PRINCETON UNIVERSITY

# LC-MS Data Analysis

## Computer Science JIW Final Report

**Adam Sanders**

**1/7/2008**

One of the most important technologies used in the analysis of proteomic data today is liquid chromatography mass spectrometry. As this technology has emerged in its modern form only recently, there is much that can be done to improve it. I demonstrate here several methods for the analysis of LC-MS raw scan data.

Adam Sanders
COS JIW
12/25/07

# LC-MS Data Analysis Final Report

**Abstract**

One of the most important technologies used in the analysis of proteomic data today is liquid chromatography mass spectrometry. As this technology has emerged in its modern form only recently, there is much that can be done to improve it. I demonstrate here several methods for the analysis of LC-MS raw scan data.

**Introduction to Proteomics**

The biological sciences are moving forward at an astonishing rate. Since the publication of Watson and Crick's famous paper proposing a structure for the protein Deoxyribose Nucleic Acid (DNA) less than a century ago, we have completely sequenced over 200 organisms, in addition to the partial sequencing countless other genomes (Watson & Crick, 1953). The wealth of genetic information generated by this work has transformed the fields of biology and medical research by allowing the large-scale interpretation of genetic data and cellular function (Lane, 2005). Tests have been developed for the rapid identification of genetic diseases such as PKU as well as the early characterization of risk factors for other genetic anomalies like breast cancer. While the advances in the field of genetics have proven an invaluable tool for the illumination of cellular function, the emphasis in molecular biology is now experiencing a shift away from sequencing DNA and characterizing genes toward a systematic evaluation of how the myriad of encoded gene *products* operate in order to sustain life (Listgarten & Emili, 2005). That is, rather than examining the differences in the nucleotide sequences that comprise individual genes, biologists are now expending an increasing amount of their time in trying to understand the protein products of those genes, as well as their interactions.

This new focus in biology has been termed proteomics, and it differs greatly from genomics in both complexity and scale. As an introduction to the field of proteomics, it is worthwhile to compare it to the more widely understood field of genomics. Genomics has been defined as "the comprehensive study of all genes in a cell or organism, how they interact and their functions" (ESA Glossary, 2007). Findings in the field of genomics could be that having a particular allele (one version of a gene) of gene A increases one's risk of cancer by 40%. This analysis can be done at birth and will remain unchanged throughout the life of the individual with gene A.

Alternatively, proteomics could be defined as the comprehensive analysis of all of the proteins found within a cell or organism, their interactions, and their functions. A result in this field could be the finding that an abnormal amount of protein A is the cause of Parkinson's disease. The definitions of these two fields of study are very similar, as are their aims. Both are concerned with studying a type of biological molecule to come to a greater understanding of how the human body functions, often with the ultimate goal of maintaining or improving that function. While genomics research seeks to reach these goals through the study of genes, proteomic research deals with proteins. However these differences, semantic as they may seem, make a tremendous difference in the scope and depth of the research.

In regards to the scope of the fields, one need only look at the molecules themselves. Genes are comprised a sequence of four separate nucleotides strung together to form a blueprint from which proteins can be translated. These genes are primarily found in either a double stranded helical structure, as is the case with DNA, or a single stranded structure, as is the case with RNAs. The types of interactions that can occur between these molecules have been well understood for decades. Proteins, on the other hand, are comprised of a sequence of twenty different amino acids that are strung together to form individual protein residues. These protein residues can then be combined, chopped up, shifted, twisted, or turned into a tremendous array of different conformations and combinations. In contrast to the regular helical structure of DNA, the physical characteristics of the individual protein residues allow the proteins residues themselves to fold into an immense number of arrangements. Taking the human genome and the human proteome as an example, the difference in scope becomes readily apparent. Though currently unknown, the number of genes in the human genome has been estimated at around 25,000. The number of proteins found in the human proteome has been projected to be over two-million.

Concerning the depth of the research, the challenges presented by proteomic analysis are far more complex than the huge but basically straightforward challenge of sequencing the genome of an organism. While genomes are static molecules that change little over the lifetime of an organism, proteomes are dynamic. This property is shown particularly well through the example of the monarch butterfly. This organism begins its life as a larva that bears more resemblance to a worm than to a grown butterfly. However, after a great deal of development, the larva experiences a metamorphosis through which it is transformed into its final winged form. From the beginning to the very end of this process, the butterfly maintains the very same set of genes. Forgetting for a moment the very likely scenario of minor genetic mutations, the genome of the larva can be said to be identical to that of the fully grown

butterfly. However, at every stage of this process, the set proteins found in the organism is different. In fact, it is very probable that every minute the complex mixture of proteins found in the organism experiences a detectable shift. Given the very palpable physical changes that the butterfly undergoes, it becomes clear that a study merely of the genetics of the butterfly would lead to an incomplete picture of how that organism really functions. Without a deep understanding of the proteins found in that organism during the different stages of its life, we cannot mentally reconcile the physical changes we witness. This dynamism found in the field of proteomics has led some researchers to modify the more traditional definition of the field to better encompass its unique challenges. One revised definition describes the study of proteomics as "the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes" (Reinert & Kohlbacher, 2005). Important to note in this new definition is the addition of differential conditions. It is in this difference that we find some of the most valuable aspects of proteomic analysis. However, with added information come added challenges.

The sheer size and complexity of proteomics make its research a very technology-driven enterprise. It is in the technology that we find a final contrast between proteomic and genomic analysis. While the technologies used to sequence the human genome (among many other genomes) are well established and fully automated, the tools used in proteomic analysis have only recently emerged. A paper in Nature cites mass spectrometry, a technology in its modern form only a decade old and the object of my research as the most important among the five pillars of proteomics analysis tools (Tyers & Mann, 2003). The relative youth and obvious importance of each of these tools them an area ripe for the innovation of new techniques and methods. In the same way that advances in genetic sequencing technology have allowed for incredible feats like the sequencing of the human genome, advances applied to the tools used in proteomic analysis promise to become incorporated into a whole generation of new scientific discoveries.

**My Project**

As mentioned above, there is a great deal of progress yet to be made in the field of proteomics. Regardless of the direction that the field takes, the instrumentation used to carry out proteomic analysis will be of great importance. It is for this reason that I elected to work on improving the raw-data analysis on one of the most important tools used in this research, liquid chromatography mass spectrometry. The use of this process allows researchers to identify and quantify not just individual proteins, but large and complex mixtures of proteins such as

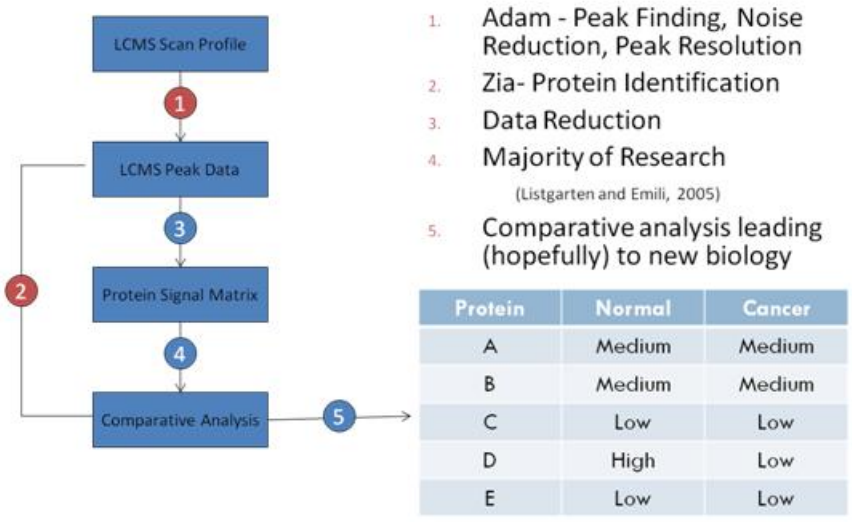
one would find in a larva, butterfly, or human being. As such, this process has become the method of choice for most proteomic analysis (Listgarten & Emili, 2005). However, there are still a great number of difficulties yet to be adequately addressed. These include: experimental noise, systematic variation between experimental runs, the extreme overall range and dynamic nature of protein levels, the huge number of protein features (peaks) in which there are an enormous number of uncorrelated features (Tyers & Mann, 2003).

The idea to develop a program to analyze LC-MS scan information as the specific data with which I worked came from the results of a previous study completed in 2007 by a team of researchers led by Lukas N. Mueller located in Zurich, Switzerland. These researchers introduced a new tool named *SuperHirn* that was to be used to analyze raw LC-MS scans. Despite their best efforts, the analysis completed by Mueller et al. was found to be sub-satisfactory by a team of researchers at Princeton University. The product of the study, a data set that could be used for cross-proteomic analysis, was found to be of low resolution and poor quality. The methods used by Mueller et. al had resulted in important signal loss, resulting in an unusable data. Compounding these difficulties, many of the methods used were very poorly documented, lacked clear reasoning, or were found to be difficult to reproduce. The product of these difficulties was the initiative to develop a new program to analyze raw LC-MS data.

This assignment was taken up by Zia Khan, a graduate student working in the lab of Mona Singh. Zia developed a large currently unnamed program used to do LC-MS analysis. The primary goal of the project was to develop a means of protein identification and quantification that had an absolute minimum of data loss. In order to accomplish this, Zia developed a program meant to work with data at a much more basic level than had previous researchers (Khan, 2007). This program is represented as step two in Figure 1 below. While this analysis worked well, some aspects of data needed thoughtful pre-processing before they were used in Khan's algorithms. It was necessary to develop algorithms that had an absolute minimum of data loss *before* the data was ported into his program. It is with that goal in mind that I was commissioned to develop a program to work with the very most basic form of LC-MS data, the LC-MS scan profile. However, in order to understand why this is important, one must first understand where my program and Zia's fit into the current flow of research.

. The LC-MS scan profile is the very first file generated by any LC-MS scan. This profile represents an enormous amount of data, as it contains the entire protein signal as well as all background noise. Ordinarily this file is then stripped of its low intensity peaks to generate LC-MS peak data. While LC-MS peak data represents a

smaller and more manageable file, it is of a lower resolution and quality than the raw scan profile. It is from this file that a protein signal matrix can be developed. The protein signal matrix is a listing of all of the protein fragments that have been identified, along with their relative intensities. The generation of this protein matrix is where the majority of the current research into LC-MS analysis is done, and is represented by step three in Figure 1 below (Lane, 2005). That the majority of research is done converting the LC-MS peak data to the protein signal matrix is understandable, as it is at this point that the information in the LC-MS scan actually becomes useful. From this protein signal matrix, one can determine what levels of protein are abnormally high or low in comparison to the other proteins in the scan. This comparison, termed comparative protein analysis, gives biologists a tremendous tool in the search for disease pathologies, most notably cancer. Through the use of protein signal matrices from the cells of several individuals, one of whom has cancer, one can determine which proteins are expressed at different level (Foss, et al.). If it was found, as in Figure 1 below, that one protein was expressed at a higher level in the individual with cancer, it would make sense to isolate this protein for further study. As an example, protein D below would be an excellent source for new research into the pathology of cancer. However, this research is by no means limited to cancer, and could conceivably diagnose any disease, among a great many other conditions.




| Protein | Normal | Cancer |
|---|---|---|
| A | Medium | Medium |
| B | Medium | Medium |
| C | Low | Low |
| D | High | Low |
| E | Low | Low |

**Figure 1: Outline demonstrating the standard flow of research**

Each arrow in the above diagram represents some form of data loss. Surprisingly, the decision to develop algorithms that work exclusively with raw scan data is surprisingly novel (Listgarten & Emili, 2005). While very basic algorithms have been developed for the processing of this raw scan data, they are mostly concerned with reducing the size of these scans. Little time has been devoted to generating a single model that does extensive

computation on this most basic form of data. One result of the lack of research in this area is that any improvements made at this most basic level of processing could, in turn, be used in all current and future research in this field. This benefit is especially true of the program currently in development by Zia Khan. By developing algorithms for the processing of raw data scans and testing my results with his program, I created a program specifically, though not exclusively, tailored to be optimal for use in his later analysis.

The first stage of this project was to develop a data viewer. This was important both for viewing the raw data as well as visually determining the effectiveness of the algorithms later tried. Once this stage was complete, several new ideas would be tried in order to filter out peaks not necessary for future analysis while maintaining as many useful peaks as possible. Finally it was necessary to develop some means of determining the success of these algorithms.

**Liquid Chromatography Mass Spectrometry (LC-MS)**

While many technologies form the base of proteomic research, mass spectrometry (MS) has been cited as one of the most important and today is the most sensitive and straightforward method for identifying and quantifying proteins (Tyers & Mann, 2003). MS measures with extreme sensitivity, the mass over charge (m/z) ratios of gas-phase ions. Since its development by J.J. Thomson and his student F.W. Aston, MS has played an increasingly significant role in proteomic analysis (Lane, 2005). Since proteins are generally large molecules, the first stage in MS is to break these proteins down into a more reasonable size. This can be accomplished through the use of an enzyme called trypsin that cuts at specific sequences. As an example let us consider the amino acid sequence for a protein involved in T-cell lymphoma-1 shown below.

MAECPTLGEAVTDHPDRLWAWEKFVYLDEKQHAWLPLTIEIK
DRLQLRVLLRREDVVLGRPMTPTQIGPSLLPIMWQLYPDGRYR
SSDSSFWRLVYHIKIDGVEDMLLELLPDD

Trypsin will cut the above sequence at any point where there is an arginine (R) or a lysine (K). If we were to digest the above sequence with trypsin, the result would be the ten smaller shown below in Table 1. I have added at left the masses of these sequences in Daltons. Comparing this table against a collection of known protein fragments allows researchers to determine the identity of a given protein, assuming this protein is already in the database.

Table 1

| Mass in Daltons | Tryptic Fragment Sequence |
|---|---|
| **3379.7540** | EDVVLGRPMTPTQIGPSLLPIMWQLYPDGR |
| **1841.8156** | MAECPTLGEAVTDHPDR |
| **1686.8142** | IDGVEDMLLELLPDD |
| **1448.8260** | QHAWLPLTIEIK |
| **971.4217** | SSDSSFWR |
| **913.4665** | FVYLDEK |
| **832.4352** | LWAWEK |
| **772.4716** | LVYHIK |
| **529.3456** | LQLR |
| **500.3555** | VLLR |

The final result of a mass spectrometry scan is a two-dimensional profile that shows the relative intensity values of each of the fragments as well as their m/z value. In general, the charge of these proteins is one, so the m/z value simplifies to just m (Khan, 2007). In Figure 2 below, we can see the MS spectrum for the T-cell lymphoma-1 protein digested with trypsin. In addition, three of the most intense peaks have been identified by their sequences. One can see that these sequences are the same as the sequences found in Table 1 above. Peaks with intensities such as these (representing a great deal of protein) would be most readily identifiable when compared with previously seen protein matrices.

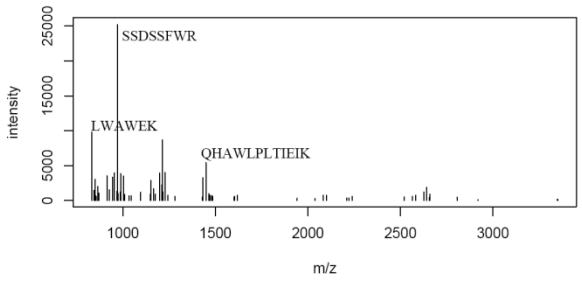



Figure 2: MS spectrum for the T-cell protein digested with trypsin

Mass spectrometry is an excellent means of determining protein identity when one can easily isolate and purify individual proteins. However, as mentioned previously, protein mixtures in the real world are tremendously variable and often it may be difficult or even impossible to isolate individual proteins from complex mixtures. It is for this reason that mass spectrometry has been paired with liquid chromatography to form liquid chromatography mass spectrometry (LC-MS). This process, which is the object of my research, also begins with the tryptic digestion of proteins. However, in this case the protein sample is not simply one isolated protein, but rather an assortment of

proteins. An excellent example of such a mixture would be a human blood sample, and in the course of my project I worked with similar data. Once digested, this mixture is put into a liquid chromatography (LC) column. The proteins are then drawn through the column at different speeds corresponding to their physical characteristics. Once they reach the end of the column (different proteins will reach this end at different points) they are immediately sent into a mass spectrometer. The result is a matrix with axes corresponding to the time at which the protein left the LC column (its retention time) the mass over charge ratio of the protein (m/z value) and the relative amount of protein found (its intensity). Figure 3 below is a very small subset of actual data from my project demonstrating the resulting three-dimensional matrix. Figure 4 demonstrates the graph of a single retention time within the data seen in Figure 3.

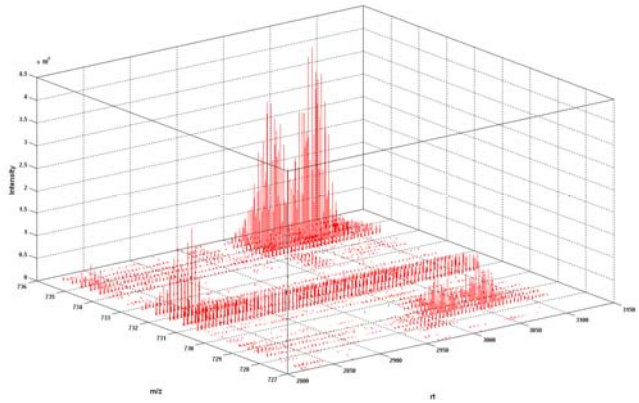


Figure 3: A small subset of the Mueller et. al data showing the dimensionality of the data

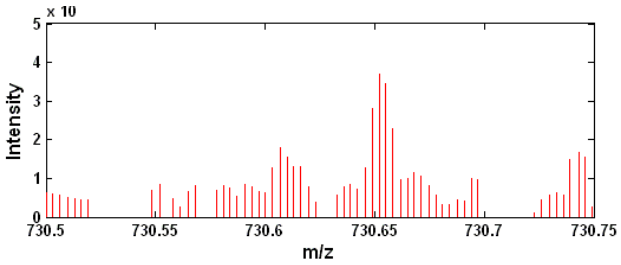



Figure 4: A single retention time from the scan in Figure 3

**The Data Set**

The specific data with which I am working was originally found in a paper by Mueller et al. published in the journal Proteomics in 2007 (Mueller, et al., 2007). The data was downloaded August 25th, 2007 from http://prottools.ethz.ch/muellelu/web/Latin_Square_Data.php. The proteins fragments represented include the tryptic

digests of several standard non-human proteins. The data is encoded in 19 xml files, each of which is roughly 1GB in size. Each of these files represents a single LC-MS scan. The data found within these files is essentially a very large list of peaks. Each of these peaks contains information about retention time, m/z value, and intensity. In addition to these basic facts, a great deal of other information was also included in these files regarding MS/MS analysis (a method by which one can obtain specific protein sequences in addition to simply their mass), as well as a great deal of information about the specific settings of the instrumentation and preparation methods of the proteins. This information was unused in my analysis, though may present an interesting direction for further research.

**Raw Signal Processing**

The first step in this research project was the find a reliable and effective way to view the data. Originally, this was anticipated to form a very large part of this project as it was assumed that the data would be very noisy. Work with previous data and the literature suggested that the peaks would have to be identified. Due to the way in which LC mass spectrometers work, the baseline for the raw data normally contains non-zero peaks at a majority of the possible positions. A complicating factor often found in LC-MS data is that often the intensity of this noise is on a gradient for every given retention time. For example, the background intensity levels are higher, on average, in areas of high mass over charge ratios than they are in areas of low m/z ratios.

Another step found previously when working with raw LC-MS scans was the need to address systematic flaws in the data that occur at a specific m/z ratio. In previous data sets, it was necessary to mask out certain m/z ranges that consistently contained false peaks. These areas can, with a little knowledge, be identified visually. They are represented as lines that span the entire LC-MS scan at an intensity level higher than surrounding regions. They differ from data points in both the fact that they exist at every retention time, as well as the fact that they do not show the characteristic isotope spacing (on the m/z axis) that is found in protein signals. Instead, these points are found in close m/z proximity to one another. In order to mask these points, all that must be done is to either remove these m/z ranges from the data set, or write some small piece of code into the analysis program that makes sure it does not consider peaks within these ranges. While visual identification and masking is not difficult, one research goal in using the Mueller data was to find a way to systematically mask these features in order to improve the automation and speed of the process.

Below is shown an example of a previous LC-MS scan using a previous data set. Though it may be difficult to identify from this small example, it contains the shifted baseline intensity values mentioned above. Also visible are the systematic errors found in many of these scans. A particularly good example of this can be identified 95% down the image where there is a band of bright peaks that stretches across the entire scan. Were this data to be analyzed, this m/z range would have to be removed. Features with high intensity are shown in bright colors (yellow) and low intensity peaks are shown as darker colors (blue). Peaks of zero intensity are shown in black.

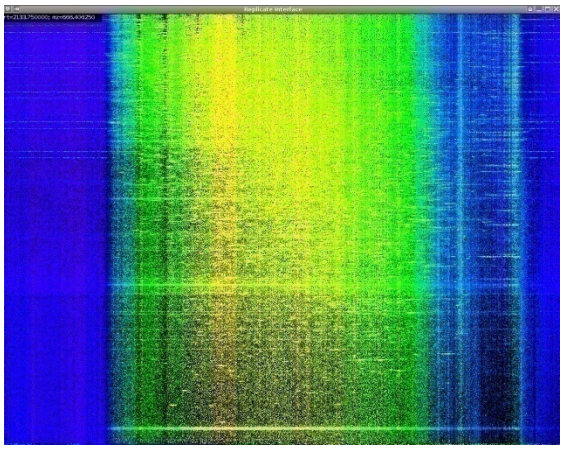


**Figure 5: Scan from a previous study done by Zia Khan**

However, when the Mueller data was examined, it was found to be a considerable amount cleaner than expected. Rather than finding the variable non-zero baseline intensity values, the Mueller raw data contained mostly zeroed peaks and real protein signal. Also not located were the systematic data aberrations mentioned above. There are a number of possible explanations for the cleanliness of the Mueller data. The first and most obvious of which is that some pre-processing had already been completed on the data before it was posted to the website from which it was downloaded. However, more likely is that the LC-MS machine in which the data was scanned was of a newer lineage and was actually more effective than previous models. This explanation fits in well with one of the largest difficulties in attempting raw LC-MS scan analysis, which is that every scanner, and thus every scan, is different (Reinert & Kohlbacher, 2005). Since scans are being done all over the world with a wide variety of different scanners and settings, there is no genuine standardized raw data. In this instance, the result of this variability was fairly positive. The result of this rather happy finding was that less work needed to be done pre-processing the data before further analysis could be done.

Given that the data was so clean, the creation of a raw data viewer was not particularly difficult. As with the rest of the code in this project, it was written in C++ and makes use of several large libraries for both reading xml files as well as creating a simple GUI in order to examine points in the data. Only minor semantic modifications were made to this program in order to view the new Mueller data. The results of these modifications are the photographs below. Here we can see the more sparse nature of the data without the background noise that characterized previous data sets. One caveat when examining pictures of the Mueller data set was that the computer with which I worked did not have the requisite memory to load an entire file into the image viewer. In order to deal with this problem, it was necessary to carved up the original data files into smaller and more manageable subsections that were less memory-intensive.

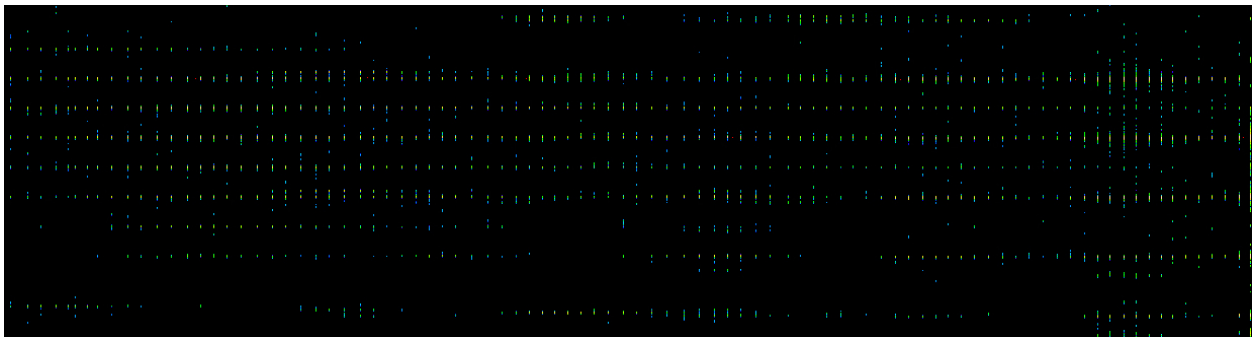


Figure 6: Subsection of data from the Mueller data set showing large areas of zeroed peaks

**2D Peak Finding**

The next step in this analysis was to find a method of deleting extraneous peaks. One of the best ways to differentiate between peaks that are simply background noise, and those that represent protein signal is to determine whether or not the signal continues over time. The logic behind this process is that, when examined over time, protein signal should form a bell curve that corresponds to the amount of any given type of protein currently coming off the LC column. It is unlikely that the entirety of a protein fragment would make it through the LC column at the exact same speed and thus show up on the mass spec at the exact same retention time. In reality, it is more likely that some small percentage of any protein fragment will arrive slightly ahead of the bulk of its peers, and those peers will in turn be followed by a small percentage of protein that made it through the column at a slower rate. In contrast to the bell curve example of real protein signal, much of the noise in the protein samples occurs as a lone

feature or peak at a single retention time. The obvious solution to this occurrence is to locate and remove those peaks that are not surrounded by a sufficient number of similar peaks.

Many different methods could be applied to this problem, the first of which would be a statistical method used to fit a bell curve on top of peaks across the retention time axis. Peaks that were found to be similar to this bell curve would be kept while those dissimilar would be removed from consideration. However, it is important to note that retention times are not sampled at an even rate (Listgarten & Emili, 2005). Since these retention times vary in rate, it becomes slightly more difficult to fit exact statistical methods to the location of extraneous peaks. Rather than use this method, another method of 2D peak finding was selected. This method involved drawing areas of influence around every single peak in the data set. These areas were rectangles that have a greater width along the retention time axis, than they do height along the m/z axis. Every peak was then given a score corresponding to the number of boxes under which it fell. Finally, a cutoff point was defined such that those peaks with a score of at least a certain level would be kept, and the remaining peaks were removed from consideration. Originally it was assumed that protein signal would maintain an exact or very close to exact m/z ratio across many retention times. Were this the case, it would be unnecessary to add any height to the boxes. This was done in order to take into account the minor variations in the m/z ratios of protein signals across different retention times. Figure 7 below is a visual demonstration of a subsection of the Mueller data before peaks were dropped. It is possible to locate several peaks in this picture alone that only contact one density box. These peaks very likely do not represent any part of a protein and would be dropped.

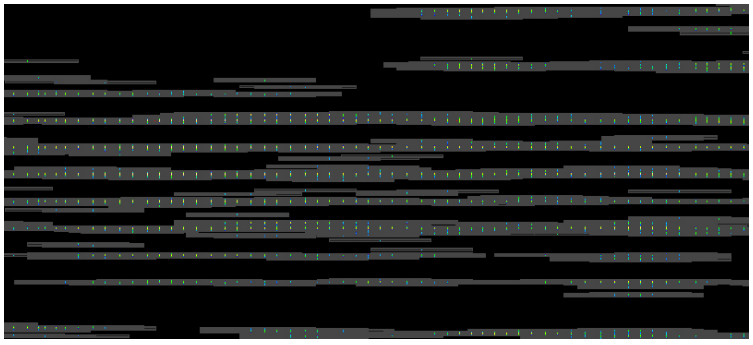


**Figure 7: Subsection of Mueller data showing density boxes**

**Single Peak Resolution**

The next challenge presented by this project came in relation to difficulties with the extreme variations in intensity values. As mentioned previously, the range of intensities found in any given LC-MS scan can vary between zero and $10^9$ (Listgarten & Emili, 2005). Peaks that are important for later stages of LC-MS analysis certainly cover higher peaks, but can also have intensity values as low as $10^3$. In Zia Khan's later analysis, it was found that these high intensity values were distorting results. This problem resulted from the fact that features of very high intensity also had peaks adjacent to them on the m/z axis of high intensity. This problem is demonstrated below at left in figure 8. The width of these segments made them difficult to process. The solution to this problem is to simply remove the features adjacent to the center peak. Because only the most intense peaks at a specific m/z range are needed to match against previously recorded proteins, the removal of the leading and trailing edges of those peaks did not reduce the ability in later analysis to identify and quantify proteins. However, it was important that the peaks maintain a reasonably accurate m/z reading. Were the features to be moved considerably, it would begin to affect the density boxes mentioned previously. Too much movement along the m/z axis could result in features being dropped from consideration because they no longer were in the small m/z range of their neighbors.

Figure 8: Example of width reduction

In an attempt to simplify this problem, a method of chunking was developed. The goal of this method was to break all the features at any given retention time into pieces roughly corresponding to important features. From that point, it would be easier to implement an algorithm that can identify the important peaks and allow removal of those features deemed unimportant. This method works by grouping together peaks that are separated by a gap greater than a certain distance. For every retention time, the m/z ratios are scanned sequentially and once gaps of a requisite distance (in the m/z axis) are found, the program defines them as a new chunk and sends them to another method for analysis. This "chunking" is demonstrated in Figure 9 below. The black lines above the image demonstrate the areas where a gap was detected, while the blue numbered lines below show the peaks that will be

sent to the next method for analysis. Once these pieces had been identified, there were a number of possible methods for removing extraneous peaks.

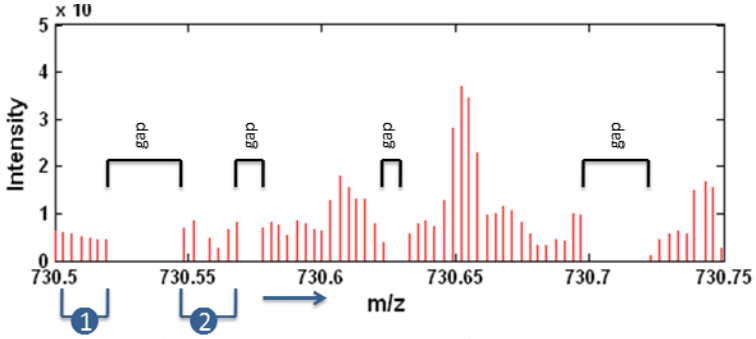



Figure 9: Data viewed at one retention time showing "chunking"

One possible method for resolving this problem would be to identify a minimum intensity below which no peaks would be considered in analysis. Declaring a minimum intensity would result in a thinning of the graph at left in Figure 8 by removing the less intense peaks to the top and bottom of the most intense peaks at the center of the figure. However, this method is not without problems. Given the tremendous variability in the intensity of important peaks, many peaks that should not be dropped would likely be removed. Conversely, there would also be a great number of peaks saved that should have been dropped. Also attempted was the method of simply returning the highest peak in the data. This method was obviously very effective at removing the width of every chunk, it had the tendency to create too much variability in the m/z axis. The result of which was poor identification in later analysis. While both of these methods were impractical for the resolution of peaks, they demonstrate some of the challenges mentioned previously when dealing with LC-MS data.

A final method applied to this problem was a determination of the mean of the m/z values weighted by their intensity. Once this weighted mean had been found, the algorithm then selects the peak closest to that location and removes all other features from the current chunk. This method is something of a compromise of previous methods. It requires some use of the intensity values to find the best peak, but the entirety of the analysis is not based on intensity. This method has several advantages over the previous methods. The primary advantage is that this method always reduces the number of peaks in a chunk to one, which results in a signal that are the most simple to analyze. In addition, it did not result in a great deal of variability in the m/z axis. Rather, it very often selected the same m/z value across multiple retention times. This method formed an important part of the later research into the area, but it was soon discovered that it did not completely cover all possible scenarios. Several very important,

though much less common, problems arose when this method was tried on larger and larger sets of data. These problems will be covered in the next section.

**Multiple Peak Resolution**

While the single peak algorithm did function, it was soon found to contain a major flaw. Ordinarily, protein signals are found to be a reasonable distance apart, but on occasion two protein fragments were found to overlap at both the m/z and retention time. The result of this overlap is two important peaks that are grouped together within the same. The close proximity of these peaks compounded difficulties in analysis for both the pre-processing stages as well as Zia's later algorithms. To make matters worse, the method used to reduce the width of peaks resulted in very poor results when two fragments were found to overlap. An example of such an occurrence can be seen below in Figure 10. At left in this figure, the two blue lines represent two separate peaks that overlap and have been caught in the same chunk. Since the current algorithm attempts to find the weighted mean of the current chunk, the peak returned will be that closest to the red line. When seen from the graph at right, it becomes clear that this may be one of the worst possible peaks to select. Not only is it not representative of the intensities of either peak, but it is exactly in the middle of both peaks in terms of its m/z ratio.

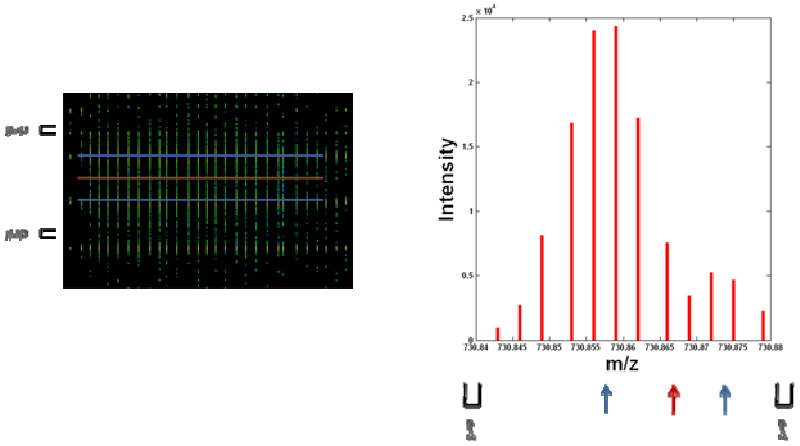



Figure 10: Demonstrates the problems with the single peak resolution methods used

In order to cope with this problem, it was necessary to develop a new algorithm that was capable of dealing with multiple peaks. The original thought was to shift a window over the data at ever retention time and record

every instance where the slope of that window shifted from positive to negative. The result of this algorithm would be something like the picture shown below. However, this too was not without problems. Figure 11 below demonstrates the results of this algorithm. The difficulty lies in the fact that many extraneous peaks lie at the leading and trailing edges of high intensity features. While the high intensity features are important, the peaks to the right in the figure below are not. However, these values shift between positive and negative slopes rapidly, resulting in many unnecessary peaks being returned. The trailing edges of the top graph in Figure 11 are blown up and shown again below. It becomes apparent that even when viewed at a higher resolution, some of the peaks are not even visible. However, these peaks would be picked up by the slope finding algorithm and returned for further analysis.

Previous research had suggested the mean shift method be used in order to group together features belonging to the same peak (Reinert & Kohlbacher, 2005). This method works by first passing a window over the data. At every single frame of this passing, the intensity weighted mean m/z value is calculated. Finally, all of the peaks in range of that window are then shifted towards the weighted mean. The result of this method is that less intense peaks travel towards more intense peaks. However, this method differs from previous examples in that it does not allow very intense peaks to effectively drown out lower intensity peaks. Once this algorithm has been run on a chunk, another gap function is used to split the chunk into several different pieces. These separate pieces can then be transferred into the original signal resolution method which calculates the intensity weighted mean m/z value and returns a single peak. This is demonstrated by Figure 12 below. In these two graphs we see two time steps from the mean shift algorithm. The first step locates each of the peaks in the current chunk by generating a new peak under them. As the window moves to the right along the higher m/z values, it moves the less intense peaks in the same direction.

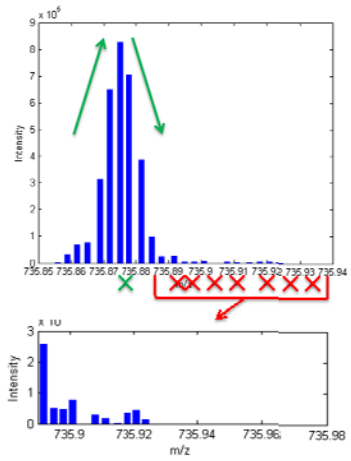



Figure 11: Demonstration of the problems with the slope finding algorithm

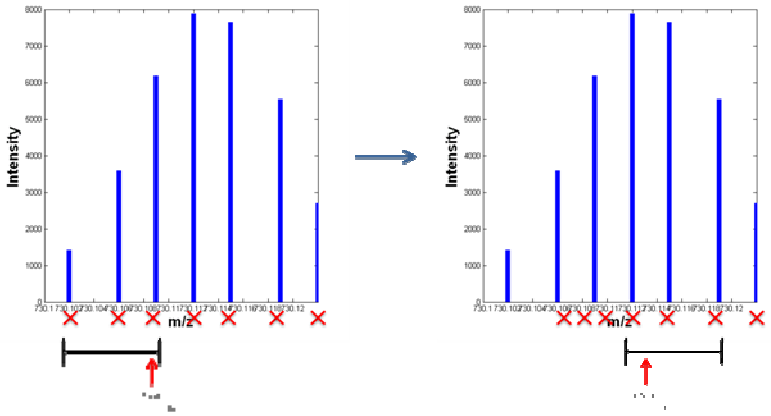



Figure 12: Demonstration of mean shift algorithm

The result of the combination of mean shift, and single peak resolution, and density boxes was a significant increase in the number of peaks identified over just chunking alone. This represents a significant improvement as without these methods those would be lost data. The goal in this project was to generate the most reliable data possible while removing as many peaks as could be reliably determined were not necessary. A brief summary of the number of peaks generated by each of the algorithms is shown in Figure 13. It is very easy from this graph to see the effects that each one of these algorithms has on the number of peaks found. Given the problem of trying to reduce the width of segments as well as deletion of extraneous peaks, the single peak method reduces the peak number by the greatest amount. This makes sense as it is the only method that only returns a single peak for ever chunk. The most ideal method found so far locates slightly more peaks than does the single peak method. This is in keeping with the finding that only a few of these chunks actually contains more than a single peak. Finally the slope search method finds entirely too many peaks. It is important to note that all of the below methods already have

density boxes applied. That is to say, these numbers are true of the number of peaks after that algorithm had been applied.

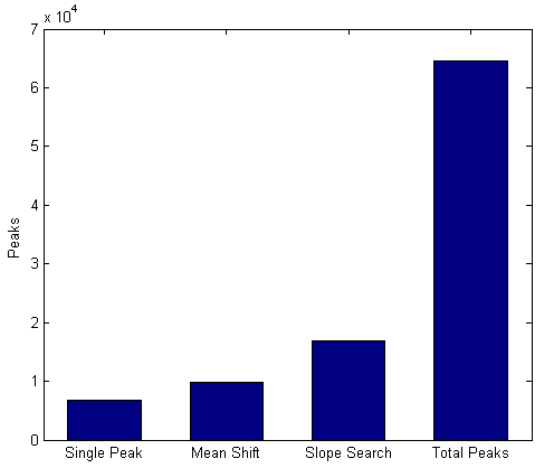



Figure 13: Summary of algorithms used to deal with multiple peaks

**Future Directions**

As mentioned previously, there is a great deal of work yet to be done in liquid chromatography mass spectrometry data analysis. In fact a number of the algorithms used in this paper could be improved. Specifically, it might be worthwhile to modify the density boxes to take into consideration intensity values. For example, those peaks that have very high intensity values could be given a greater density box weight. Since background noise peaks tend to be rather smaller than real peaks, this might result in more reliable feature detection. Further research directions can also be found for the mean shift algorithm. Presumably, if this algorithm works well enough, there is no real reason to use chunking. It might make more sense to simply run this algorithm against the entire retention time and once that is done, search for areas within the resulting graph of a minimum weighted density. These dense areas could in turn be searched for characteristic peaks. This method, in contrast to the one currently in use, would ensure that no peaks were accidentally cut in half by the chunking method. Finally, an important future research direction is a means to really quantify results. Although a great deal of effort has been devoted to examining these methods, it is extremely difficult to say quantitatively exactly how much better they fare in real tests than previous methods (Lane, 2005). This is partially a result of the specific location along the research chain that this project was located. It is difficult to say with any certainty exactly how well these algorithms work as their final results rely on a long line of research. While the results shown by pre-processing files with these methods for Zia's program look promising, this is no guarantee that they will be a positive influence on the results of other experiments. In addition,

given the variable nature of LC-MS scan files themselves, there is still more uncertainty as to their efficacy on other data files that represent proteins from other animals or that were scanned on different machines under different conditions.

**Conclusion**

The work I have completed here is hopefully one of the first among many more attempts at analyzing the most basic levels of LC-MS scan data. Despite these minor caveats found above, one can hope that more analysis will eventually follow. The benefits that we stand to gain through the exacting assessment of this important biological information are almost limitless. The development of these tools, most especially LC-MS and related technologies, stands to make a tremendous impact on the field of proteomics. The ability to rapidly and effectively identify and quantify proteins with high accuracy is an absolutely critical component of the ultimate goal of understanding the human proteome. The results accrued from this proteomic analysis have yet to be seen, but if the previous research surrounding the human genome is any indication, they stand to revolutionize the way we view ourselves and the world around us.

**Works Cited**

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature* , 198-558.

*ESA Glossary*. (2007, December 20). Retrieved December 20, 2007, from ESA Biosciences Inc.: http://www.esainc.com/

Foss, E. J., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., et al. (n.d.). Genetic Basis of Proteome Variation in Yeast.

Khan, Z. (2007, May 10). Peptide Fragmentation in Tandem MS Experiments.

Lane, C. S. (2005). Mass Spectrometry-based Proteomics in the Life Sciences. *Cellular and Molecular Life Sciences* , *62*, 848-869.

Listgarten, J., & Emili, A. (2005). Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry. *Mollecular and Cellular Proteomics* , 419-434.

Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., Bodenmiller, B., Brusniak, M.-Y., et al. (2007). SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* , 1-11.

Reinert, K., & Kohlbacher, O. (2005). *Signal Processing and Data Reduction for Differential Proteomics with HPLC/MS.* German Ministry for Edcuation and Research.

Thomson, J. J. (1913). Rays of Positive Electricity and their Applications to Chemical Analysis.

Tyers, M., & Mann, M. (2003). From Genomics to Proteomics. *Nature* , 193-197.

W., S. D. (2003). Current Challenges in Proteomics: minimizing low abundance proteins and expanding protein profiling capacities. *16th International Mass Spectrometry Conference* .

Watson, J., & Crick, H. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature* , 737-738.