Similarity among Tandem Mass Spectra from Proteomic Experiments: Detection, Significance, and Utility

David L. Tabb,† Michael J. MacCoss, Christine C. Wu, Scott D. Anderson, and John R. Yates, III*

SR11 Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Liquid chromatography paired with tandem mass spectrometry is a standard technique for identifying peptides from complex protein mixtures. Most fragment ion spectra acquired by this technique are unique, but some are repeated. Similarities among the spectra from 1D and 2D liquid chromatography experiments were calculated by the dot product algorithm. Similar spectra were grouped, and the degree of duplication was calculated for each sample. In 1D liquid chromatography data from 1D gel bands, 18% of the fragment ion spectra were duplicates. A six-cycle 2D liquid chromatographic separation of more than 200 proteins produced 28% duplicate spectra. A rat hippocampal homogenate analyzed by a 12-cycle 2D liquid chromatographic separation contained 25% duplicate spectra. Removal of these duplicate spectra, however, resulted in fewer peptides being successfully identified by SEQUEST. We propose a modification for peptide identification algorithms that would improve their performance and accuracy by explicitly recognizing and making use of spectral similarity.

Liquid chromatography paired with tandem mass spectrometry can produce thousands of fragment ion spectra for a complex mixture of peptides.[1−5] The mass spectrometer's control software catalogs intact peptide ions eluting from chromatography to produce an MS scan, isolates ions of a particular peptide for fragmentation, and collects the produced fragment ions in a fragment ion (MS/MS) spectrum. The control software attempts to prevent the repeated isolation and fragmentation of particular peptides in order to increase the diversity of spectra acquired. Thermo Finnigan's "dynamic exclusion" feature,[6] for example, maintains a list of the precursor m/z values that have been fragmented during the last several seconds. Peptide ions that are listed will not be fragmented. Fragment ion spectra may be repeated despite these features. For example, peptides that elute over a period of several minutes exceed the duration of exclusion and may be duplicated. In addition, a peptide mixture may be of sufficient complexity to yield more peptide ions within the exclusion duration than the list can hold. As a result, some degree of duplication can be expected among the MS/MS spectra for these experiments.

The spectra captured during an experiment pass through several steps before sequence identification. The instrument control software will first compose the spectra by averaging multiple microscans and centroiding the peaks. The MS/MS spectra must then be separated from the MS spectra produced in the process of liquid chromatography. For SEQUEST[7] users, this task is completed by the ExtractMS program,[8] which filters out unusable spectra on the basis of criteria such as peak count and separates singly charged peptide MS/MS spectra from those resulting from multiply charged peptides. A recently published program, 2to3, extends on ExtractMS by determining the charge state for multiply charged spectra.[9] With these processes complete, the SEQUEST algorithm can begin its task.

MS/MS spectra from the same peptide may differ from each other for several reasons. If a higher concentration of a peptide is present at one isolation than at another, the signal-to-noise difference between the spectra may cause one to be of higher quality. Variations in collision energy may produce different levels of peptide fragmentation. In addition, the mass spectrometer's detector may register higher or lower intensities due to random noise. All of these sources may yield spectra that appear different even though they represent the same peptide precursor ion.

* Corresponding author: (phone) 858 784-8876; (fax) 858 784-8883; (e-mail) jyates@scripps.edu.

† Current address: Department of Genome Sciences, University of Washington, Seattle, WA 98195.

(1) Lipton, M. S.; Paša-Tolić, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolić; Udseth, H. R.; Venkateswaran, A.; Wong, K.-K.; Zhao, R.; Smith, R. D. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11049−11054.

(2) Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hayes, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R., 3rd. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11969−11974.

(3) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd. *Nat. Biotechnol.* **2001**, *19*, 242−247.

(4) Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J. *Nature* **2002**, *419*, 520−526.

(5) VerBerkmoes, N. C.; Bundy, J. L.; Hauser, L.; Asano, K. G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J. L., Jr. *J. Proteome Res.* **2002**, *1*, 239−252.

(6) Thermo Finnigan. Product Support Bulletin 105. http://www.thermofinnigan.com.

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

(8) ExtractMS. http://fields.scripps.edu/sequest/extractms.html.

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

10.1021/ac026424o CCC: $25.00
Published on Web 00/00/0000

If spectra show significant similarity to each other, they generally represent the same peptide. The intensities of individual peaks may vary considerably from spectrum to spectrum, but the m/z values of fragment ions can be measured to within a single m/z in most mass spectrometers. If the primary fragment ions in a pair of spectra are at the same m/z locations but vary in intensity, the spectra can be judged as resulting from the same peptide. In some cases, such as the presence of labile posttranslational modifications, neutral losses from the precursor ion are the dominant peaks in spectra, reducing the amount of intensity to be found in informative fragment ion peaks. Such spectra may appear as quite similar while representing different peptides.

Identifying peptides from spectral collections is often the rate-limiting step for proteomic experiments. Peptide identification algorithms may consume several seconds or minutes for each spectrum. If duplication is ubiquitous among proteomic data, peptide identification algorithms should be modified to explicitly recognize and handle similarity. The time taken to analyze spectra could be reduced by handling similar spectra simultaneously. Likewise, spectra that are similar to each other could be combined to improve the overall signal-to-noise ratio for peptide identification. Spectral similarity has implications for both the performance and accuracy of peptide identification algorithms.

Techniques for determining the degree of similarity between spectra have been used for searching libraries of reference spectra to identify experimental ones.[10−14] The technique of cross-correlation is employed by LIBQUEST[13] with high sensitivity, but the algorithm is CPU-intensive. The approaches used in electron ionization mass spectrometry for library searching are generally faster than cross-correlation but may not be as sensitive. The dot-product comparison (also called the "spectral contrast angle") fared best in Stein and Scott's library search algorithm comparison.[10] In a study by Wan et al.,[12] the technique was shown to be effective at differentiating very similar oligonucleotide fragment ion spectra.

The dot-product comparison builds a vector in multidimensional space for each of two spectra being compared and determines the angle between the vectors. Higher angles imply greater differences between the spectra, and angles approaching zero indicate considerable similarity between the spectra. Peptide fragment ion spectra contain more peaks than the spectra typically used with dot-product comparison. The algorithm loses discriminatory power if large numbers of peaks are included in the spectra to be compared, necessitating a peak selection process to reduce spectral complexity prior to similarity analysis. When the peptide sequence corresponding to each spectrum is known, selection of significant fragment ions is relatively straightforward. Seeking similarity among uninterpreted spectra, however, requires different means for peak selection.

We adapted the dot-product algorithm to group uninterpreted peptide tandem mass spectra by similarity. The resulting software automatically selects a subset of peaks from each spectrum for use in comparison. It infers clusters of similar spectra and can retain a representative from each group while removing duplicates. The algorithm was used to analyze spectra resulting from 1D and 2D liquid chromatography (MudPIT). We show the impact of removing duplicate spectra on SEQUEST results and propose a modification for peptide identification algorithms that would improve the speed of identification while diminishing the occurrence of false positive matches.

EXPERIMENTAL SECTION

Materials. All chemicals were purchased from Sigma (St. Louis, MO) unless otherwise noted.

(1) Gel Band Samples. Cells of a stable HEK 293 (human embryonic kidney) cell line were lysed and subjected to two-step affinity chromatography to purify a multiprotein complex.[15] A 1D gel separated the proteins, and Coomassie dye stained the complex constituents. The protein bands were cut from the gel, reduced by dithiothreitol, and alkylated with iodoacetamide. An in-gel digest with trypsin (Promega sequence-grade trypsin) eluted peptides from the gel pieces.

(2) Microtubule-Associated Protein Sample. The microtubule-associated proteins (MAP) sample was purified from bovine brains by a published protocol.[16] Proteins were denatured by 8 M urea. Disulfide bridges were reduced with TCEP and alkylated with iodoacetamide. The proteins were digested initially with EndoK-C (Roche, Basel, Switzerland) and then by trypsin (Perceptive, Foster City, CA).

(3) Rat Hippocampus Sample. Three rat brains were dissected, and regions enriched in the hippocampus were pooled, homogenized, and centrifuged.[17] Proteins were denatured with 8 M urea, and disulfide bonds were reduced with dithiothreitol and alkylated with iodoacetamide. Proteinase K was used to digest the proteins to peptides as described previously.[17]

Separation and Mass Spectrometry. The liquid chromatographic separations used for each of the three described samples differed, but the same basic materials were used. Buffer A, the low-hydrophobicity buffer, was 5% acetonitrile/0.1 formic acid. Buffer B, the high-hydrophobicity buffer, was 80 acetonitrile/0.1% formic acid. Buffer C, for producing the salt steps, was 500 mM ammonium acetate/5% acetonitrile/0.1% formic acid. Capillary columns were produced from fused-silica capillaries with outer diameters of 365 μm and inner diameters of 100 μm (Polymicro, Phoenix, AZ), with tips drawn to inner diameters of 5 μm using a CO_2 laser puller (Sutter Instruments).

The gel band protein digests were analyzed by 1D liquid chromatography. The capillary columns were packed with 7−10 cm of Aqua C18 material (Phenomenex, Ventura, CA). A Surveyor pump (Thermo Finnigan, San Jose, CA) produced the 35-min gradients. An LCQ ion trap mass spectrometer (Thermo Finnigan) acquired MS/MS spectra in a data-dependent fashion as peptides eluted from the column.

The multidimensional separations used triphasic columns packed with 7 cm of Aqua C18 material, 3 cm of Partisphere SCX

(10) Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.
(11) Gan, F.; Yang, J.-H.; Liang, Y.-Z. *Anal. Sci.* **2001**, *17*, 635−638.
(12) Wan, K. X.; Vidavsky, I.; Gross, M. L. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85−88.
(13) Yates, J. R., 3rd; Morgan S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557−3565.
(14) Hansen, B. T.; Jones, J. A.; Mason, D. E.; Liebler, D. C. *Anal. Chem.* **2001**, *73*, 1676−1683.
(15) Canettieri, G.; Morantte, I.; Guzmán, E.; Asahara, H.; Herzig, S.; Anderson, S. D.; Yates, J. R., 3rd; Montminy, M. *Nat. Struct. Biol.* **2003**, *10*, 175−181.
(16) Mitchison Lab Protocols. http://mitchison.med.harvard.edu/protocols/tubprep.html.
(17) Wu, C. C.; MacCoss, M. J.; Howell, K. E.; Yates, J. R., 3rd. *Nat. Biotechnol.*, in press.

(Whatman, Clifton, NJ), and 3 additional cm of Aqua C18 material.[18] The MAP sample digest was analyzed using a 6-cycle separation as described previously,[18] whereas the peptides from the hippocampal homogenate digest were analyzed by a 12-cycle separation.[17] Peptides were electrosprayed directly from the column into either an LCQ or an LCQ Deca mass spectrometer (Thermo Finnigan) by the application of a distal voltage (2.4 kV) directly to the solvent. The instrument acquired one full-scan mass spectrum (400−1400 m/z) followed by three data-dependent MS/MS spectra at a 35% normalized collision energy continuously throughout each step of the multidimensional separation. Dynamic exclusion was configured to minimize the number of replicate MS/MS spectra by excluding the m/z of the previous 25 precursors selected for colision-induced dissociation.

Peptide Identification and Assembly. The 2to3 algorithm[9] was applied to the obtained MS/MS spectra to remove spectral copies with incorrect charge-state assignments. The normalized version of SEQUEST[19] was used to identify the spectra, using monoisotopic masses for fragment ions. Because all the samples had been reduced and alkylated, a mass of 160 was used for all cysteines rather than 103 in the SEQUEST search. The gel band spectra were analyzed against the RefSeq *Homo sapiens* sequence database.[20] The MAP sample was identified with a custom database including 1180 proteins drawn from RefSeq. The hippocampal homogenate was matched to a database consisting of RefSeq's Homo sapiens, rat, and mouse databases.

DTASelect[21] assembled and filtered the identifications, removing spectra with normalized XCorrs below 0.3, retaining spectra that matched their identifications well, and removing those that were identified poorly. This threshold passes ∼12% of the identifications in MudPIT results. In addition, DTASelect was configured to require identifications to have sequences of at least six residues, and the top sequence score for each spectrum was required to exceed the second best score by 8%. Proteins with only one peptide passing these criteria were included; multiple peptides were not required for protein inclusion. If multiple spectral copies of the same sequence, precursor charge, and modification were retained, they were counted as a single peptide.

NoDupe Algorithm. Software named "NoDupe" was created in the Java programming language to analyze spectral similarity. To reduce the complexity of the spectra, NoDupe preprocesses the spectra before comparing them. The spectra are grouped on the basis of their similarities, and a report is created for review via spreadsheet. NoDupe can optionally remove the duplicate spectra from each liquid chromatography run, keeping one spectrum from each cluster of similar spectra.

NoDupe reads spectra from SEQUEST's unified MS/MS file format.[22] All fragment ion spectra produced for a cycle of chromatography are read into memory. The fragment ions are assigned to bins 1.000 57 m/z wide to abstract away minor variations in recorded m/z values for peaks. If two succeeding peaks fall within the same m/z bin, their intensities are added together. Each peak's intensity is normalized by the sum of the peak intensities for the spectrum. To emphasize smaller peaks, the square roots of all intensities are used.[10]

Because a large proportion of the peaks in peptide fragment ion spectra are very low in intensity, NoDupe removes these peaks to prevent dilution of the similarity measurements. The sum of the intensities' square roots is calculated, and the peaks are sorted in order of decreasing intensity. Peaks are accepted into the final list of peaks until their sum is greater than half the sum of square roots. Spectra in which the intensity is concentrated in very few peaks will have a larger proportion of peaks removed than those in which intensity is spread over a larger number of major peaks. See Figure 1 for examples of the preprocessing results.

Once all spectra are preprocessed, the scans are sorted by precursor m/z. The spectral contrast angle is computed for pairs of spectra with precursors within 3 m/z of each other. The angle equation is defined as

$$\cos\theta = \sum i_A i_B / \sqrt{\sum {i_A}^2 \sum {i_B}^2}$$

where θ is the spectral contrast angle, i_A is a peak intensity from spectrum A, and i_B is a peak intensity from spectrum B. In essence, if both spectra have a peak at a particular m/z, the intensities are multiplied together and added to the first sum. For all peaks in either spectrum, the intensity is squared and added to either the second or the third sum. Only those peaks found in both spectra will contribute to the top sum.

If two spectra are identical, their angle will be zero, while two completely dissimilar spectra give a right angle ($\pi/2$ rad). The spectral contrast angle is commutative; a pair of spectra will yield the same angle whether (A) is compared to (B) or vice versa. The comparison can be written as a cosequential algorithm; the angle can be computed in time proportional to the number of peaks included. See Figure 2 for examples of this measure.

As shown in Figure 3, there was not a clear delineation between significant and insignificant spectral contrast angles. Spectra that are dissimilar to all others form a mass of high spectral contrast angles at the top of the figure, but the extent of this mass is ambiguous. Groups of spectra that SEQUEST identified as representing the same peptide sequence were analyzed visually to determine the maximum spectral contrast angle likely to indicate genuine similarity. A similarity angle cutoff of 1.1 rad (∼60°) was chosen to divide significant from insignificant spectral contrast angles. A lower cutoff would be more selective about which spectral pairs are named as similar, but the normal variation of fragment ion spectra sometimes produces pairs that yield 1.1 rad spectral contrast angles.

For each spectrum, the number of spectra matching with a similarity angle below 1.1 rad is recorded as the spectrum's "match count". A spectrum is marked as a duplicate if a similar spectrum has a higher match count. If the two most representative spectra for a group have the same match counts, the one for which the larger proportion of peaks was removed during preprocessing is retained.

(18) McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R., 3rd. *Int. J. Mass Spectrom.* **2002**, *219*, 245−251.

(19) MacCoss, M. J.; Wu, C. C.; Yates, J. R., 3rd. *Anal. Chem.* **2002**, *74*, 5593−5599.

(20) National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/.

(21) Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd. *J. Proteome Res.* **2002**, *1*, 21−26.

(22) SEQUEST Unified File Format. http://fields.scripps.edu/sequest/SQTFormat.html.

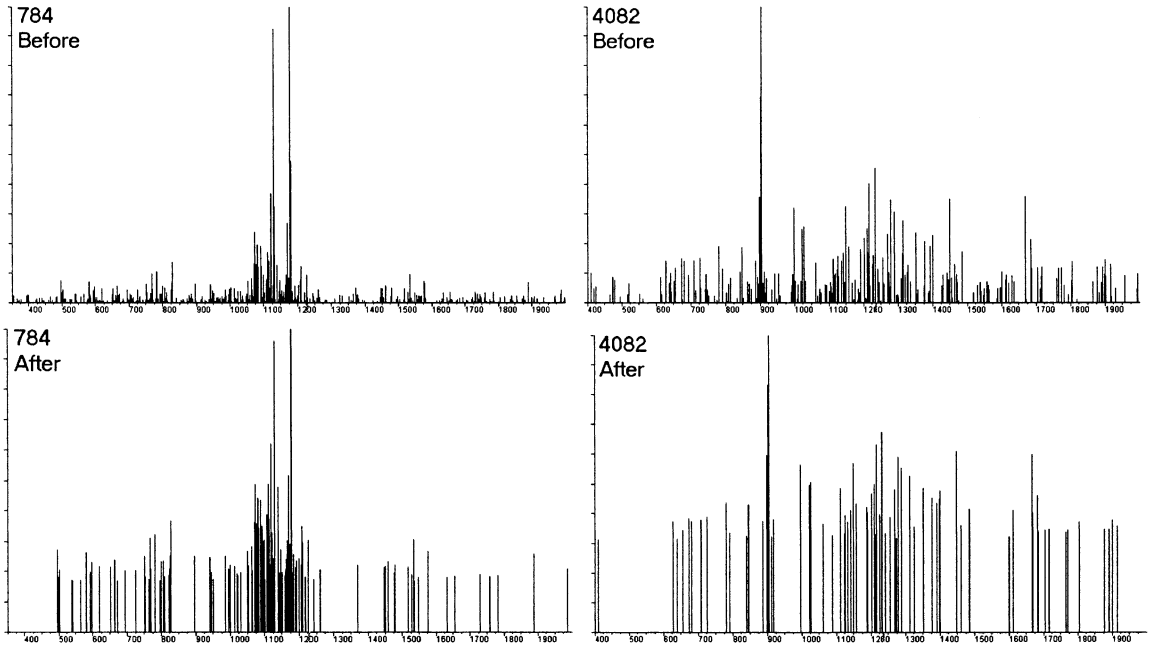


Figure 1. Peptide fragment ion spectra. Preprocessing may alter spectra considerably. The squre roots of the intensities are used, resulting in increased significance for less intense peaks. Because only the most intense peaks are retained after the intensity quota, peak counts are reduced. Scan 784 reduced from 478 peaks to 120. Scan 4082 reduced from 235 to 77 peaks. All other figures show spectra prior to preprocessing.

RESULTS AND DISCUSSION

Because the extent of duplication among acquired spectra may depend on sample type and chromatographic technique, the similarity algorithm was tried on three different sets of data. The simplest sample included 18 reversed-phase gradients on in-gel digests of 1D gel bands (see Table 3). For a sample of intermediate complexity, a six-step MudPIT identifying more than 200 microtubule-associated proteins was processed (see Table 1). To gauge results for a complex sample, an unfractionated rat hippocampal homogenate was analyzed by 12-step separation (see Table 2).

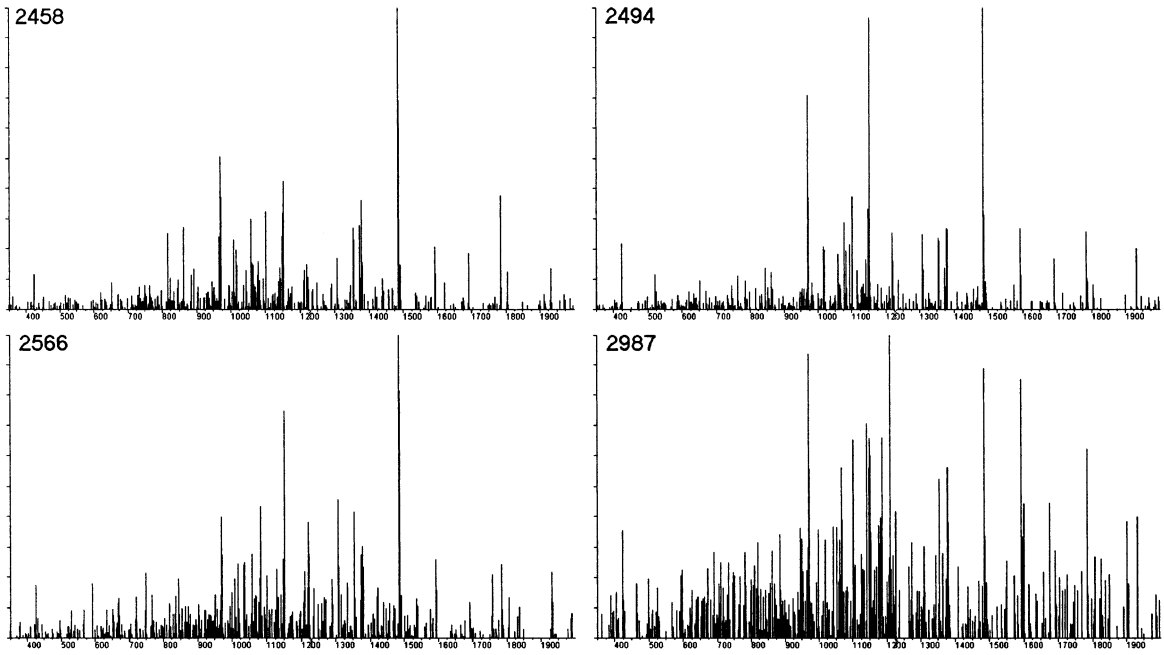
Preprocessing. The three samples differed in complexity, chromatography, and mass spectrometry. One way to compare them is by the number of tandem mass spectra that were produced per minute of separation. In the gel bands, an average of 18.3 spectra were produced for each minute of separation. This average was 24.2 for the six-cycle MudPIT of the MAP proteins. The rat hippocampus 12-cycle MudPIT yielded a higher density of sampling at 34.6 spectra/min. One cause of this diversity was the sample complexity; among the gel bands, the sample with the fewest peptides present also yielded the fewest spectra, averaging 9.6 spectra/min. The least complex MudPIT cycles were equivalent to the most complex gel band separations in spectra produced per minute.

The spectra varied considerably in the numbers of peaks they included. The doubly charged actin peptide VAPEEHPVLLTEAPLNPK appeared 35 times among the gel band data, the most for any peptide. The average number of peaks for these spectra was 303 with a standard deviation of 77. After the preprocessing in NoDupe, the average number of peaks in these spectra decreased to 74, and the relative standard deviation diminished from 26% to 20%. The preprocessing step removed ~70% of the peaks in spectra. See Figure 1 for two additional examples of the effects of preprocessing.

Spectral Similarity Characterization. Spectral duplication was common in these collections. Among the gel bands, 23% of spectra yielded angles below 1.1 rad similar to at least one other spectrum in the same band. For MudPIT results, the rate was even higher. In the MAP sample, 36% of the spectra were similar to another within the same salt cycle, and 33% of the spectra in the hippocampus sample were similar to others within the same MudPIT cycle. Individual reversed-phase gradients varied considerably from these percentages; gel band five contained only 8% similar spectra, while 60% of the spectra from the first cycle of the MAP sample were similar to others within the cycle.

Clusters were assessed among the similar spectra. On average, there were 4.2 spectra per group in the gel band data, though the mean ranged from 2.7 to 12.5 spectra in individual bands. The average cluster size in the MAP sample was 4.6 spectra, and the average group size for the hippocampus sample was 4.5 spectra. The most common type of group was the spectral pair; among spectra matching at least one other in the gel band data, 35% were in pairs. In the MAP sample, 27% of spectra showing similarity were in pairs, and 30% were members of pairs in the hippocampus spectra. Excluding groups consisting of single spectra, the gel band data averaged 36 groups per band. The larger numbers of spectra in the MudPIT separations corresponded to larger numbers of groups: 215 groups per cycle in the MAP sample and 285 in the hippocampus sample. The spectrum representing the peptide ELGGY was the most common overall, appearing 120 times in the third cycle of the hippocampus analysis.

The second cycle of the MAP sample MudPIT showed typical similarity for this sample. Of its spectra, 34% were similar to others, and 26% could be removed as duplicates. The best similarity angle for each of the 2761 spectra in this cycle is displayed in Figure 3. Approximately two-thirds of the spectra showed insignificant



Spectral Contrast Angles

	2494	2566	2987
2458	0.812	0.952	1.083
2494		0.967	1.098
2566			1.205

Figure 2. Four spectra from the sixth cycle of the MAP sample MudPIT. Scans 2494 and 2458 are the closest match with an angle of 0.812 rad. Scan 2987 is different enough from the others that it bears only marginal similarity to scans 2458 and 2494. Although 2566 rates as similar to both 2458 and 2494, it is not similar enough to scan 2987 to be grouped with it.

similarity to any spectrum and formed a mass at the top of the figure.

Peptide Identification and Similarity Clustering. NoDupe was used to test the effect of removing duplicate spectra on SEQUEST results. If retaining only the most representative spectra from spectral clusters did not result in peptide identification loss, a considerable amount of processing time could be saved. The results, however, showed that high-confidence peptide identifications are lost when duplicates are removed. Of the identifications with normalized XCorrs above 0.3, 4% were removed among the duplicates in the gel band data, 12% no longer appeared in the MAP sample MudPIT, and 14% were removed in the rat hippocampus MudPIT. A comparison of proteins identified with and without duplicate spectrum removal shows a similar trend; 5% of the gel band proteins were no longer identified, while 9% and 19% of MAP and rat hippocampus proteins were lost. Simple removal of duplicate spectra resulted in lost identifications.

Figure 4 gives an example of an identification that is lost if NoDupe removes duplicates. Scans 452 and 491 score as similar to each other with an angle of 0.847 rad. Since there are only two spectra in this group, choosing the more representative is arbitrary. NoDupe retains the spectrum with the largest proportion of peaks removed, and so scan 491 is preferred to 452 (21% of peaks remaining vs 24%). Since pairs are the most common group size, the means by which ties are broken is significant in determining the loss of peptide identifications.

The peptide IVQVVTAEAVAVLK is represented by 185 spectra in the MAP sample (see Figure 5). Five of the six cycles of chromatography include this peptide: 9 in cycle two, 11 in cycle three, 19 in cycle four, 58 in cycle five, and 88 in cycle six. Within cycle six, the spectrum assigned the best normalized XCorr (0.754) is scan 2302. NoDupe, however, selects scan 2014 as most representative (both 2014 and 2302 match to 86 other spectra in the cycle, and so the proportion of peaks retained after preprocessing is used to break the tie). Scan 4892 is assigned the same sequence as the others by SEQUEST, but NoDupe did not find it sufficiently similar to any of the other spectra for grouping. In this example, NoDupe's grouping corresponded very closely to SEQUEST's results.

Although NoDupe compares spectra within an individual liquid chromatography separation, it is apparent that spectra in multiple separation cycles can be similar to each other. In the MAP sample MudPIT (see Table 1), 1044 different peptides were observed overall, but the sum of peptide identifications for each cycle was 1253; 209 of the identifications were identical to those in other cycles. Similar results were observed for the hippocampus MudPIT (see Table 2), where 4308 different peptides were observed overall, but the sum of identifications for individual

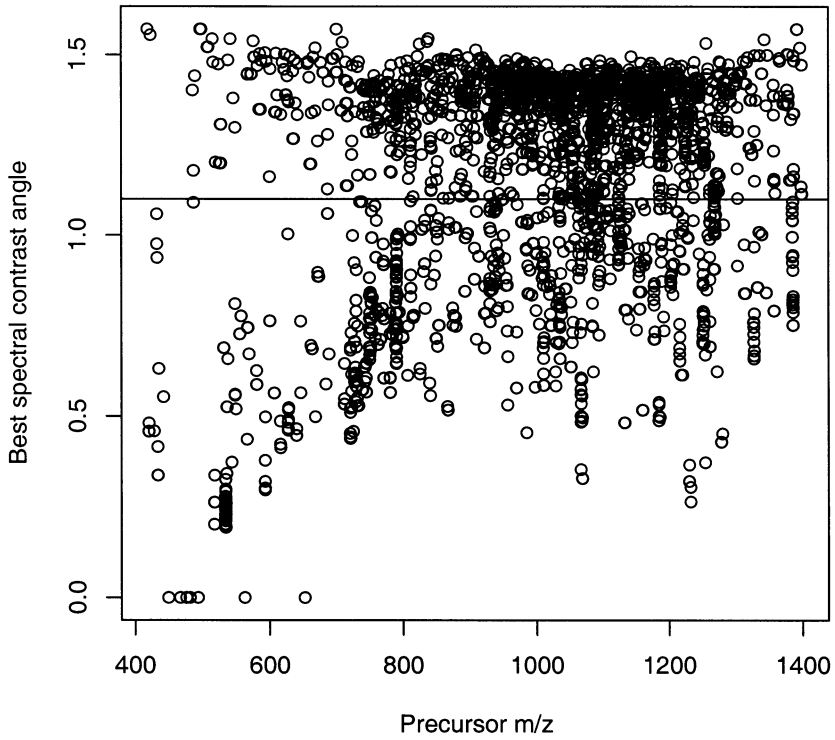


Figure 3. NoDupe analysis of the second of six cycles in the MAP sample MudPIT. Each of 2761 spectra is represented by a circle, positioned to indicate the angle to the best matching spectrum in the cycle and the *m*/*z* of the peptide ion that yielded the spectrum. Large groups of spectra for the same peptide may form vertical streaks of overlapping circles. Only the matches below the line at 1.1 rad are considered significant by NoDupe. Spectra too distant from neighboring spectra may not be compared to any spectra and are shown here as having zero angles.

Table 1. Six-Cycle MudPIT Separation of Microtubule-Associated Proteins[a]

cycle	salt concn (%)	spectra	% matched	% duplicate	Pep before	Pep after
1	0	2891	60	47	174	114
2	10	2761	34	26	204	177
3	25	2445	42	36	147	123
4	50	2770	27	19	296	288
5	80	2810	25	19	274	255
6	100	2956	26	21	158	142
all		16633	36	28	1044	916

[a] Of the 1044 identified peptides, 209 were found in multiple cycles of chromatography. Twelve percent of the identifications were lost if only one representative spectrum was retained from each group of similar spectra. The initial cycle of the MudPIT included many more duplicate spectra than other cycles.

Table 2. Twelve-Cycle MudPIT Separation of Rat Hippocampal Lysate[a]

cycle	salt concn (%)	spectra	% matched	% duplicate	Pep before	Pep after
1	0	3528	37	31	214	166
2	10	3219	39	35	154	110
3	15	3269	40	36	174	133
4	20	4041	37	29	565	503
5	25	4175	37	28	555	507
6	30	4241	30	21	467	425
7	35	4233	28	20	449	393
8	40	4191	30	21	510	449
9	45	4153	30	21	526	448
10	50	4072	26	18	467	407
11	60	4049	24	17	426	381
12	100	3894	37	32	364	252
all		47065	33	25	4308	3685

[a] Fourteen percent of the peptide identifications were lost when duplicate spectra were removed. Of the 4308 peptide identifications, 563 were found in multiple cycles. As seen in the MAP sample MudPIT, the initial cycles contained a higher proportion of duplicates. The final cycle shows an increase in duplication relative to the preceding.

cycles was 4871. Surprisingly, the gel band data also showed considerable overlap. A total of 776 different peptides were observed overall (see Table 3), but the sum of peptide counts for each gel band was 1060 peptides. Similarity may exist between spectra that resulted from different separation conditions.

Peptides may elute in multiple ammonium acetate salt steps due to several causes. A particular peptide's superabundance may result in chromatographic peak broadening. On a larger scale, overloading a column may reduce resolution for all peptides. Dead volume between the pumps and the column may result in apparent carryover between multiple steps of a MudPIT through delay of the highest hydrophobicity solvent mixture. The use of a step gradient rather than a linear gradient for the first dimension of separation may diminish its separative capacity;[23] the p*I* ranges of peptides eluting in adjacent salt steps generally overlap.[24] Taken together, these causes can explain the observation of individual peptides eluting in multiple salt steps.

CONCLUSION

This application of the dot-product algorithm reveals the degree of duplication present among spectra resulting from liquid

(23) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *J. Proteome Res.* **2003**, *2*, 43−50.
(24) Wolters, D. A.; Washburn, M. P.; Yates, J. R., III. *Anal. Chem.* **2001**, *73*, 5683−5690.

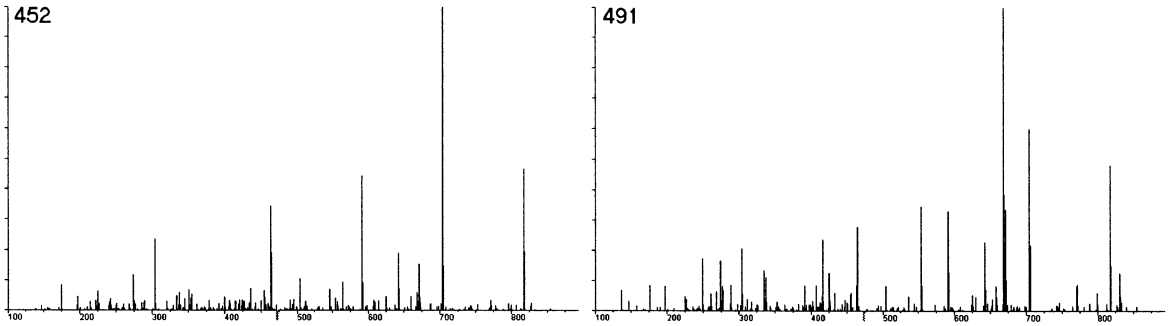


Figure 4. Scans 452 and 491. These scans yield a spectral contrast angle of 0.847 rad. Of the pair, 491 is judged by NoDupe to be the better representative. When 452 is removed from the collection, however, the peptide KLLSAEER is no longer identified (this sequence ranks third for scan 491). The differences between these spectra may be the result of a coeluting peptide in scan 491. A peptide identification algorithm that takes similarity into account could process these spectra simultaneously, saving time and retaining this peptide identification.

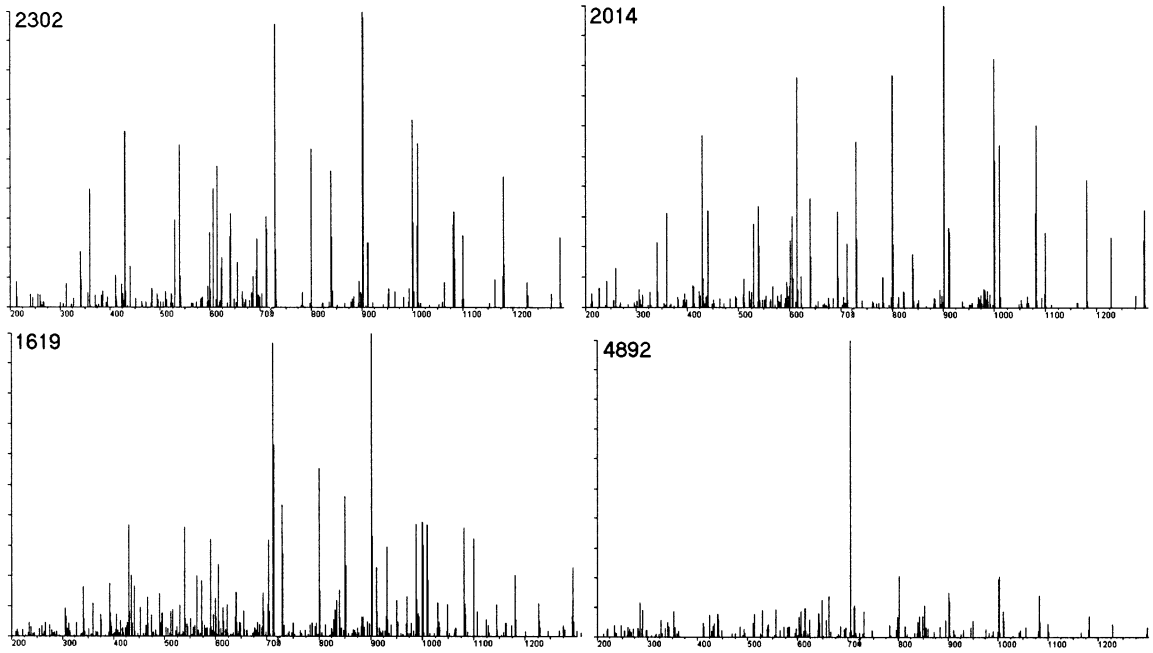


Figure 5. Scan 2302, the spectrum representing IVQVVTAEAVAVLK. This scan yields the best normalized XCorr (0.754). The spectrum NoDupe choses as most representative of the group of 88 spectra is scan 2014. Scan 1619 is shown as a variant form of this spectrum, grouped correctly by NoDupe and identified by SEQUEST as representing the same peptide sequence. SEQUEST identifies scan 4892 as being the same sequence as the others, but its similarity was insufficient for NoDupe to include it in the group.

chromatographic separations of peptides for tandem mass spectrometry. In gel bands, the proportion of duplicate spectra ranged from 5% to 45%, and MudPIT separations varied from 17% to 47%. Spectra bearing similarity to others are often members of pairs, but some clusters may be far larger in size. This use of the dot-product algorithm treats experimental spectra as the library, revealing the structure of the data before peptide identification software is employed.

Currently, SEQUEST and other peptide identification algorithms handle each spectrum as an independent entity, whether or not a spectrum is similar to others. A more efficient way to deal with duplication among proteomic spectra is to process similar spectra in parallel. If spectra are initially grouped by similarity, they can be treated as a unit during identification for substantial time savings. For example, several spectra may result from a peptide ion with a particular m/z. The precursor m/z measurements for each spectrum may vary slightly, but a similarity algorithm could note their fragment ion similarities and group them. A more accurate precursor m/z could be calculated from the multiple spectra. The peptide identification algorithm could then draw its candidate sequences from the database. Instead of finding the best sequence for each spectrum independently, though, the algorithm would find the sequence that produces the best hit against any of the spectra and then assign that sequence to all of the spectra in the group.

Such an algorithm would outperform traditional peptide identification algorithms in several ways. In a yeast database search, approximately half of SEQUEST's time is taken in searching the database for candidate peptides and calculating preliminary scores for them. This proportion increases with database size. If similar spectra are grouped, the candidate peptides can be selected from the database once for each group rather than once for each spectrum. Because the precursor m/z measurements would be more accurate, a narrower mass window could be employed for selecting candidate sequences, reducing their number. In addition, calculating preliminary scoring in

Table 3. Eighteen RPLC Separations of 1D Gel Bands[a]

gel band	spectra	% matched	% duplicate	Pep before	Pep after
1	606	28	23	65	64
2	684	27	22	59	56
3	560	26	22	64	62
4	665	15	11	62	60
5	660	8	5	66	66
6	576	22	16	62	59
7	690	18	12	44	43
8	687	11	8	53	53
9	615	12	8	90	90
10	740	22	16	55	53
11	724	27	20	60	55
12	336	49	45	22	20
13	583	23	16	56	55
14	758	30	22	62	57
15	559	19	16	42	41
16	621	20	15	84	79
17	782	32	24	78	72
18	685	38	30	36	32
All	11531	23	18	776	743

[a] A total of 776 different peptides were confidently identified from these spectra (284 were found in multiple bands). When NoDupe removed duplicate spectra prior to SEQUEST's use, 33 peptides were not identfied. This loss of 4% of the sequences resulted from removing 18% of the spectra judged to be duplicates. The retained representative spectra did not score as highly as the removed duplicates.

parallel for the group rather than serially for spectra would be more efficient. As database sizes increase, greater amounts of processing time could be saved by this technique.

An important result of processing similar spectra as groups would be that lower quality spectra are assigned sequences that correspond to the higher quality spectra in their groups. For example, if a peptide has a prominent neutral loss, some spectra may be so dominated by the precursor neutral loss that little fragment ion information is present. If these spectra are associated with spectra with more informative fragment ions, however, their sequences can be correctly assigned. As spectral collections increase in size, the chance of a hit to any protein in a database increases. Grouping the spectra by similarity before identification would help alleviate this random matching.

The differences between similar spectra may be useful for peptide identification. If one variant of a spectrum shows fragment ions more clearly at low-m/z values and another shows them more clearly in the high-m/z region, these two spectra together could yield a more accurate peptide identification than either could separately. The creation of such an algorithm would require more sophistication than the modification described above, but it may be the case that such an algorithm would yield higher accuracy in peptide identifications than is currently possible.

Similarity matching among uninterpreted spectra has other possible applications. For example, spectral libraries would be most effective if they contained representative spectra rather than randomly chosen ones. Another use would flag spectra that group by similarity but receive different sequence identifications for subsequent manual or de novo examination. Pevzner et al. suggested that similarity algorithms could match modified and unmodified variants of peptide spectra or match peptide spectra that have overlapping sequences.[25] The extent of similarity among proteomic spectral collections is a feature that proteomic software should exploit.

(25) Pevzner, P. A.; Dancik, V.; Tang, C. L. *J. Comput. Biol.* **2000**, *7*, 777−787.

ACKNOWLEDGMENT

We extend our thanks to Ianessa Morantte and Marc Montminy for allowing us to use their gel band data. We thank Ryoma Ohi and Timothy Mitchinson for supplying the MAP sample. D.L.T. thanks Michael Gross and Ilan Vidavsky for their correspondence in the creation of NoDupe. The authors acknowledge support from the National Institutes of Health in Grants RR11 823 (JRY), R33 CA81665 (JRY, DT), and F32 DK59731 (MJM). In addition, the American Cancer Society provided support in Grant PF-03-065-01-MG0 (C.C.W.).

Received for review December 13, 2002. Accepted March 12, 2003.

AC026424O