ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/8988677

Tabb, D.L., Saraf, A. & Yates, J.R. III. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. Anal. Chem. 75, 6415-6421

ARTICLE *in* ANALYTICAL CHEMISTRY · JANUARY 2004

Impact Factor: 5.64 · DOI: 10.1021/ac0347462 · Source: PubMed

CITATIONS
225

READS
32

3 AUTHORS:

David L Tabb
Vanderbilt University
97 PUBLICATIONS **6,987** CITATIONS
SEE PROFILE

Anita Saraf
Stowers Institute for Medical Research
41 PUBLICATIONS **2,158** CITATIONS
SEE PROFILE

John R Yates
The Scripps Research Institute
616 PUBLICATIONS **60,394** CITATIONS
SEE PROFILE

Available from: David L Tabb
Retrieved on: 29 January 2016

NIH Public Access
Author Manuscript
Anal Chem. Author manuscript; available in PMC 2010 August 4.

Published in final edited form as:
Anal Chem. 2003 December 1; 75(23): 6415–6421. doi:10.1021/ac0347462.

GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model

David L. Tabb[†,‡], **Anita Saraf**[§], and **John R. Yates III**[*,‖]
SR11 Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Abstract

Shotgun proteomics is a powerful tool for identifying the protein content of complex mixtures via liquid chromatography and tandem mass spectrometry. The most widely used class of algorithms for analyzing mass spectra of peptides has been database search software such as SEQUEST. A new sequence tag database search algorithm, called GutenTag, makes it possible to identify peptides with unknown posttranslational modifications or sequence variations. This software automates the process of inferring partial sequence “tags” directly from the spectrum and efficiently examines a sequence database for peptides that match these tags. When multiple candidate sequences result from the database search, the software evaluates which is the best match by a rapid examination of spectral fragment ions. We compare GutenTag’s accuracy to that of SEQUEST on a defined protein mixture, showing that both modified and unmodified peptides can be successfully identified by this approach. GutenTag analyzed 33 000 spectra from a human lens sample, identifying peptides that were missed in prior SEQUEST analysis due to sequence polymorphisms and posttranslational modifications. The software is available under license; visit http://fields.scripps.edu for information.

Tandem mass spectrometry has been established as a powerful analytical tool for proteomics. Hunt and others[1] demonstrated the ability to sequence peptides directly from tandem mass spectra in the mid-1980s. Yates and Eng[2] automated this process by comparing database-derived sequences to uninterpreted tandem mass spectra. Improved separation technologies automated the analysis of samples ranging from protein complexes[3] to whole cell lysates[4] for protein content and protein modifications. As these technologies have been refined, the emphasis in proteomics has shifted from protein identification to posttranslational modification identification[5] and protein quantitation.[6]

Tandem mass spectra of peptides reveal sequences because they show the fragmentation products of peptides.[7] Ions of a peptide eluting from liquid chromatography are isolated by the mass spectrometer and collided energetically with inert gas molecules. Each peptide precursor breaks preferentially at a peptide bond, producing two fragments (one from the N-terminal b series, and one from the C-terminal y series). The fragment ions from thousands of such breakages form a tandem mass spectrum, where each ion is represented by a peak with a recorded intensity and a mass-to-charge (*m*/*z*) ratio. Each tandem mass spectrum indicates the sequence of a peptide by two superimposed ladders of sequence ions, though typically some fragment ions from each series do not appear in the spectrum. The sequence

[*]To whom correspondence should be addressed. jyates@scripps.edu. Fax: (858)784-8883.
[†]Department of Genome Sciences, University of Washington, Seattle, WA, 98195.
[‡]Current address: Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164.
[§]Current address: Abbott Laboratories, GPRD/R4ME, 200 Abbott Park Rd., AP31/L144, Abbott Park, IL 60064-6202.
[‖]Diversa, 4955 Directors Place, San Diego, CA 92121.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of the peptide can be deduced from the *m*/*z* differences between successive b or y series peaks.

Two major approaches have evolved for identifying peptide sequences from tandem mass spectra: database searching and de novo sequence inference. Database searching, as embodied in the SEQUEST[2] and Mascot[8] algorithms, enumerates peptide sequences from a database that have masses matching that of the fragmented peptide. The fragment ions expected for each candidate sequence are compared to the observed spectrum to determine which candidate is the correct sequence. De novo algorithms use a different approach, inferring sequences directly from the spectrum without recourse to a sequence database. Sherenga,[9] SeqMS,[10] and Lutefisk[11,12] are three examples of this technique, all based upon the "spectrum graph" approach presented by Bartels.[13] The accuracy possible with de novo techniques has been limited by the accuracy of *m*/*z* measurements in most tandem mass spectrometers and by models of fragment ion peak intensity that do not take into account many of the chemical factors involved. A middle path between database and de novo algorithms is the "sequence tagging" approach;[14] a short sequence "tag" is inferred from the tandem mass spectrum, and then database lookup finds complete peptide sequences that match this sequence and the sequence masses flanking it. Because algorithms to infer tag sequences have had limited accuracy, this technique has been applicable primarily to spectra where the partial sequences can be derived either manually or by simple algorithms.

Automated identification of posttranslationally modified peptides or peptides with sequence variations has been attempted by several researchers. The SEQUEST algorithm can be configured to identify peptides with up to three different types of specified modifications.[2] The Mascot algorithm has been adapted to search spectra for modified peptides from a particular protein.[15] Algorithms such as SALSA[16] and PEDANTA[17] attempt to interrelate spectra with shared motifs among their fragment ions. SALSA attempts to find spectra for peptides containing a partial sequence supplied by the user, while PEDANTA attempts to group spectra that differ from each other by up to two mass shifts. Still, the problem of identifying modified or variant peptide sequences is a challenging one, especially when the spectrum for the unmodified sequence is not present as a standard for comparison.

Sequence inference by de novo techniques has begun focusing on tandem mass spectra produced by new types of mass analyzers that yield higher mass accuracy and resolution of fragment ions. In this work, we analyze tandem mass spectra from quadrupole ion trap instruments. Ion traps typically determine the *m*/*z* ratios of fragment ions within 0.6 *m*/*z*. These instruments are used broadly in the proteomics field, and if techniques can be developed to infer accurate sequences from these spectra, much can be learned from existing data collections generated on relatively inexpensive instruments.

We present GutenTag, a new sequence tagging algorithm with several advantages over existing implementations. By using an empirically derived model of fragment ion peak intensity, the software achieves better accuracy in automating the inference of sequence tags from the spectrum, making it possible to apply sequence tagging in large-scale analyses. When the best tag sequences for the spectrum have been inferred, the algorithm searches the sequence database for each occurrence of any of these tags in a single pass. In most cases, this database search will yield multiple candidate peptides for each spectrum. GutenTag selects the best sequence by modeling the fragment ions for each candidate sequence and comparing these to the observed spectrum. The program can assign partial sequences to spectra, enabling it to identify spectra for peptides with sequences that differ from those found in the database or peptides that have been posttranslationally modified. We compare GutenTag identifications to those of SEQUEST for a defined mixture of proteins including both modified and unmodified peptides. GutenTag identifications from a shotgun proteomic

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

analysis of a human lens sample demonstrate the importance of the capability to match peptides with unspecified modifications, sequence errors, or both. Fully automated sequence tagging algorithms comprise a powerful complement to database search algorithms for large-scale proteomic analyses.

EXPERIMENTAL SECTION

Training Set Preparation

The extraction, digestion, separation, and mass spectrometry of the peptides used as GutenTag's training set were previously described.[18] Proteins from a culture of protease-deficient *Saccharomyces cerevisiae* (strain 1560) were divided into pellet and supernatant fractions and then reduced and alkylated. The pellet fraction was digested by CNBr, endoproteinase Lys-C, and trypsin, while the supernatant fraction was digested by endoproteinase Lys-C and trypsin. Twelve-cycle MudPITs[4] analyzed the soluble fraction twice and the pellet fraction once. Thermo Finnigan (San Jose, CA) LCQ Deca ion trap mass spectrometers conducted collision-induced dissociation on the eluting peptides. The 2 to 3 algorithm[19] reduced the number of spectra to be processed by removing copies with incorrectly assigned precursor charge states, using information from the fragment ions in each spectrum to judge the correct precursor charge state. Spectra were identified with Normalized SEQUEST,[20] a modification of the algorithm that normalizes XCorr scores to range between 0 and 1. Identifications were assembled and filtered by DTASelect.[21] The 1437 identifications allowed into the training set met each of the following criteria: the precursor was doubly charged, Normalized XCorr > 0.45, and the peptide was not posttranslationally modified.

Defined Mixture Preparation

The defined protein mixture was previously described by MacCoss et al.[20] A mixture of the five proteins was reduced and alklyated before being divided into four aliquots. Aliquot 1 was digested with trypsin, aliquot 2 was digested by elastase, aliquot 3 was digested by subtilisin, and aliquot 4 was digested by proteinase K. Only aliquots 1 and 4 were used for the comparison between GutenTag and SEQUEST. Each aliquot was separated by a six-cycle MudPIT analysis. A Thermo Finnigan LCQ Deca produced tandem mass spectra for these separations. The 2 to 3 algorithm selected which precursor charge states were applicable to these spectra. A database of protein sequences identified in the sample (both contaminants and known contents) was appended to the *S. cerevisiae* ORF database, which functioned as a collection of distractor sequences. Normalized SEQUEST and GutenTag both used this database for identification.

Lens Sample Preparation

The lens sample preparation was previously described by MacCoss et al.[22] Tissues were obtained from a 4-year-old congenital cataract patient. After extraction and solubilization, proteins were reduced and carboxyamidomethylated. The mixture was digested with trypsin. The sample was analyzed via an 18-cycle MudPIT separation, and spectra were collected on a Thermo Finnigan LCQ Deca. Initial SEQUEST analysis included the following modifications: phosphorylation, oxidation, methylation, and acetylation. GutenTag identifications employed a database consisting of previously identified protein components and the *S. cerevisiae* ORF database (as distractors). Follow-up SEQUEST searches used a database consisting of previously identified protein components.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

RESULTS AND DISCUSSION

GutenTag Algorithm

Software in the Java programming language was created to preprocess spectra, infer sequence tags from them, search a database for these tag sequences, and evaluate the peptide sequences resulting from these searches (see Figure 1). Although the software was developed under Microsoft Windows 2000, it runs equally well on Linux and other operating systems that support Java Virtual Machines. With a database of 6500 proteins (such as the *S. cerevisiae* database), the software requires ~1 s per spectrum on an AMD Athlon XP 1700+, and the time required scales roughly linearly with the size of the database. GutenTag was trained to infer sequences based on a statistical analysis of tandem mass spectra produced in a tryptic digest of a yeast proteome.[18]

Each spectrum is preprocessed before sequence tags are inferred. The first step in preprocessing removes isotopic variants of each fragment ion.[23] Next, the algorithm finds an intact peptide mass that results in optimal alignment of complementary ion series in the spectrum. This optimized precursor mass is used for all subsequent operations. Finally, the software finds pairs of fragment ions separated by the masses of amino acid residues and stores these links for later sequence generation. This preprocessing helps to limit tag generation to more likely sequences and allows the program to use tighter mass tolerances for database search.

GutenTag next recurses through the links between peaks to generate candidate sequence tags. The program can be configured to accept a range of sequence tag lengths. The best-scoring tags of each length are retained. In this work, GutenTag was configured to retain the 25 best-scoring tags that were three residues in length, though the software is capable of identifying longer tag sequences. In scoring each sequence tag, GutenTag constructs a model of the peaks expected for a given tag sequence, compares the observed and expected peaks, and produces a score rating the correspondence between them. The peaks through which the recursion passed to produce the tag are assumed to be y ions, and the ions complementary to them are assessed as b ions. If a spectrum lacks sufficient contiguous y ions, no successful tag will be inferred; GutenTag uses b ions as corroborative information to y ions because y ions are more intense in tryptic peptide spectra. The intensity score (described below) for y ions is multiplied by the intensity score for b ions and the m/z score for y ions to produce a tag's final score.

The expected intensity for each fragment ion depends on the series from which it comes, its mass relative to that of the precursor, and the flanking amino acid residues.[18] A fourth-order polynomial function for each ion series models the expected proportion of the spectrum's intensity accounted for by ions of a particular relative mass (fragment ion mass divided by precursor ion mass). The amino acids N-terminal and C-terminal to the fragment ion may increase or decrease this intensity (each amino acid's characteristic "N-Bias" was drawn from a statistical analysis of the training set[18]). GutenTag estimates the percentage of ions in the appropriate series that would fall between the expected intensity and the observed intensity. In effect, GutenTag evaluates differences between expected and observed intensities via probability rather than intensity. If expected and observed intensities are very similar, the percentage of ions with intensities between these two thresholds will be very low, but if the expected peak is very intense and the observed peak is not, the percentage will be high. This percentage is subtracted from 1.0 to give a particular peak's intensity score, which is multiplied with the other peaks' scores to give the sequence tag's intensity score for that series.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The expected *m*/*z* value for each fragment ion is calculated to minimize the difference between observed and expected locations while maintaining the sequence-specific mass offsets between adjacent fragment ions. The distance separating each peak *m*/*z* from its modeled position is evaluated in light of the mass accuracy of the mass analyzer to yield a probability that it is the ion expected. The *m*/*z* score for each y ion peak is multiplied together with the others to produce the tag's *m*/*z* score. The b ions are not included in the *m*/*z* score because some of them may be absent.

The sequence tags retained for each spectrum are searched in parallel against the sequence database.[24] In some cases, multiple tags will match to the same candidate peptide sequence, as shown in Figure 1, where YVD, VDD, and DDE all match to the sequence KLLSYVDDEAFIR. When the tag matches a protein sequence, the flanking sequences are investigated to determine whether they match the masses expected based on the tag position within the spectrum. If both masses match, a complete sequence can be evaluated against the spectrum. If only one flanking mass matches, the peptide may have been posttranslationally modified, and only part of the spectrum's sequence can be compared to the spectrum. Each matching sequence, complete or partial, is added to the list of candidate sequences.

When the database search is complete, a list of peaks is generated for each candidate sequence. The *m*/*z* values of these peaks are those associated with the b and y ions projected from the sequence; if a candidate sequence is partial, only a portion of the fragment ions can be modeled. The intensity values of these theoretical peaks are estimated as they are for sequence tag inference; the model intensity is first based on the relative mass and series of the fragment ion and then modified by a multiplier that reflects the amino acids adjacent to the fragment ion. Peaks are sought within the observed spectrum at the modeled *m*/*z* values, and the normalized dot-product algorithm[25] is used to compare the theoretical and observed peak intensities:

$$d = \frac{\sum_{to}}{\sqrt{\sum t^2 \sum^{o^2}}}$$

where *d* is the normalized dot-product, *t* is the theoretical intensity, and *o* is the observed intensity. The primary score generated by GutenTag is the normalized dot-product value (which ranges from 0 to 1) multiplied by the number of modeled ions for which a matching observed peak can be found. GutenTag also determines the percentage of spectral intensity accounted for by modeled ions. The candidate sequences are sorted by the primary score and recorded to a "SQT" result file, using the Unified SEQUEST file format.

Validation on Defined Protein Mix

GutenTag's ability to identify peptides was validated against Normalized SEQUEST in its analysis of a defined five-protein mix including bovine serum albumin, horse apomyoglobin, rabbit cytochrome *c*, rabbit phosphorylase A, and bovine *β*-casein.[20] Counting only spectra from doubly charged precursor ions, a tryptic digest produced 6170 spectra, while a proteinase K digest yielded 7972 spectra. Only doubly charged peptide spectra were included in this analysis because GutenTag's intensity model is specific to this type. SEQUEST and GutenTag were run on these collections, using a sequence database including the above proteins, common contaminants, and the *S. cerevisae* ORFs. SEQUEST was not configured to identify any posttranslational modifications in this analysis. Identifications matching only yeast proteins were labeled as false, while identifications to the five proteins above and to contaminants were assessed as true identifications.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The numbers of true and false identifications from each algorithm in each set of spectra were recorded in Table 1. GutenTag assigned fewer correct, complete sequences to the spectra than did Normalized SEQUEST. When GutenTag produced a false identification, however, it generally matched the sequence partially rather than completely; 70% of tryptic digest spectra that were assigned complete sequences were assigned true identifications, while only 18% of partial matches were true identifications. SEQUEST produced a larger number of true identifications than GutenTag, but its limitation to yield only complete sequences meant that the false positives greatly outnumbered the true positives among its identifications.

GutenTag produced fewer true identifications of complete sequence than SEQUEST for a variety of reasons. First, accurately inferring a sequence tag from a spectrum depends heavily upon spectral quality; a peptide present at low quantities or that undergoes an unusual fragmentation pathway may produce a spectrum that is insufficiently informative for generation of valid tag sequences. Second, the tolerances inherent to any software for identifying spectra have been tuned in SEQUEST for almost a decade; GutenTag's performance may be improved by optimizing the mass errors tolerated in database search, the constants used to predict each amino acid's effect on fragmentation, and other values crucial to performance. Spectra that fragment by unusual pathways (such as those without mobile protons[26]) may ultimately be identifiable only by database search algorithms, but for other spectra, GutenTag's accuracy can be expected to improve through refinement.

A standard approach for finding the best identifications in proteomics data is to accept all identifications above a threshold score. The scores that would reject 95% of false positive complete sequences are reported in Table 1. The proportion of identifications above this cutoff that were true positives was substantially higher for GutenTag than for SEQUEST, even in proteinase K spectra. The percentages of all true identifications retained by this cutoff was also higher for GutenTag than SEQUEST. Even though SEQUEST yields a larger number of true positives, GutenTag's true, complete-sequence identifications are more readily separated from false ones on the basis of score (see Figure 2). The algorithm separates true and false positives among partial identifications less powerfully (see Figure 3). When GutenTag yields a partial sequence identification, a complete sequence with necessary modifications should be built from the partial identification to evaluate the match.

Receiver operating characteristic (ROC) curves can be used to determine the value of scoring functions.[27] The plots in Figure 4 show comparisons of GutenTag and SEQUEST on both tryptic and proteinase K spectra. The points on the curve all correspond to particular cutoff scores. The vertical position of the point gives the number of true identifications passing the cutoff score while the horizontal position indicates the number of false identifications over the cutoff. An ideal algorithm scores all true identifications higher than all false identifications; an ROC plot for such an algorithm would be a right angle. The number of spectra in the defined digest data that contain sufficient data for identification is unknown, and so numbers of identifications are shown rather than proportions of possible identifications. The curve for GutenTag in tryptic spectra is closer to the ideal than the curve for proteinase K spectra. Because GutenTag has the capability to attempt a partial identification rather than a complete one, GutenTag yields lower numbers of complete identifications for these two sets of spectra than does SEQUEST. If cutoffs are lowered sufficiently to begin accepting the additional complete identifications produced by SEQUEST, approximately one in four of the passing identifications will be false hits.

The top 100 identifications were examined for each algorithm on both sets of spectra. For GutenTag, only complete sequences were included in the top 100. Each identification was checked against the result for the other algorithm. GutenTag's top identifications matched SEQUEST's for all 100 of the tryptic spectra and 95 of the proteinase K spectra. Of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

SEQUEST's top 100, 88 matched GutenTag's identification among the tryptics, and 82 matched among the proteinase K spectra. The relatively low percentage of SEQUEST identifications validated by GutenTag identifications reflected that SEQUEST allows more variability in spectra than does GutenTag. GutenTag's scoring model favored peptides terminating with Arg or Lys more highly than SEQUEST; its proteinase K identifications included 18 peptides terminating in these residues, while SEQUEST's proteinase K identifications included 12. GutenTag's spectrum model is more specific to tryptic peptides than is SEQUEST's.

Identification of Modified Peptides

Several posttranslational modifications were identified among the proteins of the defined mixture. GutenTag's partial sequence identifications were analyzed to find spectra corresponding to known modifications. An acetylated peptide at the amino terminus of rabbit cytochrome *c*[28] was identified as –EKGKKIF, matching the final seven residues of the peptide. When the sequence of the protein's N-terminus was added to GutenTag's identification, the remaining mass of 42 Da indicated a likely acetylation. Likewise, an acetylation on rabbit phosphorylase A's N-terminus[29] matched to six spectra as –PLS-DQEK and three spectra as –RPLSDQEK. A phosphorylation on bovine *β*-casein's serine 50[30] matched to seven spectra as –QQQT-EDELQDK and one spectrum as –EQQQTEDELQDK, with the 80 Da remainder indicating a phosphorylation. The algorithm correctly identified spectra when modifications prevented complete sequence identification.

This capability was put to the test in an assessment of post-translational modifications in human lens tissue from a 4-year-old congenital cataract patient.[22] Shotgun proteomics captured 32 950 spectra from a tryptic digest of this complex sample. These identifications were filtered to include only the matches to crystallin α. GutenTag matched 1414 spectra to chain A and 2188 spectra to chain B; lens tissues concentrate crystallin proteins to produce their refractive capabilities, so finding a high percentage of spectra from these proteins is expected. Chain A's identifications included 375 partial sequences, while chain B's included 367. Each partial identification was compared to the database sequence to determine the amount of missing or added mass. Of the 742 partial sequences identified, 134 corresponded to mutations from threonine to serine with subsequent *β*-eliminations.

In chain A, 74 spectra for a 2553-Da peptide were all matched to similar sequences, with the N-terminal portions identified to sequences ranging from LFDQFFGEGLFEY– to LFDQFF-GEGLFEYDLLP–. The sequence database implied that these peptides had the sequence LFDQFFGEGLFEYDLLPFLSST, but the mass of the observed peptides instead indicated that the C-terminal threonine had been mutated to serine and that a *β*-elimination of one of the serines to dehydroalanine had taken place (a reaction commonly associated with phosphoserine residues). A SEQUEST search was configured to search for a mass shift of −32 Da on threonine residues (−14 Da for the mutation to serine and −18 Da for the *β*-elimination), and 137 spectra were identified to the modified sequence, with a best (nonnormalized) SEQUEST XCorr of 5.514. The spectra did not reveal which of the three serines had *β*-eliminated. In this example, GutenTag identified the peptide despite a change in protein sequence from a putative single-nucleotide polymorphism and an unanticipated modification.

Chain B showed two similar sets of spectra. Five spectra matched to the sequence LFDQFFGEHLLESDLFPTSS, where threonine 42 had become a serine and a *β*-elimination had taken place. The above SEQUEST seach was able to match 33 spectra to this sequence once this modification was specified, with a best XCorr of 5.651. GutenTag matched 48 spectra to LFDQFFGE-HLLESDLFPTSTSS with a *β*-elimination, which suggested that leucine 44 may also be mutated to a serine in this sample. A total of 308 SEQUEST

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

identifications, scoring as high as 5.7161, were found when the program was configured for this possibility. If these modifications were not specified in the SEQUEST configuration, these spectra were incorrectly identified.

A variety of other modifications were also identified. Acetylations on the N-termini of both chains were observable from large numbers of spectra. Spectra featuring the characteristic +80-Da mass difference of phosphorylation agreed with previously documented modifications on these proteins.[22] In addition, many spectra resulted from peptide ions that had lost ammonia or water before precursor ion selection. The diversity of partial identifications made it necessary to focus on the potential modifications evidenced by multiple spectra. While GutenTag automates the process of identifying the individual spectra, only partial automation was available for collating large collections of these partial identifications.

CONCLUSION

GutenTag incorporates several advantages over existing proteomic algorithms. Its fragmentation model makes use of the influence of neighboring amino acids and the relative mass of fragment ions to improve the accuracy of automated sequence tag inference. Its ability to search a database with multiple tags in a single pass allows users to increase the number of tags retained for each spectrum, resulting in increased performance and accuracy. GutenTag's ability to score multiple candidate sequences resulting from database searches based on fragment ions rather than peptide mass tolerance sets it apart from existing sequence tag search algorithms. By uniting in a single program the capability to infer tag sequences and search databases, GutenTag outperforms existing techniques and is usable on larger data sets than previously possible.

GutenTag embodies a new capability for large-scale proteomics. The automated and accurate interpretation of partial sequences from tandem mass spectra will play an important role in the identification of protein sequence features not predicted through bioinformatics. We have shown the algorithm accuracy compares quite well with database searching algorithms for peptides created by site-specific and nonspecific proteases. The algorithm can also help identify unanticipated modifications, sequence variations, and possibly alternate splice sites in proteins. By combining the best elements of de novo and database search algorithms, sequence tagging increases the scope of biological discovery possible using shotgun proteomic strategies.

Acknowledgments

The authors thank Dayin Lin for providing the training set spectra. David Goldberg of Xerox Parc made many helpful suggestions in the development of GutenTag. Matthew Wiener of Merck Research Laboratory suggested the ROC analysis. Christine Wu provided the spectral collection for the defined protein mixture. J.R.Y. was funded by NIH grant RR11823-08, while D.L.T. derived support from NIH grant R33 CA81665, and A.S.'s research was funded by R01 EY13288-03.

References

1. Hunt DF, Yates JR III, Shabanowitz J, Winston S, Hauer CR. Proc Natl Acad Sci USA 1986;83:6233–6237. [PubMed: 3462691]
2. Eng JK, McCormack AL, Yates JR III. J Am Soc Mass Spectrom 1995;67:1426–1436.
3. Link AJ. Trends Biotechnol 2002;20:S8–13. [PubMed: 12570153]
4. Washburn MP, Wolters D, Yates JR III. Nat Biotechnol 2001;19:242–247. [PubMed: 11231557]
5. Mann M, Jensen ON. Nat Biotechnol 2003;21:255–261. [PubMed: 12610572]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

6. Ranish JA, Leslie DM, Purvine SO, Goodlett DR, Eng J, Aebersold R. Nat Genet 2003;33:349–355. [PubMed: 12590263]
7. Wysocki VH, Tsaprailis G, Smith LL, Breci LA. J Mass Spectrom 2000;35:1399–1406. [PubMed: 11180630]
8. Perkins DN, Pappin JC, Creasy DM, Cottrell JS. Electrophoresis 1999;20:3551–3567. [PubMed: 10612281]
9. Dančík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. J Comput Biol 1999;6:327–342. [PubMed: 10582570]
10. Fernandez-de-Cossio J, Gonzalez J, Satomi Y, Shima T, Okumura N, Besada V, Betancourt L, Padron G, Shimonishi Y, Takao T. Electrophoresis 2000;21:1694–1699. [PubMed: 10870956]
11. Taylor JA, Johnson RS. Rapid Commun Mass Spectrom 1997:1067–1075. [PubMed: 9204580]
12. Taylor JA, Johnson RS. Anal Chem 2001;73:2594–2604. [PubMed: 11403305]
13. Bartels C. Biomed Environ Mass Spectrom 1990;19:363–368.
14. Mann M, Wilm M. Anal Chem 1994;66(6):4390–4399. [PubMed: 7847635]
15. Creasy DM, Cottrell JS. Proteomics 2002;2:1426–1434. [PubMed: 12422359]
16. Liebler DC, Hansen BT, Davey SW, Tiscareno L, Mason DE. Anal Chem 2002;74:203–210. [PubMed: 11795795]
17. Pevzner PA, Dančík V, Tang CL. J Comput Biol 2000;7:777–787. [PubMed: 11382361]
18. Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR III. Anal Chem 2003;75:1155–1163. [PubMed: 12641236]
19. Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR III. J Proteome Res 2002;1:211–215. [PubMed: 12645897]
20. MacCoss MJ, Wu CC, Yates JR 3rd. Anal Chem 2002;74:5593–9. [PubMed: 12433093]
21. Tabb DL, McDonald WH, Yates JR 3rd. J Proteome Res 2002;1:21–26. [PubMed: 12643522]
22. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, Clark JI, Yates JR III. Proc Natl Acad Sci USA 2002;99:7900–7905. [PubMed: 12060738]
23. Kubinyi H. Anal Chim Acta 1991;247:107–119.
24. Aho AV, Corasick MJ. Commun ACM 1975;18:333–340.
25. Wan KX, Vidavsky I, Gross ML. J Am Soc Mass Spectrom 2002;13:85–88. [PubMed: 11777203]
26. Huang Y, Wysocki VH, Tabb DL, Yates JR 3rd. Int J Mass Spectrom 2002;219:233–244.
27. Baker SG. J Natl Cancer Inst 2003;95:511–515. [PubMed: 12671018]
28. Krishna RG, Chin CC, Wold F. Anal Biochem 1991;199:45–50. [PubMed: 1807161]
29. Nakano K, Hwang PK, Fletterick RJ. FEBS Lett 1986;204:283–287. [PubMed: 3015680]
30. Fiat AM, Jolles P. Mol Cell Biochem 1989;87:5–30. [PubMed: 2671666]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

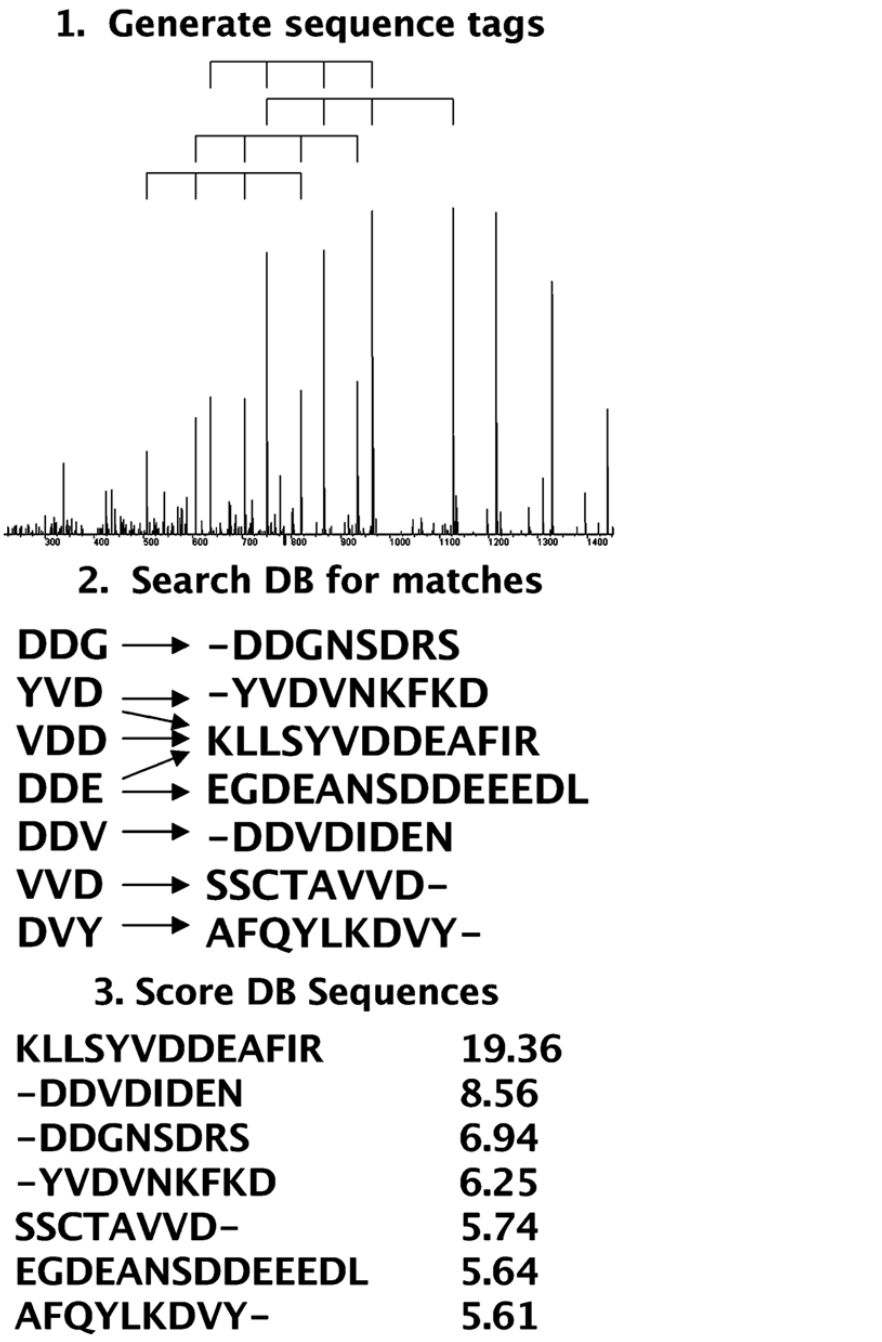

Figure 1.

GutenTag procedure. GutenTag infers short sequences directly from each spectrum. The best sequence tags are sought in a sequence database to find peptide sequences that match a tag sequence and at least one flanking sequence mass. These partial and complete peptide sequences are ranked by dot-product score. In the above example, three tag sequences match to the same complete peptide sequence. This complete sequence scores more highly than the other sequences.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

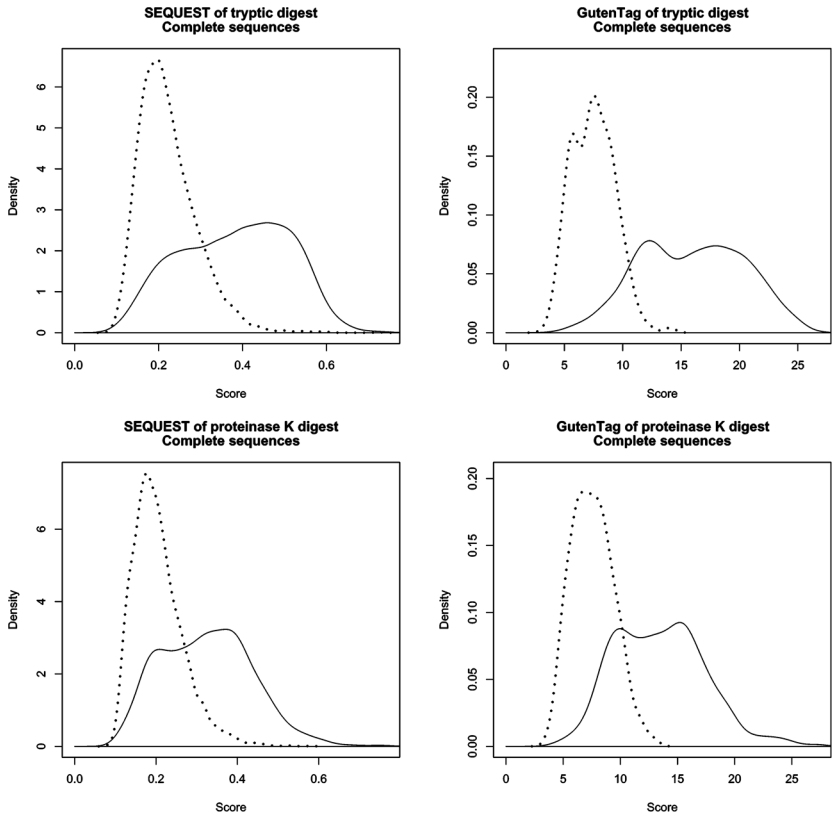

Figure 2.

Algorithm comparisons for complete sequences. Both SEQUEST and GutenTag score true identifications (solid line) more highly than false identifications (dotted line), but the degree of score overlap for complete sequences in reduced in GutenTag. Because GutenTag produces a lower number of true identifications than SEQUEST (see Table 1), GutenTag's improved separation between true and false identification scores is important in achieving parity between the algorithms. Because GutenTag's intensity model was trained on tryptic spectra, it achieves better separation for tryptic peptides than for proteinase K peptides.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

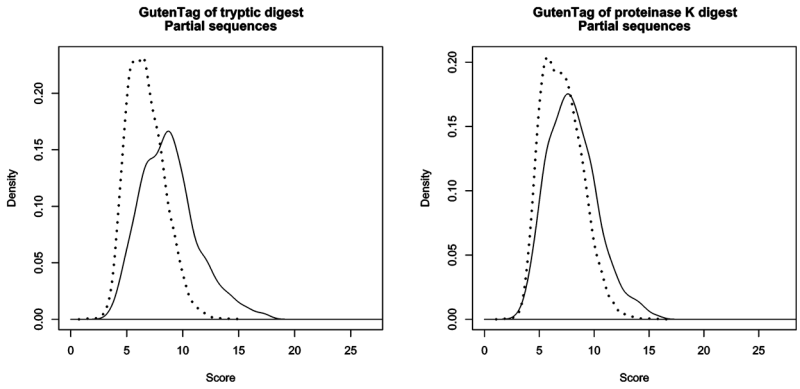

Figure 3.

Partial sequence comparisons. Partial sequences predict fewer peaks than do complete sequences, and so GutenTag does not separate true identifications from false identifications as effectively in partial sequences. In the results for the proteinase K digest, the separation is almost nonexistent due to GutenTag's trypsin-specific fragmentation model. GutenTag's partial sequence identifications should be viewed as a basis upon which to build complete sequence identifications (including whatever posttranslational modifications are necessary) rather than final answers in themselves.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

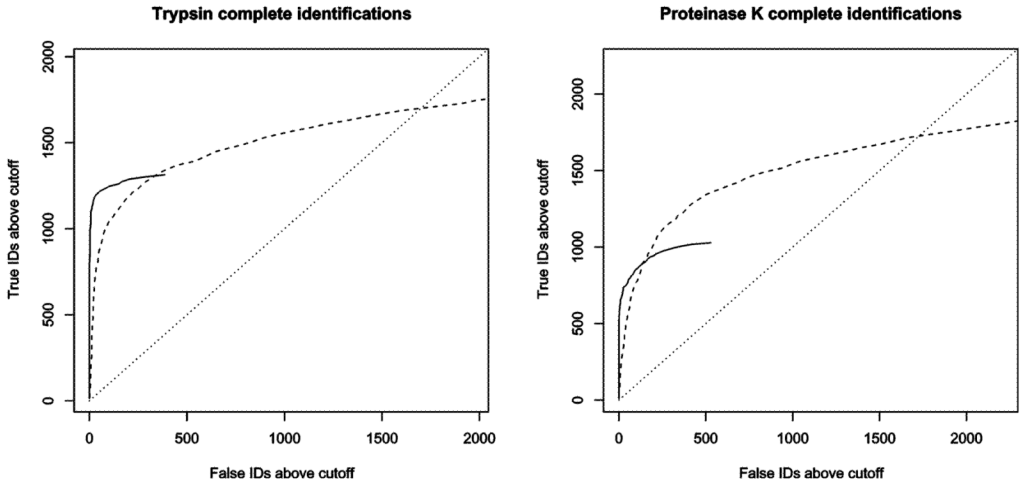

Figure 4.

Comparison of true and false positives. ROC curves compare the number of true positives to the number of false positives for various cutoffs for a particular scoring function. The above plots analyze the complete sequence identifications from GutenTag (solid lines) and SEQUEST (dashed lines). Because GutenTag can examine partial sequences in addition to complete sequences to explain each spectrum, the number of complete sequences it proposes is significantly less than SEQUEST (see Table 1). GutenTag's improved separation between true and false complete identifications is shown by its producing a curve more similar to a right angle, particularly in tryptic spectra. To make use of SEQUEST's greater numbers of correct sequences, one must lower cutoff scores to a point that substantial numbers of false positive sequences also pass the cutoff; the dotted diagonals indicate the points at which the numbers of false positives equal the true positives.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

code[a]	complete T IDs[b]	complete F IDs[b]	partial T IDs[c]	partial F IDs[c]	95%ile of F scores[d]	% T over cutoff[d]	% of T retained[d]
T-GT	1328	558	766	3515	10.4	97.7	89.2
T-SQ	1987	4183			0.348	85.1	60.2
K-GT	1039	667	621	5645	10.7	95.6	71.5
K-SQ	2230	5742			0.319	80.0	51.7

[a]The “code” column gives the set of spectra and algorithm. “T” indicates tryptic spectra; while “K” indicates proteinase K spectra. “GT” specifies GutenTag was used, and “SQ” denotes that Normalized SEQUEST identified the peptides.

[b]The “complete” columns indicate the numbers of true and false identifications among the spectra matched to complete peptide sequences.

[c]The “partial” columns indicate the numbers of true and false identifications among spectra matched to partial sequences. SEQUEST does not match spectra to partial sequences.

[d]The rightmost section of the table gives the cutoff score that removes 95% of the false positive identifications, the percentage of the identifications retained that are true, and the percentage of all true identifications that are retained by the cutoff. Both algorithms perform better on tryptic digests than proteinase K digests.

Anal Chem. Author manuscript; available in PMC 2010 August 4.