Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

research articles Journal of proteome research

Using Annotated Peptide Mass Spectrum Libraries for Protein Identification

R. Craig,[†] J. C. Cortens,[‡] D. Fenyo,[§] and R. C. Beavis*[,†,||]

Beavis Informatics Ltd., Winnipeg, MB, Canada R3B 1G7, Manitoba Centre for Systems Biology and Proteomics, University of Manitoba, Winnipeg, MB, Canada R3T 2N2, Rockefeller University, 1230 York Avenue, New York, New York 10027, and Biomedical Research Centre, University of British Columbia, Vancouver, BC, Canada V6T 1Z3

Received May 3, 2006

A system for creating a library of tandem mass spectra annotated with corresponding peptide sequences was described. This system was based on the annotated spectra currently available in the Global Proteome Machine Database (GPMDB). The library spectra were created by averaging together spectra that were annotated with the same peptide sequence, sequence modifications, and parent ion charge. The library was constructed so that experimental peptide tandem mass spectra could be compared with those in the library, resulting in a peptide sequence identification based on scoring the similarity of the experimental spectrum with the contents of the library. A software implementation that performs this type of library search was constructed and successfully used to obtain sequence identifications. The annotated tandem mass spectrum libraries for the *Homo sapiens*, *Mus musculus*, and *Saccharomyces cerevisiae* proteomes and search software were made available for download and use by other groups.

Keywords: peptide spectrum library • X! Hunter • GPM • GPMDB • protein identification

Introduction

One of the major methods for the experimental study of the proteins expressed by an organism is to use data derived from collections of tandem mass spectra to determine which proteins were present in a particular biological sample. The most common method of associating a particular tandem mass spectrum with a peptide sequence is to compare each experimental spectrum with theoretical spectra generated from a list of potential peptides, based on known, sequence-specific peptide ion fragmentation reactions. The list of peptides is derived from all of the protein sequences that could possibly be expressed by a particular organism. The comparison process generates a set of scores that indicate the similarity between any particular peptide sequence and the experimental mass spectrum. The peptide (or peptides) judged to be the most similar to the spectrum can then be associated with that spectrum and the process repeated for all of the spectra generated by the experiment. When the results of all of these spectrum-to-peptide correlations is combined, a list of candidate proteins can be generated for use by the biological researcher. The idea of selecting a chemical structure based on an enumeration of theoretical mass spectra has a long history, beginning with the work of Djerassi and Lederberg[1−3] to identify organic compounds. Its application to peptides was made possible by the sequence-specific bond cleavage rules described by Roepstroff and Folman[4] and Biemann.[5] The method has become popular for several reasons: there are practical software implementations of the idea;[6−9] it is simple to automate the analysis of large data sets; and the scores can be interpreted statistically for large-scale applications.[10,11]

Historically, the use of theoretical fragmentation patterns for analyzing small organic molecules was largely been replaced by another method commonly referred to as a "library searching". Library searching was originally formulated to improve the identification process, and it is also based on fragment mass spectra.[12,13] This type of search depended on the postulate that a particular organic molecule would fragment in a mass spectrometer in a manner that was characteristic of the detailed structure of the molecule. If this statement was true, it should be possible to generate a library of such fragmentation spectra using authentic samples of each molecule of interest. When each of these fragmentation spectra was associated with the corresponding molecular structure, the problem of assigning an experimental spectrum to a structure was reduced to the problem of determining which entry in a comprehensive library of annotated spectra corresponded best to the experimental spectrum. This method proved to be easy to automate and did not require a trained analyst to confirm the results of the match process. It was also largely hypothesis-free: it was not based on a theoretical understanding of fragmentation reactions. This feature of library searching proved to be important because even though a theoretical fragmentation pathway may be valid

* To whom correspondence should be addressed. E-mail: rbeavis@proteome.ca
[†] Beavis Informatics Ltd.
[‡] University of Manitoba.
[§] Rockefeller University.
[||] University of British Columbia.

10.1021/pr0602085 CCC: $33.50

Published on Web 07/14/2006

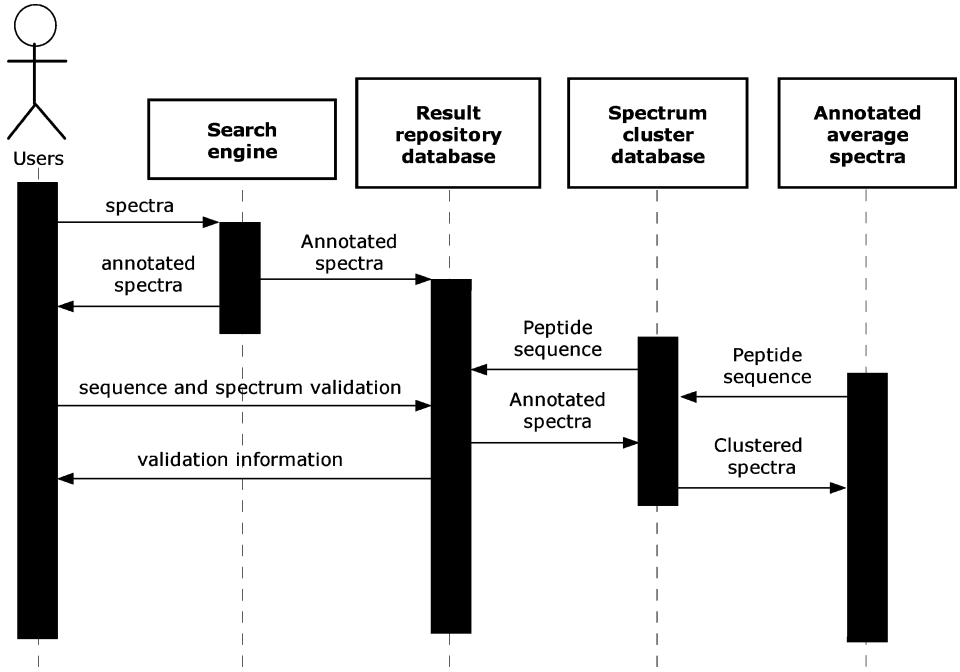

Figure 1. An activity diagram illustrating the steps required to generate an initial library of composite spectra.

in general, there were numerous special cases that deviate from the general rules. Instruments that use library searching for data interpretation have become general-purpose laboratory equipment that can be employed effectively by any user with minimal technical training. It remains the primary method of analyzing GC−MS data.[14−17]

The use of a library of authentic peptide spectra was impractical when protein identification was initially demonstrated. The main reasons that it was not employed were as follows:

1. it was unclear at the time that peptides produced similar tandem mass spectra when examined with different brands of mass spectrometer;

2. the set of possible protein sequences known for any particular species was either incomplete or changing rapidly, as new DNA and RNA sequencing methods became available; and

3. the expense of generating a comprehensive set of authentic peptide samples and measuring their tandem mass spectra.

Over the course of the past decade, the first two of these difficulties have been largely resolved. Most commercially available mass spectrometers suitable for proteomics have been found to generate a similar set of peptide fragment ions in tandem mass spectra. The variability between spectra of the same peptide has been found to be associated with the charge of the parent ion, rather than the specific brand of collisionally induced fragmentation ion source being used.[18−20] With respect to the second problem, the existence of completed genomes for a number of important laboratory model species has stabilized the lists of protein sequences used in proteomics. While there may be some variation in the predicted exon structure of some genes, the large-scale reannotations of genomes that were once common have become rare in established genome sequencing projects, such as those for human, mouse, or yeast.

The last problem, the cost of generating such a library, remains significant. Synthesizing the necessary authentic peptides and measuring their tandem mass spectra is simply too costly at present, even if a reduced set of "proteotypic" peptides is used.[21] Either a very large investment on the part of a funding agency or a dramatic reduction in peptide synthesis costs will be necessary for any such library to be populated with structures and spectra.

This paper explores the possibility of using an alternate strategy to generate such a library, which does not require the synthesis of peptide standards. This strategy involves the collection and annotation of a large number of peptide tandem mass spectrum data sets generated in the normal course of operation by various proteomics laboratories. With this collection of experimental annotations, it should be possible to develop a quality control and curation scheme resulting in a set of composite spectra, each the result of averaging together multiple observations of the same peptide. The composite spectra could then be used as a standard spectrum library, in the same manner as a library derived from synthetic peptide standards would be employed.

Describing all of the necessary components of a practical system using this strategy to identify proteins probably exceeds the scope of a single journal article. However, simply developing an annotated library of spectra using this strategy without a search engine to use it for protein identification would be a purely academic exercise. Therefore, while the work reported here focuses mainly on the process necessary to use the above scheme to create a practical annotated spectrum library, a search engine that can use those libraries to identify proteins was also developed and an illustrative example of its use described.

Experimental Methods

The overall process used to create libraries of annotated peptide MS/MS spectra was illustrated in Figure 1. The necessary software to accomplish these tasks and the annotated spectrum libraries have been made available either by file transfer protocol (FTP) or as source code in the GPM software version control system.[22] All code and databases have been written to require as little operating-system-dependent cus-

tomization as possible; that is, they could be compiled (C++ code) or run (PERL, SQL scripts, and binaries) on Microsoft Windows, LINUX, or Apple OSX operating systems. Some of the steps in the process were similar to those used to create lists of proteotypic peptides;[23] however, creating spectrum libraries has proven to be a significantly more difficult task.

Spectrum Collection and Annotation. Experimental groups were encouraged to upload groups of tandem mass spectra to a publicly available network of search servers, that used X! Tandem[9] to annotate these mass spectra with peptide sequences and assign goodness-of-fit quality measures. The user could choose to contribute these annotated sets of spectra to the GPMDB collection, in which case they were available for the library creation process. These spectra represent data from an assortment of different brands and types of tandem mass spectrometers. No attempt was made to segregate spectra based on instrument type. The GPM public servers all use a common set of sequences and accession numbers, based either on sequence collections provided by ENSEMBL[24] (*Homo sapiens* and *Mus musculus*) or SGD[25] (*Saccharomyces cerevisiae*). Any translation necessary between these basis sets of accession numbers and other systems, such as the International Protein Index,[26] was done at the last stage of the process.

Each of the individual search sites sent any new annotated data sets to the central GPMDB server daily. These data sets were loaded into the database repository and distributed to repository server computers. The data in GPMDB has been regularly examined by GPM contributors, staff, and other users. Annotations found to be in error were deleted. These annotation errors were normally caused by the use of unconventional parameter combinations by users. When unexpected behavior of the search engine was found, the X! Tandem software was corrected, tested, and redeployed. At the time of completing this manuscript, the main repository database contained approximately 1.2×10^7 annotated spectra, and the most confidently assigned subsets of these spectra were selected to construct the sequence annotated libraries.

To efficiently create the required spectrum libraries, an additional database was constructed. The database was populated by extracting the most confidently assigned spectra from the main repository and organizing them based on the accession number of the protein associated with the assigned peptide. For example, if a search was performed and a spectrum was assigned to the peptide sequence "SFQCELVMAK" and associated with the identification of protein accession number "At2g39730.1", then that spectrum would be associated with all other spectra found with that annotated sequence and accession number. This process generated a list of spectra associated with an accession number, subclassified by the peptide sequence. No attempt was made at this stage to deal with the existence of this peptide sequence in other proteins: that process was performed after the composite spectra were assembled. These database also stored information about each build of the spectrum libraries so that subsequent builds could be done incrementally rather than requiring a reclustering of the full repository database.

The algorithm for obtaining composite spectra for inclusion in a library was a straightforward, pairwise averaging process. The steps required were as follows.

1. Obtain all available spectra for a particular peptide sequence, parent ion charge state, and residue modification combination from the spectrum cluster database (e.g., obtain all of the spectra for the sequence "YHFMTWK", where the parent ion charge is +2 and the methionine residue has been oxidized).

2. Order the resulting list of spectra, from most to least confidently assigned (lowest to highest expectation value).

3. Delete duplicate spectra from the list.

4. Start with the most confident assignment. Select the next most confident assignment and identify sets of shared ions between the two spectra: a set of ions have m/z ratios within the allowed fragment ion mass tolerance.

5. Create a new m/z value for each set, by calculating a centroid of the m/z-value and intensities of the peaks in the set. Sum together the intensities of the peaks in the set and create a new spectrum made up of the summed intensities and m/z centroid pairs.

6. Take the new composite spectrum and apply the same steps to it and the next most confident spectrum, creating a new composite.

7. Continue this process until all spectra have been included into the composite.

8. Normalize the composite spectra so that the most intense peak has a relative intensity of 100.

9. Select the 20 most intense peaks from the composite spectrum and store these peaks along with the peptide sequence, parent ion charge, parent ion mass, and sequence modification information.

10. Store the resulting libraries in XML-formatted files for the modified and unmodified peptides found for a particular organism.

Once the basis set of peptide sequence annotated spectra was stored, the final step was to create a library of spectra annotated with the accession numbers associated with a particular protein sequence collection. For example, if the ENSEMBL protein sequences were used, the peptide sequences must be mapped to all ENSP-type accession numbers for protein sequences that contain those peptide sequences. This mapping was done by comparing the peptide sequence of each annotated peptide with all of the protein sequences in the appropriate sequence collection. When a particular peptide was found within a protein sequence, the accession number and position of the peptide in the protein were stored along with the annotated spectrum. Any number of accession numbers and positions could be associated with a particular spectrum. The fully annotated information for each library was stored in a formatted binary file.

Search Engine Development. A search engine to use the annotated spectrum libraries was designed and implemented, using the C++ class libraries developed for the X! Tandem project. The new search engine, X! Hunter, was designed to operate on a single thread. The search engine loaded the appropriate spectrum library and experimental spectra into memory and then scored each experimental-library spectrum pair that was within the specified parent ion mass tolerance. A binary search tree was used to accelerate the process of finding these pairs. The score used to determine the match between a library spectrum and an experimental spectrum was obtained using the 20 most intense peaks in the library and experimental spectra and eq 7 in ref 11, where the two factorial terms were replaced by a single value, $n!$; n was the number of peaks that occur in both the experimental and library spectrum. The vector inner product between the two spectrum vectors was also calculated and used as an alternative scoring method. The peaks in experimental spectra were prefiltered in an attempt

Table 1. the Current Size of Each of the Annotated Spectrum Libraries[a]

	H. sapiens ENSEMBL	*M. musculus* ENSEMBL	*S. cerevisiae* SGD
disk space (megabytes)	28.5	14.9	7.0
modified peptides	45,743	21,705	13,125
unmodified peptides	85,788	52,828	30,767
total spectra	131,531	74,033	43,892
spectra/gene	6.2	3.6	6.6

[a] Allowed modifications were methionine oxidation and cysteine alkylation.

to remove any interfering ^{13}C isotope peaks and those corresponding to neutral losses from the parent ion, such as the loss of water or ammonia. The expectation value corresponding to each score was estimated using the method described previously for X! Tandem,[11] by applying an average set of estimation parameters.

Following the scoring process, the full sequences of proteins corresponding to high scoring peptides were loaded from disk, and a full XML report of the results was written. The format of the input parameter specification and output XML files was the same as X! Tandem: a BIOML representation of sequences and parameters and GAML records of histograms. The software was compatible with mass spectrum input information in either structured text (DTA, Mascot Generic Format, or PKL) or XML (mzXML or mzData) files. A fully functional version of this search software with a user interface was made publicly available.[27]

Search Engine Comparison. The input tandem mass spectrum set (778 spectra) was generated using a Sciex QSTAR mass spectrometer, from a sample of bovine serum albumin (BSA), by the Manitoba Centre for Systems Biology and Proteomics. The default GPM parameters for QSTAR-type spectra were used. The conventional sequence and library searches were both performed using the ENSEMBL *H. sapiens* and cRAP sequences and libraries. Both searches were run on the same computer (Hewlett-Packard, model m7470n), using the GPM Web browser interface. The original search results were deposited in the GPMDB and can be retrieved for examination in detail using the following accession numbers: GPM00300004348 (conventional search) and GPM20100000113 (library search).

Results and Discussion

1. Characteristics of the Annotated Libraries. The population and curation of three libraries was performed to test the system. Table 1 lists some of the relevant characteristics of the library files. The creation of the libraries required approximately 150 h of processor time. The majority of the time was required to generate the cluster database, which required significant amounts of disk access to read the spectra from XML results files. A small library was created using the protein sequences in the GPM common Repository of Adventitious Proteins (cRAP), in addition to those of the model organisms. The cRAP sequences were selected from NCBI's nr sequence collection, and they represent artifact proteins commonly found in proteomics samples, for example, porcine trypsin or bovine serum albumin.

The distribution of peptide lengths in the *H. sapiens* library was plotted in Figure 2. As indicated by Figure 2a, the library contains very few peptide sequences with less than seven residues, reaching a maximum for peptides with approximately 13 residues. The lack of short peptides was attributed to the

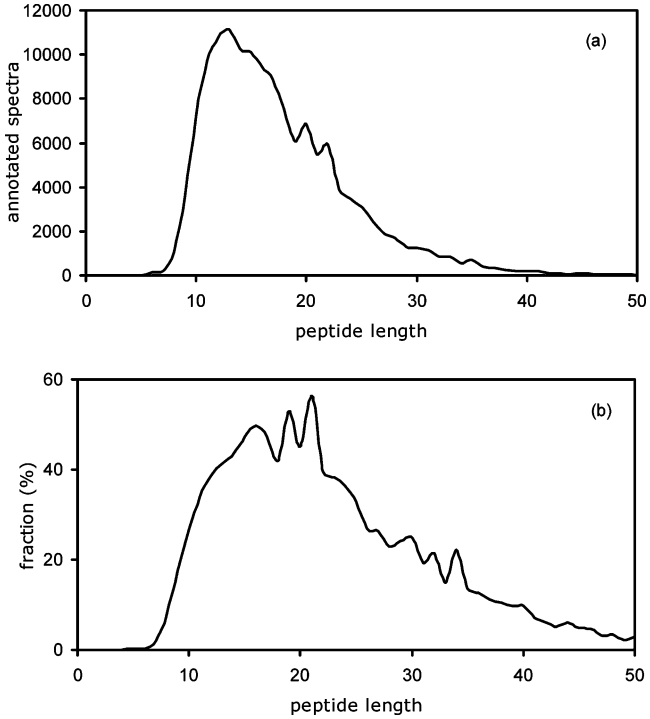

Figure 2. Graph (a) shows the composition of the *H. sapiens*-annotated peptide spectrum library as a function of peptide length, in residues. Graph (b) shows the same data as a percent fraction of the total number of unique tryptic peptides present in the human proteome with the same length.

fact that the short peptides tend to produce fragmentation signals that are difficult to recognize using "theoretical" mass spectra. These peptides have a limited number of residues, and they tend to produce few sequence-specific fragment ions, relative to longer peptides. Long peptides tend to produce fragment ions that allow the sequencing of regions of the molecule, while short peptides may only have signals characteristic of the N- and C-terminal regions of the peptide. For example, a pentamer peptide may only have strong signals corresponding to b2 and y3 ions, which may not be sufficiently characteristic to allow an unambiguous identification. Low mass peptides may be occasionally identified, but usually requiring a relatively intense strong tandem spectrum with a comparatively low signal-to-noise ratio. These peptides also tend to have low molecular masses, which result in parent ion m/z values that can be difficult to distinguish from intense chemical noise signals in that increase in intensity at low m/z values.

Figure 2b represents the same data, as a percentage of the number of unique tryptic peptide sequences of the same length, as calculated from the NCBI translation of the *H. sapiens* genome (version 36, October 2005). This curve indicated that even though the numerical size of the library may appear to be modest, the library was sufficient to provide a representative sampling of the proteome for peptides of 10−30 residues.

The amino acid composition of the peptides in the library was shown in Figure 3. Figure 3a plotted the amino acid composition of the peptides in the human library as a fraction of the total number of residues in the library. Figure 3b compared those compositions to the amino composition of the full ENSEMBL prediction of the human proteome. Both histograms were ordered from the most enriched (aspartic acid) to the most depleted (arginine). The residues lysine and arginine are

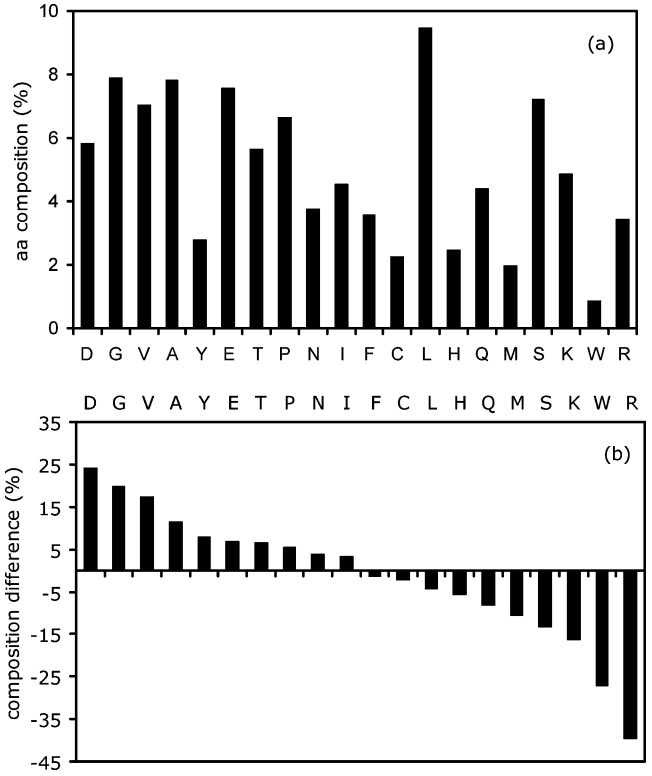

Figure 3. The amino acid composition of the human spectrum library as a fraction of the total number of residues in the peptides in the library. (a) The amino acid composition of the library, as a fraction of the total number of residues in the library. (b) The difference between the amino acid composition of the library and the overall amino acid composition of the human proteome as a fraction of the proteome composition for that particular residue.

a special case, as many of the peptides in the library were generated by the tryptic cleavage of proteins (i.e., cleavage at the peptide bond C-terminal to lysine or arginine residues). The difficulty in observing short peptides (see Figure 2) has the effect of removing those short tryptic peptides (which are relatively rich in arginine and arginine) from the library, resulting in a relative depletion of these residues in the library as a whole.

The majority of residue types were within ±10% of their proteome-wide values. The hydrophobic residue tryptophan was strongly depleted in the library compared to the proteome; however, it was also the least abundant residue in the proteome. Tryptophan residues can be sensitive to oxidation, resulting in a large number of possible degradation products that are not normally accounted for by proteomics identification searches. Tryptophan is also very hydrophobic and tends to be found in membrane spanning domains and other insoluble peptides, which can cause a depletion of tryptophan-containing peptides during sample preparation and handling.

2. Demonstration of Spectrum Library-Based Protein Identification. 2.1. General Findings. It was decided to limit the protein identification results presented here to a simple demonstration data set obtained from a relatively pure sample. The purpose of this paper was the discussion of the general properties of a library construction and library-based protein identification, rather explaining the many detailed differences that can arise between searches performed on large data sets from poorly characterized samples performed with different search algorithms.

Table 2. The Overall Performance of the Conventional Search Engine X! Tandem, Compared to the Library Search Engine, X! Hunter[a]

	unique peptides assigned	total peptides assigned	processor search time (ms/spectrum)
conventional search	27	105	21.0
library search	36	158	0.020

[a] Similar input parameters and the same set of spectra were used to perform these searches.

Table 2 compares the results of performing an identification search using the conventional search engine X! Tandem and the spectrum library search engine, X! Hunter. The complete lists of peptides identified in both cases were included as Supporting Information, and the search results themselves can be examined using the GPMDB use interface. The calculation of the time required per spectrum excluded the time required to load protein sequences or spectrum libraries and the time required to write report information to disk: it represented only the calculation time required to perform the identifications.

All of the identifications in the two result sets were manually inspected to ensure validity. A measure of the validity of the identifications, in addition to the statistical confidence assignments, was that, for the X! Tandem search, only peptides assignable to BSA were found. The X! Hunter search found only two peptides associated with non-BSA proteins, both of which were assigned with low confidence. The number of expected false positives for the X! Hunter search can be estimated by multiplying the expectation value cutoff by the total number of spectra identified, $0.01 \times 160 = 1.6$, in good agreement with the results.

In general, the spectrum library-based algorithm implementation performed significantly better than the conventional algorithm in terms of the number of peptides found, the confidence of identifications, and the time required to carry out the search.

2.2. Speed. The time required to perform the conventional search was approximately 1000× longer than the library-based search. This improvement in performance has two independent underlying causes. The first cause was simply a reduction in the number of peptides being considered by the library search. The increase in speed corresponding to this cause was similar to that found for the proteotypic peptide search engine X! P3 [23]. X! P3 performs searches approximately 10× faster than X! Tandem by limiting its initial search to only the peptide sequences that generate the best signals for a particular protein sequence. The list of peptides used by X! P3 is very similar to the list of peptides used to annotate the spectrum libraries, as both lists were composed from the same underlying data in the GPMDB.

The other underlying cause for the speed difference between two algorithms was the reduction in the complexity of the calculation. In the conventional search engine, each candidate peptide sequence must go through numerous calculations prior to the theoretical-to-experimental spectrum scoring. Initial calculations are performed to determine the mass of the peptide so that it can be compared to the parent masses of the available mass spectra, to see if it could be a viable candidate. Next, the masses of all possible modified forms of the peptide sequence must be exhaustively enumerated. For those candidate peptide sequences found to match the parent ion mass of an experimental spectrum, the theoretical spectrum must

be constructed. The mass of each residue is used to form theoretical spectra for at least b- and y- fragment ion types. Multiply charged fragment ions may also need to be calculated for parent ions with more than two charges. As a rough measure of complexity, X! Tandem uses approximately 2600 lines of code to calculate the permutations of modifications, parent ion masses, and theoretical spectra and only 50 lines of code to score the theoretical and experimental spectra. The time required to perform these calculations scales at best linearly with the number of peptides (n_p) contained in the protein sequences searched, $O(n_p)$.

The spectrum library matching implementation does not require any of these calculations. Instead, the mass of each experimental spectrum was used to look up the appropriate set of library spectra within the parent ion mass tolerance specified for the search. The lookup was performed by means of an efficient binary search tree.[28] Each library spectrum in that set was then scored and the relevant information stored, thereby eliminating a large number of calculations. This algorithm results in overall performance that scales very slowly with the number of spectra in the library (n_l), $O(log(n_l))$.

2.3. Sensitivity. An increase in sensitivity of the library search algorithm as compared to the conventional one was expected. In the test data, it was able to increase both the number of unique peptide sequences assigned by 9 (33%) and the total number of assignments by 53 (50%), while only generating two false-positive identifications. At first glance, this sort of improvement may not seem reasonable. The spectrum libraries were all created out of X! Tandem-annotated spectra, so how could it be that so many more assignments were made with the library search? Clearly, X! Tandem must have been capable of annotating spectra from these peptide sequences, or else they would not be represented in the library at all. If this is so, then why have they been missed in this case?

Manual inspection of the spectra associated with the valid assignments made by the library search that were not assigned by the theoretical spectrum approach showed that the spectra fell into two classes: (1) spectra with atypical fragment ion intensity distributions and (2) spectra with a significant number of noise peaks. Most of the spectra that fell into the first class had intense fragment ions caused by the neutral loss of ammonia (−17 Da) or water (−18 Da) from the conventional b and y ion series, with relatively small signals corresponding to the b or y ions themselves.

The reason that the library spectrum search was more sensitive at assigning these types of spectra was simple: the library search benefited from considering only a limited number of known characteristic peaks. When the results of tens of thousands of experimental data sets were combined and each peptide was treated as a special case, the library spectrum's profile was a much more accurate representation of the experimental spectrum than any theoretical spectrum generated from generally applicable rules. A conventional search engine cannot employ any such fine details of a generalized fragmentation model, as they will often generate an inappropriate choice caused by the statistical uncertainties inherent in any model. Finding a few additional assignments would be of little value if in doing so the validity of other results becomes questionable. Therefore, a library search has a considerable advantage when applied to spectra with atypical fragmentation patterns.

3. Contrasting the Capabilities of Conventional and Library Search Strategies. Even though spectrum library searches have some significant practical advantages over theoretical spectrum searches, they also have disadvantages. For example, a library search cannot find anything that has not been previously observed. Therefore, attempting to discover novel post-translational modifications using a library search would be inappropriate. The construction of spectrum libraries is not a trivial process, and it may not be possible to create a useful library for proteomes that have not been regularly examined using proteomics. Therefore, library searching will only be possible for a limited number of proteomes for the foreseeable future.

If a library can be constructed, however, a unique strength is the ability to correct assignments, based on additional information. For example, if a sequence assignment has been made in error, a skilled individual may be able to determine the correct assignment and change the annotation in the library. Once that change has been made, the library search will always produce the correct assignment, even though the conventional search will continue to make the same mistake. An application of this feature would be the annotation of spectra that simply do not contain enough sequence-specific ions to be readily assigned to any sequence. In these cases, it is often necessary to apply higher order tandem fragmentation (MS^n) to determine the identity of the sequence. Once that sequence has been determined, however, the original MS/MS spectrum can be annotated with the sequence, and it will not be necessary to repeat the more difficult (and potentially less sensitive) MS^n analysis again.

Conclusions

We have demonstrated that it was possible to create an annotated peptide tandem mass spectrum library of sufficient size to be useful for the *H. sapiens*, *M. musculus*, and *S. cerevisiae* proteomes. These libraries were composed from the most confident identifications obtained from numerous donor experimental proteomics groups using different experimental equipment and protocols. Therefore, these libraries represent the average set of ions generated from the subset of peptide sequences that produce interpretable tandem mass spectra.

These libraries were used to perform protein identifications, by comparison of library spectra with experimental spectra using an open source search engine, X! Hunter. This identification process was effective and rapid, as compared with more conventional protein identification software.

Acknowledgment. We thank all of the laboratories that were kind enough to contribute data to the GPMDB repository, which made the development of the spectrum libraries possible. In particular, we thank those responsible for the Peptide Atlas data repository, the Open Proteomics Database, and the Human Plasma Proteome Project data repository. Additional outstanding contributions have been made by the groups of Richard Smith, Leonard Foster, John Wilkins, Bill Vensel, Phil Andrews, and Tony Pawson. R.C.B. also thanks Stephen Stein (NIST) for discussions regarding the use of spectrum libraries in molecular identification.

Supporting Information Available: A complete list of peptides identified using the conventional search engine X! Tandem and the spectrum library search engine X! Hunter. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi C. Applications of artificial intelligence for chemical inference. I. The number of possible organic compounds. acyclic structures containing C, H, O, and N. *J. Am. Chem. Soc.* **1969**, *91*, 2973−2976.

(2) Duffield, A. M.; Robertson, A. V.; Djerassi, C.; Buchanan B. G.; Sutherland G. L.; Feigenbaum E. A.; Lederberg J. Applications of artificial intelligence for chemical inference. II. Interpretation of low-resolution mass spectra of ketones. *J. Am. Chem. Soc.* **1969**, *91*, 2977−2981.

(3) Schroll, G.; Duffield A. M.; Djerassi, C.; Buchanan, B. G.; Sutherland, G. L.; Feigenbaum, E. A.; Lederberg, J. Applications of artificial intelligence for chemical inference. III. Aliphatic ethers diagnosed by their low-resolution mass spectra and nuclear magnetic resonance data. *J. Am. Chem. Soc.* **1969**, *91*, 2977−2981.

(4) Roepstorff, P.; Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **1984**, *11*, 601.

(5) Biemann, K.; Martin, S. A. Mass spectrometric determination of the amino acid sequence of peptides and proteins. *Mass Spectrom. Rev.* **1987**, *6*, 1−76.

(6) Eng, J.; McCormack, A.; Yates, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994** *5*, 976.

(7) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.

(8) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J Proteome Res.* **2004**, *3*, 958−964.

(9) Craig, R.; Beavis, R. C. X! TANDEM: matching proteins with mass spectra. *Bioinformatics* **2004,** *20*, 1466−1467.

(10) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(11) Fenyö, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−774.

(12) Heller, S. Conversational mass spectral retrieval system and its use as an aid in structure determination. **1972**, *44*, 1951−1961.

(13) Heller, S. The history of the NIST/EPA/NIH mass spectral database, *Today's Chemist at Work* **1999**, *8*, 45−50.

(14) Stauffer, D. B.; McLafferty, F. W.; Ellis R. D.; Peterson, D. W. Probability-based-matching algorithm with forward searching capabilities for matching unknown mass spectra of mixtures. *Anal. Chem.* **1985**, *57*, 1056−1060.

(15) TIAFT-User Contributed Collection of EI Mass Spectra, http://www.tiaft.org/main/mslib.html.

(16) Mass Spectrometry Database Committee, http://www.ualberta.ca/~gjones/mslib.htm

(17) Aebi, B.; Bernhard, W. Advances in the use of mass spectral libraries for forensic toxicology. *J. Anal. Toxicol.* **2002**, *26*, 149−156.

(18) Gu, C.; Tsaprailis, G.; Breci, L.; Wysocki, V. H. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal. Chem.* **2000**, *72*, 5804−5813.

(19) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **2000**, *35*, 1399−1406.

(20) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R., III Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76*, 1243−1248.

(21) Aebersold R. Constellations in a cellular universe. *Nature* **2003**, *422*, 115−116.

(22) Craig, R.; Cortens; J.; Beavis, R. C. An open source system for analyzing, validating and storing protein identification data, *J. Proteome Res.* **2004**, *3*, 1234−1242; http://www.thegpm.org.

(23) Craig, R, Cortens, J.; Beavis, R. C. The use of proteotypic peptide profiling in a practical system for protein identification, *Rapid Commun. Mass Spectrom.* **2005**, *19*, 1844−1850.

(24) Hubbard, T.; et al. Ensembl 2005, *Nucleic Acids Res.*, **2005**, *33*, D447−D453.

(25) Hong, E. L.; et al. Saccharomyces Genome Database, http://www.yeastgenome.org.

(26) Kersey, P. J.; Duarte, J.; Williams, A.; Karavidopoulou, Y.; Birney, E.; Apweiler, R. The International Protein Index: An integrated database for proteomics experiments, *Proteomics* **2004**, *4*, 1985−1988.

(27) GPM X! Hunter Search page, http://h201.thegpm.org.

(28) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*, 2nd ed.; McGraw-Hill: New York, 2001; pp 273−301.

PR0602085