MS2Grouper: Group Assessment and Synthetic Replacement of Duplicate Proteomic Tandem Mass Spectra

David L. Tabb
Life Sciences Division, Oak Ridge Laboratory, Oak Ridge, Tennessee, USA

Melissa R. Thompson,* Gurusahai Khalsa-Moyers,*
Nathan C. VerBerkmoes,* and W. Hayes McDonald
Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

Shotgun proteomics experiments require the collection of thousands of tandem mass spectra; these sets of data will continue to grow as new instruments become available that can scan at even higher rates. Such data contain substantial amounts of redundancy with spectra from a particular peptide being acquired many times during a single LC-MS/MS experiment. In this article, we present MS2Grouper, an algorithm that detects spectral duplication, assesses groups of related spectra, and replaces these groups with synthetic representative spectra. Errors in detecting spectral similarity are corrected using a paraclique criterion—spectra are only assessed as groups if they are part of a clique of at least three completely interrelated spectra or are subsequently added to such cliques by being similar to all but one of the clique members. A greedy algorithm constructs a representative spectrum for each group by iteratively removing the tallest peaks from the spectral collection and matching to peaks in the other spectra. This strategy is shown to be effective in reducing spectral counts by up to 20% in LC-MS/MS datasets from protein standard mixtures and proteomes, reducing database search times without a concomitant reduction in identified peptides. (J Am Soc Mass Spectrom 2005, 16, 1250–1261) © 2005 American Society for Mass Spectrometry

The use of mass spectrometry for the analysis of proteins is a key intersection between analytical chemistry and biology. A particularly powerful strategy for analyzing proteins is the use of liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS). "Shotgun proteomics" performs this analysis on peptides from a proteolytic digestion of a mixture of proteins. This technique has found broad application in identifying both proteins and post-translation modification sites from mixtures as simple as a few proteins or as complex as entire proteomes (reviewed in [1–3]). For the most complex mixtures, additional separation is necessary. Some of the more effective forms have used an additional LC separation such as the multidimensional protein identification technology (MudPIT) strategy developed in the Yates laboratory [4–6].

Shotgun proteomics experiments work best when the mass spectrometer captures a tandem mass spectrum from each peptide present in the sample. If a small set of highly abundant peptides is repeatedly targeted for tandem mass spectrometry, the diversity of spectra collected will be reduced, preventing the detection of less concentrated peptides. One way around this problem is to improve liquid chromatographic separation efficiency, increasing the resolving capacity for the peptide mixture. The mass spectrometer's instrument control software can also play a role by maintaining a list of parent ion *m/z* values at which isolation and fragmentation have taken place; precursor ions at these *m/z* values will not be selected for a period of time (e.g., the Dynamic Exclusion option from Thermo Finnigan [7]).

Despite efforts to limit repeated collection of particular tandem mass spectra, redundancy can be found in almost any set of tandem mass spectra from a shotgun proteomic experiment. Previous work has shown that this spectral redundancy ranges from 18% for LC-MS/MS analysis of bands from 1D PAGE separation, to 25% for MudPIT proteomic analysis, to 28% for MudPIT analysis of a ~200 protein mixture [8]. This phenomenon can be exploited for several purposes. First, recognizing the extent of redundancy is useful for optimization of separation and data acquisition parameters.

Published online June 23, 2005

Address reprint requests to Dr. W. H. McDonald, Chemical Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, MS-6131, Oak Ridge, TN 37831, USA. E-mail: mcdonaldwh@ornl.gov

* Also at the Genome Science and Technology, UT-ORNL Graduate School, Oak Ridge, TN 37830-8026, USA.

1044-0305/05/$30.00
doi:10.1016/j.jasms.2005.04.010

Received February 2, 2005
Revised April 18, 2005
Accepted April 19, 2005

Next, the time required for processing spectra (such as identifying the peptide sequences fragmented in each MS/MS) can be reduced if each set of replicate spectra can be replaced by a single, representative spectrum. The process of constructing a representative spectrum can potentially yield a spectrum of higher signal-to-noise than the observed spectra because each duplicate spectrum represents an independent observation of the MS/MS fragments, making confident identification more likely. Finally, spectra that are present in multiple copies but are not identified by database searching may represent peptides with variant sequences or unanticipated post-translational modifications, making them excellent candidates for de novo sequence determination.

To leverage these potential advantages, one must first employ reliable strategies to determine which MS/MS spectra are generated from the same parent peptide. Techniques for evaluating the degree of similarity between or among spectra have broad application in analytical mass spectrometry, and multiple techniques for performing such comparisons have been reported [8–13]. LIBQUEST, one of the first algorithms published for comparing experimental peptide spectra to libraries of identified spectra, employed a variation of SEQUEST's cross correlation for comparison [14]. Faster approaches based on dot product comparisons have been more broadly used for spectrum to spectrum comparisons [8, 9, 11, 13]. These algorithms treat individual spectra as vectors in multidimensional space, where the coordinate in each dimension is the intensity at a particular *m/z* value. Computing the cosine of the angle between two such vectors gives a measurement of the similarity of two spectra. Both Tabb et al. [8] and Beer et al. [13] have reported the use of dot product-based spectral comparisons, specifically for the detection of similar peptide MS/MS spectra in shotgun proteomics data. Where systems like LIBQUEST attempt to match each experimental spectrum to an entry in a library of identified spectra, these latter systems attempt to compare each experimental spectrum to other experimental spectra with similar precursor ion *m/z* values.

Tabb et al. created "NoDupe" to reveal the amount of similarity present within individual LC-MS/MS analyses across a variety of different experimental models, showing that choosing an arbitrary representative from each group resulted in a minor loss of protein identifications [8]. Beer et al. extended the comparison across multiple LC-MS/MS experiments, pairing a system to cluster similar spectra with an algorithm to re-centroid fragment ions observed in replicate spectra to construct a single representative spectrum [13].°By identifying spectral duplication before identification, these systems can improve the efficiency of subsequent database searches. An alternative strategy would seek to identify spectral duplicates during or after identification of peptides to strengthen scoring discrimination; if multiple copies of a spectrum are captured, the resulting identifications should be identical for the set of duplicates. Whether grouping takes place before identification or after, however, the processes of scoring pairwise similarities and constructing groups from these scores are still necessary.

In this article, we describe several improvements on these strategies that we have implemented in the new MS2Grouper algorithm. We describe a technique by which the false positive and false negative rates for similarity detection can be characterized. We demonstrate the value of a paraclique-based criterion for assessing groups of similar spectra. We also describe a greedy algorithm to generate a synthetic representative spectrum for spectral clusters that obviates the need for "re-centroiding" the data, and we examine the effects of its use. In combination, these improvements make spectral grouping an effective tool for the analysis of shotgun proteomics data, reducing spectral counts by up to 20% while losing less than 1% of peptide identifications.

Methods

Chemicals

All salts, protein standards, DTT, and guanidine were obtained from Sigma Chemical Company (St. Louis, Mo). For all protein digestions, sequencing grade trypsin from Promega (Madison, WI) was used. HPLC grade water and acetonitrile from Burdick and Jackson (Muskegon, MI) and 98% formic acid from EM Science (an affiliate of Merck KGaA, Darmstadt, Germany) were used for sample cleanup and HPLC applications.

Extended Protein Standard Mixture (EPSM)

The EPSM consists of approximately equimolar amounts of carbonic anhydrase II from *Bos taurus*, conalbumin (ovotransferrin) from *Gallus gallus*, concanavalin A from *Canavalia ensiformis*, cytochrome *c* from *Bos taurus*, deoxyribonuclease I from *Bos taurus*, lysozyme *c* from *Gallus gallus*, β-lactoglobulin A from *Bos taurus*, β-lactoglobulin B from *Bos taurus*, ribonuclease A from *Bos taurus*, ribonuclease B from *Bos taurus*, thyroglobulin from *Bos taurus*, serum albumin from *Homo sapiens* and *Bos taurus*, alcohol dehydrogenase E from *Equus caballus* liver, alcohol dehydrogenase I from *Saccharomyces cerevisiae*, α-amylase from *Bacillus subtilis*, β-amylase from *Ipomoea batatas*, apomyoglobin from *Equus caballus*, hemoglobin (A and B) from *Equus caballus*, and (apo)-transferrin from *Bos taurus*. The proteins were suspended in 6 M guanidine and 10 mM DTT to form the stock solution.

Proteome Sample

Shewanella oneidensis MR-1 cells were grown in 1-L cultures to mid-log phase under aerobic conditions, pelleted by centrifugation, washed twice in ice-cold 50 mM Tris (pH 7.5), and stored at −80 °C until extraction.

For protein extraction, cell pellets were resuspended in ice-cold 50 mM Tris (pH 7.5) with 10 mM EDTA and disrupted by sonication. Unbroken cells were pelleted by centrifugation (5000 × *g* for 15 min). The supernatant was spun at 100,000 × *g* for 60 min to separate the soluble fraction from the membrane fraction. The soluble fraction was aliquoted, quantitated with BCA analysis (Pierce, Rockford, IL), and frozen at −80 °C until digestion.

Digestion of Samples

The extended protein standard mixture and *S. oneidensis* soluble proteome were both digested and processed by the following protocol. Each sample was denatured with 6 M guanidine and 5 mM DTT at 60 °C for 1 h and then diluted in 50 mM Tris (pH 7.5)/5 mM $CaCl_2$ to obtain a final guanidine concentration of 1 M. Sequencing grade trypsin was added at 1:100 (mass ratio of enzyme to protein), and digestion reactions were run for 16 h. Trypsin was added a second time at 1:100 and digestion was run for another 5 h, followed by a final reduction step with 10 mM DTT for 1 h. Samples were immediately desalted with a C18 Sep-Pak (Waters, Milford, MA) and concentrated using a centrifugal evaporator (Savant Instruments, Holbrook, NY) to ~10 μg/μl for the proteome sample and ~200 ng/μl for the extended protein standard mixture. Both samples were filtered to remove insoluble material, aliquoted and frozen at −80 °C until analysis.

LC/LC-MS/MS Analysis

The *S. oneidensis* soluble proteome and the EPSM were analyzed via a two-dimensional (2D) nano-LC-MS/MS system with a split-phase column [15]. For the LC/LC-MS/MS analysis, either an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) was interfaced with a three-dimensional (3D) quadrupole ion trap mass spectrometer (LCQ-DECA XP plus) or a Surveyor HPLC was interfaced with a linear ion trap (LTQ, ThermoElectron, San Jose, CA). The split-phase columns were constructed as follows: the upstream column was packed with ~ 3.5 cm of strong cation exchange material (Luna SCX 5 μm 100A Phenomenex, Torrance, CA) into a 100 μm i.d. fused silica capillary via a pressure cell (New Objective, Woburn, MA) followed by 3.5 cm of C-18 reverse phase (RP) material (Aqua C18 5 μm 200A Phenomenex). For each replicate analysis, ~250 μg of soluble proteome or ~25 μg EPSM was loaded off-line onto the dual phase column using the pressure cell. The loaded RP-SCX column was then positioned on the instrument behind a ~15 cm C18 RP column (Jupiter C18 5 um 300A Phenomenex) also packed via pressure cell into a Pico Frit tip (100 μm with 15 μm tip from New Objective) inline for direct microelectrospray into the mass spectrometer. The soluble proteome was analyzed in duplicate via a 24-h 12-step MudPIT analysis, and the EPSM was analyzed using a 10-h 5-step MudPIT analysis as described previously [6, 15, 16]. In essence, a salt pulse at the beginning of each MudPIT step eluted an increasingly polar set of peptides from the SCX material to the RP material for separation by hydrophobicity.

MS2Grouper Algorithm

MS2Grouper is software written in the C^{++} programming language. The algorithm encompasses three main capabilities: detecting similarity between spectra, assessing groups of related spectra on the basis of pairwise similarities, and constructing a representative spectrum for each similarity group. A flowchart of the software is shown in Figure 1. The algorithm is designed to conduct these operations on the spectra from each RPLC separation independently, though it could be adapted to run on arbitrarily large libraries of spectra. The software reads in all spectra from a separation, processes them, and then writes a new set of spectra to disk. The MS2 file format [17] allows markups to note which spectra have been synthesized from collections of others, enabling users to trace back to the spectra that gave rise to synthetics. Similarity detection in MS2Grouper is quite like the process described for NoDupe [8]. The sum of fragment ion intensities for each spectrum is computed. Each fragment ion that contributes less than 1% of this sum is removed from consideration. Each spectrum is then expressed as a series of bins, each bin 1 *m/z* wide. Each bin holds a value indicating the normalized intensity of peaks within that bin's *m/z* range. For example, a tall peak might be more intense than 80% of other fragment ions. A bin encompassing this peak alone would hold the proportion estimate 0.8. This value is approximated by this expression:

$$p = \frac{1 - e^{xi}}{1 - e^{xi},}$$

where p is the proportion estimate, x is the modeling constant (the value −520 was found to work well), and i is the intensity of a fragment divided by the intensity sum for the spectrum. All spectra are sorted by their precursor *m/z* values, and spectra are checked for similarity to any others within 3 *m/z*. The similarity value for each spectral pair is computed to be:

$$s = \frac{\sum AB}{\sqrt{\sum A^2 \sum B^2},}$$

where s is the similarity (ranging from zero to one), A is the proportion for a bin from the first spectrum, and B is the proportion for a bin from the second spectrum. If peaks are present at the same locations in both spectra, the numerator of this fraction grows to approach the value of the denominator. If two spectra have peaks in completely different locations, the denominator grows

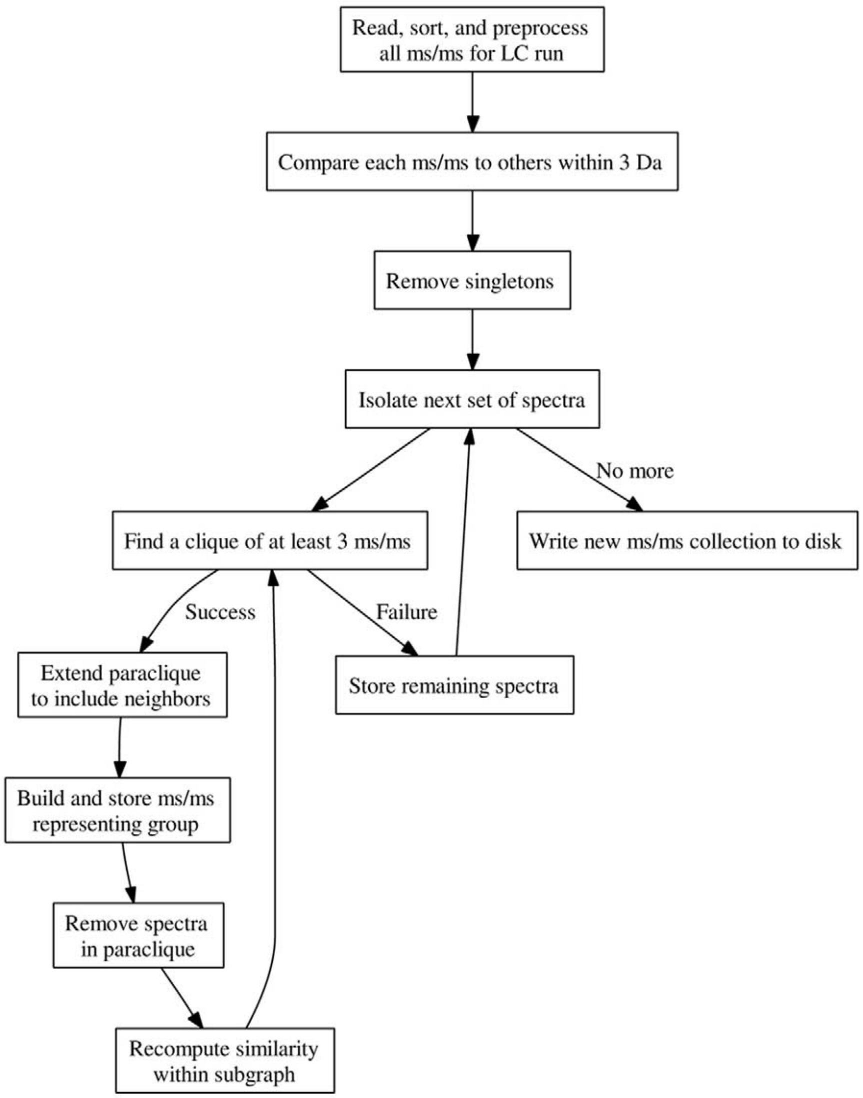
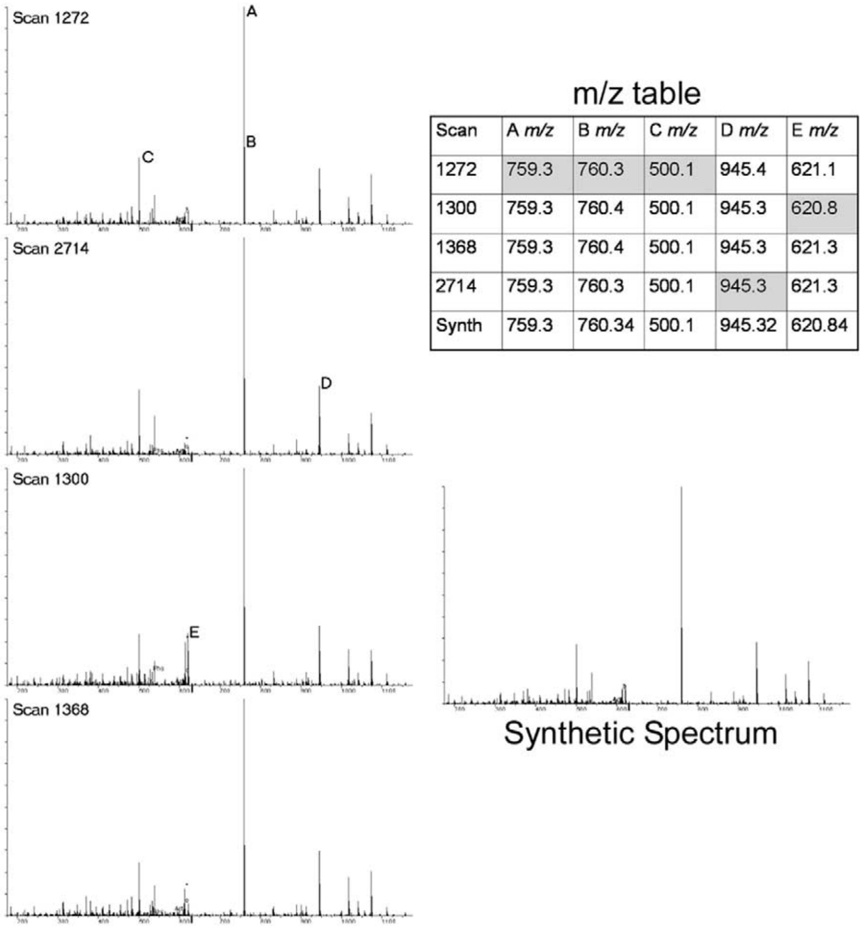

Read, sort, and preprocess all ms/ms for LC run
Compare each ms/ms to others within 3 Da
Remove singletons
Isolate next set of spectra
No more
Find a clique of at least 3 ms/ms
Write new ms/ms collection to disk
Success
Failure
Extend paraclique to include neighbors
Store remaining spectra
Build and store ms/ms representing group
Remove spectra in paraclique
Recompute similarity within subgraph

Figure 1. MS2Grouper reads spectra into memory, collapses sets of similar spectra to representatives, and writes the spectra to disk as a new file. Spectra are subdivided into batches that link internally but not externally. Group assessment for each of these sets is iterative, constructing paracliques until no clique of at least three spectra remains. Once spectra have been assigned to a paraclique, they are removed from consideration.

while the numerator remains zero. When *s* is greater than 0.7, the two spectra are reported as duplicates and connected by a similarity link. This similarity detection is symmetric because the similarity from Spectrum A to Spectrum B will always equal the similarity from Spectrum B to Spectrum A.

Group assessment is managed differently in MS2Grouper than in previous algorithms. The full set of spectra for an LC separation is divided into subsets such that no spectrum in each subset is connected to any spectrum outside the subset. These sets of spectra are visualized in undirected graphs such as Figures 5 and 6, where each spectrum is a node, and spectra bearing mutual similarity are connected by edges. This information is exported by MS2Grouper into files ready for processing by AT&T's GraphViz software [http://www.graphviz.org]. MS2Grouper employs a simplified implementation of the Bron-Kerbosch algorithm [18] to enumerate the maximal cliques from each graph. Cliques are sets of spectra in which each member is connected to every other member. The best clique is selected by choosing the clique encompassing the largest number of spectra, and ties are broken by selecting the clique incorporating the spectra with the largest sums of fragment ion intensities. If a clique of at least three spectra is identified, it is used to nucleate a *paraclique*, a process in which spectra connected to all but one member of the clique are "glommed" to the group. This technique was pioneered by Langston et al. [personal communication] and has been applied successfully for analysis of microarray data [19]. These criteria ensure that the minimum requirement for a



m/z table

Scan	A *m/z*	B *m/z*	C *m/z*	D *m/z*	E *m/z*
1272	759.3	760.3	500.1	945.4	621.1
1300	759.3	760.4	500.1	945.3	620.8
1368	759.3	760.4	500.1	945.3	621.3
2714	759.3	760.3	500.1	945.3	621.3
Synth	759.3	760.34	500.1	945.32	620.84

Figure 2. Four spectra for the doubly-charged peptide ALEGEPEWEAK formed a paraclique in the proteome data. An iterative process creates a synthetic representative for the group by repeatedly adding one new peak to the synthetic spectrum for a set of peaks removed from the input spectra. In each iteration, the most intense peak remaining among all the spectra is selected as a reference peak (the gray-shaded spaces in the table indicate which spectrum contained this reference peak). This reference peak is matched to a peak in every other spectrum, and a new peak is inserted in the synthetic spectrum at a weighted average *m/z* value. Spectra with intense fragment ions tend to steer this process; in this example, each of scan 1368's top five peaks are less intense than in other spectra. The resolution of the mass analyzer can cause some discrepancies for particular fragments. In this example, the peak labeled as "E" is a doubly-charged neutral loss from the precursor, and it appears as a single peak in some spectra and two peaks in others.

spectrum to be associated with a group is two linkages (to be part of a clique of three spectra or be glommed to one). For each set of spectra, MS2Grouper identifies the largest clique, gloms permissible spectra to the clique, collapses the formed paraclique to a group representative, removes the paraclique from the set, recomputes similarities, and then searches for the largest remaining clique. It repeats this process until no clique of at least three spectra can be found and then stores the remainder of the set as singletons. By raising the standards for group membership, MS2Grouper guards against grouping spectra improperly and thus losing identifications.

MS2Grouper replaces each group of duplicate spectra with a representative spectrum. It features two different modes for this replacement: selection of the most intense spectrum and synthesis of a summary spectrum. If the user has selected the former mode, MS2Grouper will include the spectrum with the highest intensity sum (as determined during preprocessing) as the representative for the similarity group. Since multiple, independent observations of the spectrum have been grouped, a spectrum with greater mass accuracy and signal-to-noise than the contributing spectra may be synthesized. Because each MS/MS has been subjected to centroiding by the instrument control software, some means of mapping peaks in each spectrum to peaks in the others (rather than simple spectral averaging) is necessary. MS2Grouper attempts to construct synthetic spectra by a greedy algorithm illustrated°in°Figure°2.°First,°it°uses°weighted°averaging°to determine the precursor *m/z* reported for the synthetic spectrum; the weight associated with each contributing spectrum is the intensity sum. It also uses this process to

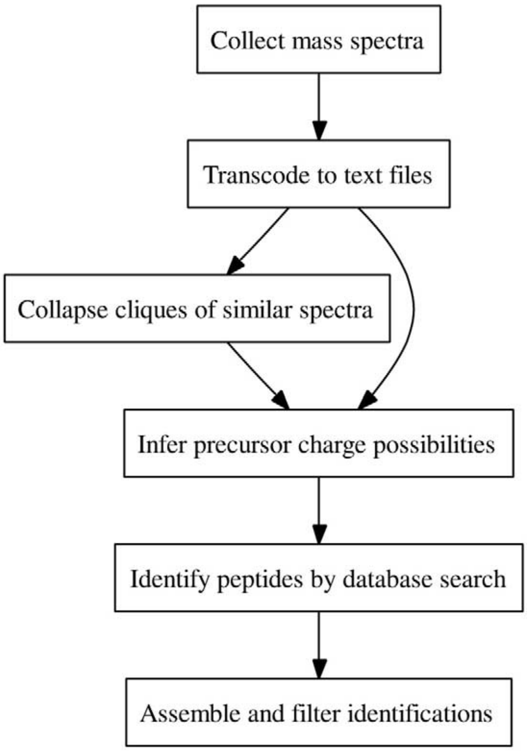

Collect mass spectra
Transcode to text files
Collapse cliques of similar spectra
Infer precursor charge possibilities
Identify peptides by database search
Assemble and filter identifications

Figure 3. Tandem mass spectra were processed through a pipeline using a series of algorithms. MS2Grouper (when used) was employed immediately after the spectra were moved to text files from the binary instrument capture files. In this way, spectral grouping could impact both precursor charge state inference and the identification process.

determine the weighted average of fragment ions observed among the spectra, and it sets this as the target number of peaks to place in the synthetic spectrum. To generate the fragment ions in the synthetic spectrum, the algorithm iteratively finds the most intense remaining fragment ion, locates the peak from each of the other spectra that matches this tallest peak best by *m/z*, and places a peak in the synthetic spectra at the weighted average *m/z* with the sum of intensities observed in the grouped spectra at this position. The software continues removing peaks in this way until the target number of peaks has been added to the synthetic spectrum. This process leads to a reduction in spectral counts because each group of duplicate spectra is replaced by a single, representative spectrum.

MS2Grouper reports changes to spectral collections in several ways. First, each spectrum written to the new MS2 file reports the number of spectra it replaces as part of its spectrum identifier; originally these identifiers contain two repeats of the scan number (as in test.748.748.2 for the doubly-charged scan no. 748 of the "test" sample), but MS2Grouper changes the second scan number to represent the group size instead (for example, to test.748.3.2 for a spectrum representing three in the original set). In addition, lines are added to the header of the spectrum to report the scan numbers replaced by this spectrum. The software also logs grouped scan numbers to both a tab-delimited text file and to an input file for GraphViz to aid visualization of spectral clustering. These reports make it clear to users which spectra represent assessed groups.

Database Search Configuration

Database search identification was employed to characterize MS2Grouper's effectiveness. All searches were conducted in Microsoft Windows XP Home on an AMD Athlon XP 2500+ "Barton" with 512 MB of RAM° operating° at° 333° MHz.° DBDigger° [20]° performed the reported searches in semi-tryptic mode, identifying peptides that represented canonical tryptic cleavage sites on one or both ends but not those that were nonstandard cleavages on both ends. The MASPIC°scorer°[20]°was°employed,°providing°high scores to indicate that matches were unlikely to have occurred by random. Searches incorporated four chemical peptide modifications that have been observed to occur in most samples: oxidation of Met (+16 Da), N-terminal Gln conversion to pyroglutamic acid (−17 Da), Asn succinimide formation when N-terminal to Gly (−17 Da), and N-terminal Cys formaldehyde adduction (+12 Da). Sequest version 27° (Release° 12)° for° Microsoft° Windows° [21]° was employed without protease specificity for time benchmarking purposes. The database used for the extended protein standard mixture included the pro-

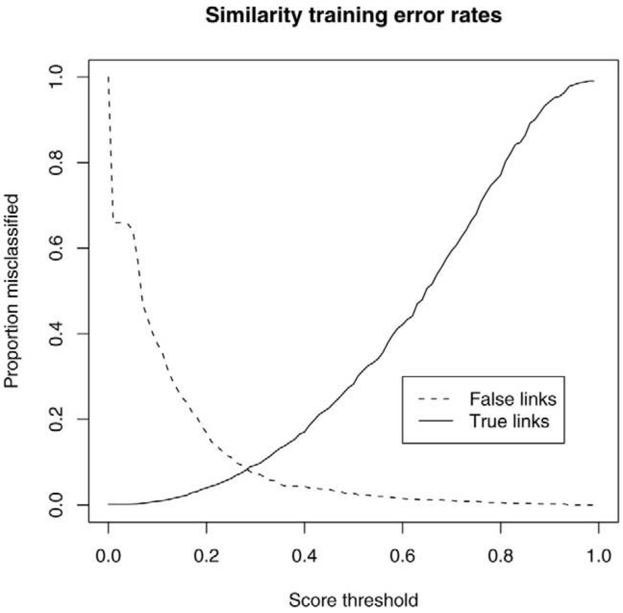


Figure 4. Two sets of spectra were produced to determine error rates in similarity detection: one in which all spectra should be linked, and one in which no spectra should be linked. The range of scores from the normalized dot product algorithm was scanned to determine the proportion of links made inappropriately (false positives) and links that it failed to make (false negatives) at each score threshold. Because the creation of inappropriate links poses a greater threat of losing identifications, we set the cutoff at 0.7, a score that should create inappropriate links only 1% of the time.

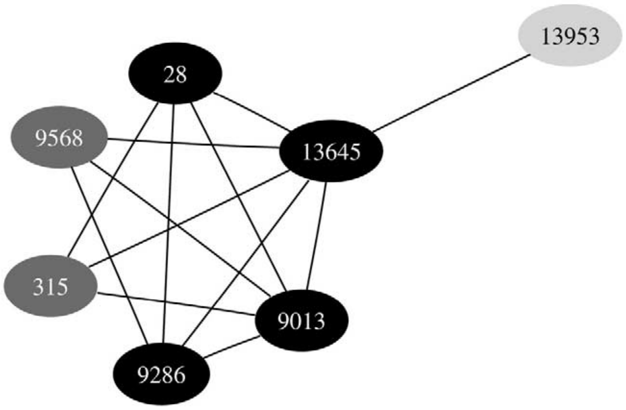


Figure 5. In this graph, each oval represents an LCQ tandem mass spectrum for the doubly-charged semitryptic myoglobin peptide YLEFISDAIIHVL. Spectra that are part of cliques are colored black, and spectra that can augment a clique by the paraclique criterion are colored slate gray. Similarity detection fails to completely interconnect these spectra despite the fact that they can all be identified to the same sequence and charge state. Group assessment can correct errors in the resulting linkages. Scans 28, 9013, 9286, and 13,645 are detected as a clique of size four. The "glomming" process adds nodes 315 and 9568 to the group because they link to three of the four members of the clique. Scan 13,953, however, is found to be similar to only one member of the clique and so is not added. When the group is replaced with a synthetic spectrum, the identification scores 80.3. If the most intense spectrum, scan 9286, represents the group, the score is 86.6. Both of these scores are well above the 95% confidence threshold of 38.5.

teins intentionally introduced to the mixture and added yeast alcohol dehydrogenase II, superoxide dismutase, trypsin, ubiquitin, as well as several keratin sequences. The 4833 ORFs of *Rhodopseudomonas palustris* [22] were added to the database to act as distracters during identification of the protein standard mixture. For proteomic identification, a database of 4798 ORFs from *Shewanella oneidensis* [23, 24] was augmented with 44 sequences for immunoglobulins, proteases, keratins, and other proteins. DTASelect [25] assembled and filtered identifications, requiring that identification scores exceeded 30.49, 38.48, and 54.37 for spectra from +1, +2, and +3 charged precursors, respectively, (these cutoffs represent 95% thresholds of confidence). Top-ranked identifications were required to exceed second-best scores by 8%. For the extended protein standard mixture analysis, *R. palustris* proteins were automatically filtered out; other proteins were allowed to appear even if only one peptide were present. DTA-Select's −DB option exported the results to Microsoft Access, which enabled differentiation of peptide identification lists.

Figure 3 illustrates the flow of data in this report. The change in performance between data that were processed by MS2Grouper and those that bypassed this step revealed the utility of the algorithm. Because MS2Grouper appears early in the pipeline, it can affect the performance of other tools, including those for inferring precursor charge states.

Results and Discussion

Similarity Detection Error Rates

The most basic capability of MS2Grouper is its ability to determine which spectra resemble each other. For this purpose, it employs a normalized dot product scorer. The third cycle of the extended protein standard mixture collected on the LTQ was observed to provide many confident identifications as well as substantial duplication. Thus, it was selected as the source of spectra for scorer validation. From the spectra with correct identifications (i.e., those that matched to proteins known to be present in the mixture), two sets of spectra were constructed. The first included 582 spectra that were captured only once by the mass spectrometer. The second set included 2237 spectra, separated into 390 groups; all were spectra appearing multiple times in this cycle of the MudPIT analysis. Because each spectrum in the first dataset was identified to a different peptide sequence, any similarities detected in this set were assessed as false positive matches. Meanwhile, the spectra in the second set were grouped with others identified to the same sequence and charge state. Any spectral pairs within these groups that were not found to be similar were assessed as false negative matches.

MS2Grouper attempted 851 spectral matches among the 582 dissimilar spectra, almost always with a low score. Among the 390 clusters of similar spectra, the software produced 11,334 matches, generally with a high score. Scaling through a set of potential threshold scores for similarity detection revealed the percentage of false positives and false negatives produced for each cutoff (see Figure 4). Because a design focus guiding MS2Grouper's development was to reduce the numbers of identifications lost, a high cutoff of 0.7 was chosen. This threshold produced a 59% false negative rate, but it yielded a very low 1% false positive rate. A more ideal way to manage separation of true and false similarities would take into account the number of comparisons expected for each spectrum, using a higher threshold for those spectra compared with larger numbers of other scans. In this case, however, we chose a simple threshold intended to work acceptably on both LCQ and LTQ data, despite the differences in numbers of spectra collected. Although this threshold neglected to link many spectra that were truly duplicates, this high cutoff prevented links between unrelated spectra. Errors in detecting similarity between spectra, however, were corrected in some cases by the system used to delineate groups of spectra (see below).

Although ion traps are among the most commonly used instruments for proteomics, mass spectrometers employing higher accuracy mass analyzers (e.g., TOF or FT-ICR) have been growing in popularity. The improved mass accuracy in tandem mass spectra

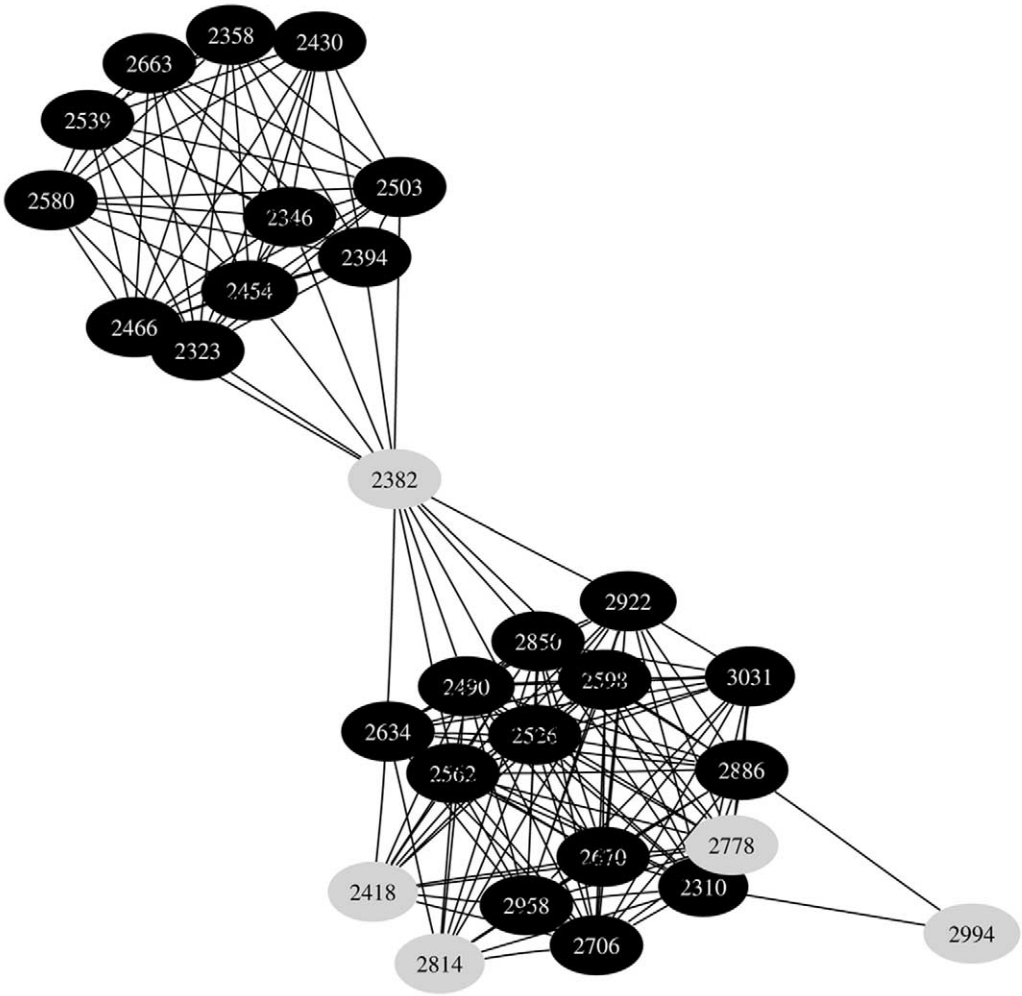


Figure 6. Two similarity groups are assessed from the LCQ spectra in this graph. The thirteen black spectra in the lower part of the figure are all identified to be the singly-charged C-terminal peptide of concanavalin: LLGLFPDAN. A synthetic spectrum for this group scores 41.8, while the best score for any single spectrum of the thirteen was 39.9. Of the eleven grouped spectra in the upper portion of the graph, only two are identified to this peptide, while the others are assigned sequences from the distractor proteins in the database search. Group assessment via clique detection prevents one group of spectra from contaminating the other.

from such instruments can reduce the extent to which fragment ions may be confused when two spectra are compared. Ideally, then, tandem mass spectra from these high mass accuracy instruments can be compared with better discrimination than ion trap spectra.

Paraclique Criterion Reduces Effect of False Negatives and False Positives

Because the threshold for similarity is set at a high value, spectra that are similar may not be linked together correctly. Figure 5 shows a group of seven spectra that have precursor *m/z* ratios that are very similar. Although their scan numbers are disparate (indicating retention times that span two hours), all seven spectra can be confidently identified to the same peptide. Despite the fact that these spectra share the same identification, similarity detection fails to connect one outlier to all but one of the other spectra. The addition of three links would cause the remaining six spectra to form a clique. The largest clique found by MS2Grouper includes four spectra. Because the other two spectra match to all but one of the four spectra in the clique, they are included in the group to form a "paraclique". This extension of the group effectively infers the links that would convert the group of six spectra into a clique. In this way, the paraclique criterion reduces the effects of false negative linkages.

Algorithms like NoDupe [8] assume that if Spectrum A is similar to Spectrum B and Spectrum B is similar to Spectrum C, A is also similar to C. In Figure 6, this policy would assess all of these spectra as duplicates of each other. MS2Grouper, however, separates the spectra into two cliques and five singletons. This mechanism prevents spectra of one peptide from being grouped

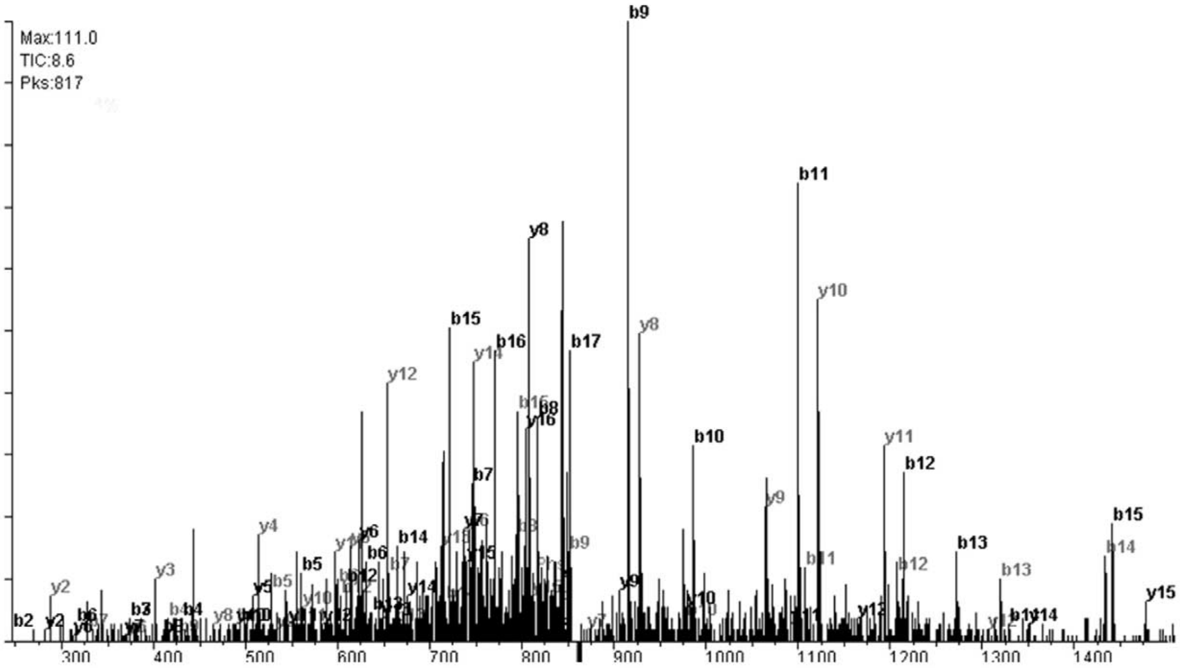


Figure 7. The fragment ions of two distinct sequences share this spectrum. Doubly-charged IMKGEADAVALDGGLVY from ovotransferrin and VEADIAGHGQEVLIRL from myoglobin have similar masses (1723.0 and 1720.97 Da, respectively), and here they have eluted from LC at the same time in the fourth cycle of the LTQ MudPIT. The ions from the ovotransferrin peptide are labeled in black while the myoglobin peptide's fragments are labeled in gray. DBDigger gives solid scores to both sequences (33.3 and 32.2), but the difference between the scores is only 3%, resulting in this identification's removal by DTASelect. Such spectra may result in spurious linkages between groups of spectra for different peptides, and so group assessment must include rules to detect these inappropriate links.

with spectra from other sequences attributable to coeluting peptides of different sequences (for an example, see° Figure° 7),° rendering° them° unidentifiable.° In° this way, requiring spectra to form paracliques protects against inappropriate similarity links.

Since the group assessment process can correct some similarity detection errors, groups assessed by MS2Grouper are more likely to contain only truly interrelated tandem mass spectra. A drawback of this system is that group assessment can only take place when all spectra are present. If a laboratory were repeatedly analyzing a particular organism, a system whereby newly observed spectra were compared to a library of identified spectra might be very beneficial. The paraclique criterion is designed to find structure in sets of unidentified spectra rather than associating unidentified spectra with known identifications.

Spectral Grouping Reduces Database Search Times

MS2Grouper's impact on overall spectral counts was evaluated by running the software on five-cycle

Table 1. Individual HPLC runs in the course of a MudPIT may vary considerably in the proportion of spectra removed as duplicates. Initial cycles appear to be most redundant, despite their reduced spectral counts. In total, the LTQ spectral counts were reduced by 15%, and LCQ spectral counts were reduced by 14.5%.

MUDPIT Cycle	Spectrum count	Spectra after grouping	Percent reduction	Singleton count	Group count
LTQ-1	7465	5704	24%	5241	463
LTQ-2	8682	6985	20%	6592	393
LTQ-3	8890	7232	19%	6834	398
LTQ-4	9246	8819	5%	8670	149
LTQ-5	8315	7683	8%	7486	197
LCQ-1	1909	1568	18%	1503	65
LCQ-2	2346	1716	27%	1603	113
LCQ-3	2215	1971	11%	1907	64
LCQ-4	2231	2078	7%	2034	44
LCQ-5	2323	2034	12%	1953	81

Table 2. Although MS2Grouper reduces the number of spectra to be identified by up to 15%, the number of peptide sequences confidently identified declines less than 1%. For the LTQ, each peptide is identified an average of 2.31 times (prior to MS2Grouper's use), while the LCQ repeats each identification an average of 1.77 times. This represents an underestimate of the true duplication rate for these spectra because more spectra for these peptides may be observed than can be confidently identified.

Instrument and MS2Grouper mode	Spectrum count	ID count	Good IDs	Peptides Identified
LTQ, no grouping	42598	63350	4330	1872
LTQ, synth spectrum	36423	55396	3401	1859
LTQ, highest signal	36423	55427	3385	1858
LCQ, no grouping	11024	16098	1774	1001
LCQ, synth spectrum	9367	14175	1556	992
LCQ, highest signal	9367	14175	1547	992
LTQ, synth spectrum	−14.5%	−12.6%	−21.5%	−0.7%
LCQ, synth spectrum	−15.0%	−11.9%	−12.3%	−0.9%

MudPIT analyses of a twenty protein mixture run once on an LCQ Deca XP 3D ion trap and once on an LTQ linear ion trap. Because of its faster scan speed, the linear ion trap collected almost four times as many spectra as the 3D ion trap during the course of the separation. The MS2Grouper analysis reveals that the five cycles of the MudPIT contained very different levels°of°redundancy°(see°Table°1);°the°reduction°in spectral counts is much higher in the initial cycles of the MudPIT than in the later cycles.

To evaluate the impact on identification of spectral grouping, the identifications of the five-cycle MudPITs were°assessed°*en masse* rather°than°separately°(see°Table 2).° Although° spectral° counts° were° reduced° by° 14.5–15.0%, the number of identifications declined by 11.9–12.6%. The reason that the number of identifications diminished less was that spectra from singly charged precursors (many of which may not be peptides) were more likely to be duplicates than those from multiply charged precursors, and the spectra from multiply charged precursors were identified twice, first assuming a +2 precursor charge and then assuming +3. The reduction in identification counts corresponded to a reduction°in°database°search°times°(see°Table°3).°The time required to run MS2Grouper (less than a minute per MudPIT cycle, even on LTQ data) was insignificant in comparison with the amount of time required for the database searches.

MS2Grouper reduced the numbers of identifiable spectra more than it reduced the numbers of other spectra. The numbers of peptides identified, however, diminished by less than 1%; for more than 99% of the identified peptides, at least one spectrum remained identifiable. This success corresponds to the chief design aim of MS2Grouper—reducing spectral redundancy without reducing peptides identified.

Synthetic Spectrum Representation Shows Both Potential and Pitfalls

As°shown°in°Table°2,°the°number°of°confident°identifications was slightly higher for MS2Grouper when it synthesized spectra to represent similarity groups than when it selected the most intense spectrum in each group as a representative. At the same time, the numbers of different peptides identified were essentially the same. This indicates that the synthetic spectrum algorithm was able, in some cases, to render low quality spectra identifiable by combining them. An examination of the identifications lost° (16)° and° gained° (3)° by° using° MS2Grouper° in° its synthetic spectrum mode illustrates the weaknesses and strengths of this algorithm.

Most identifications lost by use of MS2Grouper resulted from synthetic spectra that scored slightly lower than the best individual spectrum score in the similarity group. In the set of LTQ spectra that were not processed by MS2Grouper, LIVTQT, a singly-charged semi-tryptic N-terminal peptide of β-lactoglobulin, is observed in nine spectra, all of which list this sequence as the top match, with the best scoring spectrum receiving a score of 33.3 for this sequence. When these spectra are grouped together, the synthetic spectrum scores only 28.3, which places it below the 95% confidence thresh-

Table 3. By removing spectral redundancy, MS2Grouper reduces the time required to identify spectral collections. The times reported above for DBDigger include all five cycles of the protein standard mixture MudPITs, but the Sequest times include only the third cycle due to the longer run times of these searches. DBDigger was configured to identify semi-tryptic peptides, while Sequest conducted a search without protease specificity. DBDigger becomes more efficient as spectral collections grow, and so Sequest users can benefit more from MS2Grouper's use than DBDigger users can.

Database search	Instrument	No grouping	Synth spectrum	Reduction
DBDigger of 1–5, semitryptic	LTQ	4:20:06	3:52:20	−10.7%
DBDigger of 1–5, semitryptic	LCQ	1:20:21	1:15:13	−6.4%
Sequest of 3, unconstrained	LTQ	3:36:46	3:04:11	−15.0%
Sequest of 3, unconstrained	LCQ	0:52:59	0:48:47	−7.9%

Table 4. MS2Grouper's application to proteome MudPIT data reduces the numbers of spectra by approximately one-fifth. The protein identifications most likely to be lost by use of MS2Grouper are those for which a single peptide is identified. In this test, the number of peptides required for proteins to be identified is increased from one to three peptides. As this criterion becomes more stringent, the number of proteins lost by MS2Grouper's use diminishes.

	Proteins observed		Peptides observed	
Proteome replicate and criteria	No grouping	MS2Grouper	No grouping	MS2Grouper
Run 1: 1 pep	846	832	5085	5058
Run 1: 2 peps	563	559	4828	4806
Run 1: 3 peps	420	419	4549	4530
Run 2: 1 pep	838	821	4793	4762
Run 2: 2 peps	550	547	4536	4513
Run 2: 3 peps	420	418	4282	4260

old. Reduced scores for synthetic spectra account for twelve of the sixteen lost identifications.

The loss of four other peptide identifications resulted from different causes. One synthetic spectrum's score for the correct sequence dropped enough to reduce it to the second ranking sequence for the spectrum. A spectrum for the doubly-charged peptide GDFNADCSYVTSSQWSSIR was incorrectly grouped with ten other spectra that were not individually identified, and its score plummeted from 65.9 to 25.2, underscoring the importance of grouping spectra correctly. The other two identifications lost were from spectra that contained fragment ions from two different peptides. In one group of nine spectra, all nine matched to the same two peptide sequences in the top two ranks. The single spectrum constructed from this group did not score as well as the best of the individual spectra.

Despite the loss of sixteen identifications by grouping and synthetic spectrum production, three new identifications were gained. Five spectra were all individually identified to the singly charged semi-tryptic lysozyme *C* peptide NTDGSTDYGIL, but their best individual score was 28.3, insufficient for inclusion. A synthetic spectrum created from this quintet matched the correct sequence with a score of 31.5. Similarly, the $+2$ α-amylase peptide LLDGTVVSR improved to a score of 39.8 from three spectra, of which only one was identified correctly with a score of 27.6. One of these three spectra was incorrectly assessed as resulting from a singly charged precursor when grouping does not take place. The third gained peptide grouped twelve spectra together to receive a score of 32.169. Only four of the original dozen are identified to the correct sequence, with a best score of 26.9. Thus, while the production of synthetic spectra can lead to the loss of identifications, it can also result in the gain of peptides that would otherwise be missed.

Proteomic Samples Can Also Benefit from MS2Grouper

To test MS2Grouper's effect on biologically relevant samples, we evaluated its impact on a *Shewanella oneidensis* MR-1 proteome sample. The soluble proteins were analyzed in duplicate twelve-cycle MudPIT experiments. The two experiments generated 27,919 and 27,380 spectra, respectively. The resulting spectra were processed twice, once employing MS2Grouper and once without (see Figure 3). MS2Grouper reduced the spectral counts in the first replicate by 19.6% and in the second replicate by 21.4%. Again, some parts of the MudPIT contained more duplication than others; in this case, the second and third cycles' spectral counts for both replicates dropped by more than 34%, while the first cycle diminished by the lowest percentage, 14%, perhaps because the peptides that passed through the SCX material without being retained were more numerous in this sample than in the protein standard mixtures. Because MS2Grouper reduced these counts, the DBDigger semitryptic searches took less time, declining from an average of 5:44:45 (h:mm:ss) to 5:04:56. Because of its reorganization of the database search logic, DBDigger would be expected to benefit less from this 20% reduction in spectra than other search programs such as SEQUEST, whose search times scale approximately linearly with number of spectra.

DTASelect filtered the identifications produced for each replicate at the same score thresholds as used for the extended protein standard mixtures. Different numbers of identifications were retained as the number of peptides required for each protein to be included was varied during DTASelect filtering (see Table 4). Closer attention was directed to the identifications for proteins from which at least two peptides were observed. A total of 4841 different peptides were identified from the first replicate, with 35 found only when spectra were ungrouped and 13 appearing only if MS2Grouper was employed. These numbers are echoed by the second replicate: of 4543 identified peptides, 30 were exclusive to the ungrouped run and 7 were revealed only when MS2Grouper was used. For both replicates, the proteins listed using MS2Grouper were subsets of the proteins listed when the algorithm was not employed. Although the software reduced the numbers of spectra requiring identification by approximately one-fifth, the number of reliable peptide identifications was substantially unchanged; if the process of representative spectrum synthesis can be improved, the number of reliable identifications gained may even exceed the number that are lost.

Conclusions

Although bioinformatics research has produced powerful tools for identifying proteomic tandem mass spectra, many other areas have been less well explored. MS2Grouper focuses on one of these steps in the workflow and builds upon existing algorithms to illustrate the structure inherent in proteomic datasets. While the initial use of recognizing spectral redundancy is to reduce the number of peptide identifications to be performed, this information can be leveraged for a variety of other uses including evaluating chromatographic separations and targeting grouped spectra for more intensive algorithmic analyses. As the composition of spectra in these collections becomes better understood, the bioinformatics workflows supporting proteomics are certain to benefit.

Future directions for this research can improve the MS2Grouper algorithm and assess its findings. Although normalized dot products are both rapid and sensitive for similarity detection, it may be possible to reduce error rates by creating systems that provide probability-based assessments of similarity (like algorithms for scoring theoretical spectra against experimental ones in database identification°[26°–28].°As°noted°above,°the°algorithm°currently used to synthesize representative spectra for similarity groups often produces spectra that do not score as well as the most intense spectra in these groups. A system that takes a more holistic approach to matching peaks among multiple spectra would be likely to give better results. As the process of synthesizing such spectra improves, new opportunities will arise to improve precursor charge state detection, quality filtering, and de novo sequence inference.

Acknowledgments

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 and support by the ORNL Laboratory Director's Research and Development Program, U.S. Department of Energy, under contract no. DE-AC05-00OR22725 with Oak Ridge National Laboratory managed and operated by UT-Battelle, LLC. DLT thanks John Yates for suggesting the similarity groups may be replaced by synthetic representatives. DLT also thanks Nicole Baldwin and Michael Langston for their patient explanations of clique enumeration and extension algorithms. Conversations with Amy-Joan Ham and Sean Davey helped initiate this project. The authors thank the Dorothea Thompson Laboratory of Environmental Sciences Division, ORNL, for providing *Shewanella oneidensis* cell pellets for this study and Greg Hurst for his critical reading and suggestions for the manuscript.

References

1. Cantin, G. T.; Yates, J. R., III. *J. Chromatogr. A.* **2004,** *1053,* 7–14.
2. McDonald, W. H.; Yates, J. R., III. *Curr. Opin. Mol. Ther.* **2003,** *5,* 302–309.
3. Wu, C. C.; Yates, J. R., III *Nat. Biotechnol.* **2003,** *21,* 262–267.
4. Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999,** *17,* 676–682.
5. Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001,** *19,* 242–247.
6. Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001,** *73,* 5683–5690.
7.°ThermoElectron° Product° Support° Bulletin° 105.° http://www.thermofinnigan.com.
8. Tabb, D. L.; MacCoss, M. J.; Wu, C. C.; Anderson, S. D.; Yates, J. R., III. *Anal. Chem.* **2003,** *75,* 2470–2477.
9. Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 859–866.
10. Gan, F.; Yang, J.-H.; Liang, Y.-Z. *Anal. Sci.* **2001,** *17,* 635–638.
11. Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 85–88.
12. Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* **2001,** *73,* 1676–1683.
13. Beer, I.; Barnea, E.; Ziv, T.; Admon, A. *Proteomics* **2004,** *4,* 950–960.
14. Yates, J. R., III; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998,** *70,* 3557–3565.
15. McDonald, W. H.; Ohi, R.; Miyamoto, D.; Mitchison, T. J.; Yates, J. R., III *Int. J. Mass Spectrom.* **2002,** *219,* 245–251.
16. 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.
17. 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.
18. Bron, C.; Kerbosch, J. *Commun. ACM* **1973,** *16,* 575–577.
19. Chesler, E. J.; Lu, L.; Shou, S.; Qu, Y.; Gu, J.; Wang, J.; Hsu, H. C.; Mountz, J. D.; Baldwin, N. E.; Langston, M. A.; Hogenesch, J. B.; Threadgill, D. W.; Manly, K. F.; Williams, R. W. *Nat. Genet.* **2005,** in press.
20. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L. *Anal. Chem.* **2005,** in press.
21. Eng, J. K.; McCormack, A. L.; Yates, J. R., III *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976.
22. 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.
23. Heidelberg, J. F.; Paulsen, I. T.; Nelson, K. E.; Gaidos, E. J.; Nelson, W. C.; Read, T. D.; Eisen, J. A.; Seshadri, R.; Ward, N.; Methe, B.; Clayton, R. A.; Meyer, T.; Tsapin, A.; Scott, J.; Beanan, M.; Brinkac, L.; Daugherty, S.; DeBoy, R. T.; Dodson, R. J.; Durkin, A. S.; Haft, D. H.; Kolonay, J. F.; Madupu, R.; Peterson, J. D.; Umayam, L. A.; White, O.; Wolf, A. M.; Vamathevan, J.; Weidman, J.; Impraim, M.; Lee, K.; Berry, K.; Lee, C.; Mueller, J.; Khouri, H.; Gill, J.; Utterback, T. R.; McDonald, L. A.; Feldblyum, T. V.; Smith, H. O.; Venter, J. C.; Nealson, K. H.; Fraser, C. M. *Nat. Biotechnol.* **2002,** *20,* 1118–1123.
24. Daraselia, N.; Dernovoy, D.; Tian, Y.; Borodovsky, M.; Tatusov, R.; Tatusova, T. *Omics* **2003,** *7,* 171–175.
25. Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002,** *1,* 21–26.
26. 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.
27. Sadygov, R. G.; Yates, J. R., III. *Anal. Chem.* **2003,** *75,* 3792–3798.
28. Fridman, T.; Razumovskaya, J.; VerBerkmoes, N.; Hurst, G.; Protopopescu, V.; Xu, Y. J. *Bioinformatics Computat. Bio.* **2005,** in press.