Journal of Biomolecular Techniques
15:285–295 © 2004 ABRF

Defining Parameters for Homology-Tolerant Database Searching

J. P. Kayser,[a] J. L Vallet,[a] and R. L. Cerny[b]

[a]USDA, ARS, Roman L. Hruska US Meat Animal Research Center, Clay Center, NE; [b]Nebraska Center for Mass Spectrometry, Department of Chemistry, University of Nebraska, Lincoln, NE

De novo interpretation of tandem mass spectrometry (MS/MS) spectra provides sequences for searching protein databases when limited sequence information is present in the database. Our objective was to define a strategy for this type of homology-tolerant database search. Homology searches, using MS-Homology software, were conducted with 20, 10, or 5 of the most abundant peptides from 9 proteins, based either on precursor trigger intensity or on total ion current, and allowing for 50%, 30%, or 10% mismatch in the search. Protein scores were corrected by subtracting a threshold score that was calculated from random peptides. The highest ($p < .01$) corrected protein scores (i.e., above the threshold) were obtained by submitting 20 peptides and allowing 30% mismatch. Using these criteria, protein identification based on ion mass searching using MS/MS data (i.e., Mascot) was compared with that obtained using homology search. The highest-ranking protein was the same using Mascot, homology search using the 20 most intense peptides, or homology search using all peptides, for 63.4% of 112 spots from two-dimensional polyacrylamide gel electrophoresis gels. For these proteins, the percent coverage was greatest using Mascot compared with the use of all or just the 20 most intense peptides in a homology search (25.1%, 18.3%, and 10.6%, respectively). Finally, 35% of *de novo* sequences completely matched the corresponding known amino acid sequence of the matching peptide. This percentage increased when the search was limited to the 20 most intense peptides (44.0%). After identifying the protein using MS-Homology, a peptide mass search may increase the percent coverage of the protein identified.

Key Words: homology search, mass spectrometry, bioinformatics

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Address correspondence and reprint requests to: J. L. Vallet, USDA, ARS, RLH US Meat Animal Research Center, P. O. Box 166, Clay Center, NE 68933 (phone: 402-762-4187; fax: 402-762-4382; email: vallet@email.marc.usda.gov).

Proteomics focuses on studying the gene products, proteins, which are the active agent in the cell.[1] High-resolution two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is commonly used to separate proteins from a complex biological mixture. Analysis by 2D-PAGE provides information about the physical characteristics of a protein that affect migration within the gel, including mass and isoelectric point. The changes in protein expression due to changes in physiological conditions can be quantified using 2D-PAGE and subsequent analysis of the staining intensity of each protein spot. However, little information regarding the molecular function of the protein is obtained by 2D-PAGE alone. Since the completion of the human and mouse genome sequence projects, protein identification, which may provide clues as to function, can be undertaken by comparing the masses of the peptides of an unknown protein with those predicted from proteins derived from genomic sequence database.[1,2] Mass spectrometry provides these highly accurate peptide mass measurements and is therefore a powerful tool in identifying biologically active proteins.

TANDEM MASS SPECTROMETRY

Of the spectrometric techniques currently available, tandem mass spectrometry (MS/MS) provides the most peptide sequence information.[3] One of the difficulties in interpreting the results from this process is distinguishing between peptides that originated from the actual spot chosen and those that resulted from background contamination of the gel (e.g., protein streaking within a gel).

One solution to this problem is to select only the fragmentation information from the most abundant peptides in the chromatographic run used in the database searches. Using our quadrapole/time-of-flight mass spectrometer (Q-TOF-Ultima API, Waters/Micromass, Manchester UK), peptides resulting from a given spot are introduced into the mass spectrometer as eluent from some form of chromatography. As the peptides are introduced, precursor ions are selected during an initial survey scan if their intensity is above a predefined threshold, and this trigger intensity is recorded. The mass spectrometer then collects fragmentation spectra (collision-induced fragmentation) for each precursor ion for a predetermined period that is partially dependent on the total ion current in the fragmentation spectrum. At the end of this process, the mass spectrometer returns to the survey scan mode and additional precursor ions are selected. This process is repeated until the chromatographic separation is completed.

TRIGGER INTENSITY VERSUS TOTAL ION CURRENT

In MS/MS experiments, two different types of information are available from which to select the most abundant peptides—the trigger intensity for each precursor peptide and the total ion current for the fragments resulting from each precursor peptide. Neither the trigger intensity nor the total fragment ion current is directly reflective of the most abundant peptides present in the spot. The trigger intensity can be recorded at essentially any point in the release of that particular peptide from the chromatographic column. The total ion current is also not entirely reflective of the abundance of each peptide because collection of this information is somewhat dependent on the length of time an ion is analyzed (e.g., low-abundance peptides are examined longer by the mass spectrometer than are high-abundance peptides). Although peptide abundance cannot be accurately quantified by either method, it is more likely that higher trigger intensities and/or total fragment ion currents would be found among abundant peptides than among low abundance, potentially contaminating, peptides. To our knowledge, no one has reported the effects on subsequent database searching of peptides selected on the basis of trigger intensity or total fragment ion current.

MASS-BASED SEARCHING

High-throughput database search algorithms using peptide mass information have been developed.[4,5] These algorithms compare an experimentally obtained spectrum with those predicted for all possible peptides that are obtained from a sequence database and have a mass within a defined error tolerance. The correctness of fit between the database peptide and the spectrum is calculated by various scoring mechanisms.[5,6] Protein identification based solely on mass mapping can be achieved only if patterns of tryptic digestion and amino acid residues are generally conserved; therefore, only nearly identical homologs are likely to be found.[7,8] Thus, comparison of observed masses obtained by MS/MS with predicted masses from sequence databases does not work well for species with limited sequence information. Although it is still possible to identify proteins that possess homology in the 70–100% range, the failure rate for searches using peptide mass fingerprinting alone increases dramatically when protein similarity is less than 90%.[8,9] More accurate matching of candidate peptides to database sequences with weaker homology could be achieved by using a *de novo* interpretation of tandem mass spectrometry spectra followed by a direct homology search.[9–11] Because our interest is in the identification of proteins from livestock species, whose sequence information is incomplete, we set out to define parameters for homology searching.

HOMOLOGY-TOLERANT SEARCHING

Homology-tolerant searching first requires the generation of peptide sequence information from MS/MS results. The presence of the sequence data should increase the speed and scope of a database search, but the overall throughput is severely constrained by the interpretation step.[3,5] However, compensation for the time required to *de novo* sequence a peptide may be gained by an increase in identification of proteins from divergent species by combining peptide mass with partial sequence information[12] or using automated algorithms to derive sequence information.[10,13,14] In this series of experiments, sequence information was obtained by employing the automated *de novo* sequencing program PEAKS (Studio 2.0, Bioinformatics Solutions, Ontario, Canada). Preliminary work in our lab showed that the PEAKS algorithm provided rapid interpretation of MS/MS spectra that was more accurate than our manual interpretations of data (data not reported). In addition, PEAKS has been shown to be comparable to other *de novo* sequencing software programs using data obtained from quadrapole time-of-flight instruments.[14,15]

MS Homology

The use of *de novo* sequence information may improve the number of proteins identified from biological fluid through a homology-tolerant search of related species. One such program capable of this type of search, MS-Homology, is available through the Protein Prospector suite of programs (http://prospector.ucsf.edu). This program has a number of advantages for the analysis of sequence information resulting from MS/MS experiments. It allows one to incorporate the ambiguities routinely encountered in these experiments (e.g., the inability to distinguish between Ile and Leu). It also allows one to submit any number of peptides obtained from an MS/MS experiment, and to choose the level of amino acid homology to be used in the search. However, the effect of both the number of peptides submitted and the homology tolerated on the output (protein score) from this program is relatively uncharacterized. The most important unknown is that a significant protein score has not been established. Exploration of these issues is essential to strengthen confidence in the results obtained using this program for database searching.

Thus, these experiments were designed to develop a strategy for homology-tolerant database searching. Our objectives were to

1. examine whether selection of peptides based on precursor intensity or total ion current affects the ability to identify proteins,
2. examine how the number of peptides used in a database search affects protein identification,
3. investigate how the level of allowed amino acid mismatch affects the ability to identify proteins,
4. define what constitutes a significant match, and
5. compare homology-tolerant database searching with that obtained using mass-based searching (Mascot).

MATERIALS AND METHODS

Protein Isolation and Analysis

Porcine intrauterine proteins were collected on day 13 of pregnancy by flushing each uterine horn with 20 mL of minimum essential medium (Sigma, St. Louis, MO). Uterine flushings were dialyzed against distilled water (3 changes) to remove salts and 0.5-mL aliquots of each flushing were then lyophilized. Proteins were solubilized in 5 mM K_2CO_3, 9.6 M urea, and 50 mM dithiothreitol (DTT) and subjected to 2D gel electrophoresis.[16] Gels were then stained with Coomassie blue R-250. Protein spots were excised from gels followed by in-gel digestion with trypsin according to published procedures with some modifications.[17] Briefly, gel pieces were placed in 1.5-mL siliconized microcentrifuge tubes (no. T5040G, Marsh Biomedical Products, Rochester NY) and covered with 0.5 mL 50% methanol/5% acetic acid destain. The destain was removed and the gel pieces dehydrated in 200 μL acetonitrile for 10 min. Following a second acetonitrile wash, gel pieces were dried in a vacuum dryer (Speed-Vac, SC 110, Savant Instruments, Holbrook, NY). The dried gel pieces were rehydrated in 100 mM NH_4HCO_3 and 10 mM DTT for 30 min. This solution was then replaced with approximately the same volume of 100 mM NH_4HCO_3 and 50 mM iodoacetamide. Iodoacetamide solution was removed and gel pieces were again dehydrated with acetonitrile as described above. Gel pieces were then rehydrated in 100 mM NH_4HCO_3 for 10 min followed by dehydration in acetonitrile. Gel pieces were rehydrated in 50 μL of 50 mM NH_4HCO_3 digestion buffer containing 20 ng/μL trypsin (no. V5111, Promega, Madison, WI). After 15 min, trypsin solution was removed and replaced with digestion buffer without trypsin and incubated overnight at 37°C. Resulting peptides were then extracted with two changes, 30 min each, of 5% formic acid in 50% acetonitrile. Samples were stored at −80°C until they could be concentrated to approximately 25 μL using a vacuum dryer. The extract solution (10 μL) was injected onto a trapping column (300 μ × 1 mm) in line with a 75 μ × 15-cm C18 reversed phase LC column (Dionex, Sunnyvale, CA). Peptides were eluted from the column using a water + 0.1% formic acid (A) / 95% acetonitrile : 5% water + 0.1% formic acid (B) gradient with a flow rate of 270 nL/min. The gradient was developed with the following time profile: 0 min 5% B, 5 min 5% B, 35 min 35% B, 40 min 45% B, 42 min 60% B, 45 min 90% B, 48 min 90% B, 50 min 5% B.

A Q-TOF Ultima tandem mass spectrometer (Waters) with electrospray ionization was used to analyze the eluting peptides. The system was user controlled employing MassLynx software (v3.5, Waters) in data-dependant acquisition mode with the following parameters: 1-sec survey scan (380–1900 Da) followed by up to three 2.4-sec MS/MS acquisitions (60–1900 Da). The instrument was operated at a mass resolution of 8000. The instrument was calibrated using the fragment ion masses of doubly protonated Glu-fibrinopeptide.

Data Search

Nine proteins (Table 1) that had been previously identified using Mascot (Matrix Science, v1.9.0, London, UK) analysis of MS/MS data were selected for comparisons of data analysis methods. The raw data obtained from each MS/MS experiment was processed using ProteinLynx software (v3.5, Waters) to generate a list of centroided masses of precursor peptide ions, after Savitzky Golay smoothing of spectrum peaks and a 40% adjustment of the baseline. Furthermore, this list contained charge and intensity information for each precursor ion. This list was sorted first on predicted charge of the precursor ion followed by precursor ion intensity and only doubly charged ions were used for further analysis. In addition, the intensities of the collision-induced product ions generated during MS/MS were summed together to obtain the total ion current for each precursor ion. A second ranking of precursor ions based on this summed total fragment ion current was generated. The processed MS/MS spectra were *de novo* interpreted using PEAKS, v2.0 software (Bioinformatics Solutions) for the 20 most intense doubly charged peptides based on both precursor trigger intensity or total fragment ion current.

Homology searches were performed using the *de novo* sequences to search the NCBInr database using MS-Homology (ProteinProspector, University of San Francisco, CA; http://prospector.ucsf.edu). MS-Homology allows the incorporation of ambiguities in the search and the following common isobaric substitutions were included into the lists of peptides:

1. each instance of Leu or Ile was modified such that either was possible at that position,
2. each instance of glutamine was modified such that either a glutamine or a lysine could be present, and
3. each instance of phenylalanine was modified such that either a phenylalanine or a (oxidized) methionine could be present.

Database searching was performed with the following parameters: protein molecular weight and pI ranges were unlimited, the database was limited to those sequences from mammalian species, and the enzyme chosen was trypsin. MS-Homology allows the user to define the level of amino acid mismatch that is acceptable. To determine the effect of mismatch on the protein score, database searching was performed allowing for 50%, 30% and 10% amino acid substitution. MS-Homology derives the overall matching protein scores by adding the scores for matching peptides calculated using the default scoring matrix, Blosum 62.[18] Multiple peptide sequences may be submitted and searched simultaneously using MS-Homology. Therefore, to determine the effect of the number of peptides submitted in a search, protein scores were calculated using the 20, 10, or 5 most intense peptides based on either precursor ion intensity or total ion current. This resulted in a 2 × 3 × 3 factorial experimental design.

MS-Homology protein scores were also generated using lists of random peptides to obtain an empirical determination of a significant score. Random peptides were created using the random function of Microsoft Excel. This function returns a random number from 0 to 1. Using Excel, these numbers were multiplied by 20, rounded to the nearest integer, and then each integer was assigned at random to an amino acid, with the exception of 0 and 20, which were both assigned to the same amino acid. This resulted in a list of random amino acids. From this list, peptides were sequentially constructed with random lengths between 8 and 15 amino acids—with the exception that all internal arginine or lysine residues were removed. In addition, each random peptide was terminated with either an arginine or a lysine residue.

TABLE 1

Selected Swine Proteins Used To Derive a Homology-Tolerant Database Search Strategy

Mascot-Identified Protein	Species	Accession No.
Serum albumin	*Sus scrofa*	gi 113578
Complement component C3	*Sus scrofa*	gi 11869931
Plasma retinol binding protein	*Sus scrofa*	gi 89271
Salivary lipocalin	*Sus scrofa*	gi 20178087
N-acylsphingosine amidohydrolase	*Homo sapien*	gi 30089928
Keratin	*Homo sapien*	gi 17318569
Alpha-2-macroglobulin	*Homo sapien*	gi 4557225
Saposin	*Bos taurus*	gi 13878928
Superoxide dismutase	*Oryctolagus cuniculus*	gi 1711431

This was done because, when using trypsin, the peptides generated above are unlikely to contain internal lysine or arginine residues and terminate almost exclusively in either a lysine or an arginine residue. Twenty lists of 5, 10, and 20 amino acid peptides were then constructed, and submitted to MS-Homology for searching the mammalian database using the following search options:

1. no enzyme selection, 50% mismatch allowed;
2. trypsin enzyme selection, 50% mismatch allowed;
3. trypsin selection, 30% mismatch allowed; and
4. trypsin selection, 10% mismatch allowed.

For each search, the highest protein score was recorded. Mean protein score and standard deviations were then calculated for each number of peptide by search option combination. These were then used to calculate the protein score required to be significantly different from a list of random peptides for each number of peptides by search option combination using the following formula: Significant protein score $\geqslant$ mean protein score $\pm$ 2 standard deviations.

Finally, a comparison between homology-tolerant database search (MS-Homology) and peptide mass search (Mascot) was performed. MS/MS data from an additional 112 spots isolated on 2D-PAGE gels were processed and precursor intensity peak lists were generated as previously described using ProteinLynx. All doubly charged peptides were *de novo* sequenced using PEAKS software. Only doubly charged ions were selected because the *de novo* sequence program performs better using b- and y-ion information. Singly charged peptides generate predominantly b ions only. In addition, since the PEAKS program does not recognize charge state, it cannot process multiply charged fragment ions resulting from peptides with three or more charges. Obvious trypsin peptides were deleted from the peptide lists. Homology-tolerant searches were performed using either the top 20 peptides based on precursor ion trigger intensity or all of the doubly charged peptides for each spot. A maximum of 30% amino acid substitution was allowed. A Mascot search of the NCBInr database was also conducted using ions obtain from MS/MS data for all doubly charged tryptic peptides using the following parameters: search was limited to mammalian species, one missed cleavage was allowed, and acrylamide modification of cysteine and oxidation of methionine were permitted as modifications. Protein identifications determined by homology search of all peptides or the 20 most intense precursor ions were then compared to protein identification obtained for database search using Mascot. The percent coverage of proteins was used to directly compare the results from each search for proteins that all three search methods similarly identified.

Eleven proteins, identified as of porcine origin, were used to investigate the accuracy of PEAKS to correctly assign amino acid residues to MS/MS spectra. Of these proteins, seven were identified as having multiple isoforms based on their migration on 2D-PAGE. In these cases, all peptides from the highest and lowest scoring isoforms were used. It was assumed that isoforms with the highest score had higher quality spectra and a more complete list of tryptic peptides than the lower scoring isoforms. Therefore, by selecting both the highest and lowest scoring samples, the performance of PEAKS could be analyzed over a range of MS/MS data. Data were used from four to six isoforms of two serum proteins, transferrin and albumin, as these proteins have been well characterized and provided the greatest range of protein scores. Four proteins were represented by only one spot; therefore, only one set of peptides was used for these proteins. In total, 2417 *de novo* peptide sequences from 25 spots were submitted for database searching.

Statistical Analysis

Factors affecting protein scores were analyzed as a $2 \times 3 \times 3$ factorial experiment using PROC MIXED (SAS, Cary, NC). The model included protein (random effect), intensity source (precursor ion intensity vs. total ion current), number of peptides used in search query (5, 10, or 20), and percent allowable amino acid substitutions (10, 30 or 50%). To examine the effect of background noise attributed to random peptide matching, data were reanalyzed after all protein scores were corrected by subtracting the significant protein scores derived from random peptides from the actual protein scores. These corrected protein scores were analyzed using the same model as the uncorrected protein scores. All main effects and two- and three-way interactions were considered significant at $p < .05$ unless otherwise noted.

For the data from the larger group of protein spots, ANOVA (PROC GLM; SAS) was performed to detect differences in the mean percent of protein coverage between MS-Homology and Mascot searches. Differences were considered significant at $p < .05$.

RESULTS AND DISCUSSION

Trigger Intensity Versus Total Ion Current

In this experiment, we compared two methods of selecting peptides for homology search. Peptides were selected based on precursor ion trigger intensity or the summed total intensity of the fragment ions following

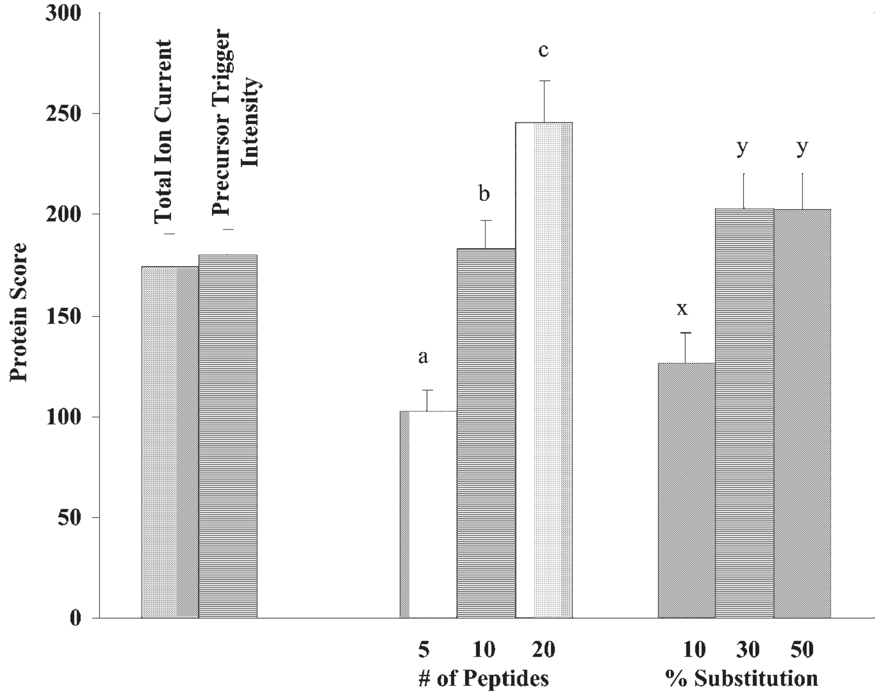

FIGURE 1

The effects of method of peptide ranking, number of peptides used, and tolerated percent of amino acid substitutions on protein scores (± SE) resulting from a homology database search using MS-Homology. Letters *a, b,* and *c* represent differences in protein scores as the number of peptides queried increases ($p < .05$). Letters *x* and *y* represent differences in protein scores as the percent of allowed amino acid substitution increases ($p < .05$).

low-energy collision-induced dissociation. The two intensities for each peptide were not well correlated ($r = 0.526$). Nevertheless, protein scores did not differ ($p < .05$) whether the top 20 peptides were selected using precursor trigger intensity (174.53 ± 15.97) or total ion current (180.36 ± 12.09), and there was no interaction with either the number of peptides used or the percent amino acid substitution (Fig. 1). By ranking peptides based on parent ion trigger intensity and selecting the 20 most intense peptides, we predicted that these peptides had a greater chance of being derived from the most abundant protein in a gel spot. That is, the likelihood increases that the resulting protein identification will correspond to the dominant protein in the spot and not be associated with proteins representing background gel streaking or contamination. In support of this prediction, Perkins et al.[5] showed that limiting a search to the most intense peaks gave the lowest probability score (best match) compared with searches that included peaks with lower intensity. Our data suggests that selection of the most intense precursor ions based on trigger intensity is sufficient to generate a legitimate protein identification.

Increasing the number of tryptic peptides or increasing the amino acid substitutions allowed for a homology search should increase protein scores. Our results indicate (Fig. 1) that this expectation was well founded. In fact, highest ($p < .05$) protein scores were obtained when the 20 most intense peptides were used with 50% amino acid substitution. Search queries with 10 peptides resulted in higher ($p < .05$) protein scores than searching with 5 peptides. Allowing 30–50% amino acid substitution resulted in higher ($p < .05$) protein scores compared with limiting substitutions to 10%. Using either of the two selection methods to rank peptide ions, it appeared that search results are most favorable when queries are performed with the 20 most abundant peptides and allowing for 30–50% amino acid mismatches. However, accepting the highest protein score as a positive match without first establishing a threshold that is both sensitive enough to accept accurate matches and selective enough to reject false positives could result in a type I error (false-positive) by considering a high protein score as a significant match when the opposite is true.[11,19,20] Thus, it was necessary to establish a threshold score for an acceptable match.

Threshold Determination

Random peptides were used to determine a threshold for an acceptable protein identification. Calculating the mean and standard deviation of protein scores for random peptides allows for the establishment of this threshold for any desired level of significance. The mean ± 2 standard deviations gives a value where searches with protein scores above this calculated threshold are significant ($p < .05$; Fig. 2). Using MS-Homology with 50% mismatch resulted in protein scores from random peptides that increased as the number of peptides in the search increased. This phe-

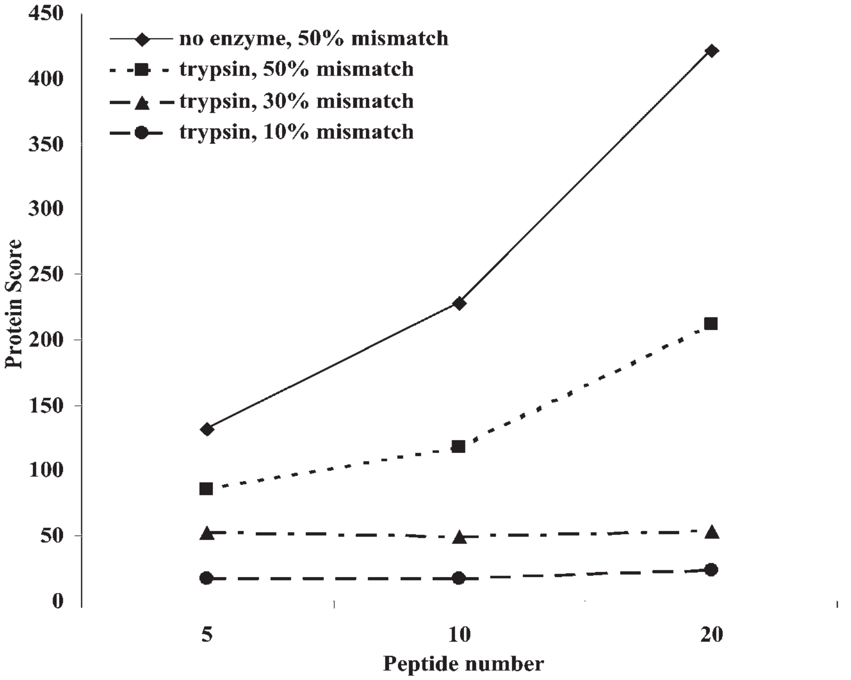

FIGURE 2

Relationship between significant protein scores (random peptide score ± 2 SD) and the number of peptide and MS-Homology search option combinations.

nomenon was made worse by selecting the no-enzyme option of this program: for 20 peptides searched, scores must be twice as great to reach significance using the no-enzyme option compared with the trypsin option. The increase in protein scores when 50% amino acid substitution was permitted illustrates the rise in false-positive matches that can occur when the error tolerance is increased—especially when the number of peptides increases from 5 to 20. Counterintuitively, when 10% or 30% substitution was allowed, the protein scores were not dependent on the number of peptides in the search (up to 20 peptides). Using these results, the difference between protein scores of the 9 known proteins used in the initial query and threshold values for random peptides could be calculated, giving a measure of signal above noise for each method of analysis. When this was done, similar to the unadjusted data, there was no difference ($p > .05$) in protein scores whether peptides were selected based on precursor ion intensity or on total ion current (113.82 ± 10.55 vs. 115.34 ± 14.31, respectively). A significant ($p < .001$) interaction between the number of peptides and the allowable percent amino acid substitution was present (Fig. 3). After subtraction of random peptide scores, permitting 30% substitution in the homology-tolerant database searches resulted in the highest ($p < .05$) protein scores when either 10 or 20 peptides were used to search the data, with 20 peptides being significantly ($p < .05$) greater than 10 peptides. Thus, the highest protein scores above those for random peptides were obtained using the following parameters: submitting the 20 most abundant peptides based on trigger intensity and allowing a maximum of 30% amino acid substitution. This suggests that at 70% homology, an optimal balance is achieved between protein scores resulting from random sequence and those resulting from specific sequence, thus providing the opportunity to identify homologs from distantly related species. According to Mackey et al.,[11] a homology search could identify amino acid sequences that share 65% identity and this would include proteins that diverged in the past 150–500 million years. However, given the constraints used in this study, this method will be unsuitable for proteins that are less than 70% identical.

Mass-Based Versus Homology-Tolerant Searching

Using the above parameters, we performed homology-tolerant searches with either the top 20 or all doubly charged peptides and compared them with Mascot searches. Of the proteins identified, 63.4% of the identifications were similar for all three analyses (Fig. 4). Of the remainder, 16% of the proteins identified as the best match by Mascot were not the highest scoring proteins from the homology search. It is in this group of spots that the potential for homology-tolerant searching can be seen. However, homology searching using the most intense peptides versus all

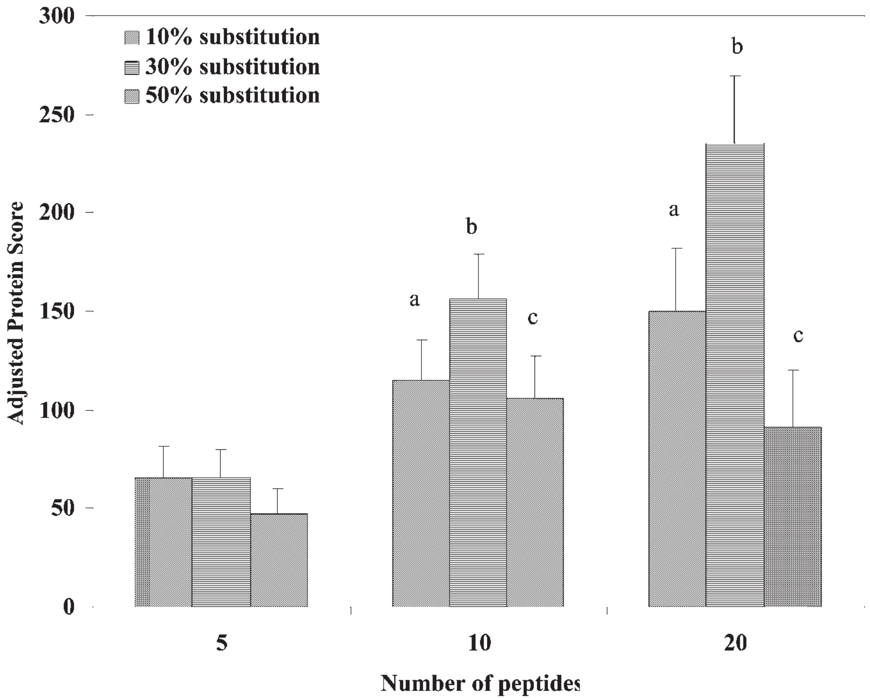

FIGURE 3

Average protein scores (± SE) from nine proteins after subtracting the significant protein score (see Fig. 2) from each protein score (see Fig. 1) obtained by MS-Homology. Letters *a, b,* and *c* indicate significant differences in protein scores among the three amino acid substitution rates within each the number of peptides in the search[Au: please clarify] (*p* <.05).

doubly charged peptides provided differing protein identifications for 6.3% of the spots. At this time, we cannot determine which identification is correct. Still, the results of the homology-tolerant search may give an alternative protein identification that is less likely to suffer from the inherit weakness in search methods using mass alone (i.e., Mascot) for cross-species identification. At the very least, if the identification is sufficiently critical, the alternative protein identifications could be further investigated using other means.

No direct comparison between MS-Homology and MASCOT can be made, because each program uses a different scoring method to determine the "best" match. However, the percent coverage of proteins matched by all three methods was highest ($p < .05$) for proteins identified using Mascot (25.1%) compared with spots analyzed using homology-tolerant searches (Fig. 4). Furthermore, using all *de novo* sequenced peptides to query the database provided greater ($p < .05$) protein coverage than when the search was limited to the 20 most intense precursor ions (18.3% vs. 10.6%, respectively). The number of peptides recognized from a digested protein is a factor in determining the confidence in a protein identification from sequence similarity search.[21] Increases in protein coverage detected by submitting more peptides to a database search may come at the expense of increasing the frequency of false-positive identifications compared with searches using only the 20 most abundant peptides, because more peptides resulting from background contamination may be incorporated in the search.

Accuracy of *De Novo* Sequence Assignments

One explanation for the reduced percent coverage of the homology-tolerant search is the performance of the *de novo* sequencing software. Although preliminary analysis indicated that PEAKS performed *de novo* sequencing better and faster than we could do manually, further analysis seemed prudent. To assess the accuracy of PEAKS to correctly *de novo* sequence MS/MS data, we selected 25 gel spots that represented 11 porcine proteins that were similarly identified by Mascot and MS-Homology. All peptide fragments were *de novo* sequenced and submitted to a database search. Combining results of MS-Homology and MASCOT searches, 18% ($n = 437$) of the 2417 peptides used in the database search matched published database peptides. Of these, 35.0% were identical to database sequences, whereas, 11.7% of the sequences had greater than 30% mismatches. Only Mascot was able to identify peptide sequences with greater than 30% incorrect amino acid assignments after analysis by PEAKS, because these peptides surpassed the set, acceptable substitution level for the homology search. This makes clear the advantage of searching with uninterpreted fragment mass lists for species in which databases are complete. Ma et al.[14] reported similar (33%) results when the quality of the spectrum, as determined by the signal-to-mass ratio, was marginal. These authors also showed a higher percentage (44%) of complete sequence matches when the signal intensity increased in relation to the peptide mass. We sub-

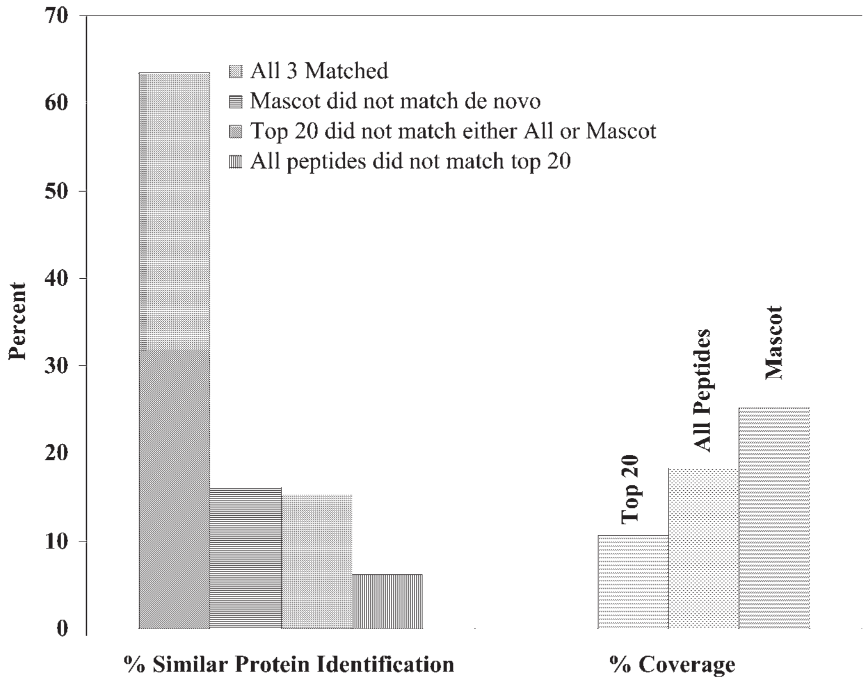

FIGURE 4

Database searches were conducted on 112 spots using *de novo* sequences of peptides in a homology-tolerant search (MS-Homology) or using peptide masses (Mascot). *Left*: Percent of similar proteins identified after searching with the 20 most intense *de novo* sequenced peptides for each spot, all *de novo* sequenced peptides for each spot, or using MASCOT. *Right:* Percent coverage for each analysis using results of those proteins identified identically by all three methods.1

mitted all *de novo* sequences irrespective of the quality of the spectrum; however, out of the 20 most intense peptides based on trigger intensity, a greater percentage (44.0%) of the *de novo* sequences matched completely to published sequences. Collectively, these results indicate that peptides with greater precursor trigger intensity have a greater proportion of spectra that are correctly interpreted by PEAKS; thus the power of homology-tolerant searching increases and the likelihood of a correct protein identification improves.

Finally, it is difficult to distinguish between isomeric and isobaric amino acids when analyzing spectra by *de novo* sequencing. Previously, the only amino acids incorrectly assigned by PEAKS were those with equivalent mass.[14] In this study, the majority of amino acid mismatches were located at the *N*-terminus (Fig. 5). Of these, the greatest percentage of substitutions were between dipeptide isobars (i.e., reversed sequence order). Much of this type of error can be overcome during the database search by setting the maximum threshold for amino acid substitutions at 30%. Allowing for this level of mismatch, MS-Homology was able to match peptides with more than one amino acid substitution; however, MS-Homology was not able to identify peptides when isomeric substitutions (i.e., Gly-Ala/Gln, Gly-Glu/Trp) occurred, resulting in insertions and deletions in the sequence. To overcome this limitation, MS-Homology allows users to set mismatching at 50% or define these changes, but as shown here, this comes at the expense of increasing the number of randomly matched peptides, elevating protein scores, and actually decreases the ability to obtain a protein identification. Another option MS-Homology provides to circumvent these limitations is to use sequence data in combination with the peptide mass information for ambiguous regions in a spectrum. This technique allows search algorithms to assign known amino acids to spectrum segments not easily interpreted *de novo* and still take advantage of the reliable *de novo* sequence information provided in an MS/MS experiment. Mann and Wilm showed that using the combination of sequence and mass data improves the discriminating power to find a matching peptide.[12] However, the limitation with this technique for cross-species searching is that the amino acid sequence of the peptide in the uninterpreted portion must be identical to that in the database. If the residues are not conserved between species, then the peptide will not be matched and will not contribute to the identification of the protein.

CONCLUSIONS

Limiting a homology search to the 20 most intense peptides based on trigger intensity may provide protection against false-positive identification because many of the most intense peptides are likely to be products from the most abundant protein from a spot. In addition, MS/MS spectra are more likely to be correctly interpreted for the 20 most intense peptides.

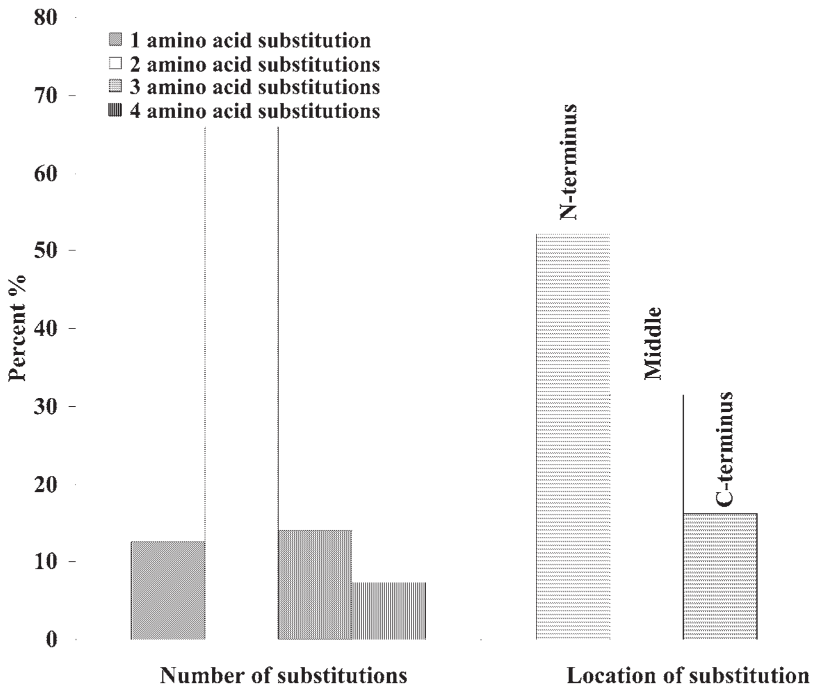

FIGURE 5

Results from 25 samples comparing peptide sequences obtained using PEAKS with the actual known peptide sequence from the database. The bars on the *left* show the percent occurrence of 1, 2, 3, or 4 mismatched amino acid in the peptides. The bars on the *right* indicate the influence of the relative location of amino acid substitutions on the incidence of mismatch.

Although high-quality spectra are associated with peptides not in the top 20, increasing the number of peptides above 20 may increase the chance that those additional peptides were generated from background proteins within the gel. Once *de novo* sequencing is performed, our results indicate that MS-Homology performs optimally allowing for 30% mismatch. Using random peptides and these settings, a protein score greater than 50 is significant. Thus, our results suggest that for cross-species comparisons, both mass-based and homology-tolerant searches should be performed and the results compared. For those proteins where the search results do not match, extra care should be taken before assigning an identification to that protein. With our data, this occurred for approximately 16% of the proteins. In addition, where search results do match, a search based on peptide mass may provide further information and assurance that a protein match is significant by increasing the percent coverage of the protein identified.

ACKNOWLEDGMENTS

The mass spectrometry facility is supported in part by NIH Grant P20 RR15635 from the COBRE Program of the National Center for Research Resources, NCI Cancer Center Support Grant P30 CA36727, and the Nebraska Research Initiative.

REFERENCES

1. Pandey A, Mann M. Proteomics to study genes and genomes. *Nature* 2000;405:837–846.
2. Standing KG. Peptide and protein *de novo* sequencing by mass spectrometry. *Curr Opin Struct Biol* 2003;13:595–601.
3. Papayannopoulos IA. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spec Rev* 1995;14:49–73.
4. Eng JK, McCormack AL, Yates JR, III. An approach to correlate tandem mass spectra data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994;5:976–989.
5. Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching databases using mass spectrometry data. *Electrophoresis* 1999;20:3551–3567.
6. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identications made by MS/MS and database search. *Anal Chem* 2002;74:5383–5392.
7. Huang L, Jacob RJ, Pegg SC-H, et al. Functional assignment of the 20 S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J Biol Chem* 2001;276:28327–28339.
8. Lester PJ, Hubbard SJ. Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics. *Proteomics* 2002;2:1392–1405.
9. Clauser KR, Baker P, Burlingame AL. Role of accurate mass measurement (± 10 ppm) in protein identification strategies employing MS of MS/MS and database searching. *Anal Chem* 1999;71:2871–2882.
10. Taylor JA, Johnson RS. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* 2001;73:2594–2604.

11. Mackey AJ, Haystead TAJ, Pearson WR. Getting more from less. *Mol Cell Proteomics* 2002;1.2:139–147.
12. Mann M, Wilm M. Error-tolerant indentification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994;66:4390–4399.
13. Taylor JA, Johnson RS. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 1997;11:1067–1075.
14. Ma B, Zhang K, Hendrie C, et al. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;17:2337–2342.
15. Johnson RS. Lutefisk1900 versus Peaks: a comparison of automated de novo sequencing programs. *J Biomol Technol* 2004;15:66.
16. Roberts RM, Baubach GA, Buhi WC, et al. Analysis of membrane polypeptides by two-dimensional polyacrylamide gel electrophoresis. In: Venter JC, Harrison LC (eds.): *Molecular and Chemical Characterization of Membrane Receptors*, vol. 3. New York: Alan R. Liss, 1984:61–113.
17. Shevchenko A, Wilm M, Vorm O, Mann M. Mass spectrometric sequencing of proteins from silver-stain polyacrylamide gels. *Anal Chem* 1996;68:850–858.
18. Henikoff S, Henikoff JG. Amino acid matrices from protein blocks. *Proc Natl Acad Sci USA* 2003; 89:10915–10919.
19. Pearson WR. Comparison of mthods for searching protein sequence databases. *Protein Sci* 1995;4:1145–1160.
20. Boguski MS, McIntosh MW. Biomedical informatics for proteomics. *Nature.* 2003;422:233–237.
21. Liska AJ, Shevcheneko A. Expanding the organismal scope of proteomics: Cross-species protein identification by mass spectrometry and its implications. *Proteomics* 2003;3:19–28.