Novel approaches to visualization and data mining reveals diagnostic information in the low amplitude region of serum mass spectra from ovarian cancer patients

Donald J. Johann, Jr.[a,*], Michael D. McGuigan[b], Stanimire Tomov[b], Vincent A. Fusaro[a], Sally Ross[a], Thomas P. Conrads[c], Timothy D. Veenstra[c], David A. Fishman[d], Gordon R. Whiteley[e], Emanuel F. Petricoin[f] and Lance A. Liotta[a]

[a] *NCI-FDA Clinical Proteomics Program, Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA*

[b] *Brookhaven National Laboratory, Information Technology Division, Upton, NY, USA*

[c] *Laboratory of Proteomics and Analytical Technologies, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD, USA*

[d] *National Ovarian Cancer Early Detection Program, Northwestern University Medical School, Chicago, IL, USA*

[e] *NCI-FDA Clinical Proteomics Program, Clinical Proteomics Reference Laboratory, SAIC Frederick, Gaithersburg, MD, USA*

[f] *NCI-FDA Clinical Proteomics Program, Office of Cell and Gene Therapy, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD, USA*

Abstract. The ability to identify patterns of diagnostic signatures in proteomic data generated by high throughput mass spectrometry (MS) based serum analysis has recently generated much excitement and interest from the scientific community. These data sets can be very large, with high-resolution MS instrumentation producing 1–2 million data points per sample. Approaches to analyze mass spectral data using unsupervised and supervised data mining operations would greatly benefit from tools that effectively allow for data reduction without losing important diagnostic information. In the past, investigators have proposed approaches where data reduction is performed by *a priori* "peak picking" and alignment/warping/smoothing components using rule-based signal-to-noise measurements. Unfortunately, while this type of system has been employed for gene microarray analysis, it is unclear whether it will be effective in the analysis of mass spectral data, which unlike microarray data, is comprised of continuous measurement operations. Moreover, it is unclear where true signal begins and noise ends. Therefore, we have developed an approach to MS data analysis using new types of data visualization and mining operations in which data reduction is accomplished by culling via the intensity of the peaks themselves instead of by location. Applying this new analysis method on a large study set of high resolution mass spectra from healthy and ovarian cancer patients, shows that all of the diagnostic information is contained within the very lowest amplitude regions of the mass spectra. This region can then be selected and studied to identify the exact location and amplitude of the diagnostic biomarkers.

Keywords: Ovarian cancer, SELDI-TOF MS, data visualization, diagnosis

1. Introduction

The serum proteome is an unexplored archive of metabolic and physiologic information that can reflect the pathologic changes within an organ system. Within

[*] Corresponding author: Dr. Donald J. Johann, Jr., NCI-FDA Clinical Proteomics Program, 8800 Rockville Pike, Building 29A, Room 2A21, Bethesda, MD 20892, USA. Tel.: +1 301 827 5194; Fax: +1 301 480 3256; E-mail: dj151o@nih.gov.

defined clinical trial study sets, experimentalists have used serum proteomic information content to detect the onset of disease at a very early stage, thus maximizing the likelihood of successful outcomes with therapeutic intervention [10,11]. However, this new source of information has its challenges, one being datasets massive in size and dimension. Current embodiments of high-resolution mass spectrometry (MS) based analysis of a single patient blood sample results in the generation of 350,000 to 400,000 records, with each record consisting of two numbers, a double precision float for the mass-to-charge (*m/z*) ratio and integer for the relative amplitude of the ion(s) being measured. The working hypothesis is that somewhere in the massive set of data there contains patterns of diagnostic information. Therefore, to maximize biomedical gain from this new and unexplained information source, innovative and robust bioinformatic methods are required.

Several clinical pilot studies have shown that a set or group of biomarkers will greatly improve the ability to discriminate the absence or presence of cancer and thus improve risk stratification and outcome. New detection methods based on pattern sets of biomarkers obtained from serum proteomic spectra have been proposed and developed [10,11], and confirmed by other centers [1,6, 12]. Current biomarkers are based on solitary proteins, i.e., Cancer Antigen 125 (CA-125) for ovarian cancer, and Prostate Specific Antigen (PSA) for prostate cancer. The predictive power of these tests to indicate the presence or absence of disease is sub-optimal. New biomarkers are needed for improvement in the early detection of cancer, and MS derived pattern based diagnostics have shown potential for superior discriminatory accuracy [1,4,6,10–12].

Massive multi-dimensional datasets containing diagnostic information bring a unique challenge in regard to the perceived clinical utility of a proteomic pattern based diagnostic system. Diagnostic information can be contained within specific sets of ions that exhibit a greater, or a lesser, abundance in the diseased state compared to the unaffected state. While earlier work utilized low-resolution mass spectrometry based platforms to generate data, recently investigators have found that increasing mass spectrometer resolution may yield greater clinical diagnostic utility [4]. Thus analysis of data streams with 15,000 to 40,000 data points is now expanded to pattern hunting on higher resolution MS platforms where data streams may be comprised of 300,000 to a few million data points per subject. These data sets are then analyzed with a set of sophisticated bioinformatics tools that render a decision on the pathophysiological condition of the patient from which the clinical sample was obtained (e.g. healthy or cancer-affected). Ultimately investigators and physicians must understand and feel comfortable with the use of such tests, or as shown by Spiegelhalter and Knill-Jones, physicians would reject a system that gave insufficient explanation even if the diagnostic model has good accuracy [13]. Therefore, significant efforts have begun to bring a more intuitive understanding of this new and complex data. The use of visualization tools to assist with experimental analysis of very large datasets has been successfully used for hypothesis testing, discovery and data reduction in the field of physics and communication theory. Analogous data mining and visualization tools have been applied to the field of diagnostic proteomics in this report. In support of these efforts, an information architecture has been formulated as illustrated in Fig. 1.

In this study, a variety of computational techniques have been used to evaluate a large and complex low molecular weight (LMW) proteomic dataset (*m/z* 700–12,000) derived from MS analysis of human serum from a cohort of 171 patients. The overall aims were as follows: a) identify global trends in the *full dataset* (171 patients spectra); b) apply these global trends to successfully isolate reduced datasets for subset analyses; and c) test whether *reduced datasets* and derived patterns can be linked to specific questions of biological significance, as well as be isolated as support for additional development and discovery activities.

2. Materials and methods

All patient samples were obtained from the same cohort as the original ovarian cancer cohort study [4, 10] that originated from the National Ovarian Cancer Early Detection Program (NOCEDP) clinic at Northwestern University Hospital (Chicago, IL, USA) and the Simone Protective Cancer Institute (Lawrenceville, NJ, USA). This cohort is composed of serum samples from 113 ovarian cancer patients and 58 controls. The datasets used in this study have been processed on a high-resolution QSTAR pulsar i hybrid triple quadrupole time-of-flight (QqTOF) MS (Applied Biosystems Inc., Framingham, MA) fitted with a PCI1000 interface (Ciphergen, Fremont, CA). Details of this instrument and associated spectrum analyzer methods have been described elsewhere [4].

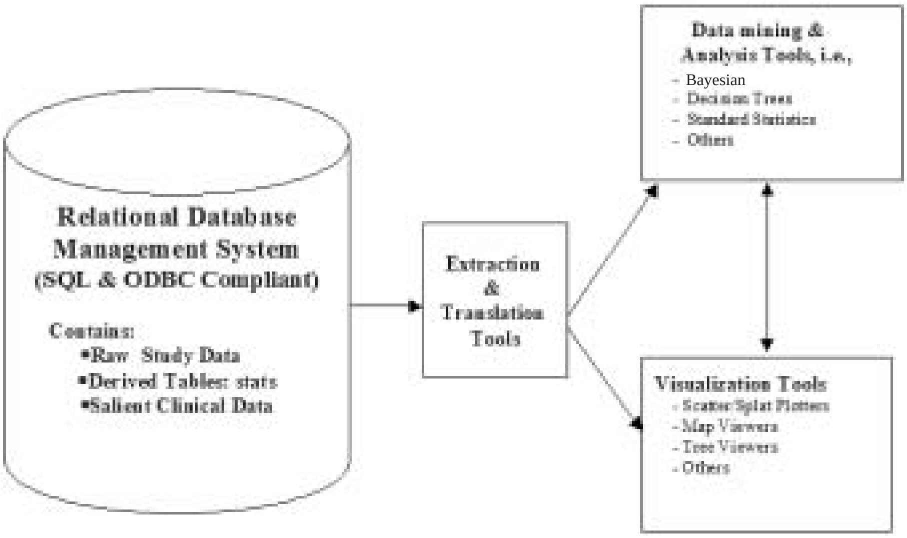


Fig. 1. Information architecture of the NCI-FDA Clinical Proteomics Program.

2.1. Analytical procedure

Following a routine blood draw, serum separation is performed using WCX2 ProteinChip arrays (Ciphergen) processed in parallel using a Biomek Laboratory workstation (Beckman-Coulter) modified to make use of a ProteinChip array bioprocessor (Ciphergen Biosystems Inc.). One hundred μl of 10 mM HCL was applied to the WCX2 protein arrays and allowed to incubate for 5 minutes. The HCl was aspirated, discarded and 100 μl of distilled, deionized water (ddH$_2$O) was applied and allowed to incubate for 1 minute. The ddH$_2$O was aspirated, discarded, and reapplied for another minute. One hundred μl of 10 mM NH$_4$HCO$_3$ with 0.1% Triton X-100 was applied to the surface and allowed to incubate for 5 minutes after which the solution was aspirated and discarded. A second application of 100 μL of 10 mM NH$_4$HCO$_3$ with 0.1% Triton X-100 was applied and allowed to incubate for 5 minutes after which the ProteinChip array bait surfaces were aspirated. Five μl of raw, undiluted serum was applied to each ProteinChip WCX2 bait surface and allowed to incubate for 55 minutes. Each ProteinChip array was washed 3 times with Dulbecco's phosphate buffered saline and ddH$_2$O. For each wash, 150 μl of either phosphate buffered saline or ddH$_2$O was sequentially dispensed, mixed by aspirating, and dispensed for a total of 10 times in the bioprocessor after which the solution was aspirated to waste. This wash process was repeated for a total of 6 washes per ProteinChip array bait surface. The ProteinChip array bait surfaces were vacuum dried to prevent cross contamination when the bioprocessor gasket was removed. After removing the bioprocessor gasket, 1.0 μl of a 30% solution of α-cyano-5-hydroxycinnamic acid in 50% (v/v) acetonitrile, 0.5% (v/v) trifluoroacetic acid was applied to each spot on the ProteinChip array twice, allowing the applied solution to dry between applications using a liquid robotic handling station Genesis Freedom 200 (TECAN, Research Triangle Park, NC).

2.2. QqTOF MS analysis

ProteinChip arrays were analyzed using a hybrid quadrupole time-of-flight mass spectrometer (QSTAR pulsar i, Applied Biosystems Inc., Framingham, Massachusetts) fitted with a ProteinChip array interface (Ciphergen Biosystems Inc.). Samples were ionized with a 337 nm pulsed nitrogen laser (ThermoLaser Sciences model VSL-337-ND-S, Waltham, Massachusetts) operating at 30 Hz. Approximately 20 mTorr of nitrogen gas was used for collisional ion cooling. Each spectrum represents 100 multi-channel averaged scans (1.667 min acquisition/spectrum). The mass spectrometer was externally calibrated using a mixture of known peptides. The output file of each serum mass spectrum contains approximately 350,000 to 400,000 records [4]. Each record consists of two numbers representing the *m/z* value of the ionic species (subsequently referred to as *mass* in this document and in graphic images) and intensity (subsequently referred to as *amplitude* or *amp* in this document and graphic images).

2.3. *Preprocessing of MS data*

Preprocessing of MS datasets, including computational quality assurance and quality controls methods are performed prior to analysis by subsequent modeling and visualization tools to screen for potential bias. Preprocessing uses binning, a standard technique used to group multivariate data. All preprocessing computational methods are performed in a Microsoft SQL Server 2000 database by computer programs composed of suites of T-SQL stored procedures. For each patient's spectra all mass values are binned and a sum is generated representing the amplitude values within that binned mass values, and sum associated amplitude values (for that bin). For all patient data in the proteomic database we normalized the summed amplitude values per mass species (following the mass bin size). Normalization is performed independent of the patient's disease/health state (labeled DzState in graphic images).

Since the resolution of the QqTOF-MS used in this study scales linearly as the *m/z* ratio increases, the binning technique uses a linear binning window that grows in size as the *m/z* value increases. A mass bucket is defined as the mass values residing within the upper and lower limits of the particular binning window. Various binning techniques have been evaluated and remain under investigation. Binning can introduce coarseness and thus subtle trends or findings can then be masked. The current binning procedure aggregates each patient spectrum to a fixed size of 7105 records. Since binning resolution can affect accuracy of subsequent analysis methods, these techniques were linked to accuracy experiments of the mass spectrum analyzer to empirically derive the current method.

The binning formula begins by using a base value of 400 parts per million. The bin window size is determined by multiplying the base value by the mass value lower limit of the current bin window. The upper limit of the binning window is simply the lower limit plus the current size of the bin window. Thus, the binning window widens in a linear manner as the *m/z* value increases.

A mass species is defined as all the mass values contained within a particular binning window or bucket. For each mass bucket, the minimum and maximum amplitude values are determined. The normalized amplitude value is converted to a double precision floating point number between zero and one, and becomes a part of the record for that database entry. The normalization method is a linear scaling transform:

- While there are mass buckets to process:
  - ∗ Normalize all raw amplitude values for the current mass bucket
    - ∗ Normalized Amp = (RawAmp – MinAmp) / (MaxAmp – MinAmp)

The database can then be queried and various types of datasets constructed. Datasets are then transferred over a network connection to an SGI graphics supercomputer for visualization and data mining operations using the MineSet visualization toolset [8,9].

3. Results

A summary of the visualization tools used in this study, the datasets to which they were applied, general type of data display, and indication for use, is shown in Table 1. Conventions used in this study for discriminating ovarian cancer patients verses normal in all the graphics include the following: the color blue or number zero (0) refers to control/normal, and the color red or number one (1) refer to ovarian cancer.

3.1. *Results from processing the full dataset*

The tools used to process and analyze the full dataset produced the set of graphics shown in Fig. 2. A global discriminating trend is present in the low amplitude regions of all these tools and can be readily identified on most of the graphics. Concordance of multiple tools using different algorithmic methods lends significant support to this finding. The following are details of the discovered global discriminating trend.

The results of the visualization tool called the Splat Visualizer are shown in Figs 2(A–C). The Splat Visualizer allows for a global view of the entire proteomic dataset. This tool allows a user to interactively visualize and explore relationships among several variables and is particularly suited for datasets containing a large number of records. The displayed landscape is three dimensional, and has the ability to be manipulated (via rotation, zoom, pan, and drill/fetch type operations) with very quick rendering speed. The Splat Visualizer shows an approximation of a three-dimensional scatter plot through the use of aggregation techniques [8].

In this example, the x-axis represents binned mass values (between *m/z* 700 and 12,000). The y-axis represents normalized amplitude values (ranging from 0 to 1). The disease state (DzState) is represented on the z-axis with the first position occupied by normal/controls,

Table 1
Summary Table of Visualization Tools and Data Sets

Visualization tool name	Data set	Display type	Indiations
Splat viz	Full, stage	3D Graph	Global trends in large datasets
Evidence visualizer	Full	2D tabular	Global trend analysis
Decision table	Full, reduced, stage	2D & 3D tabular	Subet analysis of contiguous mass & amplitude pairs
Scatter plot	Reduced	3D graph	Graphical subet analysis
Decision tree	Full	3D graph	Global analysis, search for interesting rules
Option tree	Reduced	3D graph	Subset analysis, search for interesting rules

and the second position occupied by ovarian cancer patients. Therefore, the aggregated spectra of normal and cancer patients are segregated both by voxel color and position on the z-axis, and have been normalized independent of DzState but on a bucket-by-bucket basis, thus bringing out the difference between the two groups. The red patterns on Figs 2 (A) and (B) along the x-axis are due to a relative signal void present in the low amplitude mass values of the normal dataset component allowing the red voxels to bleed through. The spectral signal homogeneity in the cancer group is shown in Fig. 2(C).

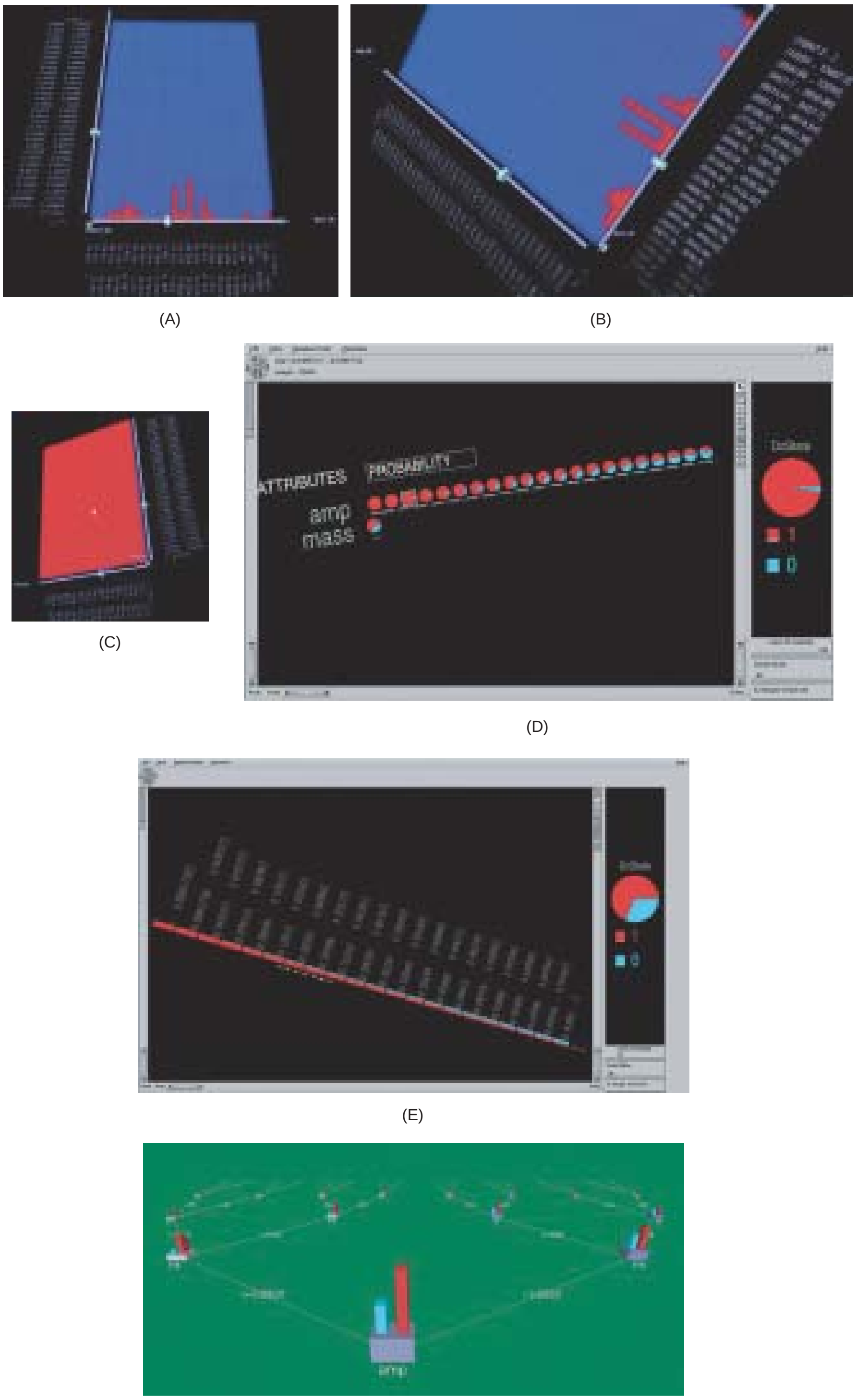
Results of the visualization tool called the Evidence Visualizer are shown in Fig. 2(D), and the Decision Table in Fig. 2(E). The Evidence Visualizer uses a Naïve Bayesian classification method, and the Decision Table uses a variant of a decision tree method. Both employ computer graphic techniques to present the findings in a comprehensible tabular form. Applying visual aids to the structure of a probabilistic relationship, algorithmic finding, or statistic results in the transformation of columns of data into an intuitive graphic picture, especially for large and complex datasets [3]. In both tool displays, the left pane contains attributes (features) that have been aggregated according to their ability to discriminate the class assignment variable, DzState. The right pane of both tools contains the class assignment variable, whose outcome metric changes dynamically as different sets of attributes are selected, thus allowing interactive "what if" type questions to be investigated and answered quickly. In the construction of these models, the inducer component of these tools examines class assignment of each dataset record using a wrapper-based approach, and eventually produces a descriptive classifier.

Visualization tools were used to probe the mass amplitude ranges containing the majority of the diagnostic information in the MS spectra of the serum samples acquired from both healthy and ovarian-cancer affected patients. Both tools (Evidence and Decision Table) show significant ability to discriminate a cancer class assignment within the low amplitude ions (predominantly red color). Ionic species in mid to high amplitude regions do not discriminate well as shown by the mixture of red and blue within this region.

A three-dimensional view of a decision tree graphic derived from the Full Dataset is shown in Fig. 2(F). Decision trees partition the data of a problem domain by progressively splitting the data space into smaller subsets. These subsets are constructed as a function of inequality-based rules that segment the data space so that the outcome indicator (DzState) is best classified. In this example approximately one-third of the dataset was reserved for testing and blind validation. The tree above reported better than a 99% accuracy following validation. Due to the size of the dataset, the depth of the tree averaged over 100 levels, thus diminishing its potential utility to interactively explore the dataset. Following exploratory analysis of the tree, it was again empirically noted the low amplitude ions had discriminatory abilities with lower probability of error verses ions having mid-range or higher amplitudes.

3.2. *Results from processing reduced datasets*

Noting the agreement in regard to the analysis of the full dataset by multiple visualization tools, and taking into account the concordant discovery of a common global discriminatory trend involving the low abundant species of many *m/z* values, the natural question is as follows: "Can a function based only on the values of low amplitude ions be assembled to effectively discriminate the ovarian cancer cohort?" To test this hypothesis reduced datasets comprised on only the low amplitude ion signals were constructed. The tools used to process and analyze the reduced datasets produced the set of graphics contained in Fig. 3. The first reduced dataset was constructed by querying the database for all records having normalized amplitude values in the range of 0 through 0.1085981. The query resulted in the exclusion of one patient from the normal group and five patients from the cancer group. The total number of records was decreased from approximately 1.2 million to 71,584, effectively decreasing the size of the ini-



(A)

(B)

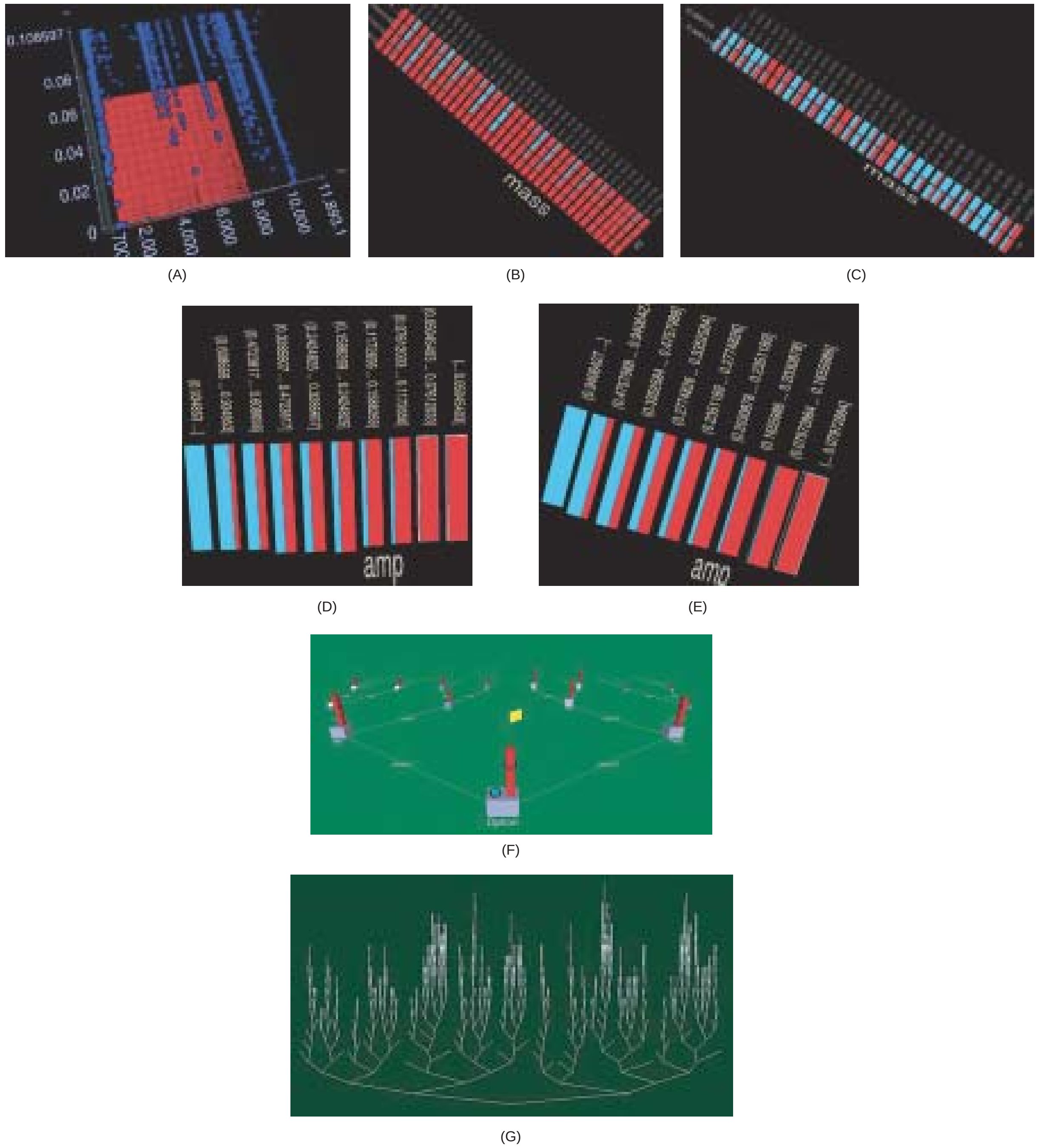
(C)

(D)

(E)

(F)

Fig. 2. Graphics from the analysis of the full dataset from the MS proteomic patterns.
A) Splat Visualizer results using the full dataset. B) Alternate view of Fig. 2(A) after rotation and zoom. C) Alternate view of Fig. 2(a) after 180° rotation. D) Evidence Visualizer analysis of full dataset. E) Decision Table analysis of full dataset. F) Decision Tree derived from the full dataset.



(A) (B) (C)

(D) (E)

(F)

(G)

Fig. 3. Plot and table graphics from the analysis of the reduced datasets.
A) Three dimensional scatter plot of a low amplitude ion signal reduced dataset. B) Decision Table analysis of a low amplitude ion signal reduced dataset. C) Decision table analysis of a high amplitude ion signal reduced dataset. D) Decision Table analysis of a reduced dataset comprised of *m/z* values between 4000 and 4283. E) Decision Table analysis of a reduced dataset comprised of *m/z* values between 7519-8028. F) Tree oriented graphics from the analysis of the reduced dataset showing the option tree derived from this dataset. G) Aerial view of the option tree derived from the reduced dataset.

tial dataset by more than an order of magnitude. This reduced dataset was then re-introduced to the visualization tool suite for spectral trend analysis.

The results of the scatter plot visualization tool produced a non-aggregated three-dimensional plot of the cancer spectra (red) verses normal/control (blue), as shown in Fig. 3A. The horizontal axis represents the *m/z* values in the range of 700 to just less than 12,000. The vertical axis displays the normalized amplitude values of this reduced dataset clipped according to the database query. The differences in the spectral patterns are evident, as a group; there is a significant discriminatory ability of numerous low abundant ion signals. Next, the following question was posed: "Can this dataset that represents a simple function of low abundant signals be mapped to find corresponding discriminating *m/z* values?"

The same reduced dataset was introduced to the Decision Table Classifier, with the results of this analysis shown in Fig. 3(B). Mass values were found to have a significant role in segregating disease states. Multiple contiguous regions of mass values show a definitive ability to discriminate cancer (red) verses normal (light blue). The horizontal axis displays a full complement of contiguous regions of mass values and the vertical axis contiguous subsets of amplitude values in the queried range. We have now discovered multiple contiguous regions of *m/z* values in the low abundant region of the cohort spectra that segregate well for the cancer cohort.

Based on these findings, the full dataset was then computationally divided as a function of contiguous graduated segments of normalized amplitude values. The low amplitude ion signals continued to show a significant discriminatory ability in identifying spectral regions associated with cancer. The mid amplitude ion signals were not found to have significant discriminatory ability. A few of the high amplitude ion signals showed discriminatory ability in identifying spectral regions associated with patients known to be normal (i.e. no evidence of cancer). A slice through the reduced dataset that isolates the high amplitude ions (0.9 to 1.0) is shown in Fig. 3(C). All patients were included in the high amplitude reduced dataset. The axes are oriented the same as for Fig. 3(B). Visualization tools have successfully isolated contiguous regions of *m/z* values with high signal intensity that effectively discriminate controls.

Considering the findings using the low and high amplitude reduced datasets, we next studied whether or not visualizations tools could assist in finding contiguous regions of *m/z* values displaying a sentinel (biomarker) role as a DzState indicator. A disease state in this case was again ovarian cancer versus unaffected. The result of intersection analysis is shown in Figs 3(D) and (E). All patients were included in these contiguous *m/z* specific reduced datasets. For both figures, the horizontal axis displays the normalized amplitude values with the lowest values appearing on the right. The vertical axis represents the contiguous masses as a group (*m/z* 4000–4283 for one extraction and *m/z* 7519–8028 for the second extraction). Both graphics show the high amplitude values discriminating the normal component of the cohort, and the low amplitude values discriminating cancer. Therefore, through iterative activities involving a series of datasets, we have progressed from a very large and complex full dataset to a multitude of reduced datasets, and finally to the initial steps consistent with biomarker discovery in two distinct spectral regions.

3.3. Results from processing the stage datasets

Visualization techniques were next applied to test if the proteomic pattern in ovarian cancer differs as a function of the stage of disease. The graphic images contained in Fig. 4 are derived from aggregate spectra, and were captured from a three-dimensional rendering of ovarian cancer patients having either stage II, III, or IV disease. The number of patients with stage II disease was 13, there were 72 patients with stage III disease, and 8 with stage IV. The complete spectral patterns were successfully grouped, examined, and contrasted as a function of stage. The x-axis represents binned mass values (*m/z* range 700 to 12,000). The y-axis displays normalized amplitude values (range 0 to 1). The stage of disease is represented by the location on the z-axis, with the first position occupied by stage II (blue voxels), the second position by stage III (green), and the third position by stage IV (red). The images of Fig. 4 are the same composite three-dimensional image undergoing a 180-degree rotation. The discriminating factors are the amplitude values, not the *m/z* values. Therefore, visualization of spectral patterns as a function of stage of disease may assist with a better understanding of the heterogeneity of cancers.

4. Discussion

It is impossible to make sense of large and complex datasets without appropriate bioinformatic tools.

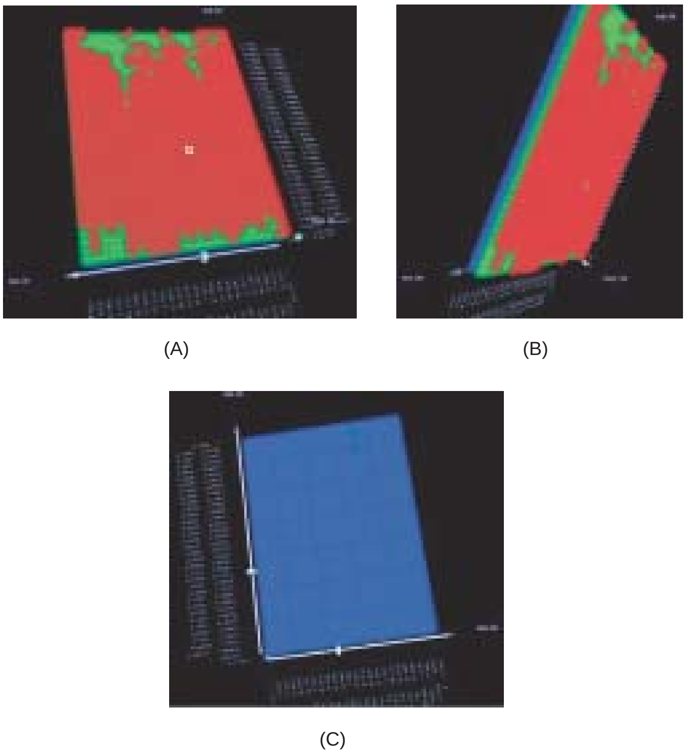


Fig. 4. Graphics from the analysis of the ovarian cancer stage dataset.
A) Splat Viz results from the stage dataset (stage IV (red), stage III (green), and stage II (blue)). B) Fig. 4A following 90° rotation. C) Fig. 4(A) following 180° rotation.

It would be impractical to analyze these datasets with traditional plotting methods and spreadsheet type programs. The process of cycling through hypothesis generation and discovery activities is greatly enabled by the bioinformatic tools as illustrated in this study. Visualization tools have been shown to greatly assist in finding and isolating regions of the proteomic MS spectrum containing important diagnostic information. In this study, the full dataset acquired from the MS spectral patterns was used to find a global discriminating pattern leading to the construction of a reduced dataset. This data reduction lead to the further subset type analysis and the construction of a series of additional reduced datasets that allowed the discovery of two small isolated regions of *m/z* values suggestive of a sentinel type role, and logical progression to further biomarker validation activities. Finally, visualization tools allowed for the examination of a cancer stage reduced dataset, and to begin considering variation in proteomic signature patterns as a function of stage of disease. Visualization tools and techniques have been found to have significant utility in this study by working through cycles involving discovery and hypothesis generation, in a large and complex dataset with focused domain questions.

All datasets used in this study are derived from a subset of ion signals found in the low molecular mass region of the serum proteome (i.e. *m/z* 700 to 12,000), a previously unexplored information archive. Mounting scientific evidence shows support that this region effectively reflects metabolic processes and can be used to detect pathogenic changes in an organ system. Therefore, there exits a pressing need for bioinformatic tools to assist in the analysis of this new and rich information archive. Challenges include not only the massive size and dimensions of these datasets, but also the ability to account for heterogeneity present in both the human population and in disease processes proper. Therefore, bioinformatic tools with the abilities of obtaining global views of aggregated spectra, along with the ability to isolated regions of interest for subset type analysis, are vital in bringing scientific understanding and eventually clinical utility to this new data source. Analytic approaches employing probabilistic or likelihood scoring strategies, begin to address the intrinsic diversity issues of patient's and their individual pathogenic processes, when examining this new and uncharted proteomic information.

Results of this study support the conclusion that discriminatory diagnostic information is enriched in the low abundance, low molecular weight region of the serum proteome. This is not unexpected, considering the pathophysiology, the physical size of an ovary harboring a stage I or II pathologic process, and the normal physiologic role of the kidney to efficiently clear peptides and cleavage products below a molecular weight of approximately 45,000 Da. Recent advances support the concept that cleavage products and small peptides that are normally cleared by the kidney exist in association with circulating carrier proteins (i.e. albumin) and thus become amplified by avoiding clearance and assuming the half-life of that carrier protein. Therefore the concentration of the ionic species (biomarker) becomes a function involving the production rate from the diseased tissues and clearance rate of the carrier protein [7]. Physiologically, these discriminators exist in the low abundant signal region of the mass spectrum.

Recently, investigators have proposed approaches where data reduction is performed by a priori "peak picking" and alignment/warping/smoothing components using rule based signal-to-noise measurement [2, 5,14]. Unfortunately, while this type of system has been employed for gene microarray analysis, it is unclear whether or not this will be effective in analyzing mass spectral data, which unlike microarray data, is comprised of a continuous measurement operation. Moreover, it is unclear where true signal begins and noise ends. Based on our findings presented here, a priori peak selection may miss most of the diagnostic information since this data is eliminated based on low amplitude attributes and it supposes that one has knowledge about what constitutes signal and real noise. Approaches to reduce data based on amplitude, perhaps only selecting out the lowest amplitude regions of the mass spectral data streams, could become an important component of data mining operations for ongoing biomarker discovery. Logical search processes that isolate *m/z* regions showing a sentinel type role via discriminating characteristics of low abundance signals for the cancer group and high abundance signals for the control group (Figs 3(D) and (E)) show promise for further biomarker discovery activities such as MS/MS sequencing.

The clinical trial for validation of diagnostic proteomic patterns is an ambitious undertaking. This technology is foreign to both the clinical laboratory as well as regulatory agencies such as the Centers for Devices and Radiological Health (CDRH), the regulatory body responsible for the notification of new diagnostic tests in the United States. As a result, it is necessary to both have tools such as pattern recognition software to present to these groups as well as mechanisms that will demonstrate the robustness and verify the selection of the selected pattern characteristics. The visualization tools presented here are capable of presenting an enormous amount of data in a format that can be easily understood. We are therefore planning on utilizing these tools to present data to both clinicians and regulatory bodies in an understandable and concise manner, and to demonstrate the pattern differences discovered by the pattern recognition software. Tools such as the Splat Visualizer add powerful evidence to be used as part of the verifications of pattern selection.

In conclusion, the data visualization tools presented in this study provide a convenient method of examining large data sets generated by MS analysis of human serum. Through the use of these tools, data can be reduced to a manageable size and presented in a concise and easily understandable format for presentation to scientists, physicians and regulatory agencies. Further analyses on larger datasets are planned for validation of MS proteomic patterns as an aid in disease diagnosis.

References

[1] B-L. Adam et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Res* **62** (2002), 3609–3614.

[2] K.A. Baggerly et al., A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples, *Proteomics* **3** (2003), 1667–1672.

[3] B. Becker, Volume Rendering for Relational Data, *Proceedings of the 1997 IEEE Symposium on Information Visualization*, (1997), 87–90.

[4] T.P. Conrads et al., Cancer diagnosis using proteomic patterns, *Expert Rev Mol Diagn* **3** (2003), 411–420.

[5] K.R. Coombes et al., Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization, *Clin Chem* **49** (2003), 1615–1623.

[6] J. Li et al., Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer, *Clin Chem* **48** (2002), 1296–1304.

[7] A. Mehta et al., Biomarker Amplification by Serum Carrier Protein Binding, *Disease Markers* **19** (2003), 1–10.

[8] MineSet User's Guide, Chapter 8 Using the Splat Visualizer, Silicon Graphics Inc.

[9] MineSet User's Guide, Chapter 10 MineSet Inducers and Classifiers, Silicon Graphics Inc.

[10] E.F. Petricoin et al., Use of proteomic patterns in serum to identify ovarian cancer, *Lancet* **359** (2002), 572–577.

[11] E.F. Petricoin et al., Serum proteomic patterns for detection of prostate cancer, *JNCI* **94** (2002), 1576–1578.

[12] Y. Qu et al., Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, *Clin Chem* **48** (2002), 1835–1843.

[13] D. Spiegelhalter and R. Knill-Jones, Statistical and knowledge-based approaches to clinical decision support systems, with an application in gastroentrerology, *Journal of the Royal Statistical Society A* **147** (1984), 35–37.

[14] A. Vlahou et al., A novel approach toward development of a rapid blood test for breast cancer, *Clin Breast Cancer* **4** (2003), 203–209.