# Filtering out MS/MS spectra of insufficient quality before database searching

John Morey, Iain Rogers, Clark Chen

Bioinformatics Solutions Inc.

## Introduction

In studying proteins using liquid chromatography coupled tandem mass spectrometry (LC-MS/MS), researchers are often faced with very large data sets. Since each data set may contain thousands of spectra, a manual inspection of each one becomes impossible. Confounding the problem, electrical noise, poor detection and contaminants scanned by the MS mean that only a small portion of these data are quality MS/MS spectra representing peptides.

The following presents a method of filtering out the poor quality spectra prior to de novo sequencing or database searching for protein identification. Database search engines and de novo sequencing tools are adequate in discarding the bad spectra; nevertheless, false positives abound, and plenty of time is wasted analyzing nothing.

## Methods

34,578 MS/MS spectra, acquired from an LCQ ion-trap mass spectrometer, were used. Of these, 2901 spectra were known to be of good quality[2]. Through an iterative process, the characteristics of goodness and badness evolved and the best determinants of spectral quality were isolated. An estimation of the performance of these seven determinants in estimating spectrum quality was conducted.

## Results

The best determinants of spectrum quality are described as follows[1]:

(1) s/n – the signal-to-noise ratio calculated over the MS/MS spectrum

(2) numPeaks – The number of peaks after centroiding, removing noise, and de-isotoping

(3) SPI – The sum of all peak intensities

(4) maxTag – The length of the longest continuous sequence tag that can be assigned to a spectrum by a simple computation that matches the masses of single residues to the mass differences between peaks in the spectrum

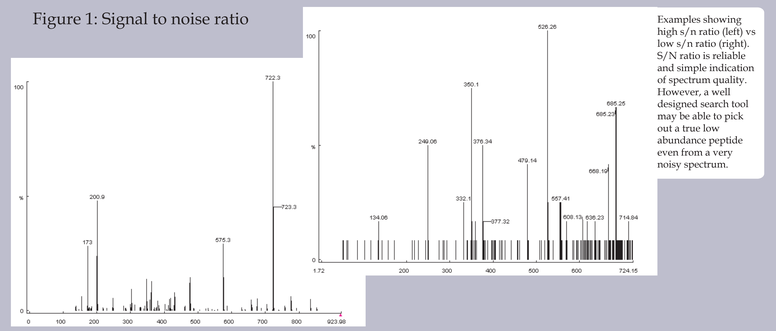
(5) numTags – The number of unique sequence tags, longer than two residues, that can be assigned to a spectrum

(6) COM – The total normalized intensity of pairs of peaks with m/z values summing to the mass of the parent ion. This value is calculated for each of parent ion charge states +1, +2, and +3, and the largest value is used.

$$\Sigma\{ \mathrm{NormI}(x) + \mathrm{NormI}(y) \mid m(x) + m(y) = \mathrm{mparent} \}$$

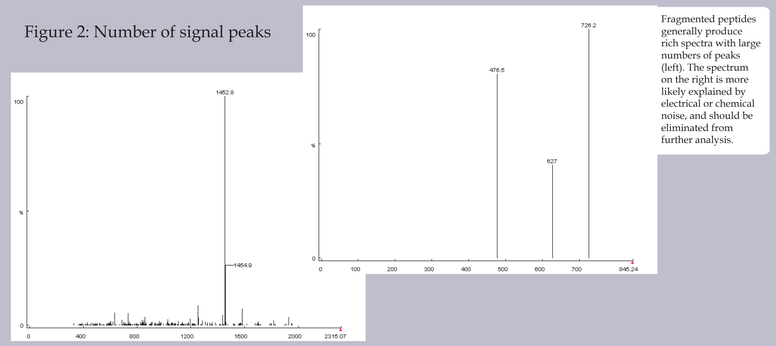
*The authors would like to thank Bin Ma, Virginia Yang, Gilles Lajoie and Weiming Zhang for their invaluable advise.*

Figure 1: Signal to noise ratio



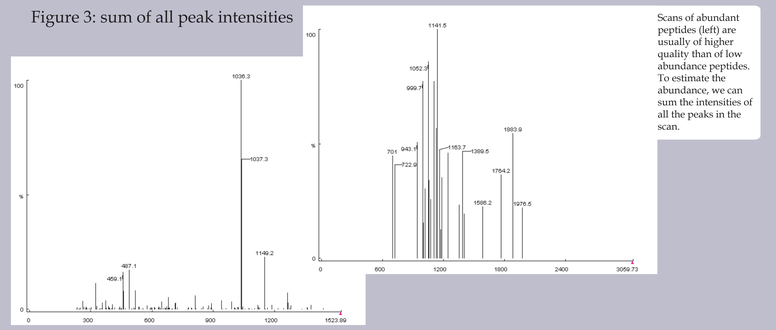

Examples showing high s/n ratio (left) vs low s/n ratio (right). S/N ratio is reliable and simple indication of spectrum quality. However, a well designed search tool may be able to pick out a true low abundance peptide even from a very noisy spectrum.

Figure 2: Number of signal peaks

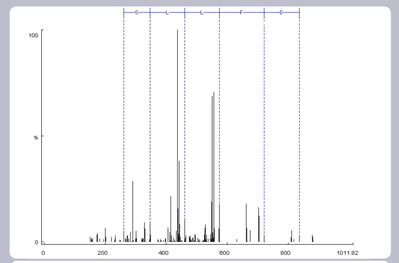



Fragmented peptides generally produce rich spectra with large numbers of peaks (left). The spectrum on the right is more likely explained by electrical or chemical noise, and should be eliminated from further analysis.

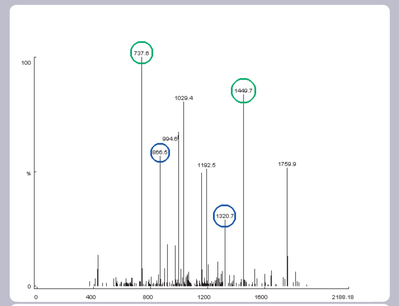
Figure 3: sum of all peak intensities



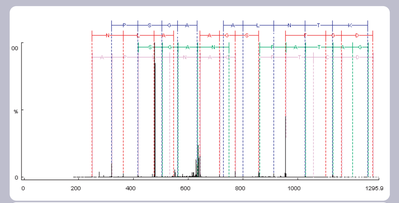

Scans of abundant peptides (left) are usually of higher quality than of low abundance peptides. To estimate the abundance, we can sum the intensities of all the peaks in the scan.




If a spectrum contains a series of peaks that differ in mass by the mass of an amino acid, then it is probable that the spectrum is of a peptide. Longer tags are less likely by random chance, and as such are a good diagnostic.




It is probable that two ions whose m/z values sum to the m/z of the precursor's m/z (2X or 3X the precursor) are an ion pair (y5 and b5 for example). The presence of several ion pairs would indicate a quality spectrum.




If a spectrum contains a number of possible short peptide sequence tags, then it is certainly worthy of further analysis.

If a threshold value for any of the determinants is chosen, then it can be used as a simple decision criterion in finding good quality spectra. Threshold values for each of these determinants were chosen empirically, based on performance in filtering the known dataset. The best threshold for each determinant is shown in Table 1, along with the resulting performance.

| Decision Criterion | insufficient spectra correctly removed (out of 31677) | | Good quality spectra removed (out of 2901) | | Overall accuracy |
|---|---|---|---|---|---|
| maxTag ≤ 4 | 9862 | 31.13% | 6 | 0.21% | 99.9392% |
| maxTag ≤ 6<br>parent ion charge ≠1 | 12201 | 38.25% | 42 | 1.45% | 99.6569% |
| numPeaks <40 | 9772 | 30.85% | 15 | 0.52% | 99.8467% |
| TIC < 600,000 | 10909 | 34.44% | 25 | 0.66% | 99.7714% |
| s/n < 0.5 | 13839 | 43.69% | 75 | 2.5% | 99.4610% |
| COM < 40 | 10022 | 31.64% | 50 | 1.72% | 99.5036% |

In addition, the efficacy of a unique combination of the above determinants was tested. Individual threshold values were more tolerant, but three conditions would have to be met: numTags ≤ 12, maxTag ≤ 9 and numPeaks < 70

| Decision Criterion | insufficient spectra correctly removed (out of 31677) | | Good quality spectra removed (out of 2901) | | Overall accuracy |
|---|---|---|---|---|---|
| numTags ≤ 12<br>maxTag ≤ 9<br>numPeaks < 70 | 10126 | 31.97% | 16 | 0.55% | 99.8422% |

All of the above determinants and threshold values were used to build an algorithm to remove poor quality MS/MS spectra from a given dataset. This algorithm was able to remove 13895/31677 or 43.865% of the poor quality spectra from the known data set while preserving 2847/2901 or 98.13% of the good quality spectra. In total the data set was reduced to 59.659% of its original size. The algorithm had 99.61% accuracy in finding spectra of insufficient quality.

## Conclusions

Removing spectra of insufficient quality will result in faster protein identification, with fewer false positive. Accuracy is increased without sacrificing sensitivity. This will be of particular use to users of ion trap instruments. The filtering process described herein is presented in the PEAKS suite of peptide MS/MS analysis software tools.

## References

1. Jaitly, D. Page Belanger, R., Faubert, D., Thibault, P., Kearney, P., MSMS Peak Identification and its Applications, Caprion Pharmaceuticals Inc, Montreal QC.

2. Keller, A., Purvine, S., Nesvizhkii, A. I., Stolyar, S., Goodlett, D. R., Kolker, E., Experimental Protein Mixture for Validating Tandem Mass Spectral Analysis (Omics, Volume 6, Number 2, 2002).