Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm

Michael J. MacCoss,[†,‡] Christine C. Wu,[†,‡] and John R. Yates, III*[,†,§]

Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, and Department of Proteomics and Metabolomics, Torrey Mesa Research Institute, 3115 Merryfield Row, San Diego, California 92121-1125

Database-searching algorithms compatible with shotgun proteomics match a peptide tandem mass spectrum to a predicted mass spectrum for an amino acid sequence within a database. SEQUEST is one of the most common software algorithms used for the analysis of peptide tandem mass spectra by using a cross-correlation (XCorr) scoring routine to match tandem mass spectra to model spectra derived from peptide sequences. To assess a match, SEQUEST uses the difference between the first- and second-ranked sequences (ΔCn). This value is dependent on the database size, search parameters, and sequence homologies. In this report, we demonstrate the use of a scoring routine (SEQUEST-NORM) that normalizes XCorr values to be independent of peptide size and the database used to perform the search. This new scoring routine is used to objectively calculate the percent confidence of protein identifications and posttranslational modifications based solely on the XCorr value.

Scientists of the postgenomic era are quickly becoming advocates of global analyses of biological systems. This excitement has been ignited by the success of microarray technology, and current trends in the field of proteomics are shifting from traditional two-dimensional gel electrophoresis (2D-GE)-based technology to more efficient global analysis methods. Shotgun proteomics has emerged as a promising strategy because it provides robust and sensitive methods to profile the protein complement within a complex biological sample.[1] This approach uses multidimensional protein identification technology (MudPIT) which incorporates the resolving power of multidimensional high-pressure liquid chromatography (LC/LC), the sequence analysis capabilities of tandem mass spectrometry, and database searching software.[1,2]

A sample for MudPIT analysis is prepared by digesting a complex protein mixture with proteases to produce an even more complex peptide mixture. The peptides are loaded directly onto an LC/LC column placed in-line with a tandem mass spectrometer. Tandem mass spectra are acquired "on the fly" as peptides are eluted from the column, ionized, and emitted into the mass spectrometer.[1,2] Respective peptide sequences are identified using software that correlates experimentally acquired tandem mass spectra against theoretical spectra predicted from amino acid sequences contained within a sequence database.[3] Peptides are then "assembled" back into proteins using a second software algorithm.[4] This methodology is becoming increasingly routine, user-friendly, and applicable to almost any biological system. However, although data can be generated and searched very rapidly, assessing the quality of these results in an objective manner can often be the analytical bottleneck especially when assessing the presence of a protein based on one tandem mass spectral match to a sequence.

Proteomic analyses are dependent on software that searches a sequence database using data derived from mass spectrometry. Current database-searching algorithms are usually divided into one of two categories: (1) software that matches peptide masses to predicted protein peptide maps[5] or (2) software that matches a peptide tandem mass spectrum to a predicted mass spectrum for an amino acid sequence within a database.[3] The first category generally assumes that the peptide masses are derived from samples composed of only a few components using a protease with selected amino acid specificity and, thus, is not compatible with complex protein mixtures. In contrast, the second category is ideally suited to protein mixtures because the data analysis is performed on the single peptide level and, later, assembled back into proteins after the peptide sequences have been identified.

The identification of proteins by MudPIT is complicated because the database correlation software must be capable of handling tandem mass spectra from peptides produced without any cleavage specificity. Even if a highly specific enzyme (e.g., trypsin) is used during the sample preparation, the presence of endogenous proteases within the experimental sample results in a significant amount of nonspecific cleavage. Recently, nonspecific proteases have been exploited to produce overlapping peptides

* To whom correspondence should be addressed. E-mail: jyates@scripps.edu. Voice: (858) 784-8862. Fax: (858) 784-8883.

† The Scripps Research Institute.

‡ These authors contributed equally to this work.

§ Torrey Mesa Research Institute.

(1) Link, A. J.; Eng, J. K.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676−82.

(2) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−7.

(3) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−89.

(4) Tabb, D. L.; McDonald, W. H.; Yates, J. R., III. *J. Proteome Res.* **2002**, *1*, 21−6.

(5) Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. *Anal. Biochem.* **1993**, *214*, 397−408.

10.1021/ac025826t CCC: $22.00
Published on Web 00/00/0000

facilitating the analysis of single-nucleotide polymorphisms[6] and covalent modifications.[7] This further accentuates the complexity of the dataset.

The database-searching algorithm SEQUEST uses a cross-correlation (XCorr) function to assess the quality of the match between a tandem mass spectrum and amino acid sequence information in a database. The XCorr value is a database-independent measure that is dependent on the quality of the tandem mass spectrum and the quality of its fit to the model spectrum.[3] SEQUEST creates a model tandem mass spectrum based on rudimentary knowledge of how peptides fragment in the CID process.[3] The protease trypsin produces, in general, doubly charged peptides whose fragmentation in the CID process can be reasonably predicted because they contain only a single basic amino acid on the C-terminus.[8] SEQUEST can match tandem mass spectra of peptides created without cleavage specificity to sequences in the database. Because peptides created by nonspecific proteases can create peptides with basic amino acids (e.g., Arg, Lys, His) anywhere within the sequence, peptide fragmentation is less predictable. The XCorr value generated during the analysis is an absolute measure of spectral quality and closeness of fit to the model spectrum. Thus, the same XCorr value for one peptide may not reflect a similar closeness of fit for another peptide with the same score (e.g., an XCorr of 3.00 for one peptide may not necessarily equate to an XCorr of 3.00 for a different peptide). SEQUEST scores are currently normalized using the difference between the first- and second-ranked sequences (ΔCn).[3,9,10] Although a $\Delta Cn > 0.1$ has been reported as an acceptable match, this value is highly dependent on the database size, the search parameters, and the sequence homologies. To create a scoring parameter that is consistent for peptides regardless of what protease was used to create them or how large the peptide is, we evaluated a procedure previously used with a cross-correlation method for library searching with tandem mass spectra of peptides.[11] In this paper, we demonstrate the use of a new SEQUEST scoring routine and validate the identification of proteins and covalent modifications using a probability-based calculation.

EXPERIMENTAL SECTION

Standard Preparation and Digestion. A known mixture of proteins containing equimolar levels of phosphorylase a (rabbit skeletal muscle), cytochrome *c* (horse), apomyoglobin (horse heart), albumin (bovine serum), and β-casein (bovine) was used for all experiments. The resulting mixture ($\sim$1 nmol/mL in water) was adjusted to 8 M urea with the addition of solid urea, reduced with dithiothreitol (20 mM final concentration at 50 °C for 20 min), and alkylated with iodoacetamide (50 mM final concentration in the dark at room temperature).

The denatured, reduced, and alkylated standard mixture was divided into four aliquots, and each was digested using a different protease. Aliquot 1 was diluted 3-fold with 100 mM Tris, pH 8.5, and $CaCl_2$ was added to a final concentration of 1 mM. Modified trypsin (Promega) was added at a 1:100 enzyme-to-substrate ratio (w/w), and the mixture was incubated overnight at 37 °C with constant mixing (Thermomixer, Eppendorf). Aliquot 2 was diluted 3-fold with 100 mM Tris, pH 8.5, elastase (Roche) was added at a 1:50 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated overnight with mixing at 37 °C. Aliquot 3 was diluted 3× with 100 mM Tris, pH 8.5, subtilisin (Sigma) was added at a 1:50 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated with mixing for 3 h at 37 °C. Aliquot 4 was adjusted to pH 11 with 1 M NaOH, proteinase K (Roche) was added at a 1:100 enzyme-to-substrate ratio (w/w), and the resultant mixture was incubated at 37 °C for 3 h with constant mixing. Each digestion was quenched with the addition of formic acid to 5% and frozen at −80 °C until analysis by MudPIT as described below.

Multidimensional Protein Identification Technology. A triphasic microcapillary column was constructed from 100-μm-i.d. fused-silica capillary tubing pulled to a 5-μm-i.d. tip using a Sutter Instruments P-2000 CO_2 laser puller (Novato, CA). Each column was packed with 7 cm of 5-μm Aqua C18 material (Phenomenex, Ventura, CA) and 3 cm of 5-μm Partisphere strong cation exchanger (Whatman, Clifton, NJ), followed by another 3 cm of Aqua C18. The columns were equilibrated with 5% acetonitrile/0.1% formic acid, and $\sim$4 pmol of each protein digest was loaded directly onto separate capillary columns using a high-pressure bomb.

After the peptide digests were loaded, the column was placed in-line with a Surveyor quaternary HPLC (ThermoFinnigan, Palo Alto, CA) and analyzed using a modified six-step separation described previously.[2] The buffer solutions used were 5% acetonitrile/0.1% formic acid (buffer A), 80% acetonitrile/0.1%formic acid (buffer B), and 500 mM ammonium acetate/5% acetonitrile/0.1% formic acid (buffer C). Step 1 consisted of a 100-min gradient from 0 to 100% buffer B. Steps 2−5 had the following profile: 3 min of 100% buffer A, 2 min of *X*% buffer C, a 10-min gradient from 0 to 15% buffer B, and a 97-min gradient from 15 to 45% buffer B. The 2-min buffer C percentages (*X*) were 10, 20, 30, 40, and 50% for the six-step analysis. For the final step, the gradient contained the following: 3 min of 100% buffer A, 20 min of 100% buffer C, a 10-min gradient from 0 to 15% buffer B, and a 107-min gradient from 15 to 70% buffer B.

As peptides eluted from the microcapillary column, they were electrosprayed directly into an LCQ-Deca mass spectrometer (ThermoFinnigan) with the application of a distal 2.4-kV spray voltage. A cycle of one full-scan mass spectrum (400−1400 m/z) followed by three data-dependent tandem mass spectra at a 35% normalized collision energy was repeated continuously throughout each step of the multidimensional separation. The application of all mass spectrometer scan functions and HPLC solvent gradients was controlled by the Xcaliber data system.

Analysis of Tandem Mass Spectra. Tandem mass spectra were analyzed using the following protocol. First, a software

(6) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R., III. *Anal. Chem.* **2000**, *72*, 757−63.
(7) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., III. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7900−5.
(8) Tang, X.; Boyd, R. K. *Rapid Commun. Mass Spectrom.* **1992**, *6*, 651−7.
(9) Yates, J. R., III; Eng, J. K.; McCormack, A. L.; Schieltz, D. M. *Anal. Chem.* **1995**, *67*, 1426−36.
(10) Yates, J. R., III; Eng, J. K.; McCormack, A. L. *Anal. Chem.* **1995**, *67*, 3202−10.
(11) Yates, J. R., III; Morgan S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K. *Anal. Chem.* **1998**, *70*, 3557−67.

algorithm called 2to3 [12] was used to determine the appropriate charge state (either +2 or +3) of multiply charged peptide mass spectra and delete spectra of poor quality. Each tandem mass spectrum after 2to3 was searched twice against the National Center for Biotechnology Information (NCBI) nonredundant protein database (downloaded 04/15/2002 with 907 654 protein sequence entries) using two different versions of SEQUEST. The first search used a previously reported cross-correlation scoring algorithm,[3] while the second search, used a normalized cross-correlation routine (SEQUEST-NORM) described below. All searches were parallelized and performed on a Beowulf computer cluster consisting of 34 1.2-GHz Athlon CPUs. Tandem mass spectra was searched without considering any enzyme cleavage specificity.

SEQUEST-NORM. SEQUEST-NORM identifies the "best" 500 amino acid sequences from a sequence database using the same preliminary scoring routine reported previously for SEQUEST.[3] The experimental tandem mass spectrum was then cross-correlated against theoretical tandem mass spectra for each 500-amino acid sequence identified in the preliminary scoring using a fast Fourier transform analysis.[3] The cross-correlation score was then normalized to the cross-correlation of the input spectrum against itself (autocorrelation). This autocorrelation value represents the best possible match between the tandem mass spectrum and a theoretical spectrum from an amino acid sequence within a database. In the final output, spectral matches were ranked from best to worst by this normalized cross-correlation, where 1.0 represents a perfect match and a value near 0.0 represent a poor match. An algorithm to convert XCorr values to normalized XCorr values will be available at the web site www.SEQUEST.org.

Validation of SEQUEST-NORM Results. The number of mass spectra at 0.005 XCorr intervals was used to calculate a relative frequency an Xcorr value results in either a correct or incorrect (*f*) match. The sum of all the incorrect *F* values in the distribution up to the acquired XCorr value was used to estimate the confidence in the identified peptide sequence (eq 1). Simply,

$$\mathrm{Pept}_{\mathrm{Prob}} \sum_{ir=0}^{\mathrm{XCorr}} f_{\mathrm{i}} \qquad (1)$$

as the XCorr value increases, the probability that the peptide sequence determined by SEQUEST is incorrect, decreases. After calculating the confidence for each peptide identification, the effect of multiple peptides on the probability of the protein identification ($\mathrm{Prot}_{\mathrm{Prob}}$) was determined using the equation

$$\mathrm{Prot}_{\mathrm{Prob}} = 1 - [(1 - \mathrm{Pept1}_{\mathrm{Prob}})(1 - \mathrm{Pept2}_{\mathrm{Prob}})(1 - \mathrm{Pept3Prob})(1 - \mathrm{Petp}n_{\mathrm{Prob}})...] \qquad (2)$$

RESULTS AND DISCUSSION

Software that searches tandem mass spectra of peptides against a sequence database usually begins by identifying strings of amino acids with the correct mass. If the cleavage specificity of a given peptide is not known, then the possible amino acid strings within a database at the respective peptide mass increase dramatically. However, given adequate computer resources, a spectrum with sufficient structural information should theoretically be able to consider all these "extra" amino acid sequences and identify the correct peptide regardless of whether the cleavage specificity is known or not.

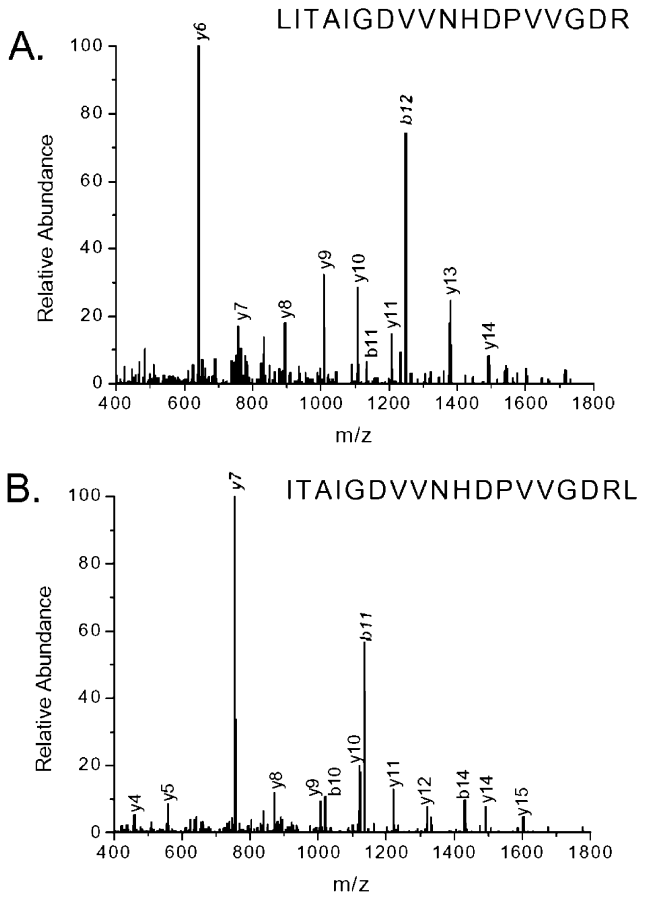

Figure 1. SEQUEST identification of peptides with no assigned enzyme cleavage specificity. Tandem mass spectra of two glycogen phosphorylase a peptides formed by cleavage with (A) trypsin and (B) proteinase K. Both peptides have identical amino acid compositions, and their sequences are the same except for the C- and N-terminal residues. Both spectra were searched against the nonredundant protein database using SEQUEST and SEQUEST-NORM with no assigned enzyme cleavage specificity. Both peptides were correctly identified to the protein by either version of SEQUEST with high XCorr scores: (A) tryptic peptide, 5.211 SEQUEST/0.566 SEQUEST-NORM; and (B) proteinase K peptide, 5.048 SEQUEST/0.511 SEQUEST-NORM.

The SEQUEST database-searching algorithm uses an XCorr routine that has never been formally validated with the use of nonspecific enzymes. Figure 1 shows the tandem mass spectra of two glycogen phosphorylase a peptides formed by cleavage with (A) trypsin (a specific protease) and (B) proteinase K (a nonspecific protease). Both peptides have identical amino acid compositions and only differ by the C- and N-terminal residues. A prominent y-ion series is present in both mass spectra. As predicted, intense peaks resulting from proline-directed fragmentation are observed in both the tryptic (y6 and b12 ions) and proteinase K (y7 and b11 ions) peptides. The spectra were searched against the nonredundant protein database using both SEQUEST and SEQUEST-NORM with no enzyme cleavage

(12) Sadygov, R. G.; Eng, J. K.; Durr, E.; Saraf, A.; McDonald, W. H.; MacCoss, M. J.; Yates, J. R., III. *J. Proteome Res.* **2003**, *3*, 211−5.

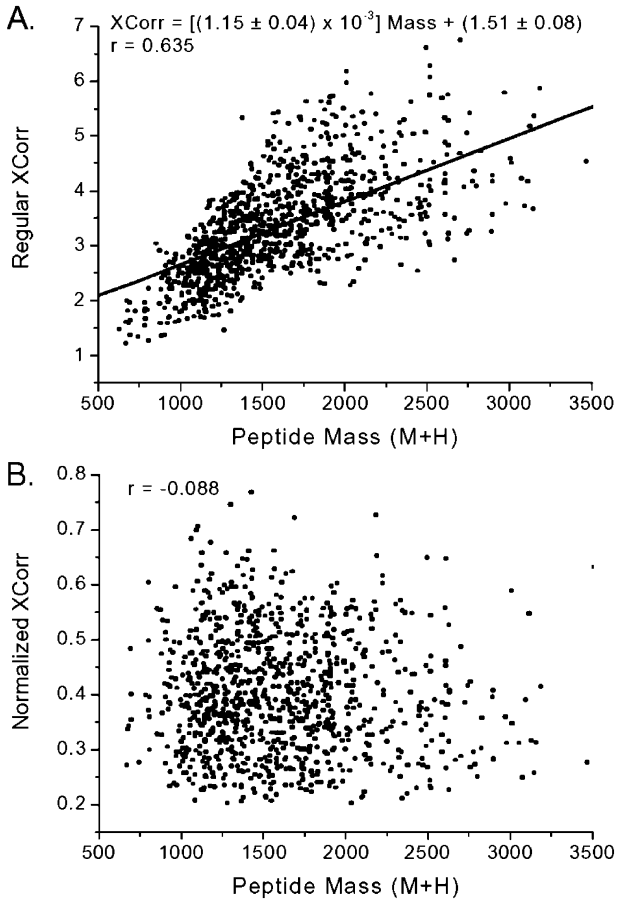

Figure 2. Elimination of the correlation between peptide mass and Xcorr value with SEQUEST-NORM. Xcorr values produced by (A) SEQUEST and (B) SEQUEST-NORM are plotted against the deconvoluted mass of peptides formed by the digestion of a standard protein mixture with trypsin, proteinase K, subtilisin, and elastase. Linear regression using a least-squares analysis shows that (A) the regular SEQUEST XCorr is correlated with peptide mass with the best-fit line having a positive slope [$(1.15 \pm 0.04) \times 10^{-3}$] while (B) the modified SEQUEST-NORM XCorr is not correlated with peptide mass ($r = -0.088$).

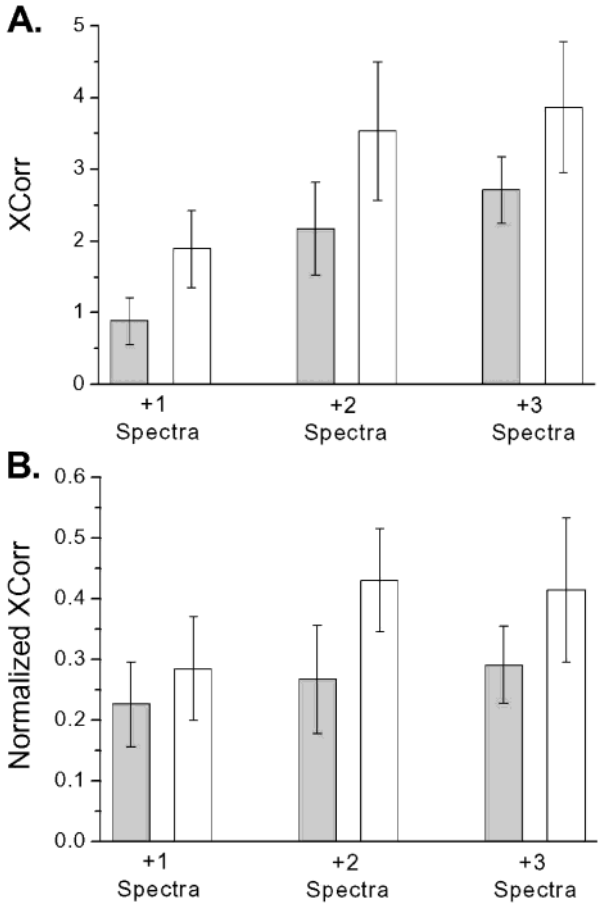

Figure 3. Minimization of the effect of precursor ion charge states of tryptic peptides on XCorr values with SEQUEST-NORM. Mean XCorr's for *correct* (unshaded) and *incorrect* (shaded) matches produced by (A) SEQUEST and (B) SEQUEST-NORM are shown for tryptic fragments of each ion charge state (+1, +2, +3).

specificity considered. Both peptides were correctly identified by both versions of SEQUEST. These spectra confirm that, given sufficient fragmentation, SEQUEST is capable of identifying the correct amino acid sequence without any assumptions concerning cleavage specificity.

Although SEQUEST can routinely identify the correct amino acid sequence, evaluation of results based on XCorr alone can be confusing because of a trend between peptide mass and XCorr. SEQUEST (Figure 2A) and SEQUEST-NORM (Figure 2B) XCorr values are plotted against the deconvoluted mass of peptides formed from the digestions with trypsin, proteinase K, subtilisin, and elastase. A linear regression using a least-squares analysis shows that the regular SEQUEST XCorr is correlated with peptide mass with the best-fit line having a positive slope [$(1.15 \pm 0.04) \times 10^{-3}$]. In contrast, the exact same tandem mass spectra show no correlation with peptide mass ($r = -0.088$) when searched with SEQUEST-NORM. These data demonstrate that the normalizing routine in SEQUEST-NORM minimizes the relationship between SEQUEST XCorr values and peptide mass.

Panels A and B of Figure 3 demonstrate the effect of the precursor ion charge state on tryptic peptides searched using SEQUEST and SEQUEST-NORM, respectively. In Figure 3A, a trend of increasing XCorr is observed with increasing precursor ion charge state in both the correctly (unshaded) and incorrectly (shaded) matching mass spectra. This observation is not surprising because larger peptides are more likely to result in multiply charged ions and, thus, follows the trend shown in Figure 2A where the XCorr increases with increasing peptide mass.

If the normalized XCorr removes dependence on peptide length, then the score should better reflect the quality of the match between the experimental and the theoretical spectrum. Because peptides fragment differently, the magnitude of the correlation will vary and all *correctly matching* peptides will not necessarily have the same scores unless the same number of fragment ions are generated for a sequence and all of those fragment ions are present in the tandem mass spectrum. However, because the match of a peptide tandem mass spectrum to an *incorrect* amino acid sequence should be a pseudorandom process, we would expect these normalized Xcorr values to fall in a similar range. As expected, the normalized Xcorr values of the incorrectly matched spectra (shaded bars; Figure 3B) all have similar mean XCorr values ($+1 = 0.23 \pm 0.03$, $+2 = 0.27 \pm 0.09$, $+3 = 0.29 \pm 0.06$). Likewise, the +2 and +3 spectra in Figure 3B also have

similar mean values and do not show the same increasing trend observed with the regular XCorr.

An important observation in Figure 3B is that the mean normalized XCorr for a +1 tryptic spectrum is significantly lower than for a tryptic +2 or +3 spectrum. Although, SEQUEST-NORM is independent of peptide size, its XCorr values are representative of the correlation between the experimental and theoretical spectra. Because the proton is localized on the C-terminal Arg or Lys of +1 peptides, charge-localized fragmentation along the peptide backbone is reduced.[13,14] Therefore, it is unlikely that a +1 tryptic peptide will provide similar sequence-specific fragmentation as multiply charged peptide precursors and thus a lower normalized XCorr value is expected because the spectrum contains less information.[15] In all cases, the mean XCorr for correctly matched tandem mass spectra is greater than for incorrectly matched spectra.[16]

Database searching software capable of handling peptides without predictable cleavage specificity is a necessary proteomic tool. A significant amount of nonspecific cleavage occurs even when a highly specific enzyme (e.g., trypsin) is used during the preparation of samples containing complex mixtures of proteins. Figure 4 shows the effect of nonspecific peptides on SEQUEST XCorr values. Three different proteases (proteinase K, elastase, subtilisin) with little or no cleavage specificity were used to produce peptides for analysis as described in the Experimental Section. The effect of charge state on the mean XCorr is shown for each nonspecific protease separately. Results identifying amino acid sequences present in the mixture are unshaded, whereas, results identifying peptides not present in the mixture are shaded. As with trypsin (Figure 2A), the mean XCorr values from the regular SEQUEST algorithm increase with increasing precursor ion charge state for both correctly and incorrectly matching spectra. The spectra searched with SEQUEST-NORM do not follow this trend. Furthermore, using the normalized XCorr, the relative difference between the correctly matching tandem mass spectra of +1 and multiply charged precursor ions is much smaller than the difference observed with tryptic peptides. Singly charged spectra of nontryptic peptides often contain more abundant fragment ions than singly charged tryptic peptides because the proton is not necessarily localized on the C-terminus. Therefore, the improved fragmentation of these +1 spectra result in a better quality match and higher normalized XCorr between the experimental and theoretical mass spectra.

Because SEQUEST-NORM is reproducible across different charge states and proteases, a data set of known peptides can be used to determine the probability an XCorr value for an unknown peptide is a correct or incorrect match. A total of 215 491 tandem mass spectra from the digestion of a known mixture of proteins (phosphorylase a, cytochrome *c*, apomyoglobin, albumin, β-casein) were analyzed by MudPIT and searched using SEQUEST-NORM. Figure 5 shows the frequency of correctly identified peptides (solid line) and incorrectly identified peptides (broken line) versus XCorr values after applying a second-order Savitsky−Golay filter.[17] The

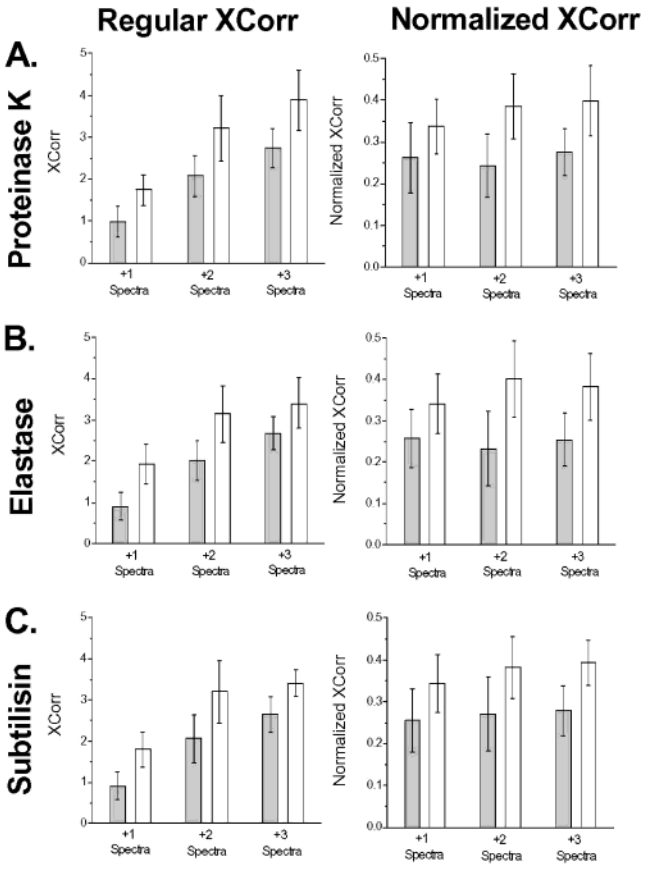

Figure 4. Effect of nonspecific peptides on XCorr values produced by SEQUEST and SEQUEST-NORM. Peptides were produced for analysis using three nonspecific enzymes: (A) proteinase K, (B) elastase, and (C) subtilisin. Mean Xcorr values for *correct* (unshaded) and *incorrect* (shaded) matches produced by SEQUEST (Regular XCorr, left side) and (B) SEQUEST-NORM (normalized XCorr, right side) are shown for each ion charge state (+1, +2, +3).

correctly matched XCorr values have a Gaussian distribution (r^2 = 0.972), whereas the incorrectly matched XCorr values are skewed toward higher XCorr values.

Assuming this data set is representative of all measurements, we can use these results to calculate the probability that a normalized XCorr identifies the correct amino acid sequence. Furthermore, the effect of multiple peptides on the confidence of a protein identification is substantial. Because the chance of a protein identification being a result of a "false positive" decreases exponentially with each peptide identified, lower XCorr's from multiple peptides can provide the same confidence as a single peptide with a high XCorr. Table 1 shows the effect of multiple peptides on the probability a protein identification is correct. A protein identification (with 95% confidence) based on a single peptide must have an XCorr > 0.40. In contrast, a protein identification based on two peptides can be awarded the same confidence if both XCorr are >0.30. Table 1 assumes that the XCorr's from multiple peptides are equal; however, different probability combinations can be calculated using eq 2 (see the Experimental Section).

Nonspecific proteases have recently been reported to facilitate the analysis of single-nucleotide polymorphisms[6] and covalent

(13) Hunt, D. F.; Yates, J. R., III; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233−7.
(14) Burlet, O.; Orkiszewski, R. S.; Ballard, K. D.; Gaskell, S. J. *Rapid Commun. Mass Spectrom.* **1992**, *6*, 658−62.
(15) Huang, E. C.; Henion, J. D. *J. Am. Soc. Mass Spectrom.* **1990**, *1*, 158−65.
(16) Griffin, P. R.; Coffman, J. A.; Hood, L. E.; Yates, J. R., III. *Int. J. Mass Spectrom. Ion Processes* **1991**, *111*, 131−49.
(17) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627−39.

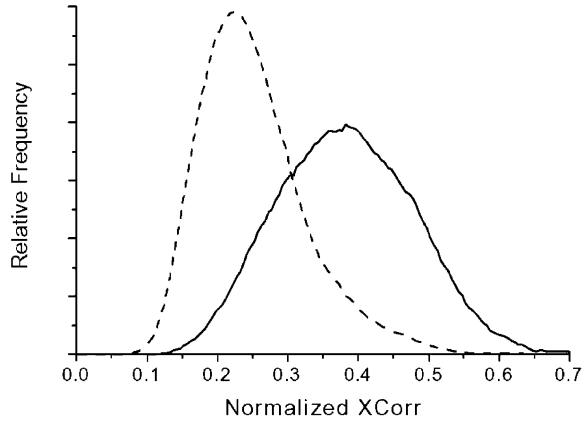

Figure 5. Relationship of frequency of correct and incorrect identifications and XCorr value produced by SEQUEST-NORM. A total of 215 491 tandem mass spectra from the digestion of a known mixture of proteins (phosphorylase a, cytochrome *c*, apomyoglobin, albumin, β-casein) were analyzed by MudPIT and searched using SEQUEST-NORM. The relative frequency of correctly identified peptides (solid line) and incorrectly identified peptides (broken line) were plotted against XCorr after smoothing with a second-order Savitsky−Golay filter.

Table 1. Effect of Multiple Peptides on the Protein Confidence[a]

	peptides per protein locus				
XCorr	1	2	3	4	5
0.25	0.552	0.799	0.910	0.960	0.982
0.26	0.605	0.844	0.938	0.976	0.990
0.27	0.654	0.880	0.959	0.986	0.995
0.28	0.699	0.909	0.973	0.992	0.998
0.29	0.739	0.932	0.982	0.995	0.999
0.30	0.775	0.949	0.989	0.997	0.999
0.31	0.806	0.962	0.993	0.999	1.000
0.32	0.833	0.972	0.995	0.999	1.000
0.33	0.856	0.979	0.997	1.000	1.000
0.34	0.876	0.985	0.998	1.000	1.000
0.35	0.893	0.989	0.999	1.000	1.000
0.36	0.908	0.991	0.999	1.000	1.000
0.37	0.921	0.994	1.000	1.000	1.000
0.38	0.932	0.995	1.000	1.000	1.000
0.39	0.942	0.997	1.000	1.000	1.000
0.40	0.951	0.998	1.000	1.000	1.000
0.41	0.958	0.998	1.000	1.000	1.000
0.42	0.965	0.999	1.000	1.000	1.000
0.43	0.970	0.999	1.000	1.000	1.000
0.44	0.975	0.999	1.000	1.000	1.000
0.45	0.979	1.000	1.000	1.000	1.000
0.46	0.983	1.000	1.000	1.000	1.000
0.47	0.986	1.000	1.000	1.000	1.000
0.48	0.989	1.000	1.000	1.000	1.000
0.49	0.991	1.000	1.000	1.000	1.000
0.50	0.993	1.000	1.000	1.000	1.000

[a] A protein is considered present when its probability exceeds 0.95. Probabilities are calculated from eq 1 and eq 2 by assuming each peptide has the same normalized Xcorr.

beta-casein (*Bos taurus*)

```
MKVLILACLV ALALARELEE LNVPGEIVES LSSSEESITR

 IEKFQS EEQQQTEDEL QDK                          S
 IEKFQS EEQQQTEDEL                              E
  EKFQS EEQQQTEDEL                              E
   KFQS EEQQQTEDEL QDK                          S
   KFQS EEQQQTEDEL QDKIHPF                      K
    FQS EEQQQTEDEL QDK                          T
    FQS EEQQQTEDEL QDKIHP                       E
INKKIEKFQS EEQQQTEDEL QDKIHPFAQT QSLVYPFPGP

IHNSLPQNIP PLTQTPVVVP PFLQPEVMGV SKVKEAMAPK

HKEMPFPKYP VEPFTESQSL TLTDVENLHL PLPLLQSWHM

QPHQPLPPTV MFPPQSVLSL SQSKVLPVPQ KAVPYPQRDM

PIQAFLLYQE PVLGPVRGPF PIIV
```

Figure 6. Overlapping peptide coverage identifies phosphorylation of serine-50 in β-casein. The complete amino acid sequence of bovine β-casein is shown. Seven overlapping peptides produced from multiple proteases (S, subtilisin; E, elastase; K, proteinase K; T, trypsin) all identify phosphorylation on S50.

modifications.[7] Overlapping peptides increase the probability that a particular modification locus is identified, reduce the ambiguity of identifications based on a single peptide, and provide multiple confirmations that increase the confidence that the detected modification is correct. Tandem mass spectra used for Figure 5 were researched to consider phosphorylations with +80 on STY. Figure 6 shows seven overlapping peptides generated by the multiple proteases (subtilisin, elastase, proteinase K, trypsin), each correctly identifying the known S50 phosphorylation of β-casein. For each peptide, when considered individually, the average statistical confidence that S50 is phosphorylated based on XCorr alone is <74.5% (range 57.9−94.2%). However, the combined data set (1) increased the confidence of a correct modification assignment to >99.9%, (2) enforces the statistical power of having multiple peptides identifying the same protein or modification site, and (3) demonstrated that the different proteases can each produce peptides that match correctly with SEQUEST-NORM.

This probability-based validation assumes that all incorrect database searches will have the same frequency distribution shown in Figure 5. This assumption is reasonable, assuming that future comparisons are made with the LCQ ion trap instrument under identical tandem mass spectrometry conditions. Yates et al. have shown that cross-correlations between tandem mass spectra acquired on different types of instruments (e.g., an ion trap and a triple quadrupole) result in scores different from tandem mass spectra acquired on the same type of instrument.[11] Furthermore, although the XCorr is independent of the database, the distribution of unmatched spectra may become more skewed toward high XCorr values if a larger database is searched because the probability of producing a random spectrum match increases with database size. A noticeable effect has not yet been observed using the current nonredundant protein database. However, if it occurs, a lookup table like Table 1 could be rapidly produced for each sequence database.

CONCLUSIONS

We have described a system for simplifying the interpretation of SEQUEST results. Although SEQUEST is a very powerful proteomics tool, validating SEQUEST output (particularly for peptides with no cleavage specificity) can be quite labor intensive. These problems were addressed by normalizing the XCorr values against the autocorrelation of the tandem mass spectrum against itself. This modified algorithm (SEQUEST-NORM) places all

results on a scale from 0.0 (no match) to 1.0 (perfect match). We conclude that (1) SEQUEST-NORM is capable of identifying peptides without any knowledge of peptide cleavage specificity and (2) the quality of protein identifications can be evaluated by assessing the probability a selected XCorr is correct or incorrect based on prior measurements. This probability-based validation of protein identifications can be easily automated and eliminates subjectivity in the evaluation of SEQUEST output.

ACKNOWLEDGMENT

The authors gratefully acknowledge financial support from National Institute of Health grants RR11823 (JRY), R33CA81665 (JRY), and F32DK59731 (MJM). We also appreciate the helpful comments of Hayes McDonald, Rovshan Sadygov, and John Venable.

NOTE ADDED AFTER ASAP POSTING

This article was inadvertently posted before final corrections were made. The corrected version was posted on September 18, 2002.

Received for review June 4, 2002. Accepted August 2, 2002.

AC025826T