Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics

ARTICLE *in* ANALYTICAL CHEMISTRY · AUGUST 2004

Impact Factor: 5.64 · DOI: 10.1021/ac035229m · Source: PubMed

Karen Jonscher
University of Colorado
30 PUBLICATIONS **1,259** CITATIONS

William Old
University of Colorado Boulder
27 PUBLICATIONS **1,962** CITATIONS

Tom H Cheung
The Hong Kong University of Science and Tec…
22 PUBLICATIONS **1,090** CITATIONS

Natalie Ahn
University of Colorado Boulder
73 PUBLICATIONS **4,979** CITATIONS

analytical chemistry

Article

Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics

Katheryn A. Resing, Karen Meyer-Arendt, Alex M. Mendoza, Lauren D. Aveline-Wolf, Karen R. Jonscher, Kevin G. Pierce, William M. Old, Hiu T. Cheung, Steven Russell, Joy L. Wattawa, Geoff R. Goehle, Robin D. Knight, and Natalie G. Ahn

Anal. Chem., **2004**, 76 (13), 3556-3568 • DOI: 10.1021/ac035229m

Downloaded from http://pubs.acs.org on December 16, 2008

Analytical Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Improving Reproducibility and Sensitivity in Identifying Human Proteins by Shotgun Proteomics

Katheryn A. Resing,*,[†] Karen Meyer-Arendt,[‡] Alex M. Mendoza,[†] Lauren D. Aveline-Wolf,[‡] Karen R. Jonscher,[†] Kevin G. Pierce,[†] William M. Old,[†] Hiu T. Cheung,[†] Steven Russell,[†] Joy L. Wattawa,[†] Geoff R. Goehle,[†] Robin D. Knight,[†] and Natalie G. Ahn*,[‡]

Department of Chemistry and Biochemistry and Howard Hughes Medical Institute, University of Colorado, Boulder, Colorado 80309-0215

Identifying proteins in cell extracts by shotgun proteomics involves digesting the proteins, sequencing the resulting peptides by data-dependent mass spectrometry (MS/MS), and searching protein databases to identify the proteins from which the peptides are derived. Manual analysis and direct spectral comparison reveal that scores from two commonly used search programs (Sequest and Mascot) validate less than half of potentially identifiable MS/MS spectra (class positive) from shotgun analyses of the human erythroleukemia K562 cell line. Here we demonstrate increased sensitivity and accuracy using a focused search strategy along with a peptide sequence validation script that does not rely exclusively on XCorr or Mowse scores generated by Sequest or Mascot, but uses consensus between the search programs, along with chemical properties and scores describing the nature of the fragmentation spectrum (ion score and RSP). The approach yielded 4.2% false positive and 8% false negative frequencies in peptide assignments. The protein profile is then assembled from peptide assignments using a novel peptide-centric protein nomenclature that more accurately reports protein variants that contain identical peptide sequences. An Isoform Resolver algorithm ensures that the protein count is not inflated by variants in the protein database, eliminating ∼25% of redundant proteins. Analysis of soluble proteins from a human K562 cells identified 5130 unique proteins, with ∼100 false positive protein assignments.

Profiling expressed proteins in a cell type (the proteome) is now possible through the convergence of genome sequencing, automated data acquisition by mass spectrometry (MS), and database search programs. One approach, called shotgun proteomics, involves proteolysis of the proteins in a sample and then sequencing peptides by MS fragmentation (MS/MS).[1] Charge competition between peptides during ionization and limitations in MS sensitivity, dynamic range, and data collection rate require prior peptide fractionation. A major advance was the introduction of a multidimensional chromatography approach, where peptides are separated first by strong cation-exchange (SCX) chromatography and then by reversed-phase chromatography,[2] coupling the reversed-phase column to a mass spectrometer (LC/MS), so that thousands of sequencing spectra can be collected in a few hours. To identify peptides, each MS/MS spectrum is compared against theoretical spectra of candidate peptide sequences represented in a protein database, and a score is assigned to rank the most likely peptide assignments.[3] However, current scoring methods are poor at distinguishing correct from incorrect sequence assignments, leading to high false positive and false negative rates.[4] Consequently, protein identification is problematic when based on a small number of peptide assignments.[5] In such cases, peptide assignments are manually validated by visual inspection of each MS/MS spectrum. This approach has been successful in characterizing the protein composition of organisms such as *Saccharomyces cerevisiae*, where 25% of ORFs contained in the genome were observed.[6]

Higher eukaryotes present a more difficult problem, because there are more proteins in their proteomes, the proteins are larger, and the protein concentration ranges are wider; consequently, more spectra must be collected to define the proteome composition. Furthermore, because protein sequence databases are larger with more sequence redundancy, each spectrum must be compared against a larger number of candidates. Manual analysis becomes a daunting task with substantial error; therefore, validation methods are needed that can be implemented computationally. A robust solution to this problem has not been achieved, although recent reports used linear discriminant analysis[7] or machine learning algorithms[8] to evaluate the scores generated by the Sequest search program or used peptide properties other than the fragmentation pattern, such as exact mass measurements,[9] to validate peptide assignments.

* Corresponding authors. E-mail: Katheryn.Resing@Colorado.edu. Phone: 303-735-4019. Fax: 303-492-2439. E-mail: Natalie.Ahn@Colorado.edu. Phone: 303-492-4799. Fax: 303-492-2439.

[†] Department of Chemistry and Biochemistry.

[‡] Howard Hughes Medical Institute.

[1] McCormack, A. L.; Schieltz, D. M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J. R. *Anal. Chem.* **1997**, *69*, 767−776.

[2] Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. *Nat. Biotechnol.* **1999**, *17*, 676−682.

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

[4] MacCoss, M. J.; Wu, C. C.; Yates, J. R. *Anal. Chem.* **2002**, *74*, 5593−5599.

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

[6] Washburn, M. P.; Wolters, D.; Yates, J. R. *Nat. Biotechnol.* **2001**, *19*, 242−247.

[7] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.

[8] Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *2*, 137−146.

10.1021/ac035229m CCC: $27.50
Published on Web 05/25/2004

Table 1. Samples of Soluble Protein Extracts from Human K562 Erythroleukemia Cells[a]

sample and search method	no. of DTAs[b]	no. of peptides identified	no. of unique peptides	no. of proteins identified[c]	no. of proteins identified by one peptide
Sample 1: 16 SCX Fractions of Soluble Lysates, Analyzed with Full Mass range[d]					
MSPlus, normal database[e]	2 117	856 (40%)	434	243	169 (69%)
MSPlus, randomized database	2 117	33			
Sequest only, normal database[f]	2 117	680 (32%)	351	209	148 (71%)
Sequest only, randomized database	2 117	22			
Mascot only, normal database[g]	2 117	702 (33%)	377	219	152 (69%)
Mascot only, randomized database	2 117	27			
Sample 2: 11 SCX Fractions, Each Analyzed in 10 Gas-Phase Fractions[h]					
MSPlus, normal database[e]	47 598	8 190 (17%)	4 387	1 757	883 (50%)
MSPlus, randomized database	47 598	259			
Sequest only, normal database[f]	47 598	5 804 (12%)	3 275	1 433	736 (51%)
Sequest only, randomized database	47 598	181			
Mascot only, normal database[g]	47 598	5 173 (11%)	2 971	1 320	689 (52%)
Mascot only, randomized database	47 598	162			
Sample 3: 7 Gel Filtration Fractions × 16 SCX Fractions, Each Analyzed in 6 Gas-Phase Fractions[i]					
MSPlus, normal database[e]	602 520	85 267 (14%)	20 675	5 130	2 323 (45%)
Sequest only, normal database[f]	602 520	64 194 (11%)	15 217	4 120	1 790 (43%)
Mascot only, normal database[g]	602 520	63 431 (11%)	16 006	3 971	1 683 (42%)

[a] Summary of samples described in this study, showing effects of varying sample fractionation by gel filtration and gas-phase fractionation. [b] Number of MS/MS files with peptide mass between 900 and 4800 Da. [c] Proteins counted after removing redundancies with Isoform Resolver (see text and Table 3). [d] Soluble extracts (1×10^8 cells) were trypsinized and peptides separated into 16 fractions by SCX-HPLC. 6.3% of each fraction was analyzed by RP-LC/MS/MS, collecting data over a full ion mass range of 350−1,500. Tables of validated peptide sequences and scores and raw data files for sample 1 are available upon request. [e] MSPlus protocol was carried out using the normal or randomized IPI database (April 10, 2003, 48 000 entries). Acceptable assignments required peptide mass between 900 and 4800 Da, up to two tryptic cleavages, no internal KK, KR, RR, or RK sequences, no covalent modifications, and either consensus between Sequest and Mascot or XCorr and Mowse scores above unequivocal thresholds (XCorr ≥2.55 (MH^+), 3.39 (MH_2^{2+}), and 3.78 (MH_3^{3+}); Mowse = 44 (MH^+), 49 (MH_2^{2+}), and 49 (MH_3^{3+})), ion ratio >25% (MH^+ or MH_2^{2+}) or >20% (MH_3^{3+}), RSP = 1, and a number of basic residues consistent with the SCX fraction the peptide was in. [f] Sequest search protocol was carried out using the normal or randomized IPI database. Acceptable assignments required XCorr values above a threshold that yielded the same ratio of the false positive assignments to the number of identified peptides (∼3.4%) as MSPlus. Sample 1: XCorr = 2.56 (MH^+), 2.98 (MH_2^{2+}), and 3.0 (MH_3^{3+}). Sample 2: XCorr = 2.31 (MH^+), 2.75 (MH_2^{2+}), and 3.0 (MH_3^{3+}). Sample 3: XCorr = 2.3 (MH^+), 2.7 (MH_2^{2+}), and 2.8 (MH_3^{3+}). [g] Mascot search protocol was carried out using the normal or randomized IPI database. Acceptable assignments required XCorr values above a threshold that yielded the same ratio of the false positive assignments to the number of identified peptides (∼3.4%) as MSPlus. Sample 1: Mowse = 44 (MH^+), 39.6 (MH_2^{2+}), and 39.6 (MH_3^{3+}). Sample 2: Mowse = 44 (MH^+), 47 (MH_2^{2+}), and 47 (MH_3^{3+}). Sample 3: Mowse = 44 (MH^+), 44 (MH_2^{2+}), and 46 (MH_3^{3+}). [h] Soluble extracts (1×10^8 cells) were trypsinized and peptides separated into 11 fractions by SCX-HPLC. 7.5% of each fraction was used in separate RP-LC/MS/MS analyses of each of 10 overlapping narrow mass ranges: 300−558, 550−678, 670−798, 790−918, 910−1038, 1030−1158, 1150−1278, 1270−1398, 1390−1558, and 1550−1718 Da, consuming 75% of the sample. [i] Soluble extracts were separated by sizing gel exclusion chromatography into 13 fractions, with pool size based on UV profiles. Nine of the gel filtration fractions (fractions 1−3, 5, 7, 9, 11−13) were trypsinized and peptides separated into 16 fractions by SCX-HPLC. 5% of each fraction was analyzed by RP-LC/MS/MS over a full mass range of 350−1500 Da, and 15% of each fraction was analyzed over six overlapping mass ranges: 300−678, 670−798, 790−918, 910−1038, 1030−1278, and 1270−1750.

Here we describe a novel programmatic approach to data analysis that significantly improves sensitivity and confidence in peptide assignments and protein representation. Key features include (i) integrating results of two database search programs to increase detection sensitivity of correctly identified peptides, (ii) implementing new filtering criteria based on peptide chemical properties to increase discrimination between correct versus incorrect sequence assignments, (iii) focusing the search strategy to improve accuracy in sequence assignments, and (iv) developing a peptide-centric nomenclature for protein profiling to accurately report ambiguities in protein identification due to sequence redundancy in the database.

METHODS

Sample Preparation. The shotgun proteomics analyses were carried out on an extract of the erythroleukemia cell line K562 grown in suspension as previously described.[10] Cells were washed twice by centrifugation, and pellets were flash frozen in liquid nitrogen. Cell pellets were suspended in lysis buffer (140 mM potassium phosphate, pH 7.4, 150 mM NaF, 1 mM Na_3VO_4, 6 mM EDTA, 6 mM EGTA, 250 mM NaCl, 4 mM DTT) containing 40 μg/mL leupeptin, 5 μg/mL pepstatin A, 4 mM benzamidine, and 20 mM PMSF and sonicated 4 × 15 s at 4 °C (Branson, microtip probe). Lysates were centrifuged at 200000*g* for 30 min at 4 °C and soluble proteins recovered in the supernatant. Typically 10^8 cells yielded ∼15 mg of protein. In this study, three different samples were analyzed (Table 1). For samples 1 and 2, proteins were alkylated with 14 mM iodoacetamide (Aldrich, Milwaukee, WI) for 30 min in the dark at room temperature. Reactions were quenched by adding 3 mM DTT, and proteins were immediately desalted on a PD10 column (Amersham, Piscataway, NJ) equilibrated with 100 mM NH_4HCO_3, followed by trypsinization at 37 °C with 3% (w/w) trypsin (Wako, Richmond, VA, Catalog No. 20709891) added in 1% aliquots at $t = 0$, 4, and 12 h. The NH_4HCO_3 was removed by repeated lyophilization and resuspension in water (usually 3 times) until the conductivity after dilution of 10 μL of sample into 3 mL of water was less than 0.004 Ω^{-1}/cm. Lyophilized peptides were dissolved in buffer A (5 mM K_2HPO_4,

[9] Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. *Proteomics* **2002**, *2*, 513−523.

[10] Whalen, A. M.; Galasinski, S. C.; Shapiro, P. S.; Nahreini, T. S.; Ahn, N. G. *Mol. Cell. Biol.* **1997**, *17*, 1947−1958.

5% acetonitrile pH 2.5) and fractionated by HPLC using a SCX column (PolySulfoethyl A, 2.1 mm i.d. × 200 mm, Poly LC), equilibrated in buffer A, and eluted using a gradient of increasing 0.5 M KCl in buffer A. For sample 3, proteins were homogenized in lysis buffer on a gel filtration column (Sephacryl HR300 (26/60), Amersham) equilibrated in lysis buffer. The column was run at 1.3 mL/min, collecting 8-mL fractions over the included volume. Proteins in the pooled fractions were then alkylated with iodoacetamide, desalted into 100 mM NH_4HCO_3, digested, lyophilized, and processed in the same manner as samples 1 and 2.

Data Collection. For samples 1 and 2, aliquots of 5−20% of the total SCX fraction were loaded onto 250-μm-i.d. reversed-phase capillary columns fabricated in-house, using a frit made from a C_{18} Empore Disk (3M, St. Paul, MN), inserting a 50-μm-i.d., 180-μm-o.d. piece of tubing as an outlet, and then joining the two pieces of tubing with epoxy glue (EPO-TEK; Epoxy Technology, Billerica, MA). Columns were packed with Jupiter C_{18} resin (10-μm particle size; Phenomenex, Torrance, CA) using a bomb pressurized with helium gas at 100−400 psi. A multistage gradient delivered by an Agilent 1100 Series HPLC (Agilent Technologies, Santa Clara, CA) was used to elute peptides into the electrospray ionization source of an LCQ Classic ion trap mass spectrometer (ThermoElectron, San Jose, CA). Columns were loaded and washed in 0.1% formic acid (buffer A) at 10 μL/min. For sample 1, peptides were eluted with a gradient into buffer B (70:30 acetonitrile/water + 0.1% formic acid) at 5 μL/min: 0−30% B in 30 min, 30−50% B in 10 min, and 50−100% B in 10 min. For samples 2 and 3, the gradient was 0−18% B in 27 min, 18−27% B in 45 min, 27−50% B in 22 min, and 50−100% B in 7 min.

The target value for the ion trap was 5×10^8 ions in full-scan mode and 2×10^7 ions in MS/MS mode and regularly reoptimized in concert with the electron multiplier voltage to enhance the mass spectrometer performance. One full-scan mass spectrum was acquired, and then MS/MS spectra were acquired for the three most intense peaks in the MS spectrum, using a normalized collision energy of 34 units. To ensure that the same high-abundance ions were not continually analyzed, dynamic exclusion was set to exclude ion mass-to-charge ratios (m/z) from MS/MS after they were analyzed twice during a 90-s interval. After 3 min, the m/z value was removed from the exclusion list. The m/z scan range was either 350−1500 for full mass range or a sequence of overlapping m/z ranges (gas-phase fractionation[11,12]), as described in Table 1.

Data Collection. DTA files were generated from the MS/MS spectra using TurboSequest, with intensity threshold of 10 000, peptide mass tolerance of 2.5 Da (average mass), allowed grouping of 1−5 scans, and minimum ion count of 35. An in-house script concatenated DTA files into a Mascot Generic File for Mascot searches. Mascot and Sequest searches were normally carried out allowing 2.5 Da (average) peptide mass tolerance and 1.0 Da (average) fragment ion mass tolerance. Static modification of Cys and tryptic cleavage were specified for the searches; because TurboSequest allows cleavages at KP or RP, the Pro cleavage trypsin setting was used for Mascot. One incomplete cleavage was allowed for TurboSequest and two incomplete cleavages for Mascot, using the IPI human protein database (http://www.ebi.ac.uk, version 2.18, updated April 10, 2003). In analysis of oxidation or phosphorylation, database searches were carried out allowing variable modification of Met and Trp, or Ser, Thr, and Tyr, respectively. The output files of Mascot were parsed into MYSQL using a modified version of DBParser[13] (provided by Jeff Kowalak, NIH) and then into an Oracle 9i database; Sequest results were parsed directly into the Oracle database using an in-house parser.

Text files summarizing results were generated by SQL for input to in-house Perl scripts that applied filtering criteria (MSPlus), generated protein profiles (Isoform Resolver), and carried out spectral comparisons (CLASP). MSPlus compares results of Sequest and Mascot searches, considering the two charge forms of multiply charged DTA files separately, and excludes or validates the assignments according to a set of rules. The rules are applied in the following order: (1) Exclude all assignments where (a) the observed MW is <900 Da, (b) the peptide length <8 amino acids long, (c) there is an internal KR, KK, RK, or RR sequence, (d) the number of basic residues is inconsistent with elution during SCX chromatography, or (e) the Sequest ion score is greater than or equal to 20 (+3) or 25% (+1 or +2). (2) Validate any assignment scoring above threshold for Mowse or above threshold for XCorr where RSP = 1 (we observed no cases in which Sequest and Mascot disagreed and both assignments were above threshold), as described in Results. (3) Validate assignments scoring below thresholds when the following conditions are met. (a) Mascot and Sequest agree (peptide isoforms with one or two substitutions of K/Q/E, D/N/I/L, E/M, V/P, or V/T are considered identical by MSPlus; the choice of peptide isoform is made in Isoform Resolver, see below), (b) RSP = 1, and (c) Sumscore is greater than or equal to 3.5, where Sumscore = XCorr + Mowse/C [C is a normalization factor = 18 (+1), 15 (+2), or 12 (+3), obtained by linear least-squares fitting of XCorr versus Mowse values for DTA files that have identical peptide assignments by Sequest and Mascot and at least one score above threshold]. (4) Resolve ambiguities regarding ion charge for DTA files (ambiguous multiply charged DTA files were searched by assuming both +2 and +3 charge); in <2% of cases, both charge forms of a DTA were found in the validated list, and MSPlus chose the form with highest Sumscore.

Isoform Resolver uses peptides validated by MSPlus to construct a protein profile. For each peptide, protein accession numbers for all protein entries containing that peptide sequence replaces the accession number(s) assigned by the search program(s). Then, peptides are grouped according to the protein variants in which they are present. The minimum number of proteins that account for all peptides in each group are computed using a greedy algorithm. Isoform Resolver also considers peptide isoforms containing one or two amino acid replacements that are often not distinguished by ion trap mass spectrometers (D/N/I/L, K/Q/E, E/M, V/T, or V/P). If a peptide isoform is present that specifies more than one protein, all peptides are reported although only one protein is counted, favoring proteins that are supported by other peptide identifications; the other protein(s) and isoform peptides are reported in a separate list.

[11] Spahr, C. S.; Davis, M. T.; McGinley, M. D.; Robinson, J. H.; Bures, E. J.; Beierle, J.; Mort, J.; Courchesne, P. L.; Chen, K.; Wahl, R. C.; Yu, W.; Luethy, R.; Patterson, S. D. *Proteomics* **2001**, *1*, 93−107.

[12] Yi, E. C.; Marelli, M.; Lee, H.; Purvine, S. O.; Aebersold, R.; Aitchison, J. D.; Goodlett, D. R. *Electrophoresis* **2002**, *23*, 3205−3216.

[13] DBParser by J. A. Kolowak is described in http://proteome.nih.gov/SymposiumII/poster9.html.

CLASP performs a pairwise comparison of DTA files in order to identify those that are plausibly derived from identical or related peptide sequences. First, DTA files are grouped by observed mass, normalized reversed-phase elution time, and SCX elution (allowable ranges defined by user). For each pair of DTA files, a similarity score is calculated as the percentage of fragment ions in common, using only the monoisotopic forms of each ion with mass tolerance of 2.0 Da. To reduce noise, CLASP only considers fragment ions with intensities greater than the mean value of all fragment ion intensities plus 0.5 times the standard deviation. The minimum acceptable similarity score is determined from cases in which one search program made an incorrect assignment due to distraction, while the other program made a correct assignment that scored below threshold and therefore was not validated by MSPlus. CLASP was used to identify cases of distraction (see Results), by comparing DTA assignments that MSPlus failed to validate, against DTA assignments that were validated by MSPlus. DTA files of peptides that eluted within 90 scans of each other, eluted in the same or adjacent two SCX fractions, and had observed mass within 9 Da of predicted (to allow for error due to space charging) were scored for spectral similarity. CLASP was also used to identify spectra of peptides that were modified by dehydration (−18 Da), deammoniation (−17 Da), Met/Trp oxidation (+16, +32 Da), or Met side chain β-elimination (−48 Da). The nonvalidated spectra were compared against the validated spectra, allowing a mass difference appropriate to each case: −14 to −20 Da (dehydrated, deammoniated), +15 to +17 Da or +31 to +33 Da (Met/Trp oxidized), or −46 to −50 Da (Met β-eliminated). Sequences were checked to ensure the amino acid composition was appropriate for each modification. A second search condition for oxidation used variable modification of Met or Trp, which allowed estimation of the number of cases in which the unmodified peptide was absent.

Manual analysis was carried out by selecting a random subset of the data and screened initially by an experienced analyst. About one-third of spectra could be evaluated based on simple criteria as incorrectly assigned (more than three major ions were not identified) or correctly assigned (all ions with signal of >10 times noise were identified, and chemically plausible cleavages C-terminal to Asp or N-terminal to Pro were observed, when present); the remainder were independently examined by a second analyst, who also evaluated internal fragment ions and other types of fragment ions not considered by Mascot and Sequest, using Protein Prospector (http://prospector.ucsf.edu) to calculate the predicted ions. Correct assignments required that the peptide sequence accounted for >95% of the fragment ion current above background. Where necessary, the manual analysis includes a detailed assessment of chemical plausibility. Examples of manually analyzed spectra are shown in Supporting Information, Figure 1.

RESULTS AND DISCUSSION

Characterizing the MS/MS Data Set. Experiments were carried out to determine the number of peptide assignments validated by conventional search methods. Initial studies compared the effectiveness of two database search programs, Sequest[3] and Mascot,[14] in assigning sequences to MS/MS data collected on a tryptic digest of extracts from human K562 cells separated into 16 SCX fractions (sample 1). For searching, MS/MS data were summarized as text files (DTA files). Two DTA files were generated for each multiply charged parent ion by assuming that the parent is either doubly or triply charged; in this discussion, these pairs are counted as one DTA file. A commonly used approach for validation of search results is to accept all peptide assignments with scores above a certain threshold; this threshold is often determined by searching against a “randomized” protein database created by inverting each protein sequence contained in the normal database.[5] Figure 1A shows the distributions of cross-correlation (XCorr) scores for doubly charged ions of a data set searched by Sequest against both the normal and randomized database. When spectra are searched against randomized databases with Sequest, the XCorr distribution peaks at low values and falls to zero at a threshold value that determines the highest scores obtained by chance. In a normal search of similar database size, peptide assignments with scores above this threshold have a very high probability of being correct.[4,7] Threshold XCorr values of 2.56, 3.22, and 3.45 for MH_1^{1+}, MH_2^{2+}, and MH_3^{3+} ion charge states, respectively, were determined from searching a randomized IPI database using Sequest (Figure 1B, set I). When searched by Sequest against a normal database (Figure 1B, set II), only 523 (25%) of 2117 DTA files generated from sample 1 showed XCorr values above these thresholds. The presence of incorrect spectra in the normal searches is revealed by the large peak centered about XCorr ~2 (Figure 1A). Similar results were obtained using Mascot and evaluating Mowse score thresholds (data not shown).

It is known that many valid peptide assignments yield scores below threshold;[4,7] therefore, we next estimated the maximum number of peptides that should be identifiable in a given MS/MS data set, regardless of score. This was assessed in three ways. First, 18 standard proteins[15] were digested using trypsin, and peptide spectra acquired during the LC/MS/MS analyses were manually validated to determine correct identifications. Analyzed by Sequest, 54% of the validated peptide assignments for this set scored below the thresholds determined by the randomized database (Figure 1B, set III). Applying this estimate from the standards to sample 1, where there are 523 DTA files with XCorr scores above threshold, predicts that an additional ~614 DTA files with scores below threshold should be identifiable, and ~1137 of the 2117 DTA files should represent identifiable tryptic peptides in this data set (class positive). Thus, an estimated 980 DTA files cannot be identified by this search protocol (class negative).

A second way to estimate the number in the class positive is based on the understanding that shotgun proteomics involves population sampling and that stochastic processes lead to low scores and incorrect assignments. Therefore, repeated analysis should eventually allow all of the identifiable peptides to score above threshold. Indeed, in replicate analyses, the number of unique peptide assignments above threshold increased by 2-fold after extrapolating to an end point (Figure 1C). This approach predicted that at least 1046 of the total DTA files should be assignable to tryptic peptides, similar to the 1137 estimated by the first method.

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

[15] Keller, A. D.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. *OMICS* **2002**, *6*, 207−21.

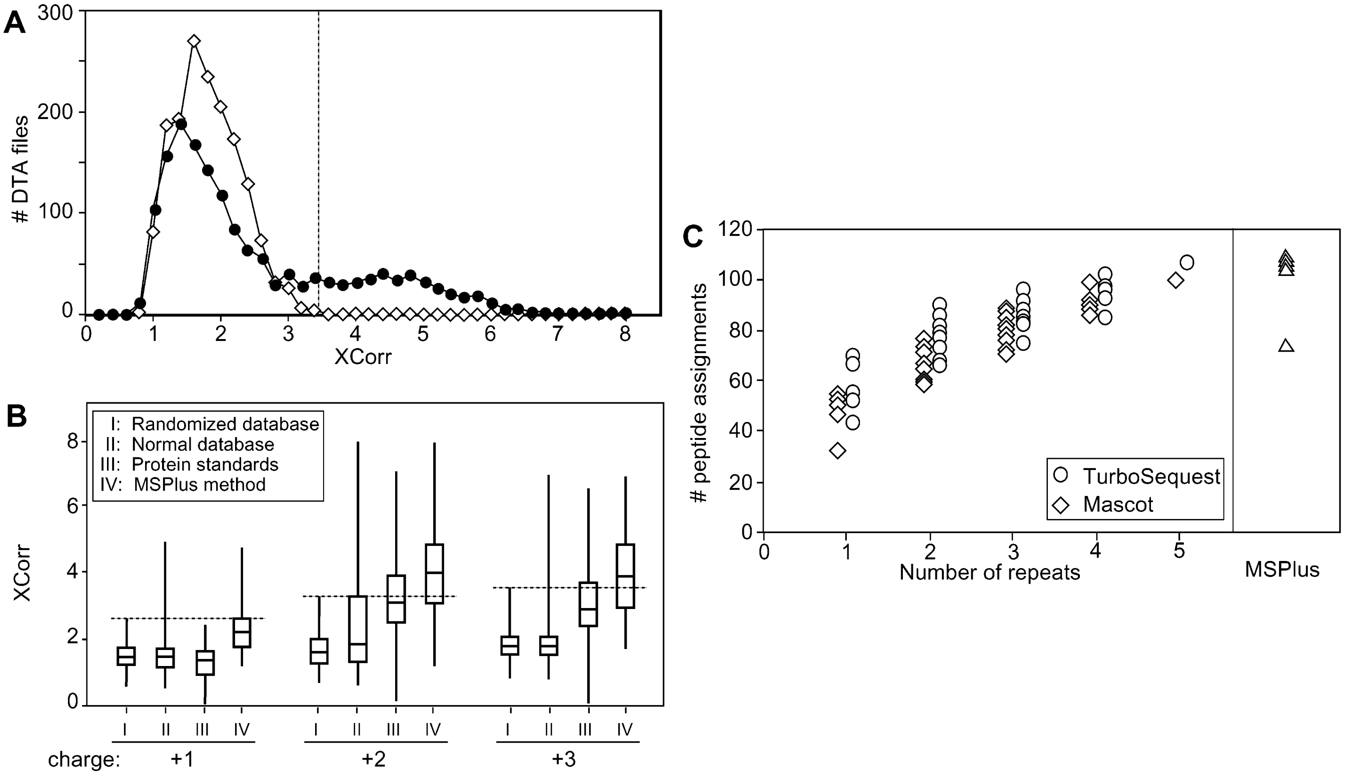

Figure 1. Randomly determined peptide scores. Only half of peptide assignments have scores greater than randomly determined. (A) Distribution of XCorr values for DTA files generated from sample 1 illustrates the method used for determining thresholds from randomized database searches. Shown are results of searching doubly charged ions (1615 files) against a normal protein database (IPI version 2.18, April 10, 2003) (closed circles) or against a randomized database in which each protein sequence in the normal database was inverted (open symbols). The highest score produced by chance in the randomized database search (XCorr = 3.22, vertical line) indicates the threshold value for doubly charged ions, above which sequence assignments can be accepted with high confidence. The highest XCorr score in the normal search was 7.93. (B) Distribution of XCorr values for singly, doubly, and triply charged ions, searched using Sequest. Box plots indicate median and quartile values (boxes) and highest and lowest scores (whiskers). (I) DTA files from sample 1 searched against the randomized database indicate thresholds of 2.56 (MH^+), 3.22 (MH_2^{2+}), and 3.45 (MH_3^{3+}) (horizontal dotted lines). (II) DTA files from sample 1 searched against the normal IPI database, where some assignments with scores below thresholds represent correct assignments. (III) DTA files representing manually confirmed peptides from a set of 18 protein standards[14] showed 46% of validated assignments with scores above thresholds from randomized searches. (IV) DTA files from sample 1 searched using MSPlus peptide sequences accepted as valid based on the combined method, showing improved discrimination from corresponding randomized database searches. Thus, the validated peptides (set IV) are a subset of set I/II. Parallel searches using Mascot indicated Mowse thresholds of 46 (MH^+), 50 (MH_2^{2+}), and 50 (MH_3^{3+}) for searches against a randomized database (not shown). Manual analysis of a random subset of peptide assignments in (II) with scores above thresholds confirmed all Sequest and Mascot identifications as valid. Similar improved discrimination was observed using MSPlus over the Mascot search alone with the normal database. The number of DTA files in each experiment were as follows: I and II (+1) 655, (+2) 1462, (+3) 1462; III (+1) 66, (+2) 201, (+3) 63; IV (+1) 167, (+2) 589, (+3) 100. (C) Repeated LC/MS/MS analyses of a single SCX fraction from sample 3, searched with Sequest (circles) or Mascot (diamonds). The number of unique peptide assignments with scores above thresholds determined from randomized searches (see panel B) was determined for each of the five analyses. The number of unique peptide assignments within 5 combinations of one data set, 10 combinations of two data sets, 10 combinations of 3 data sets, 5 combinations of 4 data sets, and one combination of 5 data sets is shown, where each data point represents one combination. The results show that the number of unique peptide assignments when considering one data set (57 for Sequest, 50 for Mascot) increases when considering all five data sets (109 for Sequest, 95 for Mascot), indicating that about twice as many peptides are identifiable in repeated experiments. In contrast, the number of peptide assignments captured by MSPlus in each individual analysis (triangles) was comparable to the number seen in aggregate by Sequest or Mascot alone, demonstrating greater sensitivity in validating assignments. Likewise, the number of proteins in each MSPlus analysis showed 92% overlap with proteins identified from five aggregate data sets by Sequest.

The third method estimated the number of class negative DTA files from combined results of several experiments, which showed that ~43% of DTA files in sample 1 represented artifacts such as oxidation during sample handling, in-source fragmentation during MS, ions too weak for search programs to assign correctly, incorrectly made DTA files, or false positives (for more detailed discussion, see below). Only 66 DTA files were not explained, and these may represent non-peptide ions, salt adducts, or other peptide modifications. This is further evidence that ~54% of DTA files would be expected as class positive. Overall, the three approaches suggest that 50–60% of DTA files should be assignable to tryptic peptides in sample 1. Of these, less than half can be validated based on having scores above thresholds. The rest are “hidden” due to their failure to be validated using a threshold approach.

Combining Sequest and Mascot To Improve Validation of Peptide Assignments. We then looked for ways to identify and validate these low-scoring peptide assignments. In studies comparing search programs, we found that a given peptide assignment might obtain a high score with one search program and a low score with another search program, presumably due to variations in the scoring algorithm between programs (Figure 2A). We

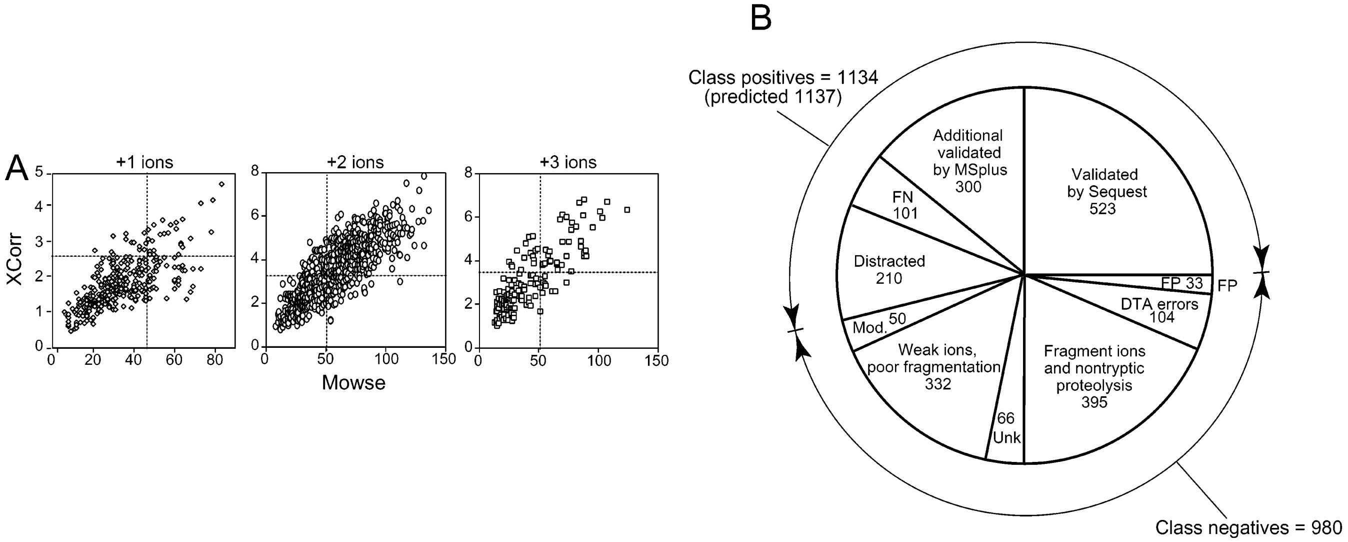

Figure 2. Combining Sequest and Mascot results to validate more DTA files. (A) Comparison of XCorr vs Mowse scores for DTA files in sample 1, in which both Sequest and Mascot identified the same peptide sequence. Thresholds for different charge states determined from randomized database searches (Figure 1B) are indicated by lines. The results show a significant number of DTA files scoring above threshold for one program but not the other, which reveals that the two programs have overlapping but nonidentical scoring criteria. (B) Summary of DTA file classifications from sample 1. The number of class positive assignments was estimated by extrapolating the number of DTA files for standard peptides above Sequest thresholds (1137 = 46% of 2117); the number of class negatives (980) was obtained by subtracting class positives from the total number of DTA files (980 = 2117−1137). The number of Sequest assignments scoring above threshold was 523. The number of assignments made by MSPlus was 856 (Table 1), which included Sequest assignments scoring above threshold (validated by Sequest = 523) and false positive assignments (FP = 33) estimated by searching against the randomized database. By difference, ~300 additional assignments were validated by MSPlus. False negative assignments (FN = 101) were estimated by manual analysis, as described in the text. The distracted class (distracted = 210) is estimated from DTA files that were incorrectly assigned by Sequest, but identifiable by CLASP, as described in the text. Other classifications in sample 1, estimated by direct counting, included the following: (i) 332 DTA files with weak ions or poor fragmentation (threshold defined by the lowest signal-to-noise ratio for fragment ions in DTA files that were manually validated), (ii) 50 posttranslationally modified peptides, e.g., containing oxidation products of methionine or tryptophan (47 DTA files), phosphate (3 DTA files), or incomplete cysteine alkylation (none observed), (iii) 395 fragment ions or nonspecific proteolysis products, identified by searching without specifying protease cleavage and extrapolating for false negatives and distracted cases (276 DTA files), or dehydrated/deammoniated fragment ions of parents identified using CLASP (119 DTA files), and (iv) 104 DTA errors. DTA errors included (a) 3.5% of DTA files (74) representing ions of charge greater than or equal to 4 (estimated by analyzing results from data collected with high-resolution scans in order to identify the charge), (b) 0.9% of DTA files (19) representing singly charged ions misassigned as multiply charged ions, due to noise peaks in the MS/MS at masses greater than the parent ion (identified by the dehydrated fragment ions of the parents in a small data set), (c) 0.5% DTA files (11) which represent incorrect combinations of MS/MS that are actually derived from different peptide ions (identified by manual analysis of those DTA files where more than one MS/MS spectrum was summed in order to enhance the signal-to-noise ratio, and the peptides were validated by MSPlus, and at least 35% of the fragment ions were unidentified). After summing all classes, 66 remained with unknown classification (Unk).

utilized parsers to capture Mascot and Sequest outputs into a database and wrote a script (MSPlus) that allowed easy comparison of results from Sequest and Mascot programs. Thus, in sample 1, 7.5% of Sequest assignments with subthreshold XCorr scores were validated when Mascot assignments of the same DTA files showed Mowse scores above threshold. Another important class included cases where both search programs made the same sequence assignment, although both XCorr and Mowse scores were below threshold. However, manual analysis indicated that only half in this class were correctly assigned; therefore, a filtering protocol was implemented in MSPlus to discriminate between correct vs incorrect assignments.

Filters tested included the Sequest scores SP and RSP (preliminary score and ranked preliminary score, where SP is the sum of the matched b and y fragment ions together with adjustment factors for matching consecutive ions in a series), ion ratio (percent of theoretical fragment ions observed in a spectrum), and ΔCN (difference in XCorr between the first and second highest ranking sequence assignments).[3,16] Examination of manually validated data revealed that 97% of correctly assigned sequences had RSP = 1, all singly or doubly charged ions had an ion ratio of >25%, and 95% of triply charged ions had an ion ratio of >20%. SP and ΔCN provided insufficient discrimination, possibly because of the large size and sequence redundancy of the human database, and were not used. We also excluded peptides with mass less than 900 Da, because we found that the frequency of incorrect assignments due to distraction (see below) was greater with small peptides. This was not a serious limitation, because peptides smaller than nine amino acids were not generally useful for uniquely specifying proteins. Also excluded were assignments with internal KR, RK, KK, or RR sequences, where trypsin would cleave efficiently.

An important filter criterion ensured that the number of basic residues in the assigned sequence was consistent with the SCX column elution. A set of rules was delineated from the SCX behavior of the high-scoring peptides. For example, only peptides with one basic residue were allowed in SCX fractions 4−8, and only peptides with one or two basic residues were allowed in fraction 9 (Figure 3A and legend). These criteria were highly discriminatory; in analyses with the randomized database, the SCX filter removed half of the false positives observed among MSPlus-validated peptides generated without the filter. Figure 3B shows

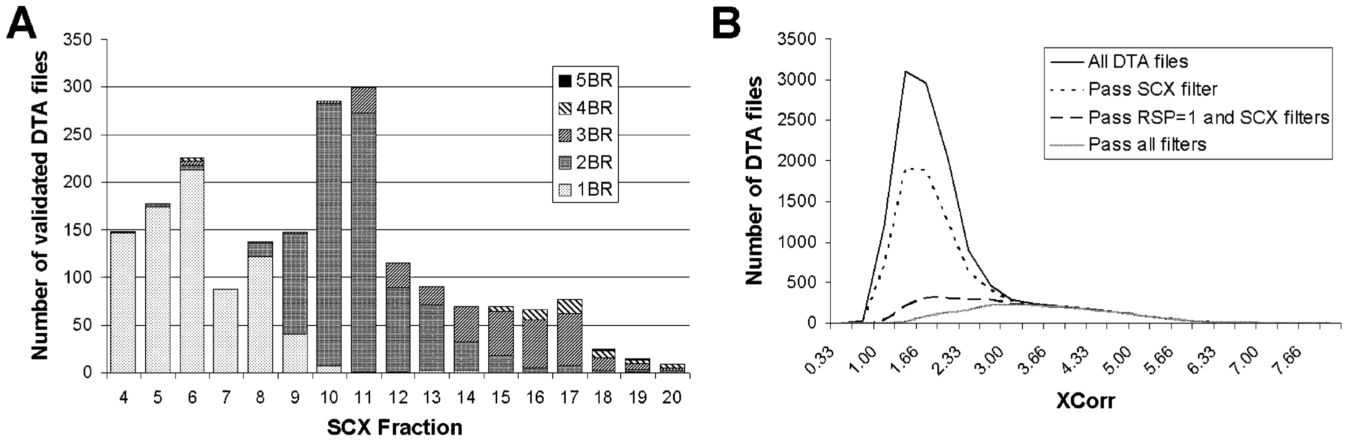

Figure 3. Using SCX chromatography as an MSPlus filtering criterion. (A) The number of basic residues (BR: Lys, Arg, His) in peptide sequences for DTA files that passed all MSPlus filters (excluding the SCX filter) are plotted vs their SCX fraction number. Early-eluting fractions primarily contain peptides with one basic residue, while later fractions contain peptides with multiple basic residues. Rules were developed for each SCX chromatographic run, because small variations were observed with column lot number and chromatographic conditions. The data shown represent one sizing gel fraction from sample 3, where the SCX filter allowed (1) 1BR in fractions 4−8, (2) 1 or 2 BR for fractions 9, (3) 2 BR for fraction 10, (4) 2 or 3 BR for fractions 11−14, and (5) 2, 3, or 4 BR for fractions 15−20. (B) The effect on the XCorr distribution of MSPlus-validated DTA files is shown after sequentially applying the SCX rules derived from panel A, the RSP = 1 filter, and the remaining MSPlus filters. Each filter removes large numbers of incorrect assignments, without affecting the high-confidence assignments with XCorr >3.5. A few DTA files with correct assignments were lost; however, many were captured in adjacent SCX fractions. Importantly, applying the SCX filter had little effect on the protein profile. Thus, the SCX filter removed only four proteins in sample 1, which in all cases were supported by only one peptide.

that the SCX filter preferentially removed incorrectly assigned spectra that were concentrated in the peak between XCorr 0.8 and 3.2. This figure also shows the effect of further applying the RSP = 1 filter as well as the complete set of filtering criteria. Although the filters removed a few validated assignments, most of these peptides were captured in other fractions or from other DTA files.

MSPlus is a heuristic or expert program that implements a set of pass/fail rules using all these criteria for acceptance/rejection of peptide assignments, as described in Methods. After filtering the combined results of Sequest and Mascot searches using these criteria, MSPlus validated 856 DTA files (40% of the total DTA files) in sample 1 (Table 1, Figure 2B). A second script (Isoform Resolver) assembled these peptide sequences into a protein profile, applying an additional filter for false positives due to sequence isoforms in the database and reporting the minimum protein profile that will account for the peptide data (discussed in detail below and in Methods). MSPlus and Isoform Resolver were applied to the five repeated LC/MS analyses described in Figure 1C (right panel) and compared to results obtained by the Sequest threshold approach. Analyses of each individual LC/MS by MSPlus/Isoform Resolver yielded an average of 53 proteins (SD 6), while combined analysis of the five data sets yielded 63 proteins by the Sequest threshold. Furthermore, an average of 92% (average 49, SD 2) of the proteins identified by MSPlus/Isoform Resolver in each data set overlapped with the proteins identified by Sequest in the combined data set. This demonstrates that the MSPlus/Isoform Resolver algorithms greatly enhance data capture from complex data sets, improving sensitivity of protein detection similar to that seen by collecting repeated data sets on the same sample.

False Positives. The accuracy of the filtering approach was tested by assessing the frequency of false positives (FP, peptides validated by MSPlus but judged incorrect by manual analysis). First, 540 peptide assignments were randomly selected for manual analysis from the sample 1 output, half of which were accepted and the other half of which were rejected by the MSPlus script, and then results were extrapolated to the full data set. The manual analysis indicated ~3.4% of the peptide assignments validated by MSPlus were incorrect. Note that this frequency represents the percentage of false positives normalized to the number of MSPlus validated assignments, rather than the statistical false positive rate.

Second, the number of false positives was estimated by searching data sets against the randomized database, where any identification is by definition incorrect. These searches yielded false positive frequencies of 3.9, 3.2, and 3.9% for sample 1, sample 2, and a data set of two gel filtration fractions from sample 3, respectively. (Samples are described in Methods and in Table 1.) Thus, the independent measurements were in good agreement with results from manual analysis.

The randomized database does not survey false positives due to peptide isoforms. For example, the presence of peptides with amino acid replacements that are often indistinguishable by ion trap mass spectrometers (D/N/I/L, K/Q/E, E/M, V/T, or V/P) may introduce false positive assignments. Many of these were detected and removed by Isoform Resolver (see below), which ensures that the protein count is not inflated by the presence of peptide isoforms. However, in the sample 1 data set, we did observe three cases in which two or more amino acid changes occurred that did not alter the peptide mass (such as QS to NT). Therefore, we estimate that the peptide false positive frequency is ~4.2%. This value is consistent with reproducibility studies discussed below.

False Negatives and Incorrect Assignments. An important goal was to minimize the number of false negative validations (FN, peptides failing the MSPlus/Isoform Resolver filters but presumably correctly assigned by Sequest). We initially assumed 314 false negative assignments from the number of class positives (1137) minus the number of MSPlus-validated DTA file assignments (856), excluding the false positives (33). However, manual analysis

Table 2. Examples Illustrating "Distraction" by Sequest and Mascot, in Which Correct Sequence Assignments Are Replaced by Incorrect Assignments as the Database Size Increases[a]

database size	first Sequest assignment	XCorr	RSP	first Mascot assignment	Mowse	second Mascot assignment	Mowse[f]
example 1							
restricted database[b]	AIGTEPDSDVLSEIMHSFAK	1.98	1	AIGTEPDSDVLSEIMHSFAK	39.5	VGLPPGKAAAKASESSSSEESR	6.3
IPI database[c]	AIGTEPDSDVLSEIMHSFAK	1.98	1	AIGTEPDSDVLSEIMHSFAK	39.5	DVKEFKPESSLTTLKAPEK	25.8
IPI, no protease[d]	AIGTEPDSDVLSEIMHSFAK	1.98	422[e]	TTIGAAGLPGRDGLPGPPGPPGPP[g]	40.0	AIGTEPDSDVLSEIMHSFAK	39.5
example 2							
restricted database[b]	EGLELPEDEEEK	2.00	1	EGLELPEDEEEK	50.4	EQVNELKEKGNK	17.1
IPI database[c]	EGLELPEDEEEK	2.00	1	EGLELPEDEEEK	50.4	GDQGIAGFPGSPGEK	36.1
IPI, no protease[d]	EGIELLLNEGSEL[g]	2.23	2	EGLELPEDEEEK	50.4	EGNLKKFQPDLK	48.0
example 3							
restricted database[b]	GDAMIMEETGK	0.74	1	GDAMIMEETGK	41.4	DGDKQRYLGK	10.8
IPI database[c]	YPILFLTQGK[g]	1.11	1	GDAMIMEETGK	41.4	TAPFFKQGRK	22.7
IPI, No protease[d]	AVYVEMLQIL[g]	1.34	12	GIMAIEMVEGE[g]	43.9	GDAMIMEETGK	41.4
example 4							
restricted database[b]	DLSLEEIQK	1.15	1	DLSLEEIQK	25.2	EQEVAELKK	13.1
IPI database[c]	DLSLKEIQK	1.64	11	IDCEAPLKK[g]	27.7	DSLKGGGALEK	25.7
IPI, no protease[d]	NSQVKELKQ[g]	1.53	243	ALASQSAGITGV[g]	31.5	ILTLDEGGSAP	31.4

[a] Data files were from sample 1, where correct sequence assignments were confirmed by manual inspection and are underlined. A total of 286 out of 856 validated DTA files showed distraction in one or both search programs. None of the distracted assignments shown here were caused by error in observed molecular mass. The spectra and manual evaluations are shown in Supporting Information, Figure 1. [b] The normal search protocol was carried out using a restricted database of 644 proteins, containing the 243 proteins and possible variants observed in sample 1. [c] The normal search protocol was carried out with the IPI database (47 306 entries), specifying peptide mass range 900−4800 Da and up to two trypsin cleavages. [d] The normal search protocol was carried out with the IPI database, specifying peptide mass range 900−4800 Da and no protease cleavage specificity. Removing protease specificity increased the effective database size by ∼10-fold. [e] Note the very large increase in RSP value; ∼20% of the peptides validated by the normal search failed the RSP filter of MSPlus when the database size was increased by removing protease specificity. [f] The scores for the second Mascot assignment showed a systematic increase with database size; this was also observed in scores resulting from searches against the IPI database randomized by sequence inversion. [g] Upon searching a larger database, the correct sequence assignment was replaced by an incorrect assignment with higher score, revealing distraction.

of 540 spectra showed that 8% of DTA files that MSPlus rejected were correctly assigned. This indicated that there were only ∼101 false negative assignments out of the 1261 DTA files that MSPlus rejected, leaving 213 of the class positives unaccounted for. Therefore, we considered the possibility that many DTA files were identifiable but incorrectly assigned by the search programs (referred to as "distracted"). Such occurrences were clearly revealed in cases where Sequest and Mascot assigned different sequences to a DTA file, only one of which was validated based on scoring and manual analysis (Table 2).

To estimate the number of incorrect assignments due to distraction, a *cl*uster *a*nalysis of *sp*ectra (CLASP) program was developed that directly compares MS/MS spectra, scoring for similarities in fragmentation patterns, parent ion mass, and reversed-phase and SCX chromatographic behavior.[17] The 1261 DTA files that failed the filtering criteria in sample 1 were compared against the 856 DTA files that passed. CLASP identified 175 additional DTA assignments with spectral similarity to previously validated peptides. Of these, 57 were correctly assigned, but failed the filtering criteria (false negatives), and 118 were incorrectly assigned by Sequest, Mascot, or both. These values represent underestimates, because some identifiable peptides failed the CLASP criteria and because CLASP can only identify cases where a peptide is sequenced more than once (and at least one case is validated). Therefore, we extrapolated from the ratio of the number of validated peptides observed multiple times to the number of peptides observed once (1.27:1). This ratio was measured by direct counting and was similar between three data sets of similar size taken on samples of similar complexity. Correcting for the total peptide number yielded 103 FN assignments, comparable to the 101 FN determined independently from manual analysis. From this, we estimated that ∼311 identifiable assignments failed the MSPlus criteria, which included ∼101 false negative assignments and ∼210 DTA files incorrectly assigned by Sequest (∼176 incorrect assignments by Mascot; some were incorrectly assigned by both search programs). The sum of all the identifiable peptides (excluding 33 false positives), distracted assignments, and false negatives was 1134 (55%), in good agreement with our initial predictions of 1046−1137 (summarized in Figure 2B).

Note that the distracted class is false negative for the search program but true negative for MSPlus. Therefore, the correct class positive size for MSPlus is 101 + 300 + 523 = 924 (Figure 2B), so that the statistical FN rate is 10.9% (101/924); presumably, if distraction could be avoided, MSPlus could validate up to 89% of the ∼210 incorrect, but identifiable assignments. When Sequest or Mascot thresholds were set to yield the same percentage of false positives as observed with MSPlus, their FN rates were 42 or 41%, respectively; thus, the 11% FN rate from MSPlus represents a significant improvement.

We evaluated possible causes for incorrect assignments. A major cause seems to be the presence of "extraneous" fragment ions not considered by the search programs; these will lower the XCorr or Mowse scores for the correct assignment and increase the probability of making an incorrect assignment. These extraneous ions include sequence-specific a, b, or y ions not considered by the search programs (e.g., triply charged fragment ions) and internal

(16) Yates, J. R.; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Anal. Chem.* **1995**, *67*, 1426−1436.

(17) Meyer-Arendt, K.; Knight, R. D.; Ahn, N. G.; Resing, K. A. University of Colorado. Manuscript in preparation.

fragment ions generated by multiple cleavages (for examples, see Supporting Information, Figure 1). Another source could be noise peaks that were not adequately removed from low-intensity spectra by the search programs, and it might be thought that this is a major source of distraction. However, comparison of the fragment ion intensities in distracted versus validated spectra showed that the distraction class was only slightly enriched for DTA files with weaker spectral intensities (Supporting Information, Figure 2). Most of the spectra with low intensities, poor fragmentation, or both were classified as class negatives and were not a significant cause of distraction, because neither the search programs nor CLASP had a strong chance of identifying them.

Effect of Database Size. Another important class of distracted assignments was that where the peptide mass shows significant error, sometimes as high as 9 Da (depending on the charge state of the peptide ion), even after minimizing space charging effects by optimizing the number of ions allowed into the trap.[18] Because the mass tolerance in our search strategy was initially narrow, we varied the mass tolerance during searching in order to examine the number of validated peptide sequence assignments (i.e., with high XCorr values). From 1.0 to 3.0 Da (average mass), 25% more assignments were validated (Figure 4A), but above 3.0 Da there was no further gain from increasing mass tolerance. Furthermore, the score distribution shifted to higher XCorr values after searching the randomized database with higher mass tolerances (data not shown), resulting in an increased percentage of scores below threshold. This suggested that the frequency of incorrect assignments due to distraction increased as the mass tolerance was set to higher values, which we ascribe to an effective increase in database size due to the increased number of peptides that must be queried as the mass tolerance is increased.

To more clearly demonstrate the effect of database size on distraction, searches were carried out under other conditions that altered the effective database size. We decreased the effective size by 75-fold upon restricting the MSPlus/Isoform Resolver search to a database consisting of only the proteins identified in sample 1 (644 proteins, including all possible variants, vs the full database of 48 000 proteins). We saw no increase in percentage FP when searching the randomized version of this database. Alternatively, we increased the effective size ~20-fold by not designating protease specificity. Table 2 shows examples in which both Sequest and Mascot programs replaced correct sequence assignments with incorrect assignments, as the database size increased from (i) the restricted set, to (ii) the normal IPI database with tryptic sites specified, to (iii) the IPI database with no protease specified. (The spectra in these examples are shown in Supporting Information, Figure 1.) Likewise, a histogram of XCorr distributions for sample 1 revealed significant increases in threshold scores as the database size increased (Figure 4B). These results demonstrate that, for spectra with low scores, better discrimination between correct versus incorrect peptide assignments can be achieved by searching against the smallest possible database.

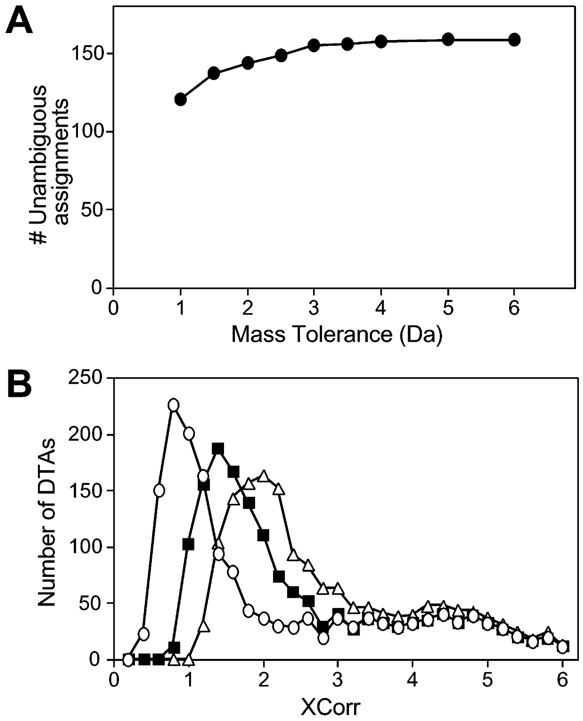

Figure 4. Increase in distraction with database size. (A) Increasing peptide mass tolerance provided only a small number of additional peptide assignments. The combined data set of five LC/MS/MS runs shown in Figure 1C were searched with increasing mass tolerance, using average parent masses. The number of high-scoring peptide assignments increased by 24%, when mass tolerance was increased from 1 to 3 Da. No further increase in high-scoring peptides was seen above 3 Da, although manual analysis identified many files with mass inaccuracies up to 9 Da. (B) Increasing the effective database size during the search increased the scoring thresholds for unequivocal assignments, providing an explanation for increased distraction. (squares) Search of data in sample 1 using the IPI database (~48 000 proteins); (circles) search using a restricted database (644 proteins, including proteins and isoforms identified by MSPlus); (triangles) search using the IPI database without limiting protease specificity, which increases the effective database size by ~10-fold.

These analyses suggested search strategies that would minimize the database size and thus minimize the distraction effect as well as computational cost. First, only fully tryptic products were considered in the searches with sample 1. When nontryptic products were considered, 147 DTA files with correct assignments in the tryptic search were assigned incorrect sequences in the nontryptic search, while 209 DTA files with lower scores in the tryptic search gave high-scoring nontryptic sequences. Of the 209 spectra, 84 yielded no new information, because they represented ions that coeluted with larger peptides that contained the same ion; thus, many nontryptic peptides appear to be fragment ions derived from in-source cleavage. More important was the number of unique peptides captured (56% of validated peptides are sampled multiple times). When unique peptide assignments were considered, more information was lost (95 unique peptide assignments) than gained (62 unique peptide assignments) in allowing nontryptic searches. Second, peptide modifications were excluded from the search. When covalently modified peptides were examined by searching for specific mass increases, unmodified forms of the same peptides were almost always present elsewhere in the data set. Therefore, accounting for peptide modifications did not significantly increase the numbers of peptides and proteins identified, and for the purposes of accurate protein identification, including modified forms in the search procedure did not justify the increased computational cost.

[18] Cleven, C. D.; Cox, K. A.; Cooks, R. G.; Bier, M. E. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 451−454.

In contrast, we considered incompletely digested peptides and found many examples of high-scoring, incompletely digested peptides with missed cleavages at K or R that were immediately adjacent to E or D (29 of 409 unique peptides), within two residues of the N- or C-terminus (25 peptides), or within highly acidic sequences (15 peptides). All were consistent with the known proteolytic specificity of trypsin.[19] We compared the number of peptides validated allowing one, two, or three missed cleavages. Allowing three missed cleavages significantly decreased the number of validated peptides, but no difference in number of validated assignments was observed when comparing one or two missed cleavages, although specific differences were observed between these cases. Allowing one or two missed cleavages with both search programs yielded 406 unique peptide assignments in all cases; 15 additional peptides were observed by specifying one missed cleavage, reflecting increased capture due to reduced distraction, and 16 additional peptides were observed by specifying two missed cleavages, reflecting increased capture of incomplete proteolytic products. Using different parameters for incomplete proteolysis in each search program should minimize distraction with one program, while capturing more incomplete proteolytic products with the other. Allowing one missed cleavage for Sequest and two missed cleavages for Mascot validated the 406 peptides along with 25 additional peptides, all of which were found in the 31 (= 15 + 16) assigned in the first two experiments (the other 6 assignments lacked consensus for capture by MSPlus). The best results were obtained by minimizing the effective database size by requiring tryptic cleavages at both N- and C-termini and excluding covalent modifications, while allowing one or two incomplete products using Sequest or Mascot, respectively.

It is important to note that the distraction is more important for the low-scoring assignments than for assignments with high, unequivocal scores. The probability of distraction for peptide assignments increases as the score falls below threshold values, because the chance that an incorrect peptide sequence will score higher than the correct assignment increases. Furthermore, the effectiveness of MSPlus was greater when large data sets were analyzed. Comparison of the number of unique peptides identified in sample 1 versus samples 2 and 3 showed that MSPlus validated 18−20% more unique peptides than Sequest or Mascot in sample 1 and 25−35% more peptides in samples 2 and 3. We assume this is because we are differentially sampling a population of ions with complex fragmentation or lower intensity ions in the gas-phase extraction; as discussed above, these will generally have lower scores and will be disproportionately represented as MSPlus-validated peptides versus those validated by the threshold approach.

Constructing an Accurate Protein Profile from Peptide Sequence Information. Assembling a protein report from validated peptide assignments is complicated by the large number of protein database entries containing redundant sequences. Furthermore, Sequest and Mascot search programs may choose different protein entries when assigning peptides to proteins with redundant sequences, without reporting that the proteins are actually indistinguishable from the supporting peptide assignments. Minimizing these ambiguities is essential for an accurate protein profile; however, it also is important to retain information about possible alternative isoforms. Therefore, we did not rely on Sequest/Mascot-designated protein accession numbers. Instead we developed a peptide-centric database that catalogues all unique peptide sequences in the human protein database and reports proteins that are redundantly associated with each peptide sequence. An Isoform Resolver algorithm was developed that links each peptide in the sample data set to all protein database entries containing that peptide, clusters peptides that are linked to the same set of accession numbers, and groups clusters that contain one or more accession numbers in common (Table 3). Finally, the minimal number of protein variants required to account for the observed peptides is determined. The resulting numbering scheme specifies all possible protein isoforms for each observed peptide and accurately reports whether the observed peptide(s) support(s) a single protein versus a set of possible variants. Consequently, closely related proteins are adjacent to each other in the final report, which is helpful for characterizing the relative contribution of isoforms in our samples.

In addition, validated peptide variants that are similar in mass due to amino acid replacements within the mass tolerance of the ion trap (D/N/I/L, K/Q/E, E/M, V/P, V/T) were treated as potentially redundant. If a peptide specifies a protein supported by only that peptide, and a related peptide variant specifies a different protein, only one protein is counted (favoring the protein supported by the largest number of peptides), although all are recorded in the output. Application of Isoform Resolver revealed that sequence redundancy caused substantial overestimation of protein counts using conventional methods; for example, Sequest predicted 309 proteins from 856 validated peptides in sample 1, whereas detection of protein and peptide variants by Isoform Resolver reduced the protein count by 24%, to 243 proteins. This reduction was slightly larger with larger data sets.

Enhancing Protein Detection. Finally, the combined approaches of examining consensus between search engines, applying filtering criteria based on peptide chemistry, and applying Isoform Resolver were used to analyze larger data sets. To evaluate effects of varying experimental conditions on peptide and protein identification, experiments compared protein profiles from sample 1, which was analyzed by SCX separation followed by RP-LC/MS/MS, sample 2, which was analyzed with fewer SCX fractions but included gas-phase fractionation, and sample 3, which fractionated proteins by sizing gel exclusion prior to digestion, SCX chromatography, and RP-LC/MS/MS, combining results from full mass range and gas-phase fractionation. To assess the effects of gas-phase fractionation, peptides in SCX fractions were analyzed over the full mass range or in 10 narrow mass ranges (Table 1, sample 1 vs 2, MSPlus/Isoform Resolver). Gas-phase fractionation yielded a >20-fold larger data set, increasing the number of proteins identified by >7-fold (from 243 to 1757 proteins) and reducing the percentage of proteins specified by only one peptide from 70 to 50%. However, gas-phase fractionation also increased the number of fragment ions observed, which lowered the percentage of validated DTA files from 40 to 17% of the total data set. Resolving proteins by gel exclusion simplified the protein complexity prior to SCX chromatography and significantly increased peptide sampling (Table 1, sample 3, MSPlus/Isoform Resolver). In all, 5130 proteins were identified in K562

[19] Hill, R. L. *Adv. Protein Chem.* **1965**, *20*, 37−107.

Table 3. Protein Variants Specified by Isoform Resolver from Observed Peptide Sequences[a]

		highest scores		no. of peptides observed			
ref no.[b]	protein accession number, annotation peptide sequence[c]	XCorr	Mowse	total	+1	+2	+3
Example 1: Only One Open Reading Frame Detected							
725	IPI00045914, nuclear receptor transcription cofactor						
725	AAPTPTPAPVPVPVPLPAPAPAPHGEAR	4.0	43	1			1
725	GNSSETSHSVPEAK	3.2	43	1		1	
725	HLWVGNLPENVR	3.4	47	2		2	
725	TYHPPAQLTHTQFPAASSVGLPSR	2.8	23	1			1
725	VDATRPEATTEVGPQIGVK	4.0	81	1		1	
725	VLQPANLGSTLTPHHPPALPSK	2.3	48	1		1	
Example 2: Two Isoforms That Cannot Be Distinguished							
770*[d]	IPI00014311, Cullin homologue 2						
770*	IPI00180783, Cullin 2						
770*	AVSTGLPHMIQELQNHIHDEGLR	4.9	67	4		2	2
770*	FVQLINTVLNGDQHFMSALDK	2.4	30	1		1	
770*	KYLHPSSYTK	2.5	29	1			1
770*	RLIHGLSMSMDSEEAMINK	1.7	30	1		1	
770*	VIHGVINSFVHVEQYK	5.2	102	4		2	2
770*	YIDDKDVFQK	1.7	37	1	1		
Example 3: One Protein Isoform Specified By One Peptide; Other Peptides Cannot Distinguish Isoforms							
775	IPI00220502, BAI1-associated protein 2 isoform 2						
a	IPI00180292, BAI1-associated protein 2 isoform 3						
b	IPI00180972, similar to BAI1-associated protein 2						
c	IPI00185159, BAI1-associated protein 2 isoform 1						
d	IPI00186231, similar to BAI1-associated protein 2						
e	IPI00217925, similar to insulin receptor substrate						
775	NPFAHVQLKPTVTNDR	3.9	72	2		2	
775_a_b_c_d_e	EGDLITLLVPEAR	3.3	46	1		1	
775_a_b_c_d_e	LHMSLQQGK	2.0	45	1	1		
775_a_b_c_d_e	MSAQESTPIMNGVTGPDGEDYSPWADRK	2.4	37	1			1
775_a_b_c_d_e	SNLVISDPIPGAKPLPVPPELAPFVGR	2.9	31	1			1
775_a_b_c_d_e	SSSTGNLLDKDDLAIPPPDYGAASR	2.1	32	1		1	
775_a_b_c_d_e	YSDKELQYIDAISNK	2.9	28	1		1	
775_a_b_c_e	SFHNELLTQLEQK	4.1	71	3	1	2	
Example 4: Complex Group with Minimum of Three Proteins, Including Bridge Peptides[f]							
787	IPI00023860, nucleosome assembly protein 1-like 1						
788*	IPI00017763, nucleosome assembly protein 1-like 4						
788*	IPI00180912, similar to NAP1						
789	IPI00184769, similar to NAP1						
a	IPI00185366, similar to NAP1						
787	EQSELDQDLDDVEEVEEEETGEETK	5.0	88	4		3	1
787	KYAVLYQPLFDK	4.0	74	8	3	5	
787	KYAVLYQPLFDKR[e]		77	6		2	4
787	LDGLVETPTGYIESLPR	5.7	113	5	1	4	1
787	YAVLYQPLFDK	4.1	58	3		2	
787	YAVLYQPLFDKR	2.7	51	3		3	
787_a	GIPEFWLTVFK	2.9	61	4		4	
787_a	NVDLLSDMVQEHDEPILK	6.4	109	3		2	1
*787_788*_789*[f]	*FYEEVHDLER*	*3.8*	*49*	*18*	*4*	*9*	*5*
788*	KYAALYQPLFDK	4.7	73	6	1	5	
788*	LDNVPHTPSSYIETLPK	4.9	66	4		1	3
788*	QVPNESFFNFFNPLK	3.4	77	5		5	
788*	YAALYQPLFDK	2.6	64	3	1	2	
*788*_789*[f]	*GIPEFWFTIFR*	*3.0*	*60*	*3*		*3*	
789	AAATAEEPNPK	2.9	56	1		1	

[a] Proteins were identified from MSPlus-validated peptide sequences in sample 1 using the Isoform Resolver script. Observed peptides were matched against a peptide-centric database, which specifies proteins that share every unique peptide sequence in the IPI database. [b] At the end of the analysis, Isoform Resolver assigns each protein a reference number or letter. Proteins that can be distinguished based on the peptide sequence information are assigned different numbers, whereas proteins that cannot be uniquely distinguished are assigned identical numbers. Those protein variants where unique peptides are identified are assigned numbers, which allow enumeration of the minimum number of proteins required to account for the peptides. Proteins that represent variants in the database which also contain one or more of the peptides are assigned letters.[c] IPI protein accession numbers from Version 2.18 update. [d] Asterisks indicate examples where peptides are found in more than one protein isoform, with no unique peptide sequence observed that specifies a single isoform. In this example, all six observed peptide sequences are found in both IPI00014211 and IPI00180783 database entries. [e] This peptide was not identified by Sequest, because it contains two missed cleavages which were not allowed by Sequest (see Methods). [f] Peptides with sequences in common between two numbered proteins are italicized.

soluble extracts, of which 55% were supported by two or more peptides.

The number of false positive proteins was estimated from the false positive frequency for MSPlus-validated peptides (~4.2% of total peptides, see above). This predicts that 4.2, 0.18, and 0.007% of proteins supported by one, two, or three peptides, respectively, should represent false positive assignments (a peptide observed in more than one charge form is counted as only one peptide). Therefore, in samples 1, 2, and 3, the number of false positive protein assignments would be approximately 7 (2.9%), 38 (2.2%), and 100 (1.9%), respectively. In sample 3, a combination of MSPlus and Isoform Resolver identified more than 5000 proteins from a single cell type in a single experiment. In comparison, conventional analyses using Sequest and Mascot, which applied XCorr and Mowse thresholds and allowed ~4% false positive peptide validations, resulted in only 4130 and 3971 proteins, respectively (Table 1).

As the number of peptides identified in each experiment increased among samples 1, 2, and 3, the number of proteins increased by a lower percentage, suggesting saturability in protein detection. Saturability was also indicated by the fact that a progressively lower percentage of proteins supported by only one peptide is observed as the total number of peptides sampled increases (Table 1).

Reproducibility. The low frequency of false positives suggested high accuracy; therefore, reproducibility should be high. To test this, the lists of identified proteins were examined to assess the degree of overlap between different experiments. To test reproducibility between identical experiments, proteins were fractionated by sizing gel exclusion and proteolyzed, and peptides were separated by SCX chromatography (sample 3). Each SCX fraction was analyzed by RP-LC/MS/MS over a full mass range as well as six gas-phase fractionation mass ranges (defined in Table 1). Triplicate experiments performed on one sizing gel fraction from sample 3, analyzed by SCX separation and gas-phase fractionation, showed 70% (SD 5%) overlap in proteins identified among all six combinations of two repeats. The nonoverlap of 30% reflects both the presence of false positives and incomplete sampling of the highly complex protein mixture.

To minimize the impact of incomplete sampling, reproducibility also was assessed by comparing smaller data sets against larger data sets (the ratio of peptides sampled in the smaller vs larger data sets was ~9). The results showed that 227 of the 243 proteins (93%) identified in sample 1 (Table 1) were also found in sample 2. A second analysis compared two subsets of data from sample 3, in which of 320 proteins identified by full mass range scanning of one size exclusion column fraction, 298 proteins (93%) were present in the larger data set collected using gas-phase fractionation. In each experiment overlapping a smaller data set onto a larger one, the ~6−7% of proteins not reproducibly detected were supported by only one peptide and could be accounted for by false positive assignments as well as incomplete sampling. Finally, sample 1 was compared to the sample 3 data set, which is 100-fold larger. Upon overlaying sample 1 onto sample 3, 234 of the 243 proteins were reproduced, 160 of which were supported by single peptides in sample 1. Of the 234 proteins, 224 were supported by 2−55 high-quality peptides in the larger data set, indicating that most of the proteins identified in the sample 1 experiment were true positives, even if the majority was observed by only one peptide. Significantly, only nine of sample 1 proteins (4%) did not reproduce in sample 3, consistent with our estimate of 4.2% peptide false positives.

CONCLUSION

This study demonstrates a novel programmatic approach for protein identification by the shotgun method, which primarily relies on peptide chemical properties, consensus between search programs, and parameters reflecting quality of the MS/MS spectra, rather than XCorr or Mowse probability threshold cutoffs for peptide validation. In three independent experiments, the number of validated peptide assignments increased by 21−58% compared with threshold methods, when allowing the same ratio of false positives to identified peptides. This represented a reduction in the false negative rate from 41 to 11%. Several lines of evidence show that the frequency of false positive peptide assignments is ~4.2% (FP/MSPlus-validated) in all three samples; thus, very few proteins are incorrectly identified, and reproducibility is high (within the limits of sampling in the system). By minimizing not only the false positive rate but also the false negative rate, we were able to achieve a high number of protein identifications with a minimum of SCX fractionation.

A major finding of this study is that ~19% of identifiable DTA files are incorrectly assigned by the search programs through a "distraction" effect; this increases to 29% using a search strategy that specifies no protease. Distraction has been noted previously;[4] however, our implementation of the CLASP algorithm allowed quantification of the effect. We found that a major cause of distraction was increased database size, which led to increased statistical thresholds for validating Sequest and Mascot searches. Effective database size was strongly affected by searching with increased mass tolerance, allowing nontryptic and missed tryptic cleavages, and allowing variable modifications (e.g., oxidation or phosphorylation). Thus, when including these in the search strategy, results should be evaluated critically to determine if there is a significant improvement in protein identification or sequence coverage to justify the increased computational cost and reduced statistical discrimination. These results also suggest that carrying out several searches on the same data set using different parameters may increase extraction of information from the data. In addition, experiments investigating physiological covalent peptide modifications should be optimized by first identifying the proteins that are present and then searching for modified peptides against a limited set of observed proteins.

An important result was that as the size of the data sets increased, the number of false positives measured by the randomized database search did not increase disproportionately. Instead, the frequency of false positives normalized to class positives remained constant at ~4%, whereas the statistical false positive rates (FP normalized to class negatives) actually decreased from sample 1 (2.6%) to sample 2 (0.7%). Furthermore, if the number of false positive peptide assignments increased disproportionately with the data set size, we should have observed an increase in the percentage of proteins supported by only one peptide. In fact, this percentage decreased from 69 to 45% between sample 1 and sample 3. This suggests that the class positive DTA files are described by different sampling statistics than the class negative DTA files, when filtered by MSPlus/Isoform Resolver. We hypothesize that the combined use of SCX, RSP, and ion ratio filters by MSPlus allows partial evaluation of peptide chemical properties and thus treats class positive versus class negative sets differently. Further studies are needed to clarify this issue.

We also describe a novel protein profiling approach that eliminates peptide and protein sequence redundancy. The removal of isoforms can also be carried out by DTA Select and Peptide Prophet programs,[7,20] but these do not allow filtering based on peptide chemical properties. Furthermore, the MSPlus/Isoform Resolver approach makes use of a new nomenclature for reporting ambiguities due to protein variants. This will facilitate comparative studies, where differential expression of specific variants may be important to understanding biological regulation. Although the protein numbering and peptide nomenclature currently implemented in the Isoform Resolver are arbitrary, numbering can be adapted to report systematic relationships such as gene loci, functional classes, or evolutionary relationships. Taken together, the analytical and computational methods described in this study provide a comprehensive set of tools for optimizing data analysis in shotgun proteomics and demonstrate high coverage of the soluble components of a mammalian cell proteome.

ACKNOWLEDGMENT

We especially thank Rich Johnson (Amgen) for his support, encouragement, and many helpful discussions on all aspects of mass spectrometry, and Dean Thompson (Agilent Labs) for assistance with computational and database issues. We are also indebted to Jeff Kowalak (NIH) for providing software used to parse Mascot data, Joel Sevinsky (U. Colorado) for help with cell culture and harvesting, Claire Haydon (HHMI) and Andy Alpert (PolyLC) for assistance with SCX chromatography, and Jimmy Eng (Institute for Systems Biology) for a script calculating peptide p*I* and for suggesting the use of randomized databases. We are also grateful to Alex Taylor (Amgen), Richard Smith (Pacific Northwest National Laboratory), and Jimmy Eng, Andrew Keller, and Andrei Nesvizhskii (Institute for Systems Biology) for helpful discussions and sharing information. Funding was provided by NIH Grants CA87648 (K.A.R.) and GM48521 (N.G.A.).

SUPPORTING INFORMATION AVAILABLE

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

Received for review October 16, 2003. Accepted March 12, 2004.

AC035229M

[20] Tabb, D. L.; McDonald, W. H.; Yates, J. R. *J. Proteome Res.* **2002**, *1*, 21–26.