Anal. Chem. **2005,** *77,* 2464−2474

DBDigger: Reorganized Proteomic Database Identification That Improves Flexibility and Speed

David L. Tabb,*,† Chandrasegaran Narasimhan,†,‡ Michael Brad Strader,§ and Robert L. Hettich§

Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6164, Genome Science and Technology, UT-ORNL Graduate School, Oak Ridge, Tennessee 37830-8026, and Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6131

Database search identification algorithms, such as Sequest and Mascot, constitute powerful enablers for proteomic tandem mass spectrometry. We introduce DBDigger, an algorithm that reorganizes the database identification process to remove a problematic bottleneck. Typically such algorithms determine which candidate sequences can be compared to each spectrum. Instead, DBDigger determines which spectra can be compared to each candidate sequence, enabling the software to generate candidate sequences only once for each HPLC separation rather than for each spectrum. This reorganization also reduces the number of times a spectrum must be predicted for a particular candidate sequence and charge state. As a result, DBDigger can accelerate some database searches by more than an order of magnitude. In addition, the software offers features to reduce the performance degradation introduced by posttranslational modification (PTM) searching. DBDigger allows researchers to specify the sequence context in which each PTM is possible. In the case of CNBr digests, for example, modified methionine residues can be limited to occur only at the C-termini of peptides. Use of "context-dependent" PTM searching reduces the performance penalty relative to traditional PTM searching. We characterize the performance possible with DBDigger, showcasing MASPIC, a new statistical scorer. We describe the implementation of these innovations in the hope that other researchers will employ them for rapid and highly flexible proteomic database search.

Tandem mass spectrometry is an essential technology for protein identification from complex mixtures. Samples ranging from electrophoretic gel spots to complete proteomes can be analyzed.[1,2] Typically, sample proteins are digested enzymatically, and the resulting peptides are separated by liquid chromatography en route to the mass spectrometer. While peptides elute from the column, the mass spectrometer repeatedly cycles between "survey" scans and "product ion" scans. The survey scans enumerate the peptide ions eluting at a particular moment in time. The mass spectrometer isolates the peptide ions with the most intense signals and subjects them to collision-induced dissociation.[3] The product ion scans catalog the fragments produced by the breakage of the peptide ions to produce tandem mass spectra. Once all spectra have been collected, database identification algorithms such as Sequest[4] or Mascot[5] identify the peptide represented by each tandem mass spectrum, and the protein content of the sample can be assessed by assembling confident peptide identifications that come from the same proteins.

The database identification algorithm was introduced with the publication of Sequest in 1994.[4] This type of software reads a tandem mass spectrum and identifies the most likely peptide sequence to have produced the spectrum from a provided protein sequence database. Three main processes are involved: candidate sequence generation, theoretical spectrum prediction, and scoring. The algorithm enumerates candidate sequences by finding portions of database protein sequences that represent peptides with masses matching that observed for a particular spectrum. Theoretical spectra for these candidate sequences are typically produced by predicting the N-terminal b series ions and C-terminal y series ions[6] that would result if such a peptide fragmented at its amide bonds; a y5 ion, for example, contains the C-terminal five amino acids, while a y4 ion contains the C-terminal four amino acids of a peptide. Each of these theoretical spectra is compared to the observed spectrum, and the best matches are reported with their associated scores for each spectrum.

Database identification algorithms have evolved since Sequest's initial publication. The first substantial change came in 1995, when Sequest was modified to handle posttranslational modification (PTM) searches.[7] To hunt for PTMs with Sequest, a user specifies the change in mass associated with the modified form of a particular residue or set of residues (such as a gain of 16 Da for

* Corresponding author: (phone) 865 241-6006; (fax) 865 576-5332; (e-mail) tabbdl@ornl.gov.

† Life Sciences Division, Oak Ridge National Laboratory.

‡ UT-ORNL Graduate School.

§ Chemical Sciences Division, Oak Ridge National Laboratory.

(1) Canettieri, G.; Morantte, I.; Guzmán, E.; Asahara, H.; Herzig, S.; Anderson, S. D.; Yates, J. R., 3rd.; Montminy, M. *Nat. Struct. Biol.* **2003**, *10*, 175−181.

(2) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(3) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. *J. Mass Spectrom.* **2000**, *35*, 1399−1406.

(4) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.

(5) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

(6) Roepstorff, P.; Fohlman, J. *Biomed. Mass Spectrom.* **1984**, *11*, 601.

(7) Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1995**, *67*, 1426−1436.

10.1021/ac0487000 CCC: $30.25 © 2005 American Chemical Society
Published on Web 03/05/2005

oxidation of Met residues). When the algorithm generates candidate sequences, it considers either of two masses possible for the specified residues. The next substantial change in these algorithms was heralded by the 1999 introduction of Mascot.[5] Instead of following Sequest's lead in producing an identification score in an arbitrary range, Mascot introduced the idea of scoring identifications by their probabilities of being false positives. Other algorithms have adopted this model as well.[8−10] Users of Sequest have responded by creating systems to calculate similar probabilities from the collection of subscores produced by the software.[11,12] A third shift in database identification software came with the publication of the TANDEM algorithm in 2003.[13,14] In this article, Craig and Beavis argued that database identification should be an iterative process in which rapid searches (such as those limited to tryptic peptides) can be used to limit the search space for complex searches (such as those incorporating multiple PTMs or sequence variants). A final way in which the database search field has changed is that accessory software for enabling proteomics has developed considerably. Tools exist to cull low-quality spectra,[15,16] assess charge states from spectra prior to identification,[17,18] parallelize database search over multiple computers,[18] and perform data mining of identifications.[19,20] Taken as a whole, the informatics behind proteomics has matured considerably in the past decade.

Several problems have not been addressed by these developments. The first is that the numbers of spectra produced in mass spectrometry facilities are growing at an increasing rate.[19] Database searches as currently run identify each spectrum independently, and so processing time scales linearly with the number of spectra. An identification strategy that decouples the processing time from the number of spectra would be beneficial for improving identification throughput. A second problem involves searches for multiple PTMs. Because most existing strategies limit the numbers of PTMs that can be sought in any one search, some users have adopted strategies where multiple searches are conducted and then collated;[21] search engines that can handle substantial numbers of PTMs simultaneously would be more efficient for finding PTMs. Problems of identification sensitivity and accuracy continue to be issues of fundamental importance for proteomics.

In this article, we characterize DBDigger, a database search algorithm designed to improve the flexibility of PTM searching and to identify large sets of spectra more efficiently. DBDigger can target particular PTMs with greater specificity and so reduce the time required to identify them. DBDigger generates candidate sequences only once from the database, decreasing the time penalty for multiple PTM searching. The algorithm is designed for single-pass, rapid scoring of potential matches. The software can identify peptides from multiple isotopic distributions in a single pass. To showcase DBDigger, we present results from MASPIC,[22] a new statistical scorer for comparing peaks between a theoretical spectrum and an observed spectrum. DBDigger's modular framework supports research on improved theoretical spectrum prediction and spectral comparison algorithms. These capabilities position DBDigger as a model database search algorithm.

EXPERIMENTAL SECTION

All proteins, salts, buffers, dithiothreitol (DTT), and guanidine hydrochloride were obtained from Sigma Chemical Co. (St. Louis, MO). Sequencing-grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (affiliate of Merck KgaA, Darmstadt, Germany). HPLC-grade acetonitrile and water were used for all LC/MS/MS analyses (Burdick and Jackson, Muskegon, MI). Ultrapure 18 MΩ water used for sample buffers was obtained from a Millipore Milli-Q system (Bedford, MA). Fused silica was purchased from Polymicro Technologies (Phoenix, AZ).

Preparation of Protein Standard Mixture Training and Test Sets. Protein standard mixture training and test sets were generated from six proteins: chicken egg lysozyme, bovine serum albumin, bovine carbonic anhydrase, horse myoglobin, bovine hemoglobin, and yeast alcohol dehydrogenase I. The proteins were mixed so that an equal weight of each was present. Hemoglobin included both α and β polypeptides, and a quantity of yeast alcohol dehydrogenase II was found in the dehydrogenase I for a total of eight polypeptides. The proteins were digested by different protocols for the two samples. Trypsin was added to an estimated 1:5 enzyme-to-substrate ratio (w/w). The training set was digested with trypsin in 80% acetonitrile for 1 h at 37 °C. The test set was digested with trypsin overnight at 37 °C in 50 mM Tris-HCl, 10 mM $CaCl_2$ (pH 7.6). After digestion, disulfide bonds between peptides were reduced by adding DTT to a final concentration of 20 mM and incubating for 1 h at 37 °C.

For both training and test mixtures, 1D LC/MS/MS experiments were performed with a Famos/Switchos/UltiMate HPLC System (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA XP Plus quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with a nanospray source. A 300 μm i.d. × 5 mm nano C_{18} precolumn (LC Packings) preconcentrated and cleaned the samples on-line with the other columns. The flow rate for reversed-phase liquid chromatography was 0.15 μL/min with a 160-min linear gradient from 100% solvent A (95% H_2O/5% ACN/0.5% formic acid) to 100% solvent B (30% H_2O/70% ACN/0.5% formic acid). The Vydac

(8) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J. Proteome Res.* **2004**, *3*, 958−964.
(9) Sadygov, R. G.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 3792−3798.
(10) Fridman, T.; Razumovskaya, J.; VerBerkmoes, N.; Hurst, G.; Protopopescu, V.; Xu, Y. *J. Bioinformatics Comput. Biol.*, in press.
(11) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(12) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 378−386.
(13) Craig, R.; Beavis, R. C. *Rapid Commun Mass Spectrom.* **2003**, *17*, 2310−2316.
(14) Craig, R.; Beavis, R. C. *Bioinformatics* **2004**, *20*, 1466−1467.
(15) Nesvizhskii, A. I.; Vogelzang, M.; Aebersold, R. *Proc. 52nd Am. Soc. Mass Spectrom. Conf.* 2004, ThPA 012.
(16) Bern, M.; Goldberg, D.; McDonald, H.; Yates, J. R., III *Proc. 12th Int. Conf. Intell. Syst. Mol. Biol.* 2004.
(17) Perez, R. E.; Asara, J. M.; Lane, W. S. *Proc. 51st Am. Soc. Mass Spectrom. Conf.* 2003, MPK 336.
(18) Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., 3rd. *J. Proteome Res.* **2002**, *1*, 211−215.
(19) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd. *J. Proteome Res.* **2002**, *1*, 21−26.
(20) Eddes, J. S.; Kapp, E. A.; Frecklington, D. F.; Connolly, L. M.; Layton, M. J.; Moritz, R. L.; Simpson, R. J. *Proteomics* **2002**, *2*, 1097−1103.
(21) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., III. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900−7905.
(22) Narasimhan, C.; Tabb, D. L.; VerBerkmoes, N. C.; Hettich, R. L.; Uberbacher, E. C., unpublished work. Oak Ridge National Laboratory, 2004.

218MS5.07515 C_{18} column (75 μm i.d. × 15 cm, 300-Å pore size, 5-μm particles) was connected to the nanospray source with 20-μm-i.d. fused-silica tubing. The LCQ was operated in the data-dependent mode with dynamic exclusion enabled,[23] where the four most abundant peaks in the 400−2000 m/z range were subjected to MS/MS analysis.

Preparation of *Rhodopseudomonas palustris* Ribosomal Proteins. 70S ribosomes from *R. palustris* were purified and fractionated using a high salt sucrose cushion and sucrose density fractionation as previously described.[24] Acid-extracted[25] proteins were denatured and reduced in 6 M guanidine hydrochloride, 50 mM Tris-HCl (pH 7.6), with 10 mM DTT at 60 °C for 45 min. Next, the proteins were digested with trypsin overnight at 37 °C. Remaining disulfide bonds were reduced with 10 mM DTT at 60 °C for 45 min. The resulting peptide mixtures were desalted using C_{18} reversed-phase extraction (Sep-Pak, Waters, Milford, MA). These mixtures were separated in 2D LC/MS/MS experiments using a Famos/Switchos/UltiMate 2D HPLC system (LC Packings) coupled to an LCQ-DECA-XP Plus ion trap MS equipped with a ThermoFinnigan nanospray source. Strong cation exchange was conducted in a 500 μm i.d. × 15 mm column (LC Packings). Peptides eluted into an LC Packings reversed-phase precolumn (300 μm i.d. × 5 mm, 300-Å PepMap). The desalted peptides flowed into a nanoscale Vydac 218MS5.07515 C_{18} analytical column (75 μm i.d. × 15 cm, 300-Å pore size, 5-μm particles). A 150-min gradient from solvent A (95% H_2O/5% ACN/0.1% formic acid) to solvent B (30% H_2O/70% ACN/0.1% formic acid) ran at 200 nL/min to elute peptides into the mass spectrometer. The LCQ was run in the data-dependent mode with dynamic exclusion enabled.[23] For these experiments, 10−30 μg of starting material was injected by the autosampler followed by 14 separate salt injections at ammonium acetate concentrations of 0, 25, 50, 100, 200, 400, 600, and 800 mM, 1 M, and five injections of 2 M.

DBDigger Software. A database search identifier was created in the C++ programming language and named "DBDigger". The code has been compiled and run under Linux for ×86, Alpha, and PowerPC architectures as well as Microsoft Windows 2000 and XP. The software is modular, allowing alternative systems for theoretical spectrum prediction and spectral comparison to be evaluated. The algorithm begins by loading the search configuration and spectra to be identified into memory. The software then processes the specified protein sequence database, iterating through all possible candidate sequences for each protein sequence entry. Theoretical spectra are predicted for candidate sequences with masses that match observed spectra, and these theoretical spectra are compared to the observed spectra by the specified scoring algorithm. Each of these steps is described in greater detail below.

Several steps are necessary before identification can begin. DBDigger is configured from a text file (an example of which appears in Supporting Information). This file allows researchers to identify peptides of multiple isotopic contributions in a single pass. The elemental composition for each amino acid residue follows; the software automatically determines the mass for each residue based on the specified isotopic distributions. The elemental compositions of posttranslationally modified residues are also specified in the configuration file along with allowable neighboring residues (with "−" indicating the peptide terminus). Researchers can specify rules defining each permissible type of protein cleavage (such as tryptic cuts after Lys and Arg residues, but not before Pro). In addition, the user can select how many conforming cleavages are required for each peptide, where "0" declares an unconstrained search, "1" requires at least one end of each peptide to conform to cleavage rules, and "2" requires both ends of each peptide to match a cleavage rule. Since extremely low mass and high mass peptides are unlikely to be useful for protein identification, researchers can stipulate a peptide mass minimum and maximum for the search. In addition, users can indicate that database protein sequences may include residues that are clipped off (such as initial methionines), allowing initial peptides from these proteins to begin at internal residues. Once the configuration is set, the software reads all spectra from a liquid chromatography run into memory (typically this will be on the order of 2000 spectra for an LCQ-series instrument). The spectra are sorted by precursor mass and are organized both as a linked list and as a balanced binary search tree for rapid retrieval. With these steps complete, the search can begin.

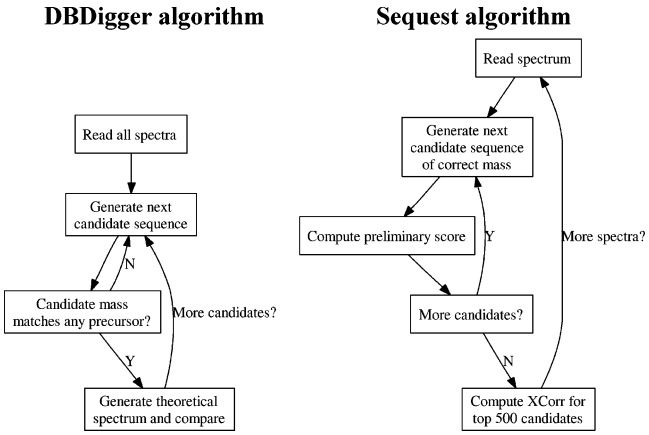


Figure 1. DBDigger reorganizes database identification so that candidate sequences are generated only once for each collection of spectra rather than once for each spectrum.

In DBDigger, candidate sequence generation is performed only once for the entire set of spectra rather than once for each spectrum (see Figure 1). Each candidate sequence is initiated at a particular position in the database sequence, extended by the succeeding residues in the sequence, compared to spectra from precursors within an acceptable mass range (usually 3 Da) of its calculated mass after each extension, and eliminated when its calculated mass exceeds that of the most massive precursor peptide for which a spectrum has been produced. As the algorithm passes through the sequence for a particular protein, these operations are managed for the set of candidate sequences initiated at prior positions. If posttranslationally modified forms of particular residues have been configured, the algorithm creates multiple copies of each candidate sequence in memory and extends one copy with each of the possible forms of the current residue. The number of modified residues for each candidate

(23) Thermo Finnigan Product Support Bulletin PSB 105.
(24) Strader, M. B.; VerBerkmoes, N. C.; Tabb, D. L.; Connelly, H. M.; Barton, J. W.; Bruce, B. D.; Pelletier, D. A.; Davison, B. H.; Hettich, R. L.; Larimer, F. W.; Hurst, G. B. *J. Proteome Res.* **2004**, *3*, 965−978.
(25) Hardy, S. J. S.; Kurland, C. G.; Voynow, P.; Mora, G. *Biochemistry* **1969**, *8*, 2897−2905.

Table 1. DBDigger Generation of a Candidate Sequences from Lysozyme[a]

sequence	mass	fate
WWCNDGRTPGSR	1435.57	no comparisons possible
NTDGSTDYGILQINSRWWCNDGRTPGSR	3171.39	compared to one +3
NLCNIPCSALLSSDITASVNCAK	2338.74	compared to seven +2s and four +3s
TPGSRNLCNIPCSALLSSDITASVNCAK	2837.27	no comparisons possible
NLCNIPCSALLSSDITASVNCAKK	2466.91	compared to two +2s and two +3s
TPGSRNLCNIPCSALLSSDITASVNCAK	2965.45	no comparisons possible
IVSDGNGMNAWVAWR	1676.90	compared to four +2s and two +3s
IVSDGNGM@NAWVAWR	1692.90	compared to two +2s and four +3s
KIVSDGNGMNAWVAWR	1805.07	compared to two +2s and four +3s
KIVSDGNGM@NAWVAWR	1821.07	compared to four +2s and three +3s

[a] When DBDigger generates a candidate sequence from the sequence database, several outcomes are possible. First, there may be no spectra from precursors near the candidate sequence's mass, and so no comparisons are possible for this candidate. Some candidate sequences can be compared to spectra from precursors of a single charge state. In these cases, a single theoretical spectrum is produced and compared to the experimental spectra. Other candidate sequences can be compared to experimental spectra of multiple charge states, and for these candidates, multiple theoretical spectra are produced and compared appropriately. The table shows consecutively produced candidate sequences from residues 62−112 of the chicken egg lysozyme sequence and their resulting comparisons. "M@" indicates an oxidized methionine residue.

sequence is tracked, and candidates are limited to contain a maximum number of modifications (usually three).

After generating a particular candidate sequence, the algorithm searches the spectra in memory to find those resulting from fragmentation of peptides within 3 Da of the candidate sequence, a process that is made efficient by the binary search tree organization of the spectra. The software creates a theoretical spectrum for the candidate sequence for each precursor charge state at which it will be compared (see Table 1). When more complex systems for predicting theoretical spectra are employed, this organization helps to minimize the impact of this process on performance. The modular structure of the program allows multiple systems for generating spectra from candidate sequences. For the purposes of this report, DBDigger was configured to use a simple spectrum generator. This system predicts only singly charged b and y fragment ions for spectra from singly and doubly charged precursors. Both doubly and singly charged fragments are predicted for spectra from triply charged precursors. After each theoretical spectrum is produced, it is simplified by merging peaks that are too similar in m/z value to be resolved by the mass analyzer. Single peaks replace these peak pairs at intermediate m/z values.

DBDigger was designed to support alternative systems to compare theoretical and experimental spectra. We present results for MASPIC, a novel statistical scorer that achieves significant accuracy and speed improvements over Sequest's cross correlation algorithm. For a complete technical description and justification of this scoring function, look for an upcoming article by Narasimhan et al.[22] The scorer makes use of extensive statistics generated during preprocessing for each spectrum to determine which candidate sequence matches are least likely to have resulted by random chance (see Figures 2 and 3). These statistics make use of three spectral principles:

1. Spectral Density Variation. Typically, existing statistical scorers assume that peaks are distributed randomly across the m/z range. MASPIC segregates each spectrum into a number of zones equal to the precursor ion charge state. This separates areas of high peak density from more sparsely occupied regions.

2. Intensity Classification of Fragment Ions. MASPIC stratifies ions into intensity classes for each spectral zone. The software accepts the smallest number of peaks into the most intense class, while larger numbers of peaks are accepted into less intense classes. As a result, it is less likely that a randomly matching peak be of a more intense class. Peaks with insignificant intensities are removed from consideration.

3. Mass Accuracy of Matching Peaks. The scorer allows for three possible classes of m/z error for each peak matching between theoretical and experimental spectra: a complete miss, a marginal hit, or a close hit. The algorithm computes the probability of randomly matching a peak of each intensity class with acceptable m/z error and with very low m/z error (see Figure 3).

This statistical evaluation of each spectrum is used to determine the probability that the overall match for each theoretical spectrum has occurred by random chance. The reported scores are higher for matches that are less likely to be random matches.

In addition to primary scores, DBDigger provides a useful secondary measure of each match. The eight most intense peaks are each assigned a power of 2; the most intense peak is assigned the weight 128, the second most intense is assigned 64, etc. The identification's secondary score is the sum of these weights for peaks that are matched between the observed and theoretical spectrum. For example, an identification that matches the most intense, second most intense, and fourth most intense experimental peaks would receive a score of $128 + 64 + 16 = 208$ out of a maximum of 255. At a glance, experienced users of the software can discern which of a spectrum's most intense eight peaks are matched by a particular identification.

DBDigger has been designed as part of a pipeline of mass spectrometry tools. Tandem mass spectra are extracted from instrument capture files into MS2 format unified spectrum files.[26] Spectra from precursors near the same m/z that contain similar fragment ions are grouped and replaced by spectra that average over the similar spectra.[27,28] The charge states of precursors for each spectrum are inferred by another algorithm under development.[29] Finally DBDigger reads in spectra from the processed

(26) McDonald, W. H.; Tabb, D. L.; Sadygov, R.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J.; Cociorva, D.; Yates, J. R., III. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 2162−2168.
(27) Beer, I.; Barnea, E.; Ziv, T.; Admon, A. *Proteomics* **2004**, *4*, 950−960.
(28) Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 2470−2477.
(29) Razumovskaya, J.; Fridman, T.; Day, R.; Uberbacher, E.; VerBerkmoes, N.; Gorin, A. *Proc. 52nd Am. Soc. Mass Spectrom. Conf.* 2004, MPE 074.

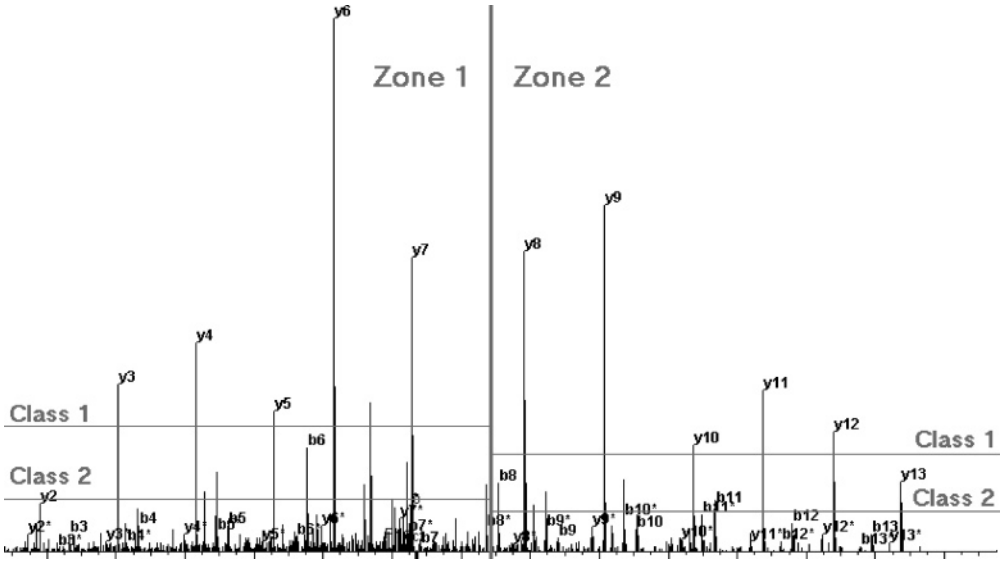


Figure 2. Spectrum representing the peptide MPCTEDYLSLILNR from bovine serum albumin in its doubly charged form. During preprocessing, the MASPIC scorer[22] separates each doubly charged peptide's tandem mass spectrum into two zones. The algorithm estimates the number of fragment ions expected to fall within each zone on the basis of the peptide's mass and charge state, and it uses this estimate to set an intensity cutoff for each zone that retains a number of observed peaks in that zone that is a multiple of the number of ions estimated to occur within it. The remaining peaks are stratified into classes on the basis of intensity, with each successive class holding twice as many peaks. This raw spectrum contains 423 peaks in the first zone and 292 in the second. Intensity filtering removed all but a fraction of these peaks from consideration in each zone. The remaining peaks were segregated into three classes on the basis of intensity.

MS2 and outputs its identifications as SQT-format unified identification files,[26] a format that enables the filtering and assembly of identifications through the DTASelect algorithm.[19] When a cluster of computers is available for identification, DBDigger separates the protein sequence database so that an equal part of it is processed by each member of the cluster. The same spectra are processed on each computer, but against different parts of the database. If the search is run on three computers, the first will match spectra against the first of each three sequence database entries, the second will draw its candidates from the second of each three protein sequences, and the third will compute candidates from the third of each three sequences. After these database searches are complete, the identifications from each computer are merged together to build a final identification file. This differs from the standard parallelization technique in which spectra are parceled out to individual computers for identification against the entire database.

Algorithm Configurations and Database Origins. Sequest and DBDigger were tested in both trypsin-specific and unconstrained modes. For tryptic peptide searches, cleavage was specified to take place after Lys or Arg but not before Pro. Sequest allows peptides to initiate at the first residue of the database sequence or at trypsin cleavage sites. DBDigger was configured to permit peptides to initiate at either the first residue or the second of the database sequence as well as the trypsin cleavage sites; this allows for more convenient searching of protein databases where N-terminal Met removal is anticipated.

The primary sequence database used for searches in this article consisted of several elements. The 4833 ORFs of the *R. palustris* proteome[30] were search targets for the ribosome search but were distractors during the protein standard mixture search. Contaminating proteins were added to this database, including 10 immunoglobulins, 12 proteases, and 14 keratin sequences. Finally, eight polypeptide sequences were added for the proteins in the standard mixture (alcohol dehydrogenase II was added to the database because this protein was sometimes observed in conjunction with the alcohol dehydrogenase I). The total count of proteins in this database was 4877, yielding a 0.16% chance of randomly hitting 1 of the 8 standard proteins.

The ribosomal sample[24] was previously analyzed by Sequest.[7] At first, no PTMs were included in the search configuration, and 317 proteins were identified by at least one high-scoring peptide; some contaminant proteins cosedimented with the 70S ribosomes during purification. These sequences were concatenated into a subset database that was used for the PTM search. In this study, the same subset database was used for the DBDigger PTM analysis. The subset database contains all of *R. palustris*' ribosomal proteins, but an identification has only a 17% chance of hitting a ribosomal protein randomly.

All database searches were run on a computer system incorporating an AMD Athlon 2500+ "Barton" processor. The computer featured 512 MB of RAM with a front side bus of 333 MHz. TurboSequest v27 and DBDigger were tested in the Microsoft Windows XP home operating system. Database indexing was not employed. Sequest allowed candidate peptides to contain zero or one missed tryptic cleavage sites in the tryptic peptide search. Identifications from the standard mixtures were considered to be true matches if their best-ranked sequences were part of the database sequences for one of the true proteins; this list included the eight polypeptides of the standard mixture, trypsin, or one of the keratins. A script was created in the *R* statistical environment[31] to automate finding the lowest score thresholds for identifications for peptides of each charge state that would

(30) Larimer, F. W.; Chain, P.; Hauser, L.; Lamerdin, J.; Malfatti, S.; Do, L.; Land, M. L.; Pelletier, D. A.; Beatty, J. T.; Lang, A. S.; Tabita, F. R.; Gibson, J. L.; Hanson, T. E.; Bobst, C.; Torres, J. L.; Peres, C.; Harrison, F. H.; Gibson, J.; Harwood, C. S. *Nat. Biotechnol.* **2004**, *22*, 55−61.
(31) Hornik, K. The R FAQ. http://www.ci.tuwien.ac.at/~hornik/R/ISBN:3-901167−51-X.

Class	Zone 1		Zone 2	
	hi-fi	lo-fi	hi-fi	lo-fi
1	0.4%	0.6%	0.3%	0.4%
2	0.8%	1.1%	0.6%	0.8%
3	2.3%	3.5%	1.7%	2.6%

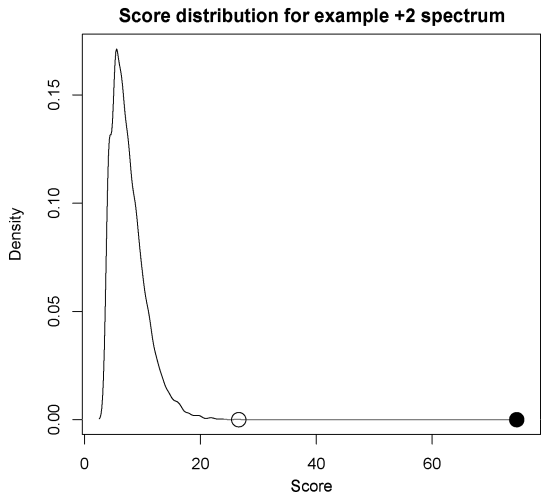


Figure 3. The MASPIC scorer attempts to match each peak from a theoretical spectrum to a peak in the experimental spectrum. During preprocessing, the algorithm determines the probability of randomly matching a peak of each intensity class for each zone at high fidelity (within 0.2 *m*/*z*) and low fidelity (more than 0.2 *m*/*z* but less than 0.5 *m*/*z*). These probabilities, shown in the table for the spectrum in Figure 2, are used as a lookup table to assess the significance of individual peak matches for a particular candidate sequence. The algorithm combines the probabilities into a score that is higher for matches that are unlikely to have occurred by random chance. A semitryptic identification of this spectrum generated 16 458 candidate sequences, almost all of which scored less than 20. The open circle, at a score of 26.63, represents the most highly scoring false match. The closed circle, at a score of 76.59, represents the single true match. A sequence receives a higher score when it matches to larger numbers of peaks, when many of those peaks are of the most intense class, and when it matches peaks at stringent *m*/*z* tolerance.

yield a mix of at least 19 true identifications for each false identification. Additional *R* scripts produced the included receiver operator characteristic (ROC) plots[32] and candidate sequence analyses. DTASelect[19] employed the thresholds reported by the *R* scripts in conjunction with the settings -a false -o -d 0 - -DB to produce the peptide counts displayed in the tables. In brief, these settings rejected identifications for which sequence ambiguity was possible (for example, Leu for Ile substitutions), eliminated degenerate protein identifications, removed the requirement that the best matching sequence score 8% better than the second best matching sequence, and produced files ready for import into a spreadsheet or database.

RESULTS AND DISCUSSION

In this study, three samples were used to characterize the performance of DBDigger. For training the scoring algorithm used in DBDigger, we employed spectra from a high-specificity trypsin digest of a mixture of six standard proteins. After algorithm development was complete, we digested the same mixture of proteins with a more traditional trypsin digestion to produce the test set. Finally, we employed spectra produced during a recent analysis of the *R. palustris* ribosome[24] to validate the software's PTM identification capabilities.

Protein Standard Mixture Training Set. A six-protein mixture was digested with trypsin under conditions that were seen to reduce nonspecific cleavage relative to standard overnight digests. A 2-h gradient of this training set produced 2401 tandem mass spectra. An assessment of spectral similarity followed by replacement of similarity sets with representative spectra reduced this count to 1638 spectra. Of these, 367 were assigned +1 precursor charge states. Of the remaining 1271 multiply charged spectra, 894 were assigned +2 precursor charge states, and 615 were assigned +3 precursor charge states.

The initial comparison between Sequest and DBDigger involved a search for tryptic peptides only. The Sequest search took 13 min, while the DBDigger search, employing MASPIC as the scorer, lasted only 23 s (see Table 2), though the algorithms processed the same numbers of spectra and produced the same numbers of identifications. Sequest compared an average of 547 candidate sequences to each spectrum, while DBDigger compared an average of 941. The primary scores for each algorithm were evaluated to determine the lowest cutoff scores in each precursor charge state for which at least 95% of the passing identifications were to proteins known to be in the mixtures. These thresholds (see Table 2) were used in DTASelect reports for the identifications. At 95% confidence, Sequest was able to identify 225 spectra while MASPIC identified 266, an 18% increase. Of the 270 spectra identified at this confidence by either algorithm, 4 were identified only by Sequest and 45 were identified only by DBDigger with the MASPIC scorer.

Next, the algorithms were compared for nonspecific searching. The Sequest search time doubled to 27 min, while DBDigger's search time increased to more than 32 min. Thresholds for 95% confidence were automatically assessed for the identifications. Because more candidate sequences were compared to each spectrum, the scores associated with the most distracting false answers increased, reducing the discrimination of the identifications. At the 95% confidence level, Sequest identified 205 spectra, while MASPIC identified 234, a 14% increase.

The time to process an unconstrained search rather than a tryptic search doubles for Sequest but increases by 85-fold for DBDigger with MASPIC as the scorer. The increased search time results from the much larger set of candidate sequences. Sequest employs two-stage scoring; all candidate sequences are evaluated by the preliminary scorer, but only the top 500 candidates are processed by the cross-correlation engine (see Figure 1). Sequest's prescoring is of no utility during database searches generating 500 or fewer candidate sequences per spectrum because every candidate sequence is subjected to both preliminary scoring and cross-correlation. Because unconstrained searching increases the number of candidate sequences by almost 2 orders of magnitude; however, the time required for prescoring the candidate sequences increases dramatically while the time required for cross-correlation is essentially unchanged. DBDigger processes every candidate sequence with its scoring function. Although the MASPIC scorer is much faster than cross-correlation, it is being employed on ~173 times as many spectra.

(32) Baker, S. G. *J. Natl. Canc. Inst.* **2003**, *95*, 511−515.

Table 2. Training Set Identifications at 95% Confidence[a]

algorithm employed	tryptic ends	time (min)	cand seqs	peptide IDs	spectral IDs	95% confidence cutoffs		
						plus1	plus2	plus3
Sequest	0	27	86295	175	205	2.968	2.485	3.348
Sequest	2	13	547	195	225	1.631	1.932	2.322
DBDigger	0	32:25	86297	211	234	29.31	33.37	44.95
DBDigger	1	6:37	16752	224	250	27.61	29.05	44.27
DBDigger	2	0:23	941	234	266	22.24	23.62	39.24

[a] A mixture of known proteins was analyzed by LC/MS/MS, producing 2401 spectra. These were grouped by similarity to reduce duplication, and likely precursor charge states for each spectrum were inferred. The resulting 1638 spectra were processed by DBDigger and by Sequest in both tryptic and unconstrained database searches. DBDigger, unlike Sequest, features a semitryptic search option. For each run, the cutoff scores for spectra of each charge state that produce at least a 19:1 ratio of true to false identifications were automatically determined; this corresponds to the 95% confidence level for each search. Because some peptides were evidenced by multiple spectra, the number of spectra identified differed from the number of peptides identified. The greater the average number of candidate sequences compared to each spectrum, the higher the false positive scores, requiring higher cutoffs for the same confidence.

Table 3. Training Set Identifications with PTMs at 95% Confidence[a]

PTM set	time (min)	cand seqs	peptide IDs	spectral IDs	95% confidence cutoffs		
					plus1	plus2	plus3
no PTMs	0:23	941	234	266	22.24	23.62	39.24
Nterm Acet, M+16	1:13	2530	243	274	22.24	24.41	40.88
Nterm Acet, M+16, Nterm Q-17	1:15	2617	247	277	20.13	25.22	41.72
Nterm Acet, M+16, Any Q-17	2:23	4177	247	276	22.19	26.02	42.42

[a] To demonstrate the value of context-dependent PTM searching, we used DBDigger to identify peptides with deamidated Gln residues, a feature know to occur predominantly at N-termini of peptides.. We included two other ubiquitous PTMs: N-terminal acetylation of peptides and oxidation of Met. Allowing Gln to lose ammonia for any occurrence of the residue required DBDigger to compare an average of 4177 candidate sequences per spectrum, but limiting these losses to Gln residues at peptide N-termini reduced this average to 2617 candidates and substantially accelerated the search.

To maximize the identifications achievable from the training set, DBDigger was configured to conduct tryptic peptide searches for several PTMs: acetylations at peptide N-termini, oxidations of Met, and ammonia loss from Gln. Alcohol dehydrogenase I and carbonic anhydrase II are known to bear acetylations at their protein N-termini. Oxidation of Met is commonly associated with sample handling.[33] Peptides with Gln at their N-termini can lose ammonia in a side-chain attack on the peptide terminus.[34] Incorporation of PTMs in searching changes the numbers of candidate sequences compared to each spectrum and highlights the importance of the context-dependent PTM searching introduced in DBDigger. The tryptic search with no PTMs examined an average of 941 candidates per spectrum (see Table 3). When Met residues were permitted to gain a mass of 16 Da and peptide N-termini were permitted to be acetylated, the candidate sequences climbed to 2530 per spectrum, with 274 spectra identifiable at 95% confidence. Two separate searches then examined the impact of context-dependent PTM searching. The first search added ammonia loss from Gln as a traditional PTM, while the second search permitted ammonia loss from Gln only when the residue was at the N-terminus of the peptide. The second, context-dependent PTM search resulted in substantially fewer candidates per spectrum. This search resulted in 5 identifications for peptides with ammonia-losing Gln at the N-terminus (see Supporting Information for an example), 10 identifications of oxidized Met residues, and 3 additional identifications for N-terminally acetylated proteins. An additional, internal, peptide was reported as being N-terminally acetylated, but this would appear to have resulted from a modified side chain (see Supporting Information). By reducing the numbers of candidate sequences compared to each spectrum, specifying required sequence contexts for PTMs can increase database search speed.

Protein Standard Mixture Test Set. Because MASPIC was optimized for the training set, a separate sample was used for validation. The same proteins were digested overnight with trypsin. Of the initial 2400 spectra, 1903 remained after grouping, almost 300 more than in the training set. Of the grouped spectra, 284 were found to have resulted from +1 precursors, and of the remaining spectra, 1293 were assigned as coming from +2 precursors while 705 were assigned as resulting from +3 precursors. Though the test sample incorporated the same proteins, the use of a different digestion protocol was observed to increase the number of cleavages after residues other than lysine and arginine, making this collection of spectra substantially different than the training set described above.

Table 4 compares DBDigger's results for this test set against those of Sequest. Using MASPIC, DBDigger achieved a considerable improvement upon Sequest's identifications in unconstrained (i.e., nontryptic) database searching, identifying 32% more spectra at the 95% confidence level. Of the 156 spectra identified by at least one algorithm, 3 were identified only by Sequest and 40 were identified only by MASPIC. When the algorithms were configured to identify only tryptic peptides, they performed similarly, with the new algorithm identifying only 3% more spectra. The utility of semitryptic searches, identifying peptides that result from tryptic

(33) Reid, G. E.; Roberts, K. D.; Kapp, E. A.; Simpson, R. J. *J. Proteome Res.* **2004**, *3*, 751−759.
(34) Krokhin, O. V.; Ens, W.; Standing, K. G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2528−34.

Table 4. Test Set Identifications at 95% Confidence[a]

algorithm employed	tryptic ends	time (min)	cand seqs	peptide IDs	spectral IDs	95% confidence cutoffs		
						plus1	plus2	plus3
Sequest	0	35	84850	110	116	2.212	3.147	3.827
Sequest	2	17	458	125	143	1.899	2.029	2.714
DBDigger	0	40:48	84854	145	153	28.66	36.42	48.36
DBDigger	1	8:16	16427	157	166	27.44	31.63	47.04
DBDigger	2	0:32	919	136	148	22.48	25.47	40.60

[a] Overnight trypsin digestion of the same standard proteins, followed by LC/MS/MS, produced 2400 spectra. Similarity grouping and charge-state inference retained 1903 spectra. Despite the larger number of spectra, fewer peptides were identified by either algorithm than in the training set. The utility of semitryptic searching for tryptic samples is demonstrated by the 166 spectra identified versus 148 for the tryptic-constrained search.

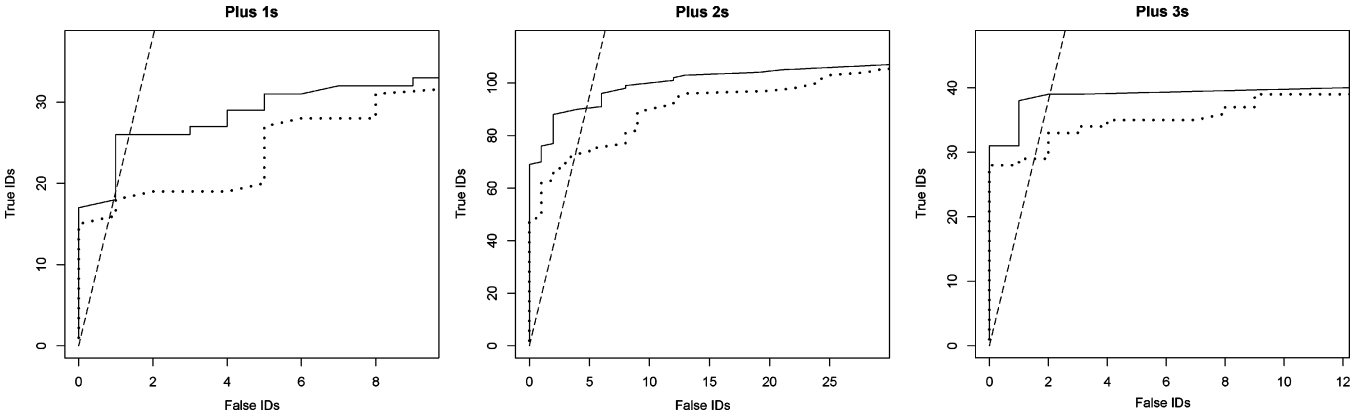


Figure 4. ROC plots demonstrate the improved accuracy available in the MASPIC scorer. The protein standard mixture test set was analyzed by Sequest and DBDigger in unconstrained (nontryptic) searches. Identifications were separated by the charge state of the precursor ions because this charge can affect identification effectiveness. A series of cutoff scores were applied to the identifications, and the numbers of true and false identifications above these thresholds were assessed. The solid lines represent MASPIC's results, and Sequest is shown by the dotted lines. DBDigger yields a larger number of true identifications with MASPIC at all points shown. The dashed diagonal in each graph intersects the curves at the 95% confidence level.

cleavages at one end and nonspecific cleavages on the other, is highlighted by DBDigger's performance in this mode; 8.5% more spectra were identified in the semitryptic search than in the nonspecific search at the 95% confidence level.

ROC curves were constructed for the nonspecific search of the protein standard mixture test set (see Figure 4). Because the number of identifiable spectra from this sample is unknown, the figures show numbers of true and false identifications rather than displaying sensitivity and specificity. These curves show the numbers of true and false identifications passing a series of score thresholds. Each threshold is represented by a point on a curve for the appropriate algorithm. An ideal scoring function separates true from false answers perfectly and would rise vertically until all true answers were included and then make a right angle for the lower scores of false answers. For one algorithm's curve to appear above another implies that the algorithm identifies a larger number of true identifications for the same number of false identifications. The curves in Figure 4 demonstrate that MASPIC exceeds Sequest's accuracy for spectra from singly, doubly, and triply charged precursors.

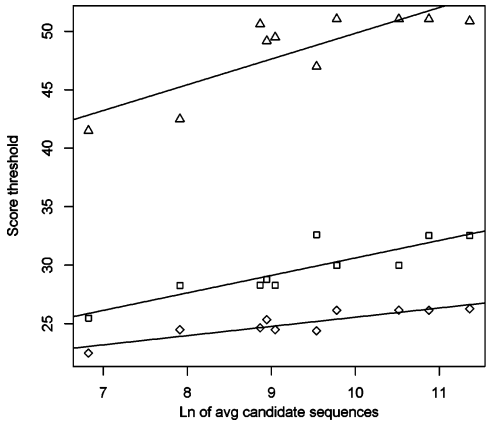
The test sample was used to estimate the effect of PTM searching on scoring thresholds. When PTMs are incorporated into a search, the number of candidate sequences compared to each spectrum can increase substantially, increasing the highest false score observed for each spectrum. PTMs can potentially increase the candidate sequences compared more than increasing sequence database size; database size increases cause a linear increase in candidate sequences compared, while each PTM added has a geometric effect on candidate sequences compared. The test set was subjected to a series of searches incorporating the following PTMs: N-terminal acetylation, N-terminal Gln deamidation, Met oxidation, Thr and Ser dehydration, and Lys and Arg methylation (see Figure 5). The average number of candidate sequences varied by 2 orders of magnitude from the search with no PTMs to the search with all of the above PTMs. Three line equations were derived from these PTM searches that allow estimation of 95% confidence thresholds for searches of unknown mixtures:

$$y_1 = 0.788x + 17.667$$

$$y_2 = 1.496x + 15.660$$

$$y_3 = 2.201x + 27.815$$

where y_n is a threshold for a particular charge state and x is the natural logarithm of the average candidate sequences compared to each spectrum. A line illustrating each of these projected thresholds is overlayed on the actual thresholds in Figure 5. An examination of the number of candidate sequences compared to each spectrum revealed that searches incorporating large numbers of potential PTMs compare many more candidates to high-mass peptides than to low-mass peptides (see Figure 6). By projecting



a	Q	M	T	S	K	R	Candidate Sequences
							919
X	X	X					2729
X	X	X	X				7093
X	X	X		X			7673
X	X	X			X		8482
X	X	X				X	13876
X	X	X	X	X			17640
X	X	X	X	X	X		36960
X	X	X	X	X		X	52856
X	X	X	X	X	X	X	85889

Figure 5. Sets of PTMs reveal threshold scaling. A series of DBDigger searches with increasing numbers of PTMs indicated the scaling of MASPIC's 95% confidence threshold scores in response to the number of candidate sequences evaluated. Each row in the table represents a different database search with PTMs. The symbols indicate which PTMs were included in each search: a, N-terminal acetylation; Q, N-terminal Gln deamidation; M, Met oxidation; T, Thr dehydration; S, Ser dehydration; K, Lys methylation; and R, Arg methylation. The average number of candidate sequences for a search with all of these PTMs is almost 2 orders of magnitude greater than for a search with no PTMs. In the accompanying graph, triangles represent 95% confidence scores for +3 precursor spectra, while squares represent thresholds for doubly charged precursors and diamonds represent cutoffs for singly charged precursors.

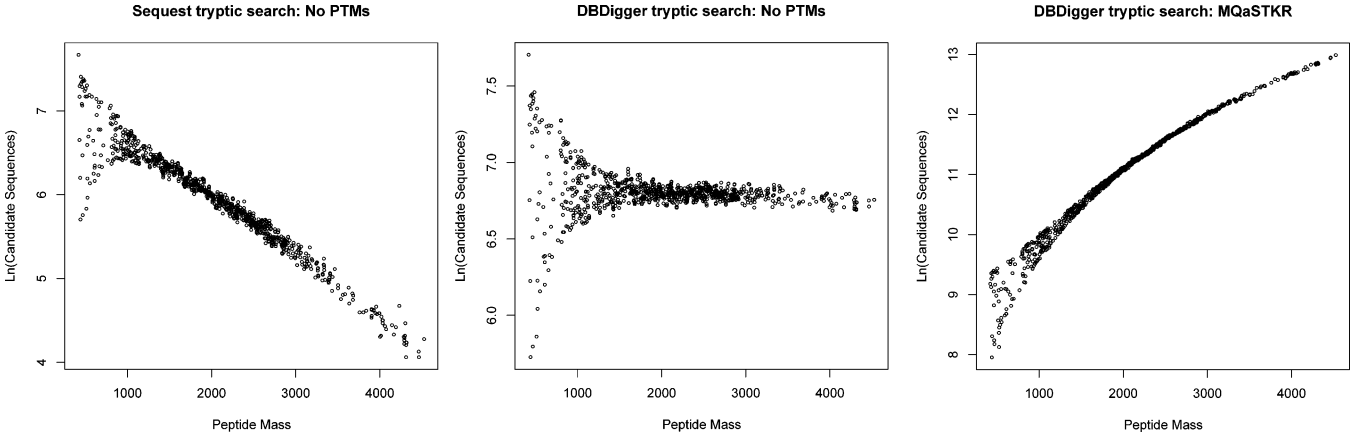


Figure 6. Precursor mass versus the natural logarithm of the number of candidate sequences compared to the spectrum. Each dot represents an identification, with the horizontal position indicating the precursor mass and the vertical position indicating the natural logarithm of the number of candidate sequences compared to the spectrum. When configured to require candidate sequences to be tryptic, Sequest and DBDigger show different tendencies as peptide mass increases. Sequest was configured to limit missed tryptic cleavage sites to one, and so more massive peptides were compared to fewer candidates than less massive peptides. DBDigger does not track missed cleavages. When PTMs are incorporated, the number of candidate sequences compared increases disproportionately for massive peptides. The title above each plot indicates which PTMs were permitted in each search (see Figure 5). In the most complex search shown, the most massive peptides are compared to roughly 70 times as many candidates as the least massive peptides. Because false identification scores increase in response to increased numbers of candidate sequences, this variation in candidate counts calls into question the common practice of applying a single cutoff score to all identifications of a particular charge state.

thresholds that reflect the average number of candidate sequences compared to each spectrum, these line equations aid users to select score thresholds that are appropriate to the search conducted. The variation in numbers of candidate sequences, however, makes it clear that, for aggressive PTM searches, scores for an individual spectrum should be evaluated in light of the number of candidate sequences compared.

Many protein sequence databases, especially those for eukaryotic proteins, contain entries that share common peptide sequences. As a result, a particular candidate sequence may repeat as a search algorithm proceeds through the database. This is most common for the shortest peptide sequences. DBDigger does not store a comprehensive list of candidate sequences already compared to each spectrum, instead storing only the sequences of the five best matches so far. As a result, DBDigger is unable to determine which new candidate sequences may be repeats of previously examined candidate sequences. While this reduces the efficiency DBDigger can achieve while searching redundant databases, it allows DBDigger to separate large databases into subsections for parallelization.

This parallelization strategy was evaluated with the protein standard test set. A semitryptic search identifying the full set of spectra against the entire database took 8:16 (minutes and seconds) on a single computer. This search was next conducted in eight separate processes. Each process identified the full set of spectra against a different eighth of the protein sequences. This led to some variation in the run times for each process; the quickest finished the search in 59.5 s, while the slowest search lasted 67.5 s. This inequity in search times is a factor to consider in launching such searches on multiple computers. Even though

Table 5. Processing Time Scales in Response to Spectral Count[a]

spectra searched	seconds elapsed	% of spectra	% of time
1903	496	100.0%	100.0%
951	308	50.0%	62.1%
475	200	25.0%	40.3%
237	120	12.5%	24.2%
118	93	6.2%	18.8%
59	45	3.1%	9.1%

[a] To measure the efficiency achieved by reorganizing database search, we constructed subsets of several sizes from the test set spectra Less time is required per identification as the number of spectra increases.

eight computers are all instructed to start a search simultaneously, their completion times are likely to vary. In this example, the eight processes took a total of 8:31 min to run. A separate algorithm merged these eight result files together in 3 s. The total amount of processing time used was 8:34 min for an eight-way parallelization, adding an overhead of 18 s to the single process search time.

The process of scoring matches can be separated from the processes of reading spectra, generating candidate sequences, and constructing theoretical spectra by running DBDigger with no scoring function. Unsurprisingly, the time required for processes other than scoring also increases when larger numbers of candidate sequences are to be compared. Three searches without scoring were conducted with the full set of spectra in the test sample being compared to the complete database in tryptic, semitryptic, and unconstrained database searches. The times required for these searches without scoring were 0:19, 4:25, and 21:46 min, respectively. If these times are compared to those in Table 4, it becomes plain that processes other than spectral match scoring can be significant in determining the amount of time required for identification.

To test the scaling efficiency of DBDigger's reorganization, we created spectral subsets of varying sizes and identified them by semitryptic search (see Table 5). In this test, identifying half the spectra took 62% as long as identifying all the spectra. As the number of spectra being identified increases, the effective number of identifications per second increases. As rapidly scanning linear ion traps become more common in proteomic laboratoriess, the number of spectra requiring identification will increase. DBDigger's organization for database search reduces the computational requirements for keeping pace with these new instruments.

Ribosomal Posttranslational Modification Search. Spectra were produced via 2D LC/MS/MS from a sample highly enriched in the ribosomal proteins of *R. palustris* and surveyed for protein content and PTMs.[24] Because of limitations in Sequest, multiple PTM searches were necessary to evaluate this sample for single, double, and triple methylations of Lys and Arg. DBDigger was configured to search for a variety of PTMs in a single pass, using the same ribosomal sequence database subset used in the initial publication; the search included single, double, and triple methylations of Lys and Arg, β-methylthiolation of Asp, deamidation of N-terminal Gln residues, and oxidation of Met residues. The PTM search against the 317 protein database produced an average of 32 579 candidate sequences for each spectrum and took 290 min to run on all 14 LC/MS/MS runs, containing a total of 19 026 spectra after spectral similarity grouping. Because a large number of PTMs were in consideration, the amount of memory required for the search quintupled with respect to the amount required for this search without PTMs; the inclusion of many PTMs geometrically increases the number of candidate sequences that must be retained simultaneously during candidate sequence production.

Comparison of PTMs from the initial publication of the ribosome data and from the DBDigger search is complicated by several factors. First, the published PTMs were required to be supported by evidence from two of three possible sources: 1D LC/FTICR MS of intact proteins, 1D LC/MS/MS of peptides, and 2D LC/MS/MS of peptides. The DBDigger analysis includes only the 2D LC/MS/MS results. Second, the published analysis stemmed from a series of unconstrained (nontryptic) Sequest searches, with each seeking a particular type of PTM. The DBDigger search included only tryptic peptides, and the single search encompassed the full set of modifications. The initial publication enumerated a methylation on RRP-L3, two separate isoforms of of RRP-L7/L12, an acetylation or trimethylation on RRP-L11, N-terminal methylations of RRP-L30 and RRP-L33, and a β-methylthiolation on RRP-S12.

RRP-L3 was initially found to carry a methylation on K155 or K158 based on an top-down protein intact mass and a single-methylation Sequest search. DBDigger identified peptides matching these PTMs, but their scores were insufficient to pass the 95% confidence criterion. Its most confident PTMs were four spectra identifying a peptide carrying a pair of double methylations, one at R205 and the other at K209 (or possibly K210), with the same peptide appearing in an unmodified form as well. An examination of the Sequest search for double methylations also reveals this pair of modifications. Because these PTMs disagree with the top-down mass found for this protein, these PTMs had previously been rejected. Since Sequest and DBDigger agree on the double methylation of RRP-L3, greater confidence can be placed in this PTM identification.

The Sequest results for RRP-L7/L12 were evaluated as revealing two isoforms. The first isoform incorporated a dimethylation at K69 and a methylation at K86, while the second isoform was trimethylated or acetylated at both K86 and K89. DBDigger's report for this protein suggests a variable single methylation at the N-terminus or K6 (some peptides are unmodified while others carry the PTM), single or double methylation at K69, and single methylation at K86 (all observations of this residue bore the PTM). Apart from the methylation of the N-terminal peptide, DBDigger's identified PTMs are consistent with Sequest's for this protein. Assessing the Sequest results for this protein was troublesome because results from three different runs had to be collated; discerning whether a combination of single methylations on a particular peptide or a smaller number of multiple methylations was the better identification led to many possible interpretations of the Sequest results. By incorporating all of the PTMs as part of a single search, DBDigger made it unnecessary to collate multiple search results, and the analysis for this protein became more straightforward.

For RRP-L11, Sequest identified a triple methylation or acetylation at K40. DBDigger backs up this claim with two different

peptides (differing by a Met oxidation), and no other methylations were reported for this protein. The same can be said of RRP-L30 and RRP-L33, where Sequest and DBDigger agree that the only methylations are to be found at the N-terminus or K3 (the spectra do not differentiate between these two possibilities).

The unusual β-methylthiolation on D88 of RRP-S12 posed a more difficult challenge. The most convincing spectra revealing this modification came from the initial publication's 1D LC/MS/MS run rather than the 2D data being reexamined here. Still, these spectra do contain evidence for this modification; the sequence VKD*LPGVRYHILR (where $*$ represents the β-methylthiolation) is identified twice from +3 precursors and once from a +2 precursor. The doubly charged identification's MASPIC score of 23.1 was insufficient to pass the 95% confidence criterion, and the better score for the spectra from triply charged precursors (43.8) was also insufficient for this level of confidence.

Important lessons can be learned from comparing the published, Sequest-based results for the ribosome sample to the new DBDigger results. First, proteomic research can benefit from second opinions; when two algorithms agree that evaluate matches by completely different techniques, one can feel greater confidence in identifications of peptides and PTMs. Second, integrating multiple PTM searches into a single search makes it much easier to analyze the resulting identifications. DBDigger can achieve both these aims.

CONCLUSION

DBDigger can be faster than Sequest for two reasons: candidate sequences need be generated only once, and DBDigger scores theoretical spectra against experimental spectra with the rapid, single-pass MASPIC scorer rather than using both a fast preliminary scorer and a slower cross-correlation comparison. Algorithms using prescoring and final scoring cannot easily be reorganized in this way, but systems such as OMSSA,[8] PepProbe,[9] and Luck[10] could benefit from adopting this structure. In brief, it is faster to determine which spectra can be compared to a particular candidate sequence than it is to determine which candidate sequences should be compared to a particular spectrum.

The second contribution DBDigger can make is the idea of context-dependent PTM seaching. By restricting the occurrence of PTMs by their locations in the peptide sequence or by the flanking amino acid residues, far fewer candidate sequences are generated than for traditional PTMs. The analysis of CNBr digests illustrates the utility of this feature. Cyanogen bromide cuts proteins at the peptide bond C-terminal to Met residues, and Met is modified to become homoserine or homoserine lactone as part of this cleavage. If one uses Sequest to search for these PTMs, the algorithm is free to insert modified Met residues in the middle of peptide sequences, even though the PTM is known to occur only at the Met where CNBr cleaves. Other PTMs also exhibit such specificity, such as Gln's loss of ammonia when at the peptide N-terminus or Asn's loss of ammonia when it is N-terminal to Gly.[34] Algorithms that implement context-dependent PTMs can substantially reduce the number of candidate sequences produced, thus accelerating the search (see Table 3) and potentially increasing discrimination.

While the literature in proteome informatics has emphasized systems for scoring theoretical spectra against experimental ones, we believe that an increased emphasis on the framework in which these scorers rest can also benefit the field. We hope to have provided sufficient detail for DBDigger's reorganization and context-dependent PTM searching to be reproduced by others. Besides the two features highlighted above, DBDigger's design incorporates many other features of great value: the use of unified file formats for spectral input and identification reporting, the ability to process multiple isotopic enrichments and multiple PTMs in a single search, the simplicity of using multiple computers for identification, the emphasis on modularity, etc. While many of these features require relatively simple software engineering to implement, their value for practical proteomics is potentially enormous.

ACKNOWLEDGMENT

The authors acknowledge funding from the United States Department of Energy's Office of Biological and Environmental Sciences through the Genomics: GTL Center for Molecular and Cellular Systems. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract DE-AC05-00OR22725. D.L.T. thanks the organizers of the NIH/NIST/ISB Peptide Fragmentation and Identification Workshop for providing a forum for many fruitful discussions. Nathan VerBerkmoes provided valuable assistance in the production of the ribosomal spectra. We thank Gregory B. Hurst and W. Hayes McDonald for critical reading of the manuscript.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 31, 2004. Accepted January 24, 2005.

AC0487000