *E* values increased by 3 orders of magnitude in cases where three tags matched and by 5 orders of magnitude in the case of 5 tags matched. Despite remarkable drop in statistical confidence, these hits would still remain significant. However, it is also apparent that weaker hits would be pushed into a "twilight zone" of marginal significance.

Although we did not observe this problem in practice, we propose the following strategy for evaluating borderline hits. For every hit, in additional to *E* values, MultiTag reports the expected number of random matches of the same combination of tags in the same way (termed predicted counts, PredCount) (Figure 2). PredCount values reflect the specificity of matches, and their ranking order is the same as the ranking order based on *E* values. However, PredCount does not reflect the expected number of false-positives when the entire query is searched against a database. Contrary to *E* values, PredCount values very weakly depend on the number of tags in the query. Therefore, in cases when most of the tags in the query were not matched, hits with low PredCount values (lower than 1×10^{-4}) deserve further analysis by manual inspection of the original spectra (as described below), even if their *E* values have only marginal significance.

An intrinsic problem to all statistical approaches to homology searches relying on average amino acid frequencies is posed by low complexity regions and other proteins or protein regions with amino acid frequencies, which strongly deviate from the database average.[19] If a MultiTag identification results in a peptide from a low complexity region or a peptide of obviously special amino acid composition, these identifications have to be interpreted with caution, since the underlying statistical model does not account for bias in amino acid composition.

An advantage of MultiTag over MS BLAST, besides its ability to represent noisy and low-intensity spectra, is that peptide sequences retrieved by sequence tag searches can be overlaid on fragment ion spectra, allowing one to determine whether the retrieved sequence is the correct sequence. This is less direct with MS BLAST or FASTS. Even though relatively weak matches can be evaluated in this way, the MultiTag approach is suited for

(19) Baudouin-Cornu, P.; Surdin-Kerjan, Y.; Marliere, P.; Thomas, D. *Science* **2001**, *293*, 297−300.

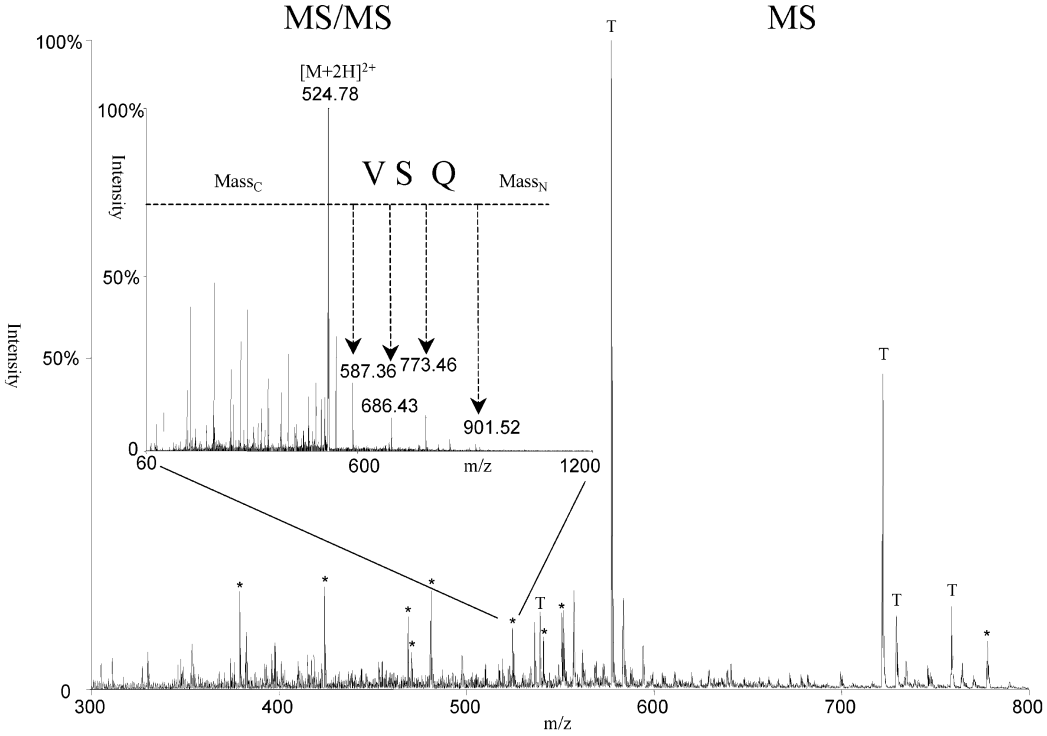

Figure 3. *Xenopus* proteins were in-gel digested and analyzed by nanoelectrospray tandem mass spectrometry. MS spectrum peaks labeled with an asterisk were fragmented, and peptide sequence tags were assembled from MS/MS spectra (inset). Abundant y-ions above the multiply charged precursor in MS/MS spectra allow direct determination of the partial amino acid sequence of a peptide and assembling a sequence tag. The resulting sequence tag from the MS/MS spectrum shown is (587.36)VSQ(901.52), peptide mass 1047.55 Peaks in the MS spectrum labeled with T are autolysis products of trypsin. All of the determined sequence tags from the analysis of this sample are presented in Table 1. The protein was identified as isoleucyl-tRNA synthetase.

a high-throughput setting, since manual evaluation is required only for easily recognizable borderline hits.

MultiTag, as well as any sequence-similarity searching method, is prone to errors if the analyzed protein contains low-complexity sequence regions, such as collagen, glycine-rich cell wall proteins, and silk proteins. Because MultiTag recognizes only a few central amino acids accurately, it would be possible that multiple sequence tags to different regions of the same proteins could not be distinguished, thus diminishing the overall score.

Identification of Proteins from *Xenopus laevis* by MultiTag Searching. We applied the MultiTag approach to the identification of proteins isolated from the African clawed frog *X. laevis* in ongoing collaboration projects. *X. laevis* is an important model organism for the study of the cell cycle (reviewed in ref 20), DNA replication,[21] and developmental biology,[22] among other biological processes. Currently, sequences of fewer than 7000 *Xenopus* proteins are present in a publicly available database, despite a public initiative in EST sequencing (fewer than 221 000 largely unannotated ESTs are available, August 16, 2002, both figures from http://www.ncbi.nlm.nih.gov/). Taken together, current database resources do not provide an adequate coverage of *Xenopus*' large 3070-megabase pseudotetraploid genome.[23] In-gel digests of *Xenopus* proteins were analyzed by PMF and by nanoelectrospray tandem mass spectrometry. Sequence-similarity searching methods were applied for protein identification, because Mascot database searching with peptide mass fingerprints and with lists of fragment masses derived from uninterpreted tandem mass spectra were unable to identify proteins by stringent matching. Two methods of sequence-similarity searching were applied in parallel to the same set of MS/MS data. Peptide sequence proposals obtained by automated de novo interpretation of tandem mass spectra were submitted to MS BLAST searching.[24] In parallel, peptide sequence tags were assembled via partial manual interpretation of spectra (Figure 3), followed by error-tolerant database searching and sorting and evaluating the results by MultiTag, as described above (Table 1). From five attempted unknown proteins, MS BLAST identified three; however, all five were identified by MultiTag. Importantly, in three cases, both MultiTag and MS BLAST identified homologous sequences from the same organism or from different species, providing an independent validation of the MultiTag approach.

These data demonstrate that MultiTag can outperform the more generic sequence-similarity searching tool, MS BLAST, when de novo sequence prediction software is unable to produce meaningful peptide sequences from noisy or low intensity spectra. On the other hand, MultiTag successfully identified the proteins, because sequence tags are easily assembled from tandem mass spectra in which complete amino acid sequence prediction is

(20) Nurse, P. *Cell* **2000**, *100*, 71−78.
(21) Herrick, J.; Stanislawski, P.; Hyrien, O.; Bensimon, A. *J. Mol. Biol.* **2000**, *300*, 1133−1142.
(22) De Robertis, E. M.; Larrain, J.; Oelgeschlager, M.; Wessely, O. *Nat. Rev. Genet.* **2000**, *1*, 171−181.
(23) Graf, J. D.; Kobel, H. R. *Methods Cell. Biol.* **1991**, *36*, 19−34.
(24) Shevchenko, A.; Sunyaev, S.; Liska, A.; Bork, P.; Shevchenko, A. *Methods Mol. Biol.* **2002**, *211*, 221−234.

impossible (Figure 2). The results suggest that three or more error-tolerantly matching sequence tags may unequivocally identify a homologous protein (Table 2), despite none of the sequenced peptides exactly matching the corresponding sequence from a database entry and sequence stretches of less than four amino acid residues being determined. Both MultiTag and MS BLAST were able to identify the proteins not identified by Mascot, because they could tolerate amino acid substitutions, resulting in an offset of the peptide's total mass and many of the fragment masses.

Homologue Identification Specificity of MultiTag Searching. By its algorithm, MultiTag is a less generic sequence-similarity searching tool, compared to MS BLAST and FASTS, since it requires identical (although short) stretches of peptide sequence for protein identification. We roughly estimated the scope of MultiTag identification from the bottom, assuming the most unfavorable model when identical amino acids between proteins are distributed uniformly along the sequence. According to our experience and Table 2, three partial matches usually produce a statistically significant hit. We have estimated the chance to obtain three partial matches and its dependence on the overall identity of the complete query sequence and the database sequence. A very simple calculation assumes that the probability that a single amino acid would match between the query and the database sequence is equal to the overall sequence identity and is independent of the sequence region and amino acid type. Assuming further a query of 10 identical tags, we estimated that the MultiTag method is able to identify almost all homologues at the level of 80% sequence identity, 75% of homologues at the level of 75% sequence identity, but only about 45% of homologues at the level of 70% sequence identity. Obviously, MultiTag cannot achieve the specificity of the methods using the knowledge of longer sequence parts. According to simulation results, sequence-based methods, such as MS BLAST and FASTS, are able to detect ~50% of homologous sequences at the sequence identity level of ~50%.[7] According to our lower limit estimate, MultiTag would require 71% sequence identity (in reality, less) to reach the same efficiency of identifications. However, simulations with MS BLAST and FASTS were performed assuming all sequence predictions are correct, which is rarely the case. Therefore, the advantage of using MultiTag is the ability to identify sequence similarities at a reasonable level with high robustness with respect to the quality of the raw data and independently of the quality of automated de novo sequence prediction techniques.

(25) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem* **2002**, *74*, 5383−5392.

CONCLUSION AND PERSPECTIVES

The MultiTag approach addresses an issue of growing prominence among the proteomics community: universal statistical evaluation of protein identifications.[25] It is a goal of the proteomics community to set a threshold for protein identifications in high-throughput settings so that a protein confidently identified in one laboratory will be confidently identified by a similar method in another institution. Consequently, the statistics of MultiTag takes a step in this direction and determines the significance of sequence tag alignments in a manner that can be adopted as a universal standard evaluation of sequence tag identifications without the need for retrospective inspection. Because the MultiTag approach could be applied to the mining of EST and genomic databases, the statistics will require alteration as a result of the size and nature of these searches. The independent statistics of MultiTag lends the method to a wide application in high-throughput proteomics.

With further software developments, it will be possible to completely automate the MultiTag method for high-throughput proteomics of organisms with unsequenced genomes or the analysis of highly modified proteins from organisms with sequenced genomes. Currently, the ability to call sequence tags automatically is available, and a scripted interface can be written to create lists of sequence tags for spectra acquired from a complete LC−MS/MS run. A corresponding scripted interface for database searching can be written that can produce a complete list of encoded retrieved database entries for submission to MultiTag for sorting and significance calculation.

Following developments in automation, MultiTag will be a good complementary method to de novo sequence prediction-based methods, such as MS BLAST and FASTS, for sequence-similarity protein identification in high-throughput settings, thus expanding the repertoire of spectra interpretation and database mining tools in the hands of mass spectrometrists. As sequence-similarity methods develop, the proteomes of organisms with unsequenced genomes will become more amenable for characterization, contributing to the development of medicine, agriculture, and the biological sciences in general.

ACKNOWLEDGMENT

We thank our collaborators, Drs. Andrei Popov and Eric Karsenti (EMBL, Heidelberg), for *X. laevis* protein gels. We are greateful to Dr. Bianca Habermann (MPI of Molecular Cell Biology and Genetics) and members of the Shevchenko's group for useful discussions and overall support.

Received for review October 4, 2002. Accepted January 3, 2003.

AC026199A