

Comparing Mass Spectra

Eric Puryear^[1], David Angulo^[2], Alex Schilling^[3], Kevin Drew^[4], Gregor von Laszewski^[5]

^[1] DePaul University, epuryear@students.depaul.edu

^[2] DePaul University, dangulo@cs.depaul.edu

^[3] The University of Chicago, aschilli@uchicago.edu

^[4] The University of Chicago, kdrew@uchicago.edu

^[5] Argonne National Laboratory gregor@mcs.anl.gov

Abstract

The comparison of mass spectra is essential to many areas of research, including bioinformatics. These comparisons are made difficult due to sample contamination and differences in equipment. Spectral Comparison addresses this problem by comparing mass spectra, and is able to compare spectra that have been obtained from multiple mass spectrometers with different mass accuracies, as well as spectra where the peaks are not perfectly aligned.

Keywords: Proteomics, Mass Spectrometry, Grid Computing, Cosine Similarity.

1 Introduction

Mass spectrometry is used in proteomics and many other disciplines. In virtually all of these scenarios, it is necessary to compare two or more spectra to determine if they represent the same sample. This can be accomplished in several ways, two of which are addressed by Spectral Comparison. The comparison of spectra is complicated by noise and slight shifts in the location of peaks due to isotopic distribution. This problem is addressed by spectral comparison through the use of ‘bins’, which allow peaks to occupy a given range without affecting the comparison.

2 Literature Review

Prior to implementing Spectral Comparison, Katty X Wan’s (et al) "Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle" and Zeev B. Alfassi’s paper entitled “On the Normalization of a Mass Spectrum for Comparison of Two Spectra” were reviewed. The methods described in these papers form the foundation of the comparison algorithms implemented in Spectral Comparison.

2.1 Katty X Wan, et al - "Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle"

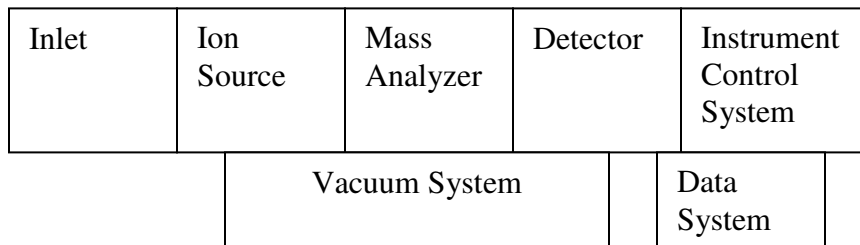
This paper [8] details two methods of comparing mass spectra: cosine similarity and the spectral contrast angle. Both of these methods are implemented in Spectral Comparison, and the test data provided by the authors was used to verify the accuracy of Spectral Comparison. Details of the formulas and how they were implemented can be seen in section 5.

2.2 Zeev B. Alfassi - “On the Normalization of a Mass Spectrum for Comparison of Two Spectra”

Zeev [1] provide a critique of the similarity index method as proposed by Wan et al [8], showing the superiority of the cosine similarity method over the similarity index. Zeev also shows methods of normalizing the spectra that allow for an accurate comparison while being less computationally expensive than the normalization techniques proposed by Wan et al [8].

3 Mass Spectrometry

Mass Spectrometry is used in proteomics to determine the mass and quantity of amino acids in a given sample. The knowledge of what amino acids are present and their concentration allows researchers to determine what peptides and proteins comprise the sample. This is done using a mass spectrometer, which consists of the following seven major components: the sample inlet, ion source, mass analyzer, a detector, a vacuum system, instrument-control system, and finally, the data system [4]. Although all of the seven major components of the mass spectrometer are important, the first three of these tend to determine the major attributes of the mass spectrometer, and will be discussed [4].



*A conceptual representation of a mass spectrometer, showing the basic components
Figure 1 [4]*

The sample inlet is where the sample to be analyzed is placed. This inlet can be one of several types including a direct stage, or a capillary column liquid chromatograph. Regardless of the type, the inlet sample is the means of introducing the sample into the mass spectrometer.

The next step in mass spectrometry is to ionize the sample. This too can be accomplished several ways, each with its own advantages and disadvantages. Early mass spectrometers used fast atom bombardment, which was then superseded by electron-ionization. Electrospray and matrix-assisted laser desorption/ionization (MALDI) are two common methods used in modern mass spectrometers that lend themselves particularly well to proteomics research [4]. Electrospray ionization involves placing the sample in an acidic,

aqueous solution and spraying that mixture through the opening of a positively charged needle. The positively charged sample is drawn into the negatively charged mass spectrometer where it is analyzed. This ionization and vaporization process can be aided by the introduction of an inert gas such as nitrogen, and by applying heat to the positively charged needle [4]. Two sub-categories of electrospray ionization that have arisen are nanospray and microspray, both of which use similar ionization techniques, but with respectively smaller needles that result in a lower flow rate, facilitating tandem mass spectrometry experiments. The disadvantages of microspray and, particularly, nanospray, are that the smaller needles are more easily obstructed and more difficult to place [4]. For these reasons, nanospray is used less frequently than electrospray, and microspray is seen as a good compromise between the ease of use of electrospray, and the benefits of nanospray.

Another form of ionization is matrix-assisted laser desorption/ionization (MALDI). MALDI, where the sample is dissolved in an ultra-violet absorbing solution (called the “matrix”) and placed on a probe or stage where it dries and crystallizes [4]. Once this solution has dried and crystallized, it is exposed to UV laser light and is vaporized, and then ionized by the acidic nature of the matrix compound, along with the addition of acid to the sample [4].

The mass analyzer is responsible for determining the mass to charge ratio (m/z) of the ions that are being analyzed. Three of the major types used in proteomics research are quadrupole mass filters, ion trap mass analyzers, and time-of-flight Mass Analyzers. Quadrupole mass filters use a set of four charged rods that are configured in such a way that only ions with the desired m/z will stay “on course” through the instrument; all ions with a different m/z will be deflected away [4].

The ion trap also uses charged rods, but instead of deflecting away the undesired m/z ions, it initially traps all the ions that enter, and over time applies an rf-voltage to the trapped ions, allowing the desired ions to exit the trap and continue through the instrument [4]. This approach results in a detector that is very accurate.

Time-of-flight mass analyzers (TOF) apply electric fields to the ions that comprise a sample, and depending on their m/z , the ions will spend a discrete amount of time “in flight”, allowing the m/z of the ions to be determined [4]. TOF mass analyzers have the added benefit of virtually no upper limit on m/z , unlike the quadrupole and ion trap methods.

4 Reasons for comparing Mass Spectra

The ability to compare mass spectra is essential in a variety of research areas, including Bioinformatics. These comparisons are made in order to identify unknown samples, calibrate equipment, and determine if experiments can be accurately reproduced [10]. For example, a biologist can attempt to identify an unknown protein that was expressed by matching it against a database of known proteins, which can be faster and more accurate than De Novo sequencing, which involves the attempt to find a protein’s

sequence without using a database of known proteins. A chemist may calibrate and test their mass spectrometer by doing a test run using a known sample, and comparing the actual output to the expected output.

5 Spectral Comparison

Spectral Comparison implements two different methods of comparing mass spectra. These are the Similarity Index method and the Cosine Similarity method. Spectral comparison uses bins of a user defined size to hold one or more peaks, which are then compared. In the event that more than one peak resides in a bin, spectral comparison will average the peaks to get a single value, which will then be compared with the value in the corresponding bin of the other spectrum. For example given the spectra in figure 2, without variable bin size, it would be difficult to get an accurate comparison between the spectra, despite their obvious similarities.

Spectrum 1 (m/z, intensity):	100, 1	200, 2	300, 5
Spectrum 2 (m/z, intensity):	103, 1	202, 2	301, 5

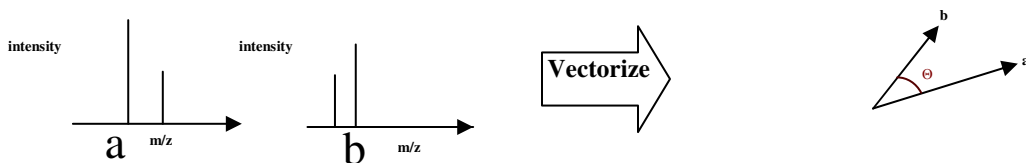
An example of how bin size can be used to compensate for slight variances in the mass to charge ratio (m/z).

Figure 2

With spectral comparison's variable bin size the spectra in figure 2 are recognized as being similar when a bin size of 5 Daltons is set. This allows peaks to shift due to differences between mass spectrometers and noise without adversely affecting the ability to compare spectra.

5.1 Cosine Similarity

Cosine similarity is the preferred method of comparing spectra in Spectral Comparison. This method implements the formula in equation 1, which consists of representing each of the spectra as a vector in n-dimensional space, where n is the number of peaks, and comparing the cosine of the angle between the vectors [4], as seen in figure 3. The formula for cosine similarity can be found in figure 4.



A simple example of how spectra are vectorized, and the resulting vectors.

Figure 3

$$\text{COS } \Theta = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2 \sum_i b_i^2}}$$

*The formula used by Spectral Comparison for Cosine Similarity, with a and b representing the two spectra to be compared
Equation 1[8]*

5.2 Similarity Index

The similarity index method, as developed by Wan et al [8], compares the difference in signal intensity by the smaller intensity for peaks that fall within a given mass range. The formula for the similarity index formula is shown in equation 2.

$$\text{SI} = \sum \frac{\sqrt{I - I_0}}{I_0} \times 100$$

*The formula used by Spectral Comparison for Similarity Index, with a and b representing the two spectra to be compared
Equation 2[8]*

5.3 Implementation

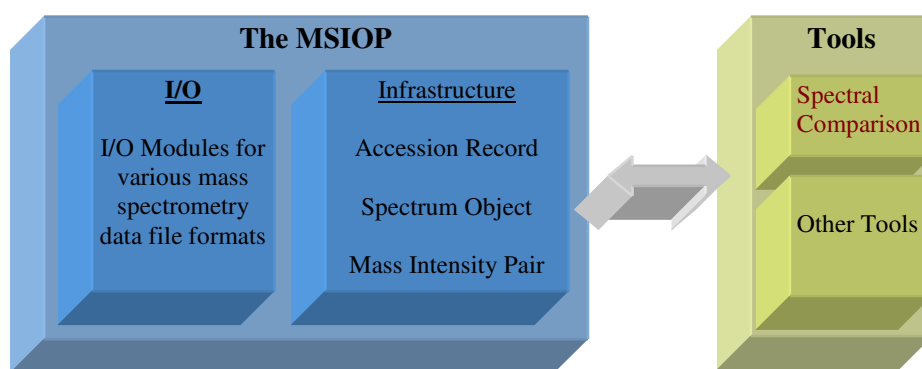
Spectral Comparison has been implemented as a C function. It is called with four arguments and returns a double ranging between 0 (completely dissimilar spectra) and 1 (completely identical spectra). The first two arguments are pointers to the spectra to be compared, the next argument is the bin size, in Daltons, to use, and the value of the final argument determines if cosine similarity (method 1) or similarity index (method 2) is to be used. The C prototype can be seen in figure 5.

```
double spectralComparison(MASS_SPECTRUM_PTR,
    MASS_SPECTRUM_PTR, double binsize, int method)
```

*The C prototype of spectral comparison, showing the parameters and return type need when using Spectral Comparison
Figure 5*

6 Infrastructure

Spectral Comparison utilizes the infrastructure provided by the Mass Spectrometry I/O Project (MSIOP)[6]. This includes the data structures for the storage of mass spectrometry data, and the input and output from these structures to and from various manufacture-specific data formats. The basic format of the MSIOP is show in figure 6.



*How Spectral Comparison interoperates with the MSIOP[6]
Figure 6*

7 Further Work

With a functional method of comparing mass spectra, the next step will be integrating this tool with a mass spectra database and web interface, allowing researchers to submit spectra for curation and then compare their unknown spectra against known samples. As there will be thousands of known samples, spectral comparison has been designed to run in a cluster or grid environment, allowing the task of comparing spectra to be divided up over multiple systems to speed computations. This is accomplished by the use of the Cactus framework, which enables cross-platform compatibility and interoperability.

Spectral Comparison's source code will also be made available for use by our collaborators and other researches wishing to compare mass spectra. In addition, the runtime efficiency of Spectral Comparison, which is currently $O(n)$, will be improved, although a major improvement is not possible as this is a comparison based program, which by definition cannot run faster than $O(n)$, where n is the input size.

8 Acknowledgements

This work was funded by the McNair scholars program and done in conjunction with the Illinois Biogrid. This material incorporates components from the Mass Spectrometry I/O Project (MSIOP) [6], which is based upon work supported by the National Science Foundation under Grant No. 0353989.

9 References

1. Alfassi, Zeev B. On the Normalization of a Mass Spectrum for Comparison of Two Spectra. Department of Nuclear Engineering, Ben Gurion University, Beer Sheva, Israel
2. Drew, Kevin; Angulo, David; Schilling, Alex; Freeman, Tim. *Mass Spectra Analysis on the Illinois Biogrid*. Proceedings of 2004 Midwest Software Engineering Conference, Chicago.
3. Krane, Dan and Raymer, Michael. *Fundamental Concepts of Bioinformatics*. Benjamin/Cummings, 2000.
4. Kinter, M., and Sherman N. Protein Sequencing and Identification Using Tandem Mass Spectrometry. Wiley-Interscience: New York, 2000.
5. Lesk, Arthur. *Introduction to protein architecture: the structural biology of proteins*. Oxford University Press. 2001.
6. Puryear, Eric; Angulo, David; Drew, Kevin; Schilling, Alex; von Laszewski, Gregor. Developing a Distributed and Scalable Foundation for Mass Spectrometry Data. DePaul University Tech Report. April 2005.
7. Steele Adam, Angulo, David S. The Illinois BioGrid: A Prototype for Industry-Academe Collaboration. Proceedings of 2003 Midwest Software Engineering Conference, Chicago.
8. Wan, Katty X, Vidavsky, Ilan, and Gross, Michael L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle.