Winter 2009

Design and implementation of a statistical analysis tool for two biological states

Gang Lu
University of New Hampshire, Durham

DESIGN AND IMPLEMENTATION OF A STATISTICAL ANALYSIS TOOL FOR TWO BIOLOGICAL STATES

BY

GANG LU

BS in Biochemistry, Jilin University, 1992

MS in Computer Science, University of New Hampshire, 2002

THESIS

Submitted to the University of New Hampshire

in Partial Fulfillment of

the Requirements for the Degree of

Master of Science

in

Biochemistry

December, 2009

This thesis has been examined and approved.

Thesis Director, Vern N. Reinhold, Research Professor in Biochemistry

Philip J. Hatcher, Professor in Computer Science

Linyuan Li, Associate Professor in Mathematics and Statistics

3 December 2009
Date

ACKNOWLEDGEMENTS

I sincerely thank Professor Reinhold's support to my work on this thesis. Other lab members, Dr. Andrew Hanneman, Dibya Himali, Kevin Bullock and Hui Zhou also gave me a lot of help especially on lab work such as how to use mass spectrometry and prepare samples.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

DESIGN AND IMPLEMENTATION OF A STATISTICAL ANALYSIS TOOL FOR TWO BIOLOGICAL STATES

by

Gang Lu

University of New Hampshire, December, 2009

The major goal of research in this thesis is to design and implement a software tool (Q5+) that can easily, quickly and reliably search biomarkers by statistically analyzing mass spectrometry data from two different biological states. Q5+ implements most of the Q5 algorithm, a very good algorithm that is used for classifying mass spectrometry data (Lilien et al). Compared Q5 Q5+ improves the usability of Q5 by incorporating a Graphic User interface and Matrix Library. Results show that by running the same data, Q5+ and Q5 showed the equivalent classification ability. Q5+ also implements the Peak Screening feature, which can be used to identify a set of peaks that may have discriminant power. Although human inspection is inevitable, it offers a way for further investigation which otherwise may not be possible only by human inspection. Overall, Q5+ is an easy and reliable tool for lab research.

CHAPTER 1

INTRODUCTION

Due to its accurate mass over charge (m/z) measurement and the application to various biological samples, Mass Spectrometry has been widely used in the studies of proteomics and glycomics. In the recent few years, Mass Spectrometry is also used for biomarker discovery and rapid clinical diagnosis (Paweletz et al). The idea is that by comparing the spectra from different groups of samples such as the samples from cancer patients vs the samples from healthy people, the difference between the two groups of samples may be discovered.

This approach will also support improvements in mass spectrometry. The introduction of SELDI (Surface Enhanced Laser Desorption Ionization) seems to make clinical diagnosis more convenient and the results more reliable. SELDI is actually a type of MALDI (Matrix-Assisted Laser Desorption Ionization) that is designed to use with pre-designed chips. As the chips have various affinities, they are used to profile part of complex biological samples.

SELDI technique suggests a new way of using mass spectrometry both in clinic and research. First, it accelerates the search of biomarkers. By comparing spectra from different types of samples such as from healthy people vs cancer

patients, a list of potential biomarkers may be discovered. Biomarkers are defined here as peaks which have discriminant power so that one or more biomarkers (either themselves or their pattern) can classify a spectrum to a certain type. The newly discovered biomarkers can also be used to define an antibody for clinical diagnosis and research. Second, the classification of an unknown spectra to a particular type (such as healthy or cancer) can also help clinic diagnosis. Maybe in the future, a drop of blood could tell whether a person has cancer. This is especially useful for early stage diagnosis and may lead towards what is called personal diagnosis.

This technique could also be useful for research especially new samples or an area in which little knowledge has gained so far. By quickly scanning proteins or glycans in the samples first, some "hot spots" may be identified before further investigation is conducted.

But this technique is still in its early stage. Generally data generated by mass spectrometry are huge and many variations are introduced during experiment processes. The variations could be caused by mass spectrometer, the sample workup, chemical reagents and samples themselves. Although mass spectrometry can be carefully calibrated, the m/z may still vary with different runs, as does the intensity. In order to detect the true differences with confidence, many algorithms have been developed to overcome the above problems.

In summary, there are three major steps which are used by most of the algorithms when processing mass spectral data, although some of the algorithms may skip one or two step(s).

The first step is peak preprocessing. The goal of this step is to reduce data points and further consolidate peaks. Many algorithms such as Ciphergen SELDI software include this step, but some algorithms skip it (Lilien et al.). The algorithms that are used in this step include baseline subtraction, mass accuracy calibration and automatic peak detection (Adam et al.). PeakMiner algorithm was also used for Peak Alignment (Adam et al).

The second step is to identify significant peaks. This step will further reduce the number of peaks. The major algorithms include MAOVA (Multiple Analysis of Variance) (Antignac et al), PCA (Principle Component Analysis) and AUC (Area under ROC Curve) (Adam et al).

The third step is to make decision based on the peaks which are chosen from step 2. The decision making algorithms mainly come from two areas, statistics and artificial intelligent (statistical learning). The algorithms from statistics include PCA (Principle Components Analysis), Disciminant Factorial Analysis (Antignac et al) and LDA (Linear Discriminant Analysis) (Lilien et al). The algorithms from artificial intelligent include Genetic Algorithm (Petricoin III et al), Neural Networks (Goodacre et al), simulated annealing and Decision Tree (Adam et al).

1.1 Introduction to Q5 algorithm

This Thesis is based on the algorithm (Q5 algorithm) published by Lilien. (Lilien et al) with further extension and improvement to the original algorithm. The Q5 algorithm includes two major steps. As shown in Figure 1-1, the algorithm doesn't contain the first step "spectra preprocessing". At the second step, PCA is used to reduce peaks. At the third step "Decision making", LDA and probabilistic classification algorithms are used for the final classification.

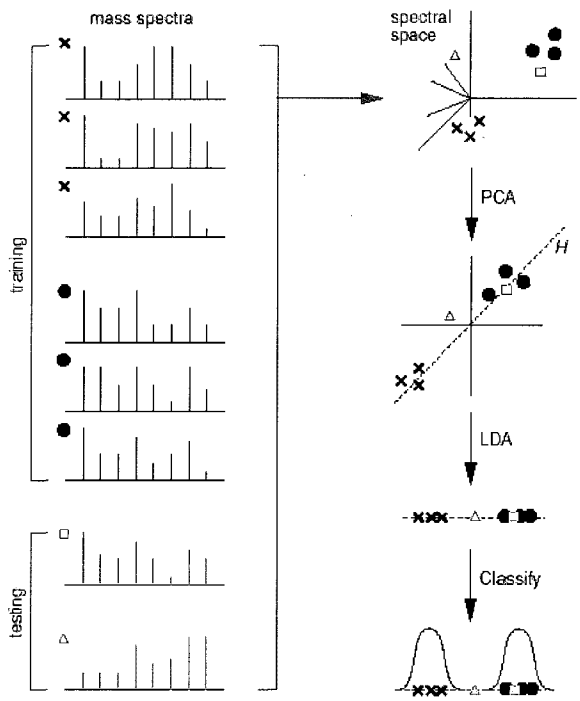


Figure 1: **Major Steps of the Q5 Algorithm.** The steps involved in building a two-class classifier are illustrated using simplified artificial spectra. On the left are training (×, ○) and testing (□, △) spectra. Shown on the right, from top to bottom, are (1) the spectral space representation of each spectrum; (2) PCA: the result of dimensionality reduction (for simplicity we show the projection onto just the top two principle components); (3) LDA: the projection of each spectrum onto the discriminant surface H; and (4) the probabilistic classification. In this example the testing spectrum denoted by □ is classified as belonging to the class denoted by ○, while the spectrum denoted by △ is unclassified.

Figure 1-1: Major steps of the Q5 algorithm (from Lilien et al)

Compared with other classification algorithms, Q5 algorithm has the following advantages. First, the algorithm has a testing step using multiple classification verification. In this step, the algorithm randomly partitions spectra to training and testing sets at each run. Because discrimination is calculated from training set, it is possible that at certain partition, the discrimination performs well to the training set and performs poorly to the testing set. By using multiple random partitions, the final performance is based on the statistics of the discrimination against multiple splits. Second, the algorithm is a combinatorial precise algorithm. Its training runtime is $O(n^3 + n^2 r)$ and testing run time is $O(mrn)$, where n is the number of training spectra, m is number of testing spectra, and r is the resolution of each mass spectrum. Third, although not implemented, it is possible for the algorithm to calculate a list of peaks which have discrimination power and contribute significantly for the classification. Finally, unlike some of the commercial software, the algorithm is free and does not bind to any type mass spectrometry.

Due to the above reasons, the Q5 algorithm seems to be a better algorithm. Q5 is implemented with MatLab. Although it makes Q5 running fast and robust, its lack of the ability of processing data from mass spectrometry directly also makes Q5 hard to use. As we know, many mass spectrometers have their own data format. Although Q5 can process several data formats including the dataset from SELDI, its ability to import data from all mass spectrometers is still limited. To use Q5, the users have to convert data from a machine format to Q5-compatible

format by themselves, which can be a very time consuming process. Other tasks that are needed to run Q5 properly such as m/z alignment among spectra and setting up running conditions also require the users to know some computer programming knowledge. In fact, as claimed by the author(s), the original Q5 implementation is mainly for demonstration purposes.

1.2 Introduction to Q5+

In order to overcome the difficulties that are mentioned above with Q5 algorithm and make Q5 algorithm a practical tool for the lab research, in this thesis, we design and develop a statistical analysis tool called Q5+ for two biological states comparison using mass spectrometry data. Compared with Q5, Q5+ is a software algorithm which implements most of the Q5 with several additional features and improvements..

Q5+ runs on Windows operating system and is implemented with C#. It incorporates the Matrix Science Library, which is a commercial COM library from Matrix Science. The library is used by Matrix Science in its own software such as "Mascot Distiller" to process various mass spectrometry data. By using this library, Q5+ can process data from several different mass spectrometers directly. Q5+ also incorporates a dialog interface from Matrix Science Library for preprocessing spectra.

Q5+ is a Graphic User Interface driven program. Importing spectral datasets is as easy as navigating via a windows dialog to the directory which contains the spectra files directly from mass spectrometry. Applying various analysis tasks to the dataset is just a few clicks from the pull-down menu.

Q5+ also implements its own matrix and math library because the math library from C# has limited functionality. The matrix and math library from Q5+ includes many commonly used matrix operations and several statistics analysis functions such as Principle Components Analysis (PCA) including eigen value and eigen vector calculation (symmetric), Linear Discriminant Analysis (LDA) including generalized eigen value and eigen vector calculations (symmetric).

Q5+ also implements several analysis functions such as "Calculate Simple Models..." which uses Q5 algorithm to verify whether there is significant difference between the two groups of spectra compared. Q5+ also implements a function to classify an unknown sample again a known spectra library, and lists peaks which may have discriminating power. In fact, Q5+ offers a platform which other algorithms can be added in easily by using Q5+'s data structure and design in the future.

Although Q5 algorithm seems to work very well on complex samples such as human serum, how it works on relatively simple samples such as the spectra with many fewer peaks remains unknown. In this study, we first ran Q5+ and Q5 with

the same dataset to show that Q5+ can do the same efficient work as its original algorithm. Then we ran Q5+ to the data from relatively simple samples with limited replications. The results show that Q5+ can successfully classify the samples and identify several peaks which may have strong discriminating powers. Because those spectra are almost identical to human eyes, the software does unearth some details which may not be found otherwise.

CHAPTER 2

DESIGN AND IMPLEMENTATION

2.1 Data Import

Q5+ is implemented with C# running on Windows. It uses the Library from Matrix Science to read in data from various mass spectrometers. First, the user needs to organize each category of spectra to a separate directory. Before the data is actually read in, the user can also preprocess spectra by modifying the preprocess settings using the dialog from Matrix Science Library, which is incorporated to Q5+. To start the dialog, choose Process > Processing Options...

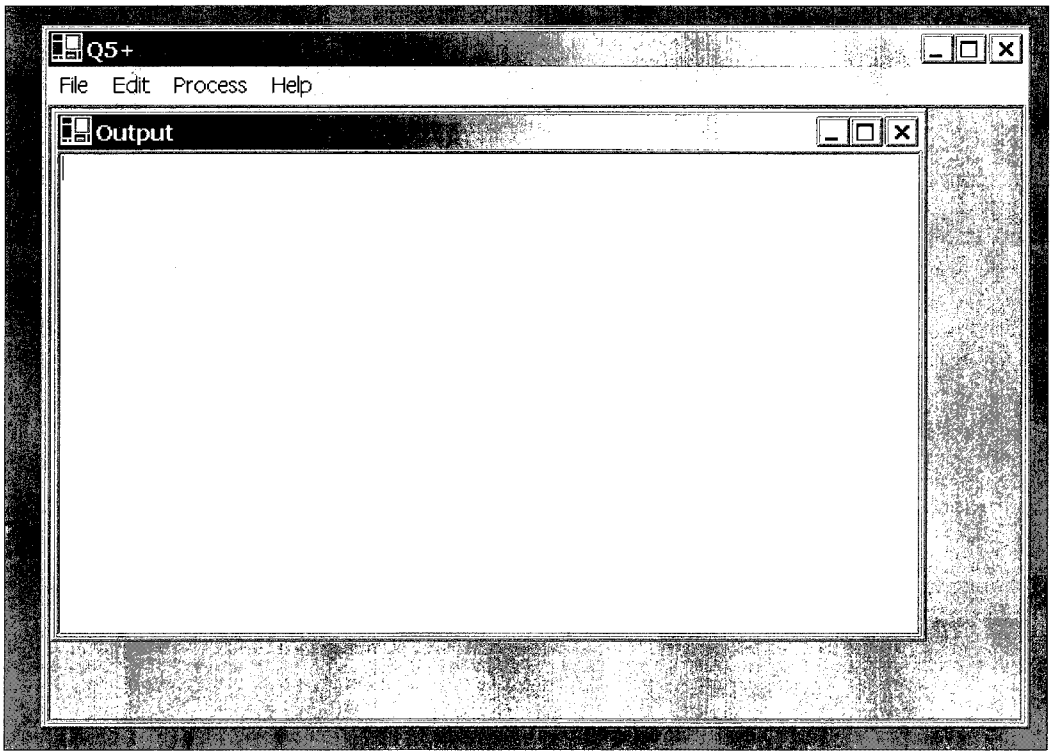


Figure 2-1: Q5+ start page

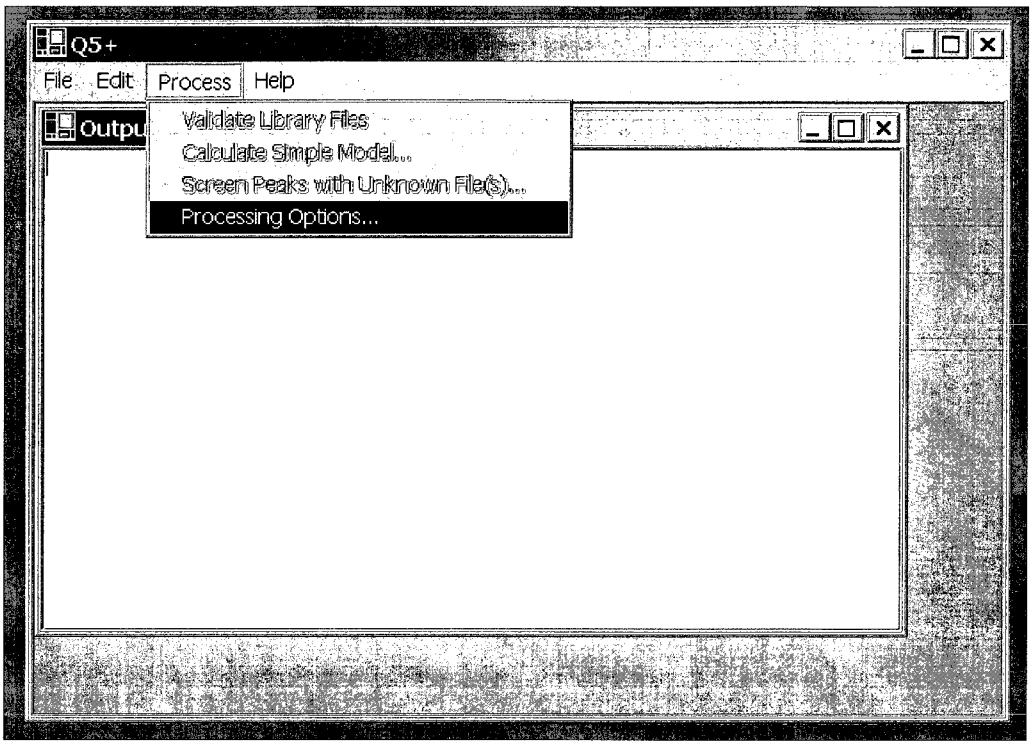


Figure 2-2: Preprocess dialog menu

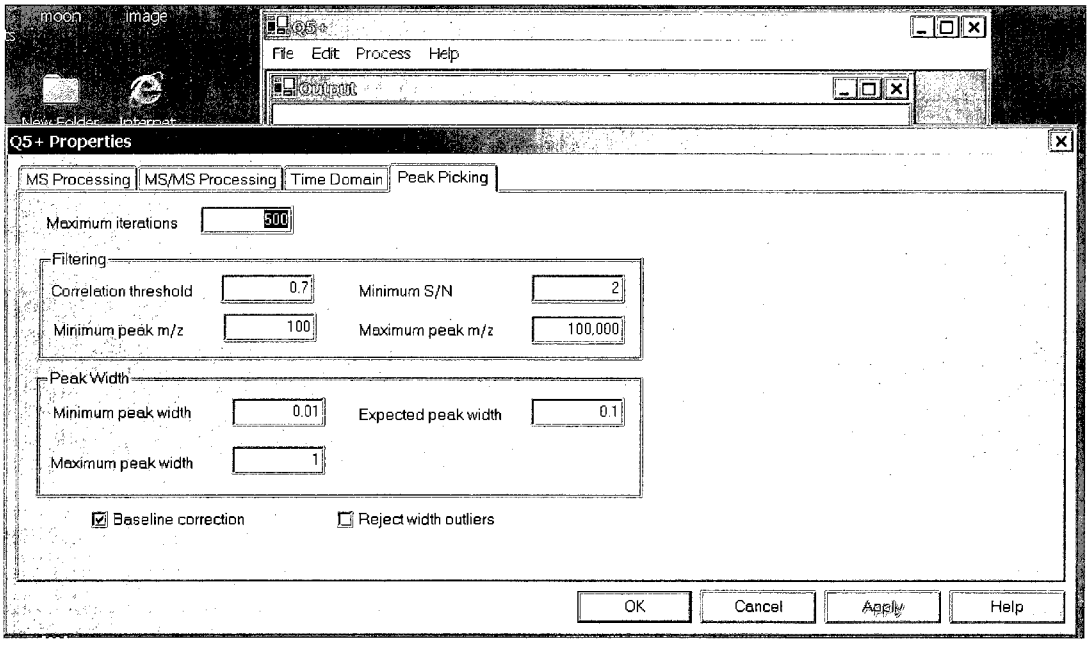


Figure 2-3: Preprocess dialog

To import spectra, choose File > New Project...

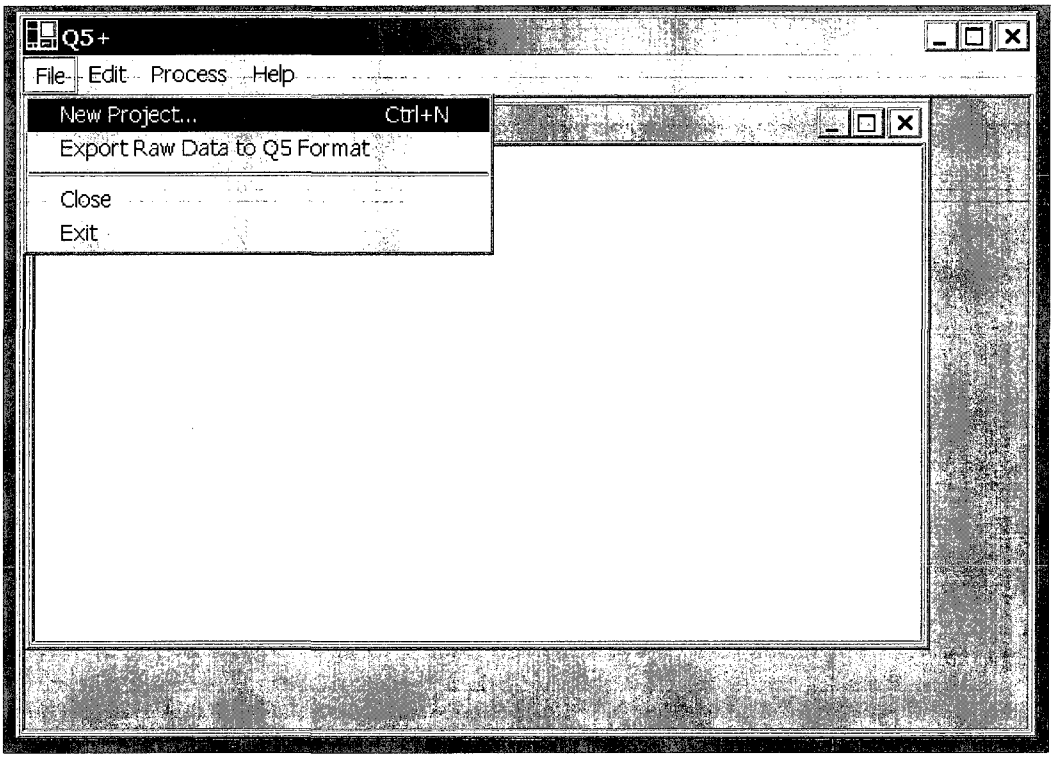


Figure 2-4: Starting a new project using Q5+

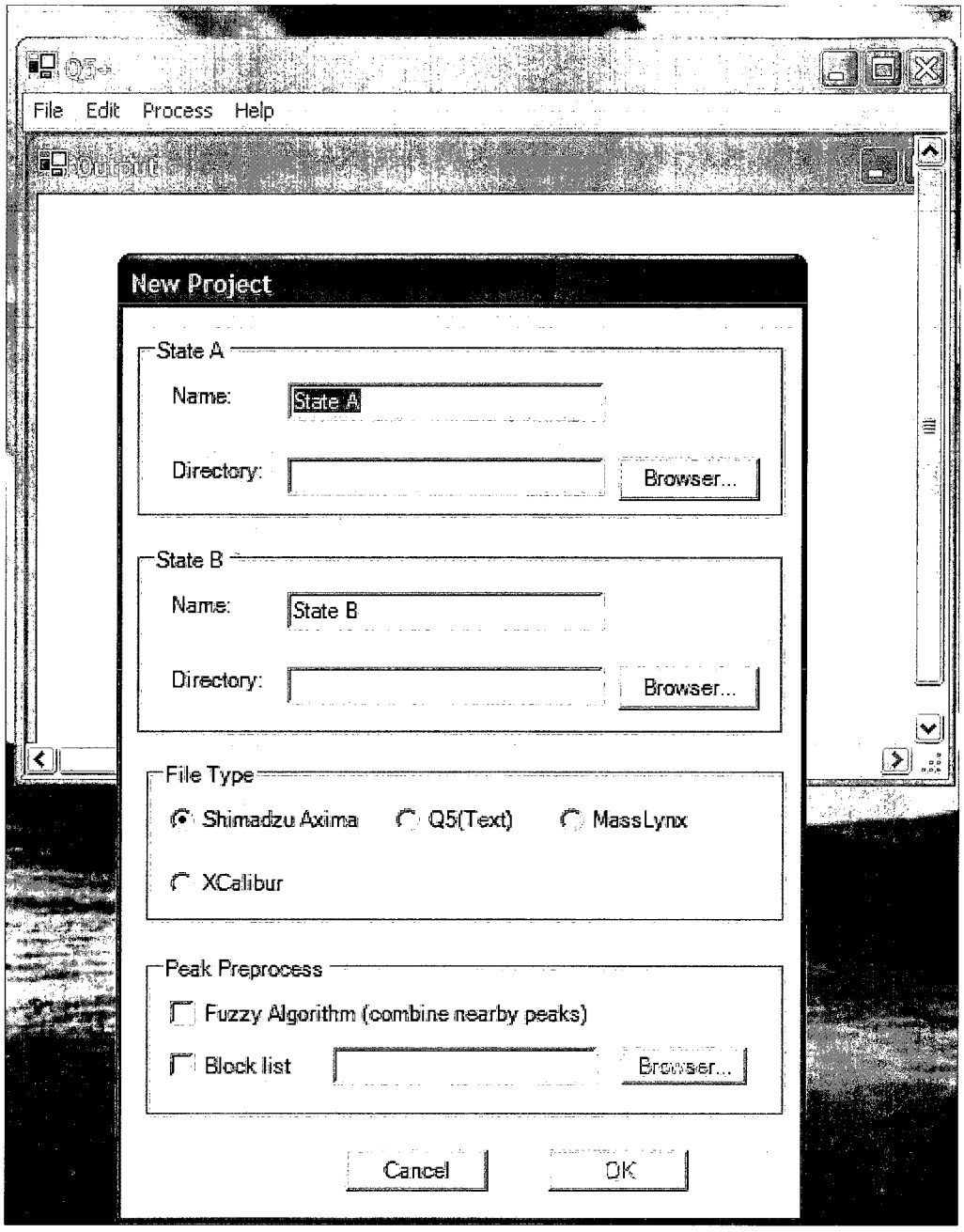


Figure 2-5: New project dialog

In the New Project dialog, the user can label the states with their own choices. The directory of each group can be filled in directly or navigated to by using the "Browser..." dialog.

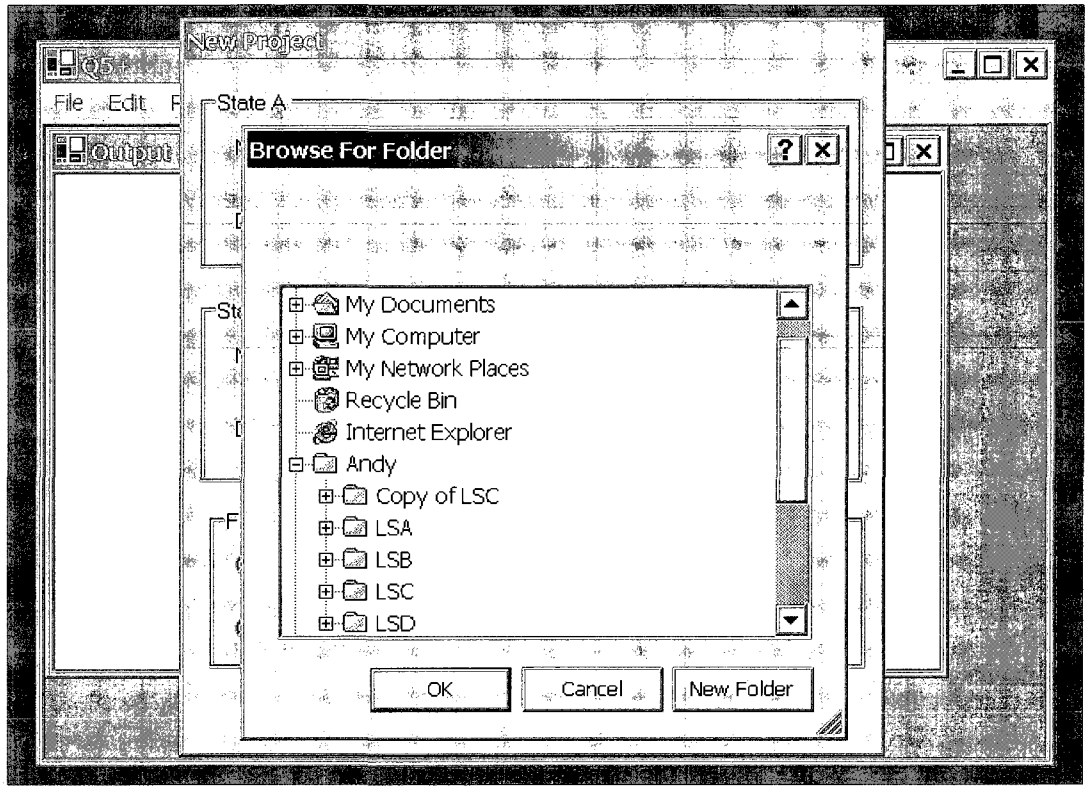


Figure 2-6: Project directory dialog

The default file type is Shimadzu Axima (MALDI). The tool can handle .run file format and .gz file format, which is the file format after spectra are processed by Kompact, the software that comes with the Shimadzu Axima instrument. Please note that if the file is in .gz format, the unzip.exe coming with Kompact has to be put to the system path and the system path only takes effect after the computer is restarted. The tool can also recognize Q5 format. Q5 format is defined here as a .txt file. The first line is a comment line and starts with %. Each line after the first line contains one pair of "m/z, intensity". The values in a pair are separated with a comma. As Q5+ depends on the Matrix Library for data import, any data format that can be handled by Matrix Library should be handled by Q5+. Figure

2-7 shows the "New Project" dialog in Mascot Distiller, a software from Matrix Science which is also use Matrix Library to handle data import.

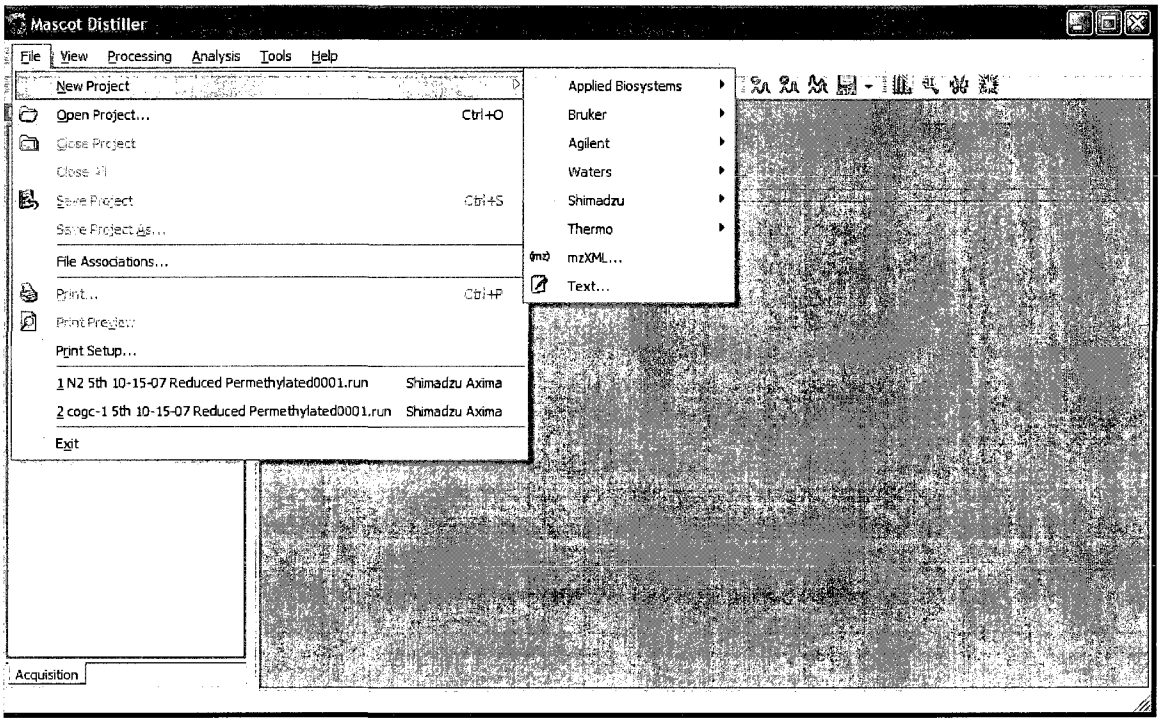


Figure 2-7: Mascot Distiller - File>New Project menu

Q5+ also offers two options for Peak Preprocessing, which are new features introduced by Q5+. By default, the two options are turned off. It is up to the user to decide whether to use these options. The first option "Fuzzy Algorithm" was introduced based on the observation that the mass spectrometry data may shift up or down a little bit at each run. The algorithm will consolidate the data points with +/-0.1dalton difference to one data point (dimension). If there is a large amount of spectra in each category, it is unnecessary to use this option because the data points nearby from all the spectra should converge to one point. But if there are not enough replications in each category, this option will help to

converge the data points. Apparently, because this feature may introduce artifacts, the user should always compare the results with the feature on and off to eliminate the potential artifacts and double-check the data points in question. The "Block list" option is used to eliminate the data points which are contained in the block list file. The file is a simple .txt file with one data point at a line. Although the "Block list" option is originally designed to eliminate internal marks (internal marks are added contaminants to the samples which are used to further overcome the m/z drifting issues at each mass spectrometry run), this option can also be used to block the predominant peaks so that the discriminating power of small peaks will be shown. Once again, the user should use with extra-caution when turning this option on.

2.2 CSBMath Library

Due to the facts that C# has a limited math library for matrix computation, Q5+ implements its own math library called CSBMath (CSBMath.dll). It contains many useful functions for Matrix computation, PCA, LDA, eigen value and eigen vectors calculation, generalized eigen value and eigen vectors calculation. Please noted that only the eigen value and eigen vectors for symmetric matrix are implemented because all the calculations are for symmetric matrix (covariant matrix). The implementation is based on the algorithms given by the book "Fundamentals of Matrix Computations" (Watkins) and is validated by running peer to peer with MatLab using the same testing samples.

2.3 Data Structure

When a "new project" is created, an internal project object will be created. The project object includes a Model object, a sorted List holding the dimensions (m/z) and the matrixes holding all the intensity values from the spectra. The indexes of matrixes are corresponding to the indexes of the dimension List.

A Model object contains a list of Run objects. The number of Runs are determined by the user input. Each Run object contains the details information about that run of calculation such as the training set and results from PCA and LDA calculation.

The data structure can also server as a platform to support other algorithms in the future.

2.4 Statistical Analysis Algorithms

The tool mainly implements two major features: Classification and Peak Screening. To start the process, go to Process >

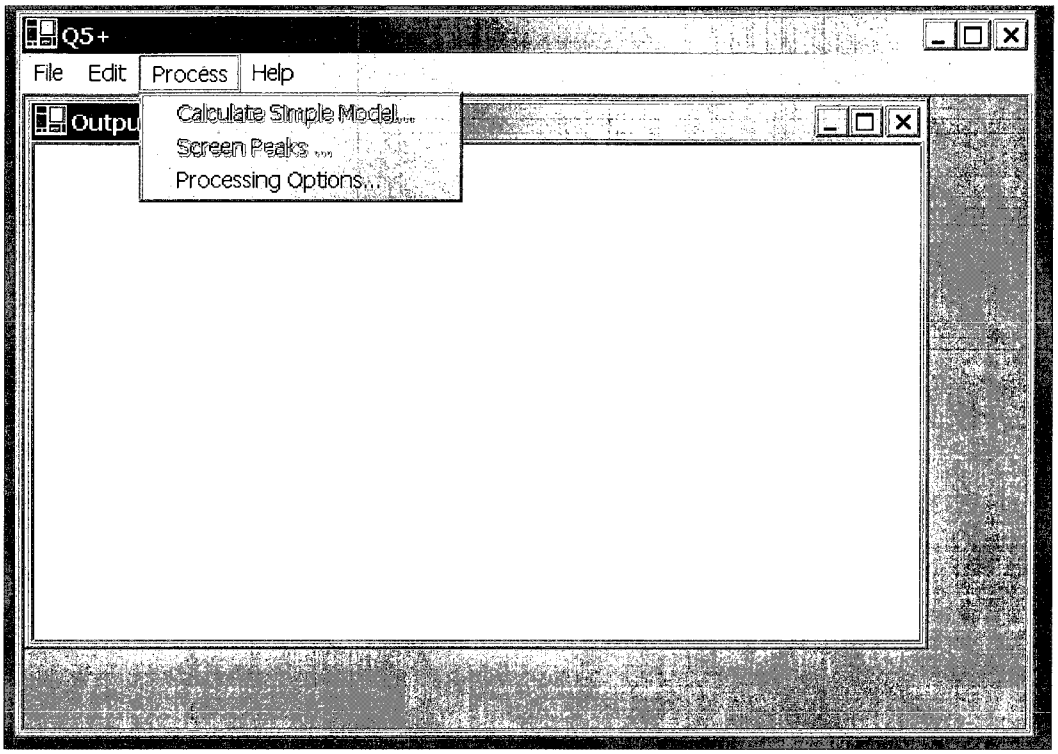


Figure 2-8: Process menu after first started

When the tool first starts, the only menu item which can be chosen from Process menu is "Processing Options..." because at that time there are not data imported to the tool (Figure 2-8). After the data are imported, the two menu items, "Calculate Simple Model ..." and "Screen Peaks...", will be enabled (Figure 2-9).

The "Calculate Simple Model ..." is used to test how well the two classes of spectra can be classified using the implemented algorithm (mainly Q5). The menu item will bring up the "Simple Model Test" Dialog (Figure 2-10) for running condition configuration. The percent of spectra for a training set and the number of tests to be run can be chosen and the default values are 85 and 10 respectively. The tool will split the spectra based on the training percentage and

run the algorithm at specified number of times. At each run, the tool will perform PCA and LDA analysis to the samples. The final results will be analyzed based on the Analysis Methods chosen. The user can have three choices, using either a simple analysis (the distance to the mean), or a statistical analysis or both. Please be noted that there are some changes to the original algorithm at this step. The simple analysis method was implemented but not used by Q5 because Q5 used statistical a method. The statistical method implemented here is different from the one that was used by Q5. Details will be discussed later in the Discuss section.

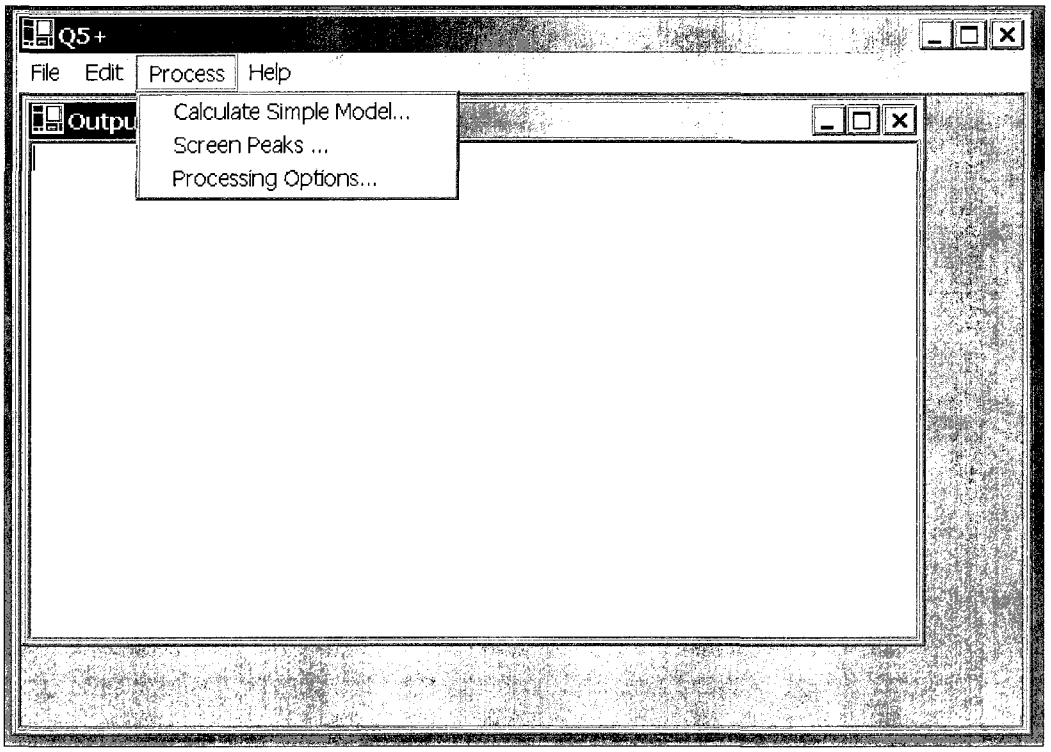


Figure 2-9: Process menu after data are imported to the tool

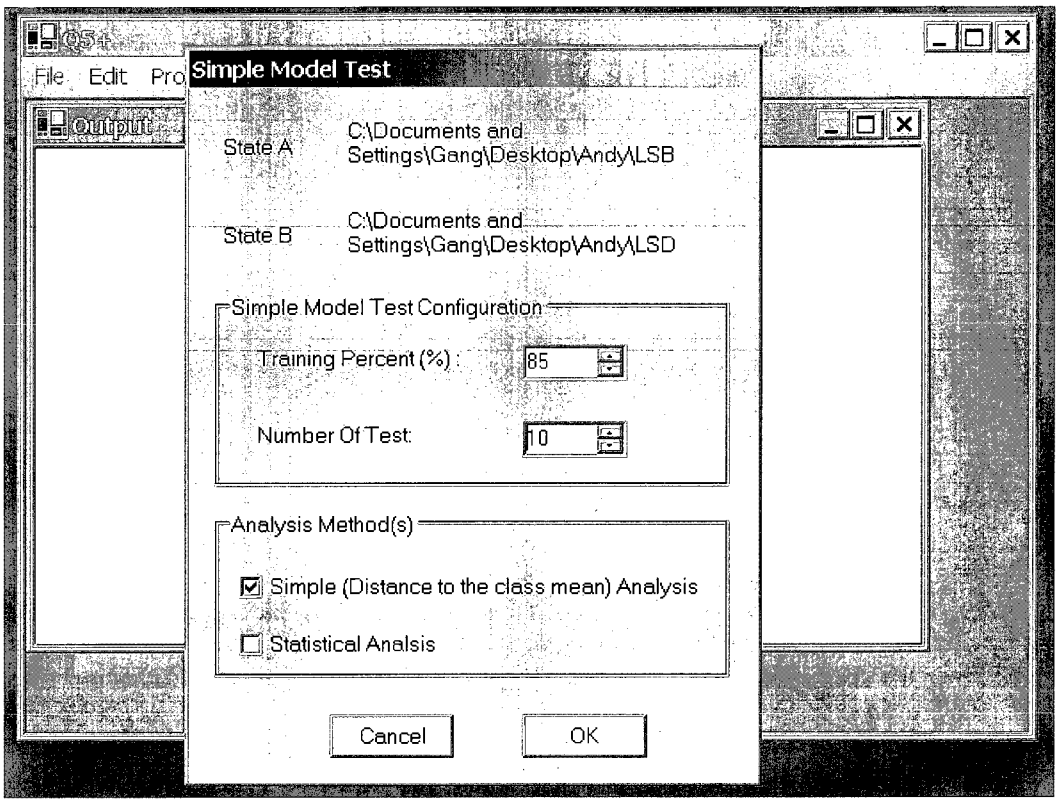


Figure 2-10: Simple model test dialog

Before using "Screen Peaks ..." function, it is strongly recommended to use the "Calculate Simple Model..." first to see how well the algorithm classifies the spectra. Only when the Algorithm works well to the classification, the peaks which are screened by the "Screen Peaks..." function seem more trustable. It is especially important to do this when using the spectra in the current library for peak screening.

The "Screen Peaks ..." menu item will bring up the "Screen Peaks" Dialog (Figure 2-11), and at this point, the user has three choices. S/he can choose to screen peaks either using spectra in current library alone, or using new files

against the current library alone, or both. The user can also choose the range of peaks to be reported using the "Peak Report" function. The first peak is always the one with the most discriminating power and will be scored as 100%. All the other peaks will be scored based on the first peak.

"Peak screening with the current library" selection will allow users to reliably analyze the spectra previous imported to the tool. One has to realize that the resulted peak list is based on the statistical average of the spectra. It may not be the case for every spectrum because each spectrum varies and may not be the typical representation for its class.

Sometimes it is useful to have information about one particular spectrum. It is especially helpful during the mass spectrometry operation because a user may want to further investigate samples such as doing MS2 to some peaks. "Peak screening with unknown file(s)" selection is designed to use for such purposes. This function will classify the unknown file first, and then print out a list of peaks that seem interesting. The reasons behind this arrangement is that an unknown file may not typically represent a class. If a file can be successfully classified to its own class, the peak list which comes from such a file is more trustable for further investigation. If not, the file can be considerable abnormal and the peak list from such file is questionable.

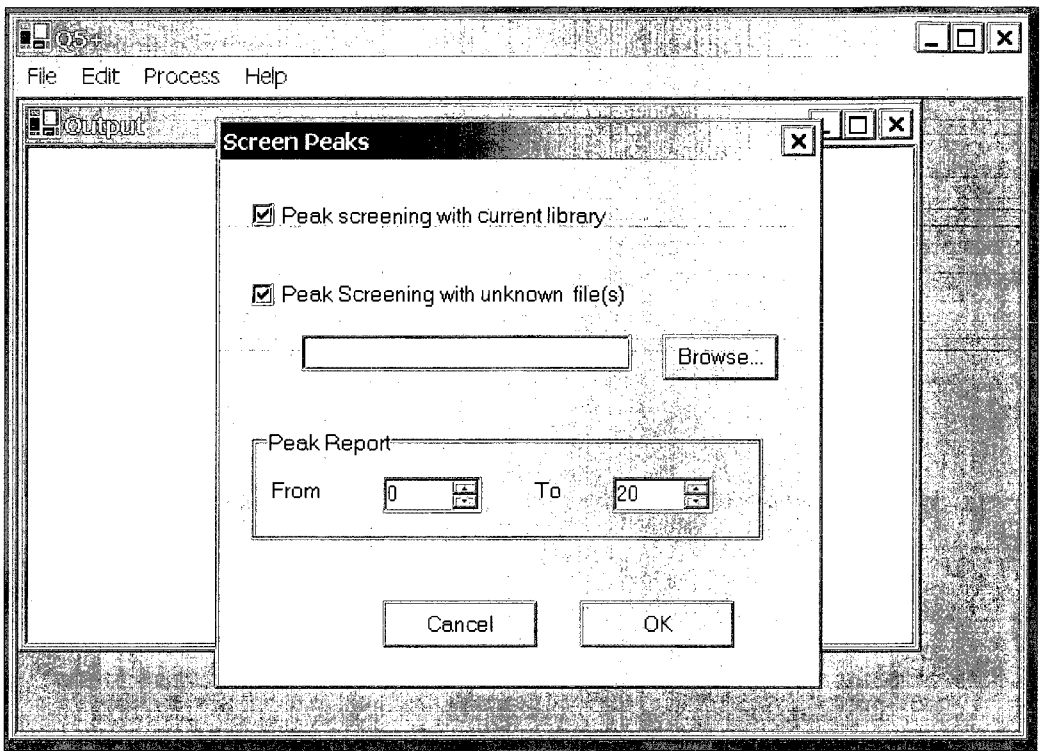


Figure 2-11: Screen peaks dialog

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Comparison with Q5 using prostate cancer dataset from Clinical Proteomics Program Databank (Petricoin. et al.)

The prostate cancer dataset from Clinical Proteomics Program Databank were used to test both Q5+ software and Q5 algorithm. The dataset includes 63 samples from healthy people with serum prostate-specific antigen [PSA] < 1 ng/mL and 43 samples from prostate cancer patients with PSA > 10 ng/mL. Sera were thawed and applied to a C16 hydrophobic interaction protein chip (Ciphergen Biosystems, Freemont, CA) and analyzed with SELDI-TOF (Emanuel F. et al.).

Table 3-1 shows the results of Q5+ running with 10 samples from each category. Overall, the results are fair as the percentages were correctly Classified, PPV, Sensitivity and Specificity were mostly around 89%. The results in Table 3-1A and Table 3-1B were run under the same conditions except that the training percentage has increased from 75% to 85%. The results suggest that the change of training percentage from 75% to 85% has little effect to the results. Table 3-1B and Table 3-1D used the same running conditions except that Table 3-1B has 10

runs and Table 3-1A has 100 runs. Each run defined here as one random partition of the samples to either training set or testing set based on the training percentage. For example, if there are 10 samples in each categories and the training percentage is 80%, in each run, 8 samples from each categories will be used as training samples to generate the model, the other 2 samples from each categories will be used as the testing/unknown samples to test the model. The selection of training samples from the 10 samples at each category is random. The results of Table 3-1C show that the percentage of correctly classified samples has improved from 87% to 90% and the sensitivity has improved from 75% to 89%. The increase of the number of runs from 10 to 100 does seem to make the results more reliable. As such, 100 runs and above are recommended for running the test. Table 3-1C and Table 3-1A used the same spectra, but the Fuzzy Algorithm option was turned on when the data were imported at Table 3-1C. Under the same running conditions, the Fuzzy Algorithm seems to improve the results a little bit (Table 3-1C vs Table 3-1A). As the spectra were from sera samples, which should contain large number of peaks, it is expected that the Fuzzy Algorithm doesn't help much to such spectra.

Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 89.67 Classified(%): 100 Positive Predictive Value(PPV%): 89.67 Sensitivity(%): 89.67 Specificity(%): 89.67 **A**	Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 90 Classified(%): 100 Positive Predictive Value(PPV%): 90.4 Sensitivity(%): 89.5 Specificity(%): 90.5 **B**

Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 90.33 Classified(%): 100 Positive Predictive Value(PPV%): 91.16 Sensitivity(%): 89.33 Specificity(%): 91.33 **C**	Result: Process 10 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 87.5 Classified(%): 100 Positive Predictive Value(PPV%): 100 Sensitivity(%): 75 Specificity(%): 100 **D**

Table 3-1: Results of Q5+ running with 10 samples from each category

A: Classification with 75% training percentage, 100 runs
B: Classification with 85% training percentage, 100 runs
C: Classification with 75% training percentage, 100 runs, fuzzy option on
D: Classification with 85% training percentage, 10 runs

Table 3-2 shows the results of using 15 samples from each category. 85% of training percentage was used for all four tests. Compared with the results of Table 3-1B, the percentages for Correctly Classified, PPV, Sensitivity and Specificity of Table 3-2B are all improved (95% vs 89% from Table 3-1B). The only difference in running condition for Table 3-1B and Table 3-2B is that Table 3-1B used 10 samples from each category and Table 3-2B used 15. The results suggest that the more samples from each category, the more reliable the results would be, which would be expected as the whole algorithm is based on statistical analysis.

One of the features of Q5+ is that once all the spectra are read in, the data can be saved/exported to Q5 format and aligned to the same MZ dimension for future use by using "File > Export Raw Data to Q5 Format" as shown by Figure 2-4. (The spectra will be export to a directory called "data" which is created if none

exists under the current directory of each category. Data from each spectrum will be saved to one file. The file name is the original spectrum file name with a ".txt" as suffix.) The data can be imported back to Q5+ later, thus cutting down the time of importing the original spectra to Q5+, and can also be directly used by Q5, thus overcoming the major obstacle of using Q5.

Result: Process 10 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 96.67 Classified(%): 100 Positive Predictive Value(PPV%): 100 Sensitivity(%): 93.33 Specificity(%): 100 **A**	Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 95.33 Classified(%): 100 Positive Predictive Value(PPV%): 97.22 Sensitivity(%): 93.33 Specificity(%): 97.33 **B**
Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 96 Classified(%): 100 Positive Predictive Value(PPV%): 98.59 Sensitivity(%): 93.33 Specificity(%): 98.67 **C**	
>> load('testout') >> processResults(res); Thresh: 0.50 : Correct: 96.17 (8.49); Classified: 100.00 (0.00); PPV: 98.60 (7.22); Sens: 94.67 (12.28); Spec: 97.67 (12.76); Thresh: 0.63 : Correct: 98.73 (5.46); Classified: 93.67 (9.70); PPV: 99.35 (4.70); Sens: 98.67 (6.56); Spec: 98.50 (11.14); Thresh: 0.75 : Correct: 99.60 (4.00); Classified: 86.33 (11.70); PPV: 99.60 (4.00); Sens: 100.00 (0.00); Spec: 98.00 (14.07); >> **D**	

Table 3-2: Results of Q5+ and Q5 running with 15 samples from each category

A: Q5+, Classification with 85% training percentage, 10 runs, data were read from spectra
B: Q5+, Classification with 85% training percentage, 100 runs, data were read from spectra
C: Q5+, Classification with 85% training percentage, 100 runs, use the exported data from B
D: Q5, Classification with 85% training percentage, 100 runs, use the exported data from B

Both Table 3-2C and Table 3-2D used the exported data from Table 3-2B. Comparing Table 3-2B and Table 3-2C, the results are equivalent, which implies that the data exported by Q5+ still consistent with the original spectra data. Comparing Table 3-2C and Table 3-2D (with 0.5 thresh value, as Q5+ doesn't implement thresh value, the results of Table 3-2D with 0.5 thresh value should be equivalent to the Q5+ results), the results are also equivalent, which implies that Q5+ has the similar classification ability as Q5.

Result: Process 100 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 95.25 Classified(%): 100 Positive Predictive Value(PPV%): 96.89 Sensitivity(%): 93.5 Specificity(%): 97 **A**	Result: Process 100 number of runs Classification Results using Statistical Analysis Method: Correctly Classified(%): 80 Classified(%): 100 Positive Predictive Value(PPV%): 72.14 Sensitivity(%): 97.75 Specificity(%): 62.25 **B**

```
>>
>>
>>
>>
>>
>>
>>
>> processResults(res)
??? Undefined function or variable 'res'.

>> processResults(res);
??? Undefined function or variable 'res'.

>> load('testout')
>> processResults(res);
Thresh: 0.50 : Correct: 95.88 ( 7.34);  Classified: 100.00 ( 0.00);
  PPV: 98.17 ( 7.54);  Sens: 94.50 (10.41);  Spec: 97.25 (11.73);
Thresh: 0.63 : Correct: 98.80 ( 4.34);  Classified: 92.25 ( 9.53);
  PPV: 99.27 ( 4.33);  Sens: 98.75 ( 5.48);  Spec: 98.67 ( 8.10);
Thresh: 0.75 : Correct: 99.83 ( 1.67);  Classified: 86.62 (10.40);
  PPV: 99.80 ( 2.00);  Sens: 100.00 ( 0.00);  Spec: 99.50 ( 5.00);
>>
```

C

Table 3-3: Results of Q5+ and Q5 running with 15 samples from each category

A: Q5+, Classification with 75% training percentage, 100 runs, data was read from spectra
B: Q5+, Classification with 75% training percentage, 100 runs, "statistical analysis' option on
C: Q5, Classification with 75% training percentage, 100 runs, use exported data from Table 3-2B

Table 3-3A and 3-3C used the same spectra as Table 3-2C and 3-2D, but with 75% training percentage. Again, the results support the above observation that the training percentage doesn't seem to affect the results much, and the results from Q5+ and Q5 are equivalent.

As mentioned above, Q5+ did not implement the thresh value. Instead, the user can choose to use either Simple Analysis Method and/or the Statistical Analysis Method option to do the classification as shown by Table 3-2B. The Simple analysis Method is closed to Q5 with thresh value of 0.5. Unlike Simple Analysis Method, the Statistical Analysis Method uses statistical decision theory and substitutes the Gaussian distributions to the maximum likelihood formula. A point is classified class A if its possibility to class A is less than or equal to its possibility to class B. To make the classification better, the "student t" distribution was used when the number of spectra in a class is less than 120.

The Q5 statistical classification method is based on a hypothetical Gaussian distribution that the probability of the midpoint to assign to the two classes is 0.5 and the probability classification threshold is chosen based on a trial basis. As such, the standard deviation of each category is based on the mean of the category and the threshold, while the intrinsic variation of each spectrum in that category is not considered. On the other hand, the Q5+ statistical classification is based on the data distribution proposed by Figure 1-1.

However, comparing the results of Table 3-3A and 3-3B, the simple method seems to have better classification overall except for sensitivity. One explanation is that because the standard deviation of each category is so small compared with the distance between the means of two categories and can even be neglected, the simple classification method seems more effective in classification.

The reasons that only part of the dataset was used were shown by Table 3-4. It took about 17 minutes to process 10 runs with 20 spectra in each category with 500 m/z as the minimum. The time will increase if more samples or runs are used. These are the two important factors to improve the classification results as discussed above.

9/12/2009 6:41:06 PM
Result: Process 10 number of runs
Classification Results using Simple Analysis Method:
Correctly Classified(%): 89 Classified(%): 100 Positive Predictive Value(PPV%): 81.97 Sensitivity(%): 100 Specificity(%): 78
9/12/2009 6:58:15 PM

Table 3-4 Results of Q5+ running with 20 spectra in each category

75% training percentage, 10 runs, preprocessed with 500M/Z as the minimum M/Z

On the other hand, it took much less time for Q5 algorithm to complete the same calculation. The difference could be caused by the different environment in which Q5+ and Q5 ran. Q5+ ran on personal computer and matlab ran on University

computing system that is a lot faster than personal computer. Another reason could be the calculation speed of the math library. Compared with matlab, which is a commercial math library that is written in C and has been greatly optimized by many developers, the math library implemented by Q5+ doesn't have that optimization feature. Q5+ is implemented with C#, an object oriented computer language which generally runs slower than C language. The Implementation of matrix with object would also take more memory and further slow the program.

To improve its performance in the future, Q5+ can switch to use commercial math library when available. The implementation of Q5+ can also be further refined by avoiding using expensive operations such as dynamic allocation.

However as shown by the above results, Q5+ has the equivalent classification ability compared with Q5 and is also very convenient to use. It can handle 15 complex spectra in each category with ease. As the number of samples and the complexity of the spectra from research lab are generally much less than those from clinic, Q5+ should be able to process the spectra from research lab without difficulties. Q5+ can also be used an adaptor to convert spectra data to Q5 format so that the data can be analyzed with Q5 easily.

A

Peak#	m/z	significant
1	**131.9**	100
2	**116.3**	81.48
3	**100.8**	65.26
4	**157.9**	51.57
5	130.7	49.24
6	**115.9**	44.87
7	116.1	42.51
8	159.6	38.14
9	**176.8**	27.39
10	156.3	25.21

B

filesC:\Documents and Settings\Gang Lu\Desktop\Test5\N1.CSV

C:\Documents and Settings\Gang Lu\Desktop\Test5\N1.CSV is classified as A

Peaks from 1 To 10:

Peak#	m/z	significant
1	**157.9**	100
2	*159.3*	75.63
3	156.5	55.86
4	103.2	54
5	**176.8**	51.11
6	**131.9**	41.49
7	**116.3**	36.75
8	191.1	34.75
9	868.9	33.1
10	**100.8**	32.17

C

filesC:\Documents and Settings\Gang Lu\Desktop\Test5\N21.CSV

C:\Documents and Settings\Gang Lu\Desktop\Test5\N21.CSV is classified as A

Peaks from 1 To 10:

Peak#	m/z	significant
1	*159.3*	100
2	156.1	75.39
3	**131.9**	66.44
4	178.2	46.08
5	184	39.44
6	160.6	38.3
7	**116.3**	31.82
8	**130.7**	30.37
9	**100.8**	27.85
10	**157.9**	24.13

D

Files C:\Documents and Settings\Gang Lu\Desktop\Test6\c1.csv

C:\Documents and Settings\Gang Lu\Desktop\Test6\c1.csv is classified as B

Peaks from 1 To 10:

Peak#	m/z	significant
1	**116.3**	100
2	*103.2*	95.7
3	125.8	87.54
4	**100.8**	79.55
5	**131.9**	71.7
6	129.3	61.8
7	132.6	57.16
8	**115.9**	47.67
9	*118*	41.03
10	**157.9**	40.23

E

Files C:\Documents and Settings\Gang Lu\Desktop\Test6\c21.csv

C:\Documents and Settings\Gang Lu\Desktop\Test6\c21.csv is classified as B

Peaks from 1 To 10:

Peak#	m/z	significant
1	**100.8**	100
2	*103.2*	72
3	221	66.52
4	197.1	59.81
5	**131.9**	53.94
6	222.9	50.34
7	219.3	44.74
8	**116.3**	39.91
9	**115.9**	35.86
10	*118*	30.87

Table 3-5: Results of Q5+ peak screening with 10 samples from each category

A. Peak list from the Model
B. and C. Screening unknown samples (healthy) with the Model
D. and E. Screening unknown samples (prostate cancer) with the Model

After the Model is set up, the Peak Screening feature of Q5+ can be used to list a series of peaks in the order of discrimination power (the power that can be used to discriminate spectra). This feature was proposed but not implemented in the original Q5 algorithm. It is still debatable whether the peak alone or the pattern of the peaks plays a role in the discrimination.

Table 3-5A shows the peaks with the current model that may have the discriminating power. The user can also classify unknown spectra with the current model as shown by Figure 2-11. However, the unknown files have to be the same file type as that of the files used to set up the model. Table 3-5B shows the results of using one of the spectra that was used to set up the Model as the unknown spectrum. The Model can classify the file to the correct category (healthy). Table 3-5C used a spectrum from the healthy category that was not used to set up the model. Again the model classified the file to the correct category. Table 3-5D is similar to Table 3-5B, a spectrum from cancer category which was used to set up the model was used as the unknown file. Table 3-5E is similar to Table 3-5C, a spectrum from cancer category which was not used to set up the model was used as unknown file. The model has classified both files correctly. The results also reveal some interesting peaks that may be worth investigating as shown by Table 3-5 (the peaks that are highlighted).

3.2 Test Q5+ with *C. elegans* samples

The *C. elegans* samples used for analysis in this Thesis are from Dr. Hanneman. The samples were prepared based on the methods described by Dr. Hanneman's report (APPENDIX A). Each sample was spotted nine times on MALDI plate and run using AXIMA-CFR MALDI-TOF Mass Spectrometry under the same power.

The data set which will be used below are from *C. elegans* Wild Type L4 vs Mutant L4 and Wild Type L1 vs Mutant L1. The spectra were preprocessed to get the best classification results. The minimum mass range was set to m/z 900 using the "Process > Process Option..." because it seems that the peaks below m/z 900 are quite noisy and the report shows that there were not many peaks of interests below m/z 900. Both Fuzzy Algorithm and Block list options were used. The block list is composed of data points 1089.4, 1089.5, 1089.6, 1293.6, 1293.5, 1293.7, 1497.7, 1497.6, 1497.8, 1701.8, 1701.7, 1701.9, which are the data points closed to the internal marks (1089.53, 1293.63, 1497.73 and 1701.83).

Table 3-6 shows the results of running both Q5+ and Q5 with the same data from *C. elegans* Wild Type L4 vs Mutant L4. Again, Q5+ seems to have equivalent classification ability compared with Q5 algorithm.

10/8/2006 2:49:56 PM Result: Process 200 number of runs Classification Results using Simple Analysis Method: Correctly Classified(%): 100 Classified(%): 100 Positive Predictive Value(PPV%): 100 Sensitivity(%): 100 Specificity(%): 100 10/8/2006 2:52:20 PM A	Thresh: 0.50 : Correct: 99.75 (3.54); Classified: 100.00 (0.00); PPV: 99.50 (7.07); Sens: 99.50 (7.07); Spec: 100.00 (0.00); Thresh: 0.63 : Correct: 100.00 (0.00); Classified: 99.75 (3.54); PPV: 99.50 (7.07); Sens: 99.50 (7.07); Spec: 100.00 (0.00); Thresh: 0.75 : Correct: 100.00 (0.00); Classified: 99.75 (3.54); PPV: 99.50 (7.07); Sens: 99.50 (7.07); Spec: 100.00 (0.00); **B**

Table 3-6: Results of Q5+ and Q5 classification with *C. elegans* Wild Type L4 vs Mutant L4

A. Classification Results of Q5+ using "Simple Analysis Method" with 85% training percentage, 200 runs
B. Classification Results of original Q5 algorithm with 85% training percentage, 200 runs

10/8/2006 11:41:19 PM

Peaks from 1 To 10:

Peak#	m/z	significant
1	912.4	100
2	917.5	98.48
3	1294	77.96
4	961.6	64.1
5	1116.5	61.93
6	1905.9	57.9
7	1323.7	57.04
8	942.4	46.15
9	1483.7	45.62
10	1187.6	42.99

10/8/2006 11:41:19 PM

Table 3-7: Results of Q5+ peak screening with *C. elegans* Wild Type L1 vs Mutant L1, pre-processing options on

10/8/2006 11:51:27 PM

Peaks from 1 To 10:

Peak#	m/z	significant
1	917.5	100
2	1279.6	84.69
3	942.4	80.62
4	961.5	74.98
5	1110.6	57.78
6	980.5	49.16
7	1146.9	37.35
8	1176.6	35.22
9	1483.8	32.83
10	925.4	30.05

10/8/2006 11:51:29 PM

Table 3-8: Results of Q5+ peak screening with *C. elegans* Wild Type L4 vs Mutant L4, pre-processing options on

Table 3-7 is the peak list results of comparing wild type L1 vs Mutant L1. Table 3-8 is the peak list results of comparing wild type L4 vs Mutant L4. Figure 3-1 shows the four spectra which are randomly picked from each category.

The peak list from the comparison of Wild Type L1 and Mutant L1 suggests that the peak 912.4 seems to be an interesting peak. It is further confirmed by visually double-checking the spectra (Figure 3-1). The m/z 912.5 is preliminary identified based on the mass as $HexNAcHex_2Fuc$ (m/z 912.5) (Appendix A).

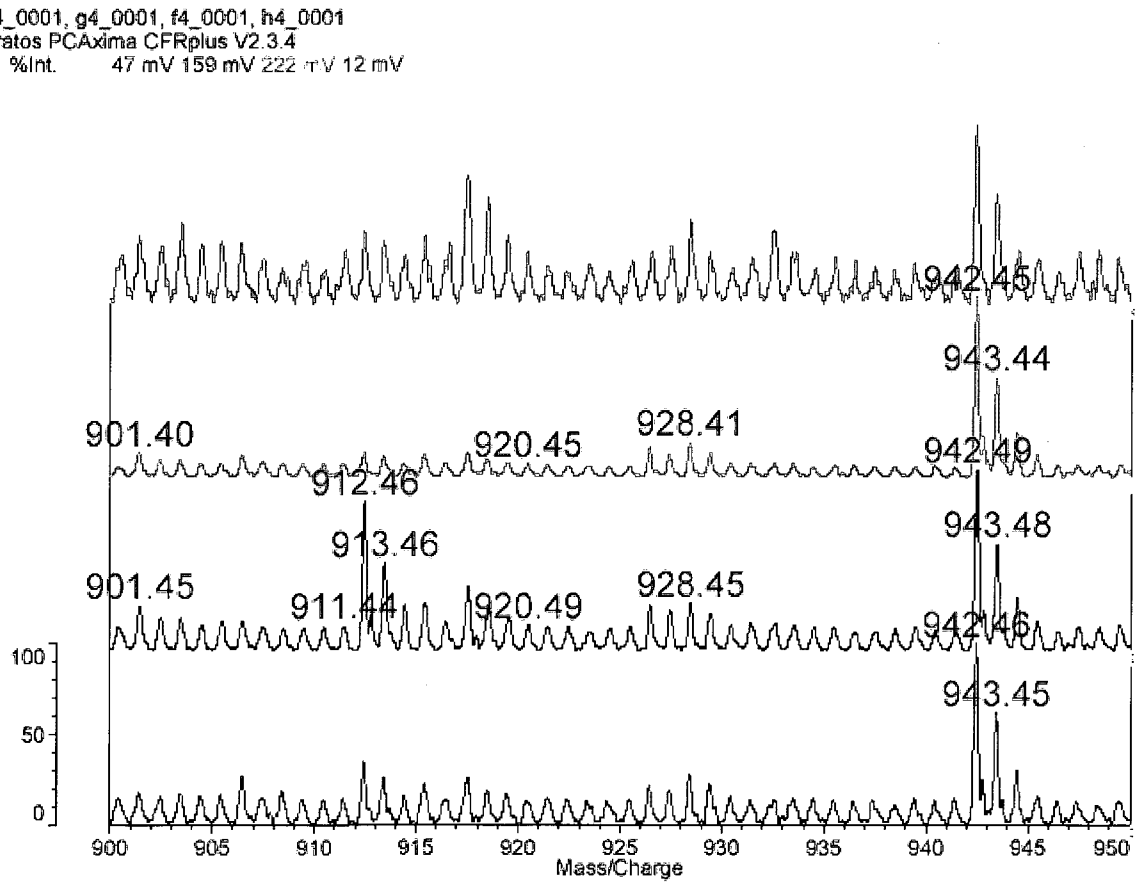


Figure 3-1: Wild vs mutant C. *elegan* glycome spectra Comparison

Red: WildType L1, Black: Mutant L1, Green: WildType L4, Orange: Mutant L4

The above results show that Q5+'s peak list function seems useful to identify a set of candidate peaks for further investigation. On the other hand, because the mathematical calculation is very sensitive, the results have to be further confirmed by the human inspection. Considering that those peaks may not be identified otherwise, Q5+ offers an easy, fast and reliable way for further investigation.

CHAPTER 4

CONCLUSION

In this thesis, I implemented a software program called Q5+ for easy, quick, yet reliably searching of biomarkers by statistically analyzing mass spectrometry data from two different biological states. First, Q5+ is very convenient to use. It has a graphical interface so that the user doesn't need the knowledge of computer programming. It uses Matrix Science library so that data can be imported directly from mass spectral data. It has its own math library so that it doesn't depend on other mathematics software, which offers a seamless integration from data import to data process/analysis. Second, it is reliable. By running the same data with Q5 peer to peer, Q5+ showed the equivalent classification ability. It can also classify unknown spectra to the correct category. Third, Q5+ can act as an adaptor to Q5. The data processed by Q5+ can be exported and used by Q5 directly. Fourth, Q5+ implemented the Peak Screening feature, which can be used to identify a set of candidate peaks having discriminating power. Although human inspection is inevitable, it offers a way for further investigation which otherwise may not be possible only by human inspection. Finally, Q5+ also implemented a few other features such as "Fuzzy algorithm" and "Block list" preprocess options that can be used to enhance the analysis. Although the time that is used by Q5+ to perform analysis is longer

than Q5, which may limit the number of samples that Q5+ can handle, as discussed above, this is generally not an issue in the research lab setting. Overall, the good feature of Q5+ is that it outperforms and can be a very useful tool for research.

REFERENCES

Bao-Ling Adam, Yinsheng Qu, John W. Davis, Michael D. Ward, Mary Ann Clements, Lisa H. Cazares, O. John semmes, Paul F. Schellhammer, Yutaka Yasui, Ziding Feng and George L. Wright, Jr. "Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men" Cancer Research 62, 3609-3614, July 1, 2002

Jean-Philippe Antignac, Bruno Le Bizec, Fabrice Monteau and Francois Andre "Differentiation of betamethasone and dexamethasone using liquid chromatography/positive electrospray tandem mass spectrometry and multivariates statistical analysis", J. Mass Spectrom. 2002, 37: 69-75

Royston Goodacre, Eadaoin M. Timmins, Rebecca Burton, Naheed Kaderbhai, Andrew M. Woodward, Douglas B. Kell and Paul J. Rooney, "Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks", Microbiology 1998, 144, 1157-1170

Ryan H. Lilien, Hany Farid, Bruce R. Donald, "Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum ", Journal of Computational Biology, 2003, 10(6): 925-946

Cloud P. Paweletz, John W. Gillespie, David K. Ornstein, Nicole L. Simone, Monica R. Brown, Kristina A. Cole, Quan-Hong Wang, Jing Huang, Nan Hu, Tai-Tung Yip, William E. Rich, Elise C. Kohn, W. Marston Linehan, Thomas Weber, Phil Taylor, Mike R. Emmert-Buck, Lance A. Liotta, and Emanuel F. Petricoin III, "Rapid Protein Display Profiling of Cancer Progression Directly From Human Tissue Using a Protein Biochip", DRUG DEVELOPMENT RESEARCH, 2000 49:34–42

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta "Use of proteomic Patterns in serum to identify ovarian cancer", The Lancet, Vol 359, 572-577, February 16, 2002

Emanuel F. Petricoin III, David K.Ornstein, Cloud P. Paweletz, Ali Ardekani, Paul S. Hackett, Ben A. Hitt, Alfredo Velassco, Christian Trucco, Laura Wiegand, Kamillah Wood, Charles B. Simone, Peter J. Levine, W. Marston Linehan, Michael R. Emmert-Buck, Seth M. Steinberg, Elise C. Kohn, Lance A. Liotta, "Serum Proteomic Patterns for Detection of Prostate Cancer", Journal of the National Cancer Institute 2002, 94(20):1576-1578

Prostate Cancer Studies Dataset:

David S. Watkins, "Fundamentals of Matrix Computations", Second Edition, 2002, John Wiley & Sons, Inc.

APPENDICES

APPENDIX A: PROCEDURE AND RESULTS FOR *C. ELEGANS* SAMPLES

By Andrew Hanneman, Hailong Zhang and Gang Lu

SAMPLE PREPARATION:

C. elegans L1 and L4 growth stages of knockout GLY002, as well as corresponding wild type L1 and L4 stages were received frozen on dry ice. The samples were briefly thawed, mixed thoroughly, and each was divided in half by volume and dispensed into two tared vials. The samples were re-frozen; one set was stored away and one was lyophilized in pulse tubes (tubes for protein lysis using the Barocycler, *Pressure Biosystems*). Freeze-dried sample aliquots were weighed after drying.

SAMPLE	Dry wt.
N2 L1 (68)	26.8
N2 L4 (69)	43.5
GLY002 L1 (70)	13.9
GLY002 L4 (71)	25.5

Lysis buffer (1.3 ml) was added to each sample (lysis buffer Recipe: 7M Urea, 2M Thiourea, 75mM C7BzO detergent, 100mM DTT). The samples were taken through the standard Barocycler program (10 cycles of 35kPa). Following centrifugation
(15 minutes 5k rcf) fresh lysis buffer was added to each sample and the pressure treatment repeated. The two lysates were combined, along with a rinsing volume, bringing the final extract volume to ~3 ml. The combined lysate was centrifuged at 12k rcf for 30 minutes and the supernatants transferred to 15 mL Falcon tubes. Protein concentration was measured by Bradford assay:

SAMPLE	Protein conc.	~Protein Yield
N2 L1 (68)	1200 ug/ml	3.6 mg (13% of dry weight)
N2 L4 (69)	2300	7.0 (16%)
GLY002 L1 (70)	1000	3.0 (21%)
GLY002 L4 (71)	1600	4.9 (19%)

Conductivity measurements ranged from 1900-3000, indicating need for clean up prior to 2DGE.

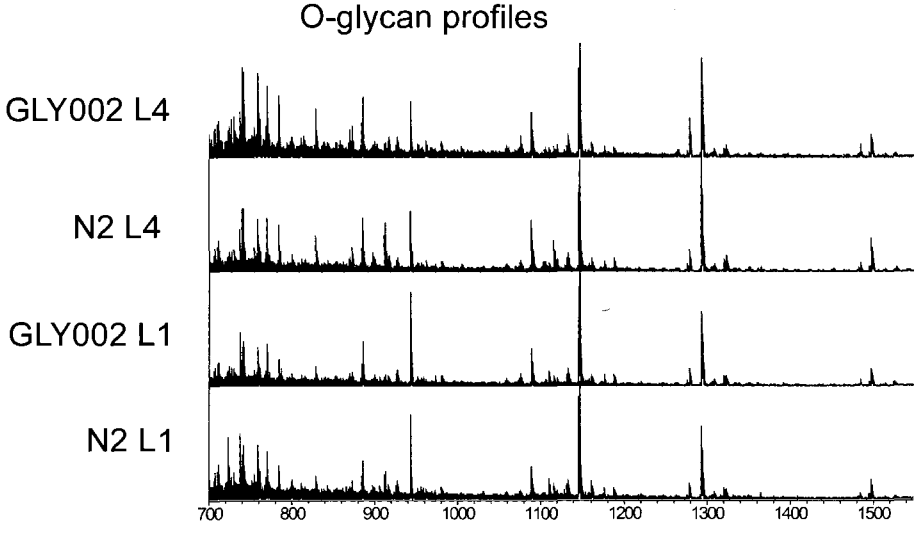
2D GEL ELECTROPHORESIS:

An aliquot of each sample containing 200 ug of protein (by Bradford) was dispensed into a 500uL 10kDa ultrafiltration (UF) device. The pH was adjusted to 8-9 by addition of TRIS to 40 mM; alkylation was carried out by addition of

acrylamide to 200 mM, and the samples allowed to sit for 1.5 hours at room temperature. The samples were ultrafiltered and brought to 250 uL with CHAPS re-suspension buffer for IEF. Each sample was run by 2DGE using pH 3-11 IPG strips. The gel images are attached.
Notes: In addition to UF, one set of samples was acetone precipitated prior to IEF- this produced gels with significantly fewer proteins than UF. A third set of duplicate gels was run using two levels of higher protein loading – these produced streaking in the acidic protein region. While not of excellent quality together the gels confirmed proteomic differences among the samples (duplicates of the same sample looked very similar to one another).

O-GLYCAN PROFILING BY MS

O-glycans were released by reductive beta elimination from 200 ug of protein and profiled by MALDI-MS. The glycans were also permethylated and profiled for confirmation and to include acidic glycans in profiles. Overall, the profiles are not apparently different from one another, or from mixed stage N2 profiles as previously observed in our lab.



Any glycan peaks observed by MALDI are indicated in the table below.

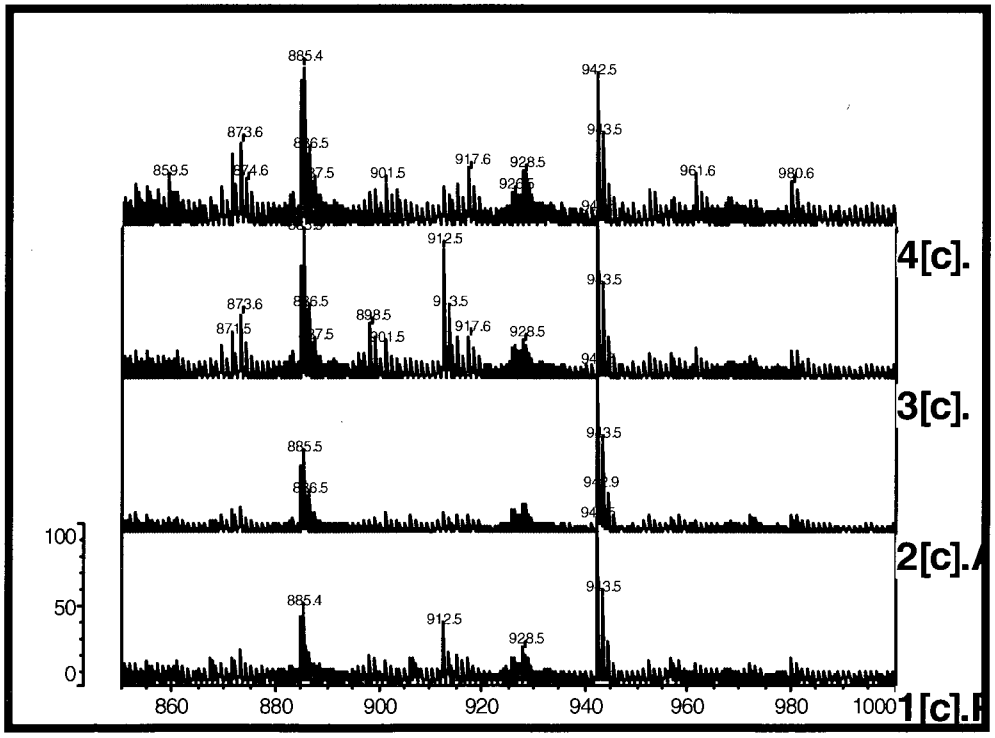
Mass (reduced)	Mass (methylated)	N2 mixed stage	N2 L1	N2 Type L4	GLY 002 L1	GLY 002 L4	Glycan Molar composition			
							Hex NAc	Hex ose	Fuc ose	Hex. Acid
570.2	738.4	*	*	*	*	*	1	2		
716.3	912.5	*	*	**?**	*	**?**	1	2	1	
732.2	942.5	*	*	*	*	*	1	3		
746.4	956.5	*	nd	nd	nd	nd	1	2		1
878.3	1116.6	*	*	*	*	*	1	3	1	

894.3	1146.6	*	*	*	*	*	1	4		
908.4	1160.6	*	*	*	*	*	1	3		1
1070.5	1364.7	*	*	*	*	*	1	4		1
1202.5	1524.8	*	*	*	*	*	1	5	1	

C. elegans *O*-glycan comparisons

To improve the spectral comparisons hexose polymer peaks -background carbohydrate contaminants- were used as internal mass standards to calibrated within 50 ppm mass accuracy. The bio-informatics tool "GlycoScreen" developed by Hailong Zhang in our lab was used to survey the adjusted spectra for possible glycan peaks that may have been missed by visual examination. The results indicated that the glycan HexNAcHexFuc (m/z 956.5), typically observed in mixed stage N2, was not observed in any of the sample spectra. Comparison with an archived spectrum of mixed stage N2 *O*-glycans indicated this glycan is typically at low abundance, and may have been below the detection limit of this analysis (indicated as n.d. in the table – also see below, *the top trace is the archived spectrum*).

The software "Q5+" was used to make spectral comparisons among samples by Gang Lu. For this analysis a set of 12 replicate MALDI spectra from each sample was used. Overall, the spectra were found to be very similar; however, intensity differences for the glycan: HexNAcHex$_2$Fuc (m/z 912.5) were noted between the L1 and L4 stages– see below.



APPENDIX B: A BRIEF EXPERIMENT DESIGN GUIDE FOR Q5+ USER

In order for Q5+ software best works for you, here are some suggestions to consider when you design the experiment.

When designing experiments for comparing different category samples, it is recommended to start with the same amount of samples from each category. For example, for *C. elegant*, the same number/weight of worms in the same growing stage from each category should be picked respectively.

Parallel operations are preferred. It is also recommended to adjust the amount of samples at some point during the experiment. For example, it is nice to adjust the amount of samples after extracting the worms because worms are not easy to be extracted and the productivity may vary quite a lot. The adjust measurement depends on the purpose of the study. For example, if the study purpose is glycome, it is better to adjust the samples using glycan. If not available, protein concentration may be used for adjustment.

When spotting to the MALDI plates, it is recommended to spots at least 2 spots from each category first for testing purpose. During the testing state, try to find an appropriate Power and concentration of samples so that the signal to noise ratio are good enough to all the samples. Because it is not known how the Power and sample concentration will affect spectra, it is recommended to use the same power to both categories of samples and try to adjust the sample concentration

to a similar level. It might be fine if the signal and noise ratio is not optimal because the Q5+ software can filter out the noise during the comparison process.

After the testing stage, it is recommended to spot multiple spots from each sample such as 10. It is OK to spot in the pattern like ABABAB..., it will be better if the samples are spotted in a random pattern such as AABABBBAA.... The reason for duplicates is that it can offset the variants which are caused by the machine at different run. Q5 software also needs multiple spectra for it training set and testing set. With duplicates, the software can make more reliable decision.

Q5+ results Interpretation. If the software can predict that the two categories of samples are significant different, which suggests that there are significant different peaks between the two categories of samples. The recommended peaks are the peaks which have discriminant power and contribute most to the discrimination of the two categories of samples.

Although the significant peaks which are screened by the two categories of samples may not be the typical representation of the two categories and appear in every spectra. More repetition of the experiments from different samples may find more significant peaks for the category. Standardizing the whole experiment procedures (starting with the same number of worm, adjusting the sample

concentration, fixing the MALDI power) could allow later to compare the results from different experiments and find the significant peaks of the category.