 DOI 10.1002/pmic.200401262 *Proteomics* 2005, *5,* 4118–4122

SHORT COMMUNICATION

Error-tolerant EST database searches by tandem mass spectrometry and multiTag software

Adam J. Liska[1], Shamil Sunyaev[2], Ignat N. Shilov[3], Dan A. Schaeffer[3] and Andrej Shevchenko[1]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
[2] Genetics Division, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, USA
[3] Applied Biosystems, Foster City, USA

The MultiTag method (Sunyaev et al., *Anal. Chem.* 2003 *15*, 1307–1315) employs multiple error-tolerant searches with peptide sequence tags (Mann and Wilm, *Anal. Chem.* 1994, *66*, 4390–4399) for the identification of proteins from organisms with unsequenced genomes. Here we demonstrate that the error-tolerant capabilities of MultiTag increased the number of peptide alignments and improved the confidence of identifications in an EST database. The MultiTag outperformed conventional database searching software that only utilizes stringent matching of tandem mass spectra to nucleotide sequences of ESTs.

Received: November 19, 2004
Revised: January 24, 2005
Accepted: January 24, 2005

Keywords:
Database searching / Error-tolerant searches / EST database / Mass spectrometry / Protein identification

Peptide sequence tags were introduced by Mann and Wilm in 1994 [1] for the error-tolerant identification of proteins by tandem mass spectra. A peptide sequence tag consists of an interpreted short amino acid sequence flanked on both sides by mass values that "lock" the sequence stretch within the length of the peptide (Fig. 1). Full or partial tags can be used for database searching. Typically, all three parts of a tag, the *C*- and *N*-terminal flanking masses and the sequence stretch, are required to match the database entry. For error-tolerant searches one of the two flanking mass regions of the tag is allowed to mismatch or an amino acid substitution in the sequence stretch is tolerated, both loosening the query specificity. Since the sequence tag is only a partial representation of a spectrum, multiple degenerate database sequences are recognized in most searches. To validate the hits, the retrieved sequences are usually manually inspected and masses of fragment ions beyond the sequence stretch used in a tag are compared to masses of fragments expected from the full sequence of a peptide candidate [1–3].

However, when peptide sequence tags are searched error-tolerantly, an overwhelming number of hits are usually produced and manual inspection of them becomes impractical. The MultiTag method [4] overcomes this problem by correlating the combined results from multiple error-tolerant searches to determine the most probable protein identification(s). Confidence of MultiTag hits is estimated by their E-values, which represent the expected number of false positives hit by a given combination of completely or partially matching sequence tags. The employed statistical model implies that the first false positive should be detected at an E-value of approximately 1, the second ranking false positive hit should have an E-value of 2, and the third false positive an E-value of 3, *etc.* Because of the imperfections of the statistical model, E-values less than 0.1 generally indicate true matches, with more significant hits having lower E-values. Along with E-values, MultiTag computes the expected number of random hits, which would have the same number of matched tags with the same degeneracy, as a true hit. This value is termed PredCount (predicted counts). Unlike E-values, PredCount values do not reflect the expected number

Correspondence: Dr. Andrej Shevchenko, Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, D-01307 Dresden, Germany
E-mail: shevchenko@mpi-cbg.de
Fax: +49-351-210-2000

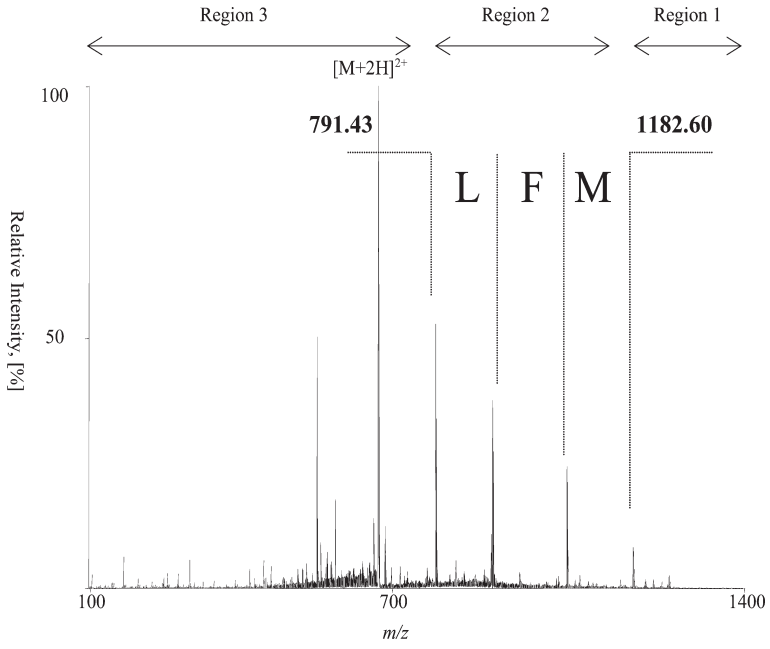

Figure 1. Assembling the peptide sequence tag: a peptide tandem mass spectrum presented in the figure was acquired from a doubly charged precursor *m/z* 676.82. A peptide sequence tag – (791.41)LFM(1182.60); precursor mass 1351.63 – was assembled from the masses of adjacent y-ions in the *m/z* region higher than *m/z* of the precursor. The peptide sequence tag effectively splits the tandem mass spectrum in three regions, which are termed, according to the convention adopted by PepSea database search program, as "Region 1" for *N*-terminal part of the spectrum (above *m/z* 1182.60), "Region 3" for *C*-terminal part of the spectrum (below *m/z* 791.43) and "Region 2" for the deduced sequence. In stringent searches, all three regions (two masses and the deduced sequence) are required to match. In error-tolerant searches, one out of the three regions is allowed to mismatch. In this analysis of a protein digest, 25 peptides were fragmented including the one shown above, 10 sequence tags were constructed and submitted for database searching and MultiTag analysis, and three complete sequence tags matched the EST sequence acc. no. 12473885, partially representing the sequence of lysyl-tRNA synthetase.

of false positive hits in the search with all submitted tags and only very weakly depend on the total number of tags in a query. Low PredCount value (typically, less than E-04) serve as a useful statistical indicator that this might be a true hit, whose E-values is high because, from many submitted tags, only a few were aligned to a database entry. For example, this often occurs in sequencing of mixtures of unknown or poorly conserved proteins.

MultiTag sorts the hits according to the E-value computed for every set of completely or partially aligned tags. Contrary to common spectra inspection routines, MultiTag does not validate hits by considering matching of other fragment ions beyond the sequence tag. Therefore, poor quality of tandem mass spectra and/or misrepresentation of fragments in certain *m/z* regions do not affect the significance of MultiTag hits and error-tolerant identification of low abundant proteins is much improved.

MultiTag was two-fold more sensitive than the conventional database searching algorithm (MASCOT) in the identification of microtubule-associated proteins in *Xenopus laevis* [5] in both cross-species protein identifications (when a database entry was used from a related organism, *i.e.*, mouse), and in direct retrieval of *Xenopus* database entries. Since a sizable number of *Xenopus* ESTs are available, we also applied MultiTag for the identification of corresponding EST database entries, however, with very limited success [5]. The MultiTag did not provide a valid estimate of the significance of error-tolerant alignments between peptide sequence tags and EST sequences and, therefore, the retrieved hits were only possible to evaluate by manual inspection of top non-error tolerant alignments.

ESTs represent nucleotide sequences, whereas amino acid polymers are analyzed by mass spectrometry and, therefore, alignments in 6× reading frames of the sequences should be considered. Also, EST sequences are generated by single-pass sequencing of cDNA clones (produced from mRNAs), which would likely result in multiple errors and frameshifts so that peptides originating from the same protein might match the sequence of the same EST clone in different frames. Sequences of ESTs are relatively short translating to less than ~150 amino acid residues and, typically, cover a relatively small piece of the full length protein sequence and, therefore, many sequenced peptides are left unaligned. Thus, even in searching sequences from the organism of origin (not cross-species), an error-tolerant method like MultiTag would be expected to be more sensitive than a method that demands exact matching, because more partial peptide sequences could be aligned to produce a higher coverage and more confident hits [6, 7].

Here we demonstrate that the MultiTag statistical model and scoring algorithm, originally developed for protein database searches, are equally applicable for error-tolerant searches in an EST database. Only minor adjustment of input settings, such as the average length of a database entry and the expected number of tryptic peptides *per* database entry, is required to produce statistically valid estimates of the EST identification confidence. Since there is often an abundance of EST sequence data *in silico*, the utilization of these sequences, either as independent resources or by applying alternate database searching strategies simultaneously [8], could facilitate protein identification.

Sequence tags were produced by manual interpretation of tandem mass spectra. A dedicated script supported the automated batch searches with a list of peptide sequence tags by PepSea engine (BioAnalyst QS software from Applied Biosystems, CA, USA) under the following settings: mass tolerance of 0.1 Da and 0.05 Da for precursor and fragment

ions, respectively; fixed modification: carboxyamidomethyl cysteine. An EST database was searched in a stringent fashion (matching regions 1, 2 and 3, see Fig. 1) 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 in the sequence tag and compiled in a text file for submission to MultiTag. MultiTag statistics relies upon an expected number of peptides originating from an "average" database entry. We previously assumed that an "average" protein, with the mass of 60 kDa, would comprise 490 amino acids, and hence its tryptic digestion would generate 41 non-overlapping peptides, each consisting of 12 amino acid residues [4]. The number of peptides expected *per* database entry has been made adjustable to account for differences in length between "average" protein and EST database entries. We assumed that the average length of an EST entry codes for 166 amino acids and we assumed that its *in silico* digestion would yield 14 peptides, also of 12 amino acids each. Since six frame translation of EST sequences should be considered, we also assumed that the database comprise 1 396 530 database entries (six frames × 232 755 *Xenopus laevis* EST entries). Importantly, no other pre-computed values used by the MultiTag scoring scheme were altered. The MultiTag software and the script supporting batch-mode searches by PepSea are available for testing upon request.

Xenopus laevis proteins, isolated by Dr. Andrei Popov from EMBL, Heidelberg, were resolved by one-dimensional SDS-PAGE, visualized by Coomassie Blue R-250 staining, digested with trypsin and recovered tryptic peptides fragmented by nanoES MS/MS. Tandem mass spectra were searched by stringent (MASCOT) and error-tolerant (MS BLAST [9] and MultiTag) methods against a protein database [5]. For better consistency with previously published protein searches, we excluded from the dataset proteins whose identification was borderline and required subsequent validation by manual inspection of tandem mass spectra. This subset of tandem mass spectra was searched against an EST database using MASCOT and the modified MultiTag software. The MASCOT identifications were made using the *Peptide Summary Report* for enhanced sensitivity and confidence of hits was estimated by MOWSE peptide scores [10]. Importantly, in this line of experiments both methods – MASCOT and MultiTag searched the same database EST others (November 27, 2002). The *Xenopus laevis* subset was used for MASCOT searches, and because MultiTag does not have species selection option, all its cross-species hits were ignored. EST hits were assigned to corresponding protein sequences by *blastx* searches at the NCBI server and were in agreement with the previously reported identifications [5]. Although the same set of MS/MS spectra was used for MultiTag and MASCOT searches, it was not always possible to interpret each spectrum and produce a reliable peptide sequence tag. On average, nine sequence tags *per* protein digest were produced and used in MultiTag searches, although all acquired MS/MS spectra were submitted to MASCOT.

From the model data set, MASCOT was able to recognize 49 peptides, making 20 identifications (Table 1). From the same dataset, MultiTag was able to recognize 87 peptides and produced 31 identifications, which included all of the MASCOT hits. In Table 1 (column "Tags matched") we listed separately the number of complete and partial tags matched to corresponding EST sequences. Out of the total of 31 identifications, in 27 cases MultiTag aligned more exactly matching peptides than MASCOT, and in nine cases additional peptides were aligned error-tolerantly. Therefore, we concluded that two factors might have contributed to the relatively better performance of the MultiTag. First, sequence tags deduced from poor quality spectra could still produce statistically reliable hits, whereas if not supported by many other matching fragment masses, these alignments would be of low confidence according to MASCOT statistics. Second, error-tolerantly aligned peptides increased the overall confidence of MultiTag identifications.

We further investigated if subsequent manual inspection could "rescue" some borderline hits, which, yet being formally non-confident, topped the output of database searches by MASCOT and MultiTag methods. Where MultiTag was able to make a significant alignment and MASCOT produced no significant matches, the top five MASCOT hits were inspected to see if the same protein had been non-confidently detected. Since it is possible that MultiTag and MASCOT might recognize different ESTs corresponding to the same cDNA sequence, these top hits were inspected to find out if alternate ESTs matched the same protein sequence. Where MultiTag was unable to make a significant alignment, the top five non-significant hits were manually inspected by comparing the masses of observed fragment ions with the fragment masses expected for the retrieved sequence, taking into consideration abundant a-, b- and y-series ions and immonium ions (see [11] for the nomenclature). In this manner, MultiTag was able to detect six additional matches; three out of these six were not in the MASCOT top five hits (note that these identifications were not included in Table 1 since they were statistically non-confident). This suggests that the MultiTag method can also retrieve true matches that are, however, not statistically significant, but could be "rescued" *via* manual inspection of a limited number of top non-confident alignments.

Thus, MultiTag proved to be a sensitive method for EST database searching, both because more identifications were made than by the conventional software and more peptides were identified in total, resulting in a higher coverage of proteins. The MultiTag balances the specificity and sensitivity of sequence tag representation of mass spectra with the sequence tags' inherent degeneracy in database searching. In cases where EST sequences could be assembled as full-length cDNA clones, we would expect even higher coverage because

Table 1. EST database searching: MASCOT *vs.* MultiTag

Mass, kDa	MASCOT hit	Peptides matched	MultiTag hit	Tags in query	Tags matched	Pred-Count	E value
175	Glutamyl-propyl-tRNA synthetase, 12748494	1	Glutamyl-propyl-tRNA synthetase, 14989013	11	2	2.70E-13	2.10E-04
175	Glutamyl-propyl-tRNA synthetase, 14989013	1	Glutamyl-propyl-tRNA synthetase, 17398490	9	2 & 1	1.10E-12	1.10E-06
165	†		Glutamyl-propyl-tRNA synthetase, 24091165	9	2	1.70E-08	5.70E-06
165	†		Glutamyl-propyl-tRNA synthetase, 12746970	4	1	5.07E-04	3.81E-03
160	Hyaluronan mediated receptor, 13252946	1	Hyaluronan mediated receptor, 13252946	9	2	5.82E-09	1.42E-04
155	†		Isoleucyl-tRNA synthetase, 24090398	9	2 & 1	2.31E-13	7.39E-08
150	†		Leucyl-tRNA synthetase, 21870435	12	2	3.36E-06	1.57E-03
150	†		Leucyl-tRNA synthetase, 21870435	6	2	3.81E-06	3.36E-04
122	Kinesin heavy chain, 17418741	3	Kinesin heavy chain, 17418741	9	2	1.42E-08	5.37E-06
122	Kinesin heavy chain, 12480559	2	Kinesin heavy chain, 12480559	7	2 & 1	6.79E-12	1.94E-08
118	Kinesin heavy chain, 17418741	2	Kinesin heavy chain, 17418741	10	3	1.27E-16	3.13E-08
	Kinesin heavy chain, 12480559	2	Kinesin heavy chain, 12480559	10	3	8.96E-16	3.13E-08
100	*Elongation factor-2, 11787464*	2	Elongation factor-2, 21875348	6	2	3.73E-06	1.27E-02
90	Heat shock protein, 10065828	2	Heat shock protein 90-beta, 21873865	13	3 & 1	1.12E-17	8.96E-07
	Glutaminyl-tRNA synthetase, 7699102	2	Glutaminyl-tRNA synthetase, 7393733	13	3	1.57E-14	8.96E-07
85	†		Cytoplasmic dynein intermediate chain, 24082627	9	3	1.12E-17	4.48E-06
70	Heat shock cognate-70, 21384290	5	Heat shock cognate, 24087000	7	2 & 1	5.00E-15	3.43E-08
68	Lysyl-tRNA synthetase, 17580417	2	Lysyl-tRNA synthetase, 24097853	4	2	5.00E-06	3.21E-04
68	Lysyl-tRNA synthetase, 12473885	3	Lysyl-tRNA synthetase, 12473885	9	3	4.03E-14	1.12E-07
68	HSP70/HSP90 organizing protein, 17395146	2	HSP70/HSP90 organizing protein, 21874237	7	2 & 1	3.73E-15	6.19E-09
52	Alpha-tublin, 12471404	2	Alpha-tublin, 21863612	16	2	7.46E-09	3.88E-03
	Formiminotransferase cyclodeaminase, 17413939	1	Formiminotransferase cyclodeaminase, 12471624	16	2	5.67E-10	3.88E-03
50	Alpha-tublin, 12471404	4	Alpha-tublin, 24097682	18	3 & 1	9.70E-20	2.01E-06
	Beta-tublin, 17425087	2	Beta-tublin, 24093819	18	3	3.66E-15	2.01E-06
50	Elongation factor-1 gamma, 17414578	6	Elongation Factor-1 gamma, 17527452	18	5	3.58E-25	3.58E-06
	Elongation factor-1 alpha, 10063988	4	Elongation factor-1 alpha, 21071694	18	3	3.28E-12	3.58E-06
36	Elongation factor-1 delta, 17397886	2	Elongation factor-1 delta, 24082682	5	4	2.24E-23	1.27E-09
34	*60S Ribosomal Protein L5B, 14181865*	1	60S Ribosomal Protein L5B, 14181865	5	1 & 1	1.12E-09	3.36E-08

Table 1. Continued

Mass, kDa	MASCOT hit	Peptides matched	MultiTag hit	Tags in query	Tags matched	Pred-Count	E value
30	40S Ribosomal protein, 14185581	1	40S Ribosomal protein S3, 24085095	6	2	2.54E-08	4.33E-06
28	Elongation factor 1-beta, 17398022	2	Elongation factor 1-beta, 24091513	7	3 & 2	5.67E-31	2.24E-08
28	†		Elongation Factor 1-beta, 21088085	5	4	6.57E-19	5.22E-09
Total peptide hits		**49**			87		
Identifications		**20**			31		

Peptides = no. of peptides matched to any single database entry; Tags = no. of complete and partial tags matching any single EST sequence, X & Y (complete tags & partial tags, respectively); Tags in Query = no. of tags submitted in query; † not in the top 5 hits; MASCOT hits below threshold score in top 5 are in *italics*; apparent molecular weights (mass) in kDa for corresponding gel bands.

often MultiTag hit multiple tags for different EST sequences of the same cDNA sequence (data not shown). However, MultiTag does require manual or software-assisted interpretation of MS/MS spectra that is not required by the conventional software. Because of the demonstrated efficiency of the conventional software, we speculate that MultiTag would be applied most efficiently in cases where stringent database searches fail to make the identification, or where they have recognized certain EST sequences only at the borderline of their scoring thresholds.

The authors are grateful to Drs. Andrei Popov and Eric Karsenti of EMBL, Heidelberg, Germany, for supplying gel separated Xenopus laevis proteins. The authors would also like to thank Ron Bonner of MDS Sciex, Concord, Canada, for his support, as well as the members of the Shevchenko group for stimulating discussions. We are grateful to Judith Nicolls for critical reading of the manuscript. Work in the Shevchenko lab was supported by BMBF grant PTJ-BIO/ 0313130.

References

[1] Mann, M., Wilm, M., *Anal Chem* 1994, *66*, 4390–4399.

[2] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[3] Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S. *et al.*, *Nature* 1996, *379*, 466–469.

[4] Sunyaev, S., Liska, A. J., Golod, A., Shevchenko, A., *Anal. Chem.* 2003, *75*, 1307–1315.

[5] Liska, A. J., Popov, A. V., Sunyaev, S., Coughlin, P. *et al.*, *Proteomics* 2004, *4*, 2707–2721.

[6] Mann, M., *Trends Biochem. Sci.* 1996, *21*, 494–495.

[7] Neubauer, G., King, A., Rappsilber, J., Calvio, C. *et al.*, *Nat. Genetic* 1998, *20*, 46–50.

[8] Liska, A. J., Shevchenko, A., *Trends Anal. Chem.* 2003, *22*, 291–298.

[9] Shevchenko, A., Sunyaev, S., Loboda, A., Bork, P. *et al.*, *Anal. Chem.* 2001, *73*, 1917–1926.

[10] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[11] Biemann, K., *Biomed. Environ. Mass Spectrom.* 1988, *16*, 99–111.

