Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries

Barbara E. Frewen,[†] Gennifer E. Merrihew,[†] Christine C. Wu,[‡] William Stafford Noble,[†,§] and Michael J. MacCoss*[,†]

Department of Genome Sciences and Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, and Department of Pharmacology, University of Colorado Health Sciences Center, Aurora, Colorado 80045

A widespread proteomics procedure for characterizing a complex mixture of proteins combines tandem mass spectrometry and database search software to yield mass spectra with identified peptide sequences. The same peptides are often detected in multiple experiments, and once they have been identified, the respective spectra can be used for future identifications. We present a method for collecting previously identified tandem mass spectra into a reference library that is used to identify new spectra. Query spectra are compared to references in the library to find the ones that are most similar. A dot product metric is used to measure the degree of similarity. With our largest library, the search of a query set finds 91% of the spectrum identifications and 93.7% of the protein identifications that could be made with a SEQUEST database search. A second experiment demonstrates that queries acquired on an LCQ ion trap mass spectrometer can be identified with a library of references acquired on an LTQ ion trap mass spectrometer. The dot product similarity score provides good separation of correct and incorrect identifications.

Shotgun proteomics has emerged as a robust and sensitive approach to profile the protein complement in a complex biological sample.[1] In this approach, a sample for analysis is prepared by digesting a protein mixture with proteases to yield a mixture of peptides. The peptides are then loaded onto a microcapillary chromatography column in-line with a mass spectrometer. Tandem mass spectra are acquired data-dependently as peptides are eluted off the column, ionized, and emitted into the mass spectrometer.[1,2] Finally, the identity of a peptide in the mixture is determined by comparing an acquired tandem mass spectrum to predicted spectra generated from amino acid sequences drawn from a database—an approach known as database searching.[3] This shotgun approach for identifying proteins in mixtures is extremely powerful and makes possible the characterization of thousands of proteins from a single 24-h mass spectrometry run.

Database searching has been invaluable for automating the characterization of uninterpreted tandem mass spectra and facilitating high-throughput proteomics; however, there remains room for improvement. First, database search algorithms make assumptions about how peptides fragment within the mass spectrometer. Although these assumptions allow for the correct identification of most peptides, better prediction of peak intensities could accommodate the identification of more peptides and introduce fewer false identifications. Second, most algorithms analyze protein modifications by iterating over all possible combinations of modified and unmodified residues in a peptide sequence. While effective, this approach is slow and impractical for considering multiple modifications for every sample analyzed. A search algorithm that made use of previously characterized mass spectra could address both of these limitations.

This type of approach has been used successfully for other types of mass spectra. For example, electron impact mass spectra are often identified based on comparisons to other experimentally generated spectra of known identity. These comparisons are robust largely because the conditions for acquiring the data have been standardized, making results highly reproducible across different laboratories and instruments. As a result of this reproducibility, large curated reference libraries have been developed for use by the general mass spectrometry community. Similarly, the development of normalized collision energy on quadrupole ion trap mass spectrometers has led to standardization in collecting spectra from collision-induced dissociation (CID). The reproducibility of CID spectra has facilitated the use of library searching for characterizing small molecules and peptide tandem mass spectra. One implementation, LIBQUEST,[4] uses a cross correlation to measure similarity between spectra. Alternatively, the dot product has been successfully used to find similar peptide spectra in clustering algorithms[5,6] and is also used in the search software

* To whom correspondence should be addressed. E-mail: maccoss@gs.washington.edu. Voice: 206-616-7451.Fax: 206-616-7301.

[†] Department of Genome Sciences, University of Washington.

[‡] Department of Pharmacology, University of Colorado Health Sciences Center.

[§] Department of Computer Science and Engineering, University of Washington.

(1) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd. *Nat. Biotechnol.* **1999**, *17*, 676−682.

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

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

(4) Yates, J. R., III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557−3565.

10.1021/ac060279n CCC: $33.50
Published on Web 07/15/2006

for the NIST library, which has recently been extended to include peptide spectra.

Just as database searching is limited to available protein sequences, library searching is limited to previously characterized spectra in an available library. High-throughput sequencing and protein prediction algorithms have provided ample protein sequence for database searches. Likewise, high-throughput proteomics experiments produce abundant spectra, which can be used as references. Modern mass spectrometers are capable of acquiring tandem mass spectra at a rate of 3−5 Hz resulting in 250 000−430 000 spectra per day. Rapid acquisition combined with computer clusters for database searching[7,8] allow even small laboratories to quickly amass a vast number of characterized spectra suitable for inclusion in a reference library. Furthermore, there already exists a rich supply of characterized peptide mass spectra in public data sets that can easily be assembled into publicly available libraries for peptide identification. The viability of using previously identified peptides as the basis of a search has already been demonstrated with the Proteotypic Peptide Profiling (P3) program.[9] As with database searching algorithms, P3 stores peptide sequences to use for identifying spectra, and like a spectrum library, it limits the basis of the search to previously observed peptides.

We have developed a method for assembling and searching tandem mass spectrum libraries. The libraries are stored in a compact binary format that is easily updated and shared between laboratories. In this report, we assemble a large library and demonstrate that it can be searched more efficiently than current database searching programs with no loss of sensitivity or specificity. Additionally, we identify previously characterized posttranslationally modified peptides without increasing the analysis time. Finally, we demonstrate that the library search can be used to compare spectra acquired on different instrument models in different laboratories. The spectrum libraries and the software for compiling and searching libraries has been made freely available for noncommercial use. Information about obtaining both can be found at http://proteome.gs.washington.edu.

METHODS

Protein Sources. Peptide mass spectra were acquired using protein from two sources, *Caenorhabditis elegans* and *Escherichia coli* cell lysates. Protein extracts from whole-worm lysate of *C. elegans* were biochemically fractionated by either solubility, density, charge, hydrophobicity, or molecular weight. Each method yielded 9−10 fractions, and each fraction was digested to peptides with trypsin. Soluble proteins from *E. coli* lysate were separated by centrifugation and digested with trypsin.

Chromatography and Mass Spectrometry. *C. elegans* MS/MS spectra were acquired on an LTQ linear ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) using either a 12-step multidimensional protein identification technology (MudPIT) protocol[10] or one-dimensional reversed-phase chromatography as described below. Samples were injected onto the column with an Agilent 1100 quaternary HPLC (Palo Alto, CA). Peptides eluting off the column electrosprayed directly into an LTQ mass spectrometer, and MS/MS spectra were acquired using data-dependent acquisition.

E. coli spectra were acquired on both an LTQ and an LCQ-XP Max ion trap mass spectrometer (ThermoFinnigan). The LTQ data were acquired identically to the *C. elegans* data. For the LCQ data, samples were loaded onto a single-phase (C18) microcapillary column and electrosprayed directly into the mass spectrometer.

BiblioSpec Software. All software were written in the C++ programming language and compiled on a Linux operating system. The software package BiblioSpec consists of several independent programs, each of which is described below. The three main programs are BlibBuild, which creates a spectral library of mass spectra, BlibFilter, which modifies an existing spectrum library to contain only one spectrum per peptide, and BlibSearch, which matches query spectra to library spectra.

Libraries are constructed from peptide MS/MS spectra that have previously been matched to peptide sequences, for example, using a database search program (e.g., SEQUEST,[3] Mascot[11]). The user prepares a list of spectra to be included in the library with the charge and sequence assigned to each as well as the file where the spectrum can be found. BlibBuild takes this list as input and extracts peak and precursor mass data from MS2[12] files. The collected spectra are sorted by precursor mass, and the new library is written to a binary file.

If a library contains multiple spectra for some peptide ions, an additional step can be taken to select the best representative among these redundant spectra. BlibFilter takes such a library and measures the similarity of all pairs of spectra for a given peptide ion. The spectrum with the highest average similarity score is chosen for inclusion in the filtered library.

The library searching function of BiblioSpec is performed by the program BlibSearch. It takes as input a library file and an MS2 file containing query spectra. The search begins by loading the library into memory. A query spectrum is read from the MS2 file and preprocessed as described in Results and Discussion. A binary search on the library returns candidate spectra whose precursor m/z is the same as that of the query spectrum within a specified tolerance. Each candidate library spectrum is preprocessed and compared to the query using a dot product. The matches are sorted by score, and the best library matches for each query spectrum are reported in a text file in SQT file format.[12]

Library Construction. The peptide identities of spectra used in the libraries were based on SEQUEST search results and stringent scoring criteria. All *C. elegans* spectra were searched using a database containing all predicted proteins from the Wormbase 130 freeze,[13] the *E. coli* proteins in RefSeq,[14] several

[5] Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 2470−2477.

[6] Tabb, D. L.; Thompson, M. R.; Khalsa-Moyers, G.; VerBerkmoes, N. C.; McDonald, W. H. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 1250−1261.

[7] Sadygov, R. G.; Eng, J.; Durr, E.; Saraf, A.; McDonald, H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1* (3), 211−215.

[8] Duncan, D. T.; Craig, R.; Link, A. J. *J. Proteome Res.* **2005**, *4*, 1842−1847.

[9] Craig, R.; Cortens, J. P.; Beavis, R. C. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1844−1850.

[10] Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001**, *73*, 5683−5690.

[11] Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.

[12] McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., III. *Rapid Commun, Mass Spectrom.* **2004**, *18*, 2162−2168.

[13] Wormbase 130 freeze; http://ws130.wormbase.org.

[14] Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res.* **2005**, *33*, D501−D504.

common contaminants, and a randomized version of each real protein for use in estimating false positives.[15,16] A SEQUEST identification was deemed correct if it met the following criteria: normalized Xcorr was at least 0.35, delta CN was at least 0.12, peptide length was at least seven residues, at least 30% of predicted b/y ions were accounted for within the spectrum, the peptide was fully tryptic, and at least one other peptide from the same protein was identified. Sets of duplicate spectra (i.e., those of the same peptide ion) were filtered by selecting the one with the highest average dot product when compared to the others in the set. For instances of two spectra per ion, the selection was random.

Spectra from each fractionated *C. elegans* sample were put into a library. Over 6 million spectra were searched, and 366 400 met the above threshold criteria for correct identification. An additional 183 spectra had two matches above the threshold, each to a peptide in a different charge states, and were not included in the dataset. There were 51 identifications made to randomized peptide sequences, suggesting that there are ~51 false positives in the library. After filtering duplicates, the library contained 26 708 spectra representing 21 264 unique peptide sequences. These sequences come from 3573 different proteins.

A second library was constructed from the *E. coli* spectra collected on the LTQ. Peptide identifications were made with SEQUEST using a database containing all *E. coli* proteins, common contaminants, and a randomized version of each real protein. The same criteria used for the *C. elegans* identifications were applied to the *E. coli* results with two exceptions. The minimum normalized Xcorr was relaxed to 0.3 and minimum delta CN was 0.1. Over 200 000 spectra were searched and 40 521 spectra were identified. These were filtered to yield 8451 spectra representing 6864 peptide sequences and 1177 proteins. There were also 23 matches to randomized sequences.

The third library consisted of the *C. elegans* library plus seven spectra that were identified as having posttranslational modifications. SEQUEST was used with the differential modifications option to search for oxidized methionine and phosphorylated serine on a set of spectra acquired from an unfractionated *C. elegans* sample. The peptide identifications were manually verified. Four spectra had a single oxidized methionine, two had a single phosphorylation, and one spectrum had two phosphorylated residues.

Query Test Sets. One test set of query spectra was taken from the unfractionated *C. elegans* sample not used in the library. Peptide identifications were made with SEQUEST using the same sequence database as above and using similar threshold criteria. The criteria differed in that the minimum Xcorr and delta CN were relaxed to 0.3 and 0.1, respectively. The query set contained 14 925 spectra from 5358 different peptide ions (907 of which are not present in the library). There were also two matches to randomized sequences. These peptides come from 1261 proteins.

A second test set was taken from the *E. coli* spectra acquired on the LCQ. A SEQUEST search with the same database used for the library and with the same criteria as for the other query set identified 924 spectra from 353 different precursor ions. There were no matches to randomized sequences with these criteria.

RESULTS AND DISCUSSION

We have developed a method for constructing and searching peptide mass spectrum libraries for high-throughput proteomics. One goal for such libraries is to serve as a convenient means for laboratories to share proteomics data. Therefore, the libraries are designed to be compact, easily assembled, and quickly updated. The libraries should also provide a searchable set of reference data for assigning peptide sequences to uncharacterized tandem mass spectra. A search produces a list of candidate matches for each query, with an associated score measuring similarity between the matched spectra. The search algorithm is designed to maximize speed while retaining good sensitivity and specificity.

The library format was structured with size and search efficiency in mind. To minimize disk space, we implemented a binary data storage format rather than a text format (e.g., MS2, mzXML). A library of 18 906 spectra takes up 111M of disk storage: 33% of a conventional MS2 file and 56% of an mzXML file of those same spectra. BiblioSpec software supports adding spectra to an existing library, merging libraries, and viewing library contents in a text format, making it easy to incorporate data from different experiments and different laboratories.

Our criteria for a successful library searching algorithm are that it be fast, return a majority of the identifications that a database search would, and that it assign scores with good discrimination between correct matches and false positives. The speed of the search was optimized by limiting the number of spectrum comparisons made, by finding candidate spectra quickly, and by using an efficient comparison technique. Comparisons were limited to library spectra with a precursor *m/z* within 3 *m/z* of the query spectrum and with the same charge state. The library stores spectra ranked by precursor *m/z* so that candidates can be efficiently located with a binary search. The dot product metric (sometimes referred to as the spectral contrast angle[17]) employed for comparing two spectra is more computationally efficient than a cross correlation and still provides a reliable measure of similarity.[17,18]

Peak Preprocessing and Parameter Optimization. The inherent noise and variability of spectra is accounted for by making adjustments to spectrum peaks prior to calculating similarity scores. Since there is no standardized method of peak preprocessing, we tried several combinations of reported normalization steps to learn what works best for our data set. However, the BiblioSpec software allows the user to modify these parameters as desired. Adjustments fall into three categories: binning peaks by *m/z*, normalizing peak intensities, and removing low-intensity noise peaks. The binning method is unvaried and involves merging peaks into bins of 1 *m/z* by summing the intensities of multiple peaks falling into the same bin.

We tried three different methods of peak normalization and two different methods of removing noise peaks. Peak normalization was done variously by taking the square root of peak intensities (SQRT),[5] by weighting the square root of the intensity by the square of the peak *m/z* (SMZ),[17] or by separating the peaks into 10 bins equally spaced across the spectrum's range of peak *m/z*'s and dividing the intensity of each peak by the maximum peak intensity for its bin (BIN).[3] To remove noise from a spectrum,

[15] Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (4), 378−386.

[16] Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43−50.

[17] Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.

[18] Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (1), 85−88.

Table 1. Results from Different Spectrum Preprocessing Methods[a]

intensity[b]		noise reduction	order	ROC score
SMZ		top 50	IF	0.991 834
SMZ		top 100	NF	0.991 486
SMZ		FC	IF	0.988 697
SMZ		top 200	NF	0.988 194
BIN	10	top 100	NF	0.988 087
SMZ		top 100	IF	0.987 283
SMZ		top 200	IF	0.986 109
SMZ		top 50	NF	0.985 885
SMZ		top 300	NF	0.985 609
BIN	10	top 200	NF	0.985 278
SMZ		top 300	IF	0.983 804
BIN	10	top 50	IF	0.982 509
BIN	10	FC	IF	0.981 097
SQRT		top 50	NF	0.980 661
BIN	10	top 100	IF	0.980 295
BIN	10	top 300	IF	0.978 778
SQRT		top 100	NF	0.978 732
BIN	10	top 200	IF	0.977 687
BIN	10	top 50	NF	0.976 924
BIN	10	top 300	NF	0.976 626
SQRT		top 300	NF	0.976 08
SQRT		FC	IF	0.975 63
SQRT		top 200	NF	0.975 081
BIN	10	FC	NF	0.963 484
SMZ		FC	NF	0.946 539
SQRT		FC	NF	0.944 237

[a] The columns from left to right are the type of peak intensity normalization, the type of noise reduction, the order in which the two steps were performed, and the area under the ROC curve. [b] Intensity normalization was done by taking the square root of the peak intensities (SQRT), by weighting the square root of the intensity by the square of the peak *m/z* (SMZ), or by dividing the peaks into 10 bins and dividing each peak by the maximum-intensity peak in the bin (BIN). To separate noise from signal, the peaks were ranked by intensity and either a fixed number of the highest intensity peaks were retained (top) or the highest intensity peaks that summed to a fraction of the total ion current (FC) (in this case, a half) were retained. These steps could be done in either order, normalizing intensity first (IF) or removing noise first (NF).

the peaks were ranked by intensity, and two different metrics were used to define the cutoff between the high-intensity signal peaks and the low-intensity noise peaks. The cutoff was either based on the number of high-intensity peaks to retain (TOP *n*)[3] or on the cumulative intensity equaling some fraction of the total ion current (FC *x*).[5] We chose to consider TOP *n* where *n* equaled 50, 100, 200, or 300 and FC *x* where *x* equaled 0.5. These two steps could be completed in either order, peak intensity normalization first (IF) or noise removal first (NF).

Twenty-five different combinations of preprocessing steps were compared on a search of a small library and test set, and the area under a receiver operator characteristic (ROC) curve was calculated for each (Table 1). The top two methods (SMZ-TOP50-IF and SMZ-TOP100-NF) had calculated areas of 0.992 and 0.991. Given the closeness of these scores, we chose to use the second-ranked method for the remaining searches with the belief that using 100 instead of 50 peaks would facilitate the identification of longer peptides.

Filtering Comparison. We examined three different ways of choosing a representative from multiple spectra for the same peptide ion. The representative spectrum was chosen as the one with the highest total ion current, the highest Xcorr obtained from a SEQUEST search, or the highest average dot product when compared to all other spectra of the same ion. Three different filtered libraries were generated from the same set of spectra, each selecting among redundant spectra in a different way. All 25 methods of preprocessing were tested on each library. The top four scoring methods are summarized in Table 2. Dot product filtering gave the best results and was the method used for the libraries used in further experiments.

Table 2. Comparison of Library Filtering To Remove Duplicate Spectra[a]

intensity[b]		noise reduction	order	ROC score
		Dot Product Filtered		
SMZ		top 50		0.990 453
SMZ		top 200	NF	0.987 918
SMZ		top 100		0.987 876
SMZ		top 100	NF	0.986 199
		Ion Current Filtered		
SMZ		top 50	NF	0.964 234
SMZ		top 100	NF	0.953 969
BIN	10	top 50	NF	0.952 338
SMZ		top 50		0.950 816
		Xcorr Filtered		
BIN	10	top 50	IF	0.982 509
BIN	10	top 100	IF	0.980 295
BIN	10	top 200	NF	0.985 278
BIN	10	top 100	NF	0.988 087

[a] A spectrum library was filtered using three different methods of selecting the representative spectra. The three resulting filtered libraries were searched using various preprocessing methods. The top four preprocessing methods for each type of filtering are presented. The first set of results are from the library filtered by dot product, the second from the library filtered by total ion current, and the third are from the library filtered by Xcorr. The columns from left to right are the type of peak intensity normalization, the type of noise reduction, the order in which the two steps were performed, and the area under the ROC curve. [b] Intensity normalization was done by taking the square root of the peak intensities (SQRT), weighting the square root of the intensity by the square of the peak *m/z* (SMZ), or by dividing the peaks into 10 bins and dividing each peak by the maximum-intensity peak in the bin (BIN). To separate noise from signal, the peaks were ranked by intensity and either a fixed number of the highest intensity peaks were retained (top) or the highest intensity peaks that summed to a fraction of the total ion current (FC) (in this case, a half) were retained. These steps could be done in either order, normalizing intensity first (IF) or removing noise first (NF).

Search Verification. Ideally, a search with a comprehensive library of tandem mass spectra will produce as many correct identifications as a database search. To test this ability, we used SEQUEST to find peptide identifications for a set of spectra and searched a library for identifications to the same spectra to see if BiblioSpec could reproduce the SEQUEST assignments. BiblioSpec found 91% (13 591) of the peptide identifications assigned by the database searching program, SEQUEST, and 93.7% of the proteins. The BiblioSpec scores ranged from 0.067 to 0.989 (Figure 1). For most of the BiblioSpec identifications that did not agree with the SEQUEST results, the respective peptide spectrum was not present in the library. Only 7% (98) of the 1334 incorrectly identified spectra had the correct spectrum in the library.

We used the SEQUEST peptide identifications as the standard and considered BiblioSpec identifications that agreed with the SEQUEST scores to be correct and those that disagreed to be incorrect. To judge how well the BiblioSpec score distinguishes between correct and incorrect matches, we plotted a ROC curve and calculated the area under it to be 0.978 (Figure 1). By setting an appropriate cutoff value for the similarity score, we can identify most of the correct matches without introducing many false

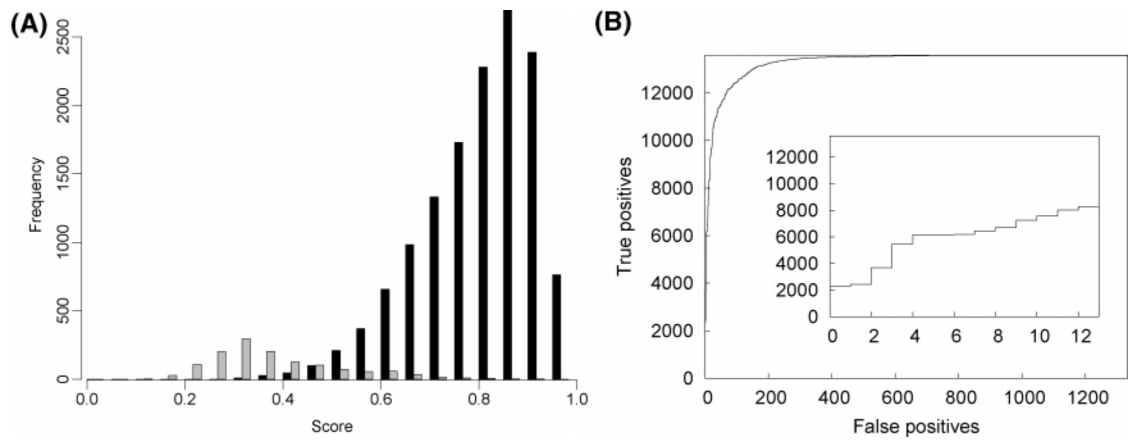

Figure 1. Discrimination between correct and incorrect peptide spectrum matches. The same data are shown here in two representations. (A) A histogram of scores returned by the library search. Scores from correct matches are show in black and incorrect matches in gray. (B) A ROC curve plotting the number of false positive matches versus the number of true positive matches for a series of score thresholds. (inset) An ROC curve for the highest scoring 1% of false positives.

positives. For example, at a threshold with a 1% false-discovery rate, 93.8% of correct matches are found. With a 5% false-discovery rate, 99.9% of correct matches are found. The query set contains a mixture of spectra at three different charge states. We further examined the discriminatory power of the similarity score for spectra of each charge state individually. The search results for singly charged spectra produced an ROC curve with an area of 0.959. For doubly and triply charged spectra, the ROC areas were 0.984 and 0.993, respectively.

For the most part, the BiblioSpec and SEQUEST results agree. To understand the sources of error, we examined some spectra for which the two methods did not agree. The errors fell into five categories: (1) a poor-quality library spectrum, (2) incorrect SEQUEST identification of the query spectrum, (3) a query spectrum containing a mixture of peptides, (4) a library spectrum containing spurious, high-intensity, high m/z peaks, and (5) very similar reference spectra for different sequences.

We found several cases in which BiblioSpec correctly identified a spectrum whose sequence was incorrectly assigned by SEQUEST. This is not surprising, because all library spectra were identified using SEQUEST, which has a measurable false-positive rate. Based on matches to randomized sequences, the library is estimated to contain 51 incorrect identifications. Even with a high degree of similarity between a query and library spectrum, if the sequence is incorrectly assigned in the library, then the query sequence assignment will also be incorrect. One example of this phenomenon is illustrated in Figure 2. In this case, the query spectrum is doubly charged and has a SEQUEST Xcorr of 3.98 (normalized Xcorr is 0.402) for the assignment to LDEQGGATAAQVEVNK. A reference spectrum for this sequence was present in the library; however, it was returned as the second-best match to the query. The best match not only has a substantially higher score, but subjectively appears to be more similar by manual inspection.

In some cases, the query spectrum appears to be a mixture of two peptides. It is not uncommon for two peptides of the same m/z to elute off of the column at the same time so that they are both isolated and fragmented together.[19,20] An example of a single spectrum with fragmentation peaks from two different peptide sequences is illustrated in Figure 3. The BiblioSpec and SEQUEST assignments for this spectrum do not agree, but fragments from both sequences appear to be present in the query spectrum. BiblioSpec found the other peptide as the second best match with a score (0.571) fairly close to the best match (0.666).

Several query spectra were incorrectly matched to library spectra due to the presence of a single peak of high intensity. This type of peak arises when a singly charged peptide experiences a neutral loss of water or ammonia. The neutral-loss peak can easily be the highest intensity peak in the spectrum and can greatly influence the similarity score between two spectra, especially in combination with SMZ preprocessing, which gives higher weight to peaks with larger m/z values. However, the neutral-loss peak is not informative, because it only indicates that the two spectra have the same precursor m/z.

Minimizing false-positive identifications is the focus of ongoing development of BiblioSpec. Poor-quality library entries can be avoided by carefully choosing SEQUEST criteria for reference identifications, verifying spectrum identities by orthogonal means (e.g., other database search algorithms or de novo sequencing), and selecting from several spectra per peptide. Furthermore, search results may benefit from tailoring the preprocessing steps to the charge state of the spectrum. For instance, it may be advantageous not to weight peak intensities by m/z for singly charged spectra. Finally, using additional features together with the similarity score could improve the discrimination between correct and incorrect sequence assignments (as in ref 21).

LTQ-LCQ Comparison. A second design goal for our library searching method was that it be robust enough to work with data collected on different mass spectrometers operating in different laboratories. To test this capability, we created a library of *E. coli* spectra acquired on an LTQ ion trap at the University of Washington and searched for peptide assignments to query spectra that were acquired on an LCQ ion trap mass spectrometer at the University of Colorado. The BiblioSpec search made correct sequence assignments for every query spectrum whose respective

[19] Zhang, N.; Li, X. J.; Ye, M.; Pan, S.; Schwikowski, B.; Aebersold, R. *Proteomics* **2005**, *5*, 4096–4106.

[20] Venable, J. D.; Dong, M. Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R., III. *Nat. Methods* **2004**, *1*, 39–45.

[21] Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2* (2), 137–146.

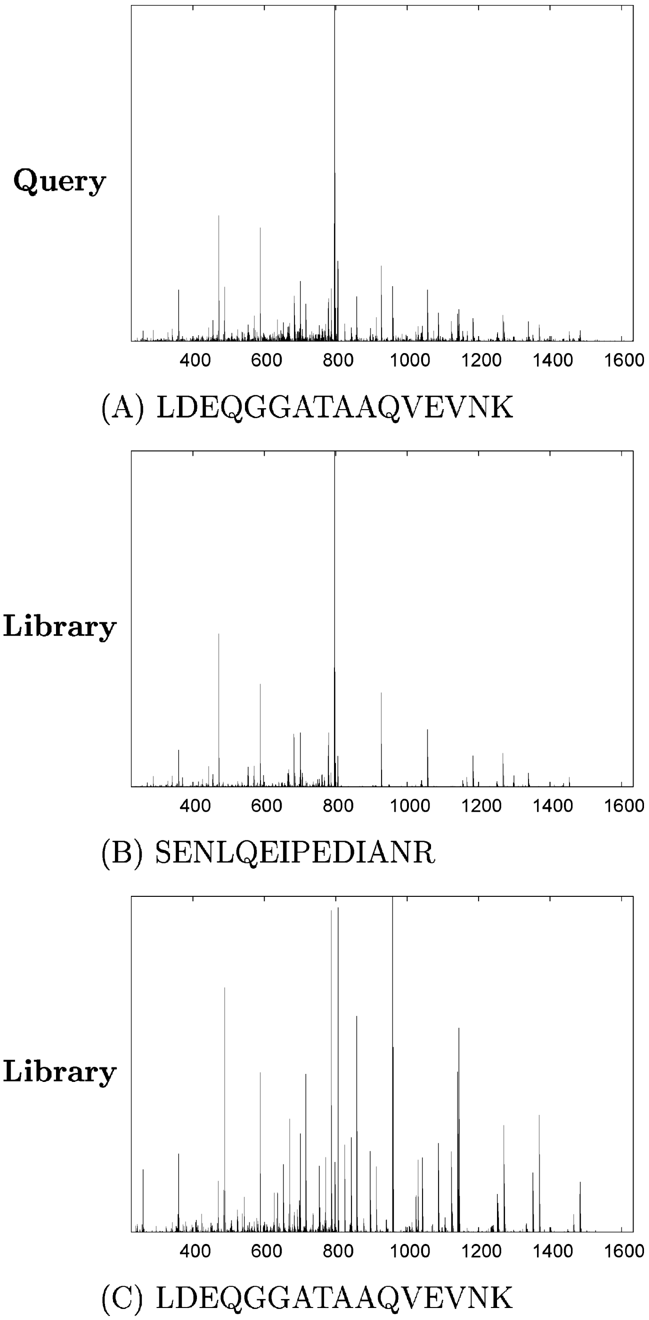

Figure 2. Example of a BiblioSpec sequence assignment that did not agree with SEQUEST. (A) Query spectrum and the sequence assigned to it by SEQUEST (Xcorr 3.980). (B) Library spectrum that best matches query (score 0.707). (C) Library spectrum with the second highest match score (0.582) and the same sequence that SEQUEST gave the query. The query spectrum looks more similar to the BiblioSpec match than to the SEQUEST match.

reference spectrum was in the library. Disagreement between BiblioSpec and SEQUEST (36% of matches) was entirely due to the limited size of the library: none of the 321 incorrect matches had the correct spectrum in the library. The area under the resulting ROC curve (Figure 4) is 0.983. Even with a much smaller library, we retain good sensitivity and specificity and are able to identify all of the sequences that are in the library. Thus, laboratories using the older LCQ ion trap should be able to take advantage of libraries of LTQ spectra.

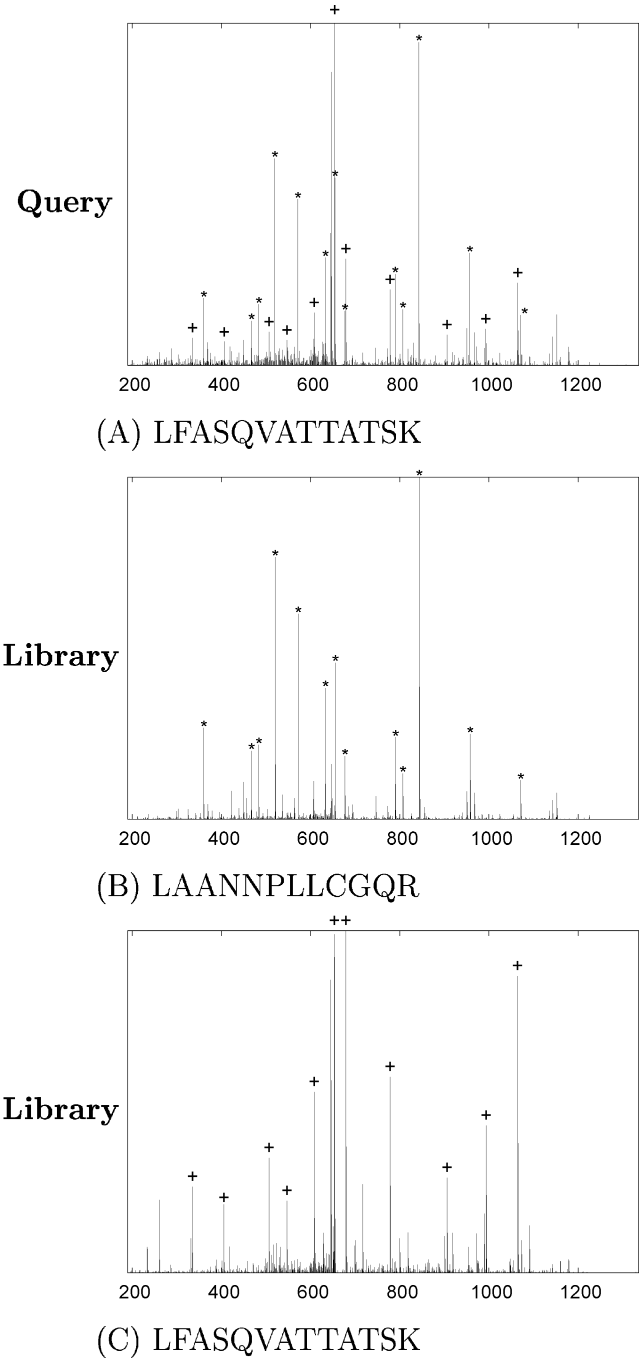

Figure 3. Example of a query spectrum for which the BiblioSpec and SEQUEST sequence assignments did not agree. The query is a mixture of two peptides, and shown below it are library spectra of the two individual components. (A) Query spectrum and the sequence assigned it by SEQUEST. Peaks in common with (B) are marked with * and peaks in common with (C) are marked with +. (B) Library spectrum that best matches the query (score 0.666). Peaks in common with query marked with a *. (C) Library spectrum with the second highest match score (0.571). Peaks in common with query marked with +.

Peptide Modification Search. BiblioSpec is able to identify spectra with posttranslational modifications (PTMs) as easily as unmodified spectra. Database searches have two limitations for PTM searches: the search time increases because many more predicted spectra must be considered, and reliability decreases

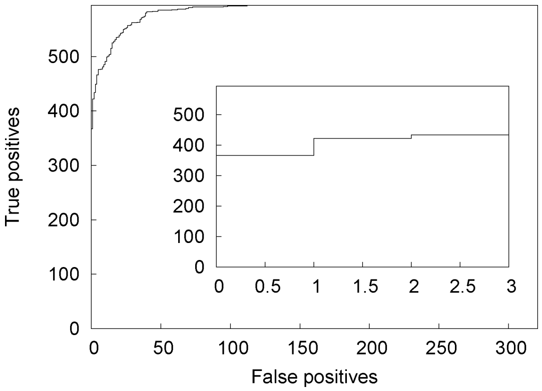

Figure 4. ROC curves of the *E. coli* search. ROC of all matches returned by the search with an area of 0.983. (Inset) ROC of the matches including the first 1% of false positives.

because a more sophisticated fragmentation model is required. However, with BiblioSpec, modified spectra may be easily included in the library and compared to queries. We demonstrated this capability by searching all spectra of the unfractionated *C. elegans* sample against a library containing modified spectra. BiblioSpec found 44 spectra that matched one of the modified library spectra. An example is shown in Figure 5.

CONCLUSIONS

Library searching offers a reliable means of peptide spectrum identification. Larger libraries increase the number of peptides that can be identified, but even with a smaller library, we can accurately identify those peptides present and discriminate those correct matches from incorrect matches based on similarity score. We have chosen reference spectra based on identifications by SEQUEST, but other sources could be used. Alternative database searching or de novo sequencing algorithms are other possible means of identifications. Spectra obtained from synthetic peptides or from recombinant proteins could also serve as references. Libraries could also be generated from previously published data that have benefited from extensive manual curation. BiblioSpec is a flexible spectrum comparison program that could be used in other applications such as for rapid detection of specific peptides of interest, for finding spectra in common between two experiments,[4] or as a preliminary step in database searching.

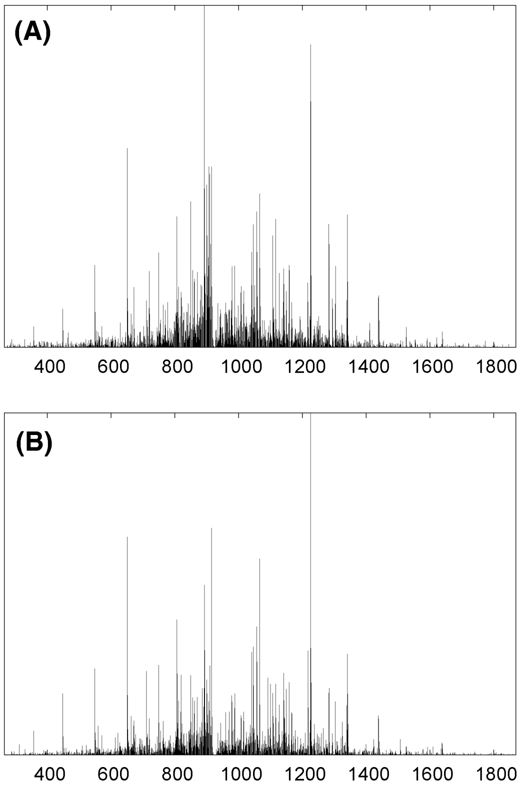

Figure 5. Example of a query spectrum match to a library spectrum with a posttranslational modification. (A) Query spectrum. (B) Library spectrum that best matches the query (score 0.803). The peptide sequence is LTNPTYGDLNHLVSLTMSGVTTCLR with an oxidized methionine.

ACKNOWLEDGMENT

Financial support was provided in part from a Basic Research Grant from Merck Research Laboratories and National Institute of Health grants P41-RR011823, R01-DK069386, and K22-AI059076.

Received for review February 13, 2006. Accepted June 7, 2006.

AC060279N