Anal. Chem. **2006,** *78,* 1071−1084

Improving Sensitivity in Shotgun Proteomics Using a Peptide-Centric Database with Reduced Complexity: Protease Cleavage and SCX Elution Rules from Data Mining of MS/MS Spectra

Chia-Yu Yen,[†] Steve Russell,[‡] Alex M. Mendoza,[‡] Karen Meyer-Arendt,[‡] Shaojun Sun,[†] Krzysztof J. Cios,[†,||] Natalie G. Ahn,[‡,⊥] and Katheryn A. Resing*[,‡]

Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center, Denver, Colorado 80217-3364, Center of Computational Pharmacology, University of Colorado at Denver and Health Sciences Center, Aurora, Colorado 80045-0511, Department of Chemistry and Biochemistry and Department of Computer Science, University of Colorado at Boulder, Boulder, Colorado 80309, and Howard Hughes Medical Institute, Boulder, Colorado 80309-0215

Correct identification of a peptide sequence from MS/MS data is still a challenging research problem, particularly in proteomic analyses of higher eukaryotes where protein databases are large. The scoring methods of search programs often generate cases where incorrect peptide sequences score higher than correct peptide sequences (referred to as distraction). Because smaller databases yield less distraction and better discrimination between correct and incorrect assignments, we developed a method for editing a peptide-centric database (PC-DB) to remove unlikely sequences and strategies for enabling search programs to utilize this peptide database. Rules for unlikely missed cleavage and nontryptic proteolysis products were identified by data mining 11 849 high-confidence peptide assignments. We also evaluated ion exchange chromatographic behavior as an editing criterion to generate subset databases. When used to search a well-annotated test data set of MS/MS spectra, we found no loss of critical information using PC-DBs, validating the methods for generating and searching against the databases. On the other hand, improved confidence in peptide assignments was achieved for tryptic peptides, measured by changes in ΔCN and RSP. Decreased distraction was also achieved, consistent with the 3−9-fold decrease in database size. Data mining identified a major class of common nonspecific proteolytic products corresponding to leucine aminopeptidase (LAP) cleavages. Large improvements in identifying LAP products were achieved using the PC-DB approach when compared with conventional searches against protein databases. These results demonstrate that peptide properties can be used to reduce database size, yielding improved accuracy and information capture due to reduced distraction, but with little loss of information compared to conventional protein database searches.

High-throughput methods for protein identification are enabled by the development of mass spectrometers capable of automated data collection, where peptide ions are fragmented to produce MS/MS spectra that provide information about the peptide sequences. Identifying peptide sequences with high accuracy and sensitivity is a challenging problem, and several search programs, such as Sequest and Mascot, have been developed that assign peptide sequences to MS/MS spectra.[1] These programs compare observed spectra with theoretical spectra derived from a protein database, assess the goodness of fit between experiment and theory, and then report the top-scoring sequences for each MS/MS spectrum.

A key problem with current programs is that their scoring methods do not always distinguish correct from incorrect sequence assignments.[2] One common solution is to identify a score threshold above which the assignment will be accepted (high confidence threshold). Typically, this threshold is identified by searching the data set against a randomized database. From this result, the number of false positive assignments at different thresholds can be evaluated.[3] However, when applying a stringent

* To whom correspondence should be addressed. Phone: 303-735-4019. Fax: 303-492-2439. Katheryn.Resing@Colorado.Edu.

[†] Department of Computer Science and Engineering, University of Colorado at Denver and Health Sciences Center.

[‡] Center of Computational Pharmacology, University of Colorado at Denver and Health Sciences Center.

[§] Department of Chemistry and Biochemistry, University of Colorado at Boulder.

[||] Department of Computer Science, University of Colorado at Boulder.

[⊥] Howard Hughes Medical Institute.

(1) Sadygov, R. G.; Cociorva, D.; Yates J. R. *Nat. Methods* **2004**, *1*, 195−202.

(2) MacCoss, M. J.; Wu, C. C.; Yates J. R. *Anal. Chem*. **2002**, *74* (21), 5593−5599.

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

10.1021/ac051127f CCC: $33.50
Published on Web 01/13/2006

threshold to searches for unmodified, tryptic peptides against a moderately large database (such as the human International Protein Index (IPI) database with ~48 000 entries), as many as half of the MS/MS spectra can yield scores below this threshold.[4] To minimize the number of false negatives, investigators often accept assignments below threshold, allowing more false positive assignments. This approach is usually combined with some method to filter the resulting list in order to reduce the number of false positives.[5] A serious limitation to this approach occurs when an incorrect assignment scores higher than the correct assignment, an event we refer to as "distraction". No current filtering approach can identify the correct assignment when distraction occurs and the search programs fails to assign the correct sequence to the top-ranking position.

Recently, several researchers noted that a lower threshold is obtained with smaller databases[2,3] and that this also leads to lower distraction.[4] This has a relatively large impact on the identification of peptides in higher eukaryotes where the database size is 30 000 open reading frames (ORFs) or more, compared with ~6000 ORFs in yeast or <1000 ORFs in some bacterial systems. The problem is exacerbated when search strategies include products of nonspecific or missed proteolytic cleavages or modified peptides in order to capture additional information; these strategies greatly increase the effective database size and the rate of distraction. One reported solution has been to carry out an initial search to identify proteins present in the sample and then use only peptides derived from those proteins in a secondary search to capture additional information.[6] This captures more peptide assignments for many of those proteins but does not allow the identification of new proteins. This illustrates the general principle that search strategies are chosen to strike a balance between greater information capture and distraction.

A major contributor to distraction is the fact that search programs consider all possible peptide sequences from full length proteins, by "predigesting" protein database entries into an indexed list of peptides. Peptides are then selected for comparison with each MS/MS spectra based on how closely they match the observed mass within a specified tolerance. This process does not enable use of information about peptide chemical properties to intelligently distinguish between those peptide candidates that are probable versus the large number that are improbable. Furthermore, the indexed files are in a proprietary format, so they cannot be edited. Edwards and Lippert[7,8] proposed algorithms to construct a compressed peptide database in order to deal with the database size problem, but the explicit connections to the original proteins and peptides were lost. These authors discussed issues regarding the generation of trypsin products, but did not consider missed cleavage and nontryptic products or chemical properties. Furthermore, these studies did not assess whether the compression algorithms and use of suffix trees to filter the sequences led to search artifacts or information loss, nor did they consider the effects of including multiple constraints.

To study the effect of using several peptide properties to select candidate peptide sequences, we developed methods for guiding the indexing algorithm to include only those specific sequences that we want to test in our search strategy. This allowed us to utilize peptide-centric databases (PC-DBs) in place of a protein database. We show that editing this peptide database with user-specified rules for exclusion of unlikely peptides provides greater discrimination between correct versus incorrect sequence assignments and also minimizes distraction, without removing or diminishing the critical information contained in the full database. In addition, the most common class of nonspecific cleavages can be accommodated without a significant increase in distraction. The approach is easily implemented and can be applied with any search program.

METHODS

The MS/MS data set used in this study was collected by MudPIT analysis of a soluble protein extract from an erythroleukemia K562 cell line, as previously described (sample 1, Resing et al.[4]). Briefly, proteins were exchanged into 100 mM $NaHCO_3$/0.5 mM $CaCl_2$ and proteolyzed by addition of three aliquots of 1% trypsin (Wako, unmodified porcine). Samples were frozen and lyophilized to remove $NaHCO_3$ and resuspended in 5 mM K_2HPO_4/5% acetonitrile for SCX chromatography. SCX fractions were then analyzed by reversed-phase LC/MS/MS using an LCQ Classic ion trap mass spectrometer (Thermo-Electron), using six overlapping narrow mass ranges.

Sequest and Mascot programs were used to search peptide sequences against the IPI human protein database (v. 2.18, April 10, 2003). The work flow for protein searching and peptide validation using the in-house program MSPlus[4] is shown in Figure 1A. MS/MS were converted to DTA text files using TurboSequest (v. 27 rev. 12), with intensity threshold of 10 000, allowed grouping of 1−5 scans, and minimum ion count of 35. An in-house script concatenated these files into Mascot Generic Files (MGFs) for searching with Mascot (v. 1.9). Searches were initiated by command line and specified a peptide mass tolerance of 2.5 Da (average mass) and fragment ion mass tolerance of 1.0 Da. When searching with the IPI protein database, up to two missed tryptic cleavages were allowed. The small, targeted PC-DBs were generated by programs using regular expression, as described below, and then converted to a FASTA formatted file, with each peptide corresponding to one FASTA entry. The FASTA file was then sent into Sequest or Mascot for indexing; to prevent the programs from further "digestion" of the peptide sequences, an imaginary enzyme was specified that would cleave C-terminal to an amino acid (X) nonexistent in the database. In-house software for parsing Sequest.OUT files and a modified version of DBParser for Mascot.DAT files[9] were used to extract and transfer information from the Sequest and Mascot results files into an Oracle 9i database.

To identify and validate peptide assignments, we used our MSPlus program,[4] which evaluates consensus between the search

(4) Resing, K. A.; Meyer-Arendt, K. E.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russel, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. *Anal. Chem.* **2004**, *76* (13), 3556−3568.

(5) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. *J. Proteome Res.* **2003**, *1*, 137−146.

(6) Craig, R. and Beavis, R. C. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2310−2316.

(7) Edwards, N. and Lippert, R. *Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI 2002)*, Springer-Verlag: New York, 2002; pp 68−81.

(8) Edwards, N. and Lippert, R. *4th Workshop on Algorithms in Bioinformatics (WABI 2004)*, Bergen, Norway, 2004.

(9) Yang, X.; Dondeti, V.; Dezube, R.; Maynard, D. M.; Geer, L. Y.; Epstein, J.; Chen, X.; Markey, S. P.; Kowalak, J. A. *J. Proteome Res.* **2004**, *3* (5), 1002−1008.

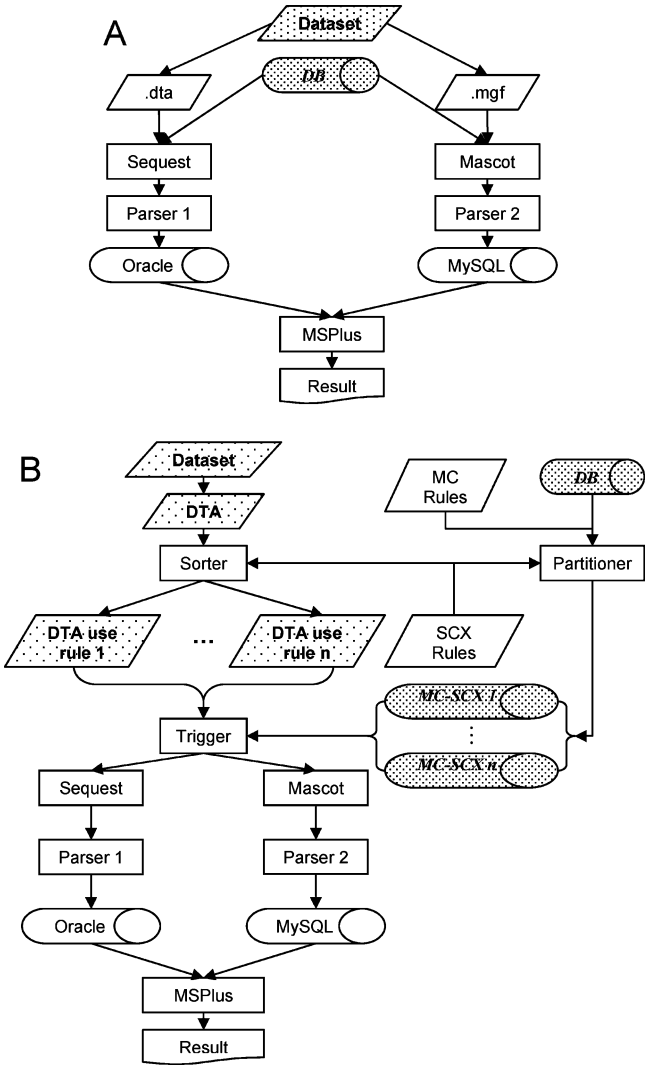

Figure 1. Conventional and PC-DB data work flows. (A) Conventional search work flow previously described by Resing et al..[4] Each MS/MS file is searched in parallel using Sequest or Mascot against the IPI protein database. Results are parsed into a relational database, and results are filtered using MSPlus, which evaluates assignments based on consensus between Sequest and Mascot, RSP = 1, and consistency of the ion charge state with sequence. (B) Prefiltered PC-DB work flow, introduced in this study. Small PC-DBs are developed from the IPI-DB, selecting for sequences that follow rules for missed cleavages (MC) and partitioning databases according to SCX rules. DTA files are matched to appropriate SCX databases using a Sorter program and then searched against that PC-DB in parallel using Sequest and Mascot. Results are filtered by MSPlus as above. The new work flow includes Partitioner and Sorter as self-organizing modules. Both programs generate sets of output according to MC and SCX rules generated from data mining, as described in the text.

programs as the primary acceptance filter and then removes some false positives by requiring RSP = 1 (RSP is a Sequest score showing the rank assigned a peptide during preliminary scoring by "SP", which evaluates plausibility of the fragment ions in the MS/MS file). In our original work flow, additional filters were used by MSPlus in order to eliminate peptide sequence assignments containing (i) KK, KR, RK, or RR residues other than those near N- or C-termini, (ii) numbers of free amines insufficient to account for ion charge, (iii) numbers of basic residues inconsistent with SCX elution, (iv) scores lower than a minimum threshold for Mowse or XCorr or for a combined score that weights Mowse and XCorr equally (SumScore), and (v) low intensity. In some analyses, we disabled one or more of these filters, as indicated in the text. The MSPlus program creates a "comma separated value" (.CSV) file that summarizes the search results for each DTA file, evaluating the highest ranked sequence from Sequest and highest two sequences from Mascot. If the two search programs agree and meet the filtering criteria, the assignment is classified as high confidence. The false discovery rate (FDR) using this approach is ~4.0%, based on searching a database of inverted sequences, or 2% based on manual analysis. For peptides with assignments classified as stringent, either XCorr (from Sequest) or Mowse (from Mascot) scores must exceed the highest scores obtained by any MS/MS spectrum in the inverted sequence search. The FDR for these stringent cases is <0.5%.

Search results were compared, and all assignments accepted by MSPlus were evaluated manually with assistance of an in-house program that summarizes all information about a spectrum, possible fragment ion assignments, and mass errors (manuscript in preparation). Criteria for manual validation required that a peptide sequence must account for >80% of the total ion intensity of the MS/MS spectrum and that the major ions must be chemically plausible. Assessment of chemical plausibility was aided by comparison with predicted spectra generated by the program, MassAnalyzer,[10] requiring a Similarity score between the predicted spectra and the observed spectra greater than 0.5. In addition, heuristic tests of chemical plausibility were imposed; examples of application of these tests in manual analysis have been described previously (Supporting Information in ref 4). In addition to normal b and y ions, multiply charged fragment ions, internal fragment ions, and possible multiple dehydrations are allowed, when consistent with the peptide sequence and lability of the peptide bonds involved. These assignments must account for >85% of the ion current in a spectrum and all moderate to intense fragment ions; lower percentages are accepted if there is evidence that a secondary peptide was captured in the MS/MS isolation window and most of the chemical noise can be accounted for by that peptide sequence. For this study, we added an additional test for possibility of a semitryptic or nontryptic peptide assignment, because a very high quality spectrum of this type can give high scores and consensus between Sequest and Mascot with an incorrect tryptic assignment. For analyses of XCorr and ΔCN thresholds for validation, ΔCN was 0.08 and XCorr thresholds were as follows: MH^+ (1.8), MH_2^{2+} (2.5), and MH_3^{3+} (3.8).[11]

In the new work flow (Figure 1B), a PC-DB was generated by predigesting all proteins in the IPI database. Peptide sizes were restricted as described in Results, and peptide sequences were written into a plain text file. A program called "MissedCleavageFilter" extracted peptides based on regular expression rules for the likely missed cleavage products.[12−14] The resulting peptide lists were output as text files and sorted by a "Partitioner" program

(10) Zhang, Z. *Anal. Chem.* **2004**, *76* (14), 3908−3922.
(11) Benzinger, A.; Muster, N.; Koch, H. B.; Yates, J. R. 3rd; Hermeking, H. *Mol. Cell. Proteomics* **2005**, *4* (6), 785−795.
(12) http://en.wikipedia.org/wiki/Regular expression.
(13) Friedl, J. E. F. *Mastering Regular Expressions*, 2nd ed; O'Reilly: Sebastopol, CA, 2002.
(14) http://dev.mysql.com/doc/mysql/en/Regexp.html.

Table 1. Summary of Protein and Peptide-Centric Databases Used in This Study[a]

database	no. of peptides	IDX size (MB) [b]
protein IPI-DB	n/a	1.44
FULL PC-DB	2 850 058	86.90
MC PC-DB	1 494 690	45.60
MC-SCX PC-DB 1	330 633	10.00
MC-SCX PC-DB 2	323 091	9.85
MC-SCX PC-DB 3	845 701	25.80
MC-SCX PC-DB 4	522 610	15.90
MC-SCX PC-DB 5	942 984	28.70

[a] Shown are the number of peptides in each PC-DB after filtering by missed cleavage (MC) and cation exchange chromatography (SCX) rules. Application of MC and SCX rules in combination leads to greatest restriction of database size. [b] Sizes of index files generated by Sequest. Although the number of peptides specified by the IPI protein database and FULL PC-DB are comparable, the sizes of their indexed files differ significantly. This accounts for the longer Sequest search times observed using PC-DBs compared to the protein DB in Figure 4.

into subsets based on rules for consistency with SCX elution behavior. A "Peptide2Fasta" script then converted each peptide list subset into FASTA format for Sequest and Mascot search engines. A "Sorter" program sorted and divided all DTA files into different groups using the same SCX rules applied by Partitioner. A "ResultCompare" program then compared the MSPlus results from two different searches, listing results in three panes in one window. The first pane shows the overlaps where the assignments were identical in each search. Different assignments are reported in the second and third panes, each representing the high-confidence assignments unique to one of the searches. In studies examining the effects of changing the stringent threshold, this new work flow was also used generate restricted peptide-centric databases from the inverted protein database.

RESULTS

Searching Peptide-Centric Databases with Sequest and Mascot. Before applying an edited PC-DB as input to an MS/MS search program, it was important to ensure that the use of a PC-DB did not introduce artifacts into the searches. Therefore, we carried out a preliminary test with a "FULL PC-DB", which contained all tryptic peptides that would be generated after indexing the IPI protein database ("IPI-DB") by Sequest or Mascot, allowing up to two missed cleavages. After in silico digestion, a limit was placed on the highest molecular mass allowed, based on the data collection range (up to $MH^+ = 4500$). A minimum peptide length of nine amino acids was specified, because previous surveys showed that smaller peptides have a higher probability of being misassigned by search programs.[1] The size of the FULL PC-DB is 2.8×10^6 peptide sequences (Table 1). The use of PC-DBs as input to Sequest and Mascot required a strategy to prevent in silico proteolysis of peptides with missed cleavage sites during the indexing process. Each peptide sequence was listed as a separate FASTA entry, and then an imaginary enzyme, which cleaves at a nonexistent amino acid, was specified to avoid further cleavage of each sequence.

We carried out performance studies using a "test" shotgun proteomics data set of proteins from a soluble cell lysate of a human erythroleukemia cell line (K562).[4] The data set is small (2117 MS/MS) because we used 17 SCX fractions, each analyzed using a short elution gradient on RP-HPLC directly coupled to a mass spectrometer (LC/MS/MS). The small size makes it ideal as a test sample, because we have manually validated or rejected >70% of the DTAs, including all "hits" where Sequest and Mascot reached consensus. (See Supporting Information, Table 1 and Figure 2, for validated assignments and all scores from Sequest and Mascot; raw data and DTA files are available from the authors upon request.) When searched against the IPI-DB using our MSPlus approach ("conventional work flow", Figure 1A), our previous studies showed that the data set yielded 826 high-confidence assignments.[4] Manual analysis showed that 10 of these assignments are false positives. We previously estimated that this data set should yield ~1134 MS/MS identifiable by a search strategy specifying two missed tryptic cleavages with no modifications. Of these, ~200 MS/MS were estimated to be misassigned due to distraction (lower than top assignment by Sequest or Mascot or Sequest RSP score greater than one) and would therefore not be identified by MSPlus. We also estimated that ~100 MS/MS were correctly assigned but rejected by MSPlus, due to low scores. About 60% of these were identified in the original study; since then, we have manually validated a total of 187 cases of distraction (manuscript in preparation). The remaining ~985 MS/MS spectra were not identifiable because they included nontryptic peptides, in-source-generated fragment ions, peptides with oxidized amino acids and other posttranslational modifications, and weak MS/MS spectra. Such peptides were not considered during searching in order to minimize the size of the database and degree of distraction.

Searches of the test data set against the FULL PC-DB versus the IPI-DB using MSPlus filtering yielded the same peptide assignments and scores (XCorr and Mowse) for most spectra (see below, Table 4). In the search against the FULL PC-DB, 829 MS/MS were validated by MSPlus; of these, no MSPlus-validated assignments were lost compared to the IPI-DB search results, and three were gained. Two of the new assignments were those where Sequest and Mascot reached consensus, but with RSP = 3 in the IPI-DB search, and therefore were rejected by MSPlus. Their masses were slightly higher than the lower mass size cutoff of 950 Da; thus, the promotion to RSP = 1 when using the FULL PC-DB was primarily due to omission of short peptides with length less than nine amino acids. The third new assignment provided interesting insight into the different ways that Mascot and Sequest handle Met at the first position, which in proteins often corresponds to an initiator Met residue. Searching the PC-DB with Mascot gave rise to a correct assignment of VNHFIAEFK (manually validated), representing loss of Met from the N-terminal position of peptide (R)MVNHFIAEFK. (The Met form was also observed several times in the data set.) By presenting this peptide to Mascot as if it were a protein FASTA entry, the N-terminal Met was removed during the indexing. This revealed a possible leucine aminopeptidase product, a class of peptides that we found was highly represented (see below). In contrast, Sequest did not yield this product, because such a modification must be user specified. Overall, the results showed that the use of a PC-DB did not introduce significant artifacts, producing results comparable to a normal protein database search, and revealing three additional correct assignments.

Identifying Missed Cleavage Rules by Mining Proteomics Data Sets. Next, we constructed a smaller database based on the assumption that not all missed cleavage (MC) products of trypsin were equally likely in a tryptic digest and that only a subset of possible products would actually be observed. This MC PC-DB included normal tryptic peptides, along with the most likely set of MC products. To determine the maximum number of missed cleavages that were actually observed in our test data set, we directed the search programs to consider nonspecific proteolytic cleavages, thus allowing an indefinite number of missed cleavages. Peptides were then filtered for trypsin specificity at both the N- and C-terminal ends, considering only those sequences identified with very high confidence, by setting thresholds of XCorr > 3.3, Mowse > 53, or both (resulting in false discovery rate of <0.5%). Approximately 700 peptides were observed, of which >99% showed two or fewer missed cleavages. No internal KR, RK, RR, or KK sequences or longer K/R strings were observed, indicating that at least one of two adjacent Arg or Lys residues would almost always be cleaved by trypsin. Alternatively, we compared searches that specified trypsin cleavages, allowing either two or three missed cleavages. No additional high-scoring peptides were found in searches allowing three missed cleavages, as compared to two. We concluded that the majority of peptides generated by our trypsin digestion protocol contained up to two incomplete cleavages and that sequences with three or more missed cleavages are far less probable.

We then examined the amino acids around each missed cleavage site, to identify sequence rules for the missed cleavage products. For this analysis, we used 11 849 high-confidence sequence assignments from a data set of 20 675 unique peptides (sample 3, described in Resing et al.[4]). Of these, 3405 peptides contained incomplete cleavage sites and were used for further analysis. Specific cases of cleavage at KP or RP were not counted in this analysis, because trypsin does not cleave at these residues. When a peptide ion was observed that would have required cleavage at a KP or RP site, it could almost always be attributed to in-source fragmentation during MS analysis. The data set excluded ~80% of the peptides observed only as MH^{1+} ions, none of which contained missed cleavage sites.

First, we examined the likelihood of observing missed cleavages within four residues of the N- or C-terminus (Figure 2). These products are generated by the inability of trypsin to act as an exopeptidase (cleaving near the end of a peptide) after an initial cleavage at a site with multiple Arg or Lys residues within four residues of each other.[15] For example, at a cleavage site such as x_nKKx_n, where x indicates any amino acid and x_n indicates a series of any amino acids, trypsin will usually cleave at only one Lys residue, often selected randomly. The resulting x_nKK or Kx_n products are then resistant to further proteolysis by trypsin (referred to as exopeptidase activity). For example, cleavages consistent with this model generate 7% of peptides with Lys located at position N, while 2.1% had Lys at position C-1 (C-terminal K/R−K/R) (where N represents the N-terminal position, and C represents the C-terminal Arg or Lys residue). In addition, trypsin often cleaves at only one of two basic residues located close together; for example, x_nKxKx_n will produce x_nKxK or xKx_n. Thus, 4.5% of peptides showed Lys at N + 1. Lower percentages were observed at the N + 2, N + 3, C − 2, and C − 3 positions (Figure 2). In general, cases caused by incomplete cleavages generated highest frequencies of Lys at positions N, N + 1, and C − 1. Thus, our initial rules allowed missed cleavages at these positions. Two peptides with three adjacent Lys residues at their C-termini were also observed.

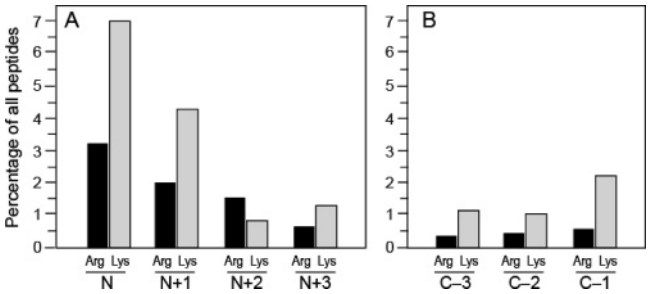

Figure 2. Locations of missed cleavages are nonrandom. Frequencies of internal Lys or Arg residues (not followed by Pro) at the four N-terminal positions or the last three positions proximal to the C-terminus in validated peptides. Data were quantified from 11 849 high-confidence sequence assignments. Missed cleavages located at the first and second positions were most frequent, for both Arg and Lys. Representation of Lys immediately before the C-terminus (C − 1) was enhanced, although surprisingly, this was not observed for Arg.

A similar analysis of Arg residues showed lower frequencies of missed cleavages near the ends of peptides at positions N, N + 1, and C − 1, compared to Lys. These differences were disproportionate to the frequency of Arg and Lys in the data set (Supplorting Information, Table 2), suggesting that the trypsin cleavage rules for Arg differ slightly from those for Lys. These results indicate that, in sequences containing x_nKRx_n or x_nRKx_n, trypsin proteolysis occurs more frequently at the Arg residue, producing x_nKR or $x_nR + Kx_n$ products, respectively, while cleavages at x_nRRx_n will produce mainly $x_nR + Rx_n$ products. Likewise, Arg was infrequent at positions N + 2 and C − 2, indicating that cleavage at Arg was efficient, even when located within two residues from another basic residue. Thus, rules for missed cleavages at Arg allow a more restricted set of sequences, compared to Lys.

Next, we considered internal cleavage sites, where the missed cleavage position was at least four residues from the N-terminal end of the peptide or three residues from the C-terminal K/R. A few sequence assignments were observed with two internal adjacent K/R residues, but were rejected following manual analysis or because the number of basic residues in the peptide sequence were chemically inconsistent with peptide SCX elution behavior.[4] Often, these represented K/Q isoforms, where the search program had chosen the wrong isoform (K and Q cannot be distinguished by ion trap MS instruments).

Distributions of the amino acids flanking internal Lys and Arg residues were then examined in order to develop additional rules defining allowable missed cleavage sites (excluding KP and RP, as described above). Table 2A shows absolute frequencies of missed cleavage at Lys. Normalized values, which adjust for the number of total occurrences of each amino acid, are also presented. Likewise, Table 2B shows absolute and normalized frequencies of amino acid residues near internal Arg missed cleavage sites. Similar patterns were seen between Lys and Arg

(15) Hill, R. L.; Craik, C. S. *Adv. Protein Chem.* **1965**, *20*, 37−107.

Table 2

AAs at indicated positions[b]	$j-4$	$j-3$	$j-2$	$j-1$	$j+1$	$j+2$	$j+3$	$j+4$	
				A. Amino Acids Proximal to Missed Cleavages at Lys[a]					
A	48 (0.92)	47 (0.91)	52 (0.99)	14 (0.27)	32 (0.61)	14 (0.27)	42 (0.80)	54 (1.03)	A
C	5 (1.46)	3 (0.88)	3 (0.87)	5 (1.46)	1 (0.29)	1 (0.29)	3 (0.88)	0 (0.00)	C
D	51 (1.25)	91 (2.26)	62 (1.52)	168 (4.12)	176 (4.32)	106 (2.60)	48 (1.18)	46 (1.13)	D
E	85 (1.47)	156 (2.73)	120(2.08)	116 (2.01)	166 (2.87)	178 (3.08)	114 (1.98)	80 (1.39)	E
F	25 (1.23)	16 (0.79)	12 (0.59)	18 (0.88)	14 (0.69)	11 (0.54)	25 (1.23)	21 (1.03)	F
G	38 (0.81)	29 (0.62)	55 (1.17)	31 (0.66)	36 (0.77)	26 (0.55)	38 (0.81)	29 (0.62)	G
H	13 (0.74)	11 (0.64)	9 (0.52)	7 (0.40)	14 (0.80)	13 (0.74)	8 (0.46)	13 (0.75)	H
I	46 (1.50)	19 (0.62)	22 (0.72)	42 (1.37)	13 (0.42)	29 (0.94)	49 (1.60)	42 (1.37)	I
K	10 (0.30)	6 (0.18)	5 (0.15)	0 (0.00)	0 (0.00)	0 (0.00)	2 (0.06)	51 (1.51)	K
L	78 (1.32)	39 (0.67)	43 (0.73)	71 (1.20)	43 (0.73)	29 (0.49)	73 (1.24)	59 (1.00)	L
M	7 (0.61)	11 (0.96)	11 (0.96)	10 (0.87)	4 (0.35)	5 (0.43)	13 (1.13)	10 (0.87)	M
N	22 (0.85)	28 (1.09)	27 (1.04)	42 (1.62)	28 (1.08)	23 (0.89)	23 (0.89)	29 (1.12)	N
P	23 (0.56)	19 (0.47)	74 (1.82)	16 (0.39)	0 (0.00)	84 (2.06)	12 (0.30)	33 (0.81)	P
Q	28 (0.87)	28 (0.87)	23 (0.71)	15 (0.46)	17 (0.53)	21 (0.65)	26 (0.81)	28 (0.87)	Q
R	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	20 (1.05)	R
S	34 (0.75)	38 (0.85)	41 (0.91)	29 (0.64)	10 (0.22)	26 (0.58)	45 (1.00)	46 (1.02)	S
T	37 (1.06)	28 (0.81)	27 (0.77)	18 (0.52)	20 (0.57)	26 (0.75)	30 (0.86)	21 (0.60)	T
V	50 (1.20)	31 (0.75)	32 (0.77)	18 (0.43)	33 (0.79)	25 (0.60)	59 (1.42)	40 (0.96)	V
W	9 (1.87)	4 (0.84)	6 (1.25)	1 (0.21)	3 (0.62)	8 (1.66)	2 (0.42)	3 (0.63)	W
Y	25 (1.70)	24 (1.65)	10 (0.68)	13 (0.89)	24 (1.63)	9 (0.61)	20 (1.37)	7 (0.48)	Y
				B. Amino Acids Proximal to Missed Cleavages at Arg[a]					
A	10 (1.27)	10 (1.27)	9 (1.15)	5 (0.64)	9 (1.15)	13 (1.65)	8 (1.02)	12 (1.53)	A
C	0 (0.00)	0 (0.00)	2 (3.89)	0 (0.00)	1 (1.94)	2 (3.89)	1 (1.94)	1 (1.94)	C
D	14 (2.29)	10 (1.64)	10 (1.64)	26 (4.26)	23 (3.77)	10 (1.64)	7 (1.15)	2 (0.33)	D
E	9 (1.04)	18 (2.08)	24 (2.77)	9 (1.04)	19 (2.19)	25 (2.89)	16 (1.85)	15 (1.73)	E
F	3 (0.98)	2 (0.66)	2 (0.66)	4 (1.31)	3 (0.98)	0 (0.00)	1 (0.33)	1 (0.33)	F
G	9 (1.28)	7 (0.99)	14 (1.99)	12 (1.70)	13 (1.85)	4 (0.57)	13 (1.85)	6 (0.85)	G
H	2 (0.76)	6 (2.29)	1 (0.38)	0 (0.00)	0 (0.00)	1 (0.38)	0 (0.00)	1 (0.38)	H
I	5 (1.09)	2 (0.43)	3 (0.65)	4 (0.87)	3 (0.65)	0 (0.00)	5 (1.09)	5 (1.09)	I
K	4 (0.79)	2 (0.39)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	7 (1.38)	K
L	6 (0.68)	3 (0.34)	1 (0.11)	12 (1.35)	3 (0.34)	3 (0.34)	13 (1.47)	6 (0.68)	L
M	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.58)	3 (1.74)	M
N	8 (2.06)	5 (1.29)	2 (0.52)	2 (0.52)	4 (1.03)	3 (0.77)	3 (0.77)	2 (0.52)	N
P	5 (0.82)	9 (1.47)	11 (1.80)	7 (1.15)	0 (0.00)	25 (4.09)	6 (0.98)	4 (0.66)	P
Q	3 (0.62)	1 (0.21)	2 (0.41)	3 (0.62)	0 (0.00)	1 (0.21)	7 (1.45)	5 (1.03)	Q
R	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	5 (1.74)	R
S	5 (0.74)	8 (1.18)	2 (0.30)	0 (0.00)	7 (1.03)	1 (0.15)	4 (0.59)	7 (1.03)	S
T	5 (0.96)	5 (0.96)	4 (0.77)	2 (0.38)	3 (0.57)	3 (0.57)	2 (0.38)	7 (1.34)	T
V	5 (0.80)	4 (0.64)	7 (1.12)	8 (1.28)	5 (0.80)	3 (0.48)	7 (1.12)	4 (0.64)	V
W	1 (1.39)	0 (0.00)	0 (0.00)	1 (1.39)	0 (0.00)	0 (0.00)	1 (1.39)	1 (1.39)	W
Y	1 (0.45)	3 (1.36)	1 (0.45)	0 (0.00)	2 (0.91)	1 (0.45)	0 (0.00)	1 (0.45)	Y

[a] Frequencies of amino acids surrounding Lys or Arg residues located at a missed cleavage site (position j). Frequency of amino acids (indicated by one-letter codes) at positions $j-1$ to $j-4$ (N-terminal to Lys/Arg), and $j+1$ to $j+4$ (C-terminal to Lys/Arg). Note that the most common Lys and Arg missed cleavages are those followed by a Pro residue, which is a known property of trypsin specificity, and have been omitted from the table. The second most common missed cleavage occurs at Lys/Arg residues in proximity to acidic amino acids Asp or Glu. [b] Frequencies of each amino acid are normalized by their frequency of occurrence in the set of 11 849 high-confidence peptide sequences. Normalized frequencies (in parentheses) were calculated by the expression $(F_{AA}/F_{Total})/P_{AA}$, where F_{AA} is the indicated frequency of each amino acid at a given position, F_{Total} is the sum of all amino acids at that position, and P_{AA} is the fraction of that amino acid in the entire data set (Supporting Information, Table 2). A normalized frequency equal to 1 is that which would be obtained randomly. A normalized frequency $\gg 1$ indicates a significantly higher probability of observing the residue at this site.

missed cleavages, although the frequencies of missed cleavages at Arg were lower, for reasons discussed above.

After normalization, we observed that Asp and Glu residues immediately adjacent to Lys or Arg missed cleavages were most highly represented ($j-1$ or $j+1$, where j is the position of the basic residue); this is consistent with the known propensity of trypsin for reduced proteolysis close to acidic residues. In addition, acidic residues were highly represented at $j \pm 2$ or $j \pm 3$, but more careful examination showed that, in such cases, at least one other Asp or Glu residue was nearby. The cleavage patterns in these cases included ExKxE, xKxEE, or EExKx, where E could be either Glu or Asp and K could be either Arg or Lys. These observations were consistent with the known lower affinity of trypsin for acidic peptides.[16]

Unexpectedly, higher frequencies of Gly residues were detected $j-2$ to Arg missed cleavages (Table 2B). In these cases, inspection of individual peptides showed Glu or Asp adjacent to the site of missed cleavage, revealing covariance between Gly and Glu/Asp residues, and indicating that proximal acidic residues represented the dominant mechanism for missed cleavage at GxR. Similarly, higher frequencies of Pro residues were detected $j \pm 2$ to Lys or Arg missed cleavages (Table 2A,B). In these cases, removal of sequences containing acidic residues adjacent to the missed cleavage site reduced the frequency of Pro at $j+2$ to Lys or Arg by 8-fold, reduced Pro at j+ $-$ 2 to Lys by 48-fold, and removed significant occurrence of Pro at $j-2$ to Arg. Thus, the

(16) Perona, J. J.; Craik, C. S. *Protein Sci.* **1995**, *4*, 337−360.

Table 3. Regular Expressions Specifying Missed Cleavage Rules

	regular expression[a]	included pattern	excluded pattern
1	[∧KR][KR][DE]	−XKE−	−KKE−
2	[DE][KR][∧KR]	−EKX−	−EKK−
3	[KR][∧KR][ED][ED]	−KXEE−	−KKEE−
4	[ED][ED][∧KR][KR]	−EEXK−	−EEKK−
5	[ED][∧KR][KR][∧KR][ED]	−EXKXE−	−EKKXE−, −EXKKE−, −EKKKE−
6	[KR]P	−KP−	
7	∧[KR][∧KR]	KX−	KK−
8	∧.[KR][∧KR]	XKX−	XKK−
9	∧.[∧KR][KR][∧KR]	XXKX−	XKKX−, XXKK−, XKKK−
10	[∧KR][KR][KR]$	−XKK	−KKK
11	[∧KR][KR][∧KR][KR]$	−XKXK	−KKXK, −XKKK, −KKKK
12	[KR][∧KR][∧KR][KR]$	−KXXK	−KKXK, −KXKK, −KKKK
13	[∧KR][KR].$	−XKX	−KKX
14	[∧KR][KR][∧KR].$	−XKXX	−KKXX, −XKKX, −KKKX
15	[DE][KR][KR][KR]$	−EKKK	
16	[DE][KR][∧KR][KR][KR]$	−EKXKK	−EKKKK
17	[∧KR][KR][DE][KR][KR]$	−XKEKK	−KKEKK

[a] A regular expression is "a string that describes a whole set of strings, according to certain syntax rules".[11] All letters in each set of brackets are considered to be only one symbol, which can be any listed amino acid. "∧" located at the first position indicates that the set of the string starts with the indicated pattern in brackets. At other positions, "∧" indicates that the listed pattern within brackets is excluded. "$" indicates the end of the pattern. "." indicates that anything can be located at that position.

tendency of Pro at $j + 2$ position to produce a missed cleavage appears to be a real effect, but small compared to the effect of acidic residues; therefore, the influence of $j \pm 2$ Pro residues was not considered further. Overall, the presence of proximal acidic residues represented the most prevalent mechanism for internal missed cleavages at Lys and Arg residues.

To summarize, the following rules were found to define the most common missed cleavage patterns, where every missed cleavage must satisfy the first rule and at least one other:

(a) The peptide contains two or fewer missed cleavages.

(b) Lys or Arg is located within the first three positions in a peptide, for example, Kx_n, xKx_n, or $xxKx_n$, where x indicates any amino acid and x_n indicates n amino acids.

(c) Lys or Arg is located within the last four residues in a peptide, and the last amino acid is Lys or Arg, for example, x_n-KxxK, x_nKxK, or x_nKK.

(d) Lys or Arg is located within the last three residues in a peptide, where the last residue is neither Lys nor Arg (e.g., peptides with missed cleavages that contain the C-terminal sequences of proteins), for example, x_nKx or x_nKxx.

(e) An internal missed cleavage at Lys or Arg has Asp or Glu immediately adjacent ($j \pm 1$) to Lys or Arg, for example, x_nEKx_n, x_nRDx_n.

(f) Two Asp/Glu residues are located at positions $j - 2$ and $j + 2$ to Lys or Arg, for example, x_nDxKxDx_n.

(g) Two Asp/Glu residues are located at positions $j + 2$ and $j + 3$ to Lys or Arg, for example, x_nKxDDx_n.

(h) EE, ED, DE, or DD residues are located at positions $j - 2$ and $j - 3$ to Lys or Arg, for example, x_nDDxKx_n.

Restriction of the PC-DB by the MC Product Rules. These MC rules were used to filter peptide sequences in the FULL PC-DB, to restrict the size of the PC-DB. A "missed cleavage PC-DB" (MC PC-DB) was generated from the FULL PC-DB by applying the rules described above. A set of regular expressions was used to represent the rules so that they could be manipulated programmatically. Regular expressions were necessary, because SQL queries did not allow selection of sequences containing two missed cleavage sites that both passed the rules. For example, the sequence KSPRLLCIEK contains two missed cleavages, the first at position N (Lys) and the second at position N + 3 (Arg). Because the missed cleavage at position N + 3 matches none of the rules, the sequence is unlikely and therefore should be excluded, even though the missed cleavage at position N matches rule b. The regular expressions representing the missed cleavages specified by rules a−h are presented in Table 3.

After removing the peptides with improbable missed cleavage patterns, the MC PC-DB size (1.5×10^6 peptides) was significantly reduced compared to the FULL PC-DB (2.8×10^6 peptides) (Table 1). Searching the test data set against the MC PC-DB provided high confidence identification of 833 DTA files (Table 4A). Of these, 826 overlapped with the 829 peptides identified in a parallel search against the FULL PC-DB (Table 4A,B). Thus, only three peptides were lost during the MC PC-DB search, while seven new identifications were added. Of the three peptides lost, two were rejected by manual analysis and thus were false positive (FP). The third was a peptide with one unusual missed cleavage site (MPSLPSYKVGDKIATR), which we believe instead represents a QV isoform that was present in the NCBInr database but absent from the IPI protein database and hence excluded from consideration. Of the seven peptides added, four were cases where the RSP score was promoted to 1, enabling them to pass the MSPlus filters (all were validated manually). Of the remaining three cases, one was too weak to enable definitive manual analysis, one was a clear false positive (later captured as a leucine aminopeptidase product), and one represented a second MS/MS of the (M)-VNHFIEFK sequence that had a Met removed from the first position of a normal tryptic peptide. Overall, two false positives were lost and one was gained; at the same time, five clearly correct assignments were gained, and one was ambiguous due to weakness of the spectra. Although this was a moderate improvement toward reducing distraction, it was consistent with the relatively small (2-fold) reduction in database size using the MC PC-DB compared to the FULL PC-DB (Table 1). More importantly, these results supported the validity of using an editing strategy

Table 4

A. Results of Searches against Protein and PC-DBs[a]

database	no. of MS/MS spectra searched	no. of high-confidence peptides
protein IPI-DB	2117	826 [b]
FULL PC-DB	2117	829
MC PC-DB	2117	833
MC-SCX PC-DB	2117	841
SCX PC-DB 1	31	16
SCX PC-DB 2	657	224
SCX PC-DB 3	245	75
SCX PC-DB 4	526	255
SCX PC-DB 5	657	271
protein IPI-DB NOE	2117	802
MC-SCX LAP PC-DB	2117	892

B. Searching against PC-DBs Improves Accuracy of Peptide Assignments[c]

database A	database B	peptides overlapping	peptides unique to A	peptides unique to B
protein IPI-DB	FULL PC-DB	826	0	3
FULL PC-DB	MC PC-DB	826	3	7
MC PC-DB	MC-SCX PC-DB	826	7	15
protein IPI-DB	MC-SCX PC-DB	816	10[d]	25[d]
MC-SCX PC-DB	MC-SCX LAP PC-DB	836	5	56
MC-SCX LAP PC-DB	protein IPI-DB NOE	682	210	120
protein IPI-DB	protein IPI-DB NOE	662	164	140

C. Specified LAP Cleavages Increase Data Capture while Maintaining Low Distraction[e]

searches	total peptides	LAP peptides	nontryptic peptides
no enzyme specificed IPI-DB	260	14	72
protein IPI-DB search	220	0	0
Full PC-DB search	222	0	0
MC PC-DB search	223	0	0
MC-SCX PC-DB search	232	0	0
MC-SCX, LAP PC-DB search	247	18	0
MC-SCX LAP PC-DB search (result to score, parameter set to 10)	250	18	0
MC-SCX LAP search (low SumScore, result to score, parameter set to 10)	290	23	0

[a] Searches of the test data set were carried out as indicated in Figure 1A (for the IPI-DB) and Figure 1B (for all PC-DBs). MSPlus was applied to evaluate high-confidence assignments based on consensus between Sequest and Mascot, RSP = 1, requirement of up to two missed cleavages, and SCX rules. Missed cleavage and SCX rules in MSPlus were turned off when they were used in the PC-DB searches. Results with MC SCX PC-DB represent the sum of results for searches with all five SCX PC-DBs, linked invidually. [b] The number of peptide identified in the IPI-DB search (826) was lower than reported previously (832)[4] because the intensity threshold for removing week spectra was slightly more stringent. [c] Pairwise comparisons of search results against protein and PC-DBs, indicating peptide assignments to the test data set that were accepted as high-confidence by MSPlus. [d] Manual analysis showed that 7 of 10 assignments unique to the search against the IPI-DB were correct. On the other hand, 24 of 35 assignments unique to the search against the PC-DB were correct. [e] All searches used only DTA files from SCX fractions 6−11, which is in the range of SCX rule 2. The "results to store" parameter for Sequest search was set to 500 except where stated. Because the false positive threshold is lowered by using PC-DB, we decreased the threshold for SumScore. With more filters applied, there are more total hits, except for no enzyme search. The reason is that no enzyme search yields more false positives.

to generate restricted databases according to trypsin specificity rules deduced from data mining.

Restriction of the Peptide Database by SCX Fractionation Rules. To further reduce the size of the PC-DB, we developed prefiltering rules based on the SCX chromatographic properties of the peptides, a filter that we had previously shown reduces the false discovery rate by ~50%.[4] During SCX chromatography at pH 2.5, peptide retention is correlated with the number of basic residues ("BR" = Lys, Arg, or His). To identify the appropriate rules, the high-confidence sequence assignments were analyzed to determine empirical relationships between BR and SCX elution behavior; these are summarized for the test data set in Table 6, with the full analysis in Supporting Information, Table 3. For example, SCX rule 1 allows 0 or 1 BR for peptides eluting in the earliest SCX fraction, rule 2 allows 1 BR for peptides eluting in the next six fractions, rule 3 allows 1 or 2 BR for peptides eluting in the next two fractions, etc. By applying these rules, the larger MC PC-DB was partitioned into five subsets ("MC SCX" 1−5), each including only those peptides that satisfied SCX rules 1, 2, 3, 4, or 5, and variously containing 27−88% fewer peptides than the MC PC-DB (Table 1).

To utilize these PC-DBs in a search, the computational work flow was modified (Figure 1B) to add the following: (1) Partitioner, a program that produces each MC-SCX PC-DB, and (2) Sorter, a preprocessing program that sorts DTA files into subsets that match the different SCX rules and are therefore searched against different MC-SCX PC-DBs (Figure 1B). Each subset of DTA files was then matched against the appropriate MC-SCX PC-

Table 5. Summary of Changes in Protein Profiles for the Peptide Assignments from Searches Shown in Table 4A

no. of support peptides	number of proteins: protein IPI-DB	FULL PC-DB	MC PC-DB	MC-SCX PC-DB	MC-SCX LAP PC-DB	protein IPI-DB NOE
all cases	294	295	295	295	303	279
two or more	128	128	130	135	142	138[c]
one only	166[a]	167	165	160	161[b]	141[d]

[a] FDR estimated at 2.6% for the IPI-DB and the PC-DBs from search against inverted DB and by manual analysis and at 6% for the NOE by manual analysis. [b] Seven have associated bridge peptides with another protein that is supported by more than one peptide. [c] Seventeen of these cases have complex changes with both gains and losses of unique peptides, 6 have losses, and 16 show gains. [d] Four have associated bridge peptides with another protein that is supported by more than one peptide.

Table 6. SCX Rules[a]

SCX fraction no.[b]	no. of BRs	SCX rule no.	SCX fraction no.[b]	no. of BRs	SCX rule no.
005	0 or 1	1	014	2	4
006	1	2	015	2	4
007	1	2	016	2	4
008	1	2	017	2	4
009	1	2	018	2 or 3	5
010	1	2	019	2 or 3	5
011	1	2	020	2 or 3	5
012	1 or 2	3	021	2 or 3	5
013	1 or 2	3			

[a] High-confidence assignments of the test data set generated by MSPlus were utilized to generate SCX rules. Details of the data analysis are shown in Supporting Information, Table 2. All assignments were partitioned by SCX fraction number. For peptides in each SCX fraction, assignments are grouped by the number of basic residues (BRs) in the assigned peptides and their frequencies tabulated. If a given partition contained fewer than four peptide assignments or the frequency of BRs is less than 5%, then the partition was rejected. Otherwise, the partition was accepted. The accepted partitions provide rules for numbers of basic residues in each SCX fraction. Groups specify five distinct SCX rules. [b] SCX fx 5 is the earliest fraction with detectable O.D. due to elution of peptides.

DB. In addition, SCX filtering rules in MSPlus were turned off. Partitioner and Sorter render the new work flow self-adjusting, because searches are tailored to data sets based on the information in each file.

DTA files were sorted into subsets that matched different SCX rules, and each subset was searched against the single MC-SCX PC-DB corresponding to each rule. Using these criteria, we found that 841 DTA files were accepted by MSPlus (Table 4A). Of the validated sequences, 816 overlapped with assignments made in IPI-DB searches, with 25 new assignments gained and 10 assignments removed (Table 4B, following the Venn diagram in Supporting Information, Figure 1). Manual analysis found 22/25 (88%) of cases unique to MC-SCX PC-DB searches to be correct. Of the three incorrect assignments, one MS/MS was later correctly identified as a leucine aminopeptidase cleavage product, and the other two would have failed a hydrophobicity filter, which eliminates peptides with hydrophobicities inconsistent with reversed-phase elution time. Of the 10 cases unique to the IPI-DB search, 7 were found correct by manual analysis and 3 were FP. Six of the validated peptides failed the SCX rules but were recaptured by MSPlus with the IPI-DB search because of their high scores. These six examples were represented by three peptide sequences, one that contained two Trp residues and eluted late from SCX, one that was very acidic and eluted early from SCX (LAADEDDDDDEEDDDEDDDDDDFDDEEAEEKAPVK), and one that eluted early from SCX (VLSDSRPAMAPGSSHLGAPASTTTAADATPSGSLAR). The latter peptide likely has a stable secondary structure created by two salt bridges between Arg and Asp residues, shielding the basic residues during SCX. This hypothesis is supported by the small number of MS/MS fragment ions generated from this triply charged ion, which should otherwise have yielded good fragmentation as the number of protons exceeds the number of Arg residues. The seventh peptide failed the missed cleavage rules and corresponds to the unusual tryptic product mentioned above, which is probably misassigned as a K isoform of a Q-containing peptide.

In summary, these experiments showed that searching with the MC-SCX PC-DB added 22 true positive (TP) assignments and removed 7 TP, for a net gain of 15 TPs, compared to the IPI-DB (Table 4B, with detailed discussion of each case provided above). The seven TPs lost were those accepted by using a high score recapture feature in MSPlus; this recapture function could be implemented in the PC-DB approach by using a secondary search to capture these cases. The MC-SCX PC-DB also led to removal of three FP assignments and added three FP for a net change of zero. The 24 new assignments captured were initially misassigned in the IPI-DB search due to distraction but were accurately assigned and captured by restricting database size using the PC-DB search strategy. Assessment of the FN cases is more complicated, because a peptide assignment may be top scoring for one or two of the three scores used by MSPlus or top scoring with too low a ΔCN value. However, we can say that the gain of 24 new high-confidence assignments enabled identification of 74% (841/1134) of the estimated number of identifiable DTA files and that almost all the expected LAP products were identified using the PC-DB approach; these results were validated both by manual analysis and by comparison with theoretical MS/MS predicted from a kinetic model of the mobile proton hypothesis.[10] Interestingly, a disproportionate number of the new assignments were from triply charged peptide ions (see Supporting Information, Table 1), ~40% of which represent missed cleavage products. Furthermore, triply charged peptide ions are almost never observed in early SCX fractions, which are predominantly peptides with one basic residue. Thus, the triply charged cases will be most affected by removing the unlikely cases from consideration.

In translating these results to a protein profile, the additional peptide assignments resulted in a net addition of one new protein (236 to 237), an increase in the number of proteins identified by

two or more peptides (71 to 79), and a decrease in the number of protein identifications supported by only one peptide (165 to 158). The results also revealed a major effect on the RSP score, promoting RSP up to one by using the smaller PC-DB. This was seen with the 4 cases described above, as well as an additional 10 cases where RSP changed to 1 with the smaller database for assignments that were already captured by MSPlus based on their high Mowse scores. Overall, the changes improved confidence for eight protein assignments made based on one peptide.

This effect on RSP suggested that improved sensitivity might be obtained when filtering out false positive assignments using the ΔCN score, generated by Sequest. This score indicates the difference between the XCorr scores of the first and second assignments. The magnitude of ΔCN for many peptide assignments should increase with the smaller database, because if the second assignment is unlikely and eliminated from the PC-DB, then the new second assignment that replaces it will have a lower score, yielding a larger ΔCN. To test this, we applied the XCorr and ΔCN threshold values (see Methods) that were reported in a recent proteomics study of mammalian cell proteins.[11] Using these criteria to evaluate searches of the test data set by Sequest, 720 MS/MS sequence assignments were validated in searches against the IPI-DB after excluding cases with observed mass of <950 Da. When using the MC-SCX PC-DB, the values of ΔCN for many peptides passing the XCorr thresholds increased noticeably, with the average (std.dev.) values of ΔCN increasing from 0.29 (0.12) to 0.37 (0.13), 0.46 (0.09) to 0.55 (0.08), and 0.45 (0.16) to 0.54 (0.15) for MH^{1+}, MH_2^{2+}, and MH_3^{3+} ions, respectively; an extreme case showed an increase in ΔCN from 0.009 to 0.651. However, only 18 additional assignments were validated after searching using the MC-SCX PC-DB filtered by XCorr and ΔCN. Thus, reducing the database size increased high-confidence assignments filtered by ΔCN by 2.5%, which was comparable to the 3% increase when filtered by MSPlus. These results clearly showed that using the smaller database significantly improved discrimination using ΔCN filtering, although it yielded only a small increase in assignments and thus did not significantly affect sensitivity.

Analysis of Score Distribution of Incorrect Assignments. Importantly, although MSPlus produced a superior overall recovery compared to the ΔCN method (841 to 738 validated cases), both approaches for capturing more data from low-scoring MS/MS recovered information on less than 8% of the 300 cases we had previously estimated should be identifiable.[4] To address this further, we examined the effect of reducing the database size on the score distribution of incorrect assignments. It is known that the score distribution of incorrect assignments shifts to lower values when using smaller databases;[2,4] a simple illustration of this effect can be seen by plotting the score distribution of all MS/MS files (Figure 3A,B). Inspection of Figure 3 shows a large peak consisting of low-confidence sequence assignments to MS/MS spectra, with a shoulder at high score values representing high-confidence assignments. The largest peak shifted to lower score values as the database size decreased with successive searches of IPI-DB, FULL, MC, and MC-SCX PC-DBs. As the database size decreases, the chances of obtaining an incorrect sequence assignment with high score also decreases; therefore, the distribution shifts toward lower scores. As a control, score distributions for false positive assignments were examined by searching the test data set against a database where peptide sequences were inverted ("inverted DBs"). These inverted DBs produced incorrect assignments for every MS/MS spectrum, while retaining amino acid compositions, peptide mass distributions, and Lys/Arg at the C-terminus that were identical with the normal database. As expected, significant reductions in XCorr and Mowse distributions were observed as the sizes of inverted databases became smaller (Figure 3C,D), but these shifts were relatively small, consistent with the 2−9-fold reduction in database size. This behavior was comparable for doubly charged and triply charged ions (Supporting Information, Figure 2).

The analysis of the inverted DBs suggested that lowering the minimum score threshold used by MSPlus might capture more information when searching against a PC-DB. This minimum score threshold is set by scaling the Mowse values to the XCorr values and then summing the two measurements to produce a SumScore.[4] In the four searches described above, SumScore was set to 3.5 based on the inverted IPI DB search. This threshold could be lowered to 2.85, in the light of the results with the inverted DB searches filtered by the MC-SCX rules. This captured an additional 7 TP assignments, increasing the yield with the MC-SCX PC-DB to 33 new assignments. On the other hand, five FP assignments were added, one later identified as a leucine aminopeptididase cleavage product (see below) and three that would have been removed using a hydrophobicity filter.[17] If a recapture function were implemented to capture the 7 additional correct MS/MS that failed the SCX and MC rules (see above), the number of new TP assignments would have increased to 40, still far below the 200−300 additional MS/MS we estimate should be identifiable.

Modified PC-DBs Enable Searching for Nontryptic Peptides without Increasing Distraction. We therefore turned our attention to another situation where distraction is an important issue in evaluating search results. A serious problem in working with species with large databases, such as human, arises when searching for peptides that are generated by nonspecific proteolysis instead of specific tryptic cleavages. By considering all possible cleavages, the effective database size increases by as much as 100-fold.[18] We have shown that this increases distraction from ~20% (allowing two missed tryptic cleavages) to ~40% (without specifying a protease), thus reducing the number of correctly assigned sequences and interfering with the sensitivity of data capture.[4]

The use of a PC-DB in this situation can be helpful for addressing this problem, assuming that nonspecific cleavages are not random and that specificity determinants can be identified in a manner similar to that used for delineating the missed cleavage rules. The first task was to identify a set of rules to identify the most probable nonspecific cleavages. The test data set was searched against the IPI-DB, using "no enzyme specified" options in Sequest and Mascot. This yielded a total of 802 peptide assignments identified by MSPlus, of which 662 were tryptic peptides seen with the tryptic search. Thus, distraction reduced the number of correct tryptic assignments by 148 [(826 − 662)/

(17) Petritis, K.; Kangas, L. J.; Ferguson, P. L.; Anderson, G. A.; Pasa-Tolic, L.; Lipton, M. S.; Auberry, K. J.; Strittmatter, E. F.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2003**, *75* (5), 1039−1048.

(18) Olsen, J. V.; Ong, S. E.; Mann, M. *Mol. Cell. Proteomics* **2004**, *3* (6), 608−614.

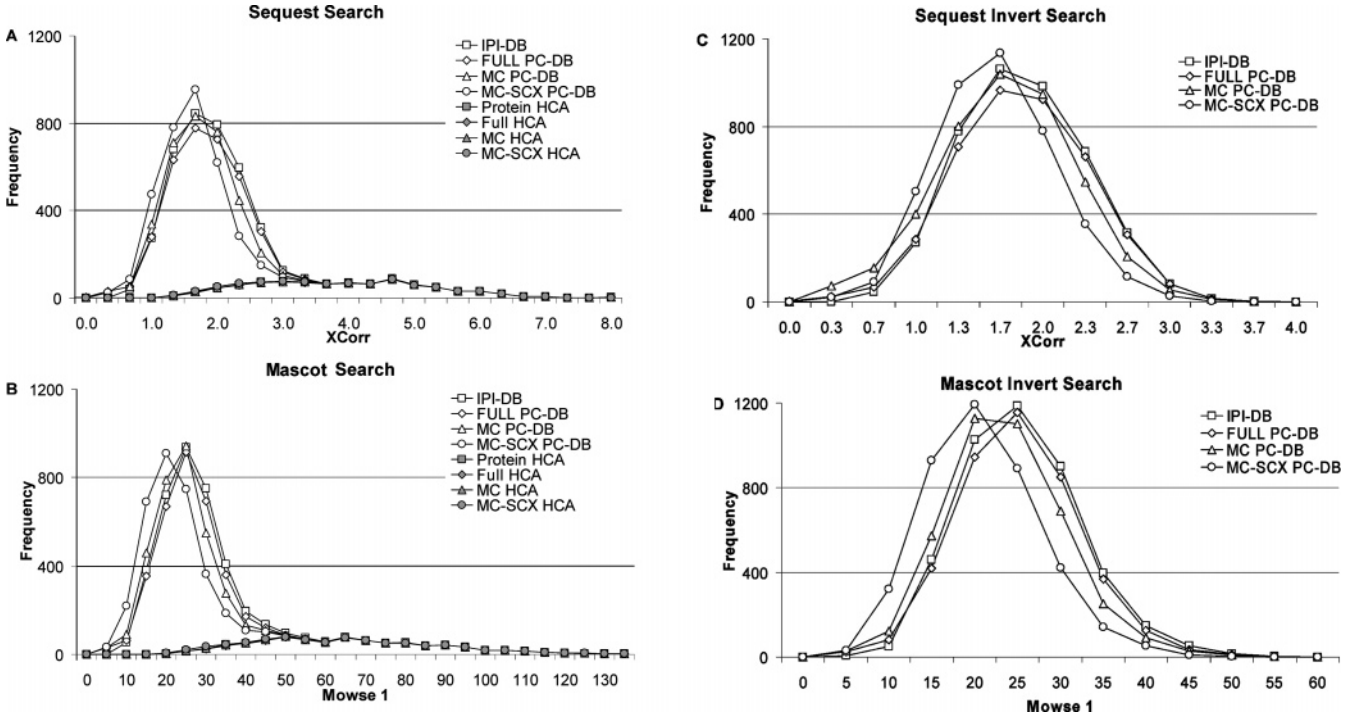

Figure 3. Reducing database size to improve discrimination between correct vs false positive assignments. (A,B) Score frequencies were evaluated using (A) Sequest or (B) Mascot to search MS/MS data sets against protein IPI-DB (squares), FULL PC-DB (diamonds), MC PC-DB (triangles), or five MC SCX PC-DBs (circles). Open symbols represent searches of all DTA files in the test data set (4245 DTAs). Closed symbols represent searches of MSPlus validated assignments. The results show that distributions of high-confidence assignments are invariant with database size, whereas distributions of low-confidence assignments decrease significantly with restriction of PC-DB size. (C,D) Score frequencies were evaluated using (C) Sequest or (D) Mascot to search the test data set against each protein or peptide database in which sequences were inverted, producing false positive assignments. Significant shifts of false positives to lower score were observed by restricting database sizes using MC and SCX rules. Many of the multiply charged MS/MS spectra are represented in the data set twice, once with a correct charge assignment and again with an incorrect charge assignment. In addition, a large portion of MS/MS represent nontryptic or covalently modified peptides, ions with misassigned charge where the MS/MS summary text file was incorrectly made, or MS/MS where two peptides of similar *m*/*z* were simultaneously fragmented. These produce a large peak of incorrect assignments with lower scores (i.e., XCorr < 3 or Mowse < 45).

826 = 20% distraction]. Searching the IPI-DB with no enzyme specified yielded 140 new nontryptic peptide assignments; a detailed analysis was carried out on this subset, to identify consensus sequence patterns that reflect known proteolysis or in-source fragmentation chemistries. Of the 140 nontryptic assignments, 21% were generated by removal of a single Leu, Ile, or occasionally Met residue from the N-terminus of a tryptic peptide. Such a pattern is consistent with leucine aminopeptidase (LAP) which shows N-terminal exopeptidase activity toward these amino acids.[19] The other two major groups were chymotryptic-type cleavages at hydrophobic residues that could not be accounted for as products of LAP (37%) and other cases (42%) that included cleavage N-terminal of proline, C-terminal of histidine, and adjacent to acidic residues, and were likely due to fragmentation in the MS source.

We therefore modified the MC PC-DB to add all possible LAP products and searched the test data set, applying the SCX rules in MSPlus. This process added 56 additional peptides, all of which were verified by manual analysis. Importantly, only five fully tryptic peptides were lost due to distraction when using the MC-SCX LAP PC-DB strategy. Thus, the PC-DB accommodates probable nonspecific cleavage products while retaining correct tryptic peptide identifications and maintaining low database size, yielding a significant lower distraction rate compared to protein DB searches. These results showed that the expected large decrease in distraction could be observed when specifying LAP products in the MC-SCX PC-DB compared to searches of the IPI-DB with no enzyme specified, due to the reduction in database size of ~100-fold. In a previous study,[4] we showed that this fold reduction in the database size produced a large shift in the score distribution of the incorrect assignments, consistent with this model.

Protein Profile Analysis. An in-house protein profiling program (IsoformResolver[4]) was used to generate the protein profiles for the six sets of peptide assignments in Table 4A (Full protein output shown in Supporting Information, Table 4). Table 5 shows the distribution of cases that have one unique peptide supporting the protein identification versus those that have several peptides supporting the identification, as recommended by Carr et al.[20] Protein specified by tryptic peptides from FULL, MC, and MC-SCX PC-DB searched were nearly identical compared to the IPI-DB search, showing only three protein identifications supported by one peptide gained and two lost. One the other hand, the PC-DB searched led to a significant gain in the number of proteins supported by two or more peptides and a small loss in the number supported by only one peptide. The protein profile

(19) Taylor, A. *FASEB J.* **1993**, *7* (2), 290−298.

(20) Carr, S.; Aebersold, R.; Baldwin, M.; Clauser, K.; Nesvizhskii, A. *Mol. Cell. Proteomics* **2004**, *3* (6), 531−533.

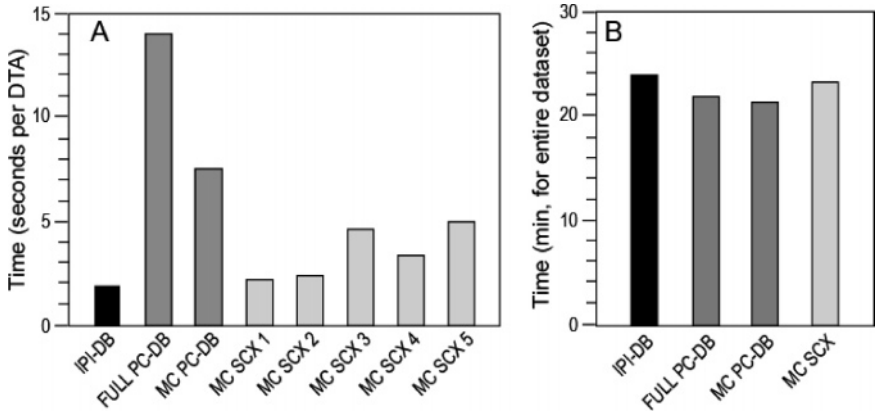

Figure 4. Running times for Sequest and Mascot searching against different databases. The running time of Sequest (A) was sensitive to the number of items instead of the number of peptides, while Mascot (B) kept the same performance for the given databases. Run times were measured for the test data set containing 4245 DTA files; Mascot reports running time for the MGF while times for Sequest were taken from each OUT file and averaged. Different computers were used for Sequest vs Mascot so a direct comparison of run time cannot be assessed.

generated from the "no enzyme specified" search showed small changes compared to the IPI-DB search, although 69 unique peptide identifications were gained and 93 lost, resulting in complex changes in the protein profile. Seventeen proteins were observed with gains and losses in peptide numbers above two peptide/proteins while several proteins were lost due to distraction, leading to a smaller number of proteins identified. Because protein identifications are more reliable when two or more peptides support them,[20,21] the strategy of reducing database size using peptide-centric databases led to increased confidence in search results.

Effect of Using PC-DBs on Run Times. Finally, we evaluated run times for searching against the PC-DBs versus IPI-DB. Comparison of the FULL PC-DB with the IPI-DB using Sequest showed an unexpected 7-fold increase in run time (Figure 4A). Examination of the index files generated from these databases revealed that the index file was 86.9 MB for the FULL PC-DB and 1.44 MB for the IPI-DB (Table 1). We attribute this increase in file size to the fact that the PC-DB has more "protein" entries, because each peptide is represented by a separate FASTA entry. Indeed, the run times of the individual PC-DBs were proportional to the number of peptide entries. On the other hand, Mascot run times were not very sensitive to the number of entries (Figure 4B). We hypothesized that the longer Sequest search times reflected how the program handles a larger number of entries in the database, rather than the actual searching process.

To test our hypothesis, we constructed a hypothetical "megaprotein" with sequence concatenated from all peptides in the FULL PC-DB, with the exception of C-terminal peptides. This database contained one large FASTA entry corresponding to the megaprotein, and 47 306 entries corresponding to the C-terminal sequences of each protein in the IPI database; thus, the megaprotein DB had almost the same number of protein entries as the IPI-DB. To force the program to index the megaprotein without cleaving peptides at missed cleavage sites, we replaced each internal Lys or Arg residue with hypothetical residues defined by Lys and Arg masses. We then specified cleavage at Lys and Arg when initiating the Sequest and Mascot searches, allowing for no missed cleavages. The search program accepted the megaprotein DB, generating an index file of 1.44 MB, which was the same size as the index file generated from the IPI-DB. Using the megaprotein DB in a Sequest search produced a run time of 1.85 s/DTA, comparable to that of the IPI-DB (1.9 s/DTA). Hence, we concluded that file management processes related to the protein information are rate limiting during searching, a problem that can be overcome using a megaprotein strategy when the source codes for indexing subroutines are not available.

DISCUSSION

In this study, we show that data mining of a large collection of peptides identified by MudPIT analyses can delineate rules defining the peptide chemical properties of those peptides, that these rules can be used to generate PC-DBs, and that using these PC-DBs with search programs Sequest and Mascot achieves lower distraction and improved confidence in results. Two approaches were evaluated using a highly annotated test data set that enabled us to readily distinguish correct versus incorrect assignments. The first approach mimicked a tryptic search allowing up to two missed cleavages, and the second provided a proof-of-concept study for a new method to identify nonspecific proteolysis in tryptic digests. The clearest effect on reducing distraction was found in the second study, where LAP products were added to the PC-DB for comparison with protein DB searches using the no enzyme specified strategy to identify nontryptic cleavage products, achieving a 100-fold difference in effective database size. Using the protein database, the distraction rate for nonspecific cleavages versus trypsin searches was almost 20% (Table 4B). However, by characterizing the type of nonspecific cleavages in our samples, we found that many of the nonspecific cleavage products were consistent with LAP activity, an abundant enzyme in mammalian cell extracts. By specifying these peptides in the PC-DB, the distraction rate was less than 1%, but 7% more peptides were identified (the LAP products). Thus, as hypothesized, the use of PC-DBs can greatly minimize distraction.

In the first study, trypsin missed cleavage rules were combined with rules governing SCX elution to generate peptide databases that were 3−9-fold smaller than the parent protein database. As expected, the PC-DBs improved the discrimination between correct and incorrect assignments when used with the search programs Sequest or Mascot, as revealed by a lower Mowse or

(21) Venable, J. D.; Yates, J. R. *Anal. Chem.* **2004**, *76* (10), 2928−2937.

XCorr score distribution of incorrect assignments (Figure 3C,D), higher ΔCN, and lower RSP scores in Sequest. Importantly, the FDR was maintained at a low level, even though smaller databases often increase FDR; this was in large part due to the removal of major classes of unlikely sequences from the input PC-DB. However, the PC-DB yielded only a 3% increase in true positive MS/MS assignments, where we estimated ~20% of the cases were subject to distraction. Because previous studies had shown a large decrease in distraction when the database was decreased by 80-fold, we believe the minimal effect seen here is because of the much smaller decrease achieved using chemical properties. This is consistent with the small shift to smaller range of scores for incorrect assignments shown in Figure 3 for this study, compared with that observed with the 80-fold smaller database in our previous study.[4]

In addition, the trypsin study targeted cases that were lower scoring, whereas the newly revealed LAP peptides mainly represented high-scoring assignments. Thus, another factor contributing to the differential effect on distraction in these two cases may be that low-scoring MS/MS spectra have properties that render them more difficult to identify and successfully score using the currently available search programs. In the previous study, we noted that distraction occurred more frequently with MS/MS spectra which contained fragment ions that were not considered by the search programs, such as multiple dehydrations, internal fragment ions, or multiply charged fragment ions.[4] We have also noted a high frequency (~30%) of MS/MS resulting from simultaneous fragmentation of two different peptides with similar m/z values (unpublished studies). Such complications likely explain the difference between analyses of low-scoring and high-scoring spectra and suggest that new scoring methods probably will be necessary to further reduce the distraction rate for MS/MS spectra in a normal tryptic search.

Increased confidence in the peptide identifications was also reflected in the protein profiles. Most of the new peptide assignments corresponded to proteins already identified from protein database searches and included both new unique peptide sequences and resampling of peptides already identified. Using the PC-DB strategy versus the conventional strategy, the average number of peptides supporting each protein identification was higher and the number of protein identifications supported by only one peptide was reduced. This result supports the argument that few false positives were created, because false positives would be expected to contribute primarily to cases where an identified protein is supported by only one peptide. It seems reasonable to exclude protein identifications based on unlikely missed cleavage or nontryptic products or peptides with aberrant chromatographic behavior. When these properties are used as a prefilter rather than as a postfilter to validate assignments, more peptides can be identified and overall confidence in the results increases, particularly for triply charged peptides and when using ΔCN criteria.

The LAP analysis addresses a major area of controversy in this field: whether to search without specifying the protease in order to capture nonspecific cleavage products or whether to specify trypsin cleavages in order to minimize incorrect assignments. Many investigators feel that the huge increase in the effective database size with nonspecific cleavage searches causes more information loss due to distraction than gain of new information. Here we show that a specific set of peptide rules can be defined to capture most of the new information; by editing the peptide list considered by the search program to include the peptides defined by these rules, distraction is minimized and in fact the yield of both tryptic and nontryptic peptides is increased significantly. Furthermore, the data mining results used to define these rules has the potential of revealing new insight into what proteolytic activities are present in the preparation. For example, the LAP peptides might account for the impression among many investigators that chymotryptic cleavages are common, because LAP products resemble chymotrypsin products when the cleavage site is reported by only the residue preceding the observed N-terminus. Further data mining studies on larger databases should reveal other classes of peptides, protease activities, and recovery biases in the MudPIT protocol. In applying this approach in other situations, it is important to confirm the specific rules, because it is likely that the specific rules for protease activity will change, depending on the sample, digestion conditions, analytical protocol, and enzyme source (in particular, modified trypsin may show differences from the unmodified trypsin used in these studies).

Although tests of the PC-DB approach showed no improvement in computational efficiency, our experiments show that the lengthy search time using PC-DBs was due to management of protein information in the indexed database, rather than the searching process. Because we have found that the available search programs do not provide accurate protein profiles, we find it more useful to reconstruct the protein assignments postsearch with our IsoformResolver program,[4] providing better control over the isoform information, greater ease in updating the protein database, and more accurate comparative analyses where isoforms are present, which solves a problem raised in a previous study regarding a peptide-centric database.[8] Thus, there is no reason to track the protein information during the search itself, and in fact, our results show that managing this information consumes a significant part of the computational expense of Sequest and Mascot, although to a lesser extent with Mascot. The strategy of using a megaprotein to encompass all the peptides except the C-termini provides an acceptable strategy to circumvent this feature of the search programs, but it would be better if search programs enabled searching with peptide lists without protein information included.

This study highlights the need to consider the relative value of using a peptide property to minimize false positives as a postsearch filter or for decreasing the size of the database prior to searching. Our study shows that using the SCX and MC rules to reduce the database size effectively reduces distraction and enables acceptance thresholds to be lowered, resulting in increased confidence in assignments. We used manual analysis to validate new assignments in this study; in a high-throughput work flow, other properties such as hydrophobicity,[16,22] scoring against improved theoretical spectra,[10] or implementing heuristic rules used in manual analysis can be used to minimize false positives by programmatic means. However, the results in this study show the potential importance of using prefiltering, and

(22) Palmblad, M.; Ramstrom, M.; Bailey, C. G.; McCutchen-Maloney, S. L.; Bergquist, J.; Zeller, L. C. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2004**, *803* (1), 131−135.

the use of a property as a prefilter versus postfilter must be evaluated.

ACKNOWLEDGMENT

We are indebted to Richard Johnson and Alex Taylor from Amgen, Seattle, for illuminating conversations in early the stages of this project, and to advice and support from William Old and Kevin Pierce, who helped assemble data for the analysis of the missed cleavage products and assisted in the analysis of the protein profiles. 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 June 23, 2005. Accepted December 9, 2005.

AC051127F