Spectrum Quality Assessment in Mass Spectrometry Proteomics

Rheanna Mainzer
Supervised by Dr. Luke Prendergast
La Trobe University

1. Background

An important research problem in *mass spectrometry* is in the identification of high quality spectra which can lead to interesting biological insights regarding peptide sequences. Mass spectrometry (MS) is a method that is used to find what proteins and peptides are in a biological sample. Mass spectrometry is very important to *proteomics* research, as discussed in the research paper by Han, X 2008. Proteomics is the study of proteins, particularly their structure and function. A mass spectrometer is a machine that is used in MS, which measures the mass to charge ratio of charged ions. The output given by the mass spectrometer is called a *spectrum.* A spectrum (or spectra for plural) is used to identify what proteins are in the sample. Currently there is no definitive way to assess the quality of the spectrum, however, Nesvizhskii et al (2006) has suggested a spectrum quality score to detect quality spectra. This score is designed for tandem MS (known as MS/MS), which is when peptides in sample have been further fragmented and the mass to charge ratios of the fragment ions measured. If we can accurately and quickly access the quality of the spectrum, we can look deeper into the good spectra to possibly identify more proteins, or discard the bad spectra without wasting time trying to identify something that is not there. This project has focused on developing R code that can report and combine several numerical features, as suggested by Nesvizhskii et al (2006), to detect quality spectra.

The mass spectrometry process can be described in a series of steps. Firstly, a sample of proteins is digested into peptides. Second, the peptides are separated based on a particular property so they are simplified when going into the mass spectrometer. The mass spectrometer then performs its analysis, separating the peptides according to their mass/charge ratio. Patterns in the mass spectrum can then be used to identify what peptides are in the sample.

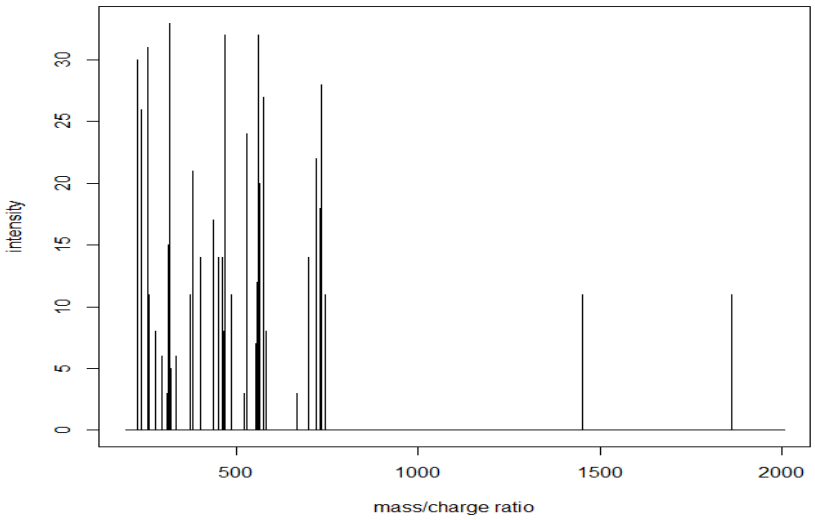

Figure 1.2 MS/MS spectrum

This is a typical spectrum from MS/MS data, and the spectrum that I focused on at the beginning of the project. This spectrum contains fragments from one of the peptides of Bovine Serum Albumin, which is a protein. Most of the peaks have a mass to charge ratio of between 0 and 800, with varying intensities.

2. The function

The Nesvizhskii paper discussed 8 different descriptive features that could be constructed from a given spectrum which would be used to create a spectrum quality score (SQS). These features aim to measure the quality of the MS/MS spectra. They characterize the overall distribution of peaks in the spectrum, as well as detect certain patterns expected to be present in high quality spectra. The first three measures are: number of peaks square root transformed, arithmetic mean of the peak intensities log transformed and standard deviation of the peak intensities log transformed. These measures are very simple to calculate using R and are to do with the size of the peaks themselves. The next two measures – smallest m/z range containing 95% of the total peak intensity, and smallest m/z range containing 50% of the total peak intensity – were much harder to calculate. The red lines in Figure 2.1 below indicate the type of interval that these measures describe. For this spectrum, we found that the smallest m/z range containing 95% of the total peak intensity is 516.286 and the smallest m/z range containing 50% of the total peak intensity is 242.1208. Measure 6 is the total ion current per m/z, which we could not calculate because we were not given this information in the dataset.

Postal Address: Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

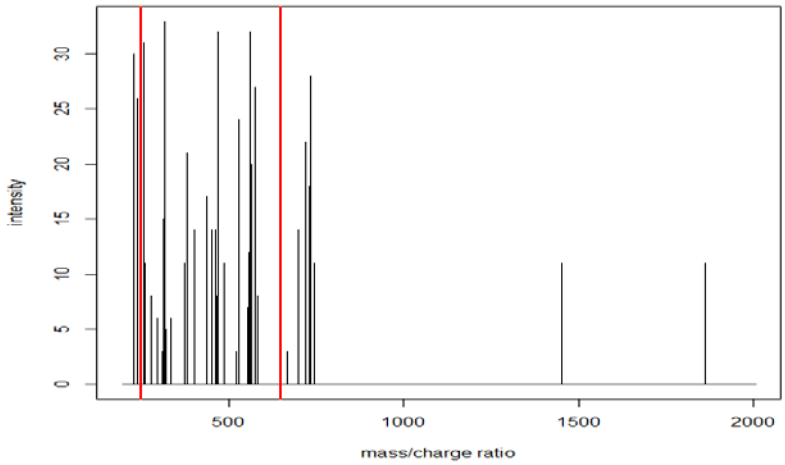

Figure 2.1 MS/MS spectrum indicating m/z interval.

The final two features are both concerned with the space between the peaks. Feature seven is the standard deviation of the consecutive m/z gaps between all peaks, log transformed (4.13 for this spectrum). Feature eight is the average number of peaks within a 2-Da interval around any peak. We found that, on average, there are 4.63 peaks within a 2-Da interval around any peak. After the code was written for each descriptive feature, we constructed a generalised function to give all the features for any spectrum.

Once we managed to get the function up and running in R, it was time to generalise it and run it on more spectra. This presented a challenge in itself because some of the descriptive measures were quite time consuming to calculate. This first spectrum that we used to write the function was relatively small in relation to the number of peaks it contains, but there were 174 MS/MS spectra in this sample, some spectra with hundreds of peaks and others with thousands. I settled on looking only at MS/MS spectra with less than 1000 peaks when performing my analysis, to overcome this challenge.

3. De Novo Sequencing

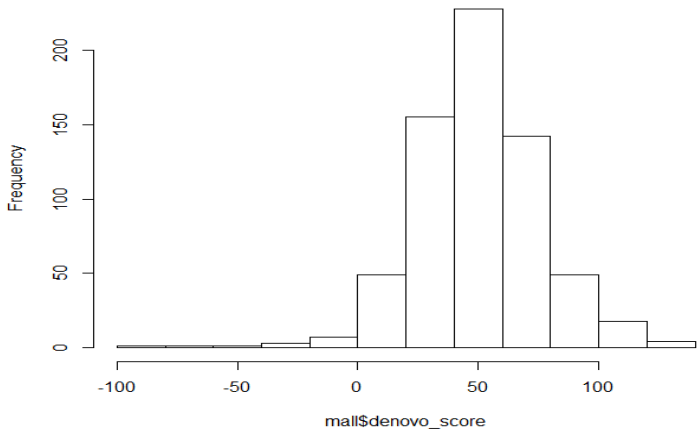

Figure 3.1 Histogram of De Novo scores for a MS/MS dataset.

Postal Address: Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

De Novo sequencing and database searching are the two main identification methods used in mass spectrometry. *De novo* sequencing derives the peptide sequence directly by using information from the MS/MS spectrum, whereas a database search queries a sequence database for the best peptide to explain the peaks in the MS/MS spectrum. The database only has known sequences in it so the search is limited to these sequences. Also, modifications can take place during the MS process and there can be a high level of noise in the spectrum. Because of this, both methods can wrongly identify proteins or not identify them at all. Further details on each procedure can be found in the article by Jing, Z 2011. On this project, we collaborated with Dr. Ira Cooke, a research fellow from La Trobe's school of molecular sciences. Dr. Cooke provided the data sets for the research. As we wanted to find out whether our function could accurately differentiate between a "good" spectrum and a "bad" spectrum, we first needed to find a way to assess the spectrum so we could place it in one of these two categories. Ira sent us a data set to work with where the spectrum had each been given a De Novo score. In Figure 3.1, a score below 0 (the left hand tail of the histogram) indicated the spectrum was "bad", i.e. it was noisy or modifications had occurred that left a number of peptides unidentified. If my function works accurately, there should be a noticable difference in the descriptive measures (as discussed in section 2) between a "good" spectrum and a "bad" spectrum.

4. Analysis

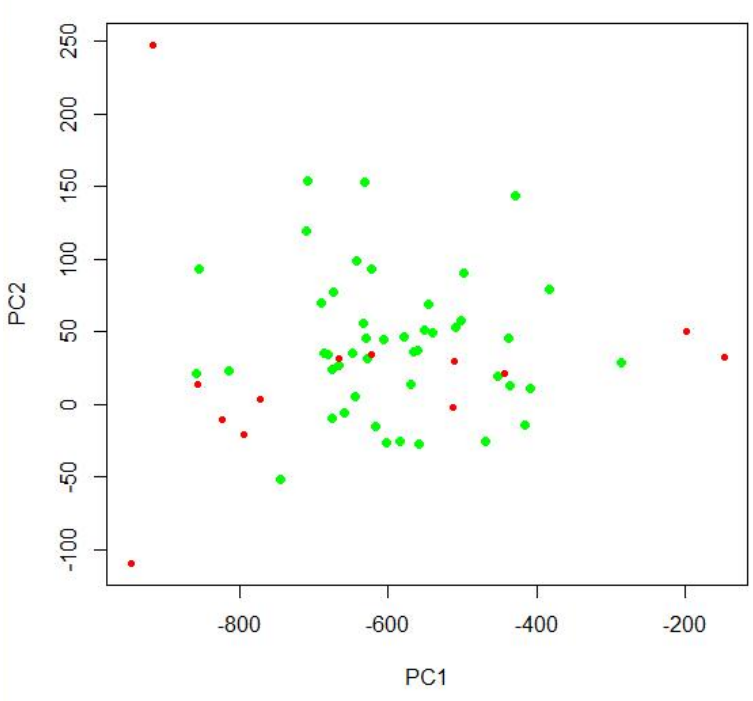

Figure 4.1 Plot of the first two principle components.

Postal Address: Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

We used *principle component analysis* (PCA) to see if there was any sort of pattern in the data. PCA is a procedure that seeks a small number of linear combinations of the variables while still retaining most of the information. What we would like to see in the PCA graph (see Figure 4.1. where we have plotted the first two sample PCA linear combinations of variables) is the good and bad data seperated into two different groups. We used a sample of 59 spectrum in our analysis, which were clearly defined as either "good" or "bad" by the value of their de novo score. We did not observe separation between the two groups in our principle component analysis so to further assess the data, we performed a *linear discriminant analysis* (LDA) which was also used by Nesvizhskii et al (2006). LDA is another classification tool that seeks an informative linear combination of the varialbes. Unlike PCA, LDA uses knowledge of the true classification of the data. Only one linear combination is available though. Our LDA was much more successful.

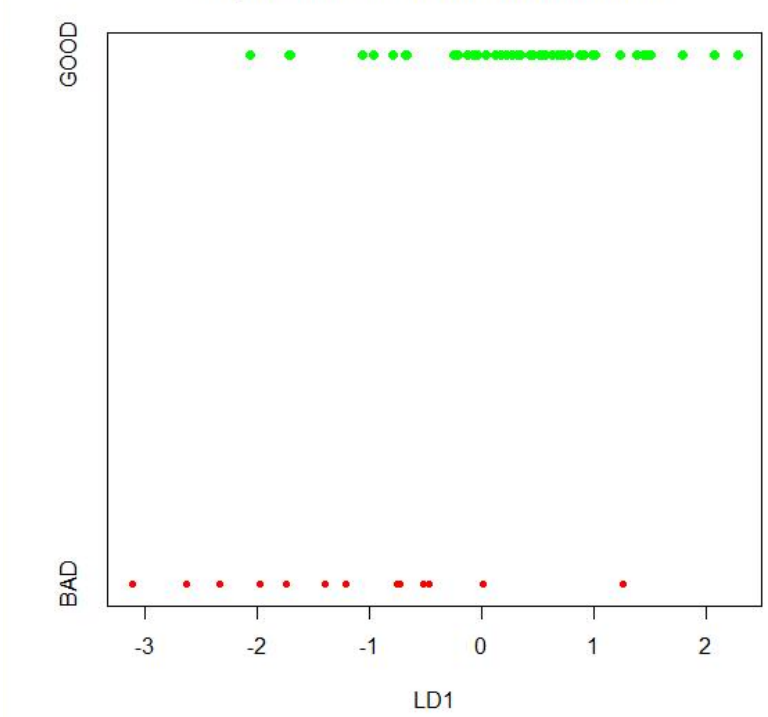

Figure 4.2 Plot of the first linear discriminant.

From Figure 4.2 we can see that the good data tend to group together, and have a higher first linear discriminant value than the bad data. When using the predicted results from the LDA we found that the correct fit rate for the "good" data was very high (93.48%), and the overall correct fit rate for the data was also high (83.05%). Even though the correct fit rate for the "bad" data was 46.15% which is much lower, this result isn't too detrimental. This is because we are much more interested in correctly fitting the good data, as these are the spectrum from which we should be able to identify peptides more easily.

5. Conclusion

Postal Address: Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au

The aim of this project was to develop R code for the spectrum quality assessment method for tandem mass spectrometry as introduced by Nesvizhskii et al (2006). The main focus of the project was writing code in R for the descriptive measures introduced in the paper, and then combining this code into a function that can be applied to MS/MS data. Over time, and with enough data collected, it could be possible to rely on a similar function based on descriptive measures to accurately assess the quality of the MS/MS data. It could also be used as a tool to reveal if there are more proteins in a sample that have not been detected by other methods, or to indicate which spectrum could contain more peptides than already identified. Dr. Ira Cooke has already indicated he may use this function, and record the results each time he uses tandem mass spectrometry to analyse a sample.

6. Acknowledgements

Thanks goes to AMSI for providing the scholarship and giving me the opportunity to research something new and interesting, the CSIRO for hosting the big day in, Dr. Ira Cooke for sharing his wealth of knowledge on mass spectrometry and sending us data to work with, and to my supervisor Dr. Luke Prendergast for his help and support throughout my project. I found the whole experience very valuable and enjoyed taking my first steps into research. I would recommend it to any student interested in pursuing mathematics and statistics in the future.

References

Han, X, Aslanian, A, Yates, JR 2008, 'Mass spectrometry for proteomics', *Current Opinion in Chemical Biology,* vol. 12, no. 5, pp. 483 – 490.

Jing, Z, Xin, L, Shan, B, Chen, W, Xie, M, Yuen, D, Zhang, W, Zhang, Z, Lajoie, G, Ma, B 2011, 'PEAKS DB: De Nove Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification', *Molecular & Cellular Proteomics,* vol. 11, not. 4, pp. 1-8.

Nesvizhskii, AI, Roos, FF, Grossmann, J, Vogelzang, M, Eddes, JS, Gruissem, W, Baginsky, S, Aebersold, R 2006, 'Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data', Molecular & Cellular Proteomics, vol. 5.4, pp. 652-670.

Noble, WS & MacCoss, MJ 2012, 'Computational and statistical analysis of protein mass spectrometry data', *PLoS Computational Biology,* vol. 8, no. 1, pp. 1 – 6.

Postal Address: Building 161
c/- The University of Melbourne
Victoria 3010 Australia

Email: enquiries@amsi.org.au
Phone: +61 3 8344 1777
Fax: +61 3 8344 6324
Web: www.amsi.org.au