research articles Journal of proteome research

Assessing Data Quality of Peptide Mass Spectra Obtained by Quadrupole Ion Trap Mass Spectrometry

Ming Xu[†], Lewis Y. Geer[*,†], Stephen H. Bryant[†], Jeri S. Roth[‡], Jeffrey A. Kowalak[‡], Dawn M. Maynard[‡], and Sanford P. Markey[‡]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland and National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland

Received August 24, 2004

An algorithm is introduced to assess spectral quality for peptide CID spectra acquired by a quadrupole ion trap mass spectrometer. The method employs a quadratic discriminant function calibrated with manually classified 'bad' and 'good' quality spectra, producing a single 'spectral quality' score. Many spectra examined that do not have significant matches are assessed to have good spectral quality, indicating that advances in search methods may yield substantial improvements in results.

Keywords: shotgun proteomics • mass spectral quality • peptide sequence • CID • quadrupole ion trap mass spectrometry

Introduction

Two-dimensional liquid chromatography coupled to quadrupole ion trap tandem mass spectrometry (ITMS)[1,2] provides an automated, high-throughput analytical method widely used to generate data for the automated assignment of peptide sequences. This technology generates large quantities of spectra that require searching against protein sequence libraries for peptide assignment. Sequence library search algorithms compare collision induced dissociation (CID) MS/MS spectra to the fragment ions predicted from a set of peptides under predefined rules and constraints. Matched peptides are scored and ranked, and top assignment(s) are reported. Besides spectral peak selection criteria, the characteristics and performance of commercial searching engines differ mainly in how to score and rank "candidate" peptides based upon matched fragment ions.[3−8] Peptide bonds are not chemically equivalent within protonated peptides, and proton retention varies between moieties after peptide bond cleavage. As a result, dissimilar intensities of fragment ions and/or incomplete fragment ion series are usually observed in CID of protonated peptides. Current sequence search engines are presently unable to incorporate fundamental mechanisms controlled by ion chemistry for peptide bond cleavage into peptide assignments. Thus, it is not unusual to find incorrect or ambiguous peptide matches to CID spectra, particularly when the spectra contain chemical or instrumental noise. Search engines may fail to assign high quality peptide CID spectra for several reasons: prominent fragmentation other than amide backbone, unanticipated post-translational modifications, or sequence-specific preferred fragmentations. A software tool that measures spectral quality is of value in reviewing and assessing the large data files acquired by ITMS, particularly for directing the attention of investigators toward high quality, unassigned spectra.

There are many literature reports about in-house or public domain packages for scoring matched peptide assignments *after* spectral searching but not prior to searching. Search engines employ some type of signal-to-noise assessment in peak selection and/or within their scoring. Mascot[5] utilizes protein sequence library dependent probability scores for candidate peptide ranking and statistically derives two acceptance thresholds: the identity score and the homology score. PeptideProphet[11] uses a statistical model to evaluate peptide assignment accuracy. A measurement of peptide CID mass spectral quality could provide an additional utility to direct more intelligent and productive search strategies. Highly scored peptide matches from noisy CID spectra may be biased to false positive assignments. Low scoring peptides matched from high quality CID spectra can be prone to false negative assignments. Keller et al. have suggested that peptide CID spectra can be filtered based upon empirical rules prior to database searching.[12] Distinguishing fragment ion signal from noise or interference is not always feasible prior to a library search for ITMS spectra. Manual inspection and sorting is too labor intensive, but chemists distinguish high vs low quality spectra easily by visual inspection. Peak rich, over-crowded peptide CID spectra are too 'noisy' for manual interpretation. Less crowded spectra with prominent peak intensity distributions suggestive of a sequence 'ladder' are more likely candidates for manual interpretation as good quality spectra. When parent ions fail to dissociate, spectra contain sparse information, having too few ions to permit interpretation. Here, we introduce a simple but robust prototype for scoring the spectral

* To whom correspondence should be addressed. Lewis Y. Geer, 38A/5S512, NCBI/NLM/NIH, Bethesda, MD 20894. Phone: (301) 435-5888. E-mail: lewisg@mail.nih.gov.

† National Center for Biotechnology Information, National Library of Medicine.

‡ National Institute of Mental Health.

10.1021/pr049844y CCC: $30.25
Published on Web 02/03/2005

quality of protonated peptide CID spectra obtained by ITMS. Spectra from samples of protein standard mixtures and yeast lysates were examined by comparing of automated search engine results to spectral quality scores. Peptide matches by commercial search engines are affected by spectral quality, and the spectral quality score is demonstrated to be useful in critically evaluating results, especially high quality spectra where further investigation is warranted.

Experimental Data

Protein standard cocktails were made of horse myoglobin (gi|2506462), bovine serum albumin (gi|1351902), and chicken lysozyme (gi|126608), all from Sigma-Aldrich Corp (St. Louis, MO), and were prepared at varied concentrations (May 2003). *S. cerevisiae* proteins were sequentially extracted from the original 5 g pellet using a modified procedure[2] to produce the following fractions: yeast soluble fraction (June 2002), light wash fraction (Sept. 2002), and heavy wash fraction (Sept. 2002). Each of the three fractions and protein standard cocktails was denatured, reduced, alkylated, and digested with endoproteinase Lys-C followed by trypsin.[13] A 20 μg aliquot of the digest was analyzed by an automated 2D-LC−MS/MS system. This system is composed of Shimadzu LC−VP series components connected directly to a ThermoFinnigan LCQ Classic ion trap mass spectrometer. This study used data from 'DTA' files, ASCII formatted files for reporting spectral data. DTA files are written by LCQ_DTA.exe utility from a raw data file acquired in centroid mode. The constraints for writing DTA files in this study were as follows: minimum number of peaks, 15; grouping tolerance, 1.4; intermediate scans, 1 or 0; and minimum scans per group, 1. The DTA files were searched against the NCBInr database or its subset using the search engine Mascot from Matrix Science. Unless otherwise indicated, Mascot searches used the NCBInr or yeast subset reference libraries with fixed carbamidomethyl modification of cysteine residues and variable modification for oxidation of methionine residues. MS/MS mass error tolerance was 0.8 Da using monoisotopic masses. Precursor mass tolerance was ± 2 Da. In Mascot, the score for a MS/MS match is based on the absolute probability (P) that the observed match between the experimental data and the database sequence is a random event. The reported 'Ions Score' is $-10\log10(P)$.[14] Mascot 'Identity Score' provides an acceptance threshold with false identification probability at a confidence level of 0.05. Peptide matches are accepted as correct if the peptide Ions Scores equal or exceed the Identity Score threshold.

DTA files were de-isotoped after normalization, where the de-isotoping window of 3.95 Da was scanned from the m/z lower boundary to the m/z upper boundary, and any m/z entry was discarded except those at maximum intensity. These re-constructed DTA files were used for both building a scoring function and for scoring spectral quality. The model for scoring spectral quality takes five parameters out of the re-constructed spectra. The training data consisted of 77 'good' spectra and 76 'bad' spectra, manually picked from multiply charged peptide spectra DTA files previously acquired from protein standards, procedural controls using the procedure described above, including about 20 extremely noisy spectral DTA files that helped to anchor the lower boundary of spectral quality scores at the value of 1.00. The upper boundary of the scores is anchored by single peak spectra at the value of 4.00. The size of the training set was determined to be over 10 times as many as the number of variants in the scoring model. The

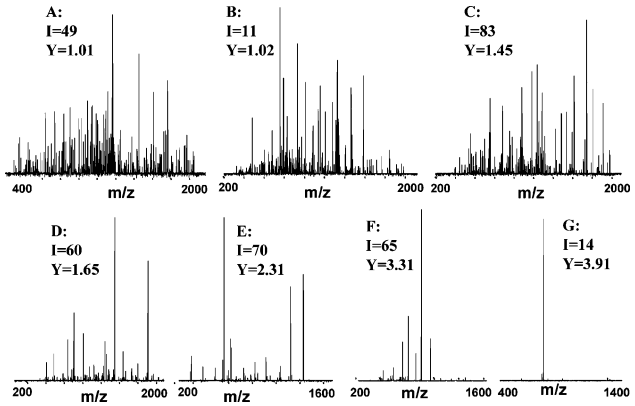

Figure 1. Yeast lysate spectra sorted by spectral quality score The scoring low boundary at $Y \approx 1.00$ is for extremely noisy spectra and the upper boundary at $Y \approx 4.00$ is for single peak spectra. Y is the spectral quality score and I is the Mascot Ions Score for the default search.

coefficients of the scoring function were calibrated using the commercial package S−Plus.[15] The calibration was done via Jackknife sampling by randomly taking 20% of the manually screened and labeled data as the calibration testing set and the remaining as the calibration training set. The process was repeated 20 times and the best result was reported. This methodology was chosen to maximally use manually validated data while avoiding overtraining as much as possible.

Results and Discussion

Spectral Quality Scoring and Distribution. The spatial distribution pattern of fragment ion peaks is differentiable between typically noisy vs good quality peptide CID spectra. The algorithm for scoring spectral quality utilizes such spatial distribution pattern of spectral peaks. By trial and error approach, we found the following quadratic discriminant function works better with the empirically selected parameters.

$$\boldsymbol{Y} = \mathbf{log}\,\{-(\boldsymbol{k}_o + \sum \boldsymbol{k}_i \boldsymbol{X}_i + \sum \boldsymbol{k}_{ii}\, \boldsymbol{X}_i\, \boldsymbol{X}_i)\}$$

$$\boldsymbol{X}_1 = \boldsymbol{C}_2/\boldsymbol{C}_1\ \%;\ \boldsymbol{X}_2 = \boldsymbol{C}_3/\boldsymbol{C}_1\ \%;\ \boldsymbol{X}_3 = \boldsymbol{C}_4;\ \boldsymbol{X}_4 = \boldsymbol{C}_5$$

where $\boldsymbol{C}_1$ is the number of peaks larger than a given peak intensity selection threshold. $\boldsymbol{C}_1$ is an adjustable parameter with a default value set at 5% of base peak intensity. $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ are the number of peaks larger than 3% TIC (total ion current) and 2% TIC, respectively. $\boldsymbol{C}_2$ is empirically accountable for mostly intense spectral peaks and $\boldsymbol{C}_3$ for less intense spectral peaks. $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ are scaled by $\boldsymbol{C}_1$ to make $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ for quantifying spectral peak intensity spatial distribution. $\boldsymbol{C}_4$ and $\boldsymbol{C}_5$ are the average peak distance along m/z for the peaks larger than 2% TIC and within 1.0∼1.5% TIC, respectively. These two parameters are empirically selected and directly used as the variables for quantifying spectral peak spatial distribution along the m/z dimension. $\boldsymbol{Y}$ is defined as the spectral quality score**.** The calibrated coefficients are k_o −11.105, k_1 0.2216, k_2 0.4229, k_3 0.011 20, k_4 0.024 44, k_{11} −0.1925, k_{22} −0.1552, k_{33} −0.000 275 5, k_{44} −0.000 147 0.[16] Spectra A−G in Figure 1 visually demonstrate spectral quality scoring. Spectrum A represents a typical noisy spectrum. Spectra are less noisy progressing from B−F. With further increments of the score beyond spectrum F, there are fewer and fewer peaks in a spectrum. "Single peak" spectra like G have scores approaching 4.0, the upper boundary of the range.

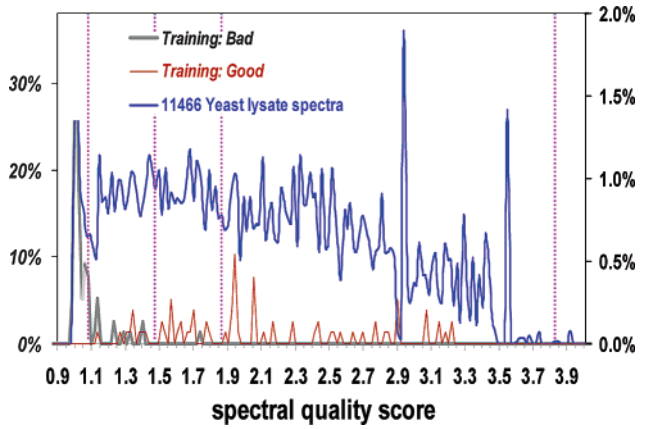

Figure 2. Spectral quality score histogram for the double and triple charged peptide spectra from yeast lysates and training data. Left vertical axis is for the training data and right for yeast lysates. The histogram is divided into five regions along *x*-axis for noise, less noisy, good, better and sparse spectral category regions from left to right.

The distribution of spectral quality scores is plotted in Figure 2 for 11 466 DTA files representing redundant doubly and triply charged peptides spectra from yeast lysates and the training data. Most of the noisy spectra in the training set and about 681 (~6%) of yeast lysate data score below 1.10. Spectra are judged to be not assignable if they have scores below 1.10. Between the score 1.10 and 1.50 is a boundary range for 'bad' and 'good' training spectra, with 2273 (~20%) of the yeast lysate spectra and a similar portion of the protein standard spectra scoring within this range. Above the score 1.50 are 'good' spectra in the training set, the majority of the protein standards, and about 74% of the yeast lysate spectra. Spectra scoring in this range are potentially interpretable ($3.8 > \boldsymbol{Y} \geq 1.5$). When the score exceeds 3.80, the spectra contain too few fragment ion peaks to assign peptide sequence information. None of the training set and protein standard spectra are found to score that high, and only 16 (<0.15%) spectra of the yeast lysate are in that range. The spectral quality score distribution is tabulated in Table 1 for the yeast lysate fractions obtained using the modified procedure[13] along with the training data. Under well-controlled situations, the spectral quality score distribution is fairly consistent from run to run, and peptide spectra categorized as 'good' and 'better' make up around 20% and 50% of the total, respectively. This categorization of the spectra quality of the yeast lysate DTA files indicates that the majority of peptide CID spectra are potentially interpretable, and that noisy or poor quality spectra may be well under 10% of the total in shotgun proteomics.

Effect of Spectral Quality Score on Peptide Matches by Search Engines. The effect of spectral quality score on searching the spectra in Figure 1 against the NCBInr database is shown in Table 2 using three commercial search engines, Mascot, Sonar, and Sequest with various search options. The noisy spectrum B, having a spectral quality score of 1.02, has dissimilar peptide matches when searched with monoisotopic and average masses at varying MS/MS mass error tolerance. Less noisy spectra such as spectrum C, using Mascot search with an increased MS/MS mass error tolerance of 1.5 Da, returns the same peptide match using either monoisotopic or average mass. Such a match seems reasonable because the error tolerance is about 1.5 times the difference in m/z between the average mass and isotopic mass of the peptide fragment ions. For the spectra scored as 'good' (spectrum *D*), the three search engines reported the same peptide hit with monoisotopic or average mass using varied fragment ion mass error windows. This suggests that a correct peptide match from a protonated peptide CID spectrum of low mass resolution and mass accuracy becomes less ambiguous when it is of high quality. Commercial search engines derive the same results with good quality spectra even though there may be variations in spectral peak selection and peptide match ranking among the search engines.

Unassigned Good Quality Spectra. The spectral quality score offers a software solution to screen for potential good quality peptide CID spectra deserving further investigation or manual interpretation. Among nonredundant multiply charged CID spectra of yeast lysates shown in Table 3, over 50% remained unassigned despite being categorized as good quality spectra. There are several reasons for lack of assignment for these spectra, such as protein modifications not considered during searches, mass error tolerance, incomplete peptide fragmentation, inadequate reference libraries, and limitations in current search algorithms and strategies.

Two spectra illustrate the value of examining high quality, unassigned mass spectra. Figure 3 shows a CID spectrum ($\boldsymbol{Y} = 2.10$) containing an obvious amino acid ladder '*LETDE*' or '*IETDE*'. The best peptide match by Mascot (Ions Score 31, Identity Score 51) suggests a peptide, *VLENTEIGDSIFDK,* found in phosphoglycerate kinase from *Saccharomyces cerevisiae* using the default search parameters. BLASTing[17] the sequencing tag '*LETDE*' finds multiple entries under *Saccharomyces cerevisiae* from the NCBInr. Allowing variable deamidation, the Mascot Ions Score increases to 85 (Identity Score 54). Alternatively, using the average mass or increasing the MS/MS m/z error tolerance to 1.0 Da increases the Mascot Ions Score to significance (88 or 74). A rational choice can be made by an investigator between allowing variable deamidation or increased m/z error tolerance based upon sample and instrument history. In a second example, peptide identification by CID can suffer from preference for amide fragmentation at the proline N-terminus[18] causing less abundant fragment ions from peptide bond cleavage elsewhere along the backbone. The default search assignment of the spectrum in Figure 4 indicated an insignificant peptide matches (Ions Score 26, Identity Score 49) when the NCBInr was searched. Low intensity ions below 10% of the base peak may require a more lenient mass error tolerance for effective assignment when compared to the major ions in a good quality spectrum. The score $\boldsymbol{Y} = 2.69$ indicates that it is likely to be an interpretable quality spectrum. Increasing the MS/MS mass tolerance to 1.2 Da raises the Ions Score to 53 for the peptide assignment shown in Figure 4 in which small peaks are reasonably assigned.

Peptide Search Assignments Versus Spectral Quality Score. Figure 5 plots the histogram for the percentage of the significant peptide hits over the spectral quality scores for nonredundant multiply charged peptide spectra for data from yeast lysate analyses. For these DTA files, charge redundant spectra are eliminated by replacing a DTA file in a doubly charge state with one in a triply charge state whenever the latter has a higher peptide value (Ions Score minus Identity Score) using the default search. The percentage of significant peptide hits starting under 5% from noisy spectral region ($\mathbf{Y} < 1.1$) goes up to over 40% after reaching the better spectral quality region ($\mathbf{Y} > 1.9$). With the spectral quality score beyond 2.9, the percent-

Table 1. Spectral Quality Score Distribution for Multiply Charged Peptide Spectra of Yeast Lysate[a]

spectral quality score Y	spectral quality category	manually validated data		soluble (Jun. 2002)		light washed (Sep. 2002)		heavy washed (Sep. 2002)		all fractions	
		'bad' (%)	'good' (%)	*n*	% total	*n*	% total	*n*	% total	*n*	% total
all		76(100)	77(100)	5096	100	3796	100	2574	100	11466	100
< 1.10	noisy	84.2	0	236	4.63	270	7.11	175	6.80	681	5.94
≥ 1.10 < 1.50	less noisy	14.5	13.0	1060	20.8	722	19.0	491	19.1	2273	19.8
≥ 1.50 < 1.90	good	1.3	24.7	1073	21.1	654	17.2	486	18.9	2213	19.3
≥ 1.90 < 3.80	better	0	62.3	2721	54.4	2144	56.5	1418	55.1	6283	54.8
≥ 3.80	sparse	0	0	6	0.12	6	0.16	4	0.16	16	0.14

[a] *n*: the number of DTA files; 'bad' and good' are the categories of manually validated data for calibrating the coefficients of the score function

Table 2. Effect of Spectral Quality on Searching against the NCBInr[a]

spectral samples	search engine	search option	top hit score	threshold	top-scored peptide match
spectrum B in Figure 1 $Y = 1.02$	Mascot	mono. 0.8 Da	11	> 50	NNFPIIKPDSFAGLRALK
		1.2 Da	22		VVIAWSVEASTEIDVAAIK
		1.5 Da	27		AADPSQGEMSADAAAGAPLPR
		ave. 0.8 Da	46		WAGNANELNAAYAADGYAR
		1.2 Da	72		
		1.5 Da	70		
	Sonar	mono. 0.8 Da	1.2	< 0.1	GSALGIGAVFGIAFLIGPII
		1.2 Da	1.2		DTLLGEEELPLTSLLPEL
		1.5 Da	0.38		MQQLQQSEAAAGDRLILK
	Sequest	mono. 1.5 Da	2.23	> 2.0	NSLGICVIDATSGPNTPRK
		ave. 1.5 Da	3.99		WAGNANELNAAYAADGYAR
spectrum C in Figure 1 $Y = 1.45$	Mascot	mono. 0.8 Da	83	> 49	VINDIFGIEEGLMTTVHSITATQK
		1.2 Da	89		
		1.5 Da	88		
		ave. 0.8 Da	15		TGVAVTGVAGIMMLLALAGISLNLWK
		1.2 Da	22		TGVAVTGVAGIMMLLALAGISLNLWK
		1.5 Da	30		VINDIFGIEEGLMTTVHSITATQK
	Sonar	mono. 0.8 Da	1.4×10^{-3}	< 0.1	VINDIFGIEEGLMTTVHSITATQK
		1.2 Da	1.9×10^{-3}		
	Sequest	mono. 1.5 Da	2.52	> 2.0	MAAWVKGGAADVDAAVEAAADLLAASR
		ave. 1.5 Da	2.51		NLQASSTGLQWYYVYDHSGEAVK
spectrum D in Figure 1 $Y = 1.65$	Mascot	mono. 0.8 Da	60	> 48	FEQASESEPTTVSYEIAGNSPNAER
		1.2 Da	75		
		1.5 Da	88		
		ave. 0.8 Da	57		FEQASESEPTTVSYEIAGNSPNAER
		1.2 Da	66		
		1.5 Da	79		
	Sonar	mono. 0.8Da	1.0×10^{-3}	< 0.1	FEQASESEPTTVSYEIAGNSPNAER
		1.2 Da	1.3×10^{-3}		
	Sequest	mono. 1.5 Da	4.59	> 2.0	FEQASESEPTTVSYEIAGNSPNAER
		ave. 1.5 Da	3.89		FEQASESEPTTVSYEIAGNSPNAER

[a] ***Y***: spectral quality score; mono: monoisotopic mass; ave.: average mass.

age of the significant peptide hits dramatically decreases as the spectral quality score further increases. This implies that peptide assignment is a function of spectral quality. Table 3 lists significant peptide hits for spectral quality categories, where varying peptide hit thresholds, ***R***, has little effect on peptide hit distribution over spectral quality categories. About 3.8% spectra of the noisy spectra ($\boldsymbol{Y} < 1.10$) have Mascot Ions Score exceeding Identity Score using the default search conditions. This value is below the false identification rate statistically derived in Mascot search engine. Over 80% of total significant peptide hits come from the good and better quality spectra, but still more than half of those good and better quality spectra fail to generate significant peptide hits. These spectra deserve further investigation in order to decrease false negative identifications. Figure 6 illustrates that the number of incorrect peptide hits above the threshold increases with lower spectral quality scores for protein standards. A protein mis-assignment may result for any identified peptide originating from protein(s) unexpected in the sample; and a correct protein assignment may be likely for any matched peptide from the proteins expected in the sample. Under default search conditions (top plot) protein mis-assignments above the threshold are 0% for noisy spectra, about 33% for less noisy spectra, 8% for good spectra, and 0% for better spectra. As shown in the lower plot of Figure 6, allowing more modifications in a search produces more peptide matches to the expected target proteins. However, expanding possible modifications introduces more arbitrary matches in peptide identification, so protein mis-assignments above threshold become 50% for 'noisy' spectra, about 55% for 'less noisy' spectra, 36% for 'good' spectra, and

Table 3. Distribution of Peptide Hit Rate at Varied Thresholds, R^a

quality score *Y*	spectral category	no. of spectra	$R \geq 1.00$ hit rate	$R \geq 1.00$ % total	$R \geq 0.85$ hit rate	$R \geq 0.85$ % total	$R \geq 0.70$ hit rate	$R \geq 0.70$ % total
all		5873	23.2%		30.7%		38.5%	
		(100%)	(1365)	100	(1800)	100	(2258)	100
< 1.10	noisy	371	3.77%		6.74%		10.2%	
		(6.31%)	(14)	1.03	(25)	1.39	(38)	1.68
≥1.10∼	less noisy	1218	18.0%		21.9%		27.6%	
< 1.50		(20.7%)	(219)	16.0	(267)	14.8	(336)	14.9
≥1.50∼	good	1156	26.0%		32.0%		37.5%	
< 1.90		(19.7%)	(300)	22.0	(370)	20.6	(433)	19.2
≥1.90∼	better	3121	26.6%		36.4%		43.4%	
< 3.80		(53.1%)	(830)	60.8	(1135)	63.0	(1448)	64.1
≥3.80	sparse	7	28.6%		42.9%		42.9%	
		(0.12%)	(2)	0.15	(3)	0.17	(3)	0.13

[a] ***R***: Mascot Ions Score divided by Ion Identity score; **hit rate**: the number of spectra at or above threshold R over total spectra; %**total**: percentage of hits over the total number of hits.

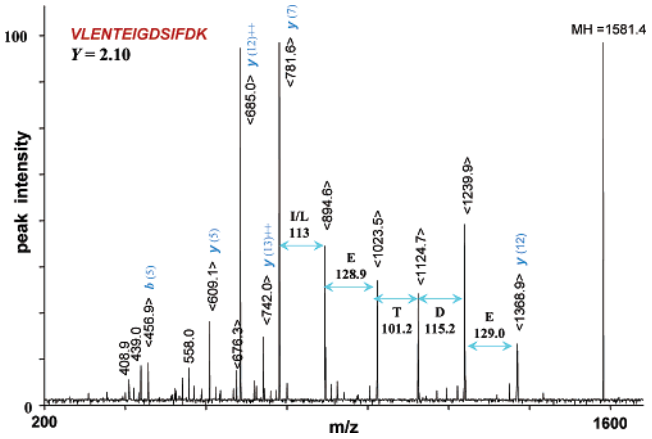

Figure 3. Peptide(2+) CID spectrum from Yeast lysate, *MH* is from first entry in DTA files for precursor peptide mass. *Y* is the spectral quality score

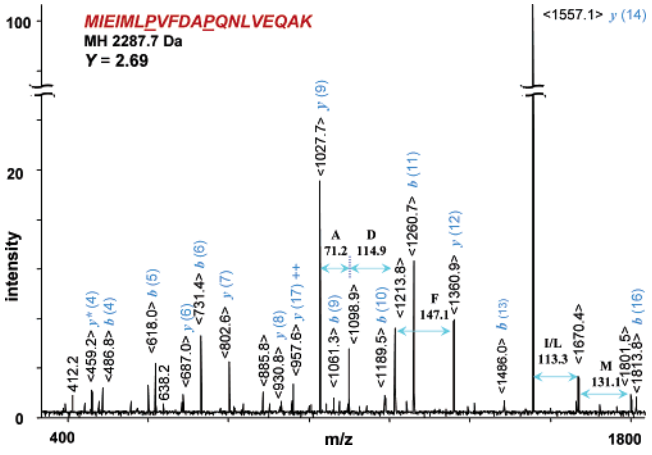

Figure 4. Peptide(2+) CID spectrum containing prolines from Yeast lysate, *MH* is from first entry in DTA files for precursor peptide mass. *Y* is the spectral quality score

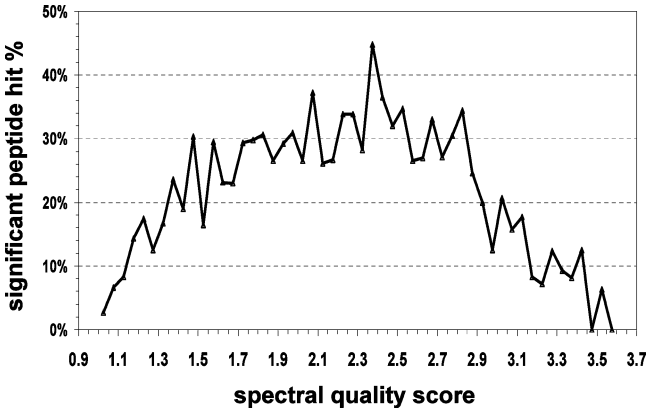

Figure 5. Histogram for the percentage of significant peptide hits (Mascot Ions Score ≥ Identity Score) versus the spectral quality score for 4008 double and 1866 triply charged peptide spectra from yeast lysates without charge redundant DTA files.

9% for 'better' spectra. Protein mis-assignments above threshold drop significantly as spectral quality score increases, indicating that peptide identification by Mascot is a function of spectral quality, as expected.

In addition to its correlation with peptide assignments by Mascot, spectral quality also impacts thresholds for peptide matches, this can be graphically illustrated by subjectively adjusting peptide assignment thresholds away from Mascot thresholds (*y* axis at 0) inversely proportionally to spectral quality score for alleviating false identifications. In the top plot under the default search, correct protein assignments are enhanced without adding any protein mis-assignment by purposely drawing the horizontal red lines further under Mascot threshold line (*y* axis at 0) as spectral quality score goes beyond 1.9. When the horizontal red lines are arbitrarily drawn above the Mascot threshold line (*y* axis at 0) for 'good' spectra (scoring between 1.5∼1.9) and 'less noisy' spectral (scoring between 1.1∼1.5) regions, false protein assignment rates drop to 0% from 8% for the 'good' spectra and to 14% from 33% for 'less noisy' spectra. Even with more variable modifications allowed in the bottom plot, the threshold adjustments used in the upper panel still apply with a significant increase in correct protein assignment with a single mis-assignment for the 'better' spectra scored at or above 1.9. An arbitrary downward red line above the Mascot threshold (*y* axis at 0) can even be drawn from the less noisy spectra to the good spectra in order to separate correct protein assignments from the region where correct and incorrect protein assignments overlap. Thus, peptide match thresholds in peptide identification may be adjusted according to spectral quality measurement.

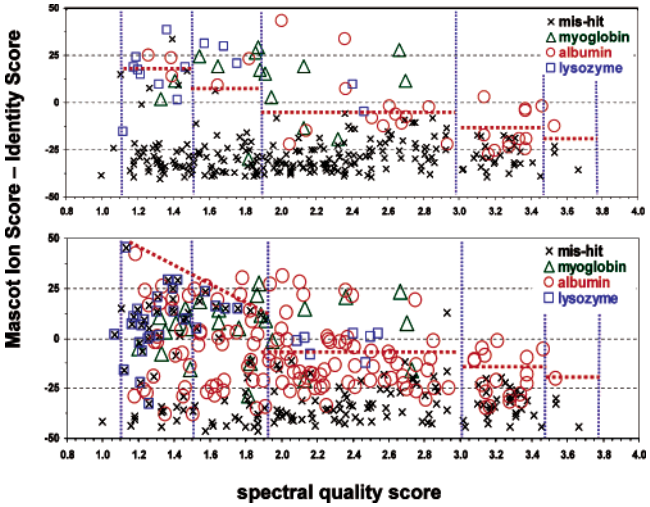

Figure 6. Mascot Ions score subtracted from Identity Score versus the spectral quality score for 295 double charged DTA files from 100 fmol protein standard cocktails. Mascot search options against the NCBInr are 1 missing tryptic cleavage, monoisotopic mass with MS/MS error tolerance 0.8 Da along with fixed modification with carbamidomethyl(C) and varied modification with oxidation (M) for the top plot and varied modification of carbamidomethyl(C), deamidation(NQ), oxidation (M) and phosphorylation (ST) for the bottom plot. Here, a mis-hit stands for any matched peptide not from the proteins expected in the sample.

Conclusions

A software prototype for scoring spectral quality employs a quadratic discriminant function calibrated from manually selected 'good' and 'bad' spectra. This score provides a simple and robust tool for a quantitative measurement of spectral data quality. The score was used to categorize protein standards, training data and yeast lysate spectra into noisy, less noisy, good and better spectra. It is evident that Mascot peptide matches and peptide hit thresholds are a function of spectral quality measurement when searching protonated peptide ITMS CID spectra against protein sequence libraries.

The spectral quality score is useful in locating false positive and false negative prone peptide identifications, especially when identifying good quality spectra warranting further investigation. For the yeast lysate data, more than 50% of nonredundant multiply charged DTA files were scored as good quality spectra but failed to generate significant peptide matches when searching against the NCBInr by the search engine Mascot. This indicates that improvement of search methods may increase the yield of significant peptide identifications. For example, lenient settings of the MS/MS m/z error and expanded modification options may be warranted for assigning peptide sequences to good quality spectra. In addition, protein standard data indicates that false peptide identification can be decreased by adjusting the acceptance thresholds inversely with spectral quality scores. The tool is shown to be useful in assessing data quality for shotgun proteomics. The scoring function can become more generic and better tuned by the addition of more training data.

References

(1) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvic, B. M.; Yates, J. R., III. *Nat. Biotechnol.* **1999**, *17*, 676−682.
(2) Washburn, M. P.; Wolters, D.; Yates, J. R., III. *Nat. Biotechnol.* **2001**, *19*, 242−247.
(3) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(4) Sadygov, R. G.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 3792−3798.
(5) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S.*Electrophoresis* **1999**, *20*, 3551−3567.
(6) Zhang, W.; Chait, B. T. *Anal. Chem.* **2000**, *72*, 2482−2489.
(7) Field, H. I.; Fenyo, D.; Beavis, R. C. *Proteomics* **2002**, *2*, 36−47.
(8) Fenyo, D.; Beavis, R. C. *Anal. Chem.* **2003**, *75*, 798−774.
(9) Zhang, N.; Abersold, R.; Schwikowski, B. *Proteomics* **2002**, *2*, 1406−1412.
(10) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. *Genome Res.* **2001**, *11*, 290−299.
(11) Moore, R. E.; Young, M. K.; Lee, T. D. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 422−426.
(12) Keller, A.; Nesvizhaskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383−5392.
(13) Maynard, D. M.; Masuda, J.; Yang, X.; Kowalak, J. A.; Markey, S. P. *J. Chromatogr B* **2004**, *810*, 69−76.
(14) http://www.matrixscience.com.
(15) S−PLUS 2000 Manaul, MathSoft, Inc., Seattle, WA.
(16) Xu, M.; Geer, L. Y.; Ross, J. S.; Kowalak, J. A.; Maynard, D. M.; Markey, S. P.; Bryant, S. H. The 51th ASMS Conference on Mass Spectrometry and Allied Topics, Montreal, Canada, June 8−12, 2003.
(17) Altschul, S. F.; Gish, W.; Millers, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−410.
(18) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates, J. R., III. *Anal. Chem.* **2003**, *75*, 1155−1163.

PR049844Y