# Computational Analysis of Mass Spectrometry Data Using Novel Combinatorial Methods

A. Fadiel†*, M. A. Langston‡*, X. Peng‡, A. D. Perkins‡, H. S. Taylor†,
O. Tuncalp†, D. Vitello†, P. Pevsner† and F. Naftolin†

†*Yale University School of Medicine, OBGYN Department and Yale Center for Research in Reproductive Biology, New Haven, CT 06520, USA*

‡*Department of Computer Science and Graduate School in Genome Science and Technology, University of Tennessee, Knoxville, TN 37996-3450, USA*

## Abstract

*The analysis of proteome profiles offers a new approach to understanding how cellular machinery functions and responds under certain conditions. By combining two-dimensional electrophoresis with mass spectrometry (MS), a snapshot of the cell's protein expression status and quantitative proteome profiling can be provided. As the cell's proteome becomes defined in normal and altered states, possible utilization of MS proteome profiling as a diagnostic tool becomes a reality. The ability of Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI-MS) to generate a spectrum with thousands of data points, necessitate the development of sophisticated analytical algorithms. In this paper, we describe how MALDI-MS can be used in monitoring proteomic profile in patients before and after treatment using a non- invasive sampling method. Because data analysis in this process possesses a challenge, we present a novel mathematical approach for analyzing data produced by MALDI MS, and discuss current applications of mass spectrometry in clinical medicine as well as challenges faced during procedures and experimentation. As a case study, we analyze protein expression patterns in premenopausal versus postmenopausal women. We also provide a proteomic profiling of premenopausal women versus postmenopausal women treated with estrogen as a hormone replacement therapy.*

***Corresponding authors:**

**Ahmed Fadiel, Ph.D., Yale University School of Medicine, FMB 328, 333 Cedar Street, New Haven, CT 06520, USA, phone: 203-737-1218, email: afadiel@yale.edu**

**Michael A. Langston, Ph.D., Department of Computer Science, University of Tennessee, 203 Claxton Complex, 1122 Volunteer Blvd, Knoxville, TN 37996, USA, phone: 865-974-3534, email: langston@cs.utk.edu**

# 1. Introduction

Traditionally, molecular mass has been measured using size exclusion chromatography (SEC), which measures the hydrodynamic volume, and indirectly the actual mass. By coupling high performance liquid chromatography with mass spectrometry, Matrix Assisted Laser Desorption/Ionization (MALDI) mass spectroscopy (MS) was developed to measure molecular weight averages with very distinct distributions and has been widely used in both drug development [1] and proteomics. In 1987, Karas and Hillenkamp successfully performed laser desorption ionization on a small organic molecule attached to a matrix [2]**.** One unexpected side effect of the matrix was that it allowed for the laser incidence spot to be refreshed between each pulse, greatly enhancing shot-to-shot reproducibility [3]. Subsequently, MALDI was performed on a wide range of biological macromolecules [4] and proved to provide a range of opportunities in clinical and scientific studies and mainly used for protein analysis.

Recent advances in mass spectrometry and materials analysis bring new capabilities to monitoring proteomic changes associated with human disease. Many of these advances incorporate technologies acting at the molecular level that facilitate the ionization, separation and identification of unknown compounds. Successful processing of unknowns can often depend on the specific properties of the sample under study, the method of mass spectroscopy employed, and the reference library used for comparison. Specifically, improvements in mass resolution, detection power and dynamic range are continuously being developed by improving time of flight detection, ion trapping efficiency and capacity [5-7]. Ultimately, physicians and researchers hope to develop instrumentation that can not only determine the composition of unknown biological samples, but also identify salient material characteristics that distinguish healthy from diseased patient tissues. Changes in peptides and proteins, at levels that can be objectively measured and evaluated as indicators of normal and pathogenic processes, provide means for molecular diagnostics [8]. Profiling proteomic changes can benefit immensely from advancements in mass spectrometry beyond simple measurements of protein molecular weights.

## 1.1 Diagnostic Proteomics

Protein changes during a disease and/or a treatment can be studied with MALDI, gel electrophoresis and bioinformatics tools [9, 10]. Identifying patterns of proteins' expressions can help reveal correlations of disease risk, disease progression and biological response to treatment. Mass spectrometry has an important role in clinical laboratories in areas that range from newborn screening [11], to detecting drug usage in athletes [12], and even toxicological and forensic applications [13]. An association between immune and oncogenic response and changes in total proteome profile has been established. In addition, correlations between specific protein markers related to specific pathways such as apoptotic cascades and the role of Toll Like Receptors in inflammation and high expression of anti-apoptotic protein, BAG-1, in invasive breast carcinomas [14]**,** and CRP, the plasma C-reactive protein, for systemic inflammation, were also recognized as strong predictors of heart attack and stroke in postmenopausal women. In cancer research, complex mass spectrometric proteomic patterns have revealed serum differences between patients with and without cancer [15]. Currently MS technologies are being utilized in investigating proteome over-expressed proteins in lung tumors, and reported two differentially expressed proteins: migration inhibitory factor (MIF) and cyclophilin A (CyP-A). Such work represents an early chapter in studying novel molecular targets for cancer diagnostics and therapeutics [16].

## 1.2. Proteomic Change in Menopause

Menopause is defined as the cessation of menstrual periods in women that occurs about 50 years of age. During menopause, many women experience symptoms including mood disorders, osteoporosis, hot flushes, cognitive dysfunction, and cardiovascular diseases. Although a vast amount of information is available on menopause, the progression of the above-mentioned conditions continues to affect the quality of life for millions of women every day. To date, little is known about proteome changes during menopause transition. This promises to open up an expanding field in which MALDI-MS techniques can assist in charting the mechanisms of hormonal activities, as well as determining how menopause may be linked with diseases related to aging and menopause [17-21]. MALDI can also be used with buffers containing high salt concentrations, i.e., at physiologically relevant conditions**,** such as those found in human blood. Such assays may facilitate analysis of specific protein activities associated in hormonal, homeostatic, disease and aging processes. For example, physiological changes resulting from menopause can now be studied in detail using mass spectrometry.

Because MS is increasingly used for protein profiling, significant challenges have arisen with regard to analyzing the data sets. These include peak identification and alignment, MS spectrum normalization, and data set visualization, among others. These pre-processing steps are arguably critical and we are currently evaluating them carefully. The final and most important step is the identification of reliable diagnostic markers that can distinguish between two subjects under different treatment regiments. Currently there are many publications that explore implementing MS in proteomic diagnosis using serum samples from patients. The need for large sample size from each patient, the need for elimination of immunoglobulin and the invasiveness of the technique, diminishes the potential of utilizing MS using serum samples. Here we utilize MALDI MS to obtain a proteomic profiling for the first time from urine samples proving the feasibility of a non- invasive sampling method. This technique substantiates the postulation that urine can be used as a source of proteins to monitor pathological changes.

## 2. Data Acquisition

We have obtained urine samples from women at three different stages of their reproductive life cycles: premenopausal women (the negative control group), postmenopausal women (the control group), and women treated with estrogen (the experimental group). The postmenopausal women were enrolled in the KEEPs trial [22, 23]. Proteins were extracted from urine samples. The samples were then subjected to automated desalting and MALDI-MS on a Micromass M@LDI-R instrument as described generally at http://info.med.yale.edu/wmkeck/prochem/biomarker.htm. This data set consists of MS spectra that extend from 800 to 3500 Da (Daltons) obtained on urine samples from random subjects selected from the aforementioned groups. Each sample was replicated with MS to average out differences due to technical errors. The MALDI analysis was done on a Waters MALDI-L/R mass spectrometer and acquired in both linear and reflectron, positive ion detection modes. The mass range acquired was dependent on the mass analyzer being used, with 700-3500 Da being used for reflectron analysis. Linear data over the 3450-28000 Da range was collected and the two spectra were merged at 3500 Da to produce a continuous reflectron + linear spectrum ranging from 700 – 28000 Da. Although the mass range was adjustable, meaningful data was not usually acquired below about 700 Da due to interference from the matrix. The linear analyzer was used for the high mass region because with the reflectron analyzer the sensitivity of detection decreases substantially above 3,500 Da. The MALDI-L/R sums 10 individual laser shots into one spectra with the laser operating at 20 Hz (i.e., 20 shots/second), acquiring new spectrum every ½ second. The laser moves in a random walk around the target well, acquiring data from a maximum of 20 different locations within each well. A spectrum was considered *acceptable* if it had a signal of greater than 2% above background noise, less than 95% of saturation, and at a minimal level of one m/z detected between 1,125 Da and 3,500 Da. Each (averaged) reflectron and linear MALDI-MS spectrum was converted to a text file listing of m/z versus intensity data points spanning the m/z range from 700-3500 and 3450 to 28,000 that was then suitable for further analysis. Data points in the linear MALDI-MS, overlapping with the 3450 to 3500 region of the reflectron spectrum were deleted, therefore resulting in the "continuous" spectrum from 700 to 28,000 daltons with the reflectron/linear "breakpoint" appearing at 3,500 Da. The expected mass resolution was 14,000 at M+H 2,465 and the mass accuracy was better than approximately +70 ppm. All spectra were visually screened. Poor quality spectra were then re-shot manually from the same targets. If the manually acquired spectrum appeared to be visually superior to the automatically acquired spectrum then the file was over-written with the manually acquired spectrum.

## 3. Pre-Processing

For this pilot study, only reflectron data was employed. In total, 16 spectra were collected from the control group, which contained older post-menopausal women without estrogen treatment. Thirteen spectra were from the experiment group, which were older post-menopausal women treated with estrogen. The negative control group contained younger pre-menopausal patients. Eight spectra were collected from this group. The mass spectra were pre-processed using the PROcess package in Bioconductor [24], with some modifications (Figure 1). In brief, both the m/z and intensity values were log-transformed. The background of each spectrum was estimated using loess method and then removed. The set of spectra was re-normalized to their median Area Under the Curve (AUC), where an AUC is calculated for all of the m/z values. Peaks were detected and aligned using default parameters. Only peaks with intensities greater than the 80th percentile were kept. Centers of the resulting intervals were defined as the locations of the aligned peaks (P-Markers). For each spectrum, actual peaks represented by an aligned peak were determined and the maximum of those was defined as the height of the (P-Markers). If there were no peaks, then the intensity was defined as zero. Only those peaks which had greater than zero intensity in 15% or more of total numbers of spectra were kept for further analysis.

## 4. Peak Scoring

We employ a scoring process to identify peaks that best distinguish one sample type from another. To accomplish this, we start with the scoring method we first developed in [25] for transcriptomic analysis, and modify it appropriately for proteomics data. See procedure *peak-score-and-select* in Appendix B for details. With this technique, we are able to assign higher scores to more highly differentiated peaks Figure 2 shows a typical histogram of peak scores. A series of thresholds (0, 0.2, 0.4 and 0.6) was applied to search for appropriate cutoff values. Visual inspection of the separation of distributions of weights between groups and within groups suggested 0.4 as an appropriate score cutoff value (Figure 3). Weights are then assigned to pairs of peaks. The assignment of such weights was performed also using methods adapted from our work in [25]. In this weighting technique the weight between samples *i* and *j* indicates the degree of similarity in their peak profiles:

$$weight(i,j) = \sum score(peak_k) \bullet (1 - |\operatorname{int}ensity\_value_{ik} - \operatorname{int}ensity\_value_{jk}|)$$

The R statistical package [26] was used to help make score and weight calculations.

This process produces three symmetric matrices containing respectively the weights between patients in the experiment and negative control, control and negative control, and control and experiment groups, respectively. These weight values are then scaled to the interval [0, 1]. Higher weighted sample pairs tend to be homogeneous, with both samples being in either experiment or control group, for example. Figures 4A and 4B closely resemble what we would expect, with the mixed sample pairs not being as highly represented on the upper end of the scale. The high frequency of mixed sample pairs on both the upper and lower end of the scale in Figure 4C indicates that there might be more difficulty in differentiating between sample types in the control versus experiment group.

## 5. Graph Algorithms

There are advantages to placing our work in a graph-theoretic framework. Not only does this make it amenable to algorithms based on decades of basic research, but this representation is known to be appropriate for probing and determining the structure of biological networks (see, for example, [27-29]). In the present context, we used each of the afore-mentioned matrices to build an edge-weighted graph, with patients represented by vertices, and with edges assigned the previously-calculated weights. Edges with weight less than a preset threshold *T* were eliminated.

### 5.1 Clique-Centric Analysis

On the resulting unweighted graph we consider **clique,** a well-known *NP*-complete problem. Clique is typically formulated as a decision problem. Its inputs are a graph $G=(V,E)$ and a positive integer $k \leq |V|$. The question asked is whether there a $V' \subseteq V$ with $|V'| \geq k$ for which every pair of vertices in $V'$ is joined by an edge in $E$. Clique is widely recognized for its relevance in bioinformatics (in fact it is employed by Bioconductor in the aforementioned default peak alignment process). In the present application, we search for all maximal cliques. A maximal clique is one to which no new vertex can be added. It need not be one of maximum size. Recent work on fast maximal clique finding methods is described in [30, 31].

A central goal in this approach to put MS into practice as a potential diagnostic too is to find a set of cliques that covers each group with minimal overlap of the cliques between the groups. Thus, in order to distinguish accurately between sample groups, edges in this unweighted graph should connect mainly members of the same group. This is achieved in part by selecting a meaningful threshold. A visual inspection of the distribution of weights within the various data sets (see Figure 4) and prior experience in handling biological data [32, 33] leads us to select the initial value $T= 0.85$. Using a modification of the analysis procedure we first devised [25] (see procedure *clique-analysis* in Appendix B for details), we iterated on each of the three graphs until a set of cliques meeting our criteria was found. The optimal threshold was found to be 0.90 for the control versus experiment and control versus negative control data, while a much lower value of 0.78 was discovered to be optimal for experiment versus negative control. Although overlap still exists in the cliques between different groups, modifications to the threshold offered no improvement.

### 5.2 Refinement with Dominating Set

Scoring, weighting and clique help us to identify a set of peaks that discriminate between patient populations. To reduce this set to a core of its most promising elements, we seek to pinpoint those peaks that best cover the patient data. We rely on our previous work on cancer microarray data to develop an effective procedure for this task using **dominating set**, another well-known NP-complete problem. Its decision version can be defined as follows. The inputs are a graph $G=(V,E)$ and a positive integer $k \leq |V|$. The question asked is whether there a $V' \subseteq V$ with $|V'| \leq k$ for which every vertex in $V$ is either in $V'$ or adjacent to a vertex in $V'$. The version we face is properly known as nonplanar red/blue dominating set, because our graphs are bipartite. (See procedure *dominating-set-winnow* in Appendix B for details).

In brief, we checked the data using the Shapiro-Wilk test for normality, based on that assumed a normal distribution of the intensity values of each peak, and estimated for it the mean and standard deviation. This was done separately for each of the two patient groups. Then, based on the estimated normal distribution, we calculated the p-values for the original individual intensity values. This approach may be illustrated by constructing the aforementioned bipartite graph. As shown in Figure 5, one set of vertices represents peaks, the opposing set represents patient samples. We place an edge between a peak and a sample if and only if the p-value of the intensity corresponding to that peak-patient pair was greater than 0.05. Following statistical convention, we considered a p-value below this cutoff to indicate an outlier. In this setting, we want to identify the peaks that dominate all (or nearly all) of the samples. Therefore, we winnow out from consideration any peak vertex not adjacent to at least 90% of the patient vertices.

Also, to remove any peaks with a low possibility of discriminating between the two groups, we calculated the p-values for tests of equal means using both the Wilcoxon and t-test methods. We used both since the t-test assumes a normal distribution, while the Wilcoxon test does not. Only peaks for which both p-values were less than 0.01 were retained. We further filtered out peaks by controlling the false discovery rate at 0.01 based on the calculation on the complete list of final p-values. Q-values were calculated using software Q-VALUE [34].

After winnowing, we take the intersection of the peaks identified via clique with those selected via dominating set. This produces our best estimate of potentially meaningful diagnostic indicators. The three matrices of weights between patients are once again calculated using this smaller set of selected peaks. We examine the weight distributions for each data set and see that there is a greater separation between the sample groups, with an increased number of homogeneous sample pairs with higher weights. See Figure 6. Using the *clique-analysis* procedure described in Appendix B, we iterate on each of the three graphs to find a suitable set of maximal cliques. In this instance, a threshold value of 0.83 was found to be optimal for each of the graphs.

## 6. Results

For the comparison between experiment and negative control groups, 90 peaks were identified by the clique analysis approach, with 67 peaks left after dominating set refinement (See Table 1 in Appendix A for the complete list of those peaks). The enumeration of maximal cliques resulted in a total of seven cliques for the experiment versus negative control graph. Five were comprised only of experiment patient samples; two were comprised only of negative control patient samples. There was no clique mixed with patients from two groups. Thus for all practical purposes the two groups of patients were completely separated based on their MS profile using our algorithm. This suggests an exceedingly high level of correlation within and separation between sample types. See Figure7A.

Similarly, we identified 182 peaks using the clique analysis approach on control and negative control groups. Dominating set trimmed the list significantly down to only 6 peaks (See Table 2 in Appendix A). There were eight maximal cliques in the control versus negative control graph. Six were comprised exclusively of control patient samples, one contained only negative control patient samples, and one was mixed with patients from both groups. This too is suggestive of highly meaningful correlation and separation. See Figure 7B.

Eleven maximal cliques were found in the control versus experiment graphs, with one being comprised of only control patient samples and one comprised of only experiment patient samples. The other nine cliques were made up of a mix of control and experiment samples, although six were biased toward control and two were comprised of mostly experiment patient samples. See Figure 7C. This is not unexpected, because only a very small number of peaks were identified. Only five peaks were selected by clique analysis, with a single peak retained after dominating set refinement (See Table 3 in Appendix A). Such a small number of peaks generally leads to unstable clustering. However, this project is intended primarily as a pilot study to help determine the viability of this general approach. Much more data is needed to obtain larger repeatable peak profiles and thus more comprehensive peak spectra between all groups. We are satisfied with our methodology given the results using peaks for experiment and negative control.

## 7. Discussion and Conclusions

The ability to monitor changes at the molecular level provides an unprecedented chance to monitor vital processes such as organ and tissue modulation over the progression of disease and the metabolization of drugs. The multi dimensionality and complexity of biological processes necessitate the usage of more holistic approaches in diagnostics tools. In addition, contemporary problems of increasing drug tolerance and current advances in genetic polymorphisms necessitate the development of individualized medicine by which patients are treated on a case by case basis. The fact that some drugs might cure some patients while harming or even killing others also substantiates the development of holistic yet individualized diagnostic schemes. Such diagnostic systems that would provide reliable markers pinpointing differences between healthy and diseased states would allow risk assessment, early detection, and more precise monitoring of these diseases and their progression.

For a scientist to obtain such a holistic view of cell machinery under certain conditions, a transcriptome or proteome needs to be profiled. The ability of microarray technology to monitor expression of thousands of genes in a single experiment has attracted tremendous interest. Microarrays by nature are directed at analyzing mRNA rather than proteins, which are the actual biological facets that drive cellular machinery. Not all transcripts are functional at all times. Moreover, proteins are often post-translationally modified. These facts diminish the potential of microarray technology as a diagnostic tool, and minimize its potential to capture many intra-cellular mechanisms. The need for a tool that can directly monitor changes at the protein levels has motivated scientists to seek alternative technologies that work directly with proteins. Mass Spectroscopy is among those technologies that are assured to help understanding diseases progression through portraying a proteome profile for any desired tissue. The ability of MS to profile a whole proteome and provide a way of comparing two states (diseased versus healthy) has moved this technology to the forefront in biomedicine. Currently, MS data sets are increasingly used for protein profiling in diseased versus normal tissues. MALDI MS was utilized in the identification of phenotypic expression patterns for many pathological condition especially cancer. There is a rapidly growing literature on the use of MS in proteomic profiling of healthy and diseased state. However, in these reports there is single reliable algorithm through which single samples can be analyzed for diagnostic purposes.

Advances in MS, protein 3D technologies, computational methods and bioinformatics have the potential to improve the way we understand the differences between healthy and diseased states. This will allow us to screen total proteome profiles in a biologic sample. Using reliable algorithms and suitable bioinformatics tools will advance our ability to diagnose diseases and identify risk factors early on that are essential for effective treatment and for the development of novel therapies.

As we develop our knowledge of the molecular basis for most human disease, mass spectrometry provides new capabilities for discovering novel proteins and processes. Mass spectrometry and specifically MALDI-MS; demonstrate great potential as a diagnostic tool in a wide range of clinical settings. Likewise, clique-centric methods are gaining rapidly in popularity due to the purity of the results they generate and the scalability now achievable on immense data sets via advances in fixed-parameter tractability [35-37]. That we have found our scoring and weighting techniques useful on both transcriptomics data as in [25] and now here on proteomics data is especially encouraging. Disease characterization and mapping by depicting specific protein profiles can aid researchers in understanding how biological processes were governed. Research pertaining to disease associated with menopause is a fertile area for further applications of molecular diagnosis exploration.

# Figures and Tables

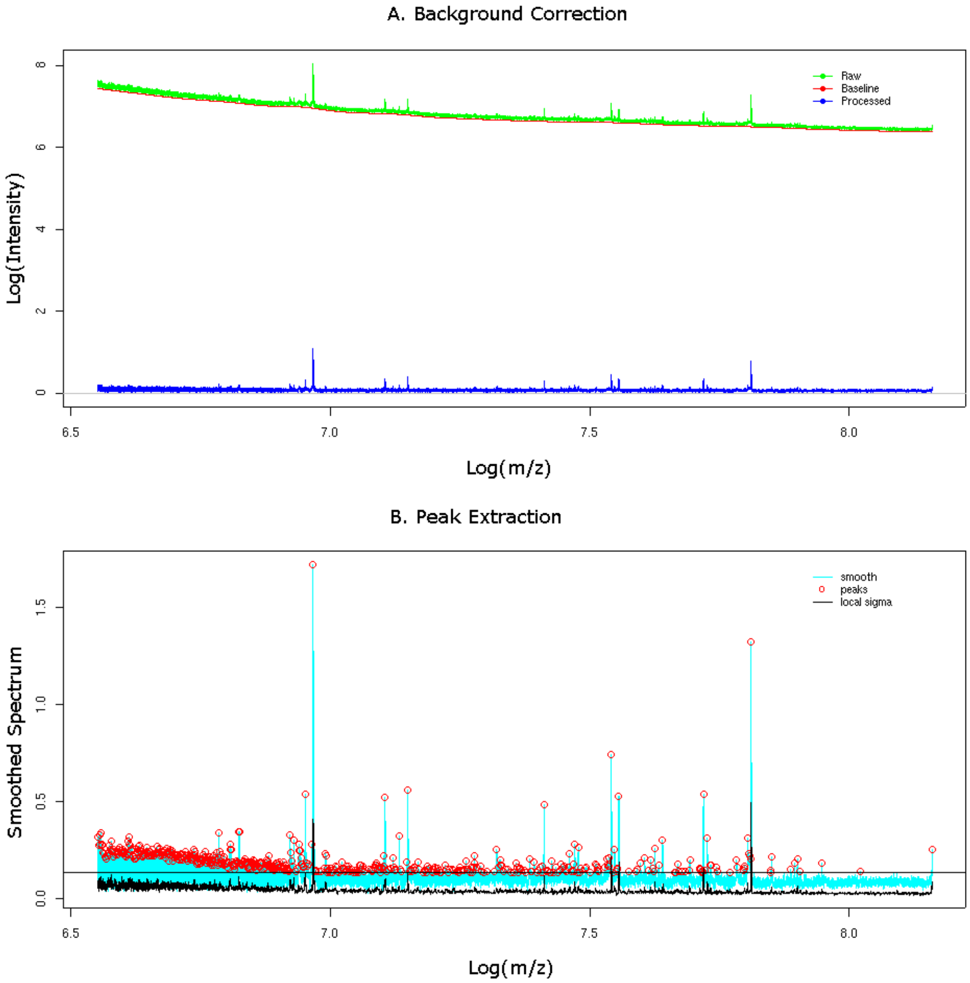

**Figure 1: A graphic example of spectrum preprocessing. A. Background correction. Horizontal axes are log-transformed m/z values and vertical axes are log-transformed intensity values. The raw, estimated and background corrected spectra are showed in color green, red and blue, respectively. B. Peak extraction. Smoothed and background corrected spectrum is in cyan and extracted peaks are labeled as red circles. Estimated local variation is in black. The 80 percent quantile is shown as the horizontal line.**

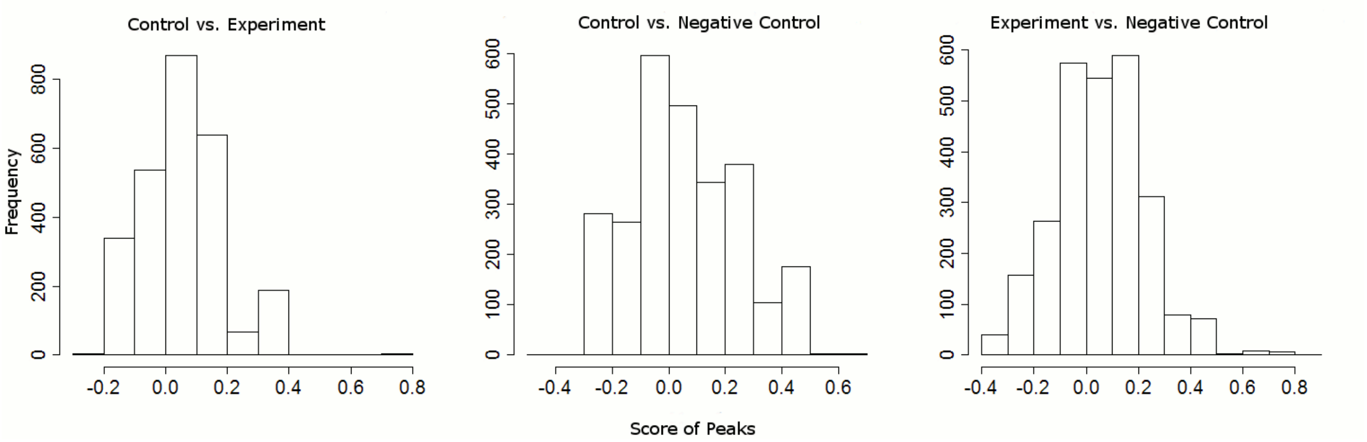

**Figure 2: Histograms of peak scores for all three pair-wise group comparisons. Peak scoring was performed as described in Appendix B.**

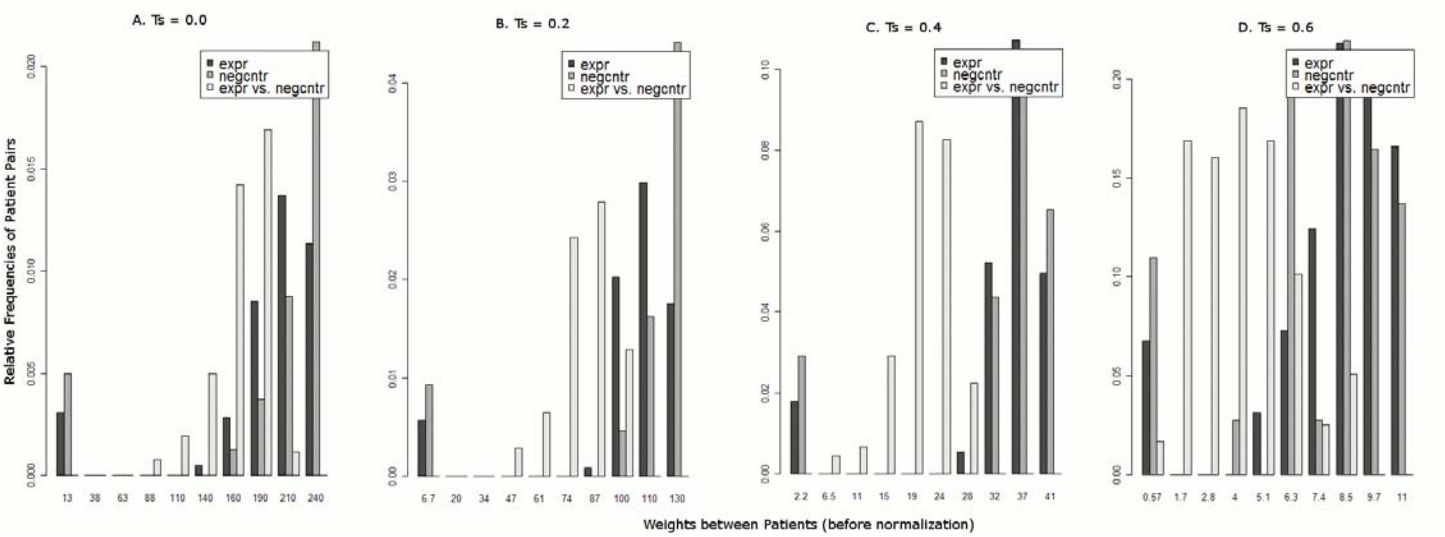

**Figure 3: Distributions of inter-subject weights under different peak score threshold values. A series of score thresholds (0, 0.2, 0.4 and 0.6) was applied to search for the appropriate cutoff values. Visual inspection of the distributions of weights between groups and within groups suggests the best separation is when 0.4 was chosen as the score cutoff value.**

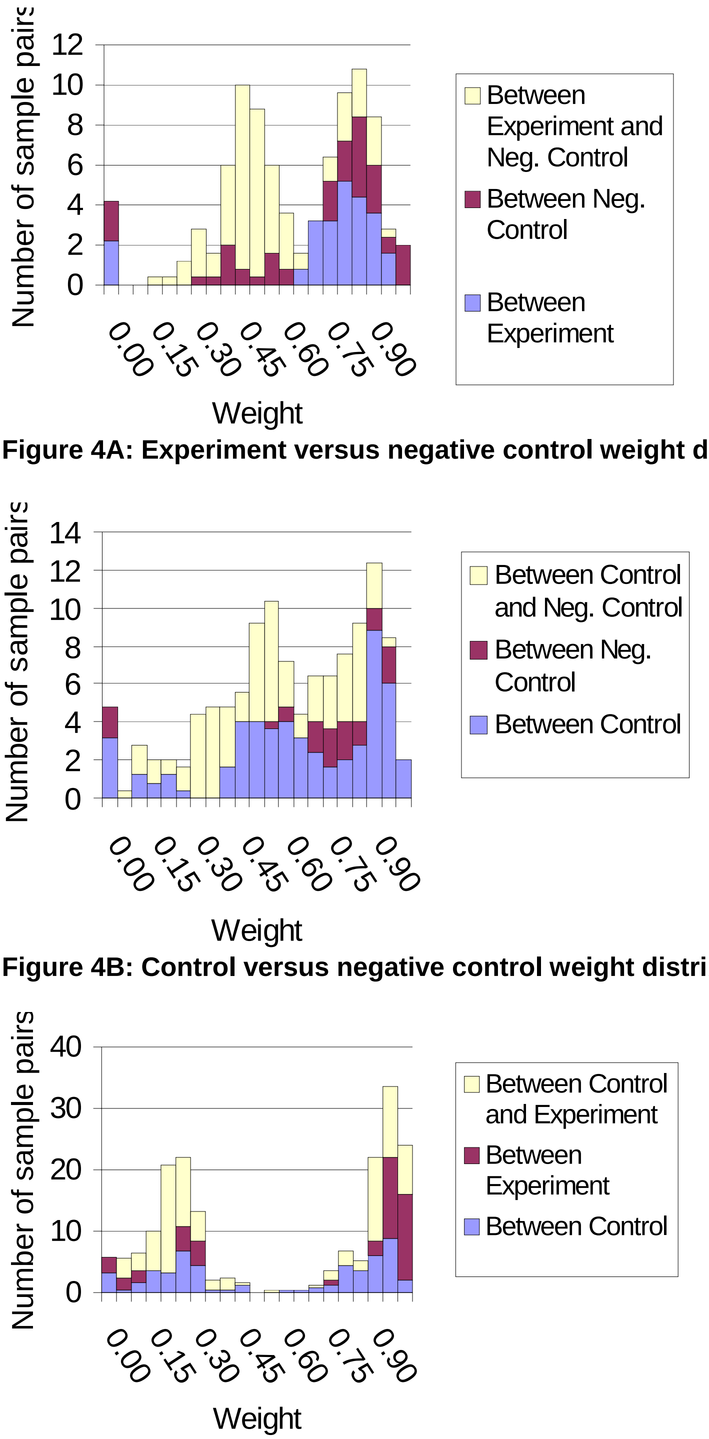

**Figure 4A: Experiment versus negative control weight distributions**

**Figure 4B: Control versus negative control weight distributions**

**Figure 4C: Control versus experiment weight distributions**

**Figure 4: Histograms of the number of sample pairs (edges) between each of the sample groups. Those pairs with high scores are predominantly homogeneous (the edges connect samples within the same group.) Visual inspection is used to select a satisfactory initial threshold.**

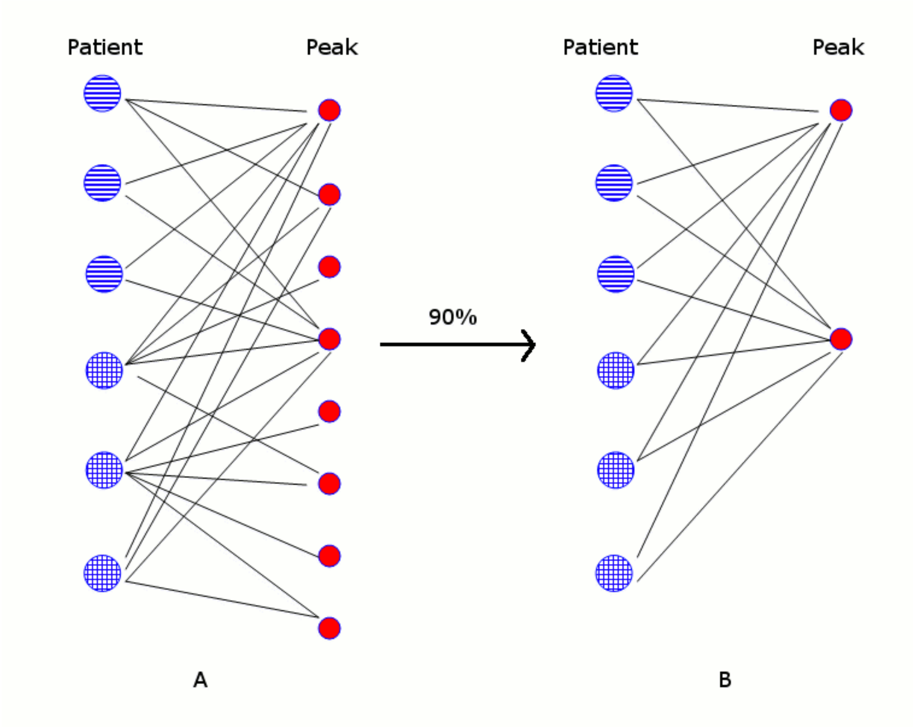

**Figure 5: Examples of patient-peak bipartite graphs. (A) Before winnowing. Peak vertices are shown in red. Patient samples are represented by blue vertices. Patient vertices shaded with horizontal lines are from one group. Those shaded with small grids are from another. Edges connect only patients and peaks (no peak-peak or patient-patient edge is possible). (B) After winnowing. Only peak vertices that dominate at least 90% of patient vertices are retained.**

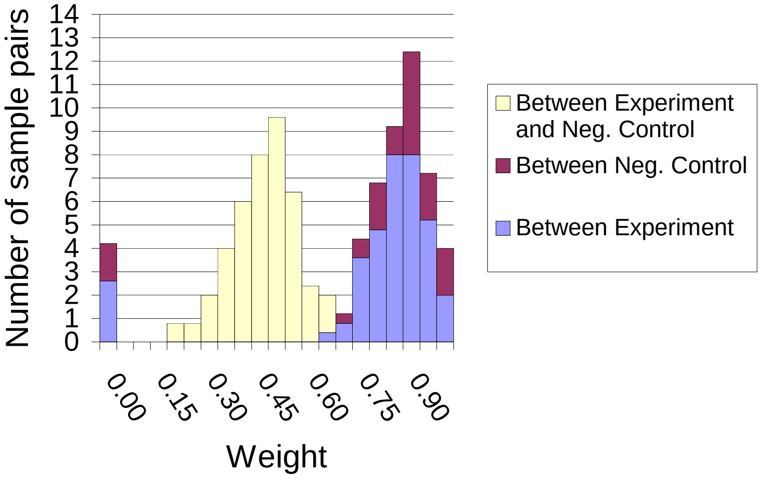

**Figure 6A: Experiment versus negative control weight distributions**

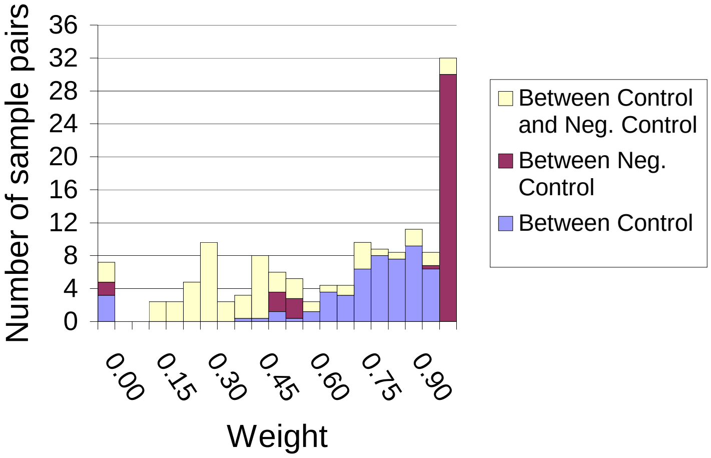

**Figure 6B: Control versus negative control weight distributions**

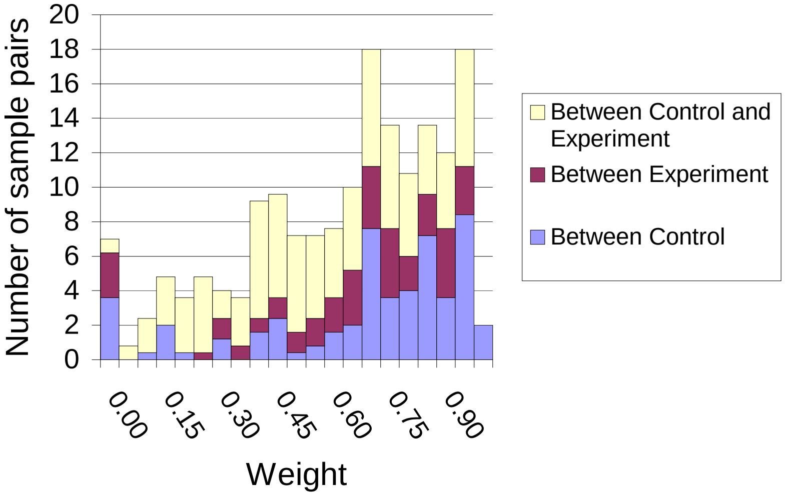

**Figure 6C: Control versus experiment weight distributions**

**Figure 6: Histograms of the number of sample pairs (edges) between each of the sample groups using weight calculated from peaks selected after dominating set refinement. The increased prevalence of homogeneous sample pairs at higher weights suggests a greater separation between the sample groups and confirms the validity of dominating set.**

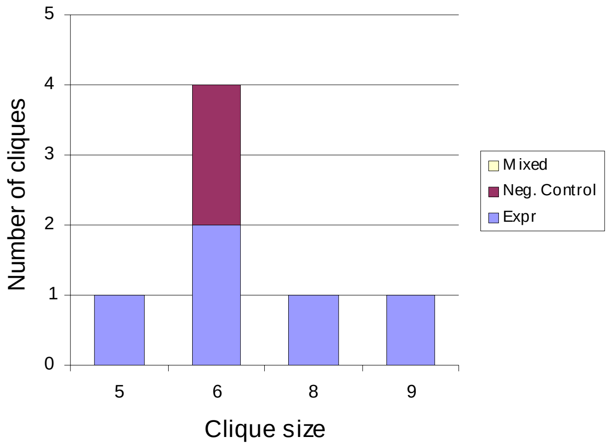

**Figure 7A: Experiment versus negative control cliques**

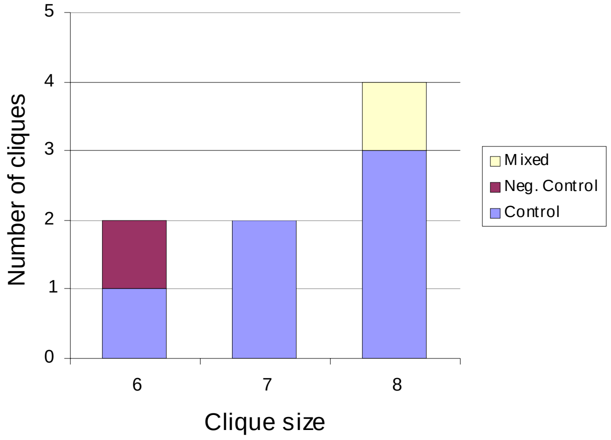

**Figure 7B: Control versus negative control cliques**

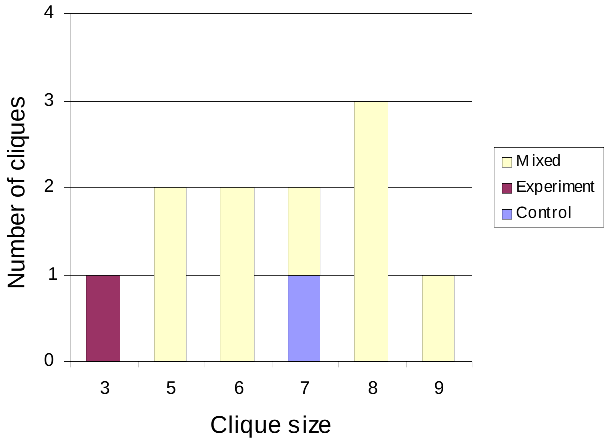

**Figure 7C: Control versus experiment cliques**

**Figure 7: Bar graphs illustrating the size and type of the maximal cliques. The experiment versus negative control graph contained all homogeneous cliques while the control versus negative control graph contained just one mixed clique. Maximal clique enumeration on the control versus experiment graph**

**resulted in one clique containing only control patient samples, one clique comprised of only experiment patient samples, and nine mixed cliques.**

## References

[1] M. Lee and E. Kerns, LC/MS APPLICATIONS IN DRUG DEVELOPMENT, *Mass Spectrometry Reviews*, vol. 18 (187-279), 1999.

[2] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, Ion Processes, *Int. J. Mass Spectrom.*, vol. 78, 1987, 53-68.

[3] R. Beavis, T. Chaudhary, and B. T. Chait, Cyano-4-hydroxycinnamic acid as a matrix for matrix-assisted laser desorption mass spectrometry., *Org. Mass Spectrom.*, vol. 27 (156), 1992, 156-158.

[4] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Y., and T. Yoshida, Protein and Polymer Analyses up to m/z 100,000 by Laser Ionization Time-of-flight Mass Spectrometry, *Rapid Communications in Mass Spectrometry*, vol. 2 (8), 1988, 151-153.

[5] H. Budzikiewicz and R. Grigsby, Mass spectrometry and isotopes: A century of research and discussion., *Mass Spectrom Rev*, 2005.

[6] K. Brown, E. Tompkins, and I. White, Applications of accelerator mass spectrometry for pharmacological and toxicological research, *Mass Spectrom Rev*, vol.
[Epub ahead of print], 2005.

[7] W. Kolch, C. Neususs, M. Pelzing, and H. Mischak, Capillary electrophoresis-mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery, *Mass Spectrom Rev*, vol. [Epub ahead of print], 2005.

[8] S. Naylor and R. Kumar, Emerging role of mass spectrometry in structural and functional proteomics., *Adv Protein Chem.*, vol. 65, 2003, 217-48.

[9] W. Blackstock and M. Weir, Proteomics: quantitative and physical mapping of cellular proteins., *Trends Biotechnol.*, vol. 17 (3), 1999, 121-7.

[10] J. Yates, Mass spectrometry. From genomics to proteomics, *Trends Genet*, vol. 16 (1), 2000, 5-8.

[11] D. Chace, T. Kalas, and E. Naylor, Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns., *Clin Chem.*, vol. 49 (11), 2003, 1797-817.

[12] M. Saugy, C. Cardis, N. Robinson, and C. Schweizer, Test methods: anabolics, *Baillieres Best Pract Res Clin Endocrinol Metab*, vol. 14 (1), 2000, 111-33.

[13] K. Dooley, Tandem mass spectrometry in the clinical chemistry laboratory., *Clin Biochem.*, vol. 36 (6), 2003, 471-81.

[14] S. Tang, N. Shehata, G. Chernenko, M. Khalifa, and X. Wang, Expression of BAG-1 in invasive breast carcinomas., *J Clin Oncol.*, vol. 17 (6), 1999, 1710-9.

[15] E. Petricoin, C. Paweletz, and L. Liotta, Clinical applications of proteomics: proteomic pattern diagnostics. J Mammary Gland Biol Neoplasia., vol. 7 (4), 2002, 433-40.

[16] M. Campa, M. Wang, B. Howard, M. Fitzgerald, and E. Patz, Protein expression profiling identifies macrophage migration inhibitory factor and cyclophilin a as potential molecular targets in non-small cell lung cancer., *Cancer Res*, vol. 63 (7), 2003, 1652-6.

[17] C. Frankenfeld, R. Patterson, N. Horner, M. Neuhouser, H. Skor, T. Kalhorn, W. Howald, and J. Lampe, Validation of a soy food-frequency questionnaire and evaluation of correlates of plasma isoflavone concentrations in postmenopausal women., *Am J Clin Nutr.*, vol. 77 (3), 2003, 674-80.

[18] C. Frankenfeld, A. McTiernan, E. Aiello, W. Thomas, K. LaCroix, J. Schramm, S. Schwartz, V. Holt, and J. Lampe, Mammographic density in relation to daidzein-metabolizing phenotypes in overweight, postmenopausal women., *Cancer Epidemiol Biomarkers Prev.*, vol. 13 (7), 2004, 1156-62.

[19] B. Simoneit, A review of current applications of mass spectrometry for biomarker/molecular tracer elucidation., *Mass Spectrom Rev*, vol. 24 (5), 2005.

[20] Y. Low, J. Taylor, P. Grace, M. Dowsett, S. Scollen, A. Dunning, A. Mulligan, A. Welch, R. Luben, K. Khaw, N. Day, N. Wareham, and S. Bingham, Phytoestrogen exposure correlation with plasma estradiol in postmenopausal women in European Prospective Investigation of Cancer and Nutrition-Norfolk may involve diet-gene nteractions., *Cancer Epidemiol Biomarkers Prev.*, vol. 14 (1), 2005, 213-20.

[21] C. Dane-Stewart, G. Watts, P. Barrett, B. Stuckey, J. Mamo, I. Martins, and T. Redgrave, Chylomicron remnant metabolism studied with a new breath test in postmenopausal women with and without type 2 diabetes mellitus. Clin Endocrinol (Oxf). vol. 58 (4), 2003, 415-20.

[22] S. M. Harman, E. A. Brinton, M. Cedars, R. Lobo, J. E. Manson, G. R. Merriam, V. M. Miller, F. Naftolin, and N. Santoro, KEEPS: The Kronos Early Estrogen Prevention Study, *Climacteric*, vol. 8 (1), 2005, 3-12.

[23] S. M. Harman, F. Naftolin, E. A. Brinton, and D. R. Judelson, Is the Estrogen Controversy Over? Deconstructing the Women's Health Initiative Study: A Critical Evaluation of the Evidence, *Ann N Y Acad Sci.*, vol. 1052, 2005, 43-56.

[24] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, and Z. J, Bioconductor: open software development for computational biology and bioinformatics, *Genome Biology*, vol. 5 (10), 2004.

[25] M. A. Langston, L. Lan, X. Peng, N. E. Baldwin, C. T. Symons, B. Zhang, and J. R. Snoddy, "A Combinatorial Approach to the Analysis of Differential Gene Expression Data: The Use of Graph Algorithms for Disease Prediction and Screening," in *Methods of Microarray Data Analysis IV, Papers from CAMDA '03*, K. F. Johnson and S. M. Lin, Eds. (Boston: Kluwer Academic Publishers, 2005) 223-238.

[26] R, "R: A Language and Environment for Statistical Computing," R Development Core Team, http://www.R-project.org.

[27] U. Alon, Biological Networks: the Tinkerer as an Engineer, *Science*, vol. 301, 2003, 1866–1867.

[28] A.-L. Barab´asi and Z. N. Oltvai, Network biology: Understanding the cell's functional organization, *Nature Reviews Genetics*, vol. 5, 2004, 101–113.

[29] Z. N. Oltvai and A.-L. Barab´asi, Systems Biology. Life's Complexity Pyramid, *Science*, vol. 298, 2002, 763–764.

[30] F. N. Abu-Khzam, N. E. Baldwin, M. A. Langston, and N. F. Samatova, On the Relative Efficiency of Maximal Clique Enumeration Algorithms, with Application to High-Throughput Computational Biology, *Proceedings, International Conference on Research Trends in Science and Technology*, Beirut, Lebanon, 2005.

[31] Y. Zhang, F. N. Abu-Khzam, N. E. Baldwin, E. J. Chesler, M. A. Langston, and N. F. Samatova, Genome-Scale Computational Approaches to Memory-Intensive Applications in Systems Biology, *Proceedings, Supercomputing*, Seattle, Washington, 2005.

[32] E. J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, J. B. Hogenesch, D. W. Threadgill, K. F. Manly, and R. W. Williams, Complex Trait Analysis of Gene Expression Uncovers Polygenic and Pleiotropic Networks that Modulate Nervous System Function, *Nature Genetics*, vol. 37 (3), 2005, 233-242.

[33] N. E. Baldwin, E. J. Chesler, S. Kirov, M. A. Langston, J. R. Snoddy, R. W. Williams, and B. Zhang, Computational, Integrative and Comparative Methods for the Elucidation of Genetic Co-Expression Networks, *Journal of Biomedicine and Biotechnology*, vol. 2, 2005, 172-180.

[34] J. D. Storey, J. E. Taylor, and D. Siegmund, Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society, Series B*, vol. 66, 2004, 187-205.

[35] F. N. Abu-Khzam, M. A. Langston, P. Shanbhag, and C. T. Symons, Scalable Parallel Algorithms for FPT Problems, *Algorithmica*, accepted for publication, 2005.

[36] F. N. Abu-Khzam, F. Cheetham, F. Dehne, M. A. Langston, S. Pitre, A. Rau-Chaplin, P. Shanbhag, and P. J. Taillon, "ClustalXP," http://ClustalXP.cgmlab.org/.

[37] M. A. Langston, Practical FPT Implementations and Applications (Plenary Talk), *Proceedings, International Workshop on Parameterized and Exact Computation*, Bergen, Norway, 2004.

## Appendix A. Tables of Peak Results (Putative Biomarkers)

### Table 1: Experiment versus negative control selected peaks

| | m/z | Score | | m/z | Score | | m/z | Score |
|---|---|---|---|---|---|---|---|---|
| M6.901420936 | 993.6857 | 0.4976 | M7.567743191 | 1934.7689 | 0.4457 | M7.7094751205 | 2229.3718 | 0.4397 |
| M6.90241193 | 994.6709 | 0.4982 | M7.5680349 | 1935.3334 | 0.4457 | M7.710392673 | 2231.4183 | 0.4397 |
| M7.371684679 | 1590.3107 | 0.4506 | M7.568072897 | 1935.4070 | 0.4417 | M7.7112443425 | 2233.3195 | 0.4397 |
| M7.372186159 | 1591.1084 | 0.4506 | M7.568409389 | 1936.0583 | 0.4417 | M7.7174674745 | 2247.2611 | 0.4397 |
| M7.3735444595 | 1593.2711 | 0.4471 | M7.569355563 | 1937.8910 | 0.4415 | M7.7213213185 | 2255.9384 | 0.4397 |
| M7.374150006 | 1594.2362 | 0.4471 | M7.572015788 | 1943.0531 | 0.4415 | M7.7219168315 | 2257.2823 | 0.4397 |
| M7.376087102 | 1597.3273 | 0.4243 | M7.57479908 | 1948.4687 | 0.4418 | M7.724766979 | 2263.7250 | 0.4397 |
| M7.377346123 | 1599.3397 | 0.4243 | M7.5753595695 | 1949.5611 | 0.4418 | M7.7283787805 | 2271.9159 | 0.4397 |
| M7.377962237 | 1600.3254 | 0.4243 | M7.5754733385 | 1949.7830 | 0.4411 | M7.734302048 | 2285.4130 | 0.4397 |
| M7.3786385435 | 1601.4080 | 0.4243 | M7.5757644945 | 1950.3507 | 0.4411 | M7.740653997 | 2299.9761 | 0.4397 |
| M7.378871598 | 1601.7813 | 0.4275 | M7.5763256745 | 1951.4455 | 0.4411 | M7.7454016575 | 2310.9215 | 0.5547 |
| M7.3792338945 | 1602.3617 | 0.4275 | M7.5767688215 | 1952.3105 | 0.4411 | M7.747771895 | 2316.4055 | 0.5547 |
| M7.379520607 | 1602.8212 | 0.4275 | M7.5772998485 | 1953.3475 | 0.4411 | M7.8578619845 | 2585.9856 | 0.6917 |
| M7.379560601 | 1602.8853 | 0.4275 | M7.577617825 | 1953.9687 | 0.4411 | M7.858987293 | 2588.8973 | 0.6917 |
| M7.379828336 | 1603.3145 | 0.4275 | M7.578369545 | 1955.4381 | 0.4411 | M7.867297957 | 2610.5024 | 0.6917 |
| M7.3801011055 | 1603.7519 | 0.4275 | M7.696364838 | 2200.3349 | 0.4386 | M7.871301935 | 2620.9757 | 0.6917 |
| M7.3804608735 | 1604.3290 | 0.4275 | M7.6970022305 | 2201.7378 | 0.4397 | M7.8725910795 | 2624.3567 | 0.6917 |
| M7.38075258 | 1604.7971 | 0.4275 | M7.701045805 | 2210.6587 | 0.4397 | M7.873368728 | 2626.3983 | 0.6917 |
| M7.381294571 | 1605.6671 | 0.4388 | M7.70132827 | 2211.2832 | 0.4397 | M8.1113965895 | 3332.2285 | 0.7658 |
| M7.5641876055 | 1927.9019 | 0.4307 | M7.704012552 | 2217.2269 | 0.4397 | M8.11354684 | 3339.4014 | 0.7658 |
| M7.566852434 | 1933.0463 | 0.4705 | M7.7076545325 | 2225.3167 | 0.4397 | M8.1155585715 | 3346.1261 | 0.7658 |
| M7.5672678455 | 1933.8495 | 0.4705 | M7.7094041295 | 2229.2135 | 0.4397 | M8.1271498955 | 3385.1378 | 0.7658 |

### Table 2: Control versus negative control selected peaks

| | m/z | score | | m/z | score | | m/z | Score |
|---|---|---|---|---|---|---|---|---|
| M7.8578619845 | 2585.9856 | 0.5393 | M7.867297957 | 2610.5024 | 0.5605 | M7.8725910795 | 2624.3567 | 0.6277 |
| M7.858987293 | 2588.8973 | 0.5393 | M7.871301935 | 2620.9757 | 0.6277 | M7.873368728 | 2626.3983 | 0.6277 |

**Table 3: Control versus experiment selected peaks**

| | mz | score |
|---|---|---|
| M7.283268882 | 1455.7389 | 0.4057 |

## Appendix B. Computational Procedures Employed

```
procedure peak-score-and-select
let $A$ be the $n$ x $m$ data matrix, where rows are patient and columns are peaks
for $j$=1 to $m$
    normalize intensity values in column $j$ to the range [0, 1]
    compute median expression value ($v_j$) and standard deviation ($\sigma_j$) on group 1 sample data for peak $j$
    repeat computation on group 2 sample data for peak j
    set score(peak $j$) = $|v_j(\text{group 1}) – v_j(\text{group 2})| – |\sigma_j(\text{group 1}) + \sigma_j(\text{group 2})|$
    delete peaks with scores not exceeding some lower limit
    return remaining peaks and their scores
```

```
procedure clique-analysis
initialize edge-weighted graph of order $n$
for $i$=1 to $n$
    for $j$ = 1 to $n$
    set the weight of each edge
for a user-specified number of iterations do
    use $T$ to delete edges with low weight
    find in resulting undirected graph all maximal cliques, $C$
    analyze $C$ to refine the choice of $T$
return $T$
```

```
procedure dominating-set-winnow
let $n$ be the number of patients
let $m$ be the number of peaks
initialize edge-weighted bipartite graph of order $n$+$m$
    for $i$=1 to m
        for $j$ = 1 to $n$
        determine the p-value (weight) of each peak ($i$,$j$)
    set threshold to 0.05 and eliminate edges of low weight
    flag peaks that dominate < 90% of patients from group 1
    flag peaks that dominate < 90% of patients from group 2
for $i$=1 to $n$
    generate p-value of equal mean using Wilcoxon and t-test
    set the final p-value of each peak as the maximum p-value of two test
flag peaks with the final p-value greater than 0.05
calculate q-vlaues of peaks based on the list of final p-values
flag peaks with q-value greater than 0.01
delete all of flagged peaks
return remaining peaks
```