MassVis: Visual Analysis of Protein Complexes Using Mass Spectrometry

Robert Kincaid*
Agilent Laboratories

Kurt Dejgaard+
McGill University

ABSTRACT

Protein complexes are formed when two or more proteins non-covalently interact to form a larger three dimensional structure with specific biological function. Understanding the composition of such complexes is vital to understanding cell biology at the molecular level. MassVis is a visual analysis tool designed to assist the interpretation of data from a new workflow for detecting the composition of such protein complexes in biological samples. The data generated by the laboratory workflow naturally lends itself to a scatter plot visualization. However, characteristics of this data give rise to some unique aspects not typical of a standard scatter plot. We are able to take the output from tandem mass spectrometry and render the data in such a way that it mimics more traditional two-dimensional gel techniques and at the same time reveals the correlated behavior indicative of protein complexes. By computationally measuring these correlated patterns in the data, membership in putative complexes can be inferred. User interactions are provided to support both an interactive discovery mode as well as an unsupervised clustering of likely complexes. The specific analysis tasks led us to design a unique arrangement of item selection and coordinated detail views in order to simultaneously view different aspects of the selected item.

KEYWORDS: information visualization, visual analysis, correlation analysis, mass spectrometry, proteomics, interactome.

INDEX TERMS: H.1.2 [User/Machine Systems]: Human information processing – Visual Analytics; I.6.9 [Visualization]: information visualization; J.3 [Life and Medical Sciences].

1 INTRODUCTION

Proteins are biological macromolecules composed primarily of amino acids translated from a cell's DNA. Some proteins are simply responsible for structural components of cells and organisms. However, many proteins participate in a variety of chemical processes such as metabolism, DNA transcription and cell-cell signalling. Frequently, for proteins to perform these functions they must exist in a protein complex where two or more proteins form a larger loosely bound unit with a specific three dimensional structure. This combined structure is necessary for the specific function the cell requires. Figure 1 shows an example of such a complex.

With the complete sequencing of the human genome and other organisms, we now have a reasonably complete inventory of what proteins a cell might produce. How these various proteins interact and combine to give rise to cellular function is far less undertood. Hence, protein-protein interactions and the composition of protein complexes is still a very active area of focus for proteomic scientists.

This paper describes the motivation and design of MassVis, a software application designed to support the analysis needs of a novel proteomics workflow currently under development for studying protein complexes separated by so-called "native electrophoresis". The nature of the data generated by this workflow is unique and not readily served by existing visualization and analysis tools. Further, the complexity of the resulting data requires additional domain-specific computational and visual support. Due to the on-going development and refinement of these laboratory and computational techniques, ad hoc visual data exploration is vital for confirming the quality of the results and effectiveness of the workflow. At the same time, rigorous results require a more computational approach for automatic extraction of putative protein complexes from the data. MassVis provides a full range of facilities from the completely visual to the completely computational. We describe enough of the biology and experimental workflow to understand the intent and function of the software. However, it is beyond the scope of this conference for a complete exposition and the reader should refer to the supplied references for further details.

As a design study, the main contributions of this paper are:

- A novel scatter plot visualization designed to support both the analysis requirements and mimic a physical view familiar to protein scientists.
- Methods to perform and visually integrate correlation analysis of the proteomic data.

2 PROBLEM DESCRIPTION

Understanding the intent of MassVis requires a brief explanation of the laboratory workflow and how it relates to the downstream analysis. First, intact protein complexes are separated using one-dimensional gel electrophoresis. In gel electrophoresis, a protein sample is placed in a special gel matrix. An electric field is ap-

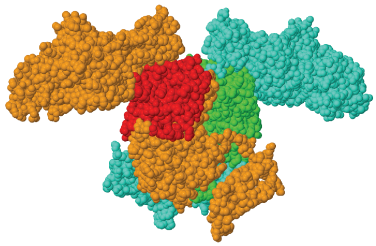

Figure 1. Space-filled model of the Phenylalanyl tRNA Synthetase complex (PDB ID: 1B70) consisting of four proteins: two copies of the alpha subunit (gold and cyan) and two copies of the beta subunit (red and green). Rendered using JMol [13]. Data obtained from PDB [5].

*email: robert_kincaid@agilent.com
+email: kurt.dejgaard@mcgill.ca

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

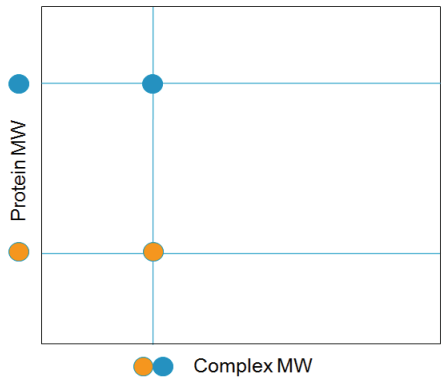

Figure 2. A schematic plot of a hypothetical complex consisting of two proteins represented by the orange and blue circles. The X coordinate represents the molecular weight range of the slice, which corresponds to the approximate molecular weight of the complex. The Y coordinates consist of the molecular weight of the individual proteins identified by mass spectrometry.

plied and proteins migrate through the gel at different rates depending on their mass-to-charge ratios. This technique is commonly used essentially to separate proteins by molecular weight. When the gel conditions are such that protein structure is maintained, it is referred to as "native" conditions and hence "native electrophoresis".

In our novel workflow, a special native gel formulation is used with conditions such that fully intact protein complexes can be subjected to electrophoresis and the intact complexes separated by molecular weight. We call this technique "WiSE Native" as an acronym of "native Wide Span Electrophoresis". Once this electrophoresis is complete, uniform slices of the gel are made. Each of these slices is then analyzed using tandem mass spectrometry to determine the proteins present in each slice.

Mass spectrometry (MS) involves sophisticated instruments used to accurately analyze the mass of the individual molecules in a sample. Tandem mass spectrometry (MS/MS) is a two-stage method of mass spectrometry that uses one stage of MS to select molecules of interest and a second stage MS to analyze collision fragments of the molecule selected by the first stage. In proteomics, this second stage results in a set of mass spectra that can often uniquely identify the molecules being analyzed [10].

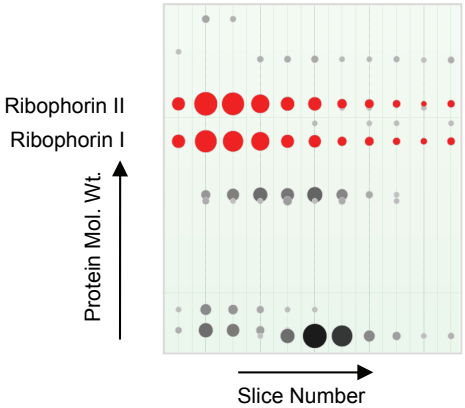

Figure 3. Slice profiles for Ribophorin (shown in red) which are members of the OST complex. The upper red profile is for RPN II while the lower profile is for the RPN I. The diameter of plotted points is proportional to the abundance of the protein found in that slice. Note the similarity in patterns between these two proteins indicating co-migration under electrophoresis.

In our workflow, each slice of the native one dimensional gel separation (along with its size separated proteins and protein complexes) is treated enzymatically in order to break the individual proteins into (tryptic) fragments. This mixture of peptides is then subjected to MS/MS. The spectra for each slice are collected and further analyzed with protein identification software (Spectrum Mill [1]). Acquired spectra are compared against a database of known spectra and statistically significant matches are determined and used for identification. Further, by summing the intensities of the various spectra attributed to the identified protein, a surrogate for the abundance of that protein can be achieved. Ultimately, a report is generated that lists all the proteins identified within each slice along with an estimate of their abundance in the form of a measurement called "total intensity". This Spectrum Mill report is a delimited text file and is read directly into MassVis. MassVis can be easily adapted to any MS/MS platform capable of generating similar data sets.

Determining the proteins contained within a given slice is important because protein complexes migrate through the 1-D gel together and are only separated by combination of MS/MS and the Spectrum Mill analysis. Our initial goal was simply to visualize this relationship between the 1-D separation of protein complexes and the constituent proteins of each slice. This is shown schematically for a single two-protein complex in Figure 2.

If the 1-D gel completely resolved complexes and individual proteins from each other, the data analysis would be trivial and one could simply determine the protein content of a slice and declare that content to represent a complex. However, complexes as well as individual proteins migrate in a more disorderly fashion and are smeared or spread out over a range of slices, generally with peak intensity at some specific slice. Thus, a given protein will have an intensity or abundance profile across the range of slices. We call this the slice profile. While these profiles frequently overlap, we can attempt to deconvolute them. We can assume that an intact complex should give rise to very similar slice profiles since the constituent proteins must co-migrate through the gel as a single intact unit. We can leverage this fact and look for proteins that have similar migration profiles and infer they are likely members of the same complex. Figure 3 shows an example of correlated slice profiles. This co-migration pattern is the central visual property we leverage in the software design. We strive to make the correlations as visible as possible, and to assist the interactive exploration and analysis of these correlations. For systematic extraction all relevant correlations, computational methods are provided for detecting and grouping similar slice profiles.

This workflow combined with the MassVis visualization offers several potential advantages over existing 2D gel techniques:

- We separate the second dimension by accurate mass measurements rather than a second gel separation. The accuracy of the mass spectrometer yields orders of magnitude better protein resolution than 2D gel methods.
- Analyzing uniform slices of the 1D gel is more comprehensive and unbiased than blotting or excising individual spots from a 2D gel and hard-to-find low abundance proteins are more likely to be detected.
- The final data set is fully digital (including protein abundance measurements) for the entire proteomic sample. No optical scanning is required as is often typical of 2D gel measurements, and the data is readily accessible to computational analysis.
- These advantages allow profiling and data analysis of entire cellular or organellar proteomes at a time (within the dynamic range of the mass spectrometer).

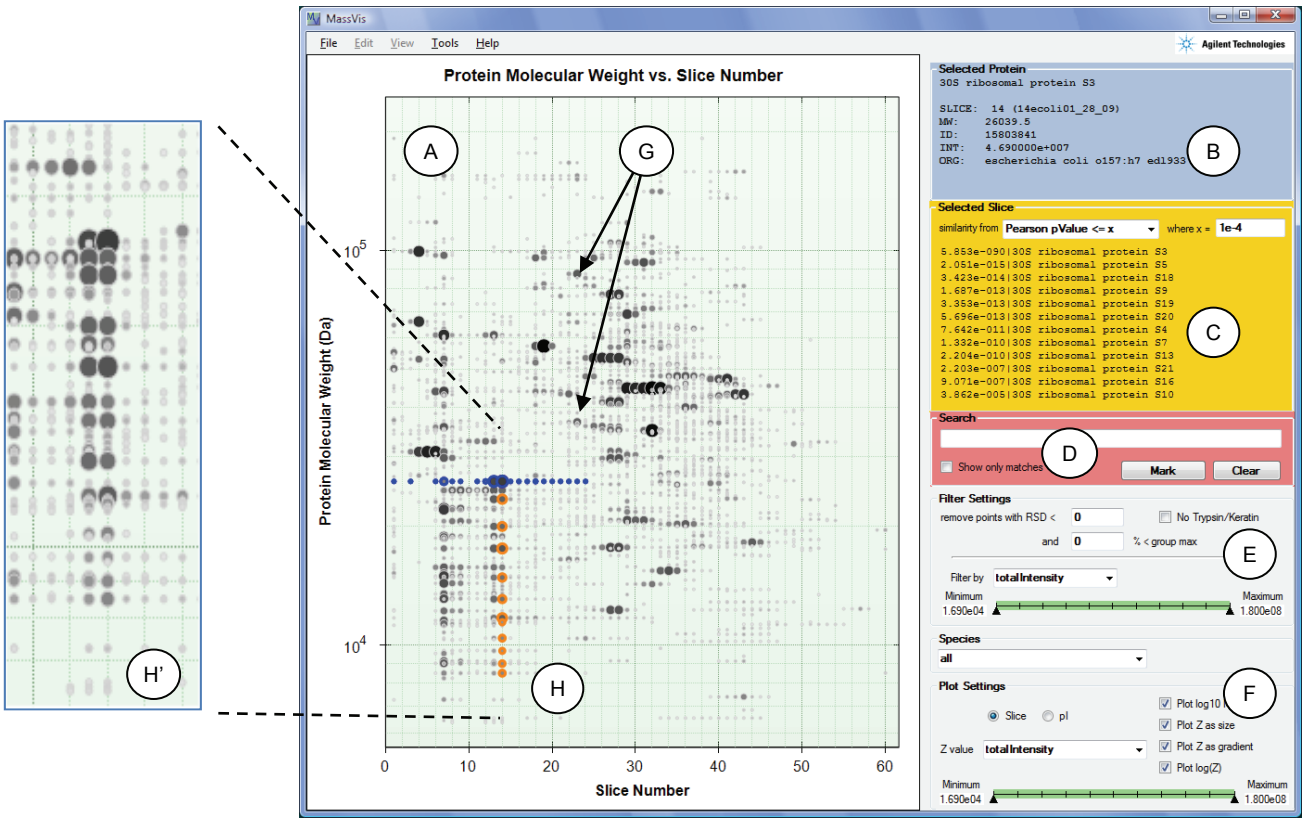

Figure 4. Overview of the MassVis user interface: (A) The main scatter plot view showing slice number on the X-axis and protein molecular weight on the Y-axis. (B) The detail pane showing the selected molecule. The blue background color matches the blue highlights shown in the scatter plot indicating where each instance of this protein occurs. (C) The content of the selected slice (12 in this case). Those proteins satisfying the similarity filtering set in pane C are shown with matching gold highlights in the scatter plot. (D) The search entry area. (E) Filter settings pane (F) Plot settings pane (G) α and β units of Phenylalanyl tRNA Synthetase (H) Indicates the region of clearly correlated patterns corresponding to the 30S ribosomal complex. (H') An enlarged version of H for clarity. Highlighting is omitted for comparison.

It is worth noting that a relatively unique property of this data is that the Y-axis is not strictly nominal, although the real valued mass measurements do generally correspond to the identity of specific proteins. One could imagine constructing a matrix view rather than a scatter plot where rows are identified by protein ID (nominal) vs. Mass (numerical). The same correlations would exist in either case. However, the spatial correspondence to a traditional 2D gel is lost in the matrix visualization. While one could sort a matrix view by molecular weight, the inability to present a vertical log scale would prevent the visual correspondence to a 2D gel. Presenting the data in a coordinate system corresponding to a 2D gel allows the proteomics scientist to locate proteins at familiar locations and landmarks.

3 Previous Work

At a fundamental level, aspects of the MassVis layout and interaction design are informed by Shneiderman's now ubiquitous mantra "overview, zoom & filter, details-on-demand" [23]. Craft and Cairns provide an updated discussion of this topic and recent related work [7].

SpotFire [3] and Tableau [2] offer flexible ad hoc scatter plot capabilities including dynamic filtering. With properly formatted data, these systems can create similar plots to that provided by MassVis. However, as we will see in the design section, they do not provide the same two-dimensional interactions or a computational means for extracting correlated slice profiles.

Li et al. [14] conducted a user study that suggests that scatter plots are more effective than parallel coordinate plots for visualizing correlations. However, while MassVis uses a scatter plot display, the correlations are not between pairs of X and Y values, but rather between slice profiles of different proteins.

Seo and Shneiderman's Hierarchical Cluster Explorer (HCE) [21] provides somewhat similar interactivity with respect to mining clustering results. HCE was developed primarily for mining large microarray data sets using heatmap representations of data matrices. HCE's ability to select similarity cut-offs for generating specific clusters is similar to methods used by MassVis, as is the ability to select individual clusters for inspection.

LifeLines2 [25] provides a system for analyzing correlated patterns across temporal data. It is somewhat reminiscent of MassVis in general layout and intent. However, LifeLines2 works with categorical data in the Y-axis and temporal data in the X-axis. Further, it leverages an alignment step that is unnecessary for MassVis since data is inherently aligned by the physical act of slicing the gel.

It is also worth noting previous efforts to apply information visualization principles to Mass Spectrometry data. SpectraMiner [27] provides visual data mining and analysis of mass spectra. Some mass spectrometry papers have appeared within the visualization community [8, 15]. Typically, these systems visualize low-

level mass spectra, sometimes as a 2D scatter plot or 3D surface. In contrast, ZoomQuant [12] uses tree maps to visualize higher-level quantitative differences between biological states organized by gene ontology categories.

View coordination is also an important aspect of the design of MassVis. A recent review is provided by Roberts [19]. Notable examples are Improvise [26] and Jigsaw [24].

Detecting correlations in complex data sets has been a common strategy for quite some time. For example, detecting correlations between gene expression profiles is widely practiced [6], as it is for protein expression. More recently, profiles of subcellular fractions have been analyzed using correlation analysis [4, 11]. Correlation measures are often recast as a distance or similarity measure and used for supervised or unsupervised clustering. These techniques are widely practiced in various data mining contexts.

4 VISUALIZATION DESIGN

Based on the domain-specific requirements described in Section 2, we can summarize our primary goals for our visualization design as:

G1: Visually represent the relationship between protein complexes and their constituent proteins in an easily understandable visualization.

G2: Allow the user to filter and adjust the display as appropriate to focus on particular proteins of interest and reduce information clutter and distracters.

G3: Provide details-on-demand to assist the user in making judgements about correlated patterns and their biological meaning.

G4: Provide computational assistance for detecting correlations in complex data sets.

We label these goals G1-G4 so that we can reference them in the sections that follow.

4.1 Main Display

The main visualization is essentially a 2D scatter plot of three-dimensional data and strives to satisfy the primary design goal G1. The independent X-axis is the number of the slice taken from the 1D gel. The dependent Y-axis is the molecular weight of individual proteins. The Y-axis label "Da" refers to "Daltons" and is the preferred mass unit for Mass Spectrometry and corresponds to the older terminology "atomic mass unit" or amu.

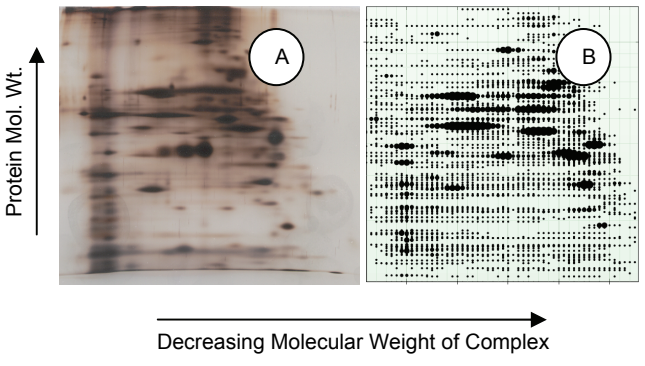

Figure 5. Comparison of (A) 2D gel separation of yeast whole cell lysate and (B) MassVis visualization of 1D Gel followed by systematic MS/MS. MassVis coloring and spot size were adjusted to maximize the similarity in appearance. Careful examination reveals many (but not all) visual features correspond between the two methods, thus verifying the ability of MassVis to mimic a 2D gel separation.

A given slice will actually correspond to a molecular weight range and increasing slice number corresponds to decreasing molecular weight of the entities in that slice. We could have labeled the slice axis with this range when it is known. However, this mapping of slice to molecular weight requires additional analysis of known proteins or standards in the sample and is somewhat time consuming. Therefore it is often convenient to simply ignore the precise molecular weight correspondence and simply work with slice numbers directly, while at the same time understanding the ordinal relationship between slice number and molecular weight. Further, since the molecular weight would typically not be a round uniform number, scaling the labelling during zooming and panning would be problematic.

A third attribute can be represented by size and/or color. Spectrum Mill provides a variety of numerical attributes compatible with a continuous mapping to size or color. Typically the user will use the attribute "totalIntensity" as this provides a useful estimation of protein abundance in each slice. All the figures in this paper use "totalIntensity" as the Z value. The correlated intensity patterns are what we use for determining which proteins are likely to be co-migrating in the gel and are thus members of the same protein complex.

Using both size and color as well as choosing to plot the Y and Z dimensions on a log scale reproduces the general layout and feel of a 2D gel. Figure 5 shows a comparison of a whole cell extract of yeast in both a 2D gel as well as the MassVis plot. While the correspondence between the 2D gel and the 1D gel + MS data is not always exact, the overall layout is familiar to proteomic scientists and the location of well-known proteins is at least qualitatively reproduced.

4.2 View Management

Various interactive settings allow the user to customize the 2D scatter plot to address design goal G2. These settings are organized in the lower right quadrant of the application window and grouped into two logical divisions: filter settings and plot settings.

Filter settings dynamically limit the data which is shown visually on the screen. The user can select from a combo box one of several possible attributes provided by the Spectrum Mill report. Generally, the user will choose "totalIntensity" to focus attention on a specific range of protein abundance. Other useful filters are spectral count or quality scores. These can be used to filter out data the user may deem unreliable. Additionally, a few specific convenience filters are provided to remove proteins that appear to vary insufficiently for a correlation analysis to be performed, or to remove molecules which are artifacts of the protein digestion process. Finally, depending on the sample and/or the database used for protein identification, results for multiple species may be returned. A combo box allows the user to select a single species and work with that data alone.

Plot settings allow the user to vary the style of the scatter plot. This includes selecting which attribute to plot as the Z value, whether to plot the Z value as a monochrome intensity map and/or size, whether to use log scales, etc.

Zoom and pan operations provide for magnification and navigation to regions of specific interest. Zooming combined with data filtering allows the user to manipulate regions plagued with data occlusion for better resolution of individual plot items. Further, the plot settings range slider controls the size and color mapping of plot items. The user can enlarge and darken smaller plot items for better visibility, or conversely reduce their size and intensity to make them less important to the display.

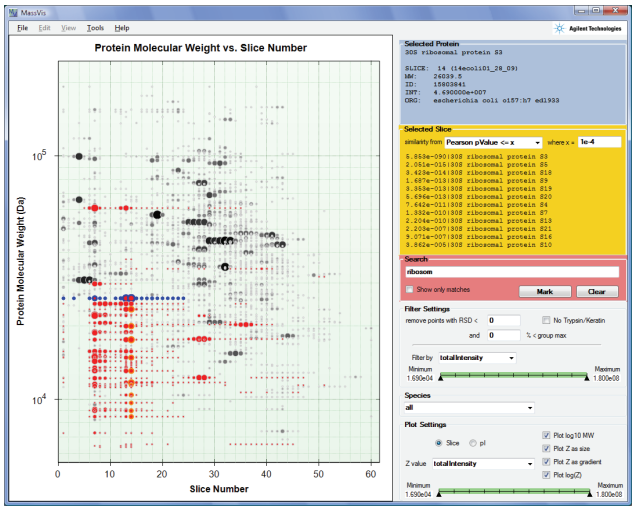

Figure 6. Searching for the stemmed term "ribosom" to find all the proteins described as "ribsome" or "ribosomal". Search hits are colored red, and we see the distinctive pattern corresponding to the 30S and 50S ribosomal subunits.

4.3 Coordinated Detail Panes

The main scatter plot view is augmented with two linked detail panes as a means to address design goal G3. When the user mouse-clicks on a scatter plot item (protein of interest) these two panes immediately update to display two sets of associated data.

4.3.1 Selected Protein Pane

The top right blue pane shows the specific protein the user selected. Various properties of the protein are shown such as the descriptive name for the protein, the slice in which this instance of the protein was found, the molecular weight, a unique database identifier, organism, and the total intensity value, etc. This allows the user to inspect a variety of relevant details for any plot item on the display. While this same information is available as a tool tip for rapid scanning of the plot, it is still useful to provide a persistent display, particularly when used in conjunction with the second detail pane related to the selected slice.

4.3.2 Selected Slice Pane

Below the blue protein pane is a gold pane titled "Selected Slice". This pane shows a list of all proteins (and similarity scores) in the currently selected slice satisfying the similarity criteria. These criteria are set via two combo boxes at the top of the pane. The first combo box sets the similarity algorithm being used (usually the modified Pearson p-value discussed in 4.6). The second combo box specifies a criterion appropriate for the similarity algorithm. For p-values we set an upper limit for what to consider strongly correlated. The user can interactively click on any spot in the scatter plot and immediately see what other proteins appear to be highly correlated and therefore putative members of the same complex.

4.3.3 Scatter Plot Encoding

The scatter plot graphically indicates content corresponding to the detail panes by adding colored borders to the rendered plot items. All instances of a selected protein have blue borders which correspond to the blue pane. For the selected slice, all proteins which satisfy the similarity criteria are rendered with gold borders. These matching colors help reinforce the correspondence to the appropriate detail panes.

As with any scatter plot, data occlusion invariably occurs. Since precise alignment of plot items is critical to visualizing correlations, we cannot rely on techniques such as dithering to reduce such occlusion and a single z-ordering strategy will always prove problematic. We rely instead on filtering and zooming to reduce visual clutter and occlusion. Further, proteins corresponding to the selected protein and slice, which are colored with blue and gold borders, are rendered last so that they are always visible.

4.4 Internal Search

When many individual measurements are plotted, it is useful to allow a direct search for relevant molecules within the MassVis data. This is accomplished with the red search pane shown in Figure 6. The user enters a search term and then presses the *mark* button. This causes all proteins whose description contains the search string to be highlighted in red. Optionally, the user can request to *only* display the matched entries to further reduce the visual clutter. Using this filtered display option the user can view the matched entries alone or in the context of the full data set as appropriate to the task at hand, enabling the user to quickly find relevant molecules of interest. Should they require finding a specific molecule, they can enter a uniquely identifying search term or a uniquely identifying GenBank accession number.

4.5 External Search

As a convenience we provide mechanisms to directly invoke web-based queries based on the currently selected protein of interest. Using a popup menu from the blue pane, the user can invoke various specialized web searches. For instance, the user can search NCBI's Entrez Protein [20] database to fetch the exact record for the molecule in question, allowing access to amino-acid sequences and additional hyperlinks to other information sources. Another useful search is provided against the Reactome database [16]. In this case the user can quickly find any existing information about complexes the molecule may be a member of and compare that to the putative clusters identified by MassVis.

4.6 Computational Methods

Visually teasing apart subtle patterns in a complex data set is a very challenging analysis task which lead to our design goal G4. We found it is often more efficient to provide computational means to group similar slice patterns together and use the visual display to let the user judge the validity of these groupings.

A common similarity measure uses the Pearson correlation coefficient, r. We have found it convenient to base our distance measure on probability rather than strictly on a correlation coefficient. It is possible to recast the Pearson coefficient as a probability via the incomplete beta function [18]. The advantage over simply relying on r is that the probability accounts for the number of points being considered (i.e. sample size), and more accurately represents the confidence that the correlation is significant. Since we use the null hypothesis that the two proteins are uncorrelated, we have a useful distance measure that states p-values near 0 are similar (short distance) and uncorrelated proteins have p-values near 1 and are dissimilar (longer distance).

While the Pearson correlation is a good measure that the patterns are similar, it does not relate any information about the relative magnitude of the two signals being compared. For our workflow, we have additional information that is relevant to the correlation analysis. We know that for a given complex, proteins must exist as intact units so that the relative abundance between members of a complex should ideally be integer multiples of their molecular weights. While experimental variance makes determining precise ratios difficult, we can still eliminate extreme ratios that make no sense for any reasonable protein complex. For example, if the abundance of protein A were 100 times the abun-

dance of protein B, even if their slice profiles looks similar, it is unlikely that these two molecules form a complex, as no biological complex would have composition $A_{100}B$.

We incorporate this knowledge with a user adjustable ratio cut-off. Any two proteins whose ratio of total normalized intensity is greater than the cut-off has its p-value artificially set to the value one. An artificial value of one has the effect of treating the pair of proteins as uncorrelated. While this heuristic is useful in practice, it alters the true underlying statistics and all subsequent p-values, including those near zero, should be considered as qualitative measures and not as strict mathematical probabilities. Even so, for proteomic scientists, these pseudo p-values and ratio cut-offs are intuitive and easy to understand in the context of the data analysis. For computing intensity ratios, the intensity values for each protein are summed for whatever slice range the user is considering.

This similarity measure can be used to interactively filter the results shown in the gold pane which shows slice content as described in section 4.3. Based on the user specified threshold, only those molecules with a similarity p-value less or equal to the cut-off are shown in the gold pane and are highlighted with gold borders (see Figure 4). The complexity in this case is simply $O(n)$ in time where n is the number of proteins in the selected slice. No additional memory overhead is required since the similarity result can be discarded after comparison to the threshold is made. Since n is usually a relatively small number and of linear complexity, these calculations can be performed interactively with virtually instantaneous updates of the display as proteins are selected.

The same similarity measure can be used as a distance measure for unsupervised clustering. In this case, a similar p-value cut-off is used to define both when to stop clustering and how to define clusters. Standard agglomerative clustering is performed until the next cluster would join a distance corresponding to a value greater than the p-value cut-off. When the process is terminated in this way, no cluster will contain any p-value distance greater than the cut-off, and the top level nodes in the resultant cluster hierarchy are used to define putative clusters. Our implementation uses standard methods with complexity $O(n^2)$ in space and worst case $O(n^2 \log n)$ in time, where n is the number of distinct proteins identified in the sample. The data sets we have analyzed so far typically cluster in less than one minute on standard consumer-grade personal computers.

5 Preliminary Evaluation and Example Results

The software described in this paper has been developed as an ongoing collaboration between Agilent and McGill University to develop a novel workflow for determining protein-protein interactions and complexes. As such, we are currently the primary users of the software and obtaining impartial evaluators requires training users in not only the use of the software but also details of the workflow itself and the specialized interpretation of the results. Thus, performing a formal user study is somewhat problematic at this time.

However, MassVis has been in active laboratory use for over a year supporting the development of the laboratory protocol described briefly in this paper. Preliminary scientific results utilizing MassVis were previously presented in conjunction with a key mass spectrometry conference [9].

In addition, MassVis has been used successfully for other proteomics applications. For example, we have examined data from subcellular fractionation followed by MS/MS analysis. The workflow is very similar to that described here. However, instead of gel electrophoresis, the scientist performs ultracentrifugation or some other method of separating cellular structures. Instead of taking gel slices, one takes fractions of each separation (the equivalent of taking a gel slice). The correlations we look for are not based on membership in a protein complex, but rather the observation that proteins from the same cellular component will have a similar distribution across the fractions. The numerical analysis is similar to profile-based methods developed by Foster et al. [11] and Andersen et al. [4].

We can further relate our own experience to date, with a few concrete examples in the following sections.

5.1 Interactive Discovery Task with E. Coli Whole Cell Extract

The initial goal of the software was to support interactive data browsing and discovery. In this mode, the user searches for visibly correlated slice profiles. For example, Figure 3 shows a typical pattern corresponding to two components of the OST48 complex (Ribophorin I and Ribophorin II). The user can freely zoom and pan the scatter plot to search for such patterns and then by clicking on specific features, inspect the details of the selected and similar proteins in the detail panes. Further, the slice details also include an automatic correlation calculation using a number of different scoring mechanisms. When a scoring method is selected from the combo box, clicking on any protein will show those proteins within the selected slice that satisfy the correlation criteria. The result is a facile interactive means to manually inspect interesting features and see if they correspond to a plausible protein complex.

Figures 4 and 6 show results from analyzing whole cell extract from a bacterial culture of E. coli. 61 slices were cut from the 1D gel electrophoresis and subjected to MS/MS analysis. Spectrum Mill identified 666 distinct proteins.

Figure 4 shows the entire data set displayed in MassVis and the result of clicking the mouse on the plot item corresponding to the protein named "30S ribosomal protein S3". All instances of this molecule are shown with blue outlines and the details about the selected molecule are shown in the blue pane at the upper right. Using the Pearson p-value score and a cut-off of $p \leq 10^{-4}$, the molecules with profiles most similar to "30S ribosomal protein S3" are listed in the gold pane, sorted by decreasing similarity to the selected protein. We notice that those proteins with similar profiles are all members of the same "30S ribosomal protein" complex. The spots in the scatter plot corresponding to these molecules are shown with gold outlines for the selected slice. We can see visually that the slice profiles for these related proteins do appear qualitatively similar.

For obvious, abundant proteins or for well-known proteins with familiar locations and patterns, this interactive task is often quite useful for quickly inspecting new data. Note, however, that the pair of Phenylalanyl tRNA Synthetase proteins are very difficult to find visually as correlated partners (see Figure 4G). Figure 7 shows a zoomed in view, which is still difficult to deconvolute visually. What is required is a more automated and systematic approach, which we discuss in the next example.

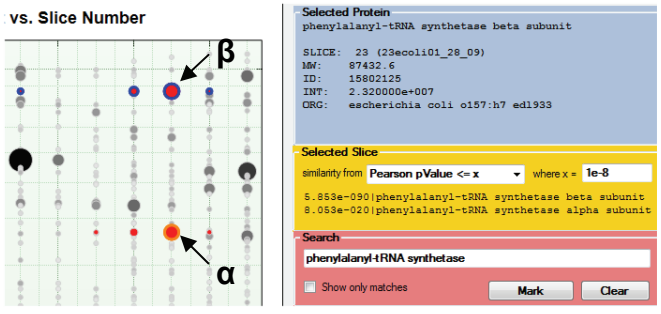

Figure 7. A zoomed in view of Phenylalanyl tRNA Synthetase from an E. coli sample. The two subunits are highlighted in red using the search function.

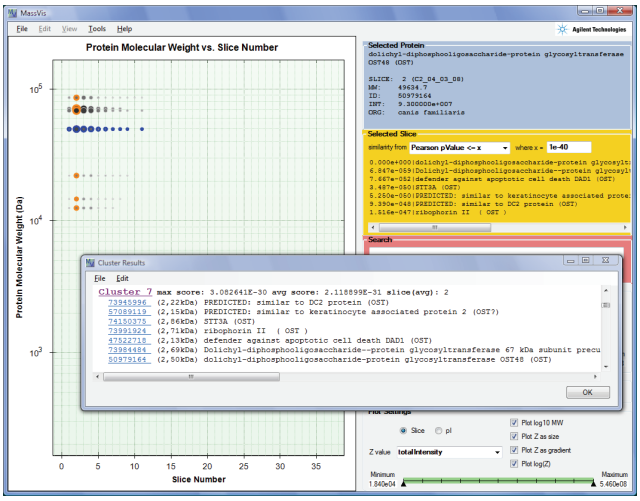

Figure 8. Detailed inspection of the OST complex. The gold pane shows the results of interactive exploration while the overlaid window shows the results of unsupervised clustering of the slice profiles. Cluster 7 corresponds to the OST complex. Clicking on the cluster hyperlink causes the display to show *only* the members of this cluster.

5.2 Unsupervised Discovery Task with an ER-Enriched Sample

For unsupervised discovery MassVis currently supports agglomerative clustering as described in section 4.6. When invoked, this function returns an interactive HTML report. The hyperlinks in this report can be used to select the cluster in the main MassVis scatter plot showing *only* the members of the cluster. This allows fast visual inspection of the cluster in isolation, without visual clutter and verification the cluster exhibits significant correlations.

We also examined a sample obtained from a dog pancreas enriched for proteins bound to the endoplasmic reticulum (ER) and is thereby enriched for proteins and complexes associated with the ribosome. The 1D gel separation was sliced into 36 individual samples and Spectrum Mill analysis of the MS/MS experiments reported a total of 506 different proteins. MassVis found the usual ribosomal complexes as shown in Figure 6. To dig deeper we employed the clustering method described in section 4.6.

Figure 8 demonstrates how the cluster report can be used to interactively inspect each cluster in isolation, in order to visually verify that the correlations look reasonable. In this case we examine the OST cluster obtained computationally from unsupervised clustering and obtain a striking result finding almost all of the proteins believed to constitute the OST complex including some

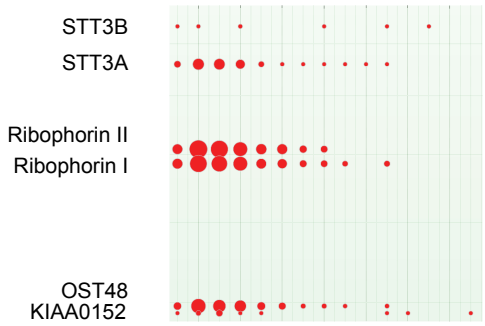

Figure 9. Members of the OST complex showing KIAA0152 and STT3A for comparison.

Protein	Shibatani et al.	Reactome	MassVis	Mol. Weight.
STT3B	✓	✓	×	93675
STT3A	×	✓	✓	85961
Ribophorin II	✓	✓	✓	70879
Ribophorin I	✓	✓	✓	68577
OST48	✓	✓	✓	49635
KIAA0152	✓	×	×	48842
DC2	✓	×	✓	22034
KCP2	✓	×	✓	14651
DAD1	✓	✓	✓	12511

Table 1. Comparison of published OST complex members from Shibatani et al. [22], Reactome [16] and MassVis.

proteins only recently reported by Shibatani et al. Table 1 compares Shibatani's results with our own as well as data from the Reactome [16] website. While there is not complete agreement between the various sources, we note that our MassVis generated result is comparable and consistent with the current literature.

5.3 Workflow Refinement Task

We notice in the previous example that we did not find every known member of the OST complex However, we can use MassVis to carefully examine the missing members to try and determine why these proteins were missed in our experiments.

Figure 9 shows a region of the OST complex that includes the missing members STT3B and KIAA0152. We can immediately see why our correlation calculations missed these molecules. STT3B does not appear to have the same distribution pattern or level of abundance, and is even missing values in a few slices. KIAA0152 appears to have a somewhat similar slice profile to the other proteins, but again has a vastly reduced overall relative abundance and is also missing data in several slices.

There are a number of possible explanations. STT3B and KIAA0152 might be weakly bound and so relatively fewer complexes survived the gel separation with these members intact. Alternatively, the parameter settings for Spectrum Mill might require modification to more accurately interpret the spectra of these molecules. This suggests further refinement in some upstream process is required (either in the laboratory workflow or the Spectrum Mill processing). We also note that the disagreement between Shibatani and Reactome for these molecules suggests they may be difficult to detect.

6 Discussion

Our data analysis requirements led us to create two detail panes showing different aspects of the selected protein of interest. However, the "selected slice" pane does not operate as a typical detail pane. Rather, it shows a computed result based on the selected item in the scatter plot. North et al. [17] describes a more general architecture for coordinated visualizations of relational multi-table databases, which includes a discussion on computed results as a variation on the relational theme. This design pattern is surprisingly rare in practice and not directly supported in visualization platforms such as Spotfire or Tableau which generally visualizes a single table or query result.

Another feature of MassVis is that a single item selection in the scatter plot invokes two orthogonal color encodings. One encoding highlights the presence of the selected protein across all slices and the other encoding highlights similar proteins within the current slice. Coordinated views and multiple color encodings are frequent visual design elements in visual analytics. However, driving multiple encodings or selections within the same view from a single selection is less common and potentially useful in other contexts. One can generalize such behavior as a multi-

dimensional selection scheme where a single item selection invokes the selection of multiple but independent sets of related data within the same view. This behaviour is different from typical linked-selection schemes described by Roberts [19] and North [17] where a single selection is projected through coordinated views.

7 Conclusion and Future Work

MassVis has proven to be a useful and vital tool for not only visualizing the results of our experiments, but also for debugging and refining both the computational algorithms as well as the laboratory workflow. The characteristics of the underlying data are unique since a seemingly standard scatter plot representation has an important secondary interpretation as a spatial analogue to a 2D gel separation. Because of this, the scatter plot provides a more natural context for a proteomic scientist compared to matrix representations using nominal row identifiers such as protein ID's.

While the software has been invaluable for the ongoing development of this new method of detecting protein-protein interactions, many areas of improvement are still possible. The current cluster exploration would benefit from functionality similar to the Seo and Shneiderman's Hierarchical Cluster Explorer [21]. Displaying the full cluster hierarchy as a dendrogram would be more informative than the current implementation of static clusters based on user set similarity cut-offs. While dealing with the added complexity of a dendrogram would place greater analytical burden on the user, methods provided by HCE to dynamically select similarity cut-offs would allow the user to make more informed cut-off selections.

Further improvements in the correlation measures are also possible. In particular, methods for matching local correlations while ignoring non-local features would find secondary protein-protein interactions where a single protein may participate in more than one complex. Finding the optimum parameter settings for clustering slice profiles still requires a certain amount of trial-and-error effort by the user. Unsupervised methods for determining optimum clustering and/or similarity parameters are still needed.

References

[1] Spectrum Mill, Agilent Technologies, Inc., 2008.

[2] Tableau, Tableau Software, http://www.tableausoftware.com, 2009.

[3] "TIBCO Spotfire," TIBCO Software Inc., http://spotfire.tibco.com, 2009.

[4] J. S. Andersen, C. J. Wilkinson, T. Mayor, et al., "Proteomic characterization of the human centrosome by protein correlation profiling," *Nature*, vol. 426, pp. 570-574, 2003.

[5] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nature Structural Biology*, vol. 10, pp. 980-980, 2003.

[6] P. Boutros and A. Okey, "Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data," *Briefings in Bioinformatics*, vol. 6, pp. 331-343, 2005.

[7] B. Craft and P. Cairns, "Beyond guidelines: what can we learn from the visual information seeking mantra?" In IV'05: Proc. of the Conf. on Information Visualization, pp. 110-118, 2005.

[8] J. de Corral and H. Pfister, "Hardware-accelerated 3D visualization of mass spectrometry data," In Proc. IEEE Vis., pp. 439-446, 2005.

[9] K. Dejgaard, "Interactomics: Improved Native Electrophoresis Protocols for Studies of Megadalton-Sized Protein Complexes," *ASMS 2008 Agilent Technologies Tech. Forum on Mass Spec.*, 2008.

[10] B. Domon and R. Aebersold, "Review - Mass spectrometry and protein analysis," *Science*, vol. 312, pp. 212-217, 2006.

[11] L. J. Foster, C. L. de Hoog, Y. L. Zhang, et al., "A mammalian organelle map by protein correlation profiling," *Cell*, vol. 125, pp. 187-199, 2006.

[12] B. Halligan, S. Mirza, M. Pellitteri-Hahn, et al., "Visualizing Quantitative Proteomics Datasets using Treemaps," In IV'07: Proc. of the Conf. on Information Vis., pp. 527-534, 2007.

[13] A. Herraez, "Biomolecules in the computer - Jmol to the rescue," *Biochem. and Mol. Biology Education*, vol. 34, pp. 255-261, 2006.

[14] J. Li, J. Martens, and J. van Wijk, "Judging correlation from scatterplots and parallel coordinate plots," *Information Visualization*, 2008.

[15] L. Linsen, J. Locherbach, M. Berth, et al., "Differential protein expression analysis via liquid-chromatography/mass-spectrometry data visualization," In Proc. IEEE Vis., pp. 447-454, 2005.

[16] L. Matthews, G. Gopinath, M. Gillespie, et al., "Reactome knowledgebase of human biological pathways and processes," *Nucleic Acids Research*, vol. 37, pp. D619-D622, 2009.

[17] C. North, N. Conklin, K. Indukuri, et al., "Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases," *Information Visualization*, vol. 1, pp. 211-228, 2002.

[18] W. H. Press, *Numerical recipes in C++*, 2nd ed. Cambridge ; New York: Cambridge University Press, 2002.

[19] J. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," In CMV '07: Proc. Conf. on Coordinated and Multiple Views in Exploratory Visualization, pp. 61–71, 2007.

[20] E. W. Sayers, T. Barrett, D. A. Benson, et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 37, pp. D5-D15, 2009.

[21] J. Seo and B. Shneiderman, "Interactively exploring hierarchical clustering results," *Computer*, vol. 35, pp. 80-86, 2002.

[22] T. Shibatani, L. David, A. McCormack, et al., "Proteomic Analysis of Mammalian Oligosaccharyltransferase Reveals Multiple Subcomplexes that Contain Sec61, TRAP, and Two Potential New Subunits," *Biochemistry*, vol. 44, pp. 5982-5992, 2005.

[23] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," In Proc. IEEE Symp. on Visual Languages, pp. 336-343, 1996.

[24] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, pp. 118-132, 2008.

[25] T. D. Wang, C. Plaisant, A. J. Quinn, et al., "Aligning temporal data by sentinel events: discovering patterns in electronic health records," in *Proc. SIGCHI conference on Human factors in computing systems*, 2008, pp. 457-466.

[26] C. Weaver, D. Fyfe, A. Robinson, et al., "Visual exploration and analysis of historic hotel visits," *Information Visualization*, vol. 6, pp. 89-103, 2007.

[27] A. Zelenyuk, D. Imre, Y. Cai, et al., "SpectraMiner, an interactive data mining and visualization software for single particle mass spectroscopy: A laboratory test case," *Int. J. Mass Spectrom.*, vol. 258, pp. 58-73, 2006.